

Polarization and Spatial Coupling: Two Techniques to Boost Performance

THÈSE N° 5706 (2013)

PRÉSENTÉE LE 3 SEPTEMBRE 2013

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

Seyed Hamed HASSANI

acceptée sur proposition du jury:

Prof. M. C. Gastpar, 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
2013

Polarization and Spatial Coupling: Two Techniques to Boost Performance

Polarization and Spatial Coupling: Two Techniques to Boost Performance

S. Hamed Hassani

EPFL - Ecole Polytechnique Fédérale de Lausanne

Thesis No. 5706 (March 2013)

Thesis presented to the faculty of computer and communication sciences for
obtaining the degree of Docteur ès Sciences

Accepted by the jury:

Nicolas Macris and **Rudiger Urbanke**
Thesis directors

Erdal Arikan
Expert

Andrea Montanari
Expert

Emre Telatar
Expert

Michael Gastpar
President of the jury

Ecole Polytechnique Fédérale de Lausanne, 2013

Abstract

During the last two decades we have witnessed considerable activity in building bridges between the fields of information theory/communications, computer science, and statistical physics. This is due to the realization that many fundamental concepts and notions in these fields are in fact related and that each field can benefit from the insight and techniques developed in the others.

For instance, the notion of channel capacity in information theory, threshold phenomena in computer science, and phase transitions in statistical physics are all expressions of the same concept. Therefore, it would be beneficial to develop a common framework that unifies these notions and that could help to leverage knowledge in one field to make progress in the others. A particularly striking example is the celebrated belief propagation algorithm. It was independently invented in each of these fields but for very different purposes. The realization of the commonality has benefited each of the areas.

We investigate *polarization* and *spatial coupling*: two techniques that were originally invented in the context of channel coding (communications) thus resulting for the first time in efficient capacity-achieving codes for a wide range of channels. As we will discuss, both techniques play a fundamental role also in computer science and statistical physics and so these two techniques can be seen as further fundamental building blocks that unite all three areas. We demonstrate applications of these techniques, as well as the fundamental phenomena they provide.

In more detail, this thesis consists of two parts. In the first part, we consider the technique of polarization and its resultant class of channel codes, called *polar codes*. Our main focus is the analysis and improvement of the behavior of polarization towards the most significant aspects of modern channel-coding theory: scaling laws, universality, and complexity (quantization). For each of these aspects, we derive fundamental laws that govern the behavior of polarization and polar codes. Even though we concentrate on applications in communications, the analysis that we provide is general and can be carried over to applications of polarization in computer science and statistical physics.

As we will show, our investigations confirm some of the inherent strengths of polar codes such as their robustness with respect to quantization. But they also make clear in which aspects further improvement of polar codes is needed.

For example, we will explain that the scaling behavior of polar codes is quite slow compared to the optimal one. Hence, further research is required in order to enhance the scaling behavior of polar codes towards optimality.

In the second part of this thesis, we investigate spatial coupling. By now, there exists already a considerable literature on spatial coupling in the realm of information theory and communications. We therefore investigate mainly the impact of spatial coupling on the fields of statistical physics and computer science. We consider two well-known models. The first is the *Curie-Weiss* model that provides us with the simplest model for understanding the mechanism of spatial coupling in the perspective of statistical physics. Many fundamental features of spatial coupling can be simply explained here. In particular, we will show how the well-known Maxwell construction in statistical physics manifests itself through spatial coupling.

We then focus on a much richer class of graphical models called *constraint satisfaction problems (CSP)* (e.g., K -SAT and Q -COL). These models are central to computer science. We follow a general framework: First, we introduce interpolation procedures for proving that the coupled and standard (un-coupled) models are fundamentally related, in that their static properties (such as their SAT/UNSAT threshold) are the same. We then use tools from spin glass theory (cavity method) to demonstrate the so-called phenomenon of threshold saturation in these coupled models. Finally, we present the algorithmic implications and argue that all these features provide a new avenue for obtaining better, provable, algorithmic lower bounds on static thresholds of the individual standard CSP models. We consider simple decimation algorithms (e.g., the unit clause propagation algorithm) for the coupled CSP models and provide a machinery to analyze these algorithms. These analyses enable us to observe that the algorithmic thresholds on the coupled model are significantly improved over the standard model. For some models (e.g., 3-SAT, 3-COL), these coupled algorithmic thresholds surpass the best lower bounds on the SAT/UNSAT threshold in the literature and provide us with a new lower bound.

We conclude by pointing out that although we only considered some specific graphical models, our results are of general nature hence applicable to a broad set of models. In particular, a main contribution of this thesis is to firmly establish both polarization, as well as spatial coupling, in the common toolbox of information theory/communication, statistical physics, and computer science.

Keywords: Channel polarization, polar codes, spatial coupling, threshold saturation, capacity achieving codes, mean field models, Curie-Weiss model, constraint satisfaction problems.

Résumé

Durant ces deux dernières décennies une activité considérable a développé des ponts entre des disciplines telles que la théorie de l'information, l'informatique et la physique statistique. Ceci est dû au fait que plusieurs concepts fondamentaux et notions reliés à ces disciplines sont en fait connectés et que chacune d'entre elles peut bénéficier d'intuitions et de techniques développées pour les autres.

Par exemple, la notion de capacité de canal en théorie de l'information, les effets de seuil en informatique et les transitions de phase en physique statistique sont des manifestations différentes d'un même concept. Par conséquent, il serait naturel de développer un cadre commun qui unifie ces notions et permettrait d'exploiter les connaissances d'un domaine pour progresser dans un autre. Un exemple particulièrement frappant est le célèbre algorithme de propagations des croyances. Il fut indépendamment inventé dans chacun de ces domaines pour des raisons à priori assez différentes, et il a été bénéfique de réaliser leur nature commune.

Nous étudions la polarisation et le couplage spatial: deux techniques qui furent inventées d'abord dans le contexte du codage pour les canaux bruités en théorie des communications, et ont permis pour la première fois d'atteindre la capacité grâce à des schémas efficaces en complexité et ceci pour une large classe de canaux. Les deux techniques jouent un rôle important aussi en informatique et physique statistique et du coup ces techniques peuvent être vues comme des éléments unificateurs dans chacun de ces trois domaines. Nous illustrons des applications de ces techniques, aussi bien que la nouvelle compréhension qu'elles apportent.

Cette thèse comporte deux parties. Dans la première, nous considérons la technique de polarisation et sa classe de codes correcteurs correspondant: les codes polaires. L'objectif principal est la compréhension et l'étude de plusieurs de leurs caractéristiques telles que leur performance à longueur finie, leur universalité et leur complexité (quantification). Pour chacun de ces aspects nous dérivons les relations fondamentales qui gouvernent la polarisation. Et même si nous nous concentrons sur les applications en communication, l'analyse que nous prodiguons est générale et peut être étendue aux applications de la polarisation en informatique ou en physique statistique.

Comme vous allons le voir, nos recherches confirment certains avantages inhérents aux codes polaires, comme leur robusticité face à la quantification. Mais elles mettent aussi en lumière quels aspects des codes polaires restent encore à améliorer. Par exemple, nous expliquerons que le comportement asymptotique des codes polaires est plutôt lent comparé au comportement optimal et que par conséquent de plus amples recherches sont nécessaires pour garantir un comportement asymptotique des codes polaires.

Dans la seconde partie de cette thèse, nous étudions le couplage spatial. Il existe dès à présent une littérature considérable sur le sujet en théorie de l'information et en communication et c'est pourquoi nous nous intéressons au couplage spatial dans la cadre de la physique statistique et de l'informatique. Nous considérons deux modèles célèbres. Le premier est le modèle de Curie-Weiss qui a l'avantage de permettre une explication des plus simples des propriétés du couplage spatial dans la cadre de la physique statistique. En particulier nous montrerons comment la fameuse construction de Maxwell se manifeste à travers le couplage spatial.

Nous nous focaliserons par la suite sur une classe de modèles plus riches appelé "problème de satisfaction de contraintes (CSP)" (p.ex. K-SAT et Q-COL), ces modèles jouant un rôle central en informatique. Notre étude suit une procédure systématique: premièrement nous introduisons le procédé d'interpolation afin de prouver que l'ensemble standard (non-couplé) et l'ensemble couplé possèdent les mêmes propriétés statiques (comme la seuil de transition de phase SAT/UNSAT).

Dans un deuxième temps nous utilisons des outils de la théorie des verres de spin (méthode de la cavité) pour démontrer dans les modèles couplés le dit phénomène de saturation de seuil.

Finalement, nous présentons les implications du point de vue algorithmique et discutons en quoi ces caractéristiques permettent la création de nouvelles démonstrations pour de meilleures bornes sur les seuils de transition des modèles CSP non-couplés. Nous considérons des algorithmes de décimation simple (comme l'algorithme de propagation de clause unité) pour des modèles CSP couplés et fournissons une méthode pour les analyser. Ainsi nous observons que le seuil de transitions de ces algorithmes est significativement amélioré sur les modèles couplés. Pour certains de ces modèles (p.ex. 3-SAT, 3-QOL), ces seuils de transitions sur les ensembles couplés dépassent les meilleures bornes inférieures connues sur les seuils de transition SAT/UNSAT et ainsi fournissent de nouvelles bornes.

Nous concluons en remarquant que même si nous avons centré notre attention sur une classe spécifique de modèles, nos résultats sont de nature générale et peuvent par conséquent s'appliquer à une classe plus large de modèles graphiques. En particulier, une contribution de cette thèse est de montrer que la polarisation, comme le couplage spatial sont désormais des outils incontournables en théorie de la communication, en physique statistique, aussi bien qu'en informatique.

Mots-clés: Polarisation de canal, code polaire, couplage spatial, saturation de seuil, modèle de Curie-Weiss, problèmes de satisfaction de contraintes.

Acknowledgements

I am truly honored and fortunate to have been advised by Dr. Nicolas Macris and Prof. Rudiger Urbanke. Over my years here, they taught me many lessons, about both life and research. I have benefited immensely from them, as patient teachers, brilliant researchers, and true friends. They have been greatly influential and supportive in all the aspects of my life. A few paragraphs would show nothing about how indebted I feel towards them. I sincerely hope that my future performance would partly acknowledge their efforts.

My first meeting with Rudiger was on February 22nd of 2008 when I was a master's student. The meeting lasted for an hour, Rudiger was the only speaker, and I could make sense out of nothing except his last sentence: "My door is always open, just pass by". This first advice from Rudiger, is the one that I took seriously the most often. As time went on, Rudiger's tasks increased exponentially, but he always fitted me through his schedule. What I will miss the most about Rudiger is our regular early morning discussions, when no one was yet in the office. Today is April 30th of 2013, the deadline for the final version of this thesis. I have been awake for most of the night and there is still a lot of small things to take care of. Of course, I met Rudiger today around 8 a.m. For me, Rudiger is an incredibly generous, wise and supportive friend, a genius, a hard worker, an extremely persistent and optimistic person. He has made every effort to teach me these properties, not by words but by action.

When I started my PhD research, my background was mainly in the areas of communications and mathematics. However, a large portion of this thesis is about statistical physics. I want to express my deepest gratitude to Nicolas for introducing and teaching me the subject of statistical physics. I learned from him how deep insight into physics is beneficial to developing ideas in other fields. Working with Nicolas has been a unique experience. We worked on many interesting topics. In particular, our collaboration during the last summer, which resulted in the last chapter of this thesis, was the most joyful and instructive period of my PhD years.

I would like to thank my thesis committee members Prof. Erdal Arıkan, Prof. Michael Gastpar, Prof. Andrea Montanari, and Prof. Emre Telatar for their very helpful comments and suggestions. I was honored to have thesis committee members who were either the inventors or the main contributors to

the topics that I worked on. I am deeply grateful to Emre for his generous support, teaching, and advice. I truly enjoyed our meetings and learned very much.

I have been very fortunate to collaborate with Prof. Dimitris Acliopas, Prof Kasra Alishahi, and Prof. Toshiyuki Tanaka. Dimitris has a great personality, sense of humor, and a mind full of innovative ideas. Working with Dimitris was another unique experience, I had to manage five tasks at the same time: learning, thinking, coding, joking, and not sleeping. He is the master of *K-SAT* and taught me many interesting things about the problem, as well as life. I hope that our collaboration will continue and we will finally step over the condensation threshold. Kasra is a truly exceptional person. I thank him for his friendship, hospitality and instructive discussions.

It has been a great pleasure working at EPFL. I would like to express gratitude to all the past and present members of the Information Processing Group (IPG). I am grateful to Prof. Bixio Rimoldi for his great tips on the writing and presentation skills. Special thanks to my colleagues Andrei, Alla, Christine, Emmanuel, Eren, Marc D., Marc V., Mohammad, Satish, Shrinivas for the many interesting scientific and non-scientific discussions and many fun conference trips. Amin Karbasi (the master of everything) was my first officemate. His valuable advice and great friendship has always been helpful to me. I am very thankful to Marjan and Amin for many nice memories. I thank Nicolaos for the wonderful office time and his guidelines about machine learning. I thank Dr. Olivier Leveque for many things including the organization of the IPG tournaments (foosball, petanque). I have greatly enjoyed our running events with Nicolas Rouzi, Prof. Suhas Diggavi, and Rudiger. My special thanks to Muriel Bardet and Francoise Behn for their every-day help and support on all sorts of different things. Chatting with Muriel has always been joyful. I thank Damir Laurenzi for the computer network management and his great patience.

I profited a lot from working with Ali Goli, Marco Mondelli, Ryuhei Mori, Ramtin Pedarsani and Liu Wei in different stages of my studies, and I am truly grateful to them. I am pretty sure that all of them will have a brilliant future.

I would like to extend my warmest gratitude to Vahid Aref for his close friendship over the past ten years. What strikes me the most about him is his amazingly caring and supportive character. I am also deeply indebted to his father, Prof. Mohammad-Reza Aref, for his well-known wisdom, support, kindness, and generosity of heart. His invaluable pieces of advice have been extremely useful throughout my life. I would also like to thank Maryam for her incredible hospitality and kindness. I will never forget my birthday cake that was carefully designed according to my thesis topics. I think she had well realized that my thesis work is in fact a piece of cake!

I am also extremely grateful to so many awesome friends who made my study in Lausanne an unforgettable stage of my life and full of memorable moments: Adel Javanmard, Ali Mousavi, Ali Naghavi, Amin Shoaie, Amir reza Rahmani, Amir Mortazavi, Azadeh and Ebrahim, Ehsan Kazemi, Ehsan Valavi, Fatemeh and Behrooz, Haleh and Mohsen, Hamed Izadi and his dear

family and uncles, Hassan Pezeshgi, Hessem Mahdaviifar, Hossein Mamaghani, Kamran, Mahboobeh and Hadi, Mahdi Aminian, Mahdi Enshayi, Mahsa and Hanif, Mani, Maryam and Ali Hormati, Maryam and Kian and Ali, Masoud Alipour, Mohammad Javad Ostadmirza, Mina and Mohammad Karzand, Mitra and Arash, Mojgan and Mohsen, Nahah and Hossein, Narges and Javad, Naghmeh and Mahdad, Nastaran and Omid, Nasibeh and Sina, Negar and Hessem, Nooshin and Pedram, Omid Etesami, Payam Delgosha, Pedram Pad, Pouya Shariatpanahi, Ramtin, Roxana, Sadegh Astanah and his family, Sara and Davood, Sharzad and Nikita and Salman, Samira and Mokhtar, Sanaz and Naser, Seyed Reza Yousefi, Simin and Mohammad Javad, Sobhan, Somayyeh, Soonaz, and Vahid Majidzadeh.

The final parts of an acknowledgement are usually for the special ones. I am deeply grateful to my three brothers and their families: Bahar and Arash and Kamran, Elham and Baran and Hamid, and Behzad. My cousin, Omid, with whom I have grown up, is also a brother to me and I thank him and his wife Behnaz. In fact, I have five more brothers: my precious friends Farid, Masih, Mohammad, Reza, and Saeed. Their friendship is one of the most valuable things I have had. I deeply thank them for all the great memories we have been sharing day and night. I will miss them the most.

I owe my deepest gratitude to Shirin for her affection, friendship, care, inspiration and support.

I am deeply indebted to my parents for their infinite love and support. Their profound wisdom has been the greatest source of inspiration throughout these years. This thesis is dedicated with love to them.

Contents

Abstract	i
Résumé	iii
Acknowledgements	v
Contents	ix
1 Introduction	1
1.1 Polarization	2
1.2 Spatial Coupling	4
1.3 Outline of this Thesis	7
I Polarization	9
2 Polarization and Polar Codes	11
2.1 Introduction	11
2.1.1 Contributions of the First Part (Chapters 3-6)	14
2.2 Basic Setting and Notations	16
2.3 Channel Polarization	17
2.3.1 Polarization Process	22
2.3.2 Polar Codes	24
2.4 Polar Codes Based on $\ell \times \ell$ Matrices	24
3 Scaling Laws for the Un-Polarized Channels	29
3.1 Problem Formulation	29
3.2 Heuristic Derivation for the BEC	32
3.2.1 Scaling Law Assumption	32
3.3 Analytical Approach: from Bounds for the BEC to Universal Bounds for BMS Channels	37
3.3.1 Characterization of μ for the BEC	37
3.3.2 Speed of Polarization for General BMS Channels	44
3.3.3 Universal Bounds on the Scaling Behavior of Polar Codes	47

3.4	Extensions and Improvements	54
3.5	Appendix: Auxiliary Lemmas and Proofs	56
4	Scaling Laws for the Polarized Channels	71
4.1	Problem Formulation	71
4.1.1	Relevant Work	73
4.2	Asymptotic Behavior of $F_n(z)$	74
4.2.1	Preliminaries	75
4.2.2	The Idea behind the Proof	75
4.2.3	A Generic Process	76
4.2.4	Proof of (4.9) in the Forward Direction	77
4.2.5	Proof of (4.9) in the Reverse Direction	78
4.3	Asymptotic Behavior of the MAP Error	80
4.4	Further Remarks	83
4.4.1	The Common Indices between Polar and Reed-Muller Codes	83
4.4.2	Selection Rule of the Rows	84
5	Efficient Construction and Universality	87
5.1	Problem Formulation	87
5.1.1	Hardness of the Construction	87
5.1.2	Is Polar Coding Universal?	89
5.2	Algorithms for Efficient Construction	90
5.2.1	Greedy Mass Transportation Algorithm	91
5.2.2	Mass Merging Algorithm	92
5.2.3	Bounds on the Approximation Loss	93
5.2.4	Exchange of Limits	97
5.2.5	Simulation Results	97
5.3	Polar Codes with SC Decoding Are Not Universal	99
5.4	Bounds on Compound Rate of BMS Channels	101
5.4.1	Trivial Bounds	101
5.4.2	A Better Universal Lower Bound	102
5.5	Extensions and Improvements	105
6	Robustness of the Successive Cancellation Decoder	107
6.1	Problem Formulation	107
6.1.1	Quantized SC Decoder	108
6.2	General Framework for the Analysis	109
6.2.1	Equivalent Tree Channel Model and Analysis of the Probability of Error for the Original SC Decoder	109
6.2.2	Quantized Density Evolution	111
6.3	Quantized SC Decoders with Different Precisions	112
6.3.1	1 Bit decoder: The Gallager Algorithm	112
6.3.2	1-Bit Decoder with Erasures	113
6.3.3	Scaling of the Gap to Capacity with Respect to the Number of Precision Bits	118

6.4	Further Remarks and Open Directions	121
6.5	Appendix: Proofs	121
II Threshold Saturation		125
7	Threshold Saturation on Coupled Graphical Models	127
7.1	Introduction	127
7.2	The Simplest Mean-Field Model: The Curie-Weiss Model	128
7.2.1	Basic Setting	129
7.2.2	Coupled Curie-Weiss Model	133
7.2.3	Contributions of Chapter 8	134
7.3	Constraint Satisfaction Problems	136
7.3.1	Basic Setting and Notation	136
7.3.2	The K -SAT Ensemble	137
7.3.3	The Coupled K -SAT Ensemble	145
7.3.4	Contributions of Chapters 9 and 10	146
8	Coupled Mean Field Models	151
8.1	Problem Formulation	151
8.2	Chain of Ising Systems on Complete Graphs	152
8.2.1	Curie-Weiss Model	152
8.2.2	Chain Curie-Weiss Model	153
8.3	A Continuum Approximation	157
8.4	Numerical Solutions	163
8.5	Further Remarks and Open Directions	168
8.6	Appendix	168
9	Coupled Constraint Satisfaction Problems	171
9.1	Problem Formulation	171
9.2	General Setting	172
9.2.1	Individual CSP Ensemble $[N, K, \alpha]$	172
9.2.2	Coupled-CSP Ensemble $[N, K, \alpha, w, L]$	173
9.2.3	K -SAT, Q -COL and K -XORSAT	174
9.3	Interpolation Arguments: From the Individual Ensemble to the Coupled Ensemble and Vice Versa	175
9.4	Zero Temperature Cavity Method and Survey Propagation Formalism	177
9.5	Coupled K -SAT Problem	179
9.5.1	Numerical Implementation	179
9.5.2	Survey Propagation for Large K	183
9.5.3	Solutions for Large K	186
9.6	Dynamical and Condensation Thresholds	190
9.7	Further Remarks and Open Directions	191
9.8	Appendix	192
9.8.1	Proofs of Theorems 9.1 and 9.2	192

9.8.2	Finite Temperature Version	196
9.8.3	Review of the Cavity Method and Survey Propagation Equations	199
10	Algorithmic Implications	203
10.1	Problem Formulation	203
10.2	Peeling Algorithms and Coupled Scalar Recursions	204
10.2.1	Pure Literal: A Peeling Algorithm for K -SAT	204
10.2.2	Peeling Algorithms for Q -COL and K -XORSAT	207
10.2.3	The Framework of Coupled Scalar Recursions	208
10.3	Unit Clause Propagation	210
10.3.1	Individual Ensemble	210
10.3.2	Description of UC Algorithm for the Coupled Formulas	211
10.3.3	Analysis of the Evolution of UC via Differential Equations	212
10.3.4	Numerical Implementation	220
10.3.5	Further Simplifications	221
10.3.6	Conserved Quantities	223
10.3.7	Slightly Modified Initial Conditions	224
10.3.8	A Potential Function	226
10.3.9	The Threshold of the UC Algorithm for the Coupled En- semble	227
10.3.10	How Does the Profile Look Like?	229
10.3.11	Why Does the UC Algorithm Work?	232
10.4	Further Remarks and Open Directions	232
10.5	Appendix: Auxiliary Lemmas and Proofs	233
10.5.1	Appendix A: A Message Passing Interpretation For UC	233
10.5.2	Appendix B: Auxiliary Lemmas and Proofs	235
	Bibliography	259
	Curriculum Vitae	269

Introduction

1

In communication and computer sciences, we typically have to design systems that are both reliable and efficient (in terms of resources and complexity). Examples include channel coding, data compression, compressive sensing, machine learning, and vision. In the last half century, a variety of ingenious designs have been conceived for each of these issues. However, despite all these efforts there is still room for improvement.

As a concrete scenario, consider the problem of channel coding: A sender desires to send K bits of information. The data is to be transmitted through a noisy channel that accepts input symbols one at a time and produces a sequence of output symbols. As the channel is noisy, the sender desires (i) to come up with an explicit (encoding) transform that generates a sequence of N symbols from the K information bits and transmits these N symbols through the channel and (ii) to reliably recover (decode) the K information bits from the N (noisy) output symbols of the channel. Shannon's channel coding theorem guarantees that these two requirements can be met, as long as K and N are sufficiently large, and as long as the "rate" $\frac{K}{N}$ is less than a fundamental quantity called the *channel capacity*. If we are interested only in the existence of such systems and have no constraints on the complexity, then the problem is solved. Shannon has already showed us how to accomplish this: just pick a random element from a suitably defined ensemble. The problem becomes much more difficult when we impose constraints on the complexity, or if we want an explicit (low-complexity) construction with the above properties.

Throughout this thesis we address two techniques that can be applied to a broad range of problems and have, roughly speaking, the following property. Starting with a "hard" problem, these two techniques enable us to transform the problem into an "easier" one and we will then solve this easier problem with a standard efficient algorithm. Of course, this "simplification" of the task

does not come entirely for free. As we will see, we lose in terms of the “dimensionality” of the problem, i.e., we have to increase the number of dimensions we work in. For instance, if we return to the concrete problem of coding, in order for the above scheme to work, we are required to work on larger instances of the problem. For some applications that are delay-sensitive (e.g., speech) this might cause problems, but for other applications the extra delay is acceptable and this trade-off is very beneficial.

The two techniques are *polarization* and *spatial coupling*. Although these transforms are fundamentally different in nature, they both have led to efficient algorithm designs that solve many seemingly hard problems. Both of these transforms were originally invented in the context of channel coding and resulted in very efficient encoding/decoding systems that achieve the capacity of a wide range of channels. In addition, despite their recent introduction, both have already had a significant impact on other areas of communications and signal processing (compressed sensing), and they have led to new insight in computer science and statistical physics. We proceed by briefly explaining each of these transforms.

1.1 Polarization

Polarization was introduced by Arikan in 2008 in the seminal paper [1]. Arikan used this technique in the context of channel coding and on the special class of channels called binary memoryless symmetric (BMS) channels. The origin of polarization can be traced back to Arikan’s efforts to improve the rates achievable by convolutional codes and sequential decoding. Let us briefly explain this technique in the context of channel coding.

Consider a BMS channel W and let $I(W)$ denote its capacity. The idea of polarization is to take two independent copies of W and to create two new channels W^0 and W^1 with the following properties:

- (i) The sum of the capacity of W^1 and W^0 is equal to twice the capacity of W , hence no information is lost.
- (ii) The channel W^1 is “better” than W and the channel W^0 is “worse” than W .

More precisely, consider Figure 1.1 where a simple transform is applied to the inputs of two independent uses of W . Here, given the output of this system, namely (y_0, y_1) , assume that we want to infer the value of the bits (u_0, u_1) . This task can be accomplished in a successive manner: in the first step infer the value of u_0 and, once this task has been accomplished, in the second step make use of the estimate of u_0 to infer the value of u_1 . In this regard, the channel W^0 is the channel that u_0 “sees”, given the observation (y_0, y_1) (see Figure 1.2), and the channel W^1 is the channel that u_1 “sees”, given the output and the actual value of u_0 . A little thought makes it clear why W^0 is worse than W and why W^1 is better. In the language of information theory, we say

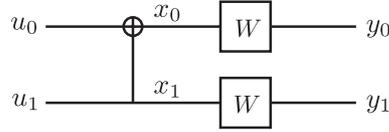


Figure 1.1: The basic channel transform where we combine two independent copies of W .

that W^1 is upgraded with respect to W and that W^0 is degraded with respect to W and we write

$$W^0 \preceq W \preceq W^1. \quad (1.1)$$

Intuitively, we think of W^1 as a less noisy version of W and W^0 as a more

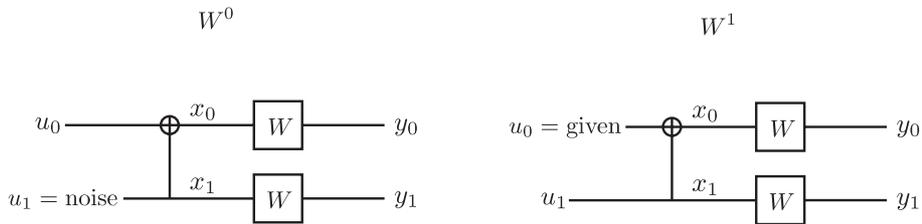


Figure 1.2: The channels W^0 (left) and W^1 (right).

noisy version of W . Degraded-ness/upgraded-ness can be thought of as a kind of majorization of densities. As a consequence, from two i.i.d. copies of W we have constructed two channels W^0 and W^1 with the property (1.1). Note that relation (1.1) already indicates some sort of polarization towards the two extremal channels: completely noiseless (perfect channel or the best channel) or completely noisy (useless channel or the worst channel). The idea of polarization is to apply recursively this simple transform to the channels W^1 and W^0 to create further polarized channels $(W^1)^1, (W^1)^0, (W^0)^1, (W^0)^0$ and so on. So in general, we begin with $N = 2^n$ independent copies of W and create 2^n new channels $W^{11\dots 1}, W^{11\dots 0}, \dots, W^{00\dots 0}$. A remarkable phenomenon here is that as n grows large, the created channels become increasingly polarized, i.e., almost all the channels are very close to one of the following two extremal states: completely noiseless (with capacity 1) or completely noisy (with capacity 0). Now, by using property (i), it is clear that, in the limit of large block lengths, the fraction of completely noiseless channels tends to $I(W)$ and the fraction of completely noisy channels tends to $1 - I(W)$. These two extremal channels are easy to deal with: For the perfect channel, we send the information uncoded; and for the useless channel, we send no information at all. Thus in a nutshell, we have introduced a transform that reduces the problem of coding over N independent uses of a channel W (a hard problem) to coding over N channels

that are either noiseless or useless (an easy problem). As we will see later, the price we pay is that N has to be taken sufficiently large in order to achieve a certain error probability.

The channel codes that are based on this technique are called polar codes. Note that polar codes achieve the capacity of any BMS channel, i.e., if we use sufficiently large block-lengths, then we can transmit reliably at any rate below capacity. In addition, these codes have low complexity, as we discuss in more detail in the next chapter.

Polar codes have opened a completely new chapter in coding theory. Much of classic coding is based on algebraic notions (Hamming distance, fields, etc.) and iterative codes are based on carefully designed bipartite graphs. The design of polar codes is based on the nature of the capacity of the individual sub-channels. This makes coding a very natural extension of information theory. The fact that polar codes can be shown in just a few paragraphs to be capacity achieving, adds to their appeal.

Soon after the invention of the technique of polarization, a large body of work generalized the basic technique to a much wider set of scenarios. For a partial list of references see [29]-[42]. We will have more to say about this later.

1.2 Spatial Coupling

It is convenient to explain the idea of spatial coupling in the general framework of graphical models. Graphical models and their associated message-passing algorithms play an increasingly significant role in communications, computer science and statistical physics. We begin by explaining the general theme of spatially coupled graphical models, then we will follow with a concrete example.

Given a graphical model that represents a “hard” problem (e.g., decoding, inference), we construct from it a larger instance of the same problem but with a particular graphical structure. This structure can be thought of as adding a spatial dimension to the graphical model at hand. Due to this additional structure, the new instance is significantly easier to solve. Mathematically speaking, this change in behavior manifests itself by a significant increase in the threshold under low-complexity processing. For instance, in inference problems this “threshold” is a measure of how much “noise” the system can tolerate and still be expected to work correctly. Naturally, under optimal processing (which is typically of exponential complexity) a system can tolerate significantly more noise than under low-complexity (message-passing) processing. The curious and important characteristic of spatial coupling is that spatially coupled systems have a threshold under low-complexity processing, which is as large as the threshold of the underlying system under optimal processing. This phenomenon is named threshold saturation in [2, 3]. The word “saturation” indicates that the threshold under low-complexity processing has increased to its largest possible value, namely the value of optimal processing. In other words, it has saturated. This threshold saturation effect has an obvious and

important operational consequence: for a properly designed graphical model, low-complexity processing suffices to reach the optimal performance.

We illustrate the idea of spatial coupling with a concrete example. The graphical model that we choose lies in the area of channel coding, where spatial coupling was first developed. We recall that the task of channel coding¹ is to generate, from K information bits, a codeword consisting of N bits. One way to do this is via a linear transform. That is, we can think of the set of K information bits as an element of the vector space $\{0, 1\}^K$. And to generate the corresponding codeword, we use a linear transform that maps $\{0, 1\}^K$ to a K -dimensional subspace of $\{0, 1\}^N$. Such codes are called *linear codes*. Given a linear code \mathcal{C} , we can associate to it a binary matrix H of size $N \times (N - K)$, such that the following holds. The set of codewords of \mathcal{C} is the set of solutions of the equation² $xH = 0$. Once a particular codeword x is sent through the channel, a noisy version of the codeword, called y , is obtained. Given y , the decoding task is to infer the transmitted codeword x .

For general matrices H , and if we require worst-case guarantees, this is a difficult task (NP-complete to be precise). But if we assume that the matrix is sparse and we only require a good performance on average, we enter the realm of sparse graph codes. In this setting, graphical models combined with message-passing decoders perform very well. Message-passing algorithms have, by definition, low complexity. They exhibit good performance for properly designed systems and map easily into hardware. The basic idea of such algorithms is to use the graphical structure and to solve the problem locally by sending messages along the edges of the graph. In the case of inference problems, these messages are probabilities that reflect the current estimate of the system. These estimates are iteratively updated until they converge or a maximum number of iterations is reached. Figure 1.3 shows and explains the natural graphical model associated with this problem.

Let us now demonstrate the idea of spatial coupling via the help of the simple graphical model of Figure 1.3. Let L be a positive integer. Given a graphical model, call it our *base graph or copy*, we start with L copies of the base graph and assign them one by one to positions $i \in \{0, 1, \dots, L - 1\}$ (see Figure 1.4). Next, we connect the neighboring copies. There are many ways in which these connections can be done, and experiments indicate that the exact type of coupling is secondary, as long as the systems are coupled “sufficiently strongly”. To be concrete, we perform the following operation (see Figure 1.4). For each copy we randomly label the edges into three equally-sized groups called left, center, and right. The edges labeled “center” are kept as is. For the edges labeled “left”, we keep one end of the edge in the current copy and connect the other end to a corresponding node on the left. In a similar manner, for the edges labeled “right” we keep one end in the current copy but connect the second end to a corresponding node on the right. In other words, we swap some of the edges between neighbors and preserve the local structure

¹We assume here that we are transmitting over a channel with binary input.

²Here, addition and multiplication is performed in the binary field $\mathbb{F}_2 = \{0, 1\}$.

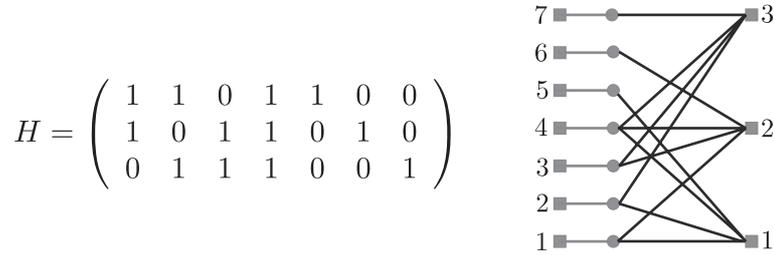


Figure 1.3: Each variable node (circle) corresponds to one code bit, i.e., one component of x . This corresponds to the columns of matrix H . Each check node (square) on the right corresponds to one linear constraint imposed by matrix H ; i.e., one row of H . Finally, each square on the left corresponds to one of the observations; i.e., one of the components of y . Of course, this is a toy example. Real applications use codes of length one thousand or even one million.

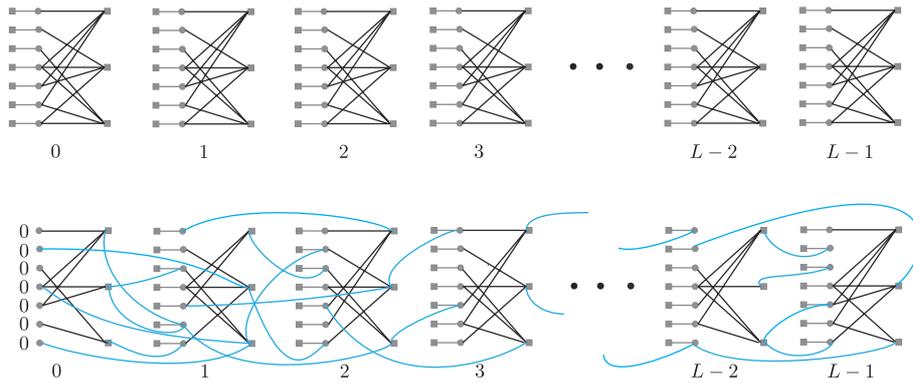


Figure 1.4: Top figure: we put L copies of a base graph in a row; each at a position $i \in \{0, 1, \dots, L-1\}$. Bottom figure: we then connect the neighboring copies (with a careful termination at the boundaries) to create the spatially coupled graph.

(node degrees) of each copy. At the boundary, some of the edges cannot be connected because either the left or right neighbor is missing. These edges are terminated in a proper way, depending on the problem at hand. The general theme is to make the problem slightly easier at the boundaries. For example, for coding, if we have an edge that goes to a missing variable node then we set the corresponding variable node to a known value. In this case, we give the system some additional information, i.e., we make the problem easier at the boundaries. Note here that if the original graph represents an error-correcting code then the coupled graph also represents an error correcting code. The

new code is L times larger and has a particular global graphical structure, but locally it looks identical to the original model (e.g., it has the same degree distribution).

As a consequence of the coupling there is a remarkable performance improvement under message-passing. This improvement is due to the special termination, as well as to the special spatial structure. For instance, consider the decoding problem. Due to the extra help at the boundary, the decoding problem can be partially solved there, even when using a suboptimal message-passing decoder. This in turn makes the decoding problem easier for the neighboring copies. This effect cascades over the whole length of the chain. A simple, albeit not very accurate, analogy is a chain of properly placed domino pieces. Once we topple a boundary piece, the whole chain is toppled. But contrary to domino pieces, where forces act in only one direction, in spatially coupled graphs there is true interaction, and information bounces back and forth between neighboring systems. The physical effect is akin to what happens when crystals grow or when super-cooled liquids are seeded. Perhaps the most surprising aspect of spatial coupling is not that the performance is simply improved, but that such systems perform under low-complexity processing as well as if they had been processed optimally, i.e., their performance saturates.

The origins of spatial coupling reach back to the area of channel coding with the work of Felstrom and Zigangirov in [4]. They introduced a class of sparse-graph codes with a convolutional structure, which they named “convolutional LDPC codes”. They observed through numerical simulations that this class of codes performs very well. In subsequent years, considerable follow-up work was done on various aspects of the performance of such codes (see [3] for a historical review). However, a rigorous reason behind the fact that spatially coupled systems perform so well in channel coding, and in particular the phenomenon of threshold saturation, was discovered only recently by Kudekar, Richardson and Urbanke [2,3]. This picture has since been completed/generalized by a vast amount of studies on graphical models in communications, computer science, and statistical physics. In each of these studies, the same fundamental phenomenon is observed when we spatially couple the underlying graphical model into a chain.

One important aspect of [2,3] is the emergence of a new class of codes, called spatially coupled LDPC codes, that achieve the capacity of the class of BMS channels. Later on, in a sequence of works (starting from [44], [45], and finally in [46]) this technique was applied to compressive sensing and yielded low-complexity compressive sensing schemes that are optimal in terms of the required number of measurements, quite a remarkable achievement.

1.3 Outline of this Thesis

This thesis consists of two parts. We focus in the first part (including Chapters 2-6) on the technique of polarization and polar codes. The results and the conclusions of the first part are summarized in Section 2.1.1. We consider the

technique of spatial coupling and the threshold saturation phenomenon in the second part (including Chapters 7-10). The results and the conclusions of the second part are summarized in Sections 7.2.3 and 7.3.4. At the end of each chapter possible extensions, improvements and implications are discussed. The discussions point toward some new research directions and open problems.

Part I

Polarization

Polarization and Polar Codes

2

2.1 Introduction

The first part of this thesis concentrates on polarization and polar codes in the context of channel coding. As mentioned in Chapter 1, channel coding is a central topic of information theory and its main purpose is for the design of efficient and reliable encoding/decoding systems that can operate at any rate below the channel capacity. Throughout the sixty years of the development of channel coding, we have witnessed many breakthroughs in understanding fundamental properties of “good” codes. This has led to various remarkable code designs. We can divide these code designs into two general groups: algebraic codes and iterative codes. Algebraic codes were the initial focus of coding theorists after the emergence of Shannon’s framework in 1948, whereas iterative codes came into serious consideration in 1993 through the invention of turbo codes. For several channels of practical interest, iterative codes provide efficient and reliable codes that operate close to the capacity [5].

Polar codes, invented by Arikan [1], are arguably the first family of low-complexity codes that provably achieve the capacity of any binary symmetric memoryless (BMS) channel. It is worth mentioning that the class of BMS channels contains some of the channels of most practical relevance (such as the binary-input additive white Gaussian noise channel, the binary symmetric channel, and the binary erasure channel). Soon after the invention of polar codes, a large body of research generalized the idea of polarization to other classes of channels, as well as other information theoretic scenarios. As a consequence, a wide range of channels, including discrete memoryless channels (DMC) and some of the well-known continuous memoryless channels (such as the AWGN channel), now have specific polar codes that achieve their capacity.

Although the main focus of coding theory has been on designing efficient

codes that achieve capacity, the recent advancements in real-world technology have brought about (or accredited) new criteria for the design of “good” codes. Such newly emerging practical criteria are the main focus of *modern* coding theory. For instance, more than half of the traffic in today’s wireless networks is video and this ratio is still growing. Such kind of real-time traffic brings about the need of low-delay coding systems.

Let us mention the most significant criteria of modern coding theory.

- (i) [Scaling Laws] In coding, the three most important parameters are rate (R), block-length (N), and block error probability (P_e). Ideally, given a family of codes, we would like to be able to describe the exact relationship between these three parameters. This however is a formidable task. It is slightly easier to fix one of the parameters and then to describe the relationship (scaling) of the remaining two. We proceed by explaining the two most relevant scaling laws.

Assume that we fix the rate and consider the relationship between the error probability and the block-length. This is the study of the classic *error exponent*. For instance, for random codes a closer look shows that $P_e = e^{-NE(R,W)+o(N)}$, where $E(R,W)$ is the so-called random error exponent [6] of the channel W .

Another option is to fix the error probability and to consider the relationship between the block-length and the rate. In other words, given a code and a desired (and fixed) error probability P_e , what is the block-length N required, in terms of the rate R , so that the code has error probability less than P_e ? This scaling is arguably more relevant (than the error exponent) from a practical point of view as we typically have a certain requirement on the error probability and are interested in using the shortest code possible to transmit at a certain rate.

In practice, the shorter a code is the better it is, since this implies small delays. As a benchmark, the shortest block-length that that we can hope for is as follows. It is not hard to see that the random variations of the channel themselves require $R \leq I(W) - \Theta(\frac{1}{\sqrt{N}})$, or equivalently, $N \geq \Theta(\frac{1}{(I(W)-R)^2})$. Indeed, a sequence of works starting from [8], then [9], and finally [10] showed that the minimum possible block-length N required to achieve a rate R with a fixed error probability P_e is roughly equal to

$$N \approx \frac{V(Q^{-1}(P_e))^2}{(I(W) - R)^2}, \quad (2.1)$$

where V is a characteristic of the channel referred to as channel dispersion, and Q is the complementary Gaussian cumulative distribution function. In other words, the best codes require a block-length equal to $\Theta(\frac{1}{(I(W)-R)^2})$.

- (ii) [Universality] In reality, no channel is exactly equal to the mathematical models that we consider. Also, depending on the conditions of the

transmission medium, the channel might vary inside a *set* of channels. This leads us to consider the following scenario. Given a set of channels, what is the maximum rate achievable simultaneously on all these channels by a fixed code? Again as a benchmark, we consider the best possible rates that can be achieved in such a setting. This is known as the *compound channel* scenario. Let \mathcal{W} denote the set of channels. The compound capacity of \mathcal{W} is defined as the rate at which we can reliably transmit irrespective of the particular channel (out of \mathcal{W}) that is chosen to transmit. In other words, the channel is not known at the transmitter and the only information about the channel is that it is inside the set \mathcal{W} , whereas at the receiver, the channel is known. The compound capacity is given by [11]

$$C(\mathcal{W}) = \max_P \inf_{W \in \mathcal{W}} I_P(W), \quad (2.2)$$

where $I_P(W)$ denotes the mutual information between the input and the output of W , with the input distribution being P . Note on one hand that the compound capacity of \mathcal{W} can be strictly smaller than the infimum of the individual capacities. This happens if the capacity-achieving input distributions for the individual channels are different. On the other hand, if the capacity-achieving input distribution is the same for all channels in \mathcal{W} , then the compound capacity is equal to the infimum of the individual capacities. This is indeed the case for us, because we restrict our attention to the class of binary-input memoryless output-symmetric (BMS) channels.

- (iii) [Complexity] Another important aspect of the design of modern codes is complexity. It is an easy task to theoretically distinguish between having polynomial or exponential complexity (in the block-length N). However, a harder and more important task for practical purposes is to see how “polynomial” a code design is. When it comes to practice, the importance of complexity goes even beyond being polynomial and it is often the case that the “constants” matter. For instance, from a practical perspective, a code with complexity $\Theta(N^{10})$ is not suitable for implementation despite having polynomial complexity, whereas “good” systems typically require linear complexity. Also, when it comes to implementation, other issues become important, such as the issue of how well a particular design maps into hardware.

Typically, the term complexity can be measured in several ways. Two well-known aspects are *algorithmic* complexity and *space* complexity. The algorithmic complexity refers to measuring the number of operations in the encoding and decoding procedures, whereas the space complexity refers to the amount of memory required for these procedures. Often, the main bottleneck in the implementation of large- and high-speed coding systems is memory (especially in the decoding part). In this regard, even a factor 2 in memory usage can make a significant difference.

Such considerations have defined a modern framework for efficient code design; its ultimate goal can be summarized as follows: design universal codes with low complexity in memory and computation (preferably linear in block-length) with block-lengths as short as possible (see (2.1)). Let us now see how well polar codes fit into this framework.

2.1.1 Contributions of the First Part (Chapters 3-6)

In the first part of this thesis, we consider polar codes and their generalizations, and investigate analytically their suitability in terms of the aforementioned framework. We often confirm our results with numerical simulations and also use the numerics to build intuition. All the results of this thesis are specific to BMS channels. In the following, we summarize the contributions of each specific chapter.

Scaling laws of polar codes are the main subject of Chapters 3 and 4. In Chapter 3 we consider the tradeoff between the rate and the block-length for a fixed error probability, i.e., we consider the finite-length scaling of polar codes. We show that the finite-length scaling of polar codes is intimately related to the dynamics of the channel polarization phenomenon. This stimulates us to derive scaling laws for the speed of polarization. Using such laws, we then provide scaling laws for polar codes that hold universally for all BMS channels.

The main results of Chapter 3 can be summarized as follows. Let W be a BMS channel with capacity $I(W)$. Fix the error probability¹ to a given value $P_e > 0$. Then the required block-length N scales in terms of the rate $R < I(W)$ as

$$N \geq \frac{\alpha}{(I(W) - R)^\mu}, \quad (2.3)$$

where α is a positive constant that depends on P_e and $I(W)$. We show that $\underline{\mu} = 3.55$ is a valid choice, and we conjecture that indeed the value of $\underline{\mu}$ can be improved to $\underline{\mu} = 3.627$, the parameter for the binary erasure channel. A comparison of (2.3) and (2.1), indicates that polar codes require a larger block-length with respect to what is optimally achievable. This gives a fundamental explanation for the numerical observations that polar codes require a larger block-length with respect to the best codes used in current practice.

Also, in Chapter 3 we show that with a fixed error probability $P_e > 0$, the block-length scales in terms of the rate as

$$N \leq \frac{\beta}{(I(W) - R)^{\bar{\mu}}}, \quad (2.4)$$

where β is a constant that depends on P_e and $I(W)$, and $\bar{\mu} = 7$. In the language of coding theory, from (2.4) we can say that for polar codes the block-length N scales “polynomially” with respect to the gap to capacity.

¹To be more precise, we consider the sum of Bhattacharyya parameters of the sub-channels chosen (by the polar coding scheme) as a proxy for the error probability.

In Chapter 4, we consider the relationship between the error probability and the block-length at a fixed rate, i.e., we consider the error exponent of polar codes. It was previously shown by Arikan and Telatar [7] that for any fixed rate $R < I(W)$, the block error probability is upper bounded by 2^{-N^β} for any $\beta < \frac{1}{2}$ and N large enough. By a careful study of the asymptotic behavior of channel polarization, we refine this result to be dependent on R , i.e., for polar codes with the successive cancellation (SC) decoder

$$P_e = 2^{-2^{\frac{n}{2} + \sqrt{\pi} Q^{-1}\left(\frac{R}{I(W)}\right) + o(n)}}, \quad (2.5)$$

where $n = \log_2 N$ and $Q(t) \triangleq \int_t^\infty e^{-z^2/2} dz / \sqrt{2\pi}$. We further show that the MAP decoder shares the the same scaling behavior as (2.5). Our results apply to general polar codes based on $\ell \times \ell$ kernel matrices. We also generalize these scaling relations for extended polar codes that are based of $\ell \times \ell$ matrices.

In Chapter 5, we consider two problems concerning the construction and the universality of polar codes. We first consider the problem of efficiently constructing polar codes over BMS channels. The complexity of designing polar codes via an exact evaluation of the polarized channels (to find which ones are good) appears to be exponential in the block length. In [25], Tal and Vardy show that if instead the evaluation is performed approximately, the construction has only linear complexity. We follow this approach and present a framework where the algorithms of [25], as well as new proposed algorithms, can be analyzed for complexity and accuracy. We provide numerical and analytical results on the efficiency of such algorithms. In particular, we show that one can find all the good channels (except a vanishing fraction) with almost linear complexity in block-length (except a poly-logarithmic factor). We then ask how much the construction of a polar code for a given channel would help in the construction of a polar code for another channel. This motivates us to consider the compound capacity of polar codes under the successive cancellation (SC) decoding for a collection of BMS channels. By deriving a sequence of upper and lower bounds, we show that in general the compound capacity under successive decoding is strictly smaller than the unrestricted compound capacity in (2.2). This indeed indicates that polar codes with the successive decoder are not universal.

Successive decoding with few messages is the subject of Chapter 6. Robustness of a decoder with respect to the number of bits per message is a key element in memory efficiency hence the practicality of that code. The original successive cancellation decoder of Arikan assumes infinite precision arithmetic. Given the successive nature of the decoding algorithm, there might be concern about the sensitivity of the performance to the precision of the computation. We show that even very coarsely quantized decoding algorithms lead to excellent performance. More concretely, we show that under successive decoding with an alphabet of cardinality only three, the decoder still has a threshold and this threshold is a sizable fraction of capacity. More generally, we show that if we are willing to transmit at a rate δ below capacity, then we universally need only $O(\log \frac{1}{\delta})$ bits of precision.

2.2 Basic Setting and Notations

Consider a memoryless channel W . Such a channel is characterized by the following quantities:

- The input alphabet \mathcal{X} .
- The output alphabet \mathcal{Y} .
- The transition probabilities $\{W(y | x) : x \in \mathcal{X}, y \in \mathcal{Y}\}$.

In this thesis we restrict ourselves to binary memoryless symmetric (BMS) channels. We say that W is binary if its input alphabet is binary, i.e., $\mathcal{X} = \{0, 1\}$. Further, we say that W is symmetric if there exists a permutation $\pi : \mathcal{Y} \rightarrow \mathcal{Y}$ such that

- $\pi = \pi^{-1}$.
- $W(y | 0) = W(\pi(y) | 1)$ for all $y \in \mathcal{Y}$.

There are some BMS channels of particular interest. We will often use them to illustrate specific concepts and properties. We now introduce these channels and give a brief illustration of them in Fig 2.1.

- Binary erasure channel (BEC) with erasure probability z , which we denote by $\text{BEC}(z)$. The value of z lies inside $[0, 1]$.
- Binary symmetric channel (BSC) with cross over probability ϵ which we denote by $\text{BSC}(\epsilon)$. The value of ϵ lies inside $[0, \frac{1}{2}]$.
- Binary additive white Gaussian noise channel (BAWGNC) with noise variance σ^2 which we denote by $\text{BAWGNC}(\sigma)$. The value of σ lies inside $(0, \infty)$.

Associated to any BMS channel are the following useful parameters.

$$I(W) = \sum_{y \in \mathcal{Y}} W(y | 1) \log \frac{W(y | 1)}{\frac{1}{2}W(y | 1) + \frac{1}{2}W(y | 0)}, \quad (2.6)$$

$$H(W) = 1 - I(W), \quad (2.7)$$

$$Z(W) = \sum_{y \in \mathcal{Y}} \sqrt{W(y | 0)W(y | 1)}, \quad (2.8)$$

$$E(W) = \frac{1}{2} \sum_{y \in \mathcal{Y}} W(y | 1) \exp\left\{-\frac{1}{2}\left(\ln \frac{W(y | 1)}{W(y | 0)} + \left|\ln \frac{W(y | 1)}{W(y | 0)}\right|\right)\right\}. \quad (2.9)$$

The parameter $I(W)$ is the capacity of W or the mutual information between the input and the output assuming a uniform distribution on the inputs. The parameter $H(W)$ is equal to the entropy of the input of W given its output when we assume a uniform distribution on the inputs, i.e., $H(W) = H(X |$

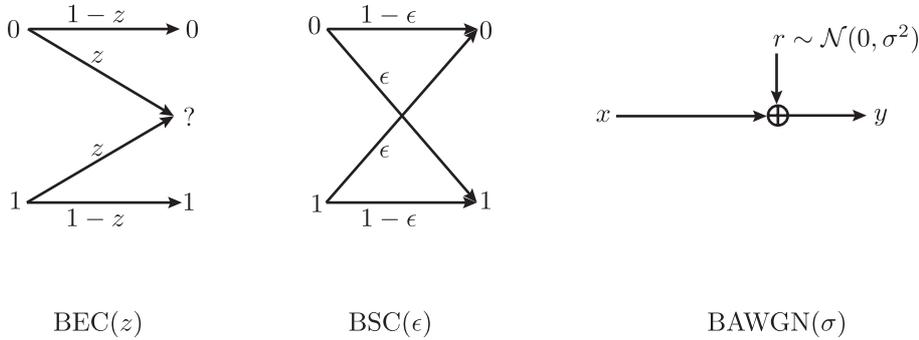


Figure 2.1: Left figure: the BEC with erasure probability z . An input symbol to this channel is erased with probability z or passed through the channel with probability $1-z$. Middle figure: the BSC with cross-over probability ϵ . An input to this channel is flipped with probability ϵ or passed through with probability $1-\epsilon$. Right figure: the BAWGN with noise variance σ^2 . The output y of this channel is the sum of its input $x \in \{0, 1\}$ and a noise value r which is a gaussian r.v. with mean 0 and variance σ^2 .

Y). Hence, we call the parameter $H(W)$ the entropy of the channel W . The parameter $Z(W)$ is called the Bhattacharyya parameter of W . Finally, $E(W)$ is called the error probability of W . It can be shown that $E(W)$ is equal to the error probability in estimating the channel input x on the basis of the channel output y via the maximum-likelihood decoding of $W(y|x)$ (with the further assumption that the input has uniform distribution).

It can be shown that the following relations hold between these parameters (see for e.g., [1] and [47, Chapter 4]):

$$0 \leq 2E(W) \leq H(W) \leq Z(W) \leq 1, \quad (2.10)$$

$$H(W) \leq h_2(E(W)), \quad (2.11)$$

$$Z(W) \leq \sqrt{1 - (1 - H(W))^2}, \quad (2.12)$$

where $h_2(\cdot)$ denotes the binary entropy function.

2.3 Channel Polarization

We start by illustrating the channel polarization phenomenon in its simplest form by using Arikan's original construction [1]. We then briefly mention generalizations of this phenomenon using other constructions. Channel polarization consists of three stages as follows.

Step 1 (Channel Splitting): Let \mathcal{W} denote the class of BMS channels. Let us define a channel transform $W \rightarrow (W^0, W^1)$, called channel splitting, that maps \mathcal{W} to $(\mathcal{W}, \mathcal{W})$. In other words, channel splitting is a transform which takes a

BMS channel W as input and outputs two BMS channels W^0 and W^1 that are constructed as follows. Having the channel $W : \{0, 1\} \rightarrow \mathcal{Y}$, the channels $W^0 : \{0, 1\} \rightarrow \mathcal{Y}^2$ and $W^1 : \{0, 1\} \rightarrow \{0, 1\} \times \mathcal{Y}^2$ are defined as

$$W^0(y_1, y_2 | u_0) = \sum_{x_2 \in \{0, 1\}} \frac{1}{2} W(y_1 | u_0 \oplus u_1) W(y_2 | u_1), \quad (2.13)$$

$$W^1(y_1, y_2, u_0 | u_1) = \frac{1}{2} W(y_1 | u_0 \oplus u_1) W(y_2 | u_1). \quad (2.14)$$

Let us now explain what is the intuitive meaning behind the formulas (2.13) and (2.14) and why W^0 is worse than W where as W^1 is better. Consider a setting as in Figure 2.2 where two independent copies of W are used for transmission. We have two input bits u_0, u_1 with i.i.d. distribution Bernoulli($\frac{1}{2}$) which are

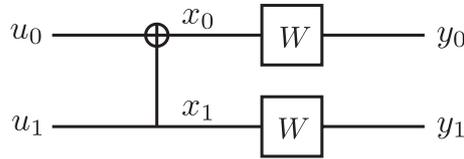


Figure 2.2: The basic channel transform where we combine two independent copies of W .

combined using a simple transform

$$(x_0, x_1) = (u_0, u_1) \underbrace{\begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}}_{G_2}. \quad (2.15)$$

The resulting bits x_0, x_1 are then transmitted through two independent copies of W to form the output values y_0, y_1 . One can easily see that the transition probability of the output (y_0, y_1) given (u_0, u_1) is

$$\begin{aligned} \Pr(y_0, y_1 | u_0, u_1) &= W(y_0 | x_0) W(y_1 | x_1) \\ &= W(y_0 | u_0 \oplus u_1) W(y_1 | u_1). \end{aligned} \quad (2.16)$$

Of course, the optimal way to infer the value of the bits (u_0, u_1) , given only the output (y_0, y_1) , is via ML decoding with transition probabilities (2.16). Now consider the following sub-optimal *successive* decoder. First decode u_0 assuming no information about u_1 (i.e., by treating u_1 as noise) and then use the resulting estimate for u_0 to decode u_1 . Figure 2.3 illustrates this successive decoder. Now for the analysis of the successive decoder lets us define the two events

$$E_0 = u_0 \text{ is decoded incorrectly}, \quad (2.17)$$

$$E_1 = u_0 \text{ is decoded correctly but } u_1 \text{ is decoded incorrectly}. \quad (2.18)$$



Figure 2.3: Successive decoding of (u_0, u_1) from the observation (y_0, y_1) : first decode u_0 assuming no information about u_1 (the left figure) and then use the resulting estimate for u_0 to decode u_1 (the right figure).

Clearly, the successive decoder fails if and only if at least one of the events E_0 or E_1 occur. Thus, one can write

$$\Pr(\text{successive decoder fails}) = \Pr(E_0 \cup E_1), \quad (2.19)$$

and as a result

$$\max(\Pr(E_0), \Pr(E_1)) \leq \Pr(\text{successive decoder fails}) \leq \Pr(E_0) + \Pr(E_1). \quad (2.20)$$

Finally, we show how one naturally derives the channels W^0 and W^1 defined in (2.13) and (2.14) from the events E_0 and E_1 respectively. A close look at the definition of W^0 in (2.13) reveals that the channel W^0 is precisely the channel between the bit u_0 and the output (y_0, y_1) when u_1 is assumed as noise. Indeed, in (2.13) the sum over u_1 corresponds to the fact that u_1 is assumed as a completely unknown random variable or as noise. The channel W^1 given in (2.14) is the channel between the bit u_1 and the vector (y_0, y_1, u_0) . In other words, for W^1 the value of u_0 is given as a part of the output. Hence, the channel W^1 models the event of decoding the bit u_1 given the true value of u_0 and the observation of (y_0, y_1) . It is now clear that the events E_0 and E_1 are precisely the event of failure in ML decoding the bits u_0 and u_1 from the channels W^0 and W^1 respectively (see Figure 2.4 for a better illustration). As a result, by (2.20) we obtain

$$\max(E(W^0), E(W^1)) \leq \Pr(\text{successive decoder fails}) \leq E(W^0) + E(W^1) \quad (2.21)$$

One can interpret W^0 as the channel that u_0 "sees" when u_1 is considered as noise and also W^1 can be interpreted as the channels that u_1 "sees" when u_0 is given. From Figure 2.4 one can also see why the channel W^0 is a "worse" channel with respect to W and W^1 is a "better" channel than W . In fact, in the channel W^0 , what the bit u_0 sees is the xor sum of two independent (noisy) observations of the bits x_0 and x_1 . Since the two observations are added, they mutually corrupt each other and the net result is less profitable (in terms of inferring about the bit u_0) than each of the individual observations. Hence, the channel W^0 is worse than W . The situation gets better for the bit u_1 on channel W^1 since for u^1 the two observations are given directly to u_1 without any further corruption. Hence, given the value of u_0 , the bit u_1 has at hand

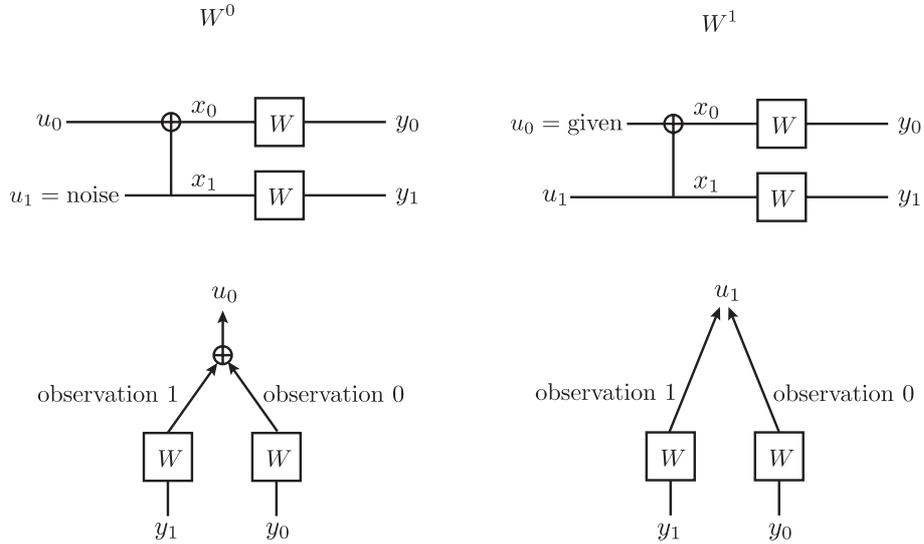


Figure 2.4: The figures on the left hand-side correspond to the channel W^0 . The figure at the bottom left explains why the channel W^0 is a worse channel than W . In the channel W^0 , what u_0 sees is intuitively the sum (XOR) of the two observations y_0 and y_1 . Each of these observations correspond to one usage of W . The fact that these observations are summed means that they mutually affect each other as some kind of a noise, making the final result worse than each individual observation. The figures on the right hand-side correspond to the channel W^1 . The figure at the bottom right explains why the channel W^1 is a better channel than W . In the channel W^1 , the (independent) observations y_0 and y_1 are separately given to u_1 . Hence, u_1 has two independent observations of itself each of which is equivalent to one usage of W . Thus, W^1 is better than W .

two independent observations of itself. So the channel W^1 is a better channel than W .

Finally, let us point out that by applying the chain rule for mutual information one can show that this transform preserves capacity [1]

$$I(W^0) + I(W^1) = 2I(W), \quad (2.22)$$

and regarding the Bhattacharyya parameter, we have

$$Z(W^1) = Z(W)^2, \quad (2.23)$$

$$Z(W) \leq Z(W^0) \leq 1 - (1 - Z(W))^2, \quad (2.24)$$

Step 2 (Infinite binary tree): Consider an infinite binary tree with the root node placed at the top. In this tree each vertex has 2 children and there are 2^n vertices at level n . Assume that we label these vertices from left to right from 1

to 2^n . Here, we intend to assign to each vertex of the tree a BMS channel. We do this by a recursive procedure. Assign to the root node the channel W itself. Now consider the channel splitting transform $W \rightarrow (W^0, W^1)$ and from left to right, assign W^0 to W^1 to the children of the root node. In general, if Q is the channel that is assigned to vertex v , we assign Q^0 to Q^1 , from left to right respectively, to the children of the node v . In this way, we recursively assign a channel to all the vertices of the tree. Figure 2.5 shows the first 2 levels of the

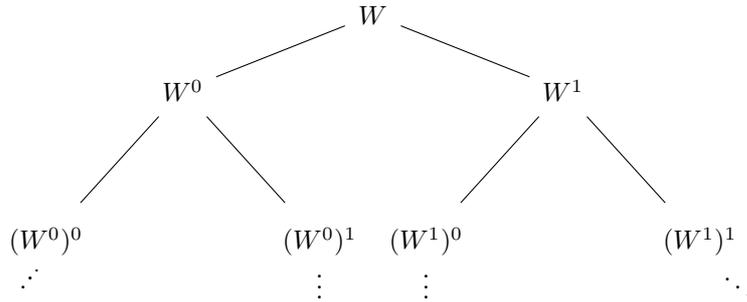


Figure 2.5: The infinite binary tree and the channels assigned to it.

binary tree. Let $W_{2^n}^{(i)}$ denote the channel that is assigned to vertex with label i at level n of the tree, $1 \leq i \leq 2^n$. As a result, one can equivalently relate the channel $W_{2^n}^{(i)}$ to W via the following procedure: let the 2-ary representation of $i - 1$ be $b_1 b_2 \cdots b_n$, where b_1 is the most significant digit. Then we have

$$W_{2^n}^{(i)} = (((W^{b_1})^{b_2}) \cdots)^{b_n}.$$

As an example, assuming $i = 7$, $n = 3$ we have $W_8^{(7)} = ((W^1)^1)^0$.

Let us recall from above that for the channels W^0 and W^1 defined in (2.13) and (2.14) there is an interesting interpretation in terms of successively decoding the bits u_0 and u_1 . As we explain very briefly here, one can generalize this interpretation to the channels $\{W_{2^n}^{(i)}\}_{1 \leq i \leq 2^n}$. We have $N = 2^n$ input bits u_0, u_1, \dots, u_{N-1} with i.i.d. distribution $\text{Bernoulli}(\frac{1}{2})$ which are combined using the transform

$$(x_0, x_1, \dots, x_{N-1}) = (u_0, u_1, \dots, u_{N-1}) G_2^{\otimes n}, \quad (2.25)$$

where G_2 is given in (2.25) and \otimes denotes kronecker power. The resulting bits x_0, x_1, \dots, x_{N-1} are then transmitted through N independent copies of W to form the output values y_0, y_1, \dots, y_{N-1} . Now, there is a pre-specified order on the bits u_0, u_1, \dots, u_{N-1} which we denote by $u_{i_0}, u_{i_1}, \dots, u_{i_{N-1}}$ such that the channel $W_N^{(j)}$ is precisely the channel that the bits u_{i_j} sees given the output y_0, \dots, y_{N-1} and previous bits $u_{i_0}, u_{i_1}, \dots, u_{i_{j-1}}$. Now, consider successively decoding the ordered bits $u_{i_0}, u_{i_1}, \dots, u_{i_{N-1}}$. In this regard, define the even

E_j by

E_j = the bit u_{i_j} is decoded incorrectly assuming that the previous bits $u_{i_0}, \dots, u_{i_{j-1}}$ are decoded correctly.

Then, we have

$$\Pr(E_j) = E(W_N^{(j)}). \quad (2.26)$$

Step 3 (Polarization property): The channels $\{W_{2^n}^{(i)}\}_{1 \leq i \leq 2^n}$ have the property that ([1]), as n grows large, a fraction close to $I(W)$ of the channels have capacity close to 1 (or Bhattacharyya parameter close to 0); and a fraction close to $1 - I(W)$ of the channels have capacity close to 0 (or Bhattacharyya parameter close to 1). In other words, as n grows large, the channels $\{W_{2^n}^{(i)}\}_{1 \leq i \leq 2^n}$ tend to become *polarized* to one of the following extremal situations: an almost perfect channel (capacity is very close to 1) or a very noisy channel (capacity is very close to 0). The basic idea behind polar codes is to use those channels that have capacity close to 1 (or equivalently have Bhattacharyya parameter close to 0) for information transmission. Before going further into the construction of polar codes using the phenomenon of channel polarization, we proceed by providing analytic grounds for justification of this phenomenon.

2.3.1 Polarization Process

Let $\{B_n, n \geq 1\}$ be a sequence of iid Bernoulli($\frac{1}{2}$) random variables. Denote by $(\mathcal{F}, \Omega, \Pr)$ the probability space generated by this sequence and let $(\mathcal{F}_n, \Omega_n, \Pr_n)$ be the probability space generated by (B_1, \dots, B_n) . For a BMS channel W , define a random sequence of channels $W_n, n \in \mathbb{N} \triangleq \{0, 1, 2, \dots\}$, as $W_0 = W$ and

$$W_n = \begin{cases} W_{n-1}^0 & \text{if } B_n = 0, \\ W_{n-1}^1 & \text{if } B_n = 1, \end{cases} \quad (2.27)$$

where the channels on the right side are given by the transform $W_{n-1} \rightarrow (W_{n-1}^0, W_{n-1}^1)$. Let us also define the random processes $\{H_n\}_{n \in \mathbb{N}}, \{I_n\}_{n \in \mathbb{N}}, \{Z_n\}_{n \in \mathbb{N}}$ and $\{E_n\}_{n \in \mathbb{N}}$ as $H_n = H(W_n), I_n = I(W_n) = 1 - H(W_n), Z_n = Z(W_n)$ and $E_n = E(W_n)$.

Example 2.1. *By a straightforward calculation one can show that for $W = \text{BEC}(z)$ we have*

$$W^0 = \text{BEC}(1 - (1 - z)^2) \quad (2.28)$$

$$W^1 = \text{BEC}(z^2). \quad (2.29)$$

Hence, when $W = \text{BEC}(z)$, the channel W_n is always a BEC. Furthermore, the processes H_n, I_n, Z_n and E_n admit simple closed form recursions as follows. We have $H_0 = z$ and for $n \geq 1$

$$H_n = \begin{cases} 1 - (1 - H_{n-1})^2, & \text{w.p. } \frac{1}{2} \\ H_{n-1}^2, & \text{w.p. } \frac{1}{2}. \end{cases} \quad (2.30)$$

Also, we have² $2E_n = H_n = 1 - I_n = Z_n$.

For channels other than the BEC, the channel W_n gets quite complicated in the sense that the cardinality of the output alphabet of the channel W_n is doubly exponential in n (or exponential in N). Thus, tracking the exact outcome of W_n seems to be a difficult task (for more detail see [?, 23]). Instead, as we will see in the sequel, one can prove many interesting properties regarding the processes H_n, Z_n and E_n .

Let us quickly review the limiting properties of the above mentioned processes [1, 7]. From (2.22) and (2.27), one can write for $n \geq 1$

$$\mathbb{E}[H(W_n) | W_{n-1}] \stackrel{(2.27)}{=} \frac{H(W_{n-1}^0) + H(W_{n-1}^1)}{2} \stackrel{(2.22)}{=} H(W_{n-1}). \quad (2.31)$$

Hence, the process H_n is a martingale. Furthermore, since H_n is also bounded (2.10), by Doob's martingale convergence theorems, the process H_n converges in \mathcal{L}^1 (and almost surely) to a limit random variable H_∞ . As the convergence is in \mathcal{L}^1 , as $n \rightarrow \infty$ we have

$$\mathbb{E}[|H_n - H_{n-1}|] = \mathbb{E}[|H(W_n^0) - H(W_n)|] \rightarrow 0.$$

As a result, we must have that $H(W_n^0) - H(W_n)$ converges to 0 almost surely (a.s.). We now claim that for a channel P , in order to have $H(P^0) = H(P)$ we must have $H(P) = 0$ (i.e., P is the noiseless channel) or $H(P) = 1$ (i.e., P is the completely noisy channel). By this claim and the fact that H_n converges a.s. to H_∞ , we conclude that H_∞ take its values in the set $\{0, 1\}$. Also, as $\mathbb{E}[H_n] = \mathbb{E}[H_\infty] = H(W)$, we obtain

$$H_\infty = \begin{cases} 0 & \text{w.p. } 1 - H(W), \\ 1 & \text{w.p. } H(W). \end{cases} \quad (2.32)$$

It remains to prove the claim mentioned above. We use the so called *extremes of information combining* inequalities [47]. Let P be an arbitrary BMS channel. To simplify notation, let $h = H(P)$ and also let $\epsilon \in [0, \frac{1}{2}]$ be such that $h_2(\epsilon) = H(P)$. We have

$$h \leq \overbrace{H(\text{BSC}(\epsilon)^0)}^{h_2(2\epsilon(1-\epsilon))} \leq H(P^0) \leq \overbrace{H(\text{BEC}(h)^0)}^{1-(1-h)^2}, \quad (2.33)$$

$$\underbrace{H(\text{BEC}(h)^1)}_{h^2} \leq H(P^1) \leq \underbrace{H(\text{BSC}(\epsilon)^1)}_{2h-h_2(2\epsilon(1-\epsilon))} \leq h. \quad (2.34)$$

Now, to prove the claim, assume that P is such that $H(P^0) = H(P)$. Using (2.33) we obtain $H(\text{BSC}(h)^0) = H(P)$ or equivalently $h_2(2\epsilon(1-\epsilon)) = h_2(\epsilon)$. As a result, ϵ must be a solution of the equation $\epsilon = 2\epsilon(1-\epsilon)$ which yields $\epsilon = 0, \frac{1}{2}$. Also, as $H(P) = h_2(\epsilon)$, then $H(P)$ can either be 0 or 1 and hence the claim is justified. Using the bounds (2.10)-(2.12) it is clear that the processes Z_n and E_n converge a.s. to H_∞ and $\frac{1}{2}H_\infty$, respectively.

²For the channel $W = \text{BEC}(z)$, it is easy to show that $2E(W) = H(W) = Z(W) = z$.

2.3.2 Polar Codes

Given the rate $R < I(W)$, polar coding is based on choosing a set of $2^n R$ rows of the matrix $G_n = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}^{\otimes n}$ to form a $2^n R \times 2^n$ matrix which is used as the generator matrix in the encoding procedure. The way this set is chosen is dependent on the channel W and is briefly explained as follows: At time $n \in \mathbb{N}$, consider a specific realization of the sequence (B_1, \dots, B_n) , and denote it by (b_1, \dots, b_n) . The random variable W_n outputs a BMS channel, according to the procedure (2.27), which we can naturally denote by $W^{(b_1, \dots, b_n)}$. Let us now identify a sequence (b_1, \dots, b_n) by an integer i in the set $\{1, \dots, N\}$ such that the binary expansion of $i-1$ is equal to the sequence (b_1, \dots, b_n) , with b_1 as the least significant bit. As an example for $n = 3$, we identify $(b_1, b_2, b_3) = (0, 0, 1)$ with 5 and $(b_1, b_2, b_3) = (1, 0, 0)$ with 2. To simplify notation, we use $W_n^{(i)}$ to denote $W^{(b_1, \dots, b_n)}$. Given the rate R , the indices of the matrix G_n are chosen as follows: Choose a subset of size NR from the set of channels $\{W_N^{(i)}\}_{1 \leq i \leq N}$ that have the least possible error probability (given in (2.9)) and choose the rows G_n with the same indices as these channels. E.g., if the channel $W_N^{(j)}$ is chosen, then the j -th row of G_n is selected. In the following, given N , we call the set of indices of NR channels with the least error probability, the set of good indices and denote it by $\mathcal{I}_{N,R}$. In the sequel, we will frequently use the term “the set of good indices” and $\mathcal{I}_{N,R}$ interchangeably.

It is proved in [1] that the block error probability of such polar coding scheme under SC decoding, denoted by $P_e(N, R)$, is bounded from both sides by³

$$\max_{i \in \mathcal{I}_{N,R}} E(W_N^{(i)}) \leq P_e(N, R) \leq \sum_{i \in \mathcal{I}_{N,R}} E(W_N^{(i)}). \quad (2.35)$$

We now briefly explain why such a code construction is reliable for any rate $R < I(W)$, provided that the block-length is large enough. Recall from Section 2.3.1 that the process $E_n = E(W_n)$ converges a.s. to a r.v. E_∞ such that $\mathbb{P}(E_\infty = 0) = 1 - H(W) = I(W)$. Hence, it is clear from the definition of the set good indices, $\mathcal{I}_{N,R}$, that the left side of (2.35) decays to 0 as n grows large. However, the story is not over yet since this is only a Lower bound on $P_e(N, R)$. Nonetheless, one can also show that the right side of (2.35) decays to 0. This was initially shown in [1] and later in [36] the authors showed that all of the three terms in (2.35) behave like $2^{-2^{\frac{n}{2}} + o(\sqrt{n})}$.

2.4 Polar Codes Based on $\ell \times \ell$ Matrices

One direct extension of polar codes is the usage of other matrices (kernels) rather than matrix G_2 . For an integer $\ell > 2$, let G be an $\ell \times \ell$ matrix. We now explain briefly how polar codes, based on the matrix G , are constructed. We

³Note here that by (2.9) the error probability of a BMS channel is less than its Bhattacharyya value. Hence, the right side of (2.35) is a better upper bound for the block error probability than the sum of the Bhattacharyya values.

note here that in order to have the polarization phenomenon, it is necessary and sufficient that G has the following property [12]: none of its column permutations are upper-triangular. We call such a matrix G a polarizing matrix and throughout this thesis we always assume that G is polarizing. Also, for $\ell > 2$ there are several polarizing $\ell \times \ell$ matrices. For future convenience, we define the ensemble of polarizing matrices as follows.

Definition 2.1. *By the ensemble of polarizing ℓ -matrices, denoted by \mathcal{G}_ℓ , we mean the set of all the polarizing matrices of size $\ell \times \ell$ endowed with uniform probability.*

We proceed by explaining the phenomenon of channel polarization a polarizing kernel G . In short, for $n \in \mathbb{N}$ the method of channel polarization takes $N = \ell^n$ copies of a BMS channel W and combines them by using the kernel matrix G to make a new set of ℓ^n channels $\{W_{\ell^n}^{(i)}\}_{1 \leq i \leq \ell^n}$. As $n \rightarrow \infty$, the set $\{W_{\ell^n}^{(i)}\}_{1 \leq i \leq \ell^n}$ tends to have extremal properties. We explain in more detail the method of channel polarization through the following three steps.

Step 1 (Channel Splitting): Let \mathcal{W} denote the class of BMS channels. Let us define a channel transform $W \rightarrow (W^0, W^1, \dots, W^{\ell-1})$, called channel splitting,

that maps \mathcal{W} to $(\overbrace{\mathcal{W}, \mathcal{W}, \dots, \mathcal{W}}^\ell)$. In other words, channel splitting is a transform which takes a BMS channel W as input and outputs ℓ BMS channels W^j , $0 \leq j \leq \ell - 1$. The channels W^j are constructed using the channel W and matrix G , according to the following rule: Consider a random row vector $U_0^{\ell-1} = (U_0, \dots, U_{\ell-1})$ that is uniformly distributed over $\{0, 1\}^\ell$. Let $X_0^{\ell-1} = U_0^{\ell-1}G$, where the arithmetic is in $\text{GF}(2)$. Also, let $Y_0^{\ell-1}$ be the result of passing each component of $X_0^{\ell-1}$ through an independent copy of W (i.e., Y_i is the outcome of passing X_i through an independent copy of W). We thus define the channel between $U_0^{\ell-1}$ and $Y_0^{\ell-1}$ by the transition probabilities

$$W_\ell(y_0^{\ell-1} | u_0^{\ell-1}) \triangleq \prod_{i=0}^{\ell-1} W(y_i | x_i) = \prod_{i=0}^{\ell-1} W(y_i | (u_0^{\ell-1}G)_i). \quad (2.36)$$

The channel $W^j : \{0, 1\} \rightarrow \mathcal{Y}^\ell \times \{0, 1\}^j$ is defined as the BMS channel with input u_j , output $(y_0^{\ell-1}, u_0^{j-1})$ and transition probabilities

$$W^j(y_0^{\ell-1}, u_0^{j-1} | u_j) = \frac{1}{2^{\ell-1}} \sum_{u_{j+1}^{\ell-1}} W_\ell(y_0^{\ell-1} | u_0^{\ell-1}). \quad (2.37)$$

Here and hereafter, u_i^j denotes the subvector (u_i, \dots, u_j) . An intuitive explanation behind the definition of W^j is as follows: Pick uniformly at random one of the 2^ℓ possible realizations of the vector $U_0^{\ell-1}$ and let it be denoted by $u_0^{\ell-1} = (u_0, \dots, u_{\ell-1})$. Construct the vector $x_0^{\ell-1} = u_0^{\ell-1}G$ and send the ℓ components of $x_0^{\ell-1}$ through ℓ parallel (and independent) copies of the channel W and finally let $y_0^{\ell-1}$ denote the output (i.e., y_i is the result of passing x_i through an independent copy of W). It is easy to see that the channel between

$u_0^{\ell-1}$ and $y_0^{\ell-1}$ is precisely the channel $W_\ell(y_0^{\ell-1} | u_0^{\ell-1})$ defined in (2.36). Figure 2.6 gives a schematic representation of this channel. Thus, given the vector

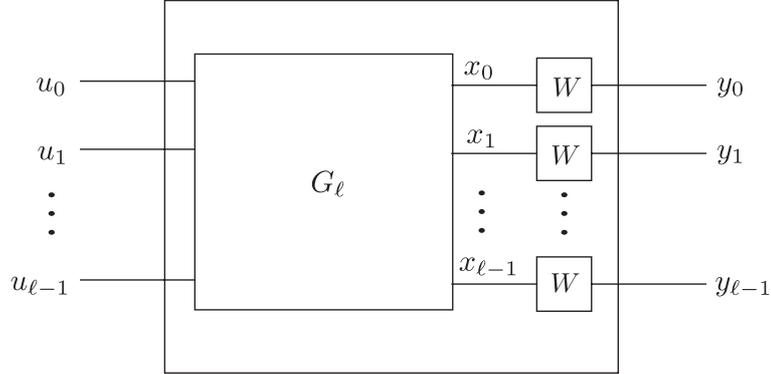


Figure 2.6: Schematic representation of the channel between the random vectors $U_0^{\ell-1}$ and $Y_0^{\ell-1}$.

$y_0^{\ell-1}$, the optimal way to infer about the value of $u_0^{\ell-1}$ is via ML decoding of $W_\ell(y_0^{\ell-1} | u_0^{\ell-1})$. Now, besides having access to $y_0^{\ell-1}$, assume that a genie also gives us the values of the bits u_0, \dots, u_{j-1} and asks us to decide on the value of u_j based on the observed vector $(y_0^{\ell-1}, u_0^{j-1})$. A little thought shows that the optimal way to do this is ML decoding of the values of u_j by using the transition probabilities $W^j(y_0^{\ell-1}, u_0^{j-1} | u_j)$ defined in (2.37). In other words, W^j is precisely the channel between u_j and $(y_0^{\ell-1}, u_0^{j-1})$ when we do not have any information about the value of $u_{j+1}, \dots, u_{\ell-1}$ (i.e., they are modeled as independent and identically-distributed (i.i.d.) random variables with a uniform distribution). Figure 2.7 gives a schematic representation of the channel W^j .

Finally, a noteworthy point to repeat is that the actual implementation of the channel splitting transform $W \rightarrow (W^0, W^1, \dots, W^{\ell-1})$ requires ℓ independent copies of W to generate $W^0, \dots, W^{\ell-1}$. Furthermore, by applying the chain rule for mutual information one can show that this transform preserves capacity [1], [12]

$$\sum_{j=0}^{\ell-1} I(W^j) = \ell I(W). \quad (2.38)$$

Step 2 (Infinite ℓ -ary tree): Consider an infinite ℓ -ary tree with the root node placed at the top. In this tree each vertex has ℓ children and there are ℓ^n vertices at level n . Assume that we label these vertices from left to right from 1 to ℓ^n . Here, we intend to assign to each vertex of the tree a BMS channel. We do this by a recursive procedure. Assign to the root node the channel W itself. Now consider the channel splitting transform $W \rightarrow (W^0, W^1, \dots, W^{\ell-1})$ and

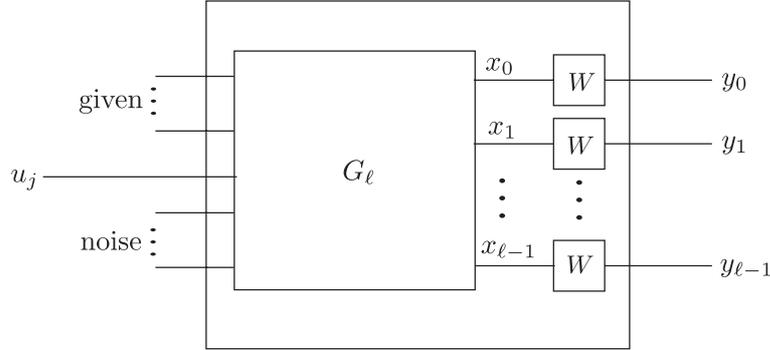


Figure 2.7: Schematic representation of the channel W^j . This is the channel that the bit u_j "sees" when the value of u_0, \dots, u_{j-1} together with $y_0^{\ell-1}$ is given as output and the bits $u_{j+1}, \dots, u_{\ell-1}$ are treated as noise (i.e., there is no information about their value and we assume their value is chosen uniformly at random).

from left to right, assign W^0 to $W^{\ell-1}$ to the children of the root node. In general, if Q is the channel that is assigned to vertex v , we assign Q^0 to $Q^{\ell-1}$, from left to right respectively, to the children of the node v . In this way, we recursively assign a channel to all the vertices of the tree. Let $W_{\ell^n}^{(i)}$ denote the channel that is assigned to vertex with label i at level n of the tree, $1 \leq i \leq \ell^n$. As a result, one can equivalently relate the channel $W_{\ell^n}^{(i)}$ to W via the following procedure: let the ℓ -ary representation of $i - 1$ be $b_1 b_2 \dots b_n$, where b_1 is the most significant digit. Then we have

$$W_{\ell^n}^{(i)} = (((W^{b_1})^{b_2}) \dots)^{b_n}.$$

As an example, assuming $i = 7$, $n = 3$ and $\ell = 2$ we have $W_8^{(7)} = ((W^1)^1)^0$.

Step 3 (Polarization property): The channels $\{W_{\ell^n}^{(i)}\}_{1 \leq i \leq \ell^n}$ have the property that ([1], [12]), as n grows large, a fraction close to $I(W)$ of the channels have capacity close to 1 (or Bhattacharyya parameter close to 0); and a fraction close to $1 - I(W)$ of the channels have capacity close to 0 (or Bhattacharyya parameter close to 1). In other words, as n grows large, the channels $\{W_{\ell^n}^{(i)}\}_{1 \leq i \leq \ell^n}$ tend to become polarized to one of the following extremal situations: an almost perfect channel (capacity is very close to 1) or a very noisy channel (capacity is very close to 0). The basic idea behind polar codes is then to use those channels that have capacity close to 1 (or equivalently have Bhattacharyya parameter close to 0) for information transmission. Accordingly, given the rate $R < I(W)$ and block-length $N = \ell^n$, the rows of the generator matrix of a polar code of block-length N correspond to a subset of the rows of the matrix $G^{\otimes n}$ whose indices are chosen with the following rule: choose a subset of size NR of the channels $\{W_{\ell^n}^{(i)}\}_{1 \leq i \leq \ell^n}$ with the least values for the Bhattacharyya

parameter and choose the rows $G^{\otimes n}$ with the indices corresponding to those of the channels. For example, if the channel $W_{\ell^n}^{(i)}$ is chosen, then the j th row of $G^{\otimes n}$ is selected, where the ℓ -ary representation of $j - 1$ is the digit-reversed version of that of $i - 1$. We decode using a successive cancellation (SC) decoder. This algorithm decodes the bits one-by-one in a prescribed order that is closely related to how the row indices of $G^{\otimes n}$ are chosen.

Scaling Laws for the Un-Polarized Channels

3

3.1 Problem Formulation

As we have seen in the previous chapter, the process Z_n polarizes in the sense that it converges a.s. to a $\{0, 1\}$ valued random variable Z_∞ . In this chapter¹, we investigate the dynamics of polarization. We begin by noting that at each time n there still exists a (small and in n vanishing) probability that the random variable Z_n takes a value far away from the endpoints of the unit interval (i.e., 0 and 1). Our primary objective is to study these small probabilities. More concretely, let $0 < a < b < 1$ be constants and consider the quantity $\Pr(Z_n \in [a, b])$. This quantity represents the fraction of sub-channels that are still un-polarized at time n . An important question is how fast the quantity $\Pr(Z_n \in [a, b])$ decays to zero. This question is intimately related to measuring the limiting properties of the sequence $\{\frac{1}{n} \log \Pr(Z_n \in [a, b])\}_{n \in \mathbb{N}}$.

Example 3.1. Assume $W = \text{BEC}(z)$. In this case the process Z_n has a simple closed form recursion as $Z_0 = z$ and

$$Z_{n+1} = \begin{cases} Z_n^2, & w.p. \frac{1}{2}, \\ 1 - (1 - Z_n)^2, & w.p. \frac{1}{2}. \end{cases} \quad (3.1)$$

Hence, it is straightforward to compute the value $\Pr(Z_n \in [a, b])$ numerically. Let $a = 1 - b = 0.1$. Figure 3.1 shows the value $\frac{1}{n} \log(\Pr(Z_n \in [a, b]))$ in terms of n for $z = 0.5, 0.6, 0.7$. This figure suggests that the sequence $\{\frac{1}{n} \log \Pr(Z_n \in [a, b])\}$ converges to a limiting value that is somewhere between -0.27 and -0.28 . Note that for different values of z , the limiting values are very close to each other.

¹The material of this chapter is based on [15], [16] and [17].

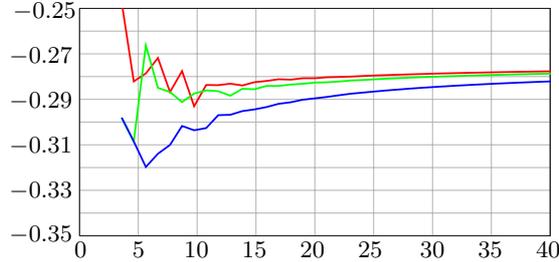


Figure 3.1: The value of $\frac{1}{n} \log(\Pr(Z_n \in [a, b]))$ versus n for $a = 1 - b = 0.1$ when W is a BEC with erasure probability $z = 0.5$ (top curve), $z = 0.6$ (middle curve) and $z = 0.7$ (bottom curve).

For other BMS channels, the process Z_n does not have a simple closed form recursion as for the BEC, and hence we need to use approximation methods (for more details see [23, 25]). Using these methods, we have plotted in Figure 3.2 the value of $\Pr(Z_n \in [a, b])$ ($a = 1 - b = 0.1$) for the channel families $\text{BSC}(\epsilon)$, and $\text{BAWGN}(\sigma)$ with different parameter values. The above numerical evi-

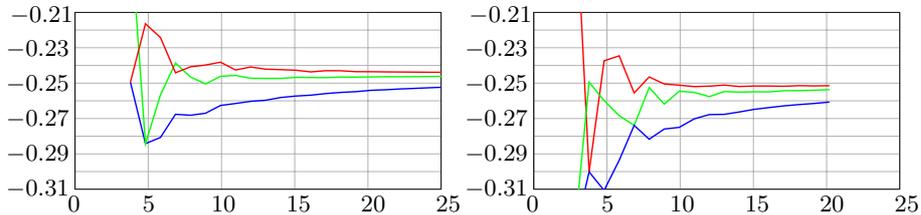


Figure 3.2: *Left figure:* The value of $\frac{1}{n} \log \Pr(Z_n \in [a, b])$ versus n for $a = 1 - b = 0.1$ and W being a BSC with cross-over probability $\epsilon = 0.11, 0.146, 0.189$. These BSC channels have capacity 0.5, 0.4 and 0.3, respectively. *Right figure:* the value of $\frac{1}{n} \log \Pr(Z_n \in [a, b])$ versus n for $a = 1 - b = 0.1$ and W is a BAWGN with noise variance $\sigma = 0.978$ (top curve), $\sigma = 1.149$ (middle curve), and $\sigma = 1.386$ (bottom curve). These BAWGN channels have capacities 0.5, 0.4 and 0.3, respectively.

dence suggests that the quantity $\Pr(Z_n \in [a, b])$ decays to zero exponentially fast in n . Further, we observe that the limiting value of this sequence is dependent on the starting channel W (e.g., from the figures it is clear that the channels BEC, BSC and BAWGN have different limiting values). Let us now be concrete and rephrase the above speculations as follows.

Question 1. *Does the quantity $\Pr(Z_n \in [a, b])$ decay exponentially in n ? If yes, what is the limiting value of $\frac{1}{n} \log \Pr(Z_n \in [a, b])$ and how is this limit*

related to the starting channel W and the choice of a and b ?

From Figures 3.1 and 3.2, we observe that the value of $\frac{1}{n} \log \Pr(Z_n \in [a, b])$ is the least when W is a BEC and this suggests that the channel BEC polarizes faster than the other BMS channels. This is intuitively justified as follows: Fix a value $z \in (0, 1)$ and assume that W is a BMS channel with Bhattacharyya parameter $Z(W) = z$. Now, consider the values $Z(W^0)$ and $Z(W^1)$. Using relations (2.24) and (2.23), it is clear that the values $Z(W^0)$ and $Z(W^1)$ are closest to the end points of the unit interval if W is a BEC. In other words, at the channel splitting transform, the channel BEC(z) polarizes faster than the other BMS channels.

Question 2. *For which set of channels does the quantity $\Pr(Z_n \in [a, b])$ decay the fastest or the slowest?*

Let us now be more ambitious and aim for our ultimate goal.

Question 3. *Can we characterize the exact behavior of $\Pr(Z_n \in [a, b])$ as a function of n , a , b and W ?*

Finally, we ask how the answers to the above questions will guide us through the understanding of the finite-length scaling behavior of polar codes. An immediate relation stems from the fact that the quantity $\Pr(Z_n \in [a, b])$ indicates the portion of the sub-channels that have not polarized at time n . In particular, all the channels in this set have a large Bhattacharyya value and hence cannot be included in the set of good indices. Therefore, the maximum reliable rate that we can achieve is restricted by the portion of this yet un-polarized channels. Consequently, the answers to the above questions will be crucial in finding answers to the following question.

Question 4. *Fix the channel W and a target block error probability P_e . To have a polar code with error probability less than P_e , how does the required block-length N scale with the rate R ?*

Finding a suitable answer to the questions 1, 3, and 4 is an easier task when the channel W is a BEC. This is due to the simple closed form expression of the process Z_n given in (3.1). In the next section (Section 3.2), we provide heuristic methods that lead to suitable numerical answers to Questions 1 and 3 for the BEC. As we will see in the next section, such heuristic derivations are in excellent compliance with numerical experiments. Using such derivations, we also give an answer to Question 4 for the BEC. The heuristic results of Section 3.2 provide us then with a concrete path to analytically tackle the above questions. In Section 3.3 we provide analytical answers to Questions 1-4 for the BEC as well as other BMS channels. Proving the full picture of Section 3.2 is beyond what we achieve in Section 3.3, nevertheless, we provide close and useful bounds.

