上海交通大學

SHANGHAI JIAO TONG UNIVERSITY

学士学位论文

BACHELOR’S THESIS

论文题目： Jensen’s Inequality, Partition Functions, and Models with Ternary Interactions

学生姓名: 王彦恒

学生学号: 517021910537

专　　业: 计算机科学与技术

指导教师: Prof. Dominik Scheder

学院(系): 电子信息与电气工程学院

上海交通大学
学位论文原创性声明

本人郑重声明：所呈交的学位论文 *Jensen's Inequality, Partition Functions, and Models with Ternary Interactions*，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的作品成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。本人完全意识到本声明的法律结果由本人承担。

学位论文作者签名：

王彦恒

日期：2021 年 6 月 7 日

上海交通大学
学位论文版权使用授权书

本学位论文作者完全了解学校有关保留、使用学位论文的规定，同意学校保留并向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅和借阅。本人授权上海交通大学可以将本学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存和汇编本学位论文。

本学位论文属于 **保　密** □，在____年解密后适用本授权书。
不保密 ☑。

（请在以上方框内打“✓”）

学位论文作者签名：
王彦恒

日期：2021 年 6 月 7 日

指导教师签名：

日期：2021 年 06月 07日

Jensen's Inequality, Partition Functions, and Models with Ternary Interactions

摘 要

PPZ 算法是一种解决 k-SAT 问题的随机算法，分析之在理论计算机科学中有重要意义。其分析可被抽象为如下的图模型：给定一张 $(k-1)$-正则有向图，设每个顶点独立取得在 $[0,1]$ 上均匀分布的随机值，若一顶点的值是其前驱邻点之中最大的，那么它得一分，否则得零分；记全体顶点的总分为 S，则 PPZ 的成功概率恰正比于 $\mathbb{E}(2^S)$ 的值。本文旨在寻找该期望的上下界。我们借助信息论、统计物理和计数算法的相关技术，证明了一系列不等式，具体如下：(1) 由变分视角和组合学方法得到了适用于二维方格图（$k=3$ 之特例）的若干下界；(2) 采用信息论和统计物理手段证明了一系列普适上界；(3) 巧妙运用条件信息熵证明了普适上界 $2^{n\cdot\Theta(\log k/k)}$，且说明了它在渐进意义上的最优性。除此以外，我们还探讨了图围长很大时 $\mathbb{E}(2^S)$ 的行为，猜想其仅与顶点数相关而与图结构无关，并为该猜想提供了很强的理论证据。这些结果加深了我们对 PPZ 的理解，亦可能有助于分析类似的复杂概率系统。

关键词：SAT 问题，PPZ 算法，配分函数，信息论

Jensen's Inequality, Partition Functions, and Models with Ternary Interactions

ABSTRACT

We study a graphical model that captures the success probability of the PPZ algorithm for k-SAT. Given a $(k-1)$-regular digraph, we put on each vertex a $[0,1]$ uniform random value independently. A vertex scores 1 if it gets the largest value amongst its predecessors. Denoting S as the total score, we aim at bounding $\mathbb{E}(2^S)$, a significant quantity in the theoretical study of PPZ. Our methodologies borrow ideas from information theory, statistical physics and counting algorithms. We proved (i) a bunch of lower bounds, tailored to the two-dimensional grid (where $k=3$), by a variational principle and a combinatorial embedding idea; (ii) a set of upper bounds that work for general k, via connections with the Boltzmann distribution in statistical physics; (iii) a tight $2^{n\cdot\Theta(\log k/k)}$ upper bound for general k by an entropy argument. We also reveal strong theoretical evidence that $\mathbb{E}(2^S)$ are essentially identical for all high-girth graphs of order n. Besides the immediate consequence in the context of PPZ, we hope the various methods presented in this thesis will benefit other fields and applications as well.

Key words: SAT, PPZ algorithm, partition functions, information theory

Contents

Chapter 1 Introduction

The satisfiability problem, or SAT for short, poses constant challenge for computer scientists from its outset. Over the decades, several ingenious algorithms have reduced the worst-case running time for k-SAT (i.e. SAT with k-CNF as input formulae) down to moderate exponentials. Among these lie the PPZ algorithm [16] and its descendant PPSZ [15] that enjoy both simplicity and efficiency. The algorithms begin by permuting (i.e. ordering) the variables randomly. Then, following this order, they assign 0/1 to each variable v by flipping fair coins, unless the value of v could be *inferred* by some fixed criterion. The choice of criterion is exactly where PPZ and PPSZ differ. PPZ's slogan reads "look for unit clause (v) or $(\neg v)$; if we find one then v must be 1/0". PPSZ uses a stronger strategy: "check if there is a set of up to h clauses that logically imply v or $\neg v$". Note that for $h = 1$ this is the same as PPZ.

Our study is motivated by the analyses of PPZ/PPSZ when the input formula has a unique satisfying assignment α. The algorithms find α successfully with probability

$$\mathbb{P}(\alpha) = \mathop{\mathbb{E}}_{\pi\sim U}\left(2^{-(n-\sum_v S_v(\pi))}\right) = 2^{-n}\cdot \mathop{\mathbb{E}}_{\pi\sim U}\left(2^{\sum_v S_v(\pi)}\right)$$

where π is sampled uniformly from all permutations on $[n] = \{1, 2, \ldots, n\}$, and $S_v(\pi) \in \{0, 1\}$ signals if we could save the coin flip at v under permutation π. If nothing is saved, then the algorithms degenerate to random guess (with success probability 2^{-n}). But the insight is: for a considerable amount of permutations we actually save a lot.

Since α is assumed unique, altering any variable v in α will violate some clause. In particular, there must be a clause containing v where all other literals evaluate to 0 under α. We call it a *critical clause* for v and assume it is unique for simplicity. Then in PPZ, $S_v(\pi) = 1$ if and only if v comes last in *the* critical clause under permutation π. Hence, we may identify the problem with a graphical model $G = (V, E)$, where $V = [n]$ is exactly the set of variables, and $uv \in E$ if u appears in the critical clause of v. Clearly $S_v(\pi) = 1$ if and only if vertex v is bigger than all its predecessors under π. This graphical model nicely captures the local nature of the algorithm and will be the subject of our study.

By a direct application of Jensen's inequality, $\mathbb{E}_\pi\left(2^{\sum_v S_v(\pi)}\right) \geq 2^{\mathbb{E}_\pi(\sum_v S_v(\pi))} = 2^{n/k}$. There are two interesting questions to ask: By what methods can we improve this lower bound? And, from the adversary's view, what is the absolute barrier or upper bound? Answering these questions will help us understand the behaviour of PPZ, and also hint about new approaches in analysing PPSZ.

As we will explain in Section 1.2, our problem has intimate connection to *partition functions*, a central object in statistical physics as well as in combinatorics. Partition functions of many physical systems have been extensively studied by statistical physicists. Although the results did not always follow mathematical rigour, many were later proved correct and inspiring; they might just inspire us equally well.

In the computer science community, efforts are devoted to efficient approximation *algorithms* for partition functions. Three genres have made major success. The first genre is based on Markov Chain Monte Carlo [12] and the general equivalence between sampling and counting [9]. It builds a Markov chain with appropriate stationary distribution and tries to show its fast convergence. The second genre depends on the correlation decay property in statistical physics, which recursively computes a marginal distribution and argues that the boundary effect decays quickly as the distance increases [2, 18]. The last and also the most analytical one relies upon Barvinok's approach, where it uses a low-degree Taylor expansion to approximate the log-partition function [3].

Despite rich literature in both fields, our problem is peculiar in two ways. First, our model differs from well-studied physical models since it involves a k-ary, asymmetric score measure and large number of spins; see Section 1.2 for details. Second, algorithmic results are not directly applicable since we are searching for *explicit* bounds. Hence, new approaches or adaptions must be made. This serves as another motivation for us: tools developed here could possibly feed back to the study of complicated systems in computer science and statistical physics.

1.1 Definitions and Conventions

Throughout the thesis we consider the abstract model defined below. Let $G = (V, E)$ be a $(k-1)$-regular directed graph where $V = [n]$. For $k = 3$ we mainly study the $\sqrt{n} \times \sqrt{n}$ square grid (with edges directing from left to right and from bottom to top, wrapping around at boundaries). Write $\flat v := \{u \in V : uv \in E\}$ for v's predecessors, $\sharp v := \{u \in V : vu \in E\}$ for its successors, and $\partial v := \flat v \cup \sharp v$. The regularity assumption means that $|\flat v| = |\sharp v| = k - 1$.

Sample a state $x \in [0,1]^n$ *uniformly* and let

$$S_v(x) := \begin{cases} 1 & x_v \geq \max x_{\flat v} \\ 0 & \text{otherwise} \end{cases}$$

be the "local score" of v under x. The total score is $S(x) := \sum_{v \in V} S_v(x)$. We aim at bounding $\mathbb{E}_{x \sim U}(2^{S(x)})$, both from below and from above. Note that sampling a uniform $x \in [0,1]^n$ is *equivalent* to generating uniform permutations π on $[n]$.

Here are some notational conventions:

Probability. The symbols μ and ν are reserved for probability distributions. The letter U stands for the uniform distribution (whose support will be clear from context). At the beginning, we will put subscripts in $\mathbb{P}$ and $\mathbb{E}$ to stress the underlying probability space. For example, subscript $x \sim \mu$ means that x is drawn with respect to distribution μ. As we go deeper, we would drop subscripts when the context is clear.

Information theory. The *entropy* of μ is defined as $H(\mu) := -\mathbb{E}_{x\sim\mu}(\log \mu(x))$, where the logarithm is always in base 2. Here is a caveat: if μ is a continuous distribution (i.e. a density) then $H(\mu)$ could be negative, so we should be alert in this case. Nevertheless, the chain rule of entropy still apply. When the underlying distribution of random variable x is clear (say μ), we also write $H(x)$ to mean $H(\mu)$. We write $h_b(p) := -p \log p - (1-p)\log(1-p)$ for the binary entropy function. The quantity $\mathrm{KL}(\mu \| \nu) := \mathbb{E}_{x\sim\mu}\left(\log \frac{\mu(x)}{\nu(x)}\right)$ is called the *Kullback-Leibler divergence* between μ and ν. It is a standard fact that $\mathrm{KL}(\mu\|\nu) \geq 0$ even when μ and ν are continuous distributions.

1.2 Connection to Partition Functions

For each $\lambda \geq 1$ we define a distribution D_λ on $[0,1]^n$ by $D_\lambda(x) := \frac{\lambda^{S(x)}}{Z(\lambda)}$, where the normalising constant $Z(\lambda)$ is exactly $\mathbb{E}_{x\sim U}(\lambda^{S(x)})$; in particular we are interested in $Z(2)$. We adopt the shorthand $D := D_2$ since we frequently work in this distribution.

In statistical physics, a distribution in the same form as D_λ is called a *Boltzmann distribution*, with the normalising constant termed *partition function*. Different definitions of the score measure give rise to different models. One of the most-studied examples is the Ising model: a state $x \in \{-1,1\}^n$ has score $S(x) := \sum_{uv \in E} x_u x_v$, and occurs with probability $\mathbb{P}(x) := \frac{\lambda^{S(x)}}{Z_{\text{Ising}}(\lambda)}$. A bulk of literature is devoted to computing the Ising partition function $Z_{\text{Ising}}(\lambda)$ and related values. However, there are some key features that distinguish our model from existing statistical physics models.

***k*-ary asymmetric interaction.** Recall that our local score measure, S_v, involves interactions among v and its $k-1$ predecessors. Even in the grid case we still have ternary interactions. Also note that v is asymmetric to $\flat v$. In contrast, well-studied statistical models (e.g. the Ising model, the hard-core model, the monomer-dimer system) usually exhibit a binary symmetric interaction.

Large number of spins. A state x in our model is a real vector in $[0,1]^n$, so there are uncountably many possible "spins" for each vertex. Even if we discretised the model, the number of possible spins would still be large. Such systems are far less understood in statistical physics than "two-spin" systems (e.g. the Ising model).

1.3 Organisation of the Thesis

We set off our journey with some breezing lower and upper bounds in Chapter 2. Specifically, we regard $Z(\lambda)$ as an maxima in a variational problem over distributions. The methods illustrate several key ingredients in later materials. In Chapter 3 we elucidate two combinatorial ideas that help us polish previous bounds. The highlight comes at Chapter 4, where we apply information-theoretic techniques to significantly sharpen the upper bound down to $2^{n \cdot \Theta(\log k/k)}$. As we will see, it is the best possible asymptotics we could expect for general graphs. However, for high-girth graphs this bound might still be loose. Chapter 5 investigates the high-girth scenario and concludes the thesis by open problems.

Instead of collecting literature at one place, we would rather devote a separate space for related work at the end of each chapter. We hope this approach will group information more effectively and serve the reader better.

Chapter 2 A Variational View

Let us prepare the ground by exploring a simple idea for both lower and upper bounds. Let μ be any distribution supported on $[0,1]^n$. By non-negativity of KL divergence,

$$\begin{aligned} 0 \le \mathrm{KL}(\mu \| D_\lambda) &= \mathop{\mathbb{E}}_{x\sim\mu}(\log \mu(x) - \log D_\lambda(x)) \\ &= -H(\mu) - \mathop{\mathbb{E}}_{x\sim\mu}(S(x)\log\lambda - \log Z(\lambda)) \\ &= \log Z(\lambda) - H(\mu) - \log\lambda \cdot \mathop{\mathbb{E}}_{x\sim\mu}(S(x)). \end{aligned}$$

Specifically for $\lambda = 2$,

$$Z(2) \ge 2^{\mathbb{E}_{x\sim\mu}(S(x))+H(\mu)} \quad =: 2^{\mathcal{F}(\mu)} \tag{2–1}$$

where the inequality is tight if and only if $\mu = D$. Hence, the distribution D is the unique maxima of the functional $\mathcal{F} : \mu \mapsto \mathbb{E}_{x\sim\mu}(S(x)) + H(\mu)$. This view has two ways of utilisation: for lower bounds, we simply make up a "good and simple" μ and evaluate the functional; for upper bounds, we turn to analyse D and bound $\mathcal{F}(D)$ from above.

2.1 Lower Bounds

2.1.1 Markovian sampling

What is an ideal choice for μ? First, it should imitate the behaviour of D so that $\mathcal{F}(\mu)$ closely estimates $\mathcal{F}(D)$. Second, the evaluation of $\mathcal{F}(\mu)$ should be simple, preferably boiled into local computable snippets. The two goals are conflicting. If we take $\mu := D$ then the first goal is perfectly satisfied but $\mathcal{F}(\mu)$ is not amenable to evaluation. On the other hand, if we choose $\mu := U$ then the converse is true.

A reasonable compromise would be sampling multiple independent Markov chains on the graph. To be concrete, we assume working in the square grid and write x_{ij} for the x value at row i, column j. We sample the first column $\{x_{i1}\}$ independently, and then drive the remaining variables by $\sqrt{n}$ independent Markov chains from left to right. That is, $\mu(x_{ij} \mid x_{i,j-1}) := \nu(x_{i,j-1}, x_{ij})$ for some Markovian transition rule ν. The joint distribution induced by this sampling scheme is our μ. See Figure 2–1 for illustration.

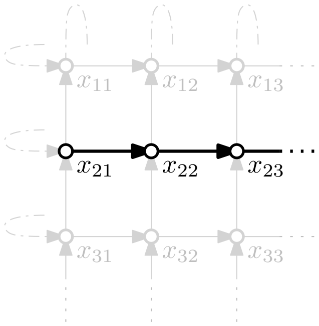

Figure 2–1 The thick line shows one of $\sqrt{n}$ Markov chains on the grid.

