CAMBRIDGE
UNIVERSITY PRESS

ARTICLE

Independent sets of a given size and structure in the hypercube

Matthew Jenssen[1], Will Perkins[2]* and Aditya Potukuchi[2]

[1]School of Mathematics, University of Birmingham, Birmingham, UK and [2]Department of Mathematics, Statistics, and Computer Science, University of Illinois at Chicago, Chicago, IL 60607, USA
*Corresponding author. E-mail: william.perkins@gmail.com

Abstract

We determine the asymptotics of the number of independent sets of size $\lfloor \beta 2^{d-1} \rfloor$ in the discrete hypercube $Q_d = \{0, 1\}^d$ for any fixed $\beta \in (0, 1)$ as $d \to \infty$, extending a result of Galvin for $\beta \in (1 - 1/\sqrt{2}, 1)$. Moreover, we prove a multivariate local central limit theorem for structural features of independent sets in Q_d drawn according to the hard-core model at any fixed fugacity $\lambda > 0$. In proving these results we develop several general tools for performing combinatorial enumeration using polymer models and the cluster expansion from statistical physics along with local central limit theorems.

1. Introduction

Let Q_d be the discrete hypercube: the graph with vertex set $\{0, 1\}^d$ in which two vectors are joined by an edge if they differ in exactly one coordinate. An independent set is a set of vertices that contains no edge. Let $\mathcal{I}(Q_d)$ be the set of independent sets of Q_d and let $i(Q_d) = |\mathcal{I}(Q_d)|$ be the number of independent sets of the hypercube. The vertices of Q_d can be divided into two sets, those whose coordinates sum to an even number and those whose coordinates sum to an odd number. This partition shows that Q_d is a bipartite graph. We let $N := 2^{d-1}$ be the number of even (or odd) vertices of Q_d. A trivial lower bound on $i(Q_d)$ is $2 \cdot 2^N - 1$ obtained by considering independent sets of only even or only odd vertices. A better lower bound is obtained by considering independent sets with an arbitrary (but constant) number of 'defect' vertices on one side of the bipartition. This increases the lower bound by a factor $\sqrt{e}$. Korshunov and Sapozhenko showed that this gives the correct asymptotics for $i(Q_d)$ as $d \to \infty$ [16].

Theorem 1 (Korshunov and Sapozhenko). *As $d \to \infty$,*

$$i(Q_d) = \left(2\sqrt{e} + o(1)\right) 2^N.$$

Galvin later studied weighted independent sets in the hypercube. For $\lambda \geq 0$, define the independence polynomial of Q_d,

$$Z_{Q_d}(\lambda) = \sum_{I \in \mathcal{I}(Q_d)} \lambda^{|I|}.$$

Taking $\lambda = 1$ recovers $i(Q_d)$. In what follows we will drop Q_d from the notation, writing $Z(\lambda)$ and $\mathcal{I}$ for $Z_{Q_d}(\lambda)$ and $\mathcal{I}(Q_d)$.

The independence polynomial $Z(\lambda)$ is also the partition function of the hard-core model on Q_d: the probability distribution μ_λ on $\mathcal{I}$ defined by $\mu_\lambda(I) = \lambda^{|I|}/Z(\lambda)$. By generalising Sapozhenko's alternative proof of Theorem 1 in [20], Galvin found the asymptotics for $Z(\lambda)$ for $\lambda > \sqrt{2} - 1$ [10] (as well as the asymptotics of $\log Z(\lambda)$ for $\lambda = \Omega(\log d/d^{1/3})$).

Theorem 2 (Galvin) *For* $\lambda > \sqrt{2} - 1$,

$$Z(\lambda) = (2 + o(1))(1 + \lambda)^N \exp\left[\lambda N \left(\frac{1}{1 + \lambda}\right)^d\right]$$

as $d \to \infty$.

Analogously to Theorem 1, the trivial lower bound for $Z(\lambda)$ is $2(1 + \lambda)^N - 1$, and the asymptotic formula in Theorem 2 includes the contribution from independent sets with a constant number of defect vertices, captured by the exponential factor.

Galvin also studied the typical structure of the defect vertices under the probability distribution μ_λ. Formally, given an independent set $I \in \mathcal{I}$, if $|I \cap \mathcal{O}| \leq |I \cap \mathcal{E}|$, we refer to the elements of $I \cap \mathcal{O}$ as the *defect* vertices of I; otherwise we say that $I \cap \mathcal{E}$ is the set of defect vertices. A natural way to describe the structure of a set $S \subset \mathcal{O}, \mathcal{E}$ is to describe the graph $Q_d^2[S]$ where Q_d^2 denotes the square of Q_d. Given an independent set I with defect vertices S, we refer to the connected components of $Q_d^2[S]$ as the *defects* of I. Galvin showed that for $\lambda > \sqrt{2} - 1$, all but a vanishing fraction of $Z(\lambda)$ comes from independent sets with defects of size at most 1.

Recently, the first two authors found the asymptotics of $Z(\lambda)$ for all fixed $\lambda > 0$ [15]. The asymptotic formula takes into account defects of arbitrary, but constant size. The smaller λ is, the larger the size of defects that must be considered.

Theorem 3 (Jenssen and Perkins) *There is a sequence of polynomials* $R_j(d, \lambda)$, $j \in \mathbb{N}$, *such that for any fixed* $t \geq 1$ *and* $\lambda > 2^{1/t} - 1$,

$$Z(\lambda) = (2 + o(1))(1 + \lambda)^N \exp\left[N \sum_{j=1}^{t-1} R_j(d, \lambda)(1 + \lambda)^{-dj}\right]$$

as $d \to \infty$. *Moreover the coefficients of the polynomial* R_j *can be computed in time* $e^{O(j \log j)}$.

In particular, $R_1 = \lambda$, recovering the formula in Theorem 2.

Given these results it is natural to ask for the asymptotics of $i_m(Q_d)$, the number of independent sets of size m in Q_d. There is a trivial lower bound of $i_m(Q_d) \geq 2\binom{N}{m}$, obtained by considering independent sets composed entirely of even or odd vertices, but depending on how large m is, we may need to take into account independent sets of size m with defects up to a given size.

Galvin [11] gave the asymptotics of $i_m(Q_d)$ in the range for which almost all independent sets of size m contain defects of size at most 1.

Theorem 4. (Galvin) *Fix* $\beta \in (1 - 1/\sqrt{2}, 1)$ *and let* $\lambda = \frac{\beta}{1-\beta}$. *Then*

$$i_{\lfloor \beta N \rfloor}(Q_d) = (2 + o(1))\binom{N}{\lfloor \beta N \rfloor} \exp\left[\lambda N \left(\frac{1}{1 + \lambda}\right)^d\right]. \tag{1}$$

as $d \to \infty$.

Note that the asymptotic formula (1) consists of the trivial lower bound $2\binom{N}{\lfloor \beta N \rfloor}$ multiplied by the same exponential correction factor in the asymptotic formula for $Z(\lambda)$ in the range $\lambda > \sqrt{2} - 1$.

We show that a similar, but more complicated, formula holds for all $\beta > 0$. In particular, when $\beta < 1 - 1/\sqrt{2}$ the formula is *not* simply the trivial bound multiplied by the appropriate exponential correction factor from Theorem 3. We explain below where the extra complexity arises.

Theorem 5. *There is a sequence of rational functions $P_j(d, \beta)$, $j \in \mathbb{N}$, such that so that for any fixed $t \geq 1$ and $\beta \in (1 - 2^{-1/t}, 1)$,*

$$i_{\lfloor \beta N \rfloor}(Q_d) = (2 + o(1)) \binom{N}{\lfloor \beta N \rfloor} \exp\left[N \sum_{j=1}^{t-1} P_j(d, \beta) \cdot (1 - \beta)^{jd} \right] \tag{2}$$

as $d \to \infty$. Moreover the coefficients of P_j can be computed in time $e^{O(j \log j)}$.

For small values of j the functions P_j can be computed by hand. For example, $P_1 = \frac{\beta}{1-\beta}$, and taking $t = 2$ in Theorem 5 recovers Galvin's Theorem 4. A more involved calculation carried out in Section 3 yields

$$P_2 = \frac{2(\beta - 1)^3 \beta - d(d - 1)(\beta - 2)\beta^3}{4(1 - \beta)^4} - \frac{\beta(1 - d\beta)^2}{2(1 - \beta)^3} .$$

By Theorem 5 this gives an explicit asymptotic formula for $i_{\lfloor \beta N \rfloor}(Q_d)$ for $\beta > 1 - 2^{-1/3}$:

$$i_{\lfloor \beta N \rfloor}(Q_d) \sim 2 \binom{N}{\lfloor \beta N \rfloor} \exp\left[N \frac{\beta}{1 - \beta}(1 - \beta)^d \right.$$
$$\left. + N \left(\frac{2(\beta - 1)^3 \beta - d(d - 1)(\beta - 2)\beta^3}{4(1 - \beta)^4} - \frac{\beta(1 - d\beta)^2}{2(1 - \beta)^3} \right) (1 - \beta)^{2d} \right] .$$

In principle, one can continue to compute $P_3, P_4, \dots$ and obtain explicit asymptotics for any fixed β. More generally, the results of [15] and of this paper hold for much smaller λ and β, tending to 0 as $d \to \infty$, as long as $\lambda \geq C \log d / d^{1/3}$ and $\beta > C \log d / d^{1/3}$ for an absolute constant C. In this case, however, the asymptotic formulas in Theorems 3 and 5 become series with a number of terms that grows with d. These series can be used to give an algorithm to approximate $Z(\lambda)$ and $i_{\lfloor \beta N \rfloor}(Q_d)$ up to a $(1 + \varepsilon)$ multiplicative factor in time polynomial in $1/\varepsilon$ and N (an FPTAS in the language of approximate counting; see e.g. [14] for such an algorithm for independent sets in expander graphs). This raises an interesting question of what it means to determine the asymptotics of a sequence $f(d)$ as $d \to \infty$. Evaluating a closed-form expression involving, say, exponentials or logarithms, might also involve truncating a power series, and so in a sense an algorithmic definition is natural. We do not pursue this further here and instead stick with β constant.