3.2 Heuristic Derivation for the BEC

3.2.1 Scaling Law Assumption

Throughout this section we assume that the channel W is the BEC(z) where $z \in [0, 1]$. To avoid cumbersome notation, let us define

$$p_n(z, a, b) = \Pr(Z_n \in [a, b]), \quad (3.2)$$

where Z_n is the Bhattacharyya process of the BEC(z). We start by noticing that by (3.1) the function $p_n(z, a, b)$ satisfies the following recursion

$$p_{n+1}(z, a, b) = \frac{p_n(z^2, a, b) + p_n(1 - (1 - z)^2, a, b)}{2}, \quad (3.3)$$

with

$$p_0(z, a, b) = \mathbb{1}_{\{z \in [a, b]\}}. \quad (3.4)$$

More generally, one can easily observe the following. Let $g : [0, 1] \rightarrow \mathbb{R}$ be an arbitrary bounded function. Define the functions $\{g_n\}_{n \in \mathbb{N}}$ as

$$g_n(z) = \mathbb{E}[g(Z_n)]. \quad (3.5)$$

Note here that in (3.5) the parameter z is the starting point of the process Z_n , i.e., $Z_0 = z$. The functions $\{g_n\}_{n \in \mathbb{N}}$ satisfy the following recursion for $n \in \mathbb{N}$

$$g_{n+1}(z) = \frac{g_n(z^2) + g_n(1 - (1 - z)^2)}{2}. \quad (3.6)$$

This observation motivates us to define the *polar operator*, call it T , as follows. Let \mathcal{B} be the space of bounded measurable functions over $[0, 1]$. The polar operator $T : \mathcal{B} \rightarrow \mathcal{B}$ maps a function $g \in \mathcal{B}$ to another function in \mathcal{B} in the following way

$$T(g) = \frac{g(z^2) + g(1 - (1 - z)^2)}{2}. \quad (3.7)$$

It is now clear that

$$\mathbb{E}[g(Z_n)] = T \circ T \circ \dots \circ T(g) \triangleq T^n(g). \quad (3.8)$$

In this new setting, our objective is to study the limiting behavior of the functions $T^n(g)$ when g is a simple function as in (3.4). This task is intimately related to studying the largest eigenvalues of the polar operator T and their corresponding eigenfunctions. In this regard, to keep things in a simple and manageable setting, we first consider finite-dimensional approximations of T . This is done by discretizing the unit interval into very small sub-intervals with the same length and by assuming that T operates on all the points of these sub-intervals in the same way. More concretely, consider a (large) number $L \in \mathbb{N}$ and let the numbers x_i , $i \in \{0, 1, \dots, L - 1\}$ be defined as $x_i = \frac{i}{L-1}$. Hence, the unit interval $[0, 1]$ can be thought of as the union of the small

sub-intervals $[x_i, x_{i+1}]$. Now, for simplicity assume that g is a (piece-wise) continuous function on $[0, 1]$. Intuitively, by assuming L to be large, we expect that the value of g is the same throughout each of the intervals $[x_i, x_{i+1}]$. Such an assumption seems also reasonable for the function $T(g)$ given in (3.7). We can approximate the function g as an L dimensional vector

$$g_L \approx [g(x_0), g(x_1), \dots, g(x_{L-1})]. \quad (3.9)$$

In this way, we expect that the function $T(g)$ can be well approximated by a matrix multiplication

$$T \approx g_L T_L, \quad (3.10)$$

where the $L \times L$ matrix T_L is defined as follows. Let $T_L(i, j)$ be an element of T_L in the i -th row and the j -th column. Define $T_L(1, 1) = T_L(L, L) = 1$ and for the other elements of T_L we let

$$T_L(i, j) = \begin{cases} \frac{1}{2}, & \text{if } j = \lfloor L(\frac{i}{L})^2 \rfloor, \\ \frac{1}{2}, & \text{if } j = \lceil L(1 - (1 - \frac{i}{L})^2) \rceil, \\ 0, & \text{o.w.} \end{cases} \quad (3.11)$$

As an example, the matrix T_L for $L = 10$ has the following form

$$T_{10} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{2} & 0 & 0 & 0 & \frac{1}{2} & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & 0 & 0 & \frac{1}{2} & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{2} & 0 & 0 & 0 & 0 & \frac{1}{2} & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{2} & 0 & 0 & 0 & \frac{1}{2} & 0 \\ 0 & 0 & 0 & 0 & 0 & \frac{1}{2} & 0 & 0 & 0 & \frac{1}{2} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

All the rows of T_L sum up to 1. Hence, an application of the Perron-Frobenius theorem [129] shows that the (absolute value of) eigenvalues of T_L are all inside the interval $[-1, +1]$. Also, it is easy to see that T_L has a trivial eigenvalue equal to $\lambda_0 = 1$ with two corresponding eigenvectors

$$\begin{aligned} v_0 &= (1, 0, \dots, 0), \\ v_1 &= (0, 0, \dots, 1). \end{aligned}$$

A little thought shows that the v_0 and v_1 correspond to the two extremal states of the polarization (i.e., the perfect channel and the useless channel). This can be justified by the fact that if we start from any initial vector e_p that has value one at position p and value zero elsewhere, then

$$e_p T_L^n \xrightarrow{n \rightarrow \infty} c_0 v_0 + c_1 v_1,$$

L	1000	2000	4000	8000
$\lambda_2(L)$	0.8227	0.8240	0.8248	0.8253
$\lambda_3(L)$	0.6878	0.6958	0.7012	0.7046

Table 3.1: Values of $\lambda_2(L)$ and $\lambda_3(L)$, which correspond to the second and third largest eigenvalues of T_L (in absolute value), are computed numerically for different values of L .

where c_0 and c_1 are positive constants. This is just a rough observation of the polarization phenomenon. In fact, by polarization one can easily guess the following. Assuming $p = zL$, we have

$$c_0 \xrightarrow{L \rightarrow \infty} 1 - z,$$

$$c_1 \xrightarrow{L \rightarrow \infty} z.$$

However, we are interested in finding out how fast such a convergence is taking place. For this purpose, we look at the second and third largest eigenvalues (in absolute value) of T_L as L grows large. We denote the second largest eigenvalue of T_L by $\lambda_2(L)$ and the third largest is denoted by $\lambda_3(L)$. Table 3.1 contains the value of these eigenvalues computed numerically for several (large) values of L . It can thus be conjectured that

$$\lim_{L \rightarrow \infty} \lambda_2(L) \approx 0.826, \quad (3.12)$$

$$\lim_{L \rightarrow \infty} \lambda_3(L) \approx 0.705. \quad (3.13)$$

This belief guides us to conclude that for L growing large, if we start from any vector g which is not a multiple of the eigenvectors of T_L , then

$$gT_L^n \approx c_0 v_0 + c_1 v_1 + c_2 \lambda_2^n v_2 + O(n \lambda_3^n). \quad (3.14)$$

The above approximate relation indicates that for large L , the distance of gT_L^n from the limiting value is roughly equal to $c_2 \lambda_2^n$.

Now, let us go back the original polar operator T defined in (3.7). As we argued above, the operators T_L , for L large, are good finite-dimensional approximations of T . The (experimental) relation (3.14) brings us to the following assumption about T .

Assumption 3.1 (Scaling Assumption). *There exists $\mu \in (0, \infty)$ such that, for any $z, a, b \in (0, 1)$ such that $a < b$, the limit $\lim_{n \rightarrow \infty} 2^{\frac{n}{\mu}} p_n(z, a, b)$ exists in $(0, \infty)$. We denote this limit by $p(z, a, b)$. In other words,*

$$\lim_{n \rightarrow \infty} 2^{\frac{n}{\mu}} \Pr(Z_n \in [a, b]) = p(z, a, b). \quad (3.15)$$

We call the value μ the scaling exponent of polar codes for the BEC.

Note here that by (3.12) we expect that

$$2^{-\frac{1}{\mu}} = \lim_{L \rightarrow \infty} \lambda_2(L) \approx 0.826 \Rightarrow \frac{1}{\mu} \approx 0.275. \quad (3.16)$$

Let us now describe a numerical method for computing μ and $p(a, b, z)$. In this regard, we follow the approach of [22]. First we note that by (3.3) and the scaling law assumption we conclude that

$$2^{-\frac{1}{\mu}}p(z, a, b) = \frac{p(z^2, a, b) + p(1 - (1 - z)^2, a, b)}{2}. \quad (3.17)$$

Equation (3.17) can be solved numerically by recursion. First of all, note that the equation is invariant under multiplicative scaling of p . Also, from the equation one can naturally guess that $p(z, a, b)$ can be factorized into

$$p(z, a, b) = c(a, b)p(z), \quad (3.18)$$

where $p(z)$ is a solution of (3.17) with $p(\frac{1}{2}) = 1$. We iteratively compute μ and $p(z)$.

Initialize $p_0(z)$ –say– with $p_0(z) = 4z(1 - z)$ and compute recursively new estimates of $p_{n+1}(z)$ by first computing

$$\hat{p}_{n+1}(z) = p_n(z^2) + p_n(1 - (1 - z)^2),$$

and then by normalizing $p_{n+1}(z) = \hat{p}_{n+1}(z)/\hat{p}_{n+1}(\frac{1}{2})$, so that $p_{n+1}(\frac{1}{2}) = 1$. We have implemented the above functional recursion numerically by discretizing the z axis. Figure 3.3 shows the resulting numerical approximation of $p_\infty(z)$ as obtained by iterating the above procedure until $\|p_{n+1}(z) - p_n(z)\|_\infty \leq 10^{-10}$ ($\forall z \in [0, 1]$) and by using a discretization with 10^6 equi-spaced values of z . From this recursion we also get a numerical estimate of the scaling expo-

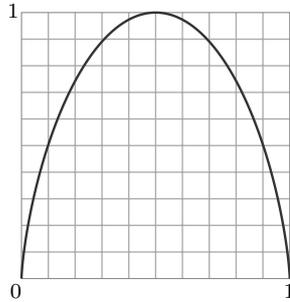


Figure 3.3: The function $p(z)$ for $z \in [0, 1]$.

nent μ . In particular, we expect $\hat{p}_n(1/2) \rightarrow 2^{\frac{1}{\mu}}$ as $n \rightarrow \infty$, or equivalently $2^{-\frac{1}{\mu}}\hat{p}_n(1/2) \rightarrow 1$. Using this method, we obtain the estimate

$$2^{-\frac{1}{\mu}} \approx 0.8260 \implies \frac{1}{\mu} \approx 0.2757. \quad (3.19)$$

As mentioned above, the function $p(a, b, z)$ differs from $p(z)$ by a multiplicative

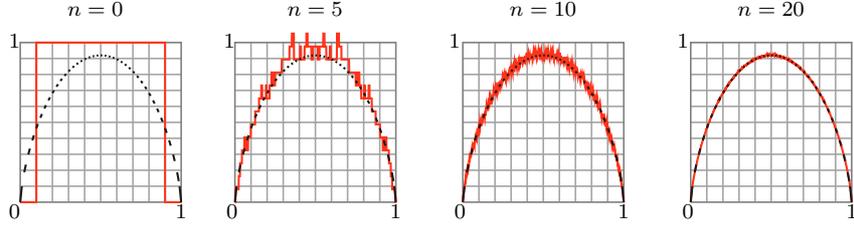


Figure 3.4: The functions $2^{\frac{n}{\mu}} p_n(a, b, z)$ for various values of n . Here we have fixed $a = 1 - b = 0.92$ and $\frac{1}{\mu} = 0.2757$. In all of the four plots, the dashed curve corresponds to $c(a, b)p(z)$ with $c(a, b) = 0.92$.

constant $c(a, b)$ that is to be found by other means. In Figure 3.4 we plot the functions $2^{\frac{n}{\mu}} p_n(z, a, b)$ for $a = 1 - b = \frac{1}{10}$ and different values of n . We observe that, as n increases these plots and the curve $c(a, b)p(z)$ with $c(a, b) = 0.92$ match very well. Even for moderate values of n (such as $n = 10$) we observe that the curves have a fairly good agreement.

Let us now see what the scaling law assumption implies about the finite-length behavior of polar codes. For simplicity, we assume that communication takes place on the BEC($\frac{1}{2}$). We are given a target error probability P_e and want to achieve a rate at least R . What block-length N should we choose?

Consider the process Z_n with $z = \frac{1}{2}$. It is easy to see that the set of possible values that Z_n takes in $[0, 1]$ is symmetric around $z = \frac{1}{2}$. Now, according to the scaling law for $x \in [0, \frac{1}{2}]$, there is a constant $p(\frac{1}{2}, x, \frac{1}{2}) \triangleq c(x)$ such that

$$\Pr(Z_n \in [x, \frac{1}{2}]) \approx c(x)2^{-\frac{n}{\mu}}, \quad (3.20)$$

As a result, noticing the fact that Z_n is symmetric around the point $z = \frac{1}{2}$ we get

$$\Pr(Z_n \in [0, x]) \leq \frac{1}{2} - c(x)2^{-\frac{n}{\mu}}. \quad (3.21)$$

From the construction procedure of polar codes (and specially relation (2.35)), we know the following. Let $z(1) \leq z(2) \leq \dots \leq z(N)$ be a re-ordering of the N possible values of Z_n in an ascending order. Then, the error probability of a polar code with rate R is bounded from below by

$$P_e \geq 1 - \sqrt{1 - z(N.R)^2} \geq \frac{z(N.R)^2}{2}. \quad (3.22)$$

So in order to achieve error probability P_e , we should certainly have $\frac{z(N.R)^2}{2} \leq P_e$ or $z(N.R) \leq \sqrt{2P_e}$. Hence, by using (3.21) we deduce that

$$R \leq \Pr(Z_n \in [0, \sqrt{2P_e}])$$

$$\begin{aligned} &\leq \frac{1}{2} - c(\sqrt{2P_e})2^{-\frac{n}{\mu}} \\ &= \frac{1}{2} - c(\sqrt{2P_e})N^{-\frac{1}{\mu}}, \end{aligned}$$

and finally,

$$N \geq \left(\frac{c(\sqrt{2P_e})}{\frac{1}{2} - R} \right)^\mu. \quad (3.23)$$

Now, from the above calculations we know that $\frac{1}{\mu} \approx 0.2757$ as a result for the channel $W = \text{BEC}(\frac{1}{2})$ we have

$$N \geq \Theta\left(\frac{1}{(I(W) - R)^{3.627}}\right). \quad (3.24)$$

In the next section, we provide methods that analytically validate the above observations. We also extend some of these observations to other BMS channels.

3.3 Analytical Approach: from Bounds for the BEC to Universal Bounds for BMS Channels

In this section we provide a rigorous basis for the observations that were derived in the previous section. Proving the full picture of Section 3.2 is beyond what we achieve here, but, we come up with close and useful bounds.

3.3.1 Characterization of μ for the BEC

We provide two approaches, that exploit different techniques, to compute the scaling exponent μ for the BEC. The first approach is based on a more careful look at equation (3.7). We observe that simple bounds can be derived on the largest nontrivial eigenvalue of the polar operator T by carefully analyzing the effect of T on some suitably chosen test functions. This approach provides us with a sequence of bounds on μ . We conjecture (and observe empirically) that these bounds indeed converge to the value of μ that is computed in Section 3.2. The second approach considers different compositions of the two operations z^2 and $2z - z^2$ and analyzes the asymptotic behavior of these compositions. This approach provides us with a close lower bound on μ .

First Approach

Consider the polar operator defined in (3.7). The objective here is to compute the largest eigenvalues of T . Specifically, we want to find the largest solutions of

$$T(f) = \lambda f. \quad (3.25)$$

A check shows that both $f(z) = z$ and $f(z) = 1$ are eigenfunctions associated to the eigenvalue $\lambda = 1$. Perhaps more interestingly, let us look at the eigenvalues

of T inside the interval $(0, 1)$. Intuitively, equation (3.17), together with the scaling law, can be reformulated as follows. The operator T has an eigenvalue $\lambda = 2^{-\frac{1}{\mu}}$ and a corresponding eigenfunction $p(z)$ such that if we take any step function $f(z) = \mathbb{1}_{\{z \in [a, b]\}}$, then

$$\lambda^{-n} T^n(f) \xrightarrow{n \rightarrow \infty} c(a, b)p(z). \quad (3.26)$$

In fact, if the scaling law is true, then we naturally expect that (3.26) holds for a much larger class of functions rather than the class of step functions. Heuristic arguments of the previous section also suggest that (3.26) holds for all (piece-wise) continuous functions $f(z)$ with $f(0) = f(1) = 0$.

Motivated by this picture, one approach to find bounds on the eigenvalue consists of the following two steps: (1) choose a suitable “test function” $f(z)$ for which we can provide good bounds on the behavior of $T^n(f)$ and (2) turn these bounds into bounds on the corresponding eigenvalue (or μ). With this in mind, for a generic test function $f(z) : [0, 1] \rightarrow [0, 1]$, let us define the sequence of functions $\{f_n(z)\}_{n \in \mathbb{N}}$ as $f_n : [0, 1] \rightarrow [0, 1]$ and for $z \in [0, 1]$,

$$f_n(z) \triangleq \mathbb{E}[f(Z_n)] = T^n(f). \quad (3.27)$$

Here, note that for $z \in [0, 1]$ the value of $f_n(z)$ is a deterministic value that is dependent on the process Z_n with the starting value $Z_0 = z$. Let us now recall once more the recursive relation of the functions f_n :

$$\begin{aligned} f_0(z) &= f(z), \\ f_n(z) &= \frac{f_{n-1}(z^2) + f_{n-1}(1 - (1 - z)^2)}{2}. \end{aligned} \quad (3.28)$$

In order to find lower and upper bounds on the speed of decay of the sequence f_n , we define sequences of numbers $\{a_m\}_{m \in \mathbb{N}}$ and $\{b_m\}_{m \in \mathbb{N}}$ as

$$a_m = \inf_{z \in [0, 1]} \frac{f_{m+1}(z)}{f_m(z)}, \quad (3.29)$$

$$b_m = \sup_{z \in [0, 1]} \frac{f_{m+1}(z)}{f_m(z)}. \quad (3.30)$$

Lemma 3.1. *Fix $m \in \mathbb{N}$. For all $n \geq m$ and $z \in [0, 1]$, we have*

$$(a_m)^{n-m} f_m(z) \leq f_n(z) \leq (b_m)^{n-m} f_m(z). \quad (3.31)$$

Furthermore, the sequence a_m is an increasing sequence and the sequence b_m is a decreasing sequence.

Proof. Here, we only prove the left-hand side of (3.31) and note that the right-hand side follows similarly. The proof goes by induction on $n - m$. For $n - m = 0$ the result is trivial. Assume that the relation (3.31) holds for a $n - m = k$, i.e., for $z \in [0, 1]$ we have

$$(a_m)^k f_m(z) \leq f_{m+k}(z). \quad (3.32)$$

We show that (3.31) is then true for $k + 1$ and $z \in [0, 1]$. We have

$$\begin{aligned}
 f_{m+k+1}(z) &\stackrel{(a)}{=} \frac{f_{m+k}(z^2) + f_{m+k}(1 - (1 - z)^2)}{2} \\
 &\stackrel{(b)}{\geq} \frac{(a_m)^k f_m(z^2) + (a_m)^k f_m(1 - (1 - z)^2)}{2} \\
 &= (a_m)^k f_{m+1}(z) \\
 &= (a_m)^k \frac{f_{m+1}(z)}{f_m(z)} f_m(z) \\
 &\geq (a_m)^k \left[\inf_{z \in [0,1]} \frac{f_{m+1}(z)}{f_m(z)} \right] f_m(z) \\
 &= (a_m)^{k+1} f_m(z).
 \end{aligned}$$

Here, (a) follows from (3.28) and (b) follows from the left-side inequality in (3.32), and hence the lemma is proved via induction. \square

Let us now begin searching for suitable test functions, i.e., candidates for $f(z)$ that provide us with good lower and upper bounds a_m and b_m . We expect that having a polynomial test function might be slightly preferable. This is due to the fact that if f is a polynomial, then $T^n(f)$ is also a polynomial and computing a_m and b_m is equivalent to finding roots of polynomials which is a manageable task. Of course the simplest polynomial that takes the value 0 on $z = 0, 1$ is $f_0(z) = z(1 - z)$. Hence, let us take our test function as $f(z) = f_0(z) = z(1 - z)$ and consider the corresponding sequence of functions $\{f_n(z)\}_{n \in \mathbb{N}}$,

$$f_n(z) = \mathbb{E}[Z_n(1 - Z_n)] = T^n(f_0). \quad (3.33)$$

A moment of thought shows that with $f_0 = z(1 - z)$ the function $2^n f_n$ is a polynomial of degree 2^{n+1} with integer coefficients. Let us first focus on computing the value of a_m for $m \in \mathbb{N}$. If the relation (3.26) holds true, then we expect that the value of a_m converges to $\lambda = 2^{-\frac{1}{\mu}}$ as m grows large.

Remark 3.1. *One can compute the value of a_m by finding the extreme points of the function $\frac{f_{m+1}}{f_m}$ (i.e., finding the roots of the polynomial $g_m = f'_{m+1}f_m - f_{m+1}f'_m$) and checking which one gives the global minimum. Assuming $f_0 = z(1 - z)$, for small values e.g., $m = 0, 1$, pen and paper suffice. For higher values of m , we can automatize the process: all these polynomials have rational coefficients and therefore it is possible to determine the number of real roots exactly and to determine their value to any desired precision. This task can be accomplished precisely by computing so-called Sturm chains (see Sturm's Theorem [18]). Computing Sturm chains is equivalent to running Euclid's algorithm starting with the second and third derivative of the original polynomial. Hence, we can find the value of a_m analytically to any desired precision. Table 3.2 contains the numerical value of a_m up to precision 10^{-4} for $m \leq 10$. As the table shows, the values a_m are increasing (see Lemma 3.1), and we conjecture*

that they converge to $2^{-0.2757} = 0.8260$, the corresponding value for the channel BEC (see (3.19)).

m	0	2	4	6	8	10
a_m	0.75	0.7897	0.8074	0.8190	0.8228	0.8239
$\log a_m$	-0.4150	-0.3406	-0.3086	-0.2880	-0.2813	-0.2794

Table 3.2: The values of a_m corresponding to the test function $f_0 = z(1-z)$ are numerically computed for several choices of m .

Let us now focus on computing the value of b_m . On the negative side, for the specific test function $f(z) = z(1-z)$ we obtain $b_m = 1$ for $m \in \mathbb{N}$ and therefore the upper bounds of (3.30) are of trivial use. In fact, it is not hard to show that if we plug in any polynomial as the test function then we get $b_m = 1$ for any m . On the positive side, we can consider other test functions that result in non-trivial values for b_m . The problem with non-polynomial functions is that methods such as the Sturm-chain method no longer apply here. Hence, finding the precise value of b_m up to a desired precision can be a difficult task and we lose the analytical tractability of b_m . As an example, choose

$$f_0(z) = z^\alpha(1-z)^\beta, \quad (3.34)$$

for some choice of $\alpha, \beta \in (0, 1)$. Then, from (3.30) we have

$$b_0 = \sup_{z \in [0,1]} \frac{f_1(z)}{f_0(z)} = \sup_{z \in [0,1]} \frac{z^\alpha(1+z)^\beta + (2-z)^\alpha(1-z)^\beta}{2}. \quad (3.35)$$

By letting $\alpha = \beta = \frac{2}{3}$, we numerically get $b_0 = 0.8312$ which is already a close bound for λ . This suggests that the test function $f_0(z) = f(z) = (z(1-z))^{\frac{2}{3}}$ is suitable candidate for obtaining good upper bounds b_m . For this specific test function, the value of b_m for various values of m has been numerically computed in Table 3.3. As we observe from Table 3.3, even for moderate values of m the (numerical) bound b_m is very close to the true “value” of λ .

m	0	2	4	6	8
b_m	0.8312	0.8294	0.8279	0.8268	0.8264
$\log b_m$	-0.2663	-0.2699	-0.2725	-0.2744	-0.2751

Table 3.3: The values of b_m corresponding to $f_0 = (z(1-z))^{\frac{2}{3}}$ are numerically computed for several choices of m .

Finally, let us relate the bounds a_m and b_m to bounds on the functions $p_n(a, b, z)$. We have

3.3. Analytical Approach: from Bounds for the BEC to Universal Bounds for BMS Channels 41

Lemma 3.2. Consider the test function $f(z) = z(1-z)$ and the corresponding sequence of function f_n defined in (3.28). Let $a, b \in (0, 1)$ be such that $\sqrt{a} \leq 1 - \sqrt{1-b}$. Then, there are constants $c_1, c_2 > 0$ such that for any $z \in (0, 1)$

$$\frac{1}{n} \log f_n(z) - \frac{c_1 \log n}{n} \leq \frac{1}{n} \log \Pr(Z_n \in [a, b]) \leq \frac{1}{n} \log f_n(z) + \frac{c_2}{n}. \quad (3.36)$$

Also, for the test function $f(z) = (z(1-z))^{\frac{2}{3}}$ and the corresponding sequence f_n , defined in (3.28), we have for $a, b \in (0, 1)$

$$\frac{1}{n} \log \Pr(Z_n \in [a, b]) \leq \frac{1}{n} \log f_n(z) + \frac{c_3}{n}, \quad (3.37)$$

where c_3 is a positive constant.

We can now easily conclude that

Corollary 3.1. Fix $m \in \mathbb{N}$. For $a, b \in [0, 1]$ such that $\sqrt{a} \leq 1 - \sqrt{1-b}$ and $n \leq m$ we have

$$\log a_m + O\left(\frac{\log n}{n}\right) \leq \frac{1}{n} \log \Pr(Z_n \in [a, b]) \leq \log b_m + O\left(\frac{1}{n}\right), \quad (3.38)$$

where a_m is defined in (3.29) with the test function $f(z) = z(1-z)$ (see Table 3.2), and b_m is defined in (3.92) with the test function $f(z) = (z(1-z))^{\frac{2}{3}}$ (see Table 3.3).

Remark 3.2. We expect that the result of Lemma 3.2 holds for any choice of a and b such that $a < b$. That is, the condition $\sqrt{a} \leq 1 - \sqrt{1-b}$ is not a serious condition and is just given to ease out the proof.

Second Approach:

Throughout this section we will prove the following theorem.

Theorem 3.1. We have

$$\liminf_{n \rightarrow \infty} \frac{1}{n} \log \left\{ \int_0^1 \Pr(Z_n \in [a, b]) dz \right\} \geq \frac{1}{2 \ln 2} - 1 \approx -0.2787. \quad (3.39)$$

Let us now explain, at the intuitive level, the main consequence of Theorem 3.1. By using the scaling law assumption, and specifically (3.14) and (3.15), we have that $\int_0^1 \Pr(Z_n \in [a, b]) dz \approx \int_0^1 2^{-\frac{a}{\mu}} p(z, a, b) dz + o(2^{-\frac{a}{\mu}})$. This relation together with (3.39) results that $\mu \geq \frac{1}{2 \ln 2} - 1 \approx -0.2787$. For the sake of brevity, we do not address here further (analytic) conclusions of Theorem 3.1 and we refer the reader to [15].

To proceed with the proof, let us recall from Section 2.3.1 the definition of Z_n (for the BEC) in terms of the sequence $\{B_n\}_{n \in \mathbb{N}}$. We start by $Z_0 = z$ and

$$Z_{n+1} = \begin{cases} Z_{n-1}^2 & ; \text{if } B_n = 1, \\ 2Z_{n-1} - Z_{n-1}^2 & ; \text{if } B_n = 0. \end{cases} \quad (3.40)$$

Hence, by considering the two maps $t_0, t_1 : [0, 1] \rightarrow [0, 1]$ defined as

$$t_0(z) = 2z - z^2, t_1(z) = z^2, \quad (3.41)$$

the value of Z_n is obtained by applying t_{B_n} on the value of Z_{n-1} , i.e.,

$$Z_n = t_{B_n}(Z_{n-1}). \quad (3.42)$$

The same rule applies for obtaining the value of Z_{n-1} from Z_{n-2} and so on. Thinking this through recursively, the value of Z_n is obtained from the starting point of the process, $Z_0 = z$, via the following (random) maps.

Definition 3.1. For each $n \in \mathbb{N}$ and a realization $(b_1, \dots, b_n) \triangleq \omega_n \in \Omega_n$ define the map ϕ_{ω_n} by

$$\phi_{\omega_n} = t_{b_n} \circ t_{b_{n-1}} \circ \dots \circ t_{b_1}. \quad (3.43)$$

Also, let Φ_n be the set of all such n -step maps.

As a result, an equivalent description of the process Z_n is as follows. At time n the value of Z_n is obtained by picking uniformly at random one of the functions $\phi_{\omega_n} \in \Phi_n$ and assigning the value $\phi_{\omega_n}(z)$ to Z_n . Consequently we have,

$$\Pr(Z_n \in [a, b]) = \sum_{\phi_{\omega_n} \in \Phi_n} \frac{1}{2^n} \mathbb{1}_{\{\phi_{\omega_n}(z) \in [a, b]\}}. \quad (3.44)$$

Using (3.44), it is apparent that in order to analyze the behavior of the quantity $\frac{1}{n} \log \Pr(Z_n \in [a, b])$ as n grows large, it is necessary to characterize the asymptotic behavior of the random maps ϕ_{ω_n} . Continuing the theme of Definition 3.1, we can assign to each realization of the infinite sequence $\{B_k\}_{k \in \mathbb{N}}$, denoted by $\{b_n\}_{n \in \mathbb{N}}$, a sequence of maps $\phi_{\omega_1}(z), \phi_{\omega_2}(z), \dots$, where $\omega_i \triangleq (b_1, \dots, b_i)$. We call the sequence $\{\phi_{\omega_k}\}_{k \in \mathbb{N}}$ the corresponding sequence of maps for the realization $\{b_k\}_{k \in \mathbb{N}}$. We also use the realization $\{b_k\}_{k \in \mathbb{N}}$ and its corresponding $\{\phi_{\omega_k}\}_{k \in \mathbb{N}}$ interchangeably. Let us now focus on the asymptotic characteristics of the functions ϕ_{ω_n} . Firstly, since $\{\phi_{\omega_n}(z)\}_{\omega_n \in \Omega_n}$ has the same law as Z_n starting at z , we conclude that for $z \in [0, 1]$ with probability one, the quantity $\lim_{k \rightarrow \infty} \phi_{\omega_k}(z)$ takes on a value in the set $\{0, 1\}$. In Figure 10.3 the functions ϕ_{ω_n} are plotted for a random realization. As it is apparent from the figure, the functions ϕ_{ω_n} seem to converge point-wise to a jump function (i.e., a sharp rise from 0 to 1). An intuitive justification of this fact is as follows. Consider a random function ϕ_{ω_n} . Due to polarization, as n grows large, almost all the values that this function takes are very close to 0 or 1. This function is also increasing and continuous (more precisely, it is a polynomial). A little thought reveals that the only choice to imagine for ϕ_{ω_n} is a very sharp rise from being almost 0 to almost 1. The formal and complete statement is given as follows.

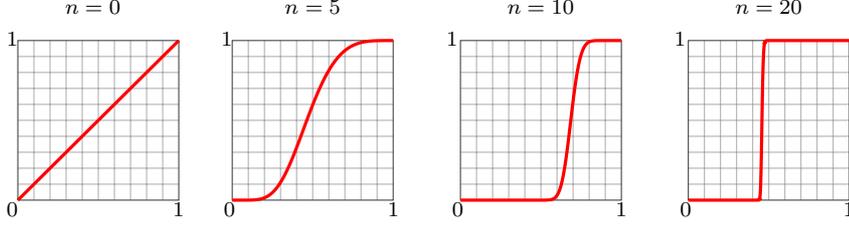


Figure 3.5: The functions ϕ_{ω_n} associated to a random realization are plotted. As we see as n grows large, the functions ϕ_{ω_n} converge point-wise to a step function.

Lemma 3.3 (Almost every realization has a threshold point). *For almost every realization of $\omega \triangleq \{b_k\}_{k \in \mathbb{N}} \in \Omega$, there exists a point $z_\omega^* \in [0, 1]$, such that*

$$\lim_{n \rightarrow \infty} \phi_{\omega_n}(z) \rightarrow \begin{cases} 0 & z \in [0, z_\omega^*) \\ 1 & z \in (z_\omega^*, 1] \end{cases}$$

Furthermore, z_ω^* has uniform distribution on $[0, 1]$. We call the point z_ω^* the threshold point of the realization $\{b_k\}_{k \in \mathbb{N}}$ or the threshold point of its corresponding sequence of maps $\{\phi_{\omega_k}\}_{k \in \mathbb{N}}$.

Looking more closely at (3.44), by the above lemma we conclude that as n grows large, the maps ϕ_{ω_n} that activate the identity function $\mathbb{1}_{\{\cdot\}}$ must have their threshold point sufficiently close to z . Let us now give an intuitive discussion about the idea behind the proof of Theorem 3.1. By using (3.44) we can write

$$\begin{aligned} \Pr(Z_n \in [a, b]) &= \sum_{\phi_{\omega_n} \in \Phi_n} \frac{1}{2^n} \mathbb{1}_{\{\phi_{\omega_n}(z) \in [a, b]\}} \\ &= \sum_{\phi_{\omega_n} \in \Phi_n} \frac{1}{2^n} \mathbb{1}_{\{z \in [\phi_{\omega_n}^{-1}(a), \phi_{\omega_n}^{-1}(b)]\}}. \end{aligned} \quad (3.45)$$

Hence by Lemma 3.3, for a large choice of n the intervals $[\phi_{\omega_n}^{-1}(a), \phi_{\omega_n}^{-1}(b)]$ have a very short length and are distributed almost uniformly along $[0, 1]$. Now, if we assume that the length of the intervals $[\phi_{\omega_n}^{-1}(a), \phi_{\omega_n}^{-1}(b)]$ is very close to their average, then we can replace the average in (3.45) by the average length of $[\phi_{\omega_n}^{-1}(a), \phi_{\omega_n}^{-1}(b)]$. That is,

$$\Pr(Z_n \in [a, b]) \approx \mathbb{E}[\phi_{\omega_n}^{-1}(b) - \phi_{\omega_n}^{-1}(a)].$$

So intuitively, all that remains is to compute the average length of the random intervals $[\phi_{\omega_n}^{-1}(a), \phi_{\omega_n}^{-1}(b)]$.

In fact we are not able to make all these heuristics precise for the point-wise values $\frac{1}{n} \log \Pr(Z_n \in [a, b])$. Nonetheless, the picture is naturally precise for

the average of $\Pr(Z_n \in [a, b])$ over $z \in [0, 1]$, i.e.,

$$\frac{1}{n} \log \left\{ \int_0^1 \Pr(Z_n \in [a, b]) dz \right\}. \quad (3.46)$$

To see this, we proceed as follows. By (3.45) we have

$$\begin{aligned} \int_0^1 \Pr(Z_n \in [a, b]) dz &= \int_0^1 \left\{ \sum_{\phi_{\omega_n}} \frac{1}{2^n} \mathbb{1}_{\{z \in \phi_{\omega_n}^{-1}[a, b]\}} \right\} dz \\ &= \sum_{\phi_{\omega_n}} \frac{1}{2^n} \left\{ \int_0^1 \mathbb{1}_{\{z \in \phi_{\omega_n}^{-1}[a, b]\}} dz \right\} \\ &= \mathbb{E}[\phi_{\omega_n}^{-1}(b) - \phi_{\omega_n}^{-1}(a)], \end{aligned}$$

and by applying $\frac{1}{n} \log(\cdot)$ to both sides we have

$$\begin{aligned} \frac{1}{n} \log \left\{ \int_0^1 \Pr(Z_n^z \in [a, b]) dz \right\} &= \frac{1}{n} \log \mathbb{E}[\phi_{\omega_n}^{-1}(b) - \phi_{\omega_n}^{-1}(a)] \\ &\geq \frac{1}{n} \mathbb{E}[\log(\phi_{\omega_n}^{-1}(b) - \phi_{\omega_n}^{-1}(a))], \end{aligned} \quad (3.47)$$

where in the last step we have used Jensen's inequality. The value of $\lim_{n \rightarrow \infty} \frac{1}{n} \mathbb{E}[\log(\phi_{\omega_n}^{-1}(b) - \phi_{\omega_n}^{-1}(a))]$ can be computed precisely.

Lemma 3.4. *We have*

$$\lim_{n \rightarrow \infty} \frac{1}{n} \mathbb{E}[\log(\phi_{\omega_n}^{-1}(b) - \phi_{\omega_n}^{-1}(a))] = \frac{1}{2 \ln 2} - 1 \approx -0.2787.$$

As a result, we have

$$\liminf_{n \rightarrow \infty} \frac{1}{n} \log \left\{ \int_0^1 \Pr(Z_n \in [a, b]) dz \right\} \geq \frac{1}{2 \ln 2} - 1.$$

The result of Theorem 3.1 provides a lower bound that is very close to the value we obtained in Section 3.2 but is not exactly equal. This is because we have used Jensen's inequality in (3.47).

3.3.2 Speed of Polarization for General BMS Channels

For a BMS channel W , there is no simple 1-dimensional recursion for the process Z_n as for the BEC. However, by using (2.24) and (2.23), we can provide bounds on how Z_n evolves:

$$Z_{n+1} \begin{cases} = Z_n^2 & ; \text{if } B_n = 1, \\ \in [Z_n \sqrt{2 - Z_n^2}, 2Z_n - Z_n^2] & ; \text{if } B_n = 0. \end{cases} \quad (3.48)$$

As a warm-up, we notice that similar techniques as used in Section 3.3.1 are applicable to provide general lower and upper bounds. For instance, to find upper bounds we can proceed as follows. For any non-negative function $g : [0, 1] \rightarrow \mathbb{R}^+$ such that $g(0) = g(1) = 0$ let

$$L_g = \sup_{z \in (0,1), y \in [z\sqrt{2-z^2}, z(2-z)]} \frac{g(z^2) + g(y)}{2g(z)}.$$

Similar to the discussion in Section 3.3.1, we can show that for the process $Z_n = Z(W_n)$ we have

$$\mathbb{E}[g(Z_n)] \leq cL_g^n, \tag{3.49}$$

where $c = \sup_{z \in [0,1]} g(z)$ is a constant. Hence, using the Markov inequality we have for $a, b \in (0, 1)$,

$$\frac{1}{n} \log \Pr(Z_n \in [a, b]) \leq \log L_g + O\left(\frac{1}{n}\right).$$

For example, assuming $g(z) = (z(1-z))^{\frac{2}{3}}$ we numerically obtain that $\log L_g = -0.169$. That is

$$\mathbb{E}[(Z_n(1-Z_n))^{\frac{2}{3}}] \leq 2^{-0.169n}, \tag{3.50}$$

and for $a, b \in (0, 1)$ we have

$$\frac{1}{n} \log \Pr(Z_n \in [a, b]) \leq -0.169 + O\left(\frac{1}{n}\right).$$

The relations of type (3.50) are upper bounds on the speed of polarization that hold *universally* over all the BMS channels. Let us now compute universal lower bounds. In the rest of this section, it is more convenient for us to consider another stochastic process related to W_n , which is the process² $H_n = H(W_n)$. The main reason to consider H_n rather than Z_n is that the process H_n is a martingale and this martingale property will help us to use the functions $\{f_n\}_{n \in \mathbb{N}}$ defined in (3.28) (with the starting function $f(z) = z(1-z)$) to provide universal lower bounds on the quantity $\mathbb{E}[H_n(1-H_n)]$. We begin by introducing one further technical condition given as follows.

Definition 3.2. *We call an integer $m \in \mathbb{N}$ suitable if the function $f_m(z)$, defined in (3.28) (with the starting function $f(z) = z(1-z)$), is concave on $[0, 1]$.*

Remark 3.3. *For small values of m , i.e., $m \leq 2$, it is easy to verify by hand that the function f_m is concave. As discussed previously, for larger values of m we can use Sturm's theorem [18] and a computer algebra system to verify this. Note that the polynomials $2^m f_m$ have integer coefficients. Hence, all the required computations can be done exactly. We have checked up to $m = 8$ that f_m is concave and we conjecture that in fact this is true for all $m \in \mathbb{N}$.*

²For the BEC the processes H_n and Z_n are identical.

We now show that for any BMS channel W , the value of a_m , defined in (3.29), is a lower bound on the speed of decay of H_n provided that m is a suitable integer.

Lemma 3.5. *Let $m \in \mathbb{N}$ be a suitable integer and W a BMS channel. We have for $n \geq m$*

$$\mathbb{E}[H_n(1 - H_n)] \geq (a_m)^{n-m} f_m(H(W)), \quad (3.51)$$

where a_m is given in (3.29).

Proof. We use induction on $n - m$: for $n - m = 0$ there is nothing to prove. Assume that the result of the lemma is correct for $n - m = k$. Hence, for any BMS channel W with $H_n = H(W_n)$ we have

$$\mathbb{E}[H_{m+k}(1 - H_{m+k})] \geq (a_m)^k f_m(H(W)). \quad (3.52)$$

We now prove the lemma for $m - n = k + 1$. For the BMS channel W , let us recall from Section 2.3 that the transform $W \rightarrow (W^0, W^1)$ yields two channels W^0 and W^1 such that (2.22) holds. Define the process $\{(W^0)_n, n \in \mathbb{N}\}$ as the channel process that starts with W^0 and evolves as in (2.27). We define $\{(W^1)_n, n \in \mathbb{N}\}$ similarly. Let us also define the two processes $H_n^0 = H((W^0)_n)$ and $H_n^1 = H((W^1)_n)$. We have,

$$\begin{aligned} & \mathbb{E}[H_{m+k+1}(1 - H_{m+k+1})] \\ & \stackrel{(a)}{=} \frac{\mathbb{E}[H_{m+k}^0(1 - H_{m+k}^0)] + \mathbb{E}[H_{m+k}^1(1 - H_{m+k}^1)]}{2} \\ & \stackrel{(b)}{\geq} (a_m)^k \frac{f_m(H(W^0)) + f_m(H(W^1))}{2} \\ & \stackrel{(c)}{\geq} (a_m)^k \frac{f_m(1 - (1 - H(W))^2) + f_m(H(W)^2)}{2} \\ & \stackrel{(d)}{=} (a_m)^k f_{m+1}(H(W)) \\ & = (a_m)^k \frac{f_{m+1}(H(W))}{f_m(H(W))} f_m(H(W)) \\ & \geq (a_m)^k \left[\inf_{h \in [0,1]} \frac{f_{m+1}(h)}{f_m(h)} \right] f_m(H(W)) \\ & \stackrel{(e)}{=} (a_m)^{m+1} f_m(H(W)). \end{aligned}$$

In the above chain of inequalities, relation (a) follows from the fact that W_m has 2^m possible values among which half of them are branched out from W^0 and the other half are branched out from W^1 . Relation (b) follows from the induction hypothesis given in (3.52). Relation (c) follows from (2.33), (2.34) and the fact that the function f_m is concave. More precisely, because f_m is concave on $[0, 1]$, we have the following inequality for any sequence of numbers

$0 \leq x' \leq x \leq y \leq y' \leq 1$ that satisfy $\frac{x+y}{2} = \frac{x'+y'}{2}$:

$$\frac{f_m(x') + f_m(y')}{2} \leq \frac{f_m(x) + f_m(y)}{2}. \quad (3.53)$$

In particular, we set $x' = H(W)^2$, $x = H(W^1)$, $y = H(W^0)$, $y' = 1 - (1 - H(W))^2$ and we know from (2.33) and (2.34) that $0 \leq x' \leq x \leq y \leq y' \leq 1$. Hence, by (3.53) we obtain (c). Relation (d) follows from the recursive definition of f_m given in (3.28). Finally, relation (e) follows from the definition of a_m given in (3.29). \square

Finally in the following two parts, we rigorously relate the results obtained in previous sections to finite-length performance of polar codes. In other words, answering Question 4 is the main focus for the remaining parts of this section.

3.3.3 Universal Bounds on the Scaling Behavior of Polar Codes

Universal Lower Bounds

Consider a BMS channel W and let us assume that a polar code with block-error probability at most a given value $P_e > 0$, is required. One way to accomplish this is to ensure that the right side of (2.35) is less than P_e . However, this is only a sufficient condition that might not be necessary. Hence, we call the right side of (2.35) *the strong reliability condition*. Numerical and analytical investigations (see [22] and [19]) suggest that once the sum of individual errors in the right side of (2.35) is less than 1, then it provides a fairly good estimate of P_e . In fact, the smaller the sum is the closer it is to P_e . Hence, the sum of individual errors can be considered as a fairly accurate proxy for P_e . Based on this measure of the block-error probability, we provide bounds on how the rate R scales in terms of the block-length N .

Theorem 3.2. *For any BMS channel W with capacity $I(W) \in (0, 1)$, there exist constants $P_e, \alpha > 0$, that depend only on $I(W)$, such that*

$$\sum_{i \in \mathcal{I}_{N,R}} E(W_N^{(i)}) \leq P_e, \quad (3.54)$$

implies

$$R < I(W) - \frac{\alpha}{N^{\frac{1}{\mu}}}, \quad (3.55)$$

where μ is a universal parameter lower bounded by 3.553.

Here, a few comments are in order:

(i) As we have seen above, we can obtain an increasing sequence of lower bounds, call this sequence $\{\mu_m\}_{m \in \mathbb{N}}$, for the universal parameter μ . For each m , in order to show the validity of the lower bound, we need to verify the concavity of a certain polynomial (defined in (3.28)) in $[0, 1]$. We explained in Remark 3.3 how we can accomplish this using the Sturm chain method. The

lower bound for μ stated in Theorem 3.2 is the one corresponding to $m = 8$, an arbitrary choice. If we increase m , we get e.g., $\mu_{16} = 3.614$. We conjecture that the sequence μ_m converges to $\mu = 3.627$, the parameter for the BEC. If such a conjecture holds, then the channel BEC polarizes the fastest among the BMS channels (see Question 9).

(ii) Let P_e, α, μ be as in Theorem 3.2. If we require the block-error probability to be less than P_e (in the sense that the condition (3.54) is fulfilled), then the block-length N should be at least

$$N > \left(\frac{\alpha}{I(W) - R} \right)^\mu. \quad (3.56)$$

(iii) From (2.1) we know that the value of μ for the random linear ensemble is $\mu = 2$, which is the optimal value since the variations of the channel itself require $\mu \geq 2$. Thus, given a rate R , reliable transmission by polar codes requires a larger block-length than the optimal value.

Proof of Theorem 3.2: To fit the bounds of Section 3.3.1 into the framework of Theorem 3.2, let us first introduce the sequence $\{\mu_m\}_{m \in \mathbb{N}}$ as

$$\mu_m = -\frac{1}{\log a_m}, \quad (3.57)$$

where a_m is defined in (3.29) with starting function $f(z) = z(1-z)$. In the previous section, we have proved that for a suitable m , the speed with which the quantity $\mathbb{E}[H_n(1-H_n)]$ decays is lower bounded by $a_m = 2^{-\frac{1}{\mu_m}}$, i.e. for $n \geq m$ we have $\mathbb{E}[H_n(1-H_n)] \geq 2^{-\frac{(n-m)}{\mu_m}} f_m(H(W))$. To relate the strong reliability condition in (3.54) to the rate bound in (3.55), we need the following lemma.

Lemma 3.6. *Consider a BMS channel W and assume that there exist positive real numbers γ, θ and $m \in \mathbb{N}$ such that $\mathbb{E}[H_n(1-H_n)] \geq \gamma 2^{-n\theta}$ for $n \geq m$. Let $\alpha, \beta \geq 0$ be such that $2\alpha + \beta = \gamma$, we have for $n \geq m$*

$$\Pr(H_n \leq \alpha 2^{-n\theta}) \leq I(W) - \beta 2^{-n\theta}. \quad (3.58)$$

Proof. The proof is by contradiction. Let us assume the contrary, i.e., we assume there exists $n \geq m$ s.t.,

$$\Pr(H_n \leq \alpha 2^{-n\theta}) > I(W) - \beta 2^{-n\theta}. \quad (3.59)$$

In the following, we show that with such an assumption we reach to a contradiction. We have

$$\begin{aligned} \mathbb{E}[H_n(1-H_n)] &= \mathbb{E}[H_n(1-H_n) \mid H_n \leq \alpha 2^{-n\theta}] \Pr(H_n \leq \alpha 2^{-n\theta}) \\ &\quad + \mathbb{E}[H_n(1-H_n) \mid H_n > \alpha 2^{-n\theta}] \Pr(H_n > \alpha 2^{-n\theta}). \end{aligned} \quad (3.60)$$

It is now easy to see that

$$\mathbb{E}[H_n(1 - H_n) \mid H_n \leq \alpha 2^{-n\theta}] \leq \alpha 2^{-n\theta},$$

and since $\mathbb{E}[H_n(1 - H_n)] \geq \gamma 2^{-n\theta}$, by using (3.60) we get

$$\mathbb{E}[H_n(1 - H_n) \mid H_n > \alpha 2^{-n\theta}] \Pr(H_n > \alpha 2^{-n\theta}) \geq 2^{-n\theta}(\gamma - \alpha). \quad (3.61)$$

We can further write

$$\begin{aligned} \mathbb{E}[(1 - H_n)] &= \mathbb{E}[1 - H_n \mid H_n \leq \alpha 2^{-n\theta}] \Pr(H_n \leq \alpha 2^{-n\theta}) \\ &\quad + \mathbb{E}[1 - H_n \mid H_n > \alpha 2^{-n\theta}] \Pr(H_n > \alpha 2^{-n\theta}), \end{aligned} \quad (3.62)$$

and noticing fact that $H_n \geq H_n(1 - H_n)$ we can plug (3.61) in (3.62) to obtain

$$\mathbb{E}[(1 - H_n)] \geq \mathbb{E}[1 - H_n \mid H_n \leq \alpha 2^{-n\theta}] \Pr(H_n \leq \alpha 2^{-n\theta}) + 2^{-n\theta}(\gamma - \alpha). \quad (3.63)$$

We now continue by using (3.59) in (3.63) to obtain

$$\begin{aligned} \mathbb{E}[(1 - H_n)] &> (I(W) - \beta 2^{-n\theta})(1 - \alpha 2^{-n\theta}) + 2^{-n\theta}(\gamma - \alpha) \\ &\geq I(W) + 2^{-n\theta}(\gamma - \alpha(1 + I(W)) - \beta), \end{aligned}$$

and since $2\alpha + \beta = \gamma$, we get $\mathbb{E}[1 - H_n] > I(W)$. This is a contradiction since H_n is a martingale and $\mathbb{E}[1 - H_n] = I(W)$. \square

Let us now use the result of Lemma 3.6 to conclude the proof of Theorem 3.2. By Lemma 3.5, we have for $n \geq m$

$$\mathbb{E}[H_n(1 - H_n)] \geq 2^{-\frac{(n-m)}{\mu_m}} f_m(H(W)).$$

Thus, if we now let $\gamma = 2^{\frac{m}{\mu_m}} f_m(H(W))$ and $2\alpha = \beta = \frac{\gamma}{2}$, then by using Lemma 3.6 we obtain

$$\Pr(H_n \leq \frac{\gamma}{4} 2^{-\frac{n}{\mu_m}}) \leq I(W) - \frac{\gamma}{2} 2^{-\frac{n}{\mu_m}}. \quad (3.64)$$

Assume that we desire to achieve a rate R equal to

$$R = I(W) - \frac{\gamma}{4} 2^{-\frac{n}{\mu_m}}. \quad (3.65)$$

Let $\mathcal{I}_{N,R}$ be the set of indices chosen for such a rate R , i.e., $\mathcal{I}_{N,R}$ includes the $2^n R$ indices of the sub-channels with the least value of error probability. Define the set A as

$$A = \{i \in \mathcal{I}_{N,R} : H(W_N^{(i)}) \geq \frac{\gamma}{4} 2^{-\frac{n}{\mu_m}}\}. \quad (3.66)$$

In this regard, note that (3.64) and (3.65) imply that $|A| \geq \frac{\gamma}{4} 2^{n(1-\frac{1}{\mu_m})}$. As a result, by using (2.10) and (2.11) we obtain

$$\begin{aligned} \sum_{i \in \mathcal{I}_{N,R}} E(W_N^{(i)}) &\geq \sum_{i \in A} E(W_N^{(i)}) \geq \frac{\gamma^2}{16} 2^{n(1-\frac{1}{\mu_m})} h_2^{-1}(2^{-\frac{n}{\mu_m}}) \\ &\geq \frac{\gamma^2}{16} \frac{2^{n(1-2\frac{1}{\mu_m})}}{8n \frac{1}{\mu_m}}, \end{aligned} \quad (3.67)$$

where the last step follows from the fact that for $x \in [0, \frac{1}{\sqrt{2}}]$, we have $h_2^{-1}(x) \geq \frac{x}{8 \log(\frac{1}{x})}$. Thus, having a block-length $N = 2^n$, in order to have error probability (measured by (2.35)) less than $\frac{\gamma^2}{16} \frac{2^{n(1-2\frac{1}{\mu_m})}}{8n \frac{1}{\mu_m}}$, the rate can be at most $I(W) - \frac{\gamma}{4} 2^{-\frac{n}{\mu_m}}$.

Finally, if we let $m = 8$ (by the discussion in Remark 3.3, we know that $m = 8$ is suitable), then $\mu_8 = \frac{1}{-\log(a_8)} = 3.553$ and choosing

$$P_e = \inf_{n \in \mathbb{N}} \left[\sum_{i \in \mathcal{I}_{N,R}} E(W_N^{(i)}) \right], \quad (3.68)$$

where R is given in (3.65), then it is easy to see from (3.67) that $P_e > 0$ (since $\frac{1}{\mu_8} < \frac{1}{2}$) and furthermore, to have block-error probability less than P_e the rate should be less than R given in (3.65).

Universal Upper Bounds

In this part, we provide upper bounds on the required block-length of Question 4. Again, the key observation here is the upper-bounds on the speed of polarization, e.g. the bounds derived in Table 3.3 for the BEC and the universal bound (3.50).

Theorem 3.3. *Let $Z_n = Z(W_n)$ be the Bhattacharyya process associated to a BMS channel W . Assume that for $n \in \mathbb{N}$ we have*

$$\mathbb{E}[(Z_n(1 - Z_n))^\alpha] \leq \beta 2^{-\rho n}, \quad (3.69)$$

where α, β, ρ are positive constants and $\alpha < 1$. Then, the block-length N required to achieve an error probability $P_e > 0$ at a given rate $R < I(W)$ is bounded above by

$$\log N \leq \left(1 + \frac{1}{\rho}\right) \log \frac{1}{d} + c_4 \left(\log \left(\log \frac{3}{d}\right)\right)^2 + c_5 \log \left(\log \left(\frac{2}{P_e}\right)\right) \log \left(\log \frac{3}{d}\right), \quad (3.70)$$

where $d = I(W) - R$ and c_4, c_5 are positive constants that depend on α, β, ρ .

Before proceeding with the proof of Theorem 3.3, let us note a few comments:

(i) In the previous sections we have computed several candidates for the value ρ required in Theorem 3.3. As an example, using the universal candidate for ρ obtained in (3.50) (i.e., $\rho = 0.169$), we obtain the following corollary.

3.3. Analytical Approach: from Bounds for the BEC to Universal Bounds for BMS Channels 51

Corollary 3.2. *For any BMS channel W , the block-length N required to achieve a rate $R < I(W)$ scales at most as*

$$N \leq \Theta\left(\frac{1}{(I(W) - R)^7}\right). \quad (3.71)$$

One important consequence of this corollary is that polar codes require a block-length that scales polynomially in terms of the gap to capacity. (ii) As we will see in the proof of Theorem 3.3, the result of this theorem is also valid if we replace P_e with the sum of Bhattacharyya values of the channels that correspond to the good indices (this sum is indeed an upper bound for P_e).

Proof of Theorem 3.3: Throughout the proof we will be using two key lemmas (Lemma 3.8 and Lemma 3.9) that are stated in the appendices. Let

$$d = I(W) - R. \quad (3.72)$$

We define $n_0 \in \mathbb{N}$ to be

$$n_0 = \left\lceil \frac{1}{\rho} \log \frac{3(1+c_1)(1+2c_2c_3)}{d} \right\rceil, \quad (3.73)$$

where the constants c_1 , c_2 and c_3 are given in Lemmas 3.8, 3.9 and 3.10, respectively. As a result of Lemma 3.8 and (3.73), we have for $n \geq n_0$

$$\Pr(Z_n \leq \frac{1}{2}) \geq R + \frac{2}{3}d. \quad (3.74)$$

We also define the set \mathcal{A} as follows. Let $N_0 = 2^{n_0}$ and

$$\mathcal{A} = \left\{ i \in \{0, \dots, N_0 - 1\} : Z(W_{N_0}^{(i)}) \leq \frac{1}{2} \right\}. \quad (3.75)$$

In other words \mathcal{A} is the set of indices at level n_0 of the corresponding infinite binary tree of W (see Section 2.3) whose Bhattacharyya parameter is not so large. Also, from (3.74) the set \mathcal{A} contains more than a fraction R of all the sub-channels at level n_0 . The idea is then to go further down through the infinite binary tree at a level $n_0 + n_1$ (the value of n_1 will be specified shortly). We then observe that the sub-channels at level $n_0 + n_1$ that are branched out from the set \mathcal{A} are polarized to a great extent in the sense that sum of their Bhattacharyya parameters is below P_e (see Figure 3.6 for a schematic illustration of the idea).

We proceed by finding a suitable candidate for n_1 . Our objective is to choose n_1 large enough s.t. there is a set of indices at level $n_0 + n_1$ with the following properties: (i) sum of the Bhattacharyya parameters of the sub-channels in this set is less than P_e and (ii) the cardinality of this set is at least $R2^{n_0+n_1}$. In what follows, we will first use the hypothesis of Lemma 3.9 to give a candidate for n_1 and then we make it clear that such a candidate is suitable for our needs. Let $\{B_m\}_{m \in \mathbb{N}}$ be a sequence of iid Bernoulli($\frac{1}{2}$) random variables. We let n_1 be the smallest integer such that the following holds

$$\Pr(2^{-2^{\sum_{i=1}^{n_1} B_i}} \leq \frac{P_e}{2^{n_0+n_1}}) \geq 1 - \frac{d}{3}. \quad (3.76)$$

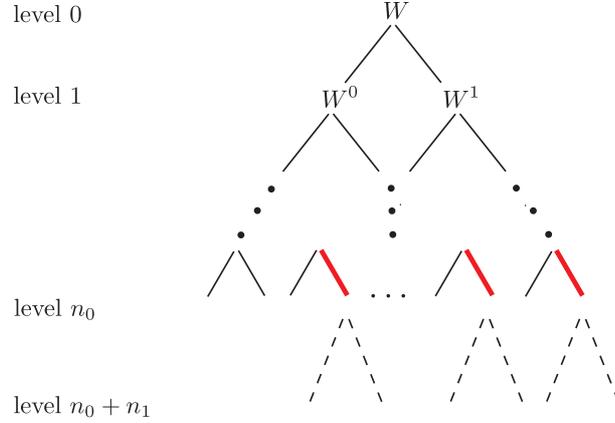


Figure 3.6: The infinite binary tree of channel W . The edges that are colored red at level n_0 of this tree correspond to the sub-channels at level n_0 whose Bhattacharyya parameter is less than $\frac{1}{2}$ (i.e., the set \mathcal{A}). The idea is then to focus on these “red” indices. We consider the sub-channels that are branched out from these red indices at a level $n_0 + n_1$ (as shown in the figure). By a careful choice of n_1 , we observe that these specific sub-channels at level $n_0 + n_1$ are greatly polarized in the sense that sum of their Bhattacharyya parameters is less than P_e . We also show that the fraction of these sub-channels is larger than R .

It is easy to see that (3.76) is equivalent to

$$\Pr\left(\sum_{i=1}^{n_1} B_i \geq \log\left(\log \frac{1}{P_e}\right) + \log(n_0 + n_1)\right) \geq 1 - \frac{d}{3}. \quad (3.77)$$

Also, as the random variables B_i are Bernoulli($\frac{1}{2}$) and iid, the relation (3.77) is equivalent to

$$\frac{\sum_{j=0}^{\log(\log \frac{1}{P_e}) + \log(n_0 + n_1)} \binom{n_1}{j}}{2^{n_1}} < \frac{d}{3}. \quad (3.78)$$

A sufficient condition for (3.78) to hold is as follows:

$$\frac{n_1^{1 + \log(\log \frac{1}{P_e}) + \log(n_0 + n_1)}}{2^{n_1}} \leq \frac{d}{3},$$

and after applying the function $\log(\cdot)$ to both sides and some further simplifications we reach to

$$n_1 - (1 + \log(\log \frac{1}{P_e}) + \log(n_0 + n_1)) \log n_1 \geq \log \frac{3}{d}. \quad (3.79)$$

3.3. Analytical Approach: from Bounds for the BEC to Universal Bounds for BMS Channels 53

It can be shown through some simple steps that there are constants $c_6, c_7 > 0$ s.t. if we choose

$$n_1 = \left\lceil \log \frac{3}{d} + c_6 (\log(\log \frac{3}{d}))^2 + c_7 \log(\log(\frac{2}{P_e})) \log(\log \frac{3}{d}) \right\rceil, \quad (3.80)$$

then the inequality (3.79) holds. Now, let $\tilde{N} = 2^{n_0+n_1}$ and consider the set \mathcal{A}_1 defined as

$$\mathcal{A}_1 = \{i \in \{0, \dots, \tilde{N} - 1\} : Z(W_{\tilde{N}}^{(i)}) \leq \frac{P_e}{\tilde{N}}\}. \quad (3.81)$$

We now show that

$$\frac{|\mathcal{A}_1|}{\tilde{N}} \geq R. \quad (3.82)$$

This relation together with (3.81) shows that block error probability of the polar code of block-length \tilde{N} and rate R is at most P_e . In order to show (3.82), we consider the sub-channels \mathcal{A}_1 that are branched out from the ones in the set \mathcal{A} . Let $i \in \mathcal{A}$ and consider the sub-channel $W_{N_0}^{(i)}$. By using the relations (3.48), Lemma 3.9 and (3.76) we conclude the following. At level $n_0 + n_1$, the number of sub-channels that are branched out from $W_{N_0}^{(i)}$ and have Bhattacharyya value less than $\frac{P_e}{\tilde{N}}$ is at least

$$2^{n_1} (1 - c_2 Z(W_{N_0}^{(i)}) (1 + \log \frac{1}{Z(W_{N_0}^{(i)})})) (1 - \frac{d}{3}).$$

Hence, by using (3.75) the total number of sub-channels at level $n_0 + n_1$ that are branched out from a sub-channel in \mathcal{A} and have Bhattacharyya value less than $\frac{P_e}{\tilde{N}}$ is

$$2^{n_0+n_1} (R + \frac{2}{3}d) (1 - \frac{d}{3}) (1 - c_2 \sum_{i \in \mathcal{A}} Z(W_{N_0}^{(i)}) (1 + \log \frac{1}{Z(W_{N_0}^{(i)})})). \quad (3.83)$$

Now, by using Lemma 3.10 we have

$$\begin{aligned} & c_2 \sum_{i \in \mathcal{A}} Z(W_{N_0}^{(i)}) (1 + \log \frac{1}{Z(W_{N_0}^{(i)})}) \\ & \leq 2c_2c_3 \sum_{i \in \mathcal{A}} (Z(W_{N_0}^{(i)}) (1 - Z(W_{N_0}^{(i)})))^\alpha \\ & \leq 2c_2c_3 \mathbb{E}[(Z_{n_0} (1 - Z_{n_0}))^\alpha] \\ & \leq 2c_2c_3 2^{-n_0\rho} \\ & \stackrel{(3.73)}{\leq} \frac{d}{3}. \end{aligned}$$

Therefore, the expression (3.83) is lower-bounded by

$$2^{n_0+n_1} (R + \frac{2}{3}d) (1 - \frac{d}{3})^2 \geq 2^{n_0+n_1} R = \tilde{N}R.$$

Hence, the relation (3.82) is proved and a block-length of size \tilde{N} is sufficient to achieve a rate R and error at most P_e . It is now easy to see that $\log \tilde{N} = n_0 + n_1$ has the form of (3.70).

3.4 Extensions and Improvements

Given the fact that polar codes do not have an optimal finite-length behavior, an important question, both from the theoretical and practical sides, is to improve the finite-length performance of these codes. We can approach this problem from two perspectives: (i) by devising better decoding algorithms and (ii) by changing the construction of polar codes (e.g., by concatenating them with other codes, use other polarizing kernels, etc). In any attempt to improve the finite-length performance, one main objective should be to improve the scaling exponent. In [26], the authors combine both of these perspectives and provide experimental evidence that the short-length performance of polar codes can be improved to be comparable to the best iterative codes used in practice. However, this improvement comes at the cost of increasing the memory usage of the decoding procedure which is undesired. It is also an interesting open question to find out how the scaling exponent changes with the list-parameter of [26]. We believe that the methods developed in this chapter can be useful in this regard.

Another approach is to consider polar codes with general $\ell \times \ell$ kernels. The objective of this section is to express hope that polar codes with larger kernels might have a better finite-length behavior. We provide analytical evidence that for large ℓ the scaling exponent tends to $\frac{1}{2}$, i.e., its optimal value. Recall from (2.1) that the optimal value of μ is $\frac{1}{2}$, and for polar codes (with $\ell = 2$) the scaling exponent is roughly $\mu = \frac{1}{3.6} \approx 0.27$ (for the BEC). We keep in mind that, in general, the decoding complexity of (extended) polar codes is $O(2^\ell N \log N)$, where N is the block-length.

Assume now that the $\ell \times \ell$ matrix G comes from the ensemble \mathcal{G}_ℓ defined in Section 2.4 and we consider a polar code based on G . We recall from Section 2.4 that the length of the polar code constructed from G is equal to $N = \ell^n$. For simplicity, we confine ourselves to the BEC. Hence, throughout this section the channel W is the BEC(z) for a fixed choice of $z \in (0, 1)$. In brief, the main result of this section is as follows. Let $\beta > 0$ be an arbitrary (small) positive number. Then, as ℓ grows large, for almost all the kernels $G \in \mathcal{G}_\ell$, we have

$$\limsup_{n \rightarrow \infty} \frac{1}{n} \log_\ell \Pr(Z_n \in [a, b]) \leq -\frac{1}{2} + \beta. \quad (3.84)$$

Here, a, b are arbitrary such that $0 < a < b < 1$ and Z_n denotes the Bhattacharyya process of a polar code based on the kernel G and channel $W = \text{BEC}(z)$.

Intuitively, Equation (3.84) indicates that when ℓ grows large, for almost all the kernels G , the ratio of the un-polarized channels scales roughly like $\ell^{-\frac{\beta}{2}}$ or equivalently $N^{-\frac{\beta}{2}}$. This in return suggests that the scaling exponent for

such polar codes tends to $\mu = 2$ as $\ell \rightarrow \infty$. In other words, in order to reach a target error probability ϵ with a block-length N , one needs to reduce the rate to $I(W) - \Theta(N^{-\frac{1}{2}})$.

In order to prove (3.84) we recall from Section 2.4 the channel splitting transform $W \rightarrow (W^0, W^1, \dots, W^{\ell-1})$: the channel $W^j : u_j \rightarrow (Y, u_0, \dots, u_{j-1})$ is the channel that the j -th bit u_j sees when it considers the bits u_0 through u_{j-1} as “known” and the bits u_{j+1} through u_ℓ as “unknown”. We state (without proof) the following facts from [13]:

- (i) Assuming that W is the BEC(z), then each of the channels W^j is also a BEC.
- (ii) The erasure probability of W^j is in general dependent of the choice of G and the value of z . This dependence is in the form of $t_j(z)$, where t_j is a polynomial of degree (at most) ℓ . Further, for $z \in [0, 1]$ we have

$$\sum_{j=0}^{\ell-1} t_j(z) = \ell z. \quad (3.85)$$

As a consequence of of these facts, the Bhattacharyya process corresponding to the channel BEC(z) and matrix G , which we denote by Z_n , has a closed form recursive expression given by $Z_0 = z$ and

$$Z_n = t_{B_n}(Z_{n-1}), \quad (3.86)$$

where $\{B_n\}_{\{n \in \mathbb{N}\}}$ is a sequence of i.i.d. random variables with uniform distribution on the set $\{0, \dots, \ell - 1\}$, i.e., $\Pr(B_1 = i) = \frac{1}{\ell}$ for $i \in \{0, \dots, \ell - 1\}$.

In order to bound the value of $\frac{1}{n} \log_\ell \Pr(Z_n \in [a, b])$, the idea is to look at the behavior of the process $Q_n = (Z_n(1 - Z_n))^\beta$ for $\beta > 0$. By the Markov inequality we have

$$\Pr(Z_n \in [a, b]) \leq \frac{\mathbb{E}[Q_n]}{\min(a, b)^\beta}.$$

Hence, it is easy to see that

$$\limsup_n \frac{1}{n} \log_\ell \Pr(Z_n \in [a, b]) \leq \limsup_n \frac{1}{n} \log_\ell \mathbb{E}[Q_n]. \quad (3.87)$$

For the rest of the proof we focus on the behavior of the process Q_n . We have

$$\begin{aligned} Q_n &= t_{B_n}(Z_{n-1})(1 - t_{B_n}(Z_{n-1})) \\ &= (Z_{n-1}(1 - Z_{n-1}))^\beta \left\{ \frac{t_{B_n}(Z_n)(1 - t_{B_n}(Z_n))}{Z_{n-1}(1 - Z_{n-1})} \right\}^\beta \\ &= Q_{n-1} \left\{ \frac{t_{B_n}(Z_n)(1 - t_{B_n}(Z_n))}{Z_{n-1}(1 - Z_{n-1})} \right\}^\beta. \end{aligned}$$

Therefore,

$$\begin{aligned}\mathbb{E}[Q_n|Z_{n-1}] &\leq Q_{n-1} \frac{\sum_{i=0}^{\ell-1} (t_i(Z_n)(1-t_i(Z_n)))^\beta}{(Z_{n-1}(1-Z_n-1))^\beta} \\ &\leq Q_{n-1} \sup_{z \in (0,1)} \frac{1}{\ell} \frac{\sum_{i=0}^{\ell-1} (t_i(z)(1-t_i(z)))^\beta}{(z(1-z))^\beta}.\end{aligned}$$

Thus, defining

$$\zeta_G^\beta = \sup_{z \in (0,1)} \frac{1}{\ell} \frac{\sum_{i=0}^{\ell-1} (t_i(z)(1-t_i(z)))^\beta}{(z(1-z))^\beta}, \quad (3.88)$$

by the chain rule of expectations, we have

$$\begin{aligned}\mathbb{E}[Q_n] &\leq \mathbb{E}[Q_0](\zeta_G^\beta)^n \\ &\leq \left(\frac{1}{2}\right)^\beta (\zeta_G^\beta)^n.\end{aligned}$$

Now, by using (3.87) for $\beta > 0$ we have

$$\limsup_n \frac{1}{n} \log_\ell \Pr(Z_n \in [a, b]) \leq \log_\ell \zeta_G^\beta. \quad (3.89)$$

The proof then follows by using the following lemma.

Lemma 3.7. *Assume that the kernel G is chosen from the ensemble \mathcal{G}_ℓ . For $0 < \beta < 1$ we have*

$$\lim_{\ell \rightarrow \infty} \mathbb{P}(\log_\ell \zeta_G^\beta < -\frac{1}{2} + 2\beta) = 1. \quad (3.90)$$

3.5 Appendix: Auxiliary Lemmas and Proofs

Proof of Lemma 3.2

The proof of the right side (3.36) and also (3.37) is an easy application of the Markov inequality. To prove the left side of (3.36), we define sequences $\{x_n\}_{n \geq 1}$ and $\{y_n\}_{n \geq 1}$ as

$$x_n = 2^{-n}, \quad (3.91)$$

$$y_n = 1 - 2^{-n}. \quad (3.92)$$

We start by noting that

$$\begin{aligned}\mathbb{E}(Z_n(1-Z_n)) &\leq \sum_{i=1}^n 2^{-i} \Pr(Z_n \in [x_{i+1}, x_i]) \\ &\quad + \sum_{i=1}^n 2^{-i} \Pr(Z_n \in [y_i, y_{i+1}]) \\ &\quad + 2^{-n}.\end{aligned}$$

As a result, there exists an index $j \in \{1, \dots, n\}$ such that at least one of the following cases occurs:

$$\mathbb{E}[Z_n(1 - Z_n)] \leq 2n[2^{-j}\Pr(Z_n \in [x_{j+1}, x_j]) + 2^{-n}], \quad (3.93)$$

or

$$\mathbb{E}[Z_n(1 - Z_n)] \leq 2n[2^{-j}\Pr(Z_n \in [y_j, y_{j+1}]) + 2^{-n}]. \quad (3.94)$$

We show that in each of these cases the statement of the lemma holds. Firstly, note that because of the symmetry of Z_n we can write

$$\Pr(Z_n^z \in [y_{j+1}, y_j]) = \Pr(Z_n^{1-z} \in [x_{j+1}, x_j]).$$

Hence, without loss of generality we can assume that (3.93) holds. We first prove the lemma for $a = 1 - b = \frac{1}{4}$. We then use this result to prove the lemma in its fullest extent. We claim that for any $1 \leq j \leq n$ we have,

$$2^{-j}\Pr(Z_n \in [x_{j+1}, x_j]) \leq 2(n+1)\Pr(Z_n \in [\frac{1}{4}, \frac{3}{4}]) + \frac{n^3}{2^n}. \quad (3.95)$$

Assuming that the above claim holds true, by using (3.93) we obtain

$$\mathbb{E}(Z_n(1 - Z_n)) \leq 2n[\Pr(Z_n \in [\frac{1}{4}, \frac{3}{4}]) + \frac{n^2 + 2}{2^n}],$$

and as a result, by taking $\frac{1}{n} \log(\cdot)$ from both sides, the first part of the lemma is proved for $a = 1 - b = \frac{1}{4}$.

We now turn to the proof of relation (3.95) for $1 \leq j \leq n$. For $j = 1$, the result of the claim is trivial. Hence, in the following we assume that $2 \leq j \leq n$. We now prove that for any fixed j such that $2 \leq j \leq n$, we have

$$2^{-j}\Pr(Z_n \in [x_{j+1}, x_j]) \leq 2(n+1)\Pr(Z_n \in [\frac{1}{4}, \frac{3}{4}]) + \frac{n^3}{2^n}, \quad (3.96)$$

and hence the relation (3.95) is also proved. We fix the index j and prove the above claim for any value of $n \in \mathbb{N}$. The proof consist of two steps.

Step1: We first show that $\forall m \in \mathbb{N}$,

$$\Pr(Z_m \in [x_{2j+2}, x_j]) \leq m\Pr(Z_m \in [x_j, \frac{3}{4}]) + \frac{1}{2^n}. \quad (3.97)$$

To prove (3.97), fix $m \in \mathbb{N}$ and define the sets A and B as

$$A = \{(b_1, \dots, b_m) \in \Omega_m : t_{b_m} \circ \dots \circ t_{b_1}(z) \in [x_{2j+2}, x_j]\},$$

$$B = \{(b_1, \dots, b_m) \in \Omega_m : t_{b_m} \circ \dots \circ t_{b_1}(z) \in [x_j, \frac{3}{4}]\}.$$

In other words, A is the set of all the paths that start from $z = Z_0$ and end up in $[x_{2j+2}, x_j]$ and B is the set of paths that start from z and end up in $[x_j, \frac{3}{4}]$. We now partition the A into the disjoint sets A_k , $k \in \{0, 1, \dots, m\}$, defined as

$$A_k = \{(b_1, \dots, b_m) \in A : b_k = 1; b_i = 0 \ \forall i > k\}. \quad (3.98)$$

It is easy to see that $|A - \cup_k A_k| \leq 1$. Our aim is now to show that for $k \in \{0, 1, \dots, m\}$,

$$|A_k| \leq |B|. \quad (3.99)$$

To do this, we show that there exists a one-to-one correspondence between A_k and a subset of B . In other words, we claim that we can map each member of A_k to a distinct member of B . Consider $(b_1, \dots, b_m) \in A_k$. We now construct a distinct member $(b'_1, \dots, b'_m) \in B$ corresponding to (b_1, \dots, b_m) . We first set $b'_i = b_i$ for $i < k$ and hence the uniqueness condition is fulfilled. Consider the number x defined as

$$x = \begin{cases} z & ; \text{if } k = 1, \\ t_{b_{k-1}} \circ \dots \circ t_{b_1}(z) & ; \text{if } k > 1. \end{cases} \quad (3.100)$$

Note that since $(b_1, \dots, b_m) \in A_k$ we have

$$t_{b_m} \circ \dots \circ t_{b_k}(x) \in [x_{2j+1}, x_j]. \quad (3.101)$$

Now, note that as $(b_1, \dots, b_m) \in A_k$, we have $b_k = 1$ and $b_i = 0$ for $i > k$. Thus, in this setting (3.101) becomes

$$\overbrace{t_0 \circ \dots \circ t_0}^{m-k \text{ times}}(x^2) \in [x_{2j+1}, x_j]. \quad (3.102)$$

Hence,

$$x_{2j+1} \leq 1 - (1 - x^2)^{2^{m-k}} \leq x_j. \quad (3.103)$$

From the left side of (3.103) and using the fact that $1 - (1 - x)^2 \leq 2x$ we obtain

$$x_{2j+1} \leq 2^{m-k} x^2 \Rightarrow 2^{-j + \frac{k-m+1}{2}} \leq x. \quad (3.104)$$

From the right side of (3.103) we have

$$\ln(1 - x_j) \leq 2^{m-k} \ln(1 - x^2),$$

and by using the inequality $-x - \frac{x^2}{2} \leq \ln(1 - x) \leq -x$ we obtain

$$x \leq 2^{-\frac{j}{2} + \frac{k-m+1}{2}}. \quad (3.105)$$

Let us recall that we let $b'_i = b_i$ for $i < k$. We now construct the remaining values b'_k, \dots, b'_m by the following algorithm: consider the number x given in (3.100). In the following, we will also construct a sequence $x = x_{k-1}, x_k, x_{k+1}, \dots, x_m$ such that for $i \geq k$ we have $x_i = t_{b'_i}(x_{i-1})$. Begin with the initial value $x_{k-1} = x$ and for $i \geq k$ recursively construct b'_i from b'_{i-1} and x_{i-1} by the following rule: if $t_{b'_{i-1}}(x_{i-1}) \leq \frac{3}{4}$, then $b'_i = 0$ and $x_i = t_0(x_{i-1})$, otherwise $b'_i = 1$ and $x_i = t_1(x_{i-1})$. We now show that the value of x_m is always in the interval $[x_j, \frac{3}{4}]$. In this regard, an important observation is that for i s.t. $k - 1 \leq i \leq m$, once the value of x_i lies in the interval $[x_j, \frac{3}{4}]$ then

for all $i \leq t \leq m$ we have $x_t \in [x_j, \frac{3}{4}]$. Hence, we only need to show that by the above algorithm, there exists an index i , s.t. $k-1 \leq i \leq m$, and the value of x_i lies in the interval $[x_j, \frac{3}{4}]$. On one hand, observe that due to (3.105) and the fact that $j \geq 2$, we have $x \leq 2^{-\frac{1}{2}} < \frac{3}{4}$. Thus, the value of x_i is definitely less than $\frac{3}{4}$ for $i \geq k$. If the value of x_{k-1} is also greater than x_j then we have nothing to prove. Else, it might be the case that $x < x_j$. We now show that in this case the algorithm moves in a way that the value of x_m falls eventually into the desired region $[x_j, \frac{3}{4}]$. To show this, a moment of thought reveals that this is equivalent to showing that we always have

$$\underbrace{t_0 \circ \dots \circ t_0}_{m-k+1 \text{ times}}(x) = 1 - (1-x)^{2^{m-k+1}} \geq x_j. \quad (3.106)$$

Note that the function $1 - (1-x)^{2^{m-k+1}}$ is a strictly increasing function of the unit interval. Thus, in order to have (3.106) it is equivalent that

$$2^{m-k+1} \ln(1-x) \leq \ln(1-x_j),$$

and after some further simplification using the inequality $-x - \frac{x^2}{2} \leq \ln(1-x) \leq -x$, we deduce that a sufficient condition to have (3.106) is

$$x_j \leq 2^{m-k} x \Rightarrow 2^{-j+k-m} \leq x. \quad (3.107)$$

But this sufficient condition is certainly met by considering the inequality (3.104) and noting the fact that $-j + \frac{k-m+1}{2} \geq -j + k - m$. Hence, the claim in (3.99) is proved and as a result, the claim in (3.97) is true.

Step 2: Firstly note that in order for Z_n to be in the interval $[x_{j+1}, x_j]$, the value of Z_{n-j} should lie in the interval $[x_{2j+1}, x_j^{2^{-2j}}]$. As a result, we can write

$$\begin{aligned} & \Pr(Z_n \in [x_{j+1}, x_j]) \\ &= \Pr(Z_n \in [x_{j+1}, x_j] \mid Z_{n-j} \in [x_{2j+1}, x_j]) \times \Pr(Z_{n-j} \in [x_{2j+1}, x_j]) \\ & \quad + \Pr(Z_n \in [x_{j+1}, x_j] \mid Z_{n-j} \in (x_j, x_j^{2^{-2j}}]) \times \Pr(Z_{n-j} \in (x_j, x_j^{2^{-2j}}]), \end{aligned} \quad (3.108)$$

and by letting $m = n - j$ in relation (3.97), we can easily obtain

$$\Pr(Z_{n-j} \in [x_{2j+1}, x_j]) \leq n \Pr(Z_{n-j} \in [x_j, \frac{3}{4}]) + \frac{n^2 + 1}{2^n}. \quad (3.109)$$

Thus, by combining (3.108) and (3.109), we obtain

$$\begin{aligned} & \Pr(Z_n \in [x_{j+1}, x_j]) \\ & \leq n \Pr(Z_{n-j} \in [x_j, \frac{3}{4}]) + \Pr(Z_{n-j} \in [x_j, x_j^{2^{-2j}}]) + \frac{n^2 + 1}{2^n}. \end{aligned} \quad (3.110)$$

Finally, in order to conclude the proof of (3.96), we prove the following relations:

$$2^{-j} \Pr(Z_{n-j} \in [x_j, \frac{3}{4}]) \leq \Pr(Z_n \in [\frac{1}{4}, \frac{3}{4}]), \quad (3.111)$$

and

$$2^{-j} \Pr(Z_{n-j} \in [x_j, x_j^{2^{-2j}}]) \leq \Pr(Z_n \in [\frac{1}{4}, \frac{3}{4}]). \quad (3.112)$$

Firstly note that since $(x_j)^{\frac{1}{2^{2j}}} \geq \frac{3}{4}$, then it is enough to prove (3.112). To prove (3.112), we only need to show that for a value x s.t. $x \in [x_j, (x_j)^{\frac{1}{2^{2j}}}]$, there exists an j -tuple $(b_1, \dots, b_j) \in \Omega_j$ such that $t_{b_1} \circ \dots \circ t_{b_j}(x) \in [\frac{1}{4}, \frac{3}{4}]$. We show this by constructing the binary values b_1, \dots, b_j in terms of x . Consider the following algorithm: start with $y_0 = x$ and for $1 \leq i \leq j$, we recursively construct b_i from y_{i-1} by the following rule: If $t_0(y_{i-1}) \leq \frac{3}{4}$, then $b_i = 0$ and $y_i = t_0(y_{i-1})$. Otherwise, let $b_i = 1$ and $y_i = t_1(x_{i-1})$. To show that this algorithm succeeds in the sense that $y_j \in [\frac{1}{4}, \frac{3}{4}]$, we first observe that once the value of y_i lies in the interval $[\frac{1}{4}, \frac{3}{4}]$ (for some $1 \leq i \leq j$), then for all $i \leq t \leq j$ we have $y_t \in [\frac{1}{4}, \frac{3}{4}]$. Hence, we only need to show that by the above algorithm, there exists an index i , s.t. $1 \leq i \leq j$, and the value of y_i lies in the interval $[\frac{1}{4}, \frac{3}{4}]$. On one hand, assume $y_0 = x \in [x_j, \frac{1}{4}]$. We can then write

$$\begin{aligned} \overbrace{t_0 \circ \dots \circ t_0}^{j \text{ times}}(x) &= 1 - (1 - x)^{2^j} \\ &\geq 1 - (1 - x_j)^{2^j} \\ &\geq \frac{1}{2}, \end{aligned}$$

where the last step follows from the fact that $x_j = 2^{-j}$. On the other hand, assume $x \in (\frac{3}{4}, (x_j)^{\frac{1}{2^j}}]$. We can write

$$\begin{aligned} \overbrace{t_1 \circ \dots \circ t_1}^{j \text{ times}}(x) &\leq ((x_j)^{\frac{1}{2^{2j}}})^{2^{2j}} \\ &\leq x_j < \frac{3}{4}. \end{aligned}$$

As a result, the above algorithm always succeeds and the lemma is proved for $a = 1 - b = \frac{1}{4}$.

We now prove the lemma for any choice of $a, b \in (0, 1)$ s.t. $\sqrt{a} \leq 1 - \sqrt{1 - b}$. Let $p_n(z, a, b)$ be defined as in (3.3). We have

$$\begin{aligned} p_{n+1}(z, a, b) &= \sum_{\phi_{\omega_{n+1}}} \frac{1}{2^{n+1}} \mathbb{1}_{\{z \in \phi_{\omega_{n+1}}^{-1}[a, b]\}} \\ &= \sum_{\phi_{\omega_n}} \frac{1}{2^n} \frac{\mathbb{1}_{\{z \in \phi_{\omega_n}^{-1}[t_0^{-1}(a), t_0^{-1}(b)]\}} + \mathbb{1}_{\{z \in \phi_{\omega_n}^{-1}[t_1^{-1}(a), t_1^{-1}(b)]\}}}{2} \\ &= \frac{1}{2} (p_n(z, t_0^{-1}(a), t_0^{-1}(b)) + p_n(z, t_1^{-1}(a), t_1^{-1}(b), z)). \end{aligned}$$

It is easy to see that if $\sqrt{a} \leq 1 - \sqrt{1 - b}$, then

$$[t_0^{-1}(a), t_1^{-1}(b)] \subseteq [t_0^{-1}(a), t_0^{-1}(b)] \cup [t_1^{-1}(a), t_1^{-1}(b)],$$

and hence,

$$2p_{n+1}(z, a, b) \geq p_n(z, t_0^{-1}(a), t_1^{-1}(b)).$$

Continuing this way, we can show that for $m \in \mathbb{N}$

$$2^m p_{n+m}(z, a, b) \geq p_n(z, \overbrace{t_0^{-1} \circ \cdots \circ t_0^{-1}}^{m \text{ times}}(a), \overbrace{t_1^{-1} \circ \cdots \circ t_1^{-1}}^{m \text{ times}}(b)). \quad (3.113)$$

As m grows large, we have

$$\begin{aligned} \overbrace{t_0^{-1} \circ \cdots \circ t_0^{-1}}^{m \text{ times}}(a) &\rightarrow 0, \\ \overbrace{t_1^{-1} \circ \cdots \circ t_1^{-1}}^{m \text{ times}}(b) &\rightarrow 1. \end{aligned}$$

Therefore, by (3.113) there exists a positive integer m_0 such that for $n \in \mathbb{N}$

$$2^{m_0} p_{n+m_0}(z, a, b) \geq p_n(z, \frac{1}{4}, \frac{3}{4}).$$

The thesis now follows from this relation together with the result of Lemma 3.31.

Proof of Lemma 3.3

Recall that for a realization $\omega = \{b_k\}_{k \in \mathbb{N}} \in \Omega$ we define $\omega_n = (b_1, \dots, b_n)$. The maps t_0 and t_1 , hence the maps ϕ_{ω_n} s, are strictly increasing maps on $[0, 1]$. Thus $\phi_{\omega_n}(z) \rightarrow 0$ implies that $\phi_{\omega_n}(z') \rightarrow 0$ for $z' \leq z$ and $\phi_{\omega_n}(z) \rightarrow 1$ implies that $\phi_{\omega_n}(z') \rightarrow 1$ for $z' \geq z$. Moreover, we know that for almost every $z \in (0, 1)$, $\lim_{n \rightarrow \infty} \phi_{\omega_n}(z)$ is either 0 or 1 for almost every realization $\{\phi_{\omega_n}\}_{n \in \mathbb{N}}$. Hence, it suffices to let

$$z_\omega^* = \inf\{z : \phi_{\omega_n}(z) \rightarrow 1\}.$$

To prove the second part of the lemma, notice that

$$\begin{aligned} z &= \Pr(Z_\infty = 1) \\ &= \Pr(\phi_{\omega_n}(z) \rightarrow 1) \\ &= \Pr(\inf\{z : \phi_{\omega_n}(z) \rightarrow 1\} \leq z) \\ &= \Pr(z_\omega^* < z). \end{aligned}$$

Which shows that z_ω^* is uniformly distributed on $[0, 1]$.

Proof of Lemma 3.4

In order to compute $\lim_{n \rightarrow \infty} \mathbb{E}[\frac{1}{n} \log(\phi_{\omega_n}^{-1}(b) - \phi_{\omega_n}^{-1}(a))]$, we first define the process $\{\bar{Z}_n\}_{n \in \mathbb{N} \cup \{0\}}$ with $\bar{Z}_0 = z \in [0, 1]$ and

$$\bar{Z}_{n+1} = \begin{cases} \sqrt{\bar{Z}_n}, & \text{w.p. } \frac{1}{2}, \\ 1 - \sqrt{1 - \bar{Z}_n}, & \text{w.p. } \frac{1}{2}. \end{cases} \quad (3.114)$$

We can think of \bar{Z}_n as the reverse stochastic process of Z_n . Equivalently, we can also define \bar{Z}_n via the inverse maps t_0^{-1}, t_1^{-1} . Consider the sequence of i.i.d. symmetric Bernoulli random variables B_1, B_2, \dots and define $\bar{Z}_n = \psi_{\omega_n}(z)$ where $\omega_n \triangleq (b_1, \dots, b_n) \in \Omega_n$ and

$$\psi_{\omega_n} = t_{b_n}^{-1} \circ t_{b_{n-1}}^{-1} \circ \dots \circ t_{b_1}^{-1}. \quad (3.115)$$

We now show that the Lebesgue measure (or the uniform probability measure) on $[0, 1]$, denoted by ν , is the unique, hence ergodic, invariant measure for the Markov process \bar{Z}_n . To prove this result, first note that if \bar{Z}_n is distributed according to the Lebesgue measure, then

$$\begin{aligned} \Pr(\bar{Z}_{n+1} < x) &= \frac{1}{2}\Pr(\bar{Z}_n < t_0(x)) + \frac{1}{2}\Pr(\bar{Z}_n < t_1(x)) \\ &= \frac{1}{2}x^2 + \frac{1}{2}(2x - x^2) = x. \end{aligned}$$

Thus, \bar{Z}_{n+1} is also distributed according to the Lebesgue measure and this implies the invariance of the Lebesgue measure for \bar{Z}_n . In order to prove the uniqueness, we will show that for any $z \in (0, 1)$, \bar{Z}_n converges weakly to a uniformly distributed random point in $[0, 1]$, i.e.,

$$\bar{Z}_n = \psi_{\omega_n}(z) \xrightarrow{d} \nu. \quad (3.116)$$

Note that with (3.116) the uniqueness of ν is proved since for any invariant measure ρ assuming \bar{Z}_n is distributed according to ρ , we have

$$\rho(\cdot) = \Pr(\bar{Z}_n \in \cdot) = \int \Pr(\bar{Z}_n \in \cdot) \rho(dz) \xrightarrow{d} \nu(\cdot). \quad (3.117)$$

To prove (3.116), note that ψ_{ω_n} has the same (probability) law as $\phi_{\omega_n}^{-1}$ and we know that $\phi_{\omega_n}^{-1}(z) \rightarrow z_\omega^*$ almost surely and hence weakly. Also, z_ω^* is distributed according to ν , which proves (3.116). We are now ready to show that

$$\lim_{n \rightarrow \infty} \mathbb{E}\left[\frac{1}{n} \log(\phi_{\omega_n}^{-1}(b) - \phi_{\omega_n}^{-1}(a))\right] = \frac{1}{2 \ln 2} - 1. \quad (3.118)$$

Using the mean value theorem, we can write

$$\psi_n(a) - \psi_n(b) = \psi'_n(c)(b - a),$$

for some $c \in (a, b)$. And by chain rule,

$$\begin{aligned} \psi'_{\omega_n}(c) &= (t_{b_n}^{-1} \circ t_{b_{n-1}}^{-1} \circ \dots \circ t_{b_1}^{-1})'(c) \\ &= t_{b_1}^{-1'}(c) \cdot t_{b_2}^{-1'}(t_{b_1}^{-1}(c)) \cdot \dots \cdot t_{b_n}^{-1'}(t_{b_{n-1}}^{-1} \circ \dots \circ t_{b_1}^{-1}(c)) \\ &= t_{b_1}^{-1'}(\psi_0(c)) \cdot t_{b_2}^{-1'}(\psi_1(c)) \cdot \dots \cdot t_{b_n}^{-1'}(\psi_{n-1}(c)), \end{aligned}$$

and after applying $\log(\cdot)$ to both sides we obtain

$$\frac{1}{n} \log(\psi'_{\omega_n}(c)) = \frac{1}{n} \sum_{j=1}^n \ln t_{b_j}^{-1'}(\psi_{j-1}(c)). \quad (3.119)$$

By the ergodic theorem, the last expression converges almost surely to the expectation of $\log t_{B_1}^{-1'}(U)$, where U is assumed to be distributed according to ν . Hence, the asymptotic value of (3.119) can be computed as

$$\begin{aligned} & \mathbb{E}[\log t_{B_1}^{-1'}(U)] \\ &= \frac{1}{2} \int_0^1 \log(\sqrt{x})' dx + \frac{1}{2} \int_0^1 \log(1 - \sqrt{1-x})' dx \\ &= \frac{1}{2 \ln 2} - 1. \end{aligned}$$

Auxiliary Lemmas

Lemma 3.8. *Consider a channel W with its Bhattacharyya process $Z_n = Z(W_n)$ and assume that for $n \in \mathbb{N}$*

$$\mathbb{E}[(Z_n(1 - Z_n))^\alpha] \leq \beta 2^{-n\rho}, \quad (3.120)$$

where α, β, ρ are positive constants with $\alpha < 1$. We then have for $n \in \mathbb{N}$

$$\Pr(Z_n \leq \frac{1}{2}) \geq I(W) - c_1 2^{-n\rho}, \quad (3.121)$$

where c_1 is a positive constant that depends on α, β, ρ .

Proof. The proof consists of three steps. First, consider an arbitrary BMS channel W and let $Z_n = Z(W_n)$. Also, consider the process $Y_n = 1 - Z_n^2$. By using the relations (2.24) and (2.23), it can easily be checked that the process E_n has the form of (3.124) and hence Lemma 3.9 is applicable to Y_n . We thus have from (3.125) that for $n \in \mathbb{N}$

$$\Pr(Y_n \geq \frac{1}{2}) \leq c_2 Y_0 (1 + \log \frac{1}{Y_0}).$$

As a consequence

$$\begin{aligned} I(W) &= \lim_{n \rightarrow \infty} \Pr(Y_n \geq \frac{1}{2}) \\ &\leq c_2 (1 - Z(W)^2) (1 + \log \frac{1}{1 - Z(W)^2}). \end{aligned} \quad (3.122)$$

In the second step, we consider a channel W for which (3.120) holds for $n \in \mathbb{N}$. By using (3.120), it is easy to see that for $n \in \mathbb{N}$

$$\begin{aligned} & \mathbb{E}[(Z_n^2(1 - Z_n^2))^\alpha \mathbb{1}_{\{Z_n \geq \frac{1}{2}\}}] \\ &= \mathbb{E}[(Z_n(1 + Z_n))^\alpha (Z_n(1 - Z_n))^\alpha \mathbb{1}_{\{Z_n \geq \frac{1}{2}\}}] \\ &\leq \sup_{z \in [\frac{1}{2}, 1]} (z(1+z))^\alpha \mathbb{E}[(Z_n(1 - Z_n))^\alpha \mathbb{1}_{\{Z_n \geq \frac{1}{2}\}}] \\ &\leq 2^\alpha \beta 2^{-n\rho} \leq \beta 2^{1-n\rho}. \end{aligned} \quad (3.123)$$

In the final step, we consider a number $n \in \mathbb{N}$ and let $N = 2^n$. We then define the set \mathcal{A} as

$$\mathcal{A} = \{i \in \{0, 1, \dots, N-1\} : Z(W_N^{(i)}) \leq \frac{1}{2}\},$$

with \mathcal{A}^c being its complement. We have

$$\begin{aligned} & \sum_{i \in \mathcal{A}^c} I(W_N^{(i)}) \\ & \stackrel{(a)}{\leq} \sum_{i \in \mathcal{A}^c} c_2 (1 - Z(W_N^{(i)}))^2 \left(1 + \log \frac{1}{1 - Z(W_N^{(i)})^2}\right) \\ & \stackrel{(b)}{\leq} \sum_{i \in \mathcal{A}^c} 4c_2 c_3 (Z(W_N^{(i)}))^2 (1 - Z(W_N^{(i)}))^2 \\ & = 4c_2 c_3 N \mathbb{E}[(Z_n^2 (1 - Z_n^2))^\alpha \mathbb{1}_{\{Z_n \geq \frac{1}{2}\}}] \\ & \stackrel{(c)}{\leq} 8c_2 c_3 N \beta 2^{-n\rho}. \end{aligned}$$

Here (a) follows from (3.122), (b) follows from Lemma 3.10 and the fact that for $x \leq \frac{3}{4}$ we have $1 + \log \frac{1}{x} \leq 4 \log \frac{1}{x}$, and (c) follows from (3.123). Now, as a consequence of the above chain of inequalities we have

$$\begin{aligned} |\mathcal{A}| & \geq \sum_{i \in \mathcal{A}} I(W_N^{(i)}) \\ & = NI(W) - \sum_{i \in \mathcal{A}^c} I(W_N^{(i)}) \\ & \geq N(I(W) - 2c_2 c_3 \beta 2^{-n\rho}), \end{aligned}$$

and consequently

$$\Pr(Z_n \leq \frac{1}{2}) = \frac{|\mathcal{A}|}{N} \geq 2c_2 c_3 \beta 2^{-n\rho}.$$

Hence, the proof follows. \square

Lemma 3.9. Consider a generic stochastic process $\{X_n\}_{n \geq 0}$ s.t. $X_0 = x$, where $x \in (0, 1)$, and for $n \geq 1$

$$X_n \leq \begin{cases} X_{n-1}^2 & ; \text{if } B_n = 1, \\ 2X_{n-1} & ; \text{if } B_n = 0. \end{cases} \quad (3.124)$$

Here, $\{B_n\}_{n \geq 1}$ is a sequence of iid random variables with distribution Bernoulli($\frac{1}{2}$). We then have for $n \in \mathbb{N}$

$$\Pr(X_n \leq 2^{-2^{\sum_{i=1}^n B_i}}) \geq 1 - c_2 x \left(1 + \log \frac{1}{x}\right), \quad (3.125)$$

where c_2 is a positive constant.

Proof. We slightly modify X_n to start with $X_0 = x$, where $x \in (0, 1)$, and for $n \geq 1$

$$X_n = \begin{cases} X_{n-1}^2 & ; \text{if } B_n = 1, \\ 2X_{n-1} & ; \text{if } B_n = 0. \end{cases} \quad (3.126)$$

It is easy to see that if we prove the lemma for this version of X_n , then the result of the lemma is valid for any generic X_n that satisfies (3.124).