To further ease our computation, we assume sampling the first column *uniformly*, and that all Markov chains have identical rule ν with *uniform stationary distribution* (equivalently, ν is doubly-stochastic). This way, (1) the marginal $\mu(x_{ij})$ is uniform for all i, j; (2) rule ν can be identified with a probability distribution on pairs, so its entropy, etc. are well-defined.

For a distribution μ defined in this fashion, we can readily evaluate $\mathcal{F}(\mu)$. Indeed, using the chain rule, we can write an explicit formula for μ:

$$\mu(x) = \prod_{i=1}^{\sqrt{n}} \prod_{j=2}^{\sqrt{n}} \nu(x_{i,j-1}, x_{ij}).$$

Hence the entropy $H(\mu)$ is simply $(n - \sqrt{n})H(\nu)$. As for the expected score, if we ignore the first column, then

$$\begin{aligned} \mathop{\mathbb{E}}_{x\sim\mu} S(x) &\geq \sum_{i=1}^{\sqrt{n}} \sum_{j=2}^{\sqrt{n}} \mathop{\mathbb{E}}_{x\sim\mu} S_{ij}(x) \\ &= \sum_{i=1}^{\sqrt{n}} \sum_{j=2}^{\sqrt{n}} \mathop{\mathbb{E}}_{\substack{(x_{i,j-1}, x_{ij})\sim\nu \\ x_{i+1,j}\sim U}} S_{ij}(x_{i,j-1}, x_{ij}, x_{i+1,j}) \\ &= (n - \sqrt{n}) \mathop{\mathbb{P}}_{\substack{(x_u, x_v)\sim\nu \\ x_w\sim U}} (x_v \geq x_u, x_w) \end{aligned}$$

where in the last line we fix an arbitrary vertex v and understand u and w as its left and lower neighbours, respectively. We see that $\mathcal{F}(\mu)$ has been boiled down to local pieces:

$$\mathcal{F}(\mu) \geq (n - \sqrt{n}) \left(\mathop{\mathbb{P}}_{\substack{(x_u, x_v)\sim\nu \\ x_w\sim U}} (x_v \geq x_u, x_w) + H(\nu) \right). \tag{2–2}$$

Now that designing Markovian rule ν is at our disposal, our hope is that some choice is powerful enough to capture the behaviour of D. From a non-rigorous view, D should display more or less "locality" as $D_1 = U$ does, so a Markovian rule might just be right.

Recall that D sets an incentive for high-score configurations (by a factor of 2 per gain). In order to simulate D locally, it is tempting to design a clear-cut rule $\nu(x_u, x_v) \propto 2^{\mathbf{1}[x_u < x_v]}$. But this chain is illegal in our framework as it doesn't have uniform stationarity. Worse still, the chain is non-reversible, leaving it painful to search for a stationary distribution. So instead we shall adopt a weaker approximation.

Definition 1. Define the *modular difference* between $a, b \in [0, 1]$ to be

$$a \ominus b := \begin{cases} a - b & a \geq b \\ a - b + 1 & a < b \end{cases}.$$

Pictorially, one may imagine bending the interval $[0, 1]$ to a ring so that $0 \equiv 1$ meet at one point. Then $a \ominus b$ is exactly the distance that we walk clockwise from b to a. Now we let

$$\nu(x_u, x_v) := \frac{1}{Z} \cdot 2^{[1-(x_v \ominus x_u)^2]/2} \tag{2–3}$$

where Z is a normalising constant. Note that ν is indeed doubly stochastic. This can be checked by a direct calculation, or more cleverly, by observing that $a \ominus b$ is a uniform random variable if any of a, b is, and thus $\mathbb{E}_{x_u\sim U}(\nu(x_u, x_v)) = \mathbb{E}_{x_v\sim U}(\nu(x_u, x_v)) = \mathbb{E}_{y\sim U}\left(2^{(1-y^2)/2}\right)/Z = Z/Z = 1$.

This definition simulates D (to some extent) by favouring $x_u < x_v$ over $x_u > x_v$. Imagine decreasing x_u gradually. At the moment it hits x_v, the density shall experience a surge. This behaviour matches D yet the form in (2–3) still looks arbitrary. We shall leave the full rationale to the chapter notes. Next we proceed to compute (2–2):

$$\begin{aligned}
\mathbb{P}(x_v \ge x_u, x_w) &= \int_0^1 \mathrm{d}x_v \int_0^{x_v} \nu(x_u, x_v)\ \mathrm{d}x_u \int_0^{x_v} 1\ \mathrm{d}x_w \\
&= \frac{1}{Z}\int_0^1 y\ \mathrm{d}y \int_0^y 2^{[1-(y-z)^2]/2}\ \mathrm{d}z \\
&= \frac{1}{Z}\int_0^1 y\ \mathrm{d}y \int_0^y 2^{(1-z^2)/2}\ \mathrm{d}z \\
&= \frac{1}{Z}\int_0^1 2^{(1-z^2)/2}\ \mathrm{d}z \int_z^1 y\ \mathrm{d}y \quad = \frac{1}{Z}\mathop{\mathbb{E}}_{z\sim U}\left[\frac{1-z^2}{2}2^{(1-z^2)/2}\right]
\end{aligned}$$

and

$$\begin{aligned}
H(\nu) &= \log Z - \frac{1}{Z}\mathop{\mathbb{E}}_{x_u,x_v\sim U}\left[\frac{1-(x_v\ominus x_u)^2}{2}\cdot 2^{[1-(x_v\ominus x_u)^2]/2}\right] \\
&= \log Z - \frac{1}{Z}\mathop{\mathbb{E}}_{z\sim U}\left[\frac{1-z^2}{2}2^{(1-z^2)/2}\right].
\end{aligned}$$

So it turns out that their summation simplifies to $\log Z$. Linking with (2–2) and (2–1) we have $Z(2) \ge 2^{\mathcal{F}(\mu)} \ge Z^{n-\sqrt{n}}$. The constant Z is larger than 1.2665 by numerical computation, so we obtain our first lower bound:

Theorem 1. $Z(2) > 1.2665^{n-\sqrt{n}} \approx 1.2665^n$ for the grid.

2.1.2 Towards optimal Markovian rule

Is there a better ν that beats the previous bound? More ambitiously, what does the optimal ν look like? In the following we develop a systematic way to answer both questions. In short, we decompose ν by a set of Fourier bases, approximate the target function (2–2) by quadratic terms, optimise over the frequency domain, and transform the solution back.

Define a set of Fourier bases $\{\phi_i(x)\}_{i\in\mathbb{Z}}$ by

$$\phi_i(x) := \begin{cases}\cos(2\pi i x) & i \ge 0\\ \sin(2\pi i x) & i < 0\end{cases}.$$

The following lemma summarises some properties that we shall make use of. The calculation is easy, so we omit the proof.

Lemma 2. For all $i, j \in \mathbb{Z}$,

(a) $\int_0^1 \phi_i(x)\ \mathrm{d}x = \mathbf{1}[i=0]$;

(b) $\int_0^1 \phi_i(x)\phi_j(x)\ \mathrm{d}x = \frac{1}{2}\cdot\mathbf{1}[i=j]$;

(c) $\int_0^1 y\cos(2\pi i y)\ \mathrm{d}y = \frac{1}{2}\cdot\mathbf{1}[i=0]$ and $\int_0^1 y\sin(2\pi i y)\ \mathrm{d}y = -\frac{1}{2\pi i}\cdot\mathbf{1}[i\ne 0]$.

Now assume ν can be decomposed into the Fourier bases, i.e.

$$\nu(x,y) = \sum_{i,j\in\mathbb{Z}} c_{ij}\cdot\phi_i(x)\phi_j(y) \tag{2–4}$$

where c_{ij} are the Fourier coefficients of ν; they are temporarily unknown. Under our framework ν should be doubly stochastic, so we have constraints $\forall x : 1 = \int_0^1 \nu(x,y)\ \mathrm{d}y$ and $\forall y : 1 = \int_0^1 \nu(x,y)\ \mathrm{d}x$. Plugging (2–4) in and using property (a), we have $\forall x, y : 1 = \sum_{i\in\mathbb{Z}} c_{i0}\phi_i(x) = \sum_{j\in\mathbb{Z}} c_{0j}\phi_j(y)$. Therefore, $c_{i0} = c_{0j} = 0$ for all $i, j \neq 0$, and $c_{00} = 1$. Equation (2–4) thus simplifies to

$$\nu(x,y) = 1 + \sum_{i,j\neq 0} c_{ij}\cdot\phi_i(x)\phi_j(y) \quad =: 1 + \delta(x,y) \tag{2–5}$$

Recall our target function is (2–2). Next we shall write $\mathbb{P}(y \geq x, z) + H(\nu)$ in terms of $\{c_{ij}\}$. Here we implicitly assume $(x,y) \sim \nu$ and $z \sim U$ and drop subscripts for brevity. The $\mathbb{P}(\cdot)$ part writes

$$\begin{aligned}
\mathbb{P}(y \geq x, z) &= \int_0^1 y\ \mathrm{d}y \int_0^y \nu(x,y)\ \mathrm{d}x \\
&= \frac{1}{3} + \sum_{i,j\neq 0} c_{ij} \int_0^1 y\phi_j(y)\ \mathrm{d}y \int_0^y \phi_i(x)\ \mathrm{d}x \\
&= \frac{1}{3} + \sum_{i,j\neq 0} \frac{c_{ij}}{2\pi i} \int_0^1 \begin{Bmatrix} y\cos(2\pi jy)\sin(2\pi iy) & i>0, j>0 \\ y\sin(2\pi jy)\sin(2\pi iy) & i>0, j<0 \\ y\cos(2\pi jy)(1-\cos(2\pi iy)) & i<0, j>0 \\ y\sin(2\pi jy)(1-\cos(2\pi iy)) & i<0, j<0 \end{Bmatrix} \mathrm{d}y.
\end{aligned}$$

Applying property (c) with much care, one may deduce

$$\mathbb{P}(y \geq x, z) = \frac{1}{3} + \sum_{i,j\neq 0} \frac{c_{ij}}{8\pi i} \begin{Bmatrix} \alpha(i,j) & i,j>0 \\ \beta(i,j) & ij<0 \\ \gamma(i,j) & i,j<0 \end{Bmatrix} \tag{2–6}$$

where

$$\alpha(i,j) = \frac{1}{\pi}\begin{cases} -\frac{1}{2i} & j = i \\ \frac{2i}{(j+i)(j-i)} & j \neq i \end{cases} \tag{2–7}$$

$$\beta(i,j) = -\mathbf{1}[i+j=0] \tag{2–8}$$

$$\gamma(i,j) = \frac{1}{\pi}\begin{cases} \frac{-3}{2i} & j = i \\ \frac{2j}{(j+i)(j-i)} - \frac{2}{j} & j \neq i \end{cases}. \tag{2–9}$$

The $H(\cdot)$ part contains a sticky logarithm, so we adopt a quadratic approximation $(1+t)\ln(1+t) \approx 1 + t + \frac{t^2}{2}$ (which originates from the Taylor expansion $\ln(1+t) = x - \frac{x^2}{2} + \frac{x^3}{3} - \cdots$). Since we use a heuristic here, the final ν we found might not be the true optimal. But arguably, it should be pretty close to the truth and behaves similarly.

$$\begin{aligned}-H(\nu) &= \frac{1}{\ln 2}\int_0^1\int_0^1 (1+\delta(x,y))\ln(1+\delta(x,y))\ \mathrm{d}x\ \mathrm{d}y \\ &\approx \frac{1}{\ln 2}\int_0^1\int_0^1 \delta(x,y)+\frac{\delta^2(x,y)}{2}\ \mathrm{d}x\ \mathrm{d}y \\ &= \frac{1}{2\ln 2}\int_0^1\int_0^1 \delta^2(x,y)\ \mathrm{d}x\ \mathrm{d}y \\ &= \frac{1}{2\ln 2}\sum_{i,j,s,t\neq 0} c_{ij}\cdot c_{st}\int_0^1 \phi_i(x)\phi_s(x)\ \mathrm{d}x\int_0^1 \phi_j(y)\phi_t(y)\ \mathrm{d}y \\ &= \frac{1}{8\ln 2}\sum_{i,j\neq 0} c_{ij}^2 \end{aligned} \tag{2–10}$$

where the third line uses property (a) and the last line uses (b). Note that we don't have such neat properties for ternary products, so a cubic approximation would be overly complicated.

It remains to express our target value by merging (2–6) with (2–10):

$$\mathbb{P}(y\geq x,z)+H(\nu)\approx \frac{1}{3}+\frac{1}{8}\sum_{i,j\neq 0}\left(\frac{\alpha|\beta|\gamma(i,j)}{\pi i}c_{ij}-\frac{1}{\ln 2}c_{ij}^2\right).$$

Observe that we could maximise the inner part for all $i,j\neq 0$ independently. Clearly the best choice for the coefficients is given by

$$c_{ij}^* := \frac{\ln 2}{2\pi i}\alpha|\beta|\gamma(i,j). \tag{2–11}$$

Conceptually we are done because this set of coefficients uniquely determines a rule ν^*. We could use numerical computation to recover the appearance of ν^* and to find out its target value. But it would be even nicer to figure out an analytic formula:

Theorem 3. The rule ν^* determined by $\{c_{ij}^*\}$ in (2–11) is exactly

$$\nu^*(x,y)=\begin{cases} cx^2-2cy^2+1-c/3 & x\geq y\\ cx^2-2cy^2+1-c/3+2cy & x<y\end{cases} \tag{2–12}$$

where $c:=\frac{\ln 2}{2}$; see Figure 2–2 for a plot. The corresponding target value is 1.268066... > 1.2680. Consequently, $Z(2)>1.2680^{n-\sqrt{n}}$ for grid.

Remark. The "optimal" coefficients $\{c_{ij}^*\}$ were derived by a heuristic (i.e. approximating KL without formal justification), so the resulting ν^* might not be the true optimal rule. However, the target value of ν^* stated above was evaluated by the proper KL formula rather than its quadratic approximation, and thus can be trusted.

