

# Overlap Analysis of the Shortest Path Problem: Local Search, Landscapes, and Franz–Parisi Potential

**Frederic Koehler**  
University of Chicago

FKOEHLER@UCHICAGO.EDU

**Joonhyung Shin**  
University of Chicago

JOONS@UCHICAGO.EDU

**Editors:** Steve Hanneke and Tor Lattimore

## Abstract

Two directions in algorithms and complexity involve: (1) classifying which optimization problems can be solved in polynomial time, and (2) understanding which computational problems are hard to solve *on average* in addition to the worst case. For many average-case problems, there does not currently exist strong evidence via reductions that they are hard. However, we can still attempt to predict their polynomial time tractability by proving lower bounds against restricted classes of algorithms. Geometric approaches to predicting tractability typically study the *optimization landscape*. For optimization problems with random objectives or constraints, ideas originating in statistical physics suggest we should study the *overlap* between approximately-optimal solutions. Formally, properties of *Gibbs measures* and the *Franz–Parisi potential* imply lower bounds against natural local search algorithms, such as Langevin dynamics. A related theory, the *Overlap Gap Property (OGP)*, proves rigorous lower bounds against classes of algorithms which are stable functions of their input. A remarkable recent work of Li and Schramm [COLT 2025] showed that the shortest path problem in random graphs, which is polynomial-time tractable, admits lower bounds against a class of stable algorithms via the OGP. We further investigate and find that: (1) via the OGP and FPP, we can show stable algorithms and natural MCMC chains fail in the optimization landscape of shortest paths, but (2) stable algorithms and local search succeed in the optimization landscape for shortest path *trees*, which agrees with OGP and FPP predictions.

**Keywords:** Overlap gap property, Franz–Parisi potential, computational-statistical gaps

## 1. Introduction

A fundamental problem in the theory of computation is to understand the complexity and tractability of natural optimization problems. In particular, from the standpoint of complexity theory we would like to understand which problems admit polynomial-time algorithms and which do not. The theory of reduction-based hardness has been very successful at this task for categorizing the complexity of different *worst-case optimization* problems; for example, many problems are known to be equivalent to 3-SAT (3-satisfiability) under polynomial time (Karp) reductions, which is the celebrated theory of NP-completeness. Since Karp reductions allow for arbitrary polynomial time computation, the reductions in this theory are allowed to be quite complex, which has allowed for many deep and nontrivial hardness results to be proven, such as in hardness of approximation and PCP theory.

However, the situation with the complexity of natural random/average-case problems, as arise when we consider problems over random graphs or random data generating processes (as often considered in, e.g., ML and statistics), is quite different. In this setting, reduction-based hardness has seen some successes, but not as comprehensively as in the worst-case setting. A key difficulty is

that reductions which work to reduce the worst-case hardness of Problem A from Problem B rarely preserve/map the natural distribution of random problem instances. See, e.g., [Brennan and Bresler \(2020\)](#); [Bresler and Harbuzova \(2025\)](#) and references within for discussion and some positive results.

For this reason, there is great interest in alternative approaches to study the complexity of random optimization problems. In general, one of the alternatives to studying reduction-based hardness is to *unconditionally rule out natural subclasses* of polynomial time algorithms. This is also a familiar strategy in the worst-case setting, for example in the context of lower bounds against decision trees, concrete families of circuits like  $AC_0$ , families of convex programs, etc. In recent years, there have been remarkable successes in proving unconditional lower bounds for random problems: for example, against the powerful Sum-Of-Squares/Lasserre hierarchy of semidefinite programs, against low-degree polynomials, against natural families of gradient-based optimization methods, against spectral algorithms, against classes of message passing algorithms, etc. (see, e.g., [Barak et al. \(2019\)](#); [Wein \(2022\)](#); [Gamarnik et al. \(2024\)](#) and references within).

A major difference between lower bounds against a concrete class like Sum-of-Squares programs versus classical reduction-based hardness is sensitivity to *polynomial-time equivalent reformulations* of the same problem. For example, although there are reductions between problems like Planted Clique and Sparse PCA (e.g., [Brennan and Bresler \(2020\)](#)), an SoS lower bound against Planted Clique does not logically imply an SoS lower bound against natural SDP formulations of Sparse PCA.

**Landscape analysis.** Another popular type of analysis for the tractability of random optimization problems is via various types of studies of *optimization geometry* and *landscape analysis*.

Landscape analysis, in both worst-case and average-case problems, can potentially be sensitive to small changes in the formulation of the problem. For example: Auer, Herbster, and Warmuth showed that for the task of learning a single sigmoidal neuron the squared loss can have exponentially many bad local minima (see [Auer et al. \(1995\)](#)), even though if we use the logistic loss<sup>1</sup> for the same problem gradient descent will be globally convergent, because it is convex. The sensitivity of landscape analysis is not necessarily a bad thing. For example, for algorithm design it is valuable to know that gradient descent on one loss function will work better than another one, even if the global minimizer of both problems is the same (so finding the global minima of either function is polynomial time equivalent).

In this paper, we compare two combinatorial optimization problems — one is over paths, and one is over trees. We can easily write both of them down as follows.

$$\begin{aligned}
 \text{(P1)} \quad & \min_{P \in \mathcal{P}_{s,t}} |P| & \text{(P2)} \quad & \min_{T \in \mathcal{T}} \sum_{v \in V} d_T(s, v)
 \end{aligned}$$

Here,  $\mathcal{P}_{s,t}$  denotes the set of paths from  $s$  to  $t$ , and  $|P|$  is the length of a path  $P$ . Meanwhile,  $\mathcal{T}$  denotes the set of spanning trees of the graph on vertex set  $V$ , and  $d_T(s, v)$  is the distance from  $s$  to a vertex  $v$  in the tree  $T$ . They are both natural formulations of the shortest path problem, and there are very simple polynomial time reductions between the two problems. While (P1) is the most obvious formulation, (P2) probably appears more in the literature (see, e.g. [Appendix K](#), for

1. With an  $\ell_2$  constraint or regularization to prevent divergence to infinity.

how it appears in [Murota and Shioura \(2014\)](#), Appendix J.2 for standard optimization references and physical justification, and the approximation algorithms literature, e.g., [Khuller et al. \(1995\)](#); [Awerbuch et al. \(1990\)](#); [Bharath-Kumar and Jaffe \(2003\)](#)). As we will see, there is a lot which can be said about the landscape of both problems.

**Surprises in shortest path optimization (P1).** Despite the great success so far of the OGP framework, in a remarkable recent work it was shown that the shortest path problem, which is definitely computationally tractable, satisfies the OGP for typical random graphs as proved in [Li and Schramm \(2025\)](#). To be precise, problem (P1) above is proven to exhibit the ensemble overlap-gap property asymptotically almost surely (a.a.s.), when  $G \sim G(n, c \log(n)/n)$  is a sparse Erdős–Rényi graph. As they showed, this arises from the fact (Theorem 2.2 of [Li and Schramm \(2025\)](#)) that shortest paths in two correlated Erdős–Rényi random graphs are either identical or have little overlap with high probability. By standard arguments in the OGP literature (see Corollary 2.3 of [Li and Schramm \(2025\)](#)), this implies that there is no stable<sup>2</sup> algorithm that computes a shortest path from 1 to 2. This is worrisome because it is the type of argument which is typically used in the OGP literature to give evidence that a problem should be computationally hard, but this problem is not actually hard.