The proof of Theorem 5 makes use of the following simple yet useful identity. Let $\mathcal{I}_m = \{I \in \mathcal{I} : |I| = m\}$ so that $i_m(Q_d) = |\mathcal{I}_m|$. For $m \in \mathbb{N}$ and $\lambda > 0$,

$$i_m(Q_d) = \frac{Z(\lambda)}{\lambda^m} \mu_\lambda(\mathcal{I}_m) . \tag{3}$$

In fact this formula follows from the definition of μ_λ and so holds for any graph, not just Q_d. To use (3) along with Theorem 3 to derive asymptotics for $i_m(Q_d)$, we must compute the asymptotics of $\mu_\lambda(\mathcal{I}_m)$. The feasibility of doing this depends very much on m and the choice of λ. By choosing λ so that the expected size of an independent set drawn from μ_λ is approximately m, we can compute the asymptotics of $\mu_\lambda(\mathcal{I}_m)$ using a local central limit theorem. In practice, we do not work

with the hard-core model directly, but with an approximating measure derived from a polymer model which describes the distribution on defects in an independent set from the hard-core model (see Section 2). This polymer model was introduced in [15].

In the next theorem we give an expansion in $(1 - \beta)^d$ for a value of the fugacity λ for which the expected size of $\mathbf{I}$, a random sample from the hard-core model, is close to $\lfloor \beta N \rfloor$. By expanding the formula (5) below and combining this with (4) we obtain Theorem 5.

Theorem 6 *There exists a sequence of rational functions $B_j(d, \beta)$, $j \in \mathbb{N}$, such that the coefficients of B_j can be computed in time $e^{O(j \log j)}$ and the following holds. Fix $\beta \in (0, 1)$ and let $t \geq 1$ be such that $\beta > 1 - 2^{-1/t}$. Then with*

$$\lambda_\beta = \frac{\beta}{1 - \beta} + \sum_{j=1}^{t-1} B_j(d, \beta) \cdot (1 - \beta)^{jd} \tag{4}$$

we have

$$i_{\lfloor \beta N \rfloor}(Q_d) = \frac{1 + o(1)}{\sqrt{2\pi N \beta (1 - \beta)}} \frac{Z(\lambda_\beta)}{\lambda_\beta^{\lfloor \beta N \rfloor}} \tag{5}$$

as $d \to \infty$. Moreover,

$$\left| \mathbb{E}_{\lambda_\beta} |\mathbf{I}| - \lfloor \beta N \rfloor \right| = o(N^{1/2}).$$

In [15] the authors prove a multivariate central limit theorem for the number of defects of different types in the polymer model. In Section 4, we establish a multivariate *local* central limit theorem for this polymer model which allows us to refine Theorems 5 and 6 further still. Given a defect S, we define the *type* of S to be the isomorphism class of the graph $Q_d^2[S]$. For a given defect type T, we let X_T be the random variable that counts the number of defects of type T in a sample from the hard-core model on Q_d. We let $m_T = \mathbb{E}_\lambda X_T$.

Given a collection $\mathcal{T}$ of types and vector of non-negative integers $\mathbf{k} = (k_T)_{T \in \mathcal{T}}$, let $i_{m,\mathbf{x}}(Q_d)$ denote the number of independent sets in Q_d of size m with exactly k_T defects of type T for all $T \in \mathcal{T}$.

Theorem 7 *Fix $\beta \in (0, 1)$ and let $\lambda = \lambda_\beta$ be as in Theorem 6. Let $\mathcal{T}_1$ be the set of defect types T such that $m_T \to \rho_T$ for some constant $\rho_T > 0$ as $d \to \infty$ and $\mathcal{T}_2$ the set of defect types T so that $m_T \to \infty$. Let $(k_T)_{T \in \mathcal{T}_1}$ be a vector of fixed non-negative integers and let $(k_T)_{T \in \mathcal{T}_2}$ be such that $k_T = \lfloor m_T + s_T \rfloor$ where $|s_T| = O(\sqrt{m_T})$ for all $T \in \mathcal{T}_2$. Let $\mathbf{k} = (k_T)_{T \in \mathcal{T}_1 \cup \mathcal{T}_2}$. Then*

$$i_{\lfloor \beta N \rfloor, \mathbf{k}}(Q_d) = \frac{1 + o(1)}{\sqrt{2\pi N \beta (1 - \beta)}} \frac{Z(\lambda_\beta)}{\lambda_\beta^{\lfloor \beta N \rfloor}} \prod_{T \in \mathcal{T}_1} \frac{\rho_T^{k_T} e^{-\rho_T}}{(k_T)!} \prod_{T \in \mathcal{T}_2} \frac{e^{-\frac{s_T^2}{2m_T}}}{\sqrt{2\pi m_T}} \tag{6}$$

as $d \to \infty$.

This formula matches that of (5) with additional factors corresponding to Poisson probabilities (for $T \in \mathcal{T}_1$) and local central limit theorem probabilities (for $T \in \mathcal{T}_2$).

1.1 Methods: maximum entropy, statistical mechanics and local central limit theorems

The methods we use here combine several different probabilistic tools, including abstract polymer models and the cluster expansion, large deviations and local central limit theorems. Counting independent sets in the hypercube is a canonical combinatorial enumeration problem, and so we hope this provides a template for using this combination of tools in other combinatorial problems.

There is a long history of using local central limit theorems in combinatorics, with many examples in analytic combinatorics and the study of integer partitions (see e.g. [7,17–19]) as well as the enumeration of contingency tables [4] and graphs with prescribed degree sequences (e.g. [5,12]). Here we show that local central limit theorems work very well in combination with two tools from statistical physics, polymer models and the cluster expansion, which have been used recently in combinatorial enumeration [2,13,15].

The connection between these methods starts with a general approach to counting via probability and the principle of maximum entropy which is laid out explicitly by Barvinok and Hartigan in [3] (and later discussed in [17]), but appears implicitly in other enumerations methods, such as the circle method (see [12] for an explanation of these connections). The main idea is that to count a subset of objects defined by a number of constraints, one considers the maximum entropy distribution on the larger set that satisfies the constraints in expectation. The size of the subset can then be expressed as the exponential of the entropy of this distribution times the probability that a random object drawn from this distribution satisfies the constraints. In the example of enumerating integer partitions, these maximum entropy distributions take the form of sequences of independent geometric random variables with different means [1,9], and asymptotic enumeration can be accomplished by solving a convex optimisation problem to find these means and then proving a local central limit theorem for linear combinations of independent geometric random variables [7,17–19].

This approach naturally leads to considering statistical physics models. For example, the maximum entropy distribution over independent sets in a graph with a given mean size is the hard-core model. The entropy of the hard-core model is the log partition function minus the expected size of an independent set: $H(\mu_\lambda) = \log Z(\lambda) - \log \lambda \cdot \mathbb{E}_\mu |\mathbf{I}|$, and so the enumeration problem for independent sets of a given size reduces to computing $\log Z$ and computing $\mu_\lambda(\mathcal{I}_k)$ (via, say, a local central limit theorem) as described above.

The complication is that the quantities of interest (say, the size of an independent set from the hard-core model) can no longer be written as the sum of independent random variables. When interactions are weak enough (or density small enough) correlations between vertices decay exponentially in distance and methods like the cluster expansion can be used to prove both central limit theorems and local central limit theorem. This type of result is closely related to the concept of equivalence of ensembles between the grand canonical ensemble (fixed mean energy) and the canonical ensemble (fixed energy). For instance, Dobrushin and Tirozzi showed that for spin models with finite-range interactions on $\mathbb{Z}^d$, a central limit theorem implies a local central limit theorem [8] (see also [6] for an extension to long-range interactions).

What we do here is prove local central limit theorems conditioned on a phase in the strong interaction, phase coexistence regime. This, in combination with using polymer models and the cluster expansion to find the asymptotics of $Z(\lambda)$, allows us to enumerate independent sets of a given size and structure.

In Section 2, we recall the even and odd polymer models introduced in [15], and state and extend some of the probabilistic estimates proved there.

In Section 3 we prove Theorems 5 and 6, finding an expansion for a fugacity λ_β so that the expected size of an independent set is sufficiently close to $\lfloor \beta N \rfloor$ that a local central limit theorem will allow us to compute asymptotics.

In Section 4 we show how local central limit theorems for polymer models follow from sufficiently fast convergence of the cluster expansion.

Finally in Section 5 we combine the above results to prove Theorem 7.