We analyze the process $A_n = -\log X_n$, i.e., $A_0 = -\log x \triangleq a_0$ and

$$A_{n+1} = \begin{cases} 2A_n & ; \text{if } B_n = 1, \\ A_n - 1 & ; \text{if } B_n = 0. \end{cases} \quad (3.127)$$

Note that in terms of the process A_n , the statement of the lemma can be phrased as

$$\Pr(A_n \geq 2^{\sum_{i=1}^n B_i}) \geq 1 - c_2 \frac{1 + a_0}{2^{a_0}}.$$

Associate to each $(b_1, \dots, b_n) \triangleq \omega_n \in \Omega_n$ a sequence of “runs” $(r_1, \dots, r_{k(\omega_n)})$. This sequence is constructed by the following procedure. We define r_1 as the smallest index $i \in \mathbb{N}$ so that $b_{i+1} \neq b_1$. In general, if $\sum_{j=1}^{k-1} r_j < n$ then

$$r_k = \min\left\{i \mid \sum_{j=1}^{k-1} r_j < i \leq n, b_{i+1} \neq b_{\sum_{j=1}^{k-1} r_j}\right\} - \sum_{j=1}^{k-1} r_j.$$

The process stops whenever the sum of the runs equals n . Denote the stopping time of the process by $k(\omega_n)$. In words, the sequence (b_1, \dots, b_n) starts with b_1 . It then repeats b_1, r_1 times. Next follow r_2 instances of \bar{b}_1 ($\bar{b}_1 := 1 - b_1$), followed again by r_3 instances of b_1 , and so on. We see that b_1 and $(r_1, \dots, r_{k(\omega_n)})$ fully describe $\omega_n = (b_1, \dots, b_n)$. Therefore, there is a one-to-one map

$$(b_1, \dots, b_n) \longleftrightarrow \{b_1, (r_1, \dots, r_{k(\omega_n)})\}. \quad (3.128)$$

Note that we can either have $b_1 = 1$ or $b_1 = 0$. We start with the first case, i.e., we first assume $B_1 = 1$. We have:

$$\sum_{i=1}^n b_i = \sum_{j \text{ odd} \leq k(\omega_n)} r_j,$$

and

$$n = \sum_{j=1}^{k(\omega_n)} r_j.$$

Analogously, for a realization $(b_1, b_2, \dots) \triangleq \omega \in \Omega$ of the infinite sequence of random variable $\{B_i\}_{i \in \mathbb{N}}$, we can associate a sequence of runs (r_1, r_2, \dots) . In this regard, considering the infinite sequence of random variables $\{B_i\}_{i \in \mathbb{N}}$ (with the extra condition $B_1 = 1$), the corresponding sequence of runs, which we denote by $\{R_k\}_{k \in \mathbb{N}}$, is an iid sequence with $\Pr(R_i = j) = \frac{1}{2^j}$. Let us now

see how we can express the A_n in terms of the $r_1, r_2, \dots, r_{k(\omega_n)}$. We begin by a simple example: Consider the sequence $(b_1 = 1, b_2, \dots, b_8)$ and the associated run sequence $(r_1, \dots, r_5) = (1, 2, 1, 3, 1)$. We have

$$\begin{aligned}
A_1 &= a_0 2^{r_1}, \\
A_3 &= a_0 2^{r_1} - r_2, \\
A_4 &= (a_0 2^{r_1} - r_2) 2^{r_3} = a_0 2^{r_1+r_3} - r_2 2^{r_3}, \\
A_7 &= (a_0 2^{r_1} - r_2) 2^{r_3} - r_4 = a_0 2^{r_1+r_3} - r_2 2^{r_3} - r_4, \\
A_8 &= ((a_0 \times 2^{r_1} - r_2) \times 2^{r_3} - r_4) \times 2^{r_5} \\
&= a_0 2^{r_1+r_3+r_5} - r_2 2^{r_3+r_5} - r_4 2^{r_5} \\
&= 2^{r_1+r_3+r_5} (a_0 - 2^{-r_1} r_2 - 2^{-(r_1+r_3)} r_4).
\end{aligned}$$

In general, for a sequence (b_1, \dots, b_n) with the associated run sequence $(r_1, \dots, r_{k(\omega_n)})$ we can write:

$$\begin{aligned}
A_n &= a_0 2^{\sum_{i \text{ odd} \leq k(\omega_n)} r_i} - \sum_{i \text{ even} \leq k(\omega_n)} r_i 2^{\sum_{i < j \text{ odd}} r_j} \\
&= a_0 2^{\sum_{i \text{ odd} \leq k(\omega_n)} r_i} - \sum_{i \text{ even} \leq k(\omega_n)} r_i 2^{(-\sum_{j \text{ odd} < i} r_j + \sum_{i \text{ odd} \leq k(\omega_n)} r_i)} \\
&= [2^{\sum_{i \text{ odd} \leq k(\omega_n)} r_i}][a_0 - (\sum_{i \text{ even} \leq k(\omega_n)} r_i 2^{-\sum_{j \text{ odd} < i} r_j})] \\
&= [2^{\sum_{i=1}^n B_i}][a_0 - (\sum_{i \text{ even} \leq k(\omega_n)} r_i 2^{-\sum_{j \text{ odd} < i} r_j})].
\end{aligned}$$

Our aim is to lower-bound

$$\begin{aligned}
&\Pr(A_n \geq 2^{\sum_{i=1}^n B_i}) \\
&= \Pr(a_0 - \sum_{i \text{ even} \leq k(\omega_n)} r_i 2^{-\sum_{j \text{ odd} < i} r_j} \geq 1),
\end{aligned}$$

or, equivalently, to upper-bound

$$\Pr\left(\sum_{i \text{ even} \leq k(\omega_n)} r_i 2^{-\sum_{j \text{ odd} < i} r_j} \geq a_0 - 1\right). \quad (3.129)$$

For $n \in \mathbb{N}$, define the set $U_n \in \mathcal{F}_n$ as

$$U_n = \{\omega_n \in \Omega_n \mid \exists l \leq k(\omega_n) : \sum_{i \text{ even} \leq l} r_i 2^{-\sum_{j \text{ odd} < i} r_j} \geq a_0 - 1\}.$$

Clearly we have:

$$\Pr\left(\sum_{i \text{ even} \leq k(\omega_n)} r_i 2^{-\sum_{j \text{ odd} < i} r_j} \geq a_0 - 1\right) \leq \Pr(U_n).$$

In the following we show that if $(b_1, \dots, b_n) \in U_n$, then for any choice of b_{n+1} , $(b_1, \dots, b_n, b_{n+1}) \in U_{n+1}$. We will only consider the case when $b_n, b_{n+1} = 1$, the other three cases can be verified similarly. Let $\omega_n = (b_1, \dots, b_{n-1}, b_n = 1) \in U_n$. Hence, $k(\omega_n)$ is an odd number (recall that $b_1 = 1$) and the quantity $\sum_{i \text{ even} \leq k(\omega_n)} r_i 2^{-\sum_{j \text{ odd} < i} r_j}$ does not depend on $r_{k(\omega_n)}$. Now consider the sequence $\omega_{n+1} = (b_1, \dots, b_n = 1, 1)$. Since the last bit (b_{n+1}) equals 1, then $r_{k(\omega_{n+1})} = r_{k(\omega_n)}$ and the value of the sum remains unchanged. As a result $(b_1, \dots, b_n, 1) \in U_{n+1}$. From above, we conclude that $\theta_i(U_i) \subseteq \theta_{i+1}(U_{i+1})$ and as a result

$$\Pr(U_i) = \Pr(\theta_i(U_i)) \leq \Pr(\theta_{i+1}(U_{i+1})) = \Pr(U_{i+1}).$$

Hence, the quantity $\lim_{n \rightarrow \infty} \Pr(U_n) = \lim_{n \rightarrow \infty} \Pr(\theta_n(U_n)) = \lim_{n \rightarrow \infty} \Pr(\cup_{i=1}^n \theta_i(U_i))$ is an upper bound on (3.129). On the other hand, consider the set

$$V = \{\omega \in \Omega \mid \exists l : \sum_{i \text{ even} \leq l} r_i 2^{-\sum_{j \text{ odd} < i} r_j} \geq a_0 - 1\}.$$

By the definition of V we have $\cup_{i=1}^{\infty} \theta_i(U_i) \subseteq V$, and as a result, $\Pr(\cup_{i=1}^{\infty} \theta_i(U_i)) \leq \Pr(V)$. In order to bound the probability of the set V , note that assuming $B_1 = 1$, the sequence $\{R_k\}_{k \in \mathbb{N}}$ (i.e., the sequence of runs when associated with the sequence $\{B_i\}_{i \in \mathbb{N}}$) is an iid sequence with $\Pr(R_i = j) = \frac{1}{2^j}$. We also have

$$\begin{aligned} & \Pr(a_0 - \sum_{i \text{ even} \leq m} R_i 2^{-\sum_{j \text{ odd} < i} R_j} \leq 1) & (3.130) \\ &= \Pr\left(\sum_{i \text{ even} \leq m} R_i 2^{-\sum_{j \text{ odd} < i} R_j} \geq a_0 - 1\right) \\ &= \Pr\left(2^{\sum_{i \text{ even} \leq m} R_i} 2^{-\sum_{j \text{ odd} < i} R_j} \geq 2^{a_0 - 1}\right) \\ &\leq \frac{\mathbb{E}[2^{\sum_{i \text{ even} \leq m} R_i} 2^{-\sum_{j \text{ odd} < i} R_j}]}{2^{a_0 - 1}}, \end{aligned}$$

where the last step follows from the Markov inequality. The idea is now to provide an upper bound on the quantity $\mathbb{E}[2^{\sum_{i \text{ even} \leq m} R_i} 2^{-\sum_{j \text{ odd} < i} R_j}]$. Let $X = \sum_{i \text{ even} \leq m} R_i 2^{-\sum_{j \text{ odd} < i} R_j}$. We have

$$\begin{aligned} & \mathbb{E}[2^X] \\ &= \sum_{l=1}^{\infty} \Pr(R_2 = l) \mathbb{E}[2^X \mid R_2 = l] \\ &\stackrel{(a)}{=} \sum_{l=1}^{\infty} \frac{1}{2^l} \mathbb{E}[2^X \mid R_2 = l] \\ &= \sum_{l=1}^{\infty} \frac{1}{2^l} \mathbb{E}[2^{\frac{R_1}{2^l}}] \mathbb{E}[2^{\frac{X}{2^l}}] \end{aligned}$$

$$\begin{aligned}
&= \sum_{l=1}^{\infty} \frac{1}{2^l (2^{1-\frac{1}{2^l}})} \mathbb{E}[2^{\frac{X}{2^l}}] \\
&\stackrel{(b)}{\leq} \sum_{l=1}^{\infty} \frac{1}{2^l (2^{1-\frac{1}{2^l}})} (\mathbb{E}[2^X])^{\frac{1}{2^l}},
\end{aligned}$$

where (a) follows from the fact that R_i s are iid and X is self-similar and (b) follows from Jensen inequality. As a result, an upper bound on the quantity $\mathbb{E}[2^X]$ can be derived as follows. We have

$$\mathbb{E}[2^X] \leq \frac{1}{2(2^{\frac{1}{2}}-1)} (\mathbb{E}[2^X])^{\frac{1}{2}} + \frac{1}{4(2^{\frac{3}{4}}-1)} (\mathbb{E}[2^X])^{\frac{1}{4}} + \frac{1}{4(2^{\frac{7}{8}}-1)} (\mathbb{E}[2^X])^{\frac{1}{8}}.$$

The equation $y = \frac{1}{2(2^{\frac{1}{2}}-1)} y^{\frac{1}{2}} + \frac{1}{4(2^{\frac{3}{4}}-1)} y^{\frac{1}{4}} + \frac{1}{4(2^{\frac{7}{8}}-1)} y^{\frac{1}{8}}$ has only one real valued solution y^* , and $y^* \leq 3$ (more precisely, $y^* \approx 2.87$). As a result, we have $\mathbb{E}[2^X] \leq y^* \leq 3$. Thus by (3.130) we obtain

$$\Pr(a_0 - \sum_{i \text{ even} \leq m} R_i 2^{-\sum_{j \text{ odd} < i} R_j} \leq 1) \leq \frac{3}{2^{a_0-1}}$$

Thus, given that $B_1 = 1$, we have:

$$\Pr(A_n \geq 2^{\sum_{i=1}^n B_i}) \geq 1 - \frac{3}{2^{a_0-1}}.$$

Or more precisely we have

$$\Pr(A_n \geq 2^{\sum_{i=1}^n B_i} \mid B_1 = 1) \geq 1 - \frac{3}{2^{a_0-1}}.$$

Now consider the case $B_1 = 0$. We show that a similar bound applies for A_n . Firstly, note that by fixing the value of n the distribution of R_1 is as follows: $\Pr(R_i) = \frac{1}{2^i}$ for $1 \leq i \leq n-1$ and $\Pr(R_1 = n) = \frac{1}{2^{n-1}}$. We have

$$\begin{aligned}
&\Pr(A_n \geq 2^{\sum_{i=1}^n B_i} \mid B_1 = 0) \\
&= \sum_{i=1}^n \Pr(A_n \geq 2^{\sum_{i=1}^n B_i} \mid R_1 = i, B_1 = 0) \Pr(R_1 = i \mid B_1 = 0) \\
&= \sum_{i \leq a_0-1, i \leq n} \Pr(A_n \geq 2^{\sum_{i=1}^n B_i} \mid R_1 = i, B_1 = 0) \Pr(R_1 = i \mid B_1 = 0) \\
&+ \sum_{i > a_0-1, i \leq n} \Pr(R_1 = i \mid B_1 = 0) \\
&\leq \sum_{i \leq a_0-1, i \leq n} \frac{1}{2^i} \frac{3}{2^{a_0-1-i}} + \frac{2}{2^{a_0-1}} \\
&\leq \frac{3a_0}{2^{a_0-1}}.
\end{aligned}$$

Hence, considering the two cases together, we have:

$$\Pr(A_n \geq 2^{\sum_{i=1}^n B_i}) \geq 1 - \frac{3(1+a_0)}{2^{a_0}}.$$

Hence, the proof follows with $c_2 = 3$. \square

Lemma 3.10. *Let $\alpha < 1$ be a constant. We have for $x \in (0, \frac{3}{4}]$*

$$x \log\left(\frac{1}{x}\right) \leq c_3 (x(1-x))^\alpha, \quad (3.131)$$

where

$$c_3 = \frac{2}{(1-\alpha) \ln 2}. \quad (3.132)$$

Proof. By applying the function $\log(\cdot)$ to both sides of (3.131) and some further simplifications, the inequality (3.131) is equivalent to the following: For $x \in (0, \frac{3}{4}]$

$$\log\left(\log\frac{1}{x}\right) \leq \log c_3 + (1-\alpha) \log\frac{1}{x} + \alpha \log(1-x).$$

As $x \leq \frac{3}{4}$, we have $\alpha \log(1-x) \geq -\log 4$. Hence, in order for the above inequality to hold it is sufficient that for $x \in (0, \frac{3}{4}]$

$$\log\left(\log\frac{1}{x}\right) \leq \log\frac{c_3}{4} + (1-\alpha) \log\frac{1}{x}.$$

Now, by letting $u = \log\frac{1}{x}$, the last inequality becomes

$$(1-\alpha)u - \log u + \log\frac{c_3}{4} \geq 0, \quad (3.133)$$

for $u \geq \log(\frac{4}{3})$. It is now easy to check that by the choice of c_3 as in (3.132), the minimum of the above expression over the range $u \geq \log(\frac{4}{3})$ is always non-negative and hence the proof follows. \square

Scaling Laws for the Polarized Channels

4

4.1 Problem Formulation

In the previous chapter, we studied scaling laws for the set of un-polarized channels, i.e., the channels whose Bhattacharyya value is bounded away from 0 and 1. The main focus of this chapter¹ is the polarized channels, i.e., the channels that the Bhattacharyya value is close to either 0 or 1. We will see in the following that different scaling laws, rather than the ones mentioned in the previous chapter, govern the behavior of the polarized channels.

We begin by recalling that for a channel W , the Bhattacharyya process $Z_n = Z(W_n)$ converges almost surely to a $\{0, 1\}$ -valued random variable Z_∞ with $\Pr(Z_\infty = 0) = I(W)$. Let $N = \ell^n$ be the block-length of the polar code². Thus, if we consider the sub-channels $\{W_N^{(i)}\}_{0 \leq i \leq N-1}$ for a sufficiently large n , then most of these sub-channels are “polarized” in the sense that their Bhattacharyya value is very close to either 0 or 1.

Consider a rate $R < I(W)$ and let $\mathcal{I}_{N,R}$ be the set of indices of the NR channels in the set $\{W_N^{(i)}\}_{0 \leq i \leq N-1}$ with the least values for the Bhattacharyya parameter. For the SC decoder, we recall that

$$\max_{i \in \mathcal{I}_{N,R}} \frac{1}{2} \left(1 - \sqrt{1 - Z(W_N^{(i)})^2} \right) \leq P_e^{\text{SC}} \leq \sum_{i \in \mathcal{I}_{N,R}} Z(W_N^{(i)}), \quad (4.1)$$

where P_e^{SC} denotes the average block error probability of the SC decoder, with block-length N and rate R . This relation shows that the distribution of the Bhattacharyya parameters of the channels $\{W_N^{(i)}\}_{0 \leq i \leq N-1}$ plays a fundamental

¹The material of this chapter is based on [20] and [21].

²Since it entails no extra work, we state all results directly for $\ell \times \ell$ kernels.

role in the analysis of polar codes. The objective of this chapter is to analyze the asymptotic behavior of

$$F_n(z) = \frac{|\{i : Z(W_{\ell^n}^{(i)}) \leq z\}|}{\ell^n}, \quad (4.2)$$

where $|A|$ denotes the cardinality of the set A . Due to the definition of Z_n , the function $F_n(z)$ is equivalent to the cumulative distribution function of the Bhattacharyya process $Z_n = Z(W_n)$, i.e.,

$$\Pr(Z_n \leq z) = F_n(z). \quad (4.3)$$

The fact that the Bhattacharyya process $\{Z_n\}_{n \in \mathbb{N}}$ converges almost surely to a $\{0, 1\}$ -valued random variable Z_∞ , with $\Pr(Z_\infty = 0) = I(W)$, implies that the functions $F_n(z)$ converge point-wise to a function $F_\infty(z)$ shown in Figure 4.1. In this chapter we intend to go one step further and zoom in (or rescale) the functions $F_n(z)$ around the points $z = 0$ and $z = 1$ to discover what the properly rescaled functions look like. The analysis of the process $\{Z_n\}_{n \in \mathbb{N}}$ around the point $z = 0$ is of particular interest, as this indicates how the “good” channels behave (i.e., how the channels that have mutual information close to 1 behave).

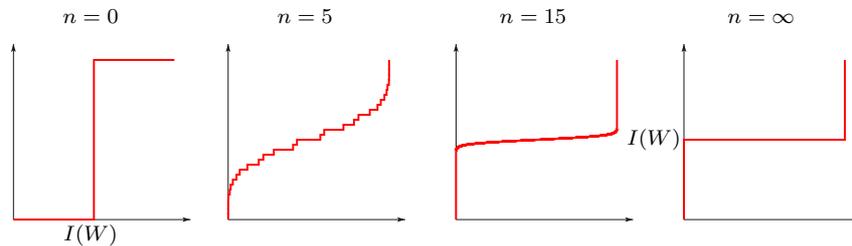


Figure 4.1: The cdf of the random variables $Z_n = Z(W_n)$ for different values of n when the channel W is $\text{BEC}(\frac{1}{2})$. The rightmost plot corresponds to the cdf of the limiting random variable Z_∞ .

We thus ask:

Question 5. *How do the cumulative distributions $\Pr(Z_n \leq z)$ vary (asymptotically) in terms of n , R and W ?*

The main contribution of this chapter is the study of asymptotics of the cumulative distribution $\Pr(Z_n \leq z)$ and its dependence on R . In more detail, for a fixed rate R , we study the scaling of $\Pr(Z_n \leq z)$ in terms of n . We will see in this chapter that for sufficiently large n , if we properly rescale the functions $F_n(z)$, we will discover the Gaussian Q function. The next question we address in this chapter is as follows.

Question 6. *What implications does the asymptotic behavior of $\Pr(Z_n \leq z)$ have on the block-error probability of the SC decoder?*

A rough answer to this question is that these two quantities share the same asymptotic behavior up to the leading exponents. We will refine this answer throughout the text. The final point that we note in this chapter is in regard to the behavior of other decoders, especially the optimal one (i.e., MAP decoder).

Question 7. *What can we say about the asymptotic behavior of the block-error probability of other decoders, especially the MAP decoder?*

In other words, we ask how much the MAP decoder is superior to the SC decoder in the asymptotic regime.

4.1.1 Relevant Work

The asymptotic behavior of the process Z_n is closely related to the “partial distances” of the kernel matrix G that the polar code is based on³:

Definition 4.1 (Partial Distances). *We define the partial distances $D_i(G)$, $i = 0, \dots, \ell - 1$, of an $\ell \times \ell$ matrix $G = \begin{bmatrix} g_0 \\ \vdots \\ g_{\ell-1} \end{bmatrix}$ (g_i 's are row vectors) as*

$$D_i(G) \triangleq d_H(\{g_i\}, \langle g_{i+1}, \dots, g_{\ell-1} \rangle), \quad i = 0, \dots, \ell - 2,$$

$$D_{\ell-1}(G) \triangleq d_H(\{g_{\ell-1}\}, \{\mathbf{0}\}).$$

Here, $\langle g_{i+1}, \dots, g_{\ell-1} \rangle$ denotes the linear space spanned by $g_{i+1}, \dots, g_{\ell-1}$. Also, $d_H(a, b)$ denotes the Hamming distance between two binary vectors a, b of equal length and more generally if A, B are two sets of binary vectors all of the same length, then $d_H(A, B) = \min_{a \in A, b \in B} d_H(a, b)$. The first exponent of G is then defined as

$$E(G) = \frac{1}{\ell} \sum_{i=0}^{\ell-1} \log_{\ell} D_i(G),$$

and the second exponent of G is defined as

$$V(G) = \frac{1}{\ell} \sum_{i=0}^{\ell-1} (\log_{\ell} D_i(G) - E(G))^2.$$

In other words, the first exponent $E(G)$ and the second exponent $V(G)$ are the mean and the variance of the random variable $\log_{\ell} D_B(G)$, where B is a random variable taking a value in $\{0, 1, \dots, \ell - 1\}$ with uniform probability. It should be noted that the invertibility of G implies that the partial distances

³See Section 2.3.

$\{D_i(G)\}$ are strictly positive, making the exponent $E(G)$ finite [12]. Note also that the condition for a matrix G to be polarizing, that none of the column permutations of G is upper triangular, implies that at least one of the $D_i(G)$'s is strictly greater than 1. This results in $E(G)$ being strictly positive.

The following theorem partially characterizes the behavior of the process $\{Z_n\}_{n \in \mathbb{N}}$ around $z = 0$.

Theorem 4.1 ([7] and [12]). *Let W be a BMS channel and assume that we are using as the kernel matrix an $\ell \times \ell$ matrix G with exponent $E(G)$. For any fixed β with $0 < \beta < E(G)$,*

$$\lim_{n \rightarrow \infty} \Pr(Z_n \leq 2^{-\ell^{n\beta}}) = I(W). \quad (4.4)$$

Conversely, if $I(W) < 1$, then for any fixed $\beta > E(G)$,

$$\lim_{n \rightarrow \infty} \Pr(Z_n \geq 2^{-\ell^{n\beta}}) = 1. \quad (4.5)$$

An important consequence of Theorem 4.1 is as follows. Let $\Pr_e^{\text{SC}}(N, R)$ be the block error probability when using polar codes with the kernel matrix G , of block-length $N = \ell^n$ and rate $R < I(W)$ under SC decoding. By using the inequality on the right-hand side in (4.1) and the limit in (4.4), we can easily conclude that for any $0 < \beta < E(G)$, the value of $P_e^{\text{SC}}(N, R)$ is less than $2^{-\ell^{n\beta}}$ for sufficiently large n . Also, by using the inequality on the left-hand side in (4.1) and (4.5), we can easily conclude that for $\beta > E(G)$ the value of $P_e^{\text{SC}}(N, R)$ is greater than $2^{-\ell^{n\beta}}$ for sufficiently large n . Hence, $P_e^{\text{SC}}(N, R)$ behaves as $2^{-\ell^{nE(G)+o(n)}}$ as n tends to infinity. Note that this result is rate-independent, provided that the rate R is less than the capacity $I(W)$. We provide here a refined estimate for $\Pr(Z_n \leq z)$. Specifically, we derive the asymptotic relation between $\Pr(Z_n \leq z)$ and the rate of transmission R . From this, we derive the asymptotic behavior of $P_e^{\text{SC}}(N, R)$ and its dependence on the rate of transmission. We further derive lower bounds on the error probability when we perform MAP decoding, instead of SC decoding.

Note that, in the following, the logarithms are in base 2 unless explicitly stated otherwise.

4.2 Asymptotic Behavior of $F_n(z)$

This section is devoted to the answer of Questions 5 and 6. In particular we show how the quantity $\Pr(Z_n \leq z)$ scales with n and R and what this implies on the scaling behavior of P_e^{SC} .

Theorem 4.2. *Consider an $\ell \times \ell$ polarizing kernel matrix $G = \begin{bmatrix} g_0 \\ \vdots \\ g_{\ell-1} \end{bmatrix}$. For a BMS channel W , let $\{Z_n = Z(W_n)\}_{n \in \mathbb{N}}$ be the Bhattacharyya process of W . Let $Q(t) \triangleq \int_t^\infty e^{-z^2/2} dz / \sqrt{2\pi}$ be the error function and $Q^{-1}(\cdot)$ be its inverse function.*

1. For $R < I(W)$,

$$\lim_{n \rightarrow \infty} \Pr \left(Z_n \leq 2^{-\ell^{nE(G) + \sqrt{nV(G)}Q^{-1} \left(\frac{R}{I(W)} \right) + f(n)}} \right) = R.$$

2. Let $H = [g_{\ell-1}^T, \dots, g_0^T]^{-1}$ (\cdot^T denotes the transpose) and assume that $D_i(H) \leq D_{i-1}(H)$ for $1 \leq i \leq \ell - 1$. Then, for $R' < 1 - I(W)$ we have

$$\lim_{n \rightarrow \infty} \Pr \left(Z_n \geq 1 - 2^{-\ell^{nE(H) + \sqrt{nV(H)}Q^{-1} \left(\frac{R'}{1-I(W)} \right) + f(n)}} \right) = R'.$$

Here, $f(n)$ is any function satisfying $f(n) = o(\sqrt{n})$.

Theorem 4.2 characterizes the asymptotic behavior of $\Pr(Z_n \leq z)$ and refines Theorem 4.1 in the following way. According to Theorem 4.1, if we transmit at a fixed rate R below the channel capacity, then the quantity $\log_\ell(-\log P_e^{\text{SC}})$ scales like $nE(G) + o(n)$. The first part of Theorem 4.2 gives us one further term by stating that $o(n)$ is in fact $\sqrt{nV(G)}Q^{-1} \left(\frac{R}{I(W)} \right) + o(\sqrt{n})$. Whereas, the second part of Theorem 4.2 characterizes the asymptotic behavior of $\Pr(Z_n \geq z)$ near $z = 1$, which is important in applications of polar codes for source coding [29, 30]. Put together, Theorem 4.2 characterizes the scaling of the error probability of polar codes in terms of the block-length when the rate is fixed. The rest of this section is devoted to providing the required machinery and intuition for proving Theorem 4.2.

4.2.1 Preliminaries

Let $\{B_n\}_{n \in \mathbb{N}}$ be a sequence of i.i.d. random variables that take their values in $\{0, 1, \dots, \ell - 1\}$ with uniform probability, i.e., $\Pr(B_0 = j) = \frac{1}{\ell}$ for $j \in \{0, 1, \dots, \ell - 1\}$. Let $(\Omega, \mathcal{F}, \Pr)$ denote the probability space generated by the sequence $\{B_n\}_{n \in \mathbb{N}}$ and let $(\Omega_n, \mathcal{F}_n, \Pr_n)$ be the probability space generated by (B_0, \dots, B_n) . Recall that $\{D_i(G)\}_{0 \leq i \leq \ell-1}$ are the partial distances of the matrix G . By using the bounds given in [13, Lemma 5.7, Lemma 5.10] we have the following relationship between the Bhattacharyya parameters of W^i and that of W . We have [13, Lemma 5.10]

$$Z(W)^{D_i(G)} \leq Z(W^i) \leq 2^{\ell-i} Z(W)^{D_i(G)}. \quad (4.6)$$

Also, let $H = [g_{\ell-1}^T, \dots, g_0^T]^{-1}$. Assuming $D_i(H) \leq D_{i-1}(H)$, we have [13, Lemma 5.7]

$$(1 - Z(W))^{D_i(H)} \leq 1 - Z(W^i) \leq 2^{2i+1} (1 - Z(W))^{D_i(H)}. \quad (4.7)$$

4.2.2 The Idea behind the Proof

We first provide an intuitive picture behind the result of Theorem 4.2. For simplicity, assume $\ell = 2$ and $G = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}$. Also, assume that W is a binary

erasure channel (BEC) with erasure probability z . The capacity of this channel is $1 - z$. For such a channel, the Bhattacharyya process has a simple closed form [1] as $Z_0 = z$ and

$$Z_{n+1} = \begin{cases} Z_n^2, & B_n = 1, \\ 2Z_n - Z_n^2, & B_n = 0. \end{cases} \quad (4.8)$$

We know from Section 4.1.1 that as n grows large, Z_n tends almost surely to a $\{0, 1\}$ -valued random variable Z_∞ with $\Pr(Z_\infty = 0) = 1 - \epsilon$. The asymptotic behavior of $\{Z_n\}_{n \in \mathbb{N}}$ can be explained roughly by considering the behavior of $\{-\log Z_n\}_{n \in \mathbb{N}}$. In particular, it is clear from (4.8) that at time $n + 1$, $-\log Z_n$ is either doubled (when $B_n = 1$), or decreased by at most 1 (when $B_n = 0$). Also, observe that once $-\log Z_n$ becomes sufficiently large, subtracting 1 from it has a negligible effect compared with the doubling operation (See Figure 4.2). Now assume that m is a sufficiently large number. Conditioned on the event that $-\log Z_m$ is a very large value (or equivalently, the value of Z_m is very close to 0: this happens with probability very close to $1 - z$), for $n > m$ the process $\{-\log Z_n\}_{n \in \mathbb{N}}$ evolves each time by being doubled if $B_n = 1$ or remaining roughly the same if $B_n = 0$. Consequently, for $n > m$ the process $\{\log(-\log Z_n)\}_{n \in \mathbb{N}}$ increases by 1 if $B_n = 1$ or remaining roughly the same if $B_n = 0$. In other words, we have $\log(-\log Z_{n+1}) = \log(-\log Z_n) + B_n$. We

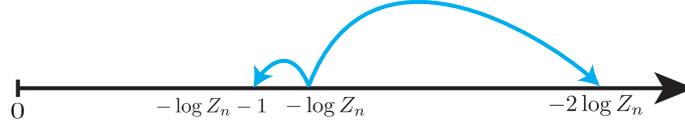


Figure 4.2: At time $n + 1$, $-\log Z_n$ is either doubled (when $B_n = 0$), or decreased by at most 1 (when $B_n = 1$). However, once Z_n becomes sufficiently small (or $-\log Z_n$ becomes sufficiently large), subtracting 1 from it has negligible effect compared with the doubling operation.

can then use the central limit theorem to characterize the asymptotic behavior of $\{-\log Z_n\}_{n \in \mathbb{N}}$ for $n \gg m$.

The proof of Theorem 4.2 is done by making the above intuitive steps rigorous for a BMS channel W and for a polarizing $\ell \times \ell$ kernel matrix G .

4.2.3 A Generic Process

In a slightly more general setting, we study the asymptotic properties of $\Pr(X_n \leq x)$ for any generic process $\{X_n\}_{n \in \mathbb{N}}$ satisfying the conditions (c1)–(c4) defined as follows.

Definition 4.2. Let S be a random variable taking values in $[1, \infty)$. Assume that the expectation and the variance of $\log S$ exist and are denoted by $\mathbb{E}[\log S]$ and $\mathbb{V}[\log S]$, respectively. We let $\{S_n\}_{n \in \mathbb{N}}$ denote i.i.d. samples of S . We also

let $\{(X_n, S_n) \in (0, 1) \times [1, \infty)\}_{n \in \mathbb{N}}$ be a random process satisfying the following conditions:

- (c1) There exists a random variable X_∞ such that $X_n \rightarrow X_\infty$ holds almost surely.
- (c2) With probability 1 we have $(X_n)^{S_n} \leq X_{n+1}$.
- (c3) There exists a constant $c \geq 1$ such that $X_{n+1} \leq c(X_n)^{S_n}$ holds with probability 1.
- (c4) S_n is independent of X_m for $m \leq n$.

The random processes $\{(Z_n, D_{B_n}(G))\}_{n \in \mathbb{N}}$ and $\{(1 - Z_n, D_{B_n}(H))\}_{n \in \mathbb{N}}$ satisfy the above four conditions. The fact that these processes satisfy the condition (c1) has been proved in [13, Lemma 5.4], and the result reads that if G is polarizing, then Z_∞ takes only 0 and 1, with probabilities $I(W)$ and $1 - I(W)$, respectively. Conditions (c2) and (c3) also hold because of (4.6) and (4.7).

Our objective now is to prove that for such a process $\{(X_n, S_n)\}_{n \in \mathbb{N}}$, we have

$$\lim_{n \rightarrow \infty} \Pr \left(X_n \leq 2^{-2^{n\mathbb{E}[\log S] + t\sqrt{n\mathbb{V}[\log S] + f(n)}}} \right) = \Pr(X_\infty = 0)Q(t), \quad (4.9)$$

where $f(n)$ is any function such that $f(n) = o(\sqrt{n})$ holds. The results of Theorem 4.2 then follow by noting that $\Pr(Z_\infty = 0) = I(W)$ and $\Pr(1 - Z_\infty = 0) = \Pr(Z_\infty = 1) = 1 - I(W)$ hold, and by substituting $t = Q^{-1}(R/I(W))$ and $t = Q^{-1}(R'/(1 - I(W)))$, respectively, into (4.9).

We prove (4.9) by showing the two inequalities obtained by replacing the equality in (4.9) by inequality in both directions.

4.2.4 Proof of (4.9) in the Forward Direction

As the first step we have

Lemma 4.1. *Let $\{(X_n, S_n)\}_{n \in \mathbb{N}}$ be a random process satisfying (c1), (c3) and (c4). For any $f(n) = o(\sqrt{n})$,*

$$\liminf_{n \rightarrow \infty} \Pr \left(X_n \leq 2^{-2^{n\mathbb{E}[\log S] + t\sqrt{n\mathbb{V}[\log S] + f(n)}}} \right) \geq \Pr(X_\infty = 0)Q(t).$$

Proof. Without loss of generality, we can assume that c in condition (c3) satisfies $c \geq 2$. Define the process $\{L_n\}_{n \in \mathbb{N}}$ as $L_n \triangleq \log X_n$. From (c3), we have

$$L_n \leq \log c + S_{n-1}L_{n-1},$$

and by applying the above relation recursively, for $m \leq n - 1$ we obtain

$$L_n \leq \left(\sum_{j=m}^{n-1} \prod_{i=j+1}^{n-1} S_i \right) \log c + \left(\prod_{i=m}^{n-1} S_i \right) L_m$$

$$\leq \left(\prod_{i=m}^{n-1} S_i \right) ((n-m) \log c + L_m). \quad (4.10)$$

Fix $\beta \in (0, \mathbb{E}[\log S])$ and let

$$m \triangleq (\log n + \log \log c) / \beta. \quad (4.11)$$

Conditioned on the event $\mathcal{D}_m(\beta) \triangleq \{X_m < 2^{-2^{\beta m}}\}$, by using (10.13) we obtain

$$L_n \leq - \left(\prod_{i=m}^{n-1} S_i \right) m \log c.$$

Let the event $\mathcal{H}_m^{n-1}(t)$ be defined as

$$\mathcal{H}_m^{n-1}(t) \triangleq \left\{ \sum_{i=m}^{n-1} \log S_i \geq (n-m) \mathbb{E}[\log S] + t \sqrt{(n-m) \mathbb{V}[\log S]} + f(n-m) \right\},$$

where f is any function such that $f(k) = o(\sqrt{k})$ holds. Conditioned on $\mathcal{D}_m(\beta)$ and $\mathcal{H}_m^{n-1}(t)$, we have

$$\log(-L_n) \geq \log m + \log \log c + (n-m) \mathbb{E}[\log S] + t \sqrt{(n-m) \mathbb{V}[\log S]} + f(n-m).$$

Hence,

$$\begin{aligned} \Pr \left(\log(-L_n) \geq \log m + \log \log c + (n-m) \mathbb{E}[\log S] + t \sqrt{(n-m) \mathbb{V}[\log S]} + f(n-m) \right) \\ \geq \Pr(\mathcal{D}_m(\beta) \cap \mathcal{H}_m^{n-1}(t)) = \Pr(\mathcal{D}_m(\beta)) \Pr(\mathcal{H}_m^{n-1}(t)). \end{aligned}$$

The last equality follows from the independence condition (c4).

Note that taking the limit $n \rightarrow \infty$ also implies $m \rightarrow \infty$ and $n-m \rightarrow \infty$ via (4.11). From the polarization theorem, we have $\lim_{n \rightarrow \infty} \Pr(\mathcal{D}_m(\beta)) = \Pr(X_\infty = 0)$. We also have $\lim_{n \rightarrow \infty} \Pr(\mathcal{H}_m^{n-1}(t)) = Q(t)$ due to the central limit theorem for $\{\log S_i\}$. We consequently have

$$\liminf_{n \rightarrow \infty} \Pr \left(\log(-\log X_n) \geq n \mathbb{E}[\log S] + t \sqrt{n \mathbb{V}[\log S]} + f(n) \right) \geq \Pr(X_\infty = 0) Q(t)$$

for any $f(n) = o(\sqrt{n})$. \square

4.2.5 Proof of (4.9) in the Reverse Direction

The second step of the proof of (4.9) is to prove the other direction of the inequality. We have

Lemma 4.2. *Let $\{(X_n, S_n)\}_{n \in \mathbb{N}}$ be a random process satisfying (c1), (c2) and (c4). For any $f(n) = o(\sqrt{n})$,*

$$\limsup_{n \rightarrow \infty} \Pr \left(X_n \leq 2^{-2^{n \mathbb{E}[\log S] + t \sqrt{n \mathbb{V}[\log S]} + f(n)}} \right) \leq \Pr(X_\infty = 0) Q(t).$$

Proof. Let $L_n \triangleq \log X_n$. From (c2), for $m \leq n-1$ we have

$$\begin{aligned} L_n &\geq S_{n-1}L_{n-1} \\ &\geq \left(\prod_{i=m}^{n-1} S_i \right) L_m, \end{aligned}$$

and thus

$$\log(-L_n) \leq \sum_{i=m}^{n-1} \log S_i + \log(-L_m). \quad (4.12)$$

Hence, for any fixed m and any $\delta \in (0, 1)$,

$$\begin{aligned} &\limsup_{n \rightarrow \infty} \Pr \left(\log(-L_n) > n\mathbb{E}[\log S] + t\sqrt{n\mathbb{V}[\log S]} + f(n) \right) \\ &\leq \limsup_{n \rightarrow \infty} \Pr \left(\log(-L_n) > n\mathbb{E}[\log S] + t\sqrt{n\mathbb{V}[\log S]} + f(n), X_m \leq \delta \right) \\ &\quad + \limsup_{n \rightarrow \infty} \Pr \left(\log(-L_n) > n\mathbb{E}[\log S] + t\sqrt{n\mathbb{V}[\log S]} + f(n), X_m > \delta \right). \end{aligned} \quad (4.13)$$

The first term in the right-hand side of (4.13) is upper bounded as

$$\begin{aligned} &\limsup_{n \rightarrow \infty} \Pr \left(\log(-L_n) > n\mathbb{E}[\log S] + t\sqrt{n\mathbb{V}[\log S]} + f(n), X_m \leq \delta \right) \\ &\stackrel{(a)}{\leq} \limsup_{n \rightarrow \infty} \Pr \left(\sum_{i=m}^{n-1} \log S_i + \log(-L_m) > n\mathbb{E}[\log S] + t\sqrt{n\mathbb{V}[\log S]} + f(n), X_m \leq \delta \right) \\ &\stackrel{(b)}{=} Q(t)\Pr(X_m \leq \delta), \end{aligned}$$

where (a) follows from (4.12), and where (b) follows from (c4) and the central limit theorem. The second term in the right-hand side of (4.13) is upper bounded as

$$\begin{aligned} &\limsup_{n \rightarrow \infty} \Pr \left(\log(-L_n) > n\mathbb{E}[\log S] + t\sqrt{n\mathbb{V}[\log S]} + f(n), X_m > \delta \right) \\ &\leq \limsup_{n \rightarrow \infty} \Pr \left(X_n \leq \frac{\delta}{2}, X_m > \delta \right) \\ &\stackrel{(a)}{\leq} \Pr \left(X_\infty \leq \frac{\delta}{2}, X_m > \delta \right), \end{aligned}$$

where (a) follows from (c1). Applying these bounds to (4.13), for any $\delta \in (0, 1)$, we have

$$\begin{aligned} &\limsup_{n \rightarrow \infty} \Pr \left(\log(-L_n) > n\mathbb{E}[\log S] + t\sqrt{n\mathbb{V}[\log S]} + f(n) \right) \\ &\leq \limsup_{m \rightarrow \infty} \left\{ Q(t)\Pr(X_m \leq \delta) + \Pr \left(X_\infty \leq \frac{\delta}{2}, X_m > \delta \right) \right\} \end{aligned}$$

$$\begin{aligned} &\leq Q(t)\Pr(X_\infty \leq \delta) + \Pr\left(X_\infty \leq \frac{\delta}{2}, X_\infty \geq \delta\right) \\ &= Q(t)\Pr(X_\infty \leq \delta). \end{aligned}$$

By letting $\delta \rightarrow 0$, we obtain the result. \square

4.3 Asymptotic Behavior of the MAP Error

In this section, we provide the relation between the MAP block error probability, block-length N at a given (fixed) rate R . Similar results as Theorem 4.2 hold for the case of the MAP decoder.

Theorem 4.3. *Let W be a BMS channel and let $R < I(W)$ be the rate of transmission. Consider an $\ell \times \ell$ kernel matrix G with $\{w_0(G), \dots, w_{\ell-1}(G)\}$ the Hamming weights of its rows and define*

$$E_w(G) = \frac{1}{\ell} \sum_{i=0}^{\ell-1} \log_\ell w_i(G), \quad V_w(G) = \frac{1}{\ell} \sum_{i=0}^{\ell-1} (\log_\ell w_i(G) - E_w(G))^2. \quad (4.14)$$

If we use polar codes of length $N = \ell^n$ and rate R for transmission, then the probability of error under MAP decoding, P_e^{MAP} , satisfies

$$\log_\ell(-\log P_e^{\text{MAP}}) \leq nE_w(G) + \sqrt{nV_w(G)}Q^{-1}\left(\frac{R}{I(W)}\right) + o(\sqrt{n}). \quad (4.15)$$

Corollary 4.1. *Let G be according to Arikan's original construction [1], i.e., $G = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}$, which is the only polarizing matrix for the case $\ell = 2$. For this G , we have $w_i(G) = D_i(G)$ for $i = 0$ and 1. Hence, the block error probability for the SC decoder and the MAP block error probability share the same asymptotic behavior according to Theorems 4.2 and 4.3.*

For a general $\ell \times \ell$ matrix G , however, we might have strict inequality $E_w(G) > E(G)$, in which case we still have an asymptotic gap between the error probability with SC decoding and the lower bound of MAP error probability. Whether or not this gap can be filled or made narrower is an open problem. We start the proof of Theorem 4.3 by stating a general fact regarding the MAP error probability of linear codes.

Lemma 4.3. *The MAP error probability of a linear code \mathcal{C} over a BMS channel W is lower bounded by $Z(W)^{2d_{\min}}/4$ where d_{\min} is the minimum distance of \mathcal{C} .*

Proof. Within this proof, the notation $\Pr(\dots)$ should be understood as generically denoting the probability of an event (\dots) . Since the MAP error probability of a linear code over a BMS channel does not depend on the transmitted codeword, we can assume without loss of generality that the transmitted codeword is the all-zero codeword, denoted by $\mathbf{0}$. Let \mathbf{Y} be the random variable corresponding to a received sequence when $\mathbf{0}$ is transmitted and let $P(y | c)$

be the likelihood of received sequence y given a codeword c . Consider an arbitrary codeword $c \in \mathcal{C} \setminus \{\mathbf{0}\}$. Since MAP and ML are equivalent for equiprobable codewords, the MAP error probability is clearly lower bounded as

$$\Pr(\cup_{c' \in \mathcal{C} \setminus \{\mathbf{0}\}} \{P(\mathbf{Y} | c') \geq P(\mathbf{Y} | \mathbf{0})\}) \geq \Pr(P(\mathbf{Y} | c) \geq P(\mathbf{Y} | \mathbf{0})).$$

That is, assuming that $\mathbf{0}$ has been sent, the MAP error probability is lower bounded by the probability that the codeword c is more likely than $\mathbf{0}$. We now provide a lower bound for the probability of the latter event. Let us consider c as a binary vector of length N , i.e., $c = (c_0, c_1, \dots, c_{N-1})$. We let A be the set of indices $i \in \{0, 1, \dots, N-1\}$ such that $c_i = 1$. Thus, the set A has cardinality equal to the Hamming weight of c which we write as $w(c)$. We thus obtain

$$\Pr(P(\mathbf{Y} | c) \geq P(\mathbf{Y} | \mathbf{0})) = \Pr\left(\prod_{i \in A} P(y_i | 1) \geq \prod_{i \in A} P(y_i | 0)\right). \quad (4.16)$$

For a positive integer m , let us define the BMS channel $W^{\otimes m} : \{0, 1\} \rightarrow \mathcal{Y}^m$ as

$$W^{\otimes m}(y_1^m | x) \triangleq \prod_{i=1}^m W(y_i | x). \quad (4.17)$$

It is now easy to see that the right-hand side of (4.16) is equal to the probability of sending the symbol 0 on the channel $W^{\otimes w(c)}$, we receive an output for which the symbol 1 is more likely than 0. Hence,

$$\begin{aligned} & \Pr\left(\prod_{i \in A} P(y_i | 1) \geq \prod_{i \in A} P(y_i | 0)\right) \\ &= P_e(W^{\otimes w(c)}) \\ &\stackrel{(a)}{\geq} \frac{1}{2} \left(1 - \sqrt{1 - Z(W^{\otimes w(c)})^2}\right) \\ &\stackrel{(b)}{=} \frac{1}{2} \left(1 - \sqrt{1 - Z(W)^{2w(c)}}\right) \\ &\geq \frac{1}{4} Z(W)^{2w(c)}, \end{aligned}$$

where step (a) follows from (4.1) and where (b) follows from the fact that for $m \geq 1$ we have $Z(W^{\otimes m}) = Z(W)^m [1]$. \square

It should be noted that the lower bound $P_e(W^{\otimes w(c)}) \geq (1/4)Z(W)^{2w(c)}$ in the proof of Lemma 4.3 is not asymptotically tight in terms of the conventional exponents. It is possible to obtain tighter lower bounds via more elaborate arguments as in [47, Chapter 4]. However, as we are only interested in the behavior of double exponents, the above bound is sufficient for the purpose of proving Theorem 4.3.

In order to prove Theorem 4.3, from Lemma 4.3 it is sufficient to prove that given any $\epsilon > 0$ there exists an integer $M \in \mathbb{N}$ such that for $n \geq M$,

$$\log_\ell(d(n, R)) \leq nE_w(G) + \sqrt{nV_w(G)} \left(Q^{-1} \left(\frac{R}{I(W)} \right) + \epsilon \right),$$

where $d(n, R)$ is the minimum distance of a polar code using the kernel matrix G , with block-length $N = \ell^n$ and rate R . We note that a row weight of the generator matrix is an upper bound of the minimum distance for a linear code, and that the weight of the i -th row of $G^{\otimes n}$ is equal to $\prod_{j=1}^n w_{i_j}(G)$, where i_j is the j th digit of the ℓ -ary representation of $i - 1$. As a result, it is sufficient to prove that, given any $\epsilon > 0$, there exists an integer $M \in \mathbb{N}$ such that for a polar code of block-length $N = \ell^n \geq \ell^M$, rate R and set of chosen indices \mathcal{I} , there exists $i \in \mathcal{I}$ for which the inequality

$$\sum_{j=1}^n \log_{\ell} w_{i_j}(G) \leq nE_w(G) + \sqrt{nV_w(G)} \left(Q^{-1} \left(\frac{R}{I(W)} \right) + \epsilon \right) \quad (4.18)$$

holds. In the proof of Theorem 4.2, we can observe that the key idea is to apply the central limit theorem for the i.i.d. sequence $\{\log S_n = \log D_{B_n}(G)\}_{n \in \mathbb{N}}$. In order to prove Theorem 4.3 we also consider the i.i.d. sequence $\{\log w_{B_n}(G)\}_{n \in \mathbb{N}}$ in addition to $\{\log D_{B_n}(G)\}_{n \in \mathbb{N}}$. Note that the two sequences $\{\log D_{B_n}(G)\}_{n \in \mathbb{N}}$ and $\{\log w_{B_n}(G)\}_{n \in \mathbb{N}}$ are in general correlated since they are both coupled to the same process $\{B_n\}_{n \in \mathbb{N}}$ and they are equal with probability one if and only if $D_i(G) = w_i(G)$ holds for all $i \in \{0, 1, \dots, \ell - 1\}$. In the same manner as the proof of Theorem 4.2, we move on to a more abstract setting. We first introduce a random variable U that takes values in $[1, \infty)$, for which we assume that the expectation and the variance of $\log U$ exist and are denoted by $\mathbb{E}[\log U]$ and $\mathbb{V}[\log U]$, respectively. We also let $\{(S_n, U_n)\}_{n \in \mathbb{N}}$ be i.i.d. drawings of (S, U) , where S is defined as in Definition 4.2. Let $\{(X_n, S_n, U_n)\}_{n \in \mathbb{N}}$ be a random process such that $\{(X_n, S_n)\}_{n \in \mathbb{N}}$ satisfies the conditions (c1) to (c4) together with the additional condition (c5) for $\{(X_n, U_n)\}_{n \in \mathbb{N}}$:

(c5) U_n is independent of X_m for $m \leq n$.

It is easy to see that the stochastic process of the triplets $\{(Z_n, D_{B_n}(G), w_{B_n}(G))\}_{n \in \mathbb{N}}$ satisfies (c1) to (c5). We first note from the proof of Theorem 4.3 that for any generic process $\{(X_n, S_n, U_n)\}_{n \in \mathbb{N}}$ satisfying (c1) to (c5), relation (4.9) holds for any function $f(n) = o(\sqrt{n})$. We also find that for real numbers v, t such that $v > t$ and for any function $g(n) = o(\sqrt{n})$ we have

$$\limsup_{n \rightarrow \infty} \Pr \left(X_n \leq 2^{-2^{n\mathbb{E}[\log S] + t\sqrt{n\mathbb{V}[\log S]} + f(n)} \right),$$

$$\sum_{i=0}^{n-1} \log U_i > n\mathbb{E}[\log U] + v\sqrt{n\mathbb{V}[\log U]} + g(n) \Big) < \Pr(X_{\infty} = 0)Q(v). \quad (4.19)$$

Before proving (4.19) let us see how it leads to the proof of Theorem 4.3. Since the stochastic process of the triplets $\{(Z_n, D_{B_n}(G), w_{B_n}(G))\}_{n \in \mathbb{N}}$ satisfies (c1) to (c5), we can use relations (4.9) and (4.19) by letting $(X_n, S_n, U_n) = (Z_n, D_{B_n}(G), w_{B_n}(G))$. Now, by (4.9) and (4.19) it is easy to see that for generator matrices of polar codes with rate R , the number of rows satisfying (4.18) is asymptotically proportional to the block-length, hence there exists at least

one row satisfying (4.18) which concludes the proof of Theorem 4.3. Thus, it remains to prove the relation (4.19).

Lemma 4.4. *Let $\{(X_n, S_n, U_n)\}_{n \in \mathbb{N}}$ be a random process satisfying (c1) to (c5). For any $f(n) = o(\sqrt{n})$ and $g(n) = o(\sqrt{n})$,*

$$\begin{aligned} \lim_{n \rightarrow \infty} \Pr \left(X_n \leq 2^{-2^{n\mathbb{E}[\log S] + t\sqrt{n\mathbb{V}[\log S]} + f(n)}}, \right. \\ \left. \sum_{i=0}^{n-1} \log U_i > n\mathbb{E}[\log U] + v\sqrt{n\mathbb{V}[\log U]} + g(n) \right) \\ = \Pr(X_\infty = 0)\Pr(A_S \geq t, A_U \geq v), \end{aligned}$$

where (A_S, A_U) are Gaussian random variables of mean zero whose covariance matrix is equal to that of

$$\left(\frac{\log S - \mathbb{E}[\log S]}{\sqrt{\mathbb{V}[\log S]}}, \frac{\log U - \mathbb{E}[\log U]}{\sqrt{\mathbb{V}[\log U]}} \right).$$

The proof of this Lemma is the same as the proofs of Lemma 4.1 and Lemma 4.2. The difference is that the central limit theorem is replaced by the two-dimensional central limit theorem. From the fact that $\Pr(A_S \geq t, A_U \geq v) \leq Q(\max\{t, v\})$, relation (4.19) is obtained for $v > t$. This completes the proof of Theorem 4.3.

4.4 Further Remarks

In this section we will explain two further (and indirect) implications of Theorems 4.2 and 4.3.

4.4.1 The Common Indices between Polar and Reed-Muller Codes

We begin this section by stating a corollary that is deduced from the proof of Theorem 4.3.

Corollary 4.2. *Assuming $G = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}$, the fraction of the common chosen row indices of $G^{\otimes n}$ between polar codes of rate R and RM codes of rate R' tends to $I(W) \min\{\frac{R}{I(W)}, R'\}$ as $n \rightarrow \infty$.*

Let $G = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}$. For this choice of G , we have $w_i(G) = D_i(G)$ for $i = 0$ and 1. Hence, the random variables $S_n = D_{B_n}(G)$ and $U_n = w_{B_n}(G)$ are equal for all $n \in \mathbb{N}$. Also note that S_n takes its value in the set $\{1, 2\}$ uniformly at random. From the proof of Theorem 4.3, the set of indices of the rows of polar codes with the kernel matrix G and rate R corresponds to the event

$$\left\{ X_n \leq 2^{-2^{n\mathbb{E}[\log S] + Q^{-1}\left(\frac{R}{I(W)}\right)\sqrt{n\mathbb{V}[\log S]} + f(n)}} \right\}.$$

Also, with the same G , the set of indices of a RM code with rate R' corresponds to the event

$$\left\{ \sum_{i=0}^{n-1} \log U_i > n\mathbb{E}[\log U] + Q^{-1}(R')\sqrt{n\mathbb{V}[\log U]} + g(n) \right\}.$$

From Lemma 4.4, it is now easy to conclude that the fraction of the common chosen row indices of $G^{\otimes n}$ between polar codes of rate R and RM codes of rate R' tends to $I(W) \min\{\frac{R}{I(W)}, R'\}$ as $n \rightarrow \infty$.

4.4.2 Selection Rule of the Rows

The proof of Lemma 4.1 suggests a way to help us select the good indices in a computationally efficient way. Let us recall that the construction of polar codes relies on finding the “quality” of the channels $W_{\ell^n}^{(i)}$, $1 \leq i \leq \ell^n$. “Quality” here either refers to large capacity (capacity close to 1) or small Bhattacharyya value. We also recall from Section 2.4 that for each i , $1 \leq i \leq N$, the channel $W_{\ell^n}^{(i)}$ is constructed from W as follows. We first compute the ℓ -ary representation of $i - 1$, which is denoted by $b_1 b_2 \cdots b_n$, with b_1 being the most significant digit and

$$W_{\ell^n}^{(i)} = (((W^{b_1})^{b_2}) \cdots)^{b_n}.$$

The goal of this section is to show that the quality of a channel $W_{\ell^n}^{(i)}$ depends on the channel W only through the first $o(n)$ digits of the sequence $b_1 b_2 \cdots b_n$. In other words, to choose the indices of the channels $W_{\ell^n}^{(i)}$, $1 \leq i \leq N$, which have the best quality, the first $o(n)$ significant digits of the ℓ -ary expansion of $i - 1$ should be determined, depending on W and the rest are determined in a Reed-Muller fashion (i.e., are chosen according to their Hamming weight).

In the proof, the ℓ -ary expansion of row indices of $G^{\otimes n}$ corresponds to realizations of B_1, \dots, B_n . The proof of Lemma 4.1 implies that it is sufficient to select the rows in $\mathcal{D}_m(\beta) \cap \mathcal{H}_m^{n-1}(t)$ in order to achieve the asymptotically optimum performance. It should be noted that the event $\mathcal{D}_m(\beta)$ applied to the Bhattacharyya process $\{Z_n = Z(W_n)\}_{n \in \mathbb{N}}$ of W depends on the channel W , whereas the event $\mathcal{H}_m^{n-1}(t)$ is channel-independent. This observation leads to the following selection rule: The first $m = s(n) \triangleq (\log n + \log \log c)/\beta$ digits of the row indices are determined in the channel-dependent way. Then, the following $(n - m)$ digits are determined in the RM way, i.e., those combinations of digits (B_m, \dots, B_{n-1}) giving large values of $\sum_{i=m}^{n-1} \log D_{B_i}(G)$ are selected. In this rule, only the first $\Theta(\log n)$ digits should be determined depending on the channel.

The above argument can further be extended in a recursive manner. Let $\mathcal{C}_m^{n-1}(\epsilon) \triangleq \{(n - m)^{-1} \sum_{i=m}^{n-1} \log S_i \geq \mathbb{E}[\log S] - \epsilon\}$. Then, it is sufficient to select rows in $\mathcal{D}_{m_0}(\beta) \cap \mathcal{C}_{m_0}^{m_1-1}(\epsilon) \cap \mathcal{H}_{m_1}^{n-1}(t)$ where $m_1 = s(n)$ and $m_0 = s(m_1)$ since $\mathcal{D}_{m_1}(\beta)$ and $\mathcal{D}_{m_0}(\beta) \cap \mathcal{C}_{m_0}^{m_1-1}(\mathbb{E}[\log S] - \beta)$ are asymptotically equal. (Use $\mathcal{C}_m^{n-1}(\epsilon)$ instead of $\mathcal{H}_m^{n-1}(t)$ in the proof of Lemma 4.1. A similar argument can be found in [1, Section IV-B].) From this observation, only $\Theta(\log \log n)$ digits

have to be determined depending on the channel. By iterating this argument, we obtain the selection rule in which only

$$\Theta(\overbrace{\log \cdots \log n}^k) \quad (4.20)$$

digits depend on the channel for any $k \in \mathbb{N}$. From the argument so far, we deduce that even though the behavior of $Z_n = Z(W_n)$ depends on the channel W , as well as the whole sequence $\{B_0, B_1, \dots, B_{n-1}\}$, whether it approaches 0 or 1 when n is large, is mostly determined by the channel W and a prefix of $\{B_0, B_1, \dots, B_{n-1}\}$ with a relatively small length. Thus, to choose the indices of the channels $W_{\ell^n}^{(i)}$ that have the best quality, the first sublinear number of significant digits of the ℓ -ary expansion of $i - 1$ are determined depending on the channel, and the rest are determined in a RM-like fashion. It should be noted that the above argument is valid in the large- n asymptotics. This does not mean that one can make arbitrarily small the number of digits that are to be determined in the channel-dependent manner.

Although the good indices of the rows of $G^{\otimes n}$ can be selected using density evolution [27], in practice storage and convolution of probability density functions is exponentially (in block-length N) costly in terms of memory and computation. Recently, several authors have considered efficient algorithms that closely approximate the density evolution procedure [25], [23]. The above-mentioned construction rule can be useful in reducing the number of convolutions and the number of levels in the quantization of channels.

Efficient Construction and Universality

5

5.1 Problem Formulation

5.1.1 Hardness of the Construction

As explained in Chapter 2 designing a polar code is equivalent to finding the set of good indices. This chapter¹ focuses on the set of good indices and how this set depends on the choice of the channel. For this purpose, given a block-length N , we need to compute certain parameters (e.g., Bhattacharyya or entropy) corresponding to each of the channels $W_N^{(i)}$, $0 \leq i \leq N - 1$. A naive way to compute such a parameter is to compute the transition probabilities of each of the sub-channels $W_N^{(i)}$ and obtain from these probabilities the relevant parameters. The main difficulty in this task is that, since the output alphabet of $W_N^{(i)}$ is $\mathcal{Y}^N \times \{0, 1\}^i$, the cardinality of the output alphabet of the channels at the level n of the infinite binary tree² is doubly exponential in n and therefore exponential in the block-length N . So computing the exact transition probabilities of these channels seems to be intractable. Therefore, we need some efficient methods to “approximate” these channels.

In [1], it is suggested to use a Monte-Carlo method for estimating the Bhattacharyya parameters. Another method in this regard is by *quantization* [24, 25, 27], [47, Appendix B]: approximating the given channel with a channel that has fewer output symbols. More precisely, given a number k , the task is to come up with efficient methods to replace the channels that have more than k outputs with “close” channels that have at most k outputs. A few comments are in order:

¹The material of this part is based on [23] and [24].

²In this chapter, we consider Arikan’s construction of polar codes (see section 2.3.).

- The term “close” depends on the definition of the quantization error, which can be different depending on the context. In our problem, in its most general setting we can define the quantization error as the difference between the true set of good indices and the approximate set of good indices. However, it seems that analyzing this type of error might be difficult and in the sequel we consider types of errors that are easier to analyze.
- As a compromise, we intuitively think of two channels as being close if they are close with respect to some given metric; typically mutual information but sometimes probability of error or Bahttacharyya. More so, we require that this closeness be in the right direction: the approximated channel must be a “pessimistic” version of the true channel so that the approximated set of good channels will be a subset of the true set.
- Intuitively, we expect that as k increases the overall error due to quantization decreases; the main art in designing quantization methods is to have a small error while using relatively small values of k . For any quantization algorithm, however, an important property is that as k grows large, the approximate set of good indices using the quantization algorithm with fixed k approaches the true set of good indices. We give a precise mathematical definition in the following.

Taking the above mentioned factors into account, a suitable formulation of the quantization problem is as follows:

Question 8. *Find procedures to replace each channel P at each level of the binary tree with another symmetric channel \tilde{P} with the number of output symbols limited to k such that*

1. *The new set of good indices obtained with this procedure is a subset of the true good indices obtained from the channel polarization, i.e., channel \tilde{P} is “polar degraded” with respect to P .*
2. *The ratio of these (new) good indices is maximized.*

More precisely, we start from channel W at the root node of the binary tree, quantize it to \tilde{W} and obtain \tilde{W}^- and \tilde{W}^+ according to (2.13) and (2.14). Then, we quantize the two new channels and continue the procedure to complete the tree. To state things mathematically, let Q_k be a quantization procedure that assigns to each channel P a binary symmetric channel \tilde{P} such that the output alphabet of \tilde{P} is limited to a constant k . We call Q_k admissible if for any i and n

$$I(\tilde{W}_N^{(i)}) \leq I(W_N^{(i)}). \quad (5.1)$$

Alternatively, we call Q_k admissible if for any i and n

$$Z(\tilde{W}_N^{(i)}) \geq Z(W_N^{(i)}). \quad (5.2)$$

Note that (5.1) and (5.2) are essentially equivalent as N grows large. Given an admissible procedure Q_k and a BMS channel W , let $\rho(Q_k, W)$ be³

$$\rho(Q_k, W) = \lim_{n \rightarrow \infty} \frac{|\{i : I(\tilde{W}_N^{(i)}) > \frac{1}{2}\}|}{N} \quad (5.3)$$

So, the quantization problem is that given a number $k \in \mathbb{N}$ and a channel W , how can we find admissible procedures Q_k such that $\rho(Q_k, W)$ is maximized and is close to the capacity of W . Can we reach the capacity of W as k goes to infinity? Are such schemes universal, in the sense that they work well for all the BMS channels? It is worth mentioning that if we first let k tend to infinity and then n to infinity, then the limit is indeed the capacity. But we are addressing a different question here, specifically we first let n tend to infinity and then k (or perhaps couple k to n). In Section 5.2.4, we prove that indeed such schemes exist.

5.1.2 Is Polar Coding Universal?

We explained in the previous section that given a BMS channel W , computing the exact transition probabilities of $W_N^{(i)}$ is exponentially hard in N . However, for the BEC this computation is easy (linear in n or logarithmic in N) and all the channels at level n of the tree are again BEC channels for which the erasure can be computed with a simple recursive numerical procedure (as in (2.30)). Knowing that finding the good indices for the BEC is an easy task, one may ask whether there is any connection between the indices of the BEC and any other BMS channels with the same capacity? More generally we can ask:

Question 9. *Given two BMS channels W and W' with equal capacity; is there any relation between the set of good indices of the two channels? What is the ratio of the intersection of the two sets (asymptotically)? Is this ratio equal (or close) to their capacity?*

The answer to Question 9 has several of relevant applications. The first one, as mentioned above, is in finding the set of good indices. There are channels, such as BEC, such that their corresponding set of good indices (or a subset of it) can be found efficiently. Knowing what relation holds between the set of good indices of these special channels and a desired channel, helps us in a more efficient construction of polar codes for the desired channel. The second application is in the design of polar codes for compound or mismatched scenarios. Consider a communication scenario where the transmitter does not know the channel. The only knowledge it has is the set of channels to which the channel belongs. This is known as the *compound channel* scenario. Let \mathcal{W} denote the set of channels. We consider the compound capacity with respect to ignorance at the transmitter, but we allow the decoder to have knowledge of the actual channel. The compound capacity of \mathcal{W} is defined as the rate at

³Instead of $\frac{1}{2}$ in (5.3) we can use any number in $(0, 1)$.

which we can reliably transmit irrespectively of the particular channel (out of \mathcal{W}) that is chosen. The compound capacity is given by [11]

$$C(\mathcal{W}) = \max_P \inf_{W \in \mathcal{W}} I_P(W),$$

where $I_P(W)$ denotes the mutual information between the input and the output of W , with the input distribution being P . Note that the compound capacity of \mathcal{W} can be strictly smaller than the infimum of the individual capacities. This happens if the capacity-achieving input distributions for the individual channels are different. Also note that if the capacity-achieving input distribution is the same for all channels in \mathcal{W} , then the compound capacity is equal to the infimum of the individual capacities. This is indeed the case here because we restrict our attention to the class of BMS channels.

We are interested in the maximum achievable rate by using polar codes and the SC decoder. Let us explain our objective in more detail as follows.

Question 10. *Given a collection \mathcal{W} of BMS channels, we are interested in constructing a polar code of rate R that works well (under SC decoding) for every channel in this collection. This means, given a target block error probability, call it P_B , we ask whether there exists a polar code of rate R such that its block error probability is at most P_B for any channel in \mathcal{W} . In particular, how large can we make R so that a construction exists for any $P_B > 0$?*

Why is this problem of practical relevance? When we design a communications system we typically start with a mathematical model. But in reality no channel is exactly equal to the assumed model. Depending on the conditions of the transmission medium, the channel will show some variations and deviations. Therefore, designing low-complexity coding schemes that are simultaneously reliable for a *set* of channels is a natural and important problem for real systems.

Sections 5.2, 5.3, and 5.4 contain our answer to Questions 8, 9, and 10, respectively. In Section 5.5, we prove that generalizations of polar codes with $\ell \times \ell$ kernels perform better, in terms of the compound rate, than the original polar codes.

5.2 Algorithms for Efficient Construction

Any discrete BMS channel can be represented as a collection of binary symmetric channels (BSC's). The binary input is given to one of these BSC's at random such that the i -th BSC is chosen with probability p_i . The output of this BSC together with its cross over probability x_i is considered as the output of the channel. Therefore, a discrete BMS channel W can be completely described by a random variable $\chi \in [0, 1/2]$. The *pdf* of χ will be of the form

$$P_\chi(x) = \sum_{i=1}^m p_i \delta(x - x_i) \quad (5.4)$$

such that $\sum_{i=1}^m p_i = 1$ and $0 \leq x_i \leq 1/2$ (see Figure 5.1). Note that $Z(W)$ and $1 - I(W)$ are expectations of the functions $f(x) = 2\sqrt{x(1-x)}$ and $g(x) = -x \log(x) - (1-x) \log(1-x)$ over the distribution P_χ , respectively. Therefore,

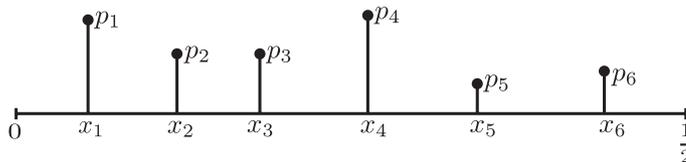


Figure 5.1: A schematic representation of a density that has the form of (5.4)

in the quantization problem we want to replace the mass distribution P_χ with another mass distribution $P_{\tilde{\chi}}$ such that the number of output symbols of $\tilde{\chi}$ is at most k , and the channel \tilde{W} is polar degraded with respect to W . We know that the following two operations imply polar degradation:

- Stochastically degrading the channel.
- Replacing the channel with a BEC channel with the same Bhattacharyya parameter.

Furthermore, note that the *stochastic dominance* of random variable $\tilde{\chi}$ with respect to χ implies \tilde{W} is stochastically degraded with respect to W . (But the reverse is not true.)

In the following, we propose different algorithms based on different methods of polar degradation of the channel. The first algorithm is a naive algorithm, called the mass transportation algorithm, based on the stochastic dominance of the random variable $\tilde{\chi}$. The second one which outperforms the first, is called greedy mass merging algorithm. For both of the algorithms, the quantized channel is stochastically degraded with respect to the original one.

5.2.1 Greedy Mass Transportation Algorithm

In the most general form of this algorithm, we basically look at the problem as a *mass transport* problem. In fact, we have non-negative masses p_i at locations $x_i, i = 1, \dots, m, x_1 < \dots < x_m$. We then require to move the masses only to the right, to concentrate them on $k < m$ locations, and try to minimize $\sum_i p_i d_i$ where $d_i = x_{i+1} - x_i$ is the amount i^{th} mass has moved. Later, we will show that this method is not optimal but useful in the theoretical analysis of the algorithms that follow.

Note that Algorithm 1 is based on the stochastic dominance of random variable $\tilde{\chi}$ with respect to χ . Furthermore, in general, we can let $d_i = f(x_{i+1}) - f(x_i)$, for an arbitrary increasing function f .

Procedure 1 Mass Transportation Algorithm

-
- 1: Start from the list $(p_1, x_1), \dots, (p_m, x_m)$.
 - 2: Repeat $m - k$ times
 - 3: Find $j = \operatorname{argmin}\{p_i d_i : i \neq m\}$
 - 4: Add p_j to p_{j+1} (i.e. move p_j to x_{j+1})
 - 5: Delete (p_j, x_j) from the list.
-

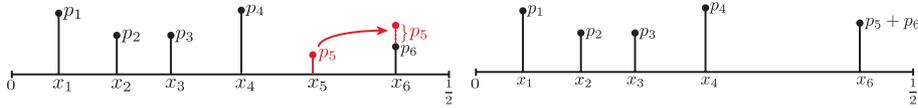


Figure 5.2: The left picture corresponds to how the the mass transportation algorithm is done on the density given in Figure 5.1.

5.2.2 Mass Merging Algorithm

The second algorithm merges the masses. Two masses p_1 and p_2 at positions x_1 and x_2 would be merged into one mass $p_1 + p_2$ at position $\bar{x}_1 = \frac{p_1}{p_1 + p_2}x_1 + \frac{p_2}{p_1 + p_2}x_2$. This algorithm is based on the stochastic degradation of the channel, but the random variable χ is not stochastically dominated by $\tilde{\chi}$. The greedy algorithm for the merging of the masses would be the following:

Procedure 2 Merging Masses Algorithm

-
- 1: Start from the list $(p_1, x_1), \dots, (p_m, x_m)$.
 - 2: Repeat $m - k$ times
 - 3: Find $j = \operatorname{argmin}\{p_i(f(\bar{x}_i) - f(x_i)) - p_{i+1}(f(x_{i+1}) - f(\bar{x}_i)) : i \neq m\}$ $\bar{x}_i = \frac{p_i}{p_i + p_{i+1}}x_i + \frac{p_{i+1}}{p_i + p_{i+1}}x_{i+1}$
 - 4: Replace the two masses (p_j, x_j) and (p_{j+1}, x_{j+1}) with a single mass $(p_j + p_{j+1}, \bar{x}_j)$.
-

Note that in practice, the function f can be any increasing concave function, for example, the entropy function (i.e. $f(x) = h_2(\frac{1-x}{2})$) or the Bhattacharyya functional (i.e. $f(x) = \sqrt{1-x^2}$). In fact, as the algorithm is greedy and suboptimal, it is in general difficult to investigate explicitly how changing the function f will affect the total error of the algorithm in the end (i.e., how far \tilde{W} is from W).

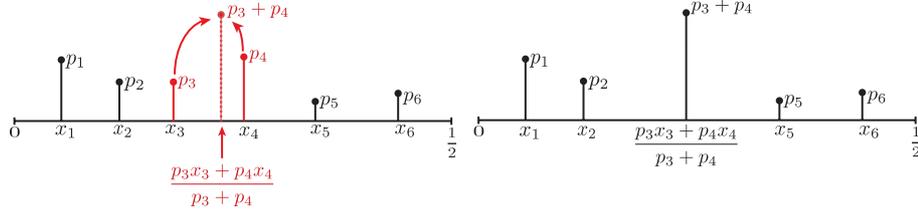


Figure 5.3: The left picture corresponds to how the mass merging algorithm is done on the density given in Figure 5.1.

5.2.3 Bounds on the Approximation Loss

In this section, we provide some bounds on the maximum approximation loss we have in the algorithms. We define the “approximation loss” to be the difference between the expectation of the function f under the true distribution P_χ and the approximated distribution $P_{\hat{\chi}}$. Note that the kind of error that is analyzed in this section is different from what was defined in Section 5.1. The connection of the approximation loss with the quantization error will be made clear in Theorem 5.1. For convenience, we will simply continue using the word “error” instead of “approximation loss” from now on.

We first find an upper bound on the error made in Algorithms 1 and 2 and then use it to provide bounds on the error made while performing operations (2.13) and (2.14).

Lemma 5.1. *The maximum error made by Algorithms 1 and 2 is upper bounded by $\mathcal{O}(\frac{1}{k})$.*

Proof. First, we derive an upper bound on the error of Algorithms 1 and 2 in each iteration, and therefore a bound on the error of the whole process. Let us consider Algorithm 1. The problem can be reduced to the following optimization problem:

$$e = \max_{p_i, x_i} \min_i (p_i d_i) \quad (5.5)$$

such that

$$\sum_i p_i = 1, \quad \sum_i d_i \leq 1, \quad (5.6)$$

where $d_i = f(x_{i+1}) - f(x_i)$, and $f(\frac{1}{2}) - f(0) = 1$ is assumed w.l.o.g. We prove the lemma by Cauchy-Schwarz inequality.

$$\min_i p_i d_i = \left(\sqrt{\min_i p_i d_i} \right)^2 = \left(\min_i \sqrt{p_i d_i} \right)^2 \quad (5.7)$$

Now by applying Cauchy-Schwarz we have

$$\sum_{i=1}^m \sqrt{p_i d_i} \leq \left(\sum_{i=1}^m p_i \right)^{1/2} \left(\sum_{i=1}^m d_i \right)^{1/2} \leq 1 \quad (5.8)$$

Since the sum of m terms $\sqrt{p_i d_i}$ is less than 1, the minimum of the terms will certainly be less than $\frac{1}{m}$. Therefore,

$$e = \left(\min \sqrt{p_i d_i} \right)^2 \leq \frac{1}{m^2}. \quad (5.9)$$

For Algorithm 2, achieving the same bound as Algorithm 1 is trivial. Denote $e^{(1)}$ the error made in Algorithm 1 and $e^{(2)}$ the error made in Algorithm 2. Then,

$$e_i^{(2)} = p_i (f(\bar{x}_i) - f(x_i)) - p_{i+1} (f(x_{i+1}) - f(\bar{x}_i)) \quad (5.10)$$

$$\leq p_i (f(\bar{x}_i) - f(x_i)) \quad (5.11)$$

$$\leq p_i (f(x_{i+1}) - f(x_i)) = e_i^{(1)}. \quad (5.12)$$

Consequently, the error generated by running the whole algorithm can be upper bounded by $\sum_{i=k+1}^m \frac{1}{i^2}$ which is $\mathcal{O}(\frac{1}{k})$. \square

What is stated in Lemma 5.1 is a loose upper bound on the error of Algorithm 2. To achieve better bounds, we upper bound the error made in each iteration of the Algorithm 2 as the following:

$$e_i = p_i (f(\bar{x}_i) - f(x_i)) - p_{i+1} (f(x_{i+1}) - f(\bar{x}_i)) \quad (5.13)$$

$$\leq p_i \frac{p_{i+1}}{p_i + p_{i+1}} \Delta x_i f'(x_i) - p_{i+1} \frac{p_i}{p_i + p_{i+1}} \Delta x_i f'(x_{i+1}) \quad (5.14)$$

$$= \frac{p_i p_{i+1}}{p_i + p_{i+1}} \Delta x_i (f'(x_i) - f'(x_{i+1})) \quad (5.15)$$

$$\leq \frac{p_i + p_{i+1}}{4} \Delta x_i^2 |f''(c_i)|, \quad (5.16)$$

where $\Delta x_i = x_{i+1} - x_i$ and (5.14) is due to concavity of function f . Furthermore, (5.16) is by the mean value theorem, where $x_i \leq c_i \leq x_{i+1}$.

If $|f''(x)|$ is bounded for $x \in (0, 1)$, then we can prove that $\min_i e_i \sim \mathcal{O}(\frac{1}{m^3})$ similarly to Lemma 5.1. Therefore the error of the whole algorithm would be $\mathcal{O}(\frac{1}{k^2})$. Unfortunately, this is not the case for either the entropy function or the Bhattacharyya function. However, we can still achieve a better upper bound for the error of Algorithm 2.

Lemma 5.2. *The maximum error made by Algorithm 2 for the entropy function $h(x)$ can be upper bounded by the order of $\mathcal{O}(\frac{\log(k)}{k^{1.5}})$.*

Proof. Let us first find an upper bound for the second derivative of the entropy function. Suppose that $h(x) = -x \log(x) - (1-x) \log(1-x)$. Then, for $0 \leq x \leq \frac{1}{2}$, we have

$$|h''(x)| = \frac{1}{x(1-x)\ln(2)} \leq \frac{2}{x\ln(2)}. \quad (5.17)$$

Using (5.17), we can further upper-bound the minimum error by

$$\min_i e_i \leq \min_i (p_i + p_{i+1}) \Delta x_i^2 \frac{1}{x_i \ln(4)}. \quad (5.18)$$

Now suppose that we have l mass points with $x_i \leq \frac{1}{\sqrt{m}}$ and $m-l$ mass points with $x_i \geq \frac{1}{\sqrt{m}}$. For the first l mass points, we use the upper bound obtained for Algorithm 1. Hence, for $1 \leq i \leq l$ we have

$$\min_i e_i \leq \min_i p_i \Delta h(x_i) \quad (5.19)$$

$$\sim \mathcal{O}\left(\frac{\log(m)}{l^2 \sqrt{m}}\right), \quad (5.20)$$

where (5.19) is due to (5.12) and (5.20) can be derived again by applying Cauchy-Schwarz inequality. Note that this time

$$\sum_{i=1}^l \Delta h(x_i) \leq h\left(\frac{1}{\sqrt{m}}\right) \sim \mathcal{O}\left(\frac{\log(m)}{\sqrt{m}}\right). \quad (5.21)$$

For the $m-l$ mass points we can write

$$\min_i e_i \leq \min_i (p_i + p_{i+1}) \Delta x_i^2 \frac{1}{x_i \ln(4)} \quad (5.22)$$

$$\leq \min_i (p_i + p_{i+1}) \Delta x_i^2 \frac{\sqrt{m}}{\ln(4)} \quad (5.23)$$

$$\sim \mathcal{O}\left(\frac{\sqrt{m}}{(m-l)^3}\right), \quad (5.24)$$

where (5.24) is due to Hölder's inequality as follows:

Let $q_i = p_i + p_{i+1}$. Therefore, $\sum_i (p_i + p_{i+1}) \leq 2$ and $\sum_i \Delta x_i \leq 1/2$.

$$\min_i q_i \Delta x_i^2 = \left(\left(\min_i q_i \Delta x_i^2 \right)^{1/3} \right)^3 = \left(\min_i (q_i \Delta x_i^2)^{1/3} \right)^3 \quad (5.25)$$

Now by applying Hölder's inequality we have

$$\sum_i (q_i \Delta x_i^2)^{1/3} \leq \left(\sum_i q_i \right)^{1/3} \left(\sum_i \Delta x_i \right)^{2/3} \leq 1 \quad (5.26)$$

Therefore,

$$\min_i e_i \leq \sqrt{m} \left(\min_i (q_i \Delta x_i^2)^{1/3} \right)^3 \sim \mathcal{O} \left(\frac{\sqrt{m}}{(m-l)^3} \right). \quad (5.27)$$

Overall, the error made in the first step of the algorithm would be

$$\min_i e_i \sim \min \left\{ \mathcal{O} \left(\frac{\log(m)}{l^2 \sqrt{m}} \right), \mathcal{O} \left(\frac{\sqrt{m}}{(m-l)^3} \right) \right\} \quad (5.28)$$

$$\sim \mathcal{O} \left(\frac{\log(m)}{m^{2.5}} \right). \quad (5.29)$$

Thus, the error generated by running the whole algorithm can be upper bounded by $\sum_{i=k+1}^m \frac{\log(i)}{i^{2.5}} \sim \mathcal{O} \left(\frac{\log(k)}{k^{1.5}} \right)$. □

We can see that the error is improved by a factor of $\frac{\log k}{\sqrt{k}}$ in comparison with Algorithm 1.

Now we use the result of Lemma 5.1 to provide bounds on the total error made in estimating the mutual information of a channel after n levels of operations (2.13) and (2.14).

Theorem 5.1. *Assume W is a BMS channel and using Algorithm 1 or 2 we quantize the channel W to a channel \tilde{W} . Taking $k = n^2$ is sufficient to give an approximation error that decays to zero.*

Proof. First notice that for any two BMS channels W and V , doing the polarization operations (2.13) and (2.14), the following is true:

$$(I(W^0) - I(V^0)) + (I(W^1) - I(V^1)) = 2(I(W) - I(V)) \quad (5.30)$$

Replacing V with \tilde{W} in (5.30) and using the result of Lemma 5.1, we conclude that after n levels of polarization the sum of the errors in approximating the mutual information of the 2^n channels is upper-bounded by $\mathcal{O}(\frac{n2^n}{k})$. In particular, taking $k = n^2$, we can say that the “average” approximation error of the 2^n channels at level n is upper-bounded by $\mathcal{O}(\frac{1}{n})$. Therefore, at least a fraction $1 - \frac{1}{\sqrt{n}}$ of the channels are distorted by at most $\frac{1}{\sqrt{n}}$ i.e., except for a negligible fraction of the channels, the error in approximating the mutual information decays to zero. □

As a result, since the overall complexity of the encoder construction is $\mathcal{O}(k^2 N)$, this leads to “almost linear” algorithms for encoder construction with arbitrary accuracy in identifying good channels.

5.2.4 Exchange of Limits

In this section, we show that there are admissible schemes such that as $k \rightarrow \infty$, the limit in (5.3) approaches $I(W)$ for any BMS channel W . We use the definition stated in (5.2) for the admissibility of the quantization procedure.

Theorem 5.2. *Given a BMS channel W and for large enough k , there exist admissible quantization schemes Q_k such that $\rho(Q_k, W)$ is arbitrarily close to $I(W)$.*

Proof. Consider the following algorithm: The algorithm starts with a quantized version of W and it does the normal channel splitting transformation followed by quantization according to Algorithm 1 or 2, but once a sub-channel is sufficiently good, in the sense that its Bhattacharyya parameter is less than an appropriately chosen parameter δ , the algorithm replaces the sub-channel with a binary erasure channel which is degraded (polar degradation) with respect to it (as the operations (2.13) and (2.14) over an erasure channel also yield an erasure channel, no further quantization is needed for the children of this sub-channel).

Since the ratio of the total good indices of $\text{BEC}(Z(P))$ is $1 - Z(P)$, then the total error that we make by replacing P with $\text{BEC}(Z(P))$ is at most $Z(P)$ which in the above algorithm is less than the parameter δ .

Now according to Theorem 5.1, for a fixed level n if we make k large enough, then the ratio of the quantized sub-channels, that their Bhattacharyya value is less than δ , approaches its original value (the one obtained with no quantization). For these sub-channels as explained above the total error made with the algorithm is δ . Now, from the polarization theorem and by sending δ to zero, we deduce that as $k \rightarrow \infty$ the number of good indices approaches the capacity of the original channel. \square

5.2.5 Simulation Results

In order to evaluate the performance of our quantization algorithms, similarly to [25], we compare the performance of the degraded quantized channel with the performance of an upgraded quantized channel. An algorithm similar to Algorithm 2 for upgrading a channel is the following. Consider three neighboring masses in positions (x_{i-1}, x_i, x_{i+1}) with probabilities (p_{i-1}, p_i, p_{i+1}) . Let $t = \frac{x_i - x_{i-1}}{x_{i+1} - x_{i-1}}$. Then, we split the middle mass at x_i to the other two masses such that the final probabilities will be $(p_{i-1} + (1-t)p_i, p_{i+1} + tp_i)$ at positions (x_{i-1}, x_{i+1}) . The greedy channel upgrading procedure is described in Algorithm 3.

The same upper bounds on the error of this algorithm can be provided similarly to Section 5.2.3 with a little bit of modification.

In the simulations, we measure the maximum achievable rate while maintaining the probability of error less than 10^{-3} . This is accomplished by finding the maximum possible number of channels with the smallest Bhattacharyya parameters such that the sum of their Bhattacharyya parameters is upper

Procedure 3 Splitting Masses Algorithm

-
- 1: Start from the list $(p_1, x_1), \dots, (p_m, x_m)$.
 - 2: Repeat $m - k$ times
 - 3: Find $j = \operatorname{argmin}\{p_i(f(x_i) - tf(x_{i+1}) - (1-t)f(x_{i-1})) : i \neq 1, m\}$
 - 4: Add $(1-t)p_j$ to p_{j-1} and tp_j to p_{j+1} .
 - 5: Delete (p_j, x_j) from the list.
-

bounded by 10^{-3} . The initial channel is a binary symmetric channel with capacity 0.5. Using Algorithms 2 and 3 for degrading and upgrading the channels with the Bhattacharyya function $f(x) = 2\sqrt{x(1-x)}$, we obtain the following results:

k	2	4	8	16	32	64
degrade	0.2895	0.3667	0.3774	0.3795	0.3799	0.3800
upgrade	0.4590	0.3943	0.3836	0.3808	0.3802	0.3801

Table 5.1: Achievable rate with error probability at most 10^{-3} vs. maximum number of output symbols k for block-length $N = 2^{15}$

We recall that the algorithm runs in complexity $\mathcal{O}(k^2N)$. Table 5.1 shows the achievable rates for Algorithms 2 and 3 when the block-length is fixed to $N = 2^{15}$ and k changes in the range of 2 to 64.

It can be seen from Table 5.1 that the difference of achievable rates within the upgraded and degraded version of the scheme is as small as 10^{-4} for $k = 64$. We expect that for a fixed k , as the block-length increases the difference will also increase (see Table 5.2).

n	5	8	11	14	17	20
degrade	0.1250	0.2109	0.2969	0.3620	0.4085	0.4403
upgrade	0.1250	0.2109	0.2974	0.3633	0.4102	0.4423

Table 5.2: Achievable rate with error probability at most 10^{-3} vs. block-length $N = 2^n$ for $k = 16$

However, in our scheme this difference will remain small even as N grows arbitrarily large, as predicted by Theorem 5.2. (see Table 5.3).

n	21	22	23	24	25
degrade	0.4484	0.4555	0.4616	0.4669	0.4715
upgrade	0.4504	0.4575	0.4636	0.4689	0.4735

Table 5.3: Achievable rate with error probability at most 10^{-3} vs. block-length $N = 2^n$ for $k = 16$

We see that the difference between the rate achievable in the degraded channel and upgraded channel gets constant 2×10^{-3} even after 25 levels of polarizations for $k = 16$.

5.3 Polar Codes with SC Decoding Are Not Universal

Consider two BMS channels P and Q . We are interested in finding the ratio of the common good indices of the two channels. In other words, we are interested in constructing a common polar code of rate R (of arbitrarily large block length) that allows reliable transmission (with the SC decoder) over both channels. Let us denote this ratio by $C_{P, SC}(P, Q)$. Trivially,

$$C_{P, SC}(P, Q) \leq \min\{I(P), I(Q)\}. \quad (5.31)$$

We will see shortly that, properly applied, this simple fact can be used to give tight bounds.

For the lower bound we claim that

$$\begin{aligned} C_{P, SC}(P, Q) &\geq C_{P, SC}(\text{BEC}(Z(P)), \text{BEC}(Z(Q))) \\ &= 1 - \max\{Z(P), Z(Q)\}. \end{aligned} \quad (5.32)$$

To see this claim, we proceed as follows. Consider a particular computation tree of height n with observations at its leaf nodes from a BMS channel with Bhattacharyya constant Z . What is the largest value that the Bhattacharyya constant of the root node can take on? From the extremes of information combining framework ([47, Chapter 4]) we can deduce that we get the largest value if we take the BMS channel to be the $\text{BEC}(Z)$. This is true, since at variable nodes the Bhattacharyya constant acts multiplicatively for any channel, and at check nodes the worst input distribution is known to be the one from the family of BEC channels. Further, BEC densities stay preserved within the computation graph.

The above considerations give rise to the following transmission scheme. We signal on the channels W^σ that are reliable for the $\text{BEC}(\max\{Z(P), Z(Q)\})$. A fortiori these channels are also reliable for the actual input distribution. In this way we can achieve a reliable transmission at rate $1 - \max\{Z(P), Z(Q)\}$.

Example 5.1 (BSC and BEC). *Let us apply the above mentioned bounds to $C_{P, SC}(P, Q)$, where $P = \text{BEC}(0.5)$ and $Q = \text{BSC}(0.11002)$. We have*

$$\begin{aligned} I(P) &= I(Q) = 0.5, \\ Z(\text{BEC}(0.5)) &= 0.5, \\ Z(\text{BSC}(0.11002)) &= 2\sqrt{0.1102(1 - 0.11002)} \approx 0.6258. \end{aligned}$$

The upper bound (5.31) and the lower bound (5.32) then translate to

$$\begin{aligned} C_{P, SC}(P, Q) &\leq \min\{0.5, 0.5\} = 0.5, \\ C_{P, SC}(P, Q) &\geq 1 - \max\{0.6258, 0.5\} = 0.3742. \end{aligned}$$

Note that the upper bound is trivial, but the lower bound is not.

In some special cases the best achievable rate is easy to determine. This happens in particular if the two channels are ordered by degradation.

Example 5.2 (BSC and BEC Ordered by Degradation). *Let $P = BEC(0.22004)$ and $Q = BSC(0.11002)$. We have $I(P) = 0.770098$ and $I(Q) = 0.5$. Further, one can check that the $BSC(0.11002)$ is degraded with respect to the $BEC(0.22004)$. This implies that any sub-channel of type σ which is good for the $BSC(0.11002)$, is also good for the $BEC(0.22004)$. Hence,*

$$C_{P,SC}(BEC(0.22004), BSC(0.11002)) = I(Q) = 0.5.$$

More generally, if the channels \mathcal{W} are such that there is a channel $W \in \mathcal{W}$ that is degraded with respect to every channel in \mathcal{W} , then $C_{P,SC}(\mathcal{W}) = C(W) = I(W)$. Moreover, the sub-channels σ that are good for W are good also for all channels in \mathcal{W} .

So far, we have looked at seemingly trivial upper and lower bounds on the compound capacity of two channels. As we will see now, it is quite simple to significantly tighten the result by considering individual branches of the computation tree separately.

Theorem 5.3 (Bounds on Pairwise Compound Rate). *Let P and Q be two BMS channels. Then for any $n \in \mathbb{N}$*

$$C_{P,SC}(P, Q) \leq \frac{1}{2^n} \sum_{i=1}^N \min\{I(P_N^{(i)}), I(Q_N^{(i)})\},$$

$$C_{P,SC}(P, Q) \geq 1 - \frac{1}{2^n} \sum_{i=1}^N \max\{Z(P_N^{(i)}), Z(Q_N^{(i)})\}.$$

Furthermore, the upper as well as the lower bounds converge to the compound capacity as n tends to infinity and the bounds are monotone with respect to n .

Proof. Consider all $N = 2^n$ sub-channels. Note that there are 2^{n-1} such channels that have $b_1 = 0$ and 2^{n-1} such channels that have a $b_1 = 1$. Recall that b_1 corresponds to the type of node at level n .

This level transforms the original channel P into P^0 and P^1 , respectively. Consider first the 2^{n-1} sub-channels that correspond to $b_1 = 1$. Instead of thinking of each tree as a tree of height n with observations from the channel P , think of each of them as a tree of height $n-1$ rooted at P^1 . By applying our previous argument, we see that if we let n tend to infinity then the common capacity for this half of channels is at most $0.5 \min\{I(P^1), I(Q^1)\}$. Clearly the same argument can be made for the second half of channels. This improves the trivial upper bound (5.31) to

$$C_{P,SC}(P, Q) \leq 0.5 \min\{I(P^1), I(Q^1)\} + 0.5 \min\{I(P^0), I(Q^0)\}.$$

Clearly the same argument can be applied to trees of any height n . This explains the upper bound on the compound capacity of the form $\min\{I(P_N^{(i)}), I(Q_N^{(i)})\}$.

In the same way, we can apply this argument to the lower bound (5.32).

From the basic polarization phenomenon we know that for every $\delta > 0$ there exists an $n \in \mathbb{N}$ so that

$$\frac{1}{2^n} |\{i \in \{1, \dots, N\} : I(P_N^{(i)}) \in [\delta, 1 - \delta]\}| \leq \delta/4.$$

Equivalent statements hold for $I(Q_N^{(i)})$, $Z(P_N^{(i)})$, and $Z(Q_N^{(i)})$.

In words, except for at most a fraction δ , all channel pairs $(P_N^{(i)}, Q_N^{(i)})$ have “polarized.” For each polarized pair both the upper and the lower bound are loose by at most δ . Therefore, the gap between the upper and lower bound is at most $(1 - \delta)2\delta + \delta$.

To see that the bounds are monotone, consider a particular index i and denote the binary expansion of $i - 1$ by the sequence σ of length n . Then we have

$$\begin{aligned} & \min\{I(P^\sigma), I(Q^\sigma)\} \\ &= \min\left\{\frac{1}{2}(I(P^{\sigma^0}) + I(P^{\sigma^1})), \frac{1}{2}(I(Q^{\sigma^0}) + I(Q^{\sigma^1}))\right\} \\ &\geq \frac{1}{2} \min\{I(P^{\sigma^0}), I(Q^{\sigma^0})\} + \frac{1}{2} \min\{I(P^{\sigma^1}), I(Q^{\sigma^1})\}. \end{aligned}$$

A similar argument applies to the lower bound. \square

Remark: In general there is no finite n such that either the upper or the lower bound agree exactly with the compound capacity. On the positive side, the lower bounds are constructive and give an actual strategy to construct polar codes of this rate.

Example 5.3 (Compound Rate of BSC(δ) and BEC(ϵ)). *Let us compute upper and lower bounds on $C_{P, SC}(BSC(0.11002), BEC(0.5))$. Note that both the BSC(0.11002) as well as the BEC(0.5) have capacity one-half. Applying the bounds of Theorem 5.3 we get*

n	0	1	2	5	10	15	20
upper bound	0.500	0.482	0.482	0.481	0.481	0.481	0.481
lower bound	0.374	0.407	0.427	0.456	0.471	0.477	0.479

These results suggest that the numerical value of $C_{P, SC}(BSC(0.11002), BEC(0.5))$ is close to 0.481.

5.4 Bounds on Compound Rate of BMS Channels

5.4.1 Trivial Bounds

In the previous example, we considered the compound capacity of two BMS channels. How does the result change if we consider a whole family of BMS channels; e.g., what is $C_{P, SC}(\{\text{BMS}(I = 0.5)\})$?

We currently do not know of a procedure (even numerical) to compute this rate. But it is easy to give some upper and lower bounds.

