Channel Comparison Methods and Statistical Problems on Graphs

by

Yuzhou Gu

B.S., Massachusetts Institute of Technology (2017)
M.Eng., Massachusetts Institute of Technology (2018)

Submitted to the Department of Electrical Engineering and Computer Science
in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2023

Authored by: Yuzhou Gu
Department of Electrical Engineering and Computer Science
May 18, 2023

Certified by: Yury Polyanskiy
Professor of Electrical Engineering and Computer Science
Thesis Supervisor

Accepted by: Leslie A. Kolodziejski
Professor of Electrical Engineering and Computer Science
Chair, Department Committee on Graduate Students

Channel Comparison Methods and Statistical Problems on Graphs

by
Yuzhou Gu

Submitted to the Department of Electrical Engineering and Computer Science
on May 18, 2023, in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Abstract

Initially driven by channel coding, information theory has developed a large collection of tools for measuring and comparing effectiveness of information channels. These tools have found applications in various fields such as statistics, probability, and theoretical computer science. This thesis explores several applications of these tools to statistical problems related to graphs.

Part I focuses on information channels and channel comparison methods, including f-divergences, strong data processing inequalities, and preorders between channels. While these theories have been well-established for binary memoryless symmetric (BMS) channels, there remains much to discover for channels with larger input alphabets. We develop a theory of q-ary input-symmetric (FMS) channels, generalizing the theory of BMS channels. We demonstrate that while FMS channels exhibit more complex behavior than BMS channels, some properties of BMS channels can be extended to FMS channels. Furthermore, we perform tight analysis on contraction properties of the Potts channels, the simplest examples of FMS channels.

In Part II, we apply the information theoretical methods established in Part I to solve problems related to random graph models with community structures. The random graph models include the stochastic block model (SBM) and its variants, which hold significance in statistics, machine learning, and network science. Central problems for these models ask about the feasibility and quality of recovering hidden community structures from unlabeled graphs. By utilizing the relationship between random graphs and random Galton-Watson trees, we demonstrate that many important problems on these graphical models can be reduced to problems on trees. We apply various channel comparison methods to solve these tree problems, demonstrating that different methods are effective for different problems and that selecting the correct tool for a problem is crucial. Problems we study include (for SBMs) weak recovery, optimal recovery algorithms, mutual information formula, and (for broadcasting on trees) reconstruction, robust reconstruction, uniqueness of belief propagation fixed points, boundary irrelevance, computation of limit information, and so on.

Thesis Supervisor: Yury Polyanskiy
Title: Professor of Electrical Engineering and Computer Science

Acknowledgments

I would like to express my deepest gratitude to my advisor, Yury Polyanskiy. He introduced me to the beauty of information theory and provided invaluable guidance and support throughout my graduate career. This thesis would not have been possible without the countless meetings and discussions we had over the years.

I would also like to thank my other thesis committee members, Guy Bresler and Elchanan Mossel.

Many other senior researchers have taught me a great deal, both research-related and beyond. I would especially like to express my appreciation to Emmanuel Abbe, David Gamarnik, Muriel Médard, and Richard Peng.

I am very grateful to Zhao Song for his help during my difficult times and for teaching me about numerical algorithms.

I would like to thank all my co-authors and collaborators. I have learned so much from them and hope to find the time to revisit all those unfinished works in the future.

My heartfelt thanks go to my parents. I cannot begin to imagine the sacrifices they have made to raise me. I owe them so much.

I would like to thank my wife, Zhulin Li, for accompanying me through this challenging and sometimes difficult period of my life.

I would like to thank all the friends who have brightened my life during these years. Among them, I especially want to acknowledge Yinzhan Xu for being both a reliable collaborator and a sincere friend; Yao Yu for introducing me to puzzle hunts; and Yuhao Du for discussions on logic puzzles.

Lastly, I would like to acknowledge all grant fundings that supported my research, including MIT-IBM Watson AI Lab No. W1771646, Air Force Research Laboratory FA8750-19-2-1000, Purdue University/NSF Subaward #10000686-015, NSF CCF-1717842, NSF CCF-1253205, Skolkovo Institute of Science and Technology No. 4015, Jacobs Family Presidential Fellowship.

Research was sponsored by the United States Air Force Research Laboratory and the Department of the Air Force Artificial Intelligence Accelerator and was accomplished under Cooperative Agreement Number FA8750-19-2-1000. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Department of the Air Force or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

Contents

List of Figures

List of Tables

List of Abbreviations

BSC	binary symmetric channel
BEC	binary erasure channel
BMS	binary memoryless symmetric (channel)
FSC	fully symmetric channel
FEC	generalized erasure channel
FMS	fully memoryless symmetric (channel)
EC	erasure channel
DPI	data processing inequality
SDPI	strong data processing inequality
LSI	log-Sobolev inequality
MLSI	modified log-Sobolev inequality
NLSI	non-linear log-Sobolev inequality
KL	Kullback-Leibler (divergence)
SKL	symmetric Kullback-Leibler
SBM	stochastic block model
HSBM	hypergraph stochastic block model
BOT	broadcasting on trees
BOHT	broadcasting on hypertrees
BOTS	broadcasting on trees with survey
BP	belief propagation
BI	boundary irrelevance
SNR	signal-to-noise ratio
LLR	log-likelihood ratio
KS	Kesten-Stigum (threshold)

Chapter 1

Introduction

1.1 Channel comparison methods

The data processing inequality (DPI) is arguably the most fundamental inequality in information theory. Intuitively, it says that processing a signal can never increase the amount of information that was originally contained in the signal. The DPI appears ubiquitously in information theory and its significance can be compared with that of the second law of thermodynamics in statistical physics.

In its most common form, the DPI states that for any Markov chain $U \to X \to Y$, we have

$$I(U;Y) \leq I(U;X), \tag{1.1}$$

where I denotes mutual information. This tells us the channel $P_{Y|X}$ contracts information. Hence, contraction is a fundamental ability of information channels.

Despite its endless applications, the DPI is a qualitative rather than a quantitative result. In many scenarios, it is not enough to know that information contracts; we also need to know the amount by which it contracts. The strong data processing inequalities (SDPIs) are introduced in order to address this issue.

Consider the same Markov chain $U \to X \to Y$. One form of the SDPI states that

$$I(U;Y) \leq \eta I(U;X), \tag{1.2}$$

where $\eta = \eta(P_{Y|X})$ is a constant that depends on the channel $P_{Y|X}$ and is commonly called the contraction coefficient of $P_{Y|X}$. The constant η is always at most one by DPI, and in many cases, it is strictly smaller than one. The value of η represents the ability of a channel to contract mutual information.

The mutual information is not the only quantity that satisfies DPI. There are many other information-like quantities, called f-information, that behave similarly, yet differently in subtle ways. For any f-information, there is a version of SDPI and a corresponding contraction coefficient that captures different aspects of contraction abilities of information channels.

Contraction coefficients are one set of tool for comparing channels. It is also possi-

ble to compare channels more directly by using several preorders defined on the space of information channels. One example is the degradation preorder, which captures the simulation relationship between channels. If a channel P is more degraded than another channel Q, it is possible to simulate P by using Q as a blackbox and then processing its output. It is intuitive that more degraded channels have stronger contraction abilities. There are other important preorders, such as the less-noisy preorder and the more-capable preorder.

1.2 Statistical problems on graphs

Channel comparison methods have found numerous applications, not only in information theory, but also in statistics, probability, and theoretical computer science. In this context, we focus on their applications in statistical problems on graphs.

Specifically, we study problems on the stochastic block model (SBM) and its variants. The SBM is a random graph model with hidden community structure and has significant applications in statistics, machine learning, and network science. In its simplest version, the SBM is defined as follows: there are n vertices, divided into several communities; for every pair of vertices, there is an edge between them with a certain probability that depends on whether the two vertices are in the same community or not. The central problem in the study of SBMs is to determine, given the graph without community information, whether and how much we can recover about the community structure.

It turns out that problems on SBMs can be reduced to problems on trees, which are often easier to solve. Problems on trees essentially involve analyzing certain belief propagation (BP) operators on the space of information channels. For example, the weak recovery problem on SBMs can be reduced to the reconstruction problem on trees, which is equivalent to asking whether the BP operator has a non-trivial fixed point.

Analyzing BP operators is a challenging task. They are defined on an infinite-dimensional space and are non-linear operators with complicated expressions. Nevertheless, from an information-theoretical perspective, BP operators can often be more concisely described using composition and $\star$-convolution of information channels.

Using this observation, channel comparison methods can be used to tackle problems about BP operators. For example, if a certain f-information behaves nicely under composition and $\star$-convolution, it may be possible to analyze the evolution of this f-information under the BP operator and deduce properties of the operator. This reduces the study of an infinite-dimensional dynamical system to a one-dimensional one. Although the details may vary in actual problems, this description captures the big picture.

1.3 Organization of the thesis

We have introduced the main theme of this thesis, and now we will provide a brief overview of the contents of the different chapters in this thesis.

The thesis is divided into two parts. In Part I, we investigate methods for comparing contraction properties of information channels. In Part II, we study statistical problems on graphs by reducing them to problems on trees, and then solving them by applying information-theoretical methods.

In most of our results, we develop new methods for channel comparison and apply them to graph problems. We include the channel comparison results in Part I when they are of independent interest, even without the context of graphical problems. We put them in Part II, along with the corresponding graph result, when it is more relevant to do so.

Part I consists of Chapter 2, 3, 4.

In Chapter 2, we review the general theory of channel comparison methods. We study f-divergences, the most important ones being Kullback-Leibler (KL) divergence, χ^2-divergence, symmetric KL (SKL) divergence, and Hellinger distance. We also study preorders between two channels, including degradation preorder, less-noisy preorder, and more-capable preorder. We examine the general behavior of these channel comparison methods under channel transformations, including composition and $\star$-convolution. We also analyze the behavior of these methods on binary memoryless symmetric (BMS) channels. Most results in this chapter are standard and have appeared in previous works.

In Chapter 3, we generalize the theory of BMS channels to symmetric channels with larger input alphabet, called q-ary fully memoryless symmetric (q-FMS) channels. We demonstrate that FMS channels can be decomposed into a mixture of simple channels, which enables us to view FMS channels as distributions on the probability simplex. Using this interpretation, we establish several results on the degradation of FMS channels. We also prove that although χ^2-information is generally not subadditive, it satisfies a form of local subadditivity for FMS channels. Results in this chapter are based on [73] and unpublished notes.

In Chapter 4, we zoom in further and examine the contraction behavior of Potts channels, which are the simplest examples of FMS channels. We establish non-linear p-log-Sobolev inequalities for the semigroup of Potts channels and compute KL contraction coefficients of the Potts channels. Results in this chapter are based on [72].

Part II consists of Chapter 5, 6, 7, 8, 9, 10, 11.

In Chapter 5, we examine a general hypergraph stochastic model (HSBM), which includes the stochastic block model and its variants that we study in this thesis as special cases. We investigate three problems on the general HSBM. The first problem is weak recovery, which asks whether it is possible to reconstruct a non-trivial fraction of the community structure given the unlabeled hypergraph. The second problem is optimal recovery, which asks what is the maximum fraction of community structure that can be recovered given the unlabeled hypergraph and possibly some noisy observations of the correct labels. The third problem is to compute the mutual information between the labels and the hypergraph, a fundamental property of the model. We demonstrate that all three problems can be reduced to problems on the broadcasting on hypertrees (BOHT) model. Specifically, non-reconstruction for the BOHT model implies the impossibility of weak recovery for the HSBM, uniqueness of BP fixed point for the BOHT model implies optimal recovery algorithms for the

HSBM, and the boundary irrelevance (BI) property for the BOHT model implies a mutual information formula for the HSBM. The reduction for weak recovery is folklore, and the reductions for optimal recovery algorithms and mutual information formula are based on generalizations of [4, 73].

In Chapter 6, we use contraction of KL divergence to prove simple yet effective non-reconstruction results for broadcasting on trees (BOT). This yields the current best results for the Potts model at small degrees and can recover the reconstruction threshold of the coloring model to the first order, which was previously proved using more complicated methods. In the assortative case, these results imply the current best results on the impossibility of weak recovery for the corresponding SBMs. Furthermore, we apply this method to a BOT model whose alphabet is continuous, establishing the exact value of the reconstruction threshold. Results in this chapter are based on [72].

In Chapter 7, we use contraction of SKL divergence and χ^2-divergence to establish results on binary simple BOHT models. We prove that when the hyperedge size r is at most four, the reconstruction threshold is at the so-called Kesten-Stigum threshold. (Note that the case $r = 4$ relies on a numerically-verified inequality.) This determines the exact value of the weak recovery threshold for the corresponding HSBMs, which has been an open question for a decade. Moreover, we demonstrate that when r is at least seven, the reconstruction threshold and the Kesten-Stigum threshold do not match, suggesting that there is an information-computation gap for the corresponding HSBMs. Results in this chapter are based on [74].

In Chapter 8, we generalize the local subadditivity result in Chapter 3 to general channels, and using this result to establish the robust reconstruction threshold for reversible BOT models. While previous work [82] has already determined the robust reconstruction threshold for general BOT models, there are certain edges cases it cannot handle, particularly including the coloring model. In addition, we prove that BI does not hold for reversible BOT models between the reconstruction threshold and the Kesten-Stigum threshold. Results in this chapter are based on [73] and unpublished notes.

In Chapter 9, we use channel preorders to provide a method to compute limit mutual information and limit probability of error in the Ising model on a tree. Our method gives rigorous bounds as opposed to conventional methods like population dynamics. By utilizing our method and analyzing experimental results, we were able to make conjectures on the limit information of the Ising model near criticality, which were later confirmed by [136]. Results in this chapter are based on [75].

In Chapter 10, we use the degradation preorder and Hellinger distance to prove uniqueness of BP fixed point and the BI property for the Ising model on a tree. These results were proven for signal-to-noise ratio (SNR) outside a small interval $(1, 3.513)$. As a consequence, we establish an optimal recovery algorithm and a mutual information formula for the corresponding SBMs, which were the best at the time. Results in this chapter are based on [4].

In Chapter 11, we generalize the method developed in the previous chapter to prove uniqueness of BP fixed point and the BI property for the Potts model. These results were shown to asymptotically approach the Kesten-Stigum threshold when

the number of communities q goes to ∞ and a parameter λ is $o(1/\log q)$. These results imply an optimal recovery algorithm and a mutual information formula for the corresponding SBMs, which are the current best results. Results in this chapter are based on [74].

To end this introduction, we present Table 1.1, which summarizes the channel comparison tools used in this thesis and their applications to statistical problems on trees. We also include a few previous results using these methods to give the reader more context. While the table is not intended to be an exhaustive list of previous works on tree problems, we believe it includes the most relevant ones.

Tool	**Application**	**Reference**
KL divergence	Reconstruction, any BOT model	[72], Ch. 6
χ^2-divergence	Reconstruction, Potts model	[126, 109]
	Reconstruction, coloring model	[125]
	Reconstruction, binary asymmetric BOT	[27, 91]
	Reconstruction, hardcore model	[19]
	Reconstruction, BOHT model	[74], Ch. 7
	Robust reconstruction, any BOT model	[82]
	Robust reconstruction, any BOT model	Ch. 8
SKL divergence	Reconstruction, any BOT model	[87, 66]
	Reconstruction, BOHT model	[74], Ch. 7
Less-noisy preorder	Reconstruction, certain tree models	[116]
	Compuation of BOT limit, Ising model	[75], Ch. 9
Degradation preorder	Uniqueness of BP fixed point, Ising model	[4, 137], Ch. 10
	Uniqueness of BP fixed point, Potts model	[73], Ch. 11

Table 1.1: Summary of channel comparison tools and applications to statistical problems on trees. When multiple tools are used in a result, we choose the one that is considered to be the most important.

Part I

Information Channels and Comparison Methods

Chapter 2

Channel comparison methods

We review the general theory of information channels and comparison methods, which have been studied extensively in literature under various contexts. The methods include f-divergences and preorders between channels. We apply these theories to binary memoryless symmetric channels which are the main examples in the chapter. We focus on the behavior of these methods under channel transformations, including composition, tensor product, and $\star$-convolution. Overall, this chapter serves as a foundation for the subsequent chapters in this thesis. The general theory is developed in more depth for special classes of information channels in Chapter 3 and 4, and are eventually applied to statistical problems on graphs.

Chapter outline In Section 2.1, we introduce several basic notions and definitions on information channels that will be used throughout the thesis. In Section 2.2, we introduce binary memoryless symmetric (BMS) channels. We explain the equivalence between BMS channels and distributions on the interval $[0, \frac{1}{2}]$, which provides a useful conceptual simplification. In Section 2.3, we introduce f-divergences and f-informations, and study the behavior of these quantities under $\star$-convolution. We focus on the Kullback-Leibler (KL) divergence, χ^2-divergence, symmetric KL (SKL) divergence, and Hellinger distance. In Section 2.4, we examine strong data processing inequalities (SDPIs) for f-divergences, which represent contraction abilities of information channels under different f-divergences. In Section 2.5, we study several preorders between two channels, including the degradation preorder, the less-noisy preorder, and the more-capable preorder. We explore their behavior under channel transformations.

2.1 Basic notations

In this section we recall some basic notations and definitions used throughout the thesis. We refer the reader to [117] for an introduction to information theory.

For $n \in \mathbb{Z}_{\geq 0}$, we use $[n]$ to denote the set $\{1, \ldots, n\}$.

Let $\mathcal{X}$, $\mathcal{Y}$ be measurable spaces. We use $P : \mathcal{X} \to \mathcal{Y}$ to denote an information channel (also known as channel, Markov kernel, probability kernel) P from $\mathcal{X}$ to

$\mathcal{Y}$. Here $\mathcal{X}$ is called the input alphabet (or the input space) and $\mathcal{Y}$ is called the output alphabet (or the output space). On input $x \in \mathcal{X}$, we use $P(\cdot|x)$ to denote the distribution of the output. For discrete $\mathcal{Y}$, we denote the transition probability from $x \in \mathcal{X}$ to $y \in \mathcal{Y}$ by $P(y|x)$, $P(x, y)$, or $P_{x,y}$.

For a measurable space $\mathcal{X}$, we use $\mathcal{P}(\mathcal{X})$ to denote the space of probability distributions on $\mathcal{X}$. For a distribution $\mu \in \mathcal{P}(\mathcal{X})$ and a channel $P : \mathcal{X} \to \mathcal{Y}$, we use μP and $P \circ \mu$ to denote the distribution of the output variable, when the input distribution is μ.

For a random variable $X \in \mathcal{X}$, we use $P_X \in \mathcal{P}(\mathcal{X})$ to denote the distribution of X.

We use Id to denote the identity channel $\mathcal{X} \to \mathcal{X}$. We use 0 to denote the trivial channel when this does not cause confusion. For $\epsilon \in [0, 1]$, we use EC_ϵ to denote the erasure channel $\mathcal{X} \to \mathcal{X} \sqcup \{*\}$ with erasure probability ϵ. Clearly $\mathrm{EC}_1 = 0$ and $\mathrm{EC}_0 = \mathrm{Id}$.

Let $P : \mathcal{X} \to \mathcal{Y}$ be a channel and $\mu \in \mathcal{P}(\mathcal{X})$. If the support of μP is equal to $\mathcal{Y}$, then we define the reverse channel P^*_μ (or P^* when μ is clear from context) by

$$\mu(x)P(x, y) = (\mu P)(y)P^*_\mu(y, x). \tag{2.1}$$

If $\mathcal{X} = \mathcal{Y}$, $\mu P = \mu$, and $P = P^*_\mu$, we say (π, P) is reversible (or P is reversible when μ is clear from context).

Let X be a discrete random variable. We use $H(X)$ to denote the entropy of X. In this thesis, we do not discuss differential entropy of continuous random variables.

Let X and Y be two possibly dependent random variables. We use $I(X; Y)$ to denote the (Shannon) mutual information between X and Y. Furthermore, for a distrbution $\mu \in \mathcal{P}(\mathcal{X})$ and a channel $P : \mathcal{X} \to \mathcal{Y}$, we define $I(\mu, P)$ as the mutual information $I(X; Y)$ where X is a random variable with distribution μ and Y is the output of P when given input X. When X is discrete, we have $I(X; X) = I(P_X, \mathrm{Id}) = H(X)$.

Let $P : \mathcal{X} \to \mathcal{Y}$ and $Q : \mathcal{Y} \to \mathcal{Z}$ be two channels where the output alphabet of P is the input alphabet of Q. We use PQ and $Q \circ P$ to denote the composition channel $\mathcal{X} \to \mathcal{Z}$.

Let $P : \mathcal{X} \to \mathcal{Y}$ and $Q : \mathcal{X} \to \mathcal{Z}$ be two channels with the same input alphabet. We say P and Q are equivalent if there exists $R : \mathcal{Y} \to \mathcal{Z}$ and $R' : \mathcal{Z} \to \mathcal{Y}$ such that $Q = R \circ P$ and $P = R' \circ Q$. Equivalent channels behave similarly in many aspects, in particular, $I(\mu, P) = I(\mu, Q)$ for any distribution $\mu \in \mathcal{P}(\mathcal{X})$.

Let $P : \mathcal{X} \to \mathcal{Y}$ and $Q : \mathcal{X}' \to \mathcal{Y}'$ be two channels. We define the tensor product channel $P \times Q : \mathcal{X} \times \mathcal{X}' \to \mathcal{Y} \times \mathcal{Y}'$ by letting P and Q acting on the two inputs independently. For $n \in \mathbb{Z}_{\geq 1}$, we use $P^{\times n} : \mathcal{X}^n \to \mathcal{Y}^n$ to denote the n-th tensor power of P.

Let $P : \mathcal{X} \to \mathcal{Y}$ and $Q : \mathcal{X} \to \mathcal{Z}$ be two channels with the same input alphabet. We define the $\star$-product (also called $\star$-convolution) $P \star Q : \mathcal{X} \to \mathcal{Y} \times \mathcal{Z}$ by letting P and Q acting on the same input independently. For $n \in \mathbb{Z}_{\geq 0}$, we use $P^{\star n} : \mathcal{X} \to \mathcal{Y}^n$ to denote the n-th $\star$-power of P.

Let $P : \mathcal{X} \to \mathcal{Y}$ and $Q : \mathcal{X} \to \mathcal{Z}$ be two channels with the same input alphabet.

A mixture of P and Q is an channel R from $\mathcal{X}$ to $\mathcal{Y} \sqcup \mathcal{Z}$ such that $R(E|x) = (1-p)P(E \cap \mathcal{Y}|x) + pQ(E \cap \mathcal{Z}|x)$ for any measurable $E \in \mathcal{Y} \sqcup \mathcal{Z}$, for some parameter $p \in [0,1]$. We denote this as $R = (1-p)P + pQ$. Similarly, if $\mathcal{A}$ is a collection of information channels with the same input alphabet, and μ is a distribution on $\mathcal{A}$, we write $\mathbb{E}_{P\sim\mu}P$ for the channel which maps input x to (P, Y), where $P \sim \mu$ and $Y \sim P(\cdot|x)$. This channel is called a mixture of $\mathcal{A}$.

2.2 Binary memoryless symmetric channels

In this section we review the theory of BMS channels, arguably the most extensively studied class of channels. We refer the reader to [120, Chapter 4] for a more complete introduction to BMS channels.

Definition 2.1 (Binary memoryless symmetric (BMS) channels)**.** A channel $P : \{\pm 1\} \to \mathcal{Y}$ is called a binary memoryless symmetric channel if there exists a measurable involution $\sigma : \mathcal{Y} \to \mathcal{Y}$ such that $P(\sigma^{-1}(E)|+) = P(E|-)$ for all measurable sets $E \subseteq \mathcal{Y}$.

Two most common examples of the BMS channels are the binary erasure channels (BECs) and binary symmetric channels (BSCs). For $\epsilon \in [0,1]$, $\mathrm{BEC}_\epsilon : \{\pm\} \to \{\pm, *\}$ denotes the channel with transition probabilities

$$\mathrm{BEC}_\epsilon(y|x) = \begin{cases} 1-\epsilon, & \text{if } y = x, \\ 0, & \text{if } y = -x, \\ \epsilon, & \text{if } y = *. \end{cases} \tag{2.2}$$

For $\delta \in [0,1]$, $\mathrm{BSC}_\delta : \{\pm\} \to \{\pm\}$ denotes the channel with transition probabilities

$$\mathrm{BSC}_\delta(y|x) = \begin{cases} 1-\delta, & \text{if } y = x, \\ \delta, & \text{if } y = -x. \end{cases} \tag{2.3}$$

One can easily verify that BSC_δ and $\mathrm{BSC}_{1-\delta}$ are equivalent for any $\delta \in [0,1]$, and that $\mathrm{BSC}_0 = \mathrm{Id}$. One may also observe that BEC_ϵ is equivalent to a mixture of BSC_0 and $\mathrm{BSC}_{\frac{1}{2}}$. This is no coincidence and hints the following general structural result on BMS channels: every BMS channel is equivalent to a mixture of BSCs.

Lemma 2.2 (Structure of BMS channels)**.** *Every BMS channel P is equivalent to a BMS channel $X \to (\Delta, Z)$, where on input $X \in \{\pm\}$ it outputs $(\Delta, Z) \in [0, \frac{1}{2}] \times \{\pm\}$ such that Δ is independent of X and $P_{Z|\Delta,X} = \mathrm{BSC}_\Delta(\cdot|X)$. Furthermore, the distribution of Δ is uniquely deteremined by P.*

In the setting of the above lemma, we call Δ the Δ-component of P, and P_Δ the Δ-distribution of P. Lemma 2.2 establishes an equivalence between a BMS channel and a probability distribution on $[0, \frac{1}{2}]$, which maps a BMS channel to its Δ-distribution. In particular, the Δ-distribution is an invariant property of a BMS channel under equivalence.

In many cases, working with $\theta = 1 - 2\Delta$ is more convenient than working with Δ itself. We call θ the θ-component of P.

Using the mixture representation of BMS channels, when dealing with BMS channels, we can often reduce to the case of BSCs.

For a BMS channel P, the composition $P \circ \mathrm{BSC}_\delta$ for $\delta \in [0,1]$ has Δ-distribution $f_*(P_\Delta)$, where P_Δ is the Δ-distrbution of P, $f(x) = |1-2\delta|x + \min\{\delta, 1-\delta\}$, and f_* is the induced pushforward map.

For two BMS channels P and Q, their $\star$-convolution $P \star Q$ has Δ-distribution

$$\mathbb{E}_{\substack{\Delta\sim P_\Delta \\ \Delta'\sim Q_\Delta}} \left[(\Delta \star (1-\Delta')) \mathbb{1}_{\frac{\Delta\Delta'}{\Delta\star(1-\Delta')}} + (\Delta \star \Delta') \mathbb{1}_{\frac{\Delta(1-\Delta')}{\Delta\star\Delta'}} \right], \tag{2.4}$$

where P_Δ (resp. Q_Δ) is the Δ-distribution of P (resp. Q), $x \star y = x(1-y) + y(1-x)$, and $\mathbb{1}_x$ denotes the point distribution on at x.

One may observe an issue in the above expression: $x = \frac{\Delta(1-\Delta')}{\Delta\star\Delta'}$ may take value in the whole interval $[0,1]$, while the Δ-distribution lives on the interval $[0, \frac{1}{2}]$. In fact, the interval $[0, \frac{1}{2}]$ can be viewed as the space $[0,1]/C_2$, where C_2 is the group of two elements, acting on $[0,1]$ by mapping x to $1-x$. Therefore in (2.4) we have omitted the step of taking the projection of $x \in [0,1]$ onto $[0,1]/C_2$. Because BSC_δ is equivalent to $\mathrm{BSC}_{1-\delta}$, in the view of equivalence, we usually do not need to distinguish a point distribution at x and a point distrbution at $1-x$.

2.3 f-divergences

The Kullback-Leibler (KL) divergence and mutual information are the most important quantities in information theory, and in some sense the most natural information measures. However, in many scenarios, it is necessary to consider other types of information measures that satisfy properties not shared by the KL divergence. A powerful class of non-KL information measures is the f-divergences.

Definition 2.3 (f-divergence and f-information)**.** Let $f : (0,\infty) \to \mathbb{R}$ be a convex function with $f(1) = 0$. For two distributions μ and ν on the same space, if $\mu \ll \nu$ (i.e., μ is absolutely continuous with respect to ν), we define the f-divergence as

$$D_f(\mu\|\nu) := \mathbb{E}_\nu \left[f\left(\frac{d\mu}{d\nu}\right) \right], \tag{2.5}$$

where $f(0) := f(0+)$ (which is possibly ∞). This definition is extended to $\mu \not\ll \nu$ by continuity (see [117, Definition 7.1] for more details).

For two possibly dependent random variables X and Y, we define the f-information as

$$I_f(X;Y) := D_f(P_{XY}\|P_X P_Y). \tag{2.6}$$

Similar to the mutual information case, for a distrbution $\mu \in \mathcal{P}(\mathcal{X})$ and a channel $P : \mathcal{X} \to \mathcal{Y}$, we define $I_f(\mu, P) := I_f(X;Y)$ where X is a random variable with

distribution μ and Y is the output of P when given input X.

The following f-divergences are the most important ones in this thesis.

- Total variation (TV) distance: $f(x) = \frac{1}{2}|x-1|$.

$$\mathrm{TV}(\mu,\nu) = \frac{1}{2}\mathbb{E}_\nu\left|\frac{d\mu}{d\nu} - 1\right| = \frac{1}{2}\int |d\mu - d\nu|\,. \tag{2.7}$$

- Kullback-Leibler (KL) divergence: $f(x) = x\log x$.

$$D(\mu\|\nu) = \mathbb{E}_\nu\left[\frac{d\mu}{d\nu}\log\frac{d\mu}{d\nu}\right] = \mathbb{E}_\mu\left[\log\frac{d\mu}{d\nu}\right] = \int d\mu\log\frac{d\mu}{d\nu}. \tag{2.8}$$

- χ^2-divergence: $f(x) = (x-1)^2$,

$$\chi^2(\mu\|\nu) = \mathbb{E}_\nu\left[\left(\frac{d\mu}{d\nu}-1\right)^2\right] = \mathbb{E}_\mu\left[\frac{d\mu}{d\nu}\right] - 1 = \int\frac{d\mu^2}{d\nu} - 1. \tag{2.9}$$

- Symmetric KL (SKL) divergence: $f(x) = (x-1)\log x$,

$$D_{\mathrm{SKL}}(\mu,\nu) = D(\mu\|\nu) + D(\nu\|\mu) = \int (d\mu - d\nu)\log\frac{d\mu}{d\nu}. \tag{2.10}$$

- Squared Hellinger distance: $f(x) = (1-\sqrt{x})^2$.

$$H^2(\mu,\nu) = \mathbb{E}_\nu\left[\left(1-\sqrt{\frac{d\mu}{d\nu}}\right)^2\right] = \int\left(\sqrt{d\mu}-\sqrt{d\nu}\right)^2. \tag{2.11}$$

We list here a few nice properties of f-divergences and f-information under tensorization and $\star$-convolution. All of these results can be found in e.g., [117].

- KL divergence is additive under tensorization:

$$D\left(\prod_{j\in[n]} P_{X_j}\|\prod_{j\in[n]} Q_{X_j}\right) = \sum_{j\in[n]} D(P_{X_j}\|Q_{X_j}). \tag{2.12}$$

- χ^2-divergence is multiplicative under tensorization:

$$1+\chi^2\left(\prod_{j\in[n]} P_{X_j}\|\prod_{j\in[n]} Q_{X_j}\right) = \prod_{j\in[n]}\left(1+\chi^2(P_{X_j}\|Q_{X_j})\right). \tag{2.13}$$

- Hellinger distance is multiplicative under tensorization:

$$1 - \frac{1}{2}H^2\left(\prod_{j\in[n]} P_{X_j} \| \prod_{j\in[n]} Q_{X_j}\right) = \prod_{j\in[n]}\left(1 - \frac{1}{2}H^2(P_{X_j}\|Q_{X_j})\right). \tag{2.14}$$

- Mutual information is subadditive under tensorization: if $P_{Y^n|X^n} = \prod_{j\in[n]} P_{Y_j|X_j}$, then

$$I(X^n;Y^n) \le \sum_{i\in[n]} I(X_i;Y_i). \tag{2.15}$$

- Mutual information is subadditive under $\star$-convolution:

$$I(X;Y^n) \le \sum_{j\in[n]} I(X;Y_j). \tag{2.16}$$

- SKL information is additive under $\star$-convolution ([87]):

$$I_{\mathrm{SKL}}(X;Y^n) = \sum_{j\in[n]} I_{\mathrm{SKL}}(X;Y_j). \tag{2.17}$$

For the rest of this section, we apply these f-informations to BMS channels. In the following, let μ be the uniform distribution on $\{\pm\}$, P be a BMS channel, Δ be the Δ-component of P, and $\theta = 1-2\Delta$ be the θ-component of P. For an f-divergence, we define $C_f(P) := I_f(\mu, P)$ and call it the f-capacity of P. In particular, we make use of the following information measures.

Definition 2.4.

$$\begin{aligned}
P_e(P) &= \mathbb{E}\Delta, && \text{(probability of error)}\\
C(P) &= \mathbb{E}\left[\log 2 - h(\Delta)\right], && \text{(capacity)}\\
\text{where}\quad h(x) &= -x\log x - (1-x)\log(1-x),\\
C_{\chi^2}(P) &= \mathbb{E}\left[(1-2\Delta)^2\right] = \mathbb{E}\theta^2, && (\chi^2\text{-capacity})\\
C_{\mathrm{SKL}}(P) &= \mathbb{E}\left[\left(\frac{1}{2}-\Delta\right)\log\frac{1-\Delta}{\Delta}\right] = \mathbb{E}\left[\theta\operatorname{arctanh}\theta\right], && \text{(SKL capacity)}\\
Z(P) &= 1 - \frac{1}{2}H^2(P(\cdot|+), P(\cdot|-)) && \text{(Bhattacharyya coefficient)}\\
&= \mathbb{E}\left[2\sqrt{\Delta(1-\Delta)}\right] = \mathbb{E}\sqrt{1-\theta^2}.
\end{aligned}$$

From general properties of f-information discussed above, for BMS channels P

and Q, we have

$$C(P \star Q) \le C(P) + C(Q), \tag{2.18}$$
$$C_{\mathrm{SKL}}(P \star Q) = C_{\mathrm{SKL}}(P) + C_{\mathrm{SKL}}(Q), \tag{2.19}$$
$$Z(P \star Q) = Z(P)Z(Q). \tag{2.20}$$

Furthermore, χ^2-capacity is subadditive under $\star$-convolution for BMS channels, a property not true for general channels.

Lemma 2.5 (Subadditivity of χ^2-capacity for BMS channels)**.** *For BMS channels P and Q, we have*

$$C_{\chi^2}(P \star Q) \le C_{\chi^2}(P) + C_{\chi^2}(Q). \tag{2.21}$$

Proof. This is essentially proved in e.g., [3]. Here we give a direct proof. By BSC mixture representation of BMS channels, it suffices to prove the case when P and Q are both BSCs. Let $P = \mathrm{BSC}_x$, $Q = \mathrm{BSC}_y$. Then

$$C_{\chi^2}(\mathrm{BSC}_x \star \mathrm{BSC}_y) = \frac{(x(1-y) - y(1-x))^2}{x(1-y) + y(1-x)} + \frac{(xy - (1-x)(1-y))^2}{xy + (1-x)(1-y)}, \tag{2.22}$$
$$C_{\chi^2}(\mathrm{BSC}_x) = (1-2x)^2, \qquad C_{\chi^2}(\mathrm{BSC}_y) = (1-2y)^2. \tag{2.23}$$

Therefore

$$\begin{aligned} & C_{\chi^2}(\mathrm{BSC}_x) + C_{\chi^2}(\mathrm{BSC}_y) - C_{\chi^2}(\mathrm{BSC}_x \star \mathrm{BSC}_y) \\ &= \frac{(1-2x)^2(1-2y)^2(x(1-x) + y(1-y))}{(x(1-y) + y(1-x))\,(xy + (1-x)(1-y))} \\ &\ge 0. \end{aligned} \tag{2.24}$$

This finishes the proof. □

2.4 Strong data processing inequalities

The strong data processing inequalities (SDPIs) are quantitative versions of the data processing inequality (DPI). They state that a fixed channel contracts information by a multiplicative factor, which is usually less than one. There are two versions of the SDPIs, depending on whether we fix the input distrbution. We first introduce the divergence form of the SDPIs.

Definition 2.6 (Strong data processing inequalities, divergence form)**.** Fix an f-divergence and a channel $P : \mathcal{X} \to \mathcal{Y}$. The input-unrestricted strong data processing inequality states that

$$D_f(\mu P \| \nu P) \le \eta_f(P) D_f(\mu \| \nu) \quad \forall \mu, \nu \in \mathcal{P}(\mathcal{X}), \tag{2.25}$$

where $\eta_f(P)$ is the smallest constant making the inequality true.

In addition to the above setting, let ν be a distribution on $\mathcal{X}$. The input-restricted strong data processing inequality states that

$$D_f(\mu P\|\nu P) \le \eta_f(\nu, P) D_f(\mu\|\nu) \quad \forall \mu \in \mathcal{P}(\mathcal{X}), \tag{2.26}$$

where $\eta_f(\nu, P)$ is the smallest constant making the inequality true.

In other words,

$$\eta_f(P) := \sup_{\substack{\mu,\nu\in\mathcal{P}(\mathcal{X}) \\ 0<D_f(\mu\|\nu)<\infty}} \frac{D_f(\mu P\|\nu P)}{D_f(\mu\|\nu)}, \tag{2.27}$$

$$\eta_f(\nu, P) := \sup_{\substack{\mu\in\mathcal{P}(\mathcal{X}) \\ 0<D_f(\mu\|\nu)<\infty}} \frac{D_f(\mu P\|\nu P)}{D_f(\mu\|\nu)}. \tag{2.28}$$

The SDPIs also have an information form, which states that

- (input-unrestricted version) for any Markov chain $U - X - Y$, we have

$$I_f(U;Y) \le \eta_f(P) I_f(U;X), \tag{2.29}$$

- (input-restricted version) for any Markov chain $U - X - Y$ such that $P_X = \nu$, we have

$$I_f(U;Y) \le \eta_f(\nu, P) I_f(U;X). \tag{2.30}$$

It is proved in e.g., [117, Chapter 33] that the divergence form and the information form of SDPIs are equivalent. Usually the divergence form is easier for computation of the contraction coefficients, and the information form is easier for applications.

As an example, for BSCs, we have ([11])

$$\begin{aligned} \eta_{\mathrm{KL}}(\mathrm{Unif}(\{\pm\}), \mathrm{BSC}_\delta) &= \eta_{\chi^2}(\mathrm{Unif}(\{\pm\}), \mathrm{BSC}_\delta) \\ &= \eta_{\mathrm{KL}}(\mathrm{BSC}_\delta) = \eta_{\chi^2}(\mathrm{BSC}_\delta) = (1-2\delta)^2, \end{aligned} \tag{2.31}$$

$$\eta_{\mathrm{TV}}(\mathrm{Unif}(\{\pm\}), \mathrm{BSC}_\delta) = \eta_{\mathrm{TV}}(\mathrm{BSC}_\delta) = |1-2\delta|. \tag{2.32}$$

This can be generalized to BMS channels. Let P be a BMS channel, Δ be its Δ-component. Then

$$\eta_{\mathrm{KL}}(\mathrm{Unif}(\{\pm\}), P) = \eta_{\chi^2}(\mathrm{Unif}(\{\pm\}), P) = \eta_{\mathrm{KL}}(P) = \eta_{\chi^2}(P) = \mathbb{E}(1-2\Delta)^2, \tag{2.33}$$

$$\eta_{\mathrm{TV}}(\mathrm{Unif}(\{\pm\}), P) = \eta_{\mathrm{TV}}(P) = \mathbb{E}|1-2\Delta|. \tag{2.34}$$

In the following, we briefly review previous works on SDPIs and contraction coefficients. We refer the reader to [118, 115, 93] for more discussions.

General properties of contraction coefficients Let us start with a few general properties of contraction coefficients.

- For any ν and P, we have

$$0 \le \eta_f(\nu, P) \le \eta_f(P) \le 1. \tag{2.35}$$

- Both $\eta_f(P)$ and $\eta_f(\nu, P)$ (for any ν) are convex in the transition matrix P [41, 118].

- Input-restricted contraction coefficients behave nicely under tensorization. [118, Theorem III.9] states that when f induces a subadditive and homogeneous f-entropy,[1]

$$\eta_f\left(\prod_{j\in[n]} \nu_j, \prod_{j\in[n]} P_j\right) = \max_{j\in[n]} \eta_f(\nu_j, P_j). \tag{2.36}$$

- Contraction coefficients are submultiplicative under composition. For a distribution $\nu \in \mathcal{P}(\mathcal{X})$ and channels $P : \mathcal{X} \to \mathcal{Y}$, $Q : \mathcal{Y} \to \mathcal{Z}$, we have

$$\eta_f(\nu, Q \circ P) \le \eta_f(\nu, P)\eta_f(\nu P, Q), \tag{2.37}$$
$$\eta_f(Q \circ P) \le \eta_f(P)\eta_f(Q). \tag{2.38}$$

Computation of input-unrestricted contraction coefficients Let us discuss computation of input-unrestricted contraction coefficients. The TV contraction coefficient $\eta_{\mathrm{TV}}(P)$ is called the Dobrushin's coefficient and [52] showed that

$$\eta_{\mathrm{TV}}(P) = \sup_{x,x'} \mathrm{TV}(P(\cdot|x), P(\cdot|x')). \tag{2.39}$$

It is known (e.g., [40, 118, 115]) that $\eta_{\chi^2}(P)$ and $\eta_{\mathrm{TV}}(P)$ are the extremal input-unrestricted contraction coefficients. Specifically, for any f-divergence we have

$$\eta_f(P) \le \eta_{\mathrm{TV}}(P), \tag{2.40}$$

and for any f-divergence where f is twice differentiable on $(0, \infty)$ and $f''(1) > 0$, we have

$$\eta_{\chi^2}(P) \le \eta_f(P). \tag{2.41}$$

[39] showed that for operator convex f, we have

$$\eta_{\chi^2}(P) = \eta_f(P). \tag{2.42}$$

Examples of operator convex divergences include the KL divergence and the squared Hellinger distance.

[1] f-entropy of a real-value random variable U with finite $\mathbb{E}[f(U)]$ is defined as $\mathrm{Ent}_f[U] := \mathbb{E}[f(U)] - f(\mathbb{E}U)$.

[111] proved that to compute $\eta_f(P)$, it suffices to consider input distributions with support size at most two. This essentially closes the problem of computation of input-unrestricted contraction coefficients.

Computation of input-restricted contraction coefficients Now let us move to the input-restricted contraction coefficients. It is known (e.g., [122, 118]) that $\eta_{\chi^2}(\nu, P)$ is the square of the maximal correlation coefficient $S(\nu, P)$, where

$$S(\nu, P) := \sup_{f,g} \mathbb{E}_{(X,Y)\sim\nu\otimes P}[f(X)g(Y)], \tag{2.43}$$

where the supremum is over all $f : \mathcal{X} \to \mathbb{R}$, $g : \mathcal{Y} \to \mathbb{R}$ such that $\mathbb{E}_\nu[f] = 0$, $\mathbb{E}_\nu[f^2] = 1$, $\mathbb{E}_{\nu P}[g] = 0$, $\mathbb{E}_{\nu P}[g^2] = 1$. Alternatively, $\eta_{\chi^2}(\nu, P)$ can be described as the largest eigenvalue of PP^*_ν viewed as a linear operator from $\mathcal{H}_0(\mathcal{X}, \nu)$, the space of zero-mean functions (under ν) on $\mathcal{X}$, to itself, or the square of the largest singular value of P^*_ν viewed as a linear operator fomr $\mathcal{H}_0(\mathcal{X}, \nu)$ to $\mathcal{H}_0(\mathcal{Y}, \nu P)$ [118]. Furthermore, [118] observed that $1 - \sqrt{\eta_{\chi^2}(\nu, P)}$ is equal to the spectral gap of (ν, P), a quantity with crucial importance in the study Markov chain mixing.

Eq. (2.41) can be strengthened (see e.g., [40, 118, 115]) to

$$\eta_{\chi^2}(\nu, P) \le \eta_f(\nu, P) \tag{2.44}$$

for any f-divergence where f is twice differentiable on $(0, \infty)$ and $f''(1) > 0$.

We remark that [11] systematically studied η_{KL}, and proved Eq. (2.42), (2.44) for the KL contraction coefficient.

[118, Theorem III.6] proved upper bounds on $\eta_f(\mu, P)$ for operator convex f, stating that

$$\eta_f(\mu, P) \le \max\left\{\eta_{\chi^2}(\mu, P), \sup_{0<\beta<1} \eta_{\mathrm{LC}_\beta}(\mu, P)\right\} \tag{2.45}$$

where LC_β is the Le-Cam divergence (see e.g., [132]), which is the f-divergence with $f(x) = \beta(1-\beta)\frac{(1-x)^2}{(1-\beta)x+\beta}$.

[94, Theorem 2.2] gave an upper bound on $\eta_f(\mu, P)$ of

$$\eta_f(\nu, P) \le \frac{f'(1) + f(0)}{f''(1)\min_{x\in\mathcal{X}} \nu(x)} \eta_{\chi^2}(\nu, P) \tag{2.46}$$

for f satisfying a series of technical conditions. These conditions are satisfied for the KL divergence, and then Eq. (2.46) gives [94, Corollary 2.1]

$$\eta_{\mathrm{KL}}(\nu, P) \le \frac{\eta_{\chi^2}(\nu, P)}{\min_{x\in\mathcal{X}} \nu(x)}. \tag{2.47}$$

Eq. (2.47) is further refined in [94, Theorem 2.3], which states that

$$\eta_{\mathrm{KL}}(\nu, P) \le \frac{2\eta_{\chi^2}(\nu, P)}{\phi\left(\max_{A\subseteq\mathcal{X}} \min\{\nu(A), 1-\nu(A)\}\right) \min_{x\in\mathcal{X}} \nu(x)}, \tag{2.48}$$
$$\phi(p) = \begin{cases} \frac{1}{1-2p}\log\frac{1-p}{p}, & 0 \le p < \frac{1}{2}, \\ 2, & p = \frac{1}{2}. \end{cases}$$

We have discussed a few general bounds on input-restricted contraction coefficients. In general, given an input-channel pair (ν, P), computing an upper bound on the KL contraction coefficient $\eta_{\mathrm{KL}}(\nu, P)$ is a daunting task. (Computing a lower bound is considerably easier: it suffices to choose one input distribution μ and compute $\frac{D(\mu P\|\nu P)}{D(\mu\|\nu)}$.) We list a few methods that can be used to compute upper bounds on $\eta_{\mathrm{KL}}(\nu, P)$.

- Compute from definition. This works for channels with very simple structures. For example, contraction coefficients of BSCs can be computed in this way. In Chapter 4, we will compute contraction coefficients of the Potts channels directly from definition.

- Compare with other contraction coefficients. The χ^2-contraction coefficient $\eta_{\chi^2}(\nu, P)$ is almost always easier to compute, and can be used to produce upper bounds. This works when $\eta_{\chi^2}(\nu, P)$ is close to $\eta_{\mathrm{KL}}(\nu, P)$, or when a very good bound is not needed.

- Use general properties such as tensorization or composition. This works when the input-channel pair has nice structures.

- Compare with log-Sobolev inequalities. [100, 118] proved that $\eta_{\mathrm{KL}}(\nu, P) \le 1 - \rho(PP^*)$, where $\rho(PP^*)$ denotes the log-Sobolev constant of PP^*, where P^* is the reverse channel. The log-Sobolev constants are sometimes easier to compute.

- Use an inductive structure of (μ, P). [29] computed the KL contraction coefficients of several Markov kernels related to the random walks on graphs with nice structures. This method, when it applies, gives better results than comparing with log-Sobolev constants.

- Use spectral independence and related conditions. Spectral independence is a condition introduced in [14] in order to study mixing time of Glauber dynamics. It has since then been successfully applied to various models, e.g., [33, 65, 32]. [34, 22] showed that spectral independence plus several technical conditions imply bounds on the KL contraction coefficient of the Glauber dynamics.

SDPIs and contraction coefficients have found numerous applications, including noisy computation [64, 115], distributed data-compression [46], statistical physics [53], portfolio theory [62], differential privacy [55], and so on. Furthermore, contraction properties of Markov kernels are also the central subject of study in Markov chain

mixing. The importance of spectral gap (which is equivalent to the input-restricted χ^2-contraction coefficient) is needless to say [88]. We mention that a lot of important results are established using contraction of KL divergence [34, 22, 13]. In this thesis, we focus on applications of SDPIs to statistical problems on graphs.

2.5 Preorders between two channels

In the previous section we discussed comparison of channels using contraction coefficients. It is also possible to compare channels directly, an idea dating back to [124]. Various preorders between channels have been studied, including channel inclusion [124], input-output degradation [40], degradation [17, 47], the less-noisy preorder [86], and the more-capable preorder [86]. In this section, we examine the properties of the last three preorders, which are the most useful ones for our purpose.

Definition 2.7. Let $P : \mathcal{X} \to \mathcal{Y}$ and $Q : \mathcal{X} \to \mathcal{Z}$ be two channels with the same input alphabet. We say

- P is a degradation of Q, denoted $P \leq_{\text{deg}} Q$, if there exists channel $R : \mathcal{Z} \to \mathcal{Y}$ such that $P = R \circ Q$;
- P is less noisy than Q, denoted $P \geq_{\text{ln}} Q$, if for every measurable space $\mathcal{W}$, distribution $\mu \in \mathcal{P}(\mathcal{W})$, and channel $R : \mathcal{W} \to \mathcal{X}$, we have $I(\mu, P \circ R) \geq I(\mu, Q \circ R)$;
- P is more capable than Q, denoted $P \geq_{\text{mc}} Q$, if for every $\mu \in \mathcal{P}(\mathcal{X})$ we have $I(\mu, P) \geq I(\mu, Q)$.

It is clear from the definition that

$$P \leq_{\text{deg}} Q \Rightarrow P \leq_{\text{ln}} Q \Rightarrow P \leq_{\text{mc}} Q. \tag{2.49}$$

Let P, Q be two channels with the same input alphabet $\mathcal{X}$. If $P \leq_{\text{deg}} Q$, then for any $\mu \in \mathcal{P}(\mathcal{X})$,

$$I_f(\mu, P) \leq I_f(\mu, Q). \tag{2.50}$$

In particular, if P and Q are BMS channels with $P \leq_{\text{deg}} Q$, then we have

$$\begin{aligned} &P_e(P) \geq P_e(Q), \quad C(P) \leq C(Q), \quad C_{\chi^2}(P) \leq C_{\chi^2}(Q), \\ &C_{\text{SKL}}(P) \leq C_{\text{SKL}}(Q), \quad Z(P) \geq Z(Q). \end{aligned} \tag{2.51}$$

These preorders behave nicely under channel transformations. We summarize these properties as follows.

Lemma 2.8. *The following holds.*

- *(Composition) Let P, Q be two channels with the same input alphabet $\mathcal{X}$. Let $R : \mathcal{W} \to \mathcal{X}$ be a channel.*

  - *If $P \le_{\deg} Q$, then $P \circ R \le_{\deg} Q \circ R$.*
  - *If $P \le_{\ln} Q$, then $P \circ R \le_{\ln} Q \circ R$.*

- *(Tensorization) Let P_1 and Q_1 be two channels with the same input alphabet $\mathcal{X}$, and P_2 and Q_2 be two channels with the same input alphabet $\mathcal{Y}$.*
  - *If $P_1 \le_{\deg} Q_1$ and $P_2 \le_{\deg} Q_2$, then $P_1 \times Q_1 \le_{\deg} P_2 \times Q_2$.*
  - *([131, 115]) If $P_1 \le_{\ln} Q_1$ and $P_2 \le_{\ln} Q_2$, then $P_1 \times Q_1 \le_{\ln} P_2 \times Q_2$.*
  - *([48]) If $P_1 \le_{\mathrm{mc}} Q_1$ and $P_2 \le_{\mathrm{mc}} Q_2$, then $P_1 \times Q_1 \le_{\mathrm{mc}} P_2 \times Q_2$.*

- *($\star$-convolution) Let P_1, P_2, Q_1, Q_2 be four channels with the same input alphabet.*
  - *If $P_1 \le_{\deg} Q_1$ and $P_2 \le_{\deg} Q_2$, then $P_1 \star Q_1 \le_{\deg} P_2 \star Q_2$.*
  - *If $P_1 \le_{\ln} Q_1$ and $P_2 \le_{\ln} Q_2$, then $P_1 \star Q_1 \le_{\ln} P_2 \star Q_2$.*

Note that the results on $\star$-convolution can be proved by combining the result on composition and tensorization: $P \star Q$ is the same channel as $(P \times Q) \circ T$, where $T : \mathcal{X} \to \mathcal{X} \times \mathcal{X}$ is the channel mapping X to (X, X).

[115] showed that for an arbitrary channel P, $\eta_{\mathrm{KL}}(P) \le \eta$ if and only if $P \le_{\ln} \mathrm{EC}_{1-\eta}$. This relates the less-noisy preorder and KL contraction coefficients.

Degradation beween BMS channels can be characterized in terms of the Δ-component. The following result is well-known and can be achieved by combining [120, Theorem 4.74] with the coupling characterization of second-order stochastic dominance.

Lemma 2.9 ([120])**.** *Let P and Q be two BMS channels. Let Δ_P (resp. Δ_Q) be the Δ-component of P (resp. Q). Then $P \ge_{\deg} Q$ if and only if there exists a coupling between Δ_P and Δ_Q so that*

$$\mathbb{E}[\Delta_P | \Delta_Q] \le \Delta_Q \tag{2.52}$$

almost surely.

The following result determines the extremal BMS channels under different constraints.

Lemma 2.10 ([121, Lemma 2])**.** *The following holds.*

1. *([120]) Among all BMS channels with the same probability of error $P_e(W) = \delta$ the least degraded one is* BEC *and the most degraded one is* BSC*, i.e.*

$$\mathrm{BSC}_\delta \le_{\deg} W \le_{\deg} \mathrm{BEC}_{2\delta}\,. \tag{2.53}$$

2. *([123] for the BSC part) Among all BMS with the same capacity C the most capable one is* BEC *and the least capable one is* BSC*, i.e.:*

$$\mathrm{BSC}_{1-h^{-1}(C)} \le_{\mathrm{mc}} W \le_{\mathrm{mc}} \mathrm{BEC}_{1-C}, \tag{2.54}$$

   where $h^{-1} : [0, \log 2] \to [0, 1/2]$ is the inverse of the binary entropy function $h : [0, 1/2] \to [0, \log 2]$.

3. *Among all BMS channels with the same χ^2-capacity $C_{\chi^2}(W) = \eta$ the least noisy one is* BEC *and the most noisy one is* BSC*, i.e.*

$$\mathrm{BSC}_{1/2-\sqrt{\eta}/2} \leq_{\ln} W \leq_{\ln} \mathrm{BEC}_{1-\eta}\,. \tag{2.55}$$

Chapter 3

Non-binary input symmetric channels

We study the class of q-ary fully memoryless symmetric (q-FMS) channels,[1] a generalization of binary memoryless symmetric (BMS) channels to q-ary input alphabets. As BMS channels are the "right" class of channels to work with when studying Ising models, FMS channels are the "right" class for Potts models. We first establish the structure of FMS channels as mixtures of simpler channels, generalizing the corresponding well-known result for BMS channels, which greatly simplifies the study of these channels. We study the relationship between FMS channels and degradation preorder and show that some but not all properties of BMS channels can be generalized to FMS channels. Furthermore, we prove a local subadditivity result for χ^2-capacity of FMS channels under $\star$-convolution, generalizing the corresponding subadditivity result for BMS channels (Lemma 2.5). Results developed in this chapter are used in Chapter 11 to study problems regarding the Potts model. This chapter is based on [73] and unpublished notes.

We remark that the local subadditivity result is further extended in Chapter 8 to general channels with slightly worse parameters.

Definition 3.1 (Fully memoryless symmetric (FMS) channels)**.** A q-ary fully memoryless symmetric (q-FMS) channel (or an FMS channel when q is obvious from context) is a channel $P : \mathcal{X} \to \mathcal{Y}$ with input alphabet $\mathcal{X} = [q]$ such that there exists a group homomorphism $\iota : \mathrm{Aut}(\mathcal{X}) \to \mathrm{Aut}(\mathcal{Y})$ such that for any measurable $E \subseteq \mathcal{Y}$, we have

$$P(E|x) = P(\iota(\tau)E|\tau(x)) \tag{3.1}$$

for all $x \in \mathcal{X}$, $\tau \in \mathrm{Aut}(\mathcal{X})$. Here $\mathrm{Aut}(\mathcal{X})$ denotes the symmetry group (also known as the automorphism group) of $\mathcal{X}$.

By definition, 2-FMS channels are exactly BMS channels.

We remark that q-FMS channels are a special case of input-symmetric channels (see e.g., [117, Chapter 19]) whose group of symmetries is the whole $\mathrm{Aut}(\mathcal{X})$.

[1] Here "fully" modifies "symmetric", and indicates that the symmetry group of the channel is the full symmetric group $\mathrm{Aut}(\mathcal{X})$ as opposed to a subgroup.

Chapter outline In Section 3.1, we prove a structural result of FMS channels (Prop. 3.3), which generalizes the BSC mixture representation of BMS channels. We use this result to derive several helpful formulas for transformations of FMS channels. In Section 3.2, we prove several properties of FMS channels related to the degradation preorder, including an equivalence condition of degradation in terms of coupling of the corresponding π-distributions, and extremal FMS channels under probability of error constraints. In Section 3.3, we prove that χ^2-capacity of FMS channels is almost subadditive when one of the channels is close to trivial.

3.1 Structure of FMS channels

The BSC mixture representation of BMS channels (Lemma 2.2) has been useful in proving results about BMS channels. Therefore it is desirable to generalize this theory to FMS channels. We define fully symmetric channels (FSCs), which generalize BSCs, and will serve as basic building blocks for FMS channels.

Definition 3.2. Let $\mathcal{X} = [q]$, $\mathcal{Y} = \mathrm{Aut}(\mathcal{X})$. For $\pi \in \mathcal{P}(\mathcal{X})/\mathrm{Aut}(\mathcal{X})$, define channel $\mathrm{FSC}_\pi : \mathcal{X} \to \mathcal{Y}$ as

$$\mathrm{FSC}_\pi(\tau|i) = \frac{1}{(q-1)!}\pi_{\tau^{-1}(i)} \qquad \forall i \in \mathcal{X}, \tau \in \mathrm{Aut}(\mathcal{X}), \tag{3.2}$$

where $\mathrm{Aut}(\mathcal{X})$ acts on $\mathcal{Y}$ by left multiplication.

We can verify that

$$\mathrm{FSC}_\pi(\eta\tau|\eta(i)) = \frac{1}{(q-1)!}\pi_{(\eta\tau)^{-1}(\eta(i))} = \frac{1}{(q-1)!}\pi_{\tau^{-1}(i)} = \mathrm{FSC}_\pi(\tau|i) \tag{3.3}$$

for $i \in \mathcal{X}$, $\eta, \tau \in \mathrm{Aut}(\mathcal{X})$. So FSCs are examples of FMS channels.

While the class of BSCs is a single-parameter family, the class of FSCs have $q-1$ parameters. Furthermore, the output alphabet of q-FSCs for $q \geq 3$ is no longer the same as the input alphabet.

The following result generalizes Lemma 2.2 to FMS channels.

Proposition 3.3 (Structure of FMS channels)**.** *Every FMS channel is equivalent to a mixture of FSCs, i.e., every FMS channel $P : \mathcal{X} \to \mathcal{Y}$ is equivalent to a channel $X \to (\pi, Z)$ where $\pi \sim P_\pi \in \mathcal{P}(\mathcal{P}(\mathcal{X})/\mathrm{Aut}(\mathcal{X}))$ is independent of X, and $Z \sim \mathrm{FSC}_\pi(\cdot|X)$ conditioned on π and X. Furthermore, P_π is uniquely deteremined by P.*

Proof. **Existence:** The proof strategy is to partition $\mathcal{Y}$ into $\mathrm{Aut}(\mathcal{X})$-orbits and show that the channel P restricted to each orbit is equivalent to an FSC.

Step 1. We first prove that we can replace P with an equivalent FMS channel whose $\mathrm{Aut}(\mathcal{X})$ action is free, so that in later steps each orbit is easier to handle. Define channel $\widetilde{P} : \mathcal{X} \to \mathcal{Y} \times \widetilde{\mathcal{Y}}$, where $\widetilde{\mathcal{Y}} = \mathrm{Aut}(\mathcal{X})$, sending X to $(Y, \widetilde{Y})$ where $Y \sim P(\cdot|X)$ and $\widetilde{Y} \sim \mathrm{Unif}(\mathrm{Aut}(\mathcal{X}))$ is independent of X. We give $\widetilde{P}$ an FMS structure where

$\mathrm{Aut}(\mathcal{X})$ acts on $\widetilde{\mathcal{Y}}$ by left multiplication. It is easy to see that P is equivalent to $\widetilde{P}$. Therefore we can replace P with $\widetilde{P}$ and wlog assume that $\mathrm{Aut}(\mathcal{X})$ action is free.

Step 2. Let $\mathcal{O} = \mathcal{Y}/\mathrm{Aut}(\mathcal{X})$ be the space of orbits of the $\mathrm{Aut}(\mathcal{X})$ action on $\mathcal{Y}$. For an orbit $o \in \mathcal{O}$, for any two elements $y_1, y_2 \in o$, the posterior distributions $\pi_1 = P_{X|Y=y_1}$ and $\pi_2 = P_{X|Y=y_2}$ (with uniform priors) differ by a permutation, by the assumption that P is FMS. In particular, π_1 and π_2 map to the same element in $\mathcal{P}(\mathcal{X})/\mathrm{Aut}(\mathcal{X})$. Therefore we can uniquely assign an element $\pi_o \in \mathcal{P}(\mathcal{X})/\mathrm{Aut}(\mathcal{X})$ for any $o \in \mathcal{O}$.

Note that by symmetry, the distribution of o does not depend on the input distribution. Let $P_o \in \mathcal{P}(\mathcal{O})$ be this distribution. Then P is equivalent to the channel $X \to (o, Z)$ where $o \sim P_o$ is independent of X, and $Z \sim \mathrm{FSC}_{\pi_o}(\cdot|X)$. (Because $\mathrm{Aut}(\mathcal{X})$ action on $\mathcal{Y}$ is free, this equivalence is in fact just renaming the output space.)

Step 3. Finally we prove that the FMS channel $X \to (o, Z)$ is equivalent to $X \to (\pi_o, Z)$. One side is easy: given (o, Z), we can generate (π_o, Z). For the other side, given (π, Z), we can generate $o' \sim P_{o|\pi_o=\pi}$. Then (o', Z) has the same distribution as (o, Z), conditioned on any input distribution. This finishes the existence proof.

Uniqueness: For any FMS channel $X \xrightarrow{Q} Y$, we can associate it with a distribution Q_π on $\mathcal{P}(\mathcal{X})/\mathrm{Aut}(\mathcal{X})$, defined as the distribution of the posterior distribution $Q_{X|Y}$, where $Y \sim Q_{Y|X} \circ \mathrm{Unif}(\mathcal{X})$ is generated with uniform prior distribution. (By definition Q_π is a distribution on $\mathcal{P}(\mathcal{X})$. However, by symmetry property of FMS, Q_π is invariant under $\mathrm{Aut}(\mathcal{X})$ action.) It is easy to see that Q_π distribution is preserved under equivalence between FMS channels. Furthermore, for an FMS channel of form $X \to (\pi, Z)$ as described in the proposition statement, this distribution of posterior distribution is equal to P_π. Therefore P_π is uniquely deteremined by P. □

In the setting of the above proposition, we call π the π-component of P, and P_π the π-distribution of P. We often use the convention that elements $\pi \in \mathcal{P}(\mathcal{X})/\mathrm{Aut}(\mathcal{X})$ satisfy $\pi_1 \geq \cdots \geq \pi_q$. Prop. 3.3 establishes an equivalence between an FMS channel and a probability distribution on $\mathcal{P}(\mathcal{X})/\mathrm{Aut}(\mathcal{X})$, which maps an FMS channel to its π-distribution. In particular, the π-distribution is an invariant property of an FMS channel under equivalence.

There are different ways to construct new FMS channels from given FMS channels. Here we present formulas for composition with Potts channels and $\star$-convolution.

Fix $q \geq 2$. For $\lambda \in [-\frac{1}{q-1}, 1]$, define Potts channel $P_\lambda : [q] \to [q]$ as

$$P_\lambda(y|x) = \lambda \mathbb{1}\{x = y\} + \frac{1-\lambda}{q} \tag{3.4}$$

for $x, y \in [q]$. Then given any q-FMS channel P, $P \circ P_\lambda$ is also a q-FMS channel. Furthermore, the π-distribution $P \circ P_\lambda$ is $f_*(P_\pi)$, where P_π is the π-distrbution of P, $f(\pi) = \lambda\pi + \frac{1-\lambda}{q}$, and f_* is the induced pushforward map.

For two q-FMS channels P and Q, their $\star$-convolution $P \star Q$ has a natural q-FMS structure. Let P_π (resp. Q_π) be the π-distrbution of P (resp. Q). Then the

π-distrbution of $P \star Q$ is

$$\mathbb{E}_{\substack{\pi\sim P_\pi \\ \pi'\sim Q_\pi}} \left[\sum_{\tau\in \mathrm{Aut}([q])} \left(\frac{1}{(q-1)!} \sum_{i\in[q]} \pi_i \pi'_{\tau(i)} \right) \mathbb{1}_{\pi\star_\tau \pi'} \right] \tag{3.5}$$

$$\text{where} \quad \pi \star_\tau \pi' := \left(\frac{\pi_i \pi'_{\tau(i)}}{\sum_{j\in[q]} \pi_j \pi'_{\tau(j)}} \right)_{i\in[q]} \in \mathcal{P}([q])/\mathrm{Aut}([q]) \tag{3.6}$$

and $\mathbb{1}_\theta \in \mathcal{P}(\mathcal{P}([q])/\mathrm{Aut}([q]))$ denotes the point distribution at $\theta \in \mathcal{P}([q])/\mathrm{Aut}([q])$.

Given any q-FMS channel P, we can restrict the input alphabet to get a q'-FMS for $q' \le q$. (Because of symmetry, the restricted channel is unique up to channel equivalence no matter what size-q' subset we choose.) In this thesis we only use the case $q' = 2$, i.e., restrict to a BMS channel. We use P^R to denote the restricted BMS channel.[2]

Let μ be the uniform distrbution on $[q]$, P be a q-FMS channel, π be its π-component. For an f-divergence, we define $C_f(P) := I_f(\mu, P)$ and call it the f-capacity of P. In particular, we make use of the following information measures.

Definition 3.4.

$$P_e(P) = \mathbb{E}\min\{1-\pi_i : i\in[q]\}, \qquad \text{(probability of error)}$$

$$C(P) = \log q - \mathbb{E}\sum_{i\in[q]} \pi_i \log\frac{1}{\pi_i}, \qquad \text{(capacity)}$$

$$C_{\chi^2}(P) = \mathbb{E}\left[q\sum_{i\in[q]} \pi_i^2 - 1 \right], \qquad (\chi^2\text{-capacity})$$

$$C_{\mathrm{SKL}}(P) = \mathbb{E}\left[\sum_{i\in[q]} \left(\pi_i - \frac{1}{q} \right) \log(\pi_i) \right]. \qquad \text{(SKL capacity)}$$

For two q-FMS channels P and Q, if $P \le_{\deg} Q$, then for any $\mu \in \mathcal{P}(\mathcal{X})$,

$$I_f(\mu, P) \le I_f(\mu, Q). \tag{3.7}$$

In particular, if P and Q are q-FMS channels with $P \le_{\deg} Q$, then we have

$$P_e(P) \ge P_e(Q), \quad C(P) \le C(Q), \quad C_{\chi^2}(P) \le C_{\chi^2}(Q), \quad C_{\mathrm{SKL}}(P) \le C_{\mathrm{SKL}}(Q). \tag{3.8}$$

From general properties of f-information discussed in Chapter 2, for q-FMS channels P and Q, we have

$$C(P\star Q) \le C(P) + C(Q), \tag{3.9}$$

$$C_{\mathrm{SKL}}(P\star Q) = C_{\mathrm{SKL}}(P) + C_{\mathrm{SKL}}(Q),. \tag{3.10}$$

[2]Here "R" stands for "restriction".

3.2 FMS channels and degradation

As shown in Lemma 2.9 and 2.10, BMS channels behave nicely with the degradation preorder (Definition 2.7). In this section, we generalize these results to FMS channels.

The following result is an equivalent condition of degradation in terms of π-components.

Proposition 3.5. *Let P, Q be two FMS channels, and π_P and π_Q be their π-components. Then $P \le_{\deg} Q$ if and only if there exists a coupling between π_P and π_Q such that*

$$\pi \le_m \mathbb{E}[\pi_Q | \pi_P = \pi] \qquad \forall \pi \in \mathcal{P}(\mathcal{X})/\operatorname{Aut}(\mathcal{X}), \tag{3.11}$$

where $\le_m$ denotes majorization (see e.g., [77, 2.18]). We use the convention that elements $\pi \in \mathcal{P}(\mathcal{X})/\operatorname{Aut}(\mathcal{X})$ are non-increasing so that the expectation is well-defined.

Proof. **Degradation $\Rightarrow$ Coupling:** Say P maps X to Y, and Q maps X to Z. Let $\pi'_P \in \mathcal{P}(\mathcal{X})$ be the posterior distribution of input X given output Y, where $Y \sim P_{Y|X} \circ \operatorname{Unif}(\mathcal{X})$ is generated with uniform prior distribution. Similarly define π'_Q. Then π_P (resp. π_Q) is the orbit of π'_P (resp. π'_Q) under permutation.

Degradation relationship $P = R \circ Q$ induces a coupling on the posterior distributions π'_P and π'_Q. One can check that this coupling is invariant under $\operatorname{Aut}(\mathcal{X})$ action and satisfies

$$\pi' = \mathbb{E}[\pi'_Q | \pi'_P = \pi'] \qquad \forall \pi' \in \mathcal{P}(\mathcal{X}). \tag{3.12}$$

For any $\pi' \in \mathcal{P}(\mathcal{X})$, let $p(\pi') \in \mathcal{P}(\mathcal{X})/\operatorname{Aut}(\mathcal{X})$ denotes its projection. Then we have

$$p(\pi') \le_m \mathbb{E}[p(\pi'_Q) | \pi'_P = \pi']. \tag{3.13}$$

Taking expectation over the orbit, we get

$$\pi \le_m \mathbb{E}[\pi_Q | \pi_P = \pi] \qquad \forall \pi \in \mathcal{P}(\mathcal{X})/\operatorname{Aut}(\mathcal{X}). \tag{3.14}$$

Coupling $\Rightarrow$ Degradation: Step 1. We prove that for $\pi, \pi' \in \mathcal{P}(\mathcal{X})/\operatorname{Aut}(\mathcal{X})$, if $\pi \le_m \pi'$, then $\operatorname{FSC}_\pi \le_{\deg} \operatorname{FSC}_{\pi'}$. Because $\pi \le_m \pi'$, there exists $a \in \mathcal{P}(\operatorname{Aut}(\mathcal{X}))$ such that (see e.g., [77, 2.20])

$$\pi_i = \sum_{\sigma \in \operatorname{Aut}(\mathcal{X})} a_\sigma \pi'_{\sigma^{-1}(i)} \qquad \forall i \in \mathcal{X}. \tag{3.15}$$

For $\rho \in \operatorname{Aut}(\mathcal{X})$, we have

$$\operatorname{FSC}_\pi(\rho | i) = \frac{1}{(q-1)!} \pi_{\rho^{-1}(i)} = \sum_{\sigma \in \operatorname{Aut}(\mathcal{X})} a_\sigma \frac{1}{(q-1)!} \pi'_{\sigma^{-1}\rho^{-1}(i)} = \sum_{\sigma \in \operatorname{Aut}(\mathcal{X})} a_\sigma \operatorname{FSC}_{\pi'}(\rho\sigma | i). \tag{3.16}$$

Therefore we can let R map $\rho\sigma$ to ρ with probability a_σ, for all $\sigma \in \operatorname{Aut}(\mathcal{X})$. This gives the desired degradation map R.

Step 2. We use the FSC mixture representation (Prop. 3.3). Suppose P maps X to (π_P, Z_P), and Q maps X to (π_Q, Z_Q). If

$$\pi = \mathbb{E}[\pi_Q | \pi_P = \pi] \qquad \forall \pi \in \mathcal{P}(\mathcal{X})/\operatorname{Aut}(\mathcal{X}), \tag{3.17}$$

then we can construct R by mapping π_Q to coupled π_P (randomly), and keeping the Z component.

Now define an FMS channel $\widetilde{P}$ whose π-component is $f(\pi_P)$, where

$$f(\pi) := \mathbb{E}[\pi_Q | \pi_P = \pi]. \tag{3.18}$$

Then by Step 1, $P \leq_{\deg} \widetilde{P}$. By Step 2, $\widetilde{P} \leq_{\deg} Q$. Therefore $P \leq_{\deg} Q$. □

To state a generalization of Lemma 2.10, we need to make a few definitions.

Definition 3.6 (Error characteristics sequence)**.** Let P be a q-FMS channel $X \to Y$. Its error characteristics sequence $\chi(P)$ is a sequence $(p_0, \ldots, p_q)$, where p_i is defined by

$$p_i = \min_{\widehat{X}: \mathcal{Y} \to \binom{\mathcal{X}}{i}} \mathbb{P}[X \notin \widehat{X}(Y)], \tag{3.19}$$

where the $\mathbb{P}$ is over $X \sim \operatorname{Unif}([q])$, $Y \sim P(\cdot|X)$. In other words, p_i is the minimum probability of error over all estimators which output a subset of $\mathcal{X}$ of size i, where an estimator succeeds if and only if the input element is contained in the output set.

Clearly, $p_0 = 1$, $p_q = 0$, and p is a non-increasing sequence. Also, for any q-FMS channel P, we have $\chi(P)_1 = P_e(P)$. In fact, the error characteristics sequence of a q-FMS channel can be computed explicitly using the FSC mixture representation (Prop. 3.3). For FSC_π with non-increasing sequence π, we have

$$\chi(\mathrm{FSC}_\pi)_i = 1 - \sum_{j \in [i]} \pi_j. \tag{3.20}$$

Therefore for an FMS channel P with π-component π, we have

$$\chi(P)_i = \mathbb{E}_\pi \left[1 - \sum_{j \in [i]} \pi_j \right]. \tag{3.21}$$

From Eq. (3.21), we see that the difference sequence $(\chi(P)_{i-1} - \chi(P)_i)_{i \in [q]}$ is always non-increasing.

In this section only, we say a sequence $(p_0, \ldots, p_q)$ is a valid sequence if $p_0 = 1$, $p_q = 0$, p is non-increasing, and the difference sequence $(p_{i-1} - p_i)_{i \in [q]}$ is non-increasing. We have shown that error characteristics sequences of FMS channels are valid sequences. The converse is also true. For any valid sequence p, we have

$$\chi(\mathrm{FSC}_{(p_{i-1} - p_i)_{i \in [q]}}) = p. \tag{3.22}$$

So the set of error characteristics sequences of FMS channels is exactly the set of valid sequences.

We define the following generalization of BECs.

Definition 3.7 (Generalized erasure channels)**.** Fix $q \in \mathbb{Z}_{\geq 2}$ and an element $b \in \mathcal{P}([q])$. We define the q-ary generalized erasure channel FEC_b as

$$\mathrm{FEC}_b = \sum_{i \in [q]} b_i \, \mathrm{FSC}_{\left(\frac{1}{i},\ldots,\frac{1}{i},0,\ldots,0\right)}, \tag{3.23}$$

where the summation notation denotes mixture of channels (i.e, the output alphabets of different FSCs in the summation are disjoint).