2. The even and odd polymer models

Let $\mathcal{E} \subset V(Q_d)$ be the set of even vertices of the hypercube, those whose coordinates sum to an even number, and let $\mathcal{O} \subset V(Q_d)$ be the odd vertices. Note that Q_d is a bipartite graph with

bipartition $(\mathcal{E}, \mathcal{O})$ and that $|\mathcal{E}| = |\mathcal{O}| = N := 2^{d-1}$. A key insight of [15] is that, for λ not too small, the hard-core measure $\mu_{Q_d, \lambda}$ can be closely approximated by a random perturbation of a random subset of either $\mathcal{O}$ or $\mathcal{E}$. The random perturbation takes the form of a polymer model with convergent cluster expansion, two notions that we introduce now.

For a set $S \subseteq \mathcal{O}$ (and analogously for $S \subseteq \mathcal{E}$), let $|S|$ denote the number of vertices of S, $N(S)$ be the set of neighbours of S and $[S] = \{v \in \mathcal{O} : N(v) \subseteq N(S)\}$ the bipartite closure of S. We call a set $S \subseteq \mathcal{O}$ an *odd polymer* if (i) the subgraph of Q_d induced by the vertex set $S \cup N(S)$ is connected and (ii) $|[S]| \le N/2$. We let $\mathcal{P}_{\mathcal{O}}$ denote the set of all odd polymers. The *weight* of an odd polymer S is

$$w(S) = \frac{\lambda^{|S|}}{(1 + \lambda)^{|N(S)|}} . \tag{7}$$

We say that two odd polymers S_1 and S_2 are *compatible* and write $S_1 \sim S_2$, if the graph distance between S_1, S_2 is Q_d is > 2. We let $\Omega_{\mathcal{O}}$ denote the set of all collections of mutually compatible odd polymers and define the following Gibbs measure on $\Omega_{\mathcal{O}}$: for $\Gamma \in \Omega_{\mathcal{O}}$,

$$\nu_{\mathcal{O}}(\Gamma) = \frac{\prod_{S \in \Gamma} w(S)}{\Xi_{\mathcal{O}}} \quad \text{where} \quad \Xi_{\mathcal{O}} = \sum_{\Gamma \in \Omega_{\mathcal{O}}} \prod_{S \in \Gamma} w(S)$$

is the odd polymer model partition function. Using $\nu_{\mathcal{O}}$ we define a measure $\mu_{\mathcal{O}, \lambda}$ on independent sets in Q_d.

Definition 8 Let $\mu_{\mathcal{O}, \lambda}$ be the measure on $\mathcal{I}$ defined by the following two-step process:

1. Choose a polymer configuration $\Gamma \in \Omega_{\mathcal{O}}$ from $\nu_{\mathcal{O}}$ and assign all vertices of $\cup_{S \in \Gamma} S$ to be occupied.
2. For each vertex v in $\mathcal{E}$ that is not blocked by an occupied vertex in $\mathcal{O}$, include v in the independent set independently with probability $\frac{\lambda}{1+\lambda}$.

Let $Z_{\mathcal{O}}(\lambda) = (1 + \lambda)^N \Xi_{\mathcal{O}}$, the independence polynomial of Q_d restricted to independent sets achievable in the odd polymer model; that is, those for which $\mu_{\mathcal{O}, \lambda}$ assigns positive probability.

We think of Step 1 in Definition 8 as a perturbation of the 'ground state' measure that simply selects a p-random subset of $\mathcal{E}$ with $p = \lambda/(1 + \lambda)$. The polymer configuration chosen in Step 1 will be typically small and so this process typically returns an independent set that is highly unbalanced with the majority of vertices even. We define even polymers, the even polymer model partition function $\Xi_{\mathcal{E}}$, and measures $\nu_{\mathcal{E}}$, $\mu_{\mathcal{E}, \lambda}$ analogously. It was shown in [15] that for $\lambda > C \log d/d^{1/3}$, the hard-core measure $\mu_{Q_d, \lambda}$ can be closely approximated by the mixture $\frac{1}{2}\mu_{\mathcal{O}, \lambda} + \frac{1}{2}\mu_{\mathcal{E}, \lambda}$.

Theorem 9 ([15]) *For $\lambda \ge C \log d/d^{1/3}$, we have*

$$\left| \log Z(\lambda) - \log \left[2 Z_{\mathcal{O}}(\lambda) \right] \right| = O \left(\exp \left(- N/d^4 \right) \right) . \tag{8}$$

Moreover, letting $\hat{\mu}_\lambda = \frac{1}{2}\mu_{\mathcal{O}, \lambda} + \frac{1}{2}\mu_{\mathcal{E}, \lambda}$, we have

$$\| \mu_\lambda - \hat{\mu}_\lambda \|_{TV} = O \left(\exp \left(- N/d^4 \right) \right) .$$

Finally, with probability at least $1 - O(\exp(-N/d^4))$ any defect vertices of I drawn from $\mu_{\mathcal{O}, \lambda}$ are on the odd side of the bipartition; that is, the defects are the polymers of the polymer configuration.

The lower bound on λ in Theorem 9 is an artefact of Sapozhenko's graph container method as implemented by Galvin [10]. Theorem 9 quite possibly remains true for $\lambda = \tilde{\Omega}(1/d)$ though proving this would require significant new ideas.

The power of Theorem 9 stems from the fact that the even and odd polymer models admit convergent cluster expansions allowing for a detailed understanding of the measures $\mu_{\mathcal{O},\lambda}, \mu_{\mathcal{E},\lambda}$. Let us now introduce the cluster expansion formally.

For a tuple Γ of odd polymers, the *incompatibility graph*, $H(\Gamma)$, is the graph with vertex set Γ and an edge between any two incompatible polymers. An odd *cluster* Γ is an ordered tuple of even polymers so that $H(\Gamma)$ is connected. The size of a cluster Γ is $\|\Gamma\| = \sum_{S \in \Gamma} |S|$. Let $\mathcal{C}$ be the set of all odd clusters. For a cluster Γ we define

$$w(\Gamma) = \phi(H(\Gamma)) \prod_{S \in \Gamma} w(S),$$

where $\phi(H)$ is the *Ursell function* of a graph H, defined by

$$\phi(H) = \frac{1}{|V(H)|!} \sum_{\substack{A \subseteq E(H) \\ \text{spanning, connected}}} (-1)^{|A|}. \tag{9}$$

The cluster expansion is the formal infinite series

$$\log \Xi_{\mathcal{O}} = \sum_{\Gamma \in \mathcal{C}} w(\Gamma). \tag{10}$$

We define the cluster expansion of $\log \Xi_{\mathcal{E}}$ analogously and note that by symmetry the expansions are identical.

In light of Theorem 9, throughout this section will we assume that $\lambda \geq C \log d / d^{1/3}$. We will also assume that $\lambda = O(1)$ as $d \to \infty$. The following result from [15] shows that for such λ the cluster expansion converges, we have good tail bounds on the expansion, and the terms of the cluster expansion can be efficiently computed. We say that a polynomial is *computable in time t*, if its coefficients can be computed in time t.

Theorem 10 *For fixed $k \geq 1$,*

$$\sum_{\substack{\Gamma \in \mathcal{C} \\ \|\Gamma\| \geq k}} |w(\Gamma)| = O\left(\frac{2^d d^{2(k-1)}}{(1+\lambda)^{dk}}\right)$$

and

$$\sum_{\substack{\Gamma \in \mathcal{C} \\ \|\Gamma\| = k}} w(\Gamma) = N \cdot R_k(\lambda, d)(1+\lambda)^{-kd} \tag{11}$$

where R_k is a polynomial in d and λ of degree at most $2k$ in d and of degree at most $3k^2$ in λ. Moreover R_k is computable in time $e^{O(k \log k)}$. In particular,

$$\log \Xi_{\mathcal{O}} = N \sum_{j=1}^{k} R_j(\lambda, d)(1+\lambda)^{-jd} + O\left(\frac{2^d d^{2k}}{(1+\lambda)^{d(k+1)}}\right).$$

We note that Theorem 3 follows in [15] from Theorems 9 and 10.

It will also be useful to record the following slightly strengthened tail bound on the cluster expansion. The following lemma is essentially contained in [15], though it does not appear explicitly and so we provide the details.

Lemma 11 *For $t, \ell \geq 1$ fixed,*

$$\sum_{\substack{\Gamma \in \mathcal{C} \\ \|\Gamma\| \geq t}} |w(\Gamma)| \|\Gamma\|^{\ell} = O\left(\frac{2^d d^{2(t-1)}}{(1+\lambda)^{dt}}\right).$$

Proof. In [15] (see Lemma 15) it is shown that

$$\sum_{\Gamma \in \mathcal{C}} |w(\Gamma)| e^{\gamma(d, \|\Gamma\|)} \leq 2^{d-1} d^{-3/2} \,,$$

where

$$\gamma(d, k) = \begin{cases} \log{(1+\lambda)}(dk - 3k^2) - 7k \log d \text{ if } k \leq \dfrac{d}{10} \\[2ex] \dfrac{d \log{(1+\lambda)}k}{20} \text{ if } \dfrac{d}{10} < k \leq d^4 \\[2ex] \dfrac{k}{d^{3/2}} \text{ if } k > d^4 \,. \end{cases}$$

Since $e^{\gamma(d,k)/2} \geq k^\ell$ for all k and d sufficiently large, it follows that

$$\sum_{\Gamma \in \mathcal{C}} |w(\Gamma)| \|\Gamma\|^\ell e^{\gamma(d, \|\Gamma\|)/2} = O(2^d d^{-3/2}) \,.$$

Keeping only terms in the above inequality corresponding to clusters of size at least k we have

$$\sum_{\substack{\Gamma \in \mathcal{C} \\ \|\Gamma\| \geq k}} |w(\Gamma)| \|\Gamma\|^\ell \leq O(2^d d^{-3/2} e^{-\gamma(d,k)/2}) \,. \tag{12}$$