As we discuss in more detail in Appendix L, the core structural result proved in [Li and Schramm \(2025\)](#) can also naturally be interpreted as a *free energy barrier* in a sense which is made precise by the *Franz–Parisi potential* (FPP) introduced in [Franz and Parisi \(1995\)](#). We will explain more about this later, but in essence the FPP is part of the general framework of studying *finite-temperature Gibbs measures* associated to the optimization problem. In the case of (P1), this would come down to studying the geometry of probability measures of the form

$$\nu_{G,\beta}(P) \propto \exp(-\bar{\beta}|P|)$$

where  $P$  ranges over the set of  $s$  to  $t$  paths as before, and  $\bar{\beta} > 0$  is an “inverse temperature” parameter that controls the level of randomness/stochasticity in related local search algorithms. Again, as we formally show in the appendix, we can easily prove the existence of a free energy barrier from the ideas in [Li and Schramm \(2025\)](#) and this implies rigorous lower bounds against natural local MCMC algorithms, which could be interpreted as evidence that the problem is hard.

Notably, free energy barriers detected by the FPP typically imply *exponential time* hitting and mixing time bounds on local MCMC algorithms — this is not possible to obtain just from lower bounds against stable algorithms (the goal of OGP) because long-time MCMC is *not* stable/Lipschitz. (See, e.g., [Ma and Schramm \(2025\)](#) for a related example even on a relatively short time horizon.)

**New insights from studying (P2).** The goal of this paper is to give a complete analysis of the optimization landscape of (P2) from both the perspective of the overlap gap (OGP) theory as well as Gibbs measures & the Franz–Parisi potential (FPP). Informally, our key findings are that:

1. Formulation (P2) *does not* have the ensemble overlap gap property<sup>3</sup>. In fact, we analytically compute an explicit formula for the limiting overlap which smoothly varies between 1 and 0.

---

2. In this particular case, any algorithm whose output changes by at most  $o(n)$  edges when we resample the edges of a graph one-by-one in random order. See also Section 1.1 below.

3. One of several reasons this is interesting is that the model does satisfy various types of *disorder chaos*, which is a shared feature with (P1) that sometimes comes together with the OGP. We discuss this more later.

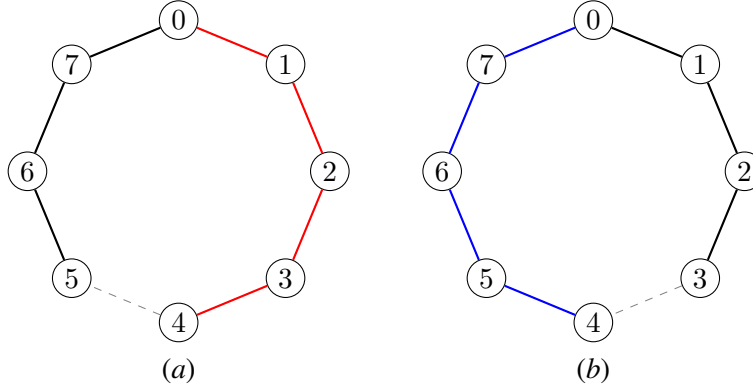


Figure 1: Illustration of Example 1. The two shortest path trees emanating from vertex 0 are very similar, but the shortest paths from  $0 \rightarrow 4$  are disjoint.

2. Formulation (P2) *does not* have free energy barriers in the sense of the Franz–Parisi potential (FPP). In fact, we obtain precise asymptotic descriptions of the finite temperature Gibbs measures and derive explicit *quasiconvex* expressions for the limiting FPP in all regimes.

In this sense, we can think of the richer tree structure having a more “stable” optimization landscape than the path structure. This phenomenon can be intuitively seen in Figure 1.

**Summary: shortest path as a model organism.** The ultimate goal of our work is *not* to prove that the shortest path problem in random graphs is polynomial tractable, which is obvious from our existing knowledge. Instead, our goal is to deepen our understanding of fundamental frameworks such as the OGP and FPP by working out one example, the shortest path problem, in great detail. This ends up to be quite mathematically interesting, but we also mention a few other reasons why we feel this is important:

- In the OGP and FPP literature, a major focus has been on proving evidence for *intractability* for problems which we currently have no algorithms for. Only in relatively few cases (e.g., [Abbe et al. \(2022a\)](#)) has the stability/overlap of an algorithm which *succeeds* been studied from the same perspective. This raises a worry that we could have misinterpreted some OGP lower bounds — we need examples to see in what kinds of special cases a problem which “appears” OGP-hard might actually be easy.
- Despite the clear historical and conceptual connections between the OGP and FPP approaches, there are few examples where both analyses have been worked out in complete detail. One reason is that the analyses are difficult for many problems.
- All commonly studied heuristics have apparent downsides and disadvantages. For example, the low-degree heuristic works reasonably well in the shortest path example from [Li and Schramm \(2025\)](#), but its predictions are “wrong” (and/or, difficult to interpret) in other settings: see [Buhai et al. \(2025\)](#); [Huang and Mossel \(2025, 2024\)](#); [Koehler and Mossel \(2022\)](#). Having several heuristics with well-understood pros and cons can help us make more reliable predictions in the future.

- Although we *do* know that the shortest path problem is tractable, this does not tell us much<sup>4</sup> about overlap concentration, stability of algorithms, the structure of the Gibbs measures, etc. We had to work it out ourselves to see what these heuristics predict about (P2).

To summarize, we consider the shortest path problem and its two landscapes (P1) and (P2) to be “model organisms” for understanding how overlap analysis works. With a better understanding of what happens in this “simple” example, we believe we will be able to better interpret these heuristics in harder problems where the accuracy of our predictions is most crucial and useful.

### 1.1. Background on the OGP and analogy with Gibbs measures

To begin to make precise some of the concepts mentioned in the introduction, We first begin by recalling the definition of the ensemble-OGP, often abbreviated as e-OGP. See the survey [Gamarnik \(2021\)](#) for a more detailed introduction.

**Definition 1** ([Gamarnik \(2021\)](#); [Chen et al. \(2019\)](#)) *Let  $(\Sigma_n, d_n)$  be a metric space and consider a collection of minimization problems, defined over  $\Sigma_n$  and parameterized by elements of a set  $\Xi_n$ , specified by a function  $L : \Sigma_n \times \Xi_n \rightarrow \mathbb{R}$ . We say the ensemble-OGP is satisfied with parameters  $\mu > 0, 0 \leq \nu_1 < \nu_2$  if for any  $\xi, \psi \in \Xi_n$ , and for any*

$$\sigma \in \{\sigma \in \Sigma_n : L(\sigma, \xi) \leq \min_s L(s, \xi) + \mu\}, \quad \tau \in \{\tau \in \Sigma_n : L(\tau, \psi) \leq \min_t L(t, \psi) + \mu\},$$

*it holds that*

$$d_n(\sigma, \tau) \in [0, \nu_1] \cup [\nu_2, \infty).$$

In the special case where  $\Xi_n$  is a singleton set, the e-OGP corresponds to a restriction on the geometric properties of the near-global minima of a single optimization problem: no pair of near-minima exist at “intermediate” distance  $(\nu_1, \nu_2)$ . In general, the e-OGP requires the same property hold for the union of all near-minima among several different optimization problems.

The e-OGP and further generalizations of the OGP have been useful for proving lower bounds against stable algorithms in a wide variety of models. A typical argument is as follows. We form an ensemble by sampling a sequence of problem instances  $\Xi_n = \{\xi_{n,0}, \dots, \xi_{n,T}\}$ , where  $\xi_{n,0}$  and  $\xi_{n,T}$  are independent and  $\xi_{n,t}$  is obtained from  $\xi_{n,t-1}$  by applying a small perturbation. Then an e-OGP argument proves the nonexistence of algorithms  $\mathcal{A}$  that are *stable* in the sense that there is a  $\kappa < \nu_2 - \nu_1$  such that  $d_n(\mathcal{A}(\xi_{n,t-1}), \mathcal{A}(\xi_{n,t})) < \kappa$  for all  $1 \leq t \leq T$ , with high probability as  $n \rightarrow \infty$ .