Proof. The formula can be verified by sending the right-hand side of (2–12) back to frequency domain again. To avoid confusion let's call the right-hand side $\nu(x,y)$. By property (b), the orthogonality of Fourier bases, we could express its coefficients as $c_{ij}=\int_0^1 \nu(x,y)\phi_i(x)\phi_j(y)\ \mathrm{d}x\ \mathrm{d}y$. Computing the integral, one may find out that $c_{ij}=c_{ij}^*$ for all $i,j\in\mathbb{Z}$. Since a set of coefficients uniquely determines the rule, we must conclude that $\nu=\nu^*$. The rest of the theorem can be verified by numerical calculation. □

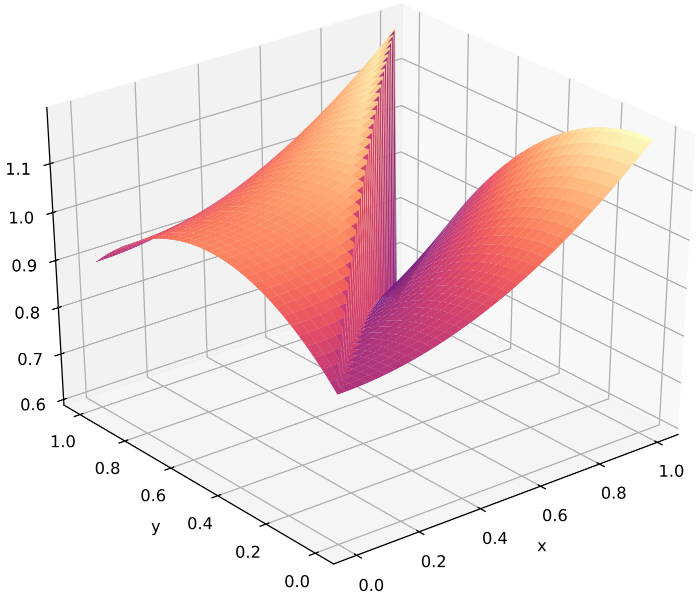

Figure 2–2 Plot of the "almost optimal" rule ν^*

2.2 Upper Bounds

Recall from (2–1) that $Z(2) = 2^{\mathcal{F}(D)}$ and we aim at understanding D. To break this monolithic distribution down to small pieces, we appeal to a lemma in information theory:

Lemma 4 (Shearer)**.** If $\mathcal{I} \subseteq 2^{[n]}$, and every $i \in n$ appears in at least k subsets in $\mathcal{I}$, then

$$H(X_1, X_2, \ldots, X_n) \le \frac{1}{k} \sum_{I \in \mathcal{I}} H(X_I).$$

Proof. We use a shorthand $X_{<a}$ to denote random vector $(X_1, X_2, \ldots, X_{a-1})$. For any $I \in \mathcal{I}$, order its members in increasing order: $i_1 < \cdots < i_t$. By chain rule,

$$H(X_I) = \sum_{j=1}^{t} H(X_{i_j} \mid X_{i_1}, \ldots, X_{i_{j-1}}) \quad \ge \sum_{j=1}^{t} H(X_{i_j} \mid X_{<i_j})$$

where the inequality follows because we have strengthened the conditions. Summing over all $I \in \mathcal{I}$, we note that for each $i \in [n]$, the term $H(X_i \mid X_{<i})$ appears at least k times in the right-hand side. Therefore,

$$\sum_{I \in \mathcal{I}} H(X_I) \ge k \sum_{i=1}^{n} H(X_i \mid X_{<i}) = kH(X_1, X_2, \ldots, X_n)$$

which completes the proof. □

Corollary 5. With D^v denoting the marginal of D on $\{v\} \cup \flat v$, we have

$$\mathcal{F}(D) \le \sum_{v \in V} \left(\mathop{\mathbb{E}}_{x \sim D^v} [S_v(x)] + \frac{1}{k} H(D^v) \right).$$

Proof. Set $\mathcal{I} := \{\{v\} \cup \flat v : v \in V\}$ and apply Shearer's lemma. □

With Corollary 5 at hand, we only have to discuss local marginal distributions of D. The following theorem gives a rough upper bound:

Theorem 6. For any $v \in V$ it holds that

$$\mathop{\mathbb{E}}_{x\sim D^v}[S_v(x)] + \frac{1}{k}H(D^v) \leq \frac{1}{k}\log\frac{2^k+k-1}{k}.$$

Consequently, $Z(2) \leq [(2^k+k-1)/k]^{n/k}$. In particular, $Z(2) < 1.4938^n$ when $k = 3$.

Proof. Let us disregard any characteristic of D^v and treat it as an arbitrary, unknown distribution κ on $\{v\} \cup \flat v$. We shall prove the theorem for *all* κ.

Write $\Omega_1 := \{x \in [0,1]^k : x_v \geq \max x_{\flat v}\}$ for the "scoring region", and likewise $\Omega_0 := \{x \in [0,1]^k : x_v < \max x_{\flat v}\}$ for the "losing region". Then $\mathrm{vol}(\Omega_1) = 1/k$ and $\mathrm{vol}(\Omega_1) = 1 - 1/k$. Denote $p := \mathbb{P}_{x\sim\kappa}(x \in \Omega_1) = \mathbb{E}_{x\sim\kappa}[S_v(x)]$. Our key observation is as follows:

Claim. If p is *fixed*, then $H(\kappa)$ could not exceed $-p\log(kp) - (1-p)\log\left(\frac{k(1-p)}{k-1}\right)$, the entropy when κ is uniform over Ω_1 and Ω_0 respectively.

It has an intuitive mental picture: κ is required to place mass p on region Ω_1 and the remaining mass $(1-p)$ on Ω_0. Since p is fixed, there's no interference between the strategies we choose for Ω_1 and Ω_0. To make the situation worst, we will of course spread p *uniformly* on Ω_1, and $(1-p)$ *uniformly* on Ω_0 at the same time.

The intuition is formalised below. Let $\kappa_1(x) := \frac{1}{p}\mathbf{1}[x \in \Omega_1]\kappa(x)$ and $\kappa_0(x) := \frac{1}{1-p}\mathbf{1}[x \in \Omega_0]\kappa(x)$ be two distributions supported on Ω_1 and Ω_0, respectively. Then

$$\begin{aligned} H(\kappa) &= -\int_{\Omega_1} p\kappa_1(x)\log(p\kappa_1(x))\,\mathrm{d}x - \int_{\Omega_0} p\kappa_0(x)\log(p\kappa_0(x))\,\mathrm{d}x \\ &= pH(\kappa_1) + (1-p)H(\kappa_0) + h_b(p) \end{aligned}$$

Because p is fixed, we can maximise $H(\kappa_1)$ and $H(\kappa_0)$ independently, and the maximizer is of course uniform distributions on both regions. The claim follows by some calculations.

Using the claim, the problem converts to a single-variate optimisation:

$$\max_{p\in[0,1]} \quad p - \frac{1}{k}\left(p\log(kp) + (1-p)\log\left(\frac{k(1-p)}{k-1}\right)\right).$$

Let us call the target function $f(p)$. By basic analysis,

$$f'(p) = 1 - \frac{1}{k}\left(\log(k-1) + \log\frac{p}{1-p}\right),$$

so $f(p)$ has a unique maximizer $p^* = 2^k/(2^k+k-1)$. With some calculation one may simplifies $f(p^*)$ to $\frac{1}{k}\log[(2^k+k-1)/k]$, proving the core of the theorem. The consequence follows directly from Corollary 5. □

Remark. Our proof dismisses the properties of D^v altogether, leading to an unrealistically large bound when $k \to \infty$. We should not even presume that the local distribution κ could be "assembled" into a global one. When $k \to \infty$ then our $p^* \to 1$, so vertex v almost surely scores under κ. Suppose there were a global distribution, say μ, such that $\mu^v = \kappa$ everywhere, then $\mathbb{E}_{x\sim\mu} S(x) \approx n$. But this is ridiculous for nearly all graphs.

Actually Theorem 6 has a one-line proof by Jensen's inequality; see chapter notes for details. We retain the above proof for two reasons. First, information theory is an elegant tool and also a recurring theme of this thesis. Second, as we will see shortly, we may bake our knowledge of D^v into the argument with little modification, whereas Jensen's inequality falls short of such generalisation.

Definition 2. We call a digraph *vertex transitive* if for all $u, v \in V$ there is an automorphism $\phi : V \to V$ such that $\phi(u) = v$.

The definition captures a strong notion of graph symmetry: any particular vertex can be identified with any other vertex. The grid is such an example. It should be clear that any vertex transitive digraph must be regular.

Lemma 7. For all λ and any vertex transitive digraph, x_v is uniformly distributed under D_λ for all v.

Proof. Let's work with permutations first. Denote by Ω be the collection of all permutations on $[n]$. Then the Boltzmann distribution D_λ naturally induces a distribution on Ω: for $\pi \in \Omega$ we have $\mathbb{P}(\pi) \propto 2^{S(\pi)}$.

Now we claim that π_v is uniformly distributed on $[n]$. That is, v does not favour any particular position in the random ordering π. Suppose it does favour i over j, i.e. $\mathbb{P}(\pi_v = i) > \mathbb{P}(\pi_v = j)$, then there is some vertex $u : \mathbb{P}(\pi_u = i) < \mathbb{P}(\pi_u = j)$ since $\sum_w \mathbb{P}(\pi_w = i) = \sum_w \mathbb{P}(\pi_w = j) = 1$. But this is absurd, as the graph is vertex transitive and thus u and v should be symmetric under the Boltzmann distribution. (More formally, we may relabel G by ϕ so that $\phi(u) = v$ and the resulting graph $G' \cong G$. On one hand π_u^G and $\pi_v^{G'}$ are identically distributed since D_λ doesn't care about labels; on the other hand π_v^G and $\pi_v^{G'}$ are identically distributed since $G \cong G'$.)

Next we return to continuous space. By the discussion above, the marginal of x_v can be written into $D_\lambda(x_v) = \frac{1}{n}\sum_{i=1}^n D_\lambda(x_v \mid \pi_v = i)$. We claim that $D_\lambda(x_v \mid \pi_v = i)$ does not depend on i (which implies $D_\lambda(x_v)$ is a constant, and it must be 1). To see the claim, write

$$D_\lambda(x_v \mid \pi_v = i) = \frac{1}{Z} \mathop{\mathbb{E}}_{x \sim U} \left(2^{S(x)} \;\middle|\; x_v, \pi_v = i\right).$$

But the condition x_v is essentially "shadowed" by the condition $\pi_v = i$. Regardless of the value of x_v, there will always $i - 1$ smaller vertices and $n - i$ bigger vertices, and every legal orderings are equally probable since the conditional space does not discriminate among vertices. Therefore, the conditional expectation can be realised equivalently by $\mathbb{E}_{\pi\sim U}(2^{S(\pi)} \mid \pi_v = i)$, which doesn't depend on x_v. □

Theorem 8. In a vertex transitive digraph, for any $v \in V$ it holds that

$$\mathop{\mathbb{E}}_{x\sim D^v}[S_v(x)] + \frac{1}{k}H(D^v) \le \frac{1}{k}\int_0^1 \log\left((2^k - 1)t^{k-1} + 1\right)\, \mathrm{d}t \qquad =: A_k.$$

Consequently, $Z(2) \le 2^{nA_k}$ for these digraphs. In particular, $Z(2) < 1.3927^n$ for grid.

Proof. Let us peel one more layer off our target:

$$\begin{aligned}
\mathop{\mathbb{E}}_{x\sim D^v}(S_v) + \frac{1}{k}H(D^v) &= \mathop{\mathbb{E}}_{x_v\sim U}\left[\mathop{\mathbb{E}}_{x_{\flat v}}(S_v \mid x_v)\right] + \frac{1}{k}H(x_{\flat v} \mid x_v) \\
&= \int_0^1 \mathop{\mathbb{E}}_{x_{\flat v}}(S_v \mid x_v = t)\, \mathrm{d}t + \frac{1}{k}\int_0^1 H(x_{\flat v} \mid x_v = t)\, \mathrm{d}t \\
&= \int_0^1 \left(\mathop{\mathbb{E}}_{x_{\flat v}}(S_v \mid x_v = t) + \frac{1}{k}H(x_{\flat v} \mid x_v = t)\right)\, \mathrm{d}t.
\end{aligned}$$

where the first line follows from Lemma 7 (noting that $H(x_v) = 0$). Our strategy is to maximise the inner part for *each* t and then integrate. The procedure is in analogue with Theorem 6. Let $\Omega_1 = \Omega_1(t) := \{x_{\flat v} : t \ge \max x_{\flat v}\}$ and $\Omega_0 = \Omega_0(t)$ be its complement.

Clearly $\mathrm{vol}(\Omega_1) = t^{k-1}$ and $\mathrm{vol}(\Omega_0) = 1 - t^{k-1}$. Denote $p = p(t) := \mathbb{P}(x_{\flat v} \in \Omega_1)$ and we shall obtain a single-variate optimisation. Solving it gives $p^*(t) = \frac{2^k t^{k-1}}{(2^k-1)t^{k-1}+1}$ and $f(p^*(t)) = \log\left((2^k - 1)t^{k-1} + 1\right)/k$. The theorem is proved by integration. □

Remark. It's hard to find an analytic formula for A_k, but we can say more about its asymptotic behaviour. For simplicity denote $g_k(t) := f(p^*(t))$. We observe that $\forall t \in [0, 1] : g_k(t) \le g_{k-1}(t)$, so A_k is monotonically decreasing with k (which is better than Theorem 6). On the other hand, $\forall k \ge 2$ we have $g_k(0) = 0$, $g_k(1) = 1$, and $g_k(t)$ is monotonically increasing in t. Also note that if $k \to \infty$ then $g_k(t)$ is basically zero for $t < 1/2$, and is almost concave for $t > 1/2$ (more precisely, for $t > 2^{O(\log k/k)}/2$). This means $g_k(t)$ is asymptotically *above* the function $t \mapsto (2t-1) \cdot \mathbf{1}[t > 1/2]$, so $A_k \ge \int_{1/2}^{1} (2t-1)\ \mathrm{d}t = 1/4$. This is not really satisfying because we wish the number converged to 0.

2.3 Notes and References

The variation view given by (2–1) is termed *Gibbs variational principle* in statistical physics; see Section 6.9.1 in [8].

The definition of "canonical rule" (2–3) is not at all arbitrary. There are other plausible choices such as $\nu(x_u, x_v) \propto 2^{x_u \ominus x_v} = 2^{1-(x_v \ominus x_u)}$, but our current choice is *the best* if ν only involves $x_v \ominus x_u$. More precisely, if ν writes $\nu(x_u, x_v) = f(x_v \ominus x_u)/Z$ for some function f and normalising constant Z, then (2–2) is maximised if and only ν equals (2–3). Indeed, repeating our computation,

$$\begin{aligned}
\mathbb{P}(x_v \ge x_u, x_w) &= \frac{1}{Z}\int_0^1 y\ \mathrm{d}y \int_0^y f(y-z)\ \mathrm{d}z = \frac{1}{Z}\int_0^1 y\ \mathrm{d}y \int_0^y f(t)\ \mathrm{d}t \\
&= \frac{1}{Z}\int_0^1 f(t)\ \mathrm{d}t \int_t^1 y\ \mathrm{d}y = \frac{1}{Z}\int_0^1 \frac{1-t^2}{2} f(t)\ \mathrm{d}t, \\
H(\nu) &= \log Z - \frac{1}{Z} \mathop{\mathbb{E}}_{x_u, x_v \sim U}[f(x_v \ominus x_u) \log f(x_v \ominus x_u)] \\
&= \log Z - \frac{1}{Z} \mathop{\mathbb{E}}_{t \sim U}[f(t) \log f(t)] = \log Z - \frac{1}{Z}\int_0^1 f(t) \log f(t)\ \mathrm{d}t.
\end{aligned}$$

Since $Z = \mathbb{E}_{x_u, x_v \sim U}[f(x_v \ominus x_u)] = \mathbb{E}_{t\sim U}[f(t)]$, we see that $f(t)/Z$ is a density function. Keep it in mind and adds up the equations, we eventually find

$$\mathbb{P}(x_v \ge x_u, x_w) + H(\nu) = \int_0^1 \frac{f(t)}{Z} \log \frac{2^{(1-t^2)/2}}{f(t)/Z}\ \mathrm{d}t.$$

But this can be interpreted as a negative KL divergence with constant shift. To be explicit, introduce normalising constant $\tilde{Z} := \mathbb{E}_{t\sim U}(2^{(1-t^2)/2})$. Then the right-hand side equals $\log \tilde{Z} - \mathrm{KL}(f/Z \| 2^{(1-t^2)/2}/\tilde{Z})$, which is maximised when $f(t) = 2^{(1-t^2)/2} Z/\tilde{Z}$.

Our last note is an alternative proof of Theorem 6 by Jensen's inequality:

$$\begin{aligned}
2^{k\,\mathbb{E}_{x\sim D^v}[S_v(x)] + H(D^v)} &= 2^{\mathbb{E}_{x\sim D^v}[kS_v(x) - \log D^v(x)]} \\
&\le \mathop{\mathbb{E}}_{x\sim D^v}\left(2^{kS_v(x) - \log D^v(x)}\right) \\
&= \mathop{\mathbb{E}}_{x\sim D^v}\left(2^{kS_v(x)} \frac{1}{D^v(x)}\right) \\
&= \mathop{\mathbb{E}}_{x\sim U}\left(2^{kS_v(x)}\right) = \frac{1}{k}2^k + \left(1 - \frac{1}{k}\right)2^0 = \frac{2^k + k - 1}{k}.
\end{aligned}$$

Chapter 3 Combinatorial Embedding and Spreading

This chapter accommodates two combinatorial ideas which assist us sharpening previous results. The first idea, *gadget embedding*, specialises in lower bounds. It packs tree-like structure into graph to reveal richer information of the model. The second idea, *weight spreading*, shows its power in upper bound contexts. It skillfully pairs the states and provides us with valuable insights into D_λ. These ideas are motivated by considerations independent of the variational view in Chapter 2.