With $t \geq 0$ fixed we have by Theorem 10 and (12) that

$$\sum_{\substack{\Gamma \in \mathcal{C} \\ \|\Gamma\| \geq t}} |w(\Gamma)| \|\Gamma\|^\ell = \sum_{\substack{\Gamma \in \mathcal{C} \\ t \leq \|\Gamma\| < 3t}} |w(\Gamma)| \|\Gamma\|^\ell + \sum_{\substack{\Gamma \in \mathcal{C} \\ \|\Gamma\| \geq 3t}} |w(\Gamma)| \|\Gamma\|^\ell$$

$$= O\left(\frac{2^d d^{2(t-1)}}{(1+\lambda)^{dt}}\right) + O\left(\frac{2^d d^{11t}}{(1+\lambda)^{3dt/2}}\right)$$

$$= O\left(\frac{2^d d^{2(t-1)}}{(1+\lambda)^{dt}}\right) \,.$$

Let $\boldsymbol{\Gamma}$ be a collection of compatible polymers sampled according to $\nu_\mathcal{O}$ (the polymer measure at Step 1 of Definition 8, the definition of $\mu_{\mathcal{O},\lambda}$). We will use the above lemma to show that $\|\boldsymbol{\Gamma}\|$ and $|N(\boldsymbol{\Gamma})|$ obey a central limit theorem. Formally, we say a sequence of random variables (X_d) obeys a central limit theorem if $(X_d - \mathbb{E}X_d)/\sqrt{\text{var}(X_d)}$ converges to $N(0,1)$ in distribution as $d \to \infty$. To prove this central limit theorem we will make use of a connection between the cluster expansion and *cumulant generating functions*.

Recall the cumulant generating function of a random variable X, is $h_t(X) = \log \mathbb{E}e^{tX}$. The ℓth *cumulant* of X is defined by taking derivatives of $h_t(X)$ and evaluating at 0:

$$\kappa_\ell(X) = \left.\frac{\partial^\ell h_t(X)}{\partial t^\ell}\right|_{t=0} \,.$$

In particular, $\kappa_1(X) = \mathbb{E}(X)$ and $\kappa_2(X) = \text{var}(X)$. The cumulants of $|N(\Gamma)|$ can be expressed in terms of the cluster expansion as follows. Consider the odd polymer model with modified weights $w_t(S) = w(S)e^{t|N(S)|}$ for $t > 0$ and let Ξ_t denote the corresponding partition function. We then have

$$h_t(|N(\boldsymbol{\Gamma})|) = \log \Xi_t - \log \Xi_\mathcal{O} \,.$$

Applying the cluster expansion to $\log \Xi_t$, taking derivatives, and evaluating at $t = 0$ shows that

$$\kappa_\ell(|N(\mathbf{\Gamma})|) = \sum_{\Gamma \in \mathcal{C}} w(\Gamma)|N(\Gamma)|^\ell \,. \tag{13}$$

Similarly $\kappa_\ell(\|\mathbf{\Gamma}\|) = \sum_{\Gamma \in \mathcal{C}} w(\Gamma)\|\Gamma\|^\ell$.

Lemma 12 *Let $\mathbf{\Gamma}$ be a collection of compatible polymers sampled according to $\nu_{\mathcal{O}}$. Then $\|\mathbf{\Gamma}\|$ and $|N(\mathbf{\Gamma})|$ both obey a central limit theorem.*

Proof. We show that $|N(\mathbf{\Gamma})|$ obeys a central limit theorem and the proof for $\|\mathbf{\Gamma}\|$ is identical. Let $Z = (|N(\mathbf{\Gamma})| - \mathbb{E}|N(\mathbf{\Gamma})|)/\sqrt{\mathrm{var}(|N(\mathbf{\Gamma})|)}$. To show that Z converges to $N(0,1)$ in distribution, it suffices to show that the cumulants of Z converge to the cumulants of a standard normal i.e. it suffices to show that $\kappa_\ell(Z) \to 0$ for $\ell \geq 3$. Now by (13) and Lemma 11,

$$\kappa_\ell(|N(\mathbf{\Gamma})|) = \sum_{\Gamma \in \mathcal{C}} w(\Gamma)|N(\Gamma)|^\ell \leq d^\ell \sum_{\Gamma \in \mathcal{C}} w(\Gamma)\|\Gamma\|^\ell = O\left(\frac{2^d d^\ell}{(1+\lambda)^d}\right) \,.$$

On the other hand, by (13) and Lemma 11 again we have

$$\mathrm{var}(|N(\mathbf{\Gamma})|) = \sum_{\Gamma \in \mathcal{C}} w(\Gamma)|N(\Gamma)|^2 \geq \sum_{\Gamma \in \mathcal{C}} w(\Gamma)\|\Gamma\|^2 = 2^{d-1}\frac{\lambda}{(1+\lambda)^d} + O\left(\frac{2^d d^2}{(1+\lambda)^{2d}}\right) \,.$$

It follows that if $\ell \geq 3$ then

$$\kappa_\ell(Z) = \mathrm{var}(|N(\mathbf{\Gamma})|)^{-\ell/2}\kappa_\ell(|N(\mathbf{\Gamma})|) \to 0$$

as desired. $\qquad\square$

Next we will use the cluster expansion to give bounds on $\mathbb{E}_{\mathcal{O},\lambda}(|\mathbf{I}|)$, the expected size of an independent set sampled according to $\mu_{\mathcal{O},\lambda}$. We begin by recording a useful expression for $\mathbb{E}_{\mathcal{O},\lambda}(|\mathbf{I}|)$ in terms of the cluster expansion.

Lemma 13

$$\mathbb{E}_{\mathcal{O},\lambda}(|\mathbf{I}|) = \frac{\lambda}{1+\lambda}N + \sum_{\Gamma \in \mathcal{C}} w(\Gamma)\left(\|\Gamma\| - \frac{\lambda}{1+\lambda}|N(\Gamma)|\right) \,.$$

Proof. Note that for an independent set I such that $\mu_{\mathcal{O},\lambda}(I) > 0$, we have

$$\mu_{\mathcal{O},\lambda}(I) = \frac{\lambda^{|I|}}{Z_{\mathcal{O}}(\lambda)} = \frac{\lambda^{|I|}}{(1+\lambda)^N \Xi_{\mathcal{O}}} \,.$$

It follows that

$$\mathbb{E}_{\mathcal{O},\lambda}(|\mathbf{I}|) = \sum_I \frac{|I|\lambda^{|I|}}{(1+\lambda)^N \Xi_{\mathcal{O}}} = \lambda\frac{d}{d\lambda}\log\left((1+\lambda)^N \Xi_{\mathcal{O}}\right) = \frac{\lambda}{1+\lambda}N + \lambda(\log \Xi_{\mathcal{O}})' \,.$$

We expand $\log \Xi_{\mathcal{O}}$ via the cluster expansion as in (10) (which converges absolutely by Theorem 10). Recalling that $w(\Gamma) = \phi(H(\Gamma))\lambda^{\|\Gamma\|}(1+\lambda)^{-|N(\Gamma)|}$ for a cluster Γ, we have

$$\mathbb{E}_{\mathcal{O},\lambda}(|\mathbf{I}|) = \frac{\lambda}{1+\lambda}N + \sum_{\Gamma \in \mathcal{C}} \phi(H(\Gamma))\frac{\lambda^{|\Gamma|}((1+\lambda)\|\Gamma\| - \lambda|N(\Gamma)|)}{(1+\lambda)^{|N(\Gamma)|+1}}$$

$$= \frac{\lambda}{1+\lambda}N + \sum_{\Gamma \in \mathcal{C}} w(\Gamma)\left(\|\Gamma\| - \frac{\lambda}{1+\lambda}|N(\Gamma)|\right) \,. \qquad\square$$

Corollary 14 *For fixed $k \geq 0$,*

$$\mathbb{E}_{\mathcal{O},\lambda}(|\mathbf{I}|) = \frac{\lambda}{1+\lambda}N + \lambda N \sum_{j=1}^{k} \frac{\frac{\partial}{\partial \lambda} R_j(\lambda, d)}{(1+\lambda)^{jd}} - \frac{jd R_j(\lambda, d)}{(1+\lambda)^{jd+1}} + O\left(\frac{2^d d^{2k+1}}{(1+\lambda)^{d(k+1)}}\right).$$

Where the $R_j(\lambda, d)$ are as in Theorem 10.

Proof. By Lemmas 11 and 13, noting that $|N(\Gamma)| \leq d\|\Gamma\|$ for any cluster Γ, we have

$$\mathbb{E}_{\mathcal{O},\lambda}(|\mathbf{I}|) = \frac{\lambda}{1+\lambda}N + \sum_{\substack{\Gamma \in \mathcal{C} \\ \|\Gamma\| \leq k}} w(\Gamma)\left(\|\Gamma\| - \frac{\lambda}{1+\lambda}|N(\Gamma)|\right) + O\left(\frac{2^d d^{2k+1}}{(1+\lambda)^{d(k+1)}}\right).$$

The result follows by recalling the definition of $R_j(\lambda, d)$ at (11) . $\qquad \square$

3. Independent sets of a given size

Theorem 6 will follow from several lemmas. The first says that almost all independent sets of size $m = \lfloor \beta N \rfloor$ are accounted for by exactly one of the two polymer distributions. The second says that if we find λ_β so that the expected size of the independent set drawn from $\mu_{\mathcal{O},\lambda_\beta}$ (as in Definition 8) is close to m then we have an asymptotic formula for the number of independent sets of size m in terms of λ_β and $Z_{\mathcal{O}}(\lambda_\beta)$. The third lemma gives an efficiently computable formula for a suitable such λ_β. We will then prove Theorem 5 by analysing expansions of $\log Z_{\mathcal{O}}(\lambda_\beta)$ and $\log \lambda_\beta$ in powers of $(1 - \beta)^d$.