One can compute that

$$\chi(\mathrm{FEC}_b)_i = \sum_{i<j\leq q} b_j \left(1 - \frac{i}{j}\right). \tag{3.24}$$

Lemma 3.8. *Fix $q \in \mathbb{Z}_{\geq 2}$. For any valid sequence $p = (p_0, \ldots, p_q)$, there exists a unique $\pi \in \mathcal{P}([q])/\operatorname{Aut}([q])$ and $b \in \mathcal{P}([q])$ such that*

$$\chi(\mathrm{FSC}_\pi) = p, \qquad \chi(\mathrm{FEC}_b) = p. \tag{3.25}$$

Proof. **FSC part.** By Eq. (3.20), the unique $\pi \in \mathcal{P}([q])/\operatorname{Aut}([q])$ satisfying $\chi(\mathrm{FSC}_\pi) = p$ is defined by $\pi_i = p_{i-1} - p_i$ for $i \in [q]$.

FEC part. By Eq. (3.24), any b satisfying $\chi(\mathrm{FEC}_b) = p$ must satisfy

$$p_i = \sum_{i<j\leq q} b_j \left(1 - \frac{i}{j}\right) \quad \forall i \in [q]. \tag{3.26}$$

Taking differences we get

$$p_{i-1} - p_i = \sum_{i\leq j\leq q} \frac{b_j}{j} \quad \forall i \in [q]. \tag{3.27}$$

Taking differences again, we get

$$b_i = i(p_{i-1} - 2p_i + p_{i+1}) \quad \forall i \in [q], \tag{3.28}$$

where we assume $p_{q+1} = 0$. This proves uniqueness. For existence, we need to prove that b defined via Eq. (3.28) satisfies $b \in \mathcal{P}([q])$. Because p is a valid sequence, we have $b_i \geq 0$ for all $i \in [q]$. Furthermore,

$$\sum_{i\in[q]} b_i = \sum_{i\in[q]} (i(p_{i-1} - p_i) - i(p_i - p_{i+1})) = \sum_{i\in[q]} (p_{i-1} - p_i) = p_0 = 1. \tag{3.29}$$

Therefore $b \in \mathcal{P}([q])$ and we finish the proof. □

Finally we can state the following generalization of Lemma 2.10.

Proposition 3.9. *Fix $q \in \mathbb{Z}_{\geq 2}$. For any q-FMS channel P, we have*

$$\mathrm{FSC}_\pi \leq_{\deg} P \leq_{\deg} \mathrm{FEC}_b, \tag{3.30}$$

where FSC_π is the unique FSC with $\chi(\mathrm{FSC}_\pi) = \chi(P)$, FEC_b is the unique FEC with $\chi(\mathrm{FEC}_b) = \chi(P)$.

Proof. **FSC part.** Let P be a q-FMS channel and π_P be its π-component. Define

$$\pi := \mathbb{E}\pi_P. \tag{3.31}$$

Then $\mathrm{FSC}_\pi \leq_{\deg} P$ by Prop. 3.5 and $\chi(\mathrm{FSC}_\pi) = \chi(P)$ by Eq. (3.20), Eq. (3.21).

FEC part. Let P be a q-FMS channel and π_P be its π-component. The class of FECs is the class of mixtures of $\mathrm{FSC}_{\left(\frac{1}{i},\ldots,\frac{1}{i},0,\ldots,0\right)}$ for $i \in [q]$. So the class of FECs is closed (up to channel equivalence) under taking mixtures. If we have proved the result for all FSCs, then by taking the mixture $P = \mathbb{E}\,\mathrm{FSC}_{\pi_P}$, we also prove the result for P. Therefore it suffices to prove the case $P = \mathrm{FSC}_\pi$, where $\pi \in \mathcal{P}([q])/\operatorname{Aut}([q])$ is non-increasing.

We define a sequence b as

$$b_i := i(\pi_i - \pi_{i+1}) \quad \forall i \in [q], \tag{3.32}$$

where we assume $\pi_{q+1} = 0$.

We prove that $b \in \mathcal{P}([q])$. Because π is non-increasing, $b_i \geq 0$ for all $i \in [q]$. Furthermore,

$$\sum_{i\in[q]} b_i = \sum_{i\in[q]} i(\pi_i - \pi_{i+1}) = \sum_{j\in[q]} \pi_j = 1. \tag{3.33}$$

Therefore $b \in \mathcal{P}([q])$, and FEC_b is well-defined.

We have

$$\pi = \sum_{i\in[q]} b_i \cdot \left(\frac{1}{i},\ldots,\frac{1}{i},0,\ldots,0\right). \tag{3.34}$$

By Prop. 3.5, this implies that $\mathrm{FSC}_\pi \leq_{\deg} \mathrm{FEC}_b$.

Finally, $\chi(\mathrm{FSC}_\pi) = \chi(\mathrm{FEC}_b)$ by Eq. (3.20), Eq. (3.24). □

Corollary 3.10. *Fix $q \in \mathbb{Z}_{\geq 2}$. For any $c \in [0, 1 - \frac{1}{q}]$, for any q-FMS channel P with $P_e(P) = c$, we have*

$$\mathrm{FSC}_\pi \leq_{\deg} P, \tag{3.35}$$

where

$$\pi = \left(1 - c, \frac{c}{q-1}, \ldots, \frac{c}{q-1}\right). \tag{3.36}$$

Proof. By Prop. 3.9 and Eq. (3.20), it suffices to prove that for all $\pi' \in \mathcal{P}([q])/\operatorname{Aut}([q])$ with $P_e(\mathrm{FSC}_{\pi'}) = c$, we have $\mathrm{FSC}_\pi \leq_{\deg} \mathrm{FSC}_{\pi'}$, where π is as defined in Eq. (3.36). The result then follows from Prop. 3.5, and the fact that $\pi \leq_m \pi'$ for all $\pi' \in \mathcal{P}([q])/\operatorname{Aut}([q])$ satisfying $\pi'_1 = 1 - c$. □

The following example shows that among FMS channels of the same probability of error, there does not necessary exist a least degraded one.

Example 3.11. Take $q = 3$ and $c = \frac{1}{4}$. We show that there does not exist a least degraded FMS channel among all FMS channels of probability of error c. Assume for the sake of contradiction that such a channel exists. By Prop. 3.9, such an channel must be of form FEC_b for some $b \in \mathcal{P}([q])$.

Let $b' = \left(\frac{1}{2}, \frac{1}{2}, 0\right)$ and $b'' = \left(\frac{5}{8}, 0, \frac{3}{8}\right)$. Then we have $\chi(\mathrm{FEC}_{b'}) = \left(1, \frac{1}{4}, 0, 0\right)$ and $\chi(\mathrm{FEC}_{b''}) = \left(1, \frac{1}{4}, \frac{1}{8}, 0\right)$. Because $\mathrm{FEC}_{b'} \leq_{\deg} \mathrm{FEC}_b$, we have $\chi(\mathrm{FEC}_b)_2 \leq \chi(\mathrm{FEC}_{b'})_2 = 0$, thus $b_3 = 0$. Because $\mathrm{FEC}_{b''} \leq_{\deg} \mathrm{FEC}_b$, we have $b_1 \geq b''_1 = \frac{5}{8}$. Therefore $b_2 = b_1 \leq \frac{3}{8}$. However, this means that $P_e(\mathrm{FEC}_b) = \frac{1}{2}b_2 + \frac{2}{3}b_3 \leq \frac{3}{16} < \frac{1}{4}$. This leads to contraction.

3.3 Local subadditivity of χ^2-capacity

As shown in Lemma 2.5, χ^2-capacity of BMS channels is subadditive under $\star$-convolution. This property is no longer true for channels with larger input alphabets. Nevertheless, we show that χ^2-capacity of FMS channels satisfies a local version of subadditivity. This result is further extended in Theorem 8.6 to general channels. The proof presented here is based on [73] and uses the FSC mixture representation of FMS channels.

Theorem 3.12 (Local subadditivity of χ^2-capacity for FMS channels)**.** *Fix $q \in \mathbb{Z}_{\geq 2}$. For any $\epsilon > 0$ and q-FMS channels P, Q with $C_{\chi^2}(P) \leq \epsilon$, we have*

$$C_{\chi^2}(P \star Q) \leq (1 + O_q(\epsilon^{1/5}))(C_{\chi^2}(P) + C_{\chi^2}(Q)), \tag{3.37}$$

where O_q hides a constant depending on q.

The remaining of this section is devoted to the proof of Theorem 3.12. We first prove the special case of FSCs.

Lemma 3.13. *Fix $q \in \mathbb{Z}_{\geq 2}$. For any $\epsilon > 0$ and $\pi, \pi' \in \mathcal{P}([q])/\operatorname{Aut}([q])$ with $C_{\chi^2}(\mathrm{FSC}_{\pi'}) \leq \epsilon$, we have*

$$C_{\chi^2}(\mathrm{FSC}_\pi \star \mathrm{FSC}_{\pi'}) \leq (1 + O_q(\epsilon^{1/2}))(C_{\chi^2}(\mathrm{FSC}_\pi) + C_{\chi^2}(\mathrm{FSC}_{\pi'})). \tag{3.38}$$

Proof of Theorem 3.12 given Lemma 3.13. Let π_P (resp. π_Q) be the π-component of P (resp. Q).

Because the constant does not depend on Q, it suffices to prove the case where Q is an FSC, i.e., π_Q is fixed.

If $C_{\chi^2}(Q) \le \epsilon^{2/5}$, then by Lemma 3.13, we have

$$\begin{aligned} C_{\chi^2}(P \star Q) &= \mathbb{E}_{\pi_P} C_{\chi^2}(\mathrm{FSC}_{\pi_P} \star \mathrm{FSC}_{\pi_Q}) \\ &\le \mathbb{E}_{\pi_P}\left[(1 + O_q(\epsilon^{1/5}))(C_{\chi^2}(\mathrm{FSC}_{\pi_P}) + C_{\chi^2}(\mathrm{FSC}_{\pi_Q}))\right] \\ &= (1 + O_q(\epsilon^{1/5}))(C_{\chi^2}(P) + C_{\chi^2}(Q)). \end{aligned} \tag{3.39}$$

In the following we assume that $C_{\chi^2}(Q) > \epsilon^{2/5}$. By Markov's inequality, we have

$$\mathbb{P}\left[C_{\chi^2}(\mathrm{FSC}_{\pi_P}) \ge \epsilon^{2/5}\right] \le \epsilon^{3/5}. \tag{3.40}$$

Write

$$\begin{aligned} C_{\chi^2}(P \star Q) &= \mathbb{E}_{\pi_P}\left[C_{\chi^2}(\mathrm{FSC}_{\pi_P} \star \mathrm{FSC}_{\pi_Q}) \mathbb{1}\{C_{\chi^2}(\mathrm{FSC}_{\pi_P}) \le \epsilon^{2/5}\}\right] \\ &\quad + \mathbb{E}_{\pi_P}\left[C_{\chi^2}(\mathrm{FSC}_{\pi_P} \star \mathrm{FSC}_{\pi_Q}) \mathbb{1}\{C_{\chi^2}(\mathrm{FSC}_{\pi_P}) > \epsilon^{2/5}\}\right] \\ &=: L + R. \end{aligned} \tag{3.41}$$

For L, by Lemma 3.13, we have

$$\begin{aligned} L &\le (1 + O_q(\epsilon^{1/5}))\mathbb{E}_{\pi_P}\left[(C_{\chi^2}(\mathrm{FSC}_{\pi_P}) + C_{\chi^2}(\mathrm{FSC}_{\pi_Q})) \mathbb{1}\{C_{\chi^2}(\mathrm{FSC}_{\pi_P}) \le \epsilon^{2/5}\}\right] \\ &\le (1 + O_q(\epsilon^{1/5}))(C_{\chi^2}(P) + C_{\chi^2}(Q)). \end{aligned} \tag{3.42}$$

For R, by Eq. (3.40) and the assumption that $C_{\chi^2}(Q) > \epsilon^{2/5}$, we have

$$\mathbb{E}_{\pi_P}\left[C_{\chi^2}(\mathrm{FSC}_{\pi_P} \star \mathrm{FSC}_{\pi_Q}) \mathbb{1}\{C_{\chi^2}(\mathrm{FSC}_{\pi_P}) > \epsilon^{2/5}\}\right] \le O_q(\epsilon^{3/5}) \le O_q(\epsilon^{1/5}) C_{\chi^2}(Q). \tag{3.43}$$

Combining Eq. (3.42) and Eq. (3.43) we finish the proof. □

Proof of Lemma 3.13. Because the statement is monotone in ϵ, we can wlog assume that $C_{\chi^2}(\mathrm{FSC}_{\pi'}) = \epsilon$.

Let $\pi'_i = \frac{1+\epsilon_i}{q}$. Then

$$\sum_i \epsilon_i = 0, \quad C_{\chi^2}(\mathrm{FSC}_{\pi'}) = \frac{1}{q}\sum_i \epsilon_i^2 = \epsilon. \tag{3.44}$$

By Eq. (3.5),

$$C_{\chi^2}(\mathrm{FSC}_\pi \star \mathrm{FSC}_{\pi'}) = q \sum_{\tau \in S_q} \frac{1}{(q-1)!} \cdot \frac{\sum_i \pi_i^2 \pi'^2_{\tau(i)}}{\sum_i \pi_i \pi'_{\tau(i)}} - 1$$
$$= q \sum_{\tau \in S_q} \frac{1}{q!} \cdot \frac{\sum_i \pi_i^2 (1+\epsilon_{\tau(i)})^2}{1+\sum_i \pi_i \epsilon_{\tau(i)}} - 1. \tag{3.45}$$

Recall the following basic equality.

$$\frac{1}{1+x} = 1 - x + x^2 - \frac{x^3}{1+x}. \tag{3.46}$$

We apply (3.46) with $x = \sum_i \pi_i \epsilon_{\tau(i)}$. Because $|x| = O_q(\epsilon^{1/2})$, we have

$$\left| \frac{x^3}{1+x} \right| = O_q(\epsilon^{3/2}). \tag{3.47}$$

So

$$\begin{aligned}
&\frac{\sum_i \pi_i^2 (1+\epsilon_{\tau(i)})^2}{1+\sum_i \pi_i \epsilon_{\tau(i)}} \\
&= \left(\sum_i \pi_i^2 (1+\epsilon_{\tau(i)})^2 \right) \left(1 - x + x^2 - \frac{x^3}{1+x} \right) \\
&\le \left(\sum_i \pi_i^2 (1+\epsilon_{\tau(i)})^2 \right) \left(1 - \sum_i \pi_i \epsilon_{\tau(i)} + \left(\sum_i \pi_i \epsilon_{\tau(i)} \right)^2 + O_q(\epsilon^{3/2}) \right) \\
&\le \left(\sum_i \pi_i^2 (1+\epsilon_{\tau(i)})^2 \right) \left(1 - \sum_i \pi_i \epsilon_{\tau(i)} + \left(\sum_i \pi_i \epsilon_{\tau(i)} \right)^2 \right) + O_q(\epsilon^{3/2}),
\end{aligned} \tag{3.48}$$

where the last step is by

$$\sum_i \pi_i^2 (1+\epsilon_{\tau(i)})^2 = O(1). \tag{3.49}$$

Let us expand the first summand in (3.48).

$$\begin{aligned}
&\left(\sum_i \pi_i^2(1+\epsilon_{\tau(i)})^2\right)\left(1-\sum_i \pi_i\epsilon_{\tau(i)}+\left(\sum_i \pi_i\epsilon_{\tau(i)}\right)^2\right)\\
&=\left(\sum_i \pi_i^2+2\sum_i \pi_i^2\epsilon_{\tau(i)}+\sum_i \pi_i^2\epsilon_{\tau(i)}^2\right)\left(1-\sum_i \pi_i\epsilon_{\tau(i)}+\left(\sum_i \pi_i\epsilon_{\tau(i)}\right)^2\right)\\
&=: (① + ② + ③)(1 - ④ + ⑤)\\
&= ① - ①④ + ①⑤ + ② - ②④ + ②⑤ + ③ - ③④ + ③⑤. \qquad (3.50)
\end{aligned}$$

Note that we have the following loose bounds:

$$① = O_q(1), \quad |②| = O_q(\epsilon^{1/2}), \quad ③ \le O_q(\epsilon), \quad |④| \le O_q(\epsilon^{1/2}), \quad ⑤ \le O_q(\epsilon). \qquad (3.51)$$

Let us study every term under $\sum_{\tau\in S_q} \frac{1}{q!}$. For simplicity, write

$$A = \sum_i \pi_i^2 = \frac{1}{q}\left(1 + C_{\chi^2}(\mathrm{FSC}_\pi)\right). \qquad (3.52)$$

①:

$$\sum_{\tau\in S_q} \frac{1}{q!}\cdot ① = \sum_{\tau\in S_q} \frac{1}{q!}\sum_i \pi_i^2 = A. \qquad (3.53)$$

①④:

$$\sum_{\tau\in S_q} \frac{1}{q!}\cdot ①④ = \sum_{\tau\in S_q} \frac{1}{q!}\left(\sum_i \pi_i^2\right)\left(\sum_i \pi_i\epsilon_{\tau(i)}\right) = 0. \qquad (3.54)$$

①⑤:

$$\begin{aligned}
\sum_{\tau\in S_q}\frac{1}{q!}\cdot ①⑤ &= \sum_{\tau\in S_q}\frac{1}{q!}\left(\sum_i \pi_i^2\right)\left(\sum_i \pi_i\epsilon_{\tau(i)}\right)^2 \\
&= A\sum_{i,j}\sum_{\tau\in S_q}\frac{1}{q!}\cdot \pi_i\pi_j\epsilon_{\tau(i)}\epsilon_{\tau(j)} \\
&= A\sum_{i,j}\epsilon_i\epsilon_j\sum_{\tau\in S_q}\frac{1}{q!}\cdot \pi_{\tau(i)}\pi_{\tau(j)} \\
&= A\sum_{i,j}\epsilon_i\epsilon_j\begin{cases} \frac{1}{q}\sum_k \pi_k^2 & i=j, \\ \frac{1}{q(q-1)}\left(1-\sum_k \pi_k^2\right) & i\neq j, \end{cases} \\
&= A\left(\sum_i \epsilon_i^2\cdot\frac{1}{q}\sum_k \pi_k^2 + \sum_i \epsilon_i(-\epsilon_i)\cdot\frac{1}{q(q-1)}\left(1-\sum_k \pi_k^2\right)\right) \\
&= A\cdot\frac{1}{q}\sum_i \epsilon_i^2\left(\sum_k \pi_k^2 - \frac{1}{q-1}\left(1-\sum_k \pi_k^2\right)\right) \\
&= \epsilon A\cdot\frac{qA-1}{q-1}. \qquad (3.55)
\end{aligned}$$

②:

$$\sum_{\tau\in S_q}\frac{1}{q!}\cdot ② = \sum_{\tau\in S_q}\frac{1}{q!}\cdot 2\sum_i \pi_i^2\epsilon_{\tau(i)} = 0. \qquad (3.56)$$

②④:

$$\begin{aligned}
\sum_{\tau\in S_q}\frac{1}{q!}\cdot ②④ &= \sum_{\tau\in S_q}\frac{1}{q!}\left(2\sum_i \pi_i^2\epsilon_{\tau(i)}\right)\left(\sum_i \pi_i\epsilon_{\tau(i)}\right)\\
&= \sum_{i,j}\sum_{\tau\in S_q}\frac{1}{q!}\cdot 2\pi_i^2\pi_j\epsilon_{\tau(i)}\epsilon_{\tau(j)}\\
&= \sum_{i,j} 2\epsilon_i\epsilon_j \sum_{\tau\in S_q}\frac{1}{q!}\cdot \pi_{\tau(i)}^2\pi_{\tau(j)}\\
&= \sum_{i,j} 2\epsilon_i\epsilon_j \begin{cases} \frac{1}{q}\sum_k \pi_k^3 & i=j,\\ \frac{1}{q(q-1)}\sum_k \pi_k^2(1-\pi_k) & i\neq j,\end{cases}\\
&= 2\sum_i \epsilon_i^2\cdot\frac{1}{q}\sum_k \pi_k^3 + 2\sum_i \epsilon_i(-\epsilon_i)\cdot\frac{1}{q(q-1)}\sum_k \pi_k^2(1-\pi_k)\\
&= 2\cdot\frac{1}{q}\sum_i \epsilon_i^2\left(\sum_k \pi_k^3 - \frac{1}{q-1}\sum_k \pi_k^2(1-\pi_k)\right)\\
&= 2\epsilon\cdot\frac{1}{q-1}\left(q\sum_k \pi_k^3 - \sum_k \pi_k^2\right)\geq 0,
\end{aligned} \tag{3.57}$$

where the last step is by

$$\sum_k \pi_k^3 = \left(\sum_k \pi_k^3\right)\left(\sum_k \pi_k\right)\geq\left(\sum_k \pi_k^2\right)^2\geq\frac{1}{q}\left(\sum_k \pi_k^2\right). \tag{3.58}$$

②⑤: By (3.51),

$$\left|\sum_{\tau\in S_q}\frac{1}{q!}\cdot ②⑤\right| = O_q(\epsilon^{3/2}). \tag{3.59}$$

③:

$$\sum_{\tau\in S_q}\frac{1}{q!}\cdot ③ = \sum_{\tau\in S_q}\frac{1}{q!}\left(\sum_i \pi_i^2\epsilon_{\tau(i)}^2\right) = \sum_i \pi_i^2\cdot\frac{1}{q}\sum_j \epsilon_j^2 = \epsilon A. \tag{3.60}$$

③④: By (3.51),

$$\left|\sum_{\tau\in S_q}\frac{1}{q!}\cdot ③④\right| = O_q(\epsilon^{3/2}). \tag{3.61}$$

③⑤: By (3.51),

$$\left|\sum_{\tau\in S_q}\frac{1}{q!}\cdot ③⑤\right| = O_q(\epsilon^2). \tag{3.62}$$

Plugging (3.53) - (3.62) into (3.50)(3.48)(3.45), we get

$$\begin{aligned}
C_{\chi^2}(\mathrm{FSC}_\pi \star \mathrm{FSC}_{\pi'}) &\le q\left(A + \epsilon A\cdot\frac{qA-1}{q-1} + \epsilon A + O_q(\epsilon^{3/2})\right) - 1\\
&= (qA-1)\left(1+\frac{q\epsilon A}{q-1}+\epsilon\right) + \epsilon + O_q(\epsilon^{3/2})\\
&= \left(1+\frac{q\epsilon A}{q-1}+\epsilon\right) C_{\chi^2}(\mathrm{FSC}_\pi) + (1+O_q(\epsilon^{1/2}))C_{\chi^2}(\mathrm{FSC}_{\pi'})\\
&= (1+O_q(\epsilon^{1/2}))(C_{\chi^2}(\mathrm{FSC}_\pi) + C_{\chi^2}(\mathrm{FSC}_{\pi'})),
\end{aligned} \tag{3.63}$$

where the last step is by $A \le 1$. □

Chapter 4

Contraction properties of the Potts semigroup

We perform a detailed analysis of contraction properties of the Potts channels, which are the simplest examples of FMS channels (Definition 3.1). The ferromagnetic Potts channels form the semigroup of random walk on a complete graph, which we call the Potts semigroup. [51] computed the maximum ratio between relative entropy and the Dirichlet form, obtaining the log-Sobolev constant α_2 in the log-Sobolev inequality (LSI, or 2-LSI) for the semigroup. We obtain the best possible non-linear inequalities, p-non-linear log-Sobolev inequalities (p-NLSIs, $p \geq 1$), relating entropy and the Dirichlet form. As an example, we show that the 1-log-Sobolev constant (also known as the modified log-Sobolev constant) satisfies $\alpha_1 = 1 + \frac{1 \pm o(1)}{\log q}$. By integrating the 1-NLSIs we obtain the tight non-linear strong data processing inequalities (SDPIs) for the Potts channels, and derive formulas and bounds on the KL contraction coefficients. These results are used in Chapter 6 to derive non-reconstruction results for the Potts model on a tree and impossibility of weak recovery for the stochastic block model with q communities. This chapter is based on [72].

Chapter outline In Section 4.1 we introduce the problem and our main results. In Section 4.2 we prove the sharpest p-NLSIs for the Potts semigroup (Theorem 4.1), and compute the input-restricted KL divergence contraction coefficients of all Potts channels (Theorem 4.3). In Section 4.3 we discuss tensorization of p-NLSIs for the Potts semigroup, and non-linear SDPI for Potts channels. In Section 4.4 we compute the input-restricted KL contraction coefficient for the coloring channel, a special case of the Potts channels. In Section 4.5 we compute the input-unrestricted KL contraction coefficient for the Potts channels. In Section 4.6 we prove upper bounds on the input-restricted KL contraction coefficient for the Potts channels. In Section 4.7 we show that the Mrs. Gerber's Lemma, a very useful result for BSCs, is not true for non-binary Potts channels. In Section 4.8 we prove a concavity property of the log-Sobolev coefficients.

4.1 Introduction

Log-Sobolev inequalities Log-Sobolev inequalities (LSIs) are a class of inequalities bounding the rate of convergence of a Markov semigroup to its stationary distribution. They upper bound certain relative entropy (KL divergence) functions via a multiple of the Dirichlet form.

Let $\mathcal{X}$ be a finite alphabet and $K : \mathcal{X} \to \mathcal{X}$ be a Markov kernel. Let $L = K - I$. We consider the semigroup $(T_t)_{t\geq 0}$, where $T_t = \exp(tL)$. Let π be a stationary measure for the semigroup. For $f, g : \mathcal{X} \to \mathbb{R}$, the Dirichlet form is defined by

$$\mathcal{E}(f, g) := -\mathbb{E}_\pi[(Lf)g] = -\sum_{x,y\in\mathcal{X}} L(x, y)f(y)g(x)\pi(x). \tag{4.1}$$

For non-zero $f : \mathcal{X} \to \mathbb{R}_{\geq 0}$, the relative entropy is defined by

$$\mathrm{Ent}_\pi(f) := \mathbb{E}_\pi\left[f \log \frac{f}{\mathbb{E}_\pi[f]}\right] = \mathbb{E}_\pi[f]D(\pi^{(f)}||\pi) \tag{4.2}$$

where $\pi^{(f)}$ is a distribution defined as $\pi^{(f)}(x) = \frac{f(x)\pi(x)}{\mathbb{E}_\pi[f]}$.

For $p > 1$, we say the semigroup $(T_t)_{t\geq 0}$ admits p-log-Sobolev inequality (p-LSI), if for some constant α_p, for all non-zero non-negative real functions f on $\mathcal{X}$, we have

$$\mathrm{Ent}_\pi(f) \leq \frac{1}{\alpha_p}\mathcal{E}\left(f^{\frac{1}{p}}, f^{1-\frac{1}{p}}\right). \tag{4.3}$$

For $p = 1$, we define 1-LSI as

$$\mathrm{Ent}_\pi(f) \leq \frac{1}{\alpha_1}\mathcal{E}(f, \log f). \tag{4.4}$$

The case $p = 2$ is the standard log-Sobolev inequality, originally studied in [71]. The case $p = 1$ is studied also under the name "modified log-Sobolev inequality" (e.g. [70, 24]).

The relationship between 1-LSI and semigroup convergence can be seen from the following identity

$$\frac{d}{dt}|_{t=0}\,\mathrm{Ent}_\pi(T_t f) = -\mathcal{E}(f, \log f). \tag{4.5}$$

Therefore

$$\mathrm{Ent}_\pi(T_t f) \leq \exp(-\alpha_1 t)\,\mathrm{Ent}_\pi(f) \tag{4.6}$$

which corresponds to a property of T_t to exponentially fast relax to equillibrium (in the sense of relative entropy).

[114] introduced non-linear p-log-Sobolev inequalities (p-NLSI), a finer description of the relationship between relative entropy and Dirichlet forms. For $p \geq 1$, we say the semigroup satisfies p-LSI if for some non-negative function[1] $\Phi_p : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$, for

[1] [114] requires the function Φ_p to be concave. We do not make this assumption initially, however to extend these inequalities to product semigroups the concavification will be necessary – see

all non-zero $f : \mathcal{X} \to \mathbb{R}_{\geq 0}$, we have

$$\frac{\mathrm{Ent}_\pi(f)}{\mathbb{E}_\pi[f]} \leq \Phi_p \left(\frac{\mathcal{E}\left(f^{\frac{1}{p}}, f^{1-\frac{1}{p}}\right)}{\mathbb{E}_\pi[f]} \right), \tag{4.7}$$

where for $p = 1$, $\mathcal{E}\left(f^{\frac{1}{p}}, f^{1-\frac{1}{p}}\right)$ should be replaced with $\mathcal{E}(f, \log f)$.

Non-linear p-log-Sobolev inequalities imply the ordinary p-log-Sobolev inequalities for

$$\alpha_p = \inf_{x>0} \frac{x}{\Phi_p(x)}. \tag{4.8}$$

When Φ_p is concave, this can be further simplified to $\alpha_p = (\Phi_p'(0))^{-1}$.

[106] proved that for reversible (π, K),

$$\frac{p^2(p'-1)}{(p')^2(p-1)}\alpha_p \leq \alpha_{p'} \leq \alpha_p \tag{4.9}$$

for $1 < p' \leq p \leq 2$. We discuss some general facts about dependence of α_p and Φ_p on p in Section 4.8.

Potts semigroup We focus on the simplest Markov semigroup, corresponding to the random walk on a complete graph. The Markov kernel is $K(x, y) = \frac{1}{q-1}\mathbb{1}\{x \neq y\}$, where $q = |\mathcal{X}|$. In the following, we always assume $\mathcal{X} = [q]$. We call it the Potts semigroup, because every operator T_t in the semigroup is a ferromagnetic Potts channel. Its stationary distribution π is uniform on $\mathcal{X}$ and its Dirichlet form is rescaled covariance:

$$\mathcal{E}(f, g) = \frac{q}{q-1}\mathrm{Cov}_\pi(f, g). \tag{4.10}$$

The Potts channel $P_\lambda : [q] \to [q]$ for $\lambda \in \left[-\frac{1}{q-1}, 1\right]$ is defined by

$$P_\lambda(x, y) = \begin{cases} \lambda + \frac{1-\lambda}{q}, & \text{if } x = y, \\ \frac{1-\lambda}{q}, & \text{if } x \neq y. \end{cases} \tag{4.11}$$

We parametrize them by λ, the second largest eigenvalue of P_λ. We say the channel is ferromagnetic if $\lambda > 0$; antiferromagnetic if $\lambda < 0$. One can see that T_t in the Potts semigroup is exactly $P_{\exp(-\frac{q}{q-1}t)}$.

[51] computed the 2-log-Sobolev constant

$$\alpha_2 = \frac{q-2}{(q-1)\log(q-1)}. \tag{4.12}$$

Section 4.3.

They observed that the infimum of the ratio $\frac{\mathcal{E}(f,f)}{\mathrm{Ent}_\pi[f^2]}$ is achieved at a two-valued function f, i.e., f takes exactly two values. In fact, the infimum is achieved at a function f where $f(1) = q-1$ and $f(i) = 1$ for $i \neq 1$. For $p \neq 2$, it seems hard to give a closed-form expression for α_p. [70] proved that

$$\frac{q}{q-1} \leq \alpha_1 \leq \left(1 + \frac{4}{\log(q-1)}\right) \frac{q}{q-1}, \tag{4.13}$$

where the upper bound is by using a two-valued function f, where $f(1) = q+1$ and $f(i) = 1$ for $i \neq 1$. [24] also discussed bounds on α_1 and α_2, proving that

$$\alpha_1 \geq \frac{q}{q-1} + \frac{2}{\sqrt{q-1}}. \tag{4.14}$$

These computations lead to the guess that for all p, the best possible p-LSI constant α_p for the Potts semigroup is achieved at a two-valued function. In Section 4.2, we prove that this is true, and in fact true for p-NLSIs for the Potts semigroup: For fixed $\frac{\mathrm{Ent}_\pi(f)}{\mathbb{E}_\pi[f]}$, the unique function (up to scalar multiplication) of the form $f(1) \geq f(2) = \cdots = f(q)$ minimizes $\frac{\mathcal{E}\left(f^{\frac{1}{p}}, f^{1-\frac{1}{p}}\right)}{\mathbb{E}_\pi[f]}$. As a result we get the sharpest p-NLSIs for the Potts semigroup for all $p \geq 1$.

We define a useful function $\psi : [0,1] \to \mathbb{R}$ as follows.

$$\psi(x) := \log q + x \log x + (1-x) \log \frac{1-x}{q-1}. \tag{4.15}$$

Note that $\psi(x)$ is the KL divergence between $\left(x, \frac{1-x}{q-1}, \ldots, \frac{1-x}{q-1}\right)$ and $\mathrm{Unif}([q])$. Simple computation shows that ψ is non-negative, convex, $\psi\left(\frac{1}{q}\right) = 0$, strictly decreasing on $[0, \frac{1}{q}]$, strictly increasing on $\left[\frac{1}{q}, 1\right]$, and takes value in $[0, \log q]$.

We define the following useful functions. For $p > 1$, define $\xi_p : [0,1] \to \mathbb{R}$ as

$$\xi_p(x) = \frac{q}{q-1}\left(1 - \frac{1}{q}\left(x^{\frac{1}{p}} + (q-1)\left(\frac{1-x}{q-1}\right)^{\frac{1}{p}}\right)\left(x^{1-\frac{1}{p}} + (q-1)\left(\frac{1-x}{q-1}\right)^{1-\frac{1}{p}}\right)\right). \tag{4.16}$$

Define $\xi_1 : [0,1] \to \mathbb{R}$ as

$$\xi_1(x) = \frac{1}{q-1}\left(-\log x - (q-1)\log\frac{1-x}{q-1} + q\left(x \log x + (1-x)\log\frac{1-x}{q-1}\right)\right). \tag{4.17}$$

For $p \geq 1$, define $b_p : [0, \log q] \to \mathbb{R}$ as[2]

$$b_p(\psi(x)) = \xi_p(x) \tag{4.18}$$

for $x \in \left[\frac{1}{q}, 1\right]$, where ψ is defined in (4.15).

Theorem 4.1 (p-NLSI for Potts semigroup)**.** *Fix $p \geq 1$. The Potts semigroup satisfies p-NLSI with $\Phi_p = b_p^{-1}$, where b_p is defined in* (4.18)*. Furthermore, this is the best possible p-NLSI.*

In other words, for any $c \in [0, \log q]$, among all functions $f : [q] \to \mathbb{R}_{\geq 0}$ with $\mathbb{E}_\pi f = 1$, $\mathrm{Ent}(f) = c$, there is a unique (up to permuting the alphabet) minimizer of $\mathcal{E}\left(f^{\frac{1}{p}}, f^{1-\frac{1}{p}}\right)$ ($\mathcal{E}(f, \log f)$ for $p = 1$), and it is of form $\left(x, \frac{q-x}{q-1}, \cdots, \frac{q-x}{q-1}\right)$ with $x \in [1, q]$.

In particular, we have

$$\alpha_p = \inf_{x \in \left(\frac{1}{q}, 1\right]} \frac{\xi_p(x)}{\psi(x)}. \tag{4.19}$$

As a corollary of our 1-NLSI, we derive the second order behavior of α_1 as q goes to ∞.

Proposition 4.2. *For $q \geq 3$, we have*

$$\frac{q}{q-1}\left(1 + \frac{1}{\log q}\right) \leq \alpha_1 \leq \frac{q}{q-1}\left(1 + \frac{1 + o(1)}{\log q}\right). \tag{4.20}$$

Strong data processing inequalities For a review of the strong data processing inequalities (SDPIs), see Section 2.4. Here we introduce a non-linear version of SDPI. That is, for all Markov chains $U \to X \to Y$ with fixed channel $P_{Y|X}$, arbitrary U, and fixed or arbitrary P_X, we have

$$I(U; Y) \leq s(I(U; X)), \tag{4.21}$$

for some non-linear function s depending on $P_{Y|X}$ and potentially P_X. See [118] for more background on SDPIs and their relationship with LSIs.

From (4.6) and Prop. 4.2 we obtain

$$\eta_{\mathrm{KL}}(\pi, P_\lambda) \leq \lambda^{\frac{q-1}{q}\alpha_1} = \lambda^{1 + \frac{1+o(1)}{\log q}} \tag{4.22}$$

for $\lambda \in [0, 1]$ and $o(1) \to 0$ as $q \to \infty$. It turns out that 1-NLSI can be seen as an infinitesimal version of the non-linear SDPIs (see [114, Theorem 2]). Thus, we can prove the best possible non-linear SDPI for the Potts channels.

[2] In the case $q = 2$, b_p differs from [114] by a constant factor due to a different parametrization of the semigroup.

Theorem 4.3 (Non-linear SDPI for Potts channel)**.** *Fix* $\lambda \in \left[-\frac{1}{q-1}, 1\right]$. *Define* $s_\lambda : [0, \log q] \to \mathbb{R}$ *as*

$$s_\lambda(\psi(x)) = \psi\left(\lambda x + \frac{1-\lambda}{q}\right), \tag{4.23}$$

for $x \in \left[\frac{1}{q}, 1\right]$, *where* ψ *is defined in* (4.15). *Let* $\widehat{s}_\lambda$ *be the concave envelope of* s_λ. *For any Markov chain* $U \to X \to Y$ *where* X *has uniform distribution and* $X \to Y$ *is the Potts channel* P_λ, *we have*

$$I(U;Y) \le \widehat{s}_\lambda(I(U;X)). \tag{4.24}$$

In particular, we have

$$\eta_{\mathrm{KL}}(\pi, P_\lambda) = \sup_{x \in \left(\frac{1}{q}, 1\right]} \frac{\psi\left(\lambda x + \frac{1-\lambda}{q}\right)}{\psi(x)}. \tag{4.25}$$

Furthermore, this is the best possible non-linear SDPI for Potts channels, in the sense that for any $c \in [0, \log q]$, *there exists a Markov chain* $U \to X \to Y$ *where* X *has uniform distribution,* $X \to Y$ *is the Potts channel* P_λ, *and* $I(U;X) = c$, *such that* $I(U;Y) = \widehat{s}_\lambda(c)$.

To compare the input-restricted η_{KL} with input-unrestricted one, in Section 4.5 we compute the exact value of $\eta_{\mathrm{KL}}(P_\lambda)$, and in Section 4.6 we prove that

$$\eta_{\mathrm{KL}}(\pi, P_\lambda) < \eta_{\mathrm{KL}}(P_\lambda) \tag{4.26}$$

for $q \ge 3$ and $\lambda \in \left[-\frac{1}{q-1}, 0\right) \cup (0, 1)$. See Section 4.2.4 for discussion on the tightness of the bound (4.22).

Let us briefly remark on the tensorization properties of the p-NLSIs and SDPIs. In Section 4.3 we extend p-NLSI and SDPIs to product spaces/channels. In these results, the functions $\check{b}_p$ (convexification of b_p) and $\widehat{s}_\lambda$ (concavification of s_λ) appear naturally. When $q = 2$, b_p is already convex, and s_λ is concave, leading to many good properties for the hypercube and for binary symmetric channels (e.g. Mrs. Gerber's Lemma [134]). However, as shown in Prop. 4.20, these properties do not hold anymore for $q \ge 3$, implying a different structure of extremal distributions that are the slowest to relax to equillibrium as $t \to \infty$ in $T_t^{\times n}$, see Section 4.3.2.

Applications One of the implications of NLSIs are improved hypercontractivity inequalities for functions in $[q]^n$ supported on subsets of cardinality $q^{(1-\epsilon)n}$ – this was established generally (for any semigroup) in [114]. Here, we show how NLSIs can be used to close the gap (between functional-analytic proofs and explicit combinatorics of [90]) in the edge-isoperimetric inequality for the $[q]^n$ – see Section 4.3.3.

Similarly, SDPIs have numerous applications. Originally introduced to study cer-

tain multi-user data-compression questions in information theory, they have been since adopted in many different scenarios. For example, [64] used SDPIs to investigate fundamental limits of fault tolerant computing. [115, 115] further developed the idea and related the amount of information transmitted in a directed or undirected graphical model in terms of the percolation probability (existence of an open path) on the same network. Other notable applications include distributed estimation [135, 28] and communication complexity [76].

In Chapter 6, we will apply SDPI and bounds on the input-restricted KL contraction coefficient to broadcasting on trees and stochastic block models.

4.2 Non-linear p-log-Sobolev inequalities for the Potts semigroup

In this section, we prove p-NLSIs for the Potts semigroup for $p \geq 1$. Because the form of the p-LSIs are slightly different for $p \neq 1$ and $p = 1$, we prove them separately.

Recall our setting. The alphabet is $\mathcal{X} = [q]$ for some positive integer $q \geq 2$. The Potts semigroup $T_t = \exp(Lt)$ for generator

$$L(x, y) = \begin{cases} -1 & \text{if } x = y, \\ \frac{1}{q-1}, & \text{if } x \neq y. \end{cases} \tag{4.27}$$

The stationary distribution is $\pi = \text{Unif}([q])$. The Dirichlet form is

$$\mathcal{E}(f, g) = -\mathbb{E}_\pi[(Lf)g] = -\frac{1}{q(q-1)} \left(\sum_x f(x) \right) \left(\sum_y g(y) \right) + \frac{1}{q-1} \sum_x f(x)g(x). \tag{4.28}$$

Relative entropy is

$$\text{Ent}_\pi(f) = \mathbb{E}_\pi \left[f \log \frac{f}{\mathbb{E}_\pi[f]} \right]. \tag{4.29}$$

The non-linear p-log-Sobolev inequality says

$$\frac{\text{Ent}_\pi(f)}{\mathbb{E}_\pi[f]} \leq \Phi_p \left(\frac{\mathcal{E}\left(f^{\frac{1}{p}}, f^{1-\frac{1}{p}} \right)}{\mathbb{E}_\pi[f]} \right) \tag{4.30}$$

for some concave Φ_p, where for $p = 1$, RHS is replaced with $\Phi_p \left(\frac{\mathcal{E}(f, \log f)}{\mathbb{E}_\pi[f]} \right)$. Because both sides of the inequality are fixed under scalar multiplication, we can wlog restrict f to be a distribution μ. Then the relative entropy is

$$\text{Ent}_\pi(\mu) = \frac{1}{q} D(\mu || \pi) = \frac{1}{q} (\log q - H(\mu)). \tag{4.31}$$

4.2.1 Non-linear p-log-Sobolev inequality for $p > 1$

We prove Theorem 4.1 for $p > 1$. Before proving the theorem we show the following.

Proposition 4.4. *Fix $r \in (0,1)$ and $c \in [0, \log q]$. Among all distributions $\mu = (p_1, \dots, p_q)$ with $H(\mu) = c$, the distribution of form $\mu = \left(x, \frac{1-x}{q-1}, \dots, \frac{1-x}{q-1}\right)$ with $x \in \left[\frac{1}{q}, 1\right]$ achieves maximum $\sum_i p_i^r$. Furthermore, up to permutation of the alphabet this is the unique maximum-achieving distribution.*

Proof. The result for $c \in \{0, \log q\}$ is obvious. In the following, assume that $c \in (0, \log q)$. Write $F(\mu) := \sum_i p_i^r$. The set $\{\mu : H(\mu) = c\}$ is compact, so the maximum value of $F(\mu)$ is achieved at some point $\mu = (p_1, \dots, p_q)$.

We prove in several steps. In Step 0, we prove that if $p_i = 0$ for some i, then there can be at most two different values of p_i's. In Step 1, we prove that if $p_i > 0$ for all i, then there can be at most two different values of p_i's. In Step 2, we prove that one of the two different values must have multiplicity one, thus finishing the proof of the proposition.

Step 0.

Claim 4.5. *Fix $a, b > 0$ and $r \in (0,1)$. Among all solutions $u, v, w \in [0,1]$ with $u + v + w = a$ and $-u \log u - v \log v - w \log w = b$, the maximum of $u^r + v^r + w^r$ is not achieved at a point where $0 = u < v < w$.*

Proof. Suppose the maximum is achieved at such a point (u_0, v_0, w_0) where $0 = u_0 < v_0 < w_0$. Extend it to a curve $(u, v = v(u), w = w(u))$ on $u \in [0, \epsilon)$ for some $\epsilon > 0$, such that $u < v < w$ for all u, satisfying

$$u + v + w = a, \tag{4.32}$$

$$-u \log u - v \log v - w \log w = b, \tag{4.33}$$

and

$$v(0) = v_0, w(0) = w_0. \tag{4.34}$$

We prove that

$$f(u) := u^r + v^r + w^r \tag{4.35}$$

decreases as u approaches 0^+, for small enough u.

By taking derivative of (4.32) and (4.33), one can compute that

$$v'(u) = \frac{\log w - \log u}{\log v - \log w}, \quad w'(u) = \frac{\log u - \log v}{\log v - \log w}. \tag{4.36}$$

Therefore

$$\begin{aligned} f'(u) &= r\left(u^{r-1} + v^{r-1}v'(u) + w^{r-1}w'(u)\right) \\ &= r\left(u^{r-1} + v^{r-1}\frac{\log w - \log u}{\log v - \log w} + w^{r-1}\frac{\log u - \log v}{\log v - \log w}\right). \end{aligned} \tag{4.37}$$

Because $0 < v_0 < w_0$, the term u^{r-1} dominates the sum, and $f'(u) > 0$ for small enough $u > 0$. Therefore the maximum of f is not achieved at $u = 0$. □

By Claim 4.5, if $p_i = 0$ for some i, then there can be at most two different values of p_i's.

Step 1.

Claim 4.6. *If $u, v, w \in (0,1)$ are all different, then*

$$\det \begin{pmatrix} 1 & \log u & u^{r-1} \\ 1 & \log v & v^{r-1} \\ 1 & \log w & w^{r-1} \end{pmatrix} \neq 0. \tag{4.38}$$

Proof of Claim. Suppose $\det = 0$. Then for some $a, b \in \mathbb{R}$, the equation $x^{r-1} + a\log x = b$ has at least three distinct solutions $x \in (0,1)$. However

$$\frac{\partial}{\partial x}\left(x^{r-1} + a\log x\right) = (r-1)x^{r-2} + \frac{a}{x} \tag{4.39}$$

is smooth on $(0,1)$, and takes zero at most once. So $x^{r-1} + a\log x$ takes each value at most once on $(0,1)$. Contradiction. □

By Lagrange multipliers, the three vectors

$$\nabla F(\mu) = \left(rp_i^{r-1}\right)_{i\in[q]}, \tag{4.40}$$

$$\nabla H(\mu) = (-1 - \log p_i)_{i\in[q]}, \tag{4.41}$$

$$\nabla \sum_{i\in[q]} p_i = \mathbb{1} \tag{4.42}$$

should be linear dependent. By Step 0 and Claim 4.6, there can be at most two different values of p_i's.

So we can assume that $p_1 = \cdots = p_m = x$, $p_{m+1} = \cdots = p_q = \frac{1-mx}{q-m}$ for some $m \in [q-1]$, $x \in \left(\frac{1}{q}, \frac{1}{m}\right]$.

Step 2. For μ of the above form, we have

$$-H(\mu) = mx\log x + (1-mx)\log\frac{1-mx}{q-m}, \tag{4.43}$$

$$F(\mu) = mx^r + (q-m)\left(\frac{1-mx}{q-m}\right)^r. \tag{4.44}$$

We smoothly continue both functions so that m can take any real value in $[1, q-1]$.

Claim 4.7. *For $m \in (1, q-1]$ and $x \in \left(\frac{1}{q}, \frac{1}{m}\right)$, we have*

$$-\frac{\partial}{\partial x} H(\mu) > 0 \tag{4.45}$$

and

$$\frac{\partial}{\partial x} H(\mu) \frac{\partial}{\partial m} F(\mu) - \frac{\partial}{\partial m} H(\mu) \frac{\partial}{\partial x} F(\mu) > 0. \tag{4.46}$$

Proof. We have

$$-\frac{\partial}{\partial x} H(\mu) = m \left(\log x - \log \frac{1-mx}{q-m} \right) > 0, \tag{4.47}$$

$$-\frac{\partial}{\partial m} H(\mu) = \frac{1-qx}{q-m} + x \left(\log x - \log \frac{1-mx}{q-m} \right), \tag{4.48}$$

$$\frac{\partial}{\partial x} F(\mu) = rm \left(x^{r-1} - \left(\frac{1-mx}{q-m} \right)^{r-1} \right), \tag{4.49}$$

$$\frac{\partial}{\partial m} F(\mu) = x^r + \left(r \frac{1-qx}{1-mx} - 1 \right) \left(\frac{1-mx}{q-m} \right)^r. \tag{4.50}$$

Let $a = \frac{qx-1}{1-mx}$. Then

$$\begin{aligned} G(\mu) &:= \frac{\partial}{\partial x} H(\mu) \frac{\partial}{\partial m} F(\mu) - \frac{\partial}{\partial m} H(\mu) \frac{\partial}{\partial x} F(\mu) \\ &= (r-1)m \left(x^r - \left(\frac{1-mx}{q-m} \right)^r \right) \left(\log x - \log \frac{1-mx}{q-m} \right) \\ &\quad - rm \frac{1-qx}{q-m} \left(x^{r-1} - \left(\frac{1-mx}{q-m} \right)^{r-1} \right) \\ &= x^{-r} m \left((r-1) \left(1 - (a+1)^{-r} \right) \log(a+1) - r \frac{a}{a+1} \left(1 - (a+1)^{1-r} \right) \right). \end{aligned} \tag{4.51}$$

The result then follows from Claim 4.8. □

Claim 4.8. *For all $r \in (0,1)$ and $a > 0$ we have*

$$(r-1) \left(1 - (a+1)^{-r} \right) \log(a+1) - r \frac{a}{a+1} \left(1 - (a+1)^{1-r} \right) > 0. \tag{4.52}$$

Proof. Let

$$f(a) := (r-1) \left(1 - (a+1)^{-r} \right) \log(a+1) - r \frac{a}{a+1} \left(1 - (a+1)^{1-r} \right). \tag{4.53}$$

Because $\lim_{a\to 0^+} f(a) = 0$, it suffices to prove that $f'(a) > 0$.

$$\begin{aligned} f'(a) &= (a+1)^{-r-1}(1-r-(a(1-r)+1)\left((a+1)^{r-1}-r\right)-(1-r)r\log(a+1)) \\ &=: (a+1)^{-r-1}g(a). \end{aligned} \tag{4.54}$$

Because $\lim_{a\to 0^+} g(a) = 0$, it suffices to prove that $g'(a) > 0$.

$$g'(a) = \frac{ar(1-r)\left(1-(a+1)^{r-1}\right)}{a+1} > 0. \tag{4.55}$$

□

Now let us return to the proof of Prop. 4.4. The set of (m, x) where $m \in [1, q-1]$, $x \in \left(\frac{1}{q}, \frac{1}{m}\right]$, and $H(\mu) = c$ can be parametrized as a curve $(m, x = x(m))$ for $m \in [1, m_c]$ for some constant m_c. Along the curve, $F(\mu)$ is continuous, and by Claim 4.7, is decreasing in m. Therefore $F(\mu)$ is maximized at $m = 1$. This finishes the proof. □

Proof of Theorem 4.1 for $p > 1$. For a distribution $\mu = (p_1, \ldots, p_q)$, we have

$$\mathcal{E}\left(\mu^{\frac{1}{p}}, \mu^{1-\frac{1}{p}}\right) = \frac{1}{q-1}\left(1 - \frac{1}{q}\left(\sum_i p_i^{\frac{1}{p}}\right)\left(\sum_i p_i^{1-\frac{1}{p}}\right)\right). \tag{4.56}$$

By Prop. 4.4, for fixed value of $\mathrm{Ent}_\pi(\mu)$, the unique distribution of the form $\left(x, \frac{1-x}{q-1}, \ldots, \frac{1-x}{q-1}\right)$ with $x \in \left[\frac{1}{q}, 1\right]$ minimizes $\mathcal{E}\left(\mu^{\frac{1}{p}}, \mu^{1-\frac{1}{p}}\right)$. Therefore for any non-zero non-negative f, we have

$$b_p\left(\frac{\mathrm{Ent}_\pi(f)}{\mathbb{E}_\pi[f]}\right) \le \frac{\mathcal{E}\left(f^{\frac{1}{p}}, f^{1-\frac{1}{p}}\right)}{\mathbb{E}_\pi[f]}. \tag{4.57}$$

So p-NLSI holds with $\Phi_p = b_p^{-1}$. The statement about optimality is immediate from the above discussions. □

4.2.2 Non-linear 1-log-Sobolev inequality

We prove Theorem 4.1 for $p = 1$. Before proving the theorem we show the following.

Proposition 4.9. *Fix $0 \le c \le \log q$. Among all distributions μ with $H(\mu) = c$, the distribution of form $\mu = \left(x, \frac{1-x}{q-1}, \ldots, \frac{1-x}{q-1}\right)$ with $x \in \left[\frac{1}{q}, 1\right]$ achieves maximum $\sum_i \log p_i$. Furthermore, up to permutation of the alphabet this is the unique minimum-achieving distribution.*

Proof. The result for $c \in \{0, \log q\}$ is obvious. In the following, assume that $0 < c < \log q$. Write $F(\mu) := \sum_i \log p_i$. The set $\{\mu : H(\mu) = c\}$ is compact, so the maximum value of $F(\mu)$ is achieved at some point $\mu = (p_1, \ldots, p_q)$.

We prove in several steps. In Step 0, we prove that $p_i > 0$ for all i. In Step 1, we prove that there can be at most two different values of p_i's. In Step 2, we prove that one of the two different values must have multiplicity one, thus finishing the proof of the proposition.

Step 0. If $p_i = 0$ for some i, then $F(\mu) = -\infty$. So $\min_{i\in[q]} p_i > 0$.

Step 1.

Claim 4.10. *If $u, v, w \in (0,1)$ are all different, then*

$$\det \begin{pmatrix} 1 & \log u & \frac{1}{u} \\ 1 & \log v & \frac{1}{v} \\ 1 & \log w & \frac{1}{w} \end{pmatrix} \neq 0. \tag{4.58}$$

Proof of Claim. Suppose $\det = 0$. Then for some $a, b \in \mathbb{R}$, the equation $\frac{1}{x} + a \log x = b$ has at least three distinct solutions $x \in (0,1)$. However, $\frac{\partial}{\partial x}\left(\frac{1}{x} + a \log x\right) = -\frac{1}{x^2} + \frac{a}{x}$ is smooth on $(0,1)$, and takes zero at most once. So $\frac{1}{x} + a \log x$ takes each value at most once on $(0,1)$. Contradiction. □

By Lagrange multipliers, the three vectors

$$\nabla F(\mu) = \left(\frac{1}{p_i}\right)_{i\in[q]}, \tag{4.59}$$

$$\nabla H(\mu) = (-1 - \log p_i)_{i\in[q]}, \tag{4.60}$$

$$\nabla \sum_{i\in[q]} p_i = \mathbb{1} \tag{4.61}$$

should be linear dependent. By Claim 4.10, there can be at most two different values of p_i's.

So we can assume that $p_1 = \cdots = p_m = x$, $p_{m+1} = \cdots = p_q = \frac{1-mx}{q-m}$ for some $m \in [q-1]$, $x \in \left(\frac{1}{q}, \frac{1}{m}\right)$.

Step 2. For μ of the above form, we have

$$-H(\mu) = mx \log x + (1 - mx) \log \frac{1 - mx}{q - m}, \tag{4.62}$$

$$F(\mu) = m \log x + (q - m) \log \frac{1 - mx}{q - m}. \tag{4.63}$$

We smoothly continue both functions so that m can take any real value in $[1, q-1]$.

Claim 4.11. *For $m \in (1, q-1]$ and $x \in \left(\frac{1}{q}, \frac{1}{m}\right)$, we have*

$$-\frac{\partial}{\partial x} H(\mu) > 0 \tag{4.64}$$

and

$$\frac{\partial}{\partial x}H(\mu)\frac{\partial}{\partial m}F(\mu) - \frac{\partial}{\partial m}H(\mu)\frac{\partial}{\partial x}F(\mu) > 0. \tag{4.65}$$

Proof of Claim. Let $f(x) = \log x - \log \frac{1-mx}{q-m}$. Then we have

$$-\frac{\partial}{\partial x}H(\mu) = mf(x) > 0, \tag{4.66}$$

$$-\frac{\partial}{\partial m}H(\mu) = \frac{1-qx}{q-m} + xf(x), \tag{4.67}$$

$$\frac{\partial}{\partial x}F(\mu) = \frac{m(1-qx)}{x(1-mx)} \tag{4.68}$$

$$\frac{\partial}{\partial m}F(\mu) = \frac{1-qx}{1-mx} + f(x). \tag{4.69}$$

So

$$\begin{aligned} G(\mu) &:= \frac{\partial}{\partial x}H(\mu)\frac{\partial}{\partial m}F(\mu) - \frac{\partial}{\partial m}H(\mu)\frac{\partial}{\partial x}F(\mu) \\ &= m\left(\frac{(1-qx)^2}{x(q-m)(1-mx)} - f(x)^2\right). \end{aligned} \tag{4.70}$$

Let $a = \frac{qx-1}{1-mx}$. Then $G(\mu) = m\left(\frac{a^2}{1+a} - \log^2(a+1)\right)$. Because $a > 0$, we have $G(\mu) > 0$ by Lemma 4.12. □

The set of (m, x) where $m \in [1, q-1]$, $x \in \left(\frac{1}{q}, \frac{1}{m}\right]$, and $H(\mu) = c$ can be parametrized as a curve $(m, x = x(m))$ for $m \in [1, m_c]$ for some constant m_c. Along the curve, $F(\mu)$ is continuous, and by Claim 4.11, is decreasing in m. Therefore $F(\mu)$ is maximized at $m = 1$. This finishes the proof. □

Lemma 4.12. *For $a \in \mathbb{R}_{>-1}$, we have $\frac{a^2}{1+a} \geq \log^2(a+1)$. Equality holds only when $a = 0$.*

Proof. We start from the well-known fact that $a \geq \log(a+1)$ for $a \in \mathbb{R}_{>-1}$ (and equality holds only when $a = 0$). Let $f(a) = a(a+2) - (2a+2)\log(a+1)$. We have $f(0) = 0$ and $f'(a) = 2(a - \log(a+1)) \geq 0$ for $a \in \mathbb{R}_{>-1}$ (and equality holds only when $a = 0$). So f is negative on $(-1, 1)$ and positive on $(1, \infty)$.

Let $g(a) = \frac{a^2}{1+a} - \log^2(a+1)$. Clearly $g(0) = 0$. Because $g'(a) = \frac{f(a)}{(a+1)^2}$, g is decreasing on $(-1, 1]$ and increasing on $[1, \infty)$. So $g(a) \geq 0$ for all $a \in \mathbb{R}_{>-1}$, and equality holds only when $a = 0$. □

Proof of Theorem 4.1 for $p = 1$. For a distribution $\mu = (p_1, \ldots, p_q)$, we have

$$\mathcal{E}(\mu, \log \mu) = \frac{1}{q-1}\sum_{i\in[q]} p_i \log p_i - \frac{1}{q(q-1)}\sum_{i\in[q]} \log p_i. \tag{4.71}$$

By Prop. 4.9, for fixed value of $\mathrm{Ent}_\pi(\mu)$, the unique distribution of the form $\left(x, \frac{1-x}{q-1}, \ldots, \frac{1-x}{q-1}\right)$ with $x \in \left[\frac{1}{q}, 1\right]$ minimizes $\mathcal{E}(\mu, \log \mu)$. Therefore for any non-zero non-negative f, we have

$$b_1\left(\frac{\mathrm{Ent}_\pi(f)}{\mathbb{E}_\pi[f]}\right) \leq \frac{\mathcal{E}(f, \log f)}{\mathbb{E}_\pi[f]}. \tag{4.72}$$

So 1-NLSI holds for $\Phi_1 = b_1^{-1}$. The statement about optimality is immediate from the above discussions. □

4.2.3 Input-restricted non-linear SDPI for Potts channels

In this section, we prove Theorem 4.3.

The subset of Potts channels corresponding to $\lambda \geq 0$ (i.e. ferromagnetic Potts channels) form a semigroup. For the semigroups, the optimal 1-NLSI is an "infinitesimal version" of the input-restricted non-linear SDPI. Consequently, by integrating the former we can get the latter (this is formalized in the first part of the proof below). Surprisingly, the result also extends beyond the semigroup to all of the Potts channels, namely we have the following.

Proposition 4.13. *Let $\lambda \in \left[-\frac{1}{q-1}, 1\right]$. Fix $0 \leq c \leq \log q$. Among all distributions μ with $H(\mu) = c$, the distribution of form $\mu = \left(x, \frac{1-x}{q-1}, \ldots, \frac{1-x}{q-1}\right)$ with $x \in \left[\frac{1}{q}, 1\right]$ achieves minimum $H(\mu P_\lambda)$. Furthermore, when $\lambda \notin \{0, 1\}$, up to permutation of the alphabet this is the unique minimum-achieving distribution.*

Proof. The result for $\lambda \in \{0, 1\}$ is obvious. In the following assume that $\lambda \notin \{0, 1\}$. The result for $c \in \{0, \log q\}$ is obvious. In the following assume that $c \notin \{0, \log q\}$. The set $\{\mu : H(\mu) = c\}$ is compact, so the minimum value of $H(\mu P_\lambda)$ is achieved at some point $\mu = (p_1, \ldots, p_q)$.

We prove in several steps. In Step 0, we prove that if $p_i = 0$ for some i, then there can be at most two different values of p_i's. In Step 1, we prove that if $p_i > 0$ for all i, then there can be at most two different values of p_i's. In Step 2, we prove that one of the two different values must have multiplicity one, thus finishing the proof of the proposition.

Step 0.

Claim 4.14. *Fix $a, b, d > 0$ and $c \in \mathbb{R}_{>-d} \backslash \{0\}$. Among all solutions $u, v, w \in [0, 1]$ with $u + v + w = a$ and $-u \log u - v \log v - w \log w = b$, the maximum of*

$$(cu + d)\log(cu + d) + (cv + d)\log(cv + d) + (cw + d)\log(cw + d) \tag{4.73}$$

is not achieved at a point where $0 = u < v < w$.

Proof. Suppose the maximum is achieved at such a point (u_0, v_0, w_0) where $0 = u_0 < v_0 < w_0$. Extend it to a curve $(u, v = v(u), w = w(u))$ on $u \in [0, \epsilon)$ for some $\epsilon > 0$,

such that $u < v < w$ for all u, satisfying

$$u + v + w = a, \tag{4.74}$$
$$-u \log u - v \log v - w \log w = b, \tag{4.75}$$

and $v(0) = v_0$, $w(0) = w_0$.

We prove that

$$f(u) := (cu + d) \log(cu + d) + (cv + d) \log(cv + d) + (cw + d) \log(cw + d) \tag{4.76}$$

decreases as u approaches 0^+ for small enough u.

By taking derivative of (4.74) and (4.75), one can compute that

$$v'(u) = \frac{\log w - \log u}{\log v - \log w}, \quad w'(u) = \frac{\log u - \log v}{\log v - \log w}. \tag{4.77}$$

Therefore

$$\begin{aligned} f'(u) &= c \left(\log(cu + d) + \log(cv + d) v'(u) + \log(cw + d) w'(u) \right) \\ &= c \left(\log(cu + d) + \log(cv + d) \frac{\log w - \log u}{\log v - \log w} + \log(cw + d) \frac{\log u - \log v}{\log v - \log w} \right). \end{aligned} \tag{4.78}$$

Because $0 < v_0 < w_0$, terms involving $\log u$ dominates the sum. The dominating term is

$$-c \log u \frac{\log(cv + d) - \log(cw + d)}{\log v - \log w} > 0. \tag{4.79}$$

Therefore the maximum of f is not achieved at $u = 0$. □

By Claim 4.14, if $p_i = 0$ for some i, then there can be at most two different values of p_i's.

Step 1.

Claim 4.15. *If $u, v, w \in (0, 1)$ are all different, then*

$$\det \begin{pmatrix} 1 & \log u & \log(\lambda u + \frac{1-\lambda}{q}) \\ 1 & \log v & \log(\lambda v + \frac{1-\lambda}{q}) \\ 1 & \log w & \log(\lambda w + \frac{1-\lambda}{q}) \end{pmatrix} \neq 0. \tag{4.80}$$

Proof of Claim. Suppose $\det = 0$. Then for some $a, b \in \mathbb{R}$, the equation

$$\log \left(\lambda x + \frac{1 - \lambda}{q} \right) + a \log x = b \tag{4.81}$$

has at least three distinct solutions $x \in (0,1)$. However,

$$\frac{\partial}{\partial x}\left(\log\left(\lambda x + \frac{1-\lambda}{q}\right) + a\log x\right) = \frac{\lambda}{\lambda x + \frac{1-\lambda}{q}} + \frac{a}{x} \tag{4.82}$$

is smooth on $(0,1)$, and takes zero at most once. So

$$\log\left(\lambda x + \frac{1-\lambda}{q}\right) + a\log x \tag{4.83}$$

takes each value at most twice on $(0,1)$. Contradiction. □

By Lagrange multipliers, the three vectors

$$\nabla H(\mu P_\lambda) = \left(-\lambda\log\left(\lambda p_i + \frac{1-\lambda}{q}\right) - \lambda\right)_{i\in[q]}, \tag{4.84}$$

$$\nabla H(\mu) = (-1 - \log p_i)_{i\in[q]}, \tag{4.85}$$

$$\nabla \sum_{i\in[q]} p_i = \mathbb{1} \tag{4.86}$$

should be linear dependent. By Claim 4.15, there can be at most two different values of p_i's.

So we can assume that $p_1 = \cdots = p_m = x$, $p_{m+1} = \cdots = p_q = \frac{1-mx}{q-m}$ for some $m \in [q-1]$, $x \in \left(\frac{1}{q}, \frac{1}{m}\right]$.

Step 2. For μ of the above form, we have

$$-H(\mu) = mx\log x + (1-mx)\log\frac{1-mx}{q-m}, \tag{4.87}$$

$$\begin{aligned} -H(\mu P_\lambda) = {} & m\left(\lambda x + \frac{1-\lambda}{q}\right)\log\left(\lambda x + \frac{1-\lambda}{q}\right) \\ & + (q-m)\left(\lambda\frac{1-mx}{q-m} + \frac{1-\lambda}{q}\right)\log\left(\lambda\frac{1-mx}{q-m} + \frac{1-\lambda}{q}\right). \end{aligned} \tag{4.88}$$

We smoothly continue both functions so that m can take any real value in $[1, q-1]$.

Claim 4.16. *For $m \in (1, q-1]$ and $x \in \left(\frac{1}{q}, \frac{1}{m}\right)$, we have*

$$-\frac{\partial}{\partial x}H(\mu) > 0 \tag{4.89}$$

and

$$\frac{\partial}{\partial m}H(\mu)\frac{\partial}{\partial x}H(\mu P_\lambda) - \frac{\partial}{\partial x}H(\mu)\frac{\partial}{\partial m}H(\mu P_\lambda) > 0. \tag{4.90}$$

Proof of Claim. Let

$$f(x) = \log x - \log \frac{1-mx}{q-m}. \tag{4.91}$$

Then

$$-\frac{\partial}{\partial x} H(\mu) = mf(x) > 0, \tag{4.92}$$

$$-\frac{\partial}{\partial m} H(\mu) = \frac{1-qx}{q-m} + xf(x), \tag{4.93}$$

$$-\frac{\partial}{\partial x} H(\mu P_\lambda) = \lambda m f\left(\lambda x + \frac{1-\lambda}{q}\right), \tag{4.94}$$

$$-\frac{\partial}{\partial m} H(\mu P_\lambda) = \lambda \frac{1-qx}{q-m} + \left(\lambda x + \frac{1-\lambda}{q}\right) f\left(\lambda x + \frac{1-\lambda}{q}\right), \tag{4.95}$$

and

$$\begin{aligned} G(\mu) :=& \frac{\partial}{\partial m} H(\mu) \frac{\partial}{\partial x} H(\mu P_\lambda) - \frac{\partial}{\partial x} H(\mu) \frac{\partial}{\partial m} H(\mu P_\lambda) \\ =& \lambda m \frac{1-qx}{q-m} \left(f\left(\lambda x + \frac{1-\lambda}{q}\right) - f(x) \right) - mf(x) f\left(\lambda x + \frac{1-\lambda}{q}\right) \frac{1-\lambda}{q}. \end{aligned} \tag{4.96}$$

$$\frac{\partial}{\partial \lambda} \frac{G(\mu)}{m\lambda f(x) f\left(\lambda x + \frac{1-\lambda}{q}\right)} = \frac{1}{q\lambda^2} + \frac{1-qx}{q-m} \frac{\frac{\partial}{\partial \lambda} f\left(\lambda x + \frac{1-\lambda}{q}\right)}{f\left(\lambda x + \frac{1-\lambda}{q}\right)^2}. \tag{4.97}$$

Note that

1. $$\frac{G(\mu)}{m\lambda f(x) f\left(\lambda x + \frac{1-\lambda}{q}\right)} \tag{4.98}$$

   is continuous for $\lambda \in \left[-\frac{1}{q-1}, 1\right]$, and takes value 0 at $\lambda = 1$;

2. $m\lambda f(x) f\left(\lambda x + \frac{1-\lambda}{q}\right) \geq 0$ for $\lambda \in \left[-\frac{1}{q-1}, 1\right]$.

So we only need to prove that

$$\frac{\partial}{\partial \lambda} \frac{G(\mu)}{m\lambda f(x) f\left(\lambda x + \frac{1-\lambda}{q}\right)} \leq 0, \tag{4.99}$$

i.e.,

$$f\left(\lambda x+\frac{1-\lambda}{q}\right)^2 \le \frac{q\lambda^2(qx-1)}{q-m}\frac{\partial}{\partial\lambda}f\left(\lambda x+\frac{1-\lambda}{q}\right). \tag{4.100}$$

Let $y=\lambda x+\frac{1-\lambda}{q}$. Then the above inequality can be rewritten as

$$f(y)^2 \le \frac{q\lambda^2(qx-1)}{q-m}\frac{\partial y}{\partial\lambda}\frac{\partial f(y)}{\partial y} = \frac{(qy-1)^2}{(q-m)y(1-my)}. \tag{4.101}$$

This is true by Lemma 4.12, applied to $a=\frac{qy-1}{1-my}$. Equality holds only when $y=\frac{1}{q}$, which cannot happen for $\lambda\neq 0$. □

The set of (m,x) where $m\in[1,q-1]$, $x\in\left(\frac{1}{q},\frac{1}{m}\right]$, and $H(\mu)=c$ can be parametrized as a curve $(m,x=x(m))$ for $m\in[1,m_c]$ for some constant m_c. Along the curve, $H(\mu P_\lambda)$ is continuous, and by Claim 4.16, is increasing in m. Therefore $H(\mu P_\lambda)$ is minimized at $m=1$. This finishes the proof. □

Proof of Theorem 4.3. Consider a Markov chain $U\to X\to Y$ where X has uniform distribution, and the channel $X\to Y$ is P_λ. Because P_X and P_Y are both uniform, for any u, we have

$$D(P_{X|U=u}||P_X)=\log q-H(P_{X|U=u}), \tag{4.102}$$
$$D(P_{Y|U=u}||P_Y)=\log q-H(P_{Y|U=u}). \tag{4.103}$$

So by Prop. 4.13 we get

$$D(P_{Y|U=u}||P_Y)\le s_\lambda(D(P_{X|U=u}||P_X)). \tag{4.104}$$

Therefore

$$I(U;Y)=D(P_{Y|U}||P_Y|P_U)\le \widehat{s}_\lambda(D(P_{X|U}||P_X|P_U))=\widehat{s}_\lambda(I(U;X)). \tag{4.105}$$

Now we prove optimality. Let $c\in[0,\log q]$. Choose $a,b\in[0,\log q]$ and $u\in[0,1]$ such that $c=(1-u)a+ub$ and $\widehat{s}_\lambda(c)=(1-u)s_\lambda(a)+us_\lambda(b)$. Choose $\rho,\tau\in[0,1]$ such that $C(P_\rho)=a$ and $C(P_\tau)=b$. Define random variable $U=(V,Z)$ such that $Z\sim\mathrm{Ber}(u)$, and conditioned on $Z=0$, $V\sim P_\rho(X)$, and conditioned on $Z=1$, $V\sim P_\tau(X)$. One can check that

$$I(U;X)=(1-u)a+ub=c, \tag{4.106}$$
$$I(U;Y)=(1-u)s_\lambda(a)+us_\lambda(b)=\widehat{s}_\lambda(c). \tag{4.107}$$

□

Let us discuss the relationship between Potts semigroup and ferromagnetic Potts channels. As discussed in the Introduction, ferromagnetic Potts channels are exactly the operators in the Potts semigroup, with $T_t=P_{\exp(-\frac{q}{q-1}t)}$. Therefore 1-LSI for the

Potts semigroup can be seen as infinitesimal SDPI for ferromagnetic Potts channels, and many results for the former can directly transfer to results for the latter.

We use Prop. 4.9 to give an alternative proof for Prop. 4.13 for ferromagnetic Potts channels.

Alternative proof of Prop. 4.13 for ferromagnetic Potts channels. Let μ and ν be two distributions with $H(\mu) = H(\nu) = c$, where μ is of form $\left(x, \frac{1-x}{q-1}, \ldots, \frac{1-x}{q-1}\right)$ for some $x \in \left[\frac{1}{q}, 1\right]$, and ν is not of this form (up to permuting the alphabet). Define $\mu_t = \mu T_t$ and $\nu_t = \nu T_t$, where $(T_t)_{t\geq 0}$ is the Potts semigroup.

We prove that $H(\mu_t) < H(\nu_t)$ for $t \in (0, \infty)$. Suppose this does not hold. Let $u = \inf\{t > 0 : H(\mu_t) \geq H(\nu_t)\}$. Then we have $H(\nu_u) = H(\mu_u)$ by continuity of semigroup. By Prop. 4.9, we have

$$\frac{\partial}{\partial t}|_{t=u} H(\nu_t) = \mathcal{E}(\nu_u, \log \nu_u) > \mathcal{E}(\mu_u, \log \mu_u) = \frac{\partial}{\partial t}|_{t=u} H(\mu_t). \tag{4.108}$$

If $u = 0$, then for some $\epsilon > 0$, $H(\nu_t) > H(\mu_t)$ for $t \in (0, \epsilon)$. If $u > 0$, then for some $\epsilon > 0$, $H(\nu_t) < H(\mu_t)$ for $t \in (u - \epsilon, u)$. Both cases lead to contradiction with definition of u. So $H(\mu_t) < H(\nu_t)$ for $t \in (0, \infty)$. This completes the proof of the result for $\lambda > 0$. □

4.2.4 Behavior for $q \to \infty$

When should one use p-NLSI instead of p-LSI? To get some insights, we consider the case of $q \to \infty$. First, we prove Prop. 4.2 that $\alpha_1 = 1 + \frac{1+o(1)}{\log q}$.

Proof of Prop. 4.2. **Lower bound.** By Theorem 4.1, we need to show that for all $x \in \left(\frac{1}{q}, 1\right]$, we have

$$\frac{q}{q-1}\left(1 + \frac{1}{\log q}\right) \leq \frac{\xi_1(x)}{\psi(x)}. \tag{4.109}$$

Noting that

$$\xi_1(x) = \frac{q}{q-1}\left(\frac{1}{q}\left(-\log x - (q-1)\log\frac{1-x}{q-1}\right) - \log q + \psi(x)\right), \tag{4.110}$$

it suffices to prove that

$$f(x) := \frac{\log q}{q}\left(-\log x - (q-1)\log\frac{1-x}{q-1}\right) - \log^2 q - \psi(x) \geq 0. \tag{4.111}$$

We have $f\left(\frac{1}{q}\right) = 0$. So it suffices to prove that $f'(x) \geq 0$ for $x \in \left[\frac{1}{q}, 1\right]$.

$$f'(x) = \frac{\log q}{q}\left(-\frac{1}{x} + \frac{q-1}{1-x}\right) - \left(\log x - \log\frac{1-x}{q-1}\right). \tag{4.112}$$

We smoothly continue this function to $\left\{(q,x)\in\mathbb{R}^2 : q\geq 3, x\in\left[\frac{1}{q},1\right]\right\}$ and prove that it is non-negative in this region.

$$\begin{aligned}\frac{\partial}{\partial q}f'(x) &= \frac{1-\log q}{q^2}\left(-\frac{1}{x}+\frac{q-1}{1-x}\right)+\frac{\log q}{q}\frac{1}{1-x}-\frac{1}{q-1}\\ &= \frac{(q-1)\log q+q(qx^2-x-1)+1}{q^2(q-1)x(1-x)}.\end{aligned} \tag{4.113}$$

The numerator is a quadratic function in x, and for fixed q, it is minimized at $x=\frac{1}{q}$, leading to

$$\frac{\partial}{\partial q}f'(x)\geq\frac{(q-1)\log q-q+1}{q^2(q-1)x(1-x)}=\frac{\log q-1}{q^2x(1-x)}\geq 0. \tag{4.114}$$

So we only need to prove $f'(x)\geq 0$ for minimum q, i.e., $q=\max\{3,\frac{1}{x}\}$. When $q=\frac{1}{x}$, on can verify that $f'(x)=0$. So the only remaining case is $q=3$. For $q=3$, we prove that f is convex in x, i.e., $f''(x)\geq 0$ for $x\in[0,1]$.

$$\begin{aligned}f''(x) &= \frac{\log q}{q}\left(\frac{1}{x^2}+\frac{q-1}{(1-x)^2}\right)-\left(\frac{1}{x}+\frac{1}{1-x}\right)\\ &= \frac{\log q((1-x)^2+(q-1)x^2)-qx(1-x)}{qx^2(1-x)^2}\\ &= \frac{(q\log q+q)x^2-(q+2\log q)x+\log q}{qx^2(1-x)^2}.\end{aligned} \tag{4.115}$$

The numerator is a quadratic function in x, and its discriminant is

$$(q+2\log q)^2-4(q\log q+q)\log q=q^2-4(q-1)\log^2 q. \tag{4.116}$$

When $q=3$, the above value is <0. So $f''(x)\geq 0$ for $q=3$ and $x\in[0,1]$. This finishes the proof of the lower bound.