**Ensemble-OGP for correlated random graphs.** One of the main applications of the e-OGP has been to optimization problems on random Erdős–Rényi graphs. In this setting, e-OGP can be applied by starting with a single random graph and then step-by-step resampling single random vertex pairs in the graph: this generates a sequence of correlated Erdős–Rényi graphs on vertex set  $V$  as elements of  $2^{\binom{V}{2}}$ . This could equivalently be thought of as simulating the trajectory of a random graph under the single-scan Glauber dynamics (see, e.g., [Wilmer et al. \(2009b\)](#)). With these definitions, satisfying the OGP (or not) becomes an event dependent on the particular realization

4. This is true even though we know for independent reasons (see Appendix K) that a natural “local search” procedure *does* converge for (P2). Perhaps unintuitively, even unimodal optimization problems can admit free energy barriers which trap finite temperature MCMC chains — see [Bandeira et al. \(2023\)](#) for a detailed example and explanation.

of the stochastic process — typically the ensemble  $\Xi_n$  (here, the collection of correlated random graphs) is *random* and we are interested in understanding whether the OGP occurs for a typical realization of the ensemble. See Section 1.2.2 for an example specific to shortest path trees.

**Aside: OGP analysis and Gibbs measures from a single perspective.** We will shortly proceed to discussion of our formal results and findings, which will involve an extensive discussion of both the OGP and Gibbs measure approaches. We believe the following simple observation helps clarify the similarities and differences between the two approaches. Starting with Gibbs measures, we know via the “Gumbel trick” (see Appendix J.5) that the analysis of Gibbs measures is equivalent to that of a *randomly perturbed optimization problem*. In particular, for any objective function  $f(x)$ ,  $\beta > 0$ , and finite set  $\mathcal{X}$ ,

$$\log \sum_{x \in \mathcal{X}} e^{-\beta f(x)} = \gamma + \mathbb{E}_G \min_{x \in \mathcal{X}} [\beta f(x) - G(x)]$$

where the expectation is over independent standard Gumbels  $G(x)$  at each site  $x \in \mathcal{X}$ . On the other hand, the ensemble-OGP studies how the optimizer of the objective  $L$  changes under *partial resampling of the underlying randomness* (random graph in our setting). So in fact, these are both natural approaches to studying *the stability of the solution of an optimization problem under random perturbation* — a kind of “sensitivity analysis” for optimization. Perhaps the key *difference* between the two is that the Gibbs measure approach is agnostic to the internal structure of  $f$ .

## 1.2. Main results

### 1.2.1. BASIC MODELS AND NOTATIONS

We denote by  $\mathcal{G}(n, q)$  the Erdős–Rényi model of random graphs, with  $n$  vertices and the edge probability  $q$ . Throughout this paper, as in Li and Schramm (2025), we consider the asymptotic regime where  $q = \Theta(\log(n)/n)$ . This means that the graph is relatively sparse and the giant component of the graph contains at least a  $1 - o(1)$  fraction of the vertices (see, e.g., Alon and Spencer (2016) for background on random graphs). As in the literature on the diameter of random graphs (Chung and Lu, 2001; Riordan and Wormald, 2010, e.g.), the vertices outside of the giant component play no role in the analysis and can be ignored.

**Single graph setup.** We will always consider a fixed sequence of  $\alpha_n \in (0, \infty)$  and define in terms of them the edge probability

$$q_n = \frac{\alpha_n \log n}{n}. \quad (1)$$

To study the asymptotic behavior of sparse random graphs, we consider a sequence  $G_1, G_2, \dots$  of independent Erdős–Rényi graphs sampled from  $G_n \sim \mathcal{G}(n, q_n)$ , where the vertices are set to be integers

$$V_n := V(G_n) = \{1, \dots, n\}.$$

Most of our single graph asymptotics results concern certain functions of a graph, which is naturally a sequence of random variables whose law is dictated by our fixed sequence  $\{\alpha_n\}$ . To guarantee  $q_n = \Theta(\log(n)/n)$ , we assume throughout the paper that

$$0 < \liminf_{n \rightarrow \infty} \alpha_n \leq \limsup_{n \rightarrow \infty} \alpha_n < \infty. \quad (2)$$

We do not require that  $\{\alpha_n\}$  converges or is above the connectivity threshold  $\liminf \alpha_n > 1$ ; our results hold without those assumptions.

**Graphs with fixed source.** We are largely concerned with single source paths, i.e., paths starting from a fixed vertex; w.l.o.g., 1 will always be the source vertex. Thus, the connected component of  $G_n$  containing 1 will be our primary focus, which we denote by  $\overline{G}_n$ . We only consider Erdős–Rényi graphs with  $\liminf \alpha_n > 0$ , so there will be one giant component of size  $n - o(n)$  with high probability. Hence, 1 will asymptotically almost surely be contained in this giant component, so  $\overline{G}_n$  can be thought of as the giant component of  $G_n$ . Whenever we talk about shortest path trees or spanning trees, we are actually referring to those of  $\overline{G}_n$ . Moreover, these spanning trees are naturally rooted at the source vertex 1 so every pair of neighboring vertices in the trees have an induced parent-child relationship; we always take this into account when comparing two trees.

**Correlated graphs setup.** In order to study the overlap of two correlated instances, we need an additional sequence of  $\rho_n \in (0, 1)$  specifying the correlation strength. We define two graphs with correlation  $\rho_n$  in the usual way. We begin with a sample  $G_n^{(1)} \sim \mathcal{G}(n, q_n)$  for each  $n$ , same with the single graph case. Then for each pair of distinct vertices of  $G_n^{(1)}$ , we resample it with probability  $1 - \rho_n$ . This gives a graph  $G_n^{(2)}$  which is also marginally  $\mathcal{G}(n, q_n)$ . Our results concern the (independent) sequence of correlated pairs  $\{(G_n^{(1)}, G_n^{(2)})\}$  obtained this way.

### 1.2.2. OVERLAP CONCENTRATION AND ENSEMBLE-OGP

When we have two graphs  $G$  and  $G'$  on the common vertex set  $V$  and their respective shortest path trees  $T$  and  $T'$  rooted at a fixed source vertex in  $V$ , we define their *overlap* by

$$R(T, T') := \frac{|T \cap T'|}{\sqrt{|T||T'|}} \quad (3)$$

where  $|T|$  denotes the number of the edges of  $T$  and  $|T \cap T'|$  denotes the number of common edges with the same parent-child relationships. Since shortest path trees are a spanning tree of the giant component which has size  $n - o_p(n)$ , the denominator of (3) can be replaced with  $n$  without affecting our asymptotic results.

Recall that in our correlated graphs setup, we have two  $\mathcal{G}(n, q_n)$  graphs  $G_n^{(1)}$  and  $G_n^{(2)}$  with correlation  $\rho_n$ . Let  $\mu_n^{(1)}$  and  $\mu_n^{(2)}$  be the uniform measures over the shortest path trees for the two graphs, respectively. We consider the optimal coupling  $\Pi_n$  of  $\mu_n^{(1)}$  and  $\mu_n^{(2)}$  defined by

$$\Pi_n := \arg \max_{\Pi} \mathbb{E}_{(T, T') \sim \Pi} [|T \cap T'|]. \quad (4)$$

Note that  $\Pi_n$  is precisely the coupling that achieves the 1-Wasserstein distance between  $\mu_n^{(1)}$  and  $\mu_n^{(2)}$  under the Hamming metric. Our main result is that the shortest path tree formulation enjoys *overlap concentration* which we formally state as follows. See Appendix D.3.1 for the proof.