3.1 Gadget Embedding

3.1.1 Motivation

We have seen in Chapter 1 an easy lower bound $\mathbb{E}_x(2^S) \ge 2^{\mathbb{E}_x(S)} = 2^{n/k}$ by Jensen's inequality. Chapter 2 took a very different route for lower bounds. Now let us step back and ask: Is it possible to refine Jensen's inequality directly?

$\mathbb{E}_x(2^S) \ge 2^{\mathbb{E}_x(S)}$ is tight if and only if $S(x)$ is a constant. More generally, it is "near tight" if $S(x)$ is concentrated. What if we artificially enforce $S(x)$ to concentrate more than usual? Namely, if we appropriately define some random variable y, then the *conditional* distribution of $S(x)$ under y may be more concentrated. Therefore, when we write

$$\mathop{\mathbb{E}}_{x}(2^S) = \mathop{\mathbb{E}}_{y}\left(\mathbb{E}(2^S \mid y)\right) \ge \mathop{\mathbb{E}}_{y}\left(2^{\mathbb{E}(S|y)}\right) = \mathop{\mathbb{E}}_{y}\left(2^{\sum_v \mathbb{E}(S_v|y)}\right) \tag{3–1}$$

we expect the Jensen's inequality here performs better.

The main technical challenge is choosing a nice conditioning variable y. It's again a game about balance. To one extreme, one may choose $y := x$ so that $(S \mid y)$ is a perfect constant. However, evaluating $\mathbb{E}_y(\cdot)$ would be as tough as before. Below we consider some feasible choices of y, ordered by increasing complexity, in the context of two-dimensional grid.

3.1.2 Horizontal embedding

Recall that the modular difference $a \ominus b$ is the clockwise distance from b to a on a circle of circumference 1. For each horizontal edge uv (except the last column) in the grid we define $y_v := x_v \ominus x_u$. Collect them all in a vector y.

Lemma 9. For the y defined above,

- The variables y_v's are uniformly and independently distributed on $[0, 1]$.
- For any particular $v \in V$, the variables y and x_v are independent.

Proof. As we saw in Chapter 2, $a \ominus b$ is uniform on $[0, 1]$ so long as a or b is uniform on $[0, 1]$. So by definition, the y_v's must be uniform. To show independence, we distinguish two cases. If v_1 and v_2 are not adjacent horizontally, then by definition y_{v_1} and y_{v_2} must be independent. If they do touch horizontally, say $v_1 \in \flat v_2$, then conditioning on y_{v_2} (i.e.

$x_{v_2} \ominus x_{v_1}$) shall not bias $y_{v_1} = x_{v_1} \ominus x_u$, as we still have a "fresh" uniform variable x_u. A similar argument extends to any subset of y_v's.

Finally, conditioned on any particular x_v, the distribution of y remain the same for a similar reason. □

We provide an alternative argument next since it illustrates a technique that we use frequently and implicitly.

Alternative proof. Let us pick one vertex from each row and call them U. Consider the following probability space: Sample $x_U \in [0,1]^{\sqrt{n}}$ uniformly. Then sample y uniformly which represents the modular differences for horizontal edges. Finally, based on the knowledge of x_U and y, fill in the correct values of $x_{V \setminus U}$ (which are uniquely determined).

It is easy to see that x generated in this way are uniformly distributed on $[0,1]^n$. Therefore, the probability space is *equivalent* to the original one. But now by construction, y is uniformly distributed and, furthermore, independent of x_U (so we actually proved more). □

Armed with these properties, we return to (3–1) and evaluate $\mathbb{E}(S_v \mid y)$ in the exponent. Call v's left neighbour u and its lower neighbour w, then

$$\mathbb{E}(S_v \mid y) = \mathop{\mathbb{E}}_{x_v \mid y} \mathbb{E}(S_v \mid y, x_v) = \mathop{\mathbb{E}}_{x_v \sim U} \mathbb{E}(S_v \mid y, x_v) = \mathop{\mathbb{E}}_{x_v \sim U} \mathbb{P}(x_u, x_w \le x_v \mid y, x_v)$$

where the second step uses the independence of x_v and y. Conditioned on (y, x_v), the value $x_u = x_v \ominus y_v$ is determined, and the value x_w is dangling independently. Therefore,

$$\begin{aligned} \mathbb{P}(x_u, x_w \le x_v \mid y, x_v) &= \mathbb{P}(x_v \ominus y_v \le x_v \mid y, x_v)\, \mathbb{P}(x_w \le x_v \mid y, x_v) \\ &= \mathbf{1}[y_v \le x_v] \cdot x_v \end{aligned}$$

and thus

$$\mathbb{E}(S_v \mid y) = \int_0^1 \mathbf{1}[y_v \le x_v] \cdot x_v \, \mathrm{d}x_v = \int_{y_v}^1 x_v \, \mathrm{d}x_v = \frac{1 - y_v^2}{2}.$$

Note that only the local term y_v takes effect though we condition on the entire y.

Finally, putting together with (3–1), we deduce

$$\begin{aligned} \mathop{\mathbb{E}}_{x}(2^S) &\ge \mathop{\mathbb{E}}_{y} \left(2^{\sum_v (1 - y_v^2)/2} \right) \\ &= \prod_v \mathop{\mathbb{E}}_{y_v} \left(2^{(1 - y_v^2)/2} \right) \\ &= \left(\int_0^1 2^{(1-t^2)/2} \, \mathrm{d}t \right)^{n - \sqrt{n}} > 1.2665^{n - \sqrt{n}}, \end{aligned}$$

where the second line uses independence of y_v's. The result matches the one we derived from canonical Markovian rule in Chapter 2, but here we have saved a significant amount of calculations.

3.1.3 Tree embedding

The horizontal embedding trick hints about a generalisation:

> *Can we choose a subgraph T in the grid such that, when revealing modular differences for all $e \in T$, our previous argument still gets through?*

We have just seen that T could be taken as $\sqrt{n}$ parallel horizontal paths. The bottleneck for this choice lies in the "dangling" lower neighbour w. Since horizontal edges reveal no information vertically, x_w is out of control and reduce the concentration of S_v. To overcome the obstacle, we hope to include vertical edges in T as well.

Not all choices of T are legal. Cyclic subgraphs (in the undirected sense) are crossed out immediately, since a cycle would deprive the edges of independence. For instance, suppose ab, ac, bd, cd constitute a cycle, then we have the identity $(x_d \ominus x_b) \oplus (x_b \ominus x_a) = (x_d \ominus x_c) \oplus (x_c \ominus x_a)$ where $\oplus$ is naturally the inverse operation of $\ominus$. Clearly the four edges (more precisely, the corresponding modular differences) cannot be independent.

But even trees/forests could be illegal. Suppose there is a long path $v_1, \ldots, v_r \in T$ such that $v_r \in \flat v_1$ in the grid (not in T, of course), then S_v will implicitly depend on all edges on this path. In this case, $\mathbb{E}(S_{v_i} \mid y)$ tangles with each other for all $1 \le i \le r$, making it impossible to factor the expectation in the final step.

Figure 3–1 shows a legal T. It is a forest consisting of trees running in diagonal direction. We may replay Lemma 9 easily for this T, so we jump directly to calculations.

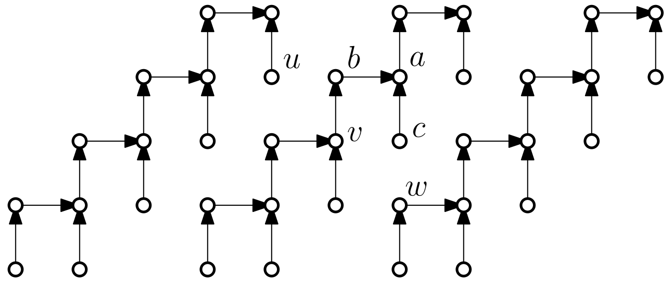

Figure 3–1 Tree embedding example. Some vertices on the grid are not shown for clarity.

There are only three types of vertices, labelled a, b and c respectively in the figure. With little modification, we could derive

$$\begin{aligned}
\mathbb{E}(S_a \mid y) &= 1 - \max(y_{ba}, y_{ca}) \\
\mathbb{E}(S_b \mid y) &= \frac{1 - y_{vb}^2}{2} \\
\mathbb{E}(S_c \mid y) &= \frac{1 - (y_{vb} \oplus y_{ba} \ominus y_{ca})^2}{2}.
\end{aligned}$$

The last one is a bit tricky. Conditioned on (y, x_c), value x_w is dangling, but x_v is determined through a path of length three! Therefore, we actually implicitly reveal an edge y_{vc}, which equals $y_{vb} \oplus y_{ba} \ominus y_{ca}$.

Summing these up, we get

$$\begin{aligned}
\mathbb{E}(S_a + S_b + S_b \mid y) &= 2 - \max(y_{ba}, y_{ca}) - \frac{y_{vb}^2 + (y_{vb} \oplus y_{ba} \ominus y_{ca})^2}{2} \\
&=: f(y_{ba}, y_{ca}, y_{vb}).
\end{aligned}$$

The point is, for this vertex group (a, b, c), only the conditions y_{ba}, y_{ca} and y_{vb} take effect. Moreover, the conditions on which different groups depend are disjoint. Therefore, we could again exploit independence of y, now at the scale of groups:

$$\begin{aligned}
\mathop{\mathbb{E}}_{x}(2^S) &\geq \mathop{\mathbb{E}}_{y}\left(\exp_2\left\{\sum_{(a,b,c)} f(y_{ba}, y_{ca}, y_{vb})\right\}\right) \\
&= \prod_{(a,b,c)} \mathbb{E}\left(\exp_2\left\{f(y_{ba}, y_{ca}, y_{vb})\right\}\right) \\
&= \left(\int_0^1\int_0^1\int_0^1 2^{f(r,s,t)}\,\mathrm{d}r\,\mathrm{d}s\,\mathrm{d}t\right)^{(n-O(\sqrt{n}))/3}
\end{aligned}$$

where the exponent counts the number of vertex groups. We subtract an $O(\sqrt{n})$ because some boundary vertices might not be covered by T. Using numerical computation, it yields the sharpest lower bound so far:

Theorem 10. $Z(2) > 1.2702^{n-O(\sqrt{n})}$ for the grid.

As is clear from our derivation, type a vertices are where the improvement took place. Through careful construction, our current T contains denser information than horizontal paths.

3.1.4 Design issues

It's time to briefly address the design issues concerning gadget embedding. Above all we remind the reader that, given unlimited computational time, one could always choose $T := G$ and handle the task by computers. People name it as hand-waving. So we'd better limit our computational power to d-dimensional numerical integration at most, say.

Fix a subgraph T. For vertex $v \in V$, define its dependency list $L_T(v)$ to contain all edges in T that (i) points to v; or (ii) lies on a path which connects a pair in $\flat v$. Here $\flat v$ is taken in the original graph. Let $\mathcal{L}_T := \{L_T(v) : v \in T\}$. We say two lists in $\mathcal{L}_T$ are *connected* if they intersect, and define *connected components* of $\mathcal{L}_T$ in the obvious way. Under this formulation, the story is all about:

> *Design a forest T such that every connected component in $\mathcal{L}_T$ contains at most d edges.*

This in principle gives an out-of-the-box lower bound method for general k, though with no performance guarantee.

3.2 Markovian Sampling Meets Embedding

At this point, we see a chance of consolidating two threads into one. The Markovian scheme in Chapter 2 sampled values horizontally, so vertical information were lost. Can we imitate the gadget embedding trick and do Markovian sampling on forests? Since we already have a stronger rule ν^* than modular difference, we expect that such scheme when armed with ν^* would produce superior lower bounds.

However, there is a twist in defining Markov chains on forests. Surely, one could first sample the roots and then work his way down to leaves by pairwise sampling via ν^*. But this leads to extremely complicated, or even undesirable, marginals at some pairs. Take the forest in Figure 3–1 as example, the marginal distribution of (x_c, x_v) would be overly tiring to write. Furthermore, it does not resemble ν^* at all and hardly reveals information about S_c. The overall effect is not very satisfactory, giving only a lower bound of about 1.2694^n.

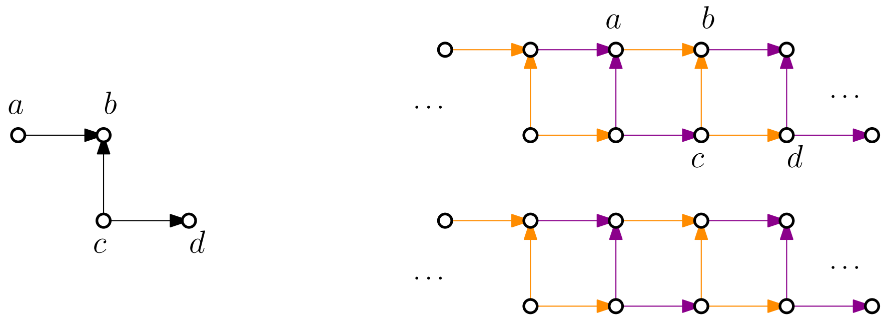

Figure 3–2 A snake and its friends.

Rather, we develop a technique that ensures correct pair marginals (at some price). Let $\delta(x,y) := \nu^*(x,y) - 1$. Since ν^* is doubly stochastic, we have $\forall x : \int_0^1 \delta(x,y)\ \mathrm{d}y = 0$ and $\forall x : \int_0^1 \delta(x,y)\ \mathrm{d}y = 0$.

Call the gadget on the left in Figure 3–2 a *snake*. As shown in the right part, we cascade snakes horizontally into chunks and embed them in the grid. Again, boundaries are not covered. Fix a small number $\epsilon > 0$, we define a distribution on snake (a,b,c,d) by

$$\nu(x_a,x_b,x_c,x_d) := 1 + \epsilon \cdot (\delta(x_a,x_b) + \delta(x_c,x_b) + \delta(x_c,x_d))$$

and extend it to a distribution on the entire graph:

$$\nu(x) := \prod_{(a,b,c,d)} \nu(x_a,x_b,x_c,x_d)$$

where the product runs over all snakes. Note if $\epsilon < \frac{1}{2\ln 2}$ then ν is positive. The lemma below justifies why it is a legal extension.

Lemma 11. $\nu(x)$ is a legal distribution. Moreover, for any embedded snake (a,b,c,d), the marginal of (x_a,x_b,x_c,x_d) is indeed $\nu(x_a,x_b,x_c,x_d)$, i.e. the distribution on snake before extension. This also justifies our usage of symbol ν for the global distribution.

Proof. For edge $e = uv$ we abbreviate $\delta_e := \delta(x_u,x_v)$. Expanding the definition of $\nu(x)$ yields

$$\nu(x) = 1 + \epsilon \sum_e \delta_e + \epsilon^2 \sum_e \sum_{e'} \delta_e \delta_{e'} + \cdots$$

where e, e', etc. pick edges from distinct snakes. In other words, each term (e.g. $\delta_e\delta_{e'}$) can be regarded as a subgraph that contains at most one edge from each snake. Note that such subgraphs never contain a cycle, by the design of our embedding.

To see $\nu(x)$ is a legal distribution, we claim that $\int \nu(x)\ \mathrm{d}x = 1$. Let us do the integration for each term separately. The integration of the unity term gives 1 certainly. For all others, say the term corresponding to subgraph T, we may always start integrating at a leaf of T, which returns a zero.