Upper bound. For the upper bound, we need find $x\in\left(\frac{1}{q},1\right]$ such that $\frac{f(x)}{\psi(x)}=o(1)$. Because the upper bound to prove is asymptotic, we assume that q is large enough. Take $x=\frac{2}{\log q}$. Then we have

$$\begin{aligned}\psi(x) &= \log q+\frac{2}{\log q}\log\frac{2}{\log q}+\left(1-\frac{2}{\log q}\right)\log\frac{1-\frac{2}{\log q}}{q-1}\\ &= \log q-\left(1-\frac{2}{\log q}\right)\log(q-1)+o(1)\\ &= 2+o(1)\end{aligned} \tag{4.117}$$

and

$$\begin{aligned} f(x) &= \frac{\log q}{q}\left(-\log\frac{2}{\log q} - (q-1)\log\frac{1-\frac{2}{\log q}}{q-1}\right) - \log^2 q - \psi(x) \\ &= \frac{\log q}{q}(q-1)\left(\log(q-1) - \log\left(1-\frac{2}{\log q}\right)\right) - \log^2 q - 2 + o(1) \\ &= \log q \cdot \left(1+O\left(\frac{1}{q}\right)\right)\cdot\left(\log q + O\left(\frac{1}{q}\right) + \frac{2}{\log q} + O\left(\frac{1}{\log^2 q}\right)\right) \\ &\quad - \log^2 q - 2 + o(1) \\ &= o(1). \end{aligned}$$

So $\frac{f(x)}{\psi(x)} = o(1)$. □

Numerical computation suggests $\frac{f(x)}{\psi(x)}$ is minimized at a point $x = \frac{2+o(1)}{\log q}$. This guides our proof of the upper bound in Prop. 4.2, but we have not attempted to prove this fact.

To understand the case $p > 1$, let us denote convexification of b_p as $\breve{b}_p$. Then NLSI lower bound, assuming $\mathbb{E}[f] = 1$, gives

$$\mathcal{E}\left(f^{\frac{1}{p}}, f^{1-\frac{1}{p}}\right) \geq \breve{b}_p(\mathrm{Ent}(f)). \tag{4.118}$$

We see that this improves upon $\alpha_p \cdot \mathrm{Ent}(f)$ the more the larger the entropy. In particular, the maximum improvement happens when $\mathrm{Ent}(f) = \log q$. That is we have for $p > 1$

$$\alpha_p \leq \frac{\breve{b}_p(x)}{x} \leq \frac{1}{\log q}. \tag{4.119}$$

Together with (4.9) and (4.12), we get $\alpha_p = \Theta\left(\frac{1}{\log q}\right)$ as $q \to \infty$. Numerical computation suggests that $\alpha_p = \frac{1+o(1)}{\log q}$.

At the same time, the improvement given by the 1-NLSI (over 1-LSI) is much stronger, since $b_1(\log q) = \infty$. To summarize, the p-NLSI should be preferred for $p = 1$ or for cases where q is small and entropy is large (i.e. functions are highly spiky).

Next, we consider SDPIs and η_{KL}. First, we show that for a fixed $\lambda \geq 0$ we have

$$\eta_{\mathrm{KL}}(\pi, P_\lambda) = \lambda - \Theta\left(\frac{1}{\log q}\right). \tag{4.120}$$

Indeed, the upper bound is given by (4.22). For the lower bound we have

$$\begin{aligned}\eta_{\mathrm{KL}}(\pi, P_\lambda) \geq \eta_{\min} &:= \frac{\psi\left(\lambda + \frac{1-\lambda}{q}\right)}{\psi(1)} \\ &= \frac{\log q + \left(\lambda + \frac{1-\lambda}{q}\right)\log\left(\lambda + \frac{1-\lambda}{q}\right) + \left(1 - \left(\lambda + \frac{1-\lambda}{q}\right)\right)\log\frac{1-\left(\lambda+\frac{1-\lambda}{q}\right)}{q-1}}{\log q} \\ &= \lambda + \frac{\lambda\log\lambda + (1-\lambda)\log(1-\lambda) + o(1)}{\log q}. \end{aligned} \tag{4.121}$$

On the other hand,

$$\eta_{\mathrm{KL}}(\pi, P_\lambda) \leq \eta_{\mathrm{KL}}(P_\lambda) \leq \eta_{\mathrm{TV}}(P_\lambda) = \lambda, \tag{4.122}$$

where η_{TV} is the contraction coefficient for the total variation distance.

Notice also that for $\widehat{s}_\lambda$ we have generally $\eta_{\min} \leq \frac{\widehat{s}_\lambda(x)}{x} \leq \eta_{\mathrm{KL}}(P_\lambda, \pi)$. Therefore we have shown that

$$\lim_{q\to\infty} \frac{\widehat{s}_\lambda(x)}{x} = \lim_{q\to\infty} \eta_{\mathrm{KL}}(P_\lambda, \pi) = \lim_{q\to\infty} \eta_{\mathrm{KL}}(P_\lambda) = \eta_{\mathrm{TV}}(P_\lambda) = \lambda. \tag{4.123}$$

The estimates of information quantities using the more sophisticated tools get improvement over simplistic coupling of at most multiplicative order $\left(1 + \Theta\left(\frac{1}{\log q}\right)\right)$.

Note, however, if λ changes with q (e.g. $\lambda = -\frac{1}{q-1}$), then the improvement over η_{TV} can be as large as a multiplicative factor of $(1+o(1))\log q$, as shown in Prop. 4.23.

4.3 Product spaces

In this section we study extensions of p-NLSIs and SDPIs to the product semigroup $(T_t^{\times n})_{t\geq 0}$ on the product space $[q]^n$ (and product channels $P_\lambda^{\times n}$). The general property of tensorization of p-NLSI was established in [114, Theorem 1], and thus we only need to concavify functions Φ_p in (4.7). Similarly, we can show that (non-linear) strong data processing inequalities tensorize if one concavifies function $s(\cdot)$ in (4.21).

After showing these extensions to product spaces, we proceed to discussing implications of p-NLSI on speed of convergence to equillibrium in terms of $\mathrm{Ent}(\nu T_t^{\times n})$ and on edge-isoperimetric inequalities.

4.3.1 Tensorization

Proposition 4.17. *Fix $p \geq 1$. Recall b_p defined in (4.18). Let $\breve{b}_p$ be the convex envelope of b_p. Then p-LSI holds for the product semigroup $(T_t^{\times n})_{t\geq 0}$ with*

$$\Phi_{n,p}(x) = n\breve{b}_p^{-1}\left(\frac{x}{n}\right). \tag{4.124}$$

Proof. By Theorem 4.1 and [114, Theorem 1]. □

As we show below $\check{b}_p \neq b_p$ (Prop. 4.20).

For non-linear SDPI, we first prove a general tensorization result.

Proposition 4.18. *Fix a probability kernel $P_{Y|X} : \mathcal{X} \to \mathcal{Y}$ and a distribution P_X on $\mathcal{X}$.*

1. *Suppose for some non-decreasing function $s : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ we have*

$$D(Q_Y||P_Y) \leq s(D(Q_X||P_X)) \tag{4.125}$$

for all distribution Q_X on $\mathcal{X}$ with $0 < D(Q_X||P_X) < \infty$. Then for all distribution Q_{X^n} on $\mathcal{X}^n$ with $0 < D(Q_{X^n}||P_X^{\times n}) < \infty$, we have

$$D(Q_{Y^n}||P_Y^{\times n}) \leq n\widehat{s}\left(\frac{1}{n}D(Q_{X^n}||P_X^{\times n})\right), \tag{4.126}$$

where $\widehat{s}$ is the concave envelope of s, and $Q_{Y^n} = P_{Y|X}^{\times n} \circ Q_{X^n}$.

2. *Suppose for some non-decreasing concave function $\widehat{s} : [0, \log|\mathcal{X}|] \to \mathbb{R}_{\geq 0}$ we have*

$$I(U;Y) \leq \widehat{s}(I(U;X)) \tag{4.127}$$

for all Markov chains $U \to X \to Y$ where the distribution of X is P_X. Then for all Markov chains $U \to X^n \to Y^n$ where the distribution of X is $P_X^{\times n}$, we have

$$I(U;Y^n) \leq n\widehat{s}\left(\frac{1}{n}I(U;X^n)\right). \tag{4.128}$$

We have separate statements for non-linear SDPI defined via KL divergence and via mutual information, because they are not equivalent in general. It is not hard to show that if KL divergence type non-linear SDPI (Inequality (4.125)) holds for some function s, then mutual information type non-linear SDPI (Inequality (4.127)) holds for $\widehat{s}$. However, it is not clear what is the best possible KL divergence type non-linear SDPI one can get starting from mutual information type non-linear SDPI. (Note the domain of function s would become larger during the translation.)

Proof of Prop. 4.18. **Proof of 1.** Perform induction on n. The base case $n = 1$ is

trivial. Now consider $n \geq 2$. We have

$$\begin{aligned}
&D(Q_{Y^n}||P_Y^{\times n}) \\
&= D\left(Q_{Y^{n-1}}||P_Y^{\times(n-1)}\right) + D(Q_{Y_n|Y^{n-1}}||P_Y|Q_{Y^{n-1}}) \\
&\leq D\left(Q_{Y^{n-1}}||P_Y^{\times(n-1)}\right) + D(Q_{Y_n|X^{n-1}}||P_Y|Q_{X^{n-1}}) \\
&\leq (n-1)\widehat{s}\left(\frac{1}{n-1}D(Q_{X^{n-1}}||P_X^{\times(n-1)})\right) + s\left(D(Q_{X_n|X^{n-1}}||P_X|Q_{X^{n-1}})\right) \\
&\leq n\widehat{s}\left(\frac{1}{n}D(Q_{X^{n-1}}||P_X^{\times(n-1)}) + \frac{1}{n}D(Q_{X_n|X^{n-1}}||P_X|Q_{X^{n-1}})\right) \\
&= n\widehat{s}\left(\frac{1}{n}D(Q_{X^n}||P_X^{\times n})\right).
\end{aligned}$$

First step is by chain rule. Second step is because we have a Markov chain $Y^{n-1} - X^{n-1} - Y_n$, and conditioning on more information does not decrease conditioned divergence. Third step is by induction hypothesis. Fourth step is by concavity. Fifth step is by chain rule.

Proof of 2. Perform induction on n. The base case $n = 1$ is trivial. Now consider $n \geq 2$. We have

$$\begin{aligned}
I(U;Y^n) &= I(U;Y^{n-1}) + I(U;Y_n|Y^{n-1}) \\
&= I(U;Y^{n-1}) + I(U,Y^{n-1};Y_n) \\
&\leq I(U;Y^{n-1}) + I(U,X^{n-1};Y_n) \\
&= I(U;Y^{n-1}) + I(U;Y_n|X^{n-1}) \\
&\leq (n-1)\widehat{s}\left(\frac{1}{n-1}I(U;X^{n-1})\right) + \widehat{s}\left(I(U;X_n|X^{n-1})\right) \\
&\leq n\widehat{s}\left(\frac{1}{n}I(U;X^{n-1}) + \frac{1}{n}I(U;X_n|X^{n-1})\right) \\
&= n\widehat{s}\left(\frac{1}{n}I(U;X^n)\right).
\end{aligned}$$

First step is by chain rule. Second step is by chain rule, and that Y_n is independent with Y^{n-1}. Third step is by data processing inequality. Fourth step is by chain rule, and that Y_n is independent with X^{n-1}. Fifth step is by induction hypothesis. Sixth step is by concavity. Seventh step is by chain rule. □

Corollary 4.19. *Recall function s_λ defined in Theorem 4.3. Let Q_{X^n} be a distribution on $[q]^n$ and $Q_{Y^n} = P_\lambda^{\times n} \circ Q_{X^n}$. Then we have*

$$\frac{1}{n}H(Y^n) \geq \log q - \widehat{s}_\lambda\left(\log q - \frac{1}{n}H(X^n)\right). \tag{4.129}$$

Furthermore, for every $c \in [0, \log q]$, there exist distributions X^n with $H(X^n) = (c + o(1))n$ such that $\frac{1}{n}H(Y^n) = \log q - \widehat{s}_\lambda(\log q - c) + o(1)$.

Proof. Inequality (4.129) follows from Prop. 4.18 and that

$$D(Q_{X^n}||\pi^{\times n}) = n\log q - H(Q_{X^n}). \tag{4.130}$$

For the second part, choose $a, b \in [0, \log q]$ and $u \in [0,1]$ such that $c = (1-u)a + ub$ and $\widehat{s}_\lambda(\log q - c) = (1-u)s_\lambda(\log q - a) + us_\lambda(\log q - b)$. Such a, b, u exist because $\widehat{s}_\lambda$ is the concave envelope of s_λ.

Let Q_A (resp. Q_B) be the unique distribution on $[q]$ of form $\left(x, \frac{1-x}{q-1}, \ldots, \frac{1-x}{q-1}\right)$ with $x \in \left[\frac{1}{q}, 1\right]$ and entropy a (resp. entropy b). Now let Q_{X^n} be the distribution $Q_A \times \cdots \times Q_A \times Q_B \times \cdots \times Q_B$, where Q_A appears $\lfloor (1-u)n \rfloor$ times and Q_B appears $\lceil un \rceil$ times. It is easy to see that this distribution satisfies the required properties. □

4.3.2 Linear piece

In Prop. 4.17 and Theorem 4.3, we make use of convexification of b_p and concavification of s_λ. When $q = 2$, we have $\breve{b}_p = b_p$ and $\widehat{s}_\lambda = s_\lambda$ (the latter fact is known as Mrs. Gerber's Lemma [134]). However, for $q \geq 3$, the situation is vastly different.

Proposition 4.20. *Recall function $b_p : [0, \log q] \to \mathbb{R}$ defined in Theorem 4.1 and $s_\lambda : [0, \log q] \to \mathbb{R}$ defined in Theorem 4.3.*

1. *For all $q \geq 3$ and $p \geq 1$, b_p is not convex near 0.*
2. *For all $q \geq 3$, $\lambda \in \left[-\frac{1}{q-1}, 0\right) \cup (0,1)$, s_λ not concave near 0.*

The proof is deferred to Section 4.7. Prop. 4.20 implies that there is a linear piece near origin in the graph of $\breve{b}_p$, $\widehat{\Phi}_p$ and $\widehat{s}_\lambda$.

This implies a curious new property distinguishing Potts semigroup with $q \geq 3$ from its binary cousin and from the Ornstein-Uhlenbeck semigroup. Both of the latter have their p-NLSI and SDPI strictly non-linear, which translates into the following fact: among all initial densities ν_0 with a given entropy $\mathrm{Ent}(\nu_0)$ a simple product distributions simultaneously maximizes $\mathrm{Ent}(\nu_0 T_t^{\times n})$ for all t. Stated differently we have (this is known as Mrs. Gerber's Lemma) when $q = 2$:

$$D(P_\lambda \circ P_{X^n}||\pi^{\times n}) \leq D(P_\lambda \circ \mathrm{Ber}(p)^{\times n}||\pi^{\times n}), \tag{4.131}$$

where π is the uniform distribution on $[q]^n$, and $\mathrm{Ber}(p)^{\times n}$ is an i.i.d. distribution on $[q]^n$ with $p \in [0, 1/2]$ solving $D(\mathrm{Ber}(p)\|\pi) = \frac{1}{n} D(P_\lambda \circ P_{X^n}||\pi^{\times n})$. That is, the slowest to relax to equillibrium is the product distribution. For the Ornstein-Uhlenbeck a similar statement holds with $\mathrm{Ber}(p)$ replaced by the $\mathcal{N}(0, \sigma^2 I_n)$.

This nice extremal property of product distributions is no longer true for $q \geq 3$ Potts semigroups, because s_λ is not concave, and the value of $\widehat{s}_\lambda$ at a point may be a mixture of two values of s_λ. More precisely, instead of (4.131), we have for every $\lambda \in \left[-\frac{1}{q-1}, 1\right]$ and every $c \in \mathbb{R}_{\geq 0}$, there exist two i.i.d distributions μ, ν on $[q]^n$ and

$t \in [0,1]$ satisfying

$$(1-t)D(\mu||\pi^{\times n}) + tD(\nu||\pi^{\times n}) = c, \tag{4.132}$$

such that for every distribution P_{X^n} on $[q]^n$ with $D(P_{X^n}||\pi^{\times n}) = c$, we have

$$D(P_\lambda \circ P_{X^n}||\pi^{\times n}) \le (1-t)D(\mu P_\lambda||\pi^{\times n}) + tD(\nu P_\lambda||\pi^{\times n}). \tag{4.133}$$

Note here μ, ν and t all depend on c and λ, and thus there is no universal distribution that is the slowest to converge to equillibrium.

Let us discuss some general implications of non-convexity of b_p and non-concavity of s_λ near 0.

Let K be a Markov kernel with stationary distribution π. Consider the tightest possible p-NLSI given by

$$b_p(x) := \inf_{\substack{f:\mathcal{X}\to\mathbb{R}_{\ge 0},\\ \mathbb{E}_\pi f=1, \mathrm{Ent}_\pi(f)=x}} \mathcal{E}\left(f^{\frac{1}{p}}, f^{1-\frac{1}{p}}\right). \tag{4.134}$$

The p-log-Sobolev constant is

$$\alpha_p := \inf_{x>0} \frac{b_p(x)}{x} = \inf_{f:\mathcal{X}\to\mathbb{R}_{\ge 0}, \mathrm{Ent}_\pi(f)>0} \frac{\mathcal{E}\left(f^{\frac{1}{p}}, f^{1-\frac{1}{p}}\right)}{\mathrm{Ent}_\pi(f)}. \tag{4.135}$$

We also define the spectral gap

$$\lambda := \inf_{f:\mathcal{X}\to\mathbb{R}_{\ge 0}, \mathrm{Var}(f)>0} \frac{\mathcal{E}(f,f)}{\mathrm{Var}(f)}, \tag{4.136}$$

where $\mathrm{Var}(f) = \mathbb{E}_\pi(f - \mathbb{E}_\pi f)^2$. For any $p > 1$, we have

$$\frac{p^2}{2(p-1)}\alpha_p \le \lambda. \tag{4.137}$$

The case $p = 2$ is proved in [51], and the general case is proved in [106]. Their proof in fact implies a stronger inequality.

Lemma 4.21.

$$\limsup_{x\to 0^+} \frac{b_p(x)}{x} \le \frac{2(p-1)}{p^2}\lambda. \tag{4.138}$$

In particular, when b_p is strictly concave near 0, we have

$$\alpha_p < \limsup_{x\to 0^+} \frac{b_p(x)}{x} \le \frac{2(p-1)}{p^2}\lambda, \tag{4.139}$$

and (4.137) *is strict.*

Proof. Take any $g : \mathcal{X} \to \mathbb{R}_{\geq 0}$ with $\mathrm{Var}(g) > 0$. Define $f_\epsilon = 1 + \epsilon g$. As $\epsilon \to 0$, we have

$$\mathcal{E}\left(f_\epsilon^{\frac{1}{p}}, f_\epsilon^{1-\frac{1}{p}}\right) = \epsilon^2 \frac{1}{p}\left(1 - \frac{1}{p}\right)\mathcal{E}(g,g) + o(\epsilon^2), \tag{4.140}$$

$$\mathrm{Ent}_\pi(f_\epsilon) = \frac{1}{2}\epsilon^2 \,\mathrm{Var}(g) + o(\epsilon^2). \tag{4.141}$$

Because $\mathrm{Ent}_\pi(f_\epsilon) \to 0$ continuously as $\epsilon \to 0$, we have

$$\limsup_{x\to 0^+} \frac{b_p(x)}{x} \leq \lim_{\epsilon \to 0} \frac{\mathcal{E}\left(f_\epsilon^{\frac{1}{p}}, f_\epsilon^{1-\frac{1}{p}}\right)}{\mathrm{Ent}_\pi(f_\epsilon)} = \frac{2(p-1)}{p^2}\frac{\mathcal{E}(g,g)}{\mathrm{Var}(g)}. \tag{4.142}$$

Lemma then follows because g is arbitrary. □

Roughly speaking, existence of a "linear piece" near 0 in $\breve{b}_p$ implies that (4.137) is strict. For the Potts semigroup with $q \geq 3$, b_p is strictly concave near 0 by proof of Prop. 4.20. So (4.137) is strict for the Potts semigroup.

The story for non-linear SDPI is very similar. Let W be a channel and ν be an input distribution. Consider the tightest possible non-linear SDPI given by

$$s(x) := \sup_{\mu : D(\mu||\nu) = x} D(\mu W||\nu W). \tag{4.143}$$

The input-restricted KL divergence contraction coefficient is

$$\eta_{\mathrm{KL}}(\nu, W) := \sup_{x>0} \frac{s(x)}{x} = \sup_{\mu : 0 < D(\mu||\nu) < \infty} \frac{D(\mu W||\nu W)}{D(\mu||\nu)}. \tag{4.144}$$

We also consider the input-restricted χ^2-divergence contraction coefficient

$$\eta_{\chi^2}(\nu, W) := \sup_{\mu : 0 < \chi^2(\mu||\nu) < \infty} \frac{\chi^2(\mu W||\nu W)}{\chi^2(\mu||\nu)}. \tag{4.145}$$

It is known ([11]) that

$$\eta_{\mathrm{KL}}(\nu, W) \geq \eta_{\chi^2}(\nu, W). \tag{4.146}$$

Similarly to the p-NLSI case, the proof of (4.146) implies a stronger inequality.

Lemma 4.22.

$$\liminf_{x\to 0^+} \frac{s(x)}{x} \geq \eta_{\chi^2}(\nu, W). \tag{4.147}$$

In particular, when s is strictly convex near 0, we have

$$\eta_{\mathrm{KL}}(\nu, W) > \liminf_{x\to 0^+} \frac{s(x)}{x} \geq \eta_{\chi^2}(\nu, W). \tag{4.148}$$

and (4.146) *is strict.*

Proof. Fix any distribution μ with $0 < \chi^2(\mu||\nu) < \infty$. Proof of [115, Theorem 2] constructs a sequence of distributions μ_ϵ satisfying

$$D(\mu_\epsilon||\nu) = \epsilon^2\chi^2(\mu||\nu) + o(\epsilon^2), \tag{4.149}$$
$$D(\mu_\epsilon W||\nu W) = \epsilon^2\chi^2(\mu W||\nu W) + o(\epsilon^2), \tag{4.150}$$

and $D(\mu_\epsilon||\nu) \to 0$ continuously as $\epsilon \to 0$. Therefore

$$\liminf_{x\to 0^+} \frac{s(x)}{x} \geq \lim_{\epsilon\to 0} \frac{D(\mu_\epsilon W||\nu W)}{D(\mu_\epsilon||\nu)} = \frac{\chi^2(\mu W||\nu W)}{\chi^2(\mu||\nu)}. \tag{4.151}$$

Lemma follows because μ is arbitrary. □

Roughly speaking, existence of a "linear piece" near 0 in $\widehat{s}$ implies that (4.146) is strict. For Potts channels P_λ with $\lambda \in \left[-\frac{1}{q-1}, 0\right) \cup (0, 1)$ and $q \geq 3$, s_λ is strictly convex near 0 by proof of Prop. 4.20. So (4.146) is strict for Potts channels.

4.3.3 Edge isoperimetric inequalities

As a toy application of the NLSIs for the product spaces, we derive an edge isoperimetric inequality for K_q^n, the graph whose vertex set is $[q]^n$, and edges connect vertex pairs with Hamming distance one. Given a graph $G = (V, E)$, edge isoperimetric inequalities solve the following combinatorial optimization problem:

$$\Psi_G(N) = \min\{|E(S, S^c)| : |S| = N\}, \tag{4.152}$$

where $|E(S, S^c)| = \#\{e \in E : |e \cap S| = 1\}$. For K_q^n, the edge isoperimetric problem has been completely solved [78, 90, 18, 79]. Specifically, [90] showed that the optimal S minimizing $|E(S, S^c)|$ for a fixed $|S|$ consists of largest elements in $[q]^n$ under a lexicographical order. In particular, we have

$$\Psi_{K_q^n}(q^m) = (n - m)(q - 1)q^m. \tag{4.153}$$

This was obtained by an explicit combinatorial argument (via a form of shifting/compression). What estimates can be obtained via LSIs and NLSIs?

Let $f = \mathbb{1}_S$ be the indicator function of a set S. Then for any $p > 1$ we have

$$\frac{\mathcal{E}(f^p, f^{1-p})}{\mathbb{E}_\pi[f]} = \frac{1}{q-1}\frac{|E(S, S^c)|}{|S|} \qquad \text{and} \qquad \frac{\mathrm{Ent}(f)}{\mathbb{E}_\pi[f]} = \log\frac{q^n}{|S|}. \tag{4.154}$$

If we relate these two ratios via the 2-LSI (note that from (4.9), of all $p > 1$ the $p = 2$ gives the best result here) and by using the known value of α_2 from (4.12) we get

$$\Psi_{K_q^n}(q^m) \geq q^m(n-m)(q-2)\frac{\log q}{\log(q-1)}. \tag{4.155}$$

Clearly the coefficient in front of $(n-m)q^m$ here is not tight.

The p-NLSI allows us to perform a better comparison. First, again via (4.9) we get the best inequality for $p = 2$, which results in

$$\Psi_{K_q^n}(q^m) \geq (q-1)q^m n\breve{b}_2\left(\frac{n-m}{n}\log q\right). \tag{4.156}$$

We know that the function $\breve{b}_2$ is continuous with $\breve{b}_2(\log q) = b_2(\log q) = 1$ (from (4.18)). Thus, for any $m = o(n)$ and $n \to \infty$ we get that (4.156) implies

$$\Psi_{K_q^n}(q^m) \geq (q-1)q^m(n-m)(1+o(1)), \tag{4.157}$$

which is tight in this regime. (However, from (4.18) we can also find that $\breve{b}_2'(1) = \infty$ and thus, even when $m = o(n)$ the right-hand side of the above inequality is $(q-1)q^m(n-\omega(m))$, implying the behavior in terms of m is not optimal.)

4.4 Input-restricted contraction coefficient of the coloring channel

In this section we compute the exact input-restrict KL contraction coefficient of the coloring channel $\mathrm{Col}_q := P_{-\frac{1}{q-1}}$.

Proposition 4.23.

$$\eta_{\mathrm{KL}}(\pi, \mathrm{Col}_q) = \frac{\log q - \log(q-1)}{\log q}. \tag{4.158}$$

Proof. By (4.25) we have

$$\begin{aligned}\eta_{\mathrm{KL}}(\pi, \mathrm{Col}_q) &= \sup_{x\in\left(\frac{1}{q},1\right]} \frac{\log q + \frac{1-x}{q-1}\log\frac{1-x}{q-1} + \frac{q+x-2}{q-1}\log\frac{q+x-2}{(q-1)^2}}{\log q + x\log x + (1-x)\log\frac{1-x}{q-1}} \\ &= \sup_{x\in\left(\frac{1}{q},1\right]} \frac{\log q - \log(q-1) + \frac{1-x}{q-1}\log(1-x) + \frac{q+x-2}{q-1}\log\frac{q+x-2}{q-1}}{\log q + x\log x + (1-x)\log\frac{1-x}{q-1}}.\end{aligned} \tag{4.159}$$

Taking $x = 1$, we get

$$\eta_{\mathrm{KL}}(\pi, \mathrm{Col}_q) \geq \frac{\log q - \log(q-1)}{\log q}. \tag{4.160}$$

To prove the proposition, we only need to prove that for $x \in \left(\frac{1}{q}, 1\right]$,

$$\frac{\frac{1-x}{q-1}\log(1-x) + \frac{q+x-2}{q-1}\log\frac{q+x-2}{q-1}}{x\log x + (1-x)\log\frac{1-x}{q-1}} \geq \frac{\log q - \log(q-1)}{\log q}. \tag{4.161}$$

(Note that both numerator and denominator in LHS are non-positive.) Define

$$g(x) = (\log q - \log(q-1))x\log x - \frac{\log q}{q-1}(1-x)\log(1-x), \tag{4.162}$$

$$h(x) = g(x) + (q-1)g\left(\frac{1-x}{q-1}\right). \tag{4.163}$$

Rearranging (4.161), we only need to prove that $h(x) \geq 0$ for $x \in \left(\frac{1}{q}, 1\right]$.

We compute that

$$g'(x) = (\log q - \log(q-1))(1+\log x) + \frac{\log q}{q-1}(1+\log(1-x)), \tag{4.164}$$

$$g''(x) = (\log q - \log(q-1))\frac{1}{x} - \frac{\log q}{q-1}\frac{1}{1-x}, \tag{4.165}$$

$$g'''(x) = -(\log q - \log(q-1))\frac{1}{x^2} - \frac{\log q}{q-1}\frac{1}{(1-x)^2} < 0. \tag{4.166}$$

Claim 4.24. *$h'''(x) < 0$ on $(0,1)$.*

Proof.

$$\begin{aligned}
h'''(x) &= g'''(x) - \frac{1}{(q-1)^2}g'''\left(\frac{1-x}{q-1}\right) \\
&= -(\log q - \log(q-1))\frac{1}{x^2} - \frac{\log q}{q-1}\frac{1}{(1-x)^2} \\
&\quad + \frac{1}{(q-1)^2}\left((\log q - \log(q-1))\frac{1}{\left(\frac{1-x}{q-1}\right)^2} + \frac{\log q}{q-1}\frac{1}{(1-\frac{1-x}{q-1})^2}\right) \\
&= (\log\frac{q}{q-1})\left(\frac{1}{(1-x)^2} - \frac{1}{x^2}\right) + \frac{\log q}{q-1}\left(\frac{1}{(q-2+x)^2} - \frac{1}{(1-x)^2}\right) \\
&= \frac{1}{(1-x)^2}\left(\left(\log\frac{q}{q-1}\right)\left(1 - \frac{(1-x)^2}{x^2}\right) + \frac{\log q}{q-1}\left(\frac{(1-x)^2}{(q-2+x)^2} - 1\right)\right) \\
&=: \frac{1}{(1-x)^2}(s(x)+t(x)).
\end{aligned}$$

We have

1. $s(x) < 0$ for $x < \frac{1}{2}$, $s(x) > 0$ for $x > \frac{1}{2}$;
2. $t(x) < 0$ for $x \in (0,1)$;

3. $s(x)$ is increasing for $x \in (0,1)$;

4. $t(x)$ is decreasing for $x \in (0,1)$.

So $h'''(x) < 0$ for $x \le \frac{1}{2}$. For $x \ge \frac{1}{2}$, we have

$$s(x) + t(x) < s(1) + t\left(\frac{1}{2}\right) = \log\frac{q}{q-1} + \frac{\log q}{q-1}\left(\frac{1}{(2q-3)^2} - 1\right). \tag{4.167}$$

It is not hard to verify that the last term is < 0 for $q \ge 3$. □

By Claim 4.24, $h'(x)$ is strictly concave. Because $h'\left(\frac{1}{q}\right) = 0$, $h\left(\frac{1}{q}\right) = h(1) = 0$, we get that $h(x) > 0$ for $x \in (1/q, 1)$. This finishes the proof. □

4.5 Input-unrestricted contraction coefficient of Potts channels

Computation of (input-restricted or input-unrestricted) contraction coefficients is often a daunting task. Previously, [94] obtained lower and upper bounds of input-unrestricted KL divergence contraction coefficients for Potts channels. In this section we compute the exact value of these contraction coefficients.

We remark that after our work, [111] proved that the input-unrestricted contraction coefficients are achieved by input distributions of support size at most two, giving an alternative (and simpler) proof for Prop. 4.25. We include our original proof here for completeness.

Proposition 4.25.

$$\eta_{\mathrm{KL}}(P_\lambda) = \frac{q\lambda^2}{(q-2)\lambda + 2}. \tag{4.168}$$

Proof. The result is obvious for $\lambda \in \{0,1\}$. In the following, assume that $\lambda \notin \{0,1\}$.

We use the following characterization of contraction coefficient using Rényi maximal correlation [119] (see e.g. [122]). For any channel M, we have

$$\eta_{\mathrm{KL}}(M) = \left(\sup_{\mu}\sup_{f,g} \mathbb{E}[f(X)g(Y)]\right)^2 \tag{4.169}$$

where μ is a distribution on $[q]$, $X \sim \mu$, $Y \sim \mu M$, $f : \mathcal{X} \to \mathbb{R}$ satisfies $\mathbb{E}_X[f] = 0$ and $\mathbb{E}_X[f^2] = 1$, and $g : \mathcal{Y} \to \mathbb{R}$ satisfies $\mathbb{E}_Y[g] = 0$ and $\mathbb{E}_Y[g^2] = 1$.

Specialize to $M = P_\lambda$. Write $\mu = (p_1, \ldots, p_q)$, $f = (f_1, \ldots, f_q)$ and $g = (g_1, \ldots, g_q)$. Then

$$\mathbb{E}[f(X)g(Y)] = \sum_{i,j} f_i p_i g_j \mathbb{P}[Y = j | X = i] = \lambda \sum f_i p_i g_i. \tag{4.170}$$

When $\lambda > 0$, we need to maximize $\sum f_i g_i p_i$. When $\lambda < 0$, we make the transform $f_i \leftarrow -f_i$, and still maximize $\sum f_i g_i p_i$. So we get the following optimization problem.

$$\begin{aligned} & \max \sum f_i g_i p_i \\ \text{s.t.} \quad & \sum f_i p_i = 0, && (4.171) \\ & \sum f_i^2 p_i = 1, && (4.172) \\ & \sum g_i \left(\lambda p_i + \frac{1-\lambda}{q} \right) = 0, && (4.173) \\ & \sum g_i^2 \left(\lambda p_i + \frac{1-\lambda}{q} \right) = 1, && (4.174) \\ & p_i \geq 0, \sum p_i = 1. && (4.175) \end{aligned}$$

Lower bound. Take

$$\mu = \left(\frac{1}{2}, \frac{1}{2}, 0, \ldots, 0 \right), \quad f = (1, -1, 0, \ldots, 0), \quad g = (u, -u, 0, \ldots, 0) \tag{4.176}$$

where

$$u = \sqrt{\frac{q}{(q-2)\lambda + 2}}. \tag{4.177}$$

Then

$$\sum f_i g_i p_i = u. \tag{4.178}$$

So

$$\eta_{\mathrm{KL}}(P_\lambda) \geq (\lambda u)^2 = \frac{q\lambda^2}{(q-2)\lambda + 2}. \tag{4.179}$$

Upper bound. Let us fix μ and maximize over f and g. Assume for the sake of contrary that $\sum f_i g_i p_i > u$. The set of possible g is bounded; some coordinates of f may be unbounded, but their values do not affect the objective function. So the maximum value of $\sum f_i g_i p_i$ is achieved at some point f and g. Let us compute the

derivatives.

$$\nabla_f \sum f_i g_i p_i = (g_i p_i)_{i\in[q]}, \tag{4.180}$$

$$\nabla_f \sum f_i p_i = (p_i)_{i\in[q]}, \tag{4.181}$$

$$\nabla_f \sum f_i^2 p_i = (2 f_i p_i)_{i\in[q]}, \tag{4.182}$$

$$\nabla_g \sum f_i g_i p_i = (f_i p_i)_{i\in[q]}, \tag{4.183}$$

$$\nabla_g \sum g_i \left(\lambda p_i + \frac{1-\lambda}{q}\right) = \left(\lambda p_i + \frac{1-\lambda}{q}\right)_{i\in[q]}, \tag{4.184}$$

$$\nabla_g \sum g_i^2 \left(\lambda p_i + \frac{1-\lambda}{q}\right) = \left(2 g_i \left(\lambda p_i + \frac{1-\lambda}{q}\right)\right)_{i\in[q]}. \tag{4.185}$$

By maximality in f, there exists some constants A and B such that

$$g_i p_i = A p_i + B f_i p_i \tag{4.186}$$

for all i. By maximality in g, there exists some constants C and D such that

$$f_i p_i = C\left(\lambda p_i + \frac{1-\lambda}{q}\right) + D g_i \left(\lambda p_i + \frac{1-\lambda}{q}\right) \tag{4.187}$$

for all i.

By (4.186),

$$\sum f_i g_i p_i = \sum f_i (A p_i + B f_i p_i) = B. \tag{4.188}$$

By (4.187),

$$\sum f_i g_i p_i = \sum g_i \left(C\left(\lambda p_i + \frac{1-\lambda}{q}\right) + D g_i \left(\lambda p_i + \frac{1-\lambda}{q}\right)\right) = D. \tag{4.189}$$

So $B = D > u > 0$.

For $p_i \neq 0$, we have $g_i = A + B f_i$ by (4.186).

If for some i, $p_i = 0$, then

$$\frac{1-\lambda}{q}(C + D g_i) = 0. \tag{4.190}$$

This means $\#\{g_i : p_i = 0\} = 1$. So we can choose f_i for such i such that

$$g_i = A + B f_i \tag{4.191}$$

for all i.

From (4.173), we get

$$\begin{aligned} 0 &= \sum g_i \left(\lambda p_i + \frac{1-\lambda}{q}\right) \\ &= \sum (A + Bf_i) \left(\lambda p_i + \frac{1-\lambda}{q}\right) \\ &= A + B\frac{1-\lambda}{q}\sum f_i. \end{aligned} \tag{4.192}$$

From (4.174), we get

$$\begin{aligned} 1 &= \sum g_i^2 \left(\lambda p_i + \frac{1-\lambda}{q}\right) \\ &= \sum (A^2 + 2ABf_i + B^2 f_i^2) \left(\lambda p_i + \frac{1-\lambda}{q}\right) \\ &= A^2 + 2AB\frac{1-\lambda}{q}\sum f_i + B^2\lambda + B^2\frac{1-\lambda}{q}\sum f_i^2 \\ &= B^2\left(\lambda + \frac{1-\lambda}{q}\sum f_i^2 - \left(\frac{1-\lambda}{q}\sum f_i\right)^2\right). \end{aligned} \tag{4.193}$$

The result then follows from Claim 4.26 because we have

$$\begin{aligned} B &= \frac{1}{\sqrt{\lambda + \frac{1-\lambda}{q}\left(\sum f_i^2 - \frac{1-\lambda}{q}\left(\sum f_i\right)^2\right)}} \\ &\le \frac{1}{\sqrt{\lambda + \frac{1-\lambda}{q}\left(\sum f_i^2 - \frac{1}{q-1}\left(\sum f_i\right)^2\right)}} \\ &\le \frac{1}{\sqrt{\lambda + \frac{1-\lambda}{q}\cdot 2}} = u. \end{aligned} \tag{4.194}$$

□

Claim 4.26. *For any distribution μ and any f satisfying* (4.171) *and* (4.172)*, we have*

$$\sum f_i^2 - \frac{1}{q-1}\left(\sum f_i\right)^2 \ge 2. \tag{4.195}$$

Proof. Let us first prove the result for f with support size two. WLOG assume that $f_1 > 0$, $f_2 < 0$, $f_3 = \cdots = f_q = 0$. One can compute that

$$f_1 = \sqrt{\frac{p_2}{p_1(p_1+p_2)}}, \quad f_2 = -\sqrt{\frac{p_1}{p_1(p_1+p_2)}}. \tag{4.196}$$

Then

$$\begin{aligned}
&f_1^2 + f_2^2 - \frac{1}{q-1}(f_1+f_2)^2 \\
&\geq f_1^2 + f_2^2 - (f_1+f_2)^2 \\
&= \frac{1}{p_1+p_2}\left(\frac{p_2}{p_1} + \frac{p_1}{p_2} - \left(\sqrt{\frac{p_2}{p_1}} - \sqrt{\frac{p_1}{p_2}}\right)^2\right) \\
&= \frac{2}{p_1+p_2} \geq 2.
\end{aligned} \tag{4.197}$$

Let us define

$$S(\mu) := \left\{ f : \sum f_i p_i = 0, \sum f_i^2 p_i = 1 \right\} \tag{4.198}$$

$$U(f) := \sum f_i^2 - \frac{1}{q-1}\left(\sum f_i\right)^2. \tag{4.199}$$

Now suppose that for some μ and $f \in S(\mu)$ we have $U(f) < 2$. The set $S(\mu)/\{\pm\}$ is continuous, and there exists $f \in S(\mu)$ with $U(f) \geq 2$ (e.g., f with support size two), so for sufficiently small $\epsilon > 0$ there exists $f \in S(\mu)$ such that $U(f) \in (2-\epsilon, 2)$.

Let $\lambda = -\frac{1}{q-1}$. Take ϵ small enough so that $\lambda + \frac{1-\lambda}{q}(2-\epsilon) > 0$ and choose $f \in S(\mu)$ with $U(f) \in (2-\epsilon, 2)$. Define

$$B = \frac{1}{\sqrt{\lambda + \frac{1-\lambda}{q}U(f)}} > u, \tag{4.200}$$

$$A = -B\frac{1-\lambda}{q}\sum f_i, \tag{4.201}$$

$$g_i = A + Bf_i \forall i. \tag{4.202}$$

One can check that g satisfies (4.173) and (4.174), and

$$\sum f_i g_i p_i = B > u. \tag{4.203}$$

By (4.169) and (4.170), this implies

$$\eta_{\mathrm{KL}}\left(P_{-\frac{1}{q-1}}\right) > \frac{1}{q-1}. \tag{4.204}$$

However, we have

$$\eta_{\mathrm{KL}}\left(P_{-\frac{1}{q-1}}\right) \leq \eta_{\mathrm{TV}}\left(P_{-\frac{1}{q-1}}\right) = \frac{1}{q-1}. \tag{4.205}$$

Contradiction. □

4.6 An upper bound for input-restricted contraction coefficient for Potts channels

In this section we prove an upper bound for the input-restricted KL divergence contraction coefficient for ferromagnetic Potts channels.

Proposition 4.27. *Fix $q \geq 3$. For all $\lambda \in [0,1]$, we have*

$$\eta_{\mathrm{KL}}(\pi, P_\lambda) \leq \frac{\lambda^2}{(1-\lambda)\frac{2(q-1)\log(q-1)}{q(q-2)} + \lambda}. \tag{4.206}$$

For all $\lambda \in [-\frac{1}{q-1}, 0]$, we have

$$\eta_{\mathrm{KL}}(\pi, P_\lambda) \leq \frac{\lambda^2}{(1+(q-1)\lambda)\frac{2(q-1)\log(q-1)}{q(q-2)} - \lambda\frac{\log q}{(q-1)(\log q - \log(q-1))}}. \tag{4.207}$$

We first prove a lemma.

Lemma 4.28. *$\frac{(qx-1)^2}{\psi(x)}$ is concave in $x \in [0,1]$.*

Proof. Let $f(x) = \frac{(qx-1)^2}{\psi(x)}$.

$$f'(x) = \frac{2q(qx-1)}{\psi(x)} - \frac{(qx-1)^2\psi'(x)}{\psi^2(x)}. \tag{4.208}$$

$$\begin{aligned} f''(x) &= \frac{2q^2}{\psi(x)} - \frac{4q(qx-1)\psi'(x)}{\psi^2(x)} - \frac{(qx-1)^2\psi''(x)}{\psi^2(x)} + \frac{2(qx-1)^2(\psi')^2(x)}{\psi^3(x)} \\ &= \frac{2}{\psi^3(x)}(q\psi(x) - (qx-1)\psi'(x))^2 - \frac{(qx-1)^2\psi''(x)}{\psi^2(x)}. \end{aligned} \tag{4.209}$$

Therefore it suffices to prove that

$$g(x) := \psi^3(x)f''(x) = 2(q\psi(x) - (qx-1)\psi'(x))^2 - (qx-1)^2\psi(x)\psi''(x) \tag{4.210}$$

is non-positive for $x \in [0,1]$. Note that $g\left(\frac{1}{q}\right) = 0$. So we only need to prove that $g'(x) \geq 0$ for $x \in [0, \frac{1}{q}]$ and $g'(x) \leq 0$ for $x \in \left[\frac{1}{q}, 1\right]$.

$$\begin{aligned} g'(x) &= -4(qx-1)\psi''(x)(q\psi(x) - (qx-1)\psi'(x)) - 2q(qx-1)\psi(x)\psi''(x) \\ &\quad - (qx-1)^2\psi'(x)\psi''(x) - (qx-1)^2\psi(x)\psi'''(x) \\ &= (qx-1)(-6q\psi(x)\psi''(x) + (qx-1)(3\psi'(x)\psi''(x) - \psi(x)\psi'''(x))). \end{aligned} \tag{4.211}$$

Therefore we would like to prove that

$$u(q,x) := -6q\psi(x)\psi''(x) + (qx-1)(3\psi'(x)\psi''(x) - \psi(x)\psi'''(x)) \tag{4.212}$$

is non-positive. We enlarge the domain of u and prove that $u(q,x) \le 0$ for real $q > 1$ and $x \in (0,1)$.

We fix $x \in (0,1)$ and consider $u_x(q) := u(q,x)$. We have $u_x\left(\frac{1}{x}\right) = 0$. So it suffices to prove that u_x is concave in q. We have

$$\psi'(x) = \log x - \log \frac{1-x}{q-1}, \tag{4.213}$$

$$\psi''(x) = \frac{1}{x} + \frac{1}{1-x}, \tag{4.214}$$

$$\psi'''(x) = \frac{1}{(1-x)^2} - \frac{1}{x^2}, \tag{4.215}$$

$$\frac{\partial}{\partial q}\psi(x) = \frac{1}{q} - \frac{1-x}{q-1}, \tag{4.216}$$

$$\frac{\partial}{\partial q}\psi'(x) = \frac{1}{q-1}, \tag{4.217}$$

$$\frac{\partial}{\partial q}\psi''(x) = \frac{\partial}{\partial q}\psi'''(x) = 0. \tag{4.218}$$

So

$$\begin{aligned} u_x'(q) = & -6\psi(x)\psi''(x) - 6q\left(\frac{1}{q} - \frac{1-x}{q-1}\right)\psi''(x) + x(3\psi'(x)\psi''(x) - \psi(x)\psi'''(x)) \\ & + (qx-1)\left(3\frac{1}{q-1}\psi''(x) - \left(\frac{1}{q} - \frac{1-x}{q-1}\right)\psi'''(x)\right). \end{aligned} \tag{4.219}$$

$$\begin{aligned} u_x''(q) = & -12\left(\frac{1}{q} - \frac{1-x}{q-1}\right)\psi''(x) - 6q\left(-\frac{1}{q^2} + \frac{1-x}{(q-1)^2}\right)\psi''(x) \\ & + 6x\frac{1}{q-1}\psi''(x) - 2x\left(\frac{1}{q} - \frac{1-x}{q-1}\right)\psi'''(x) \\ & + (qx-1)\left(-3\frac{1}{(q-1)^2}\psi''(x) - \left(-\frac{1}{q^2} + \frac{1-x}{(q-1)^2}\right)\psi'''(x)\right) \\ = & \frac{(qx-1)^2(1-2q+(q-2)x)}{q^2(q-1)^2x^2(1-x)^2} \le 0. \end{aligned} \tag{4.220}$$

We are done. □

Proof of Prop. 4.27. For fixed x, we would like to lower bound

$$f_x(\lambda) := \frac{\lambda^2\psi(x)}{\psi\left(\lambda x + \frac{1-\lambda}{q}\right)}. \tag{4.221}$$

(Value of $f_x(0)$ is defined by continuity.) Because

$$f_x(\lambda) = \frac{\psi(x)}{(qx-1)^2} \cdot \frac{(q\left(\lambda x + \frac{1-\lambda}{q}\right) - 1)^2}{\psi\left(\lambda x + \frac{1-\lambda}{q}\right)}, \tag{4.222}$$

by Lemma 4.28, $f_x(\lambda)$ is concave for $\lambda \in \left[-\frac{1}{q-1}, 1\right]$.

Let us compute lower bounds of $f_x(\lambda)$ for $\lambda = -\frac{1}{q-1}, 0, 1$.

By Prop. 4.23, we have

$$f_x\left(-\frac{1}{q-1}\right) \geq \frac{\log q}{(q-1)^2(\log q - \log(q-1))}. \tag{4.223}$$

By L'Hôpital's rule,

$$\begin{aligned} f_x(0) &= \psi(x) \lim_{\lambda \to 0} \frac{2\lambda}{\left(x - \frac{1}{q}\right)\psi'\left(\lambda x + \frac{1-\lambda}{q}\right)} \\ &= \psi(x) \lim_{\lambda \to 0} \frac{2}{\left(x - \frac{1}{q}\right)^2 \psi''\left(\lambda x + \frac{1-\lambda}{q}\right)} \\ &= \frac{2(q-1)\psi(x)}{(qx-1)^2}. \end{aligned} \tag{4.224}$$

By Lemma 4.28, $g(x) := \frac{(qx-1)^2}{\psi(x)}$ is concave in x. Also

$$g'\left(1 - \frac{1}{q}\right) = \frac{2q(q-2)}{\frac{1}{q}(q-2)\log(q-1)} - \frac{(q-2)^2 \cdot 2\log(q-1)}{\left(\frac{1}{q}(q-2)\log(q-1)\right)^2} = 0. \tag{4.225}$$

So

$$g(x) \leq g\left(1 - \frac{1}{q}\right) = \frac{q(q-2)}{\log(q-1)} \tag{4.226}$$

and

$$f_x(0) \geq \frac{2(q-1)\log(q-1)}{q(q-2)}. \tag{4.227}$$

It is easy to see that

$$f_x(1) \geq 1. \tag{4.228}$$

Because $f_x(\lambda)$ is concave in λ, Inequality (4.206) follows from (4.227) and (4.228), and Inequality (4.207) follows from (4.223) and (4.227). □

Proof of Prop. 4.27 implies the first order limit behavior of $\eta_{\mathrm{KL}}(\pi, P_\lambda)$ as $\lambda \to 0$.

$$\lim_{\lambda\to 0} \frac{\eta_{\mathrm{KL}}(\pi, P_\lambda)}{\lambda^2} = \frac{q(q-2)}{2(q-1)\log(q-1)}. \tag{4.229}$$

For all $q \geq 3$ and $\lambda \in (0, 1]$, we have

$$\begin{aligned}
\eta_{\mathrm{KL}}(\pi, P_\lambda) &\leq \frac{\lambda^2}{(1-\lambda)\frac{2(q-1)\log(q-1)}{q(q-2)} + \lambda} \\
&\leq \lambda^2(1-\lambda)\frac{q(q-2)}{2(q-1)\log(q-1)} + \lambda^3 \\
&< \lambda^2 \frac{q(q-2)}{2(q-1)\log(q-1)} \\
&< \lambda^2 \frac{q-1}{2\log(q-1)},
\end{aligned} \tag{4.230}$$

where the second step is by Cauchy inequality.

For comparison with input-unrestricted contraction coefficient

$$\eta_{\mathrm{KL}}(P_\lambda) = \frac{q\lambda^2}{(q-2)\lambda + 2}, \tag{4.231}$$

we note that $\frac{\lambda^2}{\eta_{\mathrm{KL}}(P_\lambda)}$ is linear in λ, and

$$\frac{1}{q-1} < \frac{\log q}{(q-1)^2(\log q - \log(q-1))}, \tag{4.232}$$

$$\frac{2}{q} < \frac{2(q-1)\log(q-1)}{q(q-2)}. \tag{4.233}$$

So Prop. 4.27 implies (4.26).

4.7 Non-convexity of certain functions

In this section we prove Prop. 4.20. Let us first prove a lemma.

Lemma 4.29. *Let g be a strictly increasing smooth function from $[x_0, x_1]$ to $[y_0, y_1]$, and f be a smooth function from $[x_0, x_1]$ to $\mathbb{R}$. Assume that $g'(x_0) = f'(x_0) = 0$ and $(g''f''' - f''g''')(x_0) > 0$. Then the function $h = f \circ g^{-1} : [y_0, y_1] \to \mathbb{R}$ is not concave near y_0.*

Proof. Directives of h are

$$h'(x) = \frac{f'(g^{-1}(x))}{g'(g^{-1}(x))}, \tag{4.234}$$

$$\begin{aligned} h''(x) &= \left(\frac{f''}{g'} - \frac{f'g''}{(g')^2}\right)\left(g^{-1}(x)\right)\frac{1}{g'\left(g^{-1}(x)\right)} \\ &= \left(\frac{f''}{(g')^2} - \frac{f'g''}{(g')^3}\right)\left(g^{-1}(x)\right). \end{aligned} \tag{4.235}$$

So it suffices to study the sign of $g'f'' - f'g''$ for x near x_0. Let $u = g'f'' - f'g''$. We have $u(x_0) = 0$. Let us compute the derivatives.

$$u' = g'f''' - f'g''', \tag{4.236}$$

$$u'' = g'f^{(4)} + g''f''' - f''g''' - g'g^{(4)}. \tag{4.237}$$

So $u'(x_0) = 0$ and $u''(x_0) = (g''f''' - f''g''')(x_0) > 0$. So u is positive near x_0. □

Proof of Prop. 4.20. We apply Lemma 4.29 to $g = \psi$, $x_0 = \frac{1}{q}$, $x_1 = 1$, $y_0 = 0$, $y_1 = \log q$, and various f. We have

$$\psi'\left(\frac{1}{q}\right) = 0, \tag{4.238}$$

$$\psi''\left(\frac{1}{q}\right) = \frac{q^2}{q-1}, \tag{4.239}$$

$$\psi'''\left(\frac{1}{q}\right) = -\frac{q^3(q-2)}{(q-1)^2}. \tag{4.240}$$

Part 1. For b_1, take

$$f(x) = -(q-1)\xi_1(x) = \log x + (q-1)\log\frac{1-x}{q-1} + q(\psi(x) - \log q). \tag{4.241}$$

Then

$$f'(x) = \frac{1}{x} - \frac{q-1}{1-x} + q\psi'(x), \tag{4.242}$$

$$f''(x) = -\frac{1}{x^2} - \frac{q-1}{(1-x)^2} + q\psi''(x), \tag{4.243}$$

$$f'''(x) = \frac{2}{x^3} - \frac{2(q-1)}{(1-x)^3} + q\psi'''(x). \tag{4.244}$$

So

$$f'\left(\frac{1}{q}\right) = 0, \tag{4.245}$$

$$f''\left(\frac{1}{q}\right) = -\frac{2q^3}{q-1}, \tag{4.246}$$

$$f'''\left(\frac{1}{q}\right) = \frac{3(q-2)q^4}{(q-1)^2}. \tag{4.247}$$

We have

$$\begin{aligned}(\psi''f''' - f''\psi''')\left(\frac{1}{q}\right) &= \frac{q^2}{q-1}\cdot\frac{3(q-2)q^4}{(q-1)^2} - \left(-\frac{2q^3}{q-1}\right)\left(-\frac{q^3(q-2)}{(q-1)^2}\right) \\ &= \frac{q^6(q-2)}{(q-1)^3} > 0. \end{aligned} \tag{4.248}$$

So Lemma 4.29 applies.

Part 2. For b_p, $p > 1$, take

$$\begin{aligned} f(x) &= q - (q-1)\xi_p(x) \\ &= \left(x^{\frac{1}{p}} + (q-1)\left(\frac{1-x}{q-1}\right)^{\frac{1}{p}}\right)\left(x^{1-\frac{1}{p}} + (q-1)\left(\frac{1-x}{q-1}\right)^{1-\frac{1}{p}}\right). \end{aligned} \tag{4.249}$$

For simplicity, write $r = \frac{1}{p}$ and let $u_r(x) = x^r + (q-1)\left(\frac{1-x}{q-1}\right)^r$. Then $f(x) = u_r(x)u_{1-r}(x)$. Let us compute derivatives of u_r.

$$u_r'(x) = r\left(x^{r-1} - \left(\frac{1-x}{q-1}\right)^{r-1}\right), \tag{4.250}$$

$$u_r''(x) = r(r-1)\left(x^{r-2} + \frac{1}{q-1}\left(\frac{1-x}{q-1}\right)^{r-2}\right), \tag{4.251}$$

$$u_r'''(x) = r(r-1)(r-2)\left(x^{r-3} - \frac{1}{(q-1)^2}\left(\frac{1-x}{q-1}\right)^{r-3}\right). \tag{4.252}$$

So

$$u_r\left(\frac{1}{q}\right) = q^{1-r}, \tag{4.253}$$

$$u_r'\left(\frac{1}{q}\right) = 0, \tag{4.254}$$

$$u_r''\left(\frac{1}{q}\right) = r(r-1)\frac{q}{q-1}\left(\frac{1}{q}\right)^{r-2}, \tag{4.255}$$

$$u_r'''\left(\frac{1}{q}\right) = r(r-1)(r-2)\frac{q(q-2)}{(q-1)^2}\left(\frac{1}{q}\right)^{r-3}. \tag{4.256}$$

Now we compute derivatives of f.

$$f'(x) = u_r'(x)u_{1-r}(x) + u_r(x)u_{1-r}'(x), \tag{4.257}$$

$$f''(x) = u_r''(x)u_{1-r}(x) + 2u_r'(x)u_{1-r}'(x) + u_r(x)u_{1-r}''(x), \tag{4.258}$$

$$f'''(x) = u_r'''(x)u_{1-r}(x) + 3u_r''(x)u_{1-r}'(x) + 3u_r'(x)u_{1-r}''(x) + u_r(x)u_{1-r}'''(x). \tag{4.259}$$

So

$$f'\left(\frac{1}{q}\right) = 0, \tag{4.260}$$

$$\begin{aligned} f''\left(\frac{1}{q}\right) &= r(r-1)\frac{q}{q-1}\left(\frac{1}{q}\right)^{r-2}\cdot q^r + (1-r)(-r)\frac{q}{q-1}\left(\frac{1}{q}\right)^{-r-1}\cdot q^{1-r} \\ &= 2r(r-1)\frac{q^3}{(q-1)}, \end{aligned} \tag{4.261}$$

$$\begin{aligned} f'''\left(\frac{1}{q}\right) &= r(r-1)(r-2)\frac{q(q-2)}{(q-1)^2}\left(\frac{1}{q}\right)^{r-3}\cdot q^r \\ &\quad + (1-r)(-r)(-r-1)\frac{q(q-2)}{(q-1)^2}\left(\frac{1}{q}\right)^{-r-2}\cdot q^{1-r}, \\ &= -3r(r-1)\frac{q^4(q-2)}{(q-1)^2}. \end{aligned} \tag{4.262}$$

So

$$\begin{aligned} (\psi''f''' - f''\psi''')\left(\frac{1}{q}\right) &= \frac{q^2}{q-1}\left(-3r(r-1)\frac{q^4(q-2)}{(q-1)^2}\right) \\ &\quad - 2r(r-1)\frac{q^3}{(q-1)}\left(-\frac{q^3(q-2)}{(q-1)^2}\right) \\ &= r(1-r)\frac{q^6(q-2)}{(q-1)^3} > 0. \end{aligned} \tag{4.263}$$

So Lemma 4.29 applies.

Part 3. For s_λ, take

$$f(x) = \psi\left(\lambda x + \frac{1-\lambda}{q}\right). \tag{4.264}$$

Then

$$f'(x) = \lambda\psi'\left(\lambda x + \frac{1-\lambda}{q}\right), \tag{4.265}$$

$$f''(x) = \lambda^2\psi''\left(\lambda x + \frac{1-\lambda}{q}\right), \tag{4.266}$$

$$f'''(x) = \lambda^3\psi'''\left(\lambda x + \frac{1-\lambda}{q}\right). \tag{4.267}$$

So

$$f'\left(\frac{1}{q}\right) = 0, \tag{4.268}$$

$$f''\left(\frac{1}{q}\right) = \lambda^2\psi''\left(\frac{1}{q}\right) = \lambda^2\frac{q^2}{q-1}, \tag{4.269}$$

$$f'''\left(\frac{1}{q}\right) = \lambda^3\psi'''\left(\frac{1}{q}\right) = -\lambda^3\frac{q^3(q-2)}{(q-1)^2}. \tag{4.270}$$

We have

$$\begin{aligned}(\psi''f''' - f''\psi''')\left(\frac{1}{q}\right) &= \frac{q^2}{q-1}\left(-\lambda^3\frac{q^3(q-2)}{(q-1)^2}\right) - \lambda^2\frac{q^2}{q-1}\left(-\frac{q^3(q-2)}{(q-1)^2}\right) \\ &= \frac{q^5(q-2)}{(q-1)^3}(\lambda^2-\lambda^3) > 0.\end{aligned} \tag{4.271}$$

So Lemma 4.29 applies. □

4.8 Concavity of log-Sobolev coefficients

Let K be a Markov kernel with stationary distribution π. Define Dirichlet form $\mathcal{E}(\cdot,\cdot)$ and entropy form $\mathrm{Ent}(\cdot)$ as in Section 4.1.

For $r \in \mathbb{R}$, we consider the tightest $\frac{1}{r}$-log-Sobolev inequality, corresponding to

$$b_{\frac{1}{r}}(x) := \inf_{\substack{f:\mathcal{X}\to\mathbb{R}_{\geq 0},\\ \mathbb{E}_\pi f=1, \mathrm{Ent}_\pi(f)=x}} \mathcal{E}(f^r, f^{1-r}), \tag{4.272}$$

$$\Phi_{\frac{1}{r}}(y) := \inf_{\substack{f:\mathcal{X}\to\mathbb{R}_{\geq 0},\\ \mathbb{E}_\pi f=1, \mathcal{E}(f^r,f^{1-r})=y}} \mathrm{Ent}_\pi(f). \tag{4.273}$$

The $\frac{1}{r}$-log-Sobolev constant is

$$\alpha'_{\frac{1}{r}} := \inf_{x>0} \frac{b_{\frac{1}{r}}(x)}{x} = \inf_{y>0} \frac{y}{\Phi_{\frac{1}{r}}(y)}. \tag{4.274}$$

When $r = 0$, the fraction $\frac{1}{r}$ should be understood as a formal symbol. For $r \in (0, 1)$, $\alpha'_{\frac{1}{r}}$ is the same as $\alpha_{\frac{1}{r}}$ defined in the Introduction. However, in general α'_1 is not equal to α_1. We use the superscript $'$ to emphasize the difference.

Proposition 4.30. *We have*

1. *For fixed x, $b_{\frac{1}{r}}(x)$ is concave in r.*
2. *For fixed y, $\Phi_{\frac{1}{r}}(y)$ is convex in r.*
3. *$\alpha'_{\frac{1}{r}}$ is concave in r.*

Furthermore, if (π, K) is reversible, then

1. *For fixed x, $b_{\frac{1}{r}}(x)$ is maximized at $r = \frac{1}{2}$.*
2. *For fixed y, $\Phi_{\frac{1}{r}}(y)$ is minimized at $r = \frac{1}{2}$.*
3. *$\alpha'_{\frac{1}{r}}$ is maximized at $r = \frac{1}{2}$.*

Proof. Because $\Phi_{\frac{1}{r}}$ is the inverse function of $b_{\frac{1}{r}}$, it suffices to prove statements about $b_{\frac{1}{r}}$. Because inf of concave functions is still concave, it suffices to prove that for any $f : \mathcal{X} \to \mathbb{R}_{\geq 0}$, $\mathbb{E}_\pi f = 1$, $\mathcal{E}(f^r, f^{1-r})$ is concave in r.

$$\begin{aligned}
\frac{d}{dr^2}\mathcal{E}(f^r, f^{1-r}) &= \frac{d}{dr^2} \sum_{x,y\in\mathcal{X}} (I-K)(x,y) f(y)^r f(x)^{1-r} \pi(x) \\
&= \sum_{x,y\in\mathcal{X}} (I-K)(x,y) f(y)^r f(x)^{1-r} \pi(x) (\log f(y) - \log f(x))^2 \\
&= \sum_{x\neq y\in\mathcal{X}} -K(x,y) f(y)^r f(x)^{1-r} \pi(x) (\log f(y) - \log f(x))^2 \\
&\leq 0.
\end{aligned}$$

When the Markov chain is reversible, we have $\mathcal{E}(f, g) = \mathcal{E}(g, f)$. So $b_{\frac{1}{r}}(x) = b_{\frac{1}{1-r}}(x)$ and by concavity, $b_{\frac{1}{r}}(x)$ is maximized at $r = \frac{1}{2}$. □

Part II

Statistical Problems on Graphs

Chapter 5

Graph problems and reduction to trees

Starting from this chapter, we shift our focus to statistical problems on graphs. We study a class of sparse random graphical models with hidden community structures, known as the hypergraph stochastic block model (HSBM). The stochastic block model (SBM) is the special case of the HSBM when hyperedges are of size two. Questions concerning this model revolve around recovering community structures from the unlabeled (hyper)graph. Physicists made predictions about these models using the non-rigorous cavity method [49, 15], but the program of rigorously establishing these predictions remains far from complete.

The local structure of these graphical models is closely related to a broadcasting model on Galton-Watson random (hyper)trees, a phenomenon initially proved in [103] for the two-community symmetric SBM. We refer to this tree-like model as broadcasting on hypertrees (BOHT) or, when hyperedges have size two, as broadcasting on trees (BOT). Since then, it has been discovered that many problems on the HSBM can be reduced to problems on the corresponding BOHT model. These include weak recovery [103, 109, 74], optimal recovery algorithm [104, 73], and mutual information formula [4, 73]. In this chapter we introduce the HSBM and establish these connections.

This chapter serves as a bridge between problems on graphs and problems on trees, which are tackled in later chapters using channel comparison methods established in Part I.

Chapter outline In Section 5.1, we introduce the hypergraph stochastic block model (HSBM) and problems studied on these models. In Section 5.2, we define the corresponding broadcasting on hypertrees (BOHT) model, and discuss a coupling between the HSBM and the BOHT model. In Section 5.3, we show that non-reconstruction on the BOHT model implies impossibility of weak recovery in the corresponding HSBM. In Section 5.4, we prove the boundary irrelevance (BI) property of the BOHT model implies a mutual information formula for the corresponding HSBM. In Section 5.5, we demonstrate that a property of the BOHT model, uniqueness of belief propagation (BP) fixed point, implies optimal recovery algorithms for

the corresponding HSBM.

5.1 Hypergraph stochastic block models

We start by defining the hypergraph stochastic block model.

Definition 5.1 (Hypergraph stochastic block model [15, 130])**.** Let $n \geq 1$ (number of vertices), $q \geq 2$ (number of communities), $r \geq 2$ (hyperedge size) be integers. Let $\pi \in \mathcal{P}([q])$ be a distribution with full support. Let $\mathbf{A} \in \left(\mathbb{R}^q_{\geq 0}\right)^{\otimes r}$ be a tensor satisfying

$$a_{i_1,\ldots,i_r} = a_{i_{\sigma(1)},\ldots,i_{\sigma(r)}} \tag{5.1}$$

for any $i_1, \ldots, i_r \in [q]$, $\sigma \in \mathrm{Aut}([r])$. The hypergraph stochastic block model $\mathrm{HSBM}(n, q, r, \pi, \mathbf{A})$ is defined as follows: Let $V = [n]$ be the set of vertices. Generate a random label X_u for all vertices $u \in V$ i.i.d. $\sim \pi$. Then for every $S = \{u_1, \ldots, u_r\} \in \binom{V}{r}$, add hyperedge S to the hypergraph with probability $\frac{a_{X_{u_1},\ldots,X_{u_r}}}{\binom{n}{r-1}}$. The resulting pair $(X, G = (V, E))$ is the output of the model.

In the definition, the scaling $\frac{1}{\binom{n}{r-1}}$ keeps the average degree constant with high probability as $n \to \infty$. This is called the constant degree regime. While there have been many works on SBMs and HSBMs with growing average degree (e.g., [30, 58, 25, 127, 45, 97, 21, 133, 2, 102, 6] for SBM; [67, 68, 69, 35, 36, 89, 12, 85, 44, 139] for HSBM), in this thesis, we focus on the constant degree regime.

In the HSBM, the expected degree (number of hyperedges containing a vertex) of a vertex with label $i \in [q]$ is $d_i \pm o(1)$, where

$$d_i = \sum_{i_1,\ldots,i_{r-1} \in [q]} a_{i,i_1,\ldots,i_{r-1}} \prod_{j \in [r-1]} \pi_{i_j}. \tag{5.2}$$

If $d_i \neq d_j$ for some $i, j \in [q]$, we can distinguish community i and j using a classifier based on degree, which trivially solves the weak recovery problem (Definition 5.5). Therefore, we make the following standard assumption in literature.

Condition 5.2. We say the model $\mathrm{HSBM}(n, q, r, \pi, \mathbf{A})$ is degree indistinguishable if $d_i = d_j$ for all $i, j \in [q]$, where d_i is defined in Eq. (5.2).

Under this assumption, We define a few useful derived parameters of the HSBM.

Definition 5.3 (Derived parameters for HSBM [130])**.** Consider a model $\mathrm{HSBM}(n, q, r, \pi, \mathbf{A})$ satisfying Condition 5.2.

We define the degree d as

$$d := \sum_{i_1,\ldots,i_{r-1} \in [q]} a_{i,i_1,\ldots,i_{r-1}} \prod_{j \in [r-1]} \pi_{i_j} \tag{5.3}$$

for any $i \in [q]$.

We define the signal matrix $Q \in \mathbb{R}^{q\times q}$ as

$$Q_{i,j} := \pi_j \sum_{i_1,\ldots,i_{r-2}\in[q]} a_{i,j,i_1,\ldots,i_{r-2}} \prod_{k\in[r-2]} \pi_{i_k}. \tag{5.4}$$

Because Q is self-adjoint, its eigenvalues are all real. The largest eigenvalue of Q is d. We define signal strength $\lambda := \lambda_2(Q)/d$, where $\lambda_2(Q)$ be the second-largest eigenvalue (in absolute value) of Q. We define the signal-to-noise ratio (SNR) as

$$\mathrm{SNR} := (r-1)d\lambda^2. \tag{5.5}$$

The Kesten-Stigum (KS) threshold is at $\mathrm{SNR} = 1$.

We define the following subclasses of HSBM.

Definition 5.4 (Special cases of HSBM)**.** Consider a model $\mathrm{HSBM}(n, q, r, \pi, \mathbf{A})$.

- (Stochastic block model) If $r = 2$, then we say the model is a stochastic block model (SBM), denoted as $\mathrm{SBM}(n, q, \pi, \mathbf{A})$. In this case, the tensor $\mathbf{A}$ is a symmetric matrix.
- (Symmetric HSBM) We say the model $\mathrm{HSBM}(n, q, r, \pi, \mathbf{A})$ is symmetric, if $\pi = \mathrm{Unif}([q])$ and tensor $\mathbf{A}$ satisfies

  $$a_{i_1,\ldots,i_r} = a_{\tau(i_1),\ldots,\tau(i_r)} \tag{5.6}$$

  for any $i_1, \ldots, i_r \in [q]$, $\tau \in \mathrm{Aut}([q])$.
- (Simple HSBM) If $\pi = \mathrm{Unif}([q])$ and for some $a, b \in \mathbb{R}_{\geq 0}$, we have

  $$a_{i_1,\ldots,i_r} = \begin{cases} a, & \text{if } i_1 = \cdots = i_r, \\ b, & \text{otherwise,} \end{cases} \tag{5.7}$$

  then we say the model is a simple HSBM, denoted as $\mathrm{HSBM}(n, q, r, a, b)$.
- We denote a model $\mathrm{HSBM}(n, q, r, a, b)$ with $r = 2$ as $\mathrm{SBM}(n, q, a, b)$.
- We denote a model $\mathrm{HSBM}(n, q, r, a, b)$ with $q = 2$ as $\mathrm{HSBM}(n, r, a, b)$.

For $\mathrm{HSBM}(n, q, r, a, b)$ and its subclasses, we say the model is assortative if $a > b$, and is disassortative if $a < b$.

For the SBM and the HSBM, the central question is to recover community structure given the unlabeled hypergraph. For $(X, G) \sim \mathrm{HSBM}(n, q, r, \pi, \mathbf{A})$, we would like an estimator $\widehat{X} = \widehat{X}(G)$ such that the distance between X and $\widehat{X}$ is small. In the symmetric case, it is impossible to distinguish permutations of the community labels. Therefore, the distance between the truth and the estimation is defined as

$$d_H(X, Y) := \min_{\tau\in\mathrm{Aut}([q])} \sum_{u\in V} \mathbb{1}\{X_i \neq \sigma(Y_i)\}. \tag{5.8}$$

There are several different versions of the question of recovery.

Definition 5.5 (Recovery problems for HSBM)**.** Let $(X, G) \sim \mathrm{HSBM}(n, q, r, \pi, \mathbf{A})$. We say the model admits

- exact recovery (strong consistency), if it is possible to recover the labels exactly, i.e., there exists an estimator $\widehat{X} = \widehat{X}(G)$ such that

$$\lim_{n\to\infty} \mathbb{P}[d_H(\widehat{X}, X) = 0] = 1, \tag{5.9}$$

- almost exact recovery (weak consistency), if it is possible to recover almost all labels, i.e., there exists an estimator $\widehat{X} = \widehat{X}(G)$ such that

$$\lim_{n\to\infty} \mathbb{P}[d_H(\widehat{X}, X) = o(n)] = 1, \tag{5.10}$$

- partial recovery (max-detection), if it is possible to recover a non-trivial fraction of the labels, i.e., there exists an estimator $\widehat{X} = \widehat{X}(G)$ and some $\epsilon > 0$ such that

$$\lim_{n\to\infty} \mathbb{P}[d_H(\widehat{X}, X) < 1 - \max_i \pi_i - \epsilon + o(1)] = 1, \tag{5.11}$$

- weak recovery (detection), if there exists an estimator outputting a subset $S \subseteq V$ such that for some $\epsilon > 0$, with probability $1 - o(1)$, there exists $i, j \in [q]$ such that

$$\frac{\#\{v \in S : X_v = i\}}{\#\{v \in V : X_v = i\}} - \frac{\#\{v \in S : X_v = j\}}{\#\{v \in V : X_v = j\}} \geq \epsilon. \tag{5.12}$$

Different recovery questions are interesting in different regimes. For the constant degree regime, the partial recovery and weak recovery problems are more relevant.

The definition of partial recovery and weak recovery are equivalent when π is uniform, but have subtle differences when π is non-uniform, as shown in [9]. For SBMs, [49] conjectured that partial recovery is always possible above the Kesten-Stigum (KS) threshold (Definition. 5.3). [9] gave a counterexample to the conjecture, and proved that the conjecture is true with partial recovery replaced by weak recovery. Therefore, weak recovery captures the KS threshold better than partial recovery.

We refer the reader to [1] for a survey of results on the SBM (including all four kinds of recovery problems in Definition 5.5). Here we briefly summarize works on the weak recovery problem for the SBM and the HSBM.

Works on weak recovery for SBM. [42] gave a non-trivial algorithm for the weak recovery problem. Based on the non-rigorous cavity method, [49] conjectured that for $\mathrm{SBM}(n, q, \pi, \mathbf{A})$, the algorithmic partial recovery threshold (see our previous discussion on the relationship between partial recovery and weak recovery) is at the Kesten-Stigum threshold (Definition 5.3), and for $\mathrm{SBM}(n, q, a, b)$ there is an information-computation gap for $q \geq 4$. For $\mathrm{SBM}(n, 2, a, b)$, the positive part of the

conjecture was proved in [96, 105] and the negative part was proved in [103, 105]. For $\mathrm{SBM}(n, q, \mathrm{Unif}([q]), \mathbf{A})$, [26] gave weak recovery algorithms above the KS threshold for a general of $\mathbf{A}$, but their method does not work for the $\mathrm{SBM}(n, q, a, b)$ model due to technical reasons. This technicality was overcome in [128], giving weak recovery algorithms for any $\mathrm{SBM}(n, q, \mathrm{Unif}([q]), \mathbf{A})$ model above the KS threshold. For general $\mathrm{SBM}(n, q, \pi, \mathbf{A})$, [8, 7, 9] gave algorithms for weak recovery above the KS threshold, establishing the positive part of the conjecture in full generality. For $\mathrm{SBM}(n, q, a, b)$ with $q = 3, 4$ and d large enough (Definition 5.3), [109] established the negative part of the conjecture, proving that weak recovery is impossible below the KS threshold.