Let $i_m(\mathcal{O})$ be the number of independent sets I of size m in Q_d that are achievable in the odd polymer model (i.e. $\mu_{\mathcal{O},\lambda}(I) > 0$).

Lemma 15 *For any $\beta > 0$,*

$$i_{\lfloor \beta N \rfloor}(Q_d) = (2 + o(1)) i_{\lfloor \beta N \rfloor}(\mathcal{O})$$

as $d \to \infty$.

Lemma 16 *Fix $\beta > 0$. Suppose $\lambda = \lambda(\beta, d)$ is such that*

$$\left|\mathbb{E}_{\mathcal{O},\lambda}|\mathbf{I}| - \lfloor \beta N \rfloor\right| = o(N^{1/2}). \tag{14}$$

Then

$$i_{\lfloor \beta N \rfloor}(\mathcal{O}) = (1 + o(1))\frac{(1+\lambda)Z_{\mathcal{O}}(\lambda)}{\lambda^{\lfloor \beta N \rfloor}\sqrt{2\pi N \lambda}}.$$

Lemma 17 *There exists a sequence of rational functions $B_j(d, \beta)$, $j \in \mathbb{N}$, such that B_j can be computed in time $e^{O(j \log j)}$ and the following holds. Fix $t \geq 1$ and let $r = \lceil t/2 \rceil - 1$. Suppose that $m = \lfloor \beta N \rfloor$ with $\beta > 1 - 2^{-1/t}$, then if*

$$\lambda_\beta = \frac{\beta}{1-\beta} + \sum_{j=1}^{r} B_j(\beta, d)(1 - \beta)^{jd}, \tag{15}$$

then

$$\left|\mathbb{E}_{\mathcal{O},\lambda_\beta}|\mathbf{I}| - m\right| = o(N^{1/2}).$$

We prove Lemma 16 first, for which we need the following basic binomial local central limit result.

Lemma 18 *Fix $p \in (0, 1)$ and suppose $X \sim Bin(n, p)$. Suppose $n \to \infty$ and $k = o(\sqrt{n})$, then*

$$\mathbb{P}(X = np + k) = \frac{1 + o(1)}{\sqrt{2\pi np(1-p)}}.$$

Proof of Lemma 16. Let $m = \lfloor \beta N \rfloor$. For any $\lambda > 0$, we have

$$i_m(\mathcal{O}) = \frac{Z_{\mathcal{O}}(\lambda)}{\lambda^m} \mathbb{P}_{\mathcal{O}, \lambda}[|\mathbf{I}| = m],$$

where the probability is with respect to the measure $\mu_{\mathcal{O}, \lambda}$. We therefore need to show that if λ satisfies (14), then

$$\mathbb{P}_{\mathcal{O}, \lambda}[|\mathbf{I}| = m] = (1 + o(1)) \frac{1 + \lambda}{\sqrt{2\pi N \lambda}}. \tag{16}$$

By considering first the collection of polymers $\boldsymbol{\Gamma}$ chosen at Step 1 in the definition of $\mu_{\mathcal{O}, \lambda}$ (Definition 8), and then the probability that the correct number of additional vertices are chosen at Step 2, we see that

$$\mathbb{P}_{\mathcal{O}, \lambda}[|\mathbf{I}| = m] = \sum_{\Gamma \in \Omega_{\mathcal{O}}} \mathbb{P}[\boldsymbol{\Gamma} = \Gamma] \cdot \mathbb{P}\left[\text{Bin}\left(N - |N(\Gamma)|, \frac{\lambda}{1 + \lambda}\right) = m - \|\Gamma\| \right], \tag{17}$$

where we recall that $\Omega_{\mathcal{O}}$ denotes the set of all collections of mutually compatible odd polymers. By the large deviation bound [15, Lemma 16] we have

$$\mathbb{P}\left[|N(\boldsymbol{\Gamma})| \geq \frac{2N}{d} \right] = O(\exp(-N/d^4)),$$

and so we can condition on the event $|N(\boldsymbol{\Gamma})| \leq \frac{2N}{d}$ throughout (17) and only change the resulting probability by an additive factor of $O(\exp(-N/d^4)) = o(N^{-1/2})$. Under this conditioning, the binomial probabilities in (17) are uniformly bounded by $O(1/\sqrt{N})$. To establish (16), it therefore suffices to show that with high probability in the choice of $\boldsymbol{\Gamma}$, the binomial probabilities in (17) are in fact equal to $(1 + o(1)) \frac{1 + \lambda}{\sqrt{2\pi N \lambda}}$. By Lemma 18, it suffices to show that with high probability in the choice of $\boldsymbol{\Gamma}$ we have

$$\frac{\lambda}{1 + \lambda}(N - |N(\Gamma)|) = m - \|\Gamma\| + o(N^{1/2}). \tag{18}$$

Now, by our assumption on λ, (13) and Lemma 13,

$$m = \mathbb{E}_{\mathcal{O}, \lambda}|\mathbf{I}| + o(N^{1/2}) = \mathbb{E}\|\boldsymbol{\Gamma}\| + \frac{\lambda}{1 + \lambda}(N - \mathbb{E}|N(\boldsymbol{\Gamma})|) + o(N^{1/2}).$$

It follows that to show (18) holds whp with respect to $\boldsymbol{\Gamma}$, it suffices to show that

$$\mathbb{P}\left[\|\boldsymbol{\Gamma}\| = \mathbb{E}\|\boldsymbol{\Gamma}\| + o(N^{1/2}) \right] = 1 + o(1), \tag{19}$$

and similarly for $|N(\boldsymbol{\Gamma})|$. This is an immediate consequence of Lemma 12 and the fact that

$$\mathbb{E}\|\boldsymbol{\Gamma}\| \leq \mathbb{E}|N(\boldsymbol{\Gamma})| = \sum_{\Gamma \in \mathcal{C}} w(\Gamma)|N(\Gamma)| = o(N), \tag{20}$$

where we have used (13) and Lemma 11. □

Next we prove Lemma 15.

Proof of Lemma 15. Let $\mathcal{I}_m$ denote the set of all independent sets of size m in Q_d. Then by Theorem 9 (and the symmetry between even and odd) we have

$$\left|\mu_\lambda(\mathcal{I}_m) - \hat{\mu}_\lambda(\mathcal{I}_m)\right| = \left|\frac{i_m(Q_d)\lambda^m}{Z_{Q_d}(\lambda)} - \frac{i_m(\mathcal{O})\lambda^m}{Z_\mathcal{O}(\lambda)}\right| = O\left(\exp\left(-N/d^4\right)\right).$$

By Theorem 9 again it follows that

$$\left|i_m(Q_d) - (2 + o(1))i_m(\mathcal{O})\right| = O\left(\exp\left(-N/d^4\right)\right)\frac{Z_\mathcal{O}(\lambda)}{\lambda^m}.$$

It therefore suffices to show that there is a choice of λ such that $\frac{i_m(\mathcal{O})\lambda^m}{Z_\mathcal{O}(\lambda)}$ is much larger than $\exp\left(-N/d^4\right)$. This follows from Lemma 16 by choosing λ satisfying (14). ▢

Next we prove Lemma 17. In the following, if $P(x, y)$ is a polynomial in x, y, we write $\deg_x(P)$ for the degree of P in x.

Proof of Lemma 17. Let $r = \lceil t/2 \rceil - 1$ and set

$$\lambda = \lambda_\beta = \frac{\beta}{1 - \beta} + \sum_{j=1}^r B_j(\beta, d)(1 - \beta)^{jd},$$

where the functions B_j are rational polynomials in β, d of constant degree (independent of d) to be determined later. Let $X := \sum_{j=1}^r B_j(\beta, d)(1 - \beta)^{jd+1}$ and note that $X = o(1)$. It will be useful to note that for $k = O(d)$,

$$(1 + \lambda)^{-k} = \frac{(1 - \beta)^k}{(1 + X)^k} = (1 - \beta)^k \sum_{i=0}^r \binom{-k}{i}X^i + O\left(X^{r+1}\right). \tag{21}$$

In particular, since $\beta > 1 - 2^{-1/t}$,

$$(1 + \lambda)^{-d(r+1)} = (1 + o(1))(1 - \beta)^{d(r+1)} = O\left(e^{-cd} \cdot 2^{-d(r+1)/t}\right) = O\left(e^{-cd}N^{-1/2}\right), \tag{22}$$

for some constant $c > 0$.