**Theorem 2 (Overlap concentration)** *For a given sequence  $\{\alpha_n\}$ , there is a sequence of automorphisms (i.e., strictly increasing continuous bijections)  $f_n : [0, 1] \rightarrow [0, 1]$  with the following property. For any sequence of correlations  $\{\rho_n\}$ , independently sample for each  $n$  a pair of shortest path trees  $(T_n^{(1)}, T_n^{(2)})$  from the optimal coupling  $\Pi_n$ . Then*

$$\left| R(T_n^{(1)}, T_n^{(2)}) - f_n(\rho_n) \right| \xrightarrow{p} 0.$$

**The absence of ensemble-OGP.** Theorem 2 is strong enough to prove the absence of ensemble-OGP for shortest path trees. Suppose we generate a (random) sequence of  $\mathcal{G}(n, q_n)$  graphs by a step-by-step resampling procedure. Namely, we label the pairs of vertices from 1 to  $T = \binom{n}{2}$ , sample  $G_{n,0} \sim \mathcal{G}(n, q_n)$ , and obtain  $G_{n,t}$  by resampling the  $t$ th pair of vertices in  $G_{n,t-1}$  for each  $1 \leq t \leq T$ , producing a sequence  $G_{n,0}, \dots, G_{n,T}$ . If the sequence satisfied the ensemble-OGP with parameters  $\mu > 0$  and  $0 \leq \nu_1 < \nu_2$ , it would have implied that for any pair of shortest path trees  $T_{n,i}$  and  $T_{n,j}$  of  $G_{n,i}$  and  $G_{n,j}$ , either  $R(T_{n,i}, T_{n,j}) \leq \nu_1$  or  $R(T_{n,i}, T_{n,j}) \geq \nu_2$ . Due to the overlap concentration, this does not happen: for any desired overlap  $r$ , we can choose  $\rho_n = f_n^{-1}(r)$  so that by setting  $t = \lfloor (1 - \rho_n)T \rfloor$  we can make  $R(T_{n,0}, T_{n,t})$  arbitrarily close to  $r$ .

**Stability of shortest path algorithms.** A more powerful consequence of Theorem 2 is that, in the context of the shortest path tree formulation, any algorithm that outputs a uniformly random shortest path tree is  $\kappa$ -stable for any constant  $\kappa > 0$  in the sense described in Section 1.1. Most of the well-known efficient shortest path algorithms can be easily adapted to output a uniformly random shortest path tree, so they are stable in this sense. In contrast, the OGP from Li and Schramm (2025) implies that these algorithms are *not* stable for some  $\kappa > 0$  in the shortest path formulation. More generally, the goal of most (if not all) generalizations of the OGP for random optimization problems (e.g., branching OGP, Huang and Sellke (2023)) is to *rule out the success of stable algorithms*. Since we have proved that uniformly random shortest path tree is stable, the shortest path tree problem will therefore not exhibit branching OGP, etc.

**Numerical simulations match our predictions.** To visualize the distinct differences in the behavior of the overlap, we conducted a numerical simulation (Figure 5) whose details are given in Section J.8. Figure 5(a) shows the simulation result for shortest path trees. The simulated curve is very close to the theoretical curve and the overlap smoothly decreases as the correlation drops. On the other hand, Figure 5(b) shows that the overlap of the shortest paths (contained in the corresponding trees) suddenly drops to zero, which agrees with the result of Li and Schramm (2025).

### 1.2.3. GIBBS MEASURES AND ABSENCE OF FREE ENERGY BARRIER

We extend the analysis to study the natural finite temperature analogue of shortest path trees. This is a natural way from the statistical physics perspective to analyze the structure of the set of approximate optima (in this case, approximate shortest path trees). To state the results, we first must define the natural finite temperature model at inverse temperature  $\beta \geq 0$  which is

$$\mu_{G_n, \beta}(T) = \frac{1}{Z_{G_n, \beta}} \exp \left( -\beta \log \log(n) \sum_{v \in \bar{V}_n} d_T(1, v) \right) \quad (5)$$

where  $\bar{V}_n = V(\bar{G}_n)$  is the set of vertices reachable from 1. The rationale behind this definition can be found in Appendix B.

**Remark 3** In the case  $\beta = 0$ , the measure  $\mu_{G_n, 0}$  is the uniform measure over the set of the spanning trees of  $\bar{G}_n$ . In the other extreme  $\beta = \infty$  (zero temperature limit),  $\mu_{G_n, \infty}$  is the uniform measure over the set of the shortest path trees of  $\bar{G}_n$ .

We denote the ground state energy (minimum energy) of this system by

$$E_{G_n} := \sum_{v \in \bar{V}_n} d_G(1, v)$$

which is simply the sum of distances from 1 to other vertices in the same component  $\bar{V}_n$ . We will compute several thermodynamic quantities such as free energy density and the Franz–Parisi potential relative to this ground state.

**The Franz–Parisi potential.** Suppose that the random spanning trees  $T_n$  are (independently) drawn from the Gibbs measures  $\mu_{G_n, \beta}$  conditioned on  $\{G_n\}$ . Recall the definition (3) of the overlap function for trees denoted by  $R$ . We define the *relative Franz–Parisi potential* by

$$\mathcal{F}_\beta^{\text{FP}}(r) := \lim_{\epsilon \rightarrow 0^+} \text{plim}_{n \rightarrow \infty} \frac{1}{n} \left( -\frac{1}{\beta \log \log n} \log Z_{n,r,\epsilon} - E_{G_n} \right)$$

where

$$Z_{n,r,\epsilon} := \sum_{T: R(T, T_n) \in [r-\epsilon, r+\epsilon]} \exp \left( -\beta \log \log(n) \sum_{v \in \bar{V}_n} d_T(1, v) \right)$$

is the partition function for trees overlapping with  $T_n$  by approximately  $r$ . We precisely computed the Franz–Parisi potential for all regimes where it exists; see Corollary 86. Informally speaking, the Franz–Parisi potential exists as long as the inverse temperature  $\beta > 0$  is not at the *critical temperature*, which we further explain in Section 2.3.2.

**Theorem 4 (Informal implication of Corollary 86)** *The Franz–Parisi potential  $\mathcal{F}_\beta^{\text{FP}}$ , whenever it exists, is a quasiconvex function of the overlap  $r$ . In particular, there is no free energy barrier.*

See Table 1 for a more detailed description and Figure 4 for a rough landscape of  $\mathcal{F}_\beta^{\text{FP}}$ . More precise statements computing  $\mathcal{F}_\beta^{\text{FP}}$  and justifying the landscape can be found in Appendix F.3.

## 2. Technical ingredients: the shortest path structure of sparse random graphs

The main results discussed in the preceding section are derived from core structural properties of the shortest paths in sparse Erdős–Rényi graphs. We present some of the most important ingredients which might also be of independent mathematical interest. The proofs are rather technical and are spread across several accompanying appendices — see Appendix A for an overview.

**Proxies for describing asymptotic behaviors.** The sequences  $\{\alpha_n\}$  (which determines the edge probabilities  $\{q_n\}$ ) and  $\{\rho_n\}$  (in the correlated setup) completely dictate the law of the random graphs. However, many asymptotic behaviors are often more conveniently and concisely described by certain quantities that depend on  $\{\alpha_n\}$ . Among those, we introduce a few important quantities that play a significant role throughout this paper.