Let us move to the information theoretical side. [7, 9] showed that for $q \geq 4$, there exist parameters a, b such that $\mathrm{SBM}(n, q, a, b)$ is below the KS threshold, but weak recovery is information theoretically possible. This gives evidence for the existence of an information-computation gap. [16] also crossed the KS threshold information theoretically, for $\mathrm{SBM}(n, q, a, b)$ with $q \geq 5$. For $\mathrm{SBM}(n, q, a, b)$ with $a < b$, the information theoretical weak recovery threshold has been deteremined by [43]. However their formula for the threshold is difficult to compute. In particular, it is not known whether for $q = 3$, the weak recovery threshold coincides with the KS threshold. Our work [72] (Chapter 4, 6) gave improved impossibility of weak recovery results for $\mathrm{SBM}(n, q, a, b)$.

Works on weak recovery for HSBM. For the $\mathrm{HSBM}(n, r, a, b)$, [15] conjectured that a phase transition occurs at the Kesten-Stigum threshold. [12, 56] gave non-trivial weak recovery algorithms, but the KS threshold could not be achieved using their method. We note that [56] allowed a non-uniform version of HSBM. For $\mathrm{HSBM}(n, 2, a, b)$, [112] proved the positive part of the conjecture, giving weak recovery algorithms above the KS threshold. For general $\mathrm{HSBM}(n, q, r, \mathrm{Unif}([q]), \mathbf{A})$ [130] established the positive part of the conjecture. The only result (except for the graph case) on the impossibility part is our work [74] (Chapter 7), which proved that weak recovery is impossible below the KS threshold for $\mathrm{HSBM}(n, r, a, b)$ with $r = 3, 4$ (where the $r = 4$ case depends on a numerically verified inequality).

A very useful property of HSBMs is the vanishing of long range correlations. This was first proved in [103, Lemma 4.7] for the simple binary SBM. Here we present a version for general HSBMs. In the following, we use a.a.s. (asymptotically almost surely) to denote that a event happens with probability $1 - o(1)$.

Proposition 5.6 (No long range correlations)**.** *Let $(X, G = (V, E)) \sim \mathrm{HSBM}(n, q, r, \pi, \mathbf{A})$. Let $A = A(G), B = B(G), C = C(G) \subseteq V$ be a (random) partition of V such that B separates A and C in G (i.e., there exists no hyperedges $S \in E$ intersecting both A and C). If $|A \cup B| = o(\sqrt{n})$ a.a.s., then*

$$\mathbb{P}(X_A | X_{B \cup C}, G) = (1 \pm o(1))\mathbb{P}(X_A | X_B, G) \text{ a.a.s.} \tag{5.13}$$

Proof. Our proof is a generalization of [103, Lemma 4.7]. For $S = \{u_1, \ldots, u_r\} \in \binom{V}{r}$,

define

$$\psi_S(G,X) := \begin{cases} \frac{a_{X_{u_1},\dots,X_{u_r}}}{\binom{n}{r-1}}, & \text{if } S \in E, \\ 1 - \frac{a_{X_{u_1},\dots,X_{u_r}}}{\binom{n}{r-1}}, & \text{if } S \notin E. \end{cases} \tag{5.14}$$

Then

$$\mathbb{P}(G,X) := \mathbb{P}(X)\mathbb{P}(G|X) = \mathbb{P}(X) \prod_{S \in \binom{V}{r}} \psi_S(G,X). \tag{5.15}$$

We partition $\binom{V}{r}$ into four parts. Define

$$E_1 := \left\{ S \in \binom{V}{r} : |S \cap A| \geq 1, |S \cap C| \geq 1, |S \cap (A \cup B)| \geq 2 \right\}, \tag{5.16}$$

$$E_2 := \left\{ S \in \binom{V}{r} : |S \cap A| = 1, |S \cap C| \geq 1, |S \cap B| = 0 \right\}, \tag{5.17}$$

$$E_3 := \left\{ S \in \binom{V}{r} : |S \cap C| = 0 \right\}, \tag{5.18}$$

$$E_4 := \left\{ S \in \binom{V}{r} : |S \cap A| = 0, |S \cap C| \geq 1 \right\}. \tag{5.19}$$

Then $E_1 \cup E_2 \cup E_3 \cup E_4 = \binom{V}{r}$ is a partition of $\binom{V}{r}$. Define

$$Q_i := Q_i(G,X) := \prod_{S \in E_i} \psi_S(G,X) \quad \forall i \in [4]. \tag{5.20}$$

Then

$$\mathbb{P}(G,X) = \mathbb{P}(X) Q_1 Q_2 Q_3 Q_4. \tag{5.21}$$

We prove that Q_1 and Q_2 are approximately independent of $X_{B \cup C}$ a.a.s. Let $(\alpha_n)_{n \geq 0}$ be a deterministic sequence with $\alpha_n = \omega(\sqrt{n})$ and $\alpha_n |A| = o(n)$ a.a.s. Define

$$\Omega := \{Y \in [q]^V : |N_i(Y) - \pi_i n| \leq \alpha_n \forall i \in [q]\}, \tag{5.22}$$

$$\Omega_U := \{Y \in \Omega : Y_U = X_U\}, \tag{5.23}$$

$$\text{where} \quad N_i(Y) := \#\{v \in V : Y_v = i\}. \tag{5.24}$$

By concentration of $N_i(X)$, we have $X \in \Omega$ a.a.s.

Note that $E_1 \cap E = E_2 \cap E = \emptyset$. Also,

$$|E_1| = O(|A \cup B|^2 n^{r-2}) = o(n^{r-1}) \text{ a.a.s.} \tag{5.25}$$

So

$$Q_1 = \left(1 - \frac{O(1)}{\binom{n}{r-1}}\right)^{o(n^{r-1})} = 1 - o(1) \text{ a.a.s.} \tag{5.26}$$

For Q_2, we have

$$\begin{aligned} Q_2 &= \prod_{\substack{u \in A \\ T \in \binom{C}{r-1}}} \psi_{T \cup \{u\}}(G, X) \\ &= \prod_{\substack{u \in A \\ T = \{v_1, \ldots, v_{r-1}\} \in \binom{C}{r-1}}} \left(1 - \frac{a_{X_u, X_{v_1}, \ldots, X_{v_{r-1}}}}{\binom{n}{r-1}}\right) \\ &= (1 + o(1)) \prod_{\substack{u \in A \\ T = \{v_1, \ldots, v_{r-1}\} \in \binom{C}{r-1}}} \exp\left(-\frac{a_{X_u, X_{v_1}, \ldots, X_{v_{r-1}}}}{\binom{n}{r-1}}\right) \text{ a.a.s.} \end{aligned} \tag{5.27}$$

where the third step is because $|A| = o(\sqrt{n})$ a.a.s. and

$$\exp\left(-\frac{a_{X_u, X_{v_1}, \ldots, X_{v_{r-1}}}}{\binom{n}{r-1}}\right) = \left(1 + O\left(n^{2(1-r)}\right)\right)\left(1 - \frac{a_{X_u, X_{v_1}, \ldots, X_{v_{r-1}}}}{\binom{n}{r-1}}\right). \tag{5.28}$$

For every $u \in A$ and $X \in \Omega$, we have

$$\begin{aligned} &\prod_{\{v_1, \ldots, v_{r-1}\} \in \binom{C}{r-1}} \exp\left(-\frac{a_{X_u, X_{v_1}, \ldots, X_{v_{r-1}}}}{\binom{n}{r-1}}\right) \\ =\ & \exp\left(-\sum_{\{v_1, \ldots, v_{r-1}\} \in \binom{C}{r-1}} \frac{a_{X_u, X_{v_1}, \ldots, X_{v_{r-1}}}}{\binom{n}{r-1}}\right) \\ =\ & \exp\left(-\frac{1}{(r-1)!} \sum_{v_1, \ldots, v_{r-1} \in C} \frac{a_{X_u, X_{v_1}, \ldots, X_{v_{r-1}}}}{\binom{n}{r-1}} \pm O(n^{-1})\right) \\ =\ & \exp\left(-\frac{1}{(r-1)!} \sum_{i_1, \ldots, i_{r-1} \in [q]} \frac{a_{X_u, i_1, \ldots, i_{r-1}}}{\binom{n}{r-1}} \cdot \prod_{j \in [r-1]} \left(\pi_{i_j} n \pm O(\alpha_n)\right) \pm O(n^{-1})\right) \\ =\ & \exp\left(-\sum_{i_1, \ldots, i_{r-1} \in [q]} a_{X_u, i_1, \ldots, i_{r-1}} \prod_{j \in [r-1]} \pi_{i_j} \pm O(\alpha_n n^{-1})\right) \\ =\ & \exp\left(-d_{X_u} \pm O(\alpha_n n^{-1})\right) \text{ a.a.s.} \end{aligned} \tag{5.29}$$

Combining Eq. (5.27) and Eq. (5.29) we get

$$\begin{aligned} Q_2 &= (1+o(1))\exp\left(-\sum_{u\in A} d_{X_u} \pm O(\alpha_n |A| n^{-1})\right) \\ &= (1\pm o(1))\exp\left(-\sum_{u\in A} d_{X_u}\right) =: (1\pm o(1))K(G,X_A) \text{ a.a.s.} \end{aligned} \tag{5.30}$$

By Eq. (5.26) and Eq. (5.30), we have

$$\mathbb{P}(G,X) = (1\pm o(1))\mathbb{P}(X)K(G,X_A)Q_3Q_4 \text{ a.a.s.} \tag{5.31}$$

Furthermore, for any $U = U(G) \subseteq V$ we have

$$\begin{aligned} \mathbb{P}(G,X_U) &= (1\pm o(1))\mathbb{P}(G,X_U, X\in\Omega) \\ &= (1\pm o(1))\sum_{Y\in\Omega_U} \mathbb{P}(Y)K(G,Y_A)Q_3(G,Y)Q_4(G,Y) \text{ a.a.s.} \end{aligned} \tag{5.32}$$

Therefore

$$\begin{aligned} \mathbb{P}(X_A|X_B,G) &= \frac{\mathbb{P}(X_{A\cup B},G)}{\mathbb{P}(X_B,G)} \\ &= (1\pm o(1))\frac{\sum_{Y\in\Omega_{A\cup B}} \mathbb{P}(Y)K(G,Y_A)Q_3(G,Y)Q_4(G,Y)}{\sum_{Y\in\Omega_B} \mathbb{P}(Y)K(G,Y_A)Q_3(G,Y)Q_4(G,Y)} \text{ a.a.s.} \end{aligned} \tag{5.33}$$

Note that $Q_3(G,Y)$ is a function of $(G,Y_{A\cup B})$ and $Q_4(G,Y)$ is a function of $(G,Y_{B\cup C})$. So the numerator of Eq. (5.33) is a.a.s. equal to

$$\mathbb{P}(X_A)\mathbb{P}(X_B)K(G,X_A)Q_3(G,X)\sum_{Y\in\Omega_{A\cup B}} \mathbb{P}(Y_C)Q_4(G,Y) \tag{5.34}$$

and the denominator of Eq. (5.33) is a.a.s. equal to

$$\mathbb{P}(X_B)\left(\sum_{Y\in\Omega_{B\cup C}} \mathbb{P}(Y_A)K(G,Y_A)Q_3(G,Y)\right)\left(\sum_{Y\in\Omega_{A\cup B}} \mathbb{P}(Y_C)Q_4(G,Y)\right). \tag{5.35}$$

Combining Eq. (5.33), Eq. (5.34), Eq. (5.35), we get

$$\mathbb{P}(X_A|X_B,G) = (1\pm o(1))\frac{\mathbb{P}(X_A)K(G,X_A)Q_3(G,X)}{\sum_{Y\in\Omega_{B\cup C}} \mathbb{P}(Y_A)K(G,Y_A)Q_3(G,Y)} \text{ a.a.s.} \tag{5.36}$$

Similarly, we have

$$\begin{aligned}
&\mathbb{P}(X_A|X_{B\cup C}, G) = \frac{\mathbb{P}(X, G)}{\mathbb{P}(X_{B\cup C}, G)} \\
&= (1 \pm o(1))\frac{\mathbb{P}(X)K(G, X_A)Q_3(G, X)Q_4(G, X)}{\sum_{Y\in\Omega_{B\cup C}} \mathbb{P}(Y)K(G, Y_A)Q_3(G, Y)Q_4(G, Y)} \\
&= (1 \pm o(1))\frac{\mathbb{P}(X_A)K(G, X_A)Q_3(G, X)}{\sum_{Y\in\Omega_{B\cup C}} \mathbb{P}(Y_A)K(G, Y_A)Q_3(G, Y)} \text{ a.a.s.}
\end{aligned} \tag{5.37}$$

Comparing Eq. (5.36) and Eq. (5.37) we finish the proof. □

5.2 Broadcasting on hypertrees

For the model HSBM$(n, q, r, \pi, \mathbf{A})$, we define a hypertree model which captures the local structure of the HSBM.

Definition 5.7 (Broadcasting on hypertrees)**.** Let $q \geq 2$ (number of communities), $r \geq 2$ (hyperedge size) be integers. Let $\pi \in \mathcal{P}([q])$ be a distribution with full support. Let $M : [q] \to [q]^{r-1}$ be a probability kernel, satisfying

$$M_{i,(i_1,\ldots,i_{r-1})} = M_{i,(i_{\sigma(1)},\ldots,i_{\sigma(r-1)})} \tag{5.38}$$

for all $i, i_1, \ldots, i_{r-1} \in [q]$, $\sigma \in \mathrm{Aut}([r-1])$, and

$$\sum_{k\in[q]} \pi_k \sum_{\substack{x\in[q]^{r-1} \\ x_i=j}} M_{k,(x_1,\ldots,x_{r-1})} = \pi_j \quad \forall i \in [r-1], j \in [q]. \tag{5.39}$$

Let T be a r-uniform linear[1] hypertree rooted at ρ. We define the broadcasting on hypertrees (BOHT) model BOHT(T, q, r, π, M) as follows.

1. Generate $\sigma_\rho \sim \pi$.

2. Suppose we have generated label σ_u for a vertex u. For each downward hyperedge $\{u, v_1, \ldots, v_{r-1}\}$, we generate $\sigma_{v_1}, \ldots, \sigma_{v_{r-1}}$ according to $M(\cdot|\sigma_u)$, i.e., for $i_1, \ldots, i_{r-1} \in [q]$, we have

$$\mathbb{P}[\sigma_{v_1} = i_1, \ldots, \sigma_{v_{r-1}} = i_{r-1}|\sigma_u = i] = M_{i,(i_1,\ldots,i_{r-1})}. \tag{5.40}$$

The output of the BOHT model is (T, σ).

Let D be a distribution on non-negative integers. If T is a random r-uniform linear hypertree, where every vertex independently have $b \sim D$ downward edges, then we denote the resulting model as BOHT(q, r, π, M, D). If D is a point distribution at $b \in \mathbb{Z}_{\geq 0}$, we say T is a b-regular r-uniform linear hypertree, and denote the resulting

[1]Linear means that the intersection of two distinct hyperedges has size at most one.

model as BOHT(q, r, π, M, b). If $D = \text{Pois}(d)$ for some $d \in \mathbb{R}_{\geq 0}$, we say T is a Galton-Watson r-uniform linear hypertree with expected offspring d, and denote the resulting model as BOHT$(q, r, \pi, M, \text{Pois}(d))$.

We define a few useful derived parameters for the BOHT model.

Definition 5.8 (Derived parameters for BOHT)**.** For BOHT(q, r, π, M, D), we define the expected offspring as

$$d = \mathbb{E}_{b\sim D} b. \tag{5.41}$$

For BOHT(T, q, r, π, M), the corresponding parameter should be

$$d = \text{br}(T), \tag{5.42}$$

where br denotes the branching number [92].

We define the signal matrix $\widetilde{Q} \in \mathbb{R}^{q\times q}$ as

$$\widetilde{Q}_{i,j} = \sum_{i_1,\ldots,i_{r-2}\in[q]} M_{i,(j,i_1,\ldots,i_{r-2})} \tag{5.43}$$

and define the signal strength λ as the second-largest eigenvalue (in absolute value) of matrix $\widetilde{Q}$. We define the signal-to-noise ratio (SNR) as

$$\text{SNR} := (r-1)d\lambda^2. \tag{5.44}$$

The Kesten-Stigum (KS) threshold is at SNR $= 1$.

We define the following subclasses of BOHT.

Definition 5.9 (Special cases of BOHT)**.** Consider a model BOHT(T, q, r, π, M) or BOHT(q, r, π, M, D).

- (Broadcasting on trees) If $r = 2$, then we call the model broadcasting on trees (BOT), denoted as BOT(T, q, π, M) or BOT(q, π, M, D).
- (Symmetric BOHT) We say the BOHT model is symmetric if $\pi = \text{Unif}([q])$ and the kernel M satisfies

  $$M_{i,(i_1,\ldots,i_r-1)} = M_{\tau(i),(\tau(i_1),\ldots,\tau(i_{r-1}))} \tag{5.45}$$

  for all $i_1, \ldots, i_r \in [q]$, $\tau \in \text{Aut}([q])$.
- (Simple BOHT) If $\pi = \text{Unif}([q])$ and for some $\lambda \in \left[-\frac{1}{q^{r-1}-1}, 1\right]$, we have

  $$M_{i,(i_1,\ldots,i_{r-1})} = \begin{cases} \lambda + q^{1-r}(1-\lambda), & \text{if } i = i_1 = \cdots = i_{r-1}, \\ q^{1-r}(1-\lambda), & \text{otherwise,} \end{cases} \tag{5.46}$$

  then we say the model is a simple BOHT model, denoted as BOHT(T, q, r, λ) or BOHT(q, r, λ, D).

- (Ising, Potts, random coloring) Consider a model $\mathrm{BOHT}(T, q, r, \lambda)$ or $\mathrm{BOHT}(q, r, \lambda, D)$. If $r = 2$, then we call the model the Potts model (when $q \geq 3$) or the Ising model (when $q = 2$), and denote the model as $\mathrm{BOT}(T, q, \lambda)$ or $\mathrm{BOT}(q, \lambda, D)$. The case $\lambda = -\frac{1}{q-1}$ is called the random coloring model.

- We denote a model $\mathrm{BOHT}(T, q, r, \lambda)$ (resp. $\mathrm{BOHT}(q, r, \lambda, D)$) with $q = 2$ as $\mathrm{BOHT}(T, r, \lambda)$ (resp. $\mathrm{BOHT}(r, \lambda, D)$).

For $\mathrm{BOHT}(T, q, r, \lambda)$, $\mathrm{BOHT}(q, r, \lambda, D)$, and their subclasses, we say the model is ferromagnetic if $\lambda > 0$, and is antiferromagnetic if $\lambda < 0$.

The following result establishes a relationship between HSBMs and BOHT models. This relationship was first shown in [96, 103] in the case of two-community symmetric SBMs, and later generalized to various settings [26, 31, 128, 129, 72, 37, 38, 109, 112, 130].

Theorem 5.10 (HSBM-BOHT coupling [130, Prop. 3])**.** *Let $(X, G) \sim \mathrm{HSBM}(n, q, r, \pi, \mathbf{A})$ be a model satisfying Condition 5.2. Let $v \in V$ and $k = c \log n$ for some small enough constant $c > 0$ not depending on n. Let $B(v, k)$ be the set of vertices with distance $\leq k$ to v.*

Let $(T, \sigma) \sim \mathrm{BOHT}(q, r, \pi, M, \mathrm{Pois}(d))$ where d is defined in Definition 5.3, and

$$M_{i,(i_1,\ldots,i_{r-1})} = \frac{1}{d} a_{i,i_1,\ldots,i_{r-1}} \prod_{j \in [r-1]} \pi_j. \tag{5.47}$$

Let ρ be the root of T, and T_k be the set of vertices at distance $\leq k$ to ρ.

Then $(G|_{B(v,k)}, X_{B(v,k)})$ can be coupled to (T_k, σ_{T_k}) with $o(1)$ TV distance.

In the setting of Theorem 5.10, we say the model $\mathrm{BOHT}(q, r, \pi, M, \mathrm{Pois}(d))$ is the BOHT model corresponding to $\mathrm{HSBM}(n, q, r, \pi, \mathbf{A})$. In the view of Theorem 5.10, we define the following natural condition on BOHT.

Condition 5.11. Consider the model $\mathrm{BOHT}(T, q, r, \pi, M)$ or $\mathrm{BOHT}(q, r, \pi, M, D)$. We say the model is reversible if

$$\pi_i M_{i,(j,i_1,\ldots,i_{r-2})} = \pi_j M_{j,(i,i_1,\ldots,i_{r-2})} \tag{5.48}$$

for all $i, j, i_1, \ldots, i_{r-2} \in [q]$.

One can easily verify that the BOHT model corresponding to an HSBM is always reversible.

For the BOHT model corresponding to an HSBM, parameters d, λ, SNR (Definition 5.3, 5.8) all agree for both models. Degree d agree because

$$\mathbb{E}_{b \sim \mathrm{Pois}(d)} b = d. \tag{5.49}$$

Signal strength λ agrees because

$$\begin{aligned}\widetilde{Q}_{i,j} &= \sum_{i_1,\ldots,i_{r-2}\in[q]} M_{i,(j,i_1,\ldots,i_{r-2})} \\ &= \sum_{i_1,\ldots,i_{r-2}\in[q]} \frac{1}{d} a_{i,j,i_1,\ldots,i_{r-2}} \pi_j \prod_{k\in[r-2]} \pi_k \\ &= \frac{1}{d} Q_{i,j}. \end{aligned} \tag{5.50}$$

SNR agrees because r, d, λ all agree.

We use the following notations for BOHT models.

Definition 5.12 (Notations for BOHT models)**.** Fix a BOHT model. Let T_k be the set of vertices at distance $\le k$ to ρ, and L_k be the set of vertices at distance k to ρ. Let M_k denote the channel $\sigma_\rho \to (T_k, \sigma_{L_k})$.

We define the following belief propagation (BP) operator, which is very useful in the study of the BOHT models.

Definition 5.13 (Belief propagation operator)**.** Consider the model $\mathrm{BOHT}(q, r, \pi, M, D)$. Let B denote the channel from σ_u to $\sigma_{v_1}, \ldots, \sigma_{v_{r-1}}$ in Definition 5.1. The belief propagation (BP) operator of the BOHT model is an operator from the space of information channels to itself, defined as

$$\mathrm{BP}(P) := \mathbb{E}_{b\sim D} \left(P^{\times(r-1)} \circ B\right)^{\star b}. \tag{5.51}$$

Note that the sequence $(M_k)_{k\ge 0}$ (Definition 5.12) satisfies

$$\mathrm{BP}(M_k) = M_{k+1}. \tag{5.52}$$

For a symmetric BOHT model, the BP operator sends the space of q-FMS channels to itself.

5.3 Weak recovery and reconstruction

In this section we show that for an HSBM, if the corresponding BOHT model admits non-reconstruction, then weak recovery for the HSBM is impossible.

Definition 5.14 (Reconstruction for BOHT model)**.** We say the model $\mathrm{BOHT}(T, q, r, \pi, M)$ or the model $\mathrm{BOHT}(q, r, \pi, M, D)$ admits reconstruction if

$$\lim_{k\to\infty} I(\sigma_\rho; T_k, \sigma_{L_k}) > 0. \tag{5.53}$$

We say the model admits non-reconstruction if the limit is zero.

The reconstruction problem for the $\mathrm{BOHT}(q, r, \pi, M, D)$ model can be interpreted using the BP operator: the model admits reconstruction if and only if the limit channel $\mathrm{BP}^\infty(\mathrm{Id}) := \lim_{k\to\infty} \mathrm{BP}^k(\mathrm{Id})$ is not the trivial channel.

Let us briefly discuss previous results on reconstruction for BOT and BOHT. [84] showed that for $\mathrm{BOT}(q, \pi, M, D)$, reconstruction is possible above the Kesten-Stigum threshold. [23] proved that for the Ising model $\mathrm{BOT}(2, \lambda, d)$ the reconstruction threshold coincides with the KS threshold (see also [81] for an alternative proof). [80, 63] proved the same result for general trees $\mathrm{BOT}(T, 2, \lambda)$ and $\mathrm{BOT}(2, \lambda, D)$, and [113] refined these results to criticality.

[101] proved that the reconstruction threshold does not match the KS threshold, for certain asymmetric Ising model $\mathrm{BOT}(2, \pi, M, d)$ and Potts model $\mathrm{BOT}(q, \lambda, d)$. [107] proved non-reconstruction results for the Potts model $\mathrm{BOT}(q, \lambda, d)$. [98] studied the reconstruction problem from a statistical physics point of view, and conjectured that for the Potts model $\mathrm{BOT}(q, \lambda, d)$, the KS threshold is tight when d is not too large, $q \leq 4$ (in the ferromagnetic regime $\lambda > 0$) or $q \leq 3$ (in the antiferromagnetic regime $\lambda < 0$). [126] proved that the KS threshold is tight for $\mathrm{BOT}(3, \lambda, d)$ with d large enough, and is not tight for $\mathrm{BOT}(q, \lambda, d)$ with $q \geq 5$, partially proving the conjectures of [98]. [109] proved that the KS threshold is tight for $\mathrm{BOT}(q, \lambda, D)$ for $q \leq 4$ and D satisfying mild assumptions (which allows both the regular case and the Poisson case).

[27] showed that the KS threshold is tight for $\mathrm{BOT}(2, \pi, M, d)$ when M is close enough to a binary symmetric channel. [91] determined the exact set (up to uncertainties on the boundary) of π for which the KS threshold is tight when degree d is large enough. [19] determined the reconstruction threshold for the hardcore model to the first order. [20] determined the reconstruction threshold for the random coloring model $\mathrm{BOT}(q, -\frac{1}{q-1}, d)$ to the first order, and [125] determined the threshold for $\mathrm{BOT}(q, -\frac{1}{q-1}, d)$ and $\mathrm{BOT}(q, -\frac{1}{q-1}, \mathrm{Pois}(d))$ to the third order. [59] determined the threshold for $\mathrm{BOT}(q, -\frac{1}{q-1}, D)$ for D satisfying mild conditions to the first order.

[87] gave non-reconstruction results for general $\mathrm{BOT}(q, \pi, M, D)$ via contraction of SKL information. Our work [72] (Chapter 6) gave non-reconstruction results for $\mathrm{BOT}(T, q, \pi, M)$ and $\mathrm{BOT}(q, \pi, M, D)$ via contraction of mutual information.

There are relatively fewer works on the reconstruction problems for BOHT. [112] showed that $\mathrm{BOHT}(r, \lambda, \mathrm{Pois}(d))$ admits reconstruction above the KS threshold. [130] showed that reversible $\mathrm{BOHT}(q, r, M, \mathrm{Pois}(d))$ admits reconstruction above the KS threshold. Both results are derived via relationship between weak recovery for HSBM and reconstruction for BOHT (Theorem 5.15). Our work [74] (Chapter 7) proved that the KS threshold is tight for $\mathrm{BOHT}(r, \lambda, D)$ for $r \leq 4$ (the case $r = 4$ depends on a numerically verified conjecture), and the KS threshold is not tight for $\mathrm{BOHT}(r, \lambda, d)$ and $\mathrm{BOHT}(r, \lambda, \mathrm{Pois}(d))$ for $r \geq 7$ and d large enough.

The following result was first established by [103] in the case of two-community symmetric SBMs, and later generalized to various settings [72, 109, 74]. Here we prove a general version which works for any $\mathrm{HSBM}(n, q, r, \pi, \mathbf{A})$ satisfying Condition 5.2.

Theorem 5.15 (Weak recovery for HSBM)**.** *Let* $\mathrm{HSBM}(n, q, r, \pi, \mathbf{A})$ *be a model satisfying Condition 5.2. Let* $\mathrm{BOHT}(q, r, \pi, M, \mathrm{Pois}(d))$ *be the corresponding BOHT model. If the BOHT model admits non-reconstruction, then weak recovery for the HSBM is impossible.*

Proof. For constant $k \in \mathbb{Z}_{\geq 0}$, by Theorem 5.10, for any fixed vertices $u \neq v \in V$, $\mathbb{P}[u \in B(v,k)] = o(1)$. Therefore

$$\begin{aligned} I(X_v; G, X_u) &\leq I(X_v; G, X_{\partial B(v,k)}, X_u) \\ &= I(X_v; G, X_{\partial B(v,k)}) + o(1) \\ &= I(\sigma_\rho; T_k, \sigma_{L_k}) + o(1), \end{aligned} \tag{5.54}$$

where the first step is by data processing inequality, the second step is by Prop. 5.6, and the third step is by Theorem 5.10. Taking limit $n \to \infty$, then taking limit $k \to \infty$, we get

$$\lim_{n\to\infty} I(X_v; G, X_u) \leq \lim_{k\to\infty} I(\sigma_\rho; T_k, \sigma_{L_k}). \tag{5.55}$$

RHS is zero by the assumption that the BOHT model admits non-reconstruction. By Pinsker's inequality, we have

$$I_{\mathrm{TV}}(X_v; G, X_u) \leq \sqrt{\frac{1}{2\log e} I(X_v; G, X_u)} = o(1). \tag{5.56}$$

In other words, for every $i, j \in [q]$, we have

$$\mathbb{P}(X_v = j | G, X_u = i) = \pi_j \pm o(1). \tag{5.57}$$

From this point we borrow an argument from [109]. Assume for the sake of contradiction that weak recovery is possible, i.e., there exists $S = S(G) \subseteq V$ such that for some $\epsilon > 0$, with probability $1-o(1)$, there exists $i, j \in [q]$ such that Eq. (5.12) holds. For simplicity of notation, we define $\Omega_i := \{u \in V : X_u = i\}$. Then we can write Eq. (5.12) as

$$\frac{|S \cap \Omega_i|}{|\Omega_i|} - \frac{|S \cap \Omega_j|}{|\Omega_j|} > \epsilon. \tag{5.58}$$

By concentration inequalities, we have

$$\mathbb{P}\left[\left|\frac{|\Omega_i|}{|V|} - \pi_i\right| < |V|^{-0.4} \forall i \in [q]\right] = 1 - o(1). \tag{5.59}$$

By Eq. (5.59) and Eq. (5.58), we have

$$\mathbb{P}\left[|S| \geq \epsilon_1 |V|\right] = 1 - o(1) \tag{5.60}$$

for some consant $\epsilon_1 > 0$. By Eq. (5.58), Eq. (5.60) and

$$\sum_{i\in[q]} \frac{|S \cap \Omega_i|}{|\Omega_i|} \cdot \frac{|\Omega_i|}{|V|} = \frac{|S|}{|V|}, \tag{5.61}$$

we see that for some $\epsilon_2 > 0$, with probability $1 - o(1)$, there exists $i \in [q]$ such that

$$\frac{|S \cap \Omega_i|}{|S|} - \pi_i > \epsilon_2. \tag{5.62}$$

Then

$$\begin{aligned}
&\mathbb{E}\left[\sum_{i\in[q]} \pi_i^{-1}|S\cap\Omega_i|^2 - |S|^2\right] \\
&= \mathbb{E}\left[\sum_{i\in[q]} \pi_i^{-1}\left(|S\cap\Omega_i| - \pi_i|S|\right)^2\right] \\
&\geq (\epsilon_2^2 - o(1))|V|^2.
\end{aligned} \tag{5.63}$$

Therefore

$$\begin{aligned}
\epsilon_2^2 - o(1) &\leq \frac{1}{|V|^2}\mathbb{E}\left[\sum_{i\in[q]} \pi_i^{-1}|S\cap\Omega_i|^2 - |S|^2\right] \\
&= \frac{1}{|V|^2}\sum_{u,v\in V}\mathbb{E}\left[\mathbb{1}\{u,v\in S\}\cdot\left(\sum_{i\in[q]}\pi_i^{-1}\mathbb{1}\{X_u = X_v = i\} - 1\right)\right] \\
&= \frac{1}{|V|^2}\sum_{u,v\in V}\mathbb{E}\left[\mathbb{1}\{u,v\in S\}\sum_{i\in[q]}\pi_i^{-1}\mathbb{E}\left[\mathbb{1}\{X_u = X_v = i\} - \pi_i^2|G\right]\right] \\
&= o(1),
\end{aligned}$$

which is contradiction. This finishes the proof. □

5.4 Mutual information and boundary irrelevance

In this section we show that for an HSBM, if the corresponding BOHT model admits a property called boundary irrelevance, then there is a formula for HSBM mutual information in terms of tree recursions.

The HSBM mutual information is a very natural quantity showing the amount of information about the community structure contained in the unlabeled hypergraph.

Definition 5.16 (HSBM mutual information)**.** Let $(X, G) \sim \mathrm{HSBM}(n, q, r, \pi, \mathbf{A})$. We define the HSBM mutual information as

$$\lim_{n\to\infty}\frac{1}{n}I(X;G). \tag{5.64}$$

The HSBM mutual information problem is closely related to the boundary irrelevance property for BOHT.

Definition 5.17 (Boundary irrelevance for BOHT)**.** Consider the model $\mathrm{BOHT}(T, q, r, \pi, M)$ or the model $\mathrm{BOHT}(q, r, \pi, M, D)$. Let W be a channel with input alphabet $[q]$ (called the survey channel). Let $\omega_u \sim W(\cdot|\sigma_u)$ independently for all vertices $u \in V(T)$. We say the BOHT model admits boundary irrelevance (BI) with respect to W if

$$\lim_{k\to\infty} I(\sigma_\rho; \sigma_{L_k}|T_k, \omega_{T_k}) > 0. \tag{5.65}$$

We say the BOHT model admits boundary irrelevance (BI) if it admits BI with respect to all erasure channels EC_ϵ with $\epsilon < 1$, where ϵ denotes erasure probability.

For the model $\mathrm{BOHT}(q, r, \pi, M, D)$, boundary irrelevance can be interpreted using the BP operator. For a fixed survey channel W, we define the BP_W operator as

$$\mathrm{BP}_W(P) = \mathrm{BP}(P) \star W. \tag{5.66}$$

Then the model admits BI with respect to W if and only if the BP_W operator has a unique fixed point.

The SBM mutual information problem has been studied by several works. [43] gave a mutual information formula for the disassortative simple SBM ($\mathrm{SBM}(n, q, a, b)$ with $a < b$). [54] conjectured a formula for two-community symmetric SBM $\mathrm{SBM}(n, 2, a, b)$, and proved that their formula matches the one of [43] when $a < b$, and is an lower bound of the mutual information when $a > b$. Our work [4] (Chapter 10) gave a mutual information formula for $\mathrm{SBM}(n, 2, a, b)$ when SNR is outside a finite interval $[1, 3.513]$. [137] improved the result and gave a mutual information formula for any $\mathrm{SBM}(n, 2, a, b)$. It is open whether the formula conjectured in [54] is equivalent to the one proved in [4, 137]. Our work [73] (Chapter 11) generalized the previous work and gave a mutual information formula for $\mathrm{SBM}(n, q, a, b)$ when $\mathrm{SNR} > 1 + C\max\{\lambda, q^{-1}\}\log q$ or $\mathrm{SNR} < q^{-2}$. To our knowledge, there has been no work studying the mutual information problem for models other than the simple SBM (Definition 5.4).

Results of [4, 137, 73] are based on boundary irrelevance. Surprisingly, our work [73] (Chapter 8) showed that boundary irrelevance for $\mathrm{BOT}(q, \lambda, d)$ and $\mathrm{BOT}(q, \lambda, \mathrm{Pois}(d))$ does not hold between the reconstruction threshold and the Kesten-Stigum threshold, which is known to exist when $q \geq 4$.

The following result was first established in our work [4] in the case of two-community symmetric SBMs, and generalized in our work [73] to q-community symmetric SBMs. Here we prove a general version which works for any $\mathrm{HSBM}(n, q, r, \pi, \mathbf{A})$ satisfying Condition 5.2.

Theorem 5.18 (HSBM mutual information)**.** *Let $(X, G) \sim \mathrm{HSBM}(n, q, r, \pi, \mathbf{A})$ be a model satisfying Condition 5.2. Let $\mathrm{BOHT}(q, r, \pi, M, \mathrm{Pois}(d))$ be the corresponding BOHT model. If the BOHT model admits boundary irrelevance, then*

$$\lim_{n\to\infty} \frac{1}{n} I(X; G) = \int_0^1 \lim_{k\to\infty} I(\sigma_\rho; \omega^\epsilon_{T_k\setminus\rho}|T_k) d\epsilon, \tag{5.67}$$

where ω^ϵ denotes observation through survey channel EC_ϵ.

Proof. Let $Y_v^\epsilon \sim \mathrm{EC}_\epsilon(\cdot|X_v)$ for $v \in V$ and $\epsilon \in [0,1]$. Let $u \in V$ be a fixed vertex. Define $f(\epsilon) := \frac{1}{n} I(X; G, Y^\epsilon)$. Then $f(0) = H(X_u)$ and $f(1) = \frac{1}{n} I(X; G)$. Furthermore, calculation shows that

$$f'(\epsilon) = -H(X_u | G, Y^\epsilon_{V\setminus u}). \tag{5.68}$$

Let $k \in \mathbb{Z}_{\geq 1}$ be a constant, $B(u,k)$ be the set of vertices with distance $\leq k$ to u, and $\partial B(u,k)$ be the set of vertices at distance k to u. By the data processing inequality and Prop. 5.6, we have

$$I(X_u; G, Y^\epsilon_{B(u,k)\setminus u}) \leq I(X_u; G, Y^\epsilon_{V\setminus u}) \leq I(X_u; G, Y^\epsilon_{B(u,k)\setminus u}, X_{\partial B(u,k)}) + o(1). \tag{5.69}$$

By Theorem 5.10, we have

$$I(\sigma_\rho; \omega^\epsilon_{T_k\setminus\rho} | T_k) - o(1) \leq I(X_u; G, Y^\epsilon_{V\setminus u}) \leq I(\sigma_\rho; \omega^\epsilon_{T_k\setminus\rho}, \sigma_{L_k} | T_k) + o(1). \tag{5.70}$$

Taking limit $n \to \infty$, then taking limit $k \to \infty$, we get

$$\lim_{k\to\infty} I(\sigma_\rho; \omega^\epsilon_{T_k\setminus\rho} | T_k) \leq \lim_{n\to\infty} I(X_u; G, Y^\epsilon_{V\setminus u}) \leq \lim_{k\to\infty} I(\sigma_\rho; \omega^\epsilon_{T_k\setminus\rho}, \sigma_{L_k} | T_k). \tag{5.71}$$

The first and third terms are equal by the boundary irrelevance assumption. Therefore

$$\lim_{n\to\infty} I(X_u; G, Y^\epsilon_{V\setminus u}) = \lim_{k\to\infty} I(\sigma_\rho; \omega^\epsilon_{T_k\setminus\rho} | T_k) \tag{5.72}$$

So

$$\begin{aligned} \lim_{n\to\infty} \frac{1}{n} I(X; G) &= H(X_u) - \int_0^1 \lim_{n\to\infty} H(X_u | G, Y^\epsilon_{V\setminus u}) d\epsilon \\ &= \int_0^1 \lim_{n\to\infty} I(X_u; G, Y^\epsilon_{V\setminus u}) d\epsilon \\ &= \int_0^1 \lim_{k\to\infty} I(\sigma_\rho; \omega^\epsilon_{T_k\setminus\rho} | T_k) d\epsilon. \end{aligned} \tag{5.73}$$

□

5.5 Optimal recovery and uniqueness of BP fixed point

In this section we consider the optimal recovery for HSBM, potentially with survey observations. We show that properties of the corresponding BOHT models can lead to optimal recovery algorithms for the HSBM.

Definition 5.19 (Optimal recovery for HSBM)**.** Consider the model HSBM$(n, q, r, \pi, \mathbf{A})$.

The optimal recovery problem asks to determine the optimal recovery accuracy

$$\sup_{\widehat{X}=\widehat{X}(G)} \lim_{n\to\infty} \mathbb{E}\left[1-\frac{1}{n}d_H(\widehat{X},X)\right], \tag{5.74}$$

where $\widehat{X}=\widehat{X}(G)$ goes over estimators with input G.

Let W be a channel with input alphabet $[q]$ (called the survey channel). Let $Y_u \sim W(\cdot|X_u)$ independently for all vertices $u \in V(G)$. The optimal recovery with survey problem asks to determine the optimal recovery accuracy

$$\sup_{\widehat{X}=\widehat{X}(G,Y)} \lim_{n\to\infty} \mathbb{E}\left[1-\frac{1}{n}d(\widehat{X},X)\right], \tag{5.75}$$

where $d(X,Y) := \sum_{u\in V} \mathbb{1}\{X_i \neq Y_i\}$ and $\widehat{X}=\widehat{X}(G)$ goes over estimators with input G and Y.

In the optimal recovery with survey problem, we do not need to use d_H distance because the surveys break the symmetry between communities.

[49] conjectured that belief propagation (BP) with random initialization is optimal for any SBM$(n,q,\pi,\mathbf{A})$. This conjecture is still open, even in the case of SBM$(n,2,a,b)$. Nevertheless, some progress has been made. [104] gave an algorithm for SBM$(n,2,a,b)$ and proved its optimality when SNR is larger than a constant ([4] estimated the constant to be at least 75). [4] proved that [104]'s algorithm is optimal when SNR is larger than 3.513. [137] proved that [104]'s algorithm is optimal for any SBM$(n,2,a,b)$. All these proofs are based on the uniqueness of BP fixed point.

For the q-community simple SBM (SBM(n,q,a,b)), [37] proved optimal recovery algorithms when SNR $> C_q$ for some constant C_q depending only on q. They did not give an estimate for C_q but from the proof, it is at least polynomial in q. They used a local version of uniqueness of BP fixed point, stating that the BP recursion converges to the same fixed point as long as the initial channel is close enough to the Id channel. Our work [73] proved optimal recovery algorithms when SNR $> 1+C\max\{\lambda,q^{-1}\}\log q$ for some absolute constant C.

[38] generalized [37] and proved optimal recovery algorithms for certain SBM$(n,q,\pi,\mathbf{A})$ with SNR large enough. This is the only work on optimal recovery for non-symmetric SBM. To our knowledge, there has been no work studying the optimal recovery problem for HSBM with $r \geq 3$.

[83] studied the weak recovery problem for SBM(n,q,a,b) with EC$_\epsilon$ survey. [110] proved that the local belief propagation algorithm is optimal for SBM$(n,2,a,b)$ with BEC or BSC survey, in the same parameter ranges as [104]. [4] and [137] proved optimality of the local BP algorithm for SBM$(n,2,a,b)$ with survey in the same parameter range as the results for SBM without survey.

We first prove the reduction for optimal recovery with survey. The reduction was first established in [110] in the case of two-community symmetric SBMs with BSC or BEC survey, and generalized in our work [73] to q-community symmetric SBMs. Here we prove a general version which works for any HSBM$(n,q,r,\pi,\mathbf{A})$ satisfying

Condition 5.2.

Theorem 5.20 (Optimal recovery for HSBM with survey)**.** *Let* $(X, G) \sim \mathrm{HSBM}(n, q, r, \pi, \mathbf{A})$ *be a model satisfying Condition 5.2. Let* W *be a survey channel. Let* $\mathrm{BOHT}(q, r, \pi, M, \mathrm{Pois}(d))$ *be the corresponding BOHT model. If the BOHT model admits boundary irrelevance with respect to* W*, then for the HSBM with survey, the belief propagation (Algorithm 1) achieves optimal recovery accuracy of*

$$1 - \lim_{k\to\infty} P_e(\sigma_\rho | T_k, \omega_{T_k}), \tag{5.76}$$

where ω *denotes observation through survey channel.*

Algorithm 1 Belief propagation algorithm for HSBM with survey

Input: HSBM hypergraph $G = (V, E)$, survey $Y \in \mathcal{Y}^V$
Output: $\widehat{X} \in [q]^V$
$(V, \widetilde{E}) \leftarrow$ underlying graph of G, i.e., $(u, v) \in \widetilde{E}$ if and only if $u \neq v$ and there exists a hyperedge $e \in E$ containing both u and v
$m^{(0)}_{u\to v} \leftarrow$ posterior distribution of X_u conditioned on Y_u $\forall (u, v) \in \widetilde{E}$
$r \leftarrow \lfloor \log^{0.9} n \rfloor$
for $t = 0 \to r - 1$ **do**
    for $(u, v) \in \widetilde{E}$ **do**
        $m^{(t+1)}_{u\to v} \leftarrow$ posterior distribution of X_u conditioned on Y_u and $m^{(t)}_{w\to u}$ for all $(w, u) \in \widetilde{E}$, $w \neq v$ ▷ Computation of the posterior distribution uses the hypergraph structure
    end for
end for
for $u \in V$ **do**
    $m_u \leftarrow$ posterior distribution of X_u conditioned on Y_u and $m_{w\to u}$ for $(w, u) \in \widetilde{E}$.
    $\widehat{X}_u \leftarrow \arg\max_{i\in[q]} m_u(i)$
end for
return $\widehat{X}$

Proof. We run Algorithm 1. Let $\rho \in V$ be a fixed vertex. For $k \in \mathbb{Z}_{\geq 1}$, define $B(\rho, k)$, $\partial B(\rho, k)$ as in Theorem 5.10. By Theorem 5.10 and induction on t, we see that $m^{(t)}_{u\to v}$ has the same distribution (up to $o(1)$ TV distance) as the posterior distribution of σ_ρ conditioned ω_{T_t}. Therefore $m^{(r)}_{u\to v}$ has the same distribution (up to $o(1)$ TV distance) as the posterior distribution of σ_ρ conditioned ω_{T_r}. So as $n \to \infty$, Algorithm 1 achieves accuracy

$$1 - \lim_{k\to\infty} P_e(\sigma_\rho | T_k, \omega_{T_k}). \tag{5.77}$$

On the other hand, we have

$$\begin{aligned} P_e(X_\rho|G,Y) &\geq P_e(X_\rho|G,Y,X_{\partial B(\rho,k)}) \\ &= P_e(X_\rho|G,Y_{B(\rho,k)},X_{\partial B(\rho,k)}) \pm o(1) \\ &\geq P_e(\sigma_\rho|T_k,\omega_{T_k},\sigma_{L_k}) - o(1). \\ &= P_e(\sigma_\rho|T_k,\omega_{T_k}) - o(1). \end{aligned} \tag{5.78}$$

where the first step is by data processing inequality, the second step is by Prop. 5.6, the third step is by Theorem 5.10, the fourth step is by boundary irrelevance with respect to W. Taking limit $n \to \infty$, then $k \to \infty$, we see that

$$P_e(X_\rho|G,Y) \geq \lim_{k\to\infty} P_e(\sigma_\rho|T_k,\omega_{T_k}). \tag{5.79}$$

This shows that Algorithm 1 is optimal. □

Definition 5.21 (Uniqueness of BP fixed point for BOHT)**.** Consider a symmetric BOHT model $\mathrm{BOHT}(q,r,\pi,M,D)$. We say the model admits uniqueness of BP fixed point if the corresponding BP operator (Definition 5.13) has only one non-trivial fixed point in the space of q-FMS channels.

We say the model admits (global) stability of BP fixed point if it admits uniqueness of BP fixed point, and starting from any non-trivial q-FMS channel P, the BP recursion converges to the unique non-trivial fixed point.

For HSBM without survey, optimal recovery is more difficult. The reason is that one needs a sufficiently good initial estimator to start the belief propagation. In particular, an estimator as in the definition of weak recovery (Definition 5.5) seems insufficient for the purpose of optimal recovery.

The reduction was first established in [104] in the case of two-community symmetric SBMs. [37] proved a version for q-community symmetric SBMs with strong assumptions on the initial algorithm. Our work [73] proved the case of q-community symmetric SBMs which strictly generalizes the one in [104]. Here we prove a general version which works for symmetric HSBMs.

Theorem 5.22 (Optimal recovery for symmetric HSBM)**.** *Let* $(X,G) \sim \mathrm{HSBM}(n,q,r,\pi = \mathrm{Unif}([q]),\mathbf{A})$ *be a symmetric HSBM with non-zero signal strength (Definition 5.3). Let* $\mathrm{BOHT}(q,r,\pi = \mathrm{Unif}([q]),M,\mathrm{Pois}(d))$ *be the corresponding BOHT model.*

Suppose there is an algorithm $\mathcal{A}$ *and a constant* $\epsilon > 0$ *(not depending on* n*) such that with probability* $1 - o(1)$*, the empirical transition matrix* $F \in \mathbb{R}^{q\times q}$ *defined as*

$$F_{i,j} := \frac{\#\{v \in V : X_v = i, \widehat{X}_v = j\}}{\#\{v \in V : X_v = i\}}, \qquad \widehat{X} := \mathcal{A}(G) \tag{5.80}$$

satisfies

(1) $\|F^\top \mathbb{1} - \mathbb{1}\|_\infty = o(1)$*;*

(2) $\sigma_{\min}(F) > \epsilon$*, where* $\sigma_{\min}$ *is the smallest singular value;*

(3) there exists a permutation $\tau \in \mathrm{Aut}([q])$ such that $F_{\tau(i),i} > F_{\tau(i),j} + \epsilon$ for all $i \neq j \in [q]$.

(Note that we do not assume F stays the same for different calls to $\mathcal{A}$.)

If the BOHT model admits stability of BP fixed point, then there is an algorithm (Algorithm 2) achieving the optimal recovery accuracy of

$$1 - \lim_{k\to\infty} P_e(\sigma_\rho | T_k, \sigma_{L_k}). \tag{5.81}$$

The theoretical guarantees of the initial recovery algorithm provided by previous works (that achieve the KS threshold) [9, 128] do not seem to be enough for our purpose. There are several different ways to formulate the initial point requirement. The one we state here is weaker than [37] (which required the initial point to be close enough to Id, which seems unlikely to hold near the KS threshold), and a generalization of the requirement in the $q = 2$ case used by [104]. Our initial point requirement seems more likely to hold near the KS threshold. For example, it is plausible that a balanced algorithm would achieve the empirical transition matrix F to be close to P_λ for some $|\lambda| = \Omega(1)$.

We remark that for the case $q = 2$, any algorithm that works for weak recovery can be made into an algorithm that satisfies the conditions in Theorem 5.22. Recall an algorithm for weak recovery outputs a subset S satisfying Eq. (5.12). Then for some $\epsilon' > 0$ not depending on n, with high probability we have $\epsilon' n < |S| < (1 - \epsilon')n$. If we randomly insert (if $|S| < n/2$) or delete (if $|S| > n/2$) elements of S, then in the end we can get a set of size $n/2$. This gives an estimator which always outputs a set S of size $n/2$, and satisfies Eq. (5.12) with a possibility smaller, but still constant ϵ. Then the empirical transition matrix F defined in Eq. (5.80) is within $o(1)$ distance to a non-trivial BSC channel, thus satisfies all three conditions in Theorem 5.22.

Proof of Theorem 5.22. We run Algorithm 2. The proof is a variation of the proof in [104].

We first note that because our HSBM is symmetric, the signal matrix Q is a multiple of a Potts matrix P_λ. Therefore the signal matrix $\widetilde{Q}$ of the BOHT model is a Potts channel. Note that $Q_{i,j}$ denotes the expected number of neighbors with label j for a vertex with label i.

Choice of u_i. For every $i \in [q]$, the set $\{u \in U : X_u = i\}$ has size $\frac{n}{q} \pm o(n)$. Therefore with high probability, there exists $u \in U$ with $X_u = i$ that satisfies (a). Furthermore, because Y is independent of U, we can equivalently first generate the hypergraph $G\backslash U$, then compute Y, then generate the edges adjacent to U. In this way, we see that with high probability, for all $u \in U$ satisfying (a), the empirical distribution of $\{Y_v : v \in V, (u, v) \in \widetilde{E}\}$ has $o(1)$ total variation distance to $P_\lambda F$. By assumption (1)(3) in Theorem 5.22, we have $s(P_\lambda F)_{i,\tau(i)} > s(P_\lambda F)_{i,\tau(j)} + |\lambda|\epsilon$ for $i \in [q]$, $j \in [q]\backslash i$. Therefore with high probability, for all $u \in U$ satisfying (a), we can identify X_u up to a permutation $\tau \in \mathrm{Aut}([q])$ by computing $\arg\max_{j\in[q]} sN_Y(u, j)$. Therefore with high probability we are able to choose the u_is in Line 10.

Alignment of Y with Y^v. The above discussion still holds with Y replaced by Y^v. One thing to note is that by Theorem 5.10, $|B(v, r - 1)| = n^{o(1)}$ with high

Algorithm 2 Belief propagation algorithm for symmetric HSBM

Input: HSBM hypergraph $G = (V, E)$, initial recovery algorithm $\mathcal{A}$
Output: $\widehat{X} \in [q]^V$
$(V, \widetilde{E}) \leftarrow$ underlying graph of G, i.e., $(u, v) \in \widetilde{E}$ if and only if $u \neq v$ and there exists a hyperedge $e \in E$ containing both u and v
$Q \leftarrow$ signal matrix of the model (Definition 5.3)
$s \leftarrow 1$ if $Q_{1,1} > Q_{1,2}$; $s \leftarrow -1$ if $Q_{1,1} < Q_{1,2}$
$r \leftarrow \lfloor \log^{0.9} n \rfloor$
$U \leftarrow$ random subset of V of size $\lfloor \sqrt{n} \rfloor$
$Y \leftarrow \mathcal{A}(G \backslash U)$
For $i \in [q]$, $u \in U$, compute $\widetilde{N}_Y(u, i) \leftarrow \#\{Y_v = i : v \in V, (u, v) \in \widetilde{E}\}$
For $i \in [q]$, choose $u_i \in U$ such that
  (a) u_i has at least $\sqrt{\log n}$ neighbors in $V \backslash U$, and
  (b) $s\widetilde{N}_Y(u_i, i) > s\widetilde{N}_Y(u_i, j)$ for $j \in [q] \backslash i$.
for $v \in V \backslash U$ **do**
  $Y^v \leftarrow \mathcal{A}(G \backslash B(v, r-1) \backslash U)$
  Relabel Y^v by performing a permutation $\tau \in \mathrm{Aut}([q])$, so that $s\widetilde{N}_{Y^v}(u_i, i) > s\widetilde{N}_{Y^v}(u_i, j)$ for $i \in [q]$, $j \in [q] \backslash i$. Permute randomly if this cannot be achieved.
  Define empirical transition matrix $M^v : [q] \to [q]^{r-1}$ as

$$M^v_{i,(i_1,\ldots,i_{r-1})} \leftarrow \frac{N_{Y^v}(u_i, (i_1, \ldots, i_{r-1}))}{\sum_{j_1,\ldots,j_{r-1} \in [q]} N_{Y^v}(u_i, (j_1, \ldots, j_{r-1}))}, \tag{5.82}$$

$$\text{where } N_{Y^v}(u, (j_1, \ldots, j_{r-1})) \leftarrow \frac{\sum_{\substack{(u, w_1, \ldots, w_{r-1}) \in E \\ \sigma \in \mathrm{Aut}([r-1])}} \mathbb{1}\{Y^v_{w_k} = j_{\sigma(k)} \forall k \in [r-1]\}}{\sum_{\sigma \in \mathrm{Aut}([r-1])} \mathbb{1}\{j_k = j_{\sigma(k)} \forall k \in [r-1]\}} \tag{5.83}$$

  Run belief propagation on $B(v, r-1)$ with boundary condition $Y^v_{\partial B(v,r)}$, assuming the channel from $\partial B(v, r-1)$ to $\partial B(v, r)$ is M^v
  $\widehat{X}_v \leftarrow$ maximum likelihood label according to belief propagation
end for
$\widehat{X}_v \leftarrow 1$ for all $v \in U$
return $\widehat{X}$

probability. So removing $B(v, r-1)$ from G has negligible influence to the the empirical distribution of labels of neighbors of u_is. Therefore, with high probability, we are able to permute the labels Y^v so that the empirical distributions align with that of Y. (Note that we do not assume the empirical distributions for Y and Y^v are the same; we only use that they both satisfy condition (3).) Furthermore, we can compute the transition matrix

$$M^v = (F^v)^{\times(r-1)} \circ M \pm o(1). \tag{5.84}$$

Boundary condition of BP. Because Y^v is independent of edges between $\partial B(v, r-1)$ and $\partial B(v, r)$, we can equivalently first generate the hypergraph $G \backslash B(v, r-1) \backslash U$, then compute Y^v, then generate $E(\partial B(v, r-1), \partial B(v, r))$. In this way, it is clear that $Y^v_{w_1,\ldots,w_{r-1}}$ for one $(u, w_1, \ldots, w_{r-1}) \in E(\partial B(v, r-1), \partial B(v, r))$ is equivalent to one observation of X_u through channel M^v.

Property of M^v. Note that $M^v \geq_{\deg} P_{\lambda - o(1)} F^v$ by splitting hyperedges into individual edges. By Lemma 5.23 and condition (2), we have $F_v \geq_{\deg} P_{\lambda'}$ for some constant $\lambda' > 0$ not depending on n. Therefore $M^v \geq_{\deg} P_{\lambda''}$ for some $\lambda'' > 0$ not depending on n.

Convergence of BP recursion. Because $\lambda'' > 0$ is a constant, by the stability of BP fixed point assumption, for any $\kappa > 0$, there exists some integer k_0 not depending on n such that

$$P_e(\mathrm{BP}^{k_0}(M^v)) \geq P_e(\mathrm{BP}^{k_0}(P_{\lambda''})) > \lim_{k \to \infty} P_e(\mathrm{BP}^k(\mathrm{Id})) - \kappa. \tag{5.85}$$

Because $r = \omega(1)$, belief propagation in Line 15 converges to $o(1)$ in TV distance to the fixed point. Therefore we achieve desired probability of error in Line 16. □

Lemma 5.23. *Fix $q \in \mathbb{Z}_{\geq 2}$ and $\epsilon > 0$. Then there exists $\lambda > 0$ such that for any probability kernel $U : [q] \to [q]$ with $\sigma_{\min}(U) > \epsilon$, we have $P_\lambda \leq_{\deg} U$.*

Proof. Because $\sigma_{\min}(U) > \epsilon$, we have

$$\max_{i,j \in [q]} |\left(U^{-1}\right)_{i,j}| \leq \|U^{-1}\|_2 \leq \sqrt{q}\epsilon^{-1}. \tag{5.86}$$

Let $J \in \mathbb{R}^{q \times q}$ be the all ones matrix. Because U is a stochastic matrix, we have $U^{-1}J = J$. Because the maximum (in absolute value) entry of U^{-1} is bounded by an constant, for some constant $\lambda > 0$ the matrix $U^{-1}P_\lambda$ has non-negative entries. Note that $U^{-1}P_\lambda \mathbb{1} = U^{-1}\mathbb{1} = \mathbb{1}$. So $U^{-1}P_\lambda$ is a stochastic matrix. Let $R = U^{-1}P_\lambda$. Then $R \circ U = UR = P_\lambda$, thus $P_\lambda \leq_{\deg} U$. □

Chapter 6

Reconstruction for broadcasting on trees

We establish a simple method for proving non-reconstruction results for broadcasting on trees (BOT). The method is via input-restricted KL contraction coefficients. Combined with computations of input-restricted KL contraction coefficients for the Potts channels (Chapter 4), this method gives very good non-reconstruction results in the finite d regime. A special case of the Potts model is the problem of reconstructing color of the root of a q-colored tree given knowledge of colors of all the leaves. We show that to have a non-trivial reconstruction probability the branching number of the tree should be at least

$$\frac{\log q}{\log q - \log(q-1)} = (1 - o(1))q \log q. \tag{6.1}$$

This recovers previous results of [125, 20] in (slightly) more generality, but more importantly avoids the need for any coloring-specific arguments. Combined with the reduction established in Chapter 5, we improve the state-of-the-art on the weak recovery threshold for the q-community symmetric stochastic block model (q-SBM), for all $q \geq 3$. To further show the power of our method, we prove optimal non-reconstruction results for a broadcasting on trees model with Gaussian kernels, closing a gap left open by [60]. This chapter is based on [72].

Chapter outline In Section 6.1, we present our method for proving non-reconstruction results for the BOT model. In Section 6.2, we apply our method to several examples and compare with previous results. In Section 6.3, we derive impossibility of weak recovery results for the q-SBM. In Section 6.4, we apply our method to a BOT model with continuous alphabet previously studied by [60], and establish the reconstruction threshold for this model.

6.1 Non-reconstruction for broadcasting on trees

In this section we prove a non-reconstruction result for a general class of BOT models, using input-restricted KL contraction coefficients. We recall the definition of the BOT model (Definition 5.9).

Fix an integer $q \in \mathbb{Z}_{\geq 2}$, a distribution $\pi \in \mathcal{P}([q])$ with full support, and a channel $M : [q] \to [q]$ satisfying $\pi M = \pi$ (recall Eq. (5.39)). Let T be a possibly infinite tree with a marked root ρ. We generate a label $\sigma_v \in [q]$ for every vertex $v \in T$ as follows.

1. Generate $\sigma_\rho \sim \pi$.

2. Suppose we have generated a label for a vertex u. For every child v of u, we generate σ_v according to

$$\mathbb{P}(\sigma_v = j | \sigma_u = i) = M_{i,j}. \tag{6.2}$$

Let L_k denote the set of vertices at distance k to ρ. We say the model admits reconstruction (Definition 5.14) if

$$\lim_{k \to \infty} I(\sigma_\rho; \sigma_{L_k}) > 0, \tag{6.3}$$

and the model admits non-reconstructing if the limit is equal to zero. For any vertex u, let $c(u)$ denote the set of children of u.

We recall the definition of the branching number $\mathrm{br}(T)$ of a tree T.

Definition 6.1 (Branching number [92])**.** Let T be a possibly infinite tree rooted at ρ. Define a flow to be a function $f : V(T) \to \mathbb{R}_{\geq 0}$ such that for every vertex u, we have

$$f_u = \sum_{v \in c(u)} f_v. \tag{6.4}$$

Define $\mathrm{br}(T)$ to be the sup of all numbers λ such that there exists a flow f with $f_\rho > 0$, and $f_u \leq \lambda^{-d(u,\rho)}$ for all vertices u, where $d(u, \rho)$ is the distance between u and ρ.

We are now ready to state the main theorem of this chapter.

Theorem 6.2 (Non-reconstruction for BOT)**.** *The model* $\mathrm{BOT}(T, q, \pi, M)$ *(Definition 5.9) admits non-reconstruction if*

$$\eta_{\mathrm{KL}}(\pi, M^*) \, \mathrm{br}(T) < 1, \tag{6.5}$$

where $\mathrm{br}(T)$ *is the branching number of* T *(Definition 6.1),* M^* *denotes the reverse channel of* M *with respect to* π*, and* η_{KL} *denotes the KL contraction coefficient (Definition 2.4).*

Proof. For any vertex u, let $L_{u,k}$ denote the set of descendants of u at distance k to ρ. Define

$$a_u = H(\pi)^{-1}\eta_{\mathrm{KL}}(\pi, M^*)^{d(u,\rho)} \lim_{k\to\infty} I(\sigma_u; \sigma_{L_{u,k}}). \tag{6.6}$$

By DPI, $I(\sigma_u; \sigma_{L_{u,k}})$ is non-increasing for $k \geq d(u,\rho)$, so the limit exists.

For any $v \in c(u)$, consider the Markov chain

$$\sigma_{L_{v,k}} \to \sigma_v \xrightarrow{M^*} \sigma_u. \tag{6.7}$$

Because π is an invariant distribution, the distributions of σ_v and σ_u are both π. By SDPI, we have

$$I(\sigma_u; \sigma_{L_{v,k}}) \leq \eta_{\mathrm{KL}}(\pi, M^*) I(\sigma_v; \sigma_{L_{v,k}}). \tag{6.8}$$

Because $(\sigma_{L_{v,k}})_{v\in c(u)}$ are independent conditioned on σ_u, we have

$$I(\sigma_u; \sigma_{L_{u,k}}) \leq \sum_{v\in c(u)} I(\sigma_u; \sigma_{L_{v,k}}). \tag{6.9}$$

Combine the two inequalities and let $k \to \infty$. We get that

$$a_u \leq \sum_{v\in c(u)} a_v. \tag{6.10}$$

Clearly,

$$a_u \leq \eta_{\mathrm{KL}}(\pi, M^*)^{d(u,\rho)} \tag{6.11}$$

for all vertices u. However, a is not quite a flow yet. We define a flow b from a. For a vertex u, let $u_0 = \rho, \ldots, u_\ell = u$ be the shortest path from ρ to u. Define

$$b_u = a_u \prod_{0\leq j\leq \ell-1} \frac{a_{u_j}}{\sum_{v\in c(u_j)} a_v}. \tag{6.12}$$

(If $\sum_{v\in c(u_j)} a_v = 0$ for some j, then let $b_u = 0$.) It is not hard to check that

$$b_u = \sum_{v\in c(u)} b_v, \tag{6.13}$$

and that

$$b_u \leq a_u \leq \eta_{\mathrm{KL}}(\pi, M^*)^{d(u,\rho)}. \tag{6.14}$$

By definition of branching number, we must have $b_\rho = 0$. This means

$$\lim_{k\to\infty} I(\sigma_\rho; \sigma_{L_k}) = 0, \tag{6.15}$$

and non-reconstruction holds. □

In the definition of the weak recovery problem, it is not necessary to require σ_ρ to have distribution π. Let $\sigma^i_{L_k}$ denote the leaf labels conditioned on $\sigma_\rho = i$. Then Theorem 6.1 implies that when $\eta_{\mathrm{KL}}(\pi, M^*)\,\mathrm{br}(T) < 1$, we have

$$\lim_{k\to\infty} \mathrm{TV}(\sigma^i_{L_k}, \sigma^j_{L_k}) = 0, \tag{6.16}$$

for $i \neq j \in [q]$.

Theorem 6.2 directly implies non-reconstruction results for Galton-Watson trees.

Corollary 6.3. *Consider the model* $\mathrm{BOT}(q, \lambda, D)$ *(Definition 5.9) with* $d = \mathbb{E}_{b\sim D} b$. *If*

$$\eta_{\mathrm{KL}}(\pi, M^*)d < 1, \tag{6.17}$$

then the model admits non-reconstruction.

Proof. Let T be a Galton-Watson tree with offspring distrbution D. If T extincts, then non-reconstruction obviously hold. Conditioned on non-extinction, we have $\mathrm{br}(T) = d$ almost surely by [92], thus Theorem 6.2 applies. □

[115] proved non-reconstruction results on arbitrary directed acyclic graphs (in particular trees) by reducing to percolation problems on the same graph. In the case of the BOT model, their result says that the model admits non-reconstruction if

$$\eta_{\mathrm{KL}}(M)\,\mathrm{br}(T) < 1. \tag{6.18}$$

For any channel M, we have

$$\eta_{\mathrm{KL}}(\pi, M) \le \eta_{\mathrm{KL}}(M), \tag{6.19}$$

and the inequality is often strict. So for reversible channels (i.e., $M = M^*$), Theorem 6.2 implies result (6.18). We do not know, however, how to extend Theorem 6.2 to general DAGs using input-restricted contraction coefficients.

[87] proved a non-reconstruction result very similar to Theorem 6.2. They considered the symmetrized KL divergence

$$D_{\mathrm{SKL}}(P||Q) = D(P||Q) + D(Q||P), \tag{6.20}$$

which is f-divergence with $f(x) = (x-1)\log x$. They proved that non-reconstruction holds for a Galton-Watson tree with expected offspring d if

$$\eta_{\mathrm{SKL}}(\pi, M^*)d < 1. \tag{6.21}$$

The key step is the proof is the additivity of SKL-information under $\star$-convolution (Eq. (2.17)). By slightly modifying the proof of Theorem 6.2, their result can be strengthened to that the model $\mathrm{BOT}(T, q, \pi, M)$ admits non-reconstruction if

$$\eta_{\mathrm{SKL}}(\pi, M^*)\,\mathrm{br}(T) < 1. \tag{6.22}$$

In Section 6.2 we make some comparisons with their results, showing that in the cases we consider, the KL contraction method gives better results than the SKL contraction method. We remark that the input-unrestricted KL and SKL contraction coefficients agree by Eq. (2.42) and operator convexity of the relevant f-divergences, thus differences occur only for the input-restricted contraction coefficients.

In general, if some f-information (Definition 2.3) is subadditive under $\star$-convolution (e.g., Eq. (2.16), Eq. (2.17)), then non-reconstruction holds for any tree T satisfying

$$\eta_f(\pi, M^*)\,\mathrm{br}(T) < 1, \tag{6.23}$$

by modifying the proof of Theorem 6.2. An interesting question is, given a pair (π, M), what is the smallest $\eta_f(\pi, M^*)$ over all f-information subadditive under $\star$-convolution. Solving this question would give the best possible non-reconstruction result that can be achieved by our method.

6.2 Examples

In this section we apply Theorem 6.2 to several examples and demonstrate the power of our method.

6.2.1 Ising model

Recall the Ising model (Definition 5.9), where $\pi = \mathrm{Unif}(\{\pm\})$ and $M = \mathrm{BSC}_\delta$. Because M is reversible, Eq. (2.31) implies that

$$\eta_{\mathrm{KL}}(\pi, M^*) = (1-2\delta)^2. \tag{6.24}$$

Therefore Theorem 6.2 implies non-reconstruction for $(1-2\delta)^2\,\mathrm{br}(T) < 1$, which was shown in [23] (for regular trees) and [63] (for general trees). So for the Ising model on trees our method can give the tight reconstruction threshold.

6.2.2 Potts model

Recall the Potts model (Definition 5.9), where $\pi = \mathrm{Unif}([q])$ and $M = P_\lambda$ (Eq. (4.11)). Because the Potts channels are reversible, Theorem 6.2 implies non-reconstruction for

$$\eta_{\mathrm{KL}}(\pi, P_\lambda)\,\mathrm{br}(T) < 1. \tag{6.25}$$

Let us briefly discuss previous non-reconstruction results for the Potts channel.

[107] proved non-reconstruction for

$$\frac{q\lambda^2}{(q-2)\lambda+2}\,\mathrm{br}(T) < 1. \tag{6.26}$$

By Prop. 4.25 we see that Eq. (6.26) exactly corresponds to using the input-unrestricted KL contraction coefficient $\eta_{\mathrm{KL}}(P_\lambda)$. Therefore, Theorem 6.2 is strictly stronger than [107]. [95] proved non-reconstruction for regular trees for

$$d(1-\epsilon)\frac{q\lambda^2}{(q-2)\lambda+2} < 1 \tag{6.27}$$

for some $\epsilon = \epsilon(q, d, \lambda) > 0$.

[126] obtained very sharp results for regular trees, including that Kesten-Stigum (KS) threshold is tight for $q = 3$ and large enough d, and an expression for the reconstruction threshold for larger q and $d \to \infty$. [109] improved [126] and proved that the KS threshold is tight for $q = 3, 4$ and large enough tree, for Galton-Watson random trees with mild assumptions on the offspring distribution. It is unclear what results can be achieved for small d using their method. It seems that Theorem 6.2 is not able to give tightness of the KS threshold in these cases.

[66] gave non-reconstruction results very similar to ours by using the input-restricted SKL contraction coefficients. Numerical computation suggests that Theorem 6.2 gives better results than theirs in the case of Potts models. In Figure 6-1, we compare the input-restricted SKL and KL contraction coefficients for Potts channels P_λ for $q = 5$ and $\lambda \in \left[-\frac{1}{q-1}, 1\right]$. Because a simplified expression for $\eta_{\mathrm{SKL}}(\pi, P_\lambda)$ is not known, we use a lower bound $\overline{\eta}_{\mathrm{SKL}}(\pi, P_\lambda)$, which is defined as the sup of $\frac{D_{\mathrm{SKL}}(\nu P_\lambda \| \pi)}{D_{\mathrm{SKL}}(\nu \| \pi)}$ over distributions $\nu \in \mathcal{P}([q])$ with $\nu(2) = \cdots = \nu(q)$. Clearly $\overline{\eta}_{\mathrm{SKL}}(\pi, P_\lambda) \le \eta_{\mathrm{SKL}}(\pi, P_\lambda)$. [66] conjectured that $\overline{\eta}_{\mathrm{SKL}}(\pi, P_\lambda) = \eta_{\mathrm{SKL}}(\pi, P_\lambda)$ always holds.

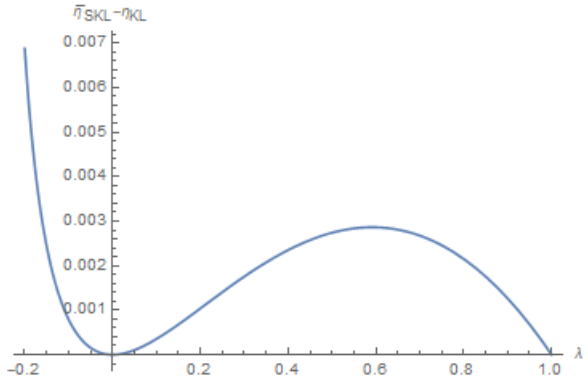

Figure 6-1: Contraction coefficient comparison for Potts channel with $q = 5$ and varying $\lambda \in \left[-\frac{1}{q-1}, 1\right]$. The figure shows $\overline{\eta}_{\mathrm{SKL}}(\pi, P_\lambda) - \eta_{\mathrm{KL}}(\pi, P_\lambda)$ is non-negative.

6.2.3 Random coloring model

The random coloring model is a special case of the Potts model where the contraction coefficient $\eta_{\mathrm{KL}}(\pi, P_\lambda)$ can be computed in closed form. It is the BOT model with $\pi = \mathrm{Unif}([q])$ and broadcasting channel $\mathrm{Col}_q := P_{-\frac{1}{q-1}}$. This channel acts on input $x \in [q]$ by outputting $y \neq x$ uniformly among all $q-1$ alternatives.

Theorem 6.2 and Prop. 4.23 together imply non-reconstruction for

$$\mathrm{br}(T) < \frac{\log q}{\log q - \log(q-1)} = (1 - o(1))q \log q. \tag{6.28}$$

This result was previously established by [125] (regular trees and Galton-Watson trees with Poisson offspring distribution), [20] (regular trees), and [59] (Galton-Watson trees with mild assumptions on the offspring distribution). Our result does not assume any conditions on the offspring distribution other than the expected offspring, and in fact works for arbitrary trees.

We remark that previous methods based on information contraction do not give the threshold $(1 - o(1))q \log q$. The information-percolation method [64, 115] implies non-reconstruction for

$$\eta_{\mathrm{KL}}(\mathrm{Col}_q)\, \mathrm{br}(T) < 1. \tag{6.29}$$

By Prop. 4.25, this gives non-reconstruction for $d < q - 1$ which is far from tight.

The SKL information contraction method [66] gives non-reconstruction for

$$\eta_{\mathrm{SKL}}(\pi, \mathrm{Col}_q)\, \mathrm{br}(T) < 1. \tag{6.30}$$

If we let $\nu_\epsilon := (1 - \epsilon, \frac{\epsilon}{q-1}, \ldots, \frac{\epsilon}{q-1})$, then

$$\eta_{\mathrm{SKL}}(\pi, \mathrm{Col}_q) \geq \lim_{\epsilon \to 0} \frac{D_{\mathrm{SKL}}(\nu_\epsilon \mathrm{Col}_q \,||\pi)}{D_{\mathrm{SKL}}(\nu_\epsilon||\pi)} = \frac{1}{q-1}. \tag{6.31}$$

Therefore this method cannot give non-reconstruction results better than for $d < q-1$.

6.2.4 Asymmetric Ising model

We consider an asymmetric version of the Ising model, which is the BOT model with

$$\pi = \left(\frac{b}{a+b}, \frac{a}{a+b}\right), \quad M = \begin{pmatrix} 1-a & a \\ b & 1-b \end{pmatrix}. \tag{6.32}$$

Note that M is reversible in this example.