In particular we have

$$\begin{aligned} C_{\text{P, SC}}(\{\text{BMS}(I = 0.5)\}) &\leq C(\text{BSC}(0.11002), \text{BEC}(0.5)) \\ &\leq 0.4817, \\ C_{\text{P, SC}}(\{\text{BMS}(I = 0.5)\}) &\geq 1 - Z(\text{BSC}(I = 0.5)) \approx 0.374. \end{aligned} \quad (5.33)$$

The upper bound is trivial. The compound rate of a whole class cannot be larger than the compound rate of two of its members. For the lower bound, note that from Theorem 5.3 we know that the achievable rate is at least as large as $1 - \max\{Z\}$, where the maximum is over all channels in the class. As the BSC has the largest Bhattacharyya parameter of all channels in the class of channels with a fixed capacity, the result follows.

5.4.2 A Better Universal Lower Bound

The universal lower bound expressed in (5.33) is rather weak. Let us therefore show how to strengthen it.

Let \mathcal{W} denote a class of BMS channels. From Theorem 5.3 we know that in order to evaluate the lower bound we have to optimize the terms $Z(P^\sigma)$ over the class \mathcal{W} .

To be specific, let \mathcal{W} be $\text{BMS}(I)$, i.e., the space of BMS channels that have capacity I . Expressed in an alternative way, this is the space of distributions that have entropy equal to $1 - I$.

The above optimization is in general a difficult problem. The first difficulty is that the space $\{\text{BMS}(I)\}$ is infinite dimensional. Hence, in order to use numerical procedures, we have to approximate this space by a finite dimensional space. Fortunately, as the space is compact, this task can be accomplished. For example, look at the densities corresponding to the class $\{\text{BMS}(I)\}$ in the $|D|$ -domain ([47, Chapter 4]). In this domain, each BMS channel W is represented by the density corresponding to the probability distribution of $|W(Y | 0) - W(Y | 1)|$, where $Y \sim W(y | 0)$. For example, the $|D|$ -density corresponding to $\text{BSC}(\epsilon)$ is $\Delta_{1-2\epsilon}$.

We quantize the interval $[0, 1]$ using real values $0 = p_1 < p_2 < \dots < p_m = 1$, $m \in \mathbb{N}$. The m -dimensional polytope approximation of $\{\text{BMS}(I)\}$, denoted by \mathcal{W}_m , is the space of all the densities which are of the form $\sum_{i=1}^m \alpha_i \Delta_{p_i}$. Let $\alpha = [\alpha_1, \dots, \alpha_m]^\top$. Then α must satisfy the following linear constraints:

$$\alpha^\top \mathbf{1}_{m \times 1} = 1, \quad \alpha^\top H_{m \times 1} = 1 - I, \quad \alpha_i \geq 0, \quad (5.34)$$

where $H_{m \times 1} = [h_2(\frac{1-p_i}{2})]_{m \times 1}$ and $\mathbf{1}_{m \times 1}$ is the all-one vector.

Due to quantization, there is in general an approximation error.

Lemma 5.3 (*m versus δ*). *Let $a \in \text{BMS}(I)$. Assume a uniform quantization of the interval $[0, 1]$ with m points $0 = p_1 < p_2 < \dots < p_m = 1$. If $m \geq 1 + \frac{1}{1 - \sqrt[4]{1 - \delta^2}}$, then there exists a density $b \in \mathcal{W}_m$ such that $|Z(a \boxtimes a) - Z(b \boxtimes b)| \leq \delta$.*

Proof. For a given density a , let $Q_u(a)(Q_d(a))$ denote the quantized density obtained by mapping the mass in the interval $(p_i, p_{i+1}]$ ($[p_i, p_{i+1})$) to p_{i+1} (p_i). Note that $Q_u(a)$ ($Q_d(a)$) is upgraded (degraded) with respect to a . Thus, $H(Q_u(a)) \leq H(a) \leq H(Q_d(a))$. The Bhattacharyya parameter $Z(a \boxtimes a)$ is given by

$$Z(a \boxtimes a) = \int_0^1 \int_0^1 \sqrt{1 - x_1^2 x_2^2} a(x_1) dx_1 a(x_2) dx_2.$$

Since $\sqrt{1 - x^2}$ is decreasing on $[0, 1]$, we have

$$\begin{aligned} & Z(Q_d(a) \boxtimes Q_d(a)) - Z(a \boxtimes a) \\ & \leq \sum_{i,j=1}^{m-1} \int_{p_i}^{p_{i+1}} \int_{p_j}^{p_{j+1}} \left(\sqrt{1 - p_i^2 p_j^2} - \sqrt{1 - x^2 y^2} \right) \\ & \quad a(x) dx a(y) dy, \\ & Z(a \boxtimes a) - Z(Q_u(a) \boxtimes Q_u(a)) \\ & \leq \sum_{i,j=1}^{m-1} \int_{p_i}^{p_{i+1}} \int_{p_j}^{p_{j+1}} \left(\sqrt{1 - x^2 y^2} - \sqrt{1 - p_{i+1}^2 p_{j+1}^2} \right) \\ & \quad a(x) dx a(y) dy. \end{aligned}$$

Now note that the maximum approximation error, call it δ , happens when xy is close to 1. This maximum error is equal to

$$\sqrt{1 - \left(1 - \left(\frac{1}{m-1}\right)\right)^4} - \sqrt{1 - 1^4}.$$

Solving for m , we see that the quantization error can be made smaller than δ by choosing m such that

$$m \geq 1 + \frac{1}{1 - \sqrt[4]{1 - \delta^2}}. \quad (5.35)$$

Note that if $a \in \mathcal{W}$ then in general neither $Q_d(a)$ nor $Q_u(a)$ are elements of \mathcal{W}_m , since their entropies do not match. In fact, as discussed above, the entropy of $Q_d(a)$ is too high, and the entropy of $Q_u(a)$ is too low. But by taking a suitable convex combination we can find an element $b \in \mathcal{W}_m$ for which $Z(b^{\boxtimes 2})$ differs from $Z(a^{\boxtimes 2})$ by at most δ .

In more detail, consider the function $f(t) = H(tQ_u(a) + (1-t)Q_d(a))$, $0 \leq t \leq 1$. Clearly, f is a continuous function on its domain. Since every density of the form of $tQ_u(a) + (1-t)Q_d(a)$ is upgraded with respect to $Q_d(a)$ and degraded with respect to $Q_u(a)$, we have $Z((Q_u(a))^{\boxtimes 2}) \leq Z((tQ_u(a) + (1-t)Q_d(a))^{\boxtimes 2}) \leq Z((Q_d(a))^{\boxtimes 2})$. As a result: $|Z((tQ_u(a) + (1-t)Q_d(a))^{\boxtimes 2}) - Z(a^{\boxtimes 2})| \leq \delta$. We further have $f(0) = H(Q_u(a)) \leq H(a) \leq H(Q_d(a)) = f(1)$. Thus there exists a $0 \leq t_0 \leq 1$ such that $f(t_0) = H(a) = I$. Hence, $t_0Q_u(a) + (1-t_0)Q_d(a) \in \text{BMS}(I)$ and $t_0Q_u(a) + (1-t_0)Q_d(a) \in \mathcal{W}_m$. Therefore $t_0Q_u(a) + (1-t_0)Q_d(a)$ is the desired density. \square

Example 5.4 (Improved Bound for $BMS(I = \frac{1}{2})$). *Let us derive an improved bound for the class $\mathcal{W} = BMS(I = \frac{1}{2})$. We choose $n = 1$, i.e., we consider channels at level 1 in Theorem 5.3.*

For $\sigma = 0$ the implied operation is \otimes . It is well known that in this case the maximum of $Z(a \otimes a)$ over all $a \in \mathcal{W}$ is achieved for $a = BSC(0.11002)$. The corresponding maximum Z value is 0.3916.

Next consider $\sigma = 1$. This corresponds to the convolution \boxtimes . Based on Lemma 5.3 consider at first the maximization of Z within the class \mathcal{W}_m :

$$\begin{aligned} \text{maximize : } & \sum_{i,j} \alpha_i \alpha_j Z(\Delta_{p_i} \boxtimes \Delta_{p_j}) = \sum_{i,j} \alpha_i \alpha_j \sqrt{1 - (p_i p_j)^2} \\ \text{subject to : } & \alpha^\top \mathbf{1}_{m \times 1} = 1, \alpha^\top H_{m \times 1} = \frac{1}{2}, \alpha_i \geq 0. \end{aligned} \quad (5.36)$$

In the above, since the p_i s are fixed, the terms $\sqrt{1 - (p_i p_j)^2}$ are also fixed. The task is to optimize the quadratic form $\alpha^\top P \alpha$ over the corresponding α polytope, where the $m \times m$ matrix P is defined as $P_{ij} = \sqrt{1 - (p_i p_j)^2}$. We find that this is a convex optimization problem.

To see this, expand $\sqrt{1 - x^2}$ as a Taylor series in the form

$$\sqrt{1 - x^2} = 1 - \sum_{l \geq 0} t_l x^{2l}, \quad (5.37)$$

where the $t_l \geq 0$. We further have

$$\alpha^\top P \alpha = \sum_{i,j} \alpha_i \alpha_j \sqrt{1 - (p_i p_j)^2} = 1 - \sum_{l \geq 0} t_l \left(\sum_i \alpha_i p_i^{2l} \right)^2. \quad (5.38)$$

Thus, since $t_l \geq 0$ and the p_i s are fixed, each of the terms $-t_l (\sum_i \alpha_i p_i^{2l})^2$ in the above sum represents a concave function. As a result the whole function is concave.

To find a bound, let us relax the condition $0 \leq \alpha_i \leq 1$ and admit $\alpha \in \mathbb{R}$. We are thus faced with solving the convex optimization problem

$$\begin{aligned} \text{maximize : } & \alpha^\top P \alpha \\ \text{subject to : } & \alpha^\top \mathbf{1}_{m \times 1} = 1, \alpha^\top H_{m \times 1} = \frac{1}{2}. \end{aligned}$$

The Kuhn-Tucker conditions for this problem yield

$$\begin{bmatrix} 2P & 1 & H \\ 1^\top & 0 & 0 \\ H^\top & 0 & 0 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_n \\ \lambda_1 \\ \lambda_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \\ \frac{1}{2} \end{bmatrix}. \quad (5.39)$$

As P is non-singular, the answer to the above set of linear equations is unique.

We can now numerically compute this upper bound and from Lemma 5.3 we have an upper bound on the estimation error due to quantization. We get an approximate value of 0.799. We conclude that

$$\begin{aligned} C_{P, SC}(\{BMS(I = 0.5)\}) &\geq 1 - \frac{1}{2}(0.392 + 0.799) \\ &= 0.404. \end{aligned}$$

This slightly improves on the value 0.374 in (5.33). In principle even better bounds can be derived by considering values of n beyond 1. But the implied optimization problems that need to be solved are non-trivial.

5.5 Extensions and Improvements

Given the fact that polar codes with successive decoding are not universal, one might wonder if this lack of universality is due to the code structure or due to the (suboptimal) successive decoding procedure. To answer this question, let us consider polar codes under MAP decoding. Let $I \in [0, 1]$ and consider the polar code (with the standard kernel G_2) designed for the BSC with capacity I . It is shown in [14] that under MAP decoding such a code achieves the capacity of any BMS channel of capacity I . Consequently, polar codes, decoded with the optimal MAP decoder, are universal. Hence, it is the suboptimal decoder that is to fault for the lack of universality. It is therefore interesting to ask (i) whether there exist universal polar codes with low-complexity decoders? and (ii) whether some suitable modification of the standard polar coding scheme allows us to construct “polar-like” codes which are universal under low-complexity decoding?

One can also consider polar codes with $\ell \times \ell$ kernels and ask whether these codes are universal under SC decoding or other decoding algorithms. Intuitively, for exactly the same reason as for the case $\ell = 2$, we expect that a polar code with a $\ell \times \ell$ kernel G and SC decoder is not universal. That is, for a collection of BMS channel with the same capacity I , there is still a gap between the value I and the best rate that a polar code with kernel G can achieve. In the rest of this section, we show that as ℓ grows large, with high probability (w.h.p) any matrix $G \in \mathcal{G}_\ell$ has a good compound rate. In this regard, the complexity of a polar code with an $\ell \times \ell$ kernel grows in general exponentially with ℓ . Therefore, finding polar codes with $\ell \times \ell$ kernels that have a reasonable complexity and a good compound rate is also an interesting open question.

Let $BMS(I)$ be the set of all BMS channels that have capacity I . Let $R < I$ be the desired rate. We recall from Section 2.4 that Given a BMS channel Q and a polar code based on a matrix $G \in \mathcal{G}_\ell$, the capacity of the channel that the i -th bit sees is $1 - H(U_i|Y_Q^n, U_1, \dots, U_{i-1})$.

Lemma 5.4. For $G \in \mathcal{G}_\ell$ we have w.h.p

$$\sup_{Q \in \text{BMS}(I)} \sum_{i=\lceil \ell(1-R) \rceil}^{\ell} H(U_i | Y_Q^n, U_1, \dots, U_{i-1}) \xrightarrow{\ell \rightarrow \infty} 0. \quad (5.40)$$

In a nutshell, the above lemma states that as $\ell \rightarrow \infty$, for any $Q \in \text{BMS}(I)$ the last ℓI channels, namely $\{Q_{\ell(1-I)}, \dots, Q_\ell\}$ turn out to be nearly noiseless channels (i.e., their capacity is tending to 1) and therefore the remaining $\ell(1-I)$ channels are almost completely noisy (i.e., their capacity is tending to 0).

From Lemma 5.4 it is easy to see that for $\delta > 0$, there exists a $L \in \mathbb{N}$ such that for $\ell \geq L$ we have the following property: if we choose $G \in \mathcal{G}_\ell$ uniformly at random, then with probability at least $1 - \delta$, uniformly for any $Q \in \text{BMS}(I)$ we have $h(Q^i) \leq \delta$ for all $i \in \{\lceil \ell(1-R) \rceil, \dots, \ell\}$, where $h(Q^i)$ denotes the entropy functional of the channel Q^i [47]. Further we have,

Lemma 5.5. Let P be a BMS channel with $h(P) \leq \delta$, then P is upgraded with respect to the BMS channel V_δ whose density in the $|D|$ domain ([47, Chapter 4]) is given by $\sqrt{\delta}\Delta_0 + (1 - \sqrt{\delta})\Delta_{1-\sqrt{\delta}}$.

In other words, by the above lemma there exist a channel V_δ that is degraded with respect to all the BMS channels with entropy less than δ . Now, recall that for $\ell \geq L$ with probability at least $1 - \delta$, uniformly for any $Q \in \text{BMS}(I)$ we have $h(Q_i) \leq \delta$ for all $i \in \{\lceil \ell(1-R) \rceil, \dots, \ell\}$ and since degradation is preserved through the channel splitting procedure, we can say the intersection of the set of good indices of all these channels contains the set of good indices of the channel V_δ . The proof now follows from the fact that $I(V_\delta) = 1 - O(\sqrt{\delta})$ and by tending δ to 0 we get the result.

Robustness of the Successive Cancellation Decoder

6

6.1 Problem Formulation

In this chapter¹, we address one further aspect of polar codes using successive decoding. We ask whether such a coding scheme is *robust*. The standard analysis of polar codes under successive decoding assumes infinite precision arithmetic. In practice, we are not able to provide decoders that perform computations with infinite precision. Nonetheless, decoders that have a smaller number of bits in precision might be preferable due to their efficiency in memory and power consumption. Given the successive nature of the decoder, one might worry how well such polar coding schemes perform under a finite precision decoder.

Question 11. *Consider polar codes with a SC decoders in which computation is performed with a few bits of precision. We ask whether such coding schemes still show any threshold behavior and, if they do, how do their thresholds scale in the number of bits of the decoder?*

In this chapter, we show that polar coding is in fact extremely robust with respect to such kinds of quantization of the SC decoder. In Figure 6.1, we plot the maximum achievable rate by using a simple successive decoder with only three messages, called the decoder with erasures, when transmission takes place over several important channel families. As it is apparent from the figure, in particular for channels with high capacity, the fraction of the capacity that is achieved by this simple decoder is close to 1, i.e., even this extremely simple decoder almost achieves capacity. We further show that, more generally, if we want to achieve a rate which is below capacity by $\delta > 0$, then we need at most $O(\log(1/\delta))$ bits of precision (all the logarithms in this chapter are in base 2).

¹The material of this chapter is based on [28].

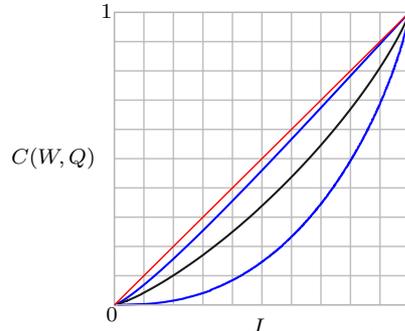


Figure 6.1: The maximum achievable rate, call it $C(W, Q)$, of a simple three message decoder, called the decoder with erasures, as a function of the capacity of the channel for different channel families. From top to bottom: the first curve corresponds to the family of binary erasure channels (BEC) where the decoder with erasures is equivalent to the original SC decoder and, hence, the maximum achievable rate is the capacity itself. The second curve corresponds to the family of binary symmetric channels (BSC). The third curve corresponds to the family of binary additive white Gaussian channels (BAWGN). The curve at the bottom corresponds to a universal lower bound on the achievable rate by the decoder with erasures.

The significance of our observations goes beyond the pure computational complexity required. Typically, the main bottleneck in the implementation of large high-speed coding systems is the memory. Therefore, if we can find decoders that work with only a few bits per message, then this can make the difference whether a coding scheme is implementable or not.

6.1.1 Quantized SC Decoder

Let $\mathbb{R}^* = \mathbb{R} \cup \{\pm\infty\}$ and consider a function $Q(x) : \mathbb{R}^* \rightarrow \mathbb{R}^*$ that is *anti-symmetric* (i.e., $Q(x) = -Q(-x)$). We define the Q -quantized SC decoder as a version of the SC decoder in which the function Q is applied to the output of any computation that the SC decoder does. We denote such a decoder by SCD_Q .

Typically, the purpose of the function Q is to model the case where we only have finite precision in our computations perhaps due to limited available memory or due to other hardware limitations. Hence, the computations are correct within a certain level of accuracy that the function Q models. Thus, let us assume that the range of Q is a finite set \mathcal{Q} with cardinality $|\mathcal{Q}|$. As a result, all the messages passed through the decoder SCD_Q belong to the set \mathcal{Q} .

Here, we consider a simple choice for the function Q that is specified by two parameters: The distance between levels Δ , and truncation threshold M .

Given a specific choice of M and Δ , we define Q as follows:

$$Q(x) = \begin{cases} \lfloor \frac{x}{\Delta} + \frac{1}{2} \rfloor \Delta, & x \in [-M, M], \\ \text{sign}(x)M, & \text{otherwise.} \end{cases} \quad (6.1)$$

Note here that $|Q| = 1 + \frac{2M}{\Delta}$.

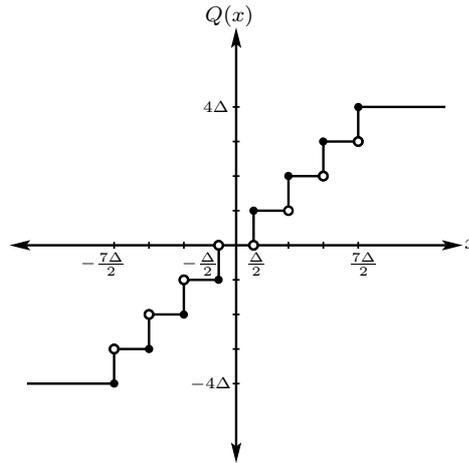


Figure 6.2: A schematic representation of the function $Q(x)$.

6.2 General Framework for the Analysis

6.2.1 Equivalent Tree Channel Model and Analysis of the Probability of Error for the Original SC Decoder

As we are dealing with a linear code, a symmetric channel and symmetric decoders throughout this chapter, without loss of generality we confine ourselves to the *all-zero codeword* (i.e., we assume that all the u_i 's are equal to 0). In order to better visualize the decoding process, the following definition is helpful.

Definition 6.1 (Tree Channels of Height n). *For each $i \in \{0, 1, \dots, N-1\}$, we introduce the notion of the i -th tree channel of height n , which is denoted by $T(i)$. Let $b_1 \dots b_n$ be the n -bit binary expansion of i . E.g., we have for $n = 3$, $0 = 000$, $1 = 001$, \dots , $7 = 111$. With a slight abuse of notation we use i and $b_1 \dots b_n$ interchangeably. Note that for our purpose it is slightly more convenient to denote the least (most) significant bit as b_n (b_1). Each tree channel consists of $n + 1$ levels, namely $0, \dots, n$. It is a complete binary tree. The root is at level n . At level j we have 2^{n-j} nodes. For $1 \leq j \leq n$, if $b_j = 0$*

then all nodes on level j are check nodes; if $b_j = 1$ then all nodes on level j are variable nodes. Finally, we give a label for each node in the tree $T(i)$: For each level j , we label the 2^{n-j} nodes at this level, respectively, from left to right by $(j, 0), (j, 1), \dots, (j, 2^{n-j} - 1)$.

All nodes at level 0 correspond to independent observations of the output of the channel W , assuming that the input is 0.

An example for $T(3)$ (that is $n = 3$, $b = 011$ and $i = 3$) is shown in Fig. 6.3.

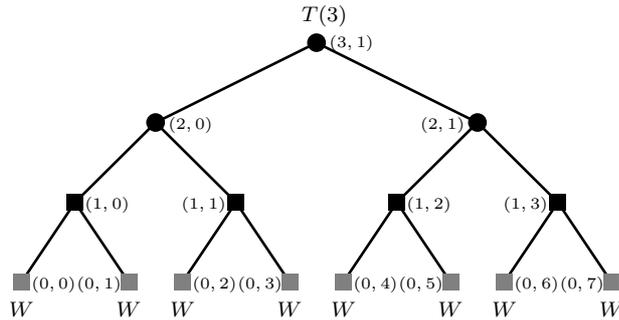


Figure 6.3: Tree representation of the tree-channel $T(3)$. The 3-bit binary expansion of 3 is $b_1 b_2 b_3 = 011$ (note that b_1 is the most significant bit). The pair beside each node is the label assigned to it.

Given the channel output vector y_0^{N-1} and assuming that the values of the bits prior to u_i are given, i.e., $u_0 = 0, \dots, u_{i-1} = 0$, we now compute the probabilities $p(y_0^{N-1}, u_0^{i-1} | u_i = 0)$ and $p(y_0^{N-1}, u_0^{i-1} | u_i = 1)$ via a simple message passing procedure on the equivalent tree channel $T(i)$. We attach to each node in $T(i)$ with label (j, k) a message² $m_{j,k}$ and we update the messages as we go up towards the root node. We start with initializing the messages at the leaf nodes of $T(i)$. For this purpose, it is convenient to represent the channel in the log-likelihood domain; i.e., for the node with label $(0, k)$ at the bottom of the tree that corresponds to an independent realization of W , we plug in the log-likelihood ratio (llr) $\log\left(\frac{W(y_k|0)}{W(y_k|1)}\right)$ as the initial message $m_{0,k}$. That is,

$$m_{0,k} = \log\left(\frac{W(y_k | 0)}{W(y_k | 1)}\right). \quad (6.2)$$

Next, the SC decoder recursively computes the messages (llr's) at each level via the following operations: If the nodes at level j are variable nodes (i.e., $b_j = 1$), we have

$$m_{j,k} = m_{j-1,2k} + m_{j-1,2k+1}, \quad (6.3)$$

²To simplify notation, we drop the dependency of the messages $m_{j,k}$ to the position i whenever it is clear from the context.

and if the nodes at level j are check nodes (i.e., $b_j = 0$), the message that is passed up is

$$m_{j,k} = 2 \tanh^{-1}(\tanh(\frac{m_{j-1,2k}}{2}) \tanh(\frac{m_{j-1,2k+1}}{2})). \quad (6.4)$$

In this way, it can be shown that ([1]) the message that we obtain at the root node is precisely the value

$$m_{n,0} = \log\left(\frac{p(y_0^{N-1}, u_0^{i-1} | u_i = 0)}{p(y_0^{N-1}, u_0^{i-1} | u_i = 1)}\right). \quad (6.5)$$

Now, given (y_0^{N-1}, u_0^{i-1}) , the value of u_i is estimated as follows. If $m_{n,0} > 0$ we let $u_i = 0$. If $m_{n,0} < 0$ we let $u_i = 1$. Finally, if $m_{n,0} = 0$ we choose the value of u_i to be either 0 or 1 with probability $\frac{1}{2}$. Thus, denoting E_i as the event that we make an error on the i -th bit within the above setting, we obtain

$$\Pr(E_i) = \Pr(m_{n,0} < 0) + \frac{1}{2}\Pr(m_{n,0} = 0). \quad (6.6)$$

Given the description of $m_{n,0}$ in terms of a tree channel, it is now clear that we can use density evolution [27] to compute the probability density function of $m_{n,0}$. In this regard, at each level j , the random variables $m_{j,k}$ are i.i.d. for $k \in \{0, 1, \dots, 2^{n-j} - 1\}$. The distribution of the leaf messages $m_{0,k}$ is the distribution of the variable $\log(\frac{W(Y|0)}{W(Y|1)})$, where $Y \sim W(y | 0)$. We can recursively compute the distribution of $m_{j,k}$ in terms of the distribution of $m_{j-1,2k}, m_{j-1,2k+1}$ and the type of the nodes at level j (variable or check) by using the relations (6.3), (6.4) with the fact that the random variables $m_{j-1,2k}$ and $m_{j-1,2k+1}$ are i.i.d.

6.2.2 Quantized Density Evolution

Let us now analyze the density evolution procedure for the quantized decoder. For each label (j, k) in $T(i)$, let $\hat{m}_{j,k}$ represent the messages at this label. The messages $\hat{m}_{j,k}$ take their values in the discrete set \mathcal{Q} (range of the function Q). It is now easy to see that for the decoder SCD_Q the messages evolve via the following relations. At the leaf nodes of the tree we plug in the message $\hat{m}_{0,k} = Q(\log(\frac{W(y_k|0)}{W(y_k|1)}))$, and the update equation for $\hat{m}_{(j,k)}$ is

$$\hat{m}_{j,k} = Q(\hat{m}_{j-1,2k} + \hat{m}_{j-1,2k+1}), \quad (6.7)$$

if the node (j, k) is a variable node and

$$\hat{m}_{j,k} = Q(2 \tanh^{-1}(\tanh(\frac{\hat{m}_{j-1,2k}}{2}) \tanh(\frac{\hat{m}_{j-1,2k+1}}{2}))), \quad (6.8)$$

if the node (j, k) is a check node. We can use the density evolution procedure to recursively obtain the densities of the messages $\hat{m}_{j,k}$.

Finally, let \hat{E}_i denote the event that we make an error in decoding the i -th bit, with a further assumption that we have correctly decoded the previous bits u_0, \dots, u_{i-1} . In a similar way as in the analysis of the original SC decoder, we get

$$\Pr(\hat{E}_i) = \Pr(\hat{m}_{n,0} < 0) + \frac{1}{2}\Pr(\hat{m}_{n,0} = 0). \quad (6.9)$$

Hence, one way to choose the information bits for the algorithm SCD_Q is to choose the bits u_i according to the least values of $\Pr(\hat{E}_i)$.

We remark here that with the decoder SCD_Q , the distribution of the messages in the trees $T(i)$ is different than the corresponding ones that result from the original SC decoder. Hence, the choice of the information indices is also specified by the choice of the function Q , as well as the channel W .

Note here that, since all of the densities takes their value in the finite alphabet \mathcal{Q} , the construction of such polar codes can be efficiently done in time $O(|\mathcal{Q}|^2 N \log N)$. We refer the reader to [1] for more details.

6.3 Quantized SC Decoders with Different Precisions

6.3.1 1 Bit decoder: The Gallager Algorithm

As our aim is to show that polar codes under successive decoding are robust against quantization, let us investigate an extreme case. Perhaps the simplest message-passing type decoder that one can envision is the Gallager algorithm. It works with single-bit messages. Does this simple decoder have a non-zero threshold? Unfortunately it does not, and this is easy to see.

We begin with the equivalent tree-channel model. For each channel i of the polar code we have such a tree of height n , and on each layer nodes are either all check nodes or all variable nodes.

Since messages are only a single bit, the “state” of the decoder at level j can be described by a single non-negative number x_j that is specifically the probability that the message at level j is incorrect. Assume that we transmit over a $\text{BSC}(p)$. Let $x_0 = p \in (0, \frac{1}{2})$. We are interested in the evolution of x_j . This evolution depends of course on the sequence of levels, i.e., it depends on which tree channel we consider.

Assume that x_j is given and that the next level consists of check nodes. In this case, the error probability increases. More precisely, $x_{j+1} = 2x_j(1 - x_j) > x_j$ when $x_j \in (0, \frac{1}{2})$. In other words, the state deteriorates. What happens if the next level consists of variable nodes instead? A little thought shows that in this case $x_{j+1} = x_j$, i.e., there is no change at all. This is true because if both incoming messages agree then we can make a decision on the outgoing message; but if they differ then we can only guess. This gives us $x_{j+1} = x_j^2 + x_j(1 - x_j) = x_j$.

As in either case, the state either becomes worse or stays unchanged, no progress in the decoding is achieved, irrespective of the given tree. In other words, this decoder has a threshold of zero. As we have seen, the problem is the processing at the variable nodes since no progress is achieved there. But since

we only have two incoming messages there are not many degrees of freedom in the processing rules. It is doubtful that any message-passing decoder with only a single-bit message can do better.

6.3.2 1-Bit Decoder with Erasures

Further to the previous example, let us now add one message to the alphabet of the Gallager decoder, i.e., we also add the possibility of having erasures. In this case $Q(x)$ becomes the sign function³, i.e.,

$$Q(x) = \begin{cases} \infty, & x > 0, \\ 0, & x = 0, \\ -\infty, & x < 0. \end{cases} \quad (6.10)$$

As a result, all messages passed by the algorithm SCD_Q take on only three possible values: $\{-\infty, 0, \infty\}$. In this regard, the decoding procedure takes a very simple form. The algorithm starts by quantizing the channel output to one of the three values in the set $\mathcal{Q} = \{-\infty, 0, \infty\}$. At a check node, we take the product of the signs of the incoming messages and, at a variable node, we have the natural addition rule ($0 \leftarrow \infty + -\infty$, $0 \leftarrow 0 + 0$ and $\infty \leftarrow \infty + \infty$, $\infty \leftarrow \infty + 0$ and $-\infty \leftarrow -\infty + -\infty$, $-\infty \leftarrow -\infty + 0$). Note that on the binary erasure channel, this algorithm is equivalent to the original SC decoder.

Our objective is now to compute the maximum reliable rate that the decoder SCD_Q can achieve for a BMS channel W . We denote this quantity by $C(W, Q)$.

Theorem 6.1. *Consider transmission over a BMS channel W of capacity $I(W)$ using polar codes and a SCD_Q with message alphabet Q . Let $C(W, Q)$ denote the maximum rate at which reliable transmission is possible for this setup. Let $|\mathcal{Q}| = 3$. Then there exists a computable decreasing sequence $\{U_n\}_{n \in \mathbb{N}}$ (see (6.22)) and a computable increasing sequence $\{L_n\}_{n \in \mathbb{N}}$ (see (6.23)), so that $L_n \leq C(W, Q) \leq U_n$ and*

$$\lim_{n \rightarrow \infty} L_n = \lim_{n \rightarrow \infty} U_n.$$

In other words, U_n is an upper bound and L_n is a lower bound on the maximum achievable rate $C(W, Q)$ and for increasing n these two bounds converge to $C(W, Q)$.

Discussion: In Figure 6.1 the value of $C(W, Q)$, $|\mathcal{Q}| = 3$, is plotted as a function of $I(W)$ for different channel families (for more details see Section 6.3.2). A universal lower bound for the maximum achievable rate is also given in Figure 6.1.

The rest of this section is devoted to providing the machinery and intuitions to prove Theorem 6.1. The methods used here are extendable to other quantized decoders. The analysis is done in three steps as we will see in the following.

³Note here that we have further assumed that $M = \Delta$ and $\Delta \rightarrow 0$.

The Density Evolution Procedure

To analyze the performance of this algorithm, first note that as all our messages take their values in the set \mathcal{Q} , then all the random variables that we consider have the following form

$$D = \begin{cases} \infty, & \text{w.p. } p, \\ 0, & \text{w.p. } e, \\ -\infty, & \text{w.p. } m. \end{cases} \quad (6.11)$$

Here, the numbers p, e, m are probability values and $p + e + m = 1$. Let us now see how the density evolves through the tree-channels. For this purpose, we should trace the output distribution of (6.7) and (6.8) when the input messages are two i.i.d. copies of a r.v. D with pdf as in (6.11).

Lemma 6.1. *Given two i.i.d. versions of a r.v. D with distribution as in (6.11), the output of a variable node operation (6.7), denoted by D^+ , has the following form*

$$D^+ = \begin{cases} \infty, & \text{w.p. } p^2 + 2pe, \\ 0, & \text{w.p. } e^2 + 2pm, \\ -\infty, & \text{w.p. } m^2 + 2em. \end{cases} \quad (6.12)$$

Also, the check node operation (6.8), yields D^- as

$$D^- = \begin{cases} \infty, & \text{w.p. } p^2 + m^2, \\ 0, & \text{w.p. } 1 - (1 - e)^2, \\ -\infty, & \text{w.p. } 2pm. \end{cases} \quad (6.13)$$

In order to compute the distribution of the messages $\hat{m}_{n,0}$ at a given level n , we use the method of [1] and define the polarization process D_n as follows. Consider the random variable $L(Y) = \log\left(\frac{W(Y|0)}{W(Y|1)}\right)$, where $Y \sim W(y|0)$. The stochastic process D_n starts from the r.v. $D_0 = Q(L(Y))$ defined as

$$D_0 = \begin{cases} \infty, & \text{w.p. } p = \Pr(L(Y) > 0), \\ 0, & \text{w.p. } e = \Pr(L(Y) = 0), \\ -\infty, & \text{w.p. } m = \Pr(L(Y) < 0), \end{cases} \quad (6.14)$$

and for $n \geq 0$

$$D_{n+1} = \begin{cases} D_n^+, & \text{w.p. } \frac{1}{2}, \\ D_n^-, & \text{w.p. } \frac{1}{2}, \end{cases} \quad (6.15)$$

where the plus and minus operations are given in (6.12), (6.13).

Analysis of the Process D_n

Note that the output of process D_n is itself a random variable of the form given in (6.11). Hence, we can equivalently represent the process D_n with a triple (m_n, e_n, p_n) , where the coupled processes m_n, e_n and p_n are evolved using the relations (6.12) and (6.13) and we always have $m_n + e_n + p_n = 1$. Following

along the same lines as the analysis of the original SC decoder in [1], we first claim that as n grows large, the process D_n will become polarized, i.e., the output of the process D_n will almost surely be a completely noiseless or a completely erasure channel.

Lemma 6.2. *The random sequence $\{D_n = (p_n, e_n, m_n), n \geq 0\}$ converges almost surely to a random variable D_∞ such that D_∞ takes its value in the set $\{(1, 0, 0), (0, 1, 0)\}$.*

Our objective is now to compute the value of $C(W, Q) = \Pr(D_\infty = (1, 0, 0))$, i.e., the highest rate that we can achieve with the 1-Bit Decoder with Erasures. In this regard, a convenient approach is to find a function $f : \mathcal{D} \rightarrow \mathbb{R}$ such that $f((0, 1, 0)) = 0$ and $f(1, 0, 0) = 1$ and for any $D \in \mathcal{D}$

$$\frac{1}{2}(f(D^+) + f(D^-)) = f(D).$$

With such a function f , the process $\{f(D_n)\}_{n \geq 0}$ is a martingale and consequently we have $\Pr(D_\infty = (1, 0, 0)) = f(D_0)$. Therefore, by computing the deterministic quantity $f(D_0)$ we obtain the value of $C(W, Q)$. However, finding a closed form for such a function seems to be a difficult task⁴. Instead, the idea is to look for alternative functions, denoted by $g : \mathcal{D} \rightarrow \mathbb{R}$, such that the process $g(D_n)$ is a super-martingale (sub-martingale) and hence we can get a sequence of upper (lower) bounds on the value of $\Pr(D_\infty = (1, 0, 0))$ as follows. Assume we have a function $g : \mathcal{D} \rightarrow \mathbb{R}$ such that $g((0, 1, 0)) = 0$ and $g(1, 0, 0) = 1$ and for any $D \in \mathcal{D}$,

$$\frac{1}{2}(g(D^+) + g(D^-)) \leq g(D). \quad (6.16)$$

Then, the process $\{g(D_n)\}_{n \geq 0}$ is a super-martingale and for $n \geq 0$ we have

$$\Pr(D_\infty = (1, 0, 0)) \leq \mathbb{E}[g(D_n)]. \quad (6.17)$$

The quantity $\mathbb{E}[g(D_n)]$ decreases by n and by using Lemma 6.2 we have

$$\Pr(D_\infty = (1, 0, 0)) = \lim_{n \rightarrow \infty} \mathbb{E}[g(D_n)]. \quad (6.18)$$

In a similar way, we can search for a function $h : \mathcal{D} \rightarrow \mathbb{R}$ such that for h with the same properties as g except that the inequality (6.16) holds in opposite direction, i.e.,

$$\frac{1}{2}(h(D^+) + h(D^-)) \geq h(D). \quad (6.19)$$

In a similar way this leads us to computable lower bounds on $C(W, Q)$. In other words, the process $\{h(D_n)\}_{n \geq 0}$ is a sub-martingale and for $n \geq 0$ we have

$$\Pr(D_\infty = (1, 0, 0)) \geq \mathbb{E}[h(D_n)]. \quad (6.20)$$

⁴The function f clearly exists as one trivial candidate for it is $f(D) = \Pr(D_\infty = (1, 0, 0))$, where D_∞ is the limiting r.v. that the process $\{D_n\}_{n \geq 0}$ with starting value $D_0 = D$ converges to.

We also obtain from Lemma 6.2 that

$$\Pr(D_\infty = (1, 0, 0)) = \lim_{n \rightarrow \infty} \mathbb{E}[h(D_n)]. \quad (6.21)$$

It remains to find some suitable candidates for g and h . Let us first note that a density D as in (6.11) can be equivalently represented as a simple BMS channel given in Fig. 6.4. This equivalence stems from the fact that for such

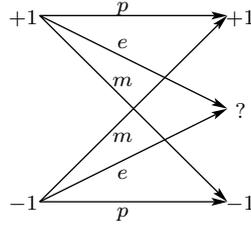


Figure 6.4: The equivalent channel for the density D given in (6.11).

a channel, conditioned on the event that the symbol $+1$ has been sent, the distribution of the output is precisely D . With a slight abuse of notation, we also denote the corresponding BMS channel by D . In particular, it is an easy exercise to show that the capacity ($I(D)$), the Bhattacharyya parameter ($Z(D)$) and the error probability ($E(D)$) of the density D are given as

$$I(D) = (m + p)(1 - h_2(\frac{p}{p + m})),$$

$$Z(D) = 2\sqrt{mp} + e, E(D) = 1 - p - \frac{e}{2},$$

where $h_2(\cdot)$ denotes the binary entropy function. Since the function Q is not an injective function, we have $\frac{I(D^+) + I(D^-)}{2} \leq I(D)$. This implies that the process $I_n = I(D_n)$ is a bounded supermartingale. Furthermore, since $I(D = (1, 0, 0)) = 1$ and $I(D = (0, 1, 0)) = 0$, we deduce from Lemma 6.2 that I_n converges a.s. to a $0 - 1$ valued r.v. I_∞ , hence

$$C(W, Q) = \Pr(D_\infty = (1, 0, 0)) = \Pr(I_\infty = 1) = \mathbb{E}(I_\infty).$$

Now, from the fact that I_n is a supermartingale, we obtain

$$C(W, Q) \leq \mathbb{E}[I_n] \triangleq U_n, \quad (6.22)$$

for $n \in \mathbb{N}$. In a similar way, we obtain a sequence of lower bounds for $C(W, Q)$.

Lemma 6.3. *Define the function $F(D)$ as $F(D) = p - 4\sqrt{pm}$ for $D \in \mathcal{D}$. We have $F(D = (1, 0, 0)) = 1$, $F(D = (0, 1, 0)) = 0$ and $\frac{F(D^+) + F(D^-)}{2} \geq F(D)$.*

Hence, the process $F_n = F(D_n)$ is a submartingale and for $n \in \mathbb{N}$ we have

$$C(W, Q) \geq \mathbb{E}[F_n] \triangleq L_n. \quad (6.23)$$

Given a BMS channel W , one can numerically compute $C(W, Q)$ with arbitrary accuracy using the sequences L_n and U_n (see Figure 6.1). Also, for a channel W with capacity $I(W)$ and error probability $E(W)$, we have

$$E(W) \leq \frac{1 - I(W)}{2}. \quad (6.24)$$

Therefore, $\inf_{\{D: E(D) = \frac{1-I(W)}{2}\}} C(D, Q) \leq C(W, Q)$, which leads to the universal lower bound obtained in Figure 6.1.

Example 6.1. Let the channel W be a BSC channel with cross over probability $\epsilon = 0.11$ (hence $I(W) \approx 0.5$). Using (6.25) we obtain

$$D_0 = \begin{cases} \infty, & \text{w.p. } 1 - \epsilon = 0.89, \\ -\infty, & \text{w.p. } \epsilon = 0.11. \end{cases} \quad (6.25)$$

Therefore, we get $L_0 = F(D_0) = -0.361$ and $U_0 = I(D_0) = 0.5$. We can also compute $L_1 = \frac{F(D_0^+) + F(D_0^-)}{2} = -0.191$, $U_1 = \frac{I(D_0^+) + I(D_0^-)}{2} = .5$ and

$$L_2 = \frac{F(D_0^{++}) + F(D_0^{+-}) + F(D_0^{-+}) + F(D_0^{--})}{4} = -0.075,$$

$$U_2 = \frac{I(D_0^{++}) + I(D_0^{+-}) + I(D_0^{-+}) + I(D_0^{--})}{4} = 0.498.$$

Continuing this way, one can find $L_{10} = 0.264, U_{10} = 0.474$ and $L_{20} = 0.398, U_{20} = 0.465$ and so on.

One can also use other functions to obtain bounds that converge faster than the bounds given above. As an example, the function $I(D)^2$ is also an other suitable choice to obtain lower bounds for the value of $C(W, Q)$. We can check numerically that for the function $I(D)^2$, the inequality (6.19) holds. As a result, the sequence $E[I(D_n)^2]$ is also a sequence of lower bounds on $C(W, Q)$. For this choice we experimentally see that the gap between the lower bound $E[I(D_n)^2]$ and the upper bound $E[I(D_n)]$ vanishes very fast. Proving the relation (6.19) for the function $I(D)^2$ seems to be a challenging task.

Scaling Behavior and the Error Exponent

In the final step, we need to show that for rates below $C(W, Q)$ the block-error probability decays to 0 for large block-lengths.

Lemma 6.4. Let $D \in \mathcal{D}$. We have

$$Z(D^-) \leq 2Z(D) \text{ and } Z(D^+) \leq 2(Z(D))^{\frac{3}{2}}.$$

Hence, for transmission rate $R < C(W, Q)$ and block-length $N = 2^n$, the probability of error of SCD_Q , denoted by $P_{e,Q}(N, R)$ satisfies $P_{e,Q}(N, R) = o(2^{-N^\beta})$ for $\beta < \frac{\log \frac{3}{2}}{2}$.

6.3.3 Scaling of the Gap to Capacity with Respect to the Number of Precision Bits

In the previous section we have considered a particular family of decoders. We have seen that a small number of messages suffice to achieve a considerable fraction of capacity. In this section we will achieve rates as large as $I(W) - d$, where d is a positive (and small) constant. Our objective is to provide bounds on the number of precision bits that are required for this purpose.

Theorem 6.2. *To achieve an additive gap $d > 0$ to capacity $I(W)$, it is sufficient to choose $\log |\mathcal{Q}| = O(\log(\frac{1}{d}))$.*

The rest of this section is devoted to providing a sketch for the proof of Theorem 6.2. Consider a BMS channel W and assume that we need an algorithm SCD_Q capable of achieving rates up to $I(W) - d$, where $d \leq \frac{1}{2}$ is a positive constant (for $d \geq \frac{1}{2}$ the 1-bit decoder with erasures is already a good choice). Our goal is to find suitable parameters M and Δ so that the algorithm SCD_Q is capable of achieving a rate at least $I(W) - d$. We denote the maximum achievable rate of the algorithm SCD_Q by $C(W, Q)$. In order to compute $C(W, Q)$, we should precisely compute the ratio of the good indices among the set $\{0, 1, \dots, N - 1\}$ when N grows large. Here, we do not intend to compute the precise value of $C(W, Q)$ but rather provide a universal lower bound on $C(W, Q)$ that is already applicable for proving the theorem.

Let us first give a very broad picture behind the proof. We first consider the original SC decoder and choose an integer n_d large enough so that for $n \geq n_d$, at least a fraction $I(W) - \frac{d}{2}$ of the sub-channels at level n have Bhattacharyya value less than e^{-2n} . More precisely, we have for $n \geq n_d$

$$\Pr(Z_n \leq e^{-2n}) \geq I(W) - \frac{d}{2}. \quad (6.26)$$

As a result, if we perform the original SC decoding, then at level n at least a fraction $I(W) - \frac{d}{2}$ of the sub-channels are very perfect. Let $\mathcal{I}_{n,d}$ denote the set of indices of these sub-channels. In the second step, we tune the parameters M and Δ for a decoder SCD_Q (with function Q given in (6.1)) in a way that the algorithm SCD_Q still decodes perfectly on the indices that belong to the set $\mathcal{I}_{n,d}$. In other words, we intend to find candidates for M and Δ in terms of n so that the messages that we get by the algorithm SCD_Q , with such candidates for M and Δ , are suitably close to their counterpart in the original SC decoder. In the last step, we show that the sub-channels branched from the indices in $\mathcal{I}_{n,d}$ are still good enough so that we can, as n grows large, achieve a fraction $I(W) - d$ of very perfect channels. The proof consists of three steps.

First step: (How to choose M and Δ ?) The primary problem we consider here is as follows: Consider a specific realization of independent uses of the channel W at each of the leaves of the tree; by using the original SC decoder, this realization results in a specific value at the root node. Now, consider the same recursive computation process with the following extra operations of the value that come out of each computation:

1. After each of the computations, we also perturb the resulting value by at most a fixed value Δ .
2. If the absolute value of the output is larger than a fixed value M , we replace the value by $\pm\infty$ according to its sign.

It is easy to see that the operations (1) and (2) are given to better analyze the algorithm SCD_Q . In this regard, how should we choose the values of M and Δ so that the final message that is computed at the top of the tree, i.e., $\hat{m}_{n,0}$ is not too far from its counterpart in the original SC decoder, i.e., $m_{n,0}$? First assume $M = \infty$. As a result, the operation (2) is not applied anymore. Straight forward computation shows that the partial derivatives of the functions $v(x, y)$ and $c(x, y)$ (which correspond to (6.3) and (6.4), respectively) are given by

$$v(x, y) := x + y, \quad (6.27)$$

$$c(x, y) := 2 \tanh^{-1}(\tanh(\frac{x}{2}) \tanh(\frac{y}{2})), \quad (6.28)$$

are always bounded above by 1. Hence, for $a, b \in \mathbb{R}$, we have

$$|v(x+a, x+b) - v(x, y)| \leq |a| + |b|, \quad (6.29)$$

$$|c(x+a, x+b) - c(x, y)| \leq |a| + |b|. \quad (6.30)$$

As a result, it is easy to see that assuming that only operation (1) is applied, the cumulative error we get on the top of the tree $T(i)$ is upper bounded by $\Delta 2^{n+1}$. Hence, the following lemma follows.

Lemma 6.5. *Consider a quantized SC algorithm in which $M = \infty$ (i.e., only operation (1) is applied). Also, consider the i -th position among the information bits with its corresponding binary tree $T(i)$. Then, for any realization of the channel outputs we have $|m_{j,k} - \hat{m}_{j,k}| \leq 2^{j+1}\Delta$ for any label $(j, k) \in T(i)$. As a result, if we choose $\Delta \leq 2^{-(n+1)}$, then $|m_{n,0} - \hat{m}_{n,0}| \leq 1$.*

Let us now assume that M is finite, hence the operation (2) is a non-trivial operation. Of course, depending on the value of M , the cumulative error varies in a large range. It seems that, in this case, providing worse case bounds as in Lemma 6.5 is a difficult task. Consequently, we seek bounds that hold with high probability.

Lemma 6.6. *Let $M = 4n$ and $\Delta = 2^{-(n+1)}$. Then with probability at least $1 - 16(n+1)(\frac{2}{e})^{2n}$, the following holds: If $\hat{m}_{n,0} \neq \infty$ then $|m_{n,0} - \hat{m}_{n,0}| \leq 1$.*

Second Step: (What happens to the almost perfect channels?) Let us now fix $n \geq n_d$ and consider the algorithm SCD_Q with parameters M and Δ as given in Lemma 6.6. In this step, we provide a lower bound on the value of $C(W, Q)$, which is equal to the final ratio of the good indices. In order to do this, we provide a lower bound only on the final ratio of the good indices that are branched out from the indices in the set $\mathcal{I}_{n,d}$. First, we consider the original SC decoder. By definition, we have for each index $i \in \mathcal{I}_{n,d}$ that

$Z(W_{N_d}^{(i)}) \leq e^{-2n}$, where $N_d = 2^{n_d}$. Assuming that the Bhattacharyya value of the distribution of $m_{n,0}$ is less than 2^{-2n} , we obtain

$$\Pr(m_{n,0} \geq 2n + 1) \geq 1 - e^{1-n}. \quad (6.31)$$

Now, by using Lemma 6.6 and (6.31), at level n with probability at least $1 - e^{1-n} - 16(n+1)(\frac{2}{e})^{2n} \geq 1 - 16(n+2)(\frac{2}{e})^{2n}$, at an index $i \in \mathcal{I}_{n,d}$, the algorithm SCD_Q outputs the $+\infty$ message. This implies that at $i \in \mathcal{I}_{n,d}$ the distribution of the messages that we get by the algorithm SCD_Q stochastically dominates the following distribution

$$D = \begin{cases} \infty & \text{w.p. } 1 - 16(n+2)(\frac{2}{e})^{2n}, \\ -\infty & \text{w.p. } 16(n+2)(\frac{2}{e})^{2n}. \end{cases} \quad (6.32)$$

Now, let C_i be the final ratio of the perfect sub-channels that are branched from $i \in \mathcal{I}_{n,d}$. It is now easy to see that C_i is lower bounded by the ratio that we get by plugging the density D , given in (6.32), into the 1-bit decoder with erasures. In this way, by using Lemma 6.3 we obtain for $i \in \mathcal{I}_{n,d}$

$$C_i \geq p - 4\sqrt{pm} \geq 1 - 16(n+2)(\frac{2}{e})^{2n} - 16\sqrt{n+2}(\frac{2}{e})^n. \quad (6.33)$$

We thus obtain from (6.26) and (6.33)

$$C(W, Q) \geq (I(W) - \frac{d}{2})(1 - 16(n+2)(\frac{2}{e})^{2n} - 16\sqrt{n+2}(\frac{2}{e})^n). \quad (6.34)$$

Third Step: (Putting things together). In the last step, we relate the values d , n_d and the lower bound (6.34) together. We first choose $n_1 \in \mathbb{N}$ such that for $n \geq n_1$ we have

$$16(n+2)(\frac{2}{e})^{2n} + 16\sqrt{n+2}(\frac{2}{e})^n \leq \frac{d}{2}. \quad (6.35)$$

One can easily see that for small values of d , a suitable candidate for n_1 is $n_1 = \frac{1}{\log(\frac{1}{2})} \log(\frac{1}{d}) + o(\log(\frac{1}{d}))$. However, to have an explicit candidate for n_1 such that (6.35) holds for all values of d , one can fix

$$n_1 = 3 \log(\frac{1}{d}) + 17. \quad (6.36)$$

Now, let $n = \max(n_1, n_d)$. From (6.34) and (6.35) it is easy to see that $C(W, Q) \geq I(W) - d$. In other words, by choosing $M = 2n$ and $\Delta = 2^{-(n+1)}$ for the function Q given in (6.1), the algorithm SCD_Q is capable of achieving rates that satisfy $C(W, Q) \geq I(W) - d$. Also, note that we have

$$|\mathcal{Q}| = 1 + \frac{2M}{\Delta} = 1 + n2^{n+2}.$$

As a result,

$$\log |\mathcal{Q}| \approx n + \log n + 2. \quad (6.37)$$

Finally, what remains to be done is to relate n_d to d .

Lemma 6.7. *In order to have (6.26) for $n \geq n_d$, it is enough to let*

$$n_d = 7 \log\left(\frac{1}{d}\right) + \log\left(\log\left(\frac{2}{d}\right)\right)^2 + 48. \quad (6.38)$$

With such a choice of n_d and n_1 as in (6.38) and (6.36), we have $n_d \geq n_1$ and $n = n_d$. Thus, we obtain from (6.37)

$$\log |\mathcal{Q}| \leq 7 \log\left(\frac{1}{d}\right) + O\left(\log\left(\log\left(\frac{1}{d}\right)\right)^2\right).$$

6.4 Further Remarks and Open Directions

There are several interesting open directions to pursue:

- (i) By using the methods developed in Chapter 3, it is not hard to compute bounds on the speed of polarization (the scaling exponent) of the 1-bit decoder with erasures. One major drawback of this decoder is that the speed of polarization is further decreased compared to the original channel polarization process. As a result, by using the 1-bit decoder with erasures, we need to construct longer codes than the standard polar codes (with the original SC decoder). Numerical implementation suggests that as the number of quantization levels grows, the speed of polarization converges very fast to the one of original SC decoder. One important question is then how this speed is related to the number of quantization levels.
- (ii) In this chapter we have considered the perhaps simplest quantization scheme. It was shown that polar codes exhibit a robust behavior with respect to this scheme in terms of the achievable rate. Other (non-uniform) quantization schemes might perform better both in terms of robustness and speed of polarization. It is therefore worth investigating the performance of various other quantization schemes.
- (iii) Implementation of such quantization schemes into hardware is also an important practical direction. We refer the interested reader to [43] for more details.

6.5 Appendix: Proofs

Proof of Lemma 6.2

We first show that the process m_n is a super-martingale which converges a.s. to 0. From (6.12) and (6.13) we obtain,

$$\begin{aligned} \mathbb{E}[m_{n+1} | m_n] &= \frac{m_n^2 + 2m_n e_n + 2m_n p_n}{2} \\ &= m_n - \frac{m_n^2}{2} \leq m_n. \end{aligned}$$

As a result, since m_n is also bounded, it converges a.s. to a r.v. m_∞ . The a.s. convergence and boundedness of m_n also imply that

$$\mathbb{E}[m_{n+1} - m_n] = -\frac{1}{2}\mathbb{E}[m_n^2] \rightarrow 0.$$

Therefore, $m_n \rightarrow 0$ almost surely. In the same way, consider the process e_n . We have

$$\mathbb{E}[e_{n+1} | e_n] = e_n + 2p_n e_n. \quad (6.39)$$

The process e_n is then a bounded sub-martingale which converges a.s. to a r.v. e_∞ . This would imply that

$$\mathbb{E}[e_{n+1} - e_n] = 2\mathbb{E}[p_n e_n] \rightarrow 0.$$

Now, since $p_n = 1 - e_n - m_n$ and $m_n \rightarrow 0$, we get

$$\mathbb{E}[e_n(1 - e_n)] \rightarrow 0.$$

Thus, e_∞ is either 0 or 1 and considering the fact that $m_\infty = 0$, the proof follows.

Proof of Lemma 6.3

The fact that $F(D = (1, 0, 0)) = 1$, $F(D = (0, 1, 0)) = 0$ is very easy to check and thus it remains to prove

$$\frac{F(D^-) + F(D^+)}{2} \geq F(D). \quad (6.40)$$

By using (6.12) and (6.13) we obtain

$$\begin{aligned} F(D^+) &= p^2 + 2pe - 4\sqrt{(p^2 + 2pe)(m^2 + 2pm)}, \\ F(D^-) &= p^2 + m^2 - 4\sqrt{2pm(p^2 + m^2)}. \end{aligned}$$

After some straight forward simplifications, we get

$$\begin{aligned} &\frac{F(D^+) + F(D^-)}{2} \\ &= p + \frac{m^2}{2} - 2\sqrt{pm}\left(\frac{\sqrt{pm}}{2} + \sqrt{(p+2e)(m+2e)} + \sqrt{2(p^2+m^2)}\right). \end{aligned}$$

Thus, in order to show (6.40), it is necessary that the right side of the above equality is less than $p - 4pm$. We now prove a slightly stronger inequality: for $p + e + m = 1$ we have

$$\frac{\sqrt{pm}}{2} + \sqrt{(p+2e)(m+2e)} + \sqrt{2(p^2+m^2)} \leq 2. \quad (6.41)$$

It is easy to see that the above inequality results (6.19). To prove (6.41), we use the fact that

$$\sqrt{(p+2e)(m+2e)} \leq \frac{p+2e+m+2e}{2} = 2 - \frac{3}{2}(p+m),$$

and apply it to (6.41). Thus, to have (6.41), it is sufficient to prove

$$\frac{\sqrt{pm}}{2} + \sqrt{2(p^2+m^2)} \leq \frac{3}{2}(p+m), \quad (6.42)$$

by squaring both sides of (6.42) and some further simplifications we get to

$$\sqrt{2pm(p^2+m^2)} \leq \frac{1}{4}(p^2+m^2) + \frac{17}{4}pm.$$

Finally, the above inequality follows by noting the fact that for $x, y \geq 0$ we have $x+y \geq 2\sqrt{xy}$ and hence

$$\frac{1}{4}(p^2+m^2) + \frac{17}{4}pm \geq 2\sqrt{\frac{17}{16}pm(p^2+m^2)} \geq \sqrt{2pm(p^2+m^2)}.$$

Proof of Lemma 6.4

Note that for $D \in \mathcal{D}$, the minus operation given in (6.13) is exactly the same as the original minus operation without any further quantization step, i.e., $D^- = D^0$. We know from (2.24) that for any BMS channel we have $Z(W^0) \leq 2Z(W)$ and hence $Z(D^0) \leq 2Z(D)$. We now prove the following

$$Z(D^+) \leq 2Z(D)^{\frac{3}{2}}. \quad (6.43)$$

Recall that $D = m\Delta_{-\infty} + e\Delta_0 + p\Delta_{\infty}$. We have from (6.12),

$$\begin{aligned} Z(D^+) &= 2\sqrt{(p^2+2pe)(m^2+2me)} + e^2 + 2pm \\ &= 2\sqrt{pm}\sqrt{(p+2e)(m+2e)} + e^2 + 2pm \\ &= 2\sqrt{pm}\sqrt{pm+4e^2+2e(m+p)} + e^2 + 2pm \\ &\stackrel{(a)}{=} 2\sqrt{pm}\sqrt{pm+2e(1+e)} + e^2 + 2pm \\ &\stackrel{(b)}{\leq} 2\sqrt{pm}(\sqrt{pm} + \sqrt{2e(1+e)}) + e^2 + 2pm \\ &= (2\sqrt{pm} + e)^2 + 2\sqrt{pm}(\sqrt{2e(1+e)} - e) \\ &= Z(D)^2 + 2\sqrt{pm}(\sqrt{2e(1+e)} - e), \end{aligned}$$

where step (a) follows from the fact that $m+e+p=1$ and step (b) follows from the inequality $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$. Following the above lines, to prove (6.43), it is enough to show that

$$2\sqrt{pm}(\sqrt{2e(1+e)} - e) \leq 2Z(D)^{\frac{3}{2}} - Z(D)^2$$

$$= Z(D)(2\sqrt{Z(D)} - Z(D)).$$

Now, by noting that $Z(D) \geq 2\sqrt{pm}$, we only need to show the following,

$$\begin{aligned} \sqrt{2e(1+e)} - e &\leq 2\sqrt{Z(D)} - Z(D) \\ &= 2\sqrt{2\sqrt{pm} + e} - 2\sqrt{pm} - e. \end{aligned}$$

Rearranging the terms, we should prove

$$\sqrt{2e(1+e)} + 2\sqrt{pm} \leq 2\sqrt{2\sqrt{pm} + e},$$

which by dividing both sides by 2 and then squaring both sides gives

$$\frac{e(1+e)}{2} + pm + \sqrt{2pme(1+e)} \leq 2\sqrt{pm} + e.$$

Now, since $e \leq 1$, we have $\frac{e(1+e)}{2} \leq 2$ and after further simplifications we finally reach the following relation to prove

$$\sqrt{pm} + \sqrt{2e(1+e)} \leq 2,$$

which by noting that $\sqrt{pm} \leq \frac{p+m}{2} = \frac{1-e}{2}$, reduces to the following inequality

$$\frac{1-e}{2} + \sqrt{2e(1+e)} \leq 2.$$

It is straight forward to show that the above inequality holds for $e \in [0, 1]$.

Part II

Threshold Saturation

Threshold Saturation on Coupled Graphical Models

7

7.1 Introduction

In the second part of this thesis, we concentrate on the technique of spatial coupling and the threshold saturation phenomenon in the broad context of graphical models. Graphical models and low-complexity message-passing algorithms play an increasingly important role in a variety of applications and branches of engineering and science. For example, most present-day physical-layer communications schemes are based on these concepts. They are also key in areas such as machine learning, vision, and social networks, where efficient inference schemes are required to deal with massive amounts of data.

As we explained earlier in Chapter 1, spatial coupling is a method that starts with a graphical model and a “hard” computational task (e.g., decoding or more generally inference) and creates from this a new graphical model for the same task that has “locally” the same structure but is computationally “easy”. The basic observation that, on spatially-coupled graphs, low-complexity (message passing) algorithms suffice to achieve optimal performance, was developed in the area of channel coding by Kudekar, Richardson and Urbanke [2, 3]. This picture has since been completed/generalized by a vast amount of studies of graphical models in communications, computer science, and statistical physics. The potential benefits of spatial coupling can be broadly classified into the following two directions.

- (i) *Spatially coupled constructions with optimal performance*: Spatially coupled graphical models have proven to be very successful in providing efficient and optimal schemes for several important scenarios (e.g., channel coding, compressed sensing, etc). Given the value that spatial coupling has brought to such standard scenarios, it is tempting to predict similar improvements due to spatial coupling for other frameworks. This

technique works whenever the structure design for the problem is in our hands. For instance, in channel coding the code design is up to us.

- (ii) *Spatial coupling as a proof technique:* There is a second potentially very fruitful application of spatial coupling, specifically to use it as a proof technique. In other words, we can use this technique as a “thought experiment”, particularly when we are not interested in designing efficient systems but rather want to analyze a given fixed system. As will be explained in more detail later, this approach allows us to attack problems that are currently out of the reach of traditional mathematical techniques. The most immediate application is to prove the existence of thresholds, such as the threshold of sparse graph codes, or the threshold of standard constraint satisfaction problems (e.g., random K -SAT). Potentially, it can also lead to better bounds for these thresholds and a better understanding of the solution spaces for these problems.

In this thesis, we investigate these applications of spatial coupling for a variety of graphical models in the areas of statistical physics and computer science. We first consider the relatively well-understood Curie-Weiss (CW) model and its spatially coupled version. This model provides us with the simplest model to understand the mechanism of spatial coupling and the phenomenon of threshold saturation in the perspective of statistical physics. In particular, we will see how the well-known method of Maxwell construction in statistical physics manifests itself through spatial coupling.

We then consider a much richer class of graphical models called constraint satisfaction models. Again, we investigate the effect of spatial coupling and in particular the mechanism behind the phenomenon of threshold saturation. Here, we will see how spatial coupling can be turned into a new powerful proof technique.

Let us begin by introducing these models together with their spatially coupled versions.

7.2 The Simplest Mean-Field Model: The Curie-Weiss Model

The Curie-Weiss (CW) model¹ was initially considered in the physics literature as a simple and exactly solvable model for a class of materials called ferromagnets. Generally speaking, a simple way to study the behavior and the interactions between the magnetic moments of the atoms in a material is to model them by variables called *spins*. A spin can be either -1 or $+1$, which represents the direction in which a magnetic moment (think of a tiny compass needle) is pointing. Depending on the material that we are modeling, these spins interact differently. Typically, the number of spins (atoms) that a (condensed) matter system contains is in the order of 10^{23} ; a huge number. One

¹This model is also known as the Ising model on a complete graph.

main objective of theoretical physics is to capture the macroscopic properties of such large systems of interacting particles by devising highly idealized but (to some extent) analyzable mathematical models. The CW model is among the simplest such models that can be exactly solved and analyzed. While being simple, this model is a suitable framework for capturing some important aspects that are typically present in more complex systems. We start with a brief review of standard material about the CW model.

7.2.1 Basic Setting

Let $G = (V, E)$ be a complete graph with N vertices. We assign to each vertex $i \in V$ an (Ising) spin $s_i \in \{-1, +1\}$. All the spins in this model interact with each other in a pair-wise manner. A configuration of such a system is given by $\underline{s} = (s_1, \dots, s_N)$. For a configuration \underline{s} of the spins, we associate an energy function or a cost function called the Hamiltonian of the system. The Hamiltonian has the form

$$H_N(\underline{s}) = -\frac{J}{N} \sum_{\langle i,j \rangle} s_i s_j, \tag{7.1}$$

where the sum over $\langle i, j \rangle$ is carried over all edges of the graph (i.e., the interaction between the spins is pair-wise). The constant J is called the *coupling strength* of the edges. Here, we assume that $J > 0$, i.e., we assume a *ferromagnetic* coupling between the spins. It is also convenient to think of J as the *inverse of the temperature* of the system, i.e., $J = \frac{1}{T}$, where T denotes the temperature. Figure 7.1 depicts a simple example of such a system with four vertices. A useful quantity for expressing the macroscopic properties of the

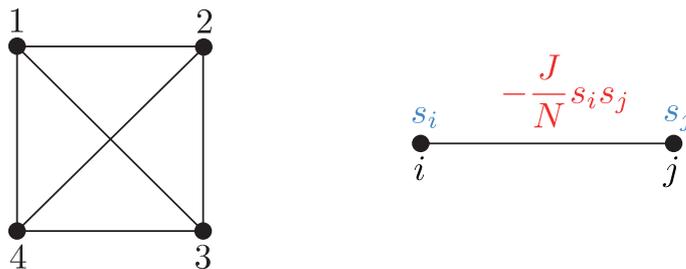


Figure 7.1: *Left:* A schematic representation of a CW model with $N = 4$, i.e., a complete graph with vertex set $V = \{1, 2, 3, 4\}$. Attached to each vertex $i \in V$, there is a spin s_i that takes its value inside the set $\{-1, +1\}$. *Right:* For any two vertices $i, j \in V$, there is an edge $\langle i, j \rangle$. The coupling strength of any such edge in the graph is equal to $J > 0$. That is, each edge bears an energy of value $-\frac{J}{N} s_i s_j$.

system is its magnetization, which is defined for a configuration \underline{s} as

$$m = \frac{1}{N} \sum_{i=1}^N s_i.$$

The free energy (in the canonical ensemble with fixed magnetization) of the system is defined as

$$\Phi_N(m) = -\frac{1}{N} \ln Z_N \quad \text{where} \quad Z_N(m) = \sum_{\underline{s}: m = \frac{1}{N} \sum_{i=1}^N s_i} e^{-H_N}. \quad (7.2)$$

The free energy is a quantity of great interest in statistical physics. As we will see in the following, we can relate the thermodynamic state variables (such as temperature, magnetization and magnetic field) by an equation that involves the derivative of the free energy. This equation is called the *equation of state* of the system.

In order to illustrate well the concepts of this section, it is helpful to establish an analogy between the variables that we introduce here and those used for liquids and gases (such as the total volume, pressure, etc). In thermodynamics, the pressure obeys the thermodynamical relation

$$p = -\frac{\partial f}{\partial v}, \quad (7.3)$$

where $f = F(T, V, N)/N$ is the thermodynamical free energy, N is the number of atoms, p is the pressure, and $v = \frac{V}{N}$ is the volume divided by N . This equation is called equation of state. For the ideal gas (where the atoms do not interact with each other), we are all familiar with the equation of state that has the form $pv = kT$, where k is the Boltzmann constant. When the particles in the gas (or liquid) interact, such a simple state equation is no longer valid. The so-called *van der Waals equation* provides an alternative that better describes real systems. It assumes that the particles have a non-zero “effective volume” modeling their molecular repulsion and that there is also a pairwise attractive inter-particle force between them. The equation is

$$p = \frac{kT}{v-b} - \frac{a}{v^2}, \quad (7.4)$$

where a is a measure for the attraction between the particles, and b denotes the effective volume occupied by a particle. Let us now plot p (vertical axis) as a function of v (horizontal axis) and the value of T is fixed. These plots are called the van der Waals *isotherms* (see figure 7.2). For higher values of T , p is a convex decreasing function of v . There is a critical temperature T_c where the curves develop an oscillatory region. Below T_c there is a part of the region where $\frac{\partial p}{\partial v} > 0$, which signals a mechanical instability. In other words, the van der Waals equation fails to describe real substances in equilibrium in this region. The final step for correcting the van der Waals equation is the *Maxwell*

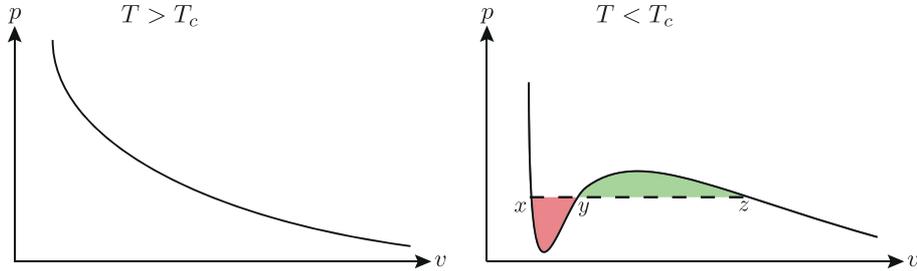


Figure 7.2: The van der Waals isotherms for two temperature values, one above the critical temperature (the left plot) and the other below the critical temperature (the right figure). For $T < T_c$, the Maxwell construction is done by replacing the part of the curve between x and z by a horizontal line positioned so that the areas of the two colored/shaded regions are equal (the dashed line in the figure).

construction (see Figure 7.2 for $T < T_c$): replace the isotherm between x and z by a horizontal line positioned so that the areas of the two hatched regions are equal.

The CW equation of state, derived below in (7.6), was originally developed by analogy to the van der Waals theory. The magnetization m corresponds to the density of the particles or equivalently v . The free energy of fixed magnetization $\Phi(m)$ is analogous to $f(T, V, N)$ and the equation of state $h = \frac{\partial \Phi}{\partial m}$ is analogous to $p = -\frac{\partial f}{\partial v}$. We see that the magnetic field h plays the same role as the pressure p .

The free energy of the CW model (given in (7.2)) has a well-defined thermodynamic limit (large N limit) that can be found as a function of m to be (we drop an irrelevant additive constant)

$$\lim_{N \rightarrow +\infty} \Phi_N(m) \equiv \Phi(m) = -\frac{J}{2}m^2 - \text{ent}(m). \quad (7.5)$$

Here, the term $-\frac{J}{2}m^2$ is the “internal energy” of the system and the function $\text{ent}(\cdot)$ is called the binary entropy function, and is defined as

$$\text{ent}(m) = -\frac{1+m}{2} \ln \frac{1+m}{2} - \frac{1-m}{2} \ln \frac{1-m}{2}.$$

The equation of state or the van der Waals curve is simply

$$h = \frac{\partial \Phi(m)}{\partial m} = -Jm + \frac{1}{2} \ln \frac{1+m}{1-m}, \quad (7.6)$$

which is equivalent to the Curie-Weiss mean field equation

$$m = \tanh(Jm + h). \quad (7.7)$$

As is well known, from the van der Waals curve $h(m)$ (7.6), we can derive an equation of state that satisfies thermodynamic stability requirements from a Maxwell construction. Similarly a physical free energy is given by the convex envelope of (7.5). For $J \leq 1$, $h(m)$ is monotone (see Figure 7.3) and the

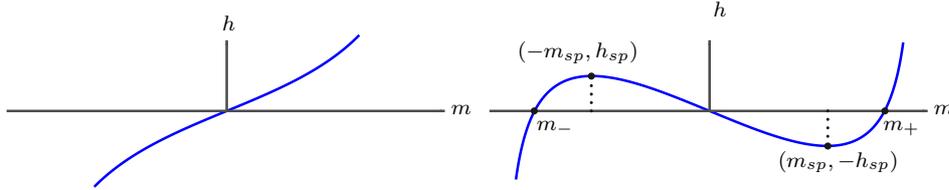


Figure 7.3: *Left plot:* the van der Waals curve in the high temperature phase $J < 1$. *Right plot:* the van der Waals curve in the low temperature phase $J > 1$. For $m \notin (m_-, m_+)$ the curve describes stable equilibrium states and for $m \in (m_-, -m_{sp}) \cup (m_{sp}, m_+)$ metastable states. For $m \in (-m_{sp}, m_{sp})$ the system is unstable. The Maxwell plateau describes superpositions of m_- and m_+ states. That is, in the Maxwell plateau the oscillatory region between m_- and m_+ is replaced by a straight line at height $h = 0$.

inverse relation $m(h)$ yields the *thermodynamic equilibrium* magnetization at a given external magnetic field h . For $J > 1$, the equations (7.6)-(7.7) might have more than one solution for a given h (see Figure 7.3). Starting with h positive and large, we follow a branch $m_+(h)$ corresponding to a *thermodynamic equilibrium state* until the point $(h = 0_+, m = m_+)$. Then we follow a lobe corresponding to a *metastable state* until the *spinodal point* $(h = -h_{sp}, m = m_{sp})$ at the minimum of the lobe. Finally from the spinodal point to the origin, the curve corresponds to an *unstable state* (where $\frac{\partial^2 \Phi(m)}{\partial m^2} < 0$). The situation is symmetric if we start on the other side of the curve with h large negative. We first follow a stable equilibrium state with magnetization equal to $m_-(h)$ until the point $(h = 0_-, m = m_-)$; we then follow a metastable state till the left spinodal point $(h = h_{sp}, m = -m_{sp})$; and finally an unstable state till the origin.

The first order phase transition line is $(h_c = 0, J > 1)$ and terminates at the critical second order phase transition point $(h_c = 0, J = 1_+)$. For $J < 1$ and $h = 0$, $m_{\pm} = 0$. We call $J = 1$ the *critical temperature* of the CW model.

For physical systems, the condition

$$\frac{\partial^2 \Phi(m)}{\partial m^2} \geq 0 \quad (7.8)$$

is an stability requirement of the system. In other words, if a physical system is at a state that the condition (7.8) is not fulfilled, then such a system is not in thermodynamical equilibrium and will re-arrange itself such that eventually the condition (7.8) is fulfilled. As a result, for a physical system at equilibrium, the van der Waals curve cannot be in the form of Figure 7.3 (the right plot). This

is because for $m \in (-m_{sp}, m_{sp})$ the stability requirement (7.8) is not fulfilled and the system is unstable. Moreover, if we initiate a physical system with an average magnetization $m \in (-m_{sp}, m_{sp})$ and initial magnetic field $h(m)$ as in Figure 7.3, then the system starts rearranging itself and will relax (after a long enough time) in a way that it ends up in thermodynamical equilibrium. The reason why the CW model predicts an unphysical isotherm is that the Hamiltonian given in (7.1) does *not* correspond to a physical (realistic) system, hence it does not possess a standard thermodynamic behavior. One main reason for non-physicality of the Hamiltonian (7.1) is the fact that its geometry (the complete graph) lacks any kind of finite-dimensional structure, whereas physical systems typically are highly finite-dimensional.

As we will see shortly, by spatially coupling the individual CW models, we provide to some extent the required finite-dimensional geometry for the CW model to be relaxed in its equilibrium state (derived from the Maxwell construction). In fact, for the coupled chain of CW models the difference between the first order phase transition and spinodal thresholds becomes much smaller, and vanishes exponentially fast with the width of the coupling along the chain.

7.2.2 Coupled Curie-Weiss Model

Here, we introduce the coupled ensemble via the simplest instance we can imagine. We postpone the general methodology of spatially coupling the CW models to Chapter 8. Consider $2L + 1$ integer positions $z = -L, \dots, +L$ on a one dimensional line. At each position, we attach a single CW spin system, i.e., a complete graph with N vertices (see Figure 7.4). Let us denote the i -th variable at position z by (i, z) and its corresponding spin by s_{iz} . We now couple the CW systems together. We connect each variable (i, z) to all the variables in positions $z - 1$ and $z + 1$, and keep its original connections inside position z . Also, for the variables at the left boundary (i.e., $z = -L$), we just connect them to all the ones at position $-L + 1$ and similarly all the variables at the right boundary (i.e., $z = L$) are connected to the variables at position $L - 1$. Thus, for positions z away from the boundary, the degree of each variable is $3N - 1$ and for the positions at the boundaries the degree of a variable is $2N - 1$. For $J > 0$, we let each edge have a coupling strength equal to $\frac{J}{3}$. Thus, for a variable away from the boundaries, the total strength that it “feels” is equal to J (assuming N is large). Also, the variables at the boundaries feel the total strength of $\frac{2J}{3}$ and hence are more free.

The coupled model introduced above is among the simplest ways to couple together the CW models placed on a chain. One possible extension is to increase the range of coupling beyond the neighboring copies, i.e., connect the variables at position z to all the variables at positions $z - w, \dots, z + w$, where w is a positive integer called the *coupling width* or the *coupling range*.

Due to this additional spatial structure (or geometry), the coupled chain of CW models tends to have a number of intriguing properties, which constitute

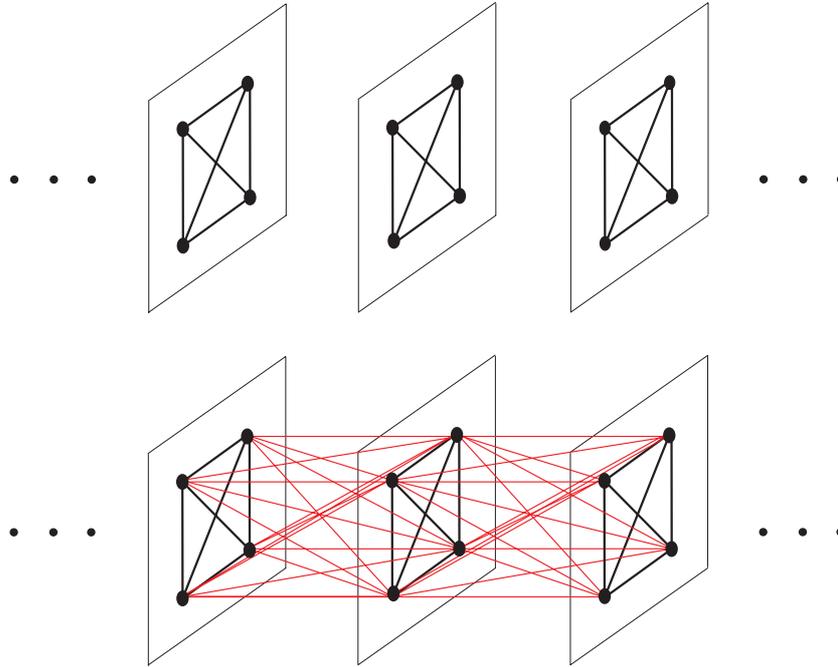


Figure 7.4: A simple version of the coupled CW models. *Top figure:* We place $2L + 1$ copies of the individual CW model on the positions $-L, \dots, L$. *Bottom figure:* We then connect the neighboring copies. Each variable at a position z , is connected to all the variables at positions $z - 1, z, z + 1$ (as long as they exist). The coupling strength of all the edges is equal to $\frac{J}{3}$.

the main subject of Chapter 8. We conclude this section by a brief summary of the results of Chapter 8.

7.2.3 Contributions of Chapter 8

The main focus of Chapter 8 is to understand the evolution of the van der Waals isotherm of the coupled chain when the individual underlying system is infinite (i.e., N is infinite), and the coupling range w together with the longitudinal length $2L + 1$ are large ($L \gg w \gg 1$), but still finite. This problem is studied for temperatures below the critical temperature of the individual system. In the limit where both L and w become infinite, the *van der Waals* isotherm of the coupled chain tends to the *Maxwell* isotherm of the individual CW system. In particular, the spinodal points of the coupled chain approach the Maxwell plateau of the individual system. This is the threshold saturation phenomenon. Correspondingly, the canonical free energy of the coupled chain is given by the convex envelope of the individual CW model.

When L and w are large but remain finite, below the critical point of the CW model, a fine structure develops around the Maxwell plateau: The straight line (Maxwell line) is replaced by an oscillatory curve with a period in the order of the inverse of the chain length with an amplitude that is exponentially small in the coupling range w (see Figure 7.5). Correspondingly, the finite-size corrections to the canonical free energy display, in addition to a “surface tension” shift, the same oscillations along the line joining the two equilibrium states of the individual system (see Figure 7.5 and formula (8.52)). The series of stable minima corresponds to kink-like magnetization density profiles, which represent the coexistence of the two stable phases of the individual system, with a well-localized interface centered at successive positions of the chain (formulas (8.44), (8.56), and Figure 8.3). A series of unstable maxima is associated with kinks centered in-between successive positions. We point out that although our analytical results are for the regime of large w , we numerically observe the same phenomena very clearly even when $w = 1$, which corresponds to nearest-neighbor coupling between individual CW systems.

One of the virtues of the present simple model is that it can, to a large extent, be treated analytically by rather explicit methods. Although our analysis is not entirely rigorous, we believe that it can be made so. We have refrained from doing so here, so that the mathematical technicalities do not obscure the main picture. Finally, let us point out that the same results hold for the CW model with random fields, which we omit for the sake of brevity and we refer to [50].

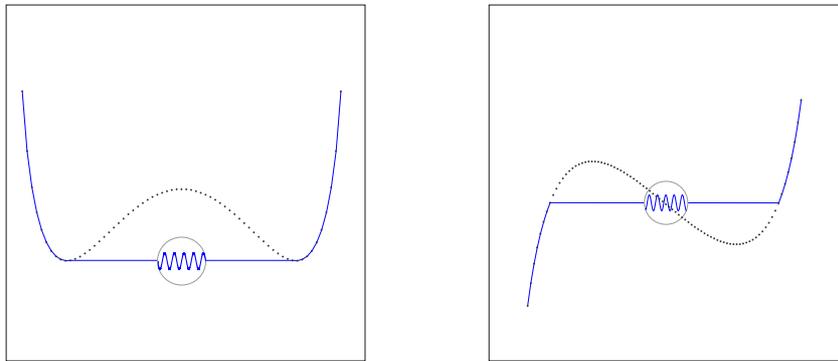


Figure 7.5: Qualitative illustration of the main result of Chapter 8. Dotted curves: free energy and van der Waals isotherm of the single system for a coupling strength $J > 1$ ($J = 1$ is the critical point). Continuous curves: free energy and van der Waals isotherm of the coupled chain for $L \gg w \gg 1$. The oscillations extend throughout the plateau with a period $M/2L$ and amplitudes $O(L^{-1}e^{-\frac{2\alpha\pi^2 w}{JM}})$ (left) and $O(e^{-\frac{2\alpha\pi^2 w}{JM}})$ (right) where $M =$ width of plateau, $\alpha = O(1)$ depends on the details of the interaction (Section 8.3). Close to the end points of the plateau, within a distance $O(L^{-1/2})$, boundary effects are important and the curves depend on the details of the boundary conditions (Section 8.4).

7.3 Constraint Satisfaction Problems

In Chapters 9 and 10 we consider the class of constraint satisfaction problems (CSP). Among the main scenarios of this class, we can mention the *satisfiability* problem (SAT) and the *graph coloring* problem (COL). Satisfiability was the first known example of an NP-complete problem. Also, a wide range of other naturally occurring decision and optimization problems can be transformed into instances of satisfiability. As a result, the problem of satisfiability lies at the heart of computational complexity theory. Speaking in terms of practical applications, the satisfiability problem is related to a vast variety of other problems, many of which have enormous practical relevance. Among these problems, are for instance computer hardware and architecture design, circuit design, verification problems, computer graphics, and image processing. The main focus of Chapters 9 and 10 is the problem of satisfiability². For the sake of brevity, we do not go into further details on other well-known instances of CSP, such as graph coloring, and we only mention the final relevant results.

7.3.1 Basic Setting and Notation

Let us begin by a brief illustration of SAT. A SAT formula consists of N Boolean variables $x_i \in \{0, 1\}$, $i \in \{1, \dots, N\}$, and a set of M logical constraints $c \in \{1, \dots, M\}$. Each logical constraint, call it a *clause*, is a disjunction (logical OR) of some variables or their negation; the negation of x_i is $\bar{x}_i = 1 - x_i$. For example, the clause $\bar{x}_1 \vee x_2$ is the logical OR operation of x_2 and the negation of x_1 . This clause is satisfied if either $x_1 = 0$ (i.e., x_1 is false) or $x_2 = 1$ (i.e., x_2 is true) or both. Analogously, the clause $x_1 \vee x_2 \vee \bar{x}_3$ is satisfied by all the configurations of the three variables except $x_1 = x_2 = \bar{x}_3 = 0$. The number of variables involved in a clause is called the length of the clause. A clause of length K is typically called a K -clause.

Given a formula consisting of M clauses and N variables, the satisfiability problem is to find a configuration for the variables such that all the clauses are satisfied (a decision problem). If such a configuration exists, we call the formula *satisfiable* and if not, we call it *un-satisfiable*. It is equally important to find a configuration that minimizes the number of violated constraints (an optimization problem). This is typically called the maximum satisfiability or the MAX-SAT problem. A formula in which all the clauses have equal length K , is called a K -SAT formula. The K -SAT problem is then the restriction of SAT to the set of K -SAT formulas. Similarly, for the K -MAX-SAT problem, the domain is confined to the set of K -SAT formulas.

It is convenient and natural to represent a SAT formula via a bipartite graph $G = (V \cup C, E)$, where we denote the set of variable nodes by V and the set of clause nodes by C . We thus have $|V| = N$ and $|C| = M$. There is an edge between a clause $c \in C$ and a variable $i \in V$ if and only if the clause c contains the variable x_i . We denote such an edge by (c, i) . Furthermore, depending on

²To be precise, we consider the problem of random satisfiability which we explain shortly.

how the variable x_i appears in the clause c (i.e., x_i or its negation \bar{x}_i) the edge (c, i) takes the form of a full edge (if x_i appears in c) or a dashed edge (if \bar{x}_i appears in c). We denote the set of edges of G by E . For a clause c in the graph, we denote by ∂c the set of variables it is connected to. Similarly, for a variable i , ∂i denotes the set of clauses it is connected to. Figure 7.6 illustrates these concepts via a simple example.

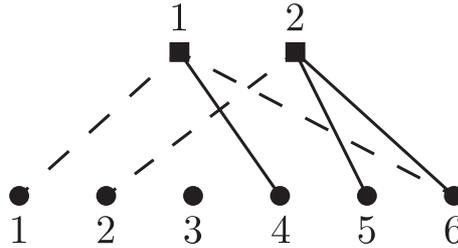


Figure 7.6: A schematic representation of a 3-SAT formula via a bipartite graph $G = (V \cup C, E)$. We have the set of variables $V = \{1, \dots, 6\}$ and the set of clauses $C = \{1, 2\}$. The set of edges of the graph is denoted by E . Hence, the formula contains 6 variables and 2 clauses of length 3. The first clause represents $\bar{x}_1 \vee x_4 \vee \bar{x}_6$ and the second clause represents $\bar{x}_2 \vee x_5 \vee x_6$.

7.3.2 The K -SAT Ensemble

As we mentioned earlier, satisfiability was the first problem proved to be NP-complete. That is, there are cleverly designed SAT formulas for which there is no known efficient algorithm to solve them, and it is not even clear whether such efficient algorithms exist or not. Consequently, these kinds of “worst case” instances are among the main challenges of computer science.

An alternative approach to the problem of satisfiability is to consider formulas that are chosen *randomly*. For instance, suppose we construct a K -SAT formula by choosing each of the clauses uniformly at random from the set of all possible K -clauses. Hence, rather than considering cleverly designed opponents (formulas), we are confronted with an *ensemble* of formulas endowed with a probabilistic structure. It is apparent that the analysis of different kinds of ensembles is deeply entangled with combinatorics and probability theory. In the following, we introduce the most famous of such probabilistic ensembles, namely the K -SAT ensemble.

Consider N Boolean variables and $M = \lfloor \alpha N \rfloor$ clauses of length K . Here, we note that α is a positive real number called the *clause density*. To choose an instance from the K -SAT ensemble, we proceed as follows: Each of the M clauses picks uniformly at random a subset of length K of the variables and flips a fair coin to decide whether or not to negate each variable. Note that all the above steps are taken independently of one another. We can easily see that

any instance of the K -SAT ensemble is chosen with uniform probability. In the following, we use $\text{SAT}(N, K, \alpha)$ to denote the ensemble of random K -SAT formulas with size N and density α .

Due to its simple probabilistic structure and the importance of the satisfiability problem, the K -SAT ensemble has become a central topic of collaboration between computer scientists, mathematicians and statistical physicists. As we will see later, random K -SAT formulas enjoy a number of intriguing mathematical properties, many of which have been discovered and many others are yet to be found or made rigorous. Also, most of the ideas and intuitions about this ensemble have been extended naturally to other CSP problems such as graph coloring (COL). We keep in mind that whether these random formulas are a good model for real-world applications or not, is a question that requires much further investigation. In fact, the ingenious structure in the real-world SAT formulas is something beyond the capability of simple probability distributions to capture. However, it is worth mentioning that random K -SAT instances are computationally hard for a certain range of densities, and this makes them popular benchmarks for testing and tuning satisfiability algorithms. In fact, some of the best practical ideas in use today come from the insight gained by studying the performance of algorithms on random K -SAT instances [70].

We proceed with a brief detour of the current state of the art for the K -SAT problem. We refer the interested reader to [71], [72] and [73] for an excellent review of these topics. We then introduce the coupled K -SAT ensemble, which is the focus of Chapter 9 and 10. Finally, we conclude this section by summarizing the main results of Chapter 9 and 10.

The Threshold Conjecture

Consider a random formula from the K -SAT ensemble. What is the probability that such a formula is satisfiable? A moment of thought shows that this probability is a non-increasing function of α . Also, for small α we expect that most of the formulas are satisfiable, whereas for α tending to infinity most of the formulas seem to be un-satisfiable. What more can we say? In particular, what happens when the size of these formulas grows unbounded, i.e., $N \rightarrow \infty$? Numerical experiments, physical arguments (as we will see later), as well as the experience from simpler CSPs, strongly indicate that as the density crosses a critical threshold, these formulas undergo a *phase transition* from becoming almost certainly satisfiable to almost certainly unsatisfiable. Despite such strong evidence, it is yet unknown if such a critical density exists for $K \geq 3$ hence has remained as a conjecture called *the satisfiability conjecture*.

Conjecture 7.1 (The satisfiability conjecture). *For $K \geq 2$, there exists a constant $\alpha_s(K)$ such that the following holds*

$$\lim_{N \rightarrow \infty} \Pr\{\text{SAT}(N, K, \alpha) \text{ is satisfiable}\} = \begin{cases} 1 & \text{if } \alpha < \alpha_s(K), \\ 0 & \text{if } \alpha > \alpha_s(K). \end{cases} \quad (7.9)$$

For $K = 2$, the satisfiability conjecture is known to be true and we have $\alpha_s(2) = 1$ [74]. The following theorem is the closest we know regarding the existence of such a threshold.

Theorem 7.1 (Friedgut [75]). *For $K \geq 3$, there exists a sequence $\{\alpha_s(K, N)\}_{N \in \mathbb{N}}$ such that for any $\epsilon > 0$ the following holds*

$$\lim_{N \rightarrow \infty} \Pr\{\text{SAT}(N, K, \alpha) \text{ is satisfiable}\} = \begin{cases} 1 & \text{if } \alpha < (1 - \epsilon)\alpha_s(K, N), \\ 0 & \text{if } \alpha > (1 + \epsilon)\alpha_s(K, N). \end{cases} \quad (7.10)$$

Theorem 7.1 comes very close to proving the satisfiability conjecture except that the sequence $\alpha_s(K, N)$ is not known to converge to a well-defined limit. In particular, there remains the possibility that such a sequence oscillates in a small window thus might not converge. From now on, we let $\alpha_s(K)$ denote both the satisfiability threshold from Conjecture 7.1 and also the threshold sequence of Theorem 7.1, and leave the corresponding interpretation to the interested reader.

The consequences of Theorem 7.1 are not confined merely to the satisfiability conjecture. Another main application of this theorem is in providing bounds on $\alpha_s(K)$ in the following way. Suppose there exists a method that proves for some density $\alpha_{\text{method}}(K)$,

$$\lim_{N \rightarrow \infty} \Pr\{\text{SAT}(N, K, \alpha_{\text{method}}(K)) \text{ is satisfiable}\} \geq C, \quad (7.11)$$

where C is a positive constant. Then, from Theorem 7.1 we conclude that for any $\alpha \leq \alpha_{\text{method}}(K)$ we have

$$\lim_{N \rightarrow \infty} \Pr\{\text{SAT}(N, K, \alpha) \text{ is satisfiable}\} = 1.$$

In particular, this would show that $\alpha_s(K) \geq \alpha_{\text{method}}(K)$. Similarly, if $\alpha_{\text{method}}(K)$ is such that the inequality (7.11) holds in the opposite direction, then the probability that a random formula is satisfiable at densities above $\alpha_{\text{method}}(K)$ tends to 0 and we obtain that $\alpha_s(K) \leq \alpha_{\text{method}}(K)$.

This consequence of Theorem 7.1 has been the main venue for providing lower bounds on $\alpha_s(K)$. We now proceed by reviewing various methods and bounds on the threshold.

Various Bounds and Asymptotic Behavior of the Threshold

Let us begin by a simple but important upper bound. For a random K -SAT formula F , we denote by $X(F)$ its number of satisfying assignments (if $X(F)$ is zero then the formula is un-satisfiable). It is an easy exercise to show that

$$\mathbb{E}[X] = 2^N \left(1 - \frac{1}{2^K}\right)^M.$$

As a result, by noticing $M = N\alpha$, if we choose

$$\alpha > \frac{-\ln 2}{\ln(1 - \frac{1}{2^K})},$$

then the value of $\mathbb{E}[X]$ is exponentially small in N . Hence, by an application of the Markov inequality we deduce that the probability of satisfiability is exponentially small. We thus have

$$\alpha_s(K) \leq \frac{-\ln 2}{\ln(1 - \frac{1}{2^K})} \leq 2^K \ln 2 - \frac{\ln 2}{2} - O(2^{-K}). \quad (7.12)$$

The above method, which is based on the first moment of X , is called the *first moment method*. In fact, this simple upper bound can be made slightly sharper [82, 83]

$$\alpha_s(K) \leq 2^K \ln 2 - \frac{1 + \ln 2}{2} - o(1). \quad (7.13)$$

where the $o(1)$ term is asymptotically vanishing in K . To obtain a lower bound, a method called the *second moment method* can be used [76, 77]. The idea is that, by an application of the Cauchy-Schwarz inequality, we can easily show that

$$\Pr(X > 0) \geq \frac{\mathbb{E}[X]^2}{\mathbb{E}[X^2]}. \quad (7.14)$$

Now, if we find densities α for which the value $\frac{\mathbb{E}[X]^2}{\mathbb{E}[X^2]}$ is bounded from below by a positive constant, it is immediate that such a value of α would be a lower bound for $\alpha_s(K)$. However, on the negative side, for the choice of $X = X(F)$ to be the number of solutions, it can be shown that for any value of α , the quantity $\frac{\mathbb{E}[X]^2}{\mathbb{E}[X^2]}$ decays to 0 by N . In other words, the number of solutions does not concentrate around its average. On the positive side, we can choose other candidates for X , rather than the number of solutions, to plug into (7.14). For instance, instead of giving an equal weight to all solutions of a formula F (as done in counting the number of solution), we can assign different weights to different solutions. This is called the *weighted second order method*. Using this method, it can be shown [76] that

$$\alpha_s(K) \geq 2^K \ln 2 - (K + 1) \frac{\ln 2}{2} - 1 - o(1). \quad (7.15)$$

Very recently, by a new version of the weighted second order method, a new lower bound has been obtained in [78]

$$\alpha_s(K) \geq 2^K \ln 2 - \frac{3 \ln 2}{2} - o(1). \quad (7.16)$$

To summarize, for large K we have

$$2^K \ln 2 - \frac{3 \ln 2}{2} - o(1) \leq \alpha_s(K) \leq 2^K \ln 2 - \frac{1 + \ln 2}{2} - o(1), \quad (7.17)$$

K	3	4	5	7	10
Upper bound from (7.12)	5.19	10.74	21.83	88.37	709.44
Best upper bound [81]	4.51	10.23	21.33	87.88	708.94
Lower bound from [76]	2.68	7.91	18.79	84.82	704.94
Best algorithmic bound	3.52	5.54	9.63	33.23	172.65

Table 7.1: Best known rigorous bounds for the location of the satisfiability threshold $\alpha_s(K)$ for small values of K . The last row gives the largest density for which a polynomial-time algorithm has been proven to find satisfying assignments. The numbers in this table are taken from [73].

where the $o(1)$ term is asymptotic in K . These bounds indicate that for large values of K , the value of $\alpha_s(K)$ is just a small constant away from $2^K \ln 2$. For smaller values of K , the bounds derived from these methods are given in Table 7.1.