First and foremost, we define

$$\ell_n^* = \frac{\log n}{\log(nq_n)} = \frac{\log n}{\log \alpha_n + \log \log n} \quad (6)$$

or equivalently,  $(nq_n)^{\ell_n^*} = n$ . This can be understood as an approximate distance of a typical pair of vertices in sparse Erdős–Rényi graphs and is the same quantity used in Li and Schramm (2025). Also, we define

$$\begin{aligned} d_n^* &:= \min \left\{ d \in \mathbb{Z} : (nq_n)^d \geq \frac{n}{(\log \log n)^2} \right\} \\ &= \left\lceil \ell_n^* - \frac{2 \log \log \log n}{\log \alpha_n + \log \log n} \right\rceil \end{aligned} \quad (7)$$

to be the smallest integer satisfying  $(nq_n)^{d_n^*} \geq n/(\log \log n)^2$ . Under the assumption (2),  $d_n^*$  is one of the two closest integers to  $\ell_n^*$  for every large  $n$ . The denominator  $(\log \log n)^2$  is somewhat arbitrary and is mostly relevant to our analysis of Gibbs measures; see Appendix A.3 and the accompanying footnotes for a brief explanation of our choice of denominator.

The gap between  $d_n^*$  and  $\ell_n^*$  is what turns out to be the key to understanding the dynamics of the graphs. Perhaps the simplest form of such a gap is their difference, which we denote by

$$\Delta_n := d_n^* - \ell_n^*. \quad (8)$$

Note that  $\Delta_n$  might be negative, but (2) implies that

$$0 \leq \liminf_{n \rightarrow \infty} \Delta_n \leq \limsup_{n \rightarrow \infty} \Delta_n \leq 1.$$

Another form, which we denote by  $\lambda_n$ , is nontrivial and less intuitive but is equally important. This is defined by the equation

$$\log(1/\lambda_n) = (nq_n)^{\Delta_n} = \frac{(nq_n)^{d_n^*}}{n}. \quad (9)$$

By definition, we have  $0 < \lambda_n < 1$  and thus

$$0 \leq \liminf_{n \rightarrow \infty} \lambda_n \leq \limsup_{n \rightarrow \infty} \lambda_n \leq 1.$$

We remind the reader that the sequences  $\{\Delta_n\}$  and  $\{\lambda_n\}$  are functions of  $n$  and  $\{\alpha_n\}$ . It turns out that the limiting behavior of  $\Delta_n$  and  $\lambda_n$  as  $n \rightarrow \infty$  is what is critically important for the asymptotics, rather than  $\alpha_n$ .

## 2.1. Asymptotics for a single graph

We first establish the distance asymptotics for  $\{G_n\}$ . The proof and related discussions are detailed in Appendix C.

**Theorem 5 (Distance asymptotics)** *Suppose that  $\lambda_n \rightarrow \lambda \in [0, 1]$ . Let  $N_{d_n^*}$  and  $N_{d_n^*+1}$  denote the number of vertices of  $G_n$  with distance  $d_n^*$  and  $d_n^* + 1$  from the source vertex 1, respectively. Then as  $n \rightarrow \infty$ , we have*

$$N_{d_n^*}/n \xrightarrow{p} 1 - \lambda, \quad \text{and} \quad N_{d_n^*+1}/n \xrightarrow{p} \lambda.$$

*So with probability  $1 - o(1)$  all vertices except for at most  $o(n)$  of them have distance either  $d_n^*$  or  $d_n^* + 1$  from vertex 1, and the proportion of vertices with distance  $d_n^*$  converges to  $1 - \lambda$  in probability.*

## 2.2. Asymptotics for correlated graphs

Recall that in the correlated setup, we consider  $\mathcal{G}(n, q_n)$  graphs  $G_n^{(1)}$  and  $G_n^{(2)}$  with correlation  $\rho_n$ . As an extension to Theorem 5, it is natural to study the asymptotics of the number of vertices that is at distance  $d_n^*$  in both  $G_n^{(1)}$  and  $G_n^{(2)}$ . This indeed turns out to be crucial in proving Theorem 2 but is more challenging to prove — see Appendix D for the details.

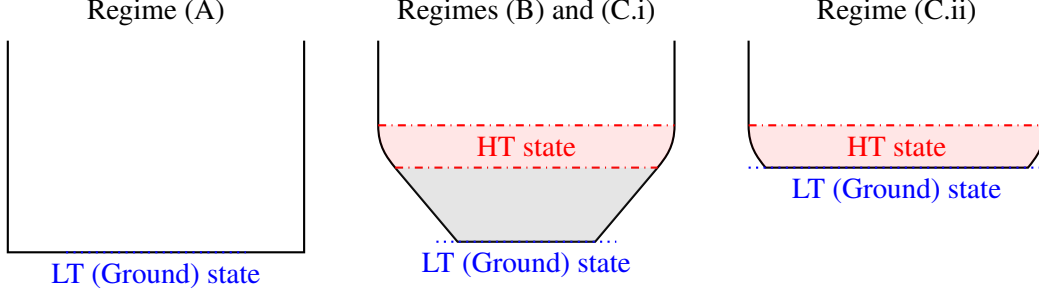


Figure 2: Visualization of the potential well for different regimes. In the middle well, the high temperature state and the low temperature state are separated by a gray region with edges of constant slope  $\beta_c^{-1}$ , resulting in a discontinuous phase transition. For the rightmost one, the two states touch each other, leading to a continuous phase transition.

Before we state the result, we define an additional proxy sequence  $\{\gamma_n\}$  by

$$\gamma_n := \rho_n^{d_n^*} \in (0, 1). \quad (10)$$

This plays an important role when  $\rho_n$  is close to 1; if  $\rho_n$  is bounded away from 1, then  $\gamma_n \rightarrow 0$  as  $n \rightarrow \infty$  since  $d_n^* \rightarrow \infty$ . When  $\rho_n$  is close to 1, we have by Taylor expansion that  $1 - \rho_n \approx -\log(\rho_n) = \log(1/\gamma_n)/d_n^*$ , so the expected number of edges resampled along a path of length  $d_n^*$  is approximately  $\log(1/\gamma_n)$ .

**Theorem 6 (Correlated distance asymptotics)** *Suppose that  $\lambda_n \rightarrow \lambda \in [0, 1]$  and*

$$\rho_n \rightarrow \rho \in [0, 1]. \quad (11)$$

*If  $\rho < 1$ , set  $\gamma = 0$  because  $\gamma_n \rightarrow 0$  necessarily. Otherwise if  $\rho = 1$ , require additionally that*

$$\gamma_n \rightarrow \gamma \in [0, 1]. \quad (12)$$

*Let  $N_{d_n^*}^I$  denote the number of vertices with distance  $d_n^*$  from the source 1 in both  $G_n^{(1)}$  and  $G_n^{(2)}$ . We have that*

$$N_{d_n^*}^I/n \xrightarrow{p} 1 - 2\lambda + \lambda^{2-\gamma},$$

*i.e., the proportion of the vertices that have distance  $d_n^*$  for both  $G_n^{(1)}$  and  $G_n^{(2)}$  converges to  $1 - 2\lambda + \lambda^{2-\gamma}$  in probability. Thus, the proportion of the vertices that have the same distance for  $G_n^{(1)}$  and  $G_n^{(2)}$  converges to  $1 - 2\lambda + 2\lambda^{2-\gamma}$  in probability.*

### 2.3. Gibbs measures, phase transitions, and energy landscape

Recall the definition of the Gibbs measures  $\mu_{G_n, \beta}$  from (5).