[27] showed that the Kesten-Stigum threshold is tight when a, b are close to $\frac{1}{2}$. [91] determined the exact set (up to uncertainties on the boundary) of π for which the KS threshold is tight when degree d is large enough. It seems that Theorem 6.2 is unable to give tightness of the KS threshold in these cases.

In Figure 6-2, we compare $\eta_{\mathrm{SKL}}(\pi, M)$ and $\eta_{\mathrm{KL}}(\pi, M)$ for $a = 0.3$ and $b \in [0, 1]$.

The plot shows that in these cases, the KL contraction method gives better results than the SKL contraction method.

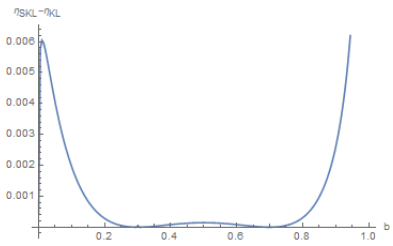

Figure 6-2: Contraction coefficient comparison for binary asymmetric channels with $a = 0.3$ and varying $b \in [0, 1]$. The figure shows $\eta_{\mathrm{SKL}}(\pi, M) - \eta_{\mathrm{KL}}(\pi, M)$ is non-negative.

6.3 Stochastic block model

In this section we study the weak recovery problem (Definition 5.5) for $\mathrm{SBM}(n, q, a, b)$ (Definition 5.4), the stochastic block model with q symmetric communities. In this model, there are n vertices, each independently and uniformly randomly assigned one of q labels. For two vertices, there is an edge between them with probability $\frac{a}{n}$ if they have the same labels, and with probability $\frac{b}{n}$ otherwise. The goal of weak recovery is to recover a non-trivial fraction of the communities given the unlabeled graph.

We refer the reader to Chapter 5 for a review of previous works on the weak recovery problem for $\mathrm{SBM}(n, q, a, b)$. Here we only mention that in the assortative regime ($a > b$), the previous best impossibility result for general q is by [16], which says weak recovery is impossible whenever

$$\frac{(a-b)^2}{a+(q-1)b} < \frac{2q\log(q-1)}{q-1}. \tag{6.33}$$

In the following, we show that non-reconstruction results based on input-restricted KL contraction coefficients lead to improved impossibility of weak recovery results for the SBM.

6.3.1 Impossibility of weak recovery via information percolation

We first give an impossibility result via an information percolation method of [115].

Proposition 6.4 ([115, Prop. 8])**.** *Weak recovery for the model* $\mathrm{SBM}(n, q, a, b)$ *is impossible, if the following tree model has non-reconstruction:*

Let $d = (\sqrt{a} - \sqrt{b})^2$*. Consider a Galton-Watson tree* T *with offspring distribution* $\mathrm{Pois}(d)$*. For each vertex, we independently and uniformly randomly choose a label* $\in [q]$*. Say vertex* v *has spin* σ_v*. We observe* $\omega_{u,v} = \mathbb{1}\{\sigma_u = \sigma_v\}$ *for each edge* (u, v)*.*

Let ρ denote the root of T and L_k denote the set of vertices at distance k to ρ. Let ω denote the set of all observations. We say the model admits non-reconstruction if

$$\lim_{k\to\infty} I(\sigma_\rho; \sigma_{L_k}|T,\omega) = 0. \tag{6.34}$$

[115] proved that the tree model has non-reconstruction when $d < \frac{q}{2}$ using a coupling argument. The following result makes an improvement.

Proposition 6.5. *The tree model in Prop. 6.4 admits non-reconstruction when*

$$d < \left(\frac{\log q - \log(q-1)}{\log q}\frac{q-1}{q} + \frac{1}{q}\right)^{-1} = q - (1+o(1))q/\log q. \tag{6.35}$$

Proof. The tree model is equivalent to the following top-down process:

1. Generate σ_ρ uniformly randomly over $[q]$.
2. Suppose we have generated a label for a vertex u. For every child v of u, we randomly choose the transition matrix M, which is the identity channel Id with probability $\frac{1}{q}$, and Col_q with probability $1-\frac{1}{q}$. Then we generate v according to $\mathbb{P}(\sigma_v = j|\sigma_u = i) = M_{i,j}$.

For any vertex u, let $L_{u,k}$ denote the set of descendants of u at distance k to ρ. Let v be a child of u.

Note that $\pi = \mathrm{Unif}([q])$ is an invariant distribution for both Id and Col_q, and the two channels are both reversible. We have

$$\mathbb{E}\left[I(\sigma_u; \sigma_{L_{v,k}}|T,\omega)|\omega_{u,v}=1\right] \le \eta_{\mathrm{KL}}(\pi,\mathrm{Id})I(\sigma_v;\sigma_{L_{v,k}}|T,\omega) = I(\sigma_v;\sigma_{L_{v,k}}|T,\omega) \tag{6.36}$$

and, by Prop. 4.23,

$$\begin{aligned}\mathbb{E}\left[I(\sigma_u; \sigma_{L_{v,k}}|T,\omega)|\omega_{u,v}=0\right] &\le \eta_{\mathrm{KL}}(\pi,\mathrm{Col}_q)I(\sigma_v;\sigma_{L_{v,k}}|T,\omega)\\ &= \frac{\log q - \log(q-1)}{\log q} I(\sigma_v;\sigma_{L_{v,k}}|T,\omega).\end{aligned} \tag{6.37}$$

Taking expectation, we get

$$I(\sigma_u; \sigma_{L_{v,k}}|T,\omega) \le \left(\frac{\log q - \log(q-1)}{\log q}\frac{q-1}{q} + \frac{1}{q}\right) I(\sigma_v;\sigma_{L_{v,k}}|T,\omega). \tag{6.38}$$

Rest of the proof is the same as Theorem 6.2. □

Prop. 6.4 and Prop. 6.5 together show that weak recovery is impossible for $\mathrm{SBM}(n,q,a,b)$ when

$$\left(\sqrt{a}-\sqrt{b}\right)^2 < \left(\frac{\log q - \log(q-1)}{\log q}\frac{q-1}{q} + \frac{1}{q}\right)^{-1}. \tag{6.39}$$

As shown in Figure 6-3, for certain parameters, (6.39) leads to slight improvement over [16].

6.3.2 Impossibility of weak recovery via Potts model

We have shown that the information percolation method together with the input-restricted KL contraction coefficients gives a simple yet strong impossibility result for weak recovery of the stochastic block model. The information percolation method can be understood as comparison with the erasure channel. However, the stochastic block model is more closely related to the Potts channel. As shown in Prop. 5.15, weak recovery for the model $\mathrm{SBM}(n, q, a, b)$ is impossible if the corresponding BOT model $\mathrm{BOT}(q, \lambda, \mathrm{Pois}(d))$ admits non-reconstruction, where

$$d = \frac{a + (q-1)b}{q}, \quad \lambda = \frac{a-b}{a+(q-1)b}. \tag{6.40}$$

Therefore, Theorem 6.2 implies the following result.

Theorem 6.6 (Impossibility of weak recovery for SBM)**.** *Weak recovery for the model* $\mathrm{SBM}(n, q, a, b)$ *is impossible when*

$$\eta_{\mathrm{KL}}(\pi, P_\lambda) d < 1, \tag{6.41}$$

where d *and* λ *are given in Eq.* (6.40)*.*

Figure 6-3 shows a comparison between the impossibility results for $q = 5$.

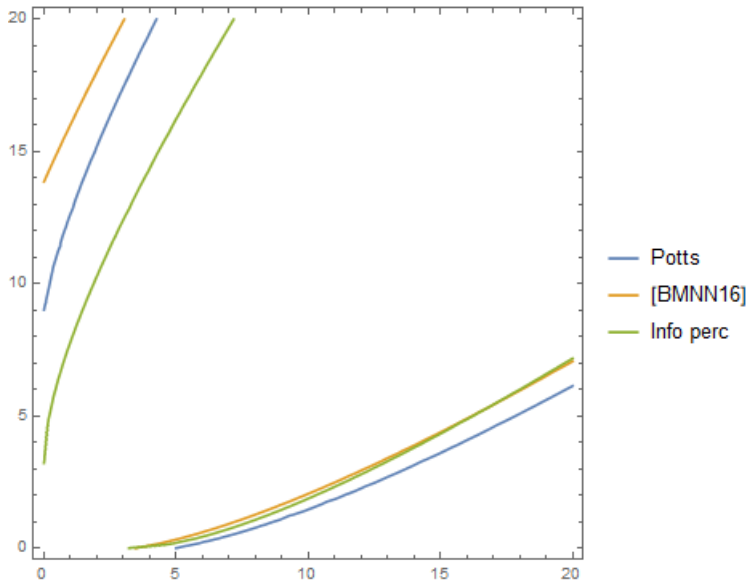

Figure 6-3: Impossibility of weak recovery results for SBM for $q = 5$. Horizontal axis is a, and vertical axis is b. In the assortative regime, (6.39) gives better results than [16] for certain parameters, and Theorem 6.6 gives the best results among the three.

Note that (6.33) is equivalent to

$$\frac{\lambda^2(q-1)}{2\log(q-1)} \cdot d < 1. \tag{6.42}$$

Comparing Theorem 6.6 and Eq. (6.42) using Eq. (4.230), we see that Theorem 6.6 strictly improves over (6.33) in the assortative regime.

6.4 Non-reconstruction for broadcasting with a Gaussian kernel

In this section, we prove optimal non-reconstruction results for a BOT model with continuous alphabet considered in [60], using our method developed in Section 6.1.

Definition 6.7 (Broadcasting on trees with a Gaussian kernel)**.** In this model, we are given a (possibly) infinite tree T with a marked root ρ. The state space $\mathcal{X}$ is the unit circle $S^1 := \mathbb{R}/2\pi\mathbb{Z}$. Let $\pi = \mathrm{Unif}(S^1)$ be the uniform distribution. Let $t > 0$ be a parameter. The transfer kernel is M_t, defined as $Y = X + Z_t$ where $Z_t \sim \mathcal{N}(0, t)$, where X is the input and Y is the output.

Now for each vertex $v \in T$, we generate a spin $\sigma_v \in \mathcal{X}$ according to the following process:

1. Generate $\sigma_\rho \sim \pi$.
2. Suppose we have generated a label for vertex u. For every child v of u, we generate v according to $\sigma_v \sim M_t(\cdot|\sigma_u)$.

Let L_k denote the set of vertices at distance k to ρ. We say the model admits non-reconstruction if and only if

$$\lim_{k\to\infty} I(\sigma_\rho; \sigma_{L_k}) = 0. \tag{6.43}$$

Let $\lambda(M_t)$ denote the second largest eigenvalue of M_t. [60] proved that for the above BOT model on a regular tree with offspring d, reconstruction holds when $d\lambda(M_t)^2 > 1$, and non-reconstruction holds for $d\lambda(M_t) < 1$. Note that there is a $\lambda(M_t)$ factor gap between the reconstruction result and the non-reconstruction result. In the following, we prove that non-reconstruction holds as long as $d\lambda(M_t)^2 < 1$, closing the gap.

We remark that [108] studied a different BOT model with Gaussian broadcasting channels, and deteremined the reconstruction threshold for their model (which happened to also coincide with the Kesten-Stigum threshold). While sharing some similarities, their and our models do not seem to be directly comparable with each other.

Theorem 6.8 (Non-reconstruction for Gaussian BOT model)**.** *Consider the BOT model defined in Definition 6.7.*

If T is an infinite rooted tree with bounded maximum degree, then the BOT model admits non-reconstruction when

$$\mathrm{br}(T)\lambda(M_t)^2 < 1. \tag{6.44}$$

If T is a Galton-Watson tree with expected offspring d, then the BOT model admits non-reconstruction when

$$d\lambda(M_t)^2 < 1. \tag{6.45}$$

The proof idea is to upper bound the input-restricted KL contraction coefficient by $\lambda(M_t)^2$, then use a tree recursion similar to that of Theorem 6.2. However, because we are working in a continuous space, we must be careful about what we mean by contraction coefficients.

We would like an inequality of form

$$I(\sigma_u; \sigma_{L_{v,k}}) \le \widetilde{\eta}_{\mathrm{KL}}(\pi, M_t) I(\sigma_v; \sigma_{L_{v,k}}) \tag{6.46}$$

where $u \in V(T)$, v is child of u, $L_{v,k}$ is the set of descendants of v at distance k to ρ, and $\widetilde{\eta}_{\mathrm{KL}}(\pi, M_t)$ is a continuous version of contraction coefficient $\eta_{\mathrm{KL}}(\pi, M_t)$.

We have

$$I(\sigma_u; \sigma_{L_{v,k}}) = \mathbb{E}_{\sigma_{L_{v,k}}} D(P_{\sigma_u|\sigma_{L_{v,k}}} \| P_{\sigma_u}) \tag{6.47}$$

$$= \mathbb{E}_{\sigma_{L_{v,k}}} D(M_t \circ P_{\sigma_v|\sigma_{L_{v,k}}} \| \pi). \tag{6.48}$$

Let us consider the distribution $P_{\sigma_u|\sigma_{L_{u,k}}}$. If $k = d(v, \rho)$, then $P_{\sigma_u|\sigma_{L_{u,k}}}$ is a point measure. However, as long as $k > d(u, \rho)$, pdf of $P_{\sigma_u|\sigma_{L_{u,k}}}$ is smooth on $\mathcal{X}$ by an induction using belief propagation equation. Therefore we make the following definition.

Definition 6.9 (Smooth contraction coefficient)**.** We define

$$\widetilde{\eta}_{\mathrm{KL}}(\pi, M_t) := \sup_{f \in \mathcal{C}} \frac{\mathrm{Ent}_\pi(M_t f)}{\mathrm{Ent}_\pi(f)}, \tag{6.49}$$

$$\mathcal{C} := \{f : \mathcal{X} \to \mathbb{R}_{\ge 0} | f \text{ smooth}, \mathbb{E}_\pi[f] = 1\}. \tag{6.50}$$

where $\mathrm{Ent}_\pi(f)$ is defined in Eq. (4.2).

Lemma 6.10.

$$\widetilde{\eta}_{\mathrm{KL}}(\pi, M_t) \le \exp(-t). \tag{6.51}$$

Proof. Note that $(M_t)_{t \ge 0}$ forms a semigroup. Therefore it suffices to prove that for all $f \in \mathcal{C}$, we have

$$\frac{d}{dt}|_{t=0} \mathrm{Ent}(f_t) \le -\mathrm{Ent}(f) \tag{6.52}$$

where $f_t = M_t f$.

We have

$$\begin{aligned}
&\frac{d}{dt}|_{t=0}\operatorname{Ent}(f_t)\\
&=\mathbb{E}\left[\frac{d}{dt}|_{t=0}(f_t\log f_t)\right]\\
&=\mathbb{E}\left[(1+\log f)\frac{d}{dt}|_{t=0}f_t\right]\\
&=\mathbb{E}\left[(\log f)\frac{d}{dt}|_{t=0}f_t\right]\\
&=\frac{1}{2}\mathbb{E}\left[f''\log f\right] && \text{(heat equation)}\\
&=-\frac{1}{2}\mathbb{E}\left[\frac{(f')^2}{f}\right] && \text{(integration by parts)}\\
&\le -\operatorname{Ent}(f). && \text{([61])}
\end{aligned}$$

This finishes the proof. □

Now we are ready to prove Theorem 6.8.

Proof of Theorem 6.8. By Lemma 6.10, we have

$$\widetilde{\eta}_{\mathrm{KL}}(\pi, M_t) \le \exp(-t) = \lambda(M_t)^2, \tag{6.53}$$

where the value of $\lambda(M_t)$ is proved in e.g., [60]. Therefore we only need to prove that $\mathrm{br}(T)\widetilde{\eta}_{\mathrm{KL}}(\pi, M_t) < 1$ implies non-reconstruction. Note that the channel M_t is reversible.

Bounded degree case: For $u \in V(T)$, define

$$r_u := \lim_{k\to\infty} I(\sigma_u; \sigma_{L_{u,k}}). \tag{6.54}$$

By data processing inequality, $I(\sigma_u; \sigma_{L_{u,k}})$ is non-increasing for $k \ge d(u,\rho)$, so the limit always exists. Because T has bounded maximum degree, we have

$$r_u \le I(\sigma_u; \sigma_{L_{u,d(u,\rho)+1}}). \tag{6.55}$$

So there exists a constant $C > 0$ such that $r_u \le C$ for all $u \in v(T)$.

Now define

$$a_u = R^{-1}\widetilde{\eta}_{\mathrm{KL}}(\pi, M_t)^{d(u,\rho)} r_u. \tag{6.56}$$

Let $c(u)$ be the set of children of u. For any $v \in c(u)$, by Markov chain

$$\sigma_{L_{v,k}} \to \sigma_v \to \sigma_u \tag{6.57}$$

and discussion before Lemma 6.10, we have

$$I(\sigma_u; \sigma_{L_{v,k}}) \le \widetilde{\eta}_{\mathrm{KL}}(\pi, M_t) I(\sigma_v, \sigma_{L_{v,k}}). \tag{6.58}$$

Because $(\sigma_{L_{v,k}})_{v \in c(u)}$ are independent conditioned on σ_u, we have

$$I(\sigma_u; \sigma_{L_{u,k}}) \le \sum_{v \in c(u)} I(\sigma_u; \sigma_{L_{v,k}}). \tag{6.59}$$

Combining the two inequalities and let $k \to \infty$, we get

$$a_u \le \sum_{v \in c(u)} a_v. \tag{6.60}$$

Furthermore, we have $a_u \le \widetilde{\eta}_{\mathrm{KL}}(\pi, M_t)^{d(u,\rho)}$.

Now define a flow b as follows. For any $u \in V(T)$, let $u_0 = \rho, \dots, u_\ell = u$ be the shortest path from ρ to u. Define

$$b_u = a_u \prod_{0 \le j \le \ell - 1} \frac{a_{u_j}}{\sum_{v \in c(u_j)} a_v}. \tag{6.61}$$

(If $\sum_{v \in c(u_j)} a_v = 0$ for some j, then let $b_u = 0$.) Then we have

$$b_u = \sum_{v \in c(u)} b_v, \tag{6.62}$$

and that

$$b_u \le a_u \le \widetilde{\eta}_{\mathrm{KL}}(\pi, M_t)^{d(u,\rho)}. \tag{6.63}$$

By definition of branching number, we must have $b_\rho = 0$. Therefore $r_\rho = 0$ and non-reconstruction holds.

Galton-Watson tree case: Let D be the offspring distribution. We have

$$\begin{aligned}
I(\sigma_\rho; \sigma_{L_k} | T) &\le \mathbb{E}_{c(\rho)} \sum_{v \in c(\rho)} I(\sigma_\rho; \sigma_{L_{v,k}} | T) \\
&\le \mathbb{E}_{c(\rho)} \sum_{v \in c(\rho)} \widetilde{\eta}_{\mathrm{KL}}(\pi, M_t) I(\sigma_v; \sigma_{L_{v,k}} | T_v) \\
&= \widetilde{\eta}_{\mathrm{KL}}(\pi, M_t) \mathbb{E}_{c(\rho)} \sum_{v \in c(\rho)} I(\sigma_v; \sigma_{L_{v,k}} | T_v) \\
&= \widetilde{\eta}_{\mathrm{KL}}(\pi, M_t) \mathbb{E}_{b \sim D} \left[b I(\sigma_\rho; \sigma_{L_{k-1}} | T) \right] \\
&= \widetilde{\eta}_{\mathrm{KL}}(\pi, M_t) d I(\sigma_\rho; \sigma_{L_{k-1}} | T).
\end{aligned}$$

Here T_v denotes the subtree rooted at v. Because $I(\sigma_\rho; \sigma_{L_1} | T) < \infty$, when $d\widetilde{\eta}_{\mathrm{KL}}(\pi, M_t) <$

1, we have

$$\lim_{k\to\infty} I(\sigma_\rho;\sigma_{L_k}|T) = 0. \tag{6.64}$$

This finishes the proof. □

Chapter 7

Reconstruction for broadcasting on hypertrees

We study the problem of weak recovery for the two-community simple HSBM (Definition 5.4). In this model, a random r-uniform hypergraph is generated by placing hyperedges with higher density if all vertices of a hyperedge share the same binary label. By analyzing contraction of symmetric KL information, we prove that for $r = 3, 4$, weak recovery is impossible below the Kesten-Stigum threshold, where the $r = 4$ case relies on a numerically verified inequality. Prior work [112] established that weak recovery in HSBM is always possible above the Kesten-Stigum threshold. Consequently, there is no information-computation gap for these r, which partially resolves a conjecture of [15]. To our knowledge this is the first impossibility result for HSBM weak recovery beyond celebrated results [96, 105] for the graph case.

As usual, we reduce the weak recovery problem for the HSBM to the study of the corresponding broadcasting on hypertrees (BOHT) model. While we show that BOHT's reconstruction threshold coincides with Kesten-Stigum for $r = 3, 4$, surprisingly, we demonstrate that for $r \geq 7$ reconstruction is possible also below the Kesten-Stigum. This shows an interesting phase transition in the parameter r, and suggests that for $r \geq 7$, there might be an information-computation gap for the HSBM. For $r = 5, 6$ and large degree we propose an approach for showing non-reconstruction below Kesten-Stigum threshold, suggesting that $r = 7$ is the correct threshold for onset of the new phase.

This chapter is based on [74].

Chapter outline In Section 7.1, we give a brief introduction to the problem and our results. In Section 7.2, we write down an explicit formula for the belief propagation operator. In Section 7.3, we prove our results on BOHT with $r = 3, 4$ (Theorem 7.1(i)(ii)). In Section 7.4, we prove our results on BOHT with $r \geq 7$ and large enough d (Theorem 7.1(iv)). In Section 7.5, we prove impossibility of weak recovery results for the HSBM (Theorem 7.2), and reconstruction for BOHT above the Kesten-Stigum threshold (Theorem 7.1(iii)). In Section 7.6, we discuss a possible approach to resolve the $r = 5, 6$ case.

7.1 Introduction

Hypergraph stochastic block model Consider the model HSBM(n, r, a, b) defined in Definition 5.4, where $n \in \mathbb{Z}_{\geq 1}$ is the number of vertices, $r \in \mathbb{Z}_{\geq 2}$ is the hyperedge size, and $a > b \in \mathbb{R}_{\geq 0}$ are two parameters. The model generates a random hypergraph as follows: Let the vertex set be $V = [n]$. Generate a random label X_u for all vertices $u \in V$ i.i.d. $\sim \mathrm{Unif}(\{\pm\})$. Then, for every $S \in \binom{V}{r}$, if all vertices in S have the same label, add hyperedge S with probability $\frac{a}{\binom{n}{r-1}}$; otherwise add hyperedge S with probability $\frac{b}{\binom{n}{r-1}}$.

For the weak recovery problem (Definition 5.5) for this model, [15] conjectured that a phase transition occurs at the Kesten-Stigum threshold. The positive part of their conjecture has been proved by [112, 130] for more general HSBMs, giving an efficient weak recovery algorithm based on above the Kesten-Stigum threshold. Despite the progress on the positive part, there has been no progress for the negative part (impossibility of weak recovery) for $r \geq 3$.

For the $r = 2$ case, the positive part was proved by [96, 105] and the negative part was established by [103, 105] via a reduction to the broadcasting on trees (BOT) model. Therefore a natural idea is to study the reconstruction problem for corresponding broadcasting on hypertrees (BOHT) model via Theorem 5.15. [138] mentioned that the difficulty in proving negative results lies in analyzing the broadcasting on hypertrees (BOHT) model. We prove impossibility of weak recovery results by proving non-reconstruction results for the BOHT model.

We define the following useful parameters. The values of d, λ and SNR agree with those defined in Definition 5.3.

- For every vertex u, the expected number of hyperedges containing u is $d \pm o(1)$, where
$$d = \frac{(a-b) + 2^{r-1}b}{2^{r-1}}. \tag{7.1}$$
- Expected number of vertices adjacent to u is $\alpha \pm o(1)$, where
$$\alpha = (r-1)d = (r-1)\frac{(a-b) + 2^{r-1}b}{2^{r-1}}. \tag{7.2}$$
- Expected number of neighbors in the same community minus the number of neighbors in the other community is $\beta \pm o(1)$ where
$$\beta = (r-1)\frac{a-b}{2^{r-1}}. \tag{7.3}$$
- The strength of the broadcasting channel is characterized by $\lambda \in [0, 1]$, defined

as

$$\lambda = \frac{\beta}{\alpha} = \frac{a-b}{a-b+2^{r-1}b}. \tag{7.4}$$

- Signal-to-noise ratio

$$\mathrm{SNR} := \alpha\lambda^2 = \frac{(r-1)(a-b)^2}{2^{r-1}((a-b)+2^{r-1}b)}. \tag{7.5}$$

The Kesten-Stigum threshold for this model is at $\mathrm{SNR} = 1$.

Broadcasting on hypertrees By Theorem 5.15, weak recovery for the HSBM can be reduced to the study of the corresponding BOHT model. For $\mathrm{HSBM}(n, r, a, b)$, the corresponding model is $\mathrm{BOHT}(r, \lambda, \mathrm{Pois}(d))$ (Definition 5.9). The model has three parameters: $r \in \mathbb{Z}_{\geq 2}$, hyperedge edge; d, expected offspring; $\lambda \in [0, 1]$, broadcasting channel strength. Let T be a linear r-uniform hypertree where either (1) every vertex has d downward hyperedges (thus $d(r-1)$ children), or (2) every vertex has b downward hyperedges (thus $b(r-1)$ children), where $b \sim \mathrm{Pois}(d)$ is i.i.d. generated from the Poisson distribution with expectation d. We call the first case the regular hypertree, and the second case the Poisson hypertree. Given a hypertree T with root ρ, we generate a label $\sigma_u \in \{\pm\}$ for every vertex u via a downward process: (1) $\sigma_\rho \sim \mathrm{Unif}(\{\pm\})$ (2) given σ_u, for every downward hyperedge $S = \{u, v_1, \ldots, v_{r-1}\}$, we generate $\sigma_{v_1}, \ldots, \sigma_{v_{r-1}}$ such that for any $x_1, \ldots, x_{r-1} \in \{\pm\}^{r-1}$,

$$\mathbb{P}[\sigma_{v_1} = x_1, \ldots, \sigma_{v_{r-1}} = x_{r-1} | \sigma_u] = \begin{cases} \lambda + \frac{1}{2^{r-1}}(1-\lambda), & \text{if } x_1 = \cdots = x_{r-1} = \sigma_u, \\ \frac{1}{2^{r-1}}(1-\lambda), & \text{otherwise.} \end{cases} \tag{7.6}$$

We denote the channel $\sigma_u \to (\sigma_{v_1}, \ldots, \sigma_{v_{r-1}})$ as $B = B_{r-1} : \{\pm\} \to \{\pm\}^{r-1}$. This is a binary memoryless symmetric (BMS) channel.

The reconstruction problem (Definition 5.14) asks whether we can gain any non-trivial information about the root given observation of far away vertices. In other words, whether the limit

$$\lim_{k \to \infty} I(\sigma_\rho; T_k, \sigma_{L_k}) \tag{7.7}$$

is non-zero, where L_k is the set of vertices at distance k to the root ρ, and T_k is the set of vertices at distance $\leq k$ to ρ. When the limit is non-zero, we say the BOHT model admits reconstruction; when the limit is zero, we say the model admits non-reconstruction. Theorem 5.15 shows that non-reconstruction for the $\mathrm{BOHT}(r, \lambda, \mathrm{Pois}(d))$ model implies impossibility of weak recovery for the $\mathrm{HSBM}(n, r, a, b)$ model.

The reconstruction problem has been studied on various BOT models, e.g., [23, 63, 101, 107, 98, 27, 19, 125, 126, 87, 91, 72, 109]. See Chapter 6 for more discussions. Nevertheless, to our knowledge, there has been no previous work studying the

reconstruction problem for BOHT.

Our results Our results on the reconstruction problem for BOHT is summarized as follows.

Theorem 7.1 (Reconstruction threshold for BOHT)**.** *Consider the* $\mathrm{BOHT}(r, \lambda, d)$ *model or the* $\mathrm{BOHT}(r, \lambda, \mathrm{Pois}(d))$ *model. We have the following non-reconstruction results for the BOHT model.*

(i) For $r = 3$*, the BOHT model admits non-reconstruction when* $(r-1)d\lambda^2 < 1$*.*

(ii) For $r = 4$*, if Conjecture 7.6 is true, then the BOHT model admits non-reconstruction when* $(r-1)d\lambda^2 < 1$*.*

We have the following reconstruction results for the BOHT model.

(iii) For $r \geq 2$*, the BOHT model admits reconstruction when* $(r-1)d\lambda^2 > 1$*.*

(iv) For $r \geq 7$*, there exists a constant* $d_0 = d_0(r)$ *such that for all* $d \geq d_0$*, there exists* $\lambda \in [0, 1]$ *such that* $(r-1)d\lambda^2 < 1$ *and the BOHT model admits reconstruction.*

We note that Conjecture 7.6 is numerically verified.

Theorem 7.1 implies the following impossibility of weak recovery results for HSBM.

Theorem 7.2 (Weak recovery threshold for HSBM)**.** *Consider the model* $\mathrm{HSBM}(n, r, a, b)$*. Recall* d *and* SNR *defined in* (7.1)(7.5)*.*

(i) For $r = 3$*, weak recovery is impossible for the HSBM when* $\mathrm{SNR} < 1$*.*

(ii) For $r = 4$*, if Conjecture 7.6 is true, weak recovery is impossible for the HSBM when* $\mathrm{SNR} < 1$*.*

Our technique We prove Theorem 7.1(i)(ii) by considering contraction of SKL capacity under belief propagation recursion. It is known that SKL capacity can be used to prove non-reconstruction results since at least [87]. Information-theoretically, SKL information is special due to its additivity (as opposed to subadditivity) under $\star$-convolution.

Interestingly, before our work, to the best of our knowledge, non-reconstruction results proved via SKL capacity could always be also shown via other information measures (χ^2-capacity [63], KL capacity [72], etc.). It appears, thus, that BOHT is the first example where contraction via SKL capacity gives better results than other information measures we have tried.

Theorem 7.1(iii) is an immediate consequence of the weak recovery results of [112].

Theorem 7.1(iv) is proved using contraction of χ^2-capacity and Gaussian approximation, which has proved successful in many different settings [125, 126, 91, 109].

Theorem 7.2 is an immediate consequence of Theorem 7.1(i)(ii) via Theorem 5.15.

7.2 Belief propagation recursion

In this section we give an explicit formula for the belief propagation (BP) operator (Definition 5.13). We take an information channel point of view, which interprets the BP recursion as an operator from the space of BMS channels (equivalently, the space of distributions on $[0, \frac{1}{2}]$, via Lemma 2.2) to itself.

Consider the model $\mathrm{BOHT}(r, \lambda, D)$ (Definition 5.9), where D is either the point distribution at d (regular hypertree case) or $\mathrm{Pois}(d)$ (Poisson hypertree case). Then the BP operator is defined as

$$\mathrm{BP}(P) := \mathbb{E}_{t\sim D}(P^{\times(r-1)} \circ B)^{\star t}, \tag{7.8}$$

where B is the channel $\sigma_u \to (\sigma_{v_1}, \ldots, \sigma_{v_{r-1}})$ for a single hyperedge, defined in (7.6). The BP operator sends the space of BMS channels to itself. For the sequence $(M_k)_{k\geq 0}$ (Definition 5.12), we have

$$M_{k+1} = \mathrm{BP}(M_k). \tag{7.9}$$

Our goal in this section is to derive a formula for $P^{\times(r-1)} \circ B$ in terms of the θ-component (recall that $\theta = 1 - 2\Delta$ where Δ is the Δ-component).

By Lemma 2.2, we only need to describe $(\mathrm{BSC}_{\Delta_1} \times \cdots \times \mathrm{BSC}_{\Delta_{r-1}}) \circ B$. Let $\theta_i := 1 - 2\Delta_i$ for $i \in [r-1]$. For $x \in \{\pm\}^{r-1}$, we have

$$\begin{aligned}
&((\mathrm{BSC}_{\Delta_1} \times \cdots \times \mathrm{BSC}_{\Delta_{r-1}}) \circ B)(x_1, \ldots, x_{r-1}|+) \\
&= \sum_{y\in\{\pm\}^{r-1}} B(y_1, \ldots, y_{r-1}|+) \prod_{i\in[r-1]} \mathrm{BSC}_{\Delta_i}(x_i|y_i) \\
&= \lambda \prod_{i\in[r-1]} \mathrm{BSC}_{\Delta_i}(x_i|+) + \frac{1}{2^{r-1}}(1-\lambda) \prod_{i\in[r-1]} \sum_{y_i\in\{\pm\}} \mathrm{BSC}_{\Delta_i}(x_i|y_i) \\
&= \lambda \prod_{i\in[r-1]} \left(\frac{1}{2} + \left(\frac{1}{2} - \Delta_i\right) x_i\right) + \frac{1}{2^{r-1}}(1-\lambda) \\
&= \lambda \prod_{i\in[r-1]} \left(\frac{1}{2} + \frac{1}{2}\theta_i x_i\right) + \frac{1}{2^{r-1}}(1-\lambda).
\end{aligned} \tag{7.10}$$

So $(\mathrm{BSC}_{\Delta_1} \times \cdots \times \mathrm{BSC}_{\Delta_{r-1}}) \circ B$ is a mixture of 2^{r-2} BSCs, indexed by the set

$$\left\{x : x \in \{\pm\}^{r-1}, x_1 = +\right\}, \tag{7.11}$$

where the BSC corresponding to x has weight (probability)

$$\left(\lambda \prod_{i\in[r-1]} \left(\frac{1}{2}+\frac{1}{2}\theta_i x_i\right)+\frac{1}{2^{r-1}}(1-\lambda)\right)+\left(\lambda \prod_{i\in[r-1]} \left(\frac{1}{2}-\frac{1}{2}\theta_i x_i\right)+\frac{1}{2^{r-1}}(1-\lambda)\right) \tag{7.12}$$

$$= \lambda\left(\prod_{i\in[r-1]} \left(\frac{1}{2}+\frac{1}{2}\theta_i x_i\right)+\prod_{i\in[r-1]} \left(\frac{1}{2}-\frac{1}{2}\theta_i x_i\right)\right)+\frac{1}{2^{r-2}}(1-\lambda)$$

and θ parameter equal to the absolute value of

$$\frac{\left(\lambda\prod_{i\in[r-1]}\left(\frac{1}{2}+\frac{1}{2}\theta_i x_i\right)+\frac{1}{2^{r-1}}(1-\lambda)\right)-\left(\lambda\prod_{i\in[r-1]}\left(\frac{1}{2}-\frac{1}{2}\theta_i x_i\right)+\frac{1}{2^{r-1}}(1-\lambda)\right)}{\left(\lambda\prod_{i\in[r-1]}\left(\frac{1}{2}+\frac{1}{2}\theta_i x_i\right)+\frac{1}{2^{r-1}}(1-\lambda)\right)+\left(\lambda\prod_{i\in[r-1]}\left(\frac{1}{2}-\frac{1}{2}\theta_i x_i\right)+\frac{1}{2^{r-1}}(1-\lambda)\right)} \tag{7.13}$$

$$= \frac{\lambda\left(\prod_{i\in[r-1]}\left(\frac{1}{2}+\frac{1}{2}\theta_i x_i\right)-\prod_{i\in[r-1]}\left(\frac{1}{2}-\frac{1}{2}\theta_i x_i\right)\right)}{\lambda\left(\prod_{i\in[r-1]}\left(\frac{1}{2}+\frac{1}{2}\theta_i x_i\right)+\prod_{i\in[r-1]}\left(\frac{1}{2}-\frac{1}{2}\theta_i x_i\right)\right)+\frac{1}{2^{r-2}}(1-\lambda)}$$

$$= \frac{\lambda\left(\prod_{i\in[r-1]}(1+\theta_i x_i)-\prod_{i\in[r-1]}(1-\theta_i x_i)\right)}{\lambda\left(\prod_{i\in[r-1]}(1+\theta_i x_i)+\prod_{i\in[r-1]}(1-\theta_i x_i)\right)+2(1-\lambda)}.$$

Although we need to take absolute value, information measures (except for P_e) in Definition 2.4 are all even functions in θ. So we do not need to worry about the sign.

7.3 Reconstruction threshold for $r = 3, 4$

In this section we prove Theorem 7.1(i)(ii). Our method is via contraction of SKL capacity. That is, we prove that

$$C_{\mathrm{SKL}}(\mathrm{BP}(P)) \le (r-1)d\lambda^2 C_{\mathrm{SKL}}(P) \tag{7.14}$$

for all BMS P.

Applying additivity of SKL capacity under $\star$-convolution (Eq. (2.19)), we get

$$C_{\mathrm{SKL}}(\mathrm{BP}(P)) = \mathbb{E}_t C_{\mathrm{SKL}}((P^{\times(r-1)}\circ B)^{\star t}) = dC_{\mathrm{SKL}}(P^{\times(r-1)}\circ B) \tag{7.15}$$

where t is the offspring ($t = d$ for regular hypertrees, $t \sim \mathrm{Pois}(d)$ for Poisson hypertrees). Therefore it suffices to prove

$$C_{\mathrm{SKL}}(P^{\times(r-1)}\circ B) \le (r-1)\lambda^2 C_{\mathrm{SKL}}(P) \tag{7.16}$$

By Lemma 2.2, we can reduce (7.16) to

$$C_{\mathrm{SKL}}\left((\mathrm{BSC}_{\Delta_1} \times \cdots \times \mathrm{BSC}_{\Delta_{r-1}}) \circ B\right) \le \lambda^2 \sum_{i \in [r-1]} C_{\mathrm{SKL}}(\mathrm{BSC}_{\Delta_i}) \tag{7.17}$$

for all $\Delta_1, \ldots, \Delta_{r-1} \in [0, \frac{1}{2}]$.

7.3.1 Case $r = 3$

For $r = 3$, (7.17) indeed holds.

Lemma 7.3. *For any $\Delta_1, \Delta_2 \in [0, \frac{1}{2}]$, we have*

$$C_{\mathrm{SKL}}((\mathrm{BSC}_{\Delta_1} \times \mathrm{BSC}_{\Delta_2}) \circ B) \le \lambda^2 (C_{\mathrm{SKL}}(\mathrm{BSC}_{\Delta_1}) + C_{\mathrm{SKL}}(\mathrm{BSC}_{\Delta_2})). \tag{7.18}$$

Proof. We expand LHS of (7.18) using the BP recursion formula established in Section 7.2. Let $\theta_i = 1 - 2\Delta_i$ for $i = 1, 2$. Then

$$\begin{aligned}
&C_{\mathrm{SKL}}((\mathrm{BSC}_{\Delta_1} \times \mathrm{BSC}_{\Delta_2}) \circ B) \qquad (7.19)\\
&= \sum_{x_1 = +, x_2 \in \{\pm\}} \frac{1}{2} \lambda(\theta_1 x_1 + \theta_2 x_2) \operatorname{arctanh} \frac{\lambda(\theta_1 x_1 + \theta_2 x_2)}{\lambda(1 + \theta_1 x_1 \theta_2 x_2) + (1 - \lambda)} \\
&= \lambda \left(\frac{1}{2}(\theta_1 + \theta_2) \operatorname{arctanh} \frac{\lambda(\theta_1 + \theta_2)}{1 + \lambda\theta_1\theta_2} + \frac{1}{2}(\theta_1 - \theta_2) \operatorname{arctanh} \frac{\lambda(\theta_1 - \theta_2)}{1 - \lambda\theta_1\theta_2} \right) \\
&= \lambda \left(\frac{1}{2}(\theta_1 + \theta_2) F_\lambda(\theta_1, \theta_2) + \frac{1}{2}(\theta_1 - \theta_2) F_\lambda(\theta_1, -\theta_2) \right)
\end{aligned}$$

where

$$F_\lambda(\theta_1, \theta_2) := \operatorname{arctanh} \frac{\lambda(\theta_1 + \theta_2)}{1 + \lambda\theta_1\theta_2}. \tag{7.20}$$

Note that by definition, $F_\lambda(\theta_1, \theta_2) = -F_\lambda(-\theta_1, -\theta_2)$ and $F_\lambda(\theta_1, \theta_2) = F_\lambda(\theta_2, \theta_1)$.

We have

$$\begin{aligned}
&\frac{1}{2}(\theta_1 + \theta_2) F_\lambda(\theta_1, \theta_2) + \frac{1}{2}(\theta_1 - \theta_2) F_\lambda(\theta_1, -\theta_2) \qquad (7.21)\\
&= \frac{1}{2}\theta_1 (F_\lambda(\theta_1, \theta_2) + F_\lambda(\theta_1, -\theta_2)) + \frac{1}{2}\theta_2 (F_\lambda(\theta_1, \theta_2) + F_\lambda(-\theta_1, \theta_2)) \\
&\le \theta_1 F_\lambda(\theta_1, 0) + \theta_2 F_\lambda(0, \theta_2) \\
&= \theta_1 \operatorname{arctanh}(\lambda\theta_1) + \theta_2 \operatorname{arctanh}(\lambda\theta_2) \\
&\le \lambda(\theta_1 \operatorname{arctanh} \theta_1 + \theta_2 \operatorname{arctanh} \theta_2) \\
&= \lambda(C_{\mathrm{SKL}}(\mathrm{BSC}_{\Delta_1}) + C_{\mathrm{SKL}}(\mathrm{BSC}_{\Delta_2})),
\end{aligned}$$

where the second step follows from Lemma 7.4, and the fourth step follows convexity of arctanh in $[0, 1]$. Combining (7.19)(7.21) we finish the proof. □

Lemma 7.4. *For $\lambda, \theta_1, \theta_2 \in [0,1]$, we have*

$$\frac{1}{2}\left(F_\lambda(\theta_1,\theta_2) + F_\lambda(\theta_1,-\theta_2)\right) \le F_\lambda(\theta_1, 0). \tag{7.22}$$

Proof. We use the formula

$$\operatorname{arctanh} x + \operatorname{arctanh} y = \operatorname{arctanh} \frac{x+y}{1+xy} \tag{7.23}$$

to expand both sides of (7.22). LHS is

$$\begin{aligned} & F_\lambda(\theta_1,\theta_2) + F_\lambda(\theta_1,-\theta_2) \\ =\ & \operatorname{arctanh} \frac{\lambda(\theta_1+\theta_2)}{1+\lambda\theta_1\theta_2} + \operatorname{arctanh} \frac{\lambda(\theta_1-\theta_2)}{1-\lambda\theta_1\theta_2} \\ =\ & \operatorname{arctanh} \frac{2\lambda\theta_1(1-\lambda\theta_2^2)}{\lambda^2(\theta_1^2-\theta_2^2)+1-\lambda^2\theta_1^2\theta_2^2}. \end{aligned} \tag{7.24}$$

RHS is

$$2F_\lambda(\theta_1, 0) = \operatorname{arctanh} \frac{2\lambda\theta_1}{1+\lambda^2\theta_1^2}. \tag{7.25}$$

By comparing (7.24)(7.25) and using monotonicity of arctanh, it suffices to prove that

$$\frac{1-\lambda\theta_2^2}{\lambda^2(\theta_1^2-\theta_2^2)+1-\lambda^2\theta_1^2\theta_2^2} \le \frac{1}{1+\lambda^2\theta_1^2}. \tag{7.26}$$

We have

$$(\lambda^2(\theta_1^2-\theta_2^2)+1-\lambda^2\theta_1^2\theta_2^2) - (1-\lambda\theta_2^2)(1+\lambda^2\theta_1^2) = \lambda(1-\lambda)(1-\lambda\theta_1^2)\theta_2^2 \ge 0. \tag{7.27}$$

This finishes the proof. □

7.3.2 Case $r = 4$

For $r = 4$, (7.17) holds conditioned on a numerically-verified conjecture.

Lemma 7.5. *If Conjecture 7.6 is true, then for all $\Delta_1, \Delta_2, \Delta_3 \in [0, \frac{1}{2}]$, we have*

$$C_{\mathrm{SKL}}((\mathrm{BSC}_{\Delta_1} \times \mathrm{BSC}_{\Delta_2} \times \mathrm{BSC}_{\Delta_3}) \circ B) \le \lambda^2 \sum_{i\in[3]} C_{\mathrm{SKL}}(\mathrm{BSC}_{\Delta_i}). \tag{7.28}$$

Conjecture 7.6. For $\lambda, \theta_1, \theta_2, \theta_3 \in [0,1]$, the following inequality holds:

$$\begin{aligned} &\frac{1}{4}(G_\lambda(\theta_1,\theta_2,\theta_3) + G_\lambda(\theta_1,-\theta_2,\theta_3) + G_\lambda(\theta_1,\theta_2,-\theta_3) + G_\lambda(\theta_1,-\theta_2,-\theta_3)) \\ \le\ &\lambda \sum_{i\in[3]} \theta_i \operatorname{arctanh} \theta_i, \end{aligned} \tag{7.29}$$

where

$$G_\lambda(\theta_1,\theta_2,\theta_3) := (\theta_1+\theta_2+\theta_3+\theta_1\theta_2\theta_3) F_\lambda(\theta_1,\theta_2,\theta_3), \tag{7.30}$$

$$F_\lambda(\theta_1,\theta_2,\theta_3) := \operatorname{arctanh} \frac{\lambda(\theta_1+\theta_2+\theta_3+\theta_1\theta_2\theta_3)}{1+\lambda(\theta_1\theta_2+\theta_2\theta_3+\theta_3\theta_1)}. \tag{7.31}$$

Proof of Lemma 7.5. We expand LHS of (7.28) using BP recursion formula established in Section 7.2. Let $\theta_i = 1-2\Delta_i$ for $i \in [3]$. Then

$$\begin{aligned} &C_{\mathrm{SKL}}((\mathrm{BSC}_{\Delta_1} \times \mathrm{BSC}_{\Delta_2} \times \mathrm{BSC}_{\Delta_3}) \circ B) \\ =\ &\sum_{x_1=+,x_2,x_3\in\{\pm\}} \frac{1}{4}\lambda(\theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \theta_1\theta_2\theta_3 x_1 x_2 x_3) \\ &\cdot \operatorname{arctanh} \frac{\lambda(\theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \theta_1\theta_2\theta_3 x_1 x_2 x_3)}{1+\lambda(\theta_1 x_1 \theta_2 x_2 + \theta_2 x_3 \theta_3 x_3 + \theta_3 x_3 \theta_1 x_1)} \\ =\ &\frac{\lambda}{4}(G_\lambda(\theta_1,\theta_2,\theta_3) + G_\lambda(\theta_1,-\theta_2,\theta_3) + G_\lambda(\theta_1,\theta_2,-\theta_3) + G_\lambda(\theta_1,-\theta_2,-\theta_3)) \end{aligned} \tag{7.32}$$

where

$$G_\lambda(\theta_1,\theta_2,\theta_3) := (\theta_1+\theta_2+\theta_3+\theta_1\theta_2\theta_3) F_\lambda(\theta_1,\theta_2,\theta_3), \tag{7.33}$$

$$F_\lambda(\theta_1,\theta_2,\theta_3) := \operatorname{arctanh} \frac{\lambda(\theta_1+\theta_2+\theta_3+\theta_1\theta_2\theta_3)}{1+\lambda(\theta_1\theta_2+\theta_2\theta_3+\theta_3\theta_1)}. \tag{7.34}$$

If Conjecture 7.6 holds, then

$$\begin{aligned} &\frac{1}{4}(G_\lambda(\theta_1,\theta_2,\theta_3) + G_\lambda(\theta_1,-\theta_2,\theta_3) + G_\lambda(\theta_1,\theta_2,-\theta_3) + G_\lambda(\theta_1,-\theta_2,-\theta_3)) \\ \le\ &\lambda \sum_{i\in[3]} \theta_i \operatorname{arctanh} \theta_i \\ =\ &\lambda \sum_{i\in[3]} C_{\mathrm{SKL}}(\mathrm{BSC}_{\Delta_i}). \end{aligned} \tag{7.35}$$

Combining (7.32)(7.35) we finish the proof. □

We remark that for $r \ge 5$ we have found counterexamples to (7.16) and (7.17). Therefore the SKL contraction method does not seem to be able to give tight reconstruction thresholds for $r \ge 5$.

7.3.3 Proof of Theorem 7.1(i)(ii)

With Lemma 7.3 and 7.5, we can prove Theorem 7.1(i)(ii).

Proof of Theorem 7.1(i)(ii). Under the conditions in Theorem 7.1(i)(ii), (7.17) holds by Lemma 7.3 and 7.5. By the above discussion, (7.14) holds. We have $C_{\mathrm{SKL}}(M_1) = C_{\mathrm{SKL}}(\mathrm{BP}(\mathrm{Id})) < \infty$. By (7.14), we have

$$C_{\mathrm{SKL}}(M_k) = C_{\mathrm{SKL}}(\mathrm{BP}^{k-1}(M_1)) = ((r-1)d\lambda^2)^{k-1} C_{\mathrm{SKL}}(M_1). \tag{7.36}$$

So

$$\lim_{k\to\infty} C_{\mathrm{SKL}}(M_k) = 0. \tag{7.37}$$

Finally,

$$\lim_{k\to\infty} I(\sigma_\rho; T_k, \sigma_{L_k}) = \lim_{k\to\infty} C(M_k) \le \lim_{k\to\infty} C_{\mathrm{SKL}}(M_k) = 0. \tag{7.38}$$

So the BOHT model admits non-reconstruction. □

7.4 Reconstruction threshold for large degree

In this section we prove Theorem 7.1(iv). Our proof is an analysis of evolution of χ^2-capacity (also called magnetization in literature) and Gaussian approximation for large degree.

7.4.1 Behavior of χ^2-capacity

Proposition 7.7 (Large degree asymptotics)**.** *Fix $r \in \mathbb{Z}_{\ge 2}$. For any $\epsilon > 0$, there exists $d_0 = d_0(r, \epsilon) > 0$ such that for any $d \ge d_0$ and $\lambda \in [0, 1]$ with $(r-1)d\lambda^2 \le 1$, for any BMS channel P we have*

$$|C_{\chi^2}(\mathrm{BP}(P)) - g_{r,d,\lambda}(C_{\chi^2}(P))| \le \epsilon, \tag{7.39}$$

where

$$g_{r,d,\lambda}(x) := \mathbb{E}_{Z\sim\mathcal{N}(0,1)} \tanh\left(s_{r,d,\lambda}(x) + \sqrt{s_{r,d,\lambda}(x)}Z\right), \tag{7.40}$$

$$s_{r,d,\lambda}(x) := d\lambda^2 \cdot \frac{1}{2}\left((1+x)^{r-1} - (1-x)^{r-1}\right). \tag{7.41}$$

The rest of this section is devoted to the proof of Prop. 7.7.

We first describe $\mathrm{BP}(P)$ in terms of the θ-component. Let P be a BMS channel and P_θ be the θ-component of P. Let t be the offspring ($t = d$ for regular hypertrees, $t \sim \mathrm{Pois}(d)$ for Poisson hypertrees). Let $(\theta_{ij})_{i\in[t], j\in[r-1]}$ generated $\overset{\text{i.i.d.}}{\sim} P_\theta$, where θ_{ij} is the θ-component of the j-th vertex in the i-th downward hyperedge. Let θ_i be

the θ-component of i-th hyperedge $P^{\times(r-1)} \circ B$. As discussed in Section 7.2, given $(\theta_{ij})_{j\in[r-1]}$, θ_i is equal to (the absolute value of)

$$\frac{\lambda\left(\prod_{j\in[r-1]}(1+\theta_{ij}x_{ij})-\prod_{j\in[r-1]}(1-\theta_{ij}x_{ij})\right)}{\lambda\left(\prod_{j\in[r-1]}(1+\theta_{ij}x_{ij})+\prod_{j\in[r-1]}(1-\theta_{ij}x_{ij})\right)+2(1-\lambda)}. \tag{7.42}$$

with probability

$$\lambda\left(\prod_{j\in[r-1]}\left(\frac{1}{2}+\frac{1}{2}\theta_{ij}x_{ij}\right)+\prod_{j\in[r-1]}\left(\frac{1}{2}-\frac{1}{2}\theta_{ij}x_{ij}\right)\right)+2^{2-r}(1-\lambda) \tag{7.43}$$

for $(x_{ij})_{j\in[r-1]} \in \{\pm\}^{r-1}$, $x_{i1} = +$.

Let $\overline{\theta}$ be the θ-component of the full channel $\mathrm{BP}(P)$. Let $P_{\overline{\theta}}$ denote the distribution of $\overline{\theta}$. Then given $(\theta_i)_{i\in[t]}$, $\overline{\theta}$ is equal to (the absolute value of)

$$\frac{\prod_{i\in[t]}(1+\theta_i x_i)-\prod_{i\in[t]}(1-\theta_i x_i)}{\prod_{i\in[t]}(1+\theta_i x_i)+\prod_{i\in[t]}(1-\theta_i x_i)} \tag{7.44}$$

with probability

$$\prod_{i\in[t]}\left(\frac{1}{2}+\frac{1}{2}\theta_i x_i\right)+\prod_{i\in[t]}\left(\frac{1}{2}-\frac{1}{2}\theta_i x_i\right) \tag{7.45}$$

for $(x_1,\dots,x_t) \in \{\pm\}^t$, $x_1 = +$. In other words,

$$\begin{aligned} & P_{\overline{\theta}|\theta_1,\dots,\theta_i} \qquad (7.46)\\ &= \sum_{(x_1,\dots,x_t)\in\{\pm\}^t}\left(\prod_{i\in[t]}\left(\frac{1}{2}+\frac{1}{2}\theta_i x_i\right)\right)\mathbb{1}\left\{\left|\frac{\prod_{i\in[t]}(1+\theta_i x_i)-\prod_{i\in[t]}(1-\theta_i x_i)}{\prod_{i\in[t]}(1+\theta_i x_i)+\prod_{i\in[t]}(1-\theta_i x_i)}\right|\right\}\\ &= \sum_{(x_1,\dots,x_t)\in\{\pm\}^t}\left(\prod_{i\in[t]}\left(\frac{1}{2}+\frac{1}{2}\theta_i x_i\right)\right)\mathbb{1}\left\{\left|\tanh\left(\sum_{i\in[t]}\operatorname{arctanh}(\theta_i x_i)\right)\right|\right\}. \end{aligned}$$

Write $\widetilde{\theta}_i = \theta_i x_i$. Then $\mathbb{P}[\widetilde{\theta}_i = s\theta_i|\theta_i] = \frac{1}{2}+\frac{1}{2}\theta_i s$ for $s \in \{\pm\}$. So $\widetilde{\theta}_i$ for $i \in [t]$ are iid generated from the same distribution. Let us call this distribution D. Then

$$P_{\overline{\theta}} = \mathbb{E}_t\mathbb{E}_{\widetilde{\theta}_1,\dots,\widetilde{\theta}_t \overset{\text{i.i.d.}}{\sim} D}\mathbb{1}\left\{\left|\tanh\left(\sum_{i\in[t]}\operatorname{arctanh}\widetilde{\theta}_i\right)\right|\right\} \tag{7.47}$$

This allows us to use central limit theorems to control the behavior of $\sum_{i\in[t]}\operatorname{arctanh}\widetilde{\theta}_i$.

Lemma 7.8. *There exists a constant $d_0 = d_0(r) > 0$ such that for any $d > d_0$,*

$\lambda \in [0,1]$ *with* $(r-1)d\lambda^2 \le 1$, *and any BMS channel* P, *we have*

$$\left|C_{\chi^2}(P^{\times(r-1)} \circ B) - s_{r,\lambda}(C_{\chi^2}(P))\right| \le O_r(\lambda^3), \tag{7.48}$$

$$s_{r,\lambda}(x) := \lambda^2 \cdot \frac{1}{2}\left((1+x)^{r-1} - (1-x)^{r-1}\right). \tag{7.49}$$

where O_r *hides a multiplicative factor depending only on* r.

Proof. We have

$$\begin{aligned}
&C_{\chi^2}(P^{\times(r-1)} \circ B) \\
&= \mathbb{E}\theta_i^2 \\
&= \mathop{\mathbb{E}}_{\theta_{i1},\ldots,\theta_{i,r-1} \overset{\text{i.i.d.}}{\sim} P_\theta} \sum_{\substack{(x_{ij})_{j\in[r-1]} \in \{\pm\}^{r-1} \\ x_{i1}=+}} \\
&\qquad \left(2^{1-r} \cdot \frac{\lambda^2\left(\prod_{j\in[r-1]}(1+\theta_{ij}x_{ij}) - \prod_{j\in[r-1]}(1-\theta_{ij}x_{ij})\right)^2}{\lambda\left(\prod_{j\in[r-1]}(1+\theta_{ij}x_{ij}) + \prod_{j\in[r-1]}(1-\theta_{ij}x_{ij})\right) + 2(1-\lambda)}\right) \\
&= \mathop{\mathbb{E}}_{\theta_{i1},\ldots,\theta_{i,r-1} \overset{\text{i.i.d.}}{\sim} P_\theta} \sum_{\substack{(x_{ij})_{j\in[r-1]} \in \{\pm\}^{r-1} \\ x_{i1}=+}} \\
&\qquad \left(2^{-r}\lambda^2\left(\prod_{j\in[r-1]}(1+\theta_{ij}x_{ij}) - \prod_{j\in[r-1]}(1-\theta_{ij}x_{ij})\right)^2\right) + O_r(\lambda^3).
\end{aligned}$$

The inner summation satisfies

$$\begin{aligned}
&\sum_{\substack{(x_{ij})_{j\in[r-1]} \in \{\pm\}^{r-1} \\ x_{i1}=+}} \left(\prod_{j\in[r-1]}(1+\theta_{ij}x_{ij}) - \prod_{j\in[r-1]}(1-\theta_{ij}x_{ij})\right)^2 \\
&= \frac{1}{2} \sum_{(x_{ij})_{j\in[r-1]} \in \{\pm\}^{r-1}} \left(\prod_{j\in[r-1]}(1+\theta_{ij}x_{ij}) - \prod_{j\in[r-1]}(1-\theta_{ij}x_{ij})\right)^2 \\
&= \frac{1}{2} \sum_{(x_{ij})_{j\in[r-1]} \in \{\pm\}^{r-1}} \left(\prod_{j\in[r-1]}\left(1+2\theta_{ij}x_{ij}+\theta_{ij}^2\right) - 2\prod_{j\in[r-1]}\left(1-\theta_{ij}^2\right)\right. \\
&\qquad \left. + \prod_{j\in[r-1]}\left(1-2\theta_{ij}x_{ij}+\theta_{ij}^2\right)\right) \\
&= 2^{r-1}\left(\prod_{j\in[r-1]}(1+\theta_{ij}^2) - \prod_{j\in[r-1]}(1-\theta_{ij}^2)\right).
\end{aligned}$$

Therefore

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\theta_{i1},\dots,\theta_{i,r-1}\overset{\text{i.i.d.}}{\sim}P_\theta} \sum_{\substack{(x_{ij})_{j\in[r-1]}\in\{\pm\}^{r-1}\\ x_{i1}=+}} \left(\prod_{j\in[r-1]}(1+\theta_{ij}x_{ij}) - \prod_{j\in[r-1]}(1-\theta_{ij}x_{ij})\right)^2 \\
&= 2^{r-1} \mathop{\mathbb{E}}_{\theta_{i1},\dots,\theta_{i,r-1}\overset{\text{i.i.d.}}{\sim}P_\theta} \left(\prod_{j\in[r-1]}(1+\theta_{ij}^2) - \prod_{j\in[r-1]}(1-\theta_{ij}^2)\right) \\
&= 2^{r-1}\left((1+C_{\chi^2}(P))^{r-1} - (1-C_{\chi^2}(P))^{r-1}\right).
\end{aligned}$$

Combining everything we finish the proof. □

Lemma 7.9. *There exists a constant $d_0 = d_0(r) > 0$ such that for any $d > d_0$, $\lambda \in [0,1]$ with $(r-1)d\lambda^2 \le 1$, and any BMS channel P, we have*

$$\left|\mathbb{E}\operatorname{arctanh}\widetilde{\theta}_i - s_{r,\lambda}(C_{\chi^2}(P))\right| = O_r(\lambda^3), \tag{7.50}$$

$$\left|\operatorname{Var}(\operatorname{arctanh}\widetilde{\theta}_i) - s_{r,\lambda}(C_{\chi^2}(P))\right| = O_r(\lambda^3). \tag{7.51}$$

Proof. Note that $\theta_i = O_r(\lambda)$ almost surely. When d is large enough, λ is small enough, and $\operatorname{arctanh}\theta_i = \theta_i + O_r(\lambda^3)$ almost surely by Taylor expansion. Then

$$\mathbb{E}\operatorname{arctanh}\widetilde{\theta}_i = \mathbb{E}[\theta_i \operatorname{arctanh}\theta_i] = \mathbb{E}\theta_i^2 + O_r(\lambda^4), \tag{7.52}$$

$$\mathbb{E}(\operatorname{arctanh}\widetilde{\theta}_i)^2 = \mathbb{E}(\operatorname{arctanh}\theta_i)^2 = \mathbb{E}\theta_i^2 + O_r(\lambda^4). \tag{7.53}$$

By Lemma 7.8, we have

$$\mathbb{E}\theta_i^2 = s_{r,\lambda}(C_{\chi^2}(P)) + O_r(\lambda^3). \tag{7.54}$$

This already implies the statement on $\mathbb{E}\operatorname{arctanh}\widetilde{\theta}_i$. For the statement on $\operatorname{Var}(\operatorname{arctanh}\widetilde{\theta}_i)$, we note that

$$\mathbb{E}\operatorname{arctanh}\widetilde{\theta}_i = s_{r,\lambda}(C_{\chi^2}(P)) + O_r(\lambda^3) = O_r(\lambda^2). \tag{7.55}$$

So

$$\begin{aligned}
\operatorname{Var}(\operatorname{arctanh}\widetilde{\theta}_i) &= \mathbb{E}(\operatorname{arctanh}\widetilde{\theta}_i)^2 - \left(\mathbb{E}\operatorname{arctanh}\widetilde{\theta}_i\right)^2 \\
&= s_{r,\lambda}(C_{\chi^2}(P)) + O_r(\lambda^3).
\end{aligned} \tag{7.56}$$

This finishes the proof. □

Now we recall a normal approximation result from [109, Prop. 5.3]. We only need the scalar version of it.

Lemma 7.10 ([109])**.** *Let $\phi : \mathbb{R} \to \mathbb{R}$ be a thrice differentiable and bounded function with bounded derivatives up to third order. Let $V_1, \dots, V_t \in \mathbb{R}$ be independent random*

real numbers. Suppose there exists deterministic numbers $\mu, \sigma \in \mathbb{R}$ such that the following holds: for some constant $C > 0$, almost surely

$$\max\left\{\left|\sum_{j\in[t]} \mathbb{E}V_j - \mu\right|, \left|\sum_{j\in[t]} \mathrm{Var}(V_j) - \sigma^2\right|\right\} \le Ct^{-1/2}, \tag{7.57}$$

$$\max\left\{|\mu|, |\sigma^2|\right\} \le C, \qquad \max_{j\in[t]} |V_j| \le Ct^{-1/2}. \tag{7.58}$$

Then for any $\epsilon > 0$, there exists $t_0 = t_0(\epsilon, \phi, C)$ such that if $t > t_0$, then

$$\left|\mathbb{E}\phi\left(\sum_{j\in[t]} V_j\right) - \mathbb{E}_{W\sim\mathcal{N}(\mu,\sigma^2)}\phi(W)\right| \le \epsilon. \tag{7.59}$$

We now have everything we need for the proof of Prop. 7.7.

Proof of Prop. 7.7. **Regular hypertree:** Define $\widetilde{\theta}$ as $\mathbb{P}[\widetilde{\theta} = s\theta|\theta] = \frac{1}{2} + \theta s$ for $s \in \{\pm\}$. Then

$$C_{\chi^2}(\mathrm{BP}(P)) = \mathbb{E}\widetilde{\theta} = \mathbb{E}_{\widetilde{\theta}_1,\dots,\widetilde{\theta}_t \overset{\text{i.i.d.}}{\sim} D} \tanh\left(\sum_{i\in[t]} \operatorname{arctanh}\widetilde{\theta}_i\right). \tag{7.60}$$

In fact, the equality is true with tanh replaced by $\tanh^2$. We use the tanh form here because it is slightly simpler.

Now we apply Lemma 7.10 with

$$\phi(x) = \tanh x, \quad V_i = \operatorname{arctanh}\widetilde{\theta}_i, \quad \mu = \sigma^2 = ds_{r,\lambda}(C_{\chi^2}(P)) = s_{r,d,\lambda}(C_{\chi^2}(P)). \tag{7.61}$$

The conditions in Lemma 7.10 are satisfied by Lemma 7.9 and because $\lambda = O(d^{-1/2})$. This finishes the proof.

Poisson hypertree: Fix $\epsilon > 0$. Let $t \sim \mathrm{Pois}(d)$. By Poisson tail bounds, we have $\mathbb{P}[|t - d| > d^{0.6}] < \epsilon/3$ for large enough d (depending only on ϵ). We apply Lemma 7.10 for every $t \in [d - d^{0.6}, d + d^{0.6}]$, with $\mu = \sigma^2 = s_{r,t,\lambda}(C_{\chi^2}(P))$ and error tolerance $\epsilon/3$. Note that

$$|s_{r,d,\lambda}(C_{\chi^2}(P)) - s_{r,t,\lambda}(C_{\chi^2}(P))| = O_r(d^{-0.4}). \tag{7.62}$$

So for d large enough (depending only on ϵ, r), we have

$$|g_{r,d,\lambda}(C_{\chi^2}(P)) - g_{r,t,\lambda}(C_{\chi^2}(P))| \le \epsilon/3 \tag{7.63}$$

by continuity of g_r (Lemma 7.11).

Therefore we have

$$\begin{aligned}
&|C_{\chi^2}(\mathrm{BP}(P)) - g_{r,d,\lambda}(C_{\chi^2}(P))| \\
&= \left|\mathbb{E}_{t\sim\mathrm{Pois}(d)} C_{\chi^2}((P^{\times(r-1)} \circ B)^{\star t}) - g_{r,d,\lambda}(C_{\chi^2}(P))\right| \\
&\le \mathbb{E}_{t\sim\mathrm{Pois}(d)} \mathbb{1}\{|t-d| \le d^{0.6}\} \left|C_{\chi^2}((P^{\times(r-1)} \circ B)^{\star t}) - g_{r,t,\lambda}(C_{\chi^2}(P))\right| \\
&\quad + \mathbb{E}_{t\sim\mathrm{Pois}(d)} \mathbb{1}\{|t-d| \le d^{0.6}\} \left|g_{r,t,\lambda}(C_{\chi^2}(P)) - g_{r,d,\lambda}(C_{\chi^2}(P))\right| \\
&\quad + \mathbb{E}_{t\sim\mathrm{Pois}(d)} \mathbb{1}\{|t-d| > d^{0.6}\} |C_{\chi^2}((P^{\times(r-1)} \circ B)^{\star t}) - g_{r,d,\lambda}(C_{\chi^2}(P))| \\
&\le \epsilon/3 + \epsilon/3 + \epsilon/3 = \epsilon.
\end{aligned}$$

Note that $C_{\chi^2}(P) \in [0,1]$ for any BMS channel P, and $g_{r,d,\lambda}(x) \in [0,1]$ for all $x \in [0,1]$. □

7.4.2 Properties of functions

In this section we state some few properties of important functions. For $r \ge 2$, we define

$$g_r(x) := \mathbb{E}_{Z\sim\mathcal{N}(0,1)} \tanh\left(s_r(x) + \sqrt{s_r(x)}Z\right), \tag{7.64}$$

$$s_r(x) := \frac{1}{2(r-1)}\left((1+x)^{r-1} - (1-x)^{r-1}\right). \tag{7.65}$$

Lemma 7.11. *For any $r \ge 2$, the function g_r is strictly increasing and continuous differentiable on $[0,1]$.*

Proof. Note that $s_r(x)$ is continuous and increasing on $[0,1]$. Therefore it suffices to prove that

$$g(s) := \mathbb{E}_{Z\sim\mathcal{N}(0,1)} \tanh\left(s + \sqrt{s}Z\right) \tag{7.66}$$

is continuous and increasing on $\mathbb{R}_{\ge 0}$. This statement is in fact equivalent to the $q = 2$ case in [126, Lemma 4.4], after a suitable change of variables. □

Lemma 7.12. *For $r \ge 7$, there exists $x \in [0,1]$ such that $g_r(x) > x$.*

Proof. We can numerically verify that $g_7(0.8) > 0.8$. Note that $s_r(0.8)$ is increasing for $r \ge 7$. Therefore for $r \ge 7$, we have $g_r(0.8) \ge g_7(0.8) > 0.8$. □

7.4.3 Proof of Theorem 7.1(iv)

In this section we prove Theorem 7.1(iv).

Proof of Theorem 7.1(iv). Choose $x \in [0,1]$ so that $g_r(x) > x$ via Lemma 7.12. By continuity of g_r (Lemma 7.11), there exists $\epsilon > 0$ such that $g_{r,d,\lambda}(x) > x + \epsilon$ for $(r-1)d\lambda^2 = 1 - \epsilon$. Note that $g_{r,d,\lambda}(x)$'s dependence on d and λ is only through $d\lambda^2$.

Take $d_0 = d_0(r, \epsilon)$ in Prop. 7.7. For any $d > d_0$, choose $\lambda \in [0, 1]$ such that $(r-1)d\lambda^2 = 1 - \epsilon$. By Prop. 7.7, choice of ϵ, and Lemma 7.11, for all BMS P with $C_{\chi^2}(P) \geq x$ we have

$$C_{\chi^2}(\mathrm{BP}(P)) \geq g_{r,d,\lambda}(C_{\chi^2}(P)) - \epsilon \geq x. \tag{7.67}$$

Therefore

$$\lim_{k\to\infty} I_{\chi^2}(\sigma_\rho; T_k, \sigma_{L_k}) = \lim_{k\to\infty} C_{\chi^2}(M_k) \geq x. \tag{7.68}$$

Finally

$$\lim_{k\to\infty} I(\sigma_\rho; T_k, \sigma_{L_k}) \geq \lim_{k\to\infty} \frac{\log e}{2} I_{\chi^2}(\sigma_\rho; T_k, \sigma_{L_k}) \geq \frac{x \log e}{2}, \tag{7.69}$$

where the first step is because $C(P) \geq \frac{\log e}{2} C_{\chi^2}(P)$ for any BMS P. □

7.5 Weak recovery threshold for HSBM

In this section we prove Theorem 7.2 and Theorem 7.1(iii).

Proof of Theorem 7.2. By combining Theorem 7.1(i)(ii) and Theorem 5.15. □

Proof of Theorem 7.1(iii). By [112], there is an algorithm for weak recovery above the Kesten-Stigum threshold. Therefore, by Theorem 5.15, the BOHT model must admit reconstruction above the Kesten-Stigum threshold. □

Theorem 7.1(iii) can also be proved directly via a majority decider. We omit the details.

7.6 Discussions

We have left the $r = 5, 6$ case open in Theorem 7.1. Our preliminary computations suggest that for $r = 5, 6$, there exists an absolute constant $d_0 \in \mathbb{R}_{\geq 0}$ such that the BOHT model has non-reconstruction when $d \geq d_0$ and $(r-1)d\lambda^2 \leq 1$. We believe that a generalization of Sly's method [126, 109] can be used to prove this. In Sly's method, we compute the first few orders of the BP recursion formula. Combined with Gaussian approximation this would imply contraction of χ^2-capacity. One technical challenge is that in the BOHT case we need a two-step application of Sly's method, in contrast with previous works.

Chapter 8

Robust reconstruction for broadcasting on trees

We study the robust reconstruction problem on the broadcasting on trees (BOT) model (Definition 5.9). Recall that the reconstruction problem (Definition 5.14) concerns the possibility of reconstructing the root label σ_ρ from the leaf labels σ_{L_k}. The robust reconstruction problem concerns the possibility of reconstructing the root label from noisy observations of the leaf labels. Seminal work [82] closed the problem for almost all BOT models on bounded degree trees, but left open the case where the broadcasting matrix M contains zeros, and the survey channel is an erasure channel. We present a method that resolves the problem for any reversible BOT model with mild offspring distributions, which include regular trees and Poisson trees. Our proof is based on local subadditivity of χ^2-information under $\star$-convolution (Theorem 8.6), which generalizes our previous result for the FMS channel case (Theorem 3.12). Using a similar method, we establish that boundary irrelevance (Definition 5.17) does not hold between the reconstruction threshold and the Kesten-Stigum threshold. This may be surprising because BI always holds for the Ising model ($q = 2$) by [137].

Chapter outline In Section 8.1, we introduce the problem and state our main results on robust reconstruction and boundary irrelevance. In Section 8.2, we prove local subadditivity of χ^2-informally for general channels. In Section 8.3, we prove our main results.

8.1 Introduction

We start with the definition of robust reconstruction for general broadcasting on hypertrees (BOHT) models.

Definition 8.1 (Robust reconstruction for BOHT model)**.** Consider the BOHT model $\mathrm{BOHT}(T, q, r, \pi, M)$ or $\mathrm{BOHT}(q, r, \pi, M, D)$ (Definition 5.7). Let W be a channel with input alphabet $[q]$. Let $\omega_u \sim W(\cdot|\sigma_u)$ independently for all vertices $u \in V(T)$. We say the BOHT model admits robust reconstruction with respect to W

if

$$\lim_{k\to\infty} I(\sigma_\rho; T_k, \omega_{L_k}) > 0. \tag{8.1}$$

Let $\mathcal{W}$ be a collection of channels with input alphabet $[q]$. We say the BOHT model admits robust reconstruction with respect to $\mathcal{W}$ if it admits robust reconstruction for every non-trivial channel $W \in \mathcal{W}$.

By definition, robust reconstruction with respect to the identity channel Id is equivalent to reconstruction.

In general, robust reconstruction with respect to the collection of all weak enough channels with input alphabet $[q]$ may have peculiarities. See the following example.

Example 8.2. We present a BOT model above the Kesten-Stigum threshold which does not admit robust reconstruction with respect to some non-trivial survey channel. Consider the model $\mathrm{BOT}(q, \pi, M, d)$ where $q = 4$, $\pi = \mathrm{Unif}([q])$, $d = 3$, and

$$M = \begin{bmatrix} 0.4 & 0.4 & 0.1 & 0.1 \\ 0.4 & 0.4 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.4 & 0.4 \\ 0.1 & 0.1 & 0.4 & 0.4 \end{bmatrix}. \tag{8.2}$$

Then $\lambda_2(M) = 0.6$ and $d\lambda_2(M)^2 = 1.08 > 1$. So we are above the Kesten-Stigum threshold. Now consider the survey channel $W : [4] \to [2]$ defined as $W(1|1) = W(1|3) = W(2|2) = W(2|4) = 1$, $W(2|1) = W(2|3) = W(1|2) = W(1|4) = 0$. The BOT model does not admit robust reconstruction with respect to W, because $W \circ M$ is equivalent to the trivial channel.

Seminal work [82] almost closed the robust reconstruction problem on BOT models. Consider the model $\mathrm{BOT}(q, \pi, M, d)$ (Definition 5.9), where $q \in \mathbb{Z}_{\geq 2}$, $\pi \in \mathcal{P}([q])$ has full support, $M : [q] \to [q]$ satisfies $\pi M = \pi$, and $d \in \mathbb{Z}_{\geq 0}$. [82] studied three collections of channels: $\mathcal{W}_1 := \{M^k : k \in \mathbb{Z}_{\geq 0}\}$, $\mathcal{W}_2 := \{P_\lambda : \lambda \in (0, 1]\}$, and $\mathcal{W}_3 := \{\mathrm{EC}_\epsilon : \epsilon \in [0, 1)\}$. Note that each of the three collections are naturally ordered and contains arbitrarily weak channels. [82] showed that

- ([107]) above the Kesten-Stigum threshold ($d\lambda^2 > 1$), the BOT model admits robust reconstruction with respect to each of the three collections;
- below the Kesten-Stigum threshold ($d\lambda^2 < 1$), the BOT model does not admit robust reconstruction with respect to $\mathcal{W}_1$, $\mathcal{W}_2$;
- below the Kesten-Stigum threshold ($d\lambda^2 < 1$), the BOT model does not admit robust reconstruction with respect to $\mathcal{W}_3$, if $M(x, y) > 0$ for all $x, y \in [q]$.