A different venue to find lower bounds is to provide algorithms capable of solving a random formula with a positive probability. We will have more to say about these algorithms and the methods used to analyze them in Chapter 10. In a nutshell, most of these algorithms act in the following way: Given a random formula, they set the variables one at a time using heuristics that use very little, and completely local, information about the variable-clause interactions. Of course, such a confinement is also what enables their analysis. Table 7.1 contains the best such algorithmic lower bounds from [79] and [80].

The MAX-SAT Version

One can also consider the MAX-SAT problem and conjecture a similar sharp thresholding behavior. Consider a random formula F . For an assignment \underline{x} of the variables in F , we define the *energy* of the assignment, denoted by $H_F(\underline{x})$, to be the number of clauses in F that the assignment \underline{x} violates. For the formula F , we define its *minimum energy level* or *ground state*, \mathcal{H}_F , to be the minimum possible energy that can be reached for F over all the assignments \underline{x} , i.e.,

$$\mathcal{H}_F = \min_{\underline{x}} H_F(\underline{x}). \quad (7.18)$$

It is more convenient to work with the normalized version of the ground state, i.e., $\frac{1}{N}\mathcal{H}_F$. In fact, it can be shown that almost surely (a.s.) for a random formula F , the *ground state per variable*, $\frac{1}{N}\mathcal{H}_F$, concentrates around its average. That is,

$$\frac{\mathcal{H}_F}{N} \xrightarrow{\text{a.s.}} \lim_{N \rightarrow \infty} \frac{\mathbb{E}[\mathcal{H}_F]}{N} \triangleq \mathcal{H}(\alpha, K). \quad (7.19)$$

The MAX-SAT conjecture is then as follows.

Conjecture 7.2 (Existence of a sharp threshold for MAX-SAT). *For $K \geq 2$ there exists a constant $\alpha_s(K)$ such that the following holds*

$$\mathcal{H}(\alpha, K) \begin{cases} = 0 & \text{if } \alpha < \alpha_s(K), \\ > 0 & \text{if } \alpha > \alpha_s(K). \end{cases} \quad (7.20)$$

In other words, the MAX-SAT conjecture states that for $\alpha < \alpha_s(K)$ there is an assignment that satisfies all the clauses except a sub-linear fraction of clauses, whereas for $\alpha > \alpha_s(K)$, any choice of the variable assignments violates a constant fraction of the clauses. With a slight abuse of notation, we have intentionally denoted both the threshold of SAT and the threshold of MAX-SAT by $\alpha_s(K)$. This is because it is widely believed (hence conjectured) that the thresholds of these two problems coincide.

The Physics Picture

Random K -SAT, together with other CSPs, have been systematically studied also by physicists during the past two decades. Such physical intuitions and derivations, which originate back to the rich theory of spin glasses, have led to the discovery of a much more refined framework for studying CSPs. As a consequence, we believe today that the extent of hardness in finding a satisfying assignment for random formulas comes from various phase transitions in the solutions space geometry of such formulas, not in their probability of satisfiability. We proceed by briefly illustrating the picture of how the geometry of the solutions space evolves as a function of the clause density. We bear in mind that such a picture yet lacks a great deal of rigor (except only a few parts [98–100]), and completing it will remain an intriguing challenge for the foreseeable future.

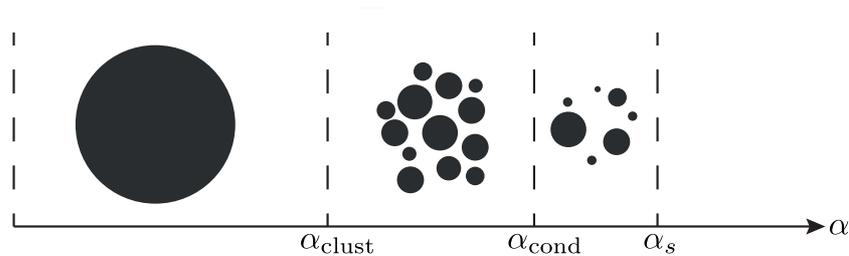


Figure 7.7: A symbolic picture of the solutions space for a random K -SAT formula.

Here, we think of the solutions of a formula as members of the Hamming cube $\{0, 1\}^N$. Figure 7.7 gives a symbolic representation of the solutions space of a typical SAT formula based on its clause density. As we observe, there are several phase transitions occurring as clauses are added. We begin by illustrating each of these phase transitions.

The Easy-SAT phase, $\alpha < \alpha_{\text{clust}}$: For low densities, the set of satisfying assignments forms a single giant *cluster*. Such a giant cluster is a well-connected object in the following sense. Consider any two satisfying assignments $\underline{x}, \underline{x}'$. Then, there exists a sequence of solutions $\underline{x} = \underline{x}_0, \underline{x}_1, \dots, \underline{x}_r = \underline{x}'$ such that the Hamming distance between the consecutive solutions x_i and x_{i+1} is very small compared to N (it is believed to be $O(\log N)$). Thus, the space of solutions can be imagined as a big cluster in which one can walk from one solution to another in steps of very small size (sub-linear in N). Such a connected structure is believed to provide the required ergodicity for the Monte-Carlo methods to sample the solutions space uniformly and in a reasonable time. Hence, this region of α is called the *Easy-SAT* phase.

The Hard-SAT phase or the clustering phase, $\alpha \in [\alpha_{\text{clust}}, \alpha_s]$: Such a single-cluster behavior of the solutions space continues up to a specific value of the clause density called the clustering transition and denoted by α_{clust} . In the regime $\alpha \in [\alpha_{\text{clust}}, \alpha_s]$ the space of solutions is believed to be fragmented into exponentially many clusters, each of which is relatively tiny and far apart from all the other clusters (like the bubbles in Figure 7.7). More precisely, for densities above α_{clust} , there are exponentially many clusters, each containing an exponentially small fraction of the solutions. The distance between any two solutions in two distinct clusters is $\Theta(N)$. Moreover, inside each cluster a constant fraction of the variables are *frozen*, i.e., take the same value in all the solutions inside that cluster.

It is widely believed that such a clustering structure is closely connected to the failure of standard “local” algorithms in finding solutions. In other words, it is believed that there is a strong connection between the “hardness” of the problem and the clustering of the solutions space. Therefore, we call this regime the *hard-SAT* regime. Let us now argue about the origins of this belief from two (equivalent) perspectives: (i) As we mentioned above, the clusters are separate by a distance of order $\Theta(N)$. Thus, in order to travel from one cluster to the other, we need to go through assignments that violate $\Theta(N)$ clauses. Now, if we think of the number of clauses violated by a generic assignment as the energy of that assignment, then the clusters are separated from each other by huge “energy barriers”. Consider a Markov chain based on “local” moves. Due to such huge energy barriers between the clusters, the Markov chain that starts from a typical high-energy assignment will be trapped inside a region around its initial point (some sort of high energy cluster), and it will take an exponential amount of time to explore the whole space. Hence, uniformly sampling a satisfying assignment becomes exponentially hard. In other words, when we enter inside the hard-SAT region, the required ergodicity of the Markov chain breaks and it takes exponential time to find a satisfying assignment. (ii) As we mentioned above, the clusters are far away from each other and contain a constant fraction of frozen variables. Let us assume that the distance between any two clusters is at least δN , where δ is a positive constant. Consider now a decimation type algorithm that sets the variables one-by-one with local decisions. Once the algorithm has set $(1 - \delta)N$ variables, it will surely be confined to at most one cluster in the solutions space. This

cluster contains some frozen variables. Hence, if any of the variables that the algorithm has already set are among the frozen ones of the cluster, and the variable has been set wrong, the algorithm already fails. To summarize, in order to perform well inside the clustering region, the algorithm must somehow “sense” these clusters and their frozen variables.

The condensation phase $\alpha \in [\alpha_{\text{cond}}, \alpha_s]$: The critical value α_{clust} signals a phase transition in the *number* of clusters from a single giant one to an exponential number of clusters. However, inside the clustering phase, there is another phase transition in the geometry of the solutions space, which is in terms of the *shape* of the clusters. For densities $\alpha \in [\alpha_{\text{cond}}, \alpha_s]$ it is predicted that almost all the solutions lie in a small (finite) number of (atypically large) clusters, while exponentially many other clusters exist.

To study the location of these phase transitions, the physicists have developed the method of *Survey Propagation* (SP)³, which is derived from the zero-temperature (level-1) cavity method of spin-glass theory [97]. Let us explain the predictions of the SP formalism for the K -SAT ensemble [101], [102]. SP is a sophisticated mean-field theory based on a set of fixed point equations. It predicts the existence of a SAT/UNSAT phase transition when α crosses a critical threshold α_s . At a lower value α_{SP} of the clause density, one finds a bifurcation from a trivial solution to non-trivial solutions of the fixed point equations. In the interval $[\alpha_{\text{SP}}, \alpha_s]$ the solutions space is fragmented into an exponentially large (in system size) number of well separated clusters of SAT solutions (ground states) in the Hamming space. The rate of growth of the number of such clusters with system size is called the *zero-energy complexity* and is positive in the interval $[\alpha_{\text{SP}}, \alpha_s]$. The complexity goes to zero at α_s and becomes formally negative above α_s .

The SP formalism says nothing about the relative sizes (internal entropy) of clusters of solutions and does not take into account which of them are “relevant” to the uniform measure over the solutions. As a result, the threshold α_{SP} does not have a clear algorithmic meaning (in the sense of being a barrier for algorithms) because SP does not take into account the size of the clusters. This issue is addressed by the entropic cavity method [84], [86], [103], [104] which allows us to compute the so-called dynamical and condensation thresholds⁴ α_d and α_c .

These methods enable us to predict such thresholds precisely. We will have more to say about these thresholds and their numerical values in Chapter 9. Let us conclude this part, by stating the large K predictions of these thresholds

$$\alpha_{\text{clust}} = \frac{2^K \ln K}{K} (1 + o(1)),$$

³We refer to [72] for a recent pedagogical account.

⁴Here, α_d denotes the dynamical threshold, which is in contrary to our previous notation for the dynamical threshold α_{clust} . We keep in mind that α_d and α_{clust} are the same concepts and we have just changed the notation to the convenient notation of the literature. The same goes with the condensation threshold, i.e., α_{cond} and α_c both denote the condensation threshold.

$$\alpha_{\text{cond}} = 2^K \ln 2 - \frac{3}{2} \ln 2 + o(1),$$

$$\alpha_s = 2^K \ln 2 - \frac{1}{2}(1 + \ln 2) + o(1).$$

where the $o(1)$ term is asymptotic in K . In particular, it is worth noticing the equivalence of the predictions of α_{cond} and α_s to the lower and upper bounds in (7.17). Let us also point out that the “algorithmic barrier” of the clustering transition, which is $\frac{2^K \ln K}{K}$. So far, all the algorithms known in the literature have fallen short in breaking this algorithmic barrier and, as we explained above, this barrier is believed to be tight for local-search algorithms.

7.3.3 The Coupled K -SAT Ensemble

This ensemble represents a chain of coupled underlying K -SAT ensembles. Figure 9.1 is a visual aid but gives only a partial view. We consider $L - w + 1$ clause positions $z \in \{0, 1, \dots, L - w\}$ and L variable positions $z \in \{0, 1, \dots, L - 1\}$. At each variable position z , we lay down N Boolean variables. Also, for each check position z , we lay down $M = \lfloor \alpha N \rfloor$ clauses of length K . So in total we have NL variables and $M(L - w + 1)$ clauses. Each clause c at a position z , chooses each of its K variables via the following procedure. We first choose a position $z + j$ with j picked uniformly at random from the window $\{0, \dots, w - 1\}$, then we pick a variable uniformly at random among all the N variables located at position $z + j$, and finally we connect the clause c and the variable. All the K variables of the clause c are chosen independently in this way. The sign of each edge is chosen independently by flipping a fair coin. This ensemble is called the (spatially) coupled K -SAT ensemble and an instance of it is called a coupled formula. We denote such an ensemble by $\text{CSAT}(N, K, \alpha, w, L)$.

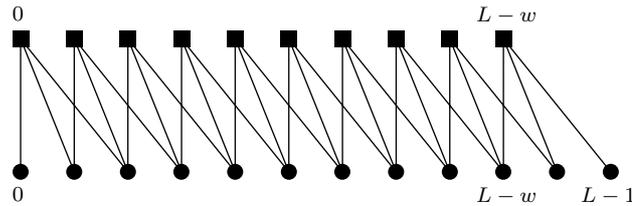


Figure 7.8: A representation of the geometry of the graphs with window size $w = 3$ along the “longitudinal chain direction” z . The “transverse direction” is viewed from the top. At each position there is a stack of N variable nodes (circles) and a stack M constraint nodes (squares). The depicted links between constraint and variable nodes represent stacks of edges.

It would be natural to extend all the concepts developed for the K -SAT ensemble to the coupled K -SAT ensemble. As the coupled ensemble is equipped with two additional parameters L and w , our notations for the coupled ensemble

will bear an additional L and w in their subscript. We denote the SAT/UNSAT threshold of the coupled ensemble by $\alpha_{s,L,w}$ and also denote the ground state per variable of the coupled ensemble (as in (7.19)) by $\mathcal{H}_{L,w}(\alpha, K)$. The overall clause density of this ensemble is $\alpha(1 - \frac{w-1}{L})$, which tends to α as L grows large. It is easy to see that at the (left and right) boundaries of the chain the variables have a smaller average degree compared to the positions in the middle. Hence, the problem is made easier at the boundaries. Moreover, the two ensembles have locally the same structure.

It seems possible to extend almost all the results of the K -SAT ensemble to the coupled ensemble. However, our main objective in this thesis is to show that due to the additional spatial structure of the coupled ensemble, a new set of remarkable aspects emerge. We now proceed by a brief explanation of the main results of Chapters 9 and 10.

7.3.4 Contributions of Chapters 9 and 10

The material of Chapters 9 and 10 can be considered as a general framework to study the (coupled version of) random ensembles of computational problems (e.g., K -SAT, Q -COL, K -XORSAT, vertex cover in random graphs, uniquely extendible CSPs). For the sake of brevity, we mainly discuss the K -SAT model here but the same approach leads to similar results for all these problems.

We begin Chapter 9 by using the tools from spin glass theory to analyze and locate the different transition regions of the coupled K -SAT ensemble. In particular, we investigate various types of threshold saturation on the coupled ensemble. The net result of our findings in this regard is that the coupled ensemble is easier to analyze or to find a satisfying assignment than the original (un-coupled) K -SAT ensemble.

In more detail, we consider the SP equations for the coupled K -SAT ensemble and solve them by the method of population dynamics. We find a positive (zero-energy) complexity in an interval $[\alpha_{\text{SP},L,w}, \alpha_{s,L,w}]$, which allows us to determine the SAT/UNSAT phase transition point $\alpha_{s,L,w}$ (where the complexity becomes formally negative). We make the following observations for the interval where the complexity is positive. We have that $\alpha_{s,L,w} > \alpha_s$ and $\alpha_{s,L,w} \downarrow \alpha_s$ as L increases (and w fixed). We find that *threshold saturation* takes place, namely $\alpha_{\text{SP},L,w} \rightarrow \alpha_s$, as L and w both increase in a way that $L \gg w \gg 1$. These findings are supported by a large K analysis of the SP fixed point equations of coupled K -SAT. In this limit, the fixed point equations reduce to one-dimensional equations, analogous to those found for the Curie-Weiss chain or coupled LDPC codes on the binary erasure channel. This enables us to study an “average total warning probability” that characterizes the phase of the system. This quantity is somewhat analogous to the average magnetization in the CW chain, or the average erasure probability for LDPC codes.

As we mentioned above, the SP formalism says nothing about the relative size (internal entropy) of clusters of solutions and does not take into account which of them are “relevant” to the uniform measure over zero-energy solu-

tions. This issue is addressed by the entropic cavity method that allows us to compute the so-called dynamical and condensation thresholds α_d and α_c . Using this method, we have computed the dynamical $\alpha_{d,L,w}$ and condensation $\alpha_{c,L,w}$ thresholds of coupled K -SAT ensemble, and observe that as L increases $\lim_{L \rightarrow +\infty} \alpha_{c,L,w} \rightarrow \alpha_c$ (w fixed) whereas $\alpha_{d,L,w} \rightarrow \alpha_c$ when both w and L increase in the regime $1 \ll w \ll L$. All these saturation phenomena indicate that for the coupled ensemble, the algorithmic barrier (or the clustering transition) is at least as much as the condensation transition. Hence, the coupled formulas are much easier to solve than the un-coupled ones.

In the second part of Chapter 9, we show how the combinatorial interpolation methods (originally introduced in [85]) can be customized to relate the coupled ensemble to the underlying uncoupled one. Using such interpolation arguments, we analytically show that as L grows, the ground state per variable $\mathcal{H}_{L,w}(\alpha, K)$ tends to its corresponding value $\mathcal{H}(\alpha, K)$ of the individual ensemble. That is,

$$\lim_{L \rightarrow \infty} \mathcal{H}_{L,w}(\alpha, K) = \mathcal{H}(\alpha, K).$$

As a consequence, we deduce that as L grows large, the satisfiability threshold (as in (7.20)) of the coupled K -SAT ensemble tends to the satisfiability threshold of the K -SAT ensemble, i.e.,

$$\lim_{L \rightarrow \infty} \alpha_{L,w,s}(K) = \alpha_s(K). \quad (7.21)$$

We notice from (7.21) that any lower bound on $\alpha_{s,L,w}$ can be turned into a lower bound for α_s by taking $L \rightarrow +\infty$. In particular, algorithmic lower bounds on $\alpha_{s,L,w}$ can be turned into lower bounds for α_s . Now, as we explained above, because of the saturation of the SP and dynamical thresholds of coupled K -SAT, the values of α for which the space of solutions is fragmented into well separated clusters are substantially larger compared to the values of individual ensembles. Therefore, we can hope that a form of *algorithmic threshold saturation*, or at least *algorithmic threshold increase*, happens when well chosen algorithms are applied to coupled K -SAT. This results in *proving* better algorithmic lower bounds on $\alpha_{s,L,w}$ and thus α_s . The proposed methodology is our main motivation for Chapter 10.

In Chapter 10, we focus on algorithmic aspects. We consider two algorithms for finding a satisfying assignment for a random coupled formulas, namely the *pure literal algorithm* and the *unit clause propagation algorithm*. The pure literal algorithm is perhaps the simplest known algorithm for solving satisfiability problems. It works up to a critical density $\alpha_{\text{pl}}(K)$ where a non-trivial 2-core emerges inside the formula. For an uncoupled formula, this critical density can be found with the help of a simple scalar fixed point equation. This algorithm extends naturally to coupled formulas. Its critical density $\alpha_{\text{pl},L,w}$, where the 2-core develops, is found by analyzing the fixed points of a set of one-dimensional fixed point equations. The recent one-dimensional theory of [126] and [127] enables us to compute the limiting value of $\alpha_{\text{pl},L,w}$ when $L \gg w \gg 1$. Let us

K	3	4	5	large K
$\alpha_{\text{cuc}}(K)$	3.67	7.81	15.76	2^{K-1}
$\alpha_{\text{uc}}(K)$	2.66	4.50	7.58	$\frac{e2^{K-1}}{K}$
$\alpha_{\text{cpl}}(K)$	1.834	1.954	1.986	2
$\alpha_{\text{pl}}(K)$	1.626	1.544	1.402	$\frac{2 \ln K}{K}$

Table 7.2: Thresholds for the peeling and unit clause propagation algorithm corresponding to the coupled and un-coupled ensembles.

denote this limit by $\alpha_{\text{cpl}}(K)$; i.e.,

$$\alpha_{\text{cpl}}(K) = \lim_{w \rightarrow \infty} \lim_{L \rightarrow \infty} \alpha_{\text{pl},L,w}(K).$$

The last two rows of Table 7.2 include the corresponding thresholds for coupled and uncoupled ensembles. For large K , we find

$$\alpha_{\text{pl}}(K) \doteq \frac{2 \ln K}{K} \quad \text{but} \quad \alpha_{\text{cpl}}(K) \doteq 2.$$

Hence, there is roughly a factor $\frac{K}{\ln K}$ of threshold improvement via spatial coupling. However, the coupled threshold of this algorithm is still far below the satisfiability threshold (which is a constant away from $2^K \ln 2$).

We next consider the unit clause propagation (UC) algorithm that is the simplest type of decimation algorithm. We first derive a suitable schedule to perform the decimation steps. We then develop the required machinery to analyze this decimation algorithm on the coupled formulas. Let us denote by $\alpha_{\text{uc}}(K)$ and $\alpha_{\text{uc},L,w}(K)$ the thresholds of the UC algorithm for the individual and coupled ensembles, respectively. We also define

$$\alpha_{\text{cuc}}(K) = \lim_{w \rightarrow \infty} \lim_{L \rightarrow \infty} \alpha_{\text{uc},L,w}(K).$$

Table 7.2 contains the corresponding thresholds for the coupled and individual ensembles. For large K we find that

$$\alpha_{\text{uc}}(K) \doteq \frac{e2^{K-1}}{K} \quad \text{but} \quad \alpha_{\text{cuc}}(K) \doteq 2^{K-1}.$$

Again, the coupled threshold is improved roughly by a factor $\frac{K}{e}$ over the individual threshold. There are a few interesting comments in order:

- (i) Comparing the numbers in Tables 7.1 and 7.2, we observe that for small K the coupled thresholds of the UC algorithm are comparable to the best lower bounds in the literature. Even for $K = 3$, the value 3.67 is a new lower bound for the K -SAT problem⁵. However, as K grows, these coupled thresholds tend to $\frac{1}{2 \ln 2}$ fraction of the best lower bounds.

⁵To be more precise, this is a new lower bound for the 3-MAX-SAT problem

- (ii) The threshold of the coupled UC algorithm is asymptotically 2^{K-1} . This intuitively confirms the fact that for the coupled ensembles, the dynamical transition is well above the dynamical transition of the individual K -SAT ensemble. In other words, the coupled formulas can be considered “easier” than the uncoupled ones in the sense that they break well through the algorithmic barrier $\frac{2^K \ln K}{K}$ of the individual ensemble.
- (iii) We believe that more sophisticated (and analyzable) algorithms for the coupled formulas can succeed all the way up to the condensation threshold. This is the topic of our current research. As we explained above, the clustering of the solutions space is believed to be the main barrier for the success of “local search” algorithms. This immediately raises the question about the mechanism behind the saturation of the dynamical threshold to the condensation threshold in the coupled formulas. In other words, how does the space of solutions change under spatial coupling and what happens to the clusters? We do not have any definitive answers for the K -SAT ensemble at the moment. However, there is a relatively “simpler” CSP ensemble called the K -XORSAT ensemble whose solutions space goes through a similar clustering transition at a well-defined dynamical threshold [92–95]. For the coupled K -XORSAT ensemble, it can be shown that at densities above the dynamical threshold of the underlying ensemble, the space of solutions has the following geometrical structure. The clusters become connected to each other and form a giant cluster. This connection is in a special form directly related to the spatial structure of the coupled formula together with the termination at the boundaries. We refer for further details to [96].
- (iv) The same results as above are obtained for the Q -COL ensemble. In particular, saturation of the SP threshold to the satisfiability threshold, as well as the saturation of dynamical threshold to the condensation threshold, can be checked. Also, an algorithm similar to the UC algorithm is proposed for the coupled Q -COL model. For $Q = 3$, this algorithm finds a proper coloring for average connectivity values up to $c_{\text{cuc}}(3) = 4.44$. We note from [86] that the condensation threshold for $Q = 3$ is $c_c(3) = 4$, which is below $c_{\text{cuc}}(3)$. Hence, for the coupled ensemble we are capable to go even above the condensation threshold. The SAT/UNSAT threshold for $Q = 3$ is $c_s(3) = 4.69$. For large Q , we find that $c_{\text{cuc}}(Q) \doteq 2Q \log(Q) - Q$. For the sake of brevity we omit the details for this model and refer to [68, 125].

8

Coupled Mean Field Models

8.1 Problem Formulation

In this chapter¹, we present in detail what we believe is the simplest and clearest situation that captures the basic underpinnings of threshold saturation. The Curie-Weiss² (CW) spin system was introduced in Section 7.2.1. Here, we introduce a one dimensional chain of $2L + 1$ CW spin systems coupled together by an interaction which is local in the longitudinal (or chain) direction and infinite range in the transverse direction. The local interaction is of Kac type with an increasing range and inversely decreasing intensity, and is ferromagnetic. This model can be viewed as an anisotropic Ising system with a Kac interaction along one longitudinal direction and a Curie-Weiss infinite range interaction along the "infinite dimensional" transverse direction.

The main focus here is to understand the evolution of the van der Waals isotherm of the coupled chain when the individual underlying system is infinite and, the range w of the Kac interaction and the longitudinal length $2L + 1$ both become large $L \gg w \gg 1$ but are still finite. This problem is studied for temperatures below the critical temperature of the individual system.

In Section 8.2, we set up our basic coupled model and give a formal solution. The asymptotic analysis for $L \gg w \gg 1$ is performed in Section 8.3 and this is supplemented by numerical simulations in Section 8.4.

¹The material of this part is based on [50].

²Ising model on a complete graph.

8.2 Chain of Ising Systems on Complete Graphs

8.2.1 Curie-Weiss Model

Let us recall from Section 7.2.1 some useful standard material about the Curie-Weiss model (CW) in the canonical ensemble (or lattice-gas interpretation) which is the natural setting for our purpose. The Hamiltonian is

$$H_N = -\frac{J}{N} \sum_{\langle i,j \rangle} s_i s_j, \quad (8.1)$$

where the spins $s_i = \pm 1$ are attached to the N vertices of a complete graph. In (9.3) the sum over $\langle i, j \rangle$ carries over all edges of the graph and we take a ferromagnetic coupling $J > 0$. In the sequel we absorb the inverse temperature in this parameter. The free energy, for a fixed magnetization $m = \frac{1}{N} \sum_{i=1}^N s_i$, is

$$\Phi_N(m) = -\frac{1}{N} \ln Z_N, \quad Z_N = \sum_{s_i: m = \frac{1}{N} \sum_{i=1}^N s_i} e^{-H_N} \quad (8.2)$$

It has a well defined thermodynamic limit as in (7.5)

$$\lim_{N \rightarrow +\infty} \Phi_N(m) \equiv \Phi(m) = -\frac{J}{2} m^2 - \text{ent}(m) \quad (8.3)$$

In the canonical formalism the equation of state is simply

$$h = \frac{\partial \Phi(m)}{\partial m} = -Jm + \frac{1}{2} \ln \frac{1+m}{1-m}, \quad (8.4)$$

which is equivalent to the Curie-Weiss mean field equation

$$m = \tanh(Jm + h). \quad (8.5)$$

As is well known, from the van der Waals curve $h(m)$ (8.4), one can derive an equation of state that satisfies thermodynamic stability requirements from a Maxwell construction. Similarly a physical free energy is given by the convex envelope of (8.3). The van der Waals curve for two values of J (below and above the critical temperature $J = 1$) is plotted in Figure 8.2.1. We refer to Section 7.2.1 for a detailed description of the different states regarding the van der Waals curve of the CW model. The following expressions valid for $J > 1$, will be useful in the sequel,

$$\begin{cases} h_{sp} &= -\sqrt{J(J-1)} + \frac{1}{2} \ln \frac{J+\sqrt{J-1}}{J-\sqrt{J-1}} \approx \frac{1}{3}(J-1)^{\frac{3}{2}}, \\ m_{sp} &= \sqrt{\frac{J-1}{J}} \approx \sqrt{J-1}, \end{cases} \quad (8.6)$$

and

$$m_{\pm} \approx \pm \sqrt{3(J-1)}. \quad (8.7)$$

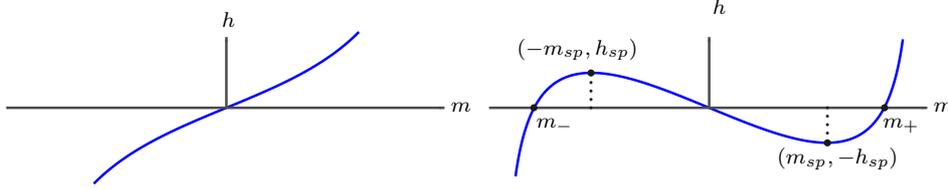


Figure 8.1: Left: van der Waals curve in the high temperature phase $J < 1$. Right: low temperature phase $J > 1$. For $m \notin (m_-, m_+)$ the curve describes stable equilibrium states and for $m \in (m_-, -m_{sp}) \cup (m_{sp}, m_+)$ it describes the metastable states. For $m \in (-m_{sp}, m_{sp})$ the system is unstable. The Maxwell plateau describes superpositions of m_- and m_+ states. That is, in the Maxwell plateau the oscillatory region between m_- and m_+ is replaced by a straight line at height $h = 0$.

In these formulas \approx means that $J \rightarrow 1_+$. The first order phase transition line is $(h_c = 0, J > 1)$ and terminates at the critical second order phase transition point $(h_c = 0, J = 1_+)$. For $J < 1$ and $h = 0$, we have $m_{\pm} = 0$. We will see that for the chain models the difference between the first order phase transition and spinodal thresholds becomes much smaller, and in fact vanishes exponentially fast with the width of the coupling along the chain.

8.2.2 Chain Curie-Weiss Model

Consider $2L + 1$ integer positions $z = -L, \dots, +L$ on a one dimensional line. At each position we attach a single CW spin system. The spins of each system are labeled as s_{iz} , $i = 1, \dots, N$, and are subjected to a magnetic field h . The spin-spin coupling is given by

$$-\frac{1}{N} J_{z,z'} s_{iz} s_{jz'} = -\frac{J}{Nw} g(w^{-1}|z - z'|) s_{iz} s_{jz'} \quad (8.8)$$

where the function $g(|x|)$ satisfies the following requirements:

a) It takes non-negative values and is independent of i, j and L . It may depend on w itself (see comments below) however we still write $g(|x|)$ instead of $g_w(|x|)$.

b) Has finite support $[-1, +1]$, i.e $g(|x|) = 0$ for $|x| > 1$.

c) It satisfies the normalization condition

$$\frac{1}{w} \sum_{z=-\infty}^{+\infty} g(w^{-1}|z|) = 1. \quad (8.9)$$

This is a purely ferromagnetic interaction which is of Kac type in the one dimensional z direction and is purely mean field in the transverse "infinite dimensional" direction. Condition a) ensures that we can find asymptotically

(as $z \rightarrow \pm\infty$) translation invariant states. Allowing for sign variations certainly leads to a richer phase diagram and is beyond the scope of this paper. Conditions b) and c) can easily be weakened without changing the main results at the expense of a slightly more technical analysis. One could allow for functions that have infinite support and decay fast enough (with finite second moment) at infinity. The normalization condition is set up so that the strength of the total coupling of one spin to the rest of the system equals J as $N \rightarrow +\infty$ (as in the individual CW system). For any given function $\tilde{g}(|x|)$ that is summable, we can always construct one that satisfies this condition $g(|x|) = w\tilde{g}(|x|)/\sum_{z=-\infty}^{+\infty}\tilde{g}(w^{-1}|z|)$. This means that in general $g(|x|)$ will depend explicitly on w ; however we could relax this slight fine tuning by taking the normalization condition to hold only asymptotically as $w \rightarrow +\infty$, namely that $\int_{-\infty}^{+\infty}g(|x|)dx = 1$.

The Hamiltonian is

$$H_{N,L} = -\frac{1}{N} \sum_{\langle iz, jz' \rangle} J_{z,z'} s_{iz} s_{jz'}. \quad (8.10)$$

The first sum carries over all pairs $\langle iz, jz' \rangle$ (counted once each) with $i, j = 1, \dots, N$ and $z, z' = -L, \dots, L$. We will adopt a canonical ensemble with

$$m = \frac{1}{(2L+1)N} \sum_{i=1, z=-L}^{N,L} s_{iz} \quad (8.11)$$

fixed. The partition function $Z_{N,L}$ is defined by summing $e^{-H_{N,L}}$ over all spin configurations $\{s_{iz} = \pm 1, i = 1, \dots, N; z = -L, \dots, L\}$ satisfying (8.11).

We now show that the free energy $f_{N,L} = -\frac{1}{N(2L+1)} \ln Z_{N,L}$ is given by a variational principle. Let us introduce a magnetization density at position z

$$m_z = \frac{1}{N} \sum_{i=1}^N s_{iz}, \quad (8.12)$$

and a matrix

$$D_{z,z'} = J_{z,z'} - J\delta_{z,z'}. \quad (8.13)$$

This matrix is symmetric and for any $z' = -L, \dots, +L$ it satisfies

$$\sum_{z=-L}^L D_{z,z'} \leq J \mathbb{I}(|z' \pm L| \leq w) \quad (8.14)$$

The important point here is that the row sum of (8.13) vanishes except for z' close to the boundaries. In this respect one may think of (8.13) as one-dimensional Laplacian matrix and, as we will see, this becomes exactly the case in an appropriate continuum limit of the model. The Hamiltonian can be re-expressed as (up to a constant)

$$H_{N,L} = -\frac{N}{2} \sum_{z,z'=-L}^L D_{z,z'} m_z m_{z'} - \frac{NJ}{2} \sum_{z=-L}^L m_z^2 \quad (8.15)$$

In the thermodynamic limit the magnetization density becomes a continuous variable $m_z \in [-1, +1]$ and the partition sum becomes (up to irrelevant prefactors)

$$Z_{N,L} = \int_{[-1,+1]^{2L+1}} \prod_{z=-L}^L dm_z \delta\left((2L+1)m - \sum_{z=-L}^L m_z\right) \times \exp -N \left(-\frac{1}{2} \sum_{z,z'=-L}^L D_{z,z'} m_z m_{z'} + \Phi(m_z) \right). \quad (8.16)$$

This integral can be interpreted as the canonical partition function of a one dimensional chain of continuous compact spins $m_z \in [-1, +1]$, at nearly zero temperature N^{-1} , with Hamiltonian

$$\Phi_L[\{m_z\}] = -\frac{1}{2} \sum_{z,z'=-L}^L D_{z,z'} m_z m_{z'} + \sum_{z=-L}^L \Phi(m_z). \quad (8.17)$$

The free energy of the finite chain obtained from (8.16) is

$$F_L(m) = - \lim_{N \rightarrow +\infty} \frac{1}{N} \ln Z_{N,L} = \min_{m_z: \sum_z m_z = (2L+1)m} \Phi_L[\{m_z\}]. \quad (8.18)$$

The solutions of this variational problem satisfy the set of equations

$$\begin{cases} \sum_{z'=-L}^L D_{z,z'} m_{z'} = \Phi'(m_z) - \lambda \\ m = \frac{1}{2L+1} \sum_{z=-L}^L m_z, \end{cases} \quad (8.19)$$

where λ is a Lagrange multiplier associated to the constraint (and where Φ' denotes the derivative of the function Φ). Denote by (λ^*, m_z^*) a solution of (8.19) for given m . The van der Waals equation of state is then given by the usual thermodynamic relation

$$h = \frac{1}{2L+1} \frac{\partial F_L(m)}{\partial m}. \quad (8.20)$$

In fact $h = \lambda^*$. Indeed, differentiating in (8.20) thanks to the chain rule and then using (8.19) yields,

$$\begin{aligned} h &= \frac{1}{2L+1} \sum_{z=-L}^L \left(- \sum_{z'=-L}^L D_{z,z'} m_{z'}^* + \Phi'(m_z^*) \right) \frac{dm_z^*}{dm} \\ &= \frac{\lambda^*}{2L+1} \sum_{z=-L}^L \frac{dm_z^*}{dm} \\ &= \lambda^* \end{aligned} \quad (8.21)$$

Let us make a few remarks on alternative forms for the above equations. First, summing over z the first equation in (8.19) we obtain thanks to (8.14)

$$h = \frac{1}{2L+1} \sum_{z=-L}^L \Phi'(m_z^*) + O\left(\frac{w}{L}\right) \quad (8.22)$$

Second, using the explicit expression for the potential $\Phi(m_z)$, equation (8.19) for the minimizing profiles can be cast in the form

$$\begin{cases} m_z^* = \tanh\{Jm_z^* + h + \sum_{z'=-L}^{+L} D_{z,z'} m_{z'}^*\}, \\ m = \frac{1}{2L+1} \sum_{z=-L}^L m_z^*. \end{cases} \quad (8.23)$$

This is a generalization of the CW equation to the chain model. We discuss a continuum version of the equation in the next section.

For $J \leq 1$ the single CW system has a unique equilibrium magnetization so we expect a unique translation invariant solution for (8.23), namely $m_z^* = m$ (neglecting boundary effect). It then follows that the van der Waals curve of the chain model is the same as that of the single CW model. On the other hand for $J > 1$ the solutions of (8.19) display non-trivial kink-like magnetization profiles. These solutions are responsible for an interesting oscillating structure in the van der Waals curve. This is investigated both numerically and to some extent analytically in the next two sections.

Before closing this section we want to point out that the same system can be analyzed in the grand-canonical ensemble (always from the lattice gas perspective) by adding an external magnetic field term $-h \sum_{i,z} s_{iz}$ to the Hamiltonian (8.15). The definition of the model is completed by imposing the boundary conditions:

$$\frac{1}{N} \sum_{i=1}^N s_{i,\pm L} = m_{\pm}(h), \quad (8.24)$$

where $m_{\pm}(h)$ are the local minima of $\Phi(m) - hm$. Note that when the minimum is unique (for $J \leq 1$ or $J > 1$ and $|h| \geq h_{sp}$) the two boundary conditions $m_{\pm}(h)$ are simply equal. The free energy (or minus the pressure of the lattice gas) is given by the variational problem

$$\min_{m_z, m_{\pm L} = m_{\pm}(h)} \left(-\frac{1}{2} \sum_{z,z'=-L}^L D_{z,z'} m_z m_{z'} + \sum_{z=-L}^L (\Phi(m_z) - hm_z) \right) \quad (8.25)$$

The critical points of this functional satisfy

$$\begin{cases} \sum_{z'=-L}^{+L} D_{z,z'} m_{z'} = \Phi'(m_z) - h \\ m_{\pm L} = m_{\pm}(h) \end{cases} \quad (8.26)$$

which is also equivalent to

$$\begin{cases} m_z = \tanh\{Jm_z + h + \sum_{z'=-L}^{+L} D_{z,z'} m_{z'}\} \\ m_{\pm L} = m_{\pm}(h). \end{cases} \quad (8.27)$$

The solutions of (8.26) or (8.27) define curves $m_z^*(h)$. Proving the existence of these curves is beyond our scope here; in general these are not single valued because the solutions are not unique for a given h . The van der Waals relation $h(m)$ can be recovered from these curves by using

$$m = \frac{1}{2L+1} \sum_{z=-L}^L m_z^*(h) \quad (8.28)$$

The magnetization profiles of the canonical and grand-canonical ensembles only differ near the boundaries. Their bulk behavior which is our interest are identical. In this paper this is verified numerically (Section 8.4). In the next section we find it more convenient to refer to the grand-canonical formalism (8.26), (8.27), (8.28).

8.3 A Continuum Approximation

The asymptotic limit of $L \gg w \gg 1$ reduces the solution of equations (8.26), (8.27), (8.28) to a problem of Newtonian mechanics. In this limit we obtain a non-linear integral equation which cannot be solved exactly; but whose solutions can be qualitatively discussed for any fixed $J > 1$ (an exact solution for all $J > 1$ is provided in a special case). Near the critical point $J \rightarrow 1_+$ this equation is solved and the solutions used to compute an approximate version of the van der Waals curve. In this way all the features of the numerical solution are reproduced. Usually continuum limits are obtained when a lattice spacing a between neighboring sites of the chain is sent to zero. This set up can also be explored for the present model and one finds that it is non trivial only near the critical point $J \rightarrow 1_+$, where it yields qualitatively identical results to the limit $w \rightarrow +\infty$, $J \rightarrow 1_+$. Away from the critical point ($J > 1$) $a \rightarrow 0$ is a trivial limit which supports only homogeneous states, contrary to the $w \rightarrow +\infty$ limit which displays non trivial features for all $J > 1$.

Asymptotics for $L \gg w \gg 1$. We set

$$z = wx, \quad m_z = m_{wx} \equiv \mu(x) \quad (8.29)$$

so equation (8.26) is equivalent to

$$\frac{J}{w} \sum_{z'=-L}^L \left\{ g\left(|x - \frac{z'}{w}\right| - w\delta_{x, \frac{z'}{w}} \right\} \mu\left(\frac{z'}{w}\right) = \Phi'(\mu(x)) - h. \quad (8.30)$$

We take the limits $L \rightarrow +\infty$ first and $w \rightarrow +\infty$ second, so that this equation becomes

$$J \int_{-\infty}^{+\infty} dx' \{g(|x'|) - \delta(x')\} \mu(x+x') = \Phi'(\mu(x)) - h. \quad (8.31)$$

which can also be cast in a more elegant form (* denotes convolution)

$$\tanh(Jg * \mu + h) = \mu. \quad (8.32)$$

We cannot solve this equation in general, except for the special case of uniform g . Equ. (8.32) for $h = 0$ appears in [62], [63] and existence plus properties of solutions has been discussed. For our purpose a qualitative discussion of its solutions suffices and we briefly outline it for the reader's convenience. For $|x| \gg 1$ we can expand $\mu(x + x')$ to second order (in (8.31)) since $g(|x|)$ vanishes for $|x| > 1$. This yields the approximate equation

$$J\kappa\mu''(x) \approx \Phi'(\mu(x)) - h, \quad \kappa = \frac{1}{2} \int_{-\infty}^{+\infty} dx' x'^2 g(|x'|). \quad (8.33)$$

We recognize here Newton's second law for a particle moving in the *inverted potential* $-\Phi(\mu(x))$ where $\mu(x)$ is the particle's position at time x and $J\kappa$ its mass. Note this is not a Cauchy problem with fixed initial position and velocity, but a boundary value problem with $\lim_{x \rightarrow \pm\infty} \mu(x) = m_{\pm}(h)$; the boundary conditions automatically fix the initial and final velocities. The nature of the solutions can be deduced by applying the conservation of mechanical energy for a ball rolling in the inverted potential. For $J < 1$ the inverted potential has a single maximum at $m_+(h) = m_-(h)$ and the only solution is $\mu(x) = m_{\pm}(h)$, corresponding to a homogeneous state. In fact this is also true for the integral equation. Now we consider $J > 1$ and $h = 0$. At time $-\infty$ the particle is on the left maximum and starts rolling down infinitely slowly, then spends a finite time in the bottom of the potential well, and finally climbs to the right maximum infinitely slowly to reach it at time $+\infty$. For the magnetization profile m_z this translates to a kink-like state. Note that the center of the kink is set by the normalization condition (8.28), and thus we have a continuum of solutions parametrized by the parameter m on the Maxwell plateau $[m_-, m_+]$. For $J > 1$ and $h > 0$, the particle starts with a positive initial velocity, rolls down the potential well, and finally reaches the right maximum infinitely slowly. Thus $\mu(x) = m_+(h)$ for all x except for an interval of width $O(1)$ near the left boundary at minus infinity. This translates into an essentially constant magnetization profile with a fast transition layer near the left boundary. For $J > 1$ and $h < 0$ the picture is similar.

These arguments imply that in a first approximation (L and w infinite) the van der Waals curve of the chain-CW system is given by the Maxwell construction of the single CW system. In order to get the finer structure around the Maxwell plateau we have to do a more careful finite size analysis.

Asymptotics for $L \gg w \gg 1$ large and $J \rightarrow 1_+$. Now we set

$$t = \sqrt{J-1}x, \quad \mu(x) = \mu\left(\frac{t}{\sqrt{J-1}}\right) \equiv \sqrt{J-1}\sigma(t) \quad (8.34)$$

and look at the regime $J \rightarrow 1_+$. A straightforward calculation shows that the

left hand side of equation (8.31) becomes

$$\frac{J(J-1)^{\frac{3}{2}}}{2} \left\{ \int_{-\infty}^{+\infty} dx g(|x|) x^2 \right\} \sigma''(t) + O((J-1)^{\frac{5}{2}}), \quad (8.35)$$

and that the right hand side becomes

$$(J-1)^{\frac{3}{2}}(-\sigma(t) + \frac{1}{3}\sigma(t)^3) - h + O((J-1)^{\frac{5}{2}}). \quad (8.36)$$

Lastly, we set $\tilde{h} = h(J-1)^{-\frac{3}{2}}$, and thus from (8.31), (8.35), (8.36)

$$\kappa \sigma''(t) = -\sigma(t) + \frac{1}{3}\sigma(t)^3 - \tilde{h}. \quad (8.37)$$

Again, this is Newton's second law for a particle of mass κ moving in the inverted potential

$$V(\sigma) = \frac{1}{2}\sigma^2 - \frac{1}{12}\sigma^4 + \tilde{h}\sigma. \quad (8.38)$$

The boundary conditions (8.26) mean that the initial and final positions of the particle for $t \rightarrow \pm\infty$ are the solutions of

$$\sigma_{\pm} - \frac{1}{3}\sigma_{\pm}^3 + \tilde{h} = 0, \quad (8.39)$$

corresponding to the local maxima of the potential. Initial and final velocities are automatically fixed by the requirement that $\lim_{t \rightarrow \pm\infty} \sigma(t) = \sigma_{\pm}$.

Summarizing, in the limit

$$\lim_{J \rightarrow 1+; h(J-1)^{-\frac{3}{2}} \text{ fixed}} \lim_{w \rightarrow +\infty} \lim_{L \rightarrow +\infty} \quad (8.40)$$

the magnetization profile is

$$m_z \approx \sqrt{J-1} \sigma \left(\sqrt{J-1} \frac{z}{w} \right) \quad (8.41)$$

where $\sigma(t)$ is a solution of (8.37).

Kink states. For $\tilde{h} = 0$ (meaning $h = 0$) (8.37) has the well known solutions

$$\sigma^{\text{kink}}(t) = \sqrt{3} \tanh \left\{ \frac{t - \tau}{\sqrt{2\kappa}} \right\} \quad (8.42)$$

The center τ of the kink is a parameter that we have to fix from the normalization condition. From (8.41) and (8.42) we have

$$\begin{aligned} \frac{1}{2L+1} \sum_{z=-L}^{+L} m_z &\approx \frac{\sqrt{3(J-1)}}{2L+1} \sum_{z=-L}^{+L} \tanh \left(L \frac{\sqrt{J-1}}{w\sqrt{2\kappa}} \left(\frac{z}{L} - \frac{w\tau}{L\sqrt{J-1}} \right) \right) \\ &\approx \frac{\sqrt{3(J-1)}}{2} \int_{-\infty}^{+\infty} dx \operatorname{sign} \left(\frac{\sqrt{J-1}}{w\sqrt{2\kappa}} \left(x - \frac{w\tau}{L\sqrt{J-1}} \right) \right) \end{aligned}$$

$$\approx \frac{\sqrt{3}w\tau}{L} \quad (8.43)$$

Since this sum must be equal to m we find $\tau \approx \frac{mL}{\sqrt{3}w}$. The net result for the magnetization profile is

$$m_z^{\text{kink}} \approx \sqrt{3(J-1)} \tanh \left\{ \frac{1}{w} \sqrt{\frac{J-1}{2\kappa}} \left(z - \frac{mL}{\sqrt{3(J-1)}} \right) \right\} \quad (8.44)$$

Homogeneous states. When $\tilde{h} \neq 0$ the solution cannot be put in closed form. To lowest order in \tilde{h} the solutions of (8.39) are $\sigma_{\pm} = \pm\sqrt{3} + \tilde{h}$. The initial velocity is (assuming the final velocity is zero) to leading order,

$$\sqrt{\frac{2}{\kappa}(V(\sigma_+) - V(\sigma_-))} \approx 2 \frac{3^{1/4}}{\tilde{h}^{1/2}} \quad (8.45)$$

Thus, roughly speaking, the particle travels with constant velocity $2 \frac{3^{1/4}}{\sqrt{\kappa}} \tilde{h}^{1/2}$ from position $-\sqrt{3} + \tilde{h}$ during a finite time $O(\frac{3^{1/4}\sqrt{\kappa}}{\tilde{h}^{1/2}})$ and then stays exponentially close to the final position $\sqrt{3} + \frac{\tilde{h}}{2}$. The magnetization profile is

$$m_z \approx \begin{cases} -\sqrt{3(J-1)} + \frac{h}{2(J-1)} + \frac{2}{\sqrt{\kappa}}(3(J-1))^{1/4}h^{1/2}\left(\frac{z+L}{w}\right), & -L \leq z \leq -L + O\left(\frac{w\sqrt{\kappa}}{2(3(J-1))^{1/4}h^{1/2}}\right) \\ \sqrt{3(J-1)} + \frac{h}{2(J-1)}, & z \geq -L + O\left(\frac{w\sqrt{\kappa}}{2(3(J-1))^{1/4}h^{1/2}}\right) \end{cases} \quad (8.46)$$

Comparison of free energies. In this paragraph we compute a naive approximation for the free energy (8.18). First consider the energy difference $F_L^{\text{kink}} - F_L^{\text{const}}$ between kink m_z^{kink} and constant $m_z^{\text{const}} = m$ states both with total magnetization $m_- < m < m_+$ on the Maxwell plateau. We write this as

$$F_L^{\text{kink}} - F_L^{\text{const}} = (F_L(m_{\pm}) - F_L^{\text{const}}) + (F_L^{\text{kink}} - F_L(m_{\pm})) \quad (8.47)$$

Because of (8.14) the first term is easily estimated as $(2L+1)(\Phi(m_{\pm}) - \Phi(m)) + O(w)$ which is negative for m on the Maxwell plateau. Since the magnetization density of the kink state tends exponentially fast to m_{\pm} for $z \rightarrow \pm\infty$ the second term is clearly $O(w)$ and therefore for L large the kink states are stable³. But our interest here is in a precise calculation of this second term which displays an interesting oscillatory structure.

$$F_L^{\text{kink}} - F_L(m_{\pm}) = -\frac{1}{2} \sum_{z,z'=-L}^L D_{z,z'} (m_z^{\text{kink}} m_{z'}^{\text{kink}} - m_{\pm}^2) \quad (8.48)$$

³this argument breaks down for $|m - m_{\pm}| = O(L^{-\frac{1}{2}})$; this is discussed in Section 8.4

$$+ \sum_{z=-L}^L (\Phi(m_z^{\text{kink}}) - \Phi(m_{\pm}))$$

Using (8.44) and (8.14) it is easy to see that, in the bulk, (8.48) is a periodic function of m with period $\frac{\sqrt{3(J-1)}}{L}$, as long as the center of the kink is in the bulk. To compute it we first extend the sums to infinity and use the Poisson summation formula

$$\sum_{z \in \mathbb{N}} F(z) = \sum_{k \in \mathbb{N}} \int_{-\infty}^{+\infty} dz e^{2\pi i k z} F(z) \quad (8.49)$$

for

$$F(z) = -\frac{1}{2} \sum_{z'=-\infty}^{+\infty} D_{z,z'} (m_z^{\text{kink}} m_{z'}^{\text{kink}} - m_{\pm}^2) + \Phi(m_z^{\text{kink}}) - \Phi(\sqrt{3(J-1)}) \quad (8.50)$$

A look at (8.44) shows that it has poles in the complex plane at $z_n = \frac{mL}{\sqrt{3(J-1)}} + i\pi(n + \frac{1}{2})w\sqrt{\frac{2\kappa}{J-1}}$, $n \in \mathbb{N}$. This suggests that the first term in (8.50) has the same pole structure. The second term involving the potential is more subtle because its exact expression involves a logarithm which induces branch cuts. However one can show, keeping the true expression for the potential, that the branch cuts are outside of a strip $|\Im(z)| < \frac{\pi}{2}w\sqrt{\frac{2\kappa}{J-1}}$, and therefore $F(z)$ is analytic in this strip. This is enough to deduce from standard Paley-Wiener theorems that for $w\sqrt{\frac{2\kappa}{J-1}}$ large $|F(k)| = O(e^{-|k|w\sqrt{\frac{2\kappa}{J-1}}(\pi^2 - \epsilon)})$. In the appendix we perform a detailed analysis to show (for $J \rightarrow 1_+$, w large and k fixed)

$$\int_{-\infty}^{+\infty} dz e^{2\pi i k z} F(z) \approx 4(J-1)\kappa w^2 \pi^2 k (1 - k^2 \frac{\pi^2 w^2 \kappa}{J-1}) \sinh^{-1} \left(k \pi^2 w \sqrt{\frac{2\kappa}{J-1}} \right) \quad (8.51)$$

Retaining the dominant terms $k = 0$ and $k = \pm 1$ in the Poisson summation formula we find for the free energy ($m_- < m < m_+$)

$$F_L^{\text{kink}}(m) \approx (2L+1)\Phi(m_{\pm}) + 4w(J-1)^{3/2} \sqrt{\frac{\kappa}{2}} - 16(\pi w)^4 \kappa^2 e^{-\pi^2 w \sqrt{\frac{2\kappa}{J-1}}} \cos \left(2\pi m \frac{L}{\sqrt{3(J-1)}} \right) \quad (8.52)$$

This result confirms the Maxwell construction, namely that the free energy per unit length converges to the convex envelope of $\Phi(m)$. The finite size corrections display an interesting structure. The first correction $O((J-1)^{3/2})$ comes from the zero mode and represents the "surface tension" of the kink

interface. The oscillatory term is a special feature of coupled mean field models. According to formula (8.44) $\frac{mL}{\sqrt{3(J-1)}}$ is the position of the kink, thus the profiles centered at integer positions correspond to minima of the periodic potential and are stable, while those centered at half-integer positions correspond to maxima and are therefore unstable states. The energy difference between a kink centered at an integer and one centered at a neighboring half-integer is a Peierls-Nabarro barrier

$$32(\pi w)^4 \kappa^2 e^{-\pi^2 w \sqrt{\frac{2\kappa}{J-1}}}. \quad (8.53)$$

This is the energy needed to displace the kink along the chain. Such energy barriers are usually derived within effective soliton like equations for the motion of defects in crystals [65]. Here the starting point was a microscopic statistical mechanics model.

Oscillations of the van der Waals curve. The van der Waals curve is easily obtained ($m_- < m < m_+$)

$$\begin{aligned} h &= \frac{1}{2L+1} \frac{\partial F_L^{\text{kink}}(m)}{\partial m} = \frac{1}{2L+1} \frac{\partial}{\partial m} (F_L^{\text{kink}} - F_L(m_{\pm})) \\ &\approx \frac{16\pi(\pi w)^4 \kappa^2}{\sqrt{3(J-1)}} e^{-\pi^2 w \sqrt{\frac{2\kappa}{J-1}}} \sin\left(2\pi m \frac{L}{\sqrt{3(J-1)}}\right) \end{aligned} \quad (8.54)$$

At this point we note that the limit $L \rightarrow +\infty$ and $\frac{\partial}{\partial h}$ do not commute. This is so because on the Maxwell plateau we have a sequence of transitions⁴ from one kink state to another. In accordance with the numerical calculations, we find a curve that oscillates around the Maxwell plateau $m \in [-\sqrt{3(J-1)}, +\sqrt{3(J-1)}]$ with a period $O(\frac{\sqrt{3(J-1)}}{L})$. The amplitude of these oscillations is exponentially small with respect to w and thus much smaller than the height $O((J-1)^{3/2})$ of the spinodal points (see (8.6)). For example for the uniform coupling function we have $\kappa = 1/6$ and the amplitude of the oscillations is $O(e^{-\pi^2 w \sqrt{\frac{1}{3(J-1)}}})$.

Uniform interaction: $h = 0$ and all J . In case of a uniform interaction along the chain $g(|x|) = \frac{1}{2}$, $|x| \leq 1$ and 0 otherwise, it turns out that equation (8.32) has the exact solution

$$\mu(x) = m_{\pm} \tanh Jm_{\pm}(x - x_0) \quad (8.55)$$

for all $h = 0$ and J . This can be checked directly by inserting the function in (8.32) and seeing that it reduces to the CW equation for m_{\pm} . Of course this solution is non trivial only for $J > 1$. Relating the center x_0 to the total magnetization we get the magnetization profile

$$m_z \approx m_+ \tanh \left\{ \frac{Jm_{\pm}}{w} \left(z - \frac{m}{m_{\pm}} L \right) \right\} \quad (8.56)$$

⁴these can be thought as first order phase transitions with infinitesimal jump discontinuities

Here \approx means that $L \gg w \gg 1$. One can check that the formula reduces to (8.44) when $J \rightarrow 1_+$. With this expression one can compute an exact formula for the exponent of the amplitude of oscillations of the van der Waals curve. Indeed as argued after (8.50) this exponent is solely determined by the location of the poles of (8.56) for $z \in \mathbb{C}$. Therefore we obtain for the case of the uniform interaction and all $J > 1$,

$$h = C(w, J)e^{-\frac{\pi^2 w}{Jm_{\pm}}} \sin\left(2\pi \frac{m}{m_{\pm}} L\right) \quad (8.57)$$

where $C(J, w)$ is a prefactor that could in principle be computed by extending the calculation of the Appendix. Up to this prefactor, the Peierls-Nabarro barrier is $e^{-\frac{\pi^2 w}{Jm_+}}$ for all $J > 1$.

Remarks. The main features of these oscillations, their period and exponentially small amplitude, are independent of the details of the exact model and its free energy. Only the prefactor will depend on such details. The period is equal to $\frac{m_+ - m_-}{2L}$ where $m_+ - m_-$ is the width of the Maxwell plateau. The wiggles have an amplitude $e^{-2\pi\Delta}$ where Δ is the width of a strip in \mathbb{C} where the kink profile is analytic (when the position variable z is continued to \mathbb{C}). In general we have $\Delta = \alpha \frac{w\pi}{2Jm_+}$ where $\alpha = O(1)$. For the uniform window $\alpha = 1$ and in general when $J \rightarrow 1_+$ we have $\alpha \rightarrow \kappa^{-1/2}$. The point here is that the amplitude of the wiggles does not depend on the details of the free energy but only on the locations of the singularities m_z^{kink} in the complex plane. If an explicit formula is not available for the kink profiles Δ can still be estimated by numerically computing the discrete Fourier transform of the kink and identifying Δ with its rate of decay. This quantity will always be proportional to the scale factor w in m_z^{kink} .

8.4 Numerical Solutions

We have carried out the numerical computations both for the equations in the canonical and grand-canonical formulations. These confirm the analytical predictions for the oscillations of the van der Waals curve. Near the end points of the Maxwell plateau the situation is not identical for the canonical and grand-canonical ensembles because boundary effects become important. For simplicity we start with the grand-canonical formulation.

Grand-canonical equations. It is convenient to solve a slightly different system of equations than (8.27) in order to eliminate boundary effects (one may think of this as a modification of the model at the boundaries of the chain)

$$\begin{cases} m_z = \tanh\left\{Jm_z + h + \sum_{z'=-L-w+1}^{+L+w-1} D_{z,z'} m_{z'}\right\}, & -L \leq z \leq +L \\ m_z = m_+(h), & L+1 \leq z \leq L+w-1 \\ m_z = m_-(h), & -L-w+1 \leq z \leq -L-1. \end{cases} \quad (8.58)$$

In other words, we force the profile to equal $m_-(h)$ at extra positions $-L-w+1$ to $-L-1$ and to $m_+(h)$ at extra positions $L+1$ to $L+w-1$. The van der Waals relation $h(m)$ is recovered from the solutions $m_z^*(h)$ of (8.58) by using (8.28). The first equation is equivalent to

$$h = -(J + D_{zz})m_z + \tanh^{-1} m_z - \sum_{z'=-L-w+1, z' \neq z}^{L+w-1} D_{zz'} m_{z'} \quad (8.59)$$

Summing over z and using (8.28) we obtain

$$h = -(J + D_{zz})m + \frac{1}{2L+1} \sum_{z=-L}^L \left\{ \tanh^{-1} m_z - \sum_{z'=-L-w+1, z' \neq z}^{L+w-1} D_{zz'} m_{z'} \right\} \quad (8.60)$$

Also, (8.59) is equivalent to

$$m_z(J + D_{z,z} - 1) = \tanh^{-1} m_z - m_z - \sum_{z'=-L-w+1, z' \neq z}^{L+w-1} D_{z,z'} m_{z'} - h \quad (8.61)$$

The last two equations are the basis of:

Procedure 4 Iterative solutions of (8.58)

- 1: Fix m . Initialize $m_z^{(0)} = m$ for $-L \leq z \leq L$ and $h^{(0)} = 0$.
- 2: From $m_z^{(t)}$ compute:

$$h^{(t+1)} \leftarrow (J + D_{z,z})m + \frac{1}{2L+1} \sum_{z=-L}^L \left\{ \tanh^{-1} m_z^{(t)} - \sum_{z'=-L-w+1, z' \neq z}^{L+w-1} D_{zz'} m_{z'}^{(t)} \right\}$$

- 3: For $-L \leq z \leq +L$, update $m_z^{(t+1)}$ as

$$m_z^{(t+1)} \leftarrow \frac{1}{J + D_{z,z} - 1} \left\{ \tanh^{-1} m_z^{(t)} - m_z^{(t)} - \sum_{z'=-L-w+1, z' \neq z}^{L+w-1} D_{z,z'} m_{z'}^{(t)} - h^{(t+1)} \right\}$$

and for a tunable value θ (for $\theta = 0.9$ the iterations are “smooth”)

$$m_z^{(t+1)} \leftarrow \theta m_z^{(t)} + (1 - \theta) m_z^{(t+1)}$$

- 4: For $-L-w+1 \leq z \leq -L-1$ let $m_z^{(t+1)} \leftarrow m_-(h^{(t+1)})$ and for $L+1 \leq z \leq L+w-1$ let $m_z^{(t+1)} \leftarrow m_+(h^{(t+1)})$.
 - 5: Continue until $t = T$ such that the ℓ_1 distance between the two consecutive profiles is less than some prescribed error δ . Output $h^{(T)}(m)$ and $m_z^{(T)}$.
-

Figures 8.2 and 8.3 show the output of this procedure for $L = 25$, $w = 1$, $g(0) = \frac{1}{2}$, $g(\pm 1) = \frac{1}{4}$. We see from Figure 8.3 that when $J = 1.1$, already for $w = 1$ the continuum approximation equ. (8.44) for the profile is good.

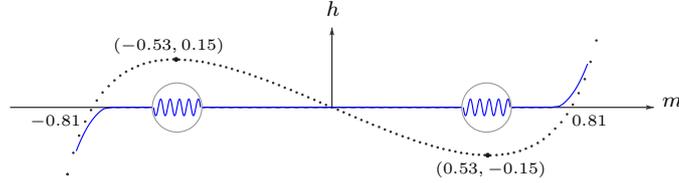


Figure 8.2: Dotted line: van der Waals curve of single system for $J = 1.4$. Continuous line: van der Waals isotherm for $J = 1.4$, $L = 25$, $w = 1$ and $g(0) = \frac{1}{2}$, $g(\pm 1) = \frac{1}{4}$. Circles: 40-fold vertical magnification. Throughout the plateau one has 50 wiggles corresponding to 50 stable kink states.

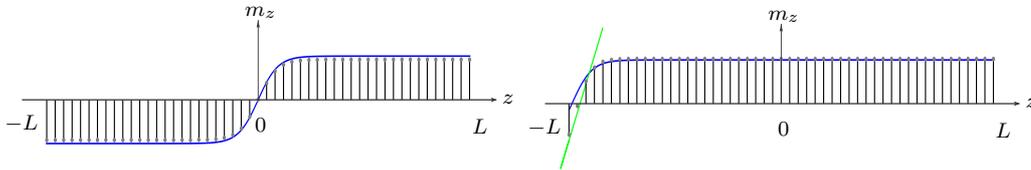


Figure 8.3: Vertical bars are the numerical values and the continuous lines (blue and green) are given by equations (8.44), (8.46). Left: kink state centered at $m = 0$ (so $h = 0$) and $J = 1.4$, $L = 25$, $w = 1$, $g(0) = \frac{1}{2}$, $g(\pm 1) = \frac{1}{4}$. Right: homogeneous solution for the same J , L , w , g and $h(m) = 0.017$.

Table 8.1 compares the numerical amplitude of the oscillations N_w for the van der Waals curve with the analytical formula (8.54)

$$\underbrace{\frac{16\pi(\pi w)^4 \kappa^2}{\sqrt{3(J-1)}}}_{C_w} \underbrace{\exp\left(-\pi^2 w \sqrt{\frac{2\kappa}{J-1}}\right)}_{E_w}. \tag{8.62}$$

We take $J = 1.05$, and the triangular window $g(|x|) = \frac{2w}{1+3w}(1 - \frac{|x|}{2})$. In order to get a stable result for $w = 3$ we have to go to lengths $L = 250$. We see that the agreement is quite good for the exponent while the pre-factor seems to be off by a constant factor $O(1)$.

For larger values of J and uniform window $g(|x|) = \frac{w}{2w+1}$ we can use formula (8.57) to compare the numerical amplitude N_w with $E_w = e^{-\frac{\pi^2 w}{Jm \pm}}$. Table 8.2 shows the results for $J = 1.4$ and $L = 80$.

Canonical equations. Let us now discuss the numerical solutions of (8.23).

Table 8.1: Amplitude of wiggles: $J = 1.05$ and triangular window.

w	N_w	E_w	C_w	$\frac{\log N_w}{\log C_w E_w}$	$\frac{\log \frac{N_w}{C_w}}{\log E_w}$	$\frac{\log \frac{N_w}{E_w}}{\log C_w}$
1	2.5×10^{-12}	2.8×10^{-14}	7.9×10^2	1.09	1.07	0.67
2	3.4×10^{-22}	9.3×10^{-25}	7.8×10^3	1.07	1.06	0.66
3	6.7×10^{-32}	5.1×10^{-35}	3.2×10^4	1.05	1.04	0.69
4	3.2×10^{-41}	3.3×10^{-45}	9.2×10^4	1.02	1.02	0.80

Table 8.2: Amplitude of wiggles: $J = 1.4$ and uniform window.

w	N_w	E_w	$\frac{\log N_w}{\log E_w}$
1	2.2×10^{-5}	1.7×10^{-4}	1.24
2	3.5×10^{-9}	3.0×10^{-8}	1.12
3	5.9×10^{-13}	5.2×10^{-12}	1.08
4	1.0×10^{-16}	9.0×10^{-16}	1.06

Here the boundary conditions are not forced at the outset and adjust themselves to non-trivial values when m is on the plateau. It turns out that for some values of m the output of iterations is greatly affected by the choice of the initial profile. Thus in order to find the correct global minimum of the canonical free energy a suitable initial condition must be chosen. A natural choice is to choose the solution of (8.58) as the initial point. The numerical procedure is as follows:

Procedure 5 Iterative solutions of (8.23)

- 1: Fix m . Initialize $m_z^{(0)}$ and $h^{(0)}$ to a solution of (8.58) given by algorithm 1.
- 2: From $m_z^{(t)}$ compute:

$$h^{(t+1)} \leftarrow (J + D_{z,z})m - \frac{1}{2L+1} \left\{ \sum_{z=-L}^L \tanh^{-1} m_z^{(t)} + \sum_{z=-L}^L \sum_{z'=-L, z' \neq z}^L D_{z,z'} m_{z'}^{(t)} \right\}$$

- 3: For $-L \leq z \leq +L$, first update $m_z^{(t+1)}$ as:

$$m_z^{(t+1)} \leftarrow \frac{1}{J + D_{z,z} - 1} \left\{ \tanh^{-1} m_z^{(t)} - m_z^{(t)} - \sum_{z'=-L, z' \neq z}^L D_{z,z'} m_{z'}^{(t)} - h^{(t+1)} \right\}$$

and for a tunable value θ (say $\theta = 0.9$),

$$m_z^{(t+1)} \leftarrow \theta m_z^{(t)} + (1 - \theta) m_z^{(t+1)}$$

- 4: Continue until $t = T$ such that the ℓ_1 distance between the two consecutive profiles is less than a prescribed error δ . Output $h^{(T)}$ and $m_z^{(T)}$.
-

Figure 8.4 shows the van der Waals curve for $J = 1.4$ with $L = 25$, $w = 1$

and $g(0) = \frac{1}{2}$, $g(\pm 1) = \frac{1}{2}$. Apart from the usual oscillations on the Maxwell plateau we observe that near the extremities (close to m_{\pm}) the curve follows the metastable branch of the single system. This can easily be explained from equ. (8.52). Indeed, the energy difference between a kink and constant state ($m_z = m$) is

$$(2L + 1)\Phi(m_{\pm}) + 4w(J - 1)^{3/2} \sqrt{\frac{\kappa}{2}} - (2L + 1)\Phi(m) \quad (8.63)$$

where we drop the exponentially small oscillatory contribution. When $|m - m_{\pm}|$ is very small this difference becomes positive because of the surface tension contribution of the kink, and the constant state is the stable state. It is easily seen that this happens for $(m - m_{\pm})^2 < \frac{2w}{2L+1} \sqrt{\frac{\kappa}{2}} (J-1)^{3/2}$. As seen in Figure 8.5 this boundary effect vanishes as L grows large. Finally Figure 8.6 displays magnetization profiles: in the bulk they are identical to the grand-canonical ones, while near the boundaries the magnetization is reduced since the effective ferromagnetic interaction is smaller.

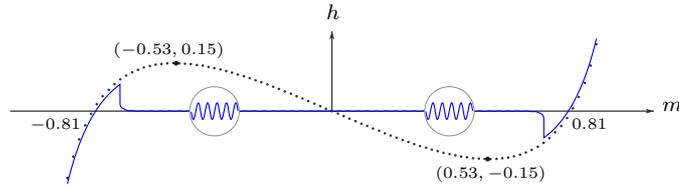


Figure 8.4: Dotted line: isotherm of single system for $J = 1.4$. Continuous line: isotherm of coupled model with $L = 25$, $w = 1$, $g(0) = \frac{1}{2}$, $g(\pm) = \frac{1}{2}$. Vertical magnification factor in the circle is 40. For $|m - m_{\pm}| = O(L^{-\frac{1}{2}})$ there is a boundary effect explained in main text.

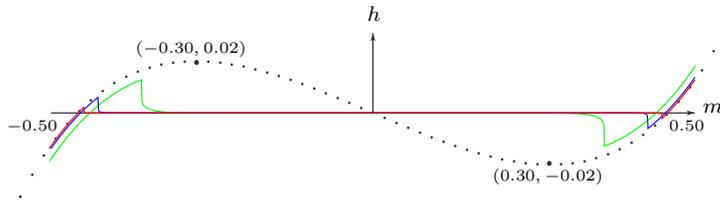


Figure 8.5: Behavior of the boundary effect for $J = 1.1$ (same w and g as above) and $L = 25, 100, 400$.

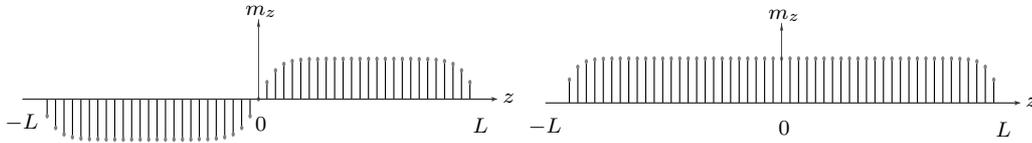


Figure 8.6: Magnetization profiles for $J = 1.1$, $L = 25$ (same w and g as above). Left: kink centered at $m = 0$. Right: homogeneous solution for $h(m) = 0.017$.

8.5 Further Remarks and Open Directions

We introduced the CW model as a “toy model” to understand the threshold saturation phenomenon. As mentioned in Chapter 7, the same phenomenon occurs in a wide variety of spatially coupled systems such as constraint satisfaction problems, compressed sensing, and communication systems.

What is the generic picture that emerges? All systems considered above are coupled chains of individual infinite dimensional systems or mean field systems. Indeed the individual systems are defined on sparse graphs or complete graphs, which are both, in some sense, infinite dimensional objects. Besides, their exact (or conjecturally exact) solutions are given by mean field equations (Curie-Weiss equation, cavity/replica equations, etc). These equations (for the individual system) have two stable fixed point solutions which describe the order parameter of the equilibrium states for the individual system. When boundary conditions are fixed such that the order parameter takes the two equilibrium values at the ends of the chain, the spatially coupled system has a series of new equilibrium states corresponding to kink profiles. Since the kink interface is well localized its free energy is close to a convex combination of the two free energies corresponding to the boundary conditions. Because of the discrete nature of the chain there are tiny free energy barriers corresponding to unstable positions for the kinks in-between two positions on the chain. This is the origin of the wiggles, both in the free energy functional (of CW or Landau or Bethe type) and in the van der Waals like curves.

There are many open questions that are worth investigating. To name a few principal ones, it is worth investigating the connections to coupled MAP systems, and discrete soliton equations and the stability of their solutions. This would allow to better understand whether the phenomenon occurs or not. Also the algorithmic implications of threshold saturation is a largely open direction.

8.6 Appendix

We give the main steps leading to formulas (8.51) and (8.52). First we notice that

$$\widehat{F}(k) \equiv \int_{-\infty}^{+\infty} dz e^{2\pi i k z} F(z) = e^{2\pi i k \frac{mL}{\sqrt{3(J-1)}}} \int_{-\infty}^{+\infty} dz e^{2\pi i k z} G(z) \quad (8.64)$$

with

$$G(z) = -\frac{1}{2} \sum_{z'=-\infty}^{+\infty} D_{z,z'}(m_z^0 m_{z'}^0 - m_{\pm}^2) + \Phi(m_z^0) - \Phi(\sqrt{3(J-1)}) \quad (8.65)$$

and m_z^0 is a kink centered at the origin,

$$m_z^0 = \sqrt{3(J-1)} \tanh\left\{\frac{1}{w} \sqrt{\frac{J-1}{2\kappa}} z\right\}. \quad (8.66)$$

Now we evaluate the sum over z' in the first term of (8.65). Setting $z' = wx'$ we have for w very large,

$$\begin{aligned} \sum_{z'=-\infty}^{+\infty} D_{z,z'} m_{z'}^0 &= \frac{J\sqrt{3(J-1)}}{w} \sum_{z'=-\infty}^{+\infty} (g(|\frac{z}{w} - \frac{z'}{w}|) - w\delta_{\frac{z}{w}, \frac{z'}{w}}) \tanh\left\{\sqrt{\frac{J-1}{2\kappa}} \frac{z'}{w}\right\} \\ &\approx J\sqrt{3(J-1)} \int_{-\infty}^{+\infty} dx' (g(|x'|) - \delta(x')) \tanh\left\{\sqrt{\frac{J-1}{2\kappa}} (x' + \frac{z}{w})\right\} \\ &\approx J\sqrt{3(J-1)} \kappa w^2 \frac{d^2}{dz^2} \tanh\left\{\frac{1}{w} \sqrt{\frac{J-1}{2\kappa}} z\right\} \end{aligned} \quad (8.67)$$

Therefore

$$\sum_{z'=-\infty}^{+\infty} D_{z,z'} m_{z'}^0 = -\sqrt{3} J(J-1)^{3/2} (1 - \tanh^2\left\{\frac{1}{w} \sqrt{\frac{J-1}{2\kappa}} z\right\}) \tanh\left\{\frac{1}{w} \sqrt{\frac{J-1}{2\kappa}} z\right\} \quad (8.68)$$

In a similar way one shows that the $-m_{\pm}^2$ term does not contribute, and one finds

$$\begin{aligned} G(z) &\approx \frac{3}{2} J(J-1)^2 (1 - \tanh^2\left\{\frac{1}{w} \sqrt{\frac{J-1}{2\kappa}} z\right\}) \tanh^2\left\{\frac{1}{w} \sqrt{\frac{J-1}{2\kappa}} z\right\} \\ &\quad + \Phi(\sqrt{3(J-1)} \tanh\left\{\frac{1}{w} \sqrt{\frac{J-1}{2\kappa}} z\right\}) - \Phi(\sqrt{3(J-1)}) \end{aligned} \quad (8.69)$$

Replacing in (8.64) we get after a scaling,

$$\hat{F}(k) = w \sqrt{\frac{2\kappa}{J-1}} e^{2\pi i k \frac{mL}{\sqrt{3(J-1)}}} \int_{-\infty}^{+\infty} dz e^{2\pi i k w \sqrt{\frac{2\kappa}{J-1}} z} \tilde{G}(z) \quad (8.70)$$

where

$$\begin{aligned} \tilde{G}(z) &\approx \frac{3}{2} J(J-1)^2 (1 - \tanh^2 z) \tanh^2 z \\ &\quad + \Phi(\sqrt{3(J-1)} \tanh z) - \Phi(\sqrt{3(J-1)}) \end{aligned} \quad (8.71)$$

As a function of $z \in \mathbb{C}$, $\tilde{G}(z)$ is analytic in the open strip $|\Im(z)| < \frac{\pi}{2}$. Indeed $\tanh z$ has poles at $z_n = (n + \frac{1}{2})i\pi$, $n \in \mathbb{Z}$ and Φ has branch cuts for

$\sqrt{3(J-1)} \tanh z \in]-\infty, -1] \cup [1, +\infty[$, or equivalently on the intervals

$$z \in \cup_{n \in \mathbb{Z}} \left[z_n, z_n - \frac{1}{2} \text{sign}(n) \ln \left| \frac{1 + \sqrt{3(J-1)}}{1 - \sqrt{3(J-1)}} \right| \right]. \quad (8.72)$$

It is easy to see that the integrand in (8.70) tends to zero exponentially fast, as $R \rightarrow +\infty$, for $z = \pm R + iu \text{sign}(k)$, $|u| \leq \frac{\pi}{2} - \delta$ (any $0 < \delta < 1$). Therefore we can shift the integration over \mathbb{R} to the line $z = t + i(\frac{\pi}{2} - \delta) \text{sign}(k)$, $t \in \mathbb{R}$, which yields,

$$\begin{aligned} \widehat{F}(k) = & w \sqrt{\frac{2\kappa}{J-1}} e^{2\pi i k \frac{mL}{\sqrt{3(J-1)}}} e^{-|k|w \sqrt{\frac{2\kappa}{J-1}} \pi(\pi-2\delta)} \\ & \times \int_{-\infty}^{+\infty} dt e^{2\pi i t w \sqrt{\frac{2\kappa}{J-1}}} \widetilde{G}(t + i(\frac{\pi}{2} - \delta) \text{sign}(k)) \end{aligned} \quad (8.73)$$

From expression (8.71) it is possible to show the estimate (for $|J-1| \ll 1$ and $0 < \delta \ll 1$ and C a numerical constant) $|\widetilde{G}(t + i \text{sign} k (\frac{\pi}{2} - \delta))| \leq C(J-1)^2 e^{-2|t|\delta^{-4}}$. Since δ can be taken as small as we wish, this allows to conclude that

$$\widehat{F}(k) = C_{\delta, J, w}(k) \delta^{-4} (J-1)^{3/2} w \sqrt{2\kappa} e^{2\pi i k \frac{mL}{\sqrt{3(J-1)}}} e^{-|k|w \sqrt{\frac{2\kappa}{J-1}} \pi(\pi-2\delta)} \quad (8.74)$$

where $C_{\delta, J}(k) < C$ for all k . This result implies that the Van der Waals curve has oscillations, around the Maxwell plateau, of period $\frac{\sqrt{3(J-1)}}{L}$ and amplitude $e^{-w \sqrt{\frac{2\kappa}{J-1}} \pi^2}$. By replacing the first terms of the expansion of Φ when $J \rightarrow 1_+$ we can obtain a completely explicit approximation for $\widehat{F}(k)$. Thanks to the exact formula

$$\int_{-\infty}^{+\infty} dz e^{ikz} (1 - \tanh^4 z) = \frac{\pi}{6} \frac{k(8-k^2)}{\sinh \frac{k\pi}{2}} \quad (8.75)$$

and using $\Phi(m) \approx -\frac{J-1}{2} m^2 + \frac{1}{12} m^4$ we get

$$\widetilde{G}(z) \approx \frac{3}{4} (J-1)^2 (1 - \tanh^4 z), \quad (8.76)$$

we find asymptotically for w large, $J \rightarrow 1_+$ and any fixed k

$$\widehat{F}(k) \approx 4(J-1) \kappa w^2 \pi^2 k (1 - k^2 \frac{\pi^2 w^2 \kappa}{J-1}) \sinh^{-1} \left(k \pi^2 w \sqrt{\frac{2\kappa}{J-1}} \right). \quad (8.77)$$

This is formula (8.51) of the main text. For the zero mode $k = 0$ we get

$$\widehat{F}(0) \approx 4(J-1)^{3/2} w \sqrt{\frac{\kappa}{2}} \quad (8.78)$$

and for the other ones $k \in \mathbb{Z}^*$

$$\widehat{F}(k) \approx -8(\pi w)^4 \kappa^2 |k|^3 e^{-|k| \pi^2 w \sqrt{\frac{2\kappa}{J-1}}} e^{2\pi i k \frac{mL}{\sqrt{3(J-1)}}} \quad (8.79)$$

Finally, for the reader's convenience, we point out that to check (8.75) one can use $\frac{1}{6}(\tanh z)''' + \frac{8}{6}(\tanh z)' = 1 - \tanh^4 z$ and $\int_{-\infty}^{+\infty} dz e^{ikz} \tanh z = i\pi(\sinh \frac{\pi k}{2})^{-1}$ [88].

Coupled Constraint Satisfaction Problems

9

9.1 Problem Formulation

As explained in Chapter 7, the K -SAT ensemble has attracted much attention in computer science, mathematics and statistical physics during the recent two decades. By now, we know that random K -SAT formulas enjoy a number of intriguing mathematical properties. Many properties have been discovered and there are many others yet to be found or made rigorous. In particular, the highly intuitive but non-rigorous tools from statistical physics have led to the discovery of a much more refined framework for studying CSPs. Such a framework predicts a series of important aspects of the solutions space of the random K -SAT formulas and is able to locate the corresponding phase transitions very precisely (for a brief review see Section 7.3.2).

The coupled K -SAT ensemble was introduced in Section 7.3.3. In this chapter and the next, we investigate various properties and aspects of the coupled K -SAT ensemble. In particular, we show how various versions of the threshold saturation phenomenon become apparent as a result of the additional spatial structure. In this chapter¹, we focus on the location of different phase transitions of the coupled ensemble and their relation to the phase transitions of the individual K -SAT ensemble. The main tools that we use are the interpolation method and the (energetic and entropic) cavity method. We adopt in this chapter the terminology of statistical physics.

The outline is as follows. In Section 9.2 we review the suitable notation and terminology required for this chapter. We also introduce the Q -COL and K -XORSAT ensembles and their coupled versions. In Section 9.3, we use combinatorial interpolation methods [85] to relate several characteristics of the coupled ensemble to the individual ensemble. In Section 9.4, we apply the zero

¹The material of this chapter is based on [68].

temperature cavity method (the survey propagation formalism) to the coupled ensemble. Finally, the entropic cavity method on the coupled ensemble will be the subject of Section 9.6.

We conclude this section by noticing that similar results hold for other CSP models (e.g. Q -COL), however, we do not address them in detail for the sake of brevity and refer the interested reader to [68].

9.2 General Setting

We define a general class of CSP that form the *individual ensemble*. Then we couple these, to form one-dimensional chains called *spatially coupled-CSP ensembles*.

9.2.1 Individual CSP Ensemble $[N, K, \alpha]$.

First, we specify an ensemble (N, K, α) of random bipartite graphs. Let $G = (V \cup C, E)$ with *variable* nodes $i \in V$, *constraint* nodes $c \in C$ and edges $\langle c, i \rangle$ connecting sets C and V . We have $|V| = N$, $|C| = M$, where $M = \lfloor \alpha N \rfloor$ (the integer part of αM) and α is a fixed number called the constraint density. We call N the size of the graph which is to be thought as large, $N \rightarrow +\infty$. All constraints c have degree K , and each edge $\langle c, i \rangle$ emanating from c is independently connected uniformly at random (u.a.r.) to a node in $i \in V$. As $N \rightarrow +\infty$, the degrees of the variable nodes tend to independent identically distributed (i.i.d.) with distribution $\text{Poisson}(\alpha K)$.

We denote by ∂i the set of constraints connected to variable node i and by ∂c the set of variable nodes connected to a constraint c .

For each graph G of the ensemble $[N, K, \alpha]$ we define a Hamiltonian (or cost function). To the variable nodes $i \in V$ we attach variables $x_i \in \mathcal{X}$ taking values in a discrete alphabet \mathcal{X} . To each constraint $c \in C$ we associate a function $\psi_c(x_{\partial c})$ which depends only on the variables $x_{\partial c} = (x_i)_{i \in \partial c}$ connected to c . For constraint satisfaction problems $\psi_c(x_{\partial c}) \in \{0, 1\}$; we say that the constraint is *satisfied* if $\psi_c(x_{\partial c}) = 1$ and *not satisfied* if $\psi_c(x_{\partial c}) = 0$. The total Hamiltonian is

$$\mathcal{H}(\underline{x}) = \sum_{c \in C} (1 - \psi_c(x_{\partial c})). \quad (9.1)$$

For many problems of interest the functions ψ_c are themselves random. This will be made precise in each specific example; the only important condition is that the functions ψ_c are i.i.d. for all $c \in C$. The ground state energy is $\min_{\underline{x}} \mathcal{H}(\underline{x})$, the minimum possible number of unsatisfiable constraints. Our main interest is in the average ground state energy per node

$$e_N(\alpha) = \frac{1}{N} \mathbb{E}[\min_{\underline{x}} \mathcal{H}(\underline{x})] \quad (9.2)$$

where the expectation is taken over the $[N, K, \alpha]$ ensemble and possibly over the randomness of ψ_c .

9.2.2 Coupled-CSP Ensemble $[N, K, \alpha, w, L]$.

This ensemble represents a chain of coupled underlying ensembles. Figure 9.1 is a visual aid but gives only a partial view. We align positions $z \in \mathbb{Z}$. On each position $z \in \mathbb{Z}$, we lay down N variable nodes labeled $(i, z) \in V_z, i = 1, \dots, N$. We also lay down $M = \lfloor \alpha N \rfloor$ check nodes labeled $(c, z) \in C_z, c = 1, \dots, M$. When the node labels are used as subscripts, say as in $a_{(i,z)}$ or $a_{(c,z)}$, we will simplify the notation to a_{iz} or a_{cz} . Let us now specify how the set of edges, E , is chosen. Each constraint (c, z) has degree K , in other words K edges emanate from it. Each of these K edges is connected to variable nodes as follows: we first pick a position $z+k$ with k uniformly random in the window $\{0, \dots, w-1\}$, then we pick a node $(i, z+k)$ u.a.r. in V_{z+k} , and finally we connect (c, z) to $(i, z+k)$. The set of edges emanating from (i, z) can be decomposed as a union $\cup_{k=0}^{w-1} \{(c, z-k), (i, z)\} \mid c \in C_z\}$. Asymptotically as $N \rightarrow +\infty$, its cardinality is Poisson(αK); and the cardinalities of each set in the union are i.i.d. Poisson($\frac{\alpha K}{w}$).

Finally, we restrict the set of constraint nodes to $\cup_{z=0, \dots, L-w} C_z$ and delete edges emanating from constraints that do not belong to this set. Restrict the set of variable nodes to $\cup_{z=0, \dots, L-1} V_z$.

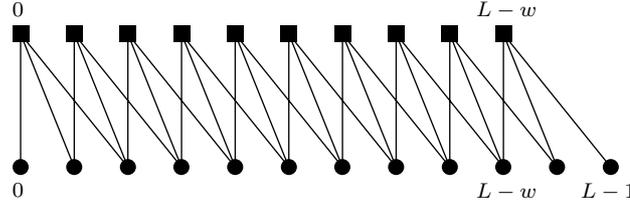


Figure 9.1: A representation of the geometry of the graphs with window size $w = 3$ along the “longitudinal chain direction” z . The “transverse direction” is viewed from the top. At each position there is a stack of N variable nodes (circles) and a stack M constraint nodes (squares). The depicted links between constraint and variable nodes represent stacks of edges.

As in subsection 9.2.1, we have a set of variables $x_{iz} \in \mathcal{X}$ and constraint functions $\psi_{cz}(x_{\partial cz})$ taking values in $\{0, 1\}$. To each coupled graph in the ensemble we associate the Hamiltonian depending on $\underline{x} = (x_{iz})$, with (i, z) being inside the set $\cup_{z=-\frac{L}{2}+1, \dots, \frac{L}{2}+w-1} V_z$,

$$\mathcal{H}_{\text{cou}}(\underline{x}) = \sum_{z=0}^{L-w} \sum_{c \in C_z} (1 - \psi_{cz}(x_{\partial cz})). \quad (9.3)$$

The minimum over \underline{x} is the ground state energy and its ensemble average per node is

$$e_{N,L,w}(\alpha) = \frac{1}{NL} \mathbb{E}[\min_{\underline{x}} \mathcal{H}_{\text{cou}}(\underline{x})], \quad (9.4)$$

where \mathbb{E} is over the $[N, K, \alpha, w, L]$ graph ensemble and on the randomness in ψ_{cz} .

Remark about the constraint density. In this paper we have adopted the constraint density of the underlying ensemble α as our control parameter. For a chain of coupled ensembles it represents the density of constraints in the bulk. More precisely, for a chain of length L the ratio of the total number of constraints to the total number of nodes is $\frac{M(L-w+1)}{NL}$ (see figure 9.1). This means that the average density of constraints is $\alpha_{\text{av}}(L, w) = \alpha \frac{L-w+1}{L} < \alpha$. This tends to α as $L \rightarrow +\infty$ so that in this limit the average density becomes insensitive to the boundary. In the present context, the spatial structure makes it more natural to take the bulk rather than the average density as a control parameter.

Remark about the boundary conditions. In the formulation above we have free boundary conditions. However, the average degree of the variable nodes close to the boundaries is reduced so that the CSP is *easier to solve close to the boundaries*. Variable nodes close to the right boundary $z = L - w, \dots, L - 1$ have degrees Poisson($\frac{\alpha K}{w}(L - 1 - z)$), and those close to the left boundary $z = 0, \dots, w - 1$ have degrees Poisson($\frac{\alpha K}{w}(z)$). It is sometimes convenient to imagine that the boundary nodes are connected to “satisfied extra constraint nodes”, and all have Poisson(αK) degree.

9.2.3 K -SAT, Q -COL and K -XORSAT

We define the main examples of constraint satisfaction problems that we analyze in this paper.

The K -SAT problem. The individual system is defined as follows. We take $x_i \in \{0, 1\}$ the Boolean alphabet. Set $n(x_i) \equiv \bar{x}_i$ for the negation operation, and define $n^d(x_i) \equiv x_i$ when $d = 0$ and $n^d(x_i) \equiv n(x_i) = \bar{x}_i$ when $d = 1$. Pick Bernoulli($\frac{1}{2}$) i.i.d. numbers $d_{\langle c, i \rangle}$ for each edge $\langle c, i \rangle \in E$. We say that an edge is *dashed* when $d_{\langle c, i \rangle} = 1$ and *full* when $d_{\langle c, i \rangle} = 0$. With this convention, a variable in a constraint is negated when it is connected to a dashed edge, and is not negated when it is connected to a full edge. We set

$$\psi_c(x_{\partial c}) = \mathbb{1}(\forall_{i \in \partial c} (n^{d_{\langle c, i \rangle}}(x_i)) = 1). \quad (9.5)$$

These definitions are extended to the coupled system in an obvious way

$$\psi_{cz}(x_{\partial(cz)}) = \mathbb{1}(\forall_{iu \in \partial(cz)} (n^{d_{\langle cz, iu \rangle}}(x_{iu})) = 1), \quad (9.6)$$

where the important point is that $d_{\langle cz, iu \rangle}$ are i.i.d. Bernoulli($\frac{1}{2}$) for all edges. The ground state energy counts the minimum possible number of unsatisfiable constraints. The instance is satisfiable iff the ground state energy is equal to zero.

The Q -COL problem. For the individual ensemble, we take $x_i \in \mathcal{X} = \{0, \dots, Q-1\}$ the Q -ary color alphabet, $K = 2$ for the constraint node degrees,

and

$$\psi_c(x_{\partial c}) = \mathbb{1}(x_i \neq x_j \text{ for } \{i, j\} = \partial c). \quad (9.7)$$

Since the constraints have degree 2 one can replace them by edges connecting directly i and j for $i, j \in \partial c$. The induced graph is, in the large size limit, equivalent to the Erdos-Rényi random graph $G(N, \frac{2c}{N})$. The constraint (9.7) forbids two neighboring nodes to have the same color.

These definitions are easily extended to the coupled system. The induced graph (obtained by replacing constraints by edges) is now a coupled chain of Erdos-Rényi graphs. In place of (9.7) we take $x_{iz} \in \mathcal{X} = \{0, \dots, Q - 1\}$ and

$$\psi_{cz}(x_{\partial(cz)}) = \mathbb{1}(x_{iu} \neq x_{jv} \text{ for } \{(i, u), (j, v)\} = \partial(c, z)). \quad (9.8)$$

Given an instance of the induced graph, the ground state energy counts the minimum possible number of edges with vertices of the same color. The graph is colorable iff this number is zero.

The K-XORSAT problem. We briefly give relevant definitions that will be used later in the thesis. For the individual system $x_i \in \{0, 1\}$ and $\psi_c(x_{\partial c}) = \mathbb{1}(\oplus_{i \in \partial c} x_i = b_c)$ with b_c being i.i.d. Bernoulli($\frac{1}{2}$). Similarly for the coupled system $\psi_{cz}(x_{\partial(cz)}) = \mathbb{1}(\oplus_{iu \in \partial(cz)} x_{iu} = b_{cz})$ with b_{cz} being i.i.d. Bernoulli($\frac{1}{2}$).

9.3 Interpolation Arguments: From the Individual Ensemble to the Coupled Ensemble and Vice Versa

For the purpose of analysis, it is useful to also consider an ensemble of coupled graphs with *periodic* boundary conditions. This ensemble is simply obtained from the $[N, K, \alpha, w, L]$ ensemble by identifying the variable nodes (i, z) at positions $z = L - w + k$ with nodes (i, z) at positions $z = k$ for each $k = 1, \dots, w - 1$. The formal expression of the Hamiltonian $\mathcal{H}_{\text{cou}}^{\text{per}}(\underline{x})$ is the same as in (9.3). Quantities pertaining to this ensemble will be denoted by a superscript "per".

Theorem 9.1 (Comparison of open and periodic chains). *For the general coupled-CSP $[N, K, \alpha, w, L]$ ensembles we have*

$$e_{N,L,w}^{\text{per}}(\alpha) - \frac{\alpha w}{L} \leq e_{N,L,w}(\alpha) \leq e_{N,L,w}^{\text{per}}(\alpha). \quad (9.9)$$

This theorem has an easy proof given in Section 9.8.1.

The next theorem does not have a trivial proof and is stated here for the special cases of K -SAT. We note here that the same result is true for the problems of Q -COL and K -XORSAT.

Theorem 9.2 (Thermodynamic limit). *For the K -SAT model the two limits $\lim_{N \rightarrow +\infty} e_N(\alpha)$ and $\lim_{N \rightarrow +\infty} e_{N,L,w}^{\text{per}}(\alpha)$ exist, are continuous, and non-decreasing in α . Moreover they are equal,*

$$\lim_{N \rightarrow +\infty} e_{N,L,w}^{\text{per}}(\alpha) = \lim_{N \rightarrow +\infty} e_N(\alpha). \quad (9.10)$$

Standard methods of statistical mechanics [109] do not allow to prove the existence of the limits because the underlying graphs have expansion properties. When the system is cut in two parts the number of edges in the cut is of the same order as the size of the two parts and is not just a “surface” term. Therefore sub-additivity of the free and ground state energies become non-trivial. However, interpolation methods allow to deal with this issue. The existence of the limit for $\lim_{N \rightarrow +\infty} e_N(\alpha)$, as well as the fact that the function is continuous and non-decreasing, is proved for a range of models including the present ones in [107], [85], and it is easy to see that the same sort of proof works for the periodic chain. This proof will not be repeated. In section 9.8.1 we provide the proof for the *equality* of the two limits. This is again based on two interpolations which provide upper and lower bounds. Note that concentration of the ground state and free energies is also implied by standard arguments not discussed here².

We are interested in the thermodynamic limit

$$\lim_{\text{therm}} \equiv \lim_{L \rightarrow +\infty} \lim_{N \rightarrow +\infty}$$

for the open chain, which captures the regime of a long one-dimensional coupled-CSP. From theorems 9.1 and 9.2 we deduce that

$$\lim_{\text{therm}} e_{N,L,w}(\alpha) = \lim_{\text{therm}} e_{N,L,w}^{\text{per}}(\alpha) = \lim_{N \rightarrow +\infty} e_N(\alpha). \quad (9.11)$$

Let pause for a moment and have a look at the satisfiability threshold in (7.20). Since the energy functions are non-decreasing we can redefine the satisfiability threshold as a natural “static phase transition” threshold as follows.