**Proxies, regimes, and assumptions.** Throughout this section, we make convergence assumptions on the proxies. First, we assume that

$$\lambda_n \rightarrow \lambda \in [0, 1] \quad \text{and} \quad \Delta_n \rightarrow \Delta \in [0, 1].$$

In the case  $\lambda = 1$  (which implies  $\Delta = 0$ ), we define an additional proxy

$$\kappa_n := (1 - \lambda_n) \log \log n$$

and further assume that  $\kappa_n \rightarrow \kappa \in [0, \infty]$ . In other words, throughout this section, we describe the Gibbs measure in one of the following regimes.

- (A)  $\Delta_n \rightarrow \Delta = 1$ . This implies  $\lambda_n \rightarrow \lambda = 0$  and  $\kappa_n \rightarrow \kappa = \infty$ .
- (B)  $\Delta_n \rightarrow \Delta \in [0, 1)$  and  $\lambda_n \rightarrow \lambda \in [0, 1)$ , which still implies  $\kappa_n \rightarrow \kappa = \infty$ .
- (C)  $\lambda_n \rightarrow \lambda = 1$ . This gives  $\Delta_n \rightarrow \Delta = 0$ , and we further split this into three sub-regimes depending on  $\kappa$ .
  - (C.i)  $\kappa_n \rightarrow \kappa = \infty$ .
  - (C.ii)  $\kappa_n \rightarrow \kappa \in (1, \infty)$ .
  - (C.iii)  $\kappa_n \rightarrow \kappa \in [0, 1]$ .

Moreover, we also assume that the inverse temperature  $\beta > 0$  is a fixed positive constant, i.e., does not change as  $n$  grows.

| Regime |         | $\beta_c$         | $\mathcal{F}_{\lambda, \Delta}(\beta)$<br>(Corollary 79)                                                                                                 | Phase transition<br>(Appendix F.2.2) | $\mathcal{F}_\beta^{\text{FP}}(r)$<br>(Corollary 86)               |
|--------|---------|-------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------|--------------------------------------------------------------------|
| (A)    |         | 0                 | $-\frac{1}{\beta}$                                                                                                                                       | N/A                                  | Increasing                                                         |
| (B)    |         | $1 - \Delta$      | $-\frac{\lambda + \Delta}{\beta} \quad (\beta \geq \beta_c)$<br>$\frac{1 - (\lambda + \Delta)}{1 - \Delta} - \frac{1}{\beta} \quad (\beta \leq \beta_c)$ | discontinuous                        | U-shaped ( $\beta > \beta_c$ )<br>Increasing ( $\beta < \beta_c$ ) |
| (C)    | (C.i)   | 1                 | $-\frac{1}{\beta}$                                                                                                                                       | discontinuous                        | Increasing                                                         |
|        | (C.ii)  | $1 - \kappa^{-1}$ |                                                                                                                                                          | continuous                           |                                                                    |
|        | (C.iii) | $\leq 0$          |                                                                                                                                                          | N/A                                  |                                                                    |

Table 1: Summary of thermodynamic properties of  $\mu_{G, \beta}$ .

### 2.3.1. FREE ENERGY DENSITY

We define the *relative free energy density* by

$$\mathcal{F}_{G_n, \beta} := \frac{1}{n} \left( -\frac{1}{\beta \log \log n} \log Z_{G_n, \beta} - E_{G_n} \right). \quad (13)$$

The following theorem computes the limiting value of  $\mathcal{F}_{G_n, \beta}$ , which turns out to exhibit distinct behaviors for different regimes. See Appendix F.2.1 for the proof.

**Theorem 7 (Limiting free energy density)** *The free energy density converges in probability to*

$$\mathcal{F}_{G_n, \beta} \xrightarrow{P} \mathcal{F}_{\lambda, \Delta}(\beta)$$

where  $\mathcal{F}_{\lambda, \Delta}$  is a continuous function of  $\beta$  as in Table 1.

### 2.3.2. PHASE TRANSITIONS

A particularly interesting regime in Theorem 7 is (B), where the free energy density has a non-differential point at  $\beta = 1 - \Delta$ . At this point, the model exhibits what is called a *first-order* or *discontinuous phase transition* in the inverse temperature  $\beta$  (see, e.g., Chaikin and Lubensky (1995); Friedli and Velenik (2017) for more background on this terminology). Here, we describe the phase transitions in our system from a different perspective; namely, we analyze how the Gibbs measure  $\mu_{G_n, \beta}$  changes with  $\beta$  as an element of the space of probability measures equipped with the 1-Wasserstein metric. This viewpoint reveals phase transition phenomena observable in Regime (C), reflected in the following result.

**Theorem 8 (Informal version of Theorem 73 and Theorem 75)** *Define*

$$\beta_c := 1 - \Delta - \kappa^{-1} \in [-\infty, 1]. \quad (14)$$

*Then  $\mu_{G_n, \beta}$  is close to the uniform measure  $\mu_{G_n, \infty}$  in 1-Wasserstein metric under the Hamming metric if  $\beta > \beta_c$ , and far from  $\mu_{G_n, \infty}$  if  $\beta < \beta_c$ .*

Theorem 8 gives the precise value of the (inverse) *critical temperature*  $\beta_c$  in Table 1, above which (lower temperature) the system behaves like the uniform measure in Wasserstein sense. We see that phase transition in Wasserstein space occurs in the Regimes (B), (C.i), and (C.ii). It turns out that, the phase transition in (C.ii) is a *continuous phase transition*, as opposed to (B) and (C.i) which exhibit a *discontinuous phase transition*. This is formalized in Appendix F.2. See Figure 3 for the phase diagrams and Figure 2 for an overall picture of the potential well, which summarize our results. Figure 2 is in a similar spirit to Figure 1 of Zdeborová and Krzakala (2010) used to illustrate “state following”, which as they explain is closely related to the Franz–Parisi potential.

### Acknowledgments

We thank David Gamarnik, Shuangping Li, Marc Mézard, Elchanan Mossel, and the reviewers for valuable discussions and suggestions.

### References

- Emmanuel Abbe, Shuangping Li, and Allan Sly. Binary perceptron: efficient algorithms can find solutions in a rare well-connected cluster. In *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing*, pages 860–873, 2022a.
- Emmanuel Abbe, Shuangping Li, and Allan Sly. Proof of the contiguity conjecture and lognormal limit for the symmetric perceptron. In *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 327–338. IEEE, 2022b.
- Dimitris Achlioptas and Amin Coja-Oghlan. Algorithmic barriers from phase transitions. In *2008 49th Annual IEEE Symposium on Foundations of Computer Science*, pages 793–802. IEEE, 2008.
- Dimitris Achlioptas and Federico Ricci-Tersenghi. On the solution-space geometry of random constraint satisfaction problems. In *Proceedings of the thirty-eighth annual ACM symposium on Theory of computing*, pages 130–139, 2006.