The zero-free condition in the last result is because their proof used contraction of an information quantity which takes infinity value when the posterior distribution given some observation is not of full support. This condition does not seem to be easily removable by modifications of their method. This is a little bit annoying, as the

random coloring model is an important BOT model [125, 20, 59, 75]. In addition, in the view of the boundary irrelevance problem (Definition 5.17), erasure observation is an important class of survey channels, with applications to the mutual information formula (Theorem 5.18).

As discussed above, it is an interesting question to remove the zero-free condition in the robust reconstruction results. Our first result shows that robust reconstruction is impossible below the Kesten-Stigum threshold for reversible BOT models with mild offspring distributions.

Condition 8.3. Let D be a distribution supported on $\mathbb{Z}_{\geq 0}$. We say D is mild if

$$\lim_{t\to\infty} \left(t^9 \mathbb{P}_{b\sim D}[b > t]\right) = 0. \tag{8.3}$$

Theorem 8.4 (Impossibility of robust reconstruction for reversible BOT)**.** *Consider the model* $\mathrm{BOT}(q, \pi, M, D)$ *(Definition 5.9) where* (π, M) *is reversible and* D *satisfies Condition 8.3. If* $d\lambda^2 < 1$*, then there exists* $\epsilon > 0$ *such that for any channel* W *with input alphabet* $[q]$ *and* $I_{\chi^2}(\pi, W) < \epsilon$*, we have*

$$\lim_{k\to\infty} I(\sigma_\rho; T_k, \omega_{L_k}) = 0. \tag{8.4}$$

Our second result is that boundary irrelevance does not hold between the reconstruction threshold and the Kesten-Stigum threshold.

Theorem 8.5 (Failure of boundary irrelevance for reversible BOT)**.** *Consider the model* $\mathrm{BOT}(q, \pi, M, D)$ *(Definition 5.9) where* (π, M) *is reversible and* D *satisfies Condition 8.3. If* $d\lambda^2 < 1$ *and reconstruction is possible, then BI does not hold for weak enough survey channel. That is, there exists* $\epsilon > 0$ *such that for any survey channel* W *with input alphabet* $[q]$ *and* $I_{\chi^2}(\pi, W) \leq \epsilon$*, we have*

$$\lim_{k\to\infty} I(\sigma_\rho; \sigma_{L_k} | T_k, \omega_{T_k}) > 0. \tag{8.5}$$

Combined with results on the reconstruction threshold of the Potts model [126, 109], Theorem 8.5 implies that for $q \geq 4$, there exists λ, d such that BI does not hold for $\mathrm{BOT}(q, \lambda, d)$. This may be surprising, because [137] proved that BI always holds for symmetric Ising models.

8.2 Local subadditivity of χ^2-information

The key ingredient in the proofs of Theorem 8.4 and Theorem 8.5 is a local subadditivity result for χ^2-information. We prove that χ^2-information is almost subadditive when one of the channels is close to trivial. This generalizes our previous result (Theorem 3.12) on FMS channels to general channels.

Theorem 8.6 (Local subadditivity of χ^2-information)**.** *Let* π *be a fixed distribution over a finite alphabet* $\mathcal{X}$ *with full support. For any* $\epsilon > 0$*, channel* $P : \mathcal{X} \to \mathcal{Y}$*,*

$Q : \mathcal{X} \to \mathcal{Z}$ *with* $I_{\chi^2}(\pi, P) \le \epsilon$ *we have*

$$I_{\chi^2}(\pi, P \star Q) \le \left(1 + O_\pi(\epsilon^{1/9})\right) \left(I_{\chi^2}(\pi, P) + I_{\chi^2}(\pi, Q)\right), \tag{8.6}$$

where O_π *hides a constant depending on* π*.*

The rest of this section is devoted to the proof of Theorem 8.6.

We view a pair (π, P) where $\pi \in \mathcal{P}(\mathcal{X})$, $P : \mathcal{X} \to \mathcal{Y}$ as a distribution of posterior distributions (i.e., the distribution of $P_{X|Y=y}$ where $y \sim P_Y$). Let μ be the posterior distribution variable of (π, P) and P_μ be its distribution. Then we have

$$I_{\chi^2}(\pi, P) = \mathbb{E}_{\mu \sim P_\mu} \chi^2(\mu \| \pi). \tag{8.7}$$

For two distributions $\mu, \nu \in \mathcal{P}(\mathcal{X})$, we define $\mu \star_\pi \nu$ as the distribution

$$\mu \star_\pi \nu = \left(\frac{\mu_i \nu_i \pi_i^{-1}}{\sum_{j \in \mathcal{X}} \mu_j \nu_j \pi_j^{-1}} \right)_{i \in \mathcal{X}}. \tag{8.8}$$

Let μ be the posterior distribution variable of (π, P) (with distribution P_μ), and ν be the posterior distribution variable of (π, Q) (with distribution Q_ν). Then the posterior distribution variable ξ of $(\pi, P \star Q)$ satisfies

$$\mathbb{P}(\xi \in \mathcal{E}) = \mathbb{E}_{\substack{\mu \sim P_\mu \\ \nu \sim Q_\nu}} \left[\left(\sum_{j \in \mathcal{X}} \mu_j \nu_j \pi_j^{-1} \right) \mathbb{1}\left\{ \mu \star_\pi \nu \in \mathcal{E} \right\} \right]. \tag{8.9}$$

for any measurable subset $\mathcal{E} \subseteq \mathcal{P}(\mathcal{X})$. Therefore, we have

$$I_{\chi^2}(\pi, P \star Q) = \mathbb{E}_{\substack{\mu \sim P_\mu \\ \nu \sim Q_\nu}} \left[\left(\sum_{j \in \mathcal{X}} \mu_j \nu_j \pi_j^{-1} \right) \chi^2(\mu \star_\pi \nu \| \pi) \right]. \tag{8.10}$$

Lemma 8.7. *Let* π *be a fixed distribution over a finite alphabet* $\mathcal{X}$ *with full support. For any* $\epsilon > 0$*, channel* $P : \mathcal{X} \to \mathcal{Y}$ *and distribution* ν *with* $\chi^2(\nu \| \pi) \le \epsilon$ *we have*

$$\mathbb{E}_{\mu \sim P_\mu} \left[\left(\sum_{j \in \mathcal{X}} \mu_j \nu_j \pi_j^{-1} \right) \chi^2(\mu \star_\pi \nu \| \pi) \right] \le \left(1 + O_\pi(\epsilon^{1/2})\right) \left(I_{\chi^2}(\pi, P) + \chi^2(\nu \| \pi)\right). \tag{8.11}$$

Proof. Write $\mu = \pi(1 + \alpha)$, $\nu = \pi(1 + \beta)$, where $\alpha \in \mathbb{R}^{\mathcal{X}}$ is a random variable, and $\beta \in \mathbb{R}^{\mathcal{X}}$ is fixed. We immediately have

$$\pi[\alpha] = 0, \quad \chi^2(\mu \| \pi) = \pi[\alpha^2], \quad \pi[\beta] = 0, \quad \chi^2(\nu \| \pi) = \pi[\beta^2], \tag{8.12}$$

$$\|\alpha\|_\infty = O_\pi(1), \quad \|\beta\|_\infty = O_\pi(\epsilon^{1/2}). \tag{8.13}$$

Note that because $\mathbb{E}[\nu] = \pi$, we have

$$\mathbb{E}[\alpha] = 0 \in \mathbb{R}^{\mathcal{X}}. \tag{8.14}$$

We have

$$\sum_{j\in\mathcal{X}} \mu_j \nu_j \pi_j^{-1} = 1 + \pi[\alpha\beta], \tag{8.15}$$

$$\mu \star_\pi \nu = \left(\frac{(1+\alpha_i)(1+\beta_i)\pi_i}{1+\pi[\alpha\beta]} \right)_{i\in\mathcal{X}}, \tag{8.16}$$

$$\chi^2(\mu \star_\pi \nu \| \pi) = \frac{\pi[(1+\alpha)^2(1+\beta)^2]}{(1+\pi[\alpha\beta])^2} - 1. \tag{8.17}$$

Therefore

$$\mathbb{E}_{\mu\sim P_\mu}\left[\left(\sum_{j\in\mathcal{X}} \mu_j \nu_j \pi_j^{-1} \right) \chi^2(\mu \star_\pi \nu \| \pi) \right] = \mathbb{E}\left[\frac{\pi[(1+\alpha)^2(1+\beta)^2] - (1+\pi[\alpha\beta])^2}{1+\pi[\alpha\beta]} \right]. \tag{8.18}$$

Note that

$$\pi[(1+\alpha)^2(1+\beta)^2] - (1+\pi[\alpha\beta])^2 = \mathrm{Var}_\pi[(1+\alpha)(1+\beta)] \tag{8.19}$$

is non-negative. By Eq. (8.13), we have

$$|\pi[\alpha\beta]| = O_\pi(\epsilon^{1/2}). \tag{8.20}$$

So

$$\begin{aligned} &\mathbb{E}\left[\frac{\pi[(1+\alpha)^2(1+\beta)^2] - (1+\pi[\alpha\beta])^2}{1+\pi[\alpha\beta]} \right] \\ \le\ & \left(1 + O_\pi(\epsilon^{1/2})\right) \mathbb{E}\left[\pi[(1+\alpha)^2(1+\beta)^2] - (1+\pi[\alpha\beta])^2 \right], \end{aligned} \tag{8.21}$$

and to prove Eq. (8.11), it suffices to prove that

$$\mathbb{E}\left[\pi[(1+\alpha)^2(1+\beta)^2] - (1+\pi[\alpha\beta])^2 \right] \le \left(1 + O_\pi(\epsilon^{1/2})\right) \left(\mathbb{E}\left[\pi[\alpha^2]\right] + \pi[\beta^2] \right). \tag{8.22}$$

We have

$$\begin{aligned} &\mathbb{E}\left[\pi[(1+\alpha)^2(1+\beta)^2] - (1+\pi[\alpha\beta])^2 \right] \\ &= \mathbb{E}\left[\pi[1 + 2\alpha + 2\beta + \alpha^2 + \beta^2 + 4\alpha\beta + 2\alpha^2\beta + 2\alpha\beta^2 + \alpha^2\beta^2] - 1 - \pi[2\alpha\beta] - (\pi[\alpha\beta])^2 \right] \\ &= \mathbb{E}\left[\pi[\alpha^2 + \beta^2 + 2\alpha\beta + 2\alpha^2\beta + 2\alpha\beta^2 + \alpha^2\beta^2] - (\pi[\alpha\beta])^2 \right] \\ &= \mathbb{E}\left[\pi[\alpha^2 + \beta^2 + 2\alpha^2\beta + \alpha^2\beta^2] - (\pi[\alpha\beta])^2 \right]. \end{aligned} \tag{8.23}$$

where the second step is by $\pi[\alpha] = \pi[\beta] = 0$, and the third step is by $\mathbb{E}[\alpha] = 0 \in \mathbb{R}^{\mathcal{X}}$.

Furthermore,

$$\mathbb{E}\left[\pi[2\alpha^2\beta+\alpha^2\beta^2]-(\pi[\alpha\beta])^2\right]\le\mathbb{E}\left[\pi[2\alpha^2|\beta|+\alpha^2|\beta|^2]\right]\le O_\pi(\epsilon^{1/2})\mathbb{E}\left[\pi[\alpha^2]\right]. \quad (8.24)$$

Combining Eq. (8.23) and Eq. (8.24) we finish the proof of Eq. (8.22). □

Lemma 8.8. *Let π be a fixed distribution over a finite alphabet $\mathcal{X}$ with full support. For any $\epsilon>0$, channel $P:\mathcal{X}\to\mathcal{Y}$ and distribution ν with $I_{\chi^2}(\pi,P)\le\epsilon$ we have*

$$\mathbb{E}_{\mu\sim P_\mu}\left[\left(\sum_{j\in\mathcal{X}}\mu_j\nu_j\pi_j^{-1}\right)\chi^2(\mu\star_\pi\nu\|\pi)\right]\le\left(1+O_\pi(\epsilon^{1/9})\right)\left(I_{\chi^2}(\pi,P)+\chi^2(\nu\|\pi)\right). \quad (8.25)$$

Proof. Following notations in proof of Lemma 8.7, write $\mu=\pi(1+\alpha)$, $\nu=\pi(1+\beta)$, where $\alpha\in\mathbb{R}^{\mathcal{X}}$ is a random variable, and $\beta\in\mathbb{R}^{\mathcal{X}}$ is fixed. Note that the bound on $\|\beta\|_\infty$ in Eq. (8.13) no longer holds.

If $\chi^2(\nu\|\pi)\le\epsilon^{2/9}$, then we can apply Lemma 8.7. In the following, assume that

$$\pi[\beta^2]=\chi^2(\nu\|\pi)\ge\epsilon^{2/9}. \quad (8.26)$$

Because

$$\mathbb{E}\left[\pi[\alpha^2]\right]=I_{\chi^2}(\pi,P)\le\epsilon, \quad (8.27)$$

by Markov inequality, we have

$$\mathbb{P}\left[\pi[\alpha^2]\ge\epsilon^{2/3}\right]\le\epsilon^{1/3}. \quad (8.28)$$

So

$$\begin{aligned}&\mathbb{E}_{\mu\sim P_\mu}\left[\left(\sum_{j\in\mathcal{X}}\mu_j\nu_j\pi_j^{-1}\right)\chi^2(\mu\star_\pi\nu\|\pi)\right]\\ =\;&\mathbb{E}_{\mu\sim P_\mu}\left[\left(\sum_{j\in\mathcal{X}}\mu_j\nu_j\pi_j^{-1}\right)\chi^2(\mu\star_\pi\nu\|\pi)\mathbb{1}\left\{\pi[\alpha^2]\le\epsilon^{2/3}\right\}\right]\\ &+\mathbb{E}_{\mu\sim P_\mu}\left[\left(\sum_{j\in\mathcal{X}}\mu_j\nu_j\pi_j^{-1}\right)\chi^2(\mu\star_\pi\nu\|\pi)\mathbb{1}\left\{\pi[\alpha^2]>\epsilon^{2/3}\right\}\right]\\ =:\;&L+R. \end{aligned} \quad (8.29)$$

For R, we have

$$R\le\epsilon^{1/3}\sup_{\mu',\nu'\in\mathcal{P}(\mathcal{X})}\left(\left(\sum_{j\in\mathcal{X}}\mu'_j\nu'_j\pi_j^{-1}\right)\chi^2(\mu'\star_\pi\nu'\|\pi)\right)=O_\pi(\epsilon^{1/3})\le O_\pi(\epsilon^{1/9})\pi[\beta^2], \quad (8.30)$$

where the last step is by Eq. (8.26).

For L, by the same computation as in proof of Lemma 8.7, we have

$$\begin{aligned} L &= \mathbb{E}_{\mu\sim P_\mu}\left[\frac{\pi[(1+\alpha)^2(1+\beta)^2]-(1+\pi[\alpha\beta])^2}{1+\pi[\alpha\beta]}\cdot \mathbb{1}\left\{\pi[\alpha^2]\le\epsilon^{2/3}\right\}\right] \\ &\le \left(1+O_\pi(\epsilon^{1/3})\right)\mathbb{E}_{\mu\sim P_\mu}\left[\left(\pi[(1+\alpha)^2(1+\beta)^2]-(1+\pi[\alpha\beta])^2\right)\mathbb{1}\left\{\pi[\alpha^2]\le\epsilon^{2/3}\right\}\right]. \end{aligned} \tag{8.31}$$

Note that $\pi[\alpha^2]\le\epsilon^{2/3}$ implies that $\|\alpha\|_\infty = O_\pi(\epsilon^{1/3})$.

So it suffices to prove that

$$\begin{aligned} &\mathbb{E}_{\mu\sim P_\mu}\left[\left(\pi[(1+\alpha)^2(1+\beta)^2]-(1+\pi[\alpha\beta])^2\right)\mathbb{1}\left\{\pi[\alpha^2]\le\epsilon^{2/3}\right\}\right] \\ \le\ &\left(1+O_\pi(\epsilon^{1/9})\right)\left(\mathbb{E}\left[\pi[\alpha^2]\right]+\pi[\beta^2]\right). \end{aligned} \tag{8.32}$$

We have

$$\begin{aligned} &\mathbb{E}_{\mu\sim P_\mu}\left[\left(\pi[(1+\alpha)^2(1+\beta)^2]-(1+\pi[\alpha\beta])^2\right)\mathbb{1}\left\{\pi[\alpha^2]\le\epsilon^{2/3}\right\}\right] \\ &= \mathbb{E}\left[\left(\pi[\alpha^2+\beta^2+2\alpha\beta+2\alpha^2\beta+2\alpha\beta^2+\alpha^2\beta^2]-(\pi[\alpha\beta])^2\right)\mathbb{1}\left\{\pi[\alpha^2]\le\epsilon^{2/3}\right\}\right] \\ &\le \mathbb{E}\left[\pi[\alpha^2]\right]+\pi[\beta^2]+\mathbb{E}\left[\left(\pi[2\alpha\beta+2\alpha^2\beta+2\alpha\beta^2+\alpha^2\beta^2]\right)\mathbb{1}\left\{\pi[\alpha^2]\le\epsilon^{2/3}\right\}\right] \\ &= \mathbb{E}\left[\pi[\alpha^2]\right]+\pi[\beta^2]+O_\pi(\epsilon^{1/3}) \\ &= \mathbb{E}\left[\pi[\alpha^2]\right]+\left(1+O_\pi(\epsilon^{1/9})\right)\pi[\beta^2], \end{aligned} \tag{8.33}$$

where where the first step is because $\pi[\alpha]=\pi[\beta]=0$, the third step is because $\|\alpha\|_\infty = O_\pi(\epsilon^{1/3})$, $\|\beta\|_\infty = O_\pi(1)$, and the fourth step is by Eq. (8.26).

By Eq. (8.31) and Eq. (8.33), we have

$$L\le\left(1+O_\pi(\epsilon^{1/9})\right)\left(\mathbb{E}\left[\pi[\alpha^2]\right]+\pi[\beta^2]\right). \tag{8.34}$$

Combining Eq. (8.34) and Eq. (8.30) we finish the proof. □

Proof of Theorem 8.6. By Eq. (8.10) and Lemma 8.8. □

8.3 Proofs of main results

In this section we prove Theorem 8.4 and Theorem 8.5. The proof uses contraction and local subadditivity properties (Theorem 3.12) of χ^2-information.

Lemma 8.9 (Contraction)**.** *Let π be a distribution on a finite alphabet $\mathcal{X}$ and $M:\mathcal{X}\to\mathcal{X}$ be a channel with invariant distribution π. If (π, M) is reversible, then*

$$\eta_{\chi^2}(\pi, M) = \lambda^2, \tag{8.35}$$

where λ is the second largest eigenvalue (in absolute value) of M. In particular, for

any channel P with input alphabet $\mathcal{X}$, we have

$$I_{\chi^2}(\pi, P \circ M) \le \lambda^2 I_{\chi^2}(\pi, P). \tag{8.36}$$

Proof. It is known (e.g., [122, 118]) that $\eta_{\chi^2}(\pi, M)$ is equal to the second largest eigenvalue of MM^*. Because M is reversible, $MM^* = M^2$ and its eigenvalues are squares of eigenvalues of M. This proves Eq. (8.35). Then Eq. (8.36) follows from reversibility of M and definition of contraction coefficient. □

Proof of Theorem 8.4. Let $C_1 > 0$ be the constant in Theorem 8.6, i.e., for all $\epsilon > 0$, channels $P : \mathcal{X} \to \mathcal{Y}$, $Q : \mathcal{X} \to \mathcal{Z}$ with $I_{\chi^2}(\pi, P) \le \epsilon$, we have

$$I_{\chi^2}(\pi, P \star Q) \le \left(1 + C_1 \epsilon^{1/9}\right) \left(I_{\chi^2}(\pi, P) + I_{\chi^2}(\pi, Q)\right). \tag{8.37}$$

Let $C_2 > 0$ be such that $I_{\chi^2}(\pi, P) \le C_2$ for all channels P with input alphabet $[q]$.

Take $c_1, c_2 > 0$ such that

$$\exp(c_1) d\lambda^2 + c_2 < 1. \tag{8.38}$$

Take b_0 (via Condition 8.3) such that for all $t \ge b_0$, we have

$$t^9 \mathbb{P}_{b \sim D}[b > t] < c_2 C_2^{-1} (c_1 C_1^{-1})^9. \tag{8.39}$$

For $\epsilon > 0$, define

$$b(\epsilon) := c_1 C_1^{-1} \epsilon^{-1/9}. \tag{8.40}$$

Take $\epsilon_0 > 0$ such that $b(\epsilon_0) > b_0$.

We prove that for any channel P with input alphabet $[q]$ and $I_{\chi^2}(\pi, P) \le \epsilon_0$, we have

$$I_{\chi^2}(\pi, \mathrm{BP}(P)) \le \left(\exp(c_1) d\lambda^2 + c_2\right) I_{\chi^2}(\pi, P). \tag{8.41}$$

Fix such a channel P. Let $\epsilon := I_{\chi^2}(\pi, P)$. By Lemma 8.9, we have

$$I_{\chi^2}(\pi, P \circ M) \le \lambda^2 \epsilon. \tag{8.42}$$

We have

$$\begin{aligned} I_{\chi^2}(\pi, \mathrm{BP}(P)) &= \mathbb{E}_{b \sim D} I_{\chi^2}(\pi, (P \circ M)^{\star b}) \\ &= \mathbb{E}_{b \sim D}\left[I_{\chi^2}(\pi, (P \circ M)^{\star b}) \mathbb{1}\{b \le b(\epsilon)\}\right] \\ &\quad + \mathbb{E}_{b \sim D}\left[I_{\chi^2}(\pi, (P \circ M)^{\star b}) \mathbb{1}\{b > b(\epsilon)\}\right] \\ &=: L + R. \end{aligned} \tag{8.43}$$

For L, by induction on b we have

$$I_{\chi^2}(\pi, (P \circ M)^{\star b}) \le (1 + C_1 \epsilon^{1/9})^b b I_{\chi^2}(\pi, P \circ M) \le \exp(C_1 \epsilon^{1/9} b) b \lambda^2 \epsilon. \tag{8.44}$$

Therefore

$$\begin{aligned} L &\le \mathbb{E}_{b\sim D}\left[\exp(C_1\epsilon^{1/9}b)b\lambda^2\epsilon\mathbb{1}\{b\le b(\epsilon)\}\right] \\ &\le \mathbb{E}_{b\sim D}\left[\exp(C_1\epsilon^{1/9}b(\epsilon))b\lambda^2\epsilon\right] \\ &\le \exp(C_1\epsilon^{1/9}b(\epsilon))d\lambda^2\epsilon \\ &\le \exp(c_1)d\lambda^2\epsilon. \end{aligned} \tag{8.45}$$

For R we have

$$R \le \mathbb{E}_{b\sim D}\left[C_2\cdot c_2C_2^{-1}\left(c_1C_1^{-1}\right)^9\cdot b(\epsilon)^{-9}\right] \le c_2\epsilon. \tag{8.46}$$

Eq. (8.43), Eq. (8.45) and Eq. (8.46) together imply Eq. (8.41).

Let M_k be the channel $\sigma_\rho \mapsto (T_k, \omega_{L_k})$. Then we have $M_{k+1} = \mathrm{BP}(M_k)$. By Eq. (8.41), Eq. (8.38) and the fact that $I_{\chi^2}(\pi, M_0) < \infty$, we finish the proof of the theorem. □

Proof of Theorem 8.5 is based on the following variant of Theorem 8.4.

Proposition 8.10. *In the setting of Theorem 8.5, for all $\delta > 0$, there exists $\epsilon > 0$ such that for any channel W with input alphabet $[q]$ and $I_{\chi^2}(\pi, W) \le \epsilon$, we have*

$$\lim_{k\to\infty} I_{\chi^2}(\sigma_\rho; \omega_{T_k}|T_k) \le \delta. \tag{8.47}$$

Proof of Theorem 8.5 given Prop. 8.10. In the reconstruction regime,

$$\lim_{k\to\infty} I(\sigma_\rho; \sigma_{L_k}, \omega_{T_k}|T_k) \ge \lim_{k\to\infty} I(\sigma_\rho; \sigma_{L_k}|T_k) > 0. \tag{8.48}$$

Take $\delta > 0$ such that $\delta\log 2 < \lim_{k\to\infty} I(\sigma_\rho; \sigma_{L_k}|T_k)$. Because $I \le I_{\chi^2}\log 2$, and by Prop. 8.10, for weak enough survey channel W we have

$$\lim_{k\to\infty} I(\sigma_\rho; \omega_{T_k}|T_k) \le \lim_{k\to\infty} I_{\chi^2}(\sigma_\rho; \omega_{T_k}|T_k)\log 2 \le \delta\log 2 < \lim_{k\to\infty} I(\sigma_\rho; \sigma_{L_k}, \omega_{T_k}|T_k). \tag{8.49}$$

Therefore

$$\begin{aligned} \lim_{k\to\infty} I(\sigma_\rho; \sigma_{L_k}|T_k, \omega_{T_k}) &= \lim_{k\to\infty}\left(I(\sigma_\rho; \sigma_{L_k}, \omega_{T_k}|T_k) - I(\sigma_\rho; \omega_{T_k}|T_k)\right) \\ &= \lim_{k\to\infty} I(\sigma_\rho; \sigma_{L_k}, \omega_{T_k}|T_k) - \lim_{k\to\infty} I(\sigma_\rho; \omega_{T_k}|T_k) > 0. \end{aligned} \tag{8.50}$$

□

Proof of Prop. 8.10. Take $C_1, C_2, c_1, c_2, b_0, b(\epsilon), \epsilon_0$ as in proof of Theorem 8.4. Take $\epsilon_1 > 0$ such that $\epsilon_1 < \min\{\epsilon_0, \delta\}$. Take $\epsilon > 0$ such that $\epsilon < \epsilon_1$ and

$$(\exp(c_1)d\lambda^2 + c_2)\epsilon_1 + \exp(c_1)\epsilon < \epsilon_1. \tag{8.51}$$

Let W (resp. P) be an arbitrary channel with input alphabet $[q]$ satisfying $I_{\chi^2}(\pi, W) \leq \epsilon$ (resp. $I_{\chi^2}(\pi, P) \leq \epsilon_1$). We have

$$\begin{aligned} I_{\chi^2}(\pi, \mathrm{BP}_W(P)) &= \mathbb{E}_{b\sim D} I_{\chi^2}(\pi, (P \circ M)^{\star b} \star W) \\ &= \mathbb{E}_{b\sim D}\left[I_{\chi^2}(\pi, (P \circ M)^{\star b} \star W)\mathbb{1}\{b \leq b(\epsilon_1)\}\right] \\ &\quad + \mathbb{E}_{b\sim D}\left[I_{\chi^2}(\pi, (P \circ M)^{\star b} \star W)\mathbb{1}\{b > b(\epsilon_1)\}\right] \\ &=: L + R. \end{aligned} \tag{8.52}$$

For L, by induction on b we have

$$I_{\chi^2}(\pi, (P \circ M)^{\star b} \star W) \leq (1 + C_1\epsilon_1)^b (b I_{\chi^2}(\pi, P \circ M) + \epsilon) \leq (1+\epsilon)^b (b\lambda^2\epsilon_1 + \epsilon). \tag{8.53}$$

Then with a computation similar to Eq. (8.45) we have

$$L \leq \exp(c_1)(d\lambda^2\epsilon_1 + \epsilon). \tag{8.54}$$

For R, with the same computation as Eq. (8.46)

$$R \leq c_2\epsilon_1. \tag{8.55}$$

Combining Eq. (8.52), Eq. (8.54), Eq. (8.55) we get

$$I_{\chi^2}(\pi, \mathrm{BP}_W(P)) \leq (\exp(c_1)d\lambda^2 + c_2)\epsilon_1 + \exp(c_1)\epsilon < \epsilon_1. \tag{8.56}$$

Let M_k be the channel $\sigma_\rho \mapsto (T_k, \omega_{T_k})$. Then $M_{k+1} = \mathrm{BP}_W(M_k)$. By Eq. (8.56) and $\epsilon \leq \epsilon_1$ we see that

$$I_{\chi^2}(\pi, M_k) \leq \epsilon_1 < \delta \tag{8.57}$$

for all k. This finishes the proof. □

Chapter 9

Computation of belief propagation limit

We consider the problem of computing the limit information, and more generally, the limit channel, in the symmetric Ising model. Computation of the belief propagation limit in BOT models is a non-trivial task. Exact computation takes time doubly exponential in depth, which is unacceptable in practice. The widely used population dynamics, while running very fast (linear in sample size), does not have sufficient correctness guarantees as depth goes large. We introduce a method for bounding the limit information based on the less-noisy preorder, which, in its most basic form, is able to recover the reconstruction threshold for the symmetric Ising model. We further refine this method using local comparison of BMS channels via channel preorders, which gives us rigorous and very good bounds on the limit information.

This chapter is based on [75].

Chapter outline In Section 9.1, we introduce the problem and our method based on channel comparison. In Section 9.2, we derive the reconstruction threshold for the symmetric Ising model using (global) less-noisy comparison. In Section 9.3, we derive bounds on the limit mutual information using global comparison. In Section 9.4, we introduce the local comparison method and present a few numerical results.

9.1 Introduction

9.1.1 Broadcasting on trees

We consider the symmetric Ising model on a regular tree with offspring $d \in \mathbb{Z}_{\geq 0}$ (denoted $\mathrm{BOT}(2, \lambda, d)$ as in Definition 5.9). In this model, a binary signal propagates from the root ρ downwards the tree through BSC_δ channels (where $\lambda = 1 - 2\delta$). Let L_k be the set of nodes at distance k to ρ and M_k be the channel $\sigma_\rho \mapsto \sigma_{L_k}$. We would like to compute the limit mutual information

$$I(\delta) := \lim_{k \to \infty} C(M_k) \tag{9.1}$$

and the limit probability of error

$$P_e(\delta) := \lim_{k\to\infty} P_e(M_k), \tag{9.2}$$

where $C(\cdot)$ and $P_e(\cdot)$ are defined in Definition 2.4. The model admits reconstruction if and only if $I(\delta) > 0$ (equivalently, $P_e(\delta) < \frac{1}{2}$). Foundational work [23, 63] established that reconstruction is possible if and only if

$$\delta < \delta_c := \frac{1}{2}\left(1 - d^{-1/2}\right). \tag{9.3}$$

We note that the positive part (i.e., $P_e < \frac{1}{2}$ when $\delta < \delta_c$) follows from the Kesten-Stigum threshold [84], which says that reconstruction can be achieved using a suboptimal estimator which outputs the majority of labels in L_k.

The computation of limit information and limit probability of error is a non-trivial task. Computing these quantities exactly takes time doubly exponential in depth k, which is totally unacceptable. A method commonly used in practice is the population dynamics [10, 99, 98]. In this method, one maintains a collection of samples from the distribution of belief propagation messages, and approximates the true BP message distribution using these samples. When the sample size is M, this method takes $O(kM)$ time and $O(M)$ space to compute an approximation up to level k. While the computation cost of the population dynamics is small (when the sample size is not too large), theoretical guarantees of the approximation accuracy as $k \to \infty$ are limited. In this section, we propose a method based on local comparison of BMS channels, which gives rigorous and quite good bounds on the limit information and limit probability of error.

To show the power of our method, we apply it to the Ising model near criticality. Various theories (starting from Ginzburg-Landau) in statistical physics predict the behavior of various quantities in he vicinity of the phase transition (called critical exponents). However, before our work [75], behavior of $I(\delta)$ and $P_e(\delta)$ near the critical point $\delta = \delta_c - \tau$ with $\tau \ll 1$ was not understood: the only known results were

$$c_1\tau + o(\tau) \le I(\delta_c - \tau) \le c_2\tau + o(\tau), \tag{9.4}$$

$$\frac{1}{2} - c_3\sqrt{\tau} + o(\sqrt{\tau}) \le P_e(\delta_c - \tau) \le \frac{1}{2} - c_4\tau + o(\tau)\,. \tag{9.5}$$

for some $0 < c_1 < c_2$ and $c_3, c_4 > 0$. Using the local comparison method, we were able to provide rigorous bounds on I and P_e, which allowed us to conjecture that on binary trees (i.e., $d = 2$)

$$I(\delta_c - \tau) = (4\sqrt{2} + o(1))\tau, \quad P_e(\delta_c - \tau) = \frac{1}{2} - \Theta(\sqrt{\tau}). \tag{9.6}$$

Our conjectures were later proved in [136].

9.1.2 Channel comparison method

The key idea of our method is very simple. Recall that the channels $(M_k)_{k\geq 0}$ satisfies the belief propagation recursion

$$M_{k+1} = \mathrm{BP}(M_k), \quad \mathrm{BP}(P) := (P \circ \mathrm{BSC}_\delta)^{\star d}. \tag{9.7}$$

By Lemma 2.8, the BP operator preserves degradation preorder and less-noisy preorder. Therefore, if we have quantization operators $\underline{Q}, \overline{Q} : \{\text{BMSs}\} \to \{\text{BMSs}\}$ satisfying

$$\underline{Q}(P) \leq_{\deg} P \leq_{\deg} \overline{Q}(P) \tag{9.8}$$

for all BMS channels P, then the quantized BP operators

$$\underline{\mathrm{BP}} := \underline{Q} \circ \mathrm{BP}, \quad \overline{\mathrm{BP}} := \overline{Q} \circ \mathrm{BP} \tag{9.9}$$

satisfy

$$\underline{\mathrm{BP}}(P) \leq_{\deg} \mathrm{BP}(P) \leq_{\deg} \overline{\mathrm{BP}}(P) \tag{9.10}$$

for all BMS channels P. By iterating, we have

$$\underline{\mathrm{BP}}^k(P) \leq_{\deg} \mathrm{BP}^k(P) \leq_{\deg} \overline{\mathrm{BP}}^k(P) \tag{9.11}$$

for all $k \geq 0$.

The above discussions still hold if we replace all $\leq_{\deg}$ by $\leq_{\ln}$.

Recall Lemma 2.10, which states that

1. among all BMS channels with χ^2-capacity c, the least noisy one is BEC_{1-c} and the most noisy one is $\mathrm{BSC}_{1/2-\sqrt{c}/2}$;

2. among all BMS channels with probability of error p, the least degraded one is BEC_{2p} and the most degraded one is BSC_p.

Therefore, one natural idea is to let $\overline{Q}$ output the unique BEC with the same probability of error (resp. χ^2-capacity), and let $\underline{Q}$ output the unique BSC with the same probability of error (resp. χ^2-capacity). These choices lead to the following results.

Proposition 9.1. *Let $\overline{Q}_{\deg}$ (resp. $\underline{Q}_{\deg}$) be the operator which maps a BMS channel P to BEC_{2p} (resp. BSC_p), where $p = P_e(P)$. Let $\underline{\mathrm{BP}}_{\deg} := \underline{Q}_{\deg} \circ \mathrm{BP}$, $\overline{\mathrm{BP}}_{\deg} := \overline{Q}_{\deg} \circ \mathrm{BP}$. Then for any $k \geq 0$,*

$$\underline{\mathrm{BP}}^k_{\deg}(\mathrm{Id}) \leq_{\deg} \mathrm{BP}^k(\mathrm{Id}) \leq_{\deg} \overline{\mathrm{BP}}^k_{\deg}(\mathrm{Id}). \tag{9.12}$$

Proposition 9.2. *Let $\overline{Q}_{\ln}$ (resp. $\underline{Q}_{\ln}$) be the operator which maps a BMS channel P to BEC_{1-c} (resp. $\mathrm{BSC}_{1/2-\sqrt{c}/2}$), where $c = C_{\chi^2}(P)$. Let $\underline{\mathrm{BP}}_{\ln} := \underline{Q}_{\ln} \circ \mathrm{BP}$, $\overline{\mathrm{BP}}_{\ln} :=$*

$\overline{Q}_{\ln} \circ \mathrm{BP}$. *Then for any* $k \geq 0$,

$$\underline{\mathrm{BP}}^k_{\ln}(\mathrm{Id}) \leq_{\ln} \mathrm{BP}^k(\mathrm{Id}) \leq_{\ln} \overline{\mathrm{BP}}^k_{\ln}(\mathrm{Id}). \tag{9.13}$$

Proofs of Prop. 9.1 and Prop. 9.2 are by combining Lemma 2.8 and Lemma 2.10.

We note that the $\underline{\mathrm{BP}}$ operators in Prop. 9.1 and Prop. 9.2 output BSC channels, and the $\overline{\mathrm{BP}}$ operators output BEC channels. Therefore computations of $\underline{\mathrm{BP}}^k(\mathrm{Id})$ and $\overline{\mathrm{BP}}^k(\mathrm{Id})$ can be considered as evolutions of a single real number. This allows us the analyze the evolutions analytically.

We call the above method the global comparison method because probability mass of the Δ-distribution (recall Lemma 2.2) is moved to a single point (BSCs) for the lower bound, and moved to 0 and $\frac{1}{2}$ (BECs) for the upper bound. In Section 9.4, we will refine this method to the local comparison method, by replacing masses in small subintervals of $[0, \frac{1}{2}]$ separately. While the global comparison method is already poweful and can recover the reconstruction threshold, we show that the local comparison method can give almost tight bounds on the limit information and limit probability of error.

9.2 The reconstruction threshold

In this section we prove the reconstruction threshold using the global less-noisy comparison method (Prop. 9.2).

Proposition 9.3. *Consider the model* $\mathrm{BOT}(2, \lambda = 1 - 2\delta, d)$. *If* $d\lambda^2 > 1$, *then reconstruction is possible.*

Proof. By Prop. 9.2, it suffices to show that there exists $\epsilon > 0$ such that

$$C_{\chi^2}(\mathrm{BP}(\mathrm{BSC}_{1/2-\epsilon})) \geq C_{\chi^2}(\mathrm{BSC}_{1/2-\epsilon}). \tag{9.14}$$

In fact, suppose that Eq. (9.14) holds. Write $\underline{\mathrm{BP}}^k_{\ln}(\mathrm{Id}) = \mathrm{BSC}_{1/2-\epsilon_k}$. By induction on k we have $\epsilon_k \geq \epsilon$ for all $k > 0$. Therefore

$$\mathrm{BP}^k(\mathrm{Id}) \geq_{\ln} \underline{\mathrm{BP}}^k_{\ln}(\mathrm{Id}) \geq_{\ln} \mathrm{BSC}_{1/2-\epsilon_k} \tag{9.15}$$

for all $k \geq 0$ and reconstruction holds.

In the following we prove Eq. (9.14). Note that $\mathrm{BSC}_{1/2-\epsilon} \circ \mathrm{BSC}_\delta = \mathrm{BSC}_\kappa$ where $\kappa = \frac{1}{2} - (1-2\delta)\epsilon$. Then we have

$$\begin{aligned} C_{\chi^2}(\mathrm{BP}(\mathrm{BSC}_{1/2-\epsilon})) &= C_{\chi^2}(\mathrm{BSC}_\kappa^{\star d}) \\ &= 2 \sum_{0 \leq i \leq d} \binom{d}{i} \cdot \frac{\kappa^{2i}(1-\kappa)^{2(d-i)}}{\kappa^i(1-\kappa)^{d-i} + \kappa^{d-i}(1-\kappa)^i} - 1. \end{aligned} \tag{9.16}$$

Using

$$\kappa^a(1-\kappa)^b+\kappa^b(1-\kappa)^a=2^{1-a-b}\left(1+\left(\binom{a}{2}+\binom{b}{2}-ab\right)4(1-2\delta)^2\epsilon^2+O(\epsilon^4)\right), \tag{9.17}$$

we can expand Eq. (9.16) in terms of ϵ and get

$$\begin{aligned} C_{\chi^2}(\mathrm{BP}(\mathrm{BSC}_{1/2-\epsilon}))=&\sum_{0\le i\le d}\binom{d}{i}2^{-d}\left(1+\left(\binom{2i}{2}+\binom{2(d-i)}{2}-4i(d-i)\right.\right.\\ &\left.\left.-\binom{i}{2}-\binom{d-i}{2}+i(d-i)\right)4(1-2\delta)^2\epsilon^2\right)+O(\epsilon^4)-1\\ =&\,4d(1-2\delta)^2\epsilon^2+O(\epsilon^4). \end{aligned} \tag{9.18}$$

Note that $C_{\chi^2}(\mathrm{BSC}_{1/2-\epsilon})=4\epsilon^2$. Therefore when $\epsilon>0$ is small enough, Eq. (9.14) holds. This finishes the proof. □

Likewise, comparison with BECs leads to tight non-reconstruction result.

Proposition 9.4. *Consider the model* $\mathrm{BOT}(2,\lambda=1-2\delta,d)$. *If* $d\lambda^2\le 1$ *and* $(d,\delta)\neq(1,0)$, *then reconstruction is impossible.*

Proof. By Prop. 9.2, it suffices to show that for all $0<\epsilon\le 1$, we have

$$C_{\chi^2}(\mathrm{BP}(\mathrm{BEC}_{1-\epsilon}))<C_{\chi^2}(\mathrm{BEC}_{1-\epsilon}). \tag{9.19}$$

In fact, suppose that Eq. (9.19) holds. Write $\underline{\mathrm{BP}}^k_{\ln}(\mathrm{Id})=\mathrm{BEC}_{1-\epsilon_k}$. Define function $g:[0,1]\to[0,1]$ as

$$g(\epsilon):=C_{\chi^2}(\mathrm{BP}(\mathrm{BEC}_{1-\epsilon})). \tag{9.20}$$

Then $\epsilon_0=1$, $\epsilon_{k+1}=g(\epsilon_k)$ for all k and Eq. (9.19) implies that

$$\lim_{k\to\infty}\epsilon_k=0, \tag{9.21}$$

which is equivalent to non-reconstruction.

Because

$$C_{\chi^2}(\mathrm{BEC}_{1-\epsilon}\circ\mathrm{BSC}_\delta)=(1-2\delta)^2\epsilon, \tag{9.22}$$

we have

$$\mathrm{BEC}_{1-\epsilon}\circ\mathrm{BSC}_\delta\le_{\ln}\mathrm{BEC}_{1-(1-2\delta)^2\epsilon}\,. \tag{9.23}$$

Therefore

$$C_{\chi^2}(\mathrm{BP}(\mathrm{BEC}_{1-\epsilon}))\le C_{\chi^2}(\mathrm{BEC}^{\star d}_{1-(1-2\delta)^2\epsilon})=1-(1-(1-2\delta)^2\epsilon)^d=:f(\epsilon). \tag{9.24}$$

Note that

$$f(0) = 0, \quad f'(\epsilon) = d(1-2\delta)^2(1-(1-2\delta)^2\epsilon)^{d-1}. \tag{9.25}$$

So $f'(\epsilon) \le 1$ for $\epsilon \in [0,1]$, and equality is achieved only when $\epsilon = 0$ and $d(1-2\delta)^2 = 1$. Therefore f has only one fixed point in $[0,1]$, which is 0. Because $g(\epsilon) \le f(\epsilon)$ for all $\epsilon \in [0,1]$, we have $g(\epsilon) < \epsilon$ for all $0 < \epsilon \le 1$. This finishes the proof. □

In the proof of Prop. 9.3, we showed that when the input information is close to 0, in the limit the information would contract to a non-zero value. Therefore our proof in fact shows that robust reconstruction (a stronger condition than reconstruction) on such trees is possible. By [82], for broadcasting on trees, the robust reconstruction threshold coincides with the Kesten-Stigum threshold. It is shown in [125] that when the alphabet size is at least five, the Kesten-Stigum threshold is never tight for the (non-robust) reconstruction problem. So for large alphabet size, the global comparison method does not yield the tight reconstruction threshold.

9.3 Bounds on mutual information

In this section we refine the computations in Section 9.2 and prove bounds on the limit mutual information.

Proposition 9.5. *Consider the model* $\mathrm{BOT}(2, \lambda = 1-2\delta, d)$*, where* $d \in \mathbb{Z}_{\ge 2}$ *and* $\delta = \delta_c - \tau$*, where* δ_c *is defined in Eq.* (9.3)*. Then*

$$\frac{2d\sqrt{d}}{d-1} \cdot \tau + o(\tau) \le \lim_{k\to\infty} C(M_k) \le \frac{4(d+1)\sqrt{d}\log 2}{d-1} \cdot \tau + o(\tau). \tag{9.26}$$

Proof. The proof is by analyzing the recursions in the proof of Prop. 9.3 and Prop. 9.4 more carefully.

Lower bound. In the setting of proof of Prop. 9.3, expanding everything to the order of ϵ^4 and computing a binomial sum, we get

$$\begin{aligned} C_{\chi^2}(\mathrm{BP}(\mathrm{BSC}_{1/2-\epsilon})) &= 4d(1-2\delta)^2\epsilon^2 - 16d(d-1)(1-2\delta)^4\epsilon^4 + O(\epsilon^6) \\ &= 4\left(1 + 4\sqrt{d}\tau + o_\tau(\tau)\right)\epsilon^2 - 16\left(\frac{d-1}{d} + o_\tau(1)\right)\epsilon^4 + O(\epsilon^6). \end{aligned} \tag{9.27}$$

The input information is $4\epsilon^2$. Solving the dynamics, we see that the largest fixed point is at

$$\underline{\epsilon}^* = \left(\sqrt{\frac{d\sqrt{d}}{d-1}} + o(1)\right)\sqrt{\tau}. \tag{9.28}$$

This gives

$$\begin{aligned}\lim_{k\to\infty} C(M_k) &\geq \lim_{k\to\infty} C\left(\underline{\mathrm{BP}}^k_{\ln}(\mathrm{Id})\right)\\ &\geq \log 2 - h\left(\frac{1}{2} - \left(\sqrt{\frac{d\sqrt{d}}{d-1}} + o(1)\right)\sqrt{\tau}\right)\\ &= \frac{2d\sqrt{d}}{d-1}\cdot\tau + o(\tau),\end{aligned} \tag{9.29}$$

where $h(x) = -x\log x - (1-x)\log(1-x)$ is the binary entropy function.

Upper bound. Following the proof of Prop. 9.4, let us consider the function $f(\epsilon) = 1-(1-(1-2\delta)^2\epsilon)^d$. Note that function $f(\epsilon)$ is concave on $[0,1]$, and there is a unique fixed point in $(0,1)$. By expanding in terms of ϵ, we have

$$\begin{aligned}f(\epsilon) &= d(1-2\delta)^2\epsilon - \binom{d}{2}(1-2\delta)^4\epsilon^2 + O(\epsilon^3)\\ &= \left(1+4\sqrt{d}\tau + o_\tau(\tau)\right)\epsilon - \left(\frac{d-1}{2d} + o_\tau(1)\right)\epsilon^2 + O(\epsilon^3).\end{aligned} \tag{9.30}$$

So the unique non-trivial fixed point of f is at

$$\overline{\epsilon}^* = \frac{8d\sqrt{d}}{d-1}\cdot\tau + o(\tau). \tag{9.31}$$

This gives

$$\lim_{k\to\infty} C(M_k) \leq \lim_{k\to\infty} C\left(\overline{\mathrm{BP}}^k_{\ln}(\mathrm{Id})\right) \leq \frac{8d\sqrt{d}\log 2}{d-1}\cdot\tau + o(\tau). \tag{9.32}$$

In fact, knowing that the limit is linear in τ, we can improve this upper bound. Instead of bounding the function f, let us bounding the function g (Eq. (9.20)) directly. We have

$$\begin{aligned}g(\epsilon) &:= C_{\chi^2}(\mathrm{BP}(\mathrm{BEC}_{1-\epsilon}))\\ &= 2\sum_{0\leq j\leq i\leq d}\binom{d}{i}\epsilon^i(1-\epsilon)^{d-i}\binom{i}{j}\cdot\frac{(1-\delta)^{2j}\delta^{2(i-j)}}{(1-\delta)^j\delta^{i-j}+(1-\delta)^{i-j}\delta^j} - 1.\end{aligned} \tag{9.33}$$

Because $g(\epsilon)\leq f(\epsilon)$ on $[0,1]$, the largest fixed point of g is upper bounded by the unique non-trivial fixed point of f, which is of order $\Theta(\tau)$. This justifies performing

series expansion in ϵ.

$$\begin{aligned} g(\epsilon) &= (1-\epsilon)^d + 2d\left((1-\delta)^2+\delta^2\right)\epsilon(1-\epsilon)^{d-1} \\ &\quad + d(d-1)\left(\frac{(1-\delta)^4+\delta^4}{(1-\delta)^2+\delta^2} + (1-\delta)\delta\right)\epsilon^2(1-\epsilon)^{d-2} + O(\epsilon^3) - 1 \\ &= d(1-2\delta)^2\epsilon - d(d-1)\cdot\frac{(1-2\delta)^4}{(1-2\delta)^2+1}\cdot\epsilon^2 + O(\epsilon^3) \\ &= \left(1+4\sqrt{d}\tau+o_\tau(\tau)\right)\epsilon - \left(\frac{d-1}{d+1}+o_\tau(1)\right)\epsilon^2 + O(\epsilon^3). \end{aligned} \tag{9.34}$$

We see that the largest fixed point of g satisfies

$$\epsilon^* = \frac{4(d+1)\sqrt{d}}{d-1}\cdot\tau + o(\tau). \tag{9.35}$$

In this way we get

$$\lim_{k\to\infty} C(M_k) \le \lim_{k\to\infty} C\left(\overline{\mathrm{BP}}_{\ln}^k(\mathrm{Id})\right) \le \frac{4(d+1)\sqrt{d}\log 2}{d-1}\cdot\tau + o(\tau). \tag{9.36}$$

□

We compare the above lower bound with (7) in [63].[1] We note that the lower bound of [63] can in the limit be simplified into

$$\lim_{k\to\infty} C_{\chi^2}(M_k) \ge \frac{1}{1+\frac{1-(1-2\delta)^2}{d(1-2\delta)^2-1}}. \tag{9.37}$$

Near the critical threshold, RHS behaves as $\frac{4d\sqrt{d}}{d-1}\cdot\tau$. So they obtained the the same χ^2-capacity capacity lower bound, thus the same mutual information lower bound, as in Prop. 9.5.

[63] did not state explicitly an upper bound on mutual information. Nonetheless, their upper bound is by comparison with percolation, and that leads to an upper bound of

$$\lim_{k\to\infty} C(M_k) \le \frac{8d\sqrt{d}\log 2}{d-1}\cdot\tau + o(\tau). \tag{9.38}$$

In this case we see that channel comparison leads to a better upper bound.

In the case of binary trees, we perform a more refined analysis to improve the upper bound.

Proposition 9.6. *Consider the model* $\mathrm{BOT}(2, \lambda = 1-2\delta, d=2)$ *where* $\delta = \delta_c - \tau$

[1] [63] contains an error stating that $C \ge C_{\chi^2}$, which should be $C \ge \frac{1}{2}C_{\chi^2}$. This leads to lower bounds on I (e.g., (4)(28) in [63]) to be off by a factor of 2. (7) in [63] is correct as stated.

and δ_c is defined in Eq. (9.3)*. Then*

$$\lim_{k\to\infty} C(M_k) \le 8(\sqrt{2}+1)\left(\log 2 - h\left(\frac{1}{2} - \sqrt{\frac{1}{\sqrt{2}} - \frac{1}{2}}\right)\right)\tau + o(\tau). \tag{9.39}$$

Proof. Suppose the input distribution is a mixture of BSC_Δ for Δ supported at $\{1/2-\alpha_t, 1/2\}$. We iterate the quantized belief propagation while finding the best (w.r.t the less-noisy order) channel within this family. This family contains all BECs (corresponding to $\alpha = 1/2$), so this approach may lead to a better bound. Define

$$\bar{\delta} := 1/2 - \alpha(1-2\delta). \tag{9.40}$$

The output distribution has support $\left\{\frac{\bar{\delta}^2}{\bar{\delta}^2+(1-\bar{\delta})^2}, \bar{\delta}, 1/2\right\}$. Using Lemma 2.10, we replace $\mathrm{BSC}_{\bar{\delta}}$ with a mixture of $\mathrm{BSC}_{1/2}$ and $\mathrm{BSC}_{\frac{\bar{\delta}^2}{\bar{\delta}^2+(1-\bar{\delta})^2}}$, while preserving χ^2-capacity. Therefore

$$1/2 - \alpha_{t+1} = \frac{\bar{\delta}^2}{\bar{\delta}^2 + (1-\bar{\delta})^2}. \tag{9.41}$$

Solving this, we get that in the $k\to\infty$ limit

$$\alpha^* = \frac{\sqrt{1-4\delta}}{2(1-2\delta)}. \tag{9.42}$$

For $\alpha = \alpha^*$, we have

$$C_{\chi^2}(\mathrm{BSC}_{\bar{\delta}}) = (1-2\delta)^2 C_{\chi^2}(\mathrm{BSC}_{\frac{\bar{\delta}^2}{\bar{\delta}^2+(1-\bar{\delta})^2}}). \tag{9.43}$$

So when applying Lemma 2.10, every unit weight for the former becomes $(1-2\delta)^2$ weight for the latter.

Let ϵ_t be the weight of $\mathrm{BSC}_{1/2}$ in iteration t. Then in the $k\to\infty$ limit ϵ should satisfy

$$1-\epsilon = (1-\epsilon)^2(\bar{\delta}^2 + (1-\bar{\delta})^2) + 2\epsilon(1-\epsilon)(1-2\delta)^2. \tag{9.44}$$

Solving this we get $\epsilon^* = 1 - 8(\sqrt{2}+1)\tau + o(\tau)$.

So an upper bound for mutual information is

$$\begin{aligned}&(1-\epsilon^*)(\log 2 - h(1/2-\alpha^*))\\ &= 8(\sqrt{2}+1)\left(\log 2 - h\left(\frac{1}{2} - \sqrt{\frac{1}{\sqrt{2}} - \frac{1}{2}}\right)\right)\tau + o(\tau).\end{aligned} \tag{9.45}$$

□

The same method can be applied to the lower bound, leading to $\alpha_* = (\sqrt{3\sqrt{2}} +$

$o(1))\sqrt{\tau}$ and $\epsilon_* = \frac{1}{3} + o(1)$, giving

$$\lim_{k\to\infty} C(M_k) \geq 4\sqrt{2}\tau + o(\tau). \tag{9.46}$$

Surprisingly, although we lower bound using a larger family, and the limiting distribution is different, we get the same lower bound as Prop. 9.5.

We have shown that for the Ising model on a binary tree, $I(\delta_c - \tau) = c\tau + o(\tau)$ for some $c \in [5.65, 9.85]$. The improvement over Prop. 9.5 can be attributed to a finer "quantization" since we try to work with less-noisy channels while staying closer to the true output of BP. We shall explore this idea further in Section 9.4 and show (numerically) that the correct slope is $c \approx 5.65$.

9.4 Improved bounds via local comparisons

One advantage of the comparison method is that it allows us to analyze BP, rather than some suboptimal algorithm. On the other hand, we incur some loss in each step of the analysis due to the crude approximations that are made to the input distribution in order to simplify the analysis. In some cases these losses can be significant. For instance, a naive application of the comparison method while matching probabilities of error (i.e., using Prop. 9.1) does not even recover the right threshold. One way to avoid this issue is to do local comparisons. We first define the local quantization operators.

Definition 9.7 (Local quantization operators)**.** Let P be a BMS channel and P_Δ be its Δ-distrbution. Let $[0, 1/2] = \bigcup_{i\in\mathcal{I}} I_i$ be a partition of $[0, 1/2]$ into a finite disjoint union of subintervals. Define the local (lower) quantization operator $\underline{Q}_{\text{deg,loc}}$ by replacing the support of P_Δ along each I_i with a single point at $\Delta_i := \frac{\int_{I_i} \Delta dP_\Delta}{\int_{I_i} dP_\Delta}$ with probability mass $\int_{I_i} dP_\Delta$ (i.e., mapping P to the BMS channel whose Δ-distrbution is the modified distribution). Likewise, define the local (upper) quantization operator $\overline{Q}_{\text{deg,loc}}$ by replacing the support of P_Δ along each I_i with two quantization points $a_i := \inf I_i$, $b_i := \sup I_i$ with probabilities $p_{a_i} = \alpha_i p_i$, $p_{b_i} = (1 - \alpha_i)p_i$, where $p_i = \int_{I_i} dP_\Delta$ and $\alpha_i = \frac{b_i - \int_{I_i} \Delta P_\Delta / \int_{I_i} dP_\Delta}{b_i - a_i}$. Furthermore, define $\underline{Q}_{\text{ln,loc}}$ (resp. $\overline{Q}_{\text{ln,loc}}$) similarly by matching the χ^2-capacity along each interval while contracting (resp. spreading) probability masses.

Using the local quantization operators we can improve Prop. 9.1 and Prop. 9.2 as follows.

Proposition 9.8. *Consider the model* $\text{BOT}(2, \lambda = 1 - 2\delta, d)$ *with* $d\lambda^2 > 1$*. Choose* $\delta_0 < 1/2$ *such that*

$$\delta_0 > \lim_{k\to\infty} P_e(M_k). \tag{9.47}$$

Define

$$\underline{\mathrm{BP}}_{\mathrm{deg,loc}} := \underline{Q}_{\mathrm{deg,loc}} \circ \mathrm{BP}, \quad \overline{\mathrm{BP}}_{\mathrm{deg,loc}} := \overline{Q}_{\mathrm{deg,loc}} \circ \mathrm{BP}, \tag{9.48}$$

$$\underline{\mathrm{BP}}_{\mathrm{ln,loc}} := \underline{Q}_{\mathrm{ln,loc}} \circ \mathrm{BP}, \quad \overline{\mathrm{BP}}_{\mathrm{ln,loc}} := \overline{Q}_{\mathrm{ln,loc}} \circ \mathrm{BP}\,. \tag{9.49}$$

Then for any $k \geq 0$ we have

$$P_e\left(\underline{\mathrm{BP}}^k_{\mathrm{deg,loc}}(\mathrm{BSC}_{\delta_0})\right) \geq P_e(\delta) \geq P_e\left(\overline{\mathrm{BP}}^k_{\mathrm{deg,loc}}(\mathrm{Id})\right), \tag{9.50}$$

$$C\left(\underline{\mathrm{BP}}^k_{\mathrm{ln,loc}}(\mathrm{BSC}_{\delta_0})\right) \leq I(\delta) \leq C\left(\overline{\mathrm{BP}}^k_{\mathrm{ln,loc}}(\mathrm{Id})\right). \tag{9.51}$$

To choose the initial δ_0 we may, for example, use a Kesten-Stigum upper bound on P_e (see e.g., [84] or Section 10.2.4).

Proof. From the construction we see that for every BMS channel P,

$$\underline{Q}_{\mathrm{deg,loc}}(P) \leq_{\mathrm{deg}} P \leq_{\mathrm{deg}} \overline{Q}_{\mathrm{deg,loc}}(P), \tag{9.52}$$

$$\underline{Q}_{\mathrm{ln,loc}}(P) \leq_{\mathrm{ln}} P \leq_{\mathrm{ln}} \overline{Q}_{\mathrm{ln,loc}}(P). \tag{9.53}$$

Therefore for all $k \geq 0$,

$$\underline{\mathrm{BP}}^k_{\mathrm{deg,loc}}(P) \leq_{\mathrm{deg}} \mathrm{BP}^k(P) \leq_{\mathrm{deg}} \overline{\mathrm{BP}}^k_{\mathrm{deg,loc}}(P), \tag{9.54}$$

$$\underline{\mathrm{BP}}^k_{\mathrm{ln,loc}}(P) \leq_{\mathrm{ln}} \mathrm{BP}^k(P) \leq_{\mathrm{ln}} \overline{\mathrm{BP}}^k_{\mathrm{ln,loc}}(P). \tag{9.55}$$

For the upper bounds, we have

$$\mathrm{BP}^{k+l}(\mathrm{Id}) \leq_{\mathrm{deg}} \mathrm{BP}^k(\mathrm{Id}) \leq_{\mathrm{deg}} \overline{\mathrm{BP}}^k_{\mathrm{deg,loc}}(\mathrm{Id}), \tag{9.56}$$

$$\mathrm{BP}^{k+l}(\mathrm{Id}) \leq_{\mathrm{ln}} \mathrm{BP}^k(\mathrm{Id}) \leq_{\mathrm{ln}} \overline{\mathrm{BP}}^k_{\mathrm{ln,loc}}(\mathrm{Id}) \tag{9.57}$$

for all $k, l \geq 0$. This proves the upper bounds in Eq. (9.50) and (9.51).

For the lower bounds, note that by Eq. (9.47), there exists l_0 such that for all $l \geq l_0$ we have

$$\mathrm{BSC}_{\delta_0} \leq_{\mathrm{deg}} M_k \tag{9.58}$$

and therefore

$$\mathrm{BSC}_{\delta_0} \leq_{\mathrm{ln}} M_k. \tag{9.59}$$

So

$$\mathrm{BP}^{k+l}(\mathrm{Id}) \geq_{\mathrm{deg}} \mathrm{BP}^k(\mathrm{BSC}_{\delta_0}) \geq_{\mathrm{deg}} \underline{\mathrm{BP}}^k_{\mathrm{deg,loc}}(\mathrm{BSC}_{\delta_0}), \tag{9.60}$$

$$\mathrm{BP}^{k+l}(\mathrm{Id}) \geq_{\mathrm{ln}} \mathrm{BP}^k(\mathrm{BSC}_{\delta_0}) \geq_{\mathrm{ln}} \underline{\mathrm{BP}}^k_{\mathrm{ln,loc}}(\mathrm{BSC}_{\delta_0}) \tag{9.61}$$

for all $k \geq 0$ and $l \geq l_0$. This proves the lower bounds in Eq. (9.50) and (9.51). □

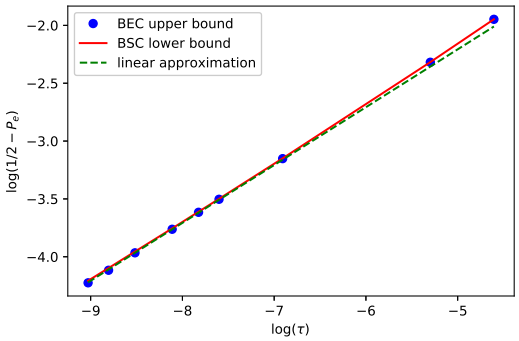

Figure 9-1: Bounds on probability of error using local comparisons for $\delta = \delta_c - \tau$. The linear approximation has a slope of $1/2$.

Using uniform quantization in the $[0, 1/2]$ interval with 1024 points, we were able to show that

$$I(\delta_c - \tau) = c\tau + o(\tau) \tag{9.62}$$

with $c \approx 5.65$. We thus conjectured that

$$I(\delta_c - \tau) = (4\sqrt{2} + o(1))\tau. \tag{9.63}$$

Using a degradation argument (or Fano's inequality), one can also show

$$1/2 - c'\sqrt{\tau} + o(\sqrt{\tau}) \le P_e(\delta_c - \tau) \le 1/2 - c\tau + o(\tau). \tag{9.64}$$

It is natural to ask what is the correct exponent for P_e. Using the same approach we were able to show (see Fig. 9-1)

$$\log(1 - 2P_e) \ge 0.504 \log \tau + c. \tag{9.65}$$

We thus conjectured that $1/2$ is the correct exponent.

We remark that both our conjectures were later proved (and strengthened) in [136].

Chapter 10

Uniqueness of BP fixed point: Ising model

We prove stability of belief propagation fixed points (Definition 5.21) and boundary irrelevance (Definition 5.17) for the Ising model $\mathrm{BOT}(2, \theta, d)$ and $\mathrm{BOT}(2, \theta, \mathrm{Pois}(d))$, when signal-to-noise ratio (SNR, Definition 5.8) is outside a finite interval $[1, 3.513]$. Via reductions established in Chapter 5, we achieve a mutual information formula and an optimal recovery algorithm for $\mathrm{SBM}(n, 2, a, b)$. Before our work, a mutual information formula was known for disassortative $\mathrm{SBM}(n, q, a, b)$ [43] and for dense binary symmetric stochastic block model [50], but open for the sparse and assortative regime. For the proof, we introduce the degradation method, which reduces the problem of boundary irrelevance to contraction of certain potential functions under BP recursion. We choose the potential function to be the Bhattacharyya coefficient (Hellinger distance) of a BMS channel, and use its contraction properties to finish the proof. This chapter is based on [4].

We remark that subsequent work [137] established stability of BP fixed points and boundary irrelevance for any $\mathrm{BOT}(T, 2, \theta)$ and $\mathrm{BOT}(2, \theta, D)$ model, thus giving a mutual information formula and an optimal recovery algorithm for any $\mathrm{SBM}(n, 2, a, b)$. Nevertheless, we believe our method is still interesting, as it can be generalized to the Potts model and $\mathrm{SBM}(n, q, a, b)$ (Chapter 11). This chapter can be seen as a warmup for Chapter 11.

Chapter outline In Section 10.1 we introduce our setting, and state our main results. In Section 10.2, we prove our main results, boundary irrelevance and uniqueness of BP fixed points for SNR outside $[1, 3.513]$. In Section 10.3, we discuss applications of uniqueness of BP fixed points and boundary irrelevance.

10.1 Introduction

Stochastic block model We consider the simplest stochastic block model, the two-community symmetric SBM, denoted $\mathrm{SBM}(n, 2, a, b)$ (Definition 5.4), where $n \in \mathbb{Z}_{\geq 1}$, $a, b \in \mathbb{R}_{\geq 0}$. The model is defined as follows. First we assign a random label $X_u \sim$

Unif($\{\pm\}$) i.i.d for $u \in V = [n]$. Then a random graph $G = (V, E)$ is generated, where $(u, v) \in E$ with probability $\frac{a}{n}$ if $X_u = X_v$, and with probability $\frac{b}{n}$ if $X_u \neq X_v$, independently for all $(u, v) \in \binom{V}{2}$.

The weak recovery problem (Definition 5.5) for SBM$(n, 2, a, b)$ was settled by [96, 103, 105], showing that weak recovery is possible above the Kesten-Stigum threshold (SNR > 1), and impossible below the KS threshold. When weak recovery is possible, the natural follow-up question is to determine the optimal recovery accuracy

$$\sup_{\hat{X}=\hat{X}(G)} \lim_{n\to\infty} \mathbb{E}\left[1 - \frac{1}{n} d_H(X, \hat{X}(G))\right], \tag{10.1}$$

$$\text{where} \quad d_H(X, Y) := \sum_{s\in\{\pm\}} \sum_{i\in[n]} \mathbb{1}\{X_i \neq sY_i\}. \tag{10.2}$$

[104] studied the problem of optimal recovery and proposed an algorithm, which they proved to be optimal when SNR is larger than a constant. They did not compute the constant, but a crude estimation shows it is at least 75 [4]. Their algorithm is conjectured to hold all the way down to the KS threshold. [110] generalized the analysis and proved the optimality of a local belief propagation algorithm for SBM$(n, 2, a, b)$ with survey, in the same parameter regime as [104].

A fundamental quantity for the SBM is the limit mutual information

$$\lim_{n\to\infty} \frac{1}{n} I(X; G). \tag{10.3}$$

Note that even the existence of the limit is non-trivial, and was proved in [5] for the disassortative regime ($a < b$). Later, an expression for the disassortative case was given by [43]. The problem of SBM mutual information for SBM$(n, 2, a, b)$ in the assortative case remained open until our work [4] presented here.

We refer the reader to Chapter 5 for a review of previous results on weak recovery, optimal recovery, and mutual information.

Broadcasting on trees In the SBM, the local neighborhood of a random vertex converges to a Galton-Watson tree with Poisson offspring distribution. Furthermore, the labels of the local neighborhood can be coupled with that of the tree model. Therefore the SBM is closely related to the BOT model, a phenomenon initially proved in [103] for SBM$(n, 2, a, b)$. In this case, the relevant BOT model is BOT$(2, \theta, \text{Pois}(d))$, defined as follows. Let T be a Galton-Watson tree with Pois(d) offspring distribution. We assign to every vertex v a label σ_v according to the following process.

1. Generate $\sigma_\rho \sim \text{Unif}(\{\pm 1\})$.
2. Suppose we have generated a label for a vertex u. For every child v of u, we genreate σ_v according to $\text{BSC}_\delta(\cdot|\sigma_u)$, where $\theta = 1 - 2\delta$.

We also consider the case where T is a regular tree (every vertex has $d \in \mathbb{Z}_{\geq 0}$ children), denoted BOT$(2, \theta, d)$.

We refer the reader to Chapter 5 for a review of previous results on the BOT model.

Boundary irrelevance Let us add side information (survey) to the BOT model. Let W be a fixed BMS channel, and for each node u we observe $\omega_u \sim W(\cdot|\sigma_u)$. We call this model broadcasting on trees with survey (BOTS). We will also denote by Δ_W the Δ-component of the BMS W (see Chapter 2 for background on BMS channels). This setting includes the one in [110], where $W = \mathrm{BSC}_\alpha$, i.e., for each node u, $\mathbb{P}[\omega_u = \sigma_u] = 1 - \mathbb{P}[\omega_u = -\sigma_u] = 1 - \alpha$; and the one in [83], where $W = \mathrm{BEC}_\epsilon$, i.e., for each node the survey reveals the correct label with probability $1 - \epsilon$ and an erasure symbol otherwise. The latter is of particular interest to us because of its application to the computation of the SBM mutual information.

We say the model admits boundary irrelevance with repsect to W if

$$\lim_{k\to\infty} I(\sigma_\rho; \sigma_{L_k}|T_k, \omega_{T_k}) = 0, \tag{10.4}$$

where L_k denotes the set of vertices at distance k to ρ, T_k denotes the set of vertices at distance $\le k$ to ρ. We say the model admits boundary irrelevance if it admits boundary irrelevance with repsect to all erasure channels BEC_ϵ with $0 \le \epsilon < 1$.

Our work [4] proved that boundary irrelevance implies a formula for SBM mutual information. This is further generalized in our work [73] and in Theorem 5.18. See Chapter 5 for more discussions.

If the model admits boundary irrelevance, then we have

$$\lim_{\epsilon\to 1^-}\lim_{k\to\infty} I(\sigma_\rho; T_k, \omega^\epsilon_{T_k}) = \lim_{k\to\infty} I(\sigma_\rho; T_k, \sigma_{L_k}), \tag{10.5}$$

where ω^ϵ denotes survey observation with $W = \mathrm{BEC}_\epsilon$. Indeed, we have

$$\begin{aligned}\lim_{\epsilon\to 1^-}\lim_{k\to\infty} I(\sigma_\rho; T_k, \omega^\epsilon_{T_k}) &= \lim_{\epsilon\to 1^-}\lim_{k\to\infty} I(\sigma_\rho; T_k, \omega^\epsilon_{T_k}, \sigma_{L_k}) \\ &= \inf_{\substack{\epsilon\in[0,1)\\ k\in\mathbb{Z}_{\ge 0}}} I(\sigma_\rho; T_k, \omega^\epsilon_{T_k}, \sigma_{L_k}) \\ &= \lim_{k\to\infty} I(\sigma_\rho; T_k, \sigma_{L_k}),\end{aligned} \tag{10.6}$$

where the first step is by boundary irrelevance, the second step is by data processing inequality, and the third step is by because for every $k \in \mathbb{Z}_{\ge 0}$, the value $I(\sigma_\rho; \omega^\epsilon_{T_k}, \sigma_{L_k})$ is continuous in $\epsilon \in [0, 1]$ including at the boundary. Property (10.5) is known as the condition for optimality of local algorithms [83, 110].

Uniqueness of BP fixed point Boundary irrelevance with respect to a channel W can be interpreted using the belief propagation operator (Definition 5.13). The BP operator for our case is an operator from the space of BMS channels to itself,

defined as

$$\mathrm{BP}(M) := \mathbb{E}_b(M \circ \mathrm{BSC}_\delta)^{\star b}, \tag{10.7}$$

where $b = d$ for the regular tree case, and $b \sim \mathrm{Pois}(d)$ for the Poisson tree case. The BP operator is relevant because if we let M_k to denote the channel $\sigma_\rho \mapsto \sigma_{L_k}$, then

$$M_{k+1} = \mathrm{BP}(M_k). \tag{10.8}$$

Given a survey BMS channel W, we consider the operator BP_W, defined as

$$\mathrm{BP}_W(M) := \left(\mathbb{E}_b(M \circ \mathrm{BSC}_\delta)^{\star b}\right) \star W. \tag{10.9}$$

Then boundary irrelevance with repsect to W is equivalent to that for any initial channel M (possibly trivial), $\mathrm{BP}_W^k(M)$ and $\mathrm{BP}_W^k(\mathrm{Id})$ goes to the same limit as $k \to \infty$. This can be understood as uniqueness of fixed point for the operator BP_W.

We can also consider uniqueness of fixed point for operator BP. We say the BOT model admits uniqueness of BP fixed point if the operator BP has only one non-trivial BMS fixed point. We sometimes need to use the stronger notion of stability of BP fixed point, which says that for any non-trivial initial BMS channel M, $\mathrm{BP}_W^k(M)$ and $\mathrm{BP}_W^k(\mathrm{Id})$ goes to the same limit as $k \to \infty$. As shown by [104], stability of BP fixed point implies an optimal recovery algorithm for SBM. See Chapter 5 for more discussions.

Our results Our main result for this chapter is as follows.

Theorem 10.1. *Consider the model* $\mathrm{BOT}(2, \theta, d)$ *or* $\mathrm{BOT}(2, \theta, \mathrm{Pois}(d))$. *Let* W *be a non-trivial BMS channel. If*

$$d\theta^2 \exp\left(-\frac{(d\theta^2 - 1)_+}{2}\right) Z(W) < 1, \tag{10.10}$$

where $Z(W)$ *is the Bhattacharyya coefficient (Definition 2.4), then the model admits BI with respect to* W.

In particular, BI holds whenever $d\theta^2 < 1$ *or* $d\theta^2 > \alpha^*$, *where* $\alpha^* \approx 3.513$ *is the unique solution in* $\mathbb{R}_{>1}$ *to the equation*

$$\exp\left(-\frac{\alpha - 1}{2}\right) \alpha = 1. \tag{10.11}$$

We remark that (10.10) is a relaxation of a sharper bound in Prop. 10.6 (e.g., for $\mathrm{BOT}(2, \theta, 2)$, BI is proven for all cases except $d\theta^2 \in (1, 1.62)$). The following corollary lists a few direct consequences of Theorem 10.1.

Corollary 10.2. *In the setting of Theorem 10.1, if any of the following is true, then the model admits BI with repsect to* W.

(i) $Z(W) < \frac{\sqrt{e}}{2} \approx 0.824$;

(ii) $P_e(W) < \frac{1}{2} - \frac{1}{4}\sqrt{4-e} \approx 0.217$*;*

(iii) $W = \mathrm{BEC}_\epsilon$ *and with* $\epsilon < \frac{\sqrt{e}}{2} \approx 0.824$*.*

Proof. For (i) we use

$$\sup_{\alpha \geq 0} \left(\alpha \exp\left(-\frac{\alpha - 1}{2} \right) \right) = \frac{2}{\sqrt{e}}. \tag{10.12}$$

For (ii) we define $p(\Delta) = 2\sqrt{\Delta(1-\Delta)}$ and notice that

$$Z(W) = \mathbb{E}[p(\Delta_W)] \leq p(\mathbb{E}\Delta_W) = p(P_e(W)) \tag{10.13}$$

because the function p is concave. So when $P_e(W) < \frac{1}{2} - \frac{1}{4}\sqrt{4-e}$, we have $Z(W) < \frac{\sqrt{e}}{2}$.

(iii) follows from (i). □

We demonstrate the region of uniqueness of BP fixed point from Corollary 10.2 on Figure 10-1. We note that taking the limit $\epsilon \to 1^-$, Theorem 10.1 implies that

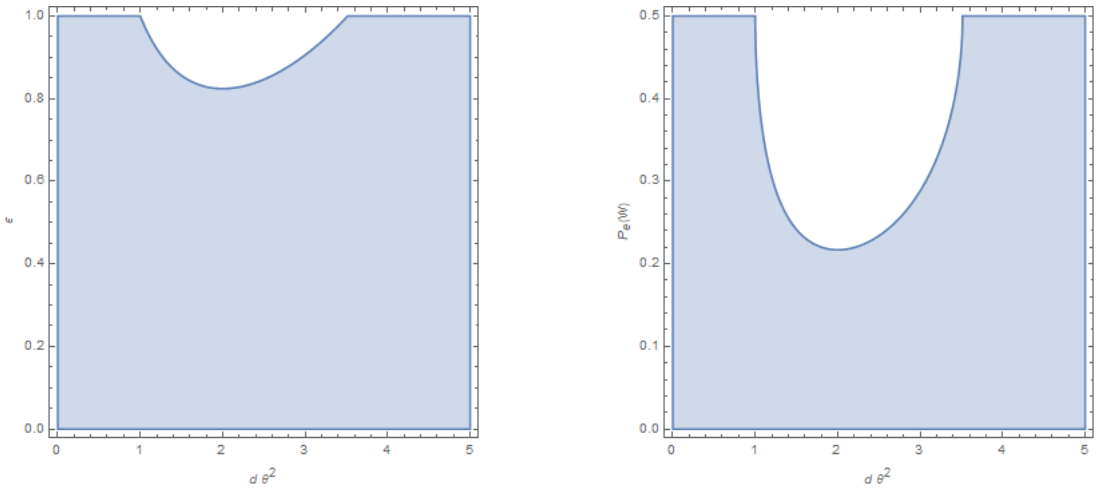

Figure 10-1: Left: Region of BP uniqueness for BEC survey from Corollary 10.2(iii). Right: Region of BP uniqueness for BMS survey from Corollary 10.2(ii).

revealing an (arbitrarily) small fraction of vertex labels gives the same information about the root bit, as revealing the whole boundary labels at large distance, even in the reconstruction regime, cf. (10.5).