The argument for marginal on a snake (a,b,c,d) is similar. We observe that, except for terms 1, δ_{ab}, δ_{cb} and δ_{cd}, all other terms must contain an "exposed" vertex $v \notin \{a,b,c,d\}$. Therefore, when we condition on (x_a,x_b,x_c,x_d), we may always start integrating at v and the term vanishes. Only $1 + \epsilon(\delta_{ab} + \delta_{cb} + \delta_{cd})$ survives, as promised. □

As a corollary, we see that the marginal of (x_a,x_b) is $1 + \epsilon\delta_{ab}$, and similarly for (x_c,x_b), (x_c,x_d). If we could take $\epsilon = 1$, then these pair marginals are guaranteed the optimal rule

ν^*. Unfortunately, ϵ cannot be 1, otherwise ν becomes negative somewhere. Nevertheless, for $\epsilon = 0.7$, say, the pair distribution still nicely mimics ν^*.

Recall from last chapter the functional $\mathcal{F}(\nu) := \mathbb{E}_{x\sim\nu}(S) + H(\nu)$. By routine calculation and ignoring boundaries,

$$H(\nu) = \sum_{(a,b,c,d)} H(\nu(x_a, x_b, x_c, x_d)) = \frac{n}{2} H(\nu(x_a, x_b, x_c, x_d)). \tag{3–2}$$

Using the fact that $(1+t)\ln(1+t) \le t + \frac{t^2}{2} - \frac{t^3}{6} + \frac{t^4}{3}$ and setting $t := \epsilon(\delta_{ab} + \delta_{cb} + \delta_{cd})$,

$$\begin{aligned} -H(\nu(x_a, x_b, x_c, x_d)) &= \frac{1}{\ln 2} \mathop{\mathbb{E}}_{x_a,x_b,x_c,x_d\sim U} ((1+t)\ln(1+t)) \\ &\le \frac{1}{\ln 2} \mathop{\mathbb{E}}_{x_a,x_b,x_c,x_d\sim U} \left(t + \frac{t^2}{2} - \frac{t^3}{6} + \frac{t^4}{3}\right) \\ &= \frac{1}{\ln 2} \mathop{\mathbb{E}}_{x_a,x_b,x_c,x_d\sim U} \left(\frac{t^2}{2} - \frac{t^3}{6} + \frac{t^4}{3}\right). \end{aligned}$$

An impatient reader may evaluate it via numerical computation. Here we perform one more step of simplification, taking $\mathbb{E}(t^4)$ as an example:

$$\mathbb{E}(t^4) = \epsilon^4\, \mathbb{E}\left((\delta_{ab} + \delta_{cb} + \delta_{cd})^4\right).$$

Once again, expanding the product will give us several terms, each corresponding to a *multiset* of four edges in snake (a, b, c, d). If a term contains exposed vertices, then it must vanish. The only surviving terms are $6(\delta_{ab}\delta_{cb})^2, 6(\delta_{cb}\delta_{cd})^2, 6(\delta_{ab}\delta_{cd})^2$ (choosing pairs twice) and $\delta_{ab}^4, \delta_{bc}^4, \delta_{cd}^4$ (choosing an edge four times). Note that some terms give identical expectations. After cleaning up, we obtain

$$\mathbb{E}(t^4) = \epsilon^4 \left(3\,\mathbb{E}(\delta_{ab}^4) + 6\left[\mathbb{E}(\delta_{ab}^2)\right]^2 + 6\,\mathbb{E}\left[(\delta_{ab}\delta_{cb})^2\right] + 6\,\mathbb{E}\left[(\delta_{cb}\delta_{cd})^2\right]\right).$$

Similarly,

$$\mathbb{E}(t^3) = 3\epsilon^3\,\mathbb{E}(\delta_{ab}^3), \qquad \mathbb{E}(t^2) = 3\epsilon^2\,\mathbb{E}(\delta_{ab}^2).$$

Putting all these into (3–2), we have

$$\begin{aligned} H(\nu) \ge \frac{-n\epsilon^2}{4\ln 2} \Big\{ & 3\,\mathbb{E}(\delta_{ab}^2) - \epsilon\,\mathbb{E}(\delta_{ab}^3) + \\ & 2\epsilon^2 \left[\mathbb{E}(\delta_{ab}^4) + 2\left[\mathbb{E}(\delta_{ab}^2)\right]^2 + 2\,\mathbb{E}\left[(\delta_{ab}\delta_{cb})^2\right] + 2\,\mathbb{E}\left[(\delta_{cb}\delta_{cd})^2\right]\right]\Big\}. \end{aligned} \tag{3–3}$$

The calculation of $\mathbb{E}_{x\sim\nu}(S)$ is much easier. Observe that there are only two types of vertices, namely b and c, in the embedding. We may find

$$\begin{aligned} \mathbb{E}(S_b) &= \frac{1}{3} + 2\epsilon \int_0^1 y\ \mathrm{d}y \int_0^y \delta(x,y)\ \mathrm{d}x = \frac{1}{3} + \frac{\epsilon\ln 2}{18} \\ \mathbb{E}(S_c) &= \frac{1}{3} + \epsilon \int_0^1 y\ \mathrm{d}y \int_0^y \delta(x,y)\ \mathrm{d}x = \frac{1}{3} + \frac{\epsilon\ln 2}{36}, \end{aligned}$$

which implies

$$\mathbb{E}(S) = \left(\frac{1}{3} + \frac{\epsilon\ln 2}{24}\right) n \tag{3–4}$$

Finally, we take $\epsilon := 0.7213 < \frac{1}{2\ln 2}$ and evaluate (3–3) (3–4). The numbers are greater than $-0.008032n$ and $0.354166n$, respectively. Therefore, $\mathcal{F}(\nu) > 0.346134n$ and the theorem below follows from the variational view (2–1):

Theorem 12. $Z(2) > 1.2711^n$ for the grid.

3.3 Weight Spreading

The weight spreading idea is motivated by a well-known relation in statistical physics.

Definition 3. For D_λ on a specific digraph we define its *occupancy* as $\alpha(\lambda) := \frac{\mathbb{E}_{D_\lambda}(S)}{n} = \frac{1}{n}\sum_{v\in V}\mathbb{E}_{D_\lambda}(S_v)$, that is the expected score per vertex under D_λ. Note that for vertex transitive graphs, $\alpha(\lambda)$ equals the scoring probability of any vertex.

Lemma 13. $\ln Z(\lambda) = n\int_1^\lambda \frac{\alpha(\lambda)}{\lambda}\,\mathrm{d}\lambda$.

Proof. By definition,

$$\begin{aligned} n\frac{\alpha(\lambda)}{\lambda} &= \int_{[0,1]^n} S(x)\frac{\lambda^{S(x)-1}}{Z(\lambda)}\,\mathrm{d}x \\ &= \frac{1}{Z(\lambda)}\int_{[0,1]^n}\frac{\mathrm{d}}{\mathrm{d}\lambda}\lambda^{S(x)}\,\mathrm{d}x \\ &= \frac{1}{Z(\lambda)}\frac{\mathrm{d}}{\mathrm{d}\lambda}\int_{[0,1]^n}\lambda^{S(x)}\,\mathrm{d}x = \frac{Z'(\lambda)}{Z(\lambda)} = (\ln Z(\lambda))'. \end{aligned}$$

The exchange in the third line is valid because we may partition the space $[0,1]^n$ into $n!$ regions according to the ordering induced by x. Points inside each region has identical scores, so $\int_{[0,1]^n}\lambda^{S(x)}\,\mathrm{d}x = \sum_\pi \frac{\lambda^{S(\pi)}}{n!}$ and the exchange is justified. After all, our model is discrete at its core. The lemma follows by integrating from 1 to λ and noting $Z(1) = 1$. □

Based on this relation, we turn to upper bound $\alpha(\lambda)$ for all $\lambda \in [1,2]$. To this end, we pair a scoring state with a non-scoring state and compare their weights.

Theorem 14. For any $\lambda \geq 1$ and vertex $v \in V$ we have

$$\mathop{\mathbb{E}}_{D_\lambda}(S_v) \leq \frac{1}{1+(k-1)/\lambda^{k-1}}.$$

As a consequence,

$$Z(\lambda) \leq \left(\frac{\lambda^{k-1}+k-1}{k}\right)^{n/(k-1)}.$$

Proof. During the proof we will solely work in D_λ, so we drop all subscripts for brevity. Denote $\Omega_1 := \{x \in [0,1]^n : x_v \geq \max x_{\flat v}\}$ and $\Omega_0 := [0,1]^n \setminus \Omega_1$. Then $\mathbb{E}(S_v) = \mathbb{P}(x \in \Omega_1)$.

For each $u \in \flat v$ we define a mapping $\phi_u : \Omega_1 \to \Omega_0$ as follows. Given state $x \in \Omega_1$ as input, ϕ_u outputs the same state but with x_v and x_u swapped. The output indeed lies in Ω_0 since vertex v now has a smaller value than its child u.

Note that the images of these mappings partition the set Ω_0. To see this more clearly, write out $\mathrm{Im}(\phi_u) = \left\{x \in [0,1]^n : x_u \geq \max x_{\flat v\cup\{v\}}\right\}$ and observe. The $\mathrm{Im}(\phi_u)$'s are disjoint and their union equals Ω_0. Figure 3–3 illustrates the idea.

Next we compare the scores before and after applying ϕ_u. After its application, vertex v surely loses score. The other $k-2$ parents of u might also lose scores since u gets a bigger assignment. Yet u itself might benefit. Finally, $\sharp v$ could as well gain scores. In the worst case, we shall lose $k-1$ scores. Writing this formally, we have $S(\phi_u(x)) \geq S(x) - (k-1)$. Therefore, $D_\lambda(\phi_u(x)) \geq D_\lambda(x)/\lambda^{k-1}$. Now we have

$$\begin{aligned} \mathbb{P}(x \in \Omega_1) &= \int_{\Omega_1} D_\lambda(x)\,\mathrm{d}x \\ &\leq \lambda^{k-1}\int_{\Omega_1} D_\lambda(\phi_u(x))\,\mathrm{d}x \\ &= \lambda^{k-1}\int_{\mathrm{Im}(\phi_u)} D_\lambda(y)\,\mathrm{d}y \quad = \lambda^{k-1}\,\mathbb{P}(x \in \mathrm{Im}(\phi_u)) \end{aligned}$$

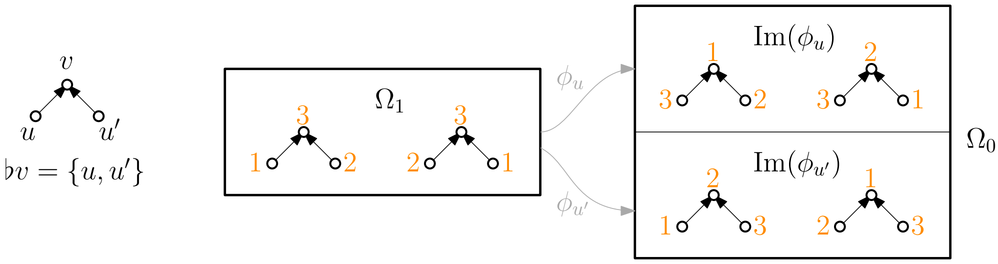

Figure 3–3 Sets and mappings in the proof when $k = 3$. The orange labels indicate the ordering induced by x in the neighbourhood.

where the last line comes from a change of variable $y := \phi_u(x)$. (The Jacobian is clearly 1 since ϕ_u behaves just like variable renaming...) Summing the inequalities for all $u \in \flat v$, we arrive at

$$(k-1) \cdot \mathbb{P}(x \in \Omega_1) \le \lambda^{k-1} \cdot \mathbb{P}(x \in \Omega_0).$$

But $\mathbb{P}(x \in \Omega_1) + \mathbb{P}(x \in \Omega_0) = 1$, so we derive at last

$$\mathbb{P}(x \in \Omega_1) \le \frac{1}{1 + (k-1)/\lambda^{k-1}}.$$

The rest of the theorem follows from Lemma 13 by noting that the primitive function of $\frac{1/\lambda}{1+(k-1)/\lambda^{k-1}}$ is exactly $\frac{1}{k-1}\ln(\lambda^{k-1} + k - 1)$. □

Remark. Our proof intentionally uses continuous space to prepare the readers for the later lemma. But the argument, at its core, is pretty discrete and combinatorial.

The weight spreading idea reveals only crude information of D_λ. Nevertheless, it provides a valuable prior knowledge of occupancy. For example, in the proof of Theorem 8 in the last chapter, we could use weight spreading to restrict the range of parameter $p = p(t)$. Below we sketch the key lemma.

Lemma 15. Assume $k = 3$ and fix $t \in (0, 1)$. For any v we have

$$p(t) := \mathop{\mathbb{E}}_{D}(S_v \mid x_v = t) \le \frac{5t^2 + 2t + 1}{8t^2}.$$

Proof. Suppose $\flat v = \{u, w\}$. Consider the spaces $A := [0,t]^2$, $B := [0,t] \times [t,1]$, $B' := [t,1] \times [0,t]$ and $C := [1,t] \times [1,t]$. It's easy to "stretch" the rectangle A into B, B' or C. Take the first as example: $(x_u, x_w) \in A$ is mapped to $\phi_{AB}(x_u, x_w) := (x_u, \frac{1-t}{t}x_w + t) \in B$. The other two mappings are defined similarly.

Now we may compare weight of $(x_u, x_w) \in A$ with $\phi_{AB}(x_u, x_w)$, $\phi_{AB'}(x_u, x_w)$, and $\phi_{AC}(x_u, x_w)$, respectively. And it is not hard to see

$$\begin{aligned} D_\lambda(x_u, x_w) &\le \lambda^2 D_\lambda(\phi_{AB}(x_u, x_w)) \\ D_\lambda(x_u, x_w) &\le \lambda^2 D_\lambda(\phi_{AB'}(x_u, x_w)) \\ D_\lambda(x_u, x_w) &\le \lambda^3 D_\lambda(\phi_{AC}(x_u, x_w)), \end{aligned}$$

which implies

$$\begin{aligned} \mathbb{P}((x_u, x_w) \in A) &\le \lambda^2 \frac{t}{1-t} \mathbb{P}((x_u, x_w) \in B) \\ \mathbb{P}((x_u, x_w) \in A) &\le \lambda^2 \frac{t}{1-t} \mathbb{P}((x_u, x_w) \in B') \\ \mathbb{P}((x_u, x_w) \in A) &\le \lambda^3 \left(\frac{t}{1-t}\right)^2 \mathbb{P}((x_u, x_w) \in C). \end{aligned}$$

where we implicitly assume the probability is conditioned on $x_v = t$. Therefore, at $\lambda = 2$,

$$\mathbb{P}(B) + \mathbb{P}(B') + \mathbb{P}(C) \geq \frac{(1-t)(1+3t)}{8t^2} \mathbb{P}(A).$$

But the left-hand side is just $1 - \mathbb{P}(A)$. Moving terms around establishes the lemma. □

Baking this prior knowledge into the optimisation of Theorem 8, one may improve the upper bound to 1.3852^n for vertex transitive graphs when $k = 3$. We leave the details to the readers. The weakness of the spreading argument lies in its assumption of worst scenario when relating $S(\phi(x))$ with $S(x)$. Conceivably, for the vast majority of scoring states, applying ϕ will not cause a radical score loss. But it seems challenging to take this aspect into account.

3.4 Notes and References

Our notion of occupancy originates from statistical physics. It is an analogue to the so-called *mean magnetization* in the hard-core model. Connection between occupancy and partition function (Lemma 13) is a well-known fact, which more or less reflects the significance of partition functions. In fact, most values of physical interest (e.g. pressure, temperature, etc) can be derived from the partition function. For an excellent introductory text to statistical physics, we recommend the lecture notes by David Tong [17]. An in-depth and rigorous treatment can be found in [8].