- Ravindra K Ahuja, Thomas L Magnanti, James B Orlin, et al. *Network flows: theory, algorithms, and applications*, volume 1. Prentice hall Englewood Cliffs, NJ, 1993.
- Ahmed El Alaoui. Near-optimal shattering in the ising pure p-spin and rarity of solutions returned by stable algorithms. *arXiv preprint arXiv:2412.03511*, 2024.
- Paul T Allen, Adam Layne, and Katharine Tsukahara. The dirichlet problem for curve shortening flow. *arXiv preprint arXiv:1208.3510*, 2012.
- Noga Alon and Joel H Spencer. *The probabilistic method*. John Wiley & Sons, 2016.
- Nima Anari, Kuikui Liu, Shayan Oveis Gharan, Cynthia Vinzant, and Thuy-Duong Vuong. Log-concave polynomials iv: approximate exchange, tight mixing times, and near-optimal sampling of forests. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 408–420, 2021.
- Gerard Ben Arous, Reza Gheissari, and Aukosh Jagannath. Algorithmic thresholds for tensor pca. *The Annals of Probability*, 48(4):2052–2087, 2020.
- G rard Ben Arous, Alexander S Wein, and Ilias Zadik. Free energy wells and overlap gap property in sparse pca. *Communications on Pure and Applied Mathematics*, 76(10):2410–2473, 2023.
- Peter Auer, Mark Herbster, and Manfred KK Warmuth. Exponentially many local minima for single neurons. *Advances in neural information processing systems*, 8, 1995.
- Fanny Augeri. A transportation approach to the mean-field approximation. *Probability Theory and Related Fields*, 180(1):1–32, 2021.
- Tim Austin. The structure of low-complexity Gibbs measures on product spaces. *The Annals of Probability*, 47(6):4002 – 4023, 2019. doi: 10.1214/19-AOP1352. URL <https://doi.org/10.1214/19-AOP1352>.
- Baruch Awerbuch, Alan Baratz, and David Peleg. Cost-sensitive analysis of communication protocols. In *Proceedings of the ninth annual ACM symposium on Principles of distributed computing*, pages 177–187, 1990.
- Carlo Baldassi, Riccardo Della Vecchia, Carlo Lucibello, and Riccardo Zecchina. Clustering of solutions in the symmetric binary perceptron. *Journal of Statistical Mechanics: Theory and Experiment*, 2020(7):073303, 2020.
- Afonso S Bandeira, Ahmed El Alaoui, Samuel Hopkins, Tselil Schramm, Alexander S Wein, and Ilias Zadik. The franz-parisi criterion and computational trade-offs in high dimensional statistics. *Advances in Neural Information Processing Systems*, 35:33831–33844, 2022.
- Afonso S Bandeira, Antoine Maillard, Richard Nickl, and Sven Wang. On free energy barriers in gaussian priors and failure of cold start mcmc for high-dimensional unimodal distributions. *Philosophical Transactions of the Royal Society A*, 381(2247):20220150, 2023.
- Boaz Barak, Samuel Hopkins, Jonathan Kelner, Pravesh K Kothari, Ankur Moitra, and Aaron Potechin. A nearly tight sum-of-squares lower bound for the planted clique problem. *SIAM Journal on Computing*, 48(2):687–735, 2019.

- A D Barbour and Aihua Xia. On stein’s factors for poisson approximation in wasserstein distance. *Bernoulli*, 12(6):943–954, 2006. ISSN 13507265. URL <http://www.jstor.org/stable/25464848>.
- A D Barbour, Lars Holst, and Svante Janson. *Poisson Approximation*. Oxford University Press, 02 1992. ISBN 9780198522355. doi: 10.1093/oso/9780198522355.001.0001. URL <https://doi.org/10.1093/oso/9780198522355.001.0001>.
- Anirban Basak and Sumit Mukherjee. Universality of the mean-field for the potts model. *Probability Theory and Related Fields*, 168:557–600, 2017.
- Reinhard Bauer and Dorothea Wagner. Batch dynamic single-source shortest-path algorithms: An experimental study. In *International Symposium on Experimental Algorithms*, pages 51–62. Springer, 2009.
- Mokhtar S Bazaraa, John J Jarvis, and Hanif D Sherali. *Linear programming and network flows*. John Wiley & Sons, 2011.
- Kadaba Bharath-Kumar and J Jaffe. Routing to multiple destinations in computer networks. *IEEE Transactions on communications*, 31(3):343–351, 2003.
- Giulio Biroli, Rémi Monasson, and Martin Weigt. A variational description of the ground state structure in random satisfiability problems. *The European Physical Journal B-Condensed Matter and Complex Systems*, 14(3):551–568, 2000.
- Giulio Biroli, Chiara Cammarota, and Federico Ricci-Tersenghi. How to iron out rough landscapes and get optimal performances: averaged gradient descent and its application to tensor pca. *Journal of Physics A: Mathematical and Theoretical*, 53(17):174003, 2020.
- Vincenzo Bonifaci. Physarum can compute shortest paths: A short proof. *Information Processing Letters*, 113(1-2):4–7, 2013.
- Vincenzo Bonifaci, Kurt Mehlhorn, and Girish Varma. Physarum can compute shortest paths. *Journal of theoretical biology*, 309:121–133, 2012.
- Alfredo Braunstein and Guilhem Semerjian. The cavity method: from exact solutions to algorithms. In *Spin Glass Theory and Far Beyond: Replica Symmetry Breaking After 40 Years*, pages 375–387. World Scientific, 2023.
- Mark Braverman and Ankur Moitra. An information complexity approach to extended formulations. In *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*, pages 161–170, 2013.
- Matthew Brennan and Guy Bresler. Reducibility and statistical-computational gaps from secret leakage. In *Conference on Learning Theory*, pages 648–847. PMLR, 2020.
- Matthew Brennan, Guy Bresler, Samuel B Hopkins, Jerry Li, and Tselil Schramm. Statistical query algorithms and low-degree tests are almost equivalent. *arXiv preprint arXiv:2009.06107*, 2020.
- Guy Bresler and Alina Harbuzova. Computational equivalence of spiked covariance and spiked wigner models via gram-schmidt perturbation. *arXiv preprint arXiv:2503.02802*, 2025.

- Guy Bresler and Brice Huang. The algorithmic phase transition of random k-sat for low degree polynomials. In *2021 IEEE 62nd annual symposium on foundations of computer science (FOCS)*, pages 298–309. IEEE, 2022.
- Rares-Darius Buhai, Jun-Ting Hsieh, Aayush Jain, and Pravesh K Kothari. The quasi-polynomial low-degree conjecture is false. *arXiv preprint arXiv:2505.17360*, 2025.
- Dmitri Burago, Yuri Burago, Sergei Ivanov, et al. *A course in metric geometry*, volume 33. American Mathematical Society Providence, 2001.
- Paul M Chaikin and Tom C Lubensky. *Principles of condensed matter physics*, volume 10. Cambridge university press Cambridge, 1995.
- Sourav Chatterjee. Disorder chaos and multiple valleys in spin glasses. *arXiv preprint arXiv:0907.3381*, 2009.
- Sourav Chatterjee and Amir Dembo. Nonlinear large deviations. *Advances in Mathematics*, 299: 396–450, 2016.
- Louis H. Y. Chen. Poisson Approximation for Dependent Trials. *The Annals of Probability*, 3(3): 534 – 545, 1975. doi: 10.1214/aop/1176996359. URL <https://doi.org/10.1214/aop/1176996359>.
- Wei-Kuo Chen, David Gamarnik, Dmitry Panchenko, and Mustazee Rahman. Suboptimality of local algorithms for a class of max-cut problems. *Annals of Probability*, 2019.
- Zongchen Chen, Conor Sheehan, and Ilias Zadik. On the low-temperature mcmc threshold: the cases of sparse tensor pca, sparse regression, and a geometric rule. *arXiv preprint arXiv:2408.00746*, 2024.
- Chi-Ning Chou, Peter J Love, Juspreet Singh Sandhu, and Jonathan Shi. Limitations of local quantum algorithms on random max-k-xor and beyond. In *49th International Colloquium on Automata, Languages, and Programming (ICALP 2022)*, volume 229, page 41. Schloss Dagstuhl–Leibniz-Zentrum für Informatik, 2022.
- Fan Chung and Linyuan Lu. The diameter of sparse random graphs. *Advances in Applied Mathematics*, 26(4):257–279, 2001.
- Amin Coja-Oghlan. The asymptotic k-sat threshold. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, pages 804–813, 2014.
- Amin Coja-Oghlan and Charilaos Efthymiou. On independent sets in random graphs. *Random Structures & Algorithms*, 47(3):436–486, 2015.
- Amin Coja-Oghlan and Konstantinos Panagiotou. Catching the k-naesat threshold. In *Proceedings of the forty-fourth annual ACM symposium on Theory of computing*, pages 899–908, 2012.
- Michele Conforti, Gérard Cornuéjols, and Giacomo Zambelli. Extended formulations in combinatorial optimization. *4OR*, 8(1):1–48, 2010.