Definition 9.1 (Satisfiability threshold or the static phase transition threshold). *We define*

$$\alpha_s = \sup\{\alpha \mid \lim_{N \rightarrow +\infty} e_N(\alpha) = 0\}, \quad (9.12)$$

and

$$\alpha_{s,L,w} = \sup\{\alpha \mid \lim_{N \rightarrow +\infty} e_{N,L,w}(\alpha) = 0\}, \quad (9.13)$$

Theorem 9.3. *We have*

$$\alpha_s = \lim_{L \rightarrow +\infty} \alpha_{s,L,w}. \quad (9.14)$$

Proof. Because of (9.11) we have

$$\begin{aligned} \alpha_s &= \sup\{\alpha \mid \lim_{N \rightarrow +\infty} e_N(\alpha) = 0\} \\ &= \sup\{\alpha \mid \lim_{\text{therm}} e_{N,L,w}^{\text{per}}(\alpha) = 0\} \\ &= \sup\{\alpha \mid \lim_{\text{therm}} e_{N,L,w}(\alpha) = 0\} \end{aligned} \quad (9.15)$$

²However concentration of the number of solutions in the SAT phase is more subtle see [108].

By using the right-hand side inequality in (9.9) and (9.10), we deduce that $\alpha_s \leq \liminf_{L \rightarrow +\infty} \alpha_{s,L,w}$. Also note that (9.15) implies $\alpha_s \geq \limsup_{L \rightarrow +\infty} \alpha_{s,L,w}$. Indeed if this was not true then one could find α_* and a sequence $L_k \uparrow +\infty$ such that $\alpha_s < \alpha_* < \alpha_{s,L_k,w}$ for k large enough; but then $\lim_{N \rightarrow +\infty} e_{N,L_k,w}(\alpha_*) = 0$ and thus $\lim_{k \rightarrow +\infty} \lim_{N \rightarrow +\infty} e_{N,L_k,w}(\alpha_*) = 0$ which, from (9.15), would mean $\alpha_* \leq \alpha_s$; a contradiction. \square

The definition of α_s implies that, for a given instance, when $\alpha < \alpha_s$ (resp. $\alpha > \alpha_s$) the number of unsatisfied constraints is $o(N)$ (resp. $O(N)$) with high probability. However it is not known how to automatically conclude that a fixed instance is SAT (resp. UNSAT) with high probability when $\alpha < \alpha_s$ (resp. $\alpha > \alpha_s$). For more details see Section 7.3.2.

Remark about finite temperatures. The theorems of this subsection have finite temperature analogs presented in appendix 9.8.2. As explained in section 9.6 these suggest that the condensation threshold obeys $\lim_{L \rightarrow +\infty} \alpha_{c,L,w} = \alpha_c$.

9.4 Zero Temperature Cavity Method and Survey Propagation Formalism

We briefly summarize the simplest form of the cavity method and survey propagation equations for the coupled-CSP. More details on the formalism are presented in appendix 9.8.3. When the graph instance is a *tree*, the minimization of (9.3) can be carried out exactly. This leads to an expression for $\min_{\underline{x}} \mathcal{H}_{\text{cou}}(\underline{x})$ in terms of energy-cost messages $E_{iu \rightarrow cz}(x_{iu})$ and $\hat{E}_{cz \rightarrow iu}(x_{iu})$ that satisfy the standard min-sum equations (see equ. (9.95) and (9.96)). These messages are normalized so that $\min_{x_{iu}} E_{iu \rightarrow cz}(x_{iu}) = \min_{x_{iu}} \hat{E}_{cz \rightarrow iu}(x_{iu}) = 0$ and they take values in $\{0, 1\}$. They may be interpreted as warning messages. Roughly speaking, nodes inform each other on the most favorable values that the variable x_{iu} should take in order to avoid energy costs. The ground state energy (on the tree) is given by the Bethe energy functional $\mathcal{E}[\{E_{iu \rightarrow cz}(\cdot), \hat{E}_{cz \rightarrow iu}(\cdot)\}]$ (see equ. (9.98)). For a *general graph instance* one considers the Bethe energy functional (9.98) as an “effective Hamiltonian” and studies the statistical mechanics of this effective system. The min-sum equations are the stationary point equations of this functional and the set of solutions $\{E_{iu \rightarrow cz}(\cdot), \hat{E}_{cz \rightarrow iu}(\cdot)\}$ characterize the *state* of the system.

It turns out that the min-sum equations may have exponentially many (in system size) solutions with infinitesimal Bethe energy per node as $N \rightarrow +\infty$. A solution $\{E_{iu \rightarrow cz}^{(p)}(\cdot), E_{cz \rightarrow iu}^{(p)}(\cdot)\}$ with infinitesimal Bethe energy defines a *pure Bethe state*³ denoted by the superscript (p) . We define the average *zero-energy complexity* as

$$\Sigma_{L,w}(\alpha) = \lim_{\epsilon \rightarrow 0} \lim_{N \rightarrow +\infty} \frac{1}{NL} \mathbb{E}[\ln(\text{number of states } p \text{ with } \frac{\mathcal{E}^{(p)}}{N} = \epsilon)]. \quad (9.16)$$

³We adopt this terminology to make a distinction with the mathematically precise notion of *pure state* for usual Ising models [109].

This quantity counts the number of pure Bethe states. The typical behavior of the complexity as a function of α is as follows. Below an *SP threshold* it vanishes, then jumps to a positive value and decreases until it becomes zero at the static phase transition threshold (and formally negative above). It therefore allows to compute

$$\alpha_{\text{SP},L,w} = \inf\{\alpha \mid \Sigma_{L,w}(\alpha) > 0\}, \quad (9.17)$$

$$\alpha_{s,L,w} = \sup\{\alpha \mid \Sigma_{L,w}(\alpha) > 0\}. \quad (9.18)$$

One expects on heuristic grounds, and it has been checked numerically for various models, that the static phase transition thresholds defined according to the energy (9.12) and complexity (9.18) coincide.

The complexity is the Boltzmann entropy (on the zero energy shell) of the effective statistical mechanical problem with Hamiltonian $\mathcal{E}[\{E_{iu \rightarrow cz}(\cdot), \hat{E}_{cz \rightarrow iu}(\cdot)\}]$. It turns out that this can be computed, thanks to an effective partition function on the same sparse graph instance, again within a message passing formalism. In this context messages are called *surveys*. They count the fraction of pure Bethe states with given warning messages. Surveys $Q_{iu \rightarrow cz}(E_{iu \rightarrow cz}(\cdot))$ and $\hat{Q}_{cz \rightarrow iu}(\hat{E}_{cz \rightarrow iu}(\cdot))$ are exchanged between variable and constraint nodes according to survey propagation equations (see (9.103) and (9.104)). The average complexity (9.16) can be computed by a Bethe type formula for the entropy of the effective model.

The survey propagation equations (9.103), (9.104) allow to compute the distribution over pure Bethe states, of the vectors $(\hat{E}_{cz \rightarrow iu}(x_{iu}), x_{iu} \in \mathcal{X})$. These are $|\mathcal{X}|$ -component vectors with components in $\{0, 1\}$. Thus the surveys are supported on an alphabet of size at most $2^{|\mathcal{X}|}$. Often the effective size of the alphabet is smaller (it is $|\mathcal{X}| + 1$ in the specific problems considered here) because the warning propagation equations (9.95), (9.96) restrict the possible values of $(\hat{E}_{cz \rightarrow iu}(x_{iu}), x_{iu} \in \mathcal{X})$. This simplification is used for each model separately in the next sections.

Let us summarize the main observations that follow from the detailed analysis in Section 9.5. As $L \rightarrow +\infty$, we find that the complexity curves $\Sigma_{L,w}(\alpha)$ supported on the interval $[\alpha_{\text{SP},L,w}, \alpha_{s,L,w}]$ converge to a limiting curve $\Sigma_w(\alpha)$ supported on the limiting interval $[\alpha_{\text{SP},w}, \alpha_s]$. Moreover, on this later interval, $\Sigma_w(\alpha)$ coincides with the complexity $\Sigma(\alpha)$ of the individual system ($L = w = 1$). This is illustrated on Figure 9.2. We observe that $\alpha_{s,L,w}$ tends to α_s from above. Also for moderate L one generally has $\alpha_{\text{SP},L,w} > \alpha_s$, but this inequality is reversed for L large enough, and $\lim_{L \rightarrow +\infty} \alpha_{\text{SP},L,w} = \alpha_{\text{SP},w} < \alpha_s$.

We observe the threshold saturation, namely $\lim_{w \rightarrow +\infty} \alpha_{\text{SP},w} \uparrow \alpha_s$. In fact we expect (from [50]) that the gap $|\alpha_{\text{SP},w} - \alpha_s|$ is exponentially small in w (K fixed) but this is hard to assess numerically. One also observes that for w fixed the gap increases with increasing K .

We point out that the complexity of the chain with periodic boundary conditions converges to that of the individual system in the infinite length limit. In other words there is no threshold saturation as long as the boundary conditions are periodic. This is easily understood by realizing that the survey

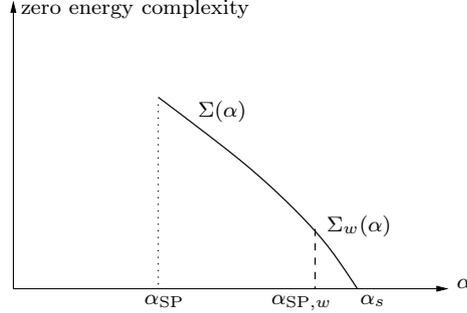


Figure 9.2: Complexity of the individual ensemble $\Sigma(\alpha)$ (i.e. $L = w = 1$) and limiting complexity $\Sigma_w(\alpha)$ of the coupled ensemble for $L \rightarrow +\infty$. We have $\alpha_{SP,w} \rightarrow \alpha_s$ as $w \rightarrow +\infty$.

propagation equations are purely local and have a translation invariant solution when the boundary conditions are periodic.

Finally, let us mention that we observe similar features for the *entropic complexity* curve. In this case α_d plays the role of α_{SP} and α_c that of α_s . We have $\alpha_{d,w} \rightarrow \alpha_c$. In particular, $\lim_{L \rightarrow +\infty} \alpha_{c,L,w} = \alpha_c$ and $\lim_{w \rightarrow +\infty} \lim_{L \rightarrow +\infty} \alpha_{d,L,w} = \alpha_c$ (see Section 9.6).

9.5 Coupled K-SAT Problem

9.5.1 Numerical Implementation

We begin with a convenient parametrization of the messages (see e.g [72]). Since $\mathcal{X} = \{0, 1\}$, the warning (energy costs) messages are two-component vectors $(E_{iu \rightarrow cz}(0), E_{iu \rightarrow cz}(1))$ and $(\bar{E}_{cz \rightarrow iu}(0), \bar{E}_{cz \rightarrow iu}(1))$ which take three possible values $(0, 1)$, $(1, 0)$ and $(0, 0)$. Warning $(0, 1)$ means that x_{iu} should take the value 0, warning $(1, 0)$ means that x_{iu} should take value 1, and warning $(0, 0)$ means that x_{iu} is free to take any value. Messages from variables to constraints can be conveniently parametrized as follows,

$$Q_{iu \rightarrow cz}^S \equiv \begin{cases} Q_{iu \rightarrow cz}(0, 1) & \text{if } x_{iu} \text{ is negated in } cz, \\ Q_{iu \rightarrow cz}(1, 0) & \text{if } x_{iu} \text{ is not negated in } cz. \end{cases}$$

This is the fraction of pure states for which the variable is forced to satisfy the constraint. Similarly,

$$Q_{iu \rightarrow cz}^U \equiv \begin{cases} Q_{iu \rightarrow cz}(0, 1) & \text{if } x_{iu} \text{ is not negated in } cz, \\ Q_{iu \rightarrow cz}(1, 0) & \text{if } x_{iu} \text{ is negated in } cz. \end{cases}$$

This is the fraction of pure states for which the variable is forced to unsatisfy the constraint. Note that $Q_{iu \rightarrow cz}(0, 0) = 1 - Q_{iu \rightarrow cz}^S - Q_{iu \rightarrow cz}^U$. Let us now

parametrize the messages from constraints to variables. If variable x_{iu} enters unnegated in constraint cz , then certainly constraint cz does not force it to take the value 0. Thus $\hat{Q}_{cz \rightarrow iu}(0, 1) = 0$, and the message can be parametrized by the single number $\hat{Q}_{cz \rightarrow iu}(1, 0)$. On the other hand, if variable x_{iu} enters negated in constraint cz , then certainly constraint cz does not force it to take the value 1. Thus $\hat{Q}_{cz \rightarrow iu}(1, 0) = 0$, and again the message can be parametrized by the single number $\hat{Q}_{cz \rightarrow iu}(0, 1)$. We set

$$\hat{Q}_{cz \rightarrow iu} \equiv \begin{cases} \hat{Q}_{cz \rightarrow iu}(0, 1) & \text{if } x_{iu} \text{ is negated in } cz, \\ \hat{Q}_{cz \rightarrow iu}(1, 0) & \text{if } x_{iu} \text{ is not negated in } cz. \end{cases}$$

Message $\hat{Q}_{cz \rightarrow iu}$ is the fraction of pure states for which cz warns iu to satisfy it. The survey propagation equations (9.103), (9.104) then become (recall $d_{\langle bv, iu \rangle} = 1$ (resp. 0) for a dashed (resp. full) edge $\langle bv, iu \rangle$),

$$\hat{Q}_{cz \rightarrow iu} = \prod_{jv \in \partial(cz) \setminus iu} Q_{jv \rightarrow cz}^U, \quad (9.19)$$

and

$$Q_{iu \rightarrow cz}^S \cong \left\{ \prod_{bv \in \partial(iu) \setminus cz}^{d_{\langle bv, iu \rangle} \neq d_{\langle iu, cz \rangle}} (1 - \hat{Q}_{bv \rightarrow iu}) \right\} \left\{ 1 - \prod_{bv \in \partial(iu) \setminus cz}^{d_{\langle bv, iu \rangle} = d_{\langle iu, cz \rangle}} (1 - \hat{Q}_{bv \rightarrow iu}) \right\}, \quad (9.20)$$

$$Q_{iu \rightarrow cz}^U \cong \left\{ \prod_{bv \in \partial(iu) \setminus cz}^{d_{\langle bv, iu \rangle} = d_{\langle iu, cz \rangle}} (1 - \hat{Q}_{bv \rightarrow iu}) \right\} \left\{ 1 - \prod_{bv \in \partial(iu) \setminus cz}^{d_{\langle bv, iu \rangle} \neq d_{\langle iu, cz \rangle}} (1 - \hat{Q}_{bv \rightarrow iu}) \right\}, \quad (9.21)$$

where \cong means that the r.h.s has to be normalized to one. Define

$$Q_{iu \rightarrow cz}^+ = \prod_{bv \in \partial(iu) \setminus cz}^{d_{\langle bv, iu \rangle} = d_{\langle iu, cz \rangle}} (1 - \hat{Q}_{bv \rightarrow iu}), \quad (9.22)$$

$$Q_{iu \rightarrow cz}^- = \prod_{bv \in \partial(iu) \setminus cz}^{d_{\langle bv, iu \rangle} \neq d_{\langle iu, cz \rangle}} (1 - \hat{Q}_{bv \rightarrow iu}). \quad (9.23)$$

Then using (9.19) and the normalized form of (9.21)

$$\hat{Q}_{cz \rightarrow iu} = \prod_{jv \in \partial(cz) \setminus iu} \frac{Q_{jv \rightarrow cz}^+ (1 - Q_{jv \rightarrow cz}^-)}{Q_{jv \rightarrow cz}^+ + Q_{jv \rightarrow cz}^- - Q_{jv \rightarrow cz}^+ Q_{jv \rightarrow cz}^-}. \quad (9.24)$$

We will work with the set of SP equations (9.22), (9.23), (9.24). The complexity becomes

$$\Sigma_{L,w}(\alpha) = \frac{1}{NL} \mathbb{E} \left[\sum_{cz} \Sigma_{cz} + \sum_{iz} \Sigma_{iz} - \sum_{\langle cz, iu \rangle} \Sigma_{cz, iu} \right], \quad (9.25)$$

with

$$\Sigma_{cz} = \ln \left\{ \prod_{iu \in \partial(cz)} (Q_{iu \rightarrow cz}^+ + Q_{iu \rightarrow cz}^- - Q_{iu \rightarrow cz}^+ Q_{iu \rightarrow cz}^-) - \prod_{iu \in \partial(cz)} Q_{iu \rightarrow cz}^+ (1 - Q_{iu \rightarrow cz}^-) \right\}, \quad (9.26)$$

$$\Sigma_{iz} = \ln \left\{ \prod_{bv \in \partial(iz)}^{d(bv, iz)=1} (1 - \hat{Q}_{bv \rightarrow iz}) + \prod_{bv \in \partial(iz)}^{d(bv, iz)=0} (1 - \hat{Q}_{bv \rightarrow iz}) - \prod_{bv \in \partial(iz)} (1 - \hat{Q}_{bv \rightarrow iz}) \right\}, \quad (9.27)$$

$$\Sigma_{cz, iu} = \ln \left\{ (Q_{iu \rightarrow cz}^+ + Q_{iu \rightarrow cz}^- - Q_{iu \rightarrow cz}^+ Q_{iu \rightarrow cz}^-) - Q_{iu \rightarrow cz}^+ (1 - Q_{iu \rightarrow cz}^-) \hat{Q}_{cz \rightarrow iu} \right\} \quad (9.28)$$

The set of SP equations (9.22), (9.23), (9.24) is solved under the following assumptions. We treat the set of messages emanating from a constraint at position z , namely $\hat{Q}_{cz \rightarrow iu}$ for $u = z, \dots, z + w - 1$, as i.i.d. copies of a r.v. \hat{Q}_z depending only on the position z . Similarly we treat the messages emanating from a variable node at position u , namely $Q_{iu \rightarrow cz}^\pm$ for $z = u - w + 1, \dots, u$, as i.i.d. copies of a r.v. Q_u^\pm . Now, fix a position z and pick p, q two independent Poisson($\frac{\alpha K}{2}$) integers. Pick k_1, \dots, k_{p+q} independently uniformly in $\{0, \dots, w - 1\}$. Similarly, pick l_1, \dots, l_{K-1} independently uniformly in $\{0, \dots, w - 1\}$. Under our assumptions the SP equations become⁴

$$Q_z^+ = \prod_{i=1}^p (1 - \hat{Q}_{z-k_i}^{(i)}), \quad (9.29)$$

$$Q_z^- = \prod_{i=p+1}^{p+q} (1 - \hat{Q}_{z-k_i}^{(i)}), \quad (9.30)$$

and

$$\hat{Q}_z = \prod_{i=1}^{K-1} \frac{Q_{z+l_i}^{+(i)} (1 - Q_{z+l_i}^{-(i)})}{Q_{z+l_i}^{+(i)} + Q_{z+l_i}^{-(i)} - Q_{z+l_i}^{+(i)} Q_{z+l_i}^{-(i)}}. \quad (9.31)$$

The boundary conditions can be taken into account by setting $\hat{Q}_z = 0$ for $z \leq -\frac{L}{2}$, $z > \frac{L}{2}$. These equations are solved by the standard method of population dynamics. It is then possible to compute the average complexity from

$$\Sigma_{L,w}(\alpha) = \frac{1}{L} \sum_{z=-\frac{L}{2}+1}^{\frac{L}{2}} (\alpha \mathbb{E}[\Sigma_z^{\text{cons}}] + \mathbb{E}[\Sigma_z^{\text{var}}] - \alpha K \mathbb{E}[\Sigma_z^{\text{edge}}]), \quad (9.32)$$

where

$$\Sigma_z^{\text{cons}} = \ln \left\{ \prod_{i=1}^K (Q_{z+l_i}^{+(i)} + Q_{z+l_i}^{-(i)} - Q_{z+l_i}^{+(i)} Q_{z+l_i}^{-(i)}) - \prod_{i=1}^K Q_{z+l_i}^{+(i)} (1 - Q_{z+l_i}^{-(i)}) \right\}, \quad (9.33)$$

⁴In (9.29), (9.30), (9.31) equalities mean that the r.v. have the same distribution.

$$\Sigma_z^{\text{var}} = \ln \left\{ \prod_i^p (1 - \hat{Q}_{z-k_i}^{(i)}) + \prod_{i=p+1}^{p+q} (1 - \hat{Q}_{z-k_i}^{(i)}) - \prod_{i=1}^{p+q} (1 - \hat{Q}_{z-k_i}^{(i)}) \right\}, \quad (9.34)$$

$$\Sigma_z^{\text{edge}} = \ln \left\{ (Q_{z+k}^+ + Q_{z+k}^- - Q_{z+k}^+ Q_{z+k}^-) - Q_{z+k}^+ (1 - Q_{z+k}^-) \hat{Q}_z \right\}. \quad (9.35)$$

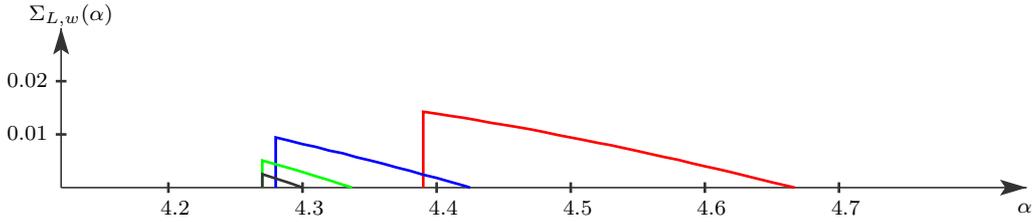


Figure 9.3: Average complexity versus α for the $[1000, 3, \alpha, 3, L]$ ensembles with $L = 10$ (rightmost curve), 20, 40, 80 (leftmost curve). Values of the corresponding thresholds are given in table 9.1.

Figure 9.5.1 shows the average complexity for the regime $N \gg L \gg w$, for $K = 3$ and $w = 3$. We find it is positive in an interval $[\alpha_{SP,L,w}, \alpha_{s,L,w}]$ whose size shrinks as L increases. The two end points of this interval are given in Table 9.1 (corresponding to Figure 9.5.1). Let us comment on the numerical findings.

First, we observe that $\alpha_{SP,L,w}$ approaches α_s as L increases. It is hard to compute more than three digits with population dynamics experiments but we expect that a small difference should remain between $\lim_{L \rightarrow \infty} \alpha_{SP,L,w}$ and α_s . This difference should decrease very fast as w grows, and in fact for $w = 3$ one does not see it in the first three digits. For the Curie-Weiss chain [50] this difference has been analytically calculated to be exponentially small⁵ in w . For the paradigmatic spatially coupled LDPC codes the difference appears only in the sixth decimal figure when state of the art density evolution numerics is used [2].

Second, we observe that $\alpha_{s,w,L}$ decreases as L increases. An extrapolation of the values suggests that as L grows larger (i.e., $L = 320, 640, \dots$) $\alpha_{s,w,L}$ should come closer to α_s . However these lengths become prohibitive for population dynamics. As discussed in Section 9.3 we expect on theoretical grounds that $\lim_{L \rightarrow +\infty} \alpha_{s,w,L} = \alpha_s$ is true for all w .

For moderate values of L we have $\alpha_s < \alpha_{SP,L,w}$. However since $\alpha_{SP,L,w} < \alpha_{s,w,L}$ and $\lim_{L \rightarrow +\infty} \alpha_{s,w,L} = \alpha_s$, for L large enough and fixed w we necessarily have $\alpha_{SP,L,w} < \alpha_s$. This turns out to be difficult to observe within population dynamics experiments, but can be checked in the large K limit.

⁵The calculation involves a non-perturbative calculation of potential energy barriers in terms of a deformation parameter $\frac{1}{w}$ when going from a continuum to a discrete model.

individual	α_{SP}	α_s
$L = 1$	3.927	4.267
coupled	$\alpha_{SP,L,3}$	$\alpha_{s,L,3}$
$L = 10$	4.386	4.663
$L = 20$	4.274	4.425
$L = 40$	4.269	4.335
$L = 80$	4.268	4.301
$L = 160$	4.267	4.284

Table 9.1: SP and static phase transition thresholds of the $[1000, 3, \alpha, 3, L]$ ensembles.

9.5.2 Survey Propagation for Large K

For large K one can derive approximations of the survey propagation equations that lend themselves to more explicit analysis [120]. We will not attempt to control the error terms, but it is known for the individual system that the approximations are excellent already for $K \geq 5$. We can check numerically that this is also the case for the coupled-CSP.

Fixed point equations. Following [120], we introduce entropic random variables

$$\hat{q}_z = -\ln(1 - \hat{Q}_z), \quad q_z^\pm = -\ln Q_z^\pm. \quad (9.36)$$

From (9.29), (9.30) and (9.31) we obtain

$$q_z^+ = \sum_{i=1}^p \hat{q}_{z-k_i}^{(i)}, \quad q_z^- = \sum_{i=p+1}^{p+q} \hat{q}_{z-k_i}^{(i)}, \quad (9.37)$$

and

$$\hat{q}_z = -\ln \left\{ 1 - \prod_{i=1}^{K-1} \frac{e^{\hat{q}_{z+l_i}^{-(i)}} - 1}{e^{\hat{q}_{z+l_i}^{-(i)}} + e^{\hat{q}_{z+l_i}^{+(i)}} - 1} \right\}, \quad (9.38)$$

we set

$$\mathbb{E}[q_z^\pm] = x_z^\pm \quad \text{and} \quad \mathbb{E}[\hat{q}_z] = y_z, \quad (9.39)$$

for the averages over the graph ensemble. The number of i.i.d. random variables in (9.37) is a $\text{Poisson}(\frac{\alpha K}{2})$ integer. Therefore we assume that for large K the r.v. q_z^\pm are self-averaging. It is reasonable to expect that they can be replaced by their expectation in (9.38) and that hence \hat{q}_z is also self-averaging. This implies a closed set of equations for the expected values of messages,

$$\begin{cases} x_z^\pm \approx \frac{\alpha K}{2w} \sum_{k=0}^{w-1} y_{z-k}, \\ y_z \approx -\sum_{k_1, \dots, k_{K-1}=0}^{w-1} \frac{1}{w^{K-1}} \ln \left\{ 1 - \prod_{i=1}^{K-1} \frac{e^{x_{z+k_i}^-} - 1}{e^{x_{z+k_i}^-} + e^{x_{z+k_i}^+} - 1} \right\}. \end{cases} \quad (9.40)$$

We further approximate (9.40). A self-consistent check with the final solution shows that $x^\pm = O(K)$ and hence the product in the log is $O(2^{-K})$. Linearizing

the logarithm yields

$$y_z \approx \sum_{k_1, \dots, k_{K-1}=0}^{w-1} \frac{1}{w^{K-1}} \prod_{i=1}^{K-1} \frac{e^{x_{z+k_i}^-} - 1}{2e^{x_{z+k_i}^-} - 1} = \left\{ \frac{1}{w} \sum_{k=0}^{w-1} \frac{e^{x_{z+k}^-} - 1}{2e^{x_{z+k}^-} - 1} \right\}^{K-1}. \quad (9.41)$$

It is convenient to introduce the rescaled parameters

$$\hat{\alpha} = 2^{-K} \alpha, \quad \varphi_z = 2^{K-1} \hat{\alpha} K y_z. \quad (9.42)$$

From (9.36) we see φ_z is a measure of the average (over the graph ensemble) probability (over pure states) that constraints at position z send warning messages. From now on we write x_z instead of x_z^\pm . The fixed point equations become

$$\begin{cases} x_z \approx \frac{1}{w} \sum_{k=0}^{w-1} \varphi_{z-k}, \\ \varphi_z \approx \hat{\alpha} K \left\{ \frac{1}{w} \sum_{l=0}^{w-1} \frac{e^{x_{z+l} - 1}}{e^{x_{z+l} - \frac{1}{2}}} \right\}^{K-1}. \end{cases} \quad (9.43)$$

Hence, the profile $\{\varphi_z\}$ satisfies

$$\varphi_z \approx \hat{\alpha} K \left\{ \frac{1}{w} \sum_{k=0}^{w-1} \frac{e^{\frac{1}{w} \sum_{l=0}^{w-1} \varphi_{z-l+k} - 1}}{e^{\frac{1}{w} \sum_{l=0}^{w-1} \varphi_{z-l+k} - \frac{1}{2}}} \right\}^{K-1}. \quad (9.44)$$

These equations have to be supplemented with the boundary condition $\varphi_z = 0$ for $z \leq -\frac{L}{2}$ and $z > \frac{L}{2}$.

The average complexity. Let us now express the complexity in terms of the fixed point profile. Let us first compute the contributions of variable and constraint nodes, and of edges.

Contribution of variable nodes. From (9.34), (9.36) and (9.39)

$$\Sigma_z^{var} = \ln \left\{ e^{-\sum_{i=1}^p \hat{q}_{z-k_i}} + e^{-\sum_{i=p+1}^q \hat{q}_{z-k_i}} - e^{-\sum_{i=1}^{p+q} \hat{q}_{z-k_i}} \right\}. \quad (9.45)$$

For K large the sums in the exponentials concentrate on their averages, so that

$$\mathbb{E}[\Sigma_z^{var}] \approx \ln \left\{ 2e^{-\frac{\alpha K}{2w} \sum_{k=0}^{w-1} y_{z-k}} - e^{-\frac{\alpha K}{w} \sum_{k=0}^{w-1} y_{z-k}} \right\}. \quad (9.46)$$

Contribution of check nodes. From (9.33), (9.36) and (9.39)

$$\begin{aligned} \mathbb{E}[\Sigma_z^{cons}] &= \mathbb{E} \left[\ln \left\{ \prod_{i=1}^K (e^{-q_{z+l_i}^+} + e^{-q_{z+l_i}^-} - e^{-q_{z+l_i}^+ - q_{z+l_i}^-}) \right. \right. \\ &\quad \left. \left. - \prod_{i=1}^K e^{-q_{z+l_i}^+} (1 - e^{-q_{z+l_i}^-}) \right\} \right] \\ &\approx \sum_{l_1, \dots, l_K=0}^{w-1} \frac{1}{w^K} \ln \left\{ \prod_{i=1}^K (2e^{-x_{z+l_i}^-} - e^{-2x_{z+l_i}^-}) - \prod_{i=1}^K e^{-x_{z+l_i}^-} (1 - e^{-x_{z+l_i}^-}) \right\}. \end{aligned} \quad (9.47)$$

Factoring the first product out of the log we get

$$\mathbb{E}[\Sigma_z^{\text{cons}}] \approx \frac{K}{w} \sum_{i=0}^{w-1} \ln\{2e^{-x_{z+i}^-} - e^{-2x_{z+i}^-}\} + \sum_{l_1, \dots, l_K=0}^{w-1} \frac{1}{w^K} \ln\left\{1 - \prod_{i=1}^K \frac{1 - e^{-x_{z+l_i}^-}}{2 - e^{-x_{z+l_i}^-}}\right\}. \quad (9.48)$$

Since the ratio in the second log is $O(2^{-K})$ we can linearize and obtain

$$\mathbb{E}[\Sigma_z^{\text{cons}}] \approx \frac{K}{w} \sum_{l=0}^{w-1} \ln\{2e^{-x_{z+l}^-} - e^{-2x_{z+l}^-}\} - \left\{\frac{1}{w} \sum_{l=0}^{w-1} \frac{1 - e^{-x_{z+l}^-}}{2 - e^{-x_{z+l}^-}}\right\}^K. \quad (9.49)$$

Contribution of edges. Similarly from (9.35), (9.36), (9.39) we have

$$\begin{aligned} \mathbb{E}[\Sigma_z^{\text{edge}}] &= \frac{1}{w} \sum_{l=0}^{w-1} \mathbb{E} \left[\ln\{ (e^{-q_{z+l}^+} + e^{-q_{z+l}^-} - e^{-q_{z+l}^+ - q_{z+l}^-}) \right. \\ &\quad \left. - e^{-q_{z+l}^+} (1 - e^{-q_{z+l}^-}) (1 - e^{-q_z}) \} \right] \\ &\approx \frac{1}{w} \sum_{l=0}^{w-1} \ln\left\{ (2e^{-x_{z+l}^-} - e^{-2x_{z+l}^-}) - e^{-x_{z+l}^-} (1 - e^{-x_{z+l}^-}) (1 - e^{-y_z}) \right\}. \end{aligned} \quad (9.50)$$

Now, using (9.40) we can express the total average complexity (9.32) in terms of rescaled variables (9.42). We find

$$\Sigma_{w,L}(\hat{\alpha}) = \frac{1}{L} \sum_{z=-\frac{L}{2}+1}^{\frac{L}{2}} \sigma_{\hat{\alpha},w,L}(z), \quad (9.51)$$

with

$$\begin{aligned} \sigma_{\hat{\alpha},w,L}(z) &\approx \ln\{2e^{-\sum_{k=0}^{w-1} \varphi_{z-k}} - e^{-\frac{2}{w} \sum_{k=0}^{w-1} \varphi_{z-k}}\} - 2^K \hat{\alpha} \left\{ \frac{1}{w} \sum_{l=0}^{w-1} \frac{e^{x_{z+l}} - 1}{2e^{x_{z+l}} - 1} \right\}^K \\ &\quad - \frac{2^K \hat{\alpha} K}{w} \sum_{l=0}^{w-1} \ln\left\{ 1 - \frac{e^{x_{z+l}} - 1}{2e^{x_{z+l}} - 1} (1 - e^{-\frac{\varphi_z}{\hat{\alpha} K 2^{K-1}}}) \right\}. \end{aligned} \quad (9.52)$$

Within our approximations the third term can be simplified further because $1 - e^{-\frac{\varphi_z}{\hat{\alpha} K 2^{K-1}}} = O(2^{-K})$ and we may linearize the log. Thus the second line in (9.52) can be replaced by

$$2\varphi_z \frac{1}{w} \sum_{l=0}^{w-1} \left\{ \frac{e^{x_{z+l}} - 1}{2e^{x_{z+l}} - 1} \right\}. \quad (9.53)$$

The complexity (9.51) can be viewed as a functional of the profiles $\{x_z, \varphi_z\}$ with boundary condition $\varphi_z = 0$ for $z \leq -\frac{L}{2}$ and $z > \frac{L}{2}$. One can check that the stationary points of this functional are given by the fixed point equations (9.43).

9.5.3 Solutions for Large K

We use the notation $f \doteq g$ to mean that $\lim_{K \rightarrow +\infty} \frac{f}{g} = 1$. The large K results for the individual system [120] are recovered by setting $L = w = 1$, in which case the fixed point equations (9.44) reduces to

$$\varphi \approx \hat{\alpha} K \left\{ \frac{e^\varphi - 1}{e^\varphi - \frac{1}{2}} \right\}^{K-1}. \quad (9.54)$$

One may easily check that this is the stationary point equation for the complexity (9.51) as a function of φ (and α fixed),

$$\Sigma_{1,1}(\hat{\alpha}, \varphi) = \ln\{2e^{-\varphi} - e^{-2\varphi}\} - 2K\hat{\alpha} \left\{ \frac{e^\varphi - 1}{2e^\varphi - 1} \right\}^K + \varphi \left\{ \frac{e^\varphi - 1}{2e^\varphi - 1} \right\}. \quad (9.55)$$

Thus, fixed points of (9.54) are stationary points of (9.55): stable fixed points correspond to minima and unstable ones to maxima.

The curve $\hat{\alpha}(\varphi)$ is shown as the dotted curve in Figure 9.5. This function is convex and has a unique minimum at $\varphi_{\text{SP}} \doteq \ln(\frac{1}{2}K \ln K)$ and $\hat{\alpha}(\varphi_{\text{SP}}) \equiv \hat{\alpha}_{\text{SP}} \doteq \frac{\ln K}{K}$. Near this minimum we have $\hat{\alpha}(\varphi) \approx (\frac{\varphi - \varphi_{\text{SP}}}{\gamma_{\text{SP}}})^2$, $\gamma_{\text{SP}} \doteq \frac{4}{3} \frac{K}{\ln K}$. For $\varphi \gg \varphi_{\text{SP}}$ we have $\hat{\alpha}(\varphi) = \frac{1}{K}(\varphi - \varphi_{\text{SP}})$ and for $0 < \varphi \ll \varphi_{\text{SP}}$ we have $\hat{\alpha}(\varphi) = \frac{1}{\varphi}$. Therefore the trivial fixed point $\varphi = 0$ is unique for $\hat{\alpha} < \hat{\alpha}_{\text{SP}}$, and there are two extra non-trivial fixed points for $\hat{\alpha} > \hat{\alpha}_{\text{SP}}$. Only one of them is stable and forms the branch $\varphi_{\text{mst}} \approx K\hat{\alpha} + \varphi_{\text{SP}}$ for $\varphi \gg \varphi_{\text{SP}}$.

For $\hat{\alpha} < \hat{\alpha}_{\text{SP}}$, the function (9.55) has a unique minimum at $\varphi = 0$. For $\hat{\alpha} > \hat{\alpha}_{\text{SP}}$ a second metastable minimum appears at $\varphi_{\text{mst}} \approx K\hat{\alpha} + \varphi_{\text{SP}}$. At this minimum we find $\Sigma_{1,1}(\hat{\alpha}, \varphi_{\text{mst}}) \doteq \ln 2 - \hat{\alpha}$ which counts the number of clusters as long as it is positive. Summarizing, the complexity vanishes for $\hat{\alpha} < \hat{\alpha}_{\text{SP}}$, and equals $(\ln 2 - \hat{\alpha})$ for $\hat{\alpha} \in [\hat{\alpha}_{\text{SP}}, \ln 2]$. In particular the static phase transition threshold is $\hat{\alpha}_s \doteq \ln 2$. Beyond the static phase transition threshold the complexity is negative and loses its meaning (one has to modify the SP formalism used here). Higher order corrections can be computed in powers of 2^{-K} , see [120].

Let us now discuss the coupled case. The picture which emerges is similar to the one for the much simpler Curie-Weiss Chain model [50] and coupled LDPC codes over the binary erasure channel [2]. Before discussing the numerical results we wish to give a heuristic argument that “explains” why threshold saturation occurs. The argument can presumably be turned into a rigorous proof using the methods in [2] for LDPC codes on the binary erasure channel.

For the sake of the argument suppose that we fix $\hat{\alpha} > \hat{\alpha}_{\text{SP}}$ and that we look for profile solutions of (9.44), on an infinite chain $L \rightarrow +\infty$, that interpolate between the (asymmetric) boundary conditions $\varphi_z = 0$, $z \rightarrow -\infty$ and $\varphi_z \rightarrow \varphi_{\text{mst}}$, $z \rightarrow +\infty$. We take as an ansatz, a kink approaching its asymptotic values (at the two ends) fast enough, with a transition region localized in a region of size $O(w)$ centered at a position $z_{\text{kink}} = \xi L$ ($\xi \in [0, 1]$). Figure 9.4 gives an

illustrative picture of the kink profile. We have

$$\bar{\varphi} \equiv \frac{1}{L} \sum_{z=0}^{L-1} \varphi_z \approx \frac{1}{L} (L - \xi L) \varphi_{\text{mst}} = (1 - \xi) \varphi_{\text{mst}}. \quad (9.56)$$

Also, it is easy to see that the associated complexity as a function of ξ , or equivalently $\bar{\varphi}$, is approximately given by a convex combination of the two minima of $\Sigma_{1,1}(\alpha, \varphi)$ (given in (9.55)) which correspond to the two points $\varphi = 0$ (with $\Sigma = 0$) and $\varphi = \varphi_{\text{st}}$ (with $\Sigma \approx \ln 2 - \hat{\alpha}$). More precisely,

$$\begin{aligned} \Sigma_{\text{kink}}(\xi) &\approx \frac{1}{L} [\xi L \times 0 + (L - \xi L) \times (\ln 2 - \hat{\alpha})] \\ &\approx \frac{\bar{\varphi}}{\varphi_{\text{mst}}} (\ln 2 - \hat{\alpha}). \end{aligned}$$

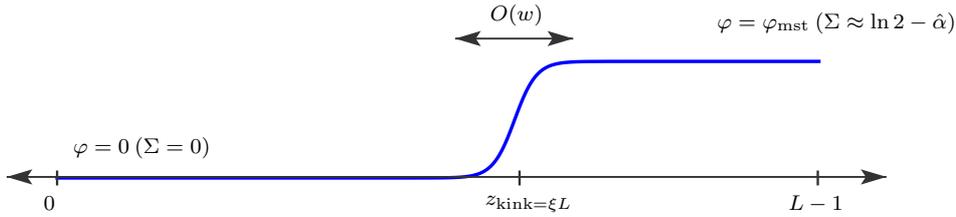


Figure 9.4: An illustrative picture of a kink-like ansatz $\{\varphi_z\}_{z=-\frac{L}{2}+1}^{\frac{L}{2}}$ for a solution of (9.44). At the right end, the kink converges to the value $\varphi = \varphi_{\text{st}}$ (with corresponding complexity $\Sigma \approx \ln 2 - \hat{\alpha}$) and at the left end it converges to $\varphi = 0$ (with $\Sigma = 0$). The transition region of size $O(w)$ which is centered at $z = z_{\text{kink}}$.

When $\hat{\alpha} < \hat{\alpha}_s$, the minimum is at $\xi = 1$ ($\bar{\varphi} = 0$). This means that the kink center will form a traveling wave through the chain, and reach its unique stable location at the right end. On the other hand when $\hat{\alpha} > \hat{\alpha}_s$ the minimum is at $\xi = 0$ ($\bar{\varphi} = \varphi_{\text{mst}}$) and the kink will travel towards the left to reach its stable location. Within the present approximation, for $\hat{\alpha} = \hat{\alpha}_s$ any position along the chain is stable for the kink center.

Summarizing, this heuristic argument suggests that for $\hat{\alpha} < \hat{\alpha}_s$ the fixed point equations (9.44) only have the trivial solution $\{\varphi_z = 0\}$, while for $\hat{\alpha} > \hat{\alpha}_s$ the only solution is $\{\varphi_z = \varphi_{\text{mst}}\}$. This means that the SP threshold coincides with $\hat{\alpha}_s$. Here, ξ has been treated as a continuous variable, which is expected to be valid only in a limit of large w . For large but finite w there will subsist a small gap between the SP and static thresholds, and for $\hat{\alpha}$ fixed in this gap only a discrete set of positions for the kink are stable. The number of such stable positions is roughly equal to $2L$.

We have solved (9.44) numerically with *symmetric* boundary conditions $\varphi_z = 0$, $z < 0$, $z \geq L$ and fixed $\bar{\varphi} \equiv \frac{1}{L} \sum_{z=0}^{L-1} \varphi_z$. In order to find a solution for *all* values of $\bar{\varphi}$ we have to let $\hat{\alpha}$ vary slightly. In other words we find a solution $(\hat{\alpha}(\bar{\varphi}); \{\varphi_z(\bar{\varphi})\})$ that is parametrized by $\bar{\varphi}$. Define the *van der Waals curve* (Figure 9.5) as the function $\hat{\alpha}(\bar{\varphi})$. The minimum of the van der Waals curve yields (as for the individual system) the SP threshold $\alpha_{SP,w,L}$ (see Table 9.2 for numerical values).

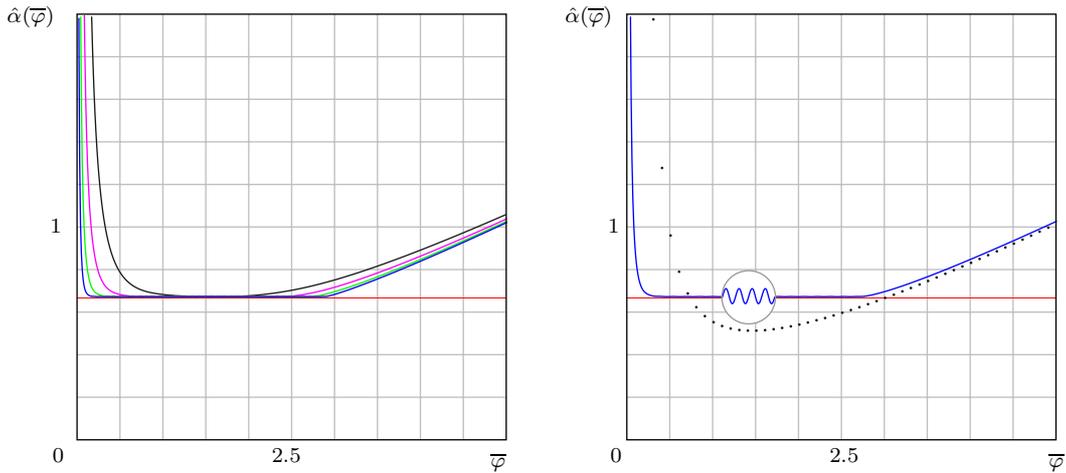


Figure 9.5: *Left:* sequence of van der Waals curves $\hat{\alpha}(\bar{\varphi})$, for $K = 5$, $w = 3$ and $L = 10, 20, 40, 80$ (top to bottom). For $\bar{\varphi} \in [\varphi_{\text{mst}}, +\infty]$ they converge to the individual system curve. *Right:* a magnification of the plateau region for $K = 5$, $w = 3$ and $L = 40$ shows the fine structure. The dotted line is the curve for the individual system and the red line shows the static phase transition threshold $\hat{\alpha}_s = 0.666$.

As L increases, the curves develop a plateau at height $\approx \hat{\alpha}_s$ for the interval $\bar{\varphi} \in [0, \varphi_{\text{mst}}]$. Moreover they converge to the van der Waals curve of the individual system for $\bar{\varphi} \in [\varphi_{\text{mst}}, +\infty]$, a fact that is consistent with theorems 9.1, 9.2. Precise enough numerics show that as long as w is finite the curves display a fine structure in the plateau interval: the magnification in Figure 9.5 shows wiggles of very small amplitude. We observe that their amplitude decays as w grows and K is fixed (we expect from [50] that this decay is exponential); and grows larger as K increases with w fixed (see Table 9.2).

Figure 9.6 illustrates the solutions of the fixed point equations for $\hat{\alpha}$ in the wiggle region for large K . The top curve is the van der Waals curve in the wiggle region. The middle left warning density profile is the fixed point solution corresponding the left point with coordinates $(\bar{\varphi}_l, \hat{\alpha}_l)$. Note that $\hat{\alpha}_l = \hat{\alpha}_{SP,L,w}$. For this point the total average complexity is approximately equal to $\frac{\bar{\varphi}_l}{\varphi_{\text{mst}}}(\hat{\alpha}_s - \hat{\alpha}_l)$. The bottom left curve shows the complexity profile. In the

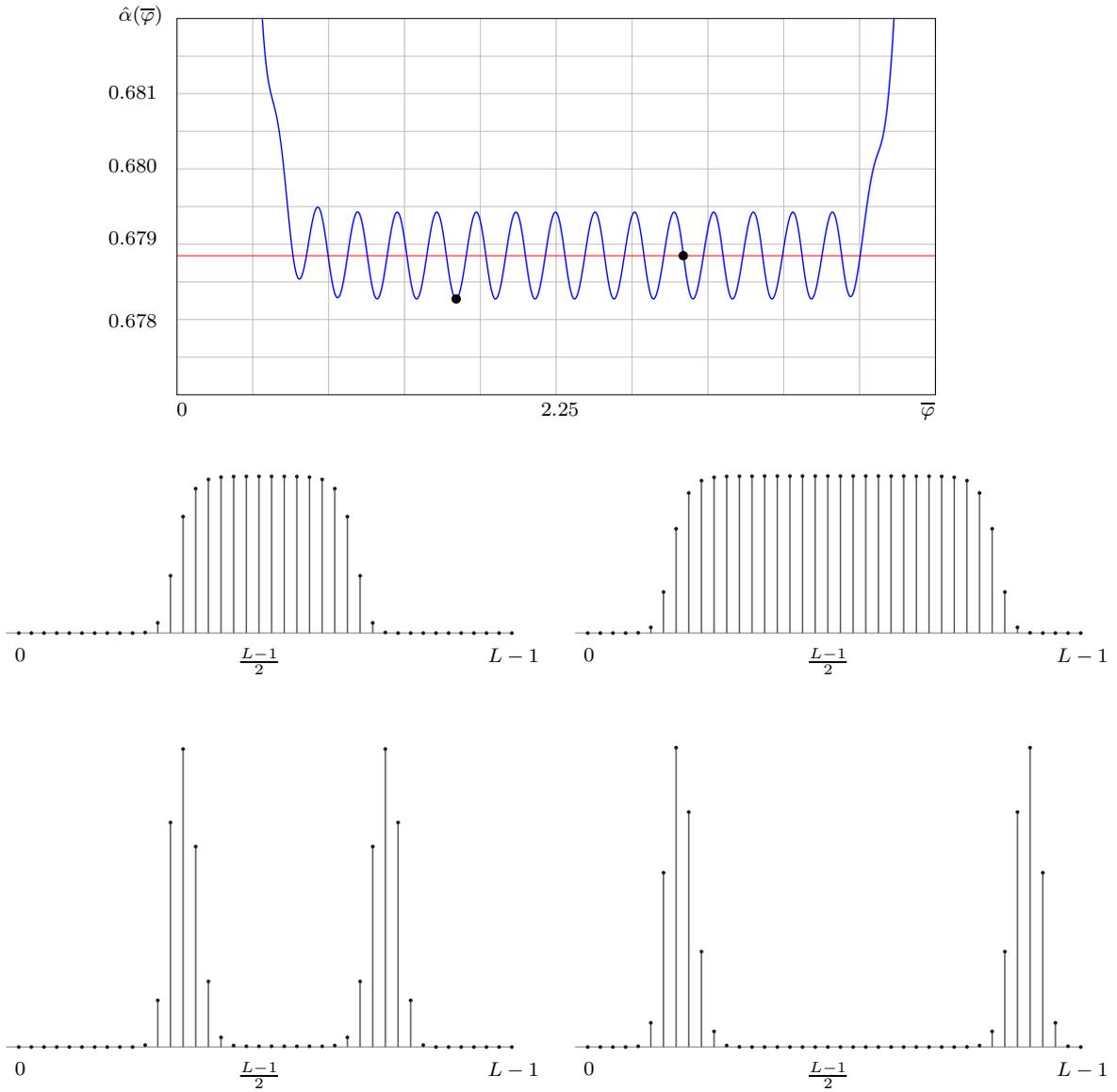


Figure 9.6: van der Waals curve in the wiggle region for the coupled system (top) for $K = 7$, $w = 4$ and $L = 40$. The red line is at the static phase transition threshold. The left point $(\bar{\varphi}_l, \hat{\alpha}_l) = (1.657, 0.678274)$ corresponds to the warning (middle left) and complexity (bottom left) profiles. In the latter the height of the middle part is $\hat{\alpha}_s - \hat{\alpha}_l \approx 0.010$. The right point $(\bar{\varphi}_r, \hat{\alpha}_r) = (3.00585, 0.688847)$ corresponds to the warning and complexity profiles on the right.

K	5	7	10
$\hat{\alpha}_s$	0.666	0.686	0.692
$\hat{\alpha}_{\text{SP}}$	0.513	0.449	0.370
$\hat{\alpha}_{\text{SP},80,3}$	0.672	0.682	0.651
$\hat{\alpha}_{\text{SP},80,5}$	0.672	0.688	0.691
$\hat{\alpha}_{\text{SP},80,7}$	0.672	0.688	0.692

Table 9.2: SP Thresholds of the individual ($L = w = 1$) and coupled ensembles ($L = 80, w = 3, 5, 7$) are found from the van der Waals curves. For $K = 10$ we clearly see that the SP threshold saturates to $\hat{\alpha}_s$ from below as w increases.

middle part, the height of this profile is approximately $(\hat{\alpha}_s - \hat{\alpha}_l) \approx 0.010$. Consider now the point on the right with coordinates $(\bar{\varphi}_r, \hat{\alpha}_r)$. Note that we take this point very close to the static phase transition threshold $\hat{\alpha}_r \approx \hat{\alpha}_s$. As a consequence the total average complexity nearly vanishes. The middle right warning density profile is flat over the whole chain, except near the ends because we enforce the boundary conditions, and the complexity density nearly vanishes except in the transition regions.

9.6 Dynamical and Condensation Thresholds

The SP formalism says nothing about the relative sizes (internal entropy) of clusters of solutions and consequently does not take into account which of them are "relevant" to the uniform measure over zero energy solutions. For similar reasons, it is not clear that the SP threshold has particular algorithmic significance. These issues are partly addressed by the more elaborate entropic cavity method [84], [86], [103]. It predicts the existence of the dynamical and condensation thresholds α_d and α_c . The dynamical threshold is believed to separate a phase ($\alpha < \alpha_d$) where the uniform measure is essentially supported on one well connected cluster of dominant entropy, and a phase ($\alpha_d < \alpha < \alpha_c$) where the measure is supported on an exponential number of clusters with equal internal entropy. For $\alpha > \alpha_c$ the measure condenses on a "handful" of clusters of dominant entropy. The condensation threshold is a static thermodynamic transition in the sense that the total ground state entropy has a non-analyticity as a function of α . These thresholds were first computed for CSP in [104], where their algorithmic significance is also discussed. See also [90], [91] for recent related algorithmic results.

We have computed the dynamical and condensation thresholds of coupled CSP. Let us denote them $\alpha_{d,L,w}$ and $\alpha_{c,L,w}$ (with w fixed). We observe that as L increases $\alpha_{c,L,w} \rightarrow \alpha_c$. This observation is consistent with the following rigorous result that we prove in appendix 9.8.2: the thermodynamic limit of the free energy (at finite temperature) of the chain is identical to that of the individual model. From the free energy one can formally obtain the entropy by differentiating the free energy with respect to temperature. The result about the free energy then suggests that the zero temperature entropy of the chain

K	α_{SP}	$\alpha_{SP,80,3}$	α_d	$\alpha_{d,80,3}$	α_c	$\alpha_{c,80,3}$	α_s	$\alpha_{s,80,3}$
3	3.927	4.268	3.86	3.86	3.86	3.86	4.267	4.268
4	8.30	9.94	9.38	9.55	9.55	9.56	9.93	10.06

Table 9.3: Thresholds of individual and coupled K -SAT model for $L = 80$ and $w = 3$. The condensation and SAT-UNSAT thresholds correspond to non analyticities of the entropy and ground state energy and remain unchanged (for $L \rightarrow +\infty$). Already for $w = 3$ the dynamical and SP thresholds saturate very close to α_c and α_s .

and individual models have the same non-analyticity points as a function of the constraint density. The second important observation is that in the regime $1 \ll w \ll L$ we find $\alpha_{d,L,w} \rightarrow \alpha_c$. Thus the dynamical threshold saturates towards the condensation threshold⁶.

The dynamical and condensation thresholds are analogous to the dynamical and condensation temperatures of p -spin glass models for $p \geq 3$, and to the glassy and Kauzmann transition temperatures in structural glasses [121], [122], [123]. One expects that a similar saturation of the dynamical towards the condensation temperature holds for coupled p -spin glass models on complete graphs for $p \geq 3$. On the other hand, for $p = 2$ the replica symmetry breaking transition is continuous, there is no dynamical temperature, and spatial coupling is not expected to modify the phase diagram.

Table 9.3 summarizes all the behaviors of the SP, SAT/UNSAT, dynamical and condensation thresholds for the K -SAT problem. The situation for coloring is similar.

9.7 Further Remarks and Open Directions

In this chapter we have developed in detail the SP formalism for coupled K -SAT ensemble. We find that the SP thresholds of spatially coupled random K -SAT ensemble nicely saturate towards the SAT/UNSAT phase transition threshold of the individual ensemble. Moreover the SAT/UNSAT phase transition threshold of the coupled and individual ensembles are identical. The saturation of the SP threshold is remarkably similar to the one of the Belief Propagation algorithmic threshold (towards the optimal one associated to the Maximum a Posteriori decoder) observed in coding theory.

Let us point out a few issues that would deserve more investigations. The large K analysis has shown that when α is in a small interval where the zero-energy complexity is strictly positive, the warning and complexity densities form kink-like profiles. These are very similar to the kink-like magnetization and free energy densities found in the CW chain of Chapter 8. A possible interpretation of the complexity density profiles is that the clusters do not only have a “size” given by their internal entropy but also have a “shape” that could be taken into account by an extension of the entropic cavity method. The

⁶Note that for $K = 3$ we already have $\alpha_d = \alpha_c$ for the individual ensemble.

simplest system where this issue could be elucidated is the XOR-SAT system for which the clusters can be precisely defined [113].

As briefly discussed above, the entropic cavity method predicts the existence of other thresholds, namely the dynamical and condensation thresholds. We have checked that the condensation one is the same for a coupled and individual ensemble (for $L \rightarrow \infty$). This observation is consistent with the result of Theorem 9.5. We also observe that the dynamical threshold of the coupled ensemble saturates towards the condensation one, for $K \geq 4$. For $K = 3$ the dynamical and condensation thresholds coincide already for the individual ensemble. We consistently observe that they remain unchanged by coupling. These observations deserve more investigations, in particular the nature of the condensed phase, the freezing of variables, the behavior of correlation functions and the possible relevance of the shape of clusters.

9.8 Appendix

9.8.1 Proofs of Theorems 9.1 and 9.2

In this section we sketch the proofs of theorems 9.1 and 9.2. The proof of theorem 9.1 is straightforward and does not depend on the details of the model at hand. On the other hand that of theorem 9.2 has to be adapted for each model at hand.

Proof of Theorem 9.1

Recall that for the Hamiltonian of the open chain $\mathcal{H}_{\text{cou}}(\underline{x})$ in (9.3), $\underline{x} = (x_{iz})$ with $(i, z) \in \cup_{z=0, \dots, L-1} V_z$. It will be convenient to set $\underline{x} = (\underline{x}', \underline{x}'')$ where $\underline{x}' = (x_{iz}; z = 0, \dots, L-w)$ and $\underline{x}'' = (x_{iz}; z = L-w+1, \dots, L-1)$. Recall also that the Hamiltonian $\mathcal{H}_{\text{cou}}^{\text{per}}(\underline{x}')$ of the periodic chain is given by the same expression (9.3) with $\underline{x}' = (x_{iz}; z = 0, \dots, L-w)$. Therefore the difference between the two Hamiltonians only comes because of the terms $\psi_{cz}(\underline{x}_{\partial cz})$ with $z = L-2w+1, \dots, L-w$. In other words,

$$|\mathcal{H}_{\text{cou}}(\underline{x}', \underline{x}'') - \mathcal{H}_{\text{cou}}^{\text{per}}(\underline{x}')| \leq Mw, \quad (9.57)$$

for all \underline{x}'' . As a result, we obtain

$$\mathcal{H}_{\text{cou}}^{\text{per}}(\underline{x}') - Mw \leq \mathcal{H}_{\text{cou}}(\underline{x}', \underline{x}'') \quad (9.58)$$

and by taking the min, dividing by NL and taking the expectation, we deduce

$$e_{N,L,w}^{\text{per}}(\alpha) - \frac{\alpha w}{L} \leq e_{N,L,w}(\alpha). \quad (9.59)$$

To obtain the right-hand side inequality of (9.9), we recall that the periodic chain is obtained from the open chain by identifying the variable nodes at the left boundary of the open chain with their corresponding ones at the right boundary. Now, since the set of the check nodes of the open and periodic chain

are identical, it is clear to get the maximum satisfying assignment, the open chain has more degrees of freedom and hence its minimum Hamiltonian is less. Hence, the right part of (9.9) is proved.

Proof of Theorem 9.2

As explained in Section 9.2 the proof that the limit exists, is continuous and non-decreasing, for the individual models is provided in [85] and is essentially the same for the coupled periodic chain. Here we prove the equality of the two limits (9.10). The following notation is convenient. For a given graph instance G (from some ensemble) we call $\mathcal{H}_G(\underline{x})$ the corresponding Hamiltonian. It always exists, as in (9.3), of a sum of terms $1 - \psi_{cz}(x_{\partial cz})$ over constraints $(c, z) \in G$. The ground state energy is equal to $\min_{\underline{x}} \mathcal{H}_G(\underline{x})$. To set up suitable interpolation procedures, it is convenient to first define three extra ensembles.

The “connected” ensemble. This is essentially the individual $[N, K, \alpha]$ ensemble scaled by L . We have a set of LN variable nodes and a set of LM constraint nodes. Each constraint node has K edges connected u.a.r. to variable nodes. Expectations with respect to this ensemble are denoted by \mathbb{E}_{conn} . Because of the existence of the limit we have

$$\lim_{N \rightarrow +\infty} \frac{1}{LN} \mathbb{E}_{\text{conn}} [\min_{\underline{x}} \mathcal{H}_G(\underline{x})] = \lim_{N \rightarrow +\infty} e_N(\alpha), \quad (9.60)$$

for any fixed L .

The “disconnected” ensemble. This is a variant of the individual $[N, K, \alpha]$ ensemble replicated L times. We place at positions $z = 0, \dots, L-1, L$ disjoint sets of variable nodes V_z containing each N nodes. Each node from the set of LM constraint nodes is affected u.a.r. to a position $z = 0, \dots, L-1$. Note that the set \tilde{C}_z of constraint nodes at position z has cardinality $M_z \sim \text{Bi}(LM, \frac{1}{L})$. Each node from \tilde{C}_z has K edges that are connected u.a.r. to nodes in V_z . Expectations are denoted by \mathbb{E}_{disc} . Since each M_z is concentrated on M with a fluctuation $O(\sqrt{M})$, we can show by an argument similar to the proof of theorem 9.1 that

$$\frac{1}{LN} \mathbb{E}_{\text{disc}} [\min_{\underline{x}} \mathcal{H}_G(\underline{x})] = e_N(\alpha) + O(N^{-1/2}), \quad (9.61)$$

where $O(N^{-1/2})$ is uniform in L .

The “ring” ensemble. This is a variant of the periodic chain in Section 9.2. We place at positions $z = 0, \dots, L-1, L$ disjoint sets of variable nodes V_z , each containing N nodes. Now we have a set of LM constraint nodes. Each constraint node is affected to a position $z = 0, \dots, L-1$ u.a.r. and (say the position is z) its K edges are connected u.a.r. to the set of variables $\cup_{k=0}^{w-1} V_{z+k \bmod L}$. Note that the sets \tilde{C}_z of constraint nodes have cardinalities $M_z \sim \text{Bi}(LM, \frac{1}{L})$. We denote by \mathbb{E}_{ring} the expectation with respect to this ensemble. Since each M_z is concentrated on M with a fluctuation $O(\sqrt{M})$, an

argument similar to the proof of theorem 9.1 shows that

$$\frac{1}{LN} \mathbb{E}_{\text{ring}}[\min_{\underline{x}} \mathcal{H}_G(\underline{x})] = e_{N,L,w}^{\text{per}}(\alpha) + O(N^{-1/2}), \quad (9.62)$$

where $O(N^{-1/2})$ is uniform in L (and depends on w).

We will show

$$\mathbb{E}_{\text{conn}}[\min_{\underline{x}} \mathcal{H}_G(\underline{x})] \leq \mathbb{E}_{\text{ring}}[\min_{\underline{x}} \mathcal{H}_G(\underline{x})] \leq \mathbb{E}_{\text{disc}}[\min_{\underline{x}} \mathcal{H}_G(\underline{x})], \quad (9.63)$$

which allows to conclude the proof of the theorem by using (9.60), (9.61), (9.62).

Left inequality in (9.63). We build a sequence of interpolating “ r -ensembles”, $r = 0, \dots, LM$, interpolating between the ring ($r = 0$) and connected ($r = LM$) ensembles. We have two sets of LM constraint and LN variable nodes. The variable nodes are organized into L disjoint sets V_z each containing N nodes, placed along the positions $z = 0, \dots, L - 1$. Expectation with respect to the r -ensemble is denoted by \mathbb{E}_r . To sample a graph G_r from this ensemble we first take r nodes - called type 1 - from the set of LM constraint nodes. Each one has K edges which are connected u.a.r. to the set of LN variable nodes. For the remaining $LM - r$ constraint nodes - called type 2 - we proceed as follows: each one is affected u.a.r. to a position z , and its K edges are then connected u.a.r. to the wN variable nodes in $\cup_{k=0}^{w-1} V_{z+k \bmod L}$. We claim that for $1 \leq r \leq LM$,

$$\mathbb{E}_r[\min_{\underline{x}} \mathcal{H}_{G_r}(\underline{x})] \leq \mathbb{E}_{r-1}[\min_{\underline{x}} \mathcal{H}_{G_{r-1}}(\underline{x})]. \quad (9.64)$$

Clearly this implies the left inequality in (9.63). Let us prove this claim. Take a random graph G_r and delete u.a.r. a constraint from the type 1 nodes: this yields an intermediate graph \tilde{G} . One can go back to a random graph G_r by adding back a type 1 node according to the above rules, or one can go to a random graph G_{r-1} by adding back a type 2 node according to the above rules. We will prove that conditioned on any realization of \tilde{G} we have

$$\mathbb{E}_r[\min_{\underline{x}} \mathcal{H}_{G_r}(\underline{x}) \mid \tilde{G}] \leq \mathbb{E}_{r-1}[\min_{\underline{x}} \mathcal{H}_{G_{r-1}}(\underline{x}) \mid \tilde{G}]. \quad (9.65)$$

Claim (9.64) follows by averaging over \tilde{G} . We now prove (9.65) for K -SAT problem.

Consider the set of “optimal assignments” \underline{x} that minimize $\mathcal{H}_{\tilde{G}}(\underline{x})$. We say that a variable is *frozen* iff it takes the same value for all optimal assignments. We call \mathcal{F} the set of variable nodes with frozen variables and $\mathcal{F}_z = \mathcal{F} \cap V_z$. Now consider adding a new constraint node n to the graph \tilde{G} . This will cost an extra energy iff the node n connects only to frozen variable nodes *and* does not satisfy them. For such an event we have

$$\min_{\underline{x}} \mathcal{H}_{\tilde{G} \cup n}(\underline{x}) - \min_{\underline{x}} \mathcal{H}_{\tilde{G}}(\underline{x}) = 1. \quad (9.66)$$

When the node n is connected u.a.r. to the LN variable nodes (n is type 1) this event has probability $\frac{1}{2^K} \left(\frac{|\mathcal{F}|}{LN}\right)^K$. Thus

$$\mathbb{E}_r[\min_{\underline{x}} \mathcal{H}_{G_r}(\underline{x}) \mid \tilde{G}] - \min_{\underline{x}} \mathcal{H}_{\tilde{G}}(\underline{x}) = \frac{1}{2^K} \left(\frac{|\mathcal{F}|}{LN}\right)^K. \quad (9.67)$$

Similarly when the node n is affected u.a.r. to a position z and then connected u.a.r. to $\cup_{k=0}^{w-1} V_{z+k \bmod L}$ (n is type 2) we get

$$\mathbb{E}_{r-1}[\min_{\underline{x}} \mathcal{H}_{G_{r-1}}(\underline{x}) \mid \tilde{G}] - \min_{\underline{x}} \mathcal{H}_{\tilde{G}}(\underline{x}) = \frac{1}{L} \sum_{z=0}^{L-1} \frac{1}{2^K} \left(\frac{1}{wN} \sum_{k=0}^{w-1} |\mathcal{F}_{z+k \bmod L}|\right)^K. \quad (9.68)$$

Claim (9.65) follows from the last two equations, convexity of x^K for $x \geq 0$, and

$$|\mathcal{F}| = \sum_{z=0}^{L-1} \frac{1}{w} \sum_{k=0}^{w-1} |\mathcal{F}_{z+k \bmod L}|. \quad (9.69)$$

Right inequality in (9.63). We construct new r -ensembles, $r = 0, \dots, LM$ that now interpolate between the disconnected ($r = 0$) and the ring ($r = LM$) ensembles. A random graph G_r is constructed as follows. We have a set of LM constraint nodes and a set of LN variable nodes organized into L disjoint sets V_z each containing N nodes, placed along positions z . We first take r constraint nodes, called type 1. Each of them is affected u.a.r. to a position z , and its K edges are connected u.a.r. to variable nodes in V_z . Next, the remaining $LM - r$ constraints nodes - called type 2 - are each affected u.a.r. to a position z and its K edges are connected u.a.r. to wN nodes in $\cup_{k=0}^{w-1} V_{z+k \bmod L}$. Note that at each position there are $\text{Bi}(r, \frac{1}{L})$ type 1 nodes and $\text{Bi}(LM - r, \frac{1}{L})$ type 2 nodes, so in total there are $\text{Bi}(LM, \frac{1}{L})$ constraint nodes. Similarly to the previous interpolation we will prove

$$\mathbb{E}_r[\min_{\underline{x}} \mathcal{H}_{G_r}(\underline{x})] \leq \mathbb{E}_{r-1}[\min_{\underline{x}} \mathcal{H}_{G_{r-1}}(\underline{x})]. \quad (9.70)$$

This inequality implies the upper bound in (9.63). To prove (9.70), as before, we consider the random graph \tilde{G} obtained by deleting u.a.r. a type 1 node from G_r . From \tilde{G} one gets a random graph G_r by adding back a type 1 node, or one gets a graph G_{r-1} by adding back a type 2 node instead. We first prove that

$$\mathbb{E}_r[\min_{\underline{x}} \mathcal{H}_{G_r}(\underline{x}) \mid \tilde{G}] \leq \mathbb{E}_{r-1}[\min_{\underline{x}} \mathcal{H}_{G_{r-1}}(\underline{x}) \mid \tilde{G}], \quad (9.71)$$

and then by averaging over graphs \tilde{G} we get (9.70). Let us briefly sketch the derivation of (9.71).

We use the same sets \mathcal{F}_z of frozen variables at position z corresponding to the ground state configurations of $\mathcal{H}_{\tilde{G}}(\underline{x})$. We have

$$\mathbb{E}_r[\min_{\underline{x}} \mathcal{H}_{G_r}(\underline{x}) \mid \tilde{G}] - \min_{\underline{x}} \mathcal{H}_{\tilde{G}}(\underline{x}) = \frac{1}{L} \sum_{z=0}^{L-1} \frac{1}{2^K} \left(\frac{|\mathcal{F}_z|}{N}\right)^K, \quad (9.72)$$

and

$$\mathbb{E}_{r-1}[\min_{\underline{x}} \mathcal{H}_{G_{r-1}}(\underline{x}) \mid \tilde{G}] - \min_{\underline{x}} \mathcal{H}_{\tilde{G}}(\underline{x}) = \frac{1}{L} \sum_{z=0}^{L-1} \frac{1}{2^K} \left(\frac{1}{wN} \sum_{k=0}^{w-1} |\mathcal{F}_{z+k \bmod L}| \right)^K. \quad (9.73)$$

Estimate (9.71) then follows by the convexity of the function x^K for $x \geq 0$.

9.8.2 Finite Temperature Version

The finite Gibbs distribution (at “inverse temperature” β) associated to the coupled CSP Hamiltonian (9.3) is

$$\mu_\beta(\underline{x}) = \frac{1}{Z_{\text{cou}}} e^{-\beta \mathcal{H}_{\text{cou}}(\underline{x})}, \quad Z_{\text{cou}} = \sum_{\underline{x}} e^{-\beta \mathcal{H}_{\text{cou}}(\underline{x})}, \quad (9.74)$$

and the average free energy per node is

$$f_{N,L,w}(\alpha, \beta) = -\frac{1}{\beta LN} \mathbb{E}[\ln Z_{\text{cou}}]. \quad (9.75)$$

The corresponding quantities $\mathcal{H}_{\text{cou}}^{\text{per}}(\underline{x})$ are associated a chain to with periodic boundary conditions (see Section 9.2); these will be denoted by a superscript “per”. Note that to get these quantities for the underlying system, one sets $L = w = 1$ in these definitions; the average free energy per node will be denoted by $f_N(\alpha, \beta)$.

We sketch the proof of the analogs of theorems 9.1 and 9.2.

Theorem 9.4 (Comparison of open and periodic chains). *For general coupled CSP $[N, K, \alpha, w, L]$ ensembles we have*

$$|f_{N,L,w}(\alpha, \beta) - f_{N,L,w}^{\text{per}}(\alpha, \beta)| \leq \frac{\alpha w}{L}. \quad (9.76)$$

Proof. We write (with the same notations than in the proof of theorem 9.1)

$$Z_{\text{cou}} = \sum_{\underline{x}} e^{-\beta(\mathcal{H}_{\text{cou}}(\underline{x}))} = \sum_{\underline{x}', \underline{x}''} e^{-\beta \mathcal{H}_{\text{cou}}^{\text{per}}(\underline{x}'')} e^{-\beta(\mathcal{H}_{\text{cou}}(\underline{x}') - \mathcal{H}_{\text{cou}}^{\text{per}}(\underline{x}''))} \quad (9.77)$$

and from (9.57) we deduce

$$e^{-\beta M w} Z_{\text{cou}}^{\text{per}} \leq Z_{\text{cou}} \leq e^{\beta M w} Z_{\text{cou}}^{\text{per}}. \quad (9.78)$$

Applying $-\frac{1}{\beta N L} \log$ on each side of this inequality, we obtain the desired estimate. \square

Theorem 9.5 (Comparison of individual and periodic ensembles). *For K -SAT and Q -coloring the limits $\lim_{N \rightarrow +\infty} f_{N,L,w}^{\text{per}}(\alpha, \beta)$ and $\lim_{N \rightarrow +\infty} f_N(\alpha, \beta)$ exist, and are continuous in (α, β) , for all L . Moreover,*

$$\lim_{N \rightarrow +\infty} f_{N,L,w}^{\text{per}}(\alpha, \beta) = \lim_{N \rightarrow +\infty} f_N(\alpha, \beta). \quad (9.79)$$

Theorems 9.4 and 9.5 yield (recall $\lim_{\text{therm}} = \lim_{L \rightarrow +\infty} \lim_{N \rightarrow +\infty}$)

$$\lim_{\text{therm}} f_{N,L,w}(\alpha, \beta) = \lim_{\text{therm}} f_{N,L,w}^{\text{per}}(\alpha, \beta) = \lim_{N \rightarrow +\infty} f_N(\alpha, \beta). \quad (9.80)$$

Proof. The proof of existence and continuity of limits for $N \rightarrow +\infty$ (L fixed) is identical to [85], so we do not repeat it here. The proof of the equality uses the same interpolating r -ensembles between the connected, ring and disconnected ensembles defined in subsection 9.8.1. The associated Gibbs measures, free energies and expectations will be denoted by scripts r , conn, ring and disc.

By an argument similar to that of theorem 9.4 we have the analogs of (9.60), (9.61), (9.62),

$$\begin{cases} -\lim_{N \rightarrow +\infty} \frac{1}{\beta LN} \mathbb{E}_{\text{conn}}[\log Z_{\text{conn}}] = \lim_{N \rightarrow +\infty} f_N(\alpha, \beta), \text{ for } L \text{ fixed,} \\ -\frac{1}{\beta LN} \mathbb{E}_{\text{disc}}[\log Z_{\text{disc}}] = f_N(\alpha, \beta) + O(N^{-1/2}), \text{ uniformly in } L, \\ -\frac{1}{\beta LN} \mathbb{E}_{\text{ring}}[\log Z_{\text{ring}}] = f_{N,L,w}^{\text{per}}(\alpha, \beta) + O(N^{-1/2}), \text{ uniformly in } L. \end{cases}$$

Thus, it is sufficient to show that

$$-\frac{1}{LN} \mathbb{E}_{\text{conn}}[\log Z_{\text{conn}}] \leq -\frac{1}{LN} \mathbb{E}_{\text{ring}}[\log Z_{\text{ring}}] \leq -\frac{1}{LN} \mathbb{E}_{\text{disc}}[\log Z_{\text{disc}}]. \quad (9.81)$$

To prove these inequalities we will use the r -ensembles. It suffices to check the analogs of (9.65) and (9.71), namely for an intermediate graph \tilde{G} ,

$$-(\mathbb{E}_r[\log Z_{G_r} | \tilde{G}] - \log Z_{\tilde{G}}) \leq -(\mathbb{E}_{r-1}[\log Z_{G_{r-1}} | \tilde{G}] - \log Z_{\tilde{G}}), \quad (9.82)$$

and then average over \tilde{G} .

Consider the graph \tilde{G} and add a new constraint node n to it. The precise way in which n is connected to the variable nodes is deferred to a later stage of the argument. We have

$$\frac{Z_{\tilde{G} \cup n}}{Z_{\tilde{G}}} = e^{-\beta} \sum_{\mathbf{x}: n \text{ is UNSAT}} \mu_{\beta, \tilde{G}}(\mathbf{x}) + \sum_{\mathbf{x}: n \text{ is SAT}} \mu_{\beta, \tilde{G}}(\mathbf{x}). \quad (9.83)$$

This is equivalent to

$$\frac{Z_{\tilde{G} \cup n}}{Z_{\tilde{G}}} = 1 - (1 - e^{-\beta}) \sum_{\mathbf{x}: n \text{ is UNSAT}} \mu_{\beta, \tilde{G}}(\mathbf{x}). \quad (9.84)$$

Taking the log and expectation over n for a given \tilde{G} , we obtain

$$-\mathbb{E}[\log \frac{Z_{\tilde{G} \cup n}}{Z_{\tilde{G}}} | \tilde{G}] = -\mathbb{E}[\log \{1 - (1 - e^{-\beta}) \sum_{\mathbf{x}: n \text{ is UNSAT}} \mu_{\beta, \tilde{G}}(\mathbf{x})\} | \tilde{G}]. \quad (9.85)$$

Note that the left hand side is identical to that of (9.82). To compute the expectation we expand $-\log(1-x) = \sum_{l \geq 1} \frac{x^l}{l}$,

$$\begin{aligned} -\mathbb{E}\left[\log \frac{Z_{\tilde{G} \cup n}}{Z_{\tilde{G}}} \mid \tilde{G}\right] &= \sum_{l \geq 1} \frac{(1 - e^{-\beta})^l}{l} \\ &\quad \times \mathbb{E}\left[\sum_{\underline{x}^{(1)}, \dots, \underline{x}^{(l)}: n \text{ is UNSAT}} \mu_{\beta, \tilde{G}}(\underline{x}^{(1)}) \dots \mu_{\beta, \tilde{G}}(\underline{x}^{(l)}) \mid \tilde{G}\right]. \end{aligned} \quad (9.86)$$

The sum over “real replicas” $\underline{x}^{(1)}, \dots, \underline{x}^{(l)}$ is over assignments such that n is UNSAT for all l of them, so the expectation in (9.86) equals

$$\frac{1}{Z_{\tilde{G}}^l} \sum_{\underline{x}^{(1)}, \dots, \underline{x}^{(l)}} e^{-\beta \sum_{\rho=1}^l \mathcal{H}_{\tilde{G}}(\underline{x}^{(\rho)})} \mathbb{E}\left[\mathbb{1}\{n \text{ UNSAT on all } \underline{x}^{(\rho)}, h = 1, \dots, l\} \mid \tilde{G}\right]. \quad (9.87)$$

Up to this stage the arguments are completely general: they apply both to coloring and satisfiability. We specialize the rest of the proof to K -SAT and leave coloring as an exercise.

We first derive (9.82) for the r -ensemble that interpolates between the *connected and ring* ensembles. This then implies the left inequality in (9.81). Given \tilde{G} and given a term $\underline{x}^{(1)}, \dots, \underline{x}^{(l)}$ in (9.87), let \mathcal{F} be the set of variable nodes with frozen bits, i.e those variable nodes such that the bit takes the same value in all assignments $\underline{x}^{(1)}$ through $\underline{x}^{(l)}$. Below we will also need the sets $\mathcal{F}_z = \mathcal{F} \cap V_z$. When n is connected u.a.r. to the LN variable nodes we go from \tilde{G} to a G_r graph and

$$\mathbb{E}_r\left[\mathbb{1}\{n \text{ UNSAT on all } \underline{x}^{(\rho)}, h = 1, \dots, l\} \mid \tilde{G}\right] = \frac{1}{2^K} \left(\frac{|\mathcal{F}|}{LN}\right)^K. \quad (9.88)$$

On other hand when n is first affected u.a.r. to a position z and then connected u.a.r. to the wN variables in $\cup_{k=0}^{w-1} V_{z+k \bmod L}$, we go from \tilde{G} to a G_{r-1} graph and

$$\begin{aligned} &\mathbb{E}_{r-1}\left[\mathbb{1}\{n \text{ UNSAT on all } \underline{x}^{(\rho)}, h = 1, \dots, l\} \mid \tilde{G}\right] \\ &= \frac{1}{L} \sum_{z=0}^{L-1} \frac{1}{2^K} \left(\frac{1}{wN} \sum_{k=0}^{w-1} |\mathcal{F}_{z+k \bmod L}|\right)^K. \end{aligned} \quad (9.89)$$

By convexity, the quantity in (9.88) is smaller than the one in (9.89). Using this fact together with (9.85), (9.86), (9.87), we obtain the final inequality (9.82). This implies the left inequality in (9.81).

The derivation of (9.82) for the r -ensemble that interpolates between the *ring and disconnected* ensembles is similar. When n is first affected u.a.r. to a position z , and then connected u.a.r. to N variable nodes in the set V_z we go from \tilde{G} to a G_{r-1} graph. Thus,

$$\mathbb{E}_{r-1}\left[\mathbb{1}\{n \text{ UNSAT on all } \underline{x}^{(\rho)}, h = 1, \dots, l\} \mid \tilde{G}\right]$$

$$= \frac{1}{L} \sum_{z=0}^{L-1} \frac{1}{2^K} \left(\frac{|\mathcal{F}_z|}{N} \right)^K. \quad (9.90)$$

Finally we notice that by convexity, (9.89) is smaller than (9.90), so that using again (9.85), (9.86) and (9.87) we obtain the final inequality (9.82). This now implies the right inequality in (9.81). \square

9.8.3 Review of the Cavity Method and Survey Propagation Equations

The main assumptions of the cavity method draw on the concept of pure (or extremal or ergodic) state. While this concept can be given a rigorous meaning for “simple” Ising-type models [109], [110], it still forms a heuristic framework in the context of disordered spin systems. We refer the interested reader to [72], [115], [116], [117], [118] for more information and various approaches.

Infinite volume Gibbs measures form a convex set whose extremal points play a special role and are called *pure states*. A crucial property of a pure state is that the correlations decay (usually exponentially fast) with the graph distance. This is not true for non-trivial convex superpositions of pure states. For “simple” Ising-type models the number of pure states is “small” and they are related by a broken symmetry. Disordered spin systems can have an exponential (in system size) number of pure states and the broken symmetry, if only there exist one, is hard to identify⁷. The growth rate of the number of pure states, is called the complexity. This is a notion analogous to the Boltzmann entropy, but at the level of pure states, instead of microscopic configurations, for which one develops a new “level” of statistical mechanics.

We assume that this picture can be taken over to CSP and even coupled-CSP. Let p index the set of pure states. The special feature about systems on random graphs is that they are locally tree-like with high probability. Thus, since for each pure state p the correlations decay sufficiently fast, the marginals of the pure state p can be computed from the sum-product (or BP) equations

$$\hat{\nu}_{cz \rightarrow iu}^{(p)}(x_{iu}) \cong \sum_{x_{\partial(cz) \setminus iu}} \psi_{cz}(x_{\partial(cz)}) \prod_{jv \in \partial(cz) \setminus iu} \nu_{jv \rightarrow cz}^{(p)}(x_{jv}), \quad (9.91)$$

$$\nu_{iu \rightarrow cz}^{(p)}(x_{iu}) \cong \prod_{bv \in \partial(iu) \setminus cz} \hat{\nu}_{bv \rightarrow iu}^{(p)}(x_{iu}). \quad (9.92)$$

In (9.91), (9.92) \cong means that the right hand side has to be divided by a normalization factor to get a true marginal on the left. The free energy of the pure state p is given by the Bethe expression,

$$\beta F^{(p)} = \sum_{cz} \ln \left\{ \sum_{x_{\partial(cz)}} \psi_{cz}(x_{\partial(cz)}) \prod_{iu \in \partial(cz)} \nu_{iu \rightarrow cz}^{(p)}(x_{iu}) \right\}$$

⁷Within the replica formalism it is a formal symmetry between “a number” of copies of the system.

$$\begin{aligned}
& + \sum_{iu} \ln \left\{ \sum_{x_{iu}} \prod_{cz \in \partial(iu)} \hat{\nu}_{cz \rightarrow iu}^{(p)}(x_{iu}) \right\} \\
& - \sum_{\langle cz, iu \rangle \in E} \ln \left\{ \sum_{x_{iu}} \nu_{iu \rightarrow cz}^{(p)}(x_{iu}) \hat{\nu}_{cz \rightarrow iu}^{(p)}(x_{iu}) \right\}. \tag{9.93}
\end{aligned}$$

To investigate the zero temperature limit $\beta \rightarrow +\infty$ we set

$$\nu_{iu \rightarrow cz}^{(p)}(x_{iu}) = \frac{e^{-\beta E_{iu \rightarrow cz}^{(p)}(x_{iu})}}{\sum_{x_{iu} \in \mathcal{X}} e^{-\beta E_{iu \rightarrow cz}^{(p)}(x_{iu})}}, \quad \hat{\nu}_{cz \rightarrow iu}^{(p)}(x_{iu}) = \frac{e^{-\beta \hat{E}_{cz \rightarrow iu}^{(p)}(x_{iu})}}{\sum_{x_{iu} \in \mathcal{X}} e^{-\beta \hat{E}_{cz \rightarrow iu}^{(p)}(x_{iu})}}. \tag{9.94}$$

When $\beta \rightarrow \infty$, the sum-product equations (9.91) and (9.92) reduce to the min-sum equations

$$\begin{aligned}
E_{iu \rightarrow cz}(x_{iu}) &= \min \left\{ 1, \sum_{bv \in \partial(iu) \setminus cz} \hat{E}_{bv \rightarrow iu}(x_{iu}) - C_{iu \rightarrow cz} \right\} \\
&\equiv \mathcal{G}_{iu \rightarrow cz} \left[\{ \hat{E}_{bv \rightarrow iu} \}_{bv \in \partial(iu) \setminus cz} \right], \tag{9.95}
\end{aligned}$$

$$\begin{aligned}
\hat{E}_{cz \rightarrow iu}(x_{iu}) &= \min_{x_{\partial cz \setminus iu}} \left\{ (1 - \psi_{cz}(x_{\partial cz})) + \sum_{jv \in \partial(cz) \setminus iu} E_{jv \rightarrow cz}(x_j) \right\} - \hat{C}_{cz \rightarrow iu} \\
&\equiv \hat{\mathcal{G}}_{cz \rightarrow iu} \left[\{ E_{jv \rightarrow cz} \}_{jv \in \partial(cz) \setminus iu} \right]. \tag{9.96}
\end{aligned}$$

Here, $C_{iu \rightarrow cz}$ and $\hat{C}_{cz \rightarrow iu}$ are normalization constants fixed so that $\min_{x_{iu}} E_{iu \rightarrow cz}(x_{iu}) = \min_{x_{iu}} \hat{E}_{cz \rightarrow iu}(x_{iu}) = 0$. The Bethe formula for the free energy of a pure state reduces to an expression for its ground-state energy

$$\lim_{\beta \rightarrow +\infty} \beta F^{(p)} = \mathcal{E} \left[\{ E_{iu \rightarrow cz}^{(p)}(\cdot), E_{cz \rightarrow iu}^{(p)}(\cdot) \} \right], \tag{9.97}$$

where the functional \mathcal{E} is given by

$$\begin{aligned}
\mathcal{E} \left[\{ E_{iu \rightarrow cz}, E_{cz \rightarrow iu} \} \right] &= \sum_{cz} \min_{x_{\partial cz}} \left\{ (1 - \psi_{cz}(x_{\partial cz})) + \sum_{iu \in \partial(cz)} E_{iu \rightarrow cz}(x_{iu}) \right\} \\
&+ \sum_{iu} \min_{x_{iu}} \left\{ \sum_{cz \in \partial iu} \hat{E}_{cz \rightarrow iu}(x_{iu}) \right\} - \sum_{\langle cz, iu \rangle} \min_{x_{iu}} \left\{ E_{iu \rightarrow cz}(x_{iu}) + \hat{E}_{cz \rightarrow iu}(x_{iu}) \right\} \\
&\equiv \sum_{cz} \mathcal{E}_{cz} \left[\{ E_{iu \rightarrow cz} \}_{iu \in \partial cz} \right] + \sum_{iu} \mathcal{E}_{iu} \left[\{ \hat{E}_{cz \rightarrow iu} \}_{cz \in \partial iu} \right] \\
&- \sum_{\langle cz, iu \rangle} \mathcal{E}_{cz, iu} \left[\{ E_{iu \rightarrow cz}, \hat{E}_{cz \rightarrow iu} \} \right]. \tag{9.98}
\end{aligned}$$

We assume that the heuristic low temperature picture carries over to the zero temperature case. In this context pure states become clusters (in Hamming space) of minimizers of the Hamiltonian. Each cluster is characterized by a set of messages $\{ E_{iu \rightarrow cz}^{(p)}(\cdot), E_{cz \rightarrow iu}^{(p)}(\cdot) \}$. At zero temperature, two minimizers

belonging to the same cluster can be connected by successive flips with infinitesimal energy cost, while for two minimizers belonging to different clusters this is not possible.

Now we wish to compute the complexity (9.16) which counts the number of clusters. For this we introduce a generating function

$$\Xi(y) = \sum_p e^{-y\mathcal{E}}\{E_{iu \rightarrow cz}^{(p)}, E_{cz \rightarrow iu}^{(p)}\}. \quad (9.99)$$

When $y \rightarrow +\infty$ the sum is dominated by solutions of the min-sum equations with minimal Bethe energy. This object can be viewed as a partition function for the effective Hamiltonian (9.98) at inverse “temperature” y (the so-called Parisi parameter). Now, if we take α in the SAT phase the minimum Bethe energy vanishes and the complexity (9.16) is given by

$$\Sigma_{L,w}(\alpha) = \lim_{y \rightarrow +\infty} \lim_{N \rightarrow +\infty} \frac{1}{NL} \ln \Xi(y). \quad (9.100)$$

A negative complexity signals that there are no zero energy states and that the system is in an UNSAT phase. When this happens one has to generalize these formulas to allow for an energy dependent complexity (obtained by the Legendre transform of $\ln \Xi(y)$) but this aspect will not concern us here. For CSP’s it can be shown that the min-sum messages take discrete values in a finite alphabet. Therefore we have

$$\begin{aligned} \Xi(y) = & \sum_{\{E_{iu \rightarrow cz}, \hat{E}_{cz \rightarrow iu}\}} \left\{ \prod_{\langle iu, cz \rangle} e^{+y\mathcal{E}_{cz, iu}} \right\} \prod_{iu} \left\{ e^{-y\mathcal{E}_{iu}} \prod_{cz \in \partial(iu)} \mathbb{1}(E_{iu \rightarrow cz} = \mathcal{G}_{iu \rightarrow cz}) \right\} \\ & \times \prod_{cz} \left\{ e^{-y\mathcal{E}_{cz}} \prod_{iu \in \partial(cz)} \mathbb{1}(\hat{E}_{cz \rightarrow iu} = \hat{\mathcal{G}}_{cz \rightarrow iu}) \right\}. \end{aligned} \quad (9.101)$$

The arguments of the functionals $\mathcal{E}_{iu}[-]$, $\mathcal{E}_{cz}[-]$, $\mathcal{E}_{iu, cz}[-]$ and $\mathcal{G}_{iu \rightarrow cz}[-]$, $\hat{\mathcal{G}}_{cz \rightarrow iu}[-]$ are the messages $\{E_{iu \rightarrow cz}(\cdot), \hat{E}_{cz \rightarrow iu}(\cdot)\}$; they are not explicitly written to ease the notation. It can easily be seen that this is the partition function of a new graphical model which is still sparse. Edges $\langle (c, z), (i, u) \rangle$ now correspond to degree two “constraint” nodes, and nodes (c, z) and (i, u) now correspond to “variable” nodes. Therefore (9.100) can be computed from the Bethe approximation for this new model. The underlying assumption here is that the new effective model has a unique “pure state” with fast decaying correlations. This is called the level-1 cavity method. If this assumption breaks down, one should repeat the whole scheme, obtaining a level-2 cavity method (and so on). At level-1, the Bethe approximation can be expressed in terms of new beliefs - called *surveys* - $Q_{iu \rightarrow cz}(E_{iu \rightarrow cz}(\cdot))$ and $\hat{Q}_{cz \rightarrow iu}(\hat{E}_{cz \rightarrow iu}(\cdot))$ that count the *fraction of clusters* p for which $E_{iu \rightarrow cz}^{(p)}(\cdot) = E_{iu \rightarrow cz}(\cdot)$ and $E_{cz \rightarrow iu}^{(p)}(\cdot) = E_{cz \rightarrow iu}(\cdot)$. Note that these are the messages on the induced graph obtained by eliminating the degree two constraint nodes of the new model. We have

$$\ln \Xi(y) = \sum_{cz} \ln \left\{ \sum_{\{E_{iu \rightarrow cz}\}_{iu \in \partial(cz)}} e^{-y\mathcal{E}_{cz}} \prod_{iu \in \partial cz} Q_{iu \rightarrow cz} \right\}$$

$$\begin{aligned}
& + \sum_{iu} \ln \left\{ \sum_{\{\hat{E}_{cz \rightarrow iu}\}_{cz \in \partial(iu)}} e^{-y \mathcal{E}_{iu}} \prod_{cz \in \partial iu} Q_{cz \rightarrow iu} \right\} \\
& - \sum_{cz, iu} \ln \left\{ \sum_{E_{iu \rightarrow cz}, \hat{E}_{cz \rightarrow iu}} e^{-y \mathcal{E}_{iu, cz}} Q_{iu \rightarrow cz} \hat{Q}_{cz \rightarrow iu} \right\}. \quad (9.102)
\end{aligned}$$

The messages satisfy the *survey propagation equations*

$$\begin{aligned}
Q_{iu \rightarrow cz}(E_{iu \rightarrow cz}) \cong & \sum_{\{\hat{E}_{bv \rightarrow iu}\}_{cz \in \partial(iu)}} \mathbb{1}(E_{iu \rightarrow cz} = \mathcal{G}_{iu \rightarrow cz}) e^{-y C_{iu \rightarrow cz}} \\
& \times \prod_{bv \in \partial(iu) \setminus cz} Q_{bv \rightarrow iu}(\hat{E}_{bv \rightarrow iu}), \quad (9.103)
\end{aligned}$$

$$\begin{aligned}
\hat{Q}_{cz \rightarrow iu}(\hat{E}_{cz \rightarrow iu}) \cong & \sum_{\{\hat{E}_{jv \rightarrow cz}\}_{jv \in \partial(cz)}} \mathbb{1}(\hat{E}_{cz \rightarrow iu} = \hat{\mathcal{G}}_{cz \rightarrow iu}) e^{-y \hat{C}_{cz \rightarrow iu}} \\
& \times \prod_{jv \in \partial(cz) \setminus iu} Q_{jv \rightarrow cz}(E_{jv \rightarrow cz}), \quad (9.104)
\end{aligned}$$

where again \cong means that the right hand side has to be normalized.

In the SAT phase one takes $y \rightarrow +\infty$ in order to compute the complexity. This has the effect of reducing the sums in (9.103), (9.104) and (9.102), to surveys such that $C_{iu \rightarrow cz} = \hat{C}_{cz \rightarrow iu} = 0$ and $\mathcal{E}_{cz} = \mathcal{E}_{iu} = \mathcal{E}_{iu, cz} = 0$.

Algorithmic Implications

10

10.1 Problem Formulation

In this chapter¹ we investigate the performance of some well-known algorithms on the coupled K -SAT ensemble². The ideas here can be naturally extended to other satisfiability problems such as Q -COL or K -XORSAT³ for which we remove the details and only mention the final results.

Let us first explain one important consequence of analyzing various algorithms on the coupled ensemble. For the individual CSP ensembles, one main direction of research is to devise algorithms for finding satisfiability assignments. Typically, for an algorithm, there exists a specific “threshold”, denoted by α_{alg} , with the following property. If we pick a random formula from the CSP ensemble with clause density less than α_{alg} , then with a positive probability the algorithm is capable of finding a solution to that formula. This then implies $\alpha_{\text{alg}} \leq \alpha_s$, where by α_s we mean the static threshold (SAT/UNSAT threshold) of the individual CSP ensemble. Consider now the coupled ensemble, with a proper extension of the same algorithm. Call $\alpha_{\text{alg},L,w}$ the algorithmic threshold for finding a satisfying assignment with positive probability, and set $\alpha_{\text{alg},w} = \lim_{L \rightarrow +\infty} \alpha_{\text{alg},L,w}$. From Theorem 9.3 we know that the coupled ensemble has the same static threshold as the individual one, when $L \rightarrow +\infty$ and w is fixed. Therefore, one certainly has the lower bound $\alpha_{\text{alg},w} \leq \alpha_s$. The point here is that for well chosen algorithms an improvement of the bound may occur, namely $\alpha_{\text{alg}} < \alpha_{\text{alg},w} \leq \alpha_s$, and one would expect to get the best lower bounds by increasing the coupling width w . A well chosen algorithm is one that shows a “threshold improvement” or even the full saturation phenomenon.

¹The material of this chapter is based on the work of [125].

²The individual and coupled K -SAT ensembles are described in detail in Chapter 7.