Davies, Jenssen, Perkins and Roberts [6, 5] developed a linear-programming approach to upper bound the occupancy. Briefly speaking, the method (i) defines some random variable Y (called "boundary") that shields a local region and establishes spatial independence; (ii) expresses the occupancy α in two ways as $\mathbb{E}(f(Y))$ and $\mathbb{E}(g(Y))$; (iii) uses a linear programme to model the distribution of Y and the constraint $\mathbb{E}(f(Y)) = \mathbb{E}(g(Y))$. It works well in settings that involve finite number of spins and hard constraints. However, it exhibits fundamental limitations in our context. Firstly, it is difficult to encode the boundary succinctly as our model does not exhibit hard constraints. Secondly, the method allows an adversary to bias the distribution of Y arbitrarily, even down to a one-point distribution. Such dramatic distributions are realistic in hard-core model etc. but unrealistic in our model.

Chapter 4 Upper Bounds via Bipartite Relaxation

In our model, every $v \in G$ is pulled by two conflicting forces. A higher value of x_v benefits S_v but unfortunately harms $S_{\sharp v}$. This chapter isolates these effects by splitting v into two copies. It leads to a bipartite model in which the interactions are more manageable and even exactly solvable in some case. Throughout the chapter we denote $d := k - 1$ to avoid clutter.

4.1 Bipartite Relaxation

Let us associate $G = (V, E)$ with a bipartite graph $G' := (A \cup B, E')$, where $A = \{v^- : v \in V\}$, $B = \{v^+ : v \in V\}$ and $E' := \{u^- v^+ : uv \in E\}$. In words, we split each vertex v into two copies, v^- and v^+. The '$-$' copy radiates edges and the '$+$' copy absorbs them; see Figure 4–1. The edge orientation does not really matter, so we are flexible to interpret G' as either directed or undirected. Clearly G' has $2n$ vertices and is d-regular in the undirected sense.

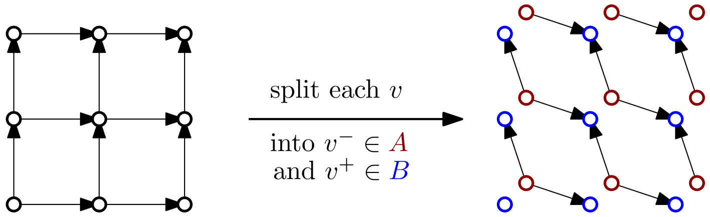

Figure 4–1 Illustration of vertex splitting.

We may define a model on G' just like before: Sample $x \in [0,1]^{2n}$ uniformly and collect score $S_b(x) := \mathbf{1}[x_b \geq \max x_{\flat b}]$ for each $b \in B$. The total score writes $S(x) := \sum_{b \in B} S_b(x)$ and we inquire about $\mathbb{E}(2^S)$. We refer to it as the *bipartite model.* The definitions are motivated by the apparent advantage that, in a bipartite model, interactions flow one-way from A to B without feedback.

The following correlation inequality is our main tool in linking the bipartite model with our old model.

Lemma 16 (Chebyshev's sum inequality)**.** For any monotonically increasing functions $f, g : \mathbb{R} \to \mathbb{R}$ and distribution μ on $\mathbb{R}$, we have

$$\mathop{\mathbb{E}}_{x \sim \mu}[f(x)] \mathop{\mathbb{E}}_{x \sim \mu}[g(x)] \leq \mathop{\mathbb{E}}_{x \sim \mu}[f(x) g(x)]$$

whenever they exist. The inequality flips sign if f is increasing and g is decreasing, or vice versa.

Proof. Since both f and g are monotonically increasing, $(f(x) - f(y)) \cdot (g(x) - g(y)) \geq 0$

for all x, y. Therefore, if we sample $x \sim \mu$, $y \sim \mu$ independently, then

$$\begin{aligned} 0 &\leq \mathop{\mathbb{E}}_{x,y} \left[(f(x) - f(y)) \cdot (g(x) - g(y)) \right] \\ &= \mathop{\mathbb{E}}_{x,y} [f(x)g(x)] - \mathop{\mathbb{E}}_{x,y} [f(x)g(y)] - \mathop{\mathbb{E}}_{x,y} [f(y)g(x)] + \mathop{\mathbb{E}}_{x,y} [f(y)g(y)] \\ &= 2 \mathop{\mathbb{E}}_{x} [f(x)g(x)] - 2 \mathop{\mathbb{E}}_{x} [f(x)] \mathop{\mathbb{E}}_{x} [g(x)]. \end{aligned}$$

The last line follows from independence of x and y. Moving terms and dividing the inequality by 2 proves the lemma. □

Lemma 17. The bipartite model is an upper bound for the old model:

$$\mathop{\mathbb{E}}_{G}(2^S) \leq \mathop{\mathbb{E}}_{G'}(2^S).$$

Proof. We may implement the splits sequentially, one vertex at a time. In an intermediate digraph, we again sample x uniformly for all vertices, and count scores for all but '$-$' vertices. Suppose, at some arbitrary step, we are dealing with G_1 and splitting vertex v into v^- and v^+. This results in digraph G_2. To establish the lemma, it suffices to show

$$\mathop{\mathbb{E}}_{G_1} \left(2^S \right) \leq \mathop{\mathbb{E}}_{G_2} \left(2^S \right). \tag{4–1}$$

Claim. Fix any *partial* assignment $y \in [0,1]^{V(G_1) \setminus \{v\}}$, it holds that $\mathbb{E}_{G_1} \left(2^S \mid y \right) \leq \mathbb{E}_{G_2} \left(2^S \mid y \right)$. Consequently, (4–1) follows by taking expectation $\mathbb{E}_{y \sim U}$ on both sides.

Under partial assignment y, all scores are fixed *except* S_v in G_1, S_{v^+} in G_2, and $S_{\sharp v}$ in both. Here we agree that $\sharp v$ is defined in the context of G_1. These undetermined scores depend on the deferred sampling of x_v (in G_1) and x_{v^+}, x_{v^-} (in G_2). Hence,

$$\mathop{\mathbb{E}}_{G_1} \left(2^S \mid y \right) = c \cdot \mathop{\mathbb{E}}_{x_v} [f(x_v) g(x_v)], \qquad \mathop{\mathbb{E}}_{G_2} \left(2^S \mid y \right) = c \cdot \mathop{\mathbb{E}}_{x_{v^+}} [f(x_{v^+})] \mathop{\mathbb{E}}_{x_{v^-}} [g(x_{v^-})]$$

where c packs the "fixed scores", f accounts of 2^{S_v} (in G_1) or $2^{S_{v^+}}$ (in G_2), and g accounts of $2^{S_{\sharp v}}$. It's important to note that both models share the same c, f, g under condition y. Now observe that f is increasing and g is decreasing. Applying Lemma 16 gives our claim. □

4.2 An Entropy Upper Bound

Retaining merely "one-way" correlation from A to B, the bipartite model is much simpler to analyse. In fact, we are able to prove a tight upper bound by an entropy argument.

Theorem 18. $\mathbb{E}_{G'}(2^S)$ is maximised when $G' = \frac{n}{d} K_{d,d}$, the disjoint union of $\frac{n}{d}$ many complete d-regular bipartite graphs. In addition, $\mathbb{E}_{G'}(2^S) < (ed)^{n/d} = 2^{n \cdot O(\log k/k)}$.

Proof. We discretise the interval $[0,1]$ into $[N]$ for some large number $N \in \mathbb{N}$, and sample x from $[N]^{2n}$ rather than from $[0,1]^{2n}$. The score calculation remains unchanged. Since the discrete scheme is coarser and gives scores more generously, there is no problem for upper bound purpose.

For state $x \in [N]^{2n}$, we abuse notation and redefine $S(x) := (S_b(x))_{b \in B}$ as a *binary vector* indicating the local scores. Let $\Omega := \{(x, R) \in [N]^{2n} \times \{0,1\}^n : R \leq S(x)\}$. In words, R has the freedom of "downgrading" scoring vertices. Then we may write

$$\mathop{\mathbb{E}}_{G'} \left(2^{|S(x)|} \right) = \sum_{x \in [N]^{2n}} \frac{1}{N^{2n}} \cdot 2^{|S(x)|} = \frac{1}{N^{2n}} \sum_{x \in [N]^{2n}} \sum_{R \leq S(x)} 1 = \frac{|\Omega|}{N^{2n}}.$$

Consider a probability space where we sample (x, R) uniformly from Ω; beware that x is *not* uniform in this space. Then $H(x, R) = \log|\Omega|$ and thus $\log \mathbb{E}_{G'}(2^{|S(x)|}) = H(x, R) - 2n \log N$. Hence it suffices to bound the entropy in our new space. By the chain rule,

$$H(x, R) = H(x_A) + H(x_B, R \mid x_A) \tag{4–2}$$

and we bound them separately. Define a random variable $M_b := \max x_{\flat b}$ for each $b \in B$. The key intuition is: If M_b is small then $x_{\flat b}$ shall concentrate at the bottom half of $[N]$ and $H(x_A)$ is lowered; if M_b is large then R_b would mostly lose freedom and $H(x_B, R \mid x_A)$ is lowered. Either way, $H(x, R)$ is constrained from above. The argument below formalises our intuition. Let $\mu(\cdot)$ be the distribution of M_b, where we have suppressed the dependence on b. Then

$$\begin{aligned} H(x_A) &\le \frac{1}{d} \sum_{b \in B} H(x_{\flat b}) \\ &= \frac{1}{d} \sum_{b \in B} H(M_b) + H(x_{\flat b} \mid M_b) \\ &= \frac{1}{d} \sum_{b \in B} \left(\sum_{m=1}^{N} \mu(m) \log \frac{1}{\mu(m)} + \sum_{m=1}^{N} \mu(m) H(x_{\flat b} \mid M_b = m) \right) \\ &\le \frac{1}{d} \sum_{b \in B} \sum_{m=1}^{N} \mu(m) \log \frac{m^d - (m-1)^d}{\mu(m)} \end{aligned} \tag{4–3}$$

where the first line follows from Lemma 4 because each $a \in A$ is covered d times. The last line follows from the fact that, given $M_b = m$, $x_{\flat b}$ is distributed over the set $[m]^d - [m-1]^d$ (either uniformly or not; we don't really know).

For the second half, we have

$$\begin{aligned} H(x_B, R \mid x_A) &\le \sum_{b \in B} H(x_b, R_b \mid x_A) \\ &= \sum_{b \in B} H(x_b, R_b \mid M_b) \\ &= \sum_{b \in B} \sum_{m=1}^{N} \mu(m) H(x_b, R_b \mid M_b = m) \\ &= \sum_{b \in B} \sum_{m=1}^{N} \mu(m) \log(2N + 1 - m) \end{aligned} \tag{4–4}$$

where the first line is due to subadditivity of entropy. The last line is justified by observing that, with the knowledge $M_b = m$, the pair (x_b, R_b) is uniformly distributed on the set $(\{1, \ldots, m-1\} \times \{0\}) \cup (\{m, \ldots, N\} \times \{0, 1\})$. (But as we care about upper bounds only, the *uniformity* is not a must. Knowing the support set is sufficient and we could write '$\le$' in place of '$=$'.)

Recall that $\log \mathbb{E}_{G'} \left(2^{|S^x|}\right) = H(x, R) - 2n \log N$. Combining it with (4–2), (4–3) and (4–4) yields

$$\log \mathop{\mathbb{E}}_{G'} \left(2^{|S^x|}\right) \le \frac{1}{d} \sum_{b \in B} \sum_{m=1}^{N} \mu(m) \log \frac{(m^d - (m-1)^d)(2N + 1 - m)^d}{N^{2d} \cdot \mu(m)}.$$

We are left with a familiar situation: the inner summation is just a negative KL divergence with constant shift. Repeating our argument in the notes of Chapter 2, the inner summation rewrites into $\log \tilde{Z} - \mathrm{KL}(\mu \| \nu)$, where

$$\nu(m) := \frac{(m^d - (m-1)^d)(2N+1-m)^d}{N^{2d} \cdot \tilde{Z}}$$

and

$$\tilde{Z} := \sum_{m=1}^{N} \frac{(m^d - (m-1)^d)(2N+1-m)^d}{N^{2d}}. \tag{4–5}$$

is its normalising constant. But $\mathrm{KL}(\mu \| \nu) \geq 0$, so

$$\mathop{\mathbb{E}}_{G'} \left(2^{|S^x|}\right) \leq \tilde{Z}^{n/d}. \tag{4–6}$$

Before evaluating $\tilde{Z}$, let us make some observations. Suppose our bipartite graph is $K_{d,d}$, then the argument above gives away nothing:

- The first inequality in (4–3) is tight because, for each $b \in B$, $x_{\flat b}$ is just x_A.
- The last inequality in (4–3) is tight since x_A is uniform conditioned on $x_A = m$. (Every $b \in B$ cares about m only.)
- The inequality in (4–4) is tight since $\{R_b\}_{b \in B}$ are independent conditioned on x_A.
- The KL divergence achieves 0 since $\mu = \nu$ for every $b \in B$. That is, the distribution of $\max x_{\flat b} = \max x_A$ is given by ν indeed. This may be verified by counting the number of $(x, R) \in \Omega$ conditioned on $\max x_A = m$.

The same observation clearly holds for $\frac{n}{d} K_{d,d}$ as well, since the disjoint components are independent. So the inequality in (4–6) is tight when $G' = \frac{n}{d} K_{d,d}$, proving the first part of the theorem.

However, it is difficult to evaluate $\tilde{Z}$ directly from its definition (4–5). We will see a workaround in the next section, but here we resort to a rough, yet asymptotically good, upper bound for $\tilde{Z}$.

Since our argument holds for any N, we could choose $N := d$ for convenience in (4–5). Dropping the $(m-1)^d$ from numerator gives

$$\tilde{Z} \leq \sum_{m=1}^{d} \left(\frac{m(2d+1-m)}{d^2} \right)^d.$$

But now the numerator is monotonically increasing when $1 \leq m \leq d$. Hence,

$$\tilde{Z} \leq d \cdot \left(\frac{d(d+1)}{d^2} \right)^d < ed,$$

proving the second part of the theorem. □

4.3 Exact Solution to $\frac{n}{d} K_{d,d}$

In this section we solve the bipartite model on $\frac{n}{d} K_{d,d}$ exactly and explicitly. It has two direct consequences: (i) provides a tight bound in Theorem 18, much sharper than $(ed)^{n/d}$; (ii) tells us the actual value of $\tilde{Z}$ as $N \to \infty$. Note that we shall work in our usual uniform space on $[0,1]^{2n}$ rather than on Ω.

Theorem 19. When $G' = \frac{n}{d} K_{d,d}$ we have

$$\mathop{\mathbb{E}}_{G'} \left(2^S\right) = \left(\frac{1}{2} + \frac{[(d-1)!]^2}{(2d-1)!} \cdot d \cdot 4^{d-1} \right)^{n/d} \sim \left(\frac{1 + \sqrt{\pi d}}{2} \right)^{n/d}.$$

Proof. Since the graph is a disjoint union, its components are independent and we could focus on one. We henceforth assume $n = d$ and the bipartite graph being exactly $K_{d,d}$.