Our method for proving boundary irrelevance also works for uniqueness of BP fixed point.

Theorem 10.3. *Consider the model* $\mathrm{BOT}(2, \theta, d)$ *or* $\mathrm{BOT}(2, \theta, \mathrm{Pois}(d))$*. If*

$$d\theta^2 \exp\left(-\frac{(d\theta^2 - 1)_+}{2} \right) < 1, , \tag{10.14}$$

then the model admits stability of BP fixed point.

Applications of Theorem 10.1 and Theorem 10.3 include a mutual information formula and an optimal recovery algorithm. We discuss these applications in Section 10.3.

Our method We believe that our proof technique offers the following improvements compared to [104, 110]: (a) it is much shorter; (b) we do not need to consider large θ, small d and small d large θ cases separately; (c) it works simultaneously for $d\theta^2 < 1$ and $d\theta^2 > 3.513$; (d) it works simultaneously with and without side information, and the side information can be any BMS, rather than specifically the BEC or BSC; (e) it closes the entire low-SNR case $d\theta^2 < 1$[1].

Our main innovation is the information-theoretic point of view: we consider BOTS with or without leaf observations as two binary input symmetric channels (BMSs) which are related to each other by a property known as degradation. This implies a certain inequality between the log-likelihood ratios (LLRs), cf. (10.24), which we exploit in the application of the potential method. These key ideas are the content of the Prop. 10.5. On the more technical side, another innovation is the choice of the potential function as $\phi(r) = e^{-\frac{1}{2}r}$.

10.2 Uniqueness of BP fixed point

In this section we prove our main results, Theorem 10.1 and Theorem 10.3.

Recall the BOTS model defined in Section 10.1. Let M_k denote the BMS channel $\sigma_\rho \to (\omega_{T_k}, \sigma_{L_k})$ and $\widetilde{M}_k$ denote the BMS channel $\sigma_\rho \to \omega_{T_k}$. Let P_{Δ_k} (resp. $P_{\widetilde{\Delta}_k}$) be distribution of Δ-component of BMS M_k (resp. $\widetilde{M}_k$). We prove the following strengthening of Theorem 10.1.

Theorem 10.4. *In the setting of Theorem 10.1, P_{Δ_k} and $P_{\widetilde{\Delta}_k}$ converge weakly to the same distribution as $k \to \infty$. In particular,*

$$\lim_{k\to\infty} P_e(M_k) = \lim_{k\to\infty} P_e(\widetilde{M}_k), \tag{10.15}$$

$$\lim_{k\to\infty} C(M_k) = \lim_{k\to\infty} C(\widetilde{M}_k). \tag{10.16}$$

Proof of Theorem 10.3 is deferred to Section 10.2.5.

10.2.1 Belief propagation recursion

The maximum a posteriori probability (MAP) decoder is the optimal decoder for this reconstruction problem. It can be implemented using belief propagation (BP) as follows.

[1]There are, however, two related low-SNR results. [110, Theorem 4.2] shows uniqueness of fixed point for $d\theta < 1$ via a simple contractivity of F_θ function in the BP recursion (10.20).[83, Theorem 3] shows (10.5) for $d\theta^2 < 1$ as an application of information contraction from [63].

For each node u, let $L_k(u)$ denote the set of nodes in subtree rooted at u that are at distance k to u. Let $T_k(u)$ denote the set of nodes in subtree rooted at u that are at distance $\leq k$ to u. Let $R_{u,k} \in \mathbb{R} \cup \{\pm\infty\}$ denote the posterior log likelihood ratio given $\omega_{T_k(u)} \cup \sigma_{L_k(u)}$:

$$R_{u,k} = \log \frac{\mathbb{P}[\sigma_u = +|\omega_{T_k(u)} \cup \sigma_{L_k(u)}]}{\mathbb{P}[\sigma_u = -|\omega_{T_k(u)} \cup \sigma_{L_k(u)}]}. \tag{10.17}$$

The initial value is

$$R_{u,0} = \sigma_u \cdot \infty. \tag{10.18}$$

Define a function $F_\theta : \mathbb{R} \cup \{\pm\infty\} \to \mathbb{R}$ as

$$F_\theta(r) = 2 \operatorname{arctanh}\left(\theta \tanh\left(\frac{1}{2}r\right)\right). \tag{10.19}$$

By definition of $R_{u,k}$ and Bayes rule (see e.g. [110]), we have

$$R_{u,k+1} = \sum_{v \in L_1(u)} F_\theta(R_{v,k}) + W_u \tag{10.20}$$

where W_u is the log likelihood ratio induced by observation, i.e.,

$$W_u = \log \frac{\mathbb{P}[\sigma_u = +|\omega_u]}{\mathbb{P}[\sigma_u = -|\omega_u]}. \tag{10.21}$$

Using (10.18)(10.20) we are able to compute $R_{\rho,k}$ recursively.

For observation without leaves, let $\widetilde{R}_{u,k}$ denote the posterior log likelihood ratio given $\omega_{T_k(u)}$. Then $\widetilde{R}_{u,k}$ satisfies the same recursion (10.20), but with a different initial value

$$\widetilde{R}_{u,0} = 0. \tag{10.22}$$

Let $M_k(u)$ denote the BMS channel $\sigma_u \to (\omega_{T_k(u)}, \sigma_{L_k(u)})$. Let $\widetilde{M}_k(u)$ denote the BMS channel $\sigma_u \to \omega_{T_k(u)}$. Let $\Delta_{u,k}$ and $\widetilde{\Delta}_{u,k}$ denote the corresponding Δ-components (both are random variables supported on $[0, \frac{1}{2}]$). They relate to log likelihood ratio via the following expression:

$$|R_{u,k}| = \log \frac{1 - \Delta_{u,k}}{\Delta_{u,k}}, \quad |\widetilde{R}_{u,k}| = \log \frac{1 - \widetilde{\Delta}_{u,k}}{\widetilde{\Delta}_{u,k}}. \tag{10.23}$$

There exists a canonical coupling between $M_k(u)$ and $\widetilde{M}_k(u)$ via forgetting $\sigma_{L_k(u)}$ (i.e., the channel $(\omega_{T_k(u)}, \sigma_{L_k(u)}) \mapsto \omega_{T_k(u)}$). So $M_k(u)$ is less degraded than $\widetilde{M}_k(u)$. Furthermore, by data processing inequality for total variation, under the canonical

coupling, we have

$$\mathbb{E}[\Delta_{u,k}|\widetilde{\Delta}_{u,k}] \le \widetilde{\Delta}_{u,k}. \tag{10.24}$$

As we will see, the core of our proof is the use of this degradation relationship.

Let μ_k^+ be the distribution of $R_{u,k}$ conditioned on $\sigma_u = +$ (and for Galton-Watson trees, without revealing structure of the subtree rooted at u), and $\widetilde{\mu}_k^+$ be the distribution of $\widetilde{R}_{u,k}$ conditioned on $\sigma_u = +$. These definitions do not depend on the choice of u. Then μ_0^+ is the point measure at $+\infty$, $\widetilde{\mu}_0^+$ is the point measure at 0.

Both distributions satisfy the same recursion. Consider the equation

$$R_{u,k+1}^+ = \sum_{v \in L_1(u)} Z_v F_\theta(R_{v,k}^+) + W_u \tag{10.25}$$

where $\{Z_v, R_{v,k}^+, W_u : v \in L_1(u)\}$ are independent, Z_v are i.i.d. Bernoulli with $\mathbb{P}[Z_v = +1] = 1 - \mathbb{P}[Z_v = -1] = 1 - \delta$, $R_{v,k}^+ \sim \mu_k^+$, and W_u distributes as log likelihood ratio corresponding to the survey BMS. Then $R_{u,k+1}^+ \sim \mu_{k+1}^+$. The same holds if we replace $R_{v,k}^+ \sim \mu_k^+$ with $\widetilde{R}_{v,k}^+ \sim \widetilde{\mu}_k^+$ and $R_{u,k+1}^+ \sim \mu_{k+1}^+$ with $\widetilde{R}_{u,k+1}^+ \sim \widetilde{\mu}_{k+1}^+$.

BP distributional fixed point A distribution μ on $\mathbb{R} \cup \{\pm\infty\}$ is called a BP fixed point of the BOTS model (θ, d, W) if taking R_i^+ i.i.d. $\sim \mu$, $i \in [d]$, Z_i and R_W as above results in

$$R^+ = \sum_{1 \le i \le d} Z_i F_\theta(R_i^+) + R_W \tag{10.26}$$

having the same distribution μ. In this work we restrict our attention to symmetric distributions, i.e., distributions associated with BMS channels. We talk below about the fixed point distribution P_Δ on $[0, \frac{1}{2}]$ that is related to μ via transformation (10.23). Namely, a distribution P_Δ is a fixed point iff the law μ of random variable R^+ is a fixed point, where R^+ is generated via sampling $\Delta \sim P_\Delta$ and then setting

$$R^+ = \begin{cases} \log \frac{1-\Delta}{\Delta}, & \text{w.p. } 1-\Delta, \\ -\log \frac{1-\Delta}{\Delta}, & \text{w.p. } \Delta. \end{cases} \tag{10.27}$$

Similarly, we define the BP fixed point for the BOTS model $(\theta, \mathrm{Pois}(d), W)$ where in (10.26) d is replaced with $b \sim \mathrm{Pois}(d)$.

10.2.2 Contraction of potential function

The technical part of our proof is contraction of certain potential functions. The next proposition shows the kind of contraction result we need.

Proposition 10.5. *Let $\phi : \mathbb{R} \cup \{\pm\infty\} \to \mathbb{R} \cup \{\pm\infty\}$ be a function such that the*

function $g : [0, \frac{1}{2}] \to \mathbb{R} \cup \{\pm\infty\}$ *defined as*

$$g(\Delta) = (1-\Delta)\phi\left(\log\frac{1-\Delta}{\Delta}\right) + \Delta\phi\left(-\log\frac{1-\Delta}{\Delta}\right) \tag{10.28}$$

is decreasing and α*-strongly convex for some* $\alpha > 0$*. If*

$$\lim_{k\to\infty} \mathbb{E}[\phi(R^+_{\rho,k}) - \phi(\widetilde{R}^+_{\rho,k})] = 0, \tag{10.29}$$

then under the canonical coupling,

$$\lim_{k\to\infty} \mathbb{E}(\Delta_{\rho,k} - \widetilde{\Delta}_{\rho,k})^2 = 0. \tag{10.30}$$

Proof. Because g is α-strongly convex, we have

$$g(\Delta_{\rho,k}) - g(\widetilde{\Delta}_{\rho,k}) \geq g'(\widetilde{\Delta}_{\rho,k})(\Delta_{\rho,k} - \widetilde{\Delta}_{\rho,k}) + \frac{\alpha}{2}(\Delta_{\rho,k} - \widetilde{\Delta}_{\rho,k})^2. \tag{10.31}$$

Then

$$\begin{aligned}
\mathbb{E}[\phi(R^+_{\rho,k}) - \phi(\widetilde{R}^+_{\rho,k})] &= \mathbb{E}_{\widetilde{\Delta}_{\rho,k}}\mathbb{E}[\phi(R^+_{\rho,k}) - \phi(\widetilde{R}^+_{\rho,k})|\widetilde{\Delta}_{\rho,k}] \\
&= \mathbb{E}_{\widetilde{\Delta}_{\rho,k}}\mathbb{E}[g(\Delta_{\rho,k}) - g(\widetilde{\Delta}_{\rho,k})|\widetilde{\Delta}_{\rho,k}] \\
&\geq \mathbb{E}_{\widetilde{\Delta}_{\rho,k}}\mathbb{E}[g'(\widetilde{\Delta}_{\rho,k})(\Delta_{\rho,k} - \widetilde{\Delta}_{\rho,k}) + \frac{\alpha}{2}(\Delta_{\rho,k} - \widetilde{\Delta}_{\rho,k})^2|\widetilde{\Delta}_{\rho,k}] \\
&= \mathbb{E}_{\widetilde{\Delta}_{\rho,k}}[g'(\widetilde{\Delta}_{\rho,k})(\mathbb{E}[\Delta_{\rho,k}|\widetilde{\Delta}_{\rho,k}] - \widetilde{\Delta}_{\rho,k})] + \frac{\alpha}{2}\mathbb{E}(\Delta_{\rho,k} - \widetilde{\Delta}_{\rho,k})^2 \\
&\geq \frac{\alpha}{2}\mathbb{E}(\Delta_{\rho,k} - \widetilde{\Delta}_{\rho,k})^2.
\end{aligned} \tag{10.32}$$

The second step is because $R^+_{\rho,k}$ and $\Delta_{\rho,k}$ (also $\widetilde{R}^+_{\rho,k}$ and $\widetilde{\Delta}_{\rho,k}$) relate via (10.27). By (10.32), we see that (10.29) implies (10.30). □

Note that (10.32) also shows that $\mathbb{E}[\phi(R^+_{\rho,k}) - \phi(\widetilde{R}^+_{\rho,k})]$ is non-negative as long as g is decreasing and convex.

We choose the potential function to be $\phi(r) = -\exp\left(-\frac{1}{2}r\right)$. This potential function is chosen so that the expectation of $\phi(R^+_{u,k+1})$ has a nice decomposition (10.41). In fact $\mathbb{E}\left[\exp\left(-\frac{1}{2}R^+\right)\right]$ is equal to the Bhattacharyya coefficient of the BMS channel, and (10.41) can be interpreted as multiplicativity of Bhattacharyya coefficients under $\star$-convolution.

The function g is given by $g(\Delta) = -2\sqrt{\Delta(1-\Delta)}$. One can check that g is decreasing and 4-strongly convex on $[0, \frac{1}{2}]$.

Proposition 10.6. *Assume that we have a non-trivial survey channel. Let*

$$C_1 = C_1(\theta, d, W) = d\theta^2\left(1 - \frac{(d\theta^2-1)_+}{d-1}\right)^{\frac{d-1}{2}} Z(W). \tag{10.33}$$

For regular trees, under the canonical coupling, for any $\epsilon > 0$, there exists k^* such that for all $k \geq k^*$,

$$\begin{aligned}&\mathbb{E}\left[\exp\left(-\frac{1}{2}\widetilde{R}^+_{\rho,k+1}\right) - \exp\left(-\frac{1}{2}R^+_{\rho,k+1}\right)\right]\\ &\leq (1+\epsilon)C_1\mathbb{E}\left[\exp\left(-\frac{1}{2}\widetilde{R}^+_{\rho,k}\right) - \exp\left(-\frac{1}{2}R^+_{\rho,k}\right)\right]. \end{aligned} \tag{10.34}$$

In particular, if $C_1 < 1$, then (10.29) *holds.*

For Galton-Watson trees with Poisson offspring distribution, the same holds with C_1 replaced by

$$C_2 = C_2(\theta, d, W) = d\theta^2 \exp\left(-d\left(1 - \sqrt{1 - \frac{(d\theta^2-1)_+}{d}}\right)\right) Z(W). \tag{10.35}$$

Proof of Prop. 10.6 is deferred to Section 10.2.3.

Prop. 10.5 and 10.6 complete the proof of Theorem 10.4, because for $i = 1, 2$, we have

$$C_i \leq d\theta^2 \exp\left(-\frac{(d\theta^2-1)_+}{2}\right) Z(W). \tag{10.36}$$

10.2.3 Proof of Prop. 10.6

Let us first deal with the regular tree case. Let u be a vertex and $v_1, \ldots, v_d$ be its children. Let $R^+_{v_1,k}, \ldots, R^+_{v_d,k}$ be i.i.d. $\sim \mu^+_k$, and $\widetilde{R}^+_{v_1,k}, \ldots, \widetilde{R}^+_{v_d,k}$ be i.i.d. $\sim \widetilde{\mu}^+_k$. Define $R^+_{u,k+1}$ and $\widetilde{R}^+_{u,k+1}$ using (10.25). Furthermore, for $0 \leq i \leq d$, define $R^+_{u,i,k+1}$ as

$$R^+_{u,i,k+1} = \sum_{1\leq j\leq i} Z_j F_\theta(\widetilde{R}^+_{v_j,k}) + \sum_{i+1\leq j\leq d} Z_j F_\theta(R^+_{v_j,k}) + W_u. \tag{10.37}$$

That is, $R^+_{u,0,k+1} = R^+_{u,k+1}$, and $R^+_{u,d,k+1} = \widetilde{R}^+_{u,k+1}$.

For $1 \leq i \leq d$ and k large enough, let us prove that

$$\begin{aligned}&\mathbb{E}\left[\exp\left(-\frac{1}{2}R^+_{u,i,k+1}\right) - \exp\left(-\frac{1}{2}R^+_{u,i-1,k+1}\right)\right]\\ &\leq (1+\epsilon)\frac{C_1}{d}\mathbb{E}\left[\exp\left(-\frac{1}{2}\widetilde{R}^+_{v_1,k}\right) - \exp\left(-\frac{1}{2}R^+_{v_1,k}\right)\right]\end{aligned} \tag{10.38}$$

where C_1 is defined in (10.33). We prove that (10.38) is true even if conditioned on $\widetilde{\Delta}_{v_i,k}$. For $\Delta \in [0, \frac{1}{2}]$, define

$$G(\Delta) = \mathbb{E}\left[\exp\left(-\frac{1}{2}R^+_{u,i,k+1}\right) - (1+\epsilon)\frac{C_1}{d}\exp\left(-\frac{1}{2}\widetilde{R}^+_{v_i,k}\right) | \widetilde{\Delta}_{v_i,k} = \Delta\right]. \tag{10.39}$$

Define $p(\Delta) = -g(\Delta) = 2\sqrt{\Delta(1-\Delta)}$ so that we work with non-negative numbers. So

$$\mathbb{E}\left[\exp\left(-\frac{1}{2}\widetilde{R}^+_{v_i,k}\right)|\widetilde{\Delta}_{v_i,k} = \Delta\right] = p(\Delta). \tag{10.40}$$

Then

$$\begin{aligned}
&\mathbb{E}\left[\exp\left(-\frac{1}{2}R^+_{u,i,k+1}\right)|\widetilde{\Delta}_{v_i,k} = \Delta\right] \\
&= \prod_{1\le j\le i-1}\mathbb{E}\left[\exp\left(-\frac{1}{2}Z_jF_\theta(\widetilde{R}^+_{v_j,k})\right)\right]\cdot\prod_{i+1\le j\le d}\mathbb{E}\left[\exp\left(-\frac{1}{2}Z_jF_\theta(R^+_{v_j,k})\right)\right] \\
&\cdot\mathbb{E}\left[\exp\left(-\frac{1}{2}Z_iF_\theta(\widetilde{R}^+_{v_i,k})\right)|\widetilde{\Delta}_{v_i,k} = \Delta\right]\cdot\mathbb{E}\left[\exp\left(-\frac{1}{2}W_u\right)\right].
\end{aligned} \tag{10.41}$$

Let us examine $\mathbb{E}\left[\exp\left(-\frac{1}{2}Z_iF_\theta(\widetilde{R}^+_{v_i,k})\right)|\widetilde{\Delta}_{v_i,k} = \Delta\right]$. We can compute that

$$\exp\left(-\frac{1}{2}Z_iF_\theta(\widetilde{R}^+_{v_i,k})\right) = \begin{cases}\exp\left(-\frac{1}{2}\log\frac{1-\Delta\star\delta}{\Delta\star\delta}\right), & \text{w.p. } 1-\Delta\star\delta,\\ \exp\left(+\frac{1}{2}\log\frac{1-\Delta\star\delta}{\Delta\star\delta}\right), & \text{w.p. } \Delta\star\delta,\end{cases} \tag{10.42}$$

where we use notation $\delta_1\star\delta_2 = \delta_1(1-\delta_2)+\delta_2(1-\delta_1)$. So

$$\mathbb{E}\left[\exp\left(-\frac{1}{2}Z_iF_\theta(\widetilde{R}^+_{v_i,k})\right)|\widetilde{\Delta}_{v_i,k} = \Delta\right] = \mathbb{E}[p(\Delta\star\delta)]. \tag{10.43}$$

Similarly,

$$\mathbb{E}\left[\exp\left(-\frac{1}{2}Z_jF_\theta(\widetilde{R}^+_{v_j,k})\right)\right] = \mathbb{E}[p(\widetilde{\Delta}_{v_1,k}\star\delta)], \tag{10.44}$$

$$\mathbb{E}\left[\exp\left(-\frac{1}{2}Z_jF_\theta(R^+_{v_j,k})\right)\right] = \mathbb{E}[p(\Delta_{v_1,k}\star\delta)]. \tag{10.45}$$

Finally,

$$\mathbb{E}\left[\exp\left(-\frac{1}{2}W_u\right)\right] = \mathbb{E}[p(\Delta_W)] = Z(W). \tag{10.46}$$

So from (10.41) we get

$$\mathbb{E}\left[\exp\left(-\frac{1}{2}R^+_{u,i,k+1}\right)|\widetilde{\Delta}_{v_i,k} = \Delta\right] = \mathbb{E}[p(\widetilde{\Delta}_{v_1,k}\star\delta)]^{i-1}\mathbb{E}[p(\Delta_{v_1,k}\star\delta)]^{d-i}p(\Delta\star\delta)Z(W). \tag{10.47}$$

So

$$G''(\Delta) = \mathbb{E}[p(\widetilde{\Delta}_{v_1,k} \star \delta)]^{i-1}\mathbb{E}[p(\Delta_{v_1,k} \star \delta)]^{d-i} Z(W) \frac{d^2}{d\Delta^2} p(\Delta \star \delta) - (1+\epsilon)\frac{C_1}{d} p''(\Delta). \tag{10.48}$$

Let us bound each factor.

$$p(\Delta \star \delta) = 2\sqrt{(\Delta \star \delta)(1 - \Delta \star \delta)} = \sqrt{1 - \theta^2(1-2\Delta)^2}. \tag{10.49}$$

So

$$\mathbb{E}[p(\widetilde{\Delta}_{v_1,k} \star \delta)] = \mathbb{E}[\sqrt{1 - \theta^2(1 - 2\widetilde{\Delta}_{v_1,k})^2}] \leq \sqrt{1 - \theta^2 \mathbb{E}(1 - 2\widetilde{\Delta}_{v_1,k})^2}. \tag{10.50}$$

By Prop. 10.7, for any $\epsilon' > 0$, for k large enough, we have

$$\mathbb{E}[(1 - 2\widetilde{\Delta}_{v_1,k})^2] \geq \left(\frac{d\theta^2 - 1}{(d-1)\theta^2} - \epsilon'\right)_+. \tag{10.51}$$

So

$$\mathbb{E}[p(\widetilde{\Delta}_{v_1,k} \star \delta)] \leq \sqrt{1 - \theta^2 \left(\frac{d\theta^2 - 1}{(d-1)\theta^2} - \epsilon'\right)_+}. \tag{10.52}$$

Similarly,

$$\mathbb{E}[p(\Delta_{v_1,k} \star \delta)] \leq \sqrt{1 - \theta^2 \left(\frac{d\theta^2 - 1}{(d-1)\theta^2} - \epsilon'\right)_+}. \tag{10.53}$$

Note that p is strictly concave on $[0, \frac{1}{2}]$, and $p'\left(\frac{1}{2}\right) = 0$. So

$$\frac{d^2}{d\Delta^2} p(\Delta \star \delta) \geq \theta^2 p''(\Delta). \tag{10.54}$$

So (10.48) gives

$$\begin{aligned} G''(\Delta) &\geq \left(1 - \theta^2 \left(\frac{d\theta^2 - 1}{(d-1)\theta^2} - \epsilon'\right)_+\right)^{\frac{d-1}{2}} \theta^2 p''(\Delta) Z(W) - (1+\epsilon)\frac{C_1}{d} p''(\Delta) \\ &= \left(\left(1 - \theta^2 \left(\frac{d\theta^2 - 1}{(d-1)\theta^2} - \epsilon'\right)_+\right)^{\frac{d-1}{2}} \theta^2 Z(W) - (1+\epsilon)\frac{C_1}{d}\right) p''(\Delta). \end{aligned} \tag{10.55}$$

Note that

$$\lim_{\epsilon' \to 0} \left(1 - \theta^2 \left(\frac{d\theta^2 - 1}{(d-1)\theta^2} - \epsilon'\right)_+\right)^{\frac{d-1}{2}} = \left(1 - \frac{(d\theta^2 - 1)_+}{d-1}\right)^{\frac{d-1}{2}}. \tag{10.56}$$

So we can take $\epsilon' > 0$ small enough so that

$$\left(1-\theta^2\left(\frac{d\theta^2-1}{(d-1)\theta^2}-\epsilon'\right)_+\right)^{\frac{d-1}{2}} < (1+\epsilon)\left(1-\frac{(d\theta^2-1)_+}{d-1}\right)^{\frac{d-1}{2}}. \tag{10.57}$$

So for k large enough, $G''(\Delta) \leq 0$ for all $\Delta \in [0, \frac{1}{2}]$ and $G(\Delta)$ is convex. Also,

$$\begin{aligned} G'\left(\frac{1}{2}\right) &= \mathbb{E}[p(\widetilde{\Delta}_{v_1,k} \star \delta)]^{i-1}\mathbb{E}[p(\Delta_{v_1,k} \star \delta)]^{d-i} Z(W) \frac{d}{d\Delta}|_{\Delta=\frac{1}{2}} p(\Delta \star \delta) \\ &\quad - (1+\epsilon)\frac{C_1}{d}p'\left(\frac{1}{2}\right) \\ &= 0. \end{aligned} \tag{10.58}$$

So G' is non-positive, thus G is decreasing on $[0, \frac{1}{2}]$. Because $M_k(v_i)$ (BMS corresponding to $R^+_{v_i,k}$) is less degraded than $\widetilde{M}_k(v_i)$ (BMS corresponding to $\widetilde{R}^+_{v_i,k}$), we get (10.38).

For Galton-Watson trees with Poisson offspring distribution, the proof is very similar to, and slightly more involved than the regular case. Let u be a vertex. Let $R^+_{v_1,k}, R^+_{v_2,k}, \ldots$ be i.i.d. $\sim \mu^+_k$, and $\widetilde{R}^+_{v_1,k}, \widetilde{R}^+_{v_2,k}, \ldots$ be i.i.d. $\sim \widetilde{\mu}^+_k$. Let $b \sim \mathrm{Pois}(d)$ and $v_1, \ldots, v_b$ be the children of u. For $i \geq 0$, define

$$R^+_{u,i,k+1} = \sum_{1 \leq j \leq \min\{i,b\}} Z_j F_\theta(\widetilde{R}^+_{v_j,k}) + \sum_{i+1 \leq j \leq b} Z_j F_\theta(R^+_{v_j,k}) + W_u. \tag{10.59}$$

For $i \geq 1$, let us prove that

$$\begin{aligned} &\mathbb{E}\left[\exp\left(-\frac{1}{2}R^+_{u,i,k+1}\right) - \exp\left(-\frac{1}{2}R^+_{u,i-1,k+1}\right)\right] \\ &\leq c_i \mathbb{E}\left[\exp\left(-\frac{1}{2}\widetilde{R}^+_{v_1,k}\right) - \exp\left(-\frac{1}{2}R^+_{v_1,k}\right)\right]. \end{aligned} \tag{10.60}$$

where c_i are constants to be chosen later. Define

$$G_i(\Delta) = \mathbb{E}\left[\exp\left(-\frac{1}{2}\widetilde{R}^+_{u,i,k+1}\right) - c_i \exp\left(-\frac{1}{2}\widetilde{R}^+_{v_i,k}\right) | \widetilde{\Delta}_{v_i,k} = \Delta\right]. \tag{10.61}$$

Let us prove that G_i is decreasing and convex on $[0, \frac{1}{2}]$. Similarly to (10.48), we have

$$G_i''(\Delta) = \mathbb{E}_b[\mathbb{1}_{b \geq i}\mathbb{E}[p(\widetilde{\Delta}_{v_1,k} \star \delta)]^{i-1}\mathbb{E}[p(\Delta_{v_1,k} \star \delta)]^{b-i} Z(W) \frac{d^2}{d\Delta^2} p(\Delta \star \delta)] - c_i p''(\Delta). \tag{10.62}$$

Let us study each term in (10.62). By (10.49) and Prop. 10.7, for any $\epsilon' > 0$, for k

large enough, we have

$$\mathbb{E}[p(\widetilde{\Delta}_{v_1,k} \star \delta)] \leq \sqrt{1-\theta^2\mathbb{E}(1-2\widetilde{\Delta}_{v_1,k})^2} \leq \sqrt{1-\theta^2\left(\frac{d\theta^2-1}{d\theta^2}-\epsilon'\right)_+}. \tag{10.63}$$

Similarly,

$$\mathbb{E}[p(\Delta_{v_1,k} \star \delta)] \leq \sqrt{1-\theta^2\left(\frac{d\theta^2-1}{d\theta^2}-\epsilon'\right)_+}. \tag{10.64}$$

(10.54) still holds in the Poisson case. So (10.62) gives

$$G_i''(\Delta) \geq \left(\mathbb{E}_b\left[\mathbb{1}_{b\geq i}\left(1-\theta^2\left(\frac{d\theta^2-1}{d\theta^2}-\epsilon'\right)_+\right)^{\frac{b-1}{2}}\right]\theta^2 Z(W) - c_i\right)p''(\Delta). \tag{10.65}$$

We can take

$$c_i = \mathbb{E}_b\left[\mathbb{1}_{b\geq i}\left(1-\theta^2\left(\frac{d\theta^2-1}{d\theta^2}-\epsilon'\right)_+\right)^{\frac{b-1}{2}}\right]\theta^2 Z(W) \tag{10.66}$$

so that $G_i''(\Delta) \geq 0$ for all $i \geq 1$ and $\Delta \in [0, \frac{1}{2}]$. Also,

$$\begin{aligned} G_i'\left(\frac{1}{2}\right) &= \mathbb{E}_b[\mathbb{1}_{b\geq i}\mathbb{E}[p(\widetilde{\Delta}_{v_1,k} \star \delta)]^{i-1}\mathbb{E}[p(\Delta_{v_1,k} \star \delta)]^{b-i}Z(W)\frac{d}{d\Delta}|_{\Delta=\frac{1}{2}}p(\Delta \star \delta)] \\ &\quad - c_i p'\left(\frac{1}{2}\right) \\ &= 0. \end{aligned} \tag{10.67}$$

So G_i is decreasing.

By summing up (10.60) for $i \geq 1$, we get

$$\begin{aligned} &\mathbb{E}\left[\exp\left(-\frac{1}{2}\widetilde{R}^+_{u,k+1}\right) - \exp\left(-\frac{1}{2}R^+_{u,k+1}\right)\right] \\ \leq\ &\left(\sum_{i\geq 1} c_i\right)\mathbb{E}\left[\exp\left(-\frac{1}{2}\widetilde{R}^+_{v_1,k}\right) - \exp\left(-\frac{1}{2}R^+_{v_1,k}\right)\right]. \end{aligned} \tag{10.68}$$

By (10.66), we have

$$\sum_{i\geq 1} c_i = \theta^2\mathbb{E}_b\left[\mathbb{1}_{b\geq i}\left(1-\theta^2\left(\frac{d\theta^2-1}{d\theta^2}-\epsilon'\right)_+\right)^{\frac{b-1}{2}}\right]Z(W) \tag{10.69}$$

$$\leq d\theta^2\exp\left(-d\left(1-\sqrt{1-\theta^2\left(\frac{d\theta^2-1}{d\theta^2}-\epsilon'\right)_+}\right)\right)Z(W). \tag{10.70}$$

We can take $\epsilon' > 0$ small enough so that

$$\exp\left(-d\left(1-\sqrt{1-\theta^2\left(\frac{d\theta^2-1}{d\theta^2}-\epsilon'\right)_+}\right)\right)$$
$$< (1+\epsilon)\exp\left(-d\left(1-\sqrt{1-\frac{(d\theta^2-1)_+}{d}}\right)\right). \tag{10.71}$$

This finishes the proof for the Poisson tree case.

10.2.4 χ^2-capacity of BOT channels

Proposition 10.7. *Consider the model* $\mathrm{BOT}(2,\theta,d)$ *or* $\mathrm{BOT}(2,\theta,\mathrm{Pois}(d))$*, with the following observation models:*

- $M_k^1 : \sigma_\rho \to \nu_{L_k}$*, where* $\nu_v \sim \mathrm{BSC}_\eta(\cdot|\sigma_v)$*;*
- $M_k^2 : \sigma_\rho \to (\sigma_{L_k}, \omega_{T_k})$*.*
- $M_k^3 : \sigma_\rho \to \sigma_{L_k}$*;*
- $M_k^4 : \sigma_\rho \to \omega_{T_k}$ *with non-trivial survey channel* W*;*
- $M_k^5 : \sigma_\rho \to \omega_{L_k}$ *with non-trivial survey channel* W*.*

For each of the above channels, we have

- *for* $\mathrm{BOT}(2,\theta,d)$ *(regular trees):*

$$\lim_{k\to\infty} C_{\chi^2}(M_k) \geq \frac{(d\theta^2-1)_+}{\theta^2(d-1)}; \tag{10.72}$$

- *for* $\mathrm{BOT}(2,\theta,\mathrm{Pois}(d))$ *(Poisson trees):*

$$\lim_{k\to\infty} C_{\chi^2}(M_k) \geq \frac{(d\theta^2-1)_+}{d\theta^2}. \tag{10.73}$$

Proof. The χ^2-capacity is always non-negative, so the $d\theta^2 \leq 1$ case is automatic. In the following we assume $d\theta^2 > 1$.

First we observe that all M_k^i's are less degraded than M_k^1 for some suitable choice of η. This is obvious for $i = 2, 3$. Clearly M_k^4 is less degraded than M_k^5. That $M_k^5 \leq_{\deg} M_k^1$ follows from [121, Lemma 2, 3], where we can take $\eta = P_e(W)$. So by Lemma 2.51, we only need to prove the result for M_k^1.

We prove the result by applying Lemma 10.8. To do this, we need to find a BMS channel more degraded than M_k^1 which takes value in $\mathbb{R}$. One natrual choice is the majority decoder. We define

$$S_k = \sum_{v\in L_k} \nu_v. \tag{10.74}$$

Then the channel $\sigma_\rho \to S_k$ is clearly more degraded than M_k^1. We apply Prop. 10.9 to conclude. □

Lemma 10.8 (Restatement of [63, Lemma 4.2(iii)])**.** *Let $P : X \to Y$ be a BMS channel with Y a real variable, and with involution $Y \mapsto -Y$. Then $C_{\chi^2}(P) \geq \frac{(\mathbb{E}^+ Y)^2}{\mathrm{Var}(Y)}$.*

Proof. Let $X \to (\Delta, Z)$ be the equivalent standard form of P. By Cauchy-Schwarz, we have

$$\mathbb{E}^+[(1-2\Delta)^2]\mathbb{E}^+[Y^2] \geq (\mathbb{E}^+[(1-2\Delta)|Y|])^2 = (\mathbb{E}^+ Y)^2. \tag{10.75}$$

This is equivalent to the desired result. □

Proposition 10.9. *Assume $d\theta^2 > 1$. Consider the channel $\sigma_\rho \to S_k$ defined in* (10.74)*.*

For $\mathrm{BOT}(2, \theta, d)$ *(regular trees),*

$$\lim_{k\to\infty} \frac{\mathrm{Var}^+ S_k}{(\mathbb{E}^+ S_k)^2} = \frac{1-\theta^2}{d\theta^2 - 1}. \tag{10.76}$$

For $\mathrm{BOT}(2, \theta, \mathrm{Pois}(d))$ *(Poisson trees),*

$$\lim_{k\to\infty} \frac{\mathrm{Var}^+ S_k}{(\mathbb{E}^+ S_k)^2} = \frac{1}{d\theta^2 - 1}. \tag{10.77}$$

Proof. The regular tree case is proved in [104, Lemma 3.4, 3.5]. (Note that the expression for $\lim_{k\to\infty} \frac{\mathrm{Var}^+ S_k}{(\mathbb{E}^+ S_k)^2}$ on top of [104, pg. 2224] is incorrect.)

Let us focus on the Poisson tree case. It is easy to see that

$$\mathbb{E}^+ S_k = (1-2\eta)(d\theta)^k. \tag{10.78}$$

Let ρ be the root, and $v_1, \ldots, v_b$ be its children. By variance decomposition, we have

$$\begin{aligned}\mathrm{Var}^+ S_{\rho,k+1} = \mathrm{Var}^+ \mathbb{E}[S_{\rho,k+1}|b] + \mathbb{E}_b \mathrm{Var}^+(\mathbb{E}[S_{\rho,k+1}|b, \sigma_{v_1}, \ldots, \sigma_{v_b}]|b) \\ + \mathbb{E}\,\mathrm{Var}^+(S_{\rho,k+1}|b, \sigma_{v_1}, \ldots, \sigma_{v_b}).\end{aligned} \tag{10.79}$$

Let us compute each summand.

$$\mathrm{Var}^+ \mathbb{E}[S_{\rho,k+1}|b] = \mathrm{Var}^+(b\theta(1-2\eta)(d\theta)^k) = d\theta^2(1-2\eta)^2(d\theta)^{2k}. \tag{10.80}$$

$$\begin{aligned}&\mathbb{E}_b \mathrm{Var}^+(\mathbb{E}[S_{\rho,k+1}|b, \sigma_{v_1}, \ldots, \sigma_{v_b}]|b) \\ &= \mathbb{E}_b \mathrm{Var}^+ \left(\sum_{i\in[b]} \sigma_{v_i}(1-2\eta)(d\theta)^k | b \right) \\ &= \mathbb{E}_b[b(1-\theta^2)(1-2\eta)^2(d\theta)^{2k}] \\ &= d(1-\theta^2)(1-2\eta)^2(d\theta)^{2k}.\end{aligned} \tag{10.81}$$

$$\mathbb{E}\operatorname{Var}^+(S_{\rho,k+1}|b,\sigma_{v_1},\dots,\sigma_{v_b}) = \mathbb{E}_b[b\sum_{i\in[b]}\operatorname{Var}^+ S_{v_i,k}] = d\operatorname{Var}^+ S_{\rho,k}. \tag{10.82}$$

Plugging (10.80)(10.81)(10.82) into (10.79), we get

$$\operatorname{Var}^+ S_{\rho,k+1} = d(1-2\eta)^2(d\theta)^{2k} + d\operatorname{Var}^+ S_{\rho,k}. \tag{10.83}$$

Solving (10.83) with initial value $S_{\rho,0} = 4\eta(1-\eta)$, we get

$$\begin{aligned}\operatorname{Var}^+ S_{\rho,k} &= 4\eta(1-\eta)d^k + \sum_{i\in[k]} d^{k-i}d(1-2\eta)^2(d\theta)^{2i-2} \\ &= 4\eta(1-\eta)d^k + (1-2\eta)^2 d^k \frac{(d\theta^2)^k - 1}{d\theta^2 - 1}.\end{aligned} \tag{10.84}$$

Putting together (10.78)(10.84), we get the desired result. □

10.2.5 Proof of Theorem 10.3

We prove the following strengthening of Theorem 10.3 (note that $\mathrm{BP}_W = \mathrm{BP}$ when W is trivial).

Theorem 10.10. *Consider the model* $\mathrm{BOT}(2,\theta,d)$. *Let* W *be a (possibility trivial) BMS channel. Then we have the following results on the fixed points of the operator* BP_W.

(i) If W *is non-trivial and* $C_1 < 1$ *(defined in Eq.* (10.33)*), there is exactly one BP fixed point. Furthermore, starting from any initial BMS channel* M, $\mathrm{BP}^k(M)$ *converges to the unique BP fixed point as* $k \to \infty$.

(ii) If W *is trivial and* $d\theta^2 \le 1$, *there is exactly one BP fixed point, which is trivial. Furthermore, starting from any initial BMS channel, the BP recursion converges to the unique BP fixed point.*

(iii) If W *is trivial,* $d\theta^2 > 1$, *and* $C_1 < 1$, *there are exactly two BP fixed points, one is trivial and the other is non-trivial. Furthermore, starting from any non-trivial (resp. trivial) BMS channel, the BP recursion converges to the non-trivial (resp. trivial) fixed point.*

The same results hold for $\mathrm{BOT}(2,\theta,\mathrm{Pois}(d))$ *with* C_1 *replaced by* C_2 *(defined in Eq.* (10.35)*).*

Proof. (i): For any channel M we have $0 \le_{\deg} M \le_{\deg} \mathrm{Id}$. Therefore

$$\mathrm{BP}_W^k(0) \le_{\deg} \mathrm{BP}_W^k(M) \le_{\deg} \mathrm{BP}_W^k(\mathrm{Id}) \tag{10.85}$$

for all $k \ge 0$. By proof of Theorem 10.4, the first and the third term converge to the same non-trivial fixed point as $k \to \infty$. Therefore the middle term also converges to that fixed point.

If there is another fixed point P, then taking initial channel to be P, the BP recursion would always stay at P. This is impossible because the choice of M in the above discussion is arbitrary.

(ii): If W is trivial and $d\theta^2 \leq 1$, we are in the non-reconstruction regime and there is a unique BP fixed point, which is trivial. Because $\mathrm{BP}^k(M) \leq_{\deg} \mathrm{BP}^k(\mathrm{Id})$ for any k and initial channel M, BP recursion always converges to the trivial channel.

(iii): There is one trivial fixed point.

Let M be a non-trivial BMS channel. Taking $r = P_e(M) < \frac{1}{2}$, then $\mathrm{BSC}_r \leq_{\deg} M$. Therefore

$$\mathrm{BP}^k(\mathrm{BSC}_r) \leq_{\deg} \mathrm{BP}^k(M) \leq_{\deg} \mathrm{BP}^k(\mathrm{Id}) \tag{10.86}$$

for all $k \geq 0$. By proof of Theorem 10.4, the first and the third term converge to the same non-trivial fixed point as $k \to \infty$. Therefore the middle term also converges to that fixed point.

If there is another non-trivial fixed point P, then taking initial channel to be P, the BP recursion would always stay at P. This is impossible because the choice of M in the above discussion is arbitrary. □

10.3 Applications

Main applications of uniqueness of BP fixed point and boundary irrelevance include a mutual information formula and an optimal recovery algorithm. In this section we prove these results via reduction established in Chapter 5.

Theorem 10.11 (Mutual information formula)**.** *Let* $(X, G) \sim \mathrm{SBM}(n, 2, a, b)$*. Let* $(\sigma, T) \sim \mathrm{BOT}(2, \theta, \mathrm{Pois}(d))$ *be the corresponding BOT model, where* $\theta = \frac{a-b}{a+b}$ *and* $d = \frac{a+b}{2}$*. Let* ω^ϵ *denote the observation of* σ *through* BEC_ϵ*. Let* $\alpha^* \approx 3.513$ *be the unique solution in* $\mathbb{R}_{>1}$ *to the equation* $\exp\left(-\frac{\alpha-1}{2}\right)\alpha = 1$*. The following hold.*

(i) For a, b *such that* $d\theta^2 \leq 1$ *or* $d\theta^2 \geq \alpha^* \approx 3.513$

$$\lim_{n\to\infty} \frac{1}{n} I(X; G) = \int_0^1 \lim_{k\to\infty} I(\sigma_\rho; \omega^\epsilon_{T_k \setminus \rho} | T_k) d\epsilon. \tag{10.87}$$

(ii) For any a, b *such that* $d\theta^2 \in (1, \alpha^*)$*, i.e., inside the gap of (i),*

$$\liminf_{n\to\infty} \frac{1}{n} I(X; G) = \int_0^1 \lim_{k\to\infty} I(\sigma_\rho; \omega^\epsilon_{T_k \setminus \rho} | T_k) d\epsilon + \xi_{\inf} \log 2, \tag{10.88}$$

$$\limsup_{n\to\infty} \frac{1}{n} I(X; G) = \int_0^1 \lim_{k\to\infty} I(\sigma_\rho; \omega^\epsilon_{T_k \setminus \rho} | T_k) d\epsilon + \xi_{\sup} \log 2, \tag{10.89}$$

where $0 \leq \xi_{\inf}, \xi_{\sup} \leq 1 - \frac{\sqrt{e}}{2} \approx 0.178$*.*

Proof. (i) hold directly by Theorem 10.1 and Theorem 5.15.

For (ii), we look into the proof of Theorem 5.15. By Corollary 10.2(iii), BI holds for all $\epsilon < \epsilon^* = \frac{\sqrt{e}}{2}$. Therefore

$$\liminf_{n\to\infty} \frac{1}{n} I(X;G) = \int_0^{\epsilon^*} \lim_{k\to\infty} I(\sigma_\rho; \omega^{\epsilon}_{T_k\setminus\rho}|T_k) + \xi_{\inf}, \tag{10.90}$$

where

$$\xi_{\inf} = \int_{\epsilon^*}^1 \liminf_{n\to\infty} I(X_u; G, Y^{\epsilon}_{V\setminus u}) \le (1-\epsilon^*)H(X_u). \tag{10.91}$$

The proof for lim sup is similar. □

Theorem 10.12 (Optimal recovery for SBM with survey)**.** *Work under the same setting as Theorem 10.1. Suppose that in addition to G, we observe survey $Y_v \sim W(\cdot|X_v)$ for all $v \in V$, where W is some non-trivial FMS channel. If $d\theta^2 \le 1$ or $d\theta^2 \ge \alpha^* \approx 3.513$, then belief propagation (Algorithm 1) achieves the optimal recovery accuracy of*

$$1 - \lim_{k\to\infty} P_e(\sigma_\rho|T_k, \omega_{T_k}). \tag{10.92}$$

Proof. By Theorem 10.1 and Theorem 5.20. □

Theorem 10.13 (Optimal recovery for SBM)**.** *Work under the same setting as Theorem 10.1. If $d\theta^2 \ge \alpha^* \approx 3.513$, then there is an algorithm (Algorithm 2) achieving the optimal recovery accuracy of*

$$1 - \lim_{k\to\infty} P_e(\sigma_\rho|T_k, \sigma_{L_k}). \tag{10.93}$$

Proof. By Theorem 10.1 and Theorem 5.22, we only need an initial weak recovery algorithm that satisfies the conditions in Theorem 5.22. By discussions before the proof of Theorem 5.22, in the two-community symmetric case, any weak recovery algorithm can be modified into one that satisfies the conditions. So we can use the weak recovery algorithms in [96] or [103]. □

Chapter 11

Uniqueness of BP fixed point: Potts model

We generalize the degradation method introduced in Chapter 10 to Potts models $\mathrm{BOT}(q, \lambda, d)$ and $\mathrm{BOT}(q, \lambda, \mathrm{Pois}(d))$. We prove that stability of BP fixed points (Definition 5.21) and boundary irrelevance (Definition 5.17) hold when $d\lambda^2 \geq 1 + C \max\{\lambda, q^{-1}\} \log q$ for some absolute constant $C > 0$ independent of q, λ, d. For large q and $\lambda = o(1/\log q)$, this is asymptotically achieving the Kesten-Stigum threshold $d\lambda^2 = 1$. These results imply mutual information formula and optimal recovery algorithms for the q-community symmetric SBM in the corresponding ranges.

This chapter is based on [73].

Chapter outline In Section 11.1, we introduce the setting and main results in this chapter. In Section 11.2, we give some preliminaries on limits of information channels. In Section 11.3, we prove Theorem 11.1, boundary irrelevance and uniqueness of BP fixed point for a wide range of parameters. In Section 11.4, we discuss applications of uniqueness of BP fixed points and boundary irrelevance. In Section 11.5, we discuss asymmetric fixed points of the BP operator.

11.1 Introduction

Stochastic block model We consider the model $\mathrm{SBM}(n, q, a, b)$ (Definition 5.4), defined as follows. The model has four parameters $n \in \mathbb{Z}_{\geq 1}$, $q \in \mathbb{Z}_{\geq 2}$, $a, b \in \mathbb{R}_{\geq 0}$. First, we assign a random label (community) $X_i \sim \mathrm{Unif}([q])$ i.i.d for $i \in V = [n]$. Then a random graph $G = (V, E)$ is generated, where $(i, j) \in E$ with probability $\frac{a}{n}$ if $X_i = X_j$, and with probability $\frac{b}{n}$ if $X_i \neq X_j$, independently for all $(i, j) \in \binom{V}{2}$. When $a > b$, we say the model is assortative. When $a < b$, we say the model is disassortative.

For the SBM, an important problem is weak recovery. We say the model admits

weak recovery if there exists an estimator $\widehat{X}(G) \in [q]^V$ such that

$$\lim_{n\to\infty} \frac{1}{n}\mathbb{E}d_H(X, \widehat{X}(G)) < 1 - \frac{1}{q}, \tag{11.1}$$

$$\text{where} \quad d_H(X, Y) := \min_{\tau\in \mathrm{Aut}([q])} \sum_{i\in[n]} \mathbb{1}\{X_i \neq \tau(Y_i)\}. \tag{11.2}$$

When weak recovery is possible, the natural follow-up question is to determine the optimal recovery accuracy

$$\sup_{\hat{X}=\hat{X}(G)} \lim_{n\to\infty} \mathbb{E}\left[1 - \frac{1}{n} d_H(X, \widehat{X}(G))\right]. \tag{11.3}$$

A fundamental quantity of the stochastic block model is its (normalized) mutual information

$$\lim_{n\to\infty} \frac{1}{n} I(X; G). \tag{11.4}$$

We refer the reader to Chapter 5 for a review of previous results on weak recovery, optimal recovery, and mutual information.

Broadcasting on trees The stochastic block model has a close relationship with the broadcasting on trees (BOT) model. The reason is that in SBM, the local neighborhood of a random vertex converges (in the sense of local weak convergence) to a Galton-Watson tree with Poisson offspring distribution. Therefore, properties of BOT can often imply corresponding results on SBM.

For the model $\mathrm{SBM}(n, q, a, b)$ we consider, the corresponding model is the Potts model $\mathrm{BOT}(q, \lambda, \mathrm{Pois}(d))$, defined as follows. The model has three parameters $q \in \mathbb{Z}_{\geq 2}$, $\lambda \in \left[-\frac{1}{q-1}, 1\right]$, $d \in \mathbb{R}_{\geq 0}$. Let T be a Galton-Watson tree with offspring distribution $\mathrm{Pois}(d)$, rooted at ρ. We assign to every vertex v a label $\sigma_v \in [q]$ according to the following process.

1. Generate $\sigma_\rho \sim \mathrm{Unif}([q])$.

2. Suppose we have generated a label for a vertex u. For every child v of u, we generate σ_v according to $P_\lambda(\cdot|\sigma_\mu)$, where P_λ is the Potts channel defined as

$$P_\lambda(j|i) = \lambda \mathbb{1}\{i = j\} + \frac{1-\lambda}{q}. \tag{11.5}$$

We also consider the case where T is a regular tree (every vertex has $d \in \mathbb{Z}_{\geq 0}$ children), denoted as $\mathrm{BOT}(q, \lambda, d)$.

An important problem on BOT is the reconstruction problem, asking whether we can gain any non-trivial information about the root given observation of far away

vertices. We say the model admits reconstruction if

$$\lim_{k\to\infty} I(\sigma_\rho; \sigma_{L_k}|T_k) > 0, \tag{11.6}$$

where L_k stands for the set of vertices at distance k to the root ρ. We say the model admits non-reconstruction if the limit is zero. It is known that non-reconstruction results for the Potts model imply impossibility of weak recovery for the corresponding SBM (Theorem 5.15), although the other side does not hold: in the case $a = 0$, there is a gap of factor 2 (as $q \to \infty$) between the BOT reconstruction threshold and the SBM weak recovery threshold.

We refer the reader to Chapter 5 for a review of previous results on reconstruction on trees.

Belief propagation Belief propagation is a powerful tool for studying the BOT model. It is usually described as an algorithm for computing posterior distribution of vertex labels given observation. Here we take an information-theoretic point of view and describe BP in terms of communication channels.

We view the BOT model as an information channel from the root label to the observation. Let M_k denote the channel $\sigma_\rho \mapsto \sigma_{L_k}$.Then $(M_k)_{k\geq 0}$ satisfies the following recursion, which we call belief propagation recursion:

$$M_{k+1} = \mathbb{E}_b(M_k \circ P_\lambda)^{\star b} \tag{11.7}$$

where b following the branching number distribution (constant in the regular tree case, Pois(d) in the Poisson tree case), and $(\cdot)^{\star b}$ denotes $\star$-convolution power. Let BP be the operator

$$\mathrm{BP}(M) := \mathbb{E}_b(M \circ P_\lambda)^{\star b} \tag{11.8}$$

defined on the space of information channels with input alphabet $[q]$. Due to symmetry in labels, we can regard BP as an operator on the space of FMS channels (Chapter 3). In terms of the BP operator, the reconstruction problem can be rephrased as whether the limit channel $\mathrm{BP}^\infty(\mathrm{Id}) := \lim_{k\to\infty} \mathrm{BP}^k(\mathrm{Id})$ is trivial or not. The problem of optimal recovery for SBM can be reduced to the following problem on trees: whether the limit

$$\lim_{n\to\infty} I(\sigma_\rho; \omega_{L_k}|T_k) \tag{11.9}$$

where ω is the observation of σ through a non-trivial channel W, stays the same for any non-trivial FMS W. Therefore, it is important to study the non-trivial fixed points of the BP operator (the trivial channel is always a fixed point).

[104] proved uniqueness of BP fixed point for $q = 2$ and large enough SNR. [4] improved to $q = 2$ and SNR $\notin [1, 3.513]$. [137] proved uniqueness of BP fixed point for $q = 2$ and any parameter d, λ, closing the question for binary symmetric models. For $q \geq 3$, [37] proved that when the initial channel U is close enough to Id, and

$d\lambda^2 > C_q$, where C_q is a constant depending on q, then $\mathrm{BP}^\infty(U) = \mathrm{BP}^\infty(\mathrm{Id})$. They did not give asymptotics for C_q, but it seems like it is at least polynomial in q. [38] generalized [37] to asymmetric models.

Boundary irrelevance [4] reduced the SBM mutual information problem to the boundary irrelevance problem, on a tree model called the broadcasting on trees with survey (BOTS) model. In the BOTS model, we observe label of every vertex through a noisy q-FMS channel W (called the survey). We say the model admits boundary irrelevance with respect to W if

$$\lim_{k\to\infty} I(\sigma_\rho; \sigma_{L_k} | T_k, \omega_{T_k}) = 0, \tag{11.10}$$

where T_k is the set of all vertices within distance at most k to the root, and ω is the observation of σ through W. We say the model admits boundary irrelevance if the model admits boundary irrelevance with respect to all erasure channels EC_ϵ with $0 \le \epsilon < 1$. Boundary irrelevance is equivalent to the condition that the operator

$$\mathrm{BP}_W(M) := \left(\mathbb{E}_b (M \circ P_\lambda)^{\star b}\right) \star W \tag{11.11}$$

has a unique fixed point in the space of q-FMS channels. Because BP and BP_W have very similar forms, the boundary irrelevance problem has a close relationship with the problem of uniqueness of BP fixed point. Indeed, these two problems can be solved using the same method.

Our main result Our main result is stability of BP fixed point and boundary irrelevance for a wide range of parameters. For a more precise statement, see Theorem 11.5 and Theorem 11.6.

Theorem 11.1 (Uniqueness of BP fixed point and boundary irrelevance)**.** *There exists an absolute constant $C > 0$ such that the following statement holds. Consider the model $\mathrm{BOT}(q, \lambda, d)$ or $\mathrm{BOT}(q, \lambda, \mathrm{Pois}(d))$. If either $d\lambda^2 < q^{-2}$ or $d\lambda^2 > 1 + C \max\{\lambda, q^{-1}\} \log q$, then boundary irrelevance holds. That is, for any non-trivial q-FMS survey channel W, we have*

$$\lim_{k\to\infty} I(\sigma_\rho; \sigma_{L_k} | T_k, \omega_{T_k}) = 0. \tag{11.12}$$

Furthermore, under the same conditions, stability of BP fixed point holds, i.e., for any non-trivial q-FMS channel P, $\mathrm{BP}^k(P)$ and $\mathrm{BP}^k(\mathrm{Id})$ converge weakly to the same limit as $k \to \infty$.

See Section 11.4 for applications to the q-community symmetric SBM.

Our technique We generalize the degradation method of [4] to q-ary symmetric channels. In this method, we find suitable potential functions Φ on the space of FMS channels, such that for two channels $M, \widetilde{M}$ are related by degradation ($\widetilde{M} \le_{\deg} M$),

we have (1) $\Phi(M)-\Phi(\widetilde{M})$ contracts to 0 under iterations of BP (2) if $\Phi(M)=\Phi(\widetilde{M})$, then $M=\widetilde{M}$. This shows that the limit channels $\mathrm{BP}^\infty(M)$ and $\mathrm{BP}^\infty(\widetilde{M})$ are equal. To carry out this method, we make use of the theory of FMS channels developed in Chapter 3.

11.2 Limit of channels

In this section we build the foundation for discussing limits of information channels. We view a channel $P:\mathcal{X}\to\mathcal{Y}$ as a distribution of posterior distributions under uniform prior, i.e., the distribution of $P_{X|Y}$ where $P_X=\mathrm{Unif}(\mathcal{X})$, $Y\sim P(\cdot|X)$. Let μ denote the posterior distribution variable and $P_\mu\in\mathcal{P}(\mathcal{P}(\mathcal{X}))$ be its distribution (called P's posterior distribution's distrbution). Note that P_μ is invariant under channel equivalence.

We often work with sequences $(P_k)_{k\geq 0}$ of channels with the same input alphabet $\mathcal{X}$. Let $P_{\pi,k}$ denote the distrbution of posterior distributions of P_k under uniform prior. Let P_∞ be a channel with input alphabet $\mathcal{X}$ and posterior distribution's distrbution $P_{\pi,\infty}$. We say $(P_k)_{k\geq 0}$ converges weakly to P_∞ if $(P_{\pi,k})_{k\geq 0}$ converges weakly to $P_{\pi,\infty}$ as distributions on $\mathcal{P}(\mathcal{X})$.

In general, given such a sequence, a limit does not necessarily exist. Nevertheless, when the channels are related to each other via degradation, a limit channel exists.

Lemma 11.2. *Let $(P_k:\mathcal{X}\to\mathcal{Y}_k)_{k\geq 0}$ be a sequence of channels with the same finite input alphabet. If $P_k\geq_{\deg} P_{k+1}$ for all k, then $(P_k)_{k\geq 0}$ converges weakly to some channel P_∞.*

Proof. By definition of degradation, there exists channel $R_k:\mathcal{Y}_k\to\mathcal{Y}_{k+1}$ such that $P_{k+1}=R_k\circ P_k$. This gives rise to an infinite Markov chain

$$X-Y_0-Y_1-Y_2-\cdots. \tag{11.13}$$

Let μ_k denote the posterior distribution variable $P_{X|Y_k}$. Then we have

$$\mathbb{E}[\mu_{k-1}|Y_k]=\mu_k. \tag{11.14}$$

Let $\mathcal{F}_k$ denote the σ-algebra generated by $(Y_i)_{i\geq k}$. Then $(\mathcal{F}_k)_{k\geq 0}$ is a reverse filtration and $(\mu_k)_{k\geq 0}$ is a reverse martingale with respect to $(\mathcal{F}_k)_{k\geq 0}$. By reverse martingale convergence theorem (e.g., [57, Theorem 4.7.1]), $\lim_{k\to\infty}\mu_k$ converges almost surely. Define $\mu_\infty:=\lim_{k\to\infty}\mu_k$. Let P_∞ be a channel with input alphabet $\mathcal{X}$ whose posterior distribution's distribution is μ_∞. Then $(P_k)_{k\geq 0}$ converges weakly to P_∞. □

Lemma 11.3. *Let $(P_k:\mathcal{X}\to\mathcal{Y}_k)_{k\geq 0}$ be a sequence of channels with the same finite input alphabet. If $P_k\leq_{\deg} P_{k+1}$ for all k, then $(P_k)_{k\geq 0}$ converges weakly to some channel P_∞.*

Proof. By definition of degradation, there exists channel $R_k:\mathcal{Y}_k\to\mathcal{Y}_{k-1}$ such that

$P_{k-1} = R_k \circ P_k$. This gives rise to an infinite Markov chain

$$X - Y_0 - Y_1 - Y_2 - \cdots . \tag{11.15}$$

Let μ_k denote the posterior distribution variable $P_{X|Y_k}$. Then we have

$$\mathbb{E}[\mu_{k+1}|Y_k] = \mu_k. \tag{11.16}$$

Let $\mathcal{F}_k$ denote the σ-algebra generated by $(Y_i)_{i\le k}$. Then $(\mathcal{F}_k)_{k\ge 0}$ is a filtration and $(\mu_k)_{k\ge 0}$ is a martingale with respect to $(\mathcal{F}_k)_{k\ge 0}$. Note that the variables μ_k take values in $\mathcal{P}(\mathcal{X})$, so are uniformly bounded. By martingale convergence theorem (e.g., [57, Theorem 4.2.11]), $\lim_{k\to\infty}\mu_k$ converges almost surely. Define $\mu_\infty := \lim_{k\to\infty}\mu_k$. Let P_∞ be a channel with input alphabet $\mathcal{X}$ whose posterior distribution's distribution is μ_∞. Then $(P_k)_{k\ge 0}$ converges weakly to P_∞. □

By symmetry, in Lemma 11.2 and Lemma 11.3, if the sequence $(P_k)_{k\ge 0}$ consists of FMS channels, then the limit P_∞ is an FMS channel.

11.3 Uniqueness and boundary irrelevance

In this section we prove uniqueness of BP fixed point and boundary irrelevance results for the Potts model for a wide range of parameters. We consider the model $\mathrm{BOT}(q,\lambda,d)$ or $\mathrm{BOT}(q,\lambda,\mathrm{Pois}(d))$ (Definition 5.9).

We state two results, one for the low SNR regime and one for the high SNR regime. We define the following constants used in the results.

Definition 11.4. For $q \in \mathbb{Z}_{\ge 2}$, $\lambda \in \left[-\frac{1}{q-1}, 1\right]$, $d \ge 0$, we define

$$C^L(q,\lambda) := \sup_{\substack{\pi\in\mathcal{P}([q])\\ v\in\mathbb{1}^\perp\subseteq\mathbb{R}^q}} \frac{f^L\left(\lambda\pi + \frac{1-\lambda}{q}, v\right)}{f^L(\pi,v)}, \tag{11.17}$$

$$\text{where } f^L(\pi,v) := \left\langle \pi^{-1} + \frac{1}{q}\pi^{-2}, v^2\right\rangle, \tag{11.18}$$

$$C^H(q,\lambda) := \sup_{\substack{\pi\in\mathcal{P}([q])\\ v\in\mathbb{1}^\perp\subseteq\mathbb{R}^q}} \frac{f^H\left(\lambda\pi + \frac{1-\lambda}{q}, v\right)}{f^H(\pi,v)}, \tag{11.19}$$

$$\text{where } f^H(\pi,v) := \|\pi^{1/4}\|_2^2\|\pi^{-3/4}v\|_2^2 - \left\langle \pi^{1/4}, \pi^{-3/4}v\right\rangle^2, \tag{11.20}$$

$$c^H(q,\lambda,d) := \left(\frac{2}{q} + \frac{q-2}{q}\cdot\frac{d\lambda^2-1}{d\lambda-1}\right)^{-1}. \tag{11.21}$$

We have the following bounds on these constants: $C^L(q,\lambda) \le q^2$ (Prop. 11.17), $C^H(q,\lambda) \le q^{5/2}$ (Prop. 11.18), $c^H(q,\lambda,d) \ge 1$ (obvious).

Theorem 11.5 (Low SNR)**.** *If*

$$d\lambda^2 C^L(q,\lambda) < 1, \tag{11.22}$$

where C^L is defined in (11.17)*, then boundary irrelevance and stability of BP fixed point hold.*

Theorem 11.6 (High SNR)**.** *If $d\lambda^2 > 1$ and*

$$d\lambda^2 \exp\left(-c^H(q,\lambda,d)\cdot\frac{d\lambda^2-1}{2}\right) C^H(q,\lambda) < 1, \tag{11.23}$$

where c^H is defined in (11.21)*, C^H is defined in* (11.19)*, then boundary irrelevance and stability of BP fixed point hold.*

Let W be a q-FMS channel. If $d\lambda^2 > 1$ and

$$d\lambda^2 \exp\left(-c^H(q,\lambda,d)\cdot\frac{d\lambda^2-1}{2}\right) C^H(q,\lambda) Z(W^R) < 1, \tag{11.24}$$

where W^R denotes the restriction of W to a BMS channel (Chapter 3), and Z denotes the Bhattacharyya coefficient (Definition 2.4). then boundary irrelevance holds with repsect to W.

Proof of Theorem 11.1 given Theorem 11.5 and 11.6. We prove the low SNR case and the high SNR case separately.

Low SNR: By Prop. 11.17, $C^L(q,\lambda) \le q^2$. If $d\lambda^2 < q^{-2}$, then (11.22) holds and Theorem 11.5 applies.

High SNR: We prove that (11.23) holds whenever $d\lambda^2 > 1+56\max\{\lambda, q^{-1}\}\log q$.

By Prop. 11.18, $C^H(q,\lambda) \le q^{5/2}$. For $d\lambda^2 > 1$, we have

$$c^H(q,\lambda,d) \ge \left(\frac{2}{q} + \frac{q-2}{q}\cdot\max\{\lambda,0\}\right)^{-1} \ge \left(\frac{2}{q}+\max\{\lambda,0\}\right)^{-1} \ge \frac{1}{4}\max\{\lambda,q^{-1}\}^{-1}. \tag{11.25}$$

Therefore

$$d\lambda^2\exp(-c^H(q,\lambda,d)\cdot\frac{d\lambda^2-1}{2})C^H(q,\lambda) \le d\lambda^2\exp\left(-\frac{d\lambda^2-1}{8\max\{\lambda,q^{-1}\}}\right)q^{5/2} =: g_{q,\lambda}(d). \tag{11.26}$$

Computing $g'_{q,\lambda}(d)$, we see that $g_{q,\lambda}(d)$ is monotone decreasing in d when $d\lambda^2 > 8\max\{\lambda,q^{-1}\}$. Therefore it suffices to prove $g_{q,\lambda}(d_0) < 1$ where $d_0\lambda^2 = 1+56\max\{\lambda,q^{-1}\}\log q$. We have

$$g_{q,\lambda}(d_0) = (1+56\max\{\lambda,q^{-1}\}\log q)\exp(-7\log q)q^{5/2} \le (1+56\log q)q^{-9/2}. \tag{11.27}$$

The last expression is < 1 for all $q \ge 3$. This finishes the proof. □

11.3.1 The degradation method

Let $(M_k)_{k\geq 0}$ and $(\widetilde{M}_k)_{k\geq 0}$ be two sequences of q-FMS channels satisfying the belief propagation recursion, i.e.,

$$M_{k+1} = \mathrm{BP}(M_k), \qquad \widetilde{M}_{k+1} = \mathrm{BP}(\widetilde{M}_k), \tag{11.28}$$

$$\mathrm{BP}(M) := \mathbb{E}_b[(M_k \circ P_\lambda)^{\star b} \star W], \tag{11.29}$$

where b follows the branching number distribution (constant if working with regular trees), and W is the survey FMS channel (trivial if there is no survey).

For the boundary irrelevance problem, we take $\widetilde{M}_0 = 0$, $M_0 = \mathrm{Id}$. For stability of BP fixed point, we take $M_0 = \mathrm{Id}$ and $\widetilde{M}_0$ be a given non-trivial FMS channel. From now on, we assume that $M_0 = \mathrm{Id}$, and either (1) W is non-trivial and $\widetilde{M}_0 = 0$, or (2) $W = 0$ and $\widetilde{M}_0$ is non-trivial. Note that in both cases, the initial channels satisfy $\widetilde{M}_0 \leq_{\deg} M_0$. Therefore $\widetilde{M}_k \leq_{\deg} M_k$ for all $k \geq 0$ because the BP operator preserves degradation preorder (Lemma 2.8). So the two channel sequences are naturally related to each other by degradation.

Because $M_0 = \mathrm{Id}$, we have $M_k \geq_{\deg} M_{k+1}$ for all $k \geq 0$. Therefore by Lemma 11.2, $M_\infty := \lim_{k\to\infty} M_k$ exists. For the boundary irrelevance problem, we also have $M_k \leq_{\deg} M_{k+1}$ for all $k \geq 0$, and by Lemma 11.3, $\widetilde{M}_\infty := \lim_{k\to\infty} \widetilde{M}_k$ exists. However, for the stability of BP fixed point problem, it is a priori unclear whether the limit $\lim_{k\to\infty} \widetilde{M}_k$ exists.

Let $\phi : \mathcal{P}([q]) \to \mathbb{R}$ be a strongly convex function invariant under $\mathrm{Aut}([q])$ action. Extend it to a function $\Phi : \{\text{FMS channels}\} \to \mathbb{R}$ as $\Phi(P) = \mathbb{E}\phi(\pi_P)$. By degradation, we have $\Phi(M_k) \geq \Phi(\widetilde{M}_k)$ for all $k \geq 0$. The following proposition shows that it suffices to prove contraction of potential function Φ.

Proposition 11.7. *Assume that $\phi : \mathcal{P}([q]) \to \mathbb{R}$ is α-strongly convex for some $\alpha > 0$, and that*

$$\lim_{k\to\infty} (\Phi(M_k) - \Phi(\widetilde{M}_k)) = 0. \tag{11.30}$$

Then under the canonical coupling, we have

$$\lim_{k\to\infty} \mathbb{E}\|\pi_k - \widetilde{\pi}_k\|_2^2 = 0, \tag{11.31}$$

where π_k (resp. $\widetilde{\pi}_k$) is the π-component of M_k (resp. $\widetilde{M}_k$). In particular, if $M_0 = \mathrm{Id}$, then both limits $\lim_{k\to\infty} M_k$ and $\lim_{k\to\infty} \widetilde{M}_k$ exist in the sense of weak convergence, and the two limits are equal.

Proof.

$$\begin{aligned}\Phi(M_k) - \Phi(\widetilde{M}_k) &= \mathbb{E}_{\widetilde{\pi}_k}\mathbb{E}[\phi(\pi_k) - \phi(\widetilde{\pi}_k)|\widetilde{\pi}_k] \\ &\geq \mathbb{E}_{\widetilde{\pi}_k}\mathbb{E}[\langle\nabla\phi(\widetilde{\pi}_k), \pi_k - \widetilde{\pi}_k\rangle + \frac{\alpha}{2}\|\pi_k - \widetilde{\pi}_k\|_2^2|\widetilde{\pi}_k] \\ &\geq \mathbb{E}_{\widetilde{\pi}_k}\langle\nabla\phi(\widetilde{\pi}_k), \mathbb{E}[\pi_k|\widetilde{\pi}_k] - \widetilde{\pi}_k\rangle + \frac{\alpha}{2}\mathbb{E}\|\pi_k - \widetilde{\pi}_k\|_2^2 \\ &\geq \frac{\alpha}{2}\mathbb{E}\|\pi_k - \widetilde{\pi}_k\|_2^2,\end{aligned} \tag{11.32}$$

where the second step is by α-strongly convexity, and the third step is because $\widetilde{\pi}_k \leq_m \mathbb{E}[\pi_k|\widetilde{\pi}_k]$ and ϕ is convex (thus Schur-convex). Taking the limit $k \to \infty$, we see that the Wasserstein W_2 distance between the π-distributions of M_k and $\widetilde{M}_k$ goes to 0. Because $\lim_{k\to\infty} M_k$ converges weakly to a limit M_∞, $\lim_{k\to\infty} \widetilde{M}_k$ also converges to the same limit. □

Proposition 11.8. *Assume that $\phi : \mathcal{P}([q]) \to \mathbb{R}$ is α-strongly convex for some $\alpha > 0$, and that*

$$\lim_{k\to\infty}(\Phi(M_k) - \Phi(\widetilde{M}_k)) = 0. \tag{11.33}$$

whenever $M_0 = \mathrm{Id}$ and (1) W is non-trivial and $\widetilde{M}_0 = 0$, or (2) $W = 0$ and $\widetilde{M}_0$ is non-trivial. Then boundary irrelevance and stability of BP fixed point hold.

Proof. **Boundary irrelevance:** Let $\widetilde{M}_0 = 0$, $M_0 = \mathrm{Id}$. By Prop. 11.7, we have

$$\lim_{k\to\infty} M_k = \lim_{k\to\infty} \widetilde{M}_k. \tag{11.34}$$

In particular,

$$\lim_{k\to\infty} C(M_k) = \lim_{k\to\infty} C(\widetilde{M}_k) \tag{11.35}$$

where C denote capacity (Definition 3.4). Note that

$$\lim_{k\to\infty} C(M_k) = \lim_{k\to\infty} I(\sigma_\rho; \sigma_{L_k}, \omega_{T_k}|T_k), \tag{11.36}$$

$$\lim_{k\to\infty} C(\widetilde{M}_k) = \lim_{k\to\infty} I(\sigma_\rho; \omega_{T_k}|T_k). \tag{11.37}$$

So this proves boundary irrelevance.

Stability of BP fixed point: Suppose there is a non-trivial fixed point FMS channel U. Let $\widetilde{M}_0 = U$, $M_0 = \mathrm{Id}$. By Prop. 11.7, we have

$$\lim_{k\to\infty} M_k = \lim_{k\to\infty} \widetilde{M}_k. \tag{11.38}$$

Because U is a fixed point, LHS is equal to U. On the other hand, RHS does not depend on U. Therefore there is a unique non-trivial FMS fixed point. □

11.3.2 Low SNR

For the low SNR case, we use SKL-capacity as the potential function. We define

$$\phi^L(\pi) = C_{\mathrm{SKL}}(\mathrm{FSC}_\pi) = \sum_{i\in[q]} \left(\pi_i - \frac{1}{q}\right) \log \pi_i. \tag{11.39}$$

We state a few properties of the function ϕ^L.

Proposition 11.9. *ϕ^L is 1-strongly convex on $\mathcal{P}([q])$.*

Proof.

$$\nabla^2 \phi^L(\pi) = \mathrm{diag}\left(\pi^{-1} + \frac{1}{q}\pi^{-2}\right) \succeq I. \tag{11.40}$$

□

Lemma 11.10. *$\Phi^L(\cdot) = C_{\mathrm{SKL}}(\cdot)$ is additive under $\star$-convolution.*

Proof. This is a restatement of Eq. (3.10). For completeness, we give a direct proof using Eq. (3.5) here.