³A detailed description of these ensembles is given in Chapter 9.

The outline of this chapter is as follows. In Section 10.2 we consider the simplest of such algorithms, namely the class of peeling algorithms and show that their threshold improves by spatial coupling. For the K -XORSAT ensemble, we observe that the coupled thresholds saturate to the static threshold. However, for the K -SAT and Q -COL ensembles, although the amount of threshold improvement is significant, due to the extreme simplicity of these peeling algorithms, the coupled thresholds are still far away from the satisfiability threshold. As a result, we consider in Section 10.3 some slightly more sophisticated algorithms. That is, we consider algorithms that involve decimation, i.e., setting the variables one-by-one according to some heuristic and reducing the graph. We show that for the simplest choice of decimation algorithms (i.e., the unit clause propagation algorithm), the coupled thresholds improve significantly and for large values of K they even reach a constant fraction ($\frac{1}{2\ln 2}$) of the satisfiability threshold.

10.2 Peeling Algorithms and Coupled Scalar Recursions

Below we illustrate a simple peeling-type algorithm applied to K -SAT, which in the literature is called the pure literal algorithm. We then discuss briefly similar algorithms for Q -COL and K -XORSAT. As we will see in the sequel, such peeling algorithms can be cast into the framework of one-dimensional coupled recursions for which a recent elegant characterization has been provided in [126] and [127].

10.2.1 Pure Literal: A Peeling Algorithm for K -SAT

We begin by a brief explanation of the algorithm. Let G be a K -SAT formula. The algorithm starts with G and in each step shortens G until we either reach the empty graph or we cannot make any further shortening. Assume now that there exists a literal (variable) i in G such that all of its incoming edges have the same sign. This literal is called a *pure literal*. One can choose a value for x_i that satisfies all of its neighboring clauses and clearly that is the optimum choice to fix the variable i in order to find a SAT assignment for G . Hence, without loss of generality, we can remove i and its neighboring clauses from G and search for a SAT assignment on the graph $G \setminus i$. In other words, finding a SAT assignment for G is equivalent to finding an assignment for $G \setminus i$. As a result, we can peel the literal i and its neighbors from G and look for new pure literals on $G \setminus i$. We continue this process until the final graph (the 2-core) has no more pure literals. If the final graph is empty then the algorithm succeeds; otherwise, it fails. This algorithm determines the 2-core of a graph G and has been analyzed by the method of differential equations [114]. Here we discuss the algorithm from the message passing point of view.

Consider the following message passing (MP) rule. As we see later, this MP rule is equivalent to the pure literal algorithm. At time $t \in \{1, 2, \dots\}$, assign to each edge $\langle i, c \rangle \in E$ two messages $\mu_{i \rightarrow c}^t$ and $\mu_{c \rightarrow i}^t$. The messages $\mu_{i \rightarrow c}^t$ represent

the messages going from literals to clauses at time t and the messages $\mu_{c \rightarrow i}^t$ are the messages from checks to variables at time t . The messages at time $t + 1$ are evolved from the ones at time t via the following procedure:

1. At time 0, initialize all the messages $\mu_{c \rightarrow i}^0$ and $\mu_{i \rightarrow c}^0$ to 0.
2. At time $t + 1$,

$$\begin{aligned}\mu_{c \rightarrow i}^{t+1} &= \mathbb{1}\left\{\sum_{j \in \partial c \setminus i} \mu_{j \rightarrow c}^t \geq 1\right\}, \\ \mu_{i \rightarrow c}^{t+1} &= \prod_{h \in \partial i \setminus c, \mu_{h \rightarrow i}^{t+1} = 0} \mathbb{1}\{J_{c,i} = J_{h,i}\}.\end{aligned}$$

Here, we recall that $J_{c,i}$ denotes the sign of the edge $\langle c, i \rangle$. The above message passing rule is equivalent to the pure literal algorithm in the following sense. When $\mu_{i \rightarrow c}^t = 1$ for at least one $c \in \partial i$, then the vertex i would have been peeled by the algorithm some time before t and if $\mu_{i \rightarrow c}^t = 0$, the vertex i would not have been peeled by the algorithm up to time t . The same statement is valid for the clauses in a way that if $\mu_{c \rightarrow i}^t = 1$ for at least one $i \in \partial c$, then the clause c would have been peeled at some time before t in the pure literal algorithm.

Define $p^t = \mathbb{P}(\mu_{c \rightarrow i}^t = 0)$ and $q^t = \mathbb{P}(\mu_{i \rightarrow c}^t = 0)$. Note that $p^0 = 1$. We derive the density evolution equations that relates p^{t+1} to p^t . Let G be randomly chosen from $\text{SAT}(N, K, \alpha)$ with N very large. Fix an edge $\langle c, i \rangle$. Observe that $\mu_{c \rightarrow i}^{t+1} = 0$ if and only if all the incoming messages to the clause c , other than the one of $\langle c, i \rangle$, take value 0. Hence, we can write

$$p^{t+1} = (q^t)^{K-1}. \quad (10.1)$$

Relating q^{t+1} to p^t is slightly more subtle. Observe that $\mu_{i \rightarrow c}^{t+1} = 1$ if and only if the sign of $\langle c, i \rangle$ is equal to the sign of every edge $\langle h, i \rangle$ such that $h \in \partial i \setminus c$ and $\mu_{h \rightarrow i}^{t+1} = 0$. Moreover, the probability that an edge $\langle c, i \rangle$ is incident to a variable i of degree d is $\frac{e^{-\alpha K} (\alpha K)^{d-1}}{(d-1)!}$. One can then write

$$\begin{aligned}1 - q^{t+1} &= \sum_{d=1}^{\infty} \frac{e^{-\alpha K} (\alpha K)^{d-1}}{(d-1)!} \sum_{j=0}^{d-1} \binom{d-1}{j} (p^{t+1})^j (1 - p^{t+1})^{d-1-j} 2^{-j} \\ &= \sum_{d=1}^{\infty} \frac{e^{-\alpha K} (\alpha K)^{d-1}}{(d-1)!} \left(1 - \frac{p^{t+1}}{2}\right)^{d-1} \\ &= \exp\left(-\frac{\alpha K}{2} p^{t+1}\right).\end{aligned}$$

Hence, from the above two relations we obtain that

$$p^{t+1} = \left(1 - \exp\left(-\frac{\alpha K}{2} p^t\right)\right)^{K-1}, \quad (10.2)$$

with $p^0 = 1$. It is more convenient to do a change of variables in the form of $x^t = \alpha p^t$, which transforms (10.2) to

$$x^{t+1} = \alpha(1 - \exp(-\frac{K}{2}x^t))^{K-1}. \quad (10.3)$$

For the pure literal algorithm to succeed, the value of x^t should tend to 0 and t increases. Now, from the recursion (10.3) and fact that $x_0 = 1$, it is easy to deduce that for t growing large, x^t tends to 0 if and only if the equation

$$x = \alpha(1 - \exp(-\frac{K}{2}x))^{K-1}. \quad (10.4)$$

has only one solution which is the trivial solution $x = 0$ on $[0, 1]$. The net result is that the pure literal rule succeeds w.h.p for $\alpha < \alpha_{\text{pl}}(K)$ such that (10.4) has a unique fixed point $x = 0$ in the unit interval $[0, 1]$. Mathematically we can define α_{pl} as

$$\alpha_{\text{pl}}(K) = \sup\{\alpha \geq 0 \mid x - \alpha(1 - \exp(-\frac{K}{2}x))^{K-1} > 0 \quad \forall x \in (0, 1]\}. \quad (10.5)$$

We now consider the coupled ensemble. The way the pure literal algorithm works on a coupled formula is similar to what we explained above and hence needs no further explanation. In order to analyze the pure literal rule we can think of extending the chain to \mathbb{Z} with “pure” variable nodes for $z < 0$ and $z > L + w - 2$. The peeling of constraints attached to pure nodes will propagate inside the chain as long as α is below the critical threshold. A similar message passing analysis as above yields a set of one-dimensional coupled recursions

$$x_z^{t+1} = \alpha \left\{ \frac{1}{w} \sum_{l=0}^{w-1} (1 - \exp(-\frac{K}{2w} \sum_{k=0}^{w-1} x_{z+k-l}^t)) \right\}^{K-1}, \quad (10.6)$$

with boundary condition $x_z^t = 0$ for z at the boundaries. This recursion results in the one-dimensional fixed point equations

$$x_z = \alpha \left\{ \frac{1}{w} \sum_{l=0}^{w-1} (1 - \exp(-\frac{K}{2w} \sum_{k=0}^{w-1} x_{z+k-l})) \right\}^{K-1}. \quad (10.7)$$

One can define $\alpha_{\text{pl},L,w}(K)$ as the largest value of α such that (10.7) has only one fixed point profile to be the all-zero profile.

Table 10.1 contains the numerical prediction of $\alpha_{\text{pl},L,w}(K)$ for $L = 80$ and $w = 5$ and different values of K . As we observe from Table 10.1 there is an improvement of the coupled threshold over the individual ensemble. For example for $K = 3$ we have $\alpha_{\text{pl}} \approx 1.636 < \alpha_{\text{pl},w=5,L=80} \approx 1.835 < \alpha_s \approx 4.26$, a modest improvement. As we will see in the sequel, the coupled thresholds $\alpha_{\text{pl},L,w}(K)$ when $L, w \rightarrow \infty$ can be precisely and analytically computed. We postpone further arguments on the amount of the improvement of the coupled thresholds to Section 10.2.3.

K	3	4	5	7
$\alpha_{\text{pl}}(K)$	1.626	1.544	1.402	1.190
$\alpha_{\text{pl},L=80,w=5}(K)$	1.834	1.954	1.986	1.998

Table 10.1: *First line:* Pure threshold for the uncoupled case. *Second line:* Pure literal threshold for a coupled chain with $w = 5$, $L = 80$. By halving these numbers, we obtain the corresponding thresholds of the leaf removal algorithm devised for the K -XORSAT problem.

10.2.2 Peeling Algorithms for Q -COL and K -XORSAT

We discuss now a similar peeling algorithm for Q -COL⁴. We start with a graph G to color with a given set of Q colors. Assume there exists a node i in G that has degree less than Q . Clearly, if we can color the graph $G \setminus i$ with Q colors, then G can also be colored with Q colors. Hence, finding a Q -coloring for G is equivalent to finding a Q -coloring for $G \setminus i$. As a result, we can peel the node i from G and continue this process until the final graph has no more nodes of degree less than Q . If the final graph is empty then the algorithm succeeds; and otherwise it fails. Such a peeling algorithm can be analyzed in the same way as above. In particular, let us define $y = cx$, where x is the fraction of nodes in the final residual graph and c is the average vertex degree. We find

$$y = cG(y), \quad (10.8)$$

where the function G is given as

$$G(y) = 1 - e^{-y} \sum_{j=0}^{Q-2} \frac{y^j}{j!}. \quad (10.9)$$

For $c < c_p$ there is a unique trivial fixed point $y = 0$ and the algorithm succeeds. Non trivial fixed points appear for $c > c_p$ which is the threshold for the emergence of a Q -core. Table 10.2 contains the numerical values of c_p for several values of Q .

We now take coupled instances from the ensemble. We can write down the density evolution equations and the corresponding one-dimensional fixed point equations. Not surprisingly, similar calculations show that the message passing algorithm is controlled by the one-dimensional fixed point equation,

$$y_z = cG\left(\frac{1}{2w-1} \sum_{k=-w+1}^{w-1} y_{z+k}\right). \quad (10.10)$$

where $y_z = cx_z$ and x_z is the fraction of remaining nodes at position z . Table 10.2 contains the numerical values of $c_{p,w=5,L=80}$ for several values of Q , and shows the threshold improvement.

⁴The Q -COL model and the K -XORSAT models are introduced in Section 9.2.3.

Q	3	4	5	7
c_s	4.69	8.90	13.69	24.46
c_p	3.35	5.14	6.79	9.87
$c_{p,L=80,w=5}$	3.58	5.74	7.84	11.92

Table 10.2: *First line:* static phase transition threshold for Q -COL. *Second line:* peeling algorithm threshold for the uncoupled case. *Third line:* peeling algorithm threshold for a coupled chain with $w = 5$, $L = 80$.

The peeling algorithm for the K -XORSAT problem is known as the “leaf removal” algorithm. As long as there is a leaf variable node, remove it and remove the attached constraint node with its emanating edges. If this process ends with an empty graph the instance is satisfiable. It is known that this algorithm is equivalent to BP message passing, and the density evolution analysis leads to the fixed point equation

$$x = (1 - \exp(-\alpha K x))^{K-1}. \quad (10.11)$$

Here, x is interpreted as the probability (when the number of iterations goes to infinity) that a constraint node is not removed. There is a threshold α_{lr} above which (10.11) has non-trivial fixed points (i.e, the fraction of remaining variables is positive), so we get a lower bound $\alpha_{lr} < \alpha_s$. For the coupled ensemble the density evolution analysis yields the one-dimensional fixed point equations

$$x_z = \left\{ \frac{1}{w} \sum_{l=0}^{w-1} (1 - \exp(-\frac{\alpha K}{w} \sum_{k=0}^{w-1} x_{z+k-l})) \right\}^{K-1}, \quad (10.12)$$

Note here that these fixed-point equations are equal to the (10.4) and (10.7) with the replacement $\alpha \rightarrow 2\alpha$. As a result, by halving the numbers in Table 10.1 we obtain the corresponding thresholds for the leaf-removal algorithm.

10.2.3 The Framework of Coupled Scalar Recursions

One-dimensional coupled recursions such as (10.6) have recently been fully characterized in [126] and [127]. Let us now briefly mention the main results in this regard.

A scalar recursion as in (10.3) can be written in the general form of

$$x^{t+1} = f(g(x^t; \alpha)), \quad (10.13)$$

where $f : [0, 1] \times \mathbb{R} \rightarrow [0, 1]$ is strictly increasing in both arguments for $x, \alpha > 0$ and $g : [0, 1] \rightarrow [0, 1]$ satisfies $g'(x) > 0$ for $x \in [0, 1]$. Such a recursion with the above mentioned properties for f and g is called a *scalar admissible system*. For example, in the recursion (10.3) we have $f(x; \alpha) = \alpha(1+x)^{K-1}$ and $g(x) = -\exp(-\frac{K}{2}x)$. Hence (10.3) is a scalar admissible system. The

threshold of such scalar system, α_{sys} , is defined as the largest value of α such that the equation $x = f(g(x; \alpha))$ has the unique $x = 0$ fixed point for $x \in [0, 1]$.

Let us now consider the coupled system of scalar recursions. Consider the position set $\mathcal{L} = \{0, 1, \dots, L + w - 1\}$ for which we assign variable $x_z^t \in [0, 1]$ to $z \in \mathcal{L}$ for time parameter $t \in \{0, 1, \dots\}$. Also, let us define $x_z^t = 0$ for $z \notin \mathcal{L}$ and all times $t \geq 0$. For the coupled system we have the recursions

$$x_z^{t+1} = \frac{1}{w} \sum_{l=0}^{w-1} f\left(\frac{1}{w} \sum_{k=0}^{w-1} g(x_{z+k-l}^t); \alpha\right), \quad (10.14)$$

for $z \in \mathcal{L}$. The threshold of the coupled system $\alpha_{\text{sys}, L, w}$ is then defined as the largest α for which the coupled system of equations

$$x_z = \frac{1}{w} \sum_{l=0}^{w-1} f\left(\frac{1}{w} \sum_{k=0}^{w-1} g(x_{z+k-l}); \alpha\right), \quad (10.15)$$

has a unique trivial fixed point, which is the trivial all-zero fixed point.

The limit of $\alpha_{\text{sys}, L, w}$ when $L, w \rightarrow \infty$ can be computed from the so called *potential function*, $\phi(x, \alpha)$, that is associated to the scalar admissible system, and is defined as

$$\phi(x, \alpha) \triangleq xg(x) - G(x) - F(g(x), \alpha), \quad (10.16)$$

where $F(x, \alpha) = \int_0^x f(z; \alpha) dz$ and $G(x) = \int_0^x g(z) dz$.

Theorem 10.1 ([126]). *We have*

$$\lim_{w \rightarrow \infty} \lim_{L \rightarrow \infty} \alpha_{\text{sys}, L, w} = \sup\{\alpha \geq 0 \mid \min_{x \in [0, 1]} \phi(x, \alpha) \geq 0\}. \quad (10.17)$$

Theorem 10.1 provides a practical way to analytically compute the coupled threshold. It is easy to numerically check that the results of Tables 10.1 and 10.2 are very close (up to the fourth decimal) to the numbers obtained from Theorem 10.1. However, we do not expect to get exactly the same numbers as (10.17). This is because the thresholds in these tables are found for finite choices of L and w .

From (10.17), we can find the asymptotic value of the thresholds when K is a very large number. For the K -SAT problem, when $K \rightarrow +\infty$ we find⁵

$$\alpha_{\text{pl}}(K) \doteq \frac{2 \ln K}{K} \quad \text{but} \quad \alpha_{\text{pl}, L, w}(K) \rightarrow 2 \quad \text{as} \quad L \gg w \rightarrow +\infty. \quad (10.18)$$

Thus, the pure literal threshold $\alpha_{\text{pl}}(K)$ is “infinitely improved” by coupling. Of course this is far away from the satisfiability threshold $\alpha_s \doteq 2^K \ln 2$.

For the problem of Q -COL we obtain from (10.17)

$$c_p(Q) \doteq Q \quad \text{but} \quad c_{p, L, w}(Q) \doteq 2Q \quad \text{as} \quad L \gg w \rightarrow +\infty. \quad (10.19)$$

⁵For two sequences $\{a_n\}$ and $\{b_n\}$, we say $a_n \doteq b_n$ if $\frac{a_n}{b_n} \rightarrow 1$ as $n \rightarrow \infty$.

This has to be compared with $c_s(Q) \doteq 2Q \ln Q$.

The leaf removal thresholds for K -XORSAT individual and coupled ensembles, $\alpha_{lr}(K)$, $\alpha_{lr,L,w}(K)$, are obtained just by halving the K -SAT pure literal thresholds. Interestingly enough, it can be shown that the coupled threshold of the leaf removal algorithm is precisely equal to the SAT/UNSAT threshold for the K -XORSAT problem. That is, $\lim_{w \rightarrow \infty} \lim_{L \rightarrow \infty} \alpha_{lr,L,w}(K) = \alpha_s(K)$. Hence, we have a full threshold saturation for the K -XORSAT problem.

For the problems of K -SAT and Q -COL, although the coupled thresholds improve significantly, they are still far below the static threshold. Hence, we need to consider slightly more sophisticated algorithms. We proceed by entering the realm of decimation algorithms, the simplest of which are the “unit constraint propagation” algorithms. For the K -SAT problem, such type of algorithm is called the *unit clause propagation* algorithm. For the rest of this chapter, our main focus is on the unit clause propagation algorithm for the coupled ensemble. We also mention the final results for the same type of algorithm for Q -COL.

10.3 Unit Clause Propagation

10.3.1 Individual Ensemble

The Unit Clause propagation algorithm, or UC for short, is a (randomized) algorithm which goes through variables one at a time, sets them permanently, and simplifies the formula as it goes along. This algorithm is like a DPLL algorithm but only explores one branch of the search tree. In brief, the algorithm works as follows: Consider a K -SAT formula which we represent by a bipartite graph G consisting of N literals or variable nodes and $M = N\alpha$ clauses or check nodes. The algorithm starts with G and in each step removes several nodes from the graph. The UC algorithm consists of two main steps:

- *Free step*: which is when all the check nodes have degree at least 2. This is the situation where the algorithm is free to do whatever it wants. However, UC does the simplest possible action. It chooses a variable uniformly at random among the currently unset variables and sets it permanently to 0 or 1 again with uniform probability.
- *Forced Step*: which is when we have a unit clause (clause with only one remaining edge). In this situation, we better try to satisfy this clause before it is too late. So in some sense we are forced to fix the variable connected to the unit clause.

Once a variable is set, it is removed from the graph together with the clauses that are satisfied by the variable. Also, there might be some clauses, connected to the variable, that are shortened. This is due to the fact that the assigned value of the variable did not satisfy these clauses. Hence, removing the variable from the graph will cause these yet unsatisfied clauses to have a less degree. A detailed description of the UC algorithm is given in Algorithm 6.

Procedure 6 Unit Clause Propagation Algorithm

-
- 1: Start with a given K -SAT formula G .
 - 2: Repeat until all the variables are set.
 - 3: If G contains unit clauses (forced step), then choose one at random and satisfy it by setting its left variable. Remove or shorten clauses containing this variable.
 - 4: If there are no unit clauses (free step), then choose one variable at random from the unset ones and set it by flipping a coin. Remove or shorten clauses containing this variable.
-

On the analysis side, the progress of UC can be modeled with differential equations. The net result is that for $\alpha < \alpha_{\text{UC}}$ with α_{UC} given as

$$\alpha_{\text{UC}} = \frac{1}{2} \left(\frac{K-1}{K-2} \right)^{K-2} \frac{2^K}{K} \doteq \frac{e}{2} \frac{2^K}{K}, \quad (10.20)$$

the UC algorithm finds an assignment that satisfies all but $o(n)$ number of clauses. It can also be shown that for densities below α_{UC} , with positive probability the output assignment satisfies all the clauses.

10.3.2 Description of UC Algorithm for the Coupled Formulas

Let us now focus on the UC algorithm for the coupled formulas. As for the un-coupled case, the UC algorithm consists of two main steps: free and forced. The operation of the algorithm at a forced step is clear: remove all the unit-clauses until no further unit-clause exists. However, at a free step, depending on how we might want to use the chain structure of the formula, we can have different *schedules* for choosing a free variable. As we will see now, for a coupled formula, the schedule within which we are choosing a variable in a free step is very important⁶.

Consider for instance the following naive schedule. At a free step, pick a variable uniformly at random from all the remaining variables and fix it by flipping a coin. Computer experiments indicate that this naive schedule gains no threshold improvement over the un-coupled ensemble. This is not surprising since this schedule does not seem to exploit the spatial (chain) structure of the formula and in some ways it greatly resembles the UC algorithm for the un-coupled ensemble. Hence, in order for the UC algorithm to have a threshold improvement over the coupled ensemble, we need to come up with schedules that exploit to some extent the additional spatial structure of the formula. We proceed by illustrating one such successful schedule.

In the very beginning of the algorithm, all the check nodes have degree K and there are no unit clauses. Hence, we are free to fix the variables in the first few steps of the algorithm. If we fix the variables from the left-most position

⁶For the peeling algorithms mentioned in Section 10.2, the performance of the algorithm is independent of the schedule of peeling steps.

(i.e., the boundary) we are somehow creating a seed at the boundary of the chain. Continuing this action at the free steps, we will eventually create unit clauses and at these forced steps a natural choice is just to clear all the unit clauses as long as they exist. However, when we are confronted with a free step, we will again try to help this seed to grow inside the chain. This can be done again by fixing a variable from the remaining left-most position. Consequently, the schedule that we apply is as follows.

- At a *free step*, pick a variable randomly from the left-most position and fix it permanently by flipping a fair coin.
- At a *forced step*, we get rid of unit clauses as long as they exist.

Computer experiments show that this schedule indeed exhibits a threshold improvement over the un-coupled ensemble. E.g., for the coupled 3-SAT problem, experiments suggest that the threshold of the UC algorithm is around 3.67. This is a significant improvement compared to the threshold of UC for the un-coupled ensemble which is $\frac{8}{3}$. Of course, one cannot be certain about these numbers until they are confirmed with analytic methods. The right tool to analyze the dynamics of the UC algorithm for the un-coupled ensemble is the method of differential equations (see [128]). The rest of this chapter focuses on writing the differential equations for the UC algorithm on the coupled ensemble and then analyzing these equations. More specifically, in the next section, we work out these differential equations in detail. Later in the subsequent sections, we simplify these equations and provide a general framework to analyze them. Using this framework, we analytically obtain the threshold of the UC algorithm on the coupled ensembles.

10.3.3 Analysis of the Evolution of UC via Differential Equations

Phases, Types, and Rounds

For the coupled ensemble, the analysis of the evolution of UC is much more involved than the un-coupled ensemble. This is because of the fact that the schedule we have used prefers the left-most variable position in a free step. Hence, the number of variables in different positions will evolve differently. As an example, one can easily see that during the algorithm, the first position that all its variables are set is the left-most position (i.e., position 0). After the evacuation of position 0, position 1 becomes the left-most position of the graph and hence, the second position that becomes empty of variables is position 1. Continuing in this manner, the last position that is evacuated is position $L + w - 2$. With these considerations, we consider $L + w - 1$ *phases* for this algorithm (see Figure 10.1). At phase $p \in \{0, 1, \dots, L + w - 2\}$, all the variables at positions prior to p have been set permanently and as a result, at a free step we will pick a variable from position p .

This statistical asymmetry in the number of variables at each position also affects the the behavior of the number of check nodes in each position. As a

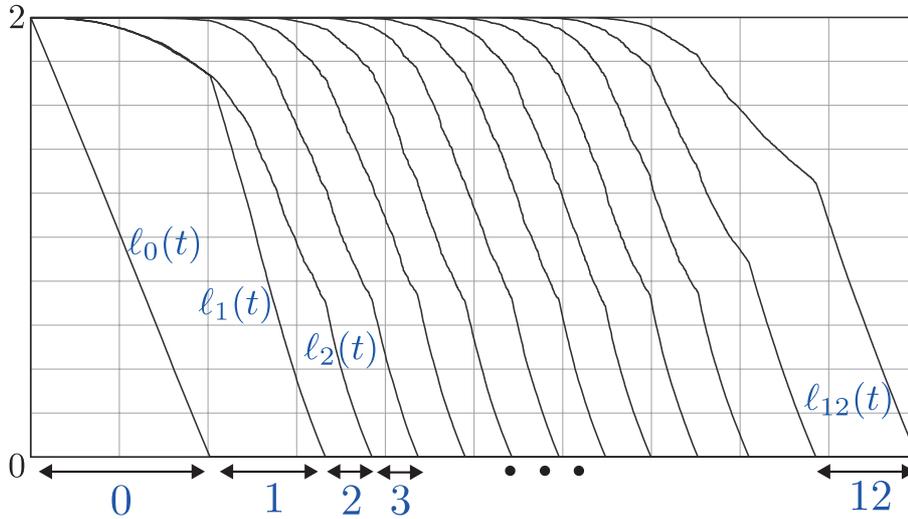


Figure 10.1: A schematic representation of how the literals at each of the positions vary in time. The horizontal axis corresponds to time t which is the number of free steps. Here we have $L = 11$ and $w = 3$. This plot corresponds to an implementation of the UC algorithm on a random coupled instance. The blue numbers below the plot are the phases of the algorithm. In the beginning of the algorithm, we are in phase 0. This phase lasts until all the literals in the first position are peeled off and as a result $\ell_0(t)$ reaches 0. We then go immediately to phase 1 and this phase lasts till $\ell_1(t)$ reaches 0 and so on. We have in total $L + w - 1 = 13$ phases.

result, we consider *types* for the check nodes. For instance, consider a degree two check node. It is easy to see that the probability that this degree two check node is hit (removed or shortened) is greatly dependent on the position of variables that it is connected to. This means that, dependent on the variable positions to which they are connected, we have different types of degree two check nodes. Clearly, the same statement holds for clauses of degree three, four, etc.

Let us now formally define the ingredients needed for the analysis. The notation we use here is slightly hard to swallow immediately. Thus, for the sake of maximum clarity, we try to uncover the details as smoothly as possible. We consider *rounds* for this algorithm. Each round consists of one free step followed by the forced steps that follow it. More precisely, at the beginning of each round we perform a free step and then we clear out all the unit-clauses as long as they exist (forced steps). We let time t be the number of rounds passed so far. This time variable will be called *round time*. The relation between t and the *natural time* (the total number of permanent fixes) is not linear. We also let $L_i(t)$ be the *number of literals* left in variable position $i \in \{0, 1, \dots, L + w - 2\}$.

We now define the check types. Consider a coupled K -SAT formula to begin with. For such a formula there are L sets of check nodes placed at positions $\{0, 1, \dots, L\}$. Let us consider a specific position $i \in \{0, 1, \dots, L\}$ and look at the check nodes at position i . Each of these check nodes can potentially be connected to any set of K variables resting in variable positions $\{i, i + 1, \dots, i + w - 1\}$. Some thought shows that there are various types of check nodes depending on the variable positions that they are connected to. For example, there is a type of check nodes for which all of the K edges go only into a single variable position $j \in \{i, i + 1, \dots, i + w - 1\}$ or there is a type for which some of its edges go to position i and the rest go to position $i + 1$ and so on. Also, as we proceed through the UC algorithm, some of these checks are shortened to create new types of checks with degrees less than K . We now explain a natural way to encode these various types.

By $C(t, i, \underline{\tau})$ we mean the number of check nodes at check position $i \in \{0, 1, \dots, L\}$ that have type $\underline{\tau}$ at round time t . The type $\underline{\tau} = (\tau_0, \dots, \tau_{w-1})$ is a w -tuple and indicates that relative to position i , how many edges the check has in (variable) positions $i, i + 1, \dots, i + w - 1$. The best way to explain $\underline{\tau}$ is through an example. Let us assume $w = 4$ and consider the set of check nodes at check position 20 that are only connected to variable positions 20, 22, 23 in the following way. For each of these check nodes there are exactly two edges going to position 20, and 1 edge going to position 22 and 1 edge going to position 23 (thus each of these checks have degree 4). Figure 10.2 illustrates a generic check node of this set. We denote the number of these checks at time

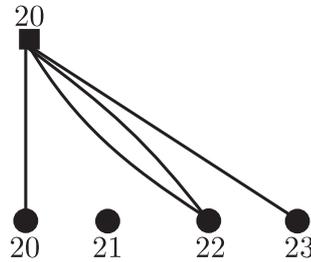


Figure 10.2: A schematic representation of checks which contribute to $C(t, 20, (1, 0, 2, 1))$. All the check nodes that contribute to $C(t, i, \underline{\tau})$, were initially (at time 0) degree K check nodes resting at check position i . However, the algorithm has evolved in a way that these check nodes have been deformed (possibly shortened or remained unchanged) to have a specific type $\underline{\tau}$.

t by $C(t, 20, (1, 0, 2, 1))$. In other words, the type is computed as follows: the check position number that the check rests in is 20. This check is connected to a variable at position 20, and 2 variables at position 22, and a variable at position 23. So, relative to the check position 20, we see the edge-tuple $(1, 0, 2, 1)$. Let us now repeat and generalize: By $C(t, i, \underline{\tau})$ we mean the number of check nodes, at time t , which rest in position i , and $\underline{\tau}$ is a w -tuple that indicates relative to

variable position i , the number of edges that go to positions $i, i+1, \dots, i+w-1$, respectively. One can easily see that by summing up elements of the w -tuple $\underline{\tau} = (\tau_0, \dots, \tau_{w-1})$, we find the degree of the corresponding check type. We denote the degree of a type $\underline{\tau}$ by $\deg(\underline{\tau})$. It is also easy to see that there are $\binom{d+w-2}{d-1}$ different types of degree d for $d \in \{2, 3, \dots, K\}$. We are now ready to write the differential equations. Our approach is as follows. Assume the phase of the algorithm is p and we are in a round t . At a free step, we fix a variable at position p (free step). This will create a number of forced steps in each of the positions $p, p+1, \dots, L+w-1$. We first compute the average of these forced fixes in each variable position as a function of the number of degree two check nodes. Using these averages, we then update the average number of check and variable nodes at each position. We proceed by explaining a key property for the analysis.

Uniform Randomness Property

The uniform randomness property means that at any round time t , for any position i and any type $\underline{\tau}$, each clause in the set $C(t, i, \underline{\tau})$ is uniformly distributed among all the possible clauses at position i with type $\underline{\tau}$. In other words, conditioned on the number of variables and check-types of different positions, the formula is uniformly random. An intuitive justification for the randomness property in our case stems from the fact that at any step (free or forced) in the UC algorithm, no information, what-so-ever, can be deduced about the structure of the remaining formula. The exact proof of the uniform randomness property in our case can be easily deduced from [128, Lemma 3].

Inside a Round

As mentioned above, a round begins with a free step and proceeds with a possible sequence of forced steps and ends when there are no more forced steps left. A crucial task in writing the differential equations is a precise characterization of the average number of forced steps taken during each round. Our objective is now to derive this average in a round t as a function of the number of degree two check nodes. In this regard, let us denote by $\beta_i(t)$ the average number of variables at position i that are set in round t . Let us also define the vector $\bar{\beta}(t)$ as

$$\bar{\beta}(t) = \begin{bmatrix} \beta_0(t) \\ \beta_1(t) \\ \vdots \\ \beta_{L+w-2}(t) \end{bmatrix}.$$

Also, it will be useful to choose a specific notation for the degree two types. For $r, s \in \{0, 1, \dots, w-1\}$ s.t. $r < s$, we define the degree two types $\pi_{r,s}$ and $\pi_{r,r}$ as the following w -tuples

$$\pi_{r,s} = (0, \dots, \overset{\text{position } r}{1}, \dots, \overset{\text{position } s}{1}, \dots, 0), \quad (10.21)$$

$$\pi_{r,r} = (0, \dots, \overset{\text{position } r}{2}, \dots, 0). \quad (10.22)$$

In other words, the tuple $\pi_{r,s}$ is zero at all its entries except the ones at positions r, s which it takes value one. Similarly, $\pi_{r,r}$ is non-zero only at position r where it takes value 2.

A critical point to consider here is the following. Consider two variable positions $i, j \in \{0, 1, \dots, L+w-2\}$. We ask ourselves if we set one variable at position j , how many immediate forced steps at position i would this create on average? We call this average number the *effect* of position j on i and denote it by $A_{i,j}$. To answer this, we should look at the degree two check nodes that are connected exactly to positions i and j . Let us now express $A_{i,j}$ in terms of the degree two check nodes that are connected to positions i and j . For simplicity, assume $i \leq j$. It is easy to see that if $j \geq i+w$, then $A_{i,j} = 0$. This is because each check node can only be connected to variable positions in a range of size at most w . Assuming $i+w > j$, we consider two cases: $i = j$ and $i < j$. When $i = j$, we should consider the degree two checks that are connected to position i twice. The possible positions of these check node lie inside the set $\{i-w+1, i-w+2, \dots, i\}$. For the checks at position $k \in \{i-w+1, i-w+2, \dots, i\}$ the corresponding type would be $\pi_{i-k, i-k}$. Hence, we obtain

$$A_{i,i} = \frac{1}{L_i(t)} \sum_{k=i-w+1}^i 2C(t, k, \pi_{i-k, i-k}). \quad (10.23)$$

In the case where $i < j$, we need the number of checks that have one edge in position i and one in position j . So the possible check positions are inside the set $\{j-w+1, \dots, i\}$. For a check position k in this set, the corresponding type would be $\pi_{i-k, j-k}$. As a result, we have

$$A_{i,j} = \frac{1}{L_j(t)} \sum_{k=j-w+1}^i C(t, k, \pi_{i-k, j-k}). \quad (10.24)$$

Note that

$$L_j(t)A_{i,j} = L_i(t)A_{j,i} = \sum_{k=j-w+1}^i C(t, k, \pi_{i-k, j-k}). \quad (10.25)$$

We now define the matrix

$$A = [A_{i,j}]_{(L+w-1) \times (L+w-1)}, \quad (10.26)$$

that contains as its entries $A_{i,j}$. The matrix A plays a key role in both writing the differential equations and their analysis. Assume now that we are in phase p . Having the matrix A , we can compute the vector $\bar{\beta}(t)$ via considering the multi-rate Galton-Watson tree starting at position p . We first note that in the beginning of each round we freely fix a variable at position p . This will create an effect (i.e., some forced fixes) in other positions. This effect is on average

equal to Ae_p , where e_p is the vector that has a 1 in its p -th position and 0 in other positions. These new (forced) fixes will also create an effect which is on average equal to A^2e_p and so on. Therefore, we obtain

$$\bar{\beta}(t) = (I + A + A^2 + \cdots)e_p = (I - A)^{-1}e_p. \quad (10.27)$$

Of course the relation (10.27) is valid if and only if the matrix A has spectral radius strictly less than 1. More precisely, we define the spectral radius of A as

$$\rho(A) = \max_{1 \leq i \leq L-w+1} |\lambda_i|,$$

where $|\lambda_i|$ denotes the absolute value of the eigenvalue λ_i of A . For (10.27) to hold, we must have

$$\rho(A) < 1. \quad (10.28)$$

Here, a few comments are in order:

- (i) We have assumed that during each round, the statistics of the formula remain constant. This condition does not completely hold, as by setting the variables the number of variables and clauses change. However, as we see in the following, by assuming $\rho(A) < 1$, the fluctuations of these statistics is $O(1)$ and when divided by L_i , their total influence would be $O(\frac{1}{N})$. As a result, they can be neglected with respect to the Wormald framework of differential equations. We thus omit such an additional factor in (10.27).
- (ii) We notice from (10.25) that the matrix A can be written in the form of $A = SL$, where S is a symmetric matrix and L is a diagonal matrix. Hence, the matrix $L^{+\frac{1}{2}}AL^{-\frac{1}{2}}$ is a symmetric matrix and hence has only real eigenvalues. However, it can easily be shown that A and $L^{+\frac{1}{2}}AL^{-\frac{1}{2}}$ have the same set of eigenvalues and thus all the eigenvalues of A are real. Further with such a representation of A , one can deduce from the Perron-Frobenius formalism [129] that

$$\rho(A) = \max_{1 \leq i \leq L-w+1} \lambda_i. \quad (10.29)$$

Consequently, for (10.27) to hold, the largest eigenvalue of A should be strictly less than 1.

The Differential Equations

Now, having the vector $\underline{\beta}$ we can find how the number of variables and checks evolve. For all $i \geq 0$,

$$\Delta L_i(t) = L_i(t+1) - L_i(t) = -2\beta_i(t). \quad (10.30)$$

To see how the check types evolve, we note that for a given check type there are two kinds of flows to be considered. A negative flow going out and a

positive flow coming in from the checks of higher degrees. In this regard, for a type $\underline{\tau} = (\tau_0, \dots, \tau_{w-1})$ with $\deg(\underline{\tau}) < K$ let $\partial\underline{\tau}$ be the set of types of degree $\deg(\underline{\tau}) + 1$ such that by removing one edge from them we reach to the type $\underline{\tau}$. The set $\partial\underline{\tau}$ consists of w types which we denote by $\underline{\tau}^d$, $d \in \{0, 1, \dots, w-1\}$, such that

$$\underline{\tau}^d = \underline{\tau} + (0, \dots, \overset{d}{1}, \dots, 0), \quad (10.31)$$

where $+$ denotes vector addition in the field of reals. Thus, if $\deg(\underline{\tau}) < K$, we obtain

$$\Delta C(t, i, \underline{\tau}) = -2 \sum_{d=0}^{w-1} \beta_{i+d}(t) \frac{\tau_d c(t, i, \underline{\tau})}{L_{i+d}(t)} + \sum_{d=0}^{w-1} (1 + \tau_d) \beta_{i+d}(t) \frac{C(t, i, \underline{\tau}^d)}{L_{i+d}(t)}. \quad (10.32)$$

The right-hand side of (10.32) has two parts. The first part corresponds to the flow that is going out of $C(t, i, \underline{\tau})$ and has negative sign. The right part is the incoming flow from the check nodes of higher degrees. In the case where $\deg(\underline{\tau}) = K$, we only have an outgoing flow since no check node with higher degrees exist. Hence, for the case $\deg(\underline{\tau}) = K$ we can write

$$\Delta C(t, i, \underline{\tau}) = -2 \sum_{d=0}^{w-1} \beta_{i+d}(t) \frac{\tau_d c(t, i, \underline{\tau})}{L_{i+d}(t)}. \quad (10.33)$$

We now write the initial conditions for the variables and check types. Firstly, note that $L_i(0) = 2N$. In the beginning of the algorithm, all checks are of degree K , thus for types $\underline{\tau}$ such that $\deg(\underline{\tau}) < K$, we have $C(0, i, \underline{\tau}) = 0$. For $\deg(\underline{\tau}) = K$ we have

$$C(0, i, \underline{\tau}) = \alpha N \frac{\binom{K}{\tau_0, \tau_1, \dots, \tau_{w-1}}}{w^K}. \quad (10.34)$$

In order to write the differential equations, we rescale the (round) time by N , i.e.

$$t \leftarrow \frac{t}{N}, \quad (10.35)$$

and also normalize all our other numbers by N , i.e.,

$$c(t, \cdot, \cdot) = \frac{C(Nt, \cdot, \cdot)}{N} \text{ and } \ell_i(t) = \frac{L_i(Nt)}{N}. \quad (10.36)$$

We then obtain for $i \in \{0, 1, \dots, L + w - 2\}$,

$$\frac{d\ell_i(t)}{dt} = -2\beta_i(t). \quad (10.37)$$

For $i \in \{0, 1, \dots, L - 1\}$ and $\deg(\underline{\tau}) < K$ we have

$$\frac{dc(t, i, \underline{\tau})}{dt} = -2 \sum_{d=0}^{w-1} \beta_{i+d}(t) \frac{\tau_d c(t, i, \underline{\tau})}{\ell_{i+d}(t)} + \sum_{d=0}^{w-1} (1 + \tau_d) \beta_{i+d}(t) \frac{c(t, i, \underline{\tau}^d)}{\ell_{i+d}(t)}, \quad (10.38)$$

and otherwise if $\deg(\underline{\tau}) = K$ we have

$$\frac{dc(t, i, \underline{\tau})}{dt} = -2 \sum_{d=0}^{w-1} \beta_{i+d}(t) \frac{\tau_d c(t, i, \underline{\tau})}{\ell_{i+d}(t)}. \quad (10.39)$$

The vector $\bar{\beta}$ is also found as follows. For p being the current phase, we have

$$\underline{\beta}(t) = (\beta_0(t), \dots, \beta_{L+w-2}(t))^T = (I - A)^{-1} e_p, \quad (10.40)$$

where $A = [A_{i,j}]_{(L+w-1)(L+w-1)}$ has the form

$$A_{i,j} = \frac{1}{\ell_j(t)} \begin{cases} \sum_{k=i-w+1}^i 2c(t, k, \pi_{i-k, i-k}) & i = j, \\ \sum_{k=j-w+1}^i c(t, k, \pi_{i-k, j-k}) & 0 < |i - j| < w, \\ 0 & \text{otherwise} \end{cases} \quad (10.41)$$

Finally, the initial conditions are given by:

$$\begin{aligned} \ell_i(0) &= 2, \text{ for } 0 \leq i \leq L + w - 2 \\ c(0, i, \underline{\tau}) &= \begin{cases} \alpha \frac{\binom{K}{\tau_0, \tau_1, \dots, \tau_{w-1}}}{w^K} & \text{if } \deg(\underline{\tau}) = K \text{ and } 0 \leq i \leq L - 1, \\ 0 & \text{otherwise} \end{cases} \end{aligned} \quad (10.42)$$

The Criterion for the UC to Succeed

We argue now that the criterion for the UC algorithm to succeed (i.e., to find a solution that satisfies almost all the constraints) is

$$\rho(A) \leq 1 - \delta, \quad (10.43)$$

for any time t and where δ is a positive constant. We give an intuitive argument here and notice that the proof can be followed similar to [128, Lemma 4] and [80, Proposition 4.9].

Consider a particular time t and assume that the condition (10.43) holds. When we fix a variable at a free step, a sequence of forced steps follows. The generation of such forced variables (or unit clauses) during the round follows the pattern of a multi-rate Galton-Watson branching process. Such a process starts with a root which is the free variable that we set in the beginning of the round. Then, at every step of the process all individuals born at the previous step generate a number of offsprings. The number of offsprings in a Galton-Watson tree may follow an arbitrary fixed distribution whose mean is dependent on the position of the variables born at the previous step (the elements of the matrix A). The net result is that if $\rho(A) < 1$, then irrespective the distribution of the offsprings and their position, the population certainly becomes extinct, eventually. Mathematically speaking, this means that the Galton-Watson process is sub-critical and with probability 1, the tree is of finite size.

Assume now that we denote the Galton-Watson tree by \mathcal{T} . Also, with a slight abuse of notation, denote by \mathcal{T}_i the set of vertices that \mathcal{T} has on position

K	3	4	5
$\alpha_{\text{UC}}(K)$	2.66	4.50	7.58
$\alpha_{\text{UC},L=50,w=3}(K)$	3.67	7.81	15.76

Table 10.3: *First line:* The thresholds for UCP on the uncoupled ensemble. *Second line:* UCP threshold for a coupled chain with $w = 3$, $L = 50$.

i . Conditioned on \mathcal{T} , the probability that at position i a variable is hit more than once by \mathcal{T} is $O(\frac{|\mathcal{T}_i|^2}{N})$. Hence, in expectation there are $O(\frac{\mathbb{E}[|\mathcal{T}_i|^2]}{N})$ unsatisfied clauses that are generated as position i and at time t . Assuming $\rho(A) \leq 1 - \delta$, $\mathbb{E}[|\mathcal{T}_i|^2]$ is uniformly bounded from above by a constant for all the times t . Hence, after the UC algorithm is completed, the expected number of un-satisfied clauses at position i is $O(1)$. In fact, with a little bit work, one can show that there is a positive probability that at a position i , there are no un-satisfied clauses.

On the other hand, if $\rho(A)$ crosses the value 1 at a time t , then the corresponding Galton-Watson process becomes super-critical and it will generate with high probability a population of size $\Theta(N)$. As a result, there are $\Theta(1)$ number of clauses unsatisfied at time t . So if the value of $\rho(A)$ stays above 1 for a notable amount of time, then $\Theta(N)$ clauses would be left unsatisfied at the end of the UC algorithm.

10.3.4 Numerical Implementation

We have implemented the above set of differential equations in C. We define the threshold $\alpha_{\text{UC},L,w}(K)$ as the highest density for which the spectral norm (largest eigenvalue) of the matrix A is strictly less than one throughout the whole algorithm. A practical point to notice here is that, for the sake of implementation, we assume a phase p finishes when its corresponding variable $\ell_p(t)$ goes below a (very) small threshold $\epsilon > 0$. In our implementations, we have typically taken $\epsilon = 10^{-5}$. However, it can be made arbitrarily small as long as the computational resources allow.

Table 10.3 shows the value of $\alpha_{\text{UC},L,w}(K)$ with $L = 50$ and $w = 3$ for different choices of K . As we observe from Table 10.3, for the UC algorithm with the specific schedule mentioned above, there is a significant threshold improvement over the un-coupled ensemble.

For $L = 50, w = 3, K = 3$ and several values of α , we have plotted in Figure 10.3 the evolution of largest eigenvalue of A as a function of round time t .

In order to characterize analytically the ultimate threshold for the UC algorithm when L and w grow large, we proceed by further analyzing the set of differential equations.

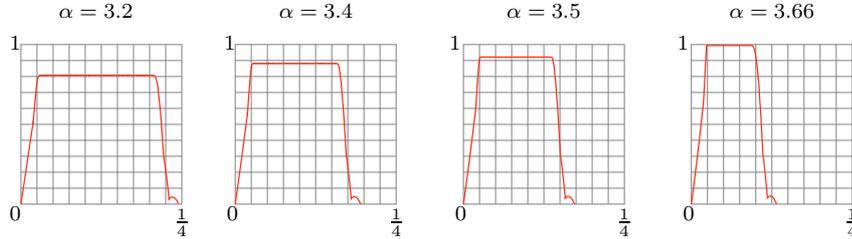


Figure 10.3: The largest eigenvalue of the matrix A , plotted versus the round time t (the number of rounds divided by the total number of variables NL). The plots correspond to an actual implementation of the UC algorithm for the 3-SAT coupled ensemble with $L = 50$ and $w = 3$. As we observe, for $\alpha < 3.67$, there is a gap between the largest eigenvalue of A and the value 1 throughout the UC algorithm. By increasing α this gap shrinks to 0. For $\alpha = 3.66$ (the right-most plot) this gap is around 0.006.

10.3.5 Further Simplifications

The set of differential equations (10.37)-(10.42) is an autonomous system of first order differential equations with fixed initial conditions. Thus, when a solution exists, it is unique. Our objective in this section is to simplify these equations and rewrite them in a new setting that only involves the profile of literals $\{\ell_i\}_{i \geq 0}$. We note here that due to the uniqueness of the solution of these equations, our methods to simplify the equations are on the safe side.

The first basic observation is that $-2\frac{\beta_i}{\ell_i} = \frac{d}{dt} \ln \ell_i$. Therefore, the equation for $c(t, i, \underline{\tau})$ with $\deg(\underline{\tau}) < K$ can be written as

$$dc(t, i, \underline{\tau}) = \left\{ \sum_{d=0}^{w-1} \tau_d (d \ln \ell_{i+d}(t)) \right\} c(t, i, \underline{\tau}) - \frac{1}{2} \sum_{d=0}^{K-1} (1 + \tau_d) d(\ln \ell_{i+d}(t)) c(t, i, \underline{\tau}^{(d)}), \quad (10.44)$$

and similarly when $\deg(\underline{\tau}) = K$ we can write

$$dc(t, i, \underline{\tau}) = \left\{ \sum_{d=0}^{w-1} \tau_d (d \ln \ell_{i+d}(t)) \right\} c(t, i, \underline{\tau}). \quad (10.45)$$

In relations (10.44) and (10.45), the term dt has been “simplified” purposely: indeed one can view this equation as a set of first order partial differential equations. The time dependence of $c(t, i, \underline{\tau})$ is not explicit but only implicit through the $\ell_i(t)$. Therefore, in the next few lines we consider $c(t, i, \underline{\tau})$ as a function of ℓ_i 's, and drop the explicit time dependence. From (10.45) one can

see that for the case $\deg(\underline{\tau}) = K$ we have

$$\begin{aligned} \frac{\partial c(t, i, \underline{\tau})}{\partial \ln \ell_{i+d}} &= \tau_d c(t, i, \underline{\tau}), \quad d = 0, \dots, w-1, \\ \frac{\partial c(t, i, \underline{\tau})}{\partial \ln \ell_{i+d}} &= 0, \quad d \neq 0, \dots, w-1. \end{aligned} \quad (10.46)$$

As a result of the above relations one can easily guess that for the case $\deg(\underline{\tau}) = K$ we have

$$c(t, i, \underline{\tau}) = p_\tau \prod_{d=0}^{w-1} (\ell_{i+d})^{\tau_d}, \quad (10.47)$$

and also by considering the initial conditions, we obtain

$$p_\tau = \frac{\binom{K}{\tau_0, \tau_1, \dots, \tau_{w-1}} \alpha}{w^K} \frac{\alpha}{2^K}. \quad (10.48)$$

Let us consider types with degree less than K . It turns out that these equations can be integrated *iteratively*. In order to represent $c(t, i, \underline{\tau})$ in terms of the literals, we first need the following definition. Consider two types $\underline{\tau}$ and $\underline{\tau}'$. We say that $\underline{\tau}'$ dominates $\underline{\tau}$ if for any $d \in \{0, 1, \dots, w-1\}$ we have $\tau_d \leq \tau'_d$. We also represent dominance by

$$\underline{\tau} \prec \underline{\tau}'. \quad (10.49)$$

Lemma 10.1. *We have for $i \in \{0, 1, \dots, L-1\}$*

$$c(t, i, \underline{\tau}) = \prod_{d=0}^{w-1} \ell_{i+d}^{\tau_d} \sum_{\underline{\tau}': \underline{\tau} \prec \underline{\tau}', \deg(\underline{\tau}')=K} p_{\underline{\tau}'} \prod_{d=0}^{w-1} \left(1 - \frac{\ell_{i+d}}{2}\right)^{\tau'_d - \tau_d}. \quad (10.50)$$

In particular, for the degree 2 types, which we need in the matrix A , we find from Lemma 10.1, and after some simple algebra, that for $i, k \in \{0, 1, \dots, L-1\}$ such that $k \in \{i-w+1, \dots, i\}$, we have

$$c(t, k, \pi_{i-k, i-k}) = \frac{\alpha}{2^K} \frac{K(K-1)}{2w^2} \ell_i^2 \left(1 - \frac{1}{w} \sum_{d=0}^{w-1} \frac{\ell_{k+d}}{2}\right)^{K-2}.$$

Also, for $i, j, k \in \{0, 1, \dots, L-1\}$ such that $i < j$ and $k \in \{j-w+1, \dots, i\}$ we have

$$c(t, k, \pi_{i-k, j-k}) = \frac{\alpha}{2^K} \frac{K(K-1)}{w^2} \ell_i \ell_j \left(1 - \frac{1}{w} \sum_{d=0}^{w-1} \frac{\ell_{k+d}}{2}\right)^{K-2}.$$

This brings us to the following conclusion.

Corollary 10.1. *For $i, j \in \{0, 1, \dots, L+w-2\}$, $A_{i,j}$ can be expressed in terms of the literals as follows. If $0 \leq j-i \leq w-1$, we have*

$$A_{i,j} = \frac{\alpha}{2^K} \frac{K(K-1)}{w} \ell_i \frac{1}{w} \sum_{k=j-i}^{w-1} \left(1 - \frac{1}{w} \sum_{d=0}^{w-1} \frac{\ell_{j-k+d}}{2}\right)^{K-2}. \quad (10.51)$$

Also, if $0 \leq i - j \leq w - 1$, we have

$$A_{i,j} = \frac{\ell_i}{\ell_j} A_{j,i}, \quad (10.52)$$

and $A_{i,j} = 0$ otherwise. More compactly, one can write

$$A_{i,j} = \mathbb{1}\{|i - j| < w\} \frac{\alpha}{2^K} \frac{K(K-1)}{w} \ell_i \frac{1}{w} \sum_{k=|j-i|}^{w-1} \left(1 - \frac{1}{w} \sum_{d=0}^{w-1} \frac{\ell_{\max(i,j)-k+d}}{2}\right)^{K-2}. \quad (10.53)$$

Let us summarize: the differential equations can be expressed, solely in terms of the literals, as an autonomous system of first order differential equations. Assuming that we are in phase p , the differential equations take the following form

$$\frac{d\bar{\ell}}{dt} = -2(I - A)^{-1} e_p, \quad (10.54)$$

where the matrix A is expressed in terms of ℓ_i 's as in Corollary 10.1 and

$$\frac{d\bar{\ell}}{dt} = \begin{bmatrix} \frac{d\ell_0}{dt} \\ \frac{d\ell_1}{dt} \\ \vdots \\ \frac{d\ell_{L+w-1}}{dt} \end{bmatrix}.$$

10.3.6 Conserved Quantities

The system of equations in (10.54) can be rewritten as

$$\sum_{j=0}^{L+w-2} (\delta_{ij} - A_{i,j}) \frac{d\ell_j}{dt} = -2\delta_{pi}, \quad i = p, p+1, \dots, L+w-2, \quad (10.55)$$

where p denotes the phase of the algorithm. Note here that $A_{i,j}$ is given as in (10.53) and

$$\frac{d\ell_j}{dt} = 0, \quad \forall j \notin \{p, p+1, \dots, L+w-2\}. \quad (10.56)$$

By multiplying $\ell_i^{-1} dt$ on both sides of (10.55) we obtain

$$d(\ln \ell_i) - \sum_{j=i-w+1}^{i+w-1} (\ell_i^{-1} A_{i,j}) d\ell_j = -2\delta_{pi} \ell_i^{-1} dt. \quad (10.57)$$

Now, after a careful manipulation one sees that the sum

$$\sum_{j=i-w+1}^{i+w-1} (\ell_i^{-1} A_{i,j}) d\ell_j =$$

$$\begin{aligned} & \frac{\alpha}{2^K} \frac{K(K-1)}{w} \sum_{j=i-w+1}^{i+w-1} d\ell_j \sum_{k=|j-i|}^{w-1} \left(1 - \frac{1}{w} \sum_{d=0}^{w-1} \frac{\ell_{\max(i,j)-k+d}}{2}\right)^{K-2} \\ &= \frac{\alpha}{2^K} \frac{K(K-1)}{w} \sum_{k=0}^{w-1} \sum_{s=0}^{w-1} d\ell_{i-k+s} \left(1 - \frac{1}{w} \sum_{d=0}^{w-1} \frac{\ell_{i-k+d}}{2}\right)^{K-2}, \end{aligned}$$

is an exact differential form. In other words, by defining

$$Q_i \triangleq -\frac{\alpha K}{2^{K-1}} \sum_{k=0}^{w-1} \left(1 - \frac{1}{w} \sum_{d=0}^{w-1} \frac{\ell_{i-k+d}}{2}\right)^{K-1}, \quad (10.58)$$

we have

$$\frac{\partial Q_i}{\partial \ell_j} = (\ell_i^{-1} A_{i,j}). \quad (10.59)$$

Equivalently, we can write

$$dQ_i = \sum_{j=i-w+1}^{i+w-1} (\ell_i^{-1} A_{i,j}) d\ell_j. \quad (10.60)$$

From (10.59) and (10.57) one gets

$$d(\ln \ell_i) - dQ_i = -2\delta_{pi} \ell_i^{-1} dt. \quad (10.61)$$

which means that

$$P_i = \ln \ell_i - Q_i + \int_0^t 2\delta_{pi} \ell_i^{-1} dt \quad (10.62)$$

is a *conserved quantity* or an *integral of motion*, i.e., the value of P_i 's is independent of the time t . The values of P_i 's hence can be found by the initial conditions $\ell_i(0) = 2$ for $i \geq 0$ and $\ell_i(0) = 0$ for $i < 0$. Consequently, we find that for $i \geq w$

$$P_i = \ln 2. \quad (10.63)$$

However, for $i < w$ the value of P_i is strictly less than $\ln 2$.

10.3.7 Slightly Modified Initial Conditions

The dependence of P_i to the position i inserts some undesired asymmetry in the analysis, and makes it quite cumbersome. As a result, we find it more convenient to slightly modify the initial conditions of the differential equations and remove such an asymmetry in the value of P_i 's. As we prove in the sequel, this modification of the initial conditions does not have any effect on the final results of this chapter and hence can be assumed without loss of any generality.

So, to summarize, the initial conditions

$$\ell_i(0) = \begin{cases} 0 & \text{If } i < 0, \\ 2 & \text{If } i \geq 0, \end{cases}$$

has the deficiency that the value of P_i is dependent on the position i . The objective here is to devise a new set of initial conditions on the problem so that the value of P_i given in (10.62) is equal to $\ln 2$ for all the positions $i \in \{0, 1, \dots, L + w - 2\}$. Let us denote the new initial values for literals by $\tilde{\ell}_i$, $0 \leq i \leq L + w - 2$. We further extend the profile $\{\tilde{\ell}_i\}$ to all the positions in \mathbb{Z} by letting

$$\tilde{\ell}_i = 0 \quad \forall i < 0 \tag{10.64}$$

$$\tilde{\ell}_i = 2 \quad \forall i > L + w - 2. \tag{10.65}$$

We now want to find the values $\tilde{\ell}_i$ such that initializing the differential equations with

$$\ell_i(0) = \tilde{\ell}_i, \quad \forall i \in \mathbb{Z} \tag{10.66}$$

will result the fact that

$$P_i = \ln 2, \quad \forall i \in \{0, 1, \dots, L + w - 2\}. \tag{10.67}$$

A close look at equations (10.58) and (10.62) shows that this objective is feasible if and only if the profile $\{\tilde{\ell}_i\}$ is the solution of the following set of equations:

$$\ln \frac{\tilde{\ell}_i}{2} = -\frac{\alpha K}{w 2^{K-1}} \sum_{k=0}^{w-1} \left(1 - \frac{1}{w} \sum_{d=0}^{w-1} \frac{\tilde{\ell}_{i-k+d}}{2}\right)^{K-1}, \quad \forall i \in \{0, 1, \dots, L + w - 2\}. \tag{10.68}$$

In order to find a solution of (10.68), we can apply the iterative procedure given in Algorithm 7.

Procedure 7 Iterative procedure to find a a solution to (10.68)

- 1: Start by initializing $\tilde{\ell}_i^0 = 2$ for $0 \leq i \leq L + w - 2$.
- 2: For $m = 1, 2, \dots$ do
 - For $i = 0, 1, \dots, L + w - 2$ do the update

$$\tilde{\ell}_i^{m+1} = 2 \exp\left\{-\frac{\alpha K}{2^{K-1} w} \sum_{k=0}^{w-1} \left(1 - \frac{1}{w} \sum_{d=0}^{w-1} \frac{\tilde{\ell}_{i-k+d}^m}{2}\right)^{K-1}\right\}. \tag{10.69}$$

Lemma 10.2. *The procedure of Algorithm 7 converges to a solution of the set of equations in (10.68). We denote such a solution by $\{\tilde{\ell}_i\}_{0 \leq i \leq L+w-2}$. Furthermore, we have*

1. The value of $\tilde{\ell}_i$ is non-decreasing on $i \in \{0, 1, \dots, L + w - 2\}$.
2. There exists constants $c_1, c_2 > 0$ such that

$$\tilde{\ell}_i \geq 2 - c_1 e^{-c_2 (\frac{i}{w})^2}. \tag{10.70}$$

Proof. Consider Algorithm 7. We first show that for $m \in \mathbb{N}$ and for $i \geq 0$, we have

$$\tilde{\ell}_i^{m+1} \leq \tilde{\ell}_i^m. \quad (10.71)$$

With this statement in mind, it is easy to see that the profile $\{\tilde{\ell}_i^m\}_{i \geq 0}$ converges to a limit as $m \rightarrow \infty$. In order to prove (10.71), we first note that the function

$$h(x) = 2 \exp\left\{-\frac{\alpha K}{2^{K-1}} \left(1 - \frac{x}{2}\right)^{K-1}\right\}, \quad (10.72)$$

is an increasing function on the domain $x \in [0, 2]$. Using this and the definition of the profile $\{\tilde{\ell}_i^0\}_{i \in \mathbb{Z}}$, it is easy to see that for $i \in \mathbb{Z}$ we have

$$\tilde{\ell}_i^1 \leq \tilde{\ell}_i^0. \quad (10.73)$$

One can then use (10.73) together with (10.69) and the fact that h is an increasing function, to show that for $i \in \mathbb{Z}$ we have $\tilde{\ell}_i^2 \leq \tilde{\ell}_i^1$. Finally, by continuing this procedure inductively, we obtain (10.71).

Furthermore, from the fact that $h(x)$ is an increasing on $[0, 2]$ and also the fact that the profile $\{\tilde{\ell}_i^0\}_{i \in \mathbb{Z}}$ is a non-decreasing profile, we deduce that the profile $\{\tilde{\ell}_i^1\}_{i \in \mathbb{Z}}$ is also a non-decreasing profile. Again, one can generalize this statement inductively to deduce that for all $m \in \mathbb{N}$, the profiles $\{\tilde{\ell}_i^m\}_{i \in \mathbb{Z}}$ are non-decreasing profiles, and hence, the same property holds for their limit. The relation (10.70) follows easily from Lemma 10.3 and we do not repeat its proof here. \square

As a consequence of Lemma 10.2, by starting with the initial conditions $\ell_i(0) = \tilde{\ell}_i$, we have $P_i = \ln 2$ for $0 \leq i \leq L - w + 2$. Also, as by (10.70) the profile $\{\tilde{\ell}_i\}_{i \geq 0}$ converges doubly exponentially fast in i to the value $\ell = 2$, we lose no generality in starting with $\{\tilde{\ell}_i\}_{i \geq 0}$ as our initial condition. Let us now summarize.

Theorem 10.2. *Starting with the initial conditions $\{\tilde{\ell}_i\}_{i \geq 0}$, for any time t during the UC algorithm, the profile of literals is a solution of the following set of equations*

$$\ln \frac{\ell_i}{2} - Q_i + \int_0^t 2\delta_{pi} \ell_i^{-1} dt = 0, \quad \forall i \in \{0, 1, \dots, L + w - 2\}. \quad (10.74)$$

10.3.8 A Potential Function

From the definition of Q_i in (10.58), one can explicitly check the following. For $j, k \geq 0$,

$$\frac{\partial Q_k}{\partial \ell_j} = \frac{\partial Q_j}{\partial \ell_k}. \quad (10.75)$$

Clearly, the same statement is also true if we replace Q_i with $\ln \frac{\ell_i}{2} - Q_i$. Now, as the space of $\{\ell_i, i \geq 0\}$ is simply connected, by the Poincaré Lemma, there exists a functional Φ such that

$$\frac{\partial \Phi}{\partial \ell_i} = \ln \frac{\ell_i}{2} - Q_i, \quad i \geq 0. \quad (10.76)$$

We call Φ the *the coupled potential function* associated to the system of differential equations or simply *the coupled potential*. Fortunately, in our case the potential Φ is easy to find. Let us first assume $L = w = 1$, i.e., the individual system. For this case we only have one literal which we denote by ℓ . From (10.76) and (10.58) we obtain

$$\frac{\partial \Phi_{\text{ind}}}{\partial \ell} = \ln \ell + \frac{\alpha K}{2^{K-1}} \left(1 - \frac{\ell}{2}\right)^{K-1} - \ln 2.$$

By integrating the above relation, we get

$$\Phi_{\text{ind}}(\ell, \alpha, K) = 2 - \ell \left(1 - \ln \frac{\ell}{2}\right) - \frac{\alpha}{2^{K-2}} \left(1 - \frac{\ell}{2}\right)^K, \quad (10.77)$$

where the constant 2 in the potential is due to the fact that, without loss of generality, we fix the potential to 0 at the point $\ell = 2$. The coupled potential can be obtained from the potential of the individual system by a simple formula derived in [126]. For our case, by using (10.77), we define the coupled potential to be

$$\Phi = \sum_{i=0}^{L+w-2} 1 - \ell_i \left(1 - \ln \frac{\ell_i}{2}\right) - \frac{\alpha}{2^{K-2}} \sum_{i=0}^{L+w-2} \frac{1}{w} \sum_{k=0}^{w-1} \left(1 - \frac{1}{w} \sum_{d=0}^{w-1} \frac{\ell_{i-k+d}}{2}\right)^K. \quad (10.78)$$

It can easily be checked that for $i \geq 0$,

$$\frac{\partial \Phi}{\partial \ell_i} = \ln \frac{\ell_i}{2} - Q_i.$$

From now on, we call the functional Φ given in (10.78) the potential associated to our system.

10.3.9 The Threshold of the UC Algorithm for the Coupled Ensemble

Let us go back for a moment and have a look at the result of Theorem 10.1 in Section 10.2.3: for coupled systems with a state described via a one-dimensional recursion (as in (10.15)), the threshold of the coupled system can be computed from the potential function of the individual system by the relation (10.17).

An educated guess for the threshold of the UC algorithm for the coupled ensemble is the one obtained by (10.17), with the potential function given in (10.77). Figure 10.4 plots the value of Φ_{ind} given in (10.77) as a function of ℓ for different values of α and for $K = 3$. The potential threshold found in this way for $K = 3$ is 3.6717 which is extremely close to the one observed in the (numerical) solution of the differential equations (see Table 10.3). Is this a coincidence? Well, let us look at the case where $K = 4$. The potential threshold is found to be 7.8146 which is again very close to the one from the numerical solution of differential equations, and so on. It thus seems that the potential threshold is equal to the threshold of the UC algorithm for the coupled instances with L, w tending to infinity. This is indeed true.

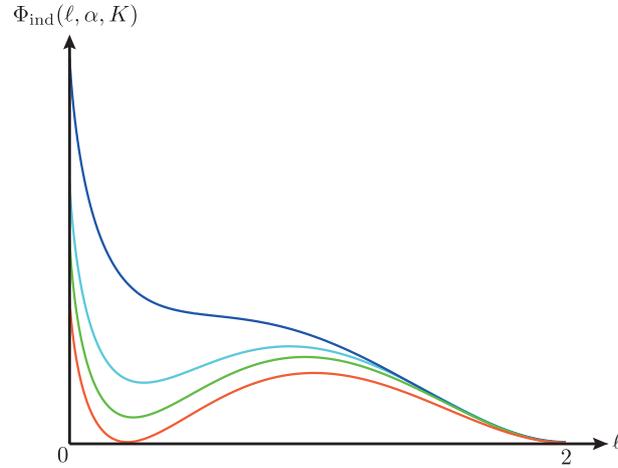


Figure 10.4: The potential function associated to the individual ensemble, $\Phi_{\text{ind}}(\ell, \alpha, K)$, plotted as a function of ℓ for $K = 3$ and different values of α . From top to bottom, the curves correspond respectively to $\alpha = 3.2, 3.5, 3.6, 3.67$. The smallest α for which the potential goes below the horizontal axis is around 3.67.

Theorem 10.3. *We have*

$$\alpha_{\text{cUC}}(K) \triangleq \lim_{w \rightarrow \infty} \lim_{L \rightarrow \infty} \alpha_{\text{UC}, L, w} = \sup\{\alpha \geq 0 \mid \min_{\ell \in [0, 2]} \Phi_{\text{ind}}(\ell, \alpha, K) \geq 0\}. \quad (10.79)$$

Remark 10.1. *For large K we find*

$$\alpha_{\text{cUC}}(K) \doteq 2^{K-1}. \quad (10.80)$$

This is roughly a factor $\frac{K}{e}$ of improvement over the threshold of the UC algorithm for the individual system that is given by

$$\alpha_{\text{UC}}(K) \doteq \frac{e2^{K-1}}{K}. \quad (10.81)$$

However, the threshold of the coupled UC is still below the SAT/UNSAT threshold $\alpha_s(K)$ which is roughly $2^K \ln 2$. It is also below the condensation threshold which is $2^K \ln 2 - 3/2 \ln 2 + o(1)$.

The rest of this chapter is devoted to the proof of Theorem 10.3. We first derive several properties of the profile of the literals. We then use these properties to prove that for $\alpha \leq \alpha_{\text{cUC}}$, there exists a constant $w_0 = w(\alpha, K) < \infty$ and a constant $\delta = \delta(\alpha, K) > 0$ such that if we choose $w \geq w_0$, for any time t during the UC algorithm, we have $\rho(A) < 1 - \delta$. Hence, the fact that the UC algorithm succeeds follows from the discussions at the end of Section 10.3.3.

10.3.10 How Does the Profile Look Like?

As the first step towards the proof of Theorem 10.3, the objective of this section is to provide various details about the way the profile looks like during the UC algorithm. Sp far, the only thing that we know for sure is that given the current phase p , the profiles rests at $\ell_i = 0$ for $i < p$ and rises up to $\ell_i = 2$ for i at its right boundary, i.e., $i > L + w - 2$. Let us define the *transition region* of the profile to be the region of positions $i \in \{p, p + 1, \dots, L + w - 2\}$ such that the value of ℓ_i is not very close to 2. One of the main results of this section is that the transition region is always $O(w)$ during the whole UC algorithm.

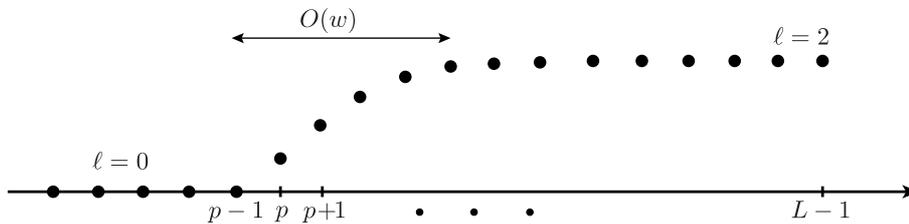


Figure 10.5: A schematic representation of the profile. The transition region is of size $O(w)$.

The idea here is to consider different regions for the position $i \in \{0, 1, \dots, L + w - 2\}$ and analyze the behavior of the profile in each of these regions. These regions are specified by the solutions the equation

$$\ln \frac{\ell}{2} = -\frac{\alpha K}{2^{K-1}} \left(1 - \frac{\ell}{2}\right)^{K-1}, \tag{10.82}$$

or equivalently, the fixed points of the equation

$$\ell = 2 \exp\left(-\frac{\alpha K}{2^{K-1}} \left(1 - \frac{\ell}{2}\right)^{K-1}\right). \tag{10.83}$$

The reason that we consider the fixed points of (10.83) stems from the conservation equations (10.74). Let us give an intuitive explanation. Assume w is a large but fixed number. We know that the profile is increasing. If the transition region of the profile is much larger than w (e.g., it is $O(w^2)$), then it is easy to see that there is a value $\ell^* \in (0, 2)$ such that the profile is close to ℓ^* for at least $O(w)$ positions. In other words, there is a small constant $\delta > 0$ and two positions i_1, i_2 such that $i_2 - i_1 \geq 2w$ and for $i \in \{i_1, i_1 + 1, \dots, i_2\}$ we have $\ell_i \in [\ell^* - \delta, \ell^* + \delta]$. Now, by looking at the conservation equations (10.74) for a position $i = \frac{i_1 + i_2}{2}$ we can easily deduce that ℓ^* should be close to a fixed point of (10.82).

We now proceed by specifying the solutions of (10.82). Let us define the function

$$f(\ell) = \ln \frac{\ell}{2} + \frac{\alpha K}{2^{K-1}} \left(1 - \frac{\ell}{2}\right)^{K-1}. \tag{10.84}$$

The first derivative of the function f is

$$f'(\ell) = \frac{1}{\ell} \left(1 - \frac{\alpha K(K-1)}{2^K} \ell \left(1 - \frac{\ell}{2} \right)^{K-2} \right). \quad (10.85)$$

Now, one can easily see that the equation $f'(\ell) = 0$ has at most two solutions on $\ell \in [0, 2]$. This is because the function $\ell \left(1 - \frac{\ell}{2} \right)^{K-2}$ is a uni-modal function on $\ell \in [0, 2]$. Hence, by the Rolle Theorem, the equation $f(\ell) = 0$ (or equivalently (10.82)) has at most three solutions on $[0, 2]$. Indeed, a bit of calculus reveals that there exists a $\alpha^* < \alpha_{\text{cUCP}}$ such that the following holds. For $\alpha < \alpha^*$ the equation (10.82) has exactly one solution which is the trivial solution $\ell = 2$ and for $\alpha > \alpha^*$ there are three distinct solutions to (10.82). In the following, we assume the harder case, i.e., we assume that $\alpha^* < \alpha < \alpha_{\text{cUCP}}$ and hence the equation (10.82) has three distinct solutions on $[0, 2]$. From what we mention in the sequel, the other case, $\alpha < \alpha^*$, is much easier to analyze and in fact will follow directly from the present analysis.

As shown in Figure 10.6, the fixed points of (10.83) are obtained by intersecting the two curves

$$y_1(\ell) = 2 \exp\left(-\frac{\alpha K}{2^{K-1}} \left(1 - \frac{\ell}{2}\right)^{K-1}\right), \quad (10.86)$$

$$y_2(\ell) = \ell, \quad (10.87)$$

on the region $\ell \in [0, 2]$.

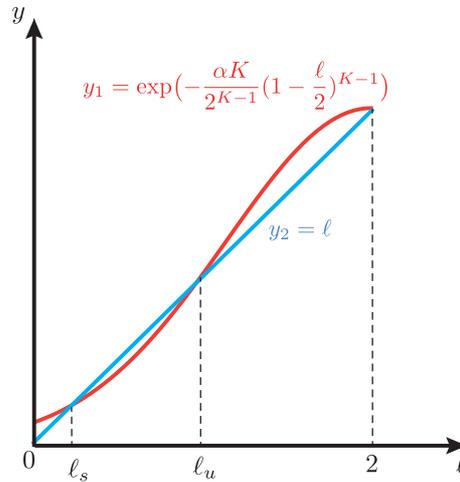


Figure 10.6: A schematic representation of the fixed points of (10.83) which are equivalently the intersection points of the two curves $y_1(\ell)$ and $y_2(\ell)$.

As mentioned above we assume there are three distinct fixed points. The largest one $\ell = 2$ is called *the trivial fixed point*. The middle one is called *the*

unstable fixed point ℓ_u and the smallest one is called *the stable fixed point* ℓ_s . Given these fixed points, we consider the following five regions for the positions $i \in \{0, 1, \dots, L+w-2\}$. Let $\delta > 0$ be a fixed constant, the value of which will be specified in the following lemma in its suitable place. For the moment, we think of δ as a fixed and given constant. As illustrated in Figure 10.7, the five regions are as follows:

- region 1 (R_1): all positions i such that $\ell_i \leq \ell_s - \delta$.
- region 2 (R_2): all positions i such that $\ell_s - \delta < \ell_i \leq \ell_s + \delta$.
- region 3 (R_3): all positions i such that $\ell_s + \delta < \ell_i \leq \ell_u - \delta$.
- region 4 (R_4): all positions i such that $\ell_u - \delta < \ell_i \leq \ell_u + \delta$.
- region 5 (R_5): all positions i such that $\ell_u + \delta < \ell_i \leq 2$.

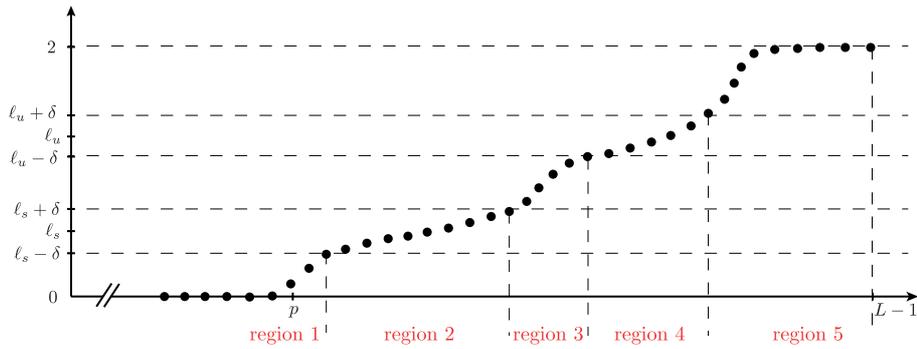


Figure 10.7: Different regions for the value of the profile.

In the following lemma, we provide bounds on the behavior of the profile in each of these regions. We note here that all the results of the following lemma (and all the other results that appear later on) are only dependent on the choice of α and K , Hence, they are independent of the time or phase of the algorithm and are valid for the profile throughout the whole UC algorithm. Let us conclude this section by summarizing its results in the following lemma.

Lemma 10.3. *The following properties hold for the profile of literals $\{\ell_i\}_{i \geq p}$ at any time t .*

(1) For $i > p$

$$\ell_i \geq 2 \exp\left(-\frac{\alpha K}{2^{K-1}}\right).$$

(2) For $i \geq p$

$$2 - \ell_{i+w} \leq \frac{\alpha K}{2^{2K-2}}(2 - \ell_i)^{K-1}.$$

(3) For $i > p$

$$\ln \frac{\ell_{i+\frac{w}{2}}}{2} \leq \frac{-\alpha}{2^{2K-1}} \left(\frac{2 - \ell_i}{2} \right)^{K-1}.$$

Also, There exist positive constants $w_0 = w_0(\alpha, K)$ and $\delta = \delta(\alpha, K)$ such that if we assume $w \geq w_0$ and define regions $R_1 - R_5$ (the regions depend on δ), then

(4) If $i \in R_1 \cup R_3$, there exists a value $\zeta = \zeta(\alpha, K) > 0$ such that $\ell_{i+w} - \ell_i \geq \zeta$. Hence, the length of the regions 1 and 3 is at most $\frac{2w}{\zeta}$.

(5) The length of the regions 2,4 is at most $2w$.

(6) Let $\epsilon > 0$ be an arbitrary positive constant. Define I_ϵ to be the first position for which the value of profile goes above $2 - \epsilon$, i.e.,

$$I_\epsilon = \operatorname{argmin}\{j \geq p \mid \forall i \geq j : \ell_i > 2 - \epsilon\}.$$

Then, there exist constants $c_1 = c_1(\alpha, K)$ and $c_2 = c_2(\alpha, K)$ such that

$$I_\epsilon \leq w(c_1 + c_2 \log(\log \frac{1}{\epsilon})).$$

10.3.11 Why Does the UC Algorithm Work?

In this section, we take the last step in order to prove Theorem 10.3. That is, we show that $\rho(A)$ will be below 1 by a strict gap during the course of the UC algorithm.

Lemma 10.4. *There exist constants $\delta = \delta(\alpha, K) > 0$ and $w_1 = w_1(\alpha, K)$, such that if we choose $w > w_1$, then for any time t , the largest eigenvalue of matrix A is less than $1 - \delta$.*

Finally, by using (10.28), (10.29), and Lemma (10.4) the result of Theorem 10.3 follows.

10.4 Further Remarks and Open Directions

This chapter was about algorithmic implications of the technique of spatial coupling. The main two consequences of this chapter are as follows: (i) For two classes of algorithms, we observed an *algorithmic threshold increase* on the coupled ensemble. This threshold increase was indeed expected from the results of Chapter 9, namely saturation of the SP and dynamical thresholds of coupled K -SAT. (ii) The satisfiability thresholds of the coupled and standard ensembles are the same (at infinite L). Hence, the algorithmic thresholds for the coupled ensemble are also lower bounds for the satisfiability threshold of the standard K -SAT ensemble. By analyzing the algorithms on the coupled ensemble we derived in this chapter new (algorithmic) lower bounds on the satisfiability threshold of the K -SAT ensemble.

Let us now mention a few open directions. We believe that more sophisticated (and analyzable) algorithms for the coupled ensemble can succeed all the way up to the condensation threshold. For an algorithm to perform well on the coupled ensemble, two key features should be carefully designed: (i) the heuristic step of fixing the variables at each step, (ii) the schedule of the algorithm and in particular how this schedule exploits the 1-dimensional structure of the formula.

In this chapter, we have mainly used the technique of spatial coupling to provide analytic results for the standard (uncoupled) ensemble. An interesting direction is to use this technique to provide *algorithms* to solve uncoupled (standard) K -SAT formulas. One approach to do it is as follows. We start from an uncoupled K -SAT formula and embed it into a coupled formula which presumably is easier to solve. Given the solution of the coupled formula, the idea is then to find a solution of the original uncoupled formula. One can think of several ways to relate a coupled formula to an uncoupled one. For instance, the interpolation ideas of Chapter 9 seem to be helpful in this regard.

10.5 Appendix: Auxiliary Lemmas and Proofs

10.5.1 Appendix A: A Message Passing Interpretation For UC

In this section we aim to express the UC algorithm in a message passing (MP) formalism. In particular, we show that with such a formalism the conservation equations (10.74) are precisely the density evolution equations that govern the dynamics of this message passing algorithm. An important consequence of this formalism is to prove that the profile of literals will always remain an strictly increasing profile during the UC algorithm.

We start by explaining the MP formalism on the individual ensemble (i.e., $L = w = 1$). For this ensemble we will analyze the dynamics of the MP procedure in a probabilistic manner and derive the so called density evolution (DE) equations of the MP procedure. We then extend the MP procedure and the DE equations to the coupled ensemble. Let us stress again the fact that the following MP procedure is designed to formulate the dynamics of the UC algorithm in a message passing fashion.

We proceed by recalling that a formula in the individual ensemble can be thought as a bipartite graph with N literals (variable nodes) and $M = N\alpha$ clauses (check nodes). We denote a check by $c, h \in \{0, 1, \dots, M-1\}$ and the variables by $i, j \in \{0, \dots, N-1\}$. Each check has K edges which randomly have chosen one of the N variables. So the graph has MK edges. An edge of the graph, between a check node c and a variable node i is also denoted by $\langle c, i \rangle$. Also, each edge has an associated sign, being -1 or $+1$ with equal probability, which we denote by $J_{c,i}$.

For each edge (c, i) of the graph we associate two types of messages: (i) the check-to-variable message $\mu_{c \rightarrow i}$ which takes its value in the set $\{0, 1\}$ (ii) the variable-to-check message, $(\mu_{i \rightarrow c}, s_{i \rightarrow c})$ which the value of $\mu_{i \rightarrow c}$ is inside $\{0, 1\}$ and the value of $s_{i \rightarrow c}$ is in $\{?, 0, 1\}$. On the intuitive level, these messages are

designed to mimic the behavior of the UC algorithm. In this regard, when a message $\mu_{c \rightarrow i}$ is 1, this means that the check c is forcing the variable i to satisfy it. This situation occurs in the course of the UC algorithm when the check c is a unit clause. Furthermore, when a message $\mu_{i \rightarrow c}$ takes the value 1, this means that the variable i tells the check c that it has a preset value. This other message $s_{i \rightarrow c}$ is the preset value of variable i that it sends to check c .

The MP procedure consists of N steps. At each step $r = 1, \dots, N$ we choose a variable, *mark* it, and update the messages. Thus, in the end of MP all the variables are marked. We now describe the MP procedure in detail through the following stages.

Initialization: In the beginning of the MP procedure, we initialize all the messages $\mu_{c \rightarrow i}$ and $\mu_{i \rightarrow c}$ to 0. This indicates that in the beginning all the checks and variables tell each other that they are essentially free. We also let all the messages $s_{i \rightarrow c}$ to be ? indicating that the variables do not give any information about their value to the checks.

During step r : In each step $r \in \{1, \dots, N\}$, we first choose one variable uniformly at random among the remaining unmarked variables. Let us denote the chosen variable by i . We do the following operations once i is chosen.

1. We first mark the variable i for later correspondence.
2. We then fix the value all the outgoing messages $\mu_{i \rightarrow c}$ to 1.
3. Let s_i be a bernoulli rv whose value is chosen by flipping a fair coin. By flipping a fair coin we specify the value of s_i and give this value to all the messages $s_{i \rightarrow c}$ whose value is equal to ? before step r . That is, if $s_{i \rightarrow c} = ?$ then we permanently fix $s_{i \rightarrow c} = s_i$.
4. Finally, we update the messages as follows. We run the following update rules until we reach a fixed state on the messages and no further updates is necessary. Consider an edge $\langle c, i \rangle$. The check to variable message we have

$$\mu_{c \rightarrow i}^{t+1} = \prod_{j \in \partial c \setminus i} \mathbb{1}\{\mu_{j \rightarrow c}^t = 1, s_{j \rightarrow c}^t = J_{c,j}\}.$$

For the variable to check messages on $\langle c, i \rangle$ we do

$$\begin{aligned} \mu_{i \rightarrow c}^{t+1} &= 1 - \prod_{h \in \partial i \setminus c} \mathbb{1}\{\mu_{h \rightarrow i}^{t+1} = 0\}, \\ s_{i \rightarrow c}^{t+1} &= (-1)^{\frac{1+J_{c,i}}{2}} \mathbb{1}\{s_{i \rightarrow c}^t = ?\}. \end{aligned}$$

10.5.2 Appendix B: Auxiliary Lemmas and Proofs

Proof of Lemma 10.3

Proof of part (1): we start by a key observation. At a phase p , the conservation equations (10.74) for a position $i > p$ reads

$$\ln \frac{\ell_i}{2} = -\frac{\alpha K}{w 2^{K-1}} \sum_{k=0}^{w-1} \left(1 - \frac{1}{w} \sum_{d=0}^{w-1} \frac{\ell_{i-k+d}}{2}\right)^{K-1}, \quad \forall i > p, \quad (10.88)$$

or equivalently,

$$\ell_i = 2 \exp\left\{-\frac{\alpha K}{w 2^{K-1}} \sum_{k=0}^{w-1} \left(1 - \frac{1}{w} \sum_{d=0}^{w-1} \frac{\ell_{i-k+d}}{2}\right)^{K-1}\right\}, \quad \forall i > p. \quad (10.89)$$

It is easy to see that (10.89) can be written in the form of

$$\ell_i = f\left(\frac{1}{w} \sum_{k=0}^{w-1} g\left(\frac{1}{w} \sum_{d=0}^{w-1} \frac{\ell_{i-k+d}}{2}\right)\right), \quad (10.90)$$

where the functions f and g are given as

$$f(x) = 2 \exp\left(\frac{\alpha K}{2^{K-1}} x\right), \quad (10.91)$$

$$g(x) = -\left(1 - \frac{x}{2}\right)^{K-1}. \quad (10.92)$$

Consider a position $i > p$. We can now write

$$\begin{aligned} \ell_i &= f\left(\frac{1}{w} \sum_{k=0}^{w-1} g\left(\sum_{d=0}^{w-1} \frac{\ell_{i-k+d}}{2}\right)\right) \\ &\stackrel{(a)}{\geq} f\left(\frac{1}{w} \sum_{k=0}^{w-1} g\left(\frac{1}{w} \sum_{d=0}^{w-1} \frac{\ell_{i-w}}{2}\right)\right) \\ &= f(g(\ell_{i-w})). \end{aligned}$$

Here, step (a) follows from the fact that the profile of literals is an increasing profile and also from the fact that the functions f and g are increasing functions on the unit interval. Using a similar argument, one can also deduce that for $i \geq p$, we have $\ell_i \leq f(g(\ell_{i+w}))$. So to summarize, we have

$$f(g(\ell_{i-w})) \leq \ell_i \leq f(g(\ell_{i+w})), \quad \forall i > p. \quad (10.93)$$

The first consequence of the above set of inequalities (10.93) is that

$$\ell_i > f(g(0)) = 2 \exp\left(-\frac{\alpha K}{2^{K-1}}\right), \quad \forall i > p. \quad (10.94)$$

Hence, part (1) of the lemma is proved.

Proof of part (2): as discussed above we have for $i \geq p$

$$\begin{aligned}\ell_{i+w} &\geq f(g(\ell_i)) \\ &= 2 \exp\left(-\frac{\alpha K}{2^{K-1}}\left(1 - \frac{\ell_i}{2}\right)^{K-1}\right) \\ &\geq 2\left(1 - \frac{\alpha K}{2^{K-1}}\left(1 - \frac{\ell_i}{2}\right)^{K-1}\right).\end{aligned}$$

Thus, by rearranging terms we get for $i \geq p$

$$2 - \ell_{i+w} \leq \frac{\alpha K}{2^{2K-2}}(2 - \ell_i)^{K-1}.$$

Proof of part (3): starting from the conservation equations (10.74), we have

$$\begin{aligned}\ln \frac{\ell_{i+\frac{w}{2}}}{2} &= -\frac{\alpha K}{2^{K-1}} \frac{1}{w} \sum_{k=0}^{w-1} \left(\frac{1}{w} \sum_{d=0}^{w-1} \frac{2 - \ell_{i+\frac{w}{2}+d-k}}{2}\right)^{K-1} \\ &\leq -\frac{\alpha K}{2^{K-1}} \frac{1}{w} \sum_{k=\frac{w}{2}}^{w-1} \left(\frac{1}{w} \sum_{d=0}^{w-1} \frac{2 - \ell_{i+\frac{w}{2}+d-k}}{2}\right)^{K-1} \\ &\leq -\frac{\alpha K}{2^{K-1}} \frac{1}{w} \sum_{s=0}^{\frac{w}{2}+1} \left(\frac{1}{w} \sum_{d=0}^{w-1} \frac{2 - \ell_{i+d-s}}{2}\right)^{K-1} \\ &\leq -\frac{\alpha K}{2^{K-1}} \frac{1}{w} \sum_{s=0}^{\frac{w}{2}+1} \left(\frac{1}{w} \sum_{d=0}^s \frac{2 - \ell_{i+d-s}}{2}\right)^{K-1} \\ &\leq -\frac{\alpha K}{2^{K-1}} \frac{1}{w} \sum_{s=0}^{\frac{w}{2}+1} \left(\frac{1}{w} \sum_{d=0}^s \frac{2 - \ell_i}{2}\right)^{K-1} \\ &\leq -\frac{\alpha K}{2^{K-1}} \frac{1}{w} \sum_{s=0}^{\frac{w}{2}+1} \left(\frac{s+1}{w} \frac{2 - \ell_i}{2}\right)^{K-1} \\ &\leq -\frac{\alpha K}{2^{K-1}} \left(\frac{2 - \ell_i}{2}\right)^{K-1} \frac{1}{w} \sum_{s=1}^{\frac{w}{2}+2} \left(\frac{s}{w}\right)^{K-1} \\ &\leq -\frac{\alpha K}{2^{K-1}} \left(\frac{2 - \ell_i}{2}\right)^{K-1} \int_0^{\frac{1}{2}} x^{K-1} dx \\ &= -\frac{\alpha}{2^{2K-1}} \left(\frac{2 - \ell_i}{2}\right)^{K-1}.\end{aligned}$$

Proof of part (4): we first consider the region 1 (R_1). In this regard, we define the sequence $\{x_j\}_{j \in \mathbb{N}}$ with $x_0 = 0$ and for $j \geq 0$

$$x_{j+1} = f(g(x_j)). \quad (10.95)$$

Now, by using the fact that for $i \geq p$ we have $\ell_{i+w} \geq f(g(\ell_i))$, we obtain

$$\ell_{p+jw} \geq x_j, \quad \forall j \geq 0. \quad (10.96)$$

We now claim that there exist a positive value $m_s \in (0, 1)$ such that for $j \geq 0$

$$\ell_s - x_{j+1} \leq m_1(\ell_s - x_j). \tag{10.97}$$

This claim can easily be deduced from Figure 10.8. A little calculus reveals that the line tangent to the curve $y_1(\ell) = f(g(\ell))$ at the point $\ell = \ell_s$ stays below the curve $y_1(\ell)$ when $\ell \in [0, \ell_s]$ (see Figure 10.8). Thus, by defining

$$m_1 = y_1'(\ell_s) < 1, \tag{10.98}$$

we have for $\ell \leq \ell_s$

$$y_1(\ell) \geq \ell_s - m_1(\ell_s - \ell). \tag{10.99}$$

Hence, one can write for $j \geq 0$

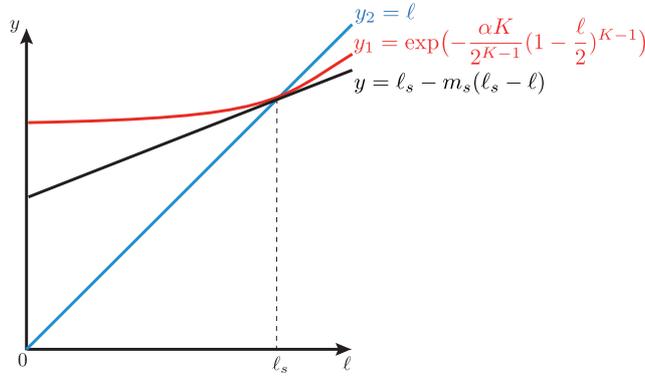


Figure 10.8: The line $y(\ell)$ is the tangent line to the curve $y_1(\ell)$ at the fixed point $\ell = \ell_s$. As we see from the figure, for $\ell \in [0, \ell_s]$, the line $y(\ell)$ stays between the curves $y_1(\ell)$ and $y_2(\ell)$.

$$x_{j+1} = y_1(x_j) \geq \ell_s - m_1(\ell_s - x_j),$$

and as a result

$$\ell_s - x_{j+1} \leq m_1(\ell_s - x_j),$$

Thus, the sequence $\{x_j\}_{j \in \mathbb{N}}$ converges exponentially fast to ℓ_s .

Another important consequence of the discussion above is that for a position $i \in R_1$ we have

$$\begin{aligned} \ell_{i+w} &\geq f(g(\ell_i)) \\ &\stackrel{(a)}{\geq} \ell_s - m_1(\ell_s - \ell_i) \\ &= (1 - m_1)\ell_s + m_1\ell_i \\ &\stackrel{(b)}{\geq} \delta(1 - m_1) + \ell_i, \end{aligned}$$

where step (a) follows from (10.99) and (b) follows from the fact that $\ell_i \leq \ell_s - \delta$. As a result, for $i \in R_1$ we have that

$$\ell_{i+w} - \ell_i \geq (1 - m_1)\delta. \quad (10.100)$$

For region 3, one can use the same ideas (as in region 1) but with a bit of more effort to justify that the same results as in region 1 hold here. That is, for $i \in R_3$ we have

$$\ell_{i+w} - \ell_i \geq (1 - m_3)\delta, \quad (10.101)$$

for a positive constant $m_3 < 1$.

Now, by letting

$$\zeta = \delta \min(1 - m_1, 1 - m_3), \quad (10.102)$$

the proof of part (4) follows from (10.100) and (10.101). We further notice that, as we will specify shortly, the value of δ is chosen to be merely dependent on α and K .

Proof of part (5): the proof for regions 2 and 4 requires different techniques than the ones presented above. In fact, one can note that in the above justifications we did not use the fact that $\alpha < \alpha_{\text{cUCP}}$. Here, we use this assumption. We also adjust the value of δ in a suitable way. In the sequel, we mainly talk about region 4 and note that the same reasoning also holds for region 2.