By Corollary 14 and Theorem 10,

$$\mathbb{E}_{\mathcal{O},\lambda}(|\mathbf{I}|) = N\frac{\lambda}{1 + \lambda} + \lambda N \sum_{j=1}^r \frac{\frac{\partial}{\partial\lambda}R_j(\lambda, d)}{(1 + \lambda)^{jd}} - \frac{jdR_j(\lambda, d)}{(1 + \lambda)^{jd+1}} + O\left(N\frac{d^{2r+1}}{(1 + \lambda)^{d(r+1)}}\right)$$

$$= \frac{\lambda}{1 + \lambda}N + N\sum_{j=1}^r F_j(\lambda, d)(1 + \lambda)^{-jd-1} + o\left(N^{1/2}\right), \tag{23}$$

where the F_j are polynomials in λ, d with $\deg_d(F_j) \leq 2j + 1$ and $\deg_\lambda(F_j) \leq 3j^2$. Our goal is to show that there exists an appropriate choice of $B_1, \ldots, B_r$ that makes this final expression (23) equal to $m + o\left(N^{1/2}\right)$.

Since $\deg_\lambda(F_j) \leq 3j^2$, and $\lambda = (\beta + X)/(1 - \beta)$ we may write

$$F_j(\lambda, d) = (1 - \beta)^{-c_j}G_j(\beta, d, X)$$

for some non-negative integer $c_j \leq 3j^2$ and G_j a polynomial in β, d, X such that $\deg_d(G_j) \leq 2j + 1$. It follows by (21) that

$$\frac{\lambda}{1+\lambda} + \sum_{j=1}^{r} F_j(\lambda, d)(1+\lambda)^{-jd-1} = (\beta + X) \sum_{i=0}^{r} (-X)^i$$

$$+ \sum_{j=1}^{r} G_j(\beta, d, X)(1-\beta)^{jd+1-c_j} \sum_{i=0}^{r} \binom{-jd-1}{i} X^i + O(d^{3r} X^{r+1}). \tag{24}$$

We now recall that $X = \sum_{j=1}^{r} B_j(1-\beta)^{jd+1}$ and we expand this final expression as a polynomial in $(1-\beta)^d$. This yields

$$\frac{\lambda}{1+\lambda} + \sum_{j=1}^{r} F_j(\lambda, d)(1+\lambda)^{-jd-1} = \beta + \sum_{j=1}^{r} Q_j(\beta, d, B_1, \dots, B_r) \cdot (1-\beta)^{jd} + O(d^{3r} X^{r+1}) \tag{25}$$

where $Q_j = Q_j(\beta, d, B_1, \dots, B_r)$ is a rational function $\beta, d, B_1, \dots, B_r$ with denominator $(1-\beta)^{b_j}$ for some $b_j \leq 3j^2$ and with $\deg_d(Q_j) \leq 2j + 1$. Moreover, by examining the expansion (24), we see that Q_j is linear in B_j where the coefficient of B_j is $(1-\beta)^2$ (in particular the coefficient is non-zero). It follows inductively that there is a choice of $B_1, \dots, B_r$ such that $Q_1 = \dots = Q_r = 0$ where B_j is a rational function of β, d of constant degree (depending on j but not d). With this choice of $B_1, \dots, B_r$ it follows from (23) and (25) that

$$\mathbb{E}_{\mathcal{O}, \lambda}(|\mathbf{I}|) = \beta N + O(Nd^{3r} X^{r+1}) = \beta N + o(N^{1/2}),$$

where for the last bound we used that $d^{3r} X^{r+1} = d^{O_r(1)}(1-\beta)^{d(r+1)} = o(N^{-1/2})$ by (22).

Finally we note that the above argument gives an algorithm for computing the B_j. Since the R_j, and so also the F_j and G_j, can be computed in $e^{O(j \log j)}$ time, we see that the Q_j can be computed in $e^{O(j \log j)}$ time. The B_j can then be computed by solving j successive linear equations. □

To give a concrete example of the algorithm above in action, we pause for a moment to calculate the rational function B_1. Using the definition of R_1 at (11), we see that $R_1 = \lambda$. In the notation of the proof of Theorem 17, it follows that $F_1 = \lambda + (1-d)\lambda^2$. Noting that $\lambda = (\beta + X)/(1-\beta)$ we have

$$F_1 = (1-\beta)^{-2} \left[(1-\beta)(\beta + X) + (1-d)(\beta + X)^2 \right]$$

and so $G_1 = (1-\beta)(\beta + X) + (1-d)(\beta + X)^2$ and $c_1 = 2$. Recalling that $X := \sum_{j=1}^{r} B_j(1-\beta)^{jd+1}$ and examining the coefficient of $(1-\beta)^d$ in (24), we see that

$$Q_1 = B_1(1-\beta)^2 + \frac{\beta(1-\beta) + (1-d)\beta^2}{1-\beta}.$$

Solving $Q_1 = 0$ yields

$$B_1 = \frac{(d\beta - 1)\beta}{(1-\beta)^3}. \tag{26}$$

3.1 Proof of Theorem 6

Combining Lemmas 15, 16 and 17 gives us the proof of Theorem 6.

3.2 Proof of Theorem 5

We now prove Theorem 5. Given the formula (5) from Theorem 6, we need to extract the binomial coefficient $\binom{N}{\lfloor \beta N \rfloor}$ and expand the logarithm of what remains.

Lemma 19 *Fix $\beta \in (0, 1)$. With $\lambda_0 = \frac{\beta}{1-\beta}$,*

$$\binom{N}{\lfloor \beta N \rfloor} = (1 + o(1)) \frac{(1 + \lambda_0)^N}{\lambda_0^{\lfloor \beta N \rfloor} \sqrt{2\pi N \beta (1 - \beta)}}$$

as $N \to \infty$.

The proof follows from Stirling's formula.

Proof of Theorem 5. Given Lemma 19, Theorems 3 and 6, we are left to compute coefficients $P_j = P_j(\beta, d)$, $j \geq 1$, so that for $t \geq 1$ and $\beta > 1 - 2^{-1/t}$ we have

$$\log\left(\frac{1 + \lambda_\beta}{1 + \lambda_0}\right) - \beta \log \frac{\lambda_\beta}{\lambda_0} + \sum_{j=1}^{t-1} R_j(d, \lambda_\beta)(1 + \lambda_\beta)^{-dj} = \sum_{j=1}^{t-1} P_j \cdot (1 - \beta)^{jd} + o(N^{-1}) \qquad (27)$$

where λ_β is given by (4). We proceed by expanding each term on the left-hand side of (27) as a power series in $(1 - \beta)^d$. As in the proof of Lemma 17 we set $X := \sum_{j=1}^r B_j(\beta, d)(1 - \beta)^{jd+1}$ and note that $X^t = o(N^{-1})$ since $\beta > 1 - 2^{-1/t}$. It follows by Taylor expansion that

$$\log\left(\frac{1 + \lambda_\beta}{1 + \lambda_0}\right) - \beta \log \frac{\lambda_\beta}{\lambda_0} = \log(1 + X) - \beta \log(1 + X/\beta)$$

$$= \sum_{i=1}^{t-1} \frac{(-1)^{1+i}}{i}(1 - \beta^{1-i})X^i + o(N^{-1}). \qquad (28)$$

We now turn to the sum on the left-hand side of (27). Since R_j is a polynomial in λ_β, d such that $\deg_{\lambda_\beta}(R_j) \leq 3j^2$ and $\deg_d(R_j) \leq 2j$ and the fact that $\lambda_\beta = (\beta + X)/(1 - \beta)$, we may write

$$R_j(\lambda_\beta, d) = (1 - \beta)^{-c_j} S_j(\beta, d, X)$$

for some non-negative integer $c_j \leq 3j^2$ and S_j a polynomial in β, d, X such that $\deg_d(S_j) \leq 2j$. It follows by (21) that

$$\sum_{j=1}^{t} R_j(\lambda_\beta, d)(1 + \lambda_\beta)^{-dj} = \sum_{j=1}^{t-1} S_j(\beta, d, X)(1 - \beta)^{jd-c_j} \sum_{i=0}^{t-1} \binom{-jd}{i} X^i + O(d^{2t}X^t). \qquad (29)$$

We note that $O(d^{2t}X^t) = O(d^{3t}(1 - \beta)^{td}) = o(N^{-1})$. To compute the P_j we simply sum (28) and (29), expand the powers of X and collect the coefficients of $(1 - \beta)^d, \ldots, (1 - \beta)^{(t-1)d}$. Finally we note that since R_j, S_j and B_j can each be computed in time $e^{O(j \log j)}$, P_j can be computed in time $e^{O(j \log j)}$. □

3.2.1 Computation of P_1, P_2

To illustrate the algorithm for computing the P_j in Theorem 5, we use it to compute P_1 and P_2. First we note that in [15] it was shown that

$$R_1 = \lambda \quad \text{and} \quad R_2 = \frac{(2\lambda^3 + \lambda^4)d(d - 1) - 2\lambda}{4}.$$

It follows that (using the notation of the proof of Theorem 5)

$$S_1 = \beta + X \quad \text{and} \quad S_2 = \frac{1}{2}(\beta - 1)^3(X + \beta) - \frac{1}{4}d(d - 1)(\beta - X - 2)(X + \beta)^3$$

and $c_1 = 1$, $c_2 = 4$. Now, to compute P_1, P_2 we compute the coefficients of $(1 - \beta)^d, (1 - \beta)^{2d}$ in the sum of (28) and (29). The coefficient of $(1 - \beta)^d$ in (28) is 0 and in (29) it is $\beta/(1 - \beta)$ and so

$$P_1 = \frac{\beta}{1 - \beta} \, .$$

The coefficient of $(1 - \beta)^{2d}$ in (28) is

$$\frac{1}{2\beta} B_1^2 (1 - \beta)^3 \, .$$

The coefficient of $(1 - \beta)^{2d}$ in (29) is

$$(1 - d\beta)B_1 + (1 - \beta)^{-4} \left(\frac{1}{2}(\beta - 1)^3\beta - \frac{1}{4}d(d - 1)(\beta - 2)\beta^3 \right) \, .$$

Recalling (26), the formula for B_1, and summing the above two expressions yields

$$P_2 = \frac{2(\beta - 1)^3\beta - d(d - 1)(\beta - 2)\beta^3}{4(1 - \beta)^4} - \frac{\beta(1 - d\beta)^2}{2(1 - \beta)^3} \, .$$

4. Local central limit theorems for polymer models

In the odd polymer model, let T be a defect type and X_T be the random variable counting the number of defects of type T in a sample from $\mu_{\mathcal{O},\lambda}$. Recall that $m_T = \mathbb{E}X_T$ and let $\sigma_T^2 = \mathrm{var}(X_T)$. Moreover, let n_T denote the number of polymers of type T and let w_T denote the weight $w(S)$ (defined at (7)) of a representative polymer S of type T.