By FSC mixture decomposition (Prop. 3.3), it suffices to prove that

$$\Phi^L(\mathrm{FSC}_\pi \star \mathrm{FSC}_{\pi'}) = \Phi^L(\mathrm{FSC}_\pi) + \Phi^L(\mathrm{FSC}_{\pi'}). \tag{11.41}$$

We have

$$
\begin{aligned}
&\Phi^L(\mathrm{FSC}_\pi \star \mathrm{FSC}_{\pi'}) \\
&= \sum_{\tau\in\mathrm{Aut}([q])} \left(\frac{1}{(q-1)!}\sum_{i\in[q]} \pi_i\pi'_{\tau(i)}\right) \phi^L(\pi \star_\tau \pi') \\
&= \sum_{\tau\in\mathrm{Aut}([q])} \left(\frac{1}{(q-1)!}\sum_{i\in[q]} \pi_i\pi'_{\tau(i)}\right) \sum_{j\in[q]} \left((\pi \star_\tau \pi')_j - \frac{1}{q}\right) \log(\pi \star_\tau \pi')_j \\
&= \sum_{\tau\in\mathrm{Aut}([q])} \left(\frac{1}{(q-1)!}\sum_{i\in[q]} \pi_i\pi'_{\tau(i)}\right) \sum_{j\in[q]} \left(\frac{\pi_j\pi'_{\tau(j)}}{\sum_{k\in[q]}\pi_k\pi'_{\tau(k)}} - \frac{1}{q}\right) \log \frac{\pi_j\pi'_{\tau(j)}}{\sum_{k\in[q]}\pi_k\pi'_{\tau(k)}} \\
&= \sum_{\tau\in\mathrm{Aut}([q])} \frac{1}{(q-1)!}\sum_{j\in[q]} \left(\pi_j\pi'_{\tau(j)} - \frac{1}{q}\sum_{i\in[q]}\pi_i\pi'_{\tau(i)}\right) \log \frac{\pi_j\pi'_{\tau(j)}}{\sum_{k\in[q]}\pi_k\pi'_{\tau(k)}} \\
&= \sum_{\tau\in\mathrm{Aut}([q])} \frac{1}{(q-1)!}\sum_{j\in[q]} \left(\pi_j\pi'_{\tau(j)} - \frac{1}{q}\sum_{i\in[q]}\pi_i\pi'_{\tau(i)}\right) \log(\pi_j\pi'_{\tau(j)}) \\
&= \sum_{j\in[q]} \left(\pi_j - \frac{1}{q}\right)\log\pi_j + \sum_{j\in[q]}\left(\pi'_j - \frac{1}{q}\right)\log\pi'_j \\
&= \Phi^L(\mathrm{FSC}_\pi) + \Phi^L(\mathrm{FSC}_{\pi'}).
\end{aligned}
$$

□

Condition (11.22) implies the desired contraction.

Proposition 11.11. *If* (11.22) *holds, then*

$$\lim_{k\to\infty}\left(\Phi^L(M_k) - \Phi^L(\widetilde{M}_k)\right) = 0. \tag{11.42}$$

Proof. Using BP equation and Lemma 11.10, we get

$$\Phi^L(M_{k+1}) = \mathbb{E}_b\left[b\Phi^L(M_k\circ P_\lambda) + \Phi^L(W)\right] = d\Phi^L(M_k\circ P_\lambda) + \Phi^L(W), \tag{11.43}$$

and the same holds with M replaced with $\widetilde{M}$.

To prove that

$$\Phi^L(M_{k+1}) - \Phi^L(\widetilde{M}_{k+1}) \le c\left(\Phi^L(M_k) - \Phi^L(\widetilde{M}_k)\right), \tag{11.44}$$

for some $c < 1$, it suffices to prove that

$$d\Phi^L(\mathrm{FSC}_\pi\circ P_\lambda) - c\Phi^L(\mathrm{FSC}_\pi) = d\phi^L\left(\lambda\pi + \frac{1-\lambda}{q}\right) - c\phi^L(\pi) \tag{11.45}$$

is concave in π.

Let $c = d\lambda^2 C^L(q,\lambda)$. Then for all $v \in \mathbb{1}^\perp \in \mathbb{R}^q$, we have

$$v^\top \nabla^2 \left(d\phi^L \left(\lambda\pi + \frac{1-\lambda}{q} \right) - c\phi^L(\pi) \right) v = d\lambda^2 f^L \left(\lambda\pi + \frac{1-\lambda}{q}, v \right) - cf^L(\pi, v) \le 0 \tag{11.46}$$

where the first step is because $v^\top \nabla^2 \phi^L(\pi) v = f^L(\pi, v)$ and the second step is by definition of C^L. Therefore contraction holds. □

Proof of Theorem 11.5. By combining Prop. 11.11, Prop. 11.9, and Prop. 11.8. □

11.3.3 High SNR

For the high SNR case, we use Bhattacharyya coefficient as the potential function. We define

$$\phi^H(\pi) = Z(\mathrm{FSC}_\pi^R) = \frac{1}{q-1} \left(\left(\sum_{i\in[q]} \sqrt{\pi_i} \right)^2 - 1 \right). \tag{11.47}$$

We state a few properties of the function ϕ^H.

Proposition 11.12. *ϕ^H is α-strongly concave on $\mathcal{P}([q])$ for some $\alpha > 0$.*

Proof. For any $\pi \in \mathcal{P}([q])$ we have

$$\nabla^2 \phi^H(\pi) = \frac{1}{q-1} \left(\frac{1}{2} \left(\pi^{-1/2}\right) \left(\pi^{-1/2}\right)^\top - \frac{1}{2} \left(\sum_{i\in[q]} \pi_i^{1/2} \right) \mathrm{diag}\left(\pi^{-3/2}\right) \right). \tag{11.48}$$

So for any $v \in \mathbb{1}^\perp \subseteq \mathbb{R}^q$,

$$v^\top \nabla^2 \phi^H(\pi) v = \frac{1}{2(q-1)} \left(\left\langle \pi^{-1/2}, v \right\rangle^2 - \left(\sum_{i\in[q]} \pi_i^{1/2} \right) \left\langle \pi^{-3/2}, v^2 \right\rangle \right). \tag{11.49}$$

Let us prove that

$$\frac{\left\langle \pi^{-1/2}, v \right\rangle^2}{\left(\sum_{i\in[q]} \pi_i^{1/2} \right) \left\langle \pi^{-3/2}, v^2 \right\rangle} \le 1 - \frac{1}{\sqrt{q}}. \tag{11.50}$$

Performing change of variable $u = \pi^{-3/4} v$, LHS of (11.50) becomes

$$\frac{\left\langle \pi^{1/4}, u \right\rangle^2}{\left(\sum_{i\in[q]} \pi_i^{1/2} \right) \|u\|_2^2}. \tag{11.51}$$

We would like to maximize this expression over the hyperplane $\left\langle \pi^{3/4}, u \right\rangle = 0$. By geometric interpretation, maximum value is achieved at projection of $\pi^{1/4}$ onto the hyperplane, i.e.,

$$u = \pi^{1/4} - \frac{\left\langle \pi^{1/4}, \pi^{3/4} \right\rangle}{\left\|\pi^{3/4}\right\|_2^2} \pi^{3/4} = \pi^{1/4} - \frac{\pi^{3/4}}{\left\|\pi^{3/4}\right\|_2^2}, \tag{11.52}$$

at which (11.51) achieves value

$$\begin{aligned} & \frac{\left(\sum_{i\in[q]} \pi_i^{1/2} - \frac{1}{\sum_{i\in[q]} \pi_i^{3/2}}\right)^2}{\left(\sum_{i\in[q]} \pi_i^{1/2}\right)\left(\sum_{i\in[q]} \pi_i^{1/2} - \frac{1}{\sum_{i\in[q]} \pi_i^{3/2}}\right)} \\ = & \frac{\sum_{i\in[q]} \pi_i^{1/2} - \frac{1}{\sum_{i\in[q]} \pi_i^{3/2}}}{\sum_{i\in[q]} \pi_i^{1/2}} \\ = & 1 - \frac{1}{\left(\sum_{i\in[q]} \pi_i^{1/2}\right)\left(\sum_{i\in[q]} \pi_i^{3/2}\right)} \\ \le & 1 - \frac{1}{\sqrt{q}} \end{aligned} \tag{11.53}$$

where the last step is because

$$\sum_{i\in[q]} \pi_i^{1/2} \le \sqrt{q}, \qquad \sum_{i\in[q]} \pi_i^{3/2} \le 1. \tag{11.54}$$

This finishes the proof of (11.50).

Therefore

$$\begin{aligned} v^\top \nabla^2 \phi^H(\pi) v &= \frac{1}{2(q-1)} \left(\left\langle \pi^{-1/2}, v \right\rangle^2 - \left(\sum_{i\in[q]} \pi_i^{1/2} \right) \left\langle \pi^{-3/2}, v^2 \right\rangle \right) \\ &\le -\frac{1}{2(q-1)\sqrt{q}} \left(\sum_{i\in[q]} \pi_i^{1/2} \right) \left\langle \pi^{-3/2}, v^2 \right\rangle \\ &\le -\frac{1}{2(q-1)\sqrt{q}} \|v\|_2^2 \end{aligned} \tag{11.55}$$

where the second step is by (11.50), and the third step is because $\sum_{i\in[q]} \pi_i^{1/2} \ge 1$ and $\pi^{-3/2} \ge 1$. □

Lemma 11.13. *$\Phi^H(\cdot) = Z(\cdot)$ is multiplicative under $\star$-convolution.*

Proof. This is a restatement of Eq. (2.20). For completeness, we give a direct proof using Eq. (3.5) here.

By FSC mixture decomposition (Prop. 3.5), it suffices to prove that

$$\Phi^H(\mathrm{FSC}_\pi \star \mathrm{FSC}_{\pi'}) = \Phi^H(\mathrm{FSC}_\pi) + \Phi^H(\mathrm{FSC}_{\pi'}). \tag{11.56}$$

We have

$$\begin{aligned}
&\Phi^H(\mathrm{FSC}_\pi \star \mathrm{FSC}_{\pi'}) \\
&= \sum_{\tau \in \mathrm{Aut}([q])} \left(\frac{1}{(q-1)!} \sum_{i \in [q]} \pi_i \pi'_{\tau(i)} \right) \phi^H(\pi \star_\tau \pi') \\
&= \sum_{\tau \in \mathrm{Aut}([q])} \left(\frac{1}{(q-1)!} \sum_{i \in [q]} \pi_i \pi'_{\tau(i)} \right) \frac{1}{q-1} \left(\left(\sum_{j \in [q]} \sqrt{(\pi \star_\tau \pi')_j} \right)^2 - 1 \right) \\
&= \sum_{\tau \in \mathrm{Aut}([q])} \left(\frac{1}{(q-1)!} \sum_{i \in [q]} \pi_i \pi'_{\tau(i)} \right) \frac{1}{q-1} \left(\left(\sum_{j \in [q]} \sqrt{\frac{\pi_j \pi'_{\tau(j)}}{\sum_{k \in [q]} \pi_k \pi'_{\tau(k)}}} \right)^2 - 1 \right) \\
&= \sum_{\tau \in \mathrm{Aut}([q])} \frac{1}{(q-1)!} \cdot \frac{1}{q-1} \left(\left(\sum_{j \in [q]} \sqrt{\pi_j \pi'_{\tau(j)}} \right)^2 - \sum_{i \in [q]} \pi_i \pi'_{\tau(i)} \right) \\
&= \sum_{\tau \in \mathrm{Aut}([q])} \frac{1}{(q-1)!} \cdot \frac{1}{q-1} \sum_{j \neq k \in [q]} \sqrt{\pi_j \pi'_{\tau(j)} \pi_k \pi'_{\tau(k)}} \\
&= \frac{1}{(q-1)^2} \left(\sum_{j \neq k \in [q]} \sqrt{\pi_j \pi_k} \right) \left(\sum_{j' \neq k' \in [q]} \sqrt{\pi'_{\tau(j')} \pi'_{\tau(k')}} \right) \\
&= \Phi^H(\mathrm{FSC}_\pi) \Phi^H(\mathrm{FSC}_{\pi'}).
\end{aligned}$$

□

Condition (11.23) implies the desired contraction.

Proposition 11.14. *If* (11.23) *or* (11.24) *holds, then*

$$\lim_{k \to \infty} \left(\Phi^H(M_k) - \Phi^H(\widetilde{M}_k) \right) = 0. \tag{11.57}$$

Proof. We treat the regular tree case and the Poisson tree case (almost) uniformly. For simplicity, in this proof, we use the following notation. Let $\mathbb{1}_R$ be 1 if we are working with regular trees, and 0 otherwise. Let $\mathbb{1}_P$ be 1 if we are working with Poisson trees, and 0 otherwise.

Using BP equation and Lemma 11.13, we have

$$\Phi^H(M_{k+1}) = \mathbb{E}_b \left[\left(\Phi^H(M_k \circ P_\lambda) \right)^b \Phi^H(W) \right] \tag{11.58}$$

and the same holds with M replaced with $\widetilde{M}$.

For $i \geq 0$, define

$$\Phi_i = \mathbb{E}_b \left[\prod_{j \in [b]} \left(\left(\Phi^H(\widetilde{M}_k \circ P_\lambda) \right)^{\mathbb{1}\{j \leq i\}} \left(\Phi^H(M_k \circ P_\lambda) \right)^{\mathbb{1}\{j > i\}} \right) \Phi^H(W) \right]. \tag{11.59}$$

Fix $i \geq 1$. Let us prove that

$$\Phi_i - \Phi_{i-1} \leq c_i \left(\Phi(\widetilde{M}_k) - \Phi(M_k) \right) \tag{11.60}$$

for some constant c_i to be determined later.

Note that

$$\begin{aligned} \Phi_i - \Phi_{i-1} = & \left(\Phi^H(\widetilde{M}_k \circ P_\lambda) - \Phi^H(M_k \circ P_\lambda) \right) \\ & \cdot \mathbb{E}_b \left[\mathbb{1}\{b \geq i\} \left(\Phi^H(\widetilde{M}_k \circ P_\lambda) \right)^{i-1} \left(\Phi^H(M_k \circ P_\lambda) \right)^{b-i} \right] \Phi^H(W). \end{aligned} \tag{11.61}$$

Note that $f^H(\pi, v) = -2(q-1)v^\top \nabla^2 \phi^H(\pi) v$. Therefore by definition of $C^H(q, \lambda)$, we have

$$\nabla^2 \left(\Phi^H(\mathrm{FSC}_\pi \circ P_\lambda) - \lambda^2 C^H(q, \lambda) \Phi^H(\mathrm{FSC}_\pi) \right) \succeq 0. \tag{11.62}$$

So by degradation,

$$\Phi^H(\widetilde{M}_k \circ P_\lambda) - \Phi^H(M_k \circ P_\lambda) \leq \lambda^2 C^H(q, \lambda) \left(\Phi^H(\widetilde{M}_k) - \Phi^H(M_k) \right). \tag{11.63}$$

By Prop. 11.16 and Lemma 11.15, for any $\epsilon > 0$, for k large enough, we have

$$\Phi^H(M_k \circ P_\lambda) \leq \left(1 - c^H(q, \lambda, d) \cdot \frac{d\lambda^2 - 1}{d - \mathbb{1}_R} + \epsilon \right)_+^{1/2}. \tag{11.64}$$

Let

$$c_i = \lambda^2 \Phi^H(W) C^H(q, \lambda) \mathbb{E}_b \left[\mathbb{1}\{b \geq i\} \left(1 - c^H(q, \lambda, d) \cdot \frac{d\lambda^2 - 1}{d - \mathbb{1}_R} + \epsilon \right)_+^{\frac{b-1}{2}} \right]. \tag{11.65}$$

Combining (11.61)(11.63)(11.64)(11.65), we get

$$\Phi_i - \Phi_{i-1} \leq c_i \left(\Phi^H(\widetilde{M}_k) - \Phi^H(M_k) \right). \tag{11.66}$$

Let us compute sum of c_i. In the regular tree case, we have

$$\sum_{i \geq 1} c_i = d\lambda^2 \Phi^H(W) C^H(q, \lambda) \left(1 - c^H(q, \lambda, d) \cdot \frac{d\lambda^2 - 1}{d - 1} + \epsilon \right)_+^{\frac{d-1}{2}}. \tag{11.67}$$

For the Poisson tree case, we have

$$\sum_{i\geq 1} c_i = d\lambda^2 \Phi^H(W) C^H(q,\lambda) \exp\left(-d\left(1-\left(1-c^H(q,\lambda,d)\cdot\frac{d\lambda^2-1}{d}+\epsilon\right)_+^{1/2}\right)\right). \tag{11.68}$$

Note that $\epsilon > 0$ can be chosen to be arbitrarily small. Therefore in both cases, for any $\epsilon' > 0$, for k large enough, we can choose c_i such that (11.66) holds and

$$\sum_{i\geq 1} c_i \leq d\lambda^2 \Phi^H(W) C^H(q,\lambda) \exp\left(-c^H(q,\lambda,d)\cdot\frac{d\lambda^2-1}{2}+\epsilon'\right). \tag{11.69}$$

Then

$$\begin{aligned}
&\Phi^H(\widetilde{M}_{k+1}) - \Phi^H(M_{k+1}) \qquad (11.70)\\
&\leq \left(\sum_{i\geq 1} c_i\right)\left(\Phi^H(\widetilde{M}_k) - \Phi^H(M_k)\right)\\
&\leq d\lambda^2 \Phi^H(W) C^H(q,\lambda) \exp\left(-c^H(q,\lambda,d)\cdot\frac{d\lambda^2-1}{2}+\epsilon'\right)\left(\Phi^H(\widetilde{M}_k) - \Phi^H(M_k)\right).
\end{aligned}$$

Because (11.23) or (11.24) holds, we can choose $\epsilon' > 0$ small enough so that

$$d\lambda^2 \Phi^H(W) C^H(q,\lambda) \exp\left(-c^H(q,\lambda,d)\cdot\frac{d\lambda^2-1}{2}+\epsilon'\right) < 1. \tag{11.71}$$

This leads to the desired contraction. □

Lemma 11.15. *For any BMS channel P, we have*

$$Z(P) \leq \sqrt{1 - C_{\chi^2}(P)}. \tag{11.72}$$

Proof. Let Δ be the Δ-component of P. Then

$$Z(P) = \mathbb{E}[2\sqrt{\Delta(1-\Delta)}] \leq \sqrt{1-\mathbb{E}[(1-2\Delta)^2]} = \sqrt{1-C_{\chi^2}(P)}. \tag{11.73}$$

The inequality step is by concavity of $\sqrt{\cdot}$. □

Proof of Theorem 11.6. Combine Prop. 11.14, 11.12, and 11.8. □

11.3.4 Majority decider

Proposition 11.16. *Consider the Potts model* $\mathrm{BOT}(q,\lambda,d)$ *or* $\mathrm{BOT}(q,\lambda,\mathrm{Pois}(d))$ *with leaf observations through a non-trivial FMS channel U. Let M_k^U denote the*

channel $\sigma_\rho \to \nu_{L_k}$ *where* $\nu_v \sim U(\cdot|\sigma_v)$. *Assume that* $d\lambda^2 > 1$. *Then*

$$\lim_{k\to\infty} C_{\chi^2}\left((M_k^U \circ P_\lambda)^R\right) \geq \begin{cases} c(q,\lambda,d)\cdot\frac{d\lambda^2-1}{d-1} & \textit{Regular tree case,} \\ c(q,\lambda,d)\cdot\frac{d\lambda^2-1}{d} & \textit{Poisson tree case.} \end{cases} \tag{11.74}$$

where

$$c(q,\lambda,d) := \left(\frac{2}{q} + \frac{q-2}{q}\cdot\frac{d\lambda^2-1}{d\lambda-1}\right)^{-1}. \tag{11.75}$$

Furthermore, $c(q,\lambda,d) \geq 1$ *for all* $\lambda \in \left[-\frac{1}{q-1}, 1\right]$ *and* $d\lambda^2 > 1$.

Proof. Let U^* be the reverse channel of U. Then the composition $U^* \circ U$ is a non-trivial ferromagnetic Potts channel. So there exists $\eta > 0$ such that $P_\eta \leq_{\deg} U$. By replacing U with P_η (and using Eq. (3.8)), we can wlog assume that $U = P_\eta$ for some $\eta > 0$.

Fix any embedding $\{\pm\} \subseteq [q]$. Let $e \in \mathbb{R}^q$ denote the vector with $e_+ = 1, e_- = -1, e_i = 0$ for $i \notin \{\pm\}$. Let

$$S_k = \sum_{v\in L_k} e_{\nu_v}. \tag{11.76}$$

We view S_k as a channel $[q] \to \mathbb{Z}$. By variational characterization of χ^2-divergence, we have

$$C_{\chi^2}((M_k^U \circ P_\lambda)^R) \geq \frac{\left(\mathbb{E}^+[S_k \circ P_\lambda]\right)^2}{\mathbb{E}^+[S_k^2 \circ P_\lambda]} \tag{11.77}$$

where $\mathbb{E}^+$ denotes expectation conditioned on root label being $+$. Similarly, we use $\mathbb{E}^-$ to denote expectation conditioned on root label being $-$, and use $\mathbb{E}^0$ to denote expectation conditioned on root label being any label not $\pm$. Same for Var^+, Var^-, Var^0.

For simplicity, in this proof, we use the following notation. Let $\mathbb{1}_R$ be 1 if we are working with regular trees, and 0 otherwise. Let $\mathbb{1}_P$ be 1 if we are working with Poisson trees, and 0 otherwise. Clearly $\mathbb{1}_P + \mathbb{1}_R = 1$.

It is easy to see that

$$\mathbb{E}^i S_k = e_i \eta (d\lambda)^k. \tag{11.78}$$

Using variance decomposition formula, we have

$$\begin{aligned} \mathrm{Var}^i(S_{k+1}) = \mathrm{Var}^i(\mathbb{E}[S_{k+1}|b]) + \mathbb{E}_b \mathrm{Var}^i(\mathbb{E}[S_{k+1}|b,\sigma_1,\ldots,\sigma_b]) \\ + \mathbb{E}\,\mathrm{Var}^i(S_{k+1}|b,\sigma_1,\ldots,\sigma_b) \end{aligned} \tag{11.79}$$

where $\sigma_1,\ldots,\sigma_b$ are labels of the children.

Let us compute each summand.

$$\mathrm{Var}^i(\mathbb{E}[S_{k+1}|b]) = \mathrm{Var}^i(b\lambda e_i \eta (d\lambda)^k) = e_i^2 d\lambda^2 \eta^2 (d\lambda)^{2k} \mathbb{1}_P, \tag{11.80}$$

$$\mathbb{E}_b \mathrm{Var}^i(\mathbb{E}[S_{k+1}|b, \sigma_1, \dots, \sigma_b]) = d\eta^2 (d\lambda)^{2k} \mathrm{Var}_{j \sim P_\lambda(\cdot|i)}(e_j), \tag{11.81}$$

$$\mathbb{E}\, \mathrm{Var}^i(S_{k+1}|b, \sigma_1, \dots, \sigma_b) = d\mathbb{E}_{j \sim P_\lambda(\cdot|i)}[\mathrm{Var}^j(S_k)]. \tag{11.82}$$

We have $\mathrm{Var}^-(S_k) = \mathrm{Var}^+(S_k)$ and

$$\begin{aligned} \mathrm{Var}^+(S_{k+1}) = {} & d\eta^2 (d\lambda)^{2k} \left(\left(\lambda + \frac{1-\lambda}{q} \cdot 2 \right) - \lambda^2 \mathbb{1}_R \right) \\ & + d \left(\left(\lambda + \frac{1-\lambda}{q} \cdot 2 \right) \mathrm{Var}^+(S_k) + \frac{1-\lambda}{q} \cdot (q-2)\, \mathrm{Var}^0(S_k) \right), \end{aligned} \tag{11.83}$$

$$\begin{aligned} \mathrm{Var}^0(S_{k+1}) = {} & d\eta^2 (d\lambda)^{2k} \left(\frac{1-\lambda}{q} \cdot 2 \right) \\ & + d \left(\frac{1-\lambda}{q} \cdot 2\, \mathrm{Var}^+(S_k) + \left(\lambda + \frac{1-\lambda}{q} \cdot (q-2) \right) \mathrm{Var}^0(S_k) \right). \end{aligned} \tag{11.84}$$

By computing linear combinations of (11.83)(11.84), we get

$$\begin{aligned} \mathrm{Var}^+(S_{k+1}) - \mathrm{Var}^0(S_{k+1}) = {} & d\lambda \left(\mathrm{Var}^+(S_k) - \mathrm{Var}^0(S_k) \right) \\ & + d\eta^2 (d\lambda)^{2k} (\lambda - \lambda^2 \mathbb{1}_R). \end{aligned} \tag{11.85}$$

$$\begin{aligned} \mathrm{Var}^+(S_{k+1}) + \frac{q-2}{2} \mathrm{Var}^0(S_{k+1}) = {} & d \left(\mathrm{Var}^+(S_k) + \frac{q-2}{2} \mathrm{Var}^0(S_{k+1}) \right) \\ & + d\eta^2 (d\lambda)^{2k} (1 - \lambda^2 \mathbb{1}_R). \end{aligned} \tag{11.86}$$

Solving (11.85)(11.86) we get

$$\begin{aligned} & \mathrm{Var}^+(S_k) - \mathrm{Var}^0(S_k) \\ = {} & \left(\mathrm{Var}^+(S_0) - \mathrm{Var}^0(S_0) \right) (d\lambda)^k + \sum_{1 \le i \le k} d\eta^2 (d\lambda)^{2i-2} (d\lambda)^{k-i} (\lambda - \lambda^2 \mathbb{1}_R) \\ = {} & O\left((d\lambda)^k \right) + d\eta^2 (d\lambda)^{k-1} \frac{(d\lambda)^k - 1}{d\lambda - 1} (\lambda - \lambda^2 \mathbb{1}_R) \\ = {} & (1 + o(1)) \frac{1 - \lambda \mathbb{1}_R}{d\lambda - 1} \eta^2 (d\lambda)^{2k} \end{aligned} \tag{11.87}$$

and

$$\begin{aligned}
&\mathrm{Var}^+(S_k) + \frac{q-2}{2}\,\mathrm{Var}^0(S_k) \qquad (11.88)\\
&= \left(\mathrm{Var}^+(S_0) + \frac{q-2}{2}\,\mathrm{Var}^0(S_0)\right) d^k + \sum_{1\le i\le k} d\eta^2 (d\lambda)^{2i-2} d^{k-i}(1-\lambda^2 \mathbb{1}_R)\\
&= O\left(d^k\right) + \eta^2 d^k \frac{(d\lambda^2)^k - 1}{d\lambda^2 - 1}(1-\lambda^2\mathbb{1}_R)\\
&= (1+o(1))\frac{1-\lambda^2\mathbb{1}_R}{d\lambda^2-1}\eta^2(d\lambda)^{2k}.
\end{aligned}$$

Combining (11.87)(11.88) we have

$$\mathrm{Var}^+(S_k) = (1+o(1))\frac{2}{q}\left(\frac{1-\lambda^2\mathbb{1}_R}{d\lambda^2-1} + \frac{1-\lambda\mathbb{1}_R}{d\lambda-1}\cdot\frac{q-2}{2}\right)\eta^2(d\lambda)^{2k}. \qquad (11.89)$$

Now we compute moments of $S_k \circ P_\lambda$.

$$\mathbb{E}^+[S_k \circ P_\lambda] = \lambda\eta(d\lambda)^k. \qquad (11.90)$$

$$\begin{aligned}
\mathbb{E}^+[S_k^2 \circ P_\lambda] &= \left(\lambda + \frac{1-\lambda}{q}\cdot 2\right)\mathbb{E}^+[S_k^2] + \frac{1-\lambda}{q}\cdot(q-2)\mathbb{E}^0[S_k^2] \qquad (11.91)\\
&= \left(\lambda + \frac{1-\lambda}{q}\cdot 2\right)\left(\mathrm{Var}^+(S_k) + (\mathbb{E}^+ S_k)^2\right) + \frac{1-\lambda}{q}\cdot(q-2)\,\mathrm{Var}^0(S_k)\\
&= \lambda(1+o(1))\frac{2}{q}\left(\frac{1-\lambda^2\mathbb{1}_R}{d\lambda^2-1} + \frac{1-\lambda\mathbb{1}_R}{d\lambda-1}\cdot\frac{q-2}{2}\right)\eta^2(d\lambda)^{2k}\\
&+ \left(\lambda + \frac{1-\lambda}{q}\cdot 2\right)\eta^2(d\lambda)^{2k} + \frac{1-\lambda}{q}\cdot 2\cdot(1+o(1))\frac{1-\lambda^2\mathbb{1}_R}{d\lambda^2-1}\eta^2(d\lambda)^{2k}\\
&= (1+o(1))\left(\frac{2}{q}\cdot\frac{d\lambda^2-\lambda^2\mathbb{1}_R}{d\lambda^2-1} + \lambda\cdot\frac{q-2}{q}\cdot\frac{d\lambda-\lambda\mathbb{1}_R}{d\lambda-1}\right)\eta^2(d\lambda)^{2k}\\
&= (1+o(1))c(q,\lambda,d)^{-1}\frac{d-\mathbb{1}_R}{d\lambda^2-1}\lambda^2\eta^2(d\lambda)^{2k}.
\end{aligned}$$

Finally,

$$C_{\chi^2}((M_k^U \circ P_\lambda)^R) \ge \frac{\left(\mathbb{E}^+[S_k \circ P_\lambda]\right)^2}{\mathbb{E}^+[S_k^2 \circ P_\lambda]} = (1+o(1))c(q,\lambda,d)\cdot\frac{d\lambda^2-1}{d-\mathbb{1}_R}. \qquad (11.92)$$

□

11.3.5 Bounds on key constants

In this section we prove bounds on key constants used in Theorem 11.5 and 11.6.

Proposition 11.17. *For $q \in \mathbb{Z}_{\geq 2}$, $\lambda \in \left[-\frac{1}{q-1}, 1\right]$, we have $C^L(q,\lambda) \leq q^2$, where $C^L(q,\lambda)$ is defined in* (11.17).

Proof. We have

$$\begin{aligned} & \frac{\left\langle \left(\lambda\pi + \frac{1-\lambda}{q}\right)^{-1} + \frac{1}{q}\left(\lambda\pi + \frac{1-\lambda}{q}\right)^{-2}, v^2 \right\rangle}{\left\langle \pi^{-1} + \frac{1}{q}\pi^{-2}, v^2 \right\rangle} \\ \leq & \max\left\{ \frac{\left\langle \left(\lambda\pi + \frac{1-\lambda}{q}\right)^{-1}, v^2 \right\rangle}{\langle \pi^{-1}, v^2\rangle}, \frac{\left\langle \left(\lambda\pi + \frac{1-\lambda}{q}\right)^{-2}, v^2 \right\rangle}{\langle \pi^{-2}, v^2\rangle} \right\} \\ \leq & \max\{q, q^2\} = q^2. \end{aligned} \tag{11.93}$$

where the second step is by Lemma 11.19. □

Proposition 11.18. *For $q \in \mathbb{Z}_{\geq 2}$, $\lambda \in \left[-\frac{1}{q-1}, 1\right]$, we have $C^H(q,\lambda) \leq q^{5/2}$, where $C^H(q,\lambda)$ is defined in* (11.19).

Proof. By (11.50),

$$f(\pi, v) \geq \frac{1}{\sqrt{q}} \left(\sum_{i\in[q]} \pi_i^{1/2} \right) \left\langle \pi^{-3/2}, v^2 \right\rangle \geq \frac{1}{\sqrt{q}} \left\langle \pi^{-3/2}, v^2 \right\rangle. \tag{11.94}$$

On the other hand,

$$\begin{aligned} f\left(\lambda\pi + \frac{1-\lambda}{q}, v\right) &\leq \left(\sum_{i\in[q]} \left(\lambda\pi_i + \frac{1-\lambda}{q}\right)^{1/2} \right) \left\langle \left(\lambda\pi + \frac{1-\lambda}{q}\right)^{-3/2}, v^2 \right\rangle \\ &\leq \sqrt{q} \cdot \left\langle \left(\lambda\pi + \frac{1-\lambda}{q}\right)^{-3/2}, v^2 \right\rangle. \end{aligned} \tag{11.95}$$

Combining (11.94)(11.95) we get

$$\frac{f\left(\lambda\pi + \frac{1-\lambda}{q}, v\right)}{f(\pi, v)} \leq q \cdot \frac{\left\langle \left(\lambda\pi + \frac{1-\lambda}{q}\right)^{-3/2}, v^2 \right\rangle}{\langle \pi^{-3/2}, v^2\rangle} \leq q^{5/2} \tag{11.96}$$

where the last step is by Lemma 11.19. □

The following lemma is the crucial step in the proof of Prop. 11.17 and 11.18.

Lemma 11.19. *For $q \in \mathbb{Z}_{\geq 2}$, $\alpha \in \mathbb{R}_{\geq 1}$, $\lambda \in \left[-\frac{1}{q-1}, 1\right]$, we have*

$$\sup_{\substack{\pi \in \mathcal{P}([q]) \\ v \in \mathbb{1}^\perp \subseteq \mathbb{R}^q}} \frac{\left\langle \left(\lambda\pi + \frac{1-\lambda}{q}\right)^{-\alpha}, v^2 \right\rangle}{\left\langle \pi^{-\alpha}, v^2 \right\rangle} \leq q^\alpha. \tag{11.97}$$

Proof. We prove the ferromagnetic case ($\lambda \in [0,1]$) and antiferromagnetic case ($\lambda \in \left[-\frac{1}{q-1}, 0\right]$) separately.

Ferromagnetic case ($\lambda \in [0,1]$)**.** In this case, we have

$$\lambda x + \frac{1-\lambda}{q} \geq \frac{x}{q} \tag{11.98}$$

for all $x \in [0,1]$. Therefore

$$\left\langle \left(\lambda\pi + \frac{1-\lambda}{q}\right)^{-\alpha}, v^2 \right\rangle \leq \left\langle \left(\frac{\pi}{q}\right)^{-\alpha}, v^2 \right\rangle = q^\alpha \left\langle \pi^{-\alpha}, v^2 \right\rangle. \tag{11.99}$$

Note that we did not use the assumption that $v \in \mathbb{1}^\perp$.

Antiferromagnetic case $\left(\lambda \in \left[-\frac{1}{q-1}, 0\right]\right)$**.** We would like to prove that

$$q^\alpha \left\langle \pi^{-\alpha}, v^2 \right\rangle - \left\langle \left(\lambda\pi + \frac{1-\lambda}{q}\right)^{-\alpha}, v^2 \right\rangle =: \left\langle b, v^2 \right\rangle \tag{11.100}$$

is non-negative for all $\pi \in \mathcal{P}([q])$, $v \in \mathbb{1}^\perp$, where

$$b := \left(\frac{\pi}{q}\right)^{-\alpha} - \left(\lambda\pi + \frac{1-\lambda}{q}\right)^{-\alpha}. \tag{11.101}$$

Step 1. We fix $\pi \in \mathcal{P}([q])$ and determine the optimal $v \in \mathbb{1}^\perp$ to plug in (11.100), reducing the statement to one involving π only.

If $\lambda x + \frac{1-\lambda}{q} \leq \frac{x}{q}$ for some $x \in [0,1]$, then $x \geq \frac{1-\lambda}{1-\lambda q} \geq \frac{q}{2q-1} > \frac{1}{2}$. So there exists at most one i such that $\lambda\pi_i + \frac{1-\lambda}{q} \leq \frac{\pi_i}{q}$ (equivalently, $b_i \leq 0$). We can wlog assume that $\pi_1 \geq \pi_2 \geq \cdots \geq \pi_q$. Then we know $\lambda\pi_i + \frac{1-\lambda}{q} > \frac{\pi_i}{q}$ (equivalently, $b_i > 0$) for all $2 \leq i \leq q$. If $b_1 \geq 0$, then $\langle b, v^2 \rangle$ is non-negative for all v and we are done. Therefore, it remains to consider the case $b_1 < 0$.

If $v_1 = 0$, then $\langle b, v^2 \rangle$ is non-negative. Therefore we can assume $v_1 \neq 0$. By rescaling, we can assume that $v_1 = 1$. So $v_2 + \cdots + v_q = -1$. Because $b_2, \ldots, b_q$ are all positive, to minimize $\sum_{2 \leq i \leq q} b_i v_i^2$ under linear constraint $v_2 + \cdots + v_q = -1$, the optimal choice is $v_i = -b_i^{-1} Z^{-1}$ for $2 \leq i \leq q$ where $Z := \sum_{2 \leq i \leq q} b_i^{-1}$. For this choice

of v, we have

$$\left\langle b, v^2 \right\rangle = b_1 + \sum_{2 \le i \le q} b_i \cdot (-b_i^{-1} Z^{-1})^2 = b_1 + Z^{-1}. \tag{11.102}$$

Therefore, it remains to prove

$$Z \le (-b_1)^{-1} \tag{11.103}$$

where $\pi_1 \ge \cdots \ge \pi_q$, $b_1 < 0$, and $b_2, \ldots, b_q > 0$.

Step 2. We reduce to the case where $\pi_3 = \cdots = \pi_q = 0$. Note that

$$Z = \sum_{2 \le i \le q} b_i^{-1} = \sum_{2 \le i \le q} \left(\left(\frac{\pi_i}{q} \right)^{-\alpha} - \left(\lambda \pi_i + \frac{1-\lambda}{q} \right)^{-\alpha} \right)^{-1}. \tag{11.104}$$

By Lemma 11.20, for fixed π_1, the optimal choice (for maximizing Z) of $\pi_2, \ldots, \pi_q$ is $\pi_3 = \cdots = \pi_q = 0$.

Write $\pi_1 = 1 - x$, $\pi_2 = x$ where $x \in \left[0, \frac{\lambda - \lambda q}{1 - \lambda q}\right]$. Then

$$b_1 = \left(\frac{1-x}{q} \right)^{-\alpha} - \left(\lambda(1-x) + \frac{1-\lambda}{q} \right)^{-\alpha}, \tag{11.105}$$

$$Z = \left(\left(\frac{x}{q} \right)^{-\alpha} - \left(\lambda x + \frac{1-\lambda}{q} \right)^{-\alpha} \right)^{-1}. \tag{11.106}$$

By rearranging terms in (11.103), we reduce to proving

$$\left(\frac{x}{q} \right)^{-\alpha} + \left(\frac{1-x}{q} \right)^{-\alpha} \ge \left(\lambda x + \frac{1-\lambda}{q} \right)^{-\alpha} + \left(\lambda(1-x) + \frac{1-\lambda}{q} \right)^{-\alpha} \tag{11.107}$$

for $q \in \mathbb{Z}_{\ge 2}$, $\alpha \in \mathbb{R}_{\ge 1}$, $\lambda \in \left[-\frac{1}{q-1}, 0 \right]$, $x \in \left[0, \frac{\lambda - \lambda q}{1 - \lambda q}\right]$.

Step 3. Let $g_{q,\alpha,x}(\lambda) := \left(\lambda x + \frac{1-\lambda}{q} \right)^{-\alpha} + \left(\lambda(1-x) + \frac{1-\lambda}{q} \right)^{-\alpha}$ be the RHS of (11.107). Then

$$\begin{aligned} g''_{q,\alpha,x}(\lambda) &= \alpha(\alpha+1) \left(x - \frac{1}{q} \right)^2 \left(\lambda x + \frac{1-\lambda}{q} \right)^{-\alpha-2} \\ &\quad + \alpha(\alpha+1) \left(1 - x - \frac{1}{q} \right)^2 \left(\lambda(1-x) + \frac{1-\lambda}{q} \right)^{-\alpha-2} \\ &> 0. \end{aligned}$$

So $g_{q,\alpha,x}$ is convex in λ. Therefore it suffices to verify (11.107) for $\lambda = 0$ and $\lambda = -\frac{1}{q-1}$.

When $\lambda = 0$, we have

$$g_{q,\alpha,x}(\lambda) = \left(\frac{1}{q}\right)^{-\alpha} + \left(\frac{1}{q}\right)^{-\alpha} \leq \left(\frac{x}{q}\right)^{-\alpha} + \left(\frac{1-x}{q}\right)^{-\alpha}. \tag{11.108}$$

When $\lambda = -\frac{1}{q-1}$, we have

$$g_{q,\alpha,x}(\lambda) = \left(\frac{1-x}{q-1}\right)^{-\alpha} + \left(\frac{x}{q-1}\right)^{-\alpha} \leq \left(\frac{x}{q}\right)^{-\alpha} + \left(\frac{1-x}{q}\right)^{-\alpha}. \tag{11.109}$$

This finishes the proof. □

Lemma 11.20. *For $q \in \mathbb{Z}_{\geq 2}$, $\alpha \in \mathbb{R}_{\geq 1}$, $\lambda \in \left[-\frac{1}{q-1}, 0\right]$, the function*

$$f(x) := \left(\left(\frac{x}{q}\right)^{-\alpha} - \left(\lambda x + \frac{1-\lambda}{q}\right)^{-\alpha}\right)^{-1}. \tag{11.110}$$

is convex in $x \in [0, \frac{1-\lambda}{1-\lambda q}]$.

Proof. Let $g(x) = \frac{1}{f(x)}$. Then

$$f''(x) = \frac{2g'(x)^2 - g(x)g''(x)}{g(x)^3}. \tag{11.111}$$

It suffices to prove that

$$2g'(x)^2 - g(x)g''(x) \geq 0. \tag{11.112}$$

We have

$$\begin{aligned} 2g'(x)^2 - g(x)g''(x) = {} & 2\alpha^2 \left(q^{-1}\left(\frac{x}{q}\right)^{-\alpha-1} - \lambda\left(\lambda x + \frac{1-\lambda}{q}\right)^{-\alpha-1}\right)^2 \\ & - \left(\left(\frac{x}{q}\right)^{-\alpha} - \left(\lambda x + \frac{1-\lambda}{q}\right)^{-\alpha}\right) \\ & \cdot \alpha(\alpha+1)\left(q^{-2}\left(\frac{x}{q}\right)^{-\alpha-2} - \lambda^2\left(\lambda x + \frac{1-\lambda}{q}\right)^{-\alpha-2}\right). \end{aligned} \tag{11.113}$$

Write $u = \frac{x}{q}$, $v = \lambda x + \frac{1-\lambda}{q}$, $c = q\lambda$. Then we have $0 \leq u \leq v \leq 1$ and $-\frac{q}{q-1} \leq c \leq 0$. It suffices to prove

$$2\alpha^2(u^{-\alpha-1} - cv^{-\alpha-1})^2 - \alpha(\alpha+1)(u^{-\alpha} - v^{-\alpha})(u^{-\alpha-2} - c^2v^{-\alpha-2}) \geq 0. \tag{11.114}$$

We have

$$\begin{aligned}
&2\alpha^2(u^{-\alpha-1}-cv^{-\alpha-1})^2-\alpha(\alpha+1)(u^{-\alpha}-v^{-\alpha})(u^{-\alpha-2}-c^2v^{-\alpha-2}) \qquad (11.115)\\
&\geq \alpha(\alpha+1)((u^{-\alpha-1}-cv^{-\alpha-1})^2-(u^{-\alpha}-v^{-\alpha})(u^{-\alpha-2}-c^2v^{-\alpha-2}))\\
&=\alpha(\alpha+1)u^{-\alpha}v^{-\alpha}(u^{-1}-cv^{-1})^2\\
&\geq 0.
\end{aligned}$$

This finishes the proof. □

11.4 Applications

Main applications of uniqueness of BP fixed point and boundary irrelevance include a mutual information formula and an optimal recovery algorithm. In this section we prove these results via reduction established in Chapter 5.

Theorem 11.21 (Mutual information formula)**.** *Let* $(X,G)\sim \mathrm{SBM}(n,q,a,b)$*. Let* $(T,\sigma)\sim \mathrm{BOT}(q,\lambda,\mathrm{Pois}(d))$ *be the corresponding Potts model (Theorem 5.10), where* $d=\frac{a+(q-1)b}{q}$ *and* $\lambda=\frac{a-b}{a+(q-1)b}$*. Let* ρ *be the root of* T*,* L_k *be the set of vertices at distance* k *to* ρ*,* T_k *be the set of vertices at distance* $\leq k$ *to* ρ*. If* (q,λ,d) *satisfies Eq.* (11.22) *or Eq.* (11.23)*, then we have*

$$\lim_{n\to\infty}\frac{1}{n}I(X;G)=\int_0^1\lim_{k\to\infty}I(\sigma_\rho;\omega^\epsilon_{T_k\setminus\rho}|T_k)d\epsilon, \qquad (11.116)$$

where ω^ϵ *denotes observation through survey channel* EC_ϵ*.*

Proof. By Theorem 11.5, Theorem 11.6 and Theorem 5.18. □

Theorem 11.22 (Optimal recovery for SBM with survey)**.** *Work under the same setting as Theorem 11.21. Suppose that in addition to* G*, we observe survey* $Y_v\sim W(\cdot|X_v)$ *for all* $v\in V$*, where* W *is some non-trivial FMS channel. If* (q,λ,d,W) *satisfies Eq.* (11.22) *or Eq.* (11.24)*, then belief propagation (Algorithm 1) achieves the optimal recovery accuracy of*

$$1-\lim_{k\to\infty}P_e(\sigma_\rho|T_k,\omega_{T_k}). \qquad (11.117)$$

Proof. By Theorem 11.5, Theorem 11.6 and Theorem 5.20. □

Theorem 11.23 (Optimal recovery for SBM)**.** *Work under the same setting as Theorem 11.21. Suppose* $d\lambda^2>1$ *and* (q,λ,d) *satisfies Eq.* (11.23)*. Suppose there is an algorithm* $\mathcal{A}$ *and a constant* $\epsilon>0$ *(not depending on* n*) such that with probability* $1-o(1)$*, the empirical transition matrix* $F\in\mathbb{R}^{q\times q}$ *defined as*

$$F_{i,j}:=\frac{\#\{v\in V:X_v=i,\widehat{X}_v=j\}}{\#\{v\in V:X_v=i\}},\qquad \widehat{X}:=\mathcal{A}(G) \qquad (11.118)$$

satisfies

(1) $\|F^\top \mathbb{1} - \mathbb{1}\|_\infty = o(1)$*;*

(2) $\sigma_{\min}(F) > \epsilon$*, where* $\sigma_{\min}$ *is the smallest singular value;*

(3) there exists a permutation $\tau \in \mathrm{Aut}([q])$ *such that* $F_{\tau(i),i} > F_{\tau(i),j} + \epsilon$ *for all* $i \neq j \in [q]$*.*

(Note that we do not asssume F *stays the same for different calls to* $\mathcal{A}$*.)*

Then there is an algorithm (Algorithm 2) achieving the optimal recovery accuracy of

$$1 - \lim_{k\to\infty} P_e(\sigma_\rho | T_k, \sigma_{L_k}). \tag{11.119}$$

Proof. By Theorem 11.6 and Theorem 5.22. □

11.5 Asymmetric fixed points

Up to now we have focused on symmetric fixed points of the BP operator. If we view the BP operator as an operator from the space of q-ary input (possibly asymmetric) channels, then a natural question to determine the (possibly asymmetric) fixed points. In the case $q = 2$, [137] showed that there is only one non-trivial fixed point, and the fixed point is symmetric. For $q \geq 3$, it is no longer the case.

Proposition 11.24. *Work under the setting of Theorem 11.1. If* $q \geq 3$ *and* $d\lambda^2 > 1$*, then the* BP *operator (Eq.* (11.8)*) at least one non-trivial asymmetric fixed point.*

Proof. Consider the channel $U : [q] \to \{\pm\}$, which maps 1 to $+$ and $2, \ldots, q$ to $-$. Because $\mathrm{BP}(U) \leq_{\deg} U$, the sequence $(\mathrm{BP}^k(U))_{k\geq 0}$ is non-increasing in degradation preorder. Therefore a limit channel $\mathrm{BP}^\infty(U)$ exists by Lemma 11.2.

We would like to show that $\mathrm{BP}^\infty(U)$ is a non-FMS non-trivial fixed point. When $d\lambda^2 > 1$, count-reconstruction is possible (see e.g. [101]). So it is possible to gain non-trivial information about whether the input is 1 by counting the number of $+$. So $\mathrm{BP}^\infty(U)$ is non-trivial.

On the other hand, $\mathrm{BP}^\infty(U)(\cdot|i)$ are the same for $i = 2, \ldots, q$. This cannot happen for any non-trivial FMS channel. Therefore $\mathrm{BP}^\infty(U)$ is not an FMS channel. □

Nevertheless, when the condition in Theorem 11.1 holds, for an open set of initial channels, it will converge to the unique FMS fixed point under BP iterations. We make the following definition.

Definition 11.25. Let $U : \mathcal{X} \to \mathcal{Y}$ be a channel where $\mathcal{X} = [q]$. We say U has full rank if there exists a partition of $\mathcal{Y}$ into q measurable subsets $\mathcal{Y} = E_1 \cup \cdots \cup E_q$ such that the $q \times q$ matrix

$$(U(E_j|i))_{i\in[q],j\in[q]} \tag{11.120}$$

is invertible.

Proposition 11.26. *Work under the setting of Theorem 11.1. If (q, λ, d) satisfies Eq.* (11.22) *or Eq.* (11.23)*, then for any q-ary input (possibly asymmetric) channel U of full rank, we have*

$$\mathrm{BP}^{\infty}(U) = \mathrm{BP}^{\infty}(\mathrm{Id}). \tag{11.121}$$

Proof. We prove that under the condition that U has full rank, there exists a channel R such that $R \circ U$ is a non-trivial Potts channel.

Because U has full rank, there exists a partition $\mathcal{Y} = E_1 \cup \cdots \cup E_q$ such that

$$(U(E_j|i))_{i\in[q],j\in[q]} \tag{11.122}$$

is invertible. Define $Q : \mathcal{Y} \to [q]$ by mapping $y \in E_i$ to i for all $i \in [q]$. Then we can replace U with $Q \circ U$ and wlog assume that $\mathcal{Y} = [q]$.

By Lemma 5.23, there exists $\lambda > 0$ such that $P_\lambda \le_{\deg} U$. Therefore $P_\lambda \le_{\deg} U \le_{\deg} \mathrm{Id}$. Degradation of q-ary input (possibly asymmetric) channels is preserved under BP operator. So by iterating the BP operator, we get

$$\mathrm{BP}^{\infty}(P_\lambda) \le_{\deg} \mathrm{BP}^{\infty}(U) \le_{\deg} \mathrm{BP}^{\infty}(\mathrm{Id}). \tag{11.123}$$

The first and third channels are equal by Theorem 11.1. Therefore $\mathrm{BP}^{\infty}(U) = \mathrm{BP}^{\infty}(\mathrm{Id})$. □

For the boundary irrelevance operator BP_W (Eq. (11.11)), the situation is simpler: when the survey FMS channel W is non-trivial, there is no asymmetric fixed point.

Proposition 11.27. *Work under the setting of Theorem 11.1. If (q, λ, d, W) satisfies Eq.* (11.22) *or Eq.* (11.24)*, then BP_W has only one fixed point.*

Proof. Suppose U is a fixed point of BP_W. We have

$$0 \le_{\deg} U \le_{\deg} \mathrm{Id}\,. \tag{11.124}$$

Degradation for q-ary input (possibly asymmetric) channels is preserved under BP_W operator. So by iterating the BP_W operator, we get

$$\mathrm{BP}_W^{\infty}(0) \le_{\deg} \mathrm{BP}_W^{\infty}(U) \le_{\deg} \mathrm{BP}_W^{\infty}(\mathrm{Id}). \tag{11.125}$$

The first and third channels are equal by Theorem 11.1. Therefore $U = \mathrm{BP}_W^{\infty}(U) = \mathrm{BP}_W^{\infty}(\mathrm{Id})$ is equal to the unique FMS fixed point. □

Bibliography

[1] Emmanuel Abbe. Community detection and stochastic block models: recent developments. *The Journal of Machine Learning Research*, 18(1):6446–6531, 2017.

[2] Emmanuel Abbe, Afonso S. Bandeira, and Georgina Hall. Exact recovery in the stochastic block model. *IEEE Transactions on Information Theory*, 62(1):471–487, 2015.

[3] Emmanuel Abbe and Enric Boix-Adserà. Subadditivity beyond trees and the chi-squared mutual information. In *2019 IEEE International Symposium on Information Theory (ISIT)*, pages 697–701. IEEE, 2019.

[4] Emmanuel Abbe, Elisabetta Cornacchia, Yuzhou Gu, and Yury Polyanskiy. Stochastic block model entropy and broadcasting on trees with survey. In *Conference on Learning Theory*, pages 1–25. PMLR, 2021.

[5] Emmanuel Abbe and Andrea Montanari. Conditional random fields, planted constraint satisfaction and entropy concentration. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques: 16th International Workshop, APPROX 2013, and 17th International Workshop, RANDOM 2013, Berkeley, CA, USA, August 21-23, 2013. Proceedings*, pages 332–346. Springer, 2013.

[6] Emmanuel Abbe and Colin Sandon. Community detection in general stochastic block models: Fundamental limits and efficient algorithms for recovery. In *2015 IEEE 56th Annual Symposium on Foundations of Computer Science*, pages 670–688. IEEE, 2015.

[7] Emmanuel Abbe and Colin Sandon. Detection in the stochastic block model with multiple clusters: proof of the achievability conjectures, acyclic bp, and the information-computation gap. *arXiv preprint arXiv:1512.09080*, 2015.

[8] Emmanuel Abbe and Colin Sandon. Achieving the KS threshold in the general stochastic block model with linearized acyclic belief propagation. *Advances in Neural Information Processing Systems*, 29, 2016.

[9] Emmanuel Abbe and Colin Sandon. Proof of the achievability conjectures for the general stochastic block model. *Communications on Pure and Applied Mathematics*, 71(7):1334–1406, 2018.

[10] Ragi Abou-Chacra, D. J. Thouless, and P. W. Anderson. A selfconsistent theory of localization. *Journal of Physics C: Solid State Physics*, 6(10):1734, 1973.

[11] Rudolf Ahlswede and Peter Gács. Spreading of sets in product spaces and hypercontraction of the markov operator. *The Annals of Probability*, pages 925–939, 1976.

[12] Kwangjun Ahn, Kangwook Lee, and Changho Suh. Hypergraph spectral clustering in the weighted stochastic block model. *IEEE Journal of Selected Topics in Signal Processing*, 12(5):959–974, 2018.

[13] Nima Anari, Vishesh Jain, Frederic Koehler, Huy Tuan Pham, and Thuy-Duong Vuong. Entropic independence i: Modified log-sobolev inequalities for fractionally log-concave distributions and high-temperature ising models. *arXiv preprint arXiv:2106.04105*, 2021.

[14] Nima Anari, Kuikui Liu, and Shayan Oveis Gharan. Spectral independence in high-dimensional expanders and applications to the hardcore model. *SIAM Journal on Computing*, 0(0):FOCS20–1–FOCS20–37, 0.

[15] Maria Chiara Angelini, Francesco Caltagirone, Florent Krzakala, and Lenka Zdeborová. Spectral detection on sparse hypergraphs. In *2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 66–73. IEEE, 2015.

[16] Jess Banks, Cristopher Moore, Joe Neeman, and Praneeth Netrapalli. Information-theoretic thresholds for community detection in sparse networks. In *Conference on Learning Theory*, pages 383–416. PMLR, 2016.

[17] P. Bergmans. Random coding theorem for broadcast channels with degraded components. *IEEE Transactions on Information Theory*, 19(2):197–207, 1973.

[18] Arthur J. Bernstein. Maximally connected arrays on the n-cube. *SIAM Journal on Applied Mathematics*, 15(6):1485–1489, 1967.

[19] Nayantara Bhatnagar, Allan Sly, and Prasad Tetali. Reconstruction threshold for the hardcore model. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques: 13th International Workshop, APPROX 2010, and 14th International Workshop, RANDOM 2010, Barcelona, Spain, September 1-3, 2010. Proceedings*, pages 434–447. Springer, 2010.

[20] Nayantara Bhatnagar, Juan Vera, Eric Vigoda, and Dror Weitz. Reconstruction for colorings on trees. *SIAM Journal on Discrete Mathematics*, 25(2):809–826, 2011.

[21] Peter J. Bickel and Aiyou Chen. A nonparametric view of network models and newman–girvan and other modularities. *Proceedings of the National Academy of Sciences*, 106(50):21068–21073, 2009.

[22] Antonio Blanca, Pietro Caputo, Zongchen Chen, Daniel Parisi, Daniel Štefankovič, and Eric Vigoda. On mixing of markov chains: Coupling, spectral independence, and entropy factorization. In *Proceedings of the 2022 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 3670–3692. SIAM, 2022.

[23] Pavel M. Bleher, Jean Ruiz, and Valentin A. Zagrebnov. On the purity of the limiting gibbs state for the ising model on the bethe lattice. *Journal of Statistical Physics*, 79:473–482, 1995.

[24] Sergey G. Bobkov and Prasad Tetali. Modified logarithmic sobolev inequalities in discrete settings. *Journal of Theoretical Probability*, 19:289–336, 2006.

[25] Ravi B. Boppana. Eigenvalues and graph bisection: An average-case analysis. In *28th Annual Symposium on Foundations of Computer Science (SFCS 1987)*, pages 280–285. IEEE, 1987.

[26] Charles Bordenave, Marc Lelarge, and Laurent Massoulié. Non-backtracking spectrum of random graphs: community detection and non-regular ramanujan graphs. In *2015 IEEE 56th Annual Symposium on Foundations of Computer Science*, pages 1347–1357. IEEE, 2015.

[27] Christian Borgs, Jennifer Chayes, Elchanan Mossel, and Sébastien Roch. The kesten-stigum reconstruction bound is tight for roughly symmetric binary channels. In *2006 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS'06)*, pages 518–530. IEEE, 2006.

[28] Mark Braverman, Ankit Garg, Tengyu Ma, Huy L. Nguyen, and David P. Woodruff. Communication lower bounds for statistical estimation problems via a distributed data processing inequality. In *Proceedings of the forty-eighth annual ACM Symposium on Theory of Computing*, pages 1011–1020, 2016.

[29] Alexandre Bristiel and Pietro Caputo. Entropy inequalities for random walks and permutations. *arXiv preprint arXiv:2109.06009*, 2021.

[30] Thang Bui, Soma Chaudhuri, Tom Leighton, and Michael Sipser. Graph bisection algorithins with good average case behavior. In *25th Annual Symposium on Foundations of Computer Science, 1984.*, pages 181–192. IEEE, 1984.

[31] Francesco Caltagirone, Marc Lelarge, and Léo Miolane. Recovering asymmetric communities in the stochastic block model. *IEEE Transactions on Network Science and Engineering*, 5(3):237–246, 2017.

[32] Yuansi Chen and Ronen Eldan. Localization schemes: A framework for proving mixing bounds for markov chains. In *2022 IEEE 63rd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 110–122. IEEE, 2022.

[33] Zongchen Chen, Andreas Galanis, Daniel Štefankovič, and Eric Vigoda. Rapid mixing for colorings via spectral independence. In *Proceedings of the 2021 ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 1548–1557. SIAM, 2021.

[34] Zongchen Chen, Kuikui Liu, and Eric Vigoda. Optimal mixing of glauber dynamics: Entropy factorization via high-dimensional expansion. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 1537–1550, 2021.

[35] I Chien, Chung-Yi Lin, and I-Hsiang Wang. Community detection in hypergraphs: Optimal statistical limit and efficient algorithms. In *International Conference on Artificial Intelligence and Statistics*, pages 871–879. PMLR, 2018.

[36] I Eli Chien, Chung-Yi Lin, and I-Hsiang Wang. On the minimax misclassification ratio of hypergraph community detection. *IEEE Transactions on Information Theory*, 65(12):8095–8118, 2019.

[37] Byron Chin and Allan Sly. Optimal recovery of block models with q communities. *arXiv preprint arXiv:2010.10672*, 2020.

[38] Byron Chin and Allan Sly. Optimal reconstruction of general sparse stochastic block models. *arXiv preprint arXiv:2111.00697*, 2021.

[39] Man-Duen Choi, Mary Beth Ruskai, and Eugene Seneta. Equivalence of certain entropy contraction coefficients. *Linear Algebra and its Applications*, 208:29–36, 1994.

[40] Joel Cohen, Johannes HB Kempermann, and Gheorghe Zbaganu. *Comparisons of stochastic matrices with applications in information theory, statistics, economics and population*. Springer Science & Business Media, 1998.

[41] Joel E. Cohen, Yoh Iwasa, Gh. Rautu, Mary Beth Ruskai, Eugene Seneta, and Gh. Zbaganu. Relative entropy under mappings by stochastic matrices. *Linear Algebra and its Applications*, 179:211–235, 1993.

[42] Amin Coja-Oghlan. Graph partitioning via adaptive spectral techniques. *Combinatorics, Probability and Computing*, 19(2):227–284, 2010.

[43] Amin Coja-Oghlan, Florent Krzakala, Will Perkins, and Lenka Zdeborová. Information-theoretic thresholds from the cavity method. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pages 146–157, 2017.

[44] Sam Cole and Yizhe Zhu. Exact recovery in the hypergraph stochastic block model: A spectral algorithm. *Linear Algebra and its Applications*, 593:45–73, 2020.

[45] Anne Condon and Richard M. Karp. Algorithms for graph partitioning on the planted partition model. *Random Structures & Algorithms*, 18(2):116–140, 2001.

[46] Thomas A. Courtade. Outer bounds for multiterminal source coding via a strong data processing inequality. In *2013 IEEE International Symposium on Information Theory*, pages 559–563. IEEE, 2013.

[47] Thomas Cover. Broadcast channels. *IEEE Transactions on Information Theory*, 18(1):2–14, 1972.

[48] Imre Csiszár and János Körner. *Information theory: coding theorems for discrete memoryless systems*. Cambridge University Press, 2011.

[49] Aurelien Decelle, Florent Krzakala, Cristopher Moore, and Lenka Zdeborová. Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications. *Physical Review E*, 84(6):066106, 2011.

[50] Yash Deshpande, Emmanuel Abbe, and Andrea Montanari. Asymptotic mutual information for the balanced binary stochastic block model. *Information and Inference: A Journal of the IMA*, 6(2):125–170, 2017.

[51] Persi Diaconis and Laurent Saloff-Coste. Logarithmic sobolev inequalities for finite markov chains. *The Annals of Applied Probability*, 6(3):695–750, 1996.

[52] Roland L. Dobrushin. Central limit theorem for nonstationary markov chains. i. *Theory of Probability & Its Applications*, 1(1):65–80, 1956.

[53] Roland L'vovich Dobrushin. Definition of random variables by conditional distributions. *Theor. Probability Appl*, 15(3):469–497, 1970.

[54] Tomas Dominguez and Jean-Christophe Mourrat. Mutual information for the sparse stochastic block model. *arXiv preprint arXiv:2209.04513*, 2022.

[55] John C. Duchi, Michael I. Jordan, and Martin J. Wainwright. Local privacy and statistical minimax rates. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, pages 429–438. IEEE, 2013.

[56] Ioana Dumitriu, Haixiao Wang, and Yizhe Zhu. Partial recovery and weak consistency in the non-uniform hypergraph stochastic block model. *arXiv preprint arXiv:2112.11671*, 2021.

[57] Rick Durrett. *Probability: theory and examples*, volume 49. Cambridge university press, 2019.

[58] Martin E. Dyer and Alan M. Frieze. The solution of some random np-hard problems in polynomial expected time. *Journal of Algorithms*, 10(4):451–489, 1989.

[59] Charilaos Efthymiou. Reconstruction/non-reconstruction thresholds for colourings of general galton-watson trees. *arXiv preprint arXiv:1406.3617*, 2014.

[60] Ronen Eldan, Dan Mikulincer, and Hester Pieters. Community detection and percolation of information in a geometric setting. *Combinatorics, Probability and Computing*, 31(6):1048–1069, 2022.

[61] Michel Émery and Joseph E. Yukich. A simple proof of the logarithmic sobolev inequality on the circle. *Séminaire de probabilités de Strasbourg*, 21:173–175, 1987.

[62] Elza Erkip and Thomas M. Cover. The efficiency of investment information. *IEEE Transactions on Information Theory*, 44(3):1026–1040, 1998.

[63] William Evans, Claire Kenyon, Yuval Peres, and Leonard J. Schulman. Broadcasting on trees and the ising model. *Annals of Applied Probability*, pages 410–433, 2000.

[64] William S. Evans and Leonard J. Schulman. Signal propagation and noisy circuits. *IEEE Transactions on Information Theory*, 45(7):2367–2373, 1999.

[65] Weiming Feng, Heng Guo, Yitong Yin, and Chihao Zhang. Rapid mixing from spectral independence beyond the boolean domain. *ACM Transactions on Algorithms (TALG)*, 18(3):1–32, 2022.

[66] Marco Formentin and Christof Külske. On the purity of the free boundary condition potts measure on random trees. *Stochastic Processes and their Applications*, 119(9):2992–3005, 2009.

[67] Debarghya Ghoshdastidar and Ambedkar Dukkipati. A provable generalized tensor spectral method for uniform hypergraph partitioning. In *International Conference on Machine Learning*, pages 400–409. PMLR, 2015.

[68] Debarghya Ghoshdastidar and Ambedkar Dukkipati. Spectral clustering using multilinear SVD: Analysis, approximations and applications. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.

[69] Debarghya Ghoshdastidar and Ambedkar Dukkipati. Consistency of spectral hypergraph partitioning under planted partition model. *The Annals of Statistics*, 45(1):289–315, 2017.

[70] Sharad Goel. Modified logarithmic sobolev inequalities for some models of random walk. *Stochastic Processes and their Applications*, 114(1):51–79, 2004.

[71] Leonard Gross. Logarithmic sobolev inequalities. *American Journal of Mathematics*, 97(4):1061–1083, 1975.

[72] Yuzhou Gu and Yury Polyanskiy. Non-linear log-Sobolev inequalities for the Potts semigroup and applications to reconstruction problems. *arXiv preprint arXiv:2005.05444*, 2020.

[73] Yuzhou Gu and Yury Polyanskiy. Uniqueness of BP fixed point for the Potts model and applications to community detection. *arXiv preprint arXiv:2303.14688*, 2023.

[74] Yuzhou Gu and Yury Polyanskiy. Weak recovery threshold for the hypergraph stochastic block model. *arXiv preprint arXiv:2303.14689*, 2023.

[75] Yuzhou Gu, Hajir Roozbehani, and Yury Polyanskiy. Broadcasting on trees near criticality. In *2020 IEEE International Symposium on Information Theory (ISIT)*, pages 1504–1509. IEEE, 2020.

[76] Uri Hadar, Jingbo Liu, Yury Polyanskiy, and Ofer Shayevitz. Communication complexity of estimating correlations. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, pages 792–803, 2019.

[77] Godfrey H. Hardy, John E. Littlewood, and George Pólya. *Inequalities*. Cambridge University Press, 1934.

[78] Lawrence Hueston Harper. Optimal assignments of numbers to vertices. *Journal of the Society for Industrial and Applied Mathematics*, 12(1):131–135, 1964.

[79] Sergiu Hart. A note on the edges of the n-cube. *Discrete Mathematics*, 14(2):157–163, 1976.

[80] Dmitry Ioffe. Extremality of the disordered state for the ising model on general trees. In *Trees: Workshop in Versailles, June 14–16 1995*, pages 3–14. Springer, 1996.

[81] Dmitry Ioffe. On the extremality of the disordered state for the ising model on the bethe lattice. *Letters in Mathematical Physics*, 37:137–143, 1996.

[82] Svante Janson and Elchanan Mossel. Robust reconstruction on trees is determined by the second eigenvalue. *The Annals of Probability*, 32(3B):2630–2649, 2004.

[83] Varun Kanade, Elchanan Mossel, and Tselil Schramm. Global and local information in clustering labeled block models. *IEEE Transactions on Information Theory*, 62(10):5906–5917, 2016.

[84] Harry Kesten and Bernt P. Stigum. Additional limit theorems for indecomposable multidimensional galton-watson processes. *The Annals of Mathematical Statistics*, 37(6):1463–1481, 1966.

[85] Chiheon Kim, Afonso S. Bandeira, and Michel X. Goemans. Stochastic block model for hypergraphs: Statistical limits and a semidefinite programming approach. *arXiv preprint arXiv:1807.02884*, 2018.

[86] J. Körner and K. Marton. Comparison of two noisy channels. *Topics in Information Theory, I. Csiszár and P. Elias, Eds., Amsterdam, The Netherlands*, pages 411–423, 1977.

[87] Christof Külske and Marco Formentin. A symmetric entropy bound on the non-reconstruction regime of Markov chains on Galton-Watson trees. *Electronic Communications in Probability*, 14:587–596, 2009.

[88] David A. Levin and Yuval Peres. *Markov chains and mixing times*, volume 107. American Mathematical Soc., 2017.

[89] Chung-Yi Lin, I Eli Chien, and I-Hsiang Wang. On the fundamental statistical limit of community detection in random hypergraphs. In *2017 IEEE International Symposium on Information Theory (ISIT)*, pages 2178–2182. IEEE, 2017.

[90] John H. Lindsey II. Assignment of numbers to vertices. *The American Mathematical Monthly*, 71(5):508–516, 1964.

[91] Wenjian Liu and Ning Ning. Large degree asymptotics and the reconstruction threshold of the asymmetric binary channels. *Journal of Statistical Physics*, 174:1161–1188, 2019.

[92] Russell Lyons. Random walks and percolation on trees. *The Annals of Probability*, 18(3):931–958, 1990.

[93] Anuran Makur. *Information contraction and decomposition*. PhD thesis, Massachusetts Institute of Technology, 2019.

[94] Anuran Makur and Yury Polyanskiy. Comparison of channels: Criteria for domination by a symmetric channel. *IEEE Transactions on Information Theory*, 64(8):5704–5725, 2018.

[95] Fabio Martinelli, Alistair Sinclair, and Dror Weitz. Fast mixing for independent sets, colorings, and other models on trees. *Random Structures & Algorithms*, 31(2):134–172, 2007.

[96] Laurent Massoulié. Community detection thresholds and the weak ramanujan property. In *Proceedings of the forty-sixth annual ACM Symposium on Theory of Computing*, pages 694–703, 2014.

[97] Frank McSherry. Spectral partitioning of random graphs. In *Proceedings 42nd IEEE Symposium on Foundations of Computer Science*, pages 529–537. IEEE, 2001.

[98] Marc Mézard and Andrea Montanari. Reconstruction on trees and spin glass transition. *Journal of statistical physics*, 124:1317–1350, 2006.

[99] Marc Mézard and Giorgio Parisi. The bethe lattice spin glass revisited. *The European Physical Journal B-Condensed Matter and Complex Systems*, 20:217–233, 2001.

[100] Laurent Miclo. *Remarques sur l'hypercontractivité et l'évolution de l'entropie pour des chaînes de Markov finies*. Springer, 1997.

[101] Elchanan Mossel. Reconstruction on trees: beating the second eigenvalue. *The Annals of Applied Probability*, 11(1):285–300, 2001.

[102] Elchanan Mossel, Joe Neeman, and Allan Sly. Consistency thresholds for binary symmetric block models. *arXiv preprint arXiv:1407.1591*, 3(5), 2014.

[103] Elchanan Mossel, Joe Neeman, and Allan Sly. Reconstruction and estimation in the planted partition model. *Probability Theory and Related Fields*, 162:431–461, 2015.

[104] Elchanan Mossel, Joe Neeman, and Allan Sly. Belief propagation, robust reconstruction and optimal recovery of block models. *The Annals of Applied Probability*, 26(4):2211–2256, 2016.

[105] Elchanan Mossel, Joe Neeman, and Allan Sly. A proof of the block model threshold conjecture. *Combinatorica*, 38(3):665–708, 2018.

[106] Elchanan Mossel, Krzysztof Oleszkiewicz, and Arnab Sen. On reverse hypercontractivity. *Geometric and Functional Analysis*, 23(3):1062–1097, 2013.

[107] Elchanan Mossel and Yuval Peres. Information flow on trees. *The Annals of Applied Probability*, 13(3):817–844, 2003.

[108] Elchanan Mossel, Sébastien Roch, and Allan Sly. Robust estimation of latent tree graphical models: Inferring hidden states with inexact parameters. *IEEE Transactions on Information Theory*, 59(7):4357–4373, 2013.

[109] Elchanan Mossel, Allan Sly, and Youngtak Sohn. Exact phase transitions for stochastic block models and reconstruction on trees. *arXiv preprint arXiv:2212.03362*, 2022.

[110] Elchanan Mossel and Jiaming Xu. Local algorithms for block models with side information. In *Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science*, pages 71–80, 2016.

[111] Or Ordentlich and Yury Polyanskiy. Strong data processing constant is achieved by binary inputs. *IEEE Transactions on Information Theory*, 68(3):1480–1481, 2021.

[112] Soumik Pal and Yizhe Zhu. Community detection in the sparse hypergraph stochastic block model. *Random Structures & Algorithms*, 59(3):407–463, 2021.

[113] Robin Pemantle and Yuval Peres. The critical Ising model on trees, concave recursions and nonlinear capacity. *The Annals of Probability*, 38(1):184 – 206, 2010.

[114] Yury Polyanskiy and Alex Samorodnitsky. Improved log-sobolev inequalities, hypercontractivity and uncertainty principle on the hypercube. *Journal of Functional Analysis*, 277(11):108280, 2019.

[115] Yury Polyanskiy and Yihong Wu. Strong data-processing inequalities for channels and bayesian networks. In *Convexity and Concentration*, pages 211–249. Springer, 2017.

[116] Yury Polyanskiy and Yihong Wu. Application of the information-percolation method to reconstruction problems on graphs. *Mathematical Statistics and Learning*, 2(1):1–24, 2020.

[117] Yury Polyanskiy and Yihong Wu. *Information Theory: From Coding to Learning*. Cambridge University Press, 2023+.

[118] Maxim Raginsky. Strong data processing inequalities and ϕ-sobolev inequalities for discrete channels. *IEEE Transactions on Information Theory*, 62(6):3355–3389, 2016.

[119] Alfréd Rényi. On measures of dependence. *Acta Mathematica Hungarica*, 10(3-4):441–451, 1959.

[120] Tom Richardson and Rüdiger Urbanke. *Modern coding theory*. Cambridge university press, 2008.

[121] Hajir Roozbehani and Yury Polyanskiy. Low density majority codes and the problem of graceful degradation. *arXiv preprint arXiv:1911.12263*, 2019.

[122] Oleg Vasil'evich Sarmanov. Maximum correlation coefficient (nonsymmetric case). *Selected translations in mathematical statistics and probability*, 2:207–210, 1963.

[123] Eren Sasoglu. *Polar coding theorems for discrete systems*. PhD thesis, EPFL, 2011.

[124] Claude E. Shannon. A note on a partial ordering for communication channels. *Information and control*, 1(4):390–397, 1958.

[125] Allan Sly. Reconstruction of random colourings. *Communications in Mathematical Physics*, 288(3):943–961, 2009.

[126] Allan Sly. Reconstruction for the Potts model. *The Annals of Probability*, 39(4):1365 – 1406, 2011.

[127] Tom A. B. Snijders and Krzysztof Nowicki. Estimation and prediction for stochastic blockmodels for graphs with latent block structure. *Journal of classification*, 14(1):75–100, 1997.

[128] Ludovic Stephan and Laurent Massoulié. Robustness of spectral methods for community detection. In *Conference on Learning Theory*, pages 2831–2860. PMLR, 2019.

[129] Ludovic Stephan and Laurent Massoulié. Non-backtracking spectra of weighted inhomogeneous random graphs. *Mathematical Statistics and Learning*, 5(3):201–271, 2022.

[130] Ludovic Stephan and Yizhe Zhu. Sparse random hypergraphs: Non-backtracking spectra and community detection. In *2022 IEEE 63rd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 567–575. IEEE, 2022.

[131] David Sutter and Joseph M. Renes. Universal polar codes for more capable and less-noisy channels and sources. In *2014 IEEE International Symposium on Information Theory*, pages 1461–1465. IEEE, 2014.

[132] Igor Vajda. On metric divergences of probability measures. *Kybernetika*, 45(6):885–900, 2009.

[133] Van Vu. A simple SVD algorithm for finding hidden partitions. *Combinatorics, Probability and Computing*, 27(1):124–140, 2018.

[134] Aaron D. Wyner and Jacob Ziv. A theorem on the entropy of certain binary sequences and applications: Part I. *IEEE Transactions on Information Theory*, 19(6):769–772, 1973.

[135] Aolin Xu and Maxim Raginsky. Converses for distributed estimation via strong data processing inequalities. In *2015 IEEE International Symposium on Information Theory (ISIT)*, pages 2376–2380. IEEE, 2015.

[136] Qian Yu and Yury Polyanskiy. Broadcasting on trees near criticality: Perturbation theory. In *2021 IEEE International Symposium on Information Theory (ISIT)*, pages 2101–2106. IEEE, 2021.

[137] Qian Yu and Yury Polyanskiy. Ising model on locally tree-like graphs: Uniqueness of solutions to cavity equations. *arXiv preprint arXiv:2211.15242*, 2022.

[138] Erchuan Zhang, David Suter, Giang Truong, and Syed Zulqarnain Gilani. Sparse hypergraph community detection thresholds in stochastic block model. In *Thirty-Sixth Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

[139] Qiaosheng Zhang and Vincent Y. F. Tan. Exact recovery in the general hypergraph stochastic block model. *IEEE Transactions on Information Theory*, 69(1):453–471, 2022.