Denote $M := \max x_A$, i.e. the maximum value on left hand side. Let us figure out its distribution analytically. With $F(t)$ notating its cumulative probability function and $f(t)$ its density function, we have

$$F(t) = \mathbb{P}(M < t) = \mathbb{P}(\forall a \in A, x_a < t) = t^d$$

since every x_a are distributed uniformly and independently. Differentiating, we derive

$$f(t) = d \cdot t^{d-1}.$$

Next we observe that the $\{S_b\}_{b\in B}$ are independent *conditioned on M*. To see this, simply note that $S_b = \mathbf{1}[x_b > M]$ depends on x_b and M only, and, x_b are by definition independent. Therefore,

$$\mathbb{E}\left(2^S \mid M\right) = \prod_{b\in B} \mathbb{E}\left(2^{S_b} \mid M\right) = (2-t)^d.$$

Taking expectation over M,

$$\begin{aligned}\mathbb{E}\left(2^S\right) = \mathbb{E}\left(\mathbb{E}(2^S \mid M)\right) &= \int_0^1 (2-t)^d \cdot d \cdot t^{d-1}\ \mathrm{d}t \\ &= d\int_0^1 \left(2t-t^2\right)^{d-1} \cdot (2-t)\ \mathrm{d}t \\ &= d\int_0^1 \left(2t-t^2\right)^{d-1} \cdot (1-t)\ \mathrm{d}t + d\int_0^1 \left(2t-t^2\right)^{d-1}\ dt \\ &=: I_1 + I_2.\end{aligned}$$

But

$$I_1 = \frac{1}{2}\left[(2t-t^2)^d\right]_0^1 = \frac{1}{2}$$

and

$$\begin{aligned}I_2 &= d\int_0^{\pi/2} \cos^{2d-1}\theta\ \mathrm{d}\theta && (\cos^2\theta = 2t-t^2) \\ &= d \cdot \frac{(2d-2)!!}{(2d-1)!!} = \frac{[(d-1)!]^2}{(2d-1)!} \cdot d \cdot 4^{d-1}. && \text{(Wallis' integral)}\end{aligned}$$

So we obtain

$$\mathbb{E}\left(2^S\right) = \frac{1}{2} + \frac{[(d-1)!]^2}{(2d-1)!} \cdot d \cdot 4^{d-1}.$$

Using Stirling's approximation $j! \sim \sqrt{2\pi j}(j/e)^j$, the quantity is asymptotically

$$\frac{1}{2} + \sqrt{2\pi} \cdot e \cdot \frac{d(d-1)}{(2d-1)^{3/2}} \left(\frac{2d-2}{2d-1}\right)^{2d-2} \sim \frac{1+\sqrt{\pi d}}{2}.$$

Raising the above results to $\frac{n}{d}K_{d,d}$ completes the proof. □

Corollary 20.

$$\mathop{\mathbb{E}}_{G'}\left(2^S\right) \le \left(\frac{1}{2} + \frac{[(d-1)!]^2}{(2d-1)!} \cdot d \cdot 4^{d-1}\right)^{n/d}.$$

Corollary 21.

$$\lim_{N\to\infty} \tilde{Z} = \frac{1}{2} + \frac{[(d-1)!]^2}{(2d-1)!} \cdot d \cdot 4^{d-1}.$$

4.4 Returning to the Old Model

Theorem 22. For any d-regular digraph G,

$$Z(2) \le \left(\frac{1}{2} + \frac{[(d-1)!]^2}{(2d-1)!} \cdot d \cdot 4^{d-1}\right)^{n/d} = 2^{n \cdot \Theta(\log k/k)}.$$

In particular, $Z(2) \le (11/6)^{n/2} < 1.3541^n$ when $k = d+1 = 3$.

Proof. A direct consequence of Lemma 17 and Corollary 20. □

Our discussion so far works by relaxing digraph G into a bipartite graph G'. Below we briefly discuss the converse procedure. Actually, G' is not uniquely decodable into its original form if we disregard vertex labels. For example, the bipartite graph $\frac{n}{d}K_{d,d}$ could result from

- a disjoint union of n/d many directed complete graphs K_d with self-loops at every vertex;
- a graph consisting of n/d layers $L_1, \ldots, L_{n/d}$, with d vertices each, such that (L_i, L_{i+1}) is a complete bipartite graph for all i. The edges direct from left to right and wrap around at boundary. (Figure 4–2)
- ...

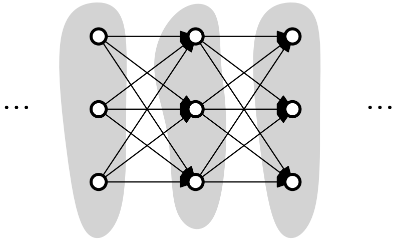

Figure 4–2 The layered graph when $d = 3$. Ignoring scores in odd layers is equivalent to the model on $\frac{n}{2d}K_{d,d}$.

Interestingly, the possibilities have disparate natures. The former contains extremely repelling vertices: only the biggest vertex scores in each K_d, and thus $\mathbb{E}(2^S) = 2^S = 2^{n/d}$. (This also shows the tightness of Jensen's lower bound for general graphs.) The latter, on the other hand, fosters friendship inside each layer; if we ignore the scores in odd layers then it is just a model on $\frac{n}{2d}K_{d,d}$ and we readily know $\mathbb{E}(2^S) \ge \left(\frac{1+\sqrt{\pi d}}{2}\right)^{n/2d} = 2^{n \cdot \Theta(\log k/k)}$ by the last section. (This also implies that the asymptotics in Theorem 22 is tight.)

4.5 Notes and References

Chebyshev's sum inequality (Lemma 16) is a special case of the following very general inequality:

Lemma 23 (The FKG inequality, [7])**.** Let L be a finite distributive lattice. Suppose $\mu : L \to \mathbb{R}^+$ satisfies $\mu(x)\mu(y) \le \mu(x \vee y)\mu(x \wedge y)$ for all $x, y \in L$. Let $f, g : L \to \mathbb{R}^+$ be monotonically increasing functions. Then

$$\left(\sum_{x\in L}\mu(x)f(x)\right)\left(\sum_{x\in L}\mu(x)g(x)\right) \le \left(\sum_{x\in L}\mu(x)f(x)g(x)\right)\left(\sum_{x\in L}\mu(x)\right).$$

An elementary proof can be found in the book [1] by Alon and Spencer.

Our proof of Theorem 18 borrows idea from Kahn's paper [10] where he used entropy to show that $\frac{n}{d}K_{d,d}$ maximises the number of independent sets among all bipartite graphs of order n. To apply argument of this kind, it seems essential to write the target value as some unweighted count, just as we did via the (x, R) trick. It can be generalised for any $\lambda \in \mathbb{N}$ by choosing $R \in \{0, \ldots, \lambda - 1\}^n$ and disregarding the difference among non-zeros.

Chapter 5 High Girth and Beyond

In a way, we have arrived at a satisfying position. The discussion so far shows that the asymptotic behaviour of both our lower and upper bounds are tight for general graphs. The gap between them, however, is intriguing. Our tight instances for both bounds have small undirected girth. What is the behaviour of $Z(2)$ when girth is large? Will it be the case that the gap closes up in the high-girth scenario? This chapter reveals strong evidence for such behaviour.

Yet many interesting problems arise and remain open. Some of these ask for the exact solution to tree-like graphs, while others concern the big picture of analysing PPSZ. The final section concludes this thesis with a list of open problems that we consider significant. Tackling them would definitely push forward our understanding of the entire picture.

5.1 Polynomial with Local Coefficients

Definition 4. Define polynomial $p_G(z) := Z_G(z+1) = \frac{1}{n!}\sum_\pi (z+1)^{S(\pi)}$, where the summation is over all permutations on $V(G)$. Let $d := \deg p_G \le n$. Denote its standard form as $p_G(z) = \sum_{i=0}^d a_i z^i$, where a_i are the standard coefficients.

What is the point for translating $Z_G(z)$ by 1? Let us massage $p_G(z)$ a little. Below we abuse notation $S(\pi)$ to denote the set of scoring vertices.

$$\begin{aligned} p_G(z) &= \frac{1}{n!}\sum_{\pi}\prod_{v\in S(\pi)}(z+1) = \frac{1}{n!}\sum_{\pi}\sum_{R\subseteq S(\pi)} z^{|R|} \\ &= \frac{1}{n!}\sum_{R\subseteq V} z^{|R|}\sum_{\pi}\prod_{\substack{uv\\ v\in R}}\mathbf{1}[\pi_u<\pi_v] \\ &= \sum_{R\subseteq V} z^{|R|}\cdot\frac{1}{|R\cup\flat R|!}\sum_{\substack{\pi\text{ on}\\ R\cup\flat R}}\prod_{\substack{uv\\ v\in R}}\mathbf{1}[\pi_u<\pi_v] \end{aligned}$$

where the factor $n!/|R\cup\flat R|!$ arise because we may choose permutation outside $R\cup\flat R$ arbitrarily, leading to $(|R\cup\flat R|+1)\cdots n$ possibilities. Now we can write explicit formulae for the standard coefficient a_i:

$$a_i = \sum_{\substack{R\subseteq V\\ |R|=i}}\frac{1}{|R\cup\flat R|!}\sum_{\substack{\pi\text{ on}\\ R\cup\flat R}}\prod_{\substack{uv\\ v\in R}}\mathbf{1}[\pi_u<\pi_v]. \tag{5–1}$$

Note that the inner evaluation cares only about $R\cup\flat R$; external structures vanish. In this sense, a_i is "locally computable". Compared to Z_G, the polynomial p_G breaks things into more manageable pieces, allowing us to establish our main theorem:

Theorem 24. *If* there exists a constant $\beta>1$ such that $p_G(z)$ has no root in the complex disc $\{z\in\mathbb{C} : |z|\le\beta\}$ for all G, *then* there is a function $\hat{p}(z)$, depending on n but independent of graph structure, such that $e^{-\epsilon n}\hat{p}(z)\le p_G(z)\le e^{\epsilon n}\hat{p}(z)$ for all graphs G and all $z\le 1$. Here $\epsilon = O(1/\beta^{\mathrm{girth}(G)})$.

Remark. We remind the reader that $p_G(-1) = Z(0) = 0$, so the condition of the theorem can never be met actually. But we believe that $z=-1$ is the only violation inside a large region including the unit disc. In that case, there is a simple workaround of this issue; see chapter notes for discussion. We insist stating the current theorem to make life easier.

The proof is presented in the following sections. One more preparation is necessary. Call a graph invariant f *multiplicative* if $f(G\dot{\cup}G') = f(G)f(G')$. Similarly, call it *additive* if $f(G\dot{\cup}G') = f(G) + f(G')$. Our polynomial $p_G(z)$, when viewed as a graph invariant, is multiplicative; just observe that

$$\begin{aligned} Z_{G\dot{\cup}G'}(z) &= \mathop{\mathbb{E}}_{\substack{x \text{ on } G \\ x' \text{ on } G'}} \left(z^{S_G(x)+S_{G'}(x')}\right) && (x, x' \sim U) \\ &= \mathop{\mathbb{E}}_{x} \left(z^{S_G(x)}\right) \cdot \mathop{\mathbb{E}}_{x'} \left(z^{S_{G'}(x')}\right) && \text{(independence)} \\ &= Z_G(z) \cdot Z_{G'}(z). \end{aligned}$$

5.2 Barvinok's Approach

Define $\ell_G(z) := \ln p_G(z)$. Let us Taylor expand it at the origin:

$$\ell_G(z) = \sum_{j=0}^{\infty} \frac{\ell^{(j)}(0)}{j!} z^j.$$

On the other hand, we could truncate the expansion at $j = m$ for some fixed number m and obtain an approximation

$$\hat{\ell}_G(z) := \sum_{j=0}^{m} \frac{\ell^{(j)}(0)}{j!} z^j.$$

What is the quality of this approximation? Barvinok's lemma asserts that $\hat{\ell}_G$ is quite close to ℓ_G under certain circumstances, even when m is arguably small.

Lemma 25 (Barvinok)**.** Let $r_1, r_2, \ldots, r_d$ be the complex roots of $p_G(z)$. If $|r_i| \geq \beta > 1$ for all i, then $|\hat{\ell}_G(z) - \ell_G(z)| < \epsilon d$ provided $|z| \leq 1$ and $m \geq \log_\beta(\frac{1}{(\beta-1)\epsilon})$.

Proof. Given the knowledge of complex roots, we could write

$$p_G(z) = c \cdot \prod_{i=1}^{d} (z - r_i) \qquad \text{and} \qquad \ell_G(z) = \ln c + \sum_{i=1}^{d} \ln(z - r_i).$$

Differentiating the second equation for j times and taking $z = 0$ gives

$$\ell_G^{(j)}(0) = -(j-1)! \sum_{i=1}^{d} r_i^{-j}. \tag{5–2}$$

So for $|z| \leq 1$ the error between $\hat{\ell}_G$ and ℓ_G is bounded by

$$|\hat{\ell}_G(z) - \ell_G(z)| = \left| \sum_{j=m+1}^{\infty} \frac{\ell_G^{(j)}(0)}{j!} z^j \right| \leq \sum_{j=m+1}^{\infty} \sum_{i=1}^{d} \frac{1}{|j| \cdot |r_i|^j} \leq \frac{1}{m+1} \sum_{j=m+1}^{\infty} \sum_{i=1}^{d} \frac{1}{|r_i|^j}.$$

But $|r_1|, \ldots, |r_d| \geq \beta$ by assumption, thus

$$|\hat{\ell}_G(z) - \ell_G(z)| \leq \frac{d}{m+1} \sum_{j=m+1}^{\infty} \frac{1}{\beta^j} = \frac{d}{(m+1)(\beta-1)\beta^m}$$

which is bounded by ϵd when $m \geq \log_\beta(\frac{1}{(\beta-1)\epsilon})$. □

Corollary 26. Let $\hat{p}_G(z) := \exp\{\hat{\ell}_G(z)\}$. If $p_G(z)$ has no complex roots in the β-disc, then $e^{-\epsilon n}\hat{p}_G(z) \le p_G(z) \le e^{\epsilon n}\hat{p}_G(z)$.

Be alarmed that we are not done yet: $\hat{p}_G$ might depend on the specific graph structure G, so it remains unclear at the moment if Theorem 24 holds. Nonetheless, it suggests a way towards the goal:

Prove that $\hat{p}_G$ does not really depend on G.

At first glance this looks impossible. But if we restrict ourselves to high-girth graphs, it has a rather simple explanation. First, observe that $\hat{p}_G = \exp\{\hat{\ell}_G\}$ only depends on the first m derivatives of ℓ_G. Second, it turns out that these derivatives depend only on $a_0, \dots, a_m$. To see this, note $\ell'_G(z) = p'_G(z)/p_G(z)$, or equivalently $p'_G(z) = p_G(z)\ell'_G(z)$. Differentiating both sides $j-1$ times and taking $z = 0$,

$$j!\ a_j = \sum_{i=0}^{j-1} \binom{j-1}{i} i!\ a_i\ \ell_G^{(j-i)}(0). \tag{5–3}$$

So we could solve $\ell_G^{(1)}(0), \dots, \ell_G^{(m)}(0)$ from $a_0, \dots, a_m$ and the boundary $\ell_G^{(0)}(0) = 0$. Third, by (5–1), the coefficient a_i merely depends on the structure $R \cup \flat R$ where $|R| = i \le m$. Finally, if the girth of G is large enough, then such $R \cup \flat R$ always forms a forest, revealing no specific graph structure at all!

5.3 Structures Disappear for High-Girth Graphs

This section makes the explanation above more precise. We begin with some definitions.

Definition 5. A coloured digraph H of maximum degree $\le k-1$ is called a *certificate* if

(1) Each $v \in H$ is coloured either black or white;

(2) Every black vertex has exactly $k-1$ predecessors in H;

(3) Every white vertex has at least one black successor.

Write $|H|$ for the number of vertices in H, and $\|H\|$ for the number of black vertices in H; clearly $|H| \le k\|H\|$. We say a digraph G is *isomorphic* to H, denoted $G \cong H$, in the sense of digraph isomorphism (i.e. disregarding colours). Denote by $\mathrm{aut}(H)$ the number of automorphisms of H, but this time taking colours into account.