Throughout this section we assume that $\lambda \geq C \log d/d^{1/3}$ as in Theorem 9 and probabilities and expectations are with respect to the odd polymer model. The main result of this section is a multivariate local central limit theorem for the number of polymers of different types, extending the multivariate central limit theorem of [15, Theorem 6].

Theorem 20 *Let $\mathcal{T}_1$ and $\mathcal{T}_2$ be two fixed sets of defect types so that for each $T \in \mathcal{T}_1$, $m_T \to \rho_T$ for some constant $\rho_T > 0$, and for each $T \in \mathcal{T}_2$, $m_T \to \infty$ as $d \to \infty$. Let $\{k_T\}_{T \in \mathcal{T}_1}$ be a collection of non-negative integers and let $\{k_T\}_{T \in \mathcal{T}_2}$ be such that $k_T = \lfloor m_T + s_T \rfloor$ where $|s_T| = O(\sqrt{m_T})$ for all $T \in \mathcal{T}_2$. Then*

$$\mathbb{P}\left(\bigcap_{T \in \mathcal{T}_1 \cup \mathcal{T}_2} X_T = k_T \right) = (1 + o(1)) \prod_{T \in \mathcal{T}_1} \frac{\rho_T^{k_T} e^{-\rho_T}}{(k_T)!} \prod_{T \in \mathcal{T}_2} \frac{e^{-\frac{s_T^2}{2m_T}}}{\sqrt{2\pi m_T}} \, .$$

The probability in the theorem statement is with respect to the odd polymer model, but the statement also holds for defects of an independent set drawn from the hard-core model on Q_d via Theorem 9.

Before we proceed it will be useful to recall a result from [15] on the cumulants of the random variables X_T. Recall that for a random variable X we use $\kappa_k(X)$ to denote the kth cumulant of X. The following result appears as Lemma 20 in [15].

Lemma 21 *For a defect type T, let $Y_T(\Gamma)$ denote the number of polymers of type T in the cluster Γ. Then for any fixed $k \geq 1$,*

$$\kappa_k(X_T) = \sum_{\Gamma \in \mathcal{C}} w(\Gamma) Y_T(\Gamma)^k = (1 + o(1))n_T w_T \, , \tag{30}$$

and

$$\sum_{\substack{\Gamma \in \mathcal{C} \\ \|\Gamma\| > |T|}} \left| w(\Gamma) Y_T(\Gamma)^k \right| = o(n_T w_T) \, . \tag{31}$$

Given a random vector $X = (X_1, \ldots, X_q) \in \mathbb{R}^q$ its characteristic function is

$$\varphi_X(t) = \mathbb{E}e^{i\langle X, t\rangle}$$

for $t \in \mathbb{R}^q$.

Lemma 22 *Fix $q \in \mathbb{N}$ and a list $T_1, \ldots, T_q$ of defect types. Let $X = (X_{T_1}, \ldots, X_{T_q})$. There exists $c > 0$ such that*

$$|\varphi_X(t)| \leq \exp\left\{-c\sum_{i=1}^q t_i^2 n_{T_i} w_{T_i}\right\},$$

for all $t \in [-\pi, \pi]^q$.

Proof. Given a cluster $\Gamma \in \mathcal{C}$, let $Y(\Gamma) = (Y_{T_1}(\Gamma), \ldots, Y_{T_q}(\Gamma))$. Using the cluster expansion we write

$$\log \mathbb{E}e^{i\langle t, X\rangle} = \sum_{\Gamma \in \mathcal{C}} w(\Gamma)e^{i\langle t, Y(\Gamma)\rangle} - \sum_{\Gamma \in \mathcal{C}} w(\Gamma)$$

$$= \sum_{\Gamma \in \mathcal{C}} w(\Gamma)(e^{i\langle t, Y(\Gamma)\rangle} - 1).$$

Let

$$\mathcal{M}_j = \left\{\Gamma \in \mathcal{C} : j = \max\{i : Y_{T_i}(\Gamma) > 0\}, \sum_{i=1}^q Y_{T_i}(\Gamma) > 1\right\}.$$

Then,

$$\mathrm{Re}\log \mathbb{E}e^{i\langle t, X\rangle} = \sum_{\Gamma \in \mathcal{C}} w(\Gamma)(\cos(\langle t, Y(\Gamma)\rangle) - 1)$$

$$= \sum_{i=1}^q n_{T_i} w_{T_i}(\cos(t_i) - 1) + \sum_{j=1}^q \sum_{\Gamma \in \mathcal{M}_j} w(\Gamma)(\cos(\langle t, Y(\Gamma)\rangle) - 1)$$

$$\leq -\frac{1}{5}\sum_{i=1}^q t_i^2 n_{T_i} w_{T_i} + \sum_{j=1}^q \sum_{\Gamma \in \mathcal{M}_j} |w(\Gamma)|\langle t, Y(\Gamma)\rangle^2$$

$$\leq -\frac{1}{5}\sum_{i=1}^q t_i^2 n_{T_i} w_{T_i} + \sum_{j=1}^q \sum_{\Gamma \in \mathcal{M}_j} |w(\Gamma)|j\sum_{i=1}^j t_i^2 Y_{T_i}(\Gamma)^2$$

$$\leq -\frac{1}{5}\sum_{i=1}^q t_i^2 n_{T_i} w_{T_i} + \sum_{i=1}^q t_i^2 o(n_{T_i} w_{T_i})$$

where for the first inequality we used that $-t^2 \leq \cos(t) - 1 \leq -t^2/5$ for $t \in [-\pi, \pi]$. For the next inequality we used Cauchy–Schwarz, and for the final inequality we used Lemma 21. □

Proof of Theorem 20. Let $\mathcal{T}_1 = \{T_1, \ldots, T_p\}$, $\mathcal{T}_2 = \{T_{p+1}, \ldots, T_q\}$ and $\mathcal{T} = \mathcal{T}_1 \cup \mathcal{T}_2$. Let $X_i = X_{T_i}$, $m_i = m_{T_i}$, $\sigma_i = \sigma_{T_i}$, $k_i = k_{T_i}$ for $i \in [q]$. Let $X = (X_1, \ldots, X_q)$ and let

$$\tilde{X} = \left(X_1, \ldots, X_p, \frac{X_{p+1} - m_{p+1}}{\sigma_{p+1}}, \ldots, \frac{X_q - m_q}{\sigma_q}\right).$$

By Fourier inversion,

$$\mathbb{P}\left(\bigcap_{T\in\mathcal{T}} X_T = k_T\right) = \frac{1}{(2\pi)^q} \int_{[-\pi,\pi]^q} \varphi_X(t) \cdot e^{-i\langle t,k\rangle} \, dt \, .$$

Making the substitution $t_i = x_i$ for $i \in [p]$ and $t_i = x_i/\sigma_i$ for $i > p$, we have

$$\mathbb{P}\left(\bigcap_{T\in\mathcal{T}} X_T = k_T\right) = \frac{1}{(2\pi)^q} \prod_{i>p} \sigma_i^{-1} \int_{B_1} \int_{B_2} \varphi_{\tilde{X}}(x) g(x) dx_q \dots dx_1 \tag{32}$$

where $B_1 = [-\pi,\pi]^p$, $B_2 = [-\pi\sigma_{p+1}, \pi\sigma_{p+1}] \times \dots \times [-\pi\sigma_q, \pi\sigma_q]$ and

$$g(x) = \exp\left\{ i \sum_{j>p} x_j \frac{m_j - k_j}{\sigma_j} - i \sum_{j\leq p} x_j k_j \right\} \, .$$

Let $Y = (Y_1, \dots, Y_q)$ where $Y_i \sim \text{Po}(\rho_i)$ for $i \in [p]$, $Y_i \sim N(0,1)$ for $i > p$, and $Y_1, \dots, Y_q$ are jointly independent. Then by Fourier inversion, we have the identity:

$$\frac{1}{(2\pi)^q} \int_{B_1} \int_{\mathbb{R}^{q-p}} \varphi_Y(x) g(x) dx_q \dots dx_1 = \prod_{T\in\mathcal{T}_1} \frac{\rho_T^{k_T} e^{-\rho_T}}{(k_T)!} \prod_{T\in\mathcal{T}_2} \frac{e^{-\frac{(k_T - m_T)^2}{2\sigma_T^2}}}{\sqrt{2\pi}} \, .$$

By (32) (noting that $m_i = (1 + o(1))\sigma_i^2$ by Lemma 21), it therefore suffices to show that

$$\int_{B_1} \int_{B_2} \varphi_{\tilde{X}}(x) g(x) dx_q \dots dx_1 = \int_{B_1} \int_{\mathbb{R}^{q-p}} \varphi_Y(x) g(x) dx_q \dots dx_1 + o(1) \, .$$