For a properly chosen δ , we intend to show that the size of R_4 is at most $2w$. For $i \geq p$ we define $\bar{\ell}_i$ as

$$\bar{\ell}_i = \frac{1}{w} \sum_{d=0}^{w-1} \ell_{i+d}. \quad (10.103)$$

Also, let us define the set \bar{R}_4 as

$$\bar{R}_4 = \{i > p \mid i - w \in R_4, i \in R_4\}. \quad (10.104)$$

We will now show that for a properly chosen δ the set \bar{R}_4 is always empty. Clearly, this proves that the set R_4 has size at most $2w$. In order to show that \bar{R}_4 is empty, we first assume the contrary, i.e., the set \bar{R}_4 is non-empty, and then reach a contradiction. So assuming \bar{R}_4 has size larger than 1, let i_0 be the smallest position in \bar{R}_4 . By using the functions f and g defined in (10.91) and (10.92) and the conservation equations (10.74) we deduce that

$$\bar{\ell}_i = \frac{1}{w} \sum_{d=0}^{w-1} f\left(\frac{1}{w} \sum_{k=0}^{w-1} g(\bar{\ell}_{i+d-k})\right), \quad \forall i \geq i_0. \quad (10.105)$$

Now, let us define the potential function

$$\bar{\Phi}(\{x_i\}) = \sum_{i \geq i_0} x_i g(x_i) - G(x_i) - F\left(\frac{1}{w} \sum_{k=0}^{w-1} g(x_{i-k})\right), \quad (10.106)$$

where the functions F, G are given as

$$F(x) = \int f(x)dx = \frac{2^K}{\alpha K} \exp\left(\frac{\alpha K}{2^{K-1}}x\right), \quad (10.107)$$

$$G(x) = \int g(x)dx = \frac{2}{K}\left(1 - \frac{x}{2}\right)^K. \quad (10.108)$$

It is easy to check that

$$(10.105) \Rightarrow \frac{\partial \bar{\Phi}}{\partial x_i} \Big|_{\bar{\ell}} = 0, \quad \forall i \geq i_0, \quad (10.109)$$

where by $\frac{\partial \bar{\Phi}}{\partial \bar{\ell}_i} \Big|_{\bar{\ell}}$ we mean the partial derivative of $\bar{\Phi}$ (with respect to x_i) computed at the profile $\{\bar{\ell}_i\}_{i \geq i_0 - w}$. The idea is now to use the relation (10.109) to get a contradiction with the fact that the set \bar{R}_4 is non-empty. In this regard, we define the shifted profile $\{S\bar{\ell}_i\}_{i \geq i_0}$ as

$$S\bar{\ell}_i = \bar{\ell}_{i+1}, \quad \forall i \geq i_0. \quad (10.110)$$

We first note that since $\{\bar{\ell}_i\}_{i \geq i_0}$ is an increasing profile, then for $i \geq i_0$ we have that $S\bar{\ell}_i \geq \bar{\ell}_i$. One can telescopically write

$$\bar{\Phi}(\{S\bar{\ell}_i\}) - \bar{\Phi}(\{\bar{\ell}_i\}) = \bar{\ell}_{i_0}g(\bar{\ell}_{i_0}) - G(\bar{\ell}_{i_0}) + F\left(\frac{1}{w} \sum_{k=0}^{w-1} g(\bar{\ell}_{i_0-k})\right) + o\left(\frac{1}{w}\right). \quad (10.111)$$

where the last (small) term comes from the difference of the two profiles at the very right end (we know from the second part of this lemma that at the right end both profiles are doubly exponentially close to 2). Now, note that since \bar{R}_4 is assumed to be non-empty, then from the definition of \bar{R}_4 it is easy to see that

$$\bar{\ell}_i \in [\ell_u - \delta, \ell_u + \delta], \quad \forall i \in \{i_0 - w, \dots, i_0\}. \quad (10.112)$$

Also, we further notice that the functions f, F, g, G have uniformly bounded first and second derivatives inside the interval $[0, 2]$, i.e., the value

$$\theta = \max_{x \in [0, 2]} \max\{|f''(x)|, |f'(x)|, |F'(x)|, |g''(x)|, |g'(x)|, |G'(x)|\}, \quad (10.113)$$

is finite. Finally by (10.111), (10.112) and (10.113) we conclude that

$$\bar{\Phi}(\{S\bar{\ell}_i\}) - \bar{\Phi}(\{\bar{\ell}_i\}) \geq \ell_u g(\ell_u) - G(\ell_u) + F(g(\ell_u)) - (1 + 2\theta + \theta + \theta^2)\delta + o\left(\frac{1}{w}\right). \quad (10.114)$$

Now, as $\alpha < \alpha_{\text{cUCP}}$, we obtain that

$$\ell_u g(\ell_u) - G(\ell_u) + F(g(\ell_u)) \geq \ell_s g(\ell_s) - G(\ell_s) + F(g(\ell_s)) \triangleq \Delta(\alpha, K) > 0. \quad (10.115)$$

Thus by fixing δ to be

$$\delta = \delta(\alpha, K) \triangleq \frac{\Delta}{2(1 + 2\theta + \theta + \theta^2)}, \quad (10.116)$$

we conclude from (10.116), (10.115) and (10.114) that

$$\bar{\Phi}(\{S\bar{\ell}_i\}) - \bar{\Phi}(\{\bar{\ell}_i\}) \geq \frac{\Delta}{2} + o\left(\frac{1}{w}\right). \quad (10.117)$$

Note here that the choice of δ is merely dependent on Δ and θ which are positive static parameters merely dependent on α and K , and hence δ is itself a static parameter independent of the dynamics of the UC algorithm.

We now show that (10.117) cannot be true provided that the conservation equations (10.109) hold. By using the mean-value theorem we know that there exists a profile $\{\ell_i^*\}_{i \geq i_0}$ such that for $i \geq i_0 - w$ we have $\bar{\ell}_i \leq \ell_i^* \leq \bar{\ell}_{i+1}$ and

$$\bar{\Phi}(\{S\bar{\ell}_i\}) - \bar{\Phi}(\{\bar{\ell}_i\}) = \sum_{i \geq i_0} \frac{\partial \bar{\Phi}}{\partial x_i} \Big|_{\bar{\ell}} (\bar{\ell}_{i+1} - \bar{\ell}_i) + \sum_{i, j \geq i_0} \frac{\partial^2 \bar{\Phi}}{\partial x_i \partial x_j} \Big|_{\ell^*} (\bar{\ell}_{i+1} - \bar{\ell}_i)(\bar{\ell}_{j+1} - \bar{\ell}_j). \quad (10.118)$$

Here, by $\frac{\partial^2 \bar{\Phi}}{\partial x_i \partial x_j} \Big|_{\ell^*}$ we mean the partial second derivative of $\bar{\Phi}$ given in (10.106) (with respect to x_i and x_j) computed at the profile $\{\ell_i^*\}_{i \geq i_0 - w}$. Now, by using (10.109) we conclude that

$$\bar{\Phi}(\{S\bar{\ell}_i\}) - \bar{\Phi}(\{\bar{\ell}_i\}) = \sum_{i, j \geq i_0} \frac{\partial^2 \bar{\Phi}}{\partial x_i \partial x_j} \Big|_{\ell^*} (\bar{\ell}_{i+1} - \bar{\ell}_i)(\bar{\ell}_{j+1} - \bar{\ell}_j). \quad (10.119)$$

We now bound the right-hand side of (10.119). For this purpose let us decompose the function $\bar{\Phi}$ given in (10.106) into two parts

$$\bar{\Phi}(\{x_i\}) = \underbrace{\sum_{i \geq i_0} x_i g(x_i) - G(x_i)}_{\bar{\Phi}_1} - \underbrace{\sum_{i \geq i_0} F\left(\frac{1}{w} \sum_{k=0}^{w-1} g(x_{i-k})\right)}_{\bar{\Phi}_2}. \quad (10.120)$$

Consequently,

$$\frac{\partial^2 \bar{\Phi}}{\partial x_i \partial x_j} = \frac{\partial^2 \bar{\Phi}_1}{\partial x_i \partial x_j} - \frac{\partial^2 \bar{\Phi}_2}{\partial x_i \partial x_j}. \quad (10.121)$$

Consider the value θ defined in (10.113). It is easy to see that

$$\frac{\partial^2 (x_i g(x_i) - G(x_i))}{\partial x_i \partial x_j} \leq 5\theta \mathbb{1}\{i = j\}. \quad (10.122)$$

Hence, we conclude that

$$\sum_{i, j \geq i_0} \frac{\partial^2 \bar{\Phi}_1}{\partial x_i \partial x_j} \Big|_{\ell^*} (\bar{\ell}_{i+1} - \bar{\ell}_i)(\bar{\ell}_{j+1} - \bar{\ell}_j) \leq 5\theta \sum_{i \geq i_0} (\bar{\ell}_{i+1} - \bar{\ell}_i)(\bar{\ell}_{i+1} - \bar{\ell}_i)$$

$$\begin{aligned}
&\stackrel{(a)}{\leq} \frac{2}{w} 5\theta \sum_{i \geq i_0} (\bar{\ell}_{i+1} - \bar{\ell}_i) \\
&\leq \frac{10\theta}{w} (2 - \bar{\ell}_{i_0}) \\
&\leq \frac{20\theta}{w}. \tag{10.123}
\end{aligned}$$

Here, step (a) follows from the fact that by definition of $\bar{\ell}_i$ given in (10.103), we can easily see that for any $i \geq i_0$ we have that $\bar{\ell}_{i+1} - \bar{\ell}_i \leq \frac{2}{w}$. Also, one can write

$$\begin{aligned}
&\frac{\partial}{\partial x_j} \frac{\partial}{\partial x_i} F\left(\frac{1}{w} \sum_{k=0}^{w-1} g(x_{i-k})\right) \\
&= \frac{g''(x_i)}{w} F'\left(\frac{1}{w} \sum_{k=0}^{w-1} g(x_{i-k})\right) \mathbb{1}\{i=j\} + \frac{g'(x_i)g'(x_j)}{w^2} F''\left(\frac{1}{w} \sum_{k=0}^{w-1} g(x_{i-k})\right) \mathbb{1}\{i-j < w\} \\
&\leq \frac{\theta^2}{w} \mathbb{1}\{i=j\} + \frac{\theta^3}{w^2} \mathbb{1}\{i-j < w\}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\sum_{i,j \geq i_0} \frac{\partial^2 \bar{\Phi}_2}{\partial x_i \partial x_j} \Big|_{\ell^*} (\bar{\ell}_{i+1} - \bar{\ell}_i)(\bar{\ell}_{j+1} - \bar{\ell}_j) \\
&\leq \sum_{i \geq i_0} \frac{\theta^2}{w} (\bar{\ell}_{i+1} - \bar{\ell}_i)^2 + \sum_{i \geq i_0} \sum_{j=i-w}^i \frac{\theta^3}{w^2} (\bar{\ell}_{i+1} - \bar{\ell}_i)(\bar{\ell}_{j+1} - \bar{\ell}_j) \\
&\leq \frac{2\theta^2}{w^2} \sum_{i \geq i_0} (\bar{\ell}_{i+1} - \bar{\ell}_i) + \frac{\theta^3}{w^2} \sum_{i \geq i_0} (\bar{\ell}_{i+1} - \bar{\ell}_i)(\bar{\ell}_{i+1} - \bar{\ell}_{i-w}) \\
&\leq \frac{4\theta^2}{w^2} + \frac{4\theta^3}{w^2}. \tag{10.124}
\end{aligned}$$

Finally, as a consequence of (10.119), (10.121), (10.123) and (10.124) we get that

$$\bar{\Phi}(\{S\bar{\ell}_i\}) - \bar{\Phi}(\{\bar{\ell}_i\}) \leq \frac{20\theta + \frac{4\theta^2}{w^2} + \frac{4\theta^3}{w^2}}{w}. \tag{10.125}$$

Hence, by choosing

$$w > w_0(\alpha, K) \triangleq \frac{4(20\theta + 4\theta^2 + 4\theta^3)}{\Delta}, \tag{10.126}$$

we get that

$$\bar{\Phi}(\{S\bar{\ell}_i\}) - \bar{\Phi}(\{\bar{\ell}_i\}) < \frac{\Delta}{4}. \tag{10.127}$$

The above inequality contradicts (10.117). Hence, the assumption that \bar{R}_4 is non-empty contradicts the conservation equation (10.109) and part (5) is proved.

Proof of part (6): our objective here is to find a suitable candidate for the constants c_1 and c_2 such that the result of part (5) holds. Let us first prove part (5) for $\epsilon = \epsilon_0$ given as

$$\epsilon_0 = \min\left\{1, \left(\frac{2^{2K-3}}{\alpha K}\right)^{\frac{1}{K-2}}\right\} \implies \frac{\alpha K}{2^{2K-2}}(\epsilon_0)^{K-2} \leq \frac{1}{2}. \quad (10.128)$$

We first show that there exists a constant c_0 such that $I_{\epsilon_0} \leq c_0 w$. We consider two cases. In the first case, we assume that $\ell_u + \delta \geq 2 - \epsilon_0$. In this case, it is easy to see that $I_{\epsilon_0} - 1$ falls in one of the regions $R_1 - R_4$. Now, from parts (4)-(5) we have

$$|R_1| + |R_2| + |R_3| + |R_4| \leq \zeta w + 2w + \zeta w + 2w = (4 + 2\zeta)w. \quad (10.129)$$

Hence, by choosing

$$c_0 \geq (4 + 2\zeta) + 1, \quad (10.130)$$

we deduce the $I_{\epsilon_0} \leq c_0 w$ for the first case. In the second case, we assume that $\ell_u + \delta < 2 - \epsilon_0$. In this case I_{ϵ_0} falls inside the region R_5 . Define the region of positions D as

$$D = \{i \geq p \mid \ell_u + \delta \leq \ell_i \leq 2 - \epsilon_0\}. \quad (10.131)$$

We further have

$$I_{\epsilon_0} \leq |R_1| + |R_2| + |R_3| + |R_4| + |D| \leq (4 + 2\zeta)w + |D|. \quad (10.132)$$

Here, we note that exactly as in the proof of part (4), there exists a constant $r = r(\alpha, K)$ such that

$$|D| \leq r w. \quad (10.133)$$

As a result, it is clear from (10.132) and (10.133) that we have for both of the above cases

$$I_{\epsilon_0} \leq (2\zeta + 5 + r)w. \quad (10.134)$$

Let us now prove part (6) for an arbitrary choice of ϵ . If $\epsilon > \epsilon_0$, then (10.134) is also a bound for I_ϵ because I_ϵ is clearly decreasing in ϵ . Assume now that $\epsilon < \epsilon_0$ and consider the region of positions E defined as

$$E = \{i \geq p \mid 2 - \epsilon_0 \leq \ell_i \leq 2 - \epsilon\}. \quad (10.135)$$

We have from (10.134)

$$I_\epsilon = |R_1| + I_{\epsilon_0} + |E| \leq (5 + 2\zeta + r)w + |E|. \quad (10.136)$$

It thus remains to find an upper bound on the size of E . Consider a position $i \in E$, we now note that for $m \in \mathbb{N}$

$$2 - \ell_{i+mw} \leq (2 - \ell_i)^{(K-1)^m} \left(\frac{\alpha K}{2^{2K-2}}\right)^{(K-1)^{m-1}}.$$

This relation follows easily by m times using the inequality in part (2) of the lemma. Using this bound, we have

$$\begin{aligned}
2 - \ell_{i+mw} &\leq (2 - \ell_i)^{(K-1)^m} \left(\frac{\alpha K}{2^{2K-2}}\right)^{(K-1)^{m-1}} \\
&= (2 - \ell_i)^{(K-1)^{m-1}} \left(\frac{\alpha K}{2^{2K-2}}(2 - \ell_i)\right)^{(K-1)^m - (K-1)^{m-1}} \\
&\stackrel{(a)}{\leq} \left(\frac{\alpha K}{2^{2K-2}}(2 - \ell_i)\right)^{(K-1)^m - (K-1)^{m-1}} \\
&\stackrel{(b)}{\leq} \left(\frac{1}{2}\right)^{(K-1)^m - (K-1)^{m-1}} \\
&\stackrel{(c)}{\leq} \left(\frac{1}{2}\right)^{2^{m-1}}. \tag{10.137}
\end{aligned}$$

Here, step (a) follows from the fact that by using (10.128) and (10.135) we deduce that for $i \in E$ we have $\ell_i \geq 1$ and hence $2 - \ell_i \leq 1$. Again by using (10.128) and (10.135) we deduce that for $i \in E$ we have $\frac{\alpha K}{2^{2K-2}}(2 - \ell_i) \leq \frac{1}{2}$ and hence step (b) follows. Also, step (c) follows from the fact that $K \geq 3$. Now, for a position $i \in E$, it is apparent from (10.137) that if we choose

$$m = \lceil 1 + \log_2(\log_2 \frac{1}{\epsilon}) \rceil \implies \left(\frac{1}{2}\right)^{2^{m-1}} \leq \epsilon,$$

then we have $2 - \ell_{i+mw} < \epsilon$ and hence $i + mw \notin E$. Therefore we have

$$|E| \leq \lceil w(1 + \log_2(\log_2 \frac{1}{\epsilon})) \rceil,$$

and as a result we obtain from (10.136) that

$$\begin{aligned}
I_\epsilon &\leq w(5 + 2\zeta + r + 2 + \log_2(\log_2 \frac{1}{\epsilon})) \\
&\leq \underbrace{w(7 + 2\zeta + r - \frac{\log(\log 2)}{\log 2})}_{\triangleq c_1(\alpha, K)} + \underbrace{\frac{1}{\log 2}}_{\triangleq c_2(\alpha, K)} \log(\log \frac{1}{\epsilon}).
\end{aligned}$$

Proof of Lemma 10.4

Let us first explain what is the main approach behind the proof. Consider the vector $\vec{d} = \{d_i\}_{0 \leq i \leq L+w-1}$ defined as follows. For $i \geq 0$,

$$d_i = w(\ell_{i+1} - \ell_i). \tag{10.138}$$

We first note that all the entries of the vector d are strictly positive. In what follows, we show that there exists a constant $\delta > 0$, such that the vector $A^n d$ decays to the all-zero vector faster than $(1 - \delta)^n$. Using this, we conclude that the largest eigenvalue of the matrix A should be less than $(1 - \delta)$.

We begin by defining for $n \in \mathbb{N}$, the *ratio profile* $\bar{\gamma}_n = \{\gamma_{n,i}\}_{0 \leq i \leq L+w-1}$ as

$$\gamma_{n,i} = \begin{cases} 0 & i < p, \\ \frac{(A^n d)_i}{d_i} & i \geq p. \end{cases} \quad (10.139)$$

where by $(A^n d)_i$ we mean the i -th entry of the vector $A^n d$. Also for $n = 0$, we let

$$\gamma_{0,i} = \mathbb{1}_{\{i \geq p\}}. \quad (10.140)$$

The proof consists of four steps. For technical reasons that will be clear later on, we begin by confining our analysis to positions i that are inside a specific region defined as follows. Let $\epsilon > 0$ be a (small) constant, the value of which we will specify later in the proof. Define the integer T_ϵ as

$$T_\epsilon = \operatorname{argmin}\{i \geq p \mid \forall j \geq i : \ell_{j-w} \geq 2 - \epsilon\}. \quad (10.141)$$

We first confine our analysis to the region $0 \leq i \leq T_\epsilon$. Note here that the smaller we make ϵ , the larger the value of T_ϵ will be. Also, in the last step of the proof, we show that by choosing a sufficiently small ϵ , all the final results are valid for the complete region $0 \leq i \leq L + w - 1$.

Step 1: Recursive bounds for the ratio profiles: In this step, we intend to bound the ratio profile $\bar{\gamma}_{n+1}$, in terms of the profile $\bar{\gamma}_n$.

We first write

$$\begin{aligned} (A^{n+1}d)_i &= (A(A^n d))_i \\ &= \sum_{j=0}^{L+w-1} A_{i,j} (A^n d)_j \\ &= \sum_{j=0}^{L+w-1} A_{i,j} \gamma_{n,j} d_j. \end{aligned}$$

Now, on one hand, by using (10.53) we have for $i \geq p$,

$$\begin{aligned} (A^{n+1}d)_i &= \ell_i \frac{\alpha K(K-1)}{2^K w} \sum_{j=i-w+1}^{i+w-1} \frac{\gamma_{n,j} d_j}{w} \sum_{k=|j-i|}^{w-1} \left(1 - \frac{1}{w} \sum_{d=0}^{w-1} \frac{\ell_{\max(i,j)-k+d}}{2}\right)^{K-2} \\ &= \ell_i \frac{\alpha K(K-1)}{2^K} \frac{1}{w} \sum_{k=0}^{w-1} \left(\sum_{d=0}^{w-1} \gamma_{n,i-k+d} d_{i-k+d}\right) \left(1 - \frac{1}{w} \sum_{d=0}^{w-1} \frac{\ell_{i-k+d}}{2}\right)^{K-2}. \end{aligned} \quad (10.142)$$

On the other hand, from the conservation equations (10.74) we can write

$$\begin{aligned} \ln \frac{\ell_{i+1}}{\ell_i} &= \frac{\alpha K}{2^{K-1}} \frac{1}{w} \sum_{k=0}^{w-1} \left\{ \left(1 - \frac{1}{w} \sum_{d=0}^{w-1} \frac{\ell_{i-k+d}}{2}\right)^{K-1} - \left(1 - \frac{1}{w} \sum_{d=0}^{w-1} \frac{\ell_{i+1-k+d}}{2}\right)^{K-1} \right\} \\ &\quad + \int_0^t 2\delta_{pi} \ell_i^{-1} dt, \end{aligned} \quad (10.143)$$

We now intend to bound the expression of (10.143). To simplify notation, let us first define

$$x_{i,k} = 1 - \frac{1}{w} \sum_{d=0}^{w-1} \frac{\ell_{i-k+d}}{2},$$

$$t_{i,k} = x_{i,k} - x_{i+1,k} = \frac{1}{w} \sum_{d=0}^{w-1} \frac{\ell_{i+1-k+d} - \ell_{i-k+d}}{2} = \frac{1}{w^2} \sum_{d=0}^{w-1} \frac{d_{i-k+d}}{2}.$$

We first note that for $i \leq T_\epsilon$ we have

$$\begin{aligned} \frac{t_{i,k}}{x_{i,k}} &= \frac{\sum_{d=0}^{w-1} \frac{\ell_{i+1-k+d} - \ell_{i-k+d}}{2}}{\sum_{d=0}^{w-1} \frac{2 - \ell_{i-k+d}}{2}} \\ &= \frac{\frac{\ell_{i-k+w} - \ell_{i-k}}{2}}{\sum_{d=0}^{w-1} \frac{2 - \ell_{i-k+d}}{2}} \\ &\leq \frac{2}{w\epsilon}, \end{aligned}$$

where the last step follows from the definition of T_ϵ in (10.141) and the fact that $i \leq T_\epsilon$. Let us assume here that the value of w is chosen large enough so that

$$w \geq \frac{2^K}{\epsilon}. \quad (10.144)$$

Then, we clearly have $\frac{t_k}{x_k} < 1$ and after some simple manipulation we can write

$$\begin{aligned} &\frac{1}{w} \sum_{k=0}^{w-1} x_{i,k}^{K-1} - (x_{i,k} - t_k)^{K-1} \\ &\geq \frac{1}{w} \sum_{k=0}^{w-1} x_k^{K-2} t_{i,k} \left(1 - \frac{2^K}{w\epsilon}\right). \end{aligned} \quad (10.145)$$

As a result of inequalities (10.143)-(10.145) we obtain for $i \leq T_\epsilon$

$$\begin{aligned} &\ln \frac{\ell_{i+1}}{\ell_i} \\ &\geq \frac{\alpha K(K-1)}{2^K} \left(1 - \frac{2^K}{w\epsilon}\right) \frac{1}{w} \sum_{k=0}^{w-1} \left(\frac{1}{w} \sum_{d=0}^{w-1} d_{i-k+d}\right) \left(1 - \frac{1}{w} \sum_{d=0}^{w-1} \frac{\ell_{i-k+d}}{2}\right)^{K-2} \\ &\quad + \int_0^t 2\delta_{pi} \ell_i^{-1} dt. \end{aligned} \quad (10.146)$$

Now, using the fact that $\frac{\ell_{i+1}}{\ell_i} = 1 + \frac{d_i}{w\ell_i}$ and the inequality $\ln(1+x) \leq x$, we have

$$\ln \frac{\ell_{i+1}}{\ell_i} \leq \frac{d_i}{w\ell_i}. \quad (10.147)$$

This inequality together with (10.146) yields

$$d_i \geq \frac{\alpha K(K-1)}{2^K} \left(1 - \frac{2^K}{w\epsilon}\right) \frac{1}{w} \sum_{k=0}^{w-1} \left(\sum_{d=0}^{w-1} d_{i-k+d}\right) \left(1 - \frac{1}{w} \sum_{d=0}^{w-1} \frac{\ell_{i-k+d}}{2}\right)^{K-2} + w\ell_i \int_0^t 2\delta_{pi}\ell_i^{-1} dt. \quad (10.148)$$

Finally, by using (10.142) and (10.148), we obtain for $0 \leq i \leq T_\epsilon$

$$\gamma_{n+1,i} \leq \frac{1}{\left(1 - \frac{2^K}{w\epsilon}\right)} \frac{\sum_{k=0}^{w-1} \left(1 - \frac{1}{w} \sum_{d=0}^{w-1} \frac{\ell_{i-k+d}}{2}\right)^{K-2} \left(\sum_{d=0}^{w-1} \gamma_{n,i-k+d} d_{i-k+d}\right)}{\sum_{k=0}^{w-1} \left(1 - \frac{1}{w} \sum_{d=0}^{w-1} \frac{\ell_{i-k+d}}{2}\right)^{K-2} \left(\sum_{d=0}^{w-1} d_{i-k+d}\right)} \quad (10.149)$$

Step2: Upper bounds on $\bar{\gamma}_1$ and $\bar{\gamma}_2$: As we see in the following, and this is also confirmed by numerical experiments, the behavior of the ration profile $\bar{\gamma}_1$ depends heavily on the value of ℓ_p . Consequently, we consider the following two cases which depend on the value of ℓ_p and provide an upper bound on $\bar{\gamma}_1$ for each case. We then use this upper bound together with the recursion (10.149) to provide an upper bound for $\bar{\gamma}_2$ (independent of the cases). We then use this upper bound in later steps of the proof.

Case1: We assume in this case that

$$\ell_{p+1} \leq 3\ell_p. \quad (10.150)$$

By using the recursion (10.149) and the fact that $\bar{\gamma}_0$ is the profile given in (10.140), we obtain after some simple manipulations that

$$\begin{aligned} \gamma_{1,i} &\stackrel{(a)}{\leq} \frac{1}{\left(1 - \frac{2^K}{w\epsilon}\right)} \left(1 - \frac{\sum_{k=0}^{w-1} \left(1 - \frac{1}{w} \sum_{d=0}^{w-1} \frac{\ell_{i-k+d}}{2}\right)^{K-2} \left(\sum_{d=0}^{w-1} d_{i-k+d} \mathbb{1}_{\{i-k+d < 0\}}\right)}{\sum_{k=0}^{w-1} \left(1 - \frac{1}{w} \sum_{d=0}^{w-1} \frac{\ell_{i-k+d}}{2}\right)^{K-2} \left(\sum_{d=0}^{w-1} d_{i-k+d}\right)}\right) \\ &\stackrel{(b)}{=} \frac{1}{\left(1 - \frac{2^K}{w\epsilon}\right)} \left(1 - \frac{\sum_{k=i-p+1}^{w-1} \left(1 - \frac{1}{w} \sum_{d=0}^{w-1} \frac{\ell_{i-k+d}}{2}\right)^{K-2} w\ell_p}{\sum_{k=0}^{w-1} \left(1 - \frac{1}{w} \sum_{d=0}^{w-1} \frac{\ell_{i-k+d}}{2}\right)^{K-2} \left(\sum_{d=0}^{w-1} d_{i-k+d}\right)}\right) \\ &\stackrel{(c)}{\leq} \frac{1}{\left(1 - \frac{2^K}{w\epsilon}\right)} \left(1 - \frac{w-1-(i-p)}{2w} \ell_p\right) \\ &\stackrel{(d)}{\leq} \frac{1}{\left(1 - \frac{2^K}{w\epsilon}\right)} \left(1 - \frac{w-1-(i-p)}{6w} \ell_{p+1}\right) \\ &\stackrel{(e)}{\leq} \frac{1}{\left(1 - \frac{2^K}{w\epsilon}\right)} \left(1 - \frac{w-1-(i-p)}{3w} \exp\left(-\frac{\alpha K}{2^{K-1}}\right)\right). \end{aligned} \quad (10.151)$$

Here, the relation (a) follows from (10.149) and (10.140). The relation (b) follows from the fact that $d_{p-1} = w\ell_p$ and $d_j = 0$ for $j < p-1$. The relation (c) follows from the inequality $\sum_{d=0}^{w-1} d_{i-k+d} \leq 2w$ and also from the fact that

the sequence

$$y_k = \left(1 - \frac{1}{w} \sum_{d=0}^{w-1} \frac{\ell_{i-k+d}}{2}\right)^{K-2}$$

is an increasing sequence in k , and hence,

$$\frac{\sum_{k=i-p+1}^{w-1} y_k}{\sum_{k=0}^{w-1} y_k} \geq \frac{w-1-(i-p)}{w}.$$

The relation (d) follows from (10.150) and finally the relation (e) follows from the first part of Lemma 10.3. As a result of the above series of inequalities, we deduce the following. For case 1, we have

$$\gamma_{1,i} \leq \frac{1}{\left(1 - \frac{2^K}{w\epsilon}\right)} \begin{cases} 1 - \frac{1}{12} \exp\left(-\frac{\alpha K}{2^{K-1}}\right) & p \leq i \leq p + \frac{w-1}{2}, \\ 1 & p + \frac{w-1}{2} < i \leq T_\epsilon. \end{cases} \quad (10.152)$$

A moment of thought reveals that with exactly the same argument as above we can also show that

$$\gamma_{2,i} \leq \frac{1}{\left(1 - \frac{2^K}{w\epsilon}\right)^2} \begin{cases} 1 - \frac{1}{12} \exp\left(-\frac{\alpha K}{2^{K-1}}\right) & p \leq i \leq p + \frac{w-1}{2}, \\ 1 & p + \frac{w-1}{2} < i \leq T_\epsilon. \end{cases} \quad (10.153)$$

Case 2: Contrary to the first case, in this case we assume that

$$\ell_{p+1} > 3\ell_p. \quad (10.154)$$

As a result, we have that

$$\frac{\ell_{p+1} - \ell_p}{\ell_p} \geq 2.$$

Now, by using the fact that for $x \geq 2$ we have $\ln(1+x) \leq x - \frac{1}{2}$, we obtain that

$$\ln \frac{\ell_{p+1}}{\ell_p} = \ln\left(1 + \frac{\ell_{p+1} - \ell_p}{\ell_p}\right) \leq \frac{\ell_{p+1} - \ell_p}{\ell_p} - \frac{1}{2},$$

and hence

$$\ln \frac{\ell_{p+1}}{\ell_p} \leq \frac{d_p}{w\ell_p} - \frac{1}{2}. \quad (10.155)$$

Now, by using (10.155) and (10.146) we obtain

$$d_p \geq \frac{w\ell_p}{2} + \ell_p \frac{\alpha K(K-1)}{2^K} \left(1 - \frac{2^K}{w\epsilon}\right) \frac{1}{w} \sum_{k=0}^{w-1} \left(\sum_{d=0}^{w-1} d_{i-k+d}\right) \left(1 - \frac{1}{w} \sum_{d=0}^{w-1} \frac{\ell_{i-k+d}}{2}\right)^{K-2}. \quad (10.156)$$

Now, by using (10.156) and (10.142) we reach to the following bound on $\gamma_{1,p}$

$$\gamma_{1,p} \leq \frac{1}{\left(1 - \frac{2^K}{w\epsilon}\right)} \frac{\sum_{k=0}^{w-1} \left(\sum_{d=0}^{w-1} \gamma_{0,i-k+d} d_{i-k+d}\right) \left(1 - \frac{1}{w} \sum_{d=0}^{w-1} \frac{\ell_{i-k+d}}{2}\right)^{K-2}}{\frac{w}{2} + \sum_{k=0}^{w-1} \left(\sum_{d=0}^{w-1} d_{i-k+d}\right) \left(1 - \frac{1}{w} \sum_{d=0}^{w-1} \frac{\ell_{i-k+d}}{2}\right)^{K-2}}$$

$$\begin{aligned} & \stackrel{(a)}{\leq} \frac{1}{\left(1 - \frac{2^K}{w\epsilon}\right)} \frac{2w}{\frac{w}{2} + 2w} \\ & = \frac{4}{5\left(1 - \frac{2^K}{w\epsilon}\right)}, \end{aligned}$$

where (a) follows from the fact that $\sum_{d=0}^{w-1} \gamma_{0,i-k+d} d_{i-k+d} \leq \sum_{d=0}^{w-1} d_{i-k+d} \leq 2$. Therefore, for the second case, the ratio profile $\bar{\gamma}_0$ is bounded from above as follows

$$\gamma_{1,i} \leq \frac{1}{\left(1 - \frac{2^K}{w\epsilon}\right)} \begin{cases} \frac{4}{5} & i = p, \\ 1 & p < i \leq T_\epsilon. \end{cases} \quad (10.157)$$

We now find an upper bound on the profile $\bar{\gamma}_2$ for this case. As the details of how we reach such an upper bound are exactly the same as in (10.158), we omit them here to avoid repeating long expressions. We have

$$\begin{aligned} \gamma_{2,i} & \leq \\ & \frac{1}{\left(1 - \frac{2^K}{w\epsilon}\right)} \left(\frac{\sum_{k=0}^{w-1} \left(1 - \frac{1}{w} \sum_{d=0}^{w-1} \frac{\ell_{i-k+d}}{2}\right)^{K-2} \left(\sum_{d=0}^{w-1} d_{i-k+d} \gamma_{1,i-k+d}\right)}{\sum_{k=0}^{w-1} \left(1 - \frac{1}{w} \sum_{d=0}^{w-1} \frac{\ell_{i-k+d}}{2}\right)^{K-2} \left(\sum_{d=0}^{w-1} d_{i-k+d}\right)} \right) \\ & \stackrel{(a)}{\leq} \frac{1}{\left(1 - \frac{2^K}{w\epsilon}\right)^2} \left(1 - \frac{\sum_{k=0}^{w-1} \left(1 - \frac{1}{w} \sum_{d=0}^{w-1} \frac{\ell_{i-k+d}}{2}\right)^{K-2} \sum_{d=0}^{w-1} \frac{1}{5} d_p \mathbb{1}_{\{i-k+d=p\}}}{\sum_{k=0}^{w-1} \left(1 - \frac{1}{w} \sum_{d=0}^{w-1} \frac{\ell_{i-k+d}}{2}\right)^{K-2} \left(\sum_{d=0}^{w-1} d_{i-k+d}\right)} \right) \\ & \stackrel{(b)}{\leq} \frac{1}{\left(1 - \frac{2^K}{w\epsilon}\right)^2} \left(1 - \frac{\sum_{k=i-p}^{w-1} \left(1 - \frac{1}{w} \sum_{d=0}^{w-1} \frac{\ell_{i-k+d}}{2}\right)^{K-2} \frac{2}{15} w \ell_{p+1}}{\sum_{k=0}^{w-1} \left(1 - \frac{1}{w} \sum_{d=0}^{w-1} \frac{\ell_{i-k+d}}{2}\right)^{K-2} \left(\sum_{d=0}^{w-1} d_{i-k+d}\right)} \right) \\ & \stackrel{(c)}{\leq} \frac{1}{\left(1 - \frac{2^K}{w\epsilon}\right)^2} \left(1 - \frac{w - (i-p)}{2w} \ell_p \right) \\ & \stackrel{(d)}{\leq} \frac{1}{\left(1 - \frac{2^K}{w\epsilon}\right)^2} \left(1 - \frac{w - (i-p)}{w} \frac{4}{15} \ell_{p+1} \right) \\ & \stackrel{(e)}{\leq} \frac{1}{\left(1 - \frac{2^K}{w\epsilon}\right)^2} \left(1 - \frac{w - (i-p)}{w} \frac{4}{15} \exp\left(-\frac{\alpha K}{2^{K-1}}\right) \right). \end{aligned} \quad (10.158)$$

Here, step (a) follows from (10.157). Step (b) follows from (10.154). The other steps follow exactly similar to the proof of (10.158) and hence we omit further explanations about them. As a result of the above set of inequalities, we have

$$\gamma_{2,i} \leq \frac{1}{\left(1 - \frac{2^K}{w\epsilon}\right)^2} \begin{cases} 1 - \frac{2}{15} \exp\left(-\frac{\alpha K}{2^{K-1}}\right) & p \leq i \leq p + \frac{w-1}{2}, \\ 1 & p + \frac{w-1}{2} < i \leq T_\epsilon. \end{cases} \quad (10.159)$$

Let us now finalize the result of Step 2 of the proof. By using (10.153) and (10.159) we deduce that the ratio profile $\bar{\gamma}_2$ is bounded above as follows:

$$\gamma_{2,i} \leq \frac{1}{\left(1 - \frac{2^K}{w\epsilon}\right)^2} \begin{cases} 1 - \frac{1}{12} \exp\left(-\frac{\alpha K}{2^{K-1}}\right) & p \leq i \leq p + \frac{w-1}{2}, \\ 1 & p + \frac{w-1}{2} < i \leq T_\epsilon. \end{cases} \quad (10.160)$$

Step3: A recursive bound on $\bar{\gamma}_n$: In this step we prove the following bound on $\bar{\gamma}_n$: for $n \geq 2$, we have

$$\gamma_{n,i} \leq \frac{1}{\left(1 - \frac{2^K}{w\epsilon}\right)^n} \begin{cases} 1 - c_1 c_2^{n-2} & p \leq i \leq p + (n-1)\frac{w-1}{4}, \\ 1 & p + (n-1)\frac{w-1}{4} < i \leq T_\epsilon, \end{cases} \quad (10.161)$$

where the constants c_3 and c_4 , that depend on α and K , are given as

$$c_3 = \frac{1}{12} \exp\left(-\frac{\alpha K}{2^{K-1}}\right), \quad (10.162)$$

$$c_4 = \frac{1}{8c}, \quad (10.163)$$

and also the constant c is given in Lemma 10.5. We prove this statement by induction on n . For $n = 2$ the result is clear due to (10.160). Let us assume (10.161) holds for $n = m$. We now show that it also holds for $n = m + 1$. Consider a position $p \leq i \leq p + m\frac{w-1}{4}$. We can write

$$\begin{aligned} & \left(1 - \frac{2^K}{w\epsilon}\right)^{m+1} \gamma_{m+1,i} \\ & \stackrel{(a)}{\leq} \frac{\sum_{k=0}^{w-1} \left(1 - \frac{1}{w} \sum_{d=0}^{w-1} \frac{\ell_{i-k+d}}{2}\right)^{K-2} \sum_{d=0}^{w-1} d_{i-k+d} \gamma_{m,i-k+d} \left(1 - \frac{2^K}{w\epsilon}\right)^m}{\sum_{k=0}^{w-1} \left(1 - \frac{1}{w} \sum_{d=0}^{w-1} \frac{\ell_{i-k+d}}{2}\right)^{K-2} \left(\sum_{d=0}^{w-1} d_{i-k+d}\right)} \\ & \stackrel{(b)}{\leq} 1 - \frac{\sum_{k=0}^{w-1} \left(1 - \frac{1}{w} \sum_{d=0}^{w-1} \frac{\ell_{i-k+d}}{2}\right)^{K-2} \sum_{d=0}^{w-1} d_{i-k+d} c_3 c_4^{m-2} \mathbb{1}_{\{i-k+d \leq p + (m-1)\frac{w-1}{4}\}}}{\sum_{k=0}^{w-1} \left(1 - \frac{1}{w} \sum_{d=0}^{w-1} \frac{\ell_{i-k+d}}{2}\right)^{K-2} \left(\sum_{d=0}^{w-1} d_{i-k+d}\right)} \\ & \stackrel{(c)}{\leq} 1 - \frac{\sum_{k=\frac{w-1}{2}}^{w-1} \left(1 - \frac{1}{w} \sum_{d=0}^{w-1} \frac{\ell_{i-k+d}}{2}\right)^{K-2} \sum_{d=0}^{\frac{w-1}{4}} d_{i-k+d} c_3 c_4^{m-2}}{\sum_{k=0}^{w-1} \left(1 - \frac{1}{w} \sum_{d=0}^{w-1} \frac{\ell_{i-k+d}}{2}\right)^{K-2} \left(\sum_{d=0}^{w-1} d_{i-k+d}\right)} \\ & \stackrel{(d)}{\leq} 1 - c_3 c_4^{m-2} \frac{1}{4c} \frac{\sum_{k=\frac{w-1}{2}}^{w-1} \left(1 - \frac{1}{w} \sum_{d=0}^{w-1} \frac{\ell_{i-k+d}}{2}\right)^{K-2}}{\sum_{k=0}^{w-1} \left(1 - \frac{1}{w} \sum_{d=0}^{w-1} \frac{\ell_{i-k+d}}{2}\right)^{K-2}} \\ & \leq 1 - c_3 c_4^{m-2} \frac{1}{8c} = 1 - c_3 c_4^{m-1}. \end{aligned}$$

Here, step (a) follows from (10.149). Step (b) follows from the induction hypothesis and (10.161). Step (c) follows from the fact that $i \leq p + m\frac{w-1}{4}$. Step (d) follows by noticing from Lemma 10.5 that for integers $k_1 > k_2$ we have

$$\sum_{d=0}^{w-1} d_{i-k_1+d} \leq 4c \sum_{d=0}^{w-1} d_{i-k_2+d}.$$

Finally, step (e) follows from the fact that the sequence

$$y_k = \left(1 - \frac{1}{w} \sum_{d=0}^{w-1} \frac{\ell_{i-k+d}}{2}\right)^{K-2}$$

is an increasing sequence in k .

Step 4: Putting things together: Finally in this step we will complete the proof of the lemma. The main tools that we use here are the bound (10.161) and the following facts deduced from the the Perron-Frobenius formalism [129]. Consider an $r \times r$ matrix $X = [X_{i,j}]_{1 \leq i,j \leq r}$ with non-negative entries.

(F1) There exist a number $\lambda_X \geq 0$ such that λ_X is itself an eigenvalue of X and any other eigenvalue λ of X (possibly complex) is smaller than λ_X in absolute value, $|\lambda| \leq \lambda_X$. We call λ_X the *largest eigenvalue* of X .

(F2) We have

$$\lambda_X \leq \max_{1 \leq i \leq r} \sum_{j=1}^r X_{i,j}. \quad (10.164)$$

(F3) In addition, if X is symmetric, the value of λ_X can be computed as follows

$$\lambda_X = \max_{\{y=(y_1, \dots, y_r) \text{ s.t. } \forall j: y_j \geq 0\}} \frac{y^T X y}{y^T y}. \quad (10.165)$$

(F4) If there exist integers $m \in \mathbb{N}$ such that all the entries of X^m are strictly positive, then λ_X is a simple eigenvalue of X . If such an assumption holds, we let v_X denote the eigenvector corresponding to λ_X .

(F5) With the assumptions of (F4), let d be a vector of size r such that $d_j > 0$ for $1 \leq j \leq r$. Then, there exists a constant $e > 0$ such that for $n \in \mathbb{N}$ we have

$$X^n d = e \lambda_X^n v_X + o(r \lambda_X^n). \quad (10.166)$$

We now proceed with the proof. Let B be a $(L + w - 1) \times (L + w - 1)$ matrix whose entries are given as follows.

$$B_{i,j} = \begin{cases} A_{i,j} & 0 \leq i, j \leq T_\epsilon, \\ 0 & \text{o.w.} \end{cases} \quad (10.167)$$

Also, let D be a $(L + w - 1) \times (L + w - 1)$ matrix whose entries are given as follows.

$$D_{i,j} = \begin{cases} A_{i,j} & i, j \geq T_\epsilon - (w - 1), \\ 0 & \text{o.w.} \end{cases} \quad (10.168)$$

A moment of thought shows that B and D have non-negative entries and for $0 \leq i, j \leq L + w - 1$ we have

$$A_{i,j} \leq B_{i,j} + D_{i,j}. \quad (10.169)$$

Let us now denote the largest eigenvalue of A , B and D by λ_A , λ_B and λ_D , respectively. Let us now show that

$$\lambda_A \leq \lambda_B + \lambda_D. \quad (10.170)$$

Form (10.25), it is easy to see that $A = LS$, where L is a diagonal matrix and S is a symmetric matrix. This is also true for the matrices B and D with the same diagonal matrix L , i.e., $B = LS_1$ and $D = LS_2$, where S_1 and S_2 are symmetric. Further, a moment of thought reveals that (i) the matrix $L^{-\frac{1}{2}}AL^{\frac{1}{2}} = L^{\frac{1}{2}}SL^{\frac{1}{2}}$ is a symmetric matrix and (ii) the matrices A and $L^{-\frac{1}{2}}AL^{\frac{1}{2}}$ have the same set of eigenvalues. Consequently, from the fact (F3) we obtain

$$\lambda_A = \max_{\{y=(y_1, \dots, y_r) \text{ s.t. } \forall j: y_j \geq 0\}} \frac{y^T L^{-\frac{1}{2}} A L^{\frac{1}{2}} y}{y^T y}.$$

By repeating the exact same argument for B and D , we obtain that

$$\lambda_B = \max_{\{y=(y_1, \dots, y_r) \text{ s.t. } \forall j: y_j \geq 0\}} \frac{y^T L^{-\frac{1}{2}} B L^{\frac{1}{2}} y}{y^T y},$$

$$\lambda_D = \max_{\{y=(y_1, \dots, y_r) \text{ s.t. } \forall j: y_j \geq 0\}} \frac{y^T L^{-\frac{1}{2}} D L^{\frac{1}{2}} y}{y^T y}.$$

Now, by using the above relations for $\lambda_A, \lambda_B, \lambda_D$ and also the relation (10.169), the relation (10.170) follows easily.

The idea is now to show that if the value of ϵ is chosen suitably in terms of α and K , then there exists a constant $\delta = \delta(\alpha, K)$ such that

$$\lambda_B + \lambda_D \leq 1 - \delta, \quad (10.171)$$

and as a result, the whole proof is complete by noting the relation (10.170). Therefore, what remains to be done is to provide suitable upper bounds of λ_B and λ_D . We start with the matrix D . Note here that the value of ϵ is at our hands to choose. In other words, in order to prove that $\lambda_B + \lambda_D$ is strictly below 1 by a constant gap, we should “choose” a suitable value of ϵ . Hence, in the following we will gradually give several upper bounds (that only depend on α and K) on the value of ϵ . We then choose in the end a value of ϵ that satisfies all these bounds and show that for this value of ϵ the quantity $\lambda_B + \lambda_D$ is strictly below 1 by a gap that is only dependent of α and K .

To find an upper bound on λ_D , the idea here is to find an upper bound on the sum of the components of each row of D and then use the fact (F2). In this regard, we need to find suitable bounds on the entries of D . Consider a position $j \geq T_\epsilon$. From (10.141) we deduce that

$$\ell_{j+w} > 2 - \epsilon.$$

Assume here that we choose the value of ϵ to be

$$\epsilon < 1. \quad (10.172)$$

As a result,

$$\ln \frac{\ell_{j+w}}{2} > \ln\left(1 - \frac{\epsilon}{2}\right) > -\epsilon.$$

Now, from this inequality and part (3) of Lemma 10.3 we have

$$\frac{-\alpha}{2^{2K-1}} \left(\frac{2 - \ell_{j+\frac{w}{2}}}{2} \right)^{K-1} > -\epsilon \implies 2 - \ell_{j+\frac{w}{2}} < \left(\epsilon \frac{2^{2K-1}}{\alpha} \right)^{\frac{1}{K-1}}.$$

Now, again by choosing

$$\left(\epsilon \frac{2^{2K-1}}{\alpha} \right)^{\frac{1}{K-1}} \leq 1 \implies \epsilon \leq \frac{\alpha}{2^{2K-1}}, \quad (10.173)$$

we obtain that

$$\ln \frac{\ell_{j+\frac{w}{2}}}{2} > \ln \left(1 - \frac{\left(\epsilon \frac{2^{2K-1}}{\alpha} \right)^{\frac{1}{K-1}}}{2} \right) > - \left(\epsilon \frac{2^{2K-1}}{\alpha} \right)^{\frac{1}{K-1}},$$

and by using this inequality and part (3) of Lemma 10.3 we deduce that for $j \geq T_\epsilon$ we have

$$2 - \ell_j \leq \epsilon^{\frac{1}{(K-1)^2}} \left(\frac{2^{2K-1}}{\alpha} \right)^{\frac{K}{(K-1)^2}}. \quad (10.174)$$

One can repeat the above argument once more to get a similar upper bound, as in the form of (10.174), for $2 - \ell_{j-w}$ and then for $2 - \ell_{j-2w}$ and so on. Since, the arguments are exactly as above, we do not repeat the details and only mention the net result: by choosing

$$\epsilon \leq \left(\frac{\alpha}{2^{2K-1}} \right)^{K+K(K-1)^2+(K-1)^4} \triangleq c_5(\alpha, K). \quad (10.175)$$

we have for $j \geq T_\epsilon$

$$2 - \ell_{j-2w} \leq \epsilon^{\frac{1}{(K-1)^6}} \underbrace{\left(\frac{2^{2K-1}}{\alpha} \right)^{\frac{K}{(K-1)^6} + \frac{K}{(K-1)^4} + \frac{K}{(K-1)^2}}_{\triangleq c_6(\alpha, K)} = \epsilon^{\frac{1}{(K-1)^6}} c_6. \quad (10.176)$$

Equivalently, we can say that for $i \geq T_\epsilon - 2w$ we have

$$2 - \ell_i \leq \epsilon^{\frac{1}{(K-1)^6}} c_6. \quad (10.177)$$

Now, recall the definition of the matrix D from (10.168). For $i < T_\epsilon - w$, all the entries in the i -th row of D are equal to 0. For $i \geq T_\epsilon - w$, we first note from (10.53) that the sum of the entries of the i -th row is

$$\begin{aligned} \sum_{j=0}^{L+w-1} D_{i,j} &= \sum_{j=i-w+1}^{i-w+1} \frac{\alpha}{2^K} \frac{K(K-1)}{w} \ell_i \frac{1}{w} \sum_{k=|j-i|}^{w-1} \left(1 - \frac{1}{w} \sum_{d=0}^{w-1} \frac{\ell_{\max(i,j)-k+d}}{2} \right)^{K-2} \\ &\leq \frac{\alpha}{2^K} \frac{K(K-1)}{w} \ell_i \sum_{k=0}^{w-1} \left(1 - \frac{1}{w} \sum_{d=0}^{w-1} \frac{\ell_{i-k+d}}{2} \right)^{K-2} \\ &\leq \frac{\alpha K(K-1)}{2^K} \ell_i \left(1 - \frac{\ell_{i-w}}{2} \right)^{K-2}, \end{aligned}$$

where the last step follows from the fact that the profile of literals $\{\ell_j\}_{j \geq 0}$ is non-decreasing in j . As a result, by using (10.177), provided that (10.175) holds, we obtain for $i \geq T_\epsilon - w$

$$\sum_{j=0}^{L+w-1} D_{i,j} \leq \epsilon^{\frac{K-2}{(K-1)^6}} \underbrace{\left(\frac{c_6}{2}\right)^{K-2} \frac{K(K-1)}{2^{K-1}}}_{\triangleq c_7(\alpha, K)}. \quad (10.178)$$

Now, from the fact that for $i < T_\epsilon - w$ the i -th row of D has all its entries equal to 0, we deduce that

$$\max_{0 \leq i \leq L+w-1} \sum_{j=0}^{L+w-1} D_{i,j} \leq \epsilon^{\frac{K-2}{(K-1)^6}} c_7. \quad (10.179)$$

Finally, by using the fact (F2) we conclude the following. Provided that (10.175) holds, we have

$$\lambda_D \leq \epsilon^{\frac{K-2}{(K-1)^6}} c_7. \quad (10.180)$$

We proceed by finding an upper bound on λ_B . Consider the vector \bar{d} defined in (10.138). Note here for $i \geq p$ we have $d_i > 0$. The idea here is first to show that the vector $B^n d$ converges exponentially to 0 in n . From (10.167) we have for $n \in \mathbb{N}$

$$B^n d \leq A^n d, \quad (10.181)$$

and as a result, we obtain from (10.139) that for $i \geq 0$

$$(B^n d)_i \leq \gamma_{n,i} d_i. \quad (10.182)$$

From part (5) of Lemma 10.3 we can easily deduce that

$$T_\epsilon - p \leq w(c_1 + c_2 \log(\log \frac{1}{\epsilon})). \quad (10.183)$$

Consider now the integer m defined as

$$m_\epsilon = \lceil 8(c_1 + c_2 \log(\log \frac{1}{\epsilon})) \rceil. \quad (10.184)$$

It is easy to see from (10.183) that

$$m_\epsilon \geq \frac{4(T_\epsilon - p)}{w - 1}. \quad (10.185)$$

Now, from the relations (10.182), (10.161) and the fact that $B_{i,j} = 0$ for $i, j > T_\epsilon$, we deduce that

$$B^{m_\epsilon} d \leq \frac{1 - c_3 c_4^{m_\epsilon - 2}}{(1 - \frac{2^K}{w\epsilon})^{m_\epsilon}} d. \quad (10.186)$$

Let us proceed with further simplification of the above bound. Assume now that w is chosen to be

$$\frac{2^K}{w\epsilon} \leq \frac{1}{2} \implies w \geq \frac{2^{K+1}}{\epsilon}. \quad (10.187)$$

By using the relation $\frac{1}{1-x} \leq 1 + 2x$ (for $x \leq \frac{1}{2}$), we obtain

$$\frac{1}{(1 - \frac{2^K}{w\epsilon})^{m_\epsilon}} \leq (1 + \frac{2^{K+1}}{w\epsilon})^{m_\epsilon} \leq 1 + 2^{m_\epsilon} \frac{2^{K+1}}{w\epsilon}.$$

As a result, we obtain from (10.186)

$$B^{m_\epsilon} d \leq (1 + \frac{2^{m_\epsilon+K+1}}{w\epsilon})(1 - c_3 c_4^{m_\epsilon-2}). \quad (10.188)$$

Now, by choosing

$$\frac{2^{m_\epsilon+K+1}}{w\epsilon} \leq c_3 c_4^{m_\epsilon-2} \implies w \geq \frac{2^{m_\epsilon+K+1}}{\epsilon c_3 c_4^{m_\epsilon-2}}, \quad (10.189)$$

we obtain from (10.188)

$$B^{m_\epsilon} d \leq 1 - (c_3 c_4^{m_\epsilon-2})^2 d. \quad (10.190)$$

Consequently, for any integer $n = m_\epsilon u$ we have

$$\begin{aligned} B^n d &= (B^{m_\epsilon})^u d \\ &\leq (1 - (c_3 c_4^{m_\epsilon-2})^2)^u d \\ &= ((1 - (c_3 c_4^{m_\epsilon-2})^2)^{\frac{1}{m_\epsilon}})^n d. \end{aligned} \quad (10.191)$$

We are now ready to use the result of the facts (F4) and (F5). But before that, the assumptions of the fact (F4) should be checked. Let \tilde{B} be a $(T_\epsilon - p) \times (T_\epsilon - p)$ matrix defined as follows. For $i, j \leq T_\epsilon - p$

$$\tilde{B}_{i,j} = B_{i+p,j+p} \stackrel{(10.167)}{=} A_{i+p,j+p}. \quad (10.192)$$

It is easy to see from the definition of B in (10.167) that

$$B = \left[\begin{array}{c|c|c} \mathbf{0}_{p \times p} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \tilde{B} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0}_{(L-w-T_\epsilon) \times (L-w-T_\epsilon)} \end{array} \right]. \quad (10.193)$$

Here, by $\mathbf{0}$ we mean a matrix whose entries are all zero. It is also easy to check that for the integer $R = \lceil \frac{T_\epsilon - p}{w} \rceil$, all the entries of the matrix B^R are strictly positive. Hence, the results of the facts (F4) and (F5) apply to the matrix \tilde{B} . It is now easy to see from (10.193) that the result of the facts apply also to the matrix B . That is, from fact (F4) we know that λ_B is a simple, and from the fact (F5) we know that for the vector d and $n \in \mathbb{N}$ we have

$$B^n d = c_8 \lambda_B^n v + o((T_\epsilon - p) \lambda_X^n), \quad (10.194)$$

where c_8 is a positive constant and v is the unique right eigenvector of B corresponding to λ_B . Now, from (10.191) and (10.194) we can easily see that

$$\lambda_B \leq (1 - (c_3 c_4^{m_\epsilon - 2})^2)^{\frac{1}{m_\epsilon}}. \quad (10.195)$$

Now, by using (10.180) and (10.195), we have

$$\begin{aligned} \lambda_B + \lambda_D &\leq (1 - (c_3 c_4^{m_\epsilon - 2})^2)^{\frac{1}{m_\epsilon}} + \epsilon^{\frac{K-2}{(K-1)^6}} c_7 \\ &\leq 1 - \frac{1}{m_\epsilon} (c_3 c_4^{m_\epsilon - 2})^2 + \epsilon^{\frac{K-2}{(K-1)^6}} c_7, \end{aligned} \quad (10.196)$$

where in the last step we have used the fact that for numbers $x, y \in [0, 1]$, we have $(1-x)^y \leq 1-xy$. We proceed by showing that there exists a constant $c_9 \triangleq c_9(\alpha, K)$ such that by choosing

$$\epsilon \leq c_9(\alpha, K), \quad (10.197)$$

we have

$$\epsilon^{\frac{K-2}{(K-1)^6}} c_7 \leq \frac{1}{2} \left(\frac{1}{m_\epsilon} (c_3 c_4^{m_\epsilon - 2})^2 \right). \quad (10.198)$$

To find such a candidate for c_9 , we note that in order for (10.198) to hold, we must have

$$\frac{K-2}{(K-1)^6} \log \epsilon + \log c_7 \leq -\log 2m_\epsilon + 2 \log c_3 + 2(m_\epsilon - 2) \log c_4.$$

By rearranging the terms we get to

$$\log \frac{1}{\epsilon} \geq \frac{(K-1)^6}{K-2} \left(\log \frac{2c_7 c_4^4}{c_3^2} - 2m_\epsilon \log c_4 + \log m_\epsilon \right),$$

and by using (10.184), we deduce that in order for (10.198) to hold, it is sufficient to have

$$\begin{aligned} \log \frac{1}{\epsilon} &\geq \frac{(K-1)^6}{K-2} \left(\log \frac{16c_7 c_4^4}{c_3^2} - 16(1+c_1) \log c_4 \right. \\ &\quad \left. - c_2 \log c_4 \log \left(\log \frac{1}{\epsilon} \right) + \log \left(c_1 + c_2 \log \left(\log \frac{1}{\epsilon} \right) \right) \right). \end{aligned} \quad (10.199)$$

Now, note here that all the constants $c_1 - c_8$ defined above are positive constants which only depend on α and K . Also, it is easy to see that if ϵ is sufficiently small, the relation (10.199) holds true. This proves the existence of a constant $c_9 = c_9(\alpha, K)$ such that if choose ϵ according to (10.197), then the relation (10.199), and hence the relation (10.198), hold true. Now, from (10.175), (10.197), let us choose the value of ϵ to be

$$\epsilon = \min(c_5, c_9) \triangleq c_{10}(\alpha, K). \quad (10.200)$$

For this value of ϵ , by plugging (10.198) into (10.196) we obtain

$$\lambda_B + \lambda_D \leq 1 - \underbrace{\frac{1}{16(c_1 + c_2 \log(\log(\frac{1}{c_{10}})))} (c_3 c_4)^{8(c_1 + c_2 \log(\log(\frac{1}{c_{10}}))) - 2}}_{\triangleq \delta(\alpha, K)} = 1 - \delta.$$

Of course, the above relation holds true provided that from (10.187), (10.189) and the value of $w_0(\alpha, K)$ given in Lemma 10.3, we have

$$w \geq \max\left(\frac{2^{K+1}}{c_{10}}, \frac{2^{K+1+8(c_1+c_2 \log(\log(\frac{1}{c_{10}})))}}{c_{10}c_1c_2^{8(c_1+c_2 \log(\log(\frac{1}{c_{10}}))) - 2}}, w_0(\alpha, K)\right) \triangleq w_1(\alpha, K).$$

An Auxiliary Lemma

Lemma 10.5. *Define*

$$d_i \triangleq w(\ell_{i+1} - \ell_i).$$

Then, there exist a constant $c = c(\alpha, K)$, which only depends on α and K , such that the following holds

$$\max_{p \leq i \leq j} \frac{d_j}{d_i} \leq c. \tag{10.201}$$

Proof. We first note that using the conservation equations (10.74) we have

$$\ln \frac{\ell_{i+1}}{\ell_i} = Q_i - Q_{i+1} + \int_0^t 2\delta_{pi} \ell_i^{-1} dt,$$

and by using (10.58), we obtain

$$\begin{aligned} \ln \frac{\ell_{i+1}}{\ell_i} &= \frac{\alpha K}{2^{K-1}} \left(\left(1 - \frac{1}{w} \sum_{d=0}^{w-1} \frac{\ell_{i+1-w+d}}{2}\right)^{K-1} - \left(1 - \frac{1}{w} \sum_{d=0}^{w-1} \frac{\ell_{i+1+d}}{2}\right)^{K-1} \right) \\ &\quad + \int_0^t 2\delta_{pi} \ell_i^{-1} dt, \end{aligned} \tag{10.202}$$

The idea of the proof is to consider two different regions for the positions $i \geq p$ and provide a suitable candidate for the value c based on a careful analysis on these regions. Let us begin by defining i_0 to be

$$i_0 = \operatorname{argmin}\{i \geq p \mid \ell_{i+1} > \max\{2 - 4(\frac{2}{\alpha K})^{\frac{1}{K-2}}, 1\}\}. \tag{10.203}$$

Note here that by using the second part of Lemma 10.3 we deduce that

$$\forall i \geq i_0 : 2 - \ell_{i+w} < \frac{1}{2}(2 - \ell_i). \tag{10.204}$$

In order to find a suitable candidate for c in (10.201), we consider two cases for the positions i, j .

Case 1: We consider positions i, j such that $j \geq i \geq i_0 + w$. For such positions, on one hand we can write

$$\begin{aligned} \ln \frac{\ell_{i+1}}{\ell_i} &= \frac{\alpha K}{2^{K-1}} \left(\left(\frac{1}{w} \sum_{d=0}^{w-1} \frac{2^{-\ell_{i+1+d-w}}}{2} \right)^{K-1} - \left(\frac{1}{w} \sum_{d=0}^{w-1} \frac{2^{-\ell_{i+1+d}}}{2} \right)^{K-1} \right) \\ &\geq \frac{\alpha K}{2^{K-1}} \left(\left(\frac{1}{w} \sum_{d=0}^{w-1} \frac{2^{-\ell_{i+1+d-w}}}{2} \right)^{K-1} - \left(\frac{1}{w} \sum_{d=0}^{w-1} \frac{2^{-\ell_{i+1+d-w}}}{4} \right)^{K-1} \right) \\ &= \frac{\alpha K}{2^{K-1}} \left(1 - \frac{1}{2^{K-1}} \right) \left(\frac{1}{w} \sum_{d=0}^{w-1} \frac{2^{-\ell_{i+1+d-w}}}{2} \right)^{K-1}. \end{aligned}$$

As a result, by noticing that $\frac{\ell_{i+1}}{\ell_i} = 1 + \frac{d_i}{w\ell_i}$ and using the first part of Lemma 10.3, we obtain

$$d_i \geq w\ell_i \frac{\alpha K}{2^{K-1}} \left(1 - \frac{1}{2^{K-1}} \right) \left(1 - \frac{1}{w} \sum_{d=0}^{w-1} \frac{\ell_{i+1+d-w}}{2} \right)^{K-1}. \quad (10.205)$$

On the other hand, by using (10.202) for any position j , we have

$$\ln \frac{\ell_{j+1}}{\ell_j} \leq \frac{\alpha K}{2^{K-1}} \left(1 - \frac{1}{w} \sum_{d=0}^{w-1} \frac{\ell_{j+1-w+d}}{2} \right)^{K-1}.$$

Now, as $j \geq i > i_0$, then by definition we have $\ell_j \geq 1$, hence $\ell_{j+1} - \ell_j \leq 1$ and hence $\frac{\ell_{j+1} - \ell_j}{\ell_j} \leq 1$. We thus have $\ln(1 + \frac{\ell_{j+1} - \ell_j}{\ell_j}) \geq \frac{1}{2} \frac{\ell_{j+1} - \ell_j}{\ell_j}$. As a result,

$$d_j \leq w\ell_j \frac{2\alpha K}{2^{K-1}} \left(1 - \frac{1}{w} \sum_{d=0}^{w-1} \frac{\ell_{j+1-w+d}}{2} \right)^{K-1}. \quad (10.206)$$

Finally, for $j \geq i \geq i_0 + w$, we obtain from (10.205) and (10.206)

$$\frac{d_j}{d_i} \leq \frac{\ell_j}{\ell_i} \left(\frac{2^{K-1}}{2^{K-1} - 1} \right) \frac{\left(1 - \frac{1}{w} \sum_{d=0}^{w-1} \frac{\ell_{j+1-w+d}}{2} \right)^{K-1}}{\left(1 - \frac{1}{w} \sum_{d=0}^{w-1} \frac{\ell_{i+1-w+d}}{2} \right)^{K-1}},$$

which easily results in the following simple inequality for $j \geq i \geq i_0 + w$

$$\frac{d_j}{d_i} \leq 4. \quad (10.207)$$

Case 2: In this case, we consider positions i, j such that $j \geq i$ and $i < i_0 + w$. We first note that by using the relation (10.203) and part (3) of Lemma 10.3, we get that

$$\ell_{i_0+w} \leq 2 \exp(-\alpha K 2^{-2K} \exp(-\alpha K 2^{-3K})). \quad (10.208)$$

Now, using the results of Lemma 10.3 and (10.208), we deduce that there exists a value $\zeta \triangleq \zeta(\alpha, K) > 0$ such that for $p \leq i \leq i_0 + w$ we have

$$\ell_{i+w} - \ell_i \geq \zeta. \quad (10.209)$$

Furthermore, by using (10.202) and the fact that for $x \geq y$ we have $x^{K-1} - y^{K-1} \geq (x - y)^{K-1}$, we can write for $i \geq p$

$$\begin{aligned} \ln \frac{\ell_{i+1}}{\ell_i} &\geq \frac{\alpha K}{2^{K-1}} \left(\frac{1}{w} \sum_{d=0}^{w-1} \frac{\ell_{i+1+d} - \ell_{i+1+d-w}}{2} \right)^{K-1} \\ &\geq \frac{\alpha K}{2^{K-1}} \zeta^{K-1}, \end{aligned}$$

and as a result, by using the first part of Lemma 10.3 we get for $p \leq i \leq i_0 + w$

$$d_i \geq w \frac{\alpha K}{2^{K-2}} \zeta^{K-1} \exp\left(-\frac{\alpha K}{2^{K-1}}\right) \triangleq D(\alpha, K), \quad (10.210)$$

and by noting that $d_j = \ell_{j+1} - \ell_j \leq 2$, we have for positions i, j such that $j \geq i$ and $i < i_0 + w$

$$\frac{d_j}{d_i} \leq \frac{2}{D(\alpha, K)}. \quad (10.211)$$

Finally, a candidate for the value of c in the lemma is $c = \max\{4, \frac{2}{D(\alpha, K)}\}$. \square

Bibliography

- [1] E. Arıkan, “Channel polarization: A method for constructing capacity-achieving codes for symmetric binary-input memoryless channels,” *IEEE Trans. Inform. Theory*, 55(7), pp. 3051–3073, (2009).
- [2] S. Kudekar, T. Richardson, R. Urbanke, “Threshold saturation via spatial coupling: why convolutional LDPC ensembles perform so well over the BEC”, *IEEE Trans. Inform. Theory*, 57(2), pp.803-834, (2011).
- [3] S. Kudekar, T. Richardson, R. Urbanke, “Spatially coupled ensembles universally achieve capacity under belief propagation”, [online] Available: [arXiv:1201.2999 \[cs.IT\]](https://arxiv.org/abs/1201.2999).
- [4] A. J. Felstrom, K. S. Zigangirov, “Time-varying periodic convolutional codes with low density parity check matrix”, *IEEE Trans. Inform. Theory*, 45 (5), pp.2181-2190 (1999).
- [5] S. Chung, G. D. Forney, T. J. Richardson, and R. Urbanke, “ On the design of low-density parity-check codes within 0.0045 dB of the Shannon limit,” *IEEE Communications Letter*, 5(2), pp. 58-60, (2011).
- [6] R.G. Gallager, “A simple derivation of the coding theorem and some applications”, *IEEE Transactions on Information Theory*, vol 11, no. 1, pp. 3-18, (1965).
- [7] E. Arıkan and E. Telatar, “On the rate of channel polarization,” in *Proc. ISIT*, Seoul, South Korea, pp.1493-1495, (2009).
- [8] R. L. Dobrushin, “Mathematical problems in the Shannon theory of optimal coding of information”, in *Proc. 4th Berkeley Symp. Mathematics, Statistics, and Probability*, vol 1, pp. 211-252,(1961).
- [9] V. Strassen, “Asymptotische absch atzungen in Shannon informationstheorie”, in *Trans. 3d Prague Conf. Inf. Theory*, Prague, pp. 689-723, (1962).
- [10] Y. Polyanskiy, H. V. Poor, and S. Verdu, “A channel coding rate in the finite block-length regime”, *IEEE Trans. Inf. Theory*, 56 (5), pp. 2307-2359, (2010).

- [11] D. Blackwell, L. Breiman, and A. J. Thomasian, "The capacity of a class of channels", *The Annals of Mathematical Statistics*, 3 (4), pp. 1229–1241, (1959).
- [12] S. B. Korada and E. Şaşoğlu and R. Urbanke, "Polar codes: Characterization of Exponent, Bounds, and Constructions", *IEEE Trans. Inform. Theory*, 56 (12), pp.6253 - 6264, (2010).
- [13] S. B. Korada, "Polar codes for channel and source coding," Ph.D. dissertation, EPFL, Lausanne, Switzerland, July 2009.
- [14] E. Sasoglu, "Polar coding theorems for discrete systems," Ph.D. dissertation, EPFL, Lausanne, Switzerland, September 2011.
- [15] S. H. Hassani, K. and R. Urbanke, "On the scaling of polar codes: II. The behavior of un-polarized channels," *in proc. ISIT*, Austin, Texas, USA, pp. 879–883, (2010).
- [16] A. Goli, S. H. Hassani, and R. Urbanke, "Universal bounds on the scaling behavior of polar codes," *in proc. ISIT*, Boston, USA, pp. 1957–1961, (2012).
- [17] S. H. Hassani, K. Alishahi, and R. Urbanke, "On the finite-length scaling of polar codes," submitted.
- [18] http://en.wikipedia.org/wiki/Sturms_theorem
- [19] M. Bastani Parizi and E. Telatar, "On correlation between polarized BECs," [online] Available: [arXiv:1301.5536](https://arxiv.org/abs/1301.5536) [cs.IT].
- [20] S. H. Hassani and R. Urbanke, "On the scaling of polar codes: I. The behavior of polarized channels," *in proc. ISIT*, Austin, Texas, USA, pp. 874–878, (2010).
- [21] S. H. Hassani, R. Mori, T. Tanaka and R. Urbanke, "Rate dependent analysis of the asymptotic behavior of channel polarization", *IEEE Trans. Inform. Theory*, accepted for publication, 2012.
- [22] S. B. Korada, A. Montanari, E. Telatar and R. Urbanke, "An empirical scaling law for polar codes", in *Proc. ISIT*, pp.884-888, (2010).
- [23] R. Pedarsani, H. Hassani, I. Tal and E. Telatar, "On the construction of polar codes," in *Proc. ISIT*, St. Petersburg, Russia, pp. 11–15, (2011).
- [24] S. H. Hassani, S. B. Korada and R. Urbanke, "The compound capacity of polar codes," in *Proc. 47th Annual Allerton Conference on Communication, Control, and Computing*, pp. 16–21, (2009).
- [25] I. Tal and A. Vardy, "How to construct polar codes," presented at 2010 IEEE Info. Theory Workshop, Dublin, Ireland, (2010). [online] Available: [arXiv:1105.6164v1](https://arxiv.org/abs/1105.6164v1) [cs.IT].

-
- [26] I. Tal and A. Vardy, “List decoding of polar codes,” [online] Available: [arXiv:1206.0050](https://arxiv.org/abs/1206.0050) [cs.IT].
- [27] R. Mori and T. Tanaka, “Performance and construction of polar codes on symmetric binary-input memoryless channels”, in *Proc. ISIT*, pp.1496–1500,(2009).
- [28] S. H. Hassani, and R. Urbanke, “Polar codes: Robustness of the successive cancellation decoder with respect to quantization,” in *Proc. ISIT*, Boston, USA, pp. 1962–1966, (2012).
- [29] E. Arıkan, “Source polarization,” in *ISIT*, Austin, Texas, USA, pp. 899–903, (2010).
- [30] S. B. Korada and R. Urbanke, “Polar codes are optimal for lossy source coding,” *IEEE Trans. Info. Theory*, 56 (12), pp. 1751–1768, (2010).
- [31] E. Sasoglu, E. Telatar and E. Arıkan, “Polarization for arbitrary discrete memoryless channels”, [online] Available [arXiv:0908.0302](https://arxiv.org/abs/0908.0302)(cs.IT).
- [32] D. Sutter, J. M. Renes, F. Dupuis, R. Renner, “Achieving the capacity of any DMC using only polar codes,” [online] Available: [arXiv:1205.3756](https://arxiv.org/abs/1205.3756) [cs.IT].
- [33] H. Mahdavifar and A. Vardy, “Achieving the secrecy capacity of wire-tap channels using polar codes,” [online] Available: [arXiv:1001.0210v2](https://arxiv.org/abs/1001.0210v2) [cs.IT].
- [34] E. Hof and S. Shamai, “Secrecy-achieving polar-coding for binary-input memoryless symmetric wire-tap channels,” [online] Available: [arXiv:1005.2759v2](https://arxiv.org/abs/1005.2759v2) [cs.IT].
- [35] E. Hof, I. Sason and S. Shamai, “Polar coding for reliable communications over parallel channels,” [online] Available: [arXiv:1005.2770v1](https://arxiv.org/abs/1005.2770v1) [cs.IT].
- [36] E. Abbe and E. Telatar, “Polar codes for the m -user MAC,” [online] Available: [arXiv:1006.4255](https://arxiv.org/abs/1006.4255) [cs.IT].
- [37] E. Sasoglu, E. Telatar and E. Yeh, “Polar codes for the two-user multiple-access channel,” [online] Available: [arXiv:1002.0777v2](https://arxiv.org/abs/1002.0777v2) [cs.IT].
- [38] N. Goela, E. Abbe and M. Gastpar, “Polar codes for deterministic broadcast channels,” in *Proc. 2012 International Zurich Seminar on Communications*, Zurich, Switzerland, 2012.
- [39] J. M. Renes, M. M. Wilde, “Polar codes for private and quantum communication over arbitrary channels,” [online] Available: [arXiv:1212.2537](https://arxiv.org/abs/1212.2537) [cs.IT].
- [40] M. R. Bloch, L. Luzzi, J. Kliewer, “Strong coordination with polar codes,” [online] Available: [arXiv:1210.2159](https://arxiv.org/abs/1210.2159) [cs.IT].

- [41] D. Burshtein, A. Strugatski, “Polar write once memory codes,” [online] Available: [arXiv:1207.0782](https://arxiv.org/abs/1207.0782) [cs.IT].
- [42] S. Haghshatshoar, E. Abbe and E. Telatar, “Adaptive sensing using deterministic partial Hadamard matrices”, [online] Available: [arXiv:1202.6555](https://arxiv.org/abs/1202.6555) [cs.IT].
- [43] C. Leroux, I. Tal, A. Vardy and W. J. Gross, “Hardware architectures for Successive Cancellation Decoding of Polar Codes,” in *Proc. ICASSP*, Prague, Czech Republic, pp. 1665-1668, (2011).
- [44] S. Kudekar and H. D. Pfister, “The effect of spatial coupling on compressive sensing,” in *Proc. of the Allerton Conf. on Communications, Control, and Computing*, Monticello, IL, USA, (2010).
- [45] F. Krzakala, M. Mézard, F. Sausset, Y. Sun, and L. Zdeborova, “Statistical physics-based reconstruction in compressed sensing,” [online] Available: [arXiv:1109.4424](https://arxiv.org/abs/1109.4424) [cs.IT].
- [46] D. Donoho, A. Javanmard, and A. Montanari, “Information-theoretically optimal compressed sensing via spatial coupling and approximate message passing,” [online] Available: [arXiv:1112.0708](https://arxiv.org/abs/1112.0708) [cs.IT].
- [47] T. Richardson and R. Urbanke, *Modern Coding Theory*. Cambridge University Press, 2008.
- [48] T. Tanaka and R. Mori, “Refined rate of channel polarization,” in *Proc. 2010 IEEE Int. Symp. Info. Theory*, Austin, Texas, USA, pp. 889–893, 2010.
- [49] S. H. Hassani, N. Macris, R. Urbanke, “Coupled graphical models and their threshold”, *Information Theory Workshop (ITW)*, Dublin, (2010); also in [lanl.arxiv.org no 1105.0785](https://arxiv.org/abs/1105.0785)[cs.IT]
- [50] S. H. Hassani, N. Macris, R. Urbanke, “Chains of mean field models”, *Journal of Statistical Physics: Theory and Experiments*, (2012).
- [51] K. Takeuchi, T. Tanaka, and T. Kawabata, “Improvement of BP-based CDMA multiuser detection by spatial coupling”, in [lanl.arxiv.org no 1102.3061](https://arxiv.org/abs/1102.3061) [cs.IT]
- [52] S. Kudekar, K. Kasai, “Spatially coupled codes over the multiple access channel”, in [lanl.arxiv.org no 1102.2856](https://arxiv.org/abs/1102.2856)[cs.IT]
- [53] J. L. Lebowitz, O. Penrose, “Rigorous treatment of the van der Waals-Maxwell theory of the liquid-vapor transition”, *J. Math. Phys* vol 7, pp. 98-113 ,(1966).
- [54] L.W.J. den Ouden, H.W. Capel and J.H.H. Perk, “Systems with separable many-particle interactions. II”, *Physica A*, vol 85, pp. 425-456, (1976).

- [55] M. Suzuki, "Coherent-Anomaly Method: Mean Field, Fluctuations and Systematics", *World Scientific*, (1995).
- [56] H. Falk, Th. W. Ruijgrok, "Deterioration of the molecular field", *Physica*, vol 78, pp. 73-90,(1974).
- [57] C. J. Thompson, "On a model of molecular field deterioration", *Physica*, vol 79(A), pp. 113-119,(1975).
- [58] P. Bak, J. von Boehm, "Ising model with solitons, phasons and the Devil's staircase", in *Phys. Rev B*, vol 21, pp. 5297-5308,(1980).
- [59] P. Bak, "Physics in One Dimension", edited by J. Bernasconi and T. Schneider, Berlin (1980).
- [60] J. Krug, J. L. Lebowitz, H. Spohn, M. Q. Zhang, "The fast rate limit of driven diffusive systems", *J. Stat. Phys.*, vol 44, pp. 535-565, (1986).
- [61] H. van Beijeren, L. S. Schulman, "Phase transitions in lattice-gas models far from equilibrium", *Phys. Rev. Lett.*, vol 53, pp. 806-809, (1984).
- [62] M. Cassandro, E. Orlandi, E. Presutti, "Interfaces and typical Gibbs configurations for one dimensional potentials", *Probab. Theory. Relat. Fields*, vol 96, pp. 57-96,(1993).
- [63] A. De Masi, E. Orlandi, E. Presutti, L. Triolo, "Glauber evolution with Kac potentials: I. Microscopic equations and fluctuation theory", *Nonlinearity*, vol 7, pp. 633-696, (1994).
- [64] O. Penrose, "A mean field equation of motion for the dynamic Ising model", *J. Stat. Phys.*, vol 63(5/6), pp. 975-986,(1991).
- [65] F. R. N. Nabarro, *Theory of crystal dislocations*, Dover, New-York, (1987).
- [66] S. H. Hassani, N. Macris, R. Mori, "Near concavity of the growth rate for coupled LDPC chains", in *Proc. ISIT*, St Petersburg, ,(2011).
- [67] M. Mézard, G. Parisi, "The cavity method at zero temperature", *J. Stat. Phys.*, vol 111(1-2) pp. 1-34, (2003).
- [68] S. H. Hassani, N. Macris, R. Urbanke, "Threshold saturation in spatially coupled constraint satisfaction problems", *J. Stat. Phys.: theory and experiments*,(2012).
- [69] K. Takeuchi, T. Tanaka, and T. Kawabata, "A phenomenological study on threshold improvement via spatial coupling", in [lanl.arXiv.org no 1102.3056](https://arxiv.org/abs/1102.3056) [cs.IT]
- [70] C. P. Gomes, B. Selman, N. Crato, and H. Kautz, "Heavy-tailed phenomena in satisfiability and constraint satisfaction problems", *J. Automat. Reason.*, 24(1-2), pp. 67-100, (2000).

-
- [71] C. Moore and S. Mertens, *The nature of computation*, Oxford University Press, (2011)
- [72] M. Mézard and A. Montanari, *Information, physics and computation*, Oxford University Press, (2009).
- [73] D. Achlioptas, *Random satisfiability*, Handbook of satisfiability, Eds. A. Biere et al., IOS Press, pp. 243-268 (2009).
- [74] V. Chvatal and B. Reed, "Mick gets some (the odds are on his side)", *Proc. 33rd IEEE Symp. Foundations of Computer Science (FOCS)*, pp. 620627 (1992).
- [75] E. Friedgut, "Sharp thresholds of graph properties, and the K -SAT problem", *Journal of the American Mathematical Society*, 12 (4), pp. 1017-1054, (1999).
- [76] D. Achlioptas and Y. Peres, "The threshold for random K -SAT is $2^K \log 2 - O(K)$ ", *J. Amer. Math. Soc.*, 17(4), pp. 947973, (2004).
- [77] D. Achlioptas and C. Moore, "Random K -SAT: Two moments suffice to cross a sharp threshold", *SIAM J. Comput.*, 36(3), pp. 740762, (2006).
- [78] A. Coja-Oghlan and K. Panagiotou, "Going after the K -SAT threshold", [online] Available: [arXiv:1212.1682](https://arxiv.org/abs/1212.1682) [math.CO].
- [79] M. T. Hajiaghayi and G. B. Sorkin, "The satisfiability threshold of random 3-SAT is at least 3.52", *volume RC22942 of IBM Research Report*, (2003).
- [80] A. C. Kaporis, L. M. Kirousis, and E. G. Lalas, "The probabilistic analysis of a greedy satisfiability algorithm", *Random Structures and Algorithms*, 28(4), pp. 444480, (2006).
- [81] O. Dubois, Y. Boufkhad, and J. Mandler, "Typical random 3-SAT formulae and the satisfiability threshold", *Electronic Colloquium on Computational Complexity (ECCC)*, 10(007), (2003).
- [82] O. Dubois, and Y. Boufkhad, "A general upper bound for the satisfiability threshold of random r -SAT formulae", *J. Algorithms*, 24(2), pp. 395-420, (1997).
- [83] L. M. Kirousis, E. Kranakis, D. Krizanc, and Y. Stamatiou, "Approximating the unsatisfiability threshold of random formulas", *Random Structures Algorithms*, 12(3), pp. 253-269, (1998).
- [84] M. Mézard, M. Palassini and O. Rivoire, "Landscape of solutions in constraint satisfaction problems", *Phys. Rev. Lett.*, vol 95, pp. 200-202, (2005).

- [85] M. Bayati, D. Gamarnik, and P. Tetali, “Combinatorial approach to the interpolation method and scaling limits in sparse random graphs”, *in proc. STOC*, pp. 105-114, (2010).
- [86] L. Zdeborova, F. Krzakala, “Phase transitions in the coloring of random graphs”, *Phys. Rev. E*, vol 76, 031131, (2007).
- [87] A. Montanari, F. Ricci-Tersenghi and G. Semerjian, “Clusters of solutions and replica symmetry breaking in random K -satisfiability”, *Journal of Statistical Mechanics: theory and experiment*, P04004, (2008).
- [88] I. S. Gradshteyn, I. M Ryzhik, “Table of integrals, series and products”, Academic Press, Fifth edition, ed. Allan Jeffreys, (1994).
- [89] S. H. Hassani, N. Macris, R. Urbanke, “Coupled graphical models and their threshold”, *Information Theory Workshop (ITW)*, Dublin, (2010).
- [90] F. Ricci-Tersenghi, G. Semerjian, “On the cavity method for decimated random constraint satisfaction problems and the analysis of belief propagation guided decimation algorithms”, *J. Stat. Mech.*, P09001, (2009).
- [91] A. Coja-Oglhan, “On belief propagation guided decimation for random K -SAT”, *in proc. SODA*, pp. 957-966, (2011).
- [92] O. Dubois and J. Mandler, “The 3-XORSAT threshold”, *In proc. of 43rd IEEE FOCS*, pp.769-778, (2002).
- [93] M. Mezard, F. Ricci-Tersenghi, R. Zecchina, “Alternative solutions to diluted p-spin models and XORSAT Problems”, *J. Stat. Phys*, vol 111, pp.505-533 (2003)
- [94] D. Achlioptas and M. Molloy, “The solution space geometry of random linear equations”, [arXiv:1107.5550 \[cs.DS\]](https://arxiv.org/abs/1107.5550).
- [95] M. Ibrahimi, Y. Kanoria, M. Kranning, and A. Montanari “The Set of Solutions of Random XORSAT Formulae,” *in proc. SODA 2012*.
- [96] S. H. Hassani, N. Macris, R. Urbanke, “The space of solutions of coupled XORSAT formulae”, *submitted to ISIT*, (2013).
- [97] M. Mézard, G. Parisi, “The cavity method at zero temperature”, *J. Stat. Phys*, 111 (1-2), pp. 1-34, (2003).
- [98] M. Mézard, T. Mora, and R. Zecchina, “Clustering of solutions in the random satisfiability problem”, *Phys. Rev. Lett.* vol 94, 197205, (2005).
- [99] D. Achlioptas and F. Ricci-Tersenghi, “On the solution space geometry of random constraint satisfaction problems”, *in proc. STOC*, pp. 130139, (2006).

- [100] D. Achlioptas and A. Coja-Oghlan, “Algorithmic barriers from phase transitions”, *in proc. FOCS*, pp. 793-802, (2008).
- [101] M. Mézard, G. Parisi, R. Zecchina, “Analytic and algorithmic solution of random satisfiability problems”, *Science*, vol 297, pp. 812-815, (2002).
- [102] M. Mézard, R. Zecchina, “Random K -satisfiability problem: from an analytic solution to an efficient algorithm”, *Phys. Rev. E*, vol 66, 056126-1, (2002).
- [103] A. Montanari, F. Ricci-Tersenghi and G. Semerjian, “Clusters of solutions and replica symmetry breaking in random K -satisfiability”, *Journal of Statistical Mechanics: theory and experiment*, P04004 (2008).
- [104] F. Krzakala, A. Montanari, F. Ricci-Tersenghi, G. Semerjian, L. Zdeborova, “Gibbs states and the set of solutions of random constraint satisfaction problems”, *Proc. National Academy of Sciences*, vol 104, pp. 10318-10323, (2007).
- [105] F. Guerra, F. Toninelli, “The thermodynamic limit in mean field spin glass models”, *Comm. Math. Phys.*, vol 203, pp. 71-79, (2002).
- [106] F. Guerra and F. Toninelli, “The high temperature region of the Viana-Bray diluted spin glass model”, *J. Stat. Phys.*, vol 115, pp. 531-555, (2004)
- [107] S. Franz, M. Leone, “Replica bounds for optimization problems and diluted spin glass problems”, *J. stat. Phys.*, vol 111, pp. 535-564, (2003).
- [108] E. Abbe, A. Montanari, “On the concentration of the number of solutions of random satisfiability formulas”, [online] Available: [arXiv: 1006.3786v1](https://arxiv.org/abs/1006.3786v1).
- [109] D. Ruelle, “Statistical mechanics: rigorous results”, Mathematical Series Monograph Series”, W. A. Benjamin, Inc (1983).
- [110] H. O. Georgii, “Gibbs Measures and Phase Transitions”, *De Gruyter Studies in Mathematics*, vol 9, Berlin: de Gruyter, (1988).
- [111] J. Franco, “Probabilistic analysis of the pure literal heuristic for the satisfiability problem”, *Ann. Oper. Res.*, vol 1, pp. 273-289, (1984).
- [112] A. Z. Broder, A. M. Frieze, and E. Upfal, “On the satisfiability and maximum satisfiability of random 3-CNF formulas”, *in proc. 4th Annual ACM-SIAM Symposium on Discrete Algorithms*, New York, pp. 322-330, (1993).
- [113] M. Mézard, F. Ricci-Tersenghi, and R. Zecchina, “Two Solutions to diluted p -Spin models and XORSAT problems”, *Journal of Statistical Physics*, vol 111, pp. 505-533, (2003).

- [114] B. Pittel, J. Spencer, and N. Wormald, “Sudden emergence of a giant k -core in a random graph”, *J. Comb. Theory. B*, vol 67, pp. 111-151, (1996).
- [115] M. Mézard, G. Parisi, M. A. Virasoro, “Spin glass theory and beyond”, World Scientific, (1987).
- [116] M. Talagrand, “Spin glasses: a challenge for mathematicians”, Springer-Verlag, (2000).
- [117] Stein, C. Newman, “Thermodynamic chaos and the structure of short-range spin glasses”, in *A. Bovier and P. Picco. Mathematics of Spin Glasses and Neural Networks*, Boston: Birkhauser, pp. 243-247, (1998).
- [118] M. Aizenman, J. Wehr, (1990), “Rounding effects of quenched randomness on first-order phase transitions”, *Commun. Math. Phys.*, vol 130, pp. 489-528, (1990).
- [119] M. Luby, M. Mitzenmacher, A. Shokrollahi, “Analysis of random processes via and-or trees”, in *proc. of the ninth annual ACM-SIAM Symposium on Discrete Algorithms*, (1998).
- [120] S. Mertens, M. Mézard, and R. Zecchina, “Threshold values of random K -SAT from the cavity method”, *Random Struct. Algorithms*, vol 28. pp. 340-373, (2006).
- [121] T. R. Kirkpatrick, P. G. Wolynes, “Connections between some kinetic and equilibrium theories of the glass transition”, *Phys. Rev. A*, vol 35. pp. 3072-3080, (1987).
- [122] T. R. Kirkpatrick, D. Thirumalai, “ p -spin interaction spin glass models: Connections with the structural glass problem”, *Phys. Rev. B*, vol 36 pp. 5388-5397, (1987).
- [123] L. Berthier, G. Biroli, “Theoretical perspective on the glass transition and amorphous materials”, *Rev. Mod. Phys.*, vol 83, pp. 587- 645, (2011).
- [124] R. Mulet, A. Pagnani, M. Weigt and R. Zecchina, “Coloring random graphs”, *Phys. Rev. Lett.*, vol 89. pp. 268701-268704, (2002).
- [125] D. Achliotas, S. H. Hassani, N. Macris and R. Urbanke, “Spatial coupling: A useful technique to boost performance,” *In preparation*.
- [126] A. Yedla, Y. Jian, P. S. Nguyen and H. D. Pfister, “A simple proof of threshold saturation for coupled scalar recursions,” in *proc. ISTC*, Gothenburg, Sweden (2011).
- [127] S. Kudekar, T. Richardson and R. Urbanke, “Wave-like solutions of general one-dimensional spatially coupled systems,” [online] Available: [arXiv:1208.5273 \[cs.IT\]](https://arxiv.org/abs/1208.5273).

-
- [128] D. Achlioptas, "Lower bounds for random 3-SAT via differential equations," *Theoretical Computer Science* 265, pp. 159-185, (2001) .
- [129] C. Godsil and G. Royle, *Algebraic Graph Theory*, Springer, (2001).

S. Hamed Hassani

CONTACT Address: INR 031 (Bâtiment INR), Station 14, CH-1015 Lausanne
Email: seyedhamed.hassani@epfl.ch
Web page: <http://people.epfl.ch/seyedhamed.hassani>

RESEARCH INTERESTS Machine Learning, Graphical models, Coding and Information Theory

EDUCATION **Ecole Polytechnique Fédérale de Lausanne (EPFL)**

- Ph.D. in Communication and Computer Sciences (2013)
Dissertation title: Polarization and Spatial Coupling: Two Techniques to Boost Performance
Thesis advisors: Nicolas Macris, Rudiger Urbanke

Sharif University of Technology

- B.Sc. in Communication Systems (2002-2007)
- B.Sc. in Pure Mathematics (2004-2007)

HONORS

- Nominated for the student paper award, ISIT 2010
- EPFL departmental fellowship, 2007-2008 and doctoral fellowship, 2008
- MIT presidential fellowship, 2008
- Ranked among the top 100 in the nationwide university entrance exam, 2002

SELECTED GRADUATE COURSES

Mathematics
Statistical Theory, Measure Theory, Probability Theory, Long Time Behavior of Reversible Markov Chains, Functional Analysis, Algebraic Geometry, Algebraic Topology, Non-Commutative Algebra, Galois Theory, Combinatorial Optimization

Communication and Computer Sciences
Modern Coding Theory, Information Theory and Coding, Advanced Algorithms, Models and Methods for Random Networks, Mathematical Principles of Signal Processing, Advanced Digital Communications, TCP/IP Networks

PUBLICATIONS

Journal Papers

- D. Achlioptas, S. H. Hassani, N. Macris, R. Urbanke, “Spatial Coupling: A Useful Technique to Boost Performance”, *in preparation*.
- M. Mondelli, S. H. Hassani, R. Urbanke, “Scaling Exponent of List Decoders with Applications to Polar Codes”, *submitted to IEEE Information Transactions on Information Theory*, 2013
- S. H. Hassani, K. Alishahi, R. Urbanke, “Finite-length Scaling of Polar Codes”, *submitted to IEEE Transactions on Information Theory*, 2013

- S. H. Hassani, R. Mori, T. Tanaka, R. Urbanke, “Rate-Dependent Analysis of the Asymptotic Behavior of Channel Polarization”, *IEEE Transactions on Information Theory*, 2013
- S. H. Hassani, N. Macris, R. Urbanke, “Threshold Saturation in Spatially Coupled Constraint Satisfaction Problems”, *Journal of Statistical Mechanics-Theory and Experiment*, 2012
- S. H. Hassani, N. Macris, R. Urbanke, “Chain of Mean Field Models”, *Journal of Statistical Mechanics-Theory and Experiment*, 2012

Conference Papers

- M. Mondelli, S. H. Hassani, R. Urbanke, “Scaling Exponent of List Decoders with Applications to Polar Codes”, *submitted to IEEE Information Theory Workshop (ITW)*, Seville, 2013
- W. Liu, S. H. Hassani, R. Urbanke, “The Least Degraded and the Least Upgraded Channel with respect to a Channel Family”, *submitted to IEEE Information Theory Workshop (ITW)*, Seville, 2013
- S. H. Hassani, N. Macris, R. Urbanke, “The Space of Solutions of Coupled XORSAT Formulae”, *to be presented at IEEE International Symposium on Information Theory (ISIT)*, Istanbul, 2013
- S. H. Hassani, R. Urbanke, “Polar Codes: Robustness of the Successive Cancellation Decoder with Respect to Quantization”, *IEEE International Symposium on Information Theory (ISIT)*, Boston, 2012
- Ali Goli, S. Hamed Hassani, Rudiger Urbanke, “Universal Bounds on the Scaling Behavior of Polar Codes”, *IEEE International Symposium on Information Theory*, Boston, 2012
- R. Pedarsani, S. H. Hassani, I. Tal, E. Telatar, “On the Construction of Polar Codes”, *IEEE International Symposium on Information Theory*, St. Petersburg, 2011
- S. H. Hassani, N. Macris, R. Mori, “Near-Concavity of the Growth Rate for Coupled LDPC Chains”, *IEEE International Symposium on Information Theory*, St. Petersburg, 2011
- S. H. Hassani, N. Macris, R. Urbanke, “Coupled Graphical Models and Their Thresholds”, *IEEE Information Theory Workshop (ITW)*, Dublin, 2010
- S. H. Hassani, R. Urbanke, “On the Scaling of Polar Codes: I. The Behavior of Polarized Channels”, *IEEE International Symposium on Information Theory (ISIT)*, Texas, 2010
- S. H. Hassani, K. Alishahi, R. Urbanke, “On the Scaling of Polar Codes: II. The Behavior of Unpolarized Channels”, *IEEE International Symposium on Information Theory (ISIT)*, Texas, 2010
- S. H. Hassani, S. B. Korada, R. Urbanke, “The Compound Capacity of Polar Codes”, *Allerton Conference on Communications, Controlled Computing*, Texas, 2009
- S. Feizi, F. Ashtiani, S. H. Hassani, “Modeling the Behavior of Contending Opportunistic Cognitive Radios”, *IEEE International Conference on Signal Processing and Communications (ICSPC)*, Dubai, 2007.
- S. H. Hassani, P. Tehrani, F. Ashtiani, “A New Model for the Analysis of IEEE 802.11 MAC Protocol Based on Queuing Networks”, *IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, Athens, 2007

- S. H. Hassani, M. R. Aref, “A New (t, n) Multi-Secret Sharing Scheme Based on Linear Algebra”, *IEEE International Conference on Security and Cryptography*, Sebutal, Portugal, 2006

EXPERIENCE Teaching Assistant at EPFL

Principles of Digital Communications (Spring 2011, Spring 2012)

Information Theory and Coding (Fall 2008, Fall 2010)

Quantum Information Theory (Spring 2010)

Signal Processing for Communications (Spring 2009)