Figure 5–1 illustrates the idea. Intuitively, the black vertices encode score information, while the white vertices serve as (additional) witnesses. Item (2) enforces the soundness of witnessing, and item (3) rules out redundant witnesses.

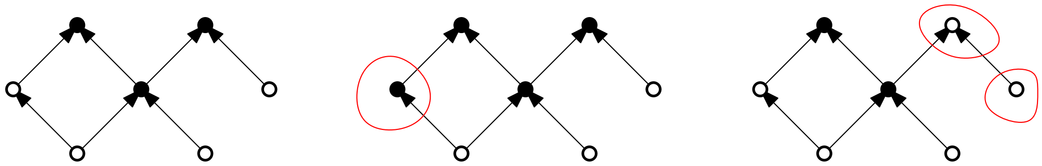

Figure 5–1 Examples when $k = 3$. The leftmost graph is a certificate while the other two are not. The red circles mark violation of the definition.

Let $\mathcal{H}$ be the class of all certificates. Define its subclass, $\mathcal{H}_s$, by admitting those H with $|H| \le s$. Further restrict this subclass to $\mathcal{C}_s$ by choosing only connected certificates (in the sense of undirected connectivity).

Definition 6. For $H \in \mathcal{H}$, define $\operatorname{ind}(H, G) := |\{S \subseteq V(G) : G[S] \cong H\}|$, that is the number of sites we could embed H into G as an *induced* subgraph (disregarding colours). Note that $\operatorname{ind}(H, G) = 0$ when $|H| > |G|$.

Using the definitions above, the coefficients a_i (5–1) may be rewritten as

$$\begin{aligned} a_i &= \sum_{\substack{H\in\mathcal{H}\\ \|H\|=i}} \frac{1}{|H|!} \operatorname{ind}(H,G)\cdot \operatorname{aut}(H) \sum_{\pi \text{ on } H} \prod_{\substack{uv\in H\\ v \text{ black}}} \mathbf{1}[\pi_u < \pi_v] \\ &= \sum_{H\in\mathcal{H}_{ki}} \frac{1}{|H|!} \operatorname{ind}(H,G)\cdot \operatorname{aut}(H)\cdot \mathbf{1}[\|H\| = i] \sum_{\pi \text{ on } H} \prod_{\substack{uv\in H\\ v \text{ black}}} \mathbf{1}[\pi_u < \pi_v] \\ &=: \sum_{H\in\mathcal{H}_{ki}} \lambda_i(H)\cdot \operatorname{ind}(H,G) \end{aligned} \tag{5–4}$$

Observe that the coefficients $\lambda_i(H)$ may be predetermined without even knowing G. Hence, the structure of G pops up at $\operatorname{ind}(H, G)$ only.

Our next step is to make the summation effectively running over $H \in \mathcal{C}_{ki}$ instead of $H \in \mathcal{H}_{ki}$. To this end, we develop some simple properties of $\operatorname{ind}(H, G)$.

Lemma 27. Suppose $H_1 \in \mathcal{H}_r$, $H_2 \in \mathcal{H}_s$, then

$$\operatorname{ind}(H_1, G)\cdot \operatorname{ind}(H_2, G) = \sum_{H\in\mathcal{H}_{r+s}} c(H_1, H_2; H)\cdot \operatorname{ind}(H, G)$$

where $c(H_1, H_2; H) := |\{(S_1, S_2) : S_1 \cup S_2 = V(H), H[S_1] \cong H_1, H[S_2] \cong H_2\}|$.

Proof. The left-hand side counts the size of the set $\{(S_1, S_2) : G[S_1] \cong H_1, G[S_2] \cong H_2\}$. We could imagine moving two templates, H_1 and H_2, around the graph G and count every time when both of them found a match.

The right-hand side does the same thing in two stages: (i) it enumerates H as a candidate structure for $G[S_1] \cup G[S_2]$; (ii) it decomposes H into S_1, S_2 and count. The $c(H_1, H_2; H)$ takes care of possible decompositions, while the $\operatorname{ind}(H, G)$ accounts for possible locations of the bulk $G[S_1] \cup G[S_2]$. □

This lemma is useful since it decomposes a product to linear combination. As we will see later, applying it iteratively would kill higher-order terms that are difficult to manipulate.

The following crucial lemma gives us the power for simplifying $\mathcal{H}$ to $\mathcal{C}$.

Lemma 28. Let $f(G) := \sum_{H\in\mathcal{H}} \gamma(H)\cdot \operatorname{ind}(H, G)$ be a graph invariant, then: $f(G)$ is additive $\iff$ $\gamma(H) = 0$ for all disconnected H.

Proof. ($\Leftarrow$) For any disjoint graphs G, G' and connected certificate H, it holds that $\operatorname{ind}(H, G\dot\cup G') = \operatorname{ind}(H, G) + \operatorname{ind}(H, G')$ since H cannot span two disjoint components. Hence $f(G\dot\cup G') = f(G) + f(G')$.

($\Rightarrow$) Assume without loss of generality that $\gamma(H) = 0$ for all $H \in \mathcal{C}$. (If not, then subtract $\sum_{H\in\mathcal{C}} \gamma(H)\cdot \operatorname{ind}(H, G)$ from $f(G)$. The resulting invariant preserves additivity as well as the disconnected coefficients.) We claim that $\gamma(H) = 0$ for all $H \in \mathcal{H}$.

Proceed by induction on $|H|$. The base case $|H| = 1$ is vacuously true as H must be connected. Now assume the claim holds for $|H| \leq s$ and we step to $|H| = s + 1$. If H is connected then we are done. If H is disconnected, then partition it into two components, say H_1 and H_2. But by induction hypothesis, $f(H_i) = \sum_{|H'|>s} \gamma(H')\cdot \operatorname{ind}(H', H_i)$, which is 0 since $|H_i| \leq s < |H'|$. Hence $f(H) = f(H_1) + f(H_2) = 0$ by additivity. On the other hand, again by hypothesis we have $f(H) = \gamma(H) \operatorname{ind}(H, H) = \gamma(H)$. Therefore $\gamma(H) = 0$, finishing the induction. □

Write σ_j as a shorthand for $\sum_{i=1}^{d} r_i^{-j}$; see (5–2). Plug (5–2) into (5–3) and simplify, we would get a recursive formula

$$\sigma_j = -ja_j - \sum_{i=1}^{j-1} a_i \sigma_{j-i}.$$

The recursion is non-linear due to the product terms $a_i\sigma_{j-i}$. But we could spread the product with the help of Lemma 27. It's easy to prove by induction on j that

$$\sigma_j = \sum_{H \in \mathcal{H}_{kj}} \gamma_j(H) \cdot \mathrm{ind}(H, G).$$

for *some* constants $\gamma_j(H)$ independent of G.

Because $p_G(z)$ is multiplicative, the power sum of roots, σ_j, must be additive. Then by Lemma 28, $\gamma_j(H) = 0$ whenever H is disconnected. Therefore, the above equation simplifies to

$$\sigma_j = \sum_{H \in \mathcal{C}_{kj}} \gamma_k(H) \cdot \mathrm{ind}(H, G).$$

Finally, recall $\ell_G^{(j)}(0) = -(j-1)!\ \sigma_j$ by (5–2), so the derivatives $\ell_G^{(j)}(0)$ only depend on $\mathrm{ind}(H, G)$ for $H \in \mathcal{C}_{kj}$. But the next easy fact reminds us that such $\mathrm{ind}(H, G)$ never depend on G provided the girth is large:

Lemma 29. Suppose $H \in \mathcal{C}_{kj}$. For all $(k-1)$-regular digraphs G with girth $> kj$, the count $\mathrm{ind}(H, G)$ only depends on n.

Proof. Since $\mathrm{girth}(G) > |H|$, we have $\mathrm{ind}(H, G) = 0$ whenever H has a loop. So $\mathrm{ind}(H, G) > 0$ only when H is a tree. But the number of ways we could embed a small tree into G does not depend on the structure of G, as we always (effectively) see a k-regular tree everywhere in G. So $\mathrm{ind}(H, G)$ only depends on the number of sites, i.e. the order n. □

Corollary 30. There is a universal $\hat{p}(z)$ such that $\hat{p}_G(z) = \hat{p}(z)$ for all graphs G of girth $> km$ and order n.

Corollary 26 along with Corollary 30 directly imply Theorem 24.

5.4 Open Problems

Problem 1. For the time being, we are unable to verify the zero-free condition of Theorem 24. However, numerical computations on some toy instances are in strong favour of it. We have tried 2 × 4, 2 × 5, 3 × 3 and 3 × 4 square grids, as well as some other make-up 2-regular graphs. The root pattern of $p_G(z)$ for these graphs are depicted in Figure 5–2. Based on graphical observations, we propose the bold conjecture

$p_G(z) \neq 0$ for all G and $z : \Re z > -1$.

Try to prove it. If we succeed, then by Theorem 24, $Z(2)$ indeed disregards structures for high-girth graphs.

Problem 2. Determine the asymptotic behaviour of $Z(2)$ when girth is large. Is it $2^{n\cdot\Theta(1/k)}$, $2^{n\cdot\Theta(\log k/k)}$, or somewhere in between? Note that Theorem 24 gives no clues to the answer. To tackle this problem, one may wish to solve the limiting model on an infinite k-regular tree. That is, an infinite tree where every vertex has k successors and k predecessors.

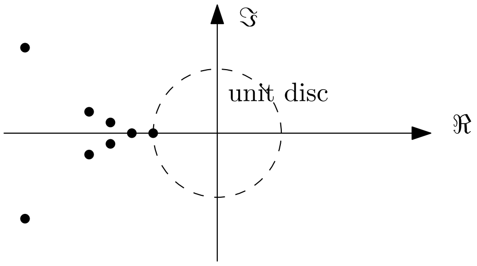

Figure 5–2 The root pattern of $p_G(z)$ on the complex plane

Problem 3. Suppose we fix the spins of some vertices, which we call a "boundary condition". Does the effect on $\mathbb{E}(S_v)$ decay quickly when the distance between boundary and v increases? (In statistical physics, such property is named *correlation decay.*)

Problem 4. Generalise the model and analysis to PPSZ. Our current model has score measure $S_v := \mathbf{1}[x_v \geq \max x_{\flat v}]$ that corresponds to PPZ. For the more complicated PPSZ, let us introduce more terminologies. Fix parameter $h \in \mathbb{N}$ which corresponds to the "power" of PPSZ. Define ∇v to be all ancestors of v within distance h. Set $U \subseteq \nabla v$ is called a v-cut if, starting from v, every maximal simple path in ∇v bumps into U. Finally, $S_v := \mathbf{1}[\exists v\text{-cut } U : x_v \geq \max x_U]$.

5.5 Notes and References

Our polynomial $p_G(z)$ is an extension of the (x, R) trick in Chapter 4. Taking $z := 1$ restores the counting result.

The study of complex zeros of partition functions dates back to the Lee-Yang circle theorem [11] on Ising model. They used the non-analyticity of $\log Z$ as a notion of phase transition, and proved that the Ising model exhibits at most one phase transitions if there is any.

Recently, Barvinok [4, 3] initiated the use of zero-free property in approximating partition functions. Later, the work by Patel and Regts [14] speeds up his method by transforming the power sum, σ_j, into a summation of connected induced subgraph counts. Our exposition loosely follows their method, but with special focus on analytic properties rather than algorithmic implementations. We would like to mention that Liu, Sinclair and Srivastava [13] generalised Patel and Regts' work to the Ising model. Their paper introduced a structure named "insects" which serves the same purpose as our "certificates".

The condition of Theorem 24 can be modified to avoid the root $z = -1$. Roughly speaking, if one could find a $\delta > 0$ such that $p_G(z) \neq 0$ for $(\Re z, \Im z) \in [-\delta, 1+\delta] \times [-\delta, \delta]$, then the conclusion of Theorem 24 still holds. The proof idea is to construct a suitable polynomial $q(z)$ that maps the disc to the aforementioned strip. A detailed proof can be found in Lemma 2.2.3 of Barvinok's book [3].

Bibliography

[1] N. Alon and J. H. Spencer. *The probabilistic method*. John Wiley & Sons, 2004.

[2] A. Bandyopadhyay and D. Gamarnik. Counting without sampling: asymptotics of the log-partition function for certain statistical physics models. *Random Structures & Algorithms*, 33(4):452–479, 2008.

[3] A. Barvinok. *Combinatorics and complexity of partition functions*, volume 9. Springer, 2016.

[4] A. Barvinok. Computing the permanent of (some) complex matrices. *Foundations of Computational Mathematics*, 16(2):329–342, 2016.

[5] E. Cohen, W. Perkins, and P. Tetali. On the widom–rowlinson occupancy fraction in regular graphs. *Combinatorics, Probability and Computing*, 26(2):183–194, 2017.

[6] E. Davies, M. Jenssen, W. Perkins, and B. Roberts. Independent sets, matchings, and occupancy fractions. *Journal of the London Mathematical Society*, 96(1):47–66, 2017.

[7] C. M. Fortuin, P. W. Kasteleyn, and J. Ginibre. Correlation inequalities on some partially ordered sets. *Communications in Mathematical Physics*, 22(2):89–103, 1971.

[8] S. Friedli and Y. Velenik. *Statistical mechanics of lattice systems: a concrete mathematical introduction*. Cambridge University Press, 2017.

[9] M. R. Jerrum, L. G. Valiant, and V. V. Vazirani. Random generation of combinatorial structures from a uniform distribution. *Theoretical computer science*, 43:169–188, 1986.

[10] J. Kahn. An entropy approach to the hard-core model on bipartite graphs. *Combinatorics, Probability & Computing*, 10(3):219, 2001.

[11] T.-D. Lee and C.-N. Yang. Statistical theory of equations of state and phase transitions. ii. lattice gas and ising model. *Physical Review*, 87(3):410, 1952.

[12] D. A. Levin and Y. Peres. *Markov chains and mixing times*, volume 107. American Mathematical Soc., 2017.

[13] J. Liu, A. Sinclair, and P. Srivastava. The ising partition function: zeros and deterministic approximation. *Journal of Statistical Physics*, 174(2):287–315, 2019.

[14] V. Patel and G. Regts. Deterministic polynomial-time approximation algorithms for partition functions and graph polynomials. *SIAM Journal on Computing*, 46(6):1893–1919, 2017.

[15] R. Paturi, P. Pudlák, M. E. Saks, and F. Zane. An improved exponential-time algorithm for k-sat. *Journal of the ACM (JACM)*, 52(3):337–364, 2005.

[16] R. Paturi, P. Pudlák, and F. Zane. Satisfiability coding lemma. In *Proceedings 38th Annual Symposium on Foundations of Computer Science*, pages 566–574. IEEE, 1997.

[17] D. Tong. Lecture notes on statistical physics. http://www.damtp.cam.ac.uk/user/tong/statphys.html.

[18] D. Weitz. Counting independent sets up to the tree threshold. In *Proceedings of the thirty-eighth annual ACM symposium on Theory of computing*, pages 140–149, 2006.

Acknowledgements

I would like to give my foremost thanks to my advisor, Professor Dominik Scheder, for his enthusiastic guidance over the year. Through our discussions, I had a unique journey of experiencing interactive proof *in person*. I always learn something new in his office live show, seeing him grab a piece of paper and begin explaining ideas. The scenes gave me valuable, first-hand examples about how things pop up and how analyses are carried through by a computer scientist.

I am also grateful to Professor Chihao Zhang for guiding me through sampling and counting algorithms last year. His seminar marked a special start for my study in partition functions.

The formatting of this final version partly owes to a template by the *SJTUThesis* project; many thanks to their community work! Finally, my thanks go to Otfried Cheong, the author of a handy drawing editor called Ipe. Without it, creating illustrations for this thesis would be far less enjoyable.