Since, $m_i \to \infty$ for $i > p$, Lemma 21 implies that $\sigma_i \to \infty$ for $i > p$ also. It follows that

$$\int_{B_1} \int_{\mathbb{R}^{q-p} \setminus B_2} \varphi_Y(x) g(x) dx_q \dots dx_1 = o(1)$$

and so it suffices to show that

$$\int_{B_1} \int_{B_2} |\varphi_{\tilde{X}}(x) - \varphi_Y(x)| dx_q \dots dx_1 = o(1) \, .$$

In [15, Theorem 6] it was shown that $\tilde{X}$ converges to Y in distribution and so $\varphi_{\tilde{X}} \to \varphi_Y$ pointwise. It therefore suffices, by dominated convergence, to show that $|\varphi_{\tilde{X}}(x) - \varphi_Y(x)|$ is bounded by an integrable function. By Lemmas 21 and 22,

$$|\varphi_{\tilde{X}}(x)| = |\varphi_X(x_1, \dots, x_p, x_{p+1}/\sigma_{p+1}, \dots, x_q/\sigma_q)| \leq e^{-\Theta(\sum_{i=1}^q x_i^2)} \, .$$

We have a similar bound for $|\varphi_Y(x)|$ and so we are done. $\qquad\square$

5. Independent sets of a given size and structure

Here we prove Theorem 7. Recall that for a set of defect types $\mathcal{T}$ and a vector of integers $\mathbf{x} = (x_T)_{T\in\mathcal{T}}$, we let $i_{m,\mathbf{x}}(Q_d)$ denote the number of independent sets in Q_d of size m with exactly x_T defects of type T for each $T \in \mathcal{T}$.

We use the following identity, an easy extension of (3). For any $\lambda > 0$,

$$i_{m,\mathbf{x}}(Q_d) = \frac{Z(\lambda)}{\lambda^m} \mathbb{P}_\lambda \left[|\mathbf{I}| = m, (X_T)_{T\in\mathcal{T}} = \mathbf{x} \right] \, , \tag{33}$$

where X_T be the random variable counting the number of defects of type T in a sample from μ_λ. Theorem 7 follows immediately from (33), Theorem 6 and the following lemma.

Lemma 23 *Fix $\lambda > 0$ and let $m = m(d)$ be such that that $|\mathbb{E}_\lambda|\mathbf{I}| - m| = o(N^{1/2})$. Let $\mathcal{T}_1, \mathcal{T}_2$ be the sets of defect types such that $m_T \to \rho_T$ for some fixed $\rho_T > 0$ as $d \to \infty$ for all $T \in \mathcal{T}_1$ and $m_T \to \infty$ for all $T \in \mathcal{T}_2$. Let $(k_T)_{T \in \mathcal{T}_1}$ be a vector of fixed non-negative integers and let $(k_T)_{T \in \mathcal{T}_2}$ be such that $k_T = \lfloor m_T + s_T \rfloor$ where $|s_T| = O(\sqrt{m_T})$ for all $T \in \mathcal{T}_2$. Let $\mathbf{x} = (k_T)_{T \in \mathcal{T}_1 \cup \mathcal{T}_2}$. Then*

$$\mathbb{P}_\lambda\left[|\mathbf{I}| = m, (X_T)_{T \in \mathcal{T}_1 \cup \mathcal{T}_2} = \mathbf{x}\right] = (1 + o(1))\mathbb{P}_\lambda[|\mathbf{I}| = m] \prod_{T \in \mathcal{T}_1} \frac{\rho_T^{k_T} e^{-\rho_T}}{(k_T)!} \prod_{T \in \mathcal{T}_2} \frac{e^{-\frac{s_T^2}{2m_T}}}{\sqrt{2\pi m_T}}.$$

Proof. Using Theorems 9 and 20, it is enough to show that

$$\mathbb{P}_{\mathcal{O},\lambda}\left[|\mathbf{I}| = m \middle| (X_T)_{T \in \mathcal{T}_1 \cup \mathcal{T}_2} = \mathbf{x}\right] = (1 + o(1))\mathbb{P}_{\mathcal{O},\lambda}\left[|\mathbf{I}| = m\right],$$

where the above probabilities are with respect to $\mu_{\mathcal{O},\lambda}$ and X_T is now the random variable counting the number of defects of type T in a sample from $\mu_{\mathcal{O},\lambda}$. Let $\boldsymbol{\Gamma}$ be the random collection of compatible polymers chosen at Step 1 in the definition of $\mu_{\mathcal{O},\lambda}$ (Definition 8) and let Γ be a fixed collection of compatible polymers such that

$$\|\Gamma\| = \mathbf{E}[\|\boldsymbol{\Gamma}\|] + o(N^{1/2}) \tag{34}$$

and

$$|N(\Gamma)| = \mathbf{E}[|N(\boldsymbol{\Gamma})|] + o(N^{1/2}). \tag{35}$$

Then the proof of Lemma 16 gives us that

$$\mathbb{P}_{\mathcal{O},\lambda}\left[|\mathbf{I}| = m \middle| \boldsymbol{\Gamma} = \Gamma\right] = (1 + o(1))\mathbb{P}_{\mathcal{O},\lambda}\left[|\mathbf{I}| = m\right]$$

and so in particular, if Γ is also consistent with $\mathbf{x}$ (i.e. Γ has precisely k_T polymers of type T for all $T \in \mathcal{T}_1 \cup \mathcal{T}_2$),

$$\mathbb{P}_{\mathcal{O},\lambda}\left[|\mathbf{I}| = m \middle| (X_T)_{T \in \mathcal{T}_1 \cup \mathcal{T}_2} = \mathbf{x},\ \boldsymbol{\Gamma} = \Gamma\right] = (1 + o(1))\mathbb{P}_{\mathcal{O},\lambda}\left[|\mathbf{I}| = m\right].$$

Finally, Lemma 12 gives us that (34) and (35) both hold with probability $1 - o(1)$, completing the proof. □

Acknowledgements

References

[1] Arratia, R. and Tavaré, S. (1994) Independent process approximations for random combinatorial structures. *Adv. Math.* **104**(1) 90–154.

[2] Balogh, J., Garcia, R. I. and Li, L. (2021) Independent sets in the middle two layers of Boolean lattice. *J. Comb. Theory Ser. A* **178** 105341.

[3] Barvinok, A. and Hartigan, J. (2010) Maximum entropy gaussian approximations for the number of integer points and volumes of polytopes. *Adv. Appl. Math.* **45**(2) 252–289.

[4] Barvinok, A. and Hartigan, J. (2012) An asymptotic formula for the number of non-negative integer matrices with prescribed row and column sums. *Trans. Am. Math. Soc.* **364**(8) 4323–4368.

[5] Barvinok, A. and Hartigan, J. A. (2013) The number of graphs and a random graph with a given degree sequence. *Random Struct. Algorithms* **42**(3) 301–348.

[6] Campanino, M., Capocaccia, D. and Tirozzi, B. (1979) The local central limit theorem for a Gibbs random field. *Commun. Math. Phys.* **70**(2) 125–132.

[7] DeSalvo, S. and Menz, G. (2016) A robust quantitative local central limit theorem with applications to enumerative combinatorics and random combinatorial structures. arXiv preprint arXiv:1610.07664.

[8] Dobrushin, R. and Tirozzi, B. (1977) The central limit theorem and the problem of equivalence of ensembles. *Commun. Math. Phys.* **54**(2) 173–192.

[9] Fristedt, B. (1993) The structure of random partitions of large integers. *Trans. Am. Math. Soc.* **337**(2) 703–735.

[10] Galvin, D. (2011) A threshold phenomenon for random independent sets in the discrete hypercube. *Comb. Prob. Comput.* **20**(1) 27–51.

[11] Galvin, D. (2012) The independent set sequence of regular bipartite graphs. *Discrete Math.* **312**(19) 2881–2892.

[12] Isaev, M. and McKay, B. D. (2018) Complex martingales and asymptotic enumeration. *Random Struct. Algorithms* **52**(4) 617–661.

[13] Jenssen, M. and Keevash, P. (2020) Homomorphisms from the torus. arXiv preprint arXiv:2009.08315.

[14] Jenssen, M., Keevash, P. and Perkins, W. (2020) Algorithms for #BIS-hard problems on expander graphs. *SIAM J. Comput.* **49**(4) 681–710.

[15] Jenssen, M. and Perkins, W. (2020) Independent sets in the hypercube revisited. *J. London Math. Soc.* **102**(2) 645–669.

[16] Korshunov, A. and Sapozhenko, A. (1983) The number of binary codes with distance 2. *Problemy Kibernet* **40** 111–130.

[17] McKinley, G., Michelen, M. and Perkins, W. (2020) Maximum entropy and integer partitions. arXiv preprint arXiv:2012.14498.

[18] Melczer, S., Panova, G. and Pemantle, R. (2020) Counting partitions inside a rectangle. *SIAM J. Discrete Math.* **34**(4) 2388–2410.

[19] Romik, D. (2005) Partitions of n into $t\sqrt{n}$ parts. *Eur. J. Comb.* **26**(1) 1–17.

[20] Sapozhenko, A. (1987) On the number of connected subsets with given cardinality of the boundary in bipartite graphs. *Metody Diskret Analiz* **45** 42–70.