- Amir Dembo and Ofer Zeitouni. *Large Deviations Techniques and Applications*. Stochastic Modelling and Applied Probability. Springer Berlin Heidelberg, 2009. ISBN 9783642033117. URL <https://link.springer.com/book/10.1007/978-3-642-03311-7>.
- Amir Dembo, Thomas M Cover, and Joy A Thomas. Information theoretic inequalities. *IEEE Transactions on Information theory*, 37(6):1501–1518, 1991.
- Yash Deshpande and Andrea Montanari. Finding hidden cliques of size  $n/e$  in nearly linear time. *Foundations of Computational Mathematics*, 15(4):1069–1128, 2015.
- Jian Ding, Allan Sly, and Nike Sun. Proof of the satisfiability conjecture for large  $k$ . In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pages 59–68, 2015.
- Ahmed El Alaoui and Florent Krzakala. Estimation in the spiked wigner model: a short proof of the replica formula. In *2018 IEEE International Symposium on Information Theory (ISIT)*, pages 1874–1878. IEEE, 2018.
- Ahmed El Alaoui, Andrea Montanari, and Mark Sellke. Sampling from the sherrington-kirkpatrick gibbs measure via algorithmic stochastic localization. In *2022 IEEE 63rd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 323–334. IEEE, 2022.
- Ahmed El Alaoui, Andrea Montanari, and Mark Sellke. Shattering in pure spherical spin glasses. *Communications in Mathematical Physics*, 406(5):1–36, 2025.
- Ronen Eldan. Gaussian-width gradient complexity, reverse log-sobolev inequalities and nonlinear large deviations. *Geometric and Functional Analysis*, 28(6):1548–1596, 2018.
- Ronen Eldan. Taming correlations through entropy-efficient measure decompositions with applications to mean-field approximation. *Probability Theory and Related Fields*, 176(3):737–755, 2020.
- Ronen Eldan and Renan Gross. Decomposition of mean-field gibbs distributions into product measures. *Electronic Journal of Probability*, 23, 2018.
- Richard S Ellis. *Entropy, large deviations, and statistical mechanics*. Springer, 2007.
- Silvio Franz and Giorgio Parisi. Recipes for metastable states in spin glasses. *Journal de Physique I*, 5(11):1401–1415, 1995.
- Sacha Friedli and Yvan Velenik. *Statistical mechanics of lattice systems: a concrete mathematical introduction*. Cambridge University Press, 2017.
- David Gamarnik. The overlap gap property: A topological barrier to optimizing over random structures. *Proceedings of the National Academy of Sciences*, 118(41):e2108492118, 2021.
- David Gamarnik and Madhu Sudan. Limits of local algorithms over sparse random graphs. In *Proceedings of the 5th conference on Innovations in theoretical computer science*, pages 369–376, 2014.
- David Gamarnik and Madhu Sudan. Performance of sequential local algorithms for the random nae-k-sat problem. *SIAM Journal on Computing*, 46(2):590–619, 2017.

- David Gamarnik and Ilias Zadik. The landscape of the planted clique problem: Dense subgraphs and the overlap gap property. *The Annals of Applied Probability*, 34(4):3375–3434, 2024.
- David Gamarnik, Devavrat Shah, and Yehua Wei. Belief propagation for min-cost network flow: Convergence and correctness. *Operations research*, 60(2):410–428, 2012.
- David Gamarnik, Eren C Kızıldağ, Will Perkins, and Changji Xu. Algorithms and barriers in the symmetric binary perceptron model. In *2022 IEEE 63rd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 576–587. IEEE, 2022.
- David Gamarnik, Aukosh Jagannath, and Alexander S Wein. Hardness of random optimization problems for boolean circuits, low-degree polynomials, and langevin dynamics. *SIAM Journal on Computing*, 53(1):1–46, 2024.
- Reza Gheissari, Aukosh Jagannath, and Yiming Xu. Finding planted cliques using gradient descent. *SIAM Journal on Mathematics of Data Science*, 7(2):643–669, 2025.
- Michel X Goemans. Smallest compact formulation for the permutahedron. *Mathematical Programming*, 153(1):5–11, 2015.
- Herbert Goldstein. *Classical Mechanics*. Addison-Wesley, 1980.
- Mika Goos, Rahul Jain, and Thomas Watson. Extension complexity of independent set polytopes. *SIAM Journal on Computing*, 47(1):241–269, 2018.
- Matthew A Grayson. Shortening embedded curves. *Annals of Mathematics*, 129(1):71–111, 1989.
- Thomas Dueholm Hansen, Haim Kaplan, and Uri Zwick. Dantzig’s pivoting rule for shortest paths, deterministic mdps, and minimum cost to time ratio cycles. In *Proceedings of the Twenty-Fifth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 847–860. SIAM, 2014.
- Justin Holmgren and Alexander S Wein. Counterexamples to the low-degree conjecture. In *12th Innovations in Theoretical Computer Science Conference (ITCS 2021)*, volume 185, 2021.
- Samuel B Hopkins, Jonathan Shi, and David Steurer. Tensor principal component analysis via sum-of-square proofs. In *Conference on Learning Theory*, pages 956–1006. PMLR, 2015.
- Samuel B Hopkins, Pravesh K Kothari, Aaron Potechin, Prasad Raghavendra, Tselil Schramm, and David Steurer. The power of sum-of-squares for detecting hidden structures. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 720–731. IEEE, 2017.
- Brice Huang and Mark Sellke. Algorithmic threshold for multi-species spherical spin glasses. *arXiv preprint arXiv:2303.12172*, 2023.
- Brice Huang and Mark Sellke. Strong low degree hardness for stable local optima in spin glasses. *arXiv preprint arXiv:2501.06427*, 2025.
- Han Huang and Elchanan Mossel. Low degree hardness for broadcasting on trees. In *Advances in Neural Information Processing Systems 37*, 2024.

- Han Huang and Elchanan Mossel. Optimal low degree hardness for broadcasting on trees. *arXiv preprint arXiv:2502.04861*, 2025.
- Peter J Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101, 1964.
- Vishesh Jain, Frederic Koehler, and Andrej Risteski. Mean-field approximation, convex hierarchies, and the optimality of correlation rounding: a unified perspective. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, pages 1226–1236, 2019.
- Samir Khuller, Balaji Raghavachari, and Neal Young. Balancing minimum spanning trees and shortest-path trees. *Algorithmica*, 14(4):305–321, 1995.
- Jeong Han Kim and James R Roche. Covering cubes by random half cubes, with applications to binary neural networks. *Journal of Computer and System Sciences*, 56(2):223–252, 1998.
- Gustav Kirchhoff. On the solution of the equations obtained from the investigation of the linear distribution of galvanic currents. *IRE transactions on circuit theory*, 5(1):4–7, 1958.
- Eren C Kızıldağ. Sharp phase transition for multi overlap gap property in ising  $p$ -spin glass and random  $k$ -sat models. *arXiv preprint arXiv:2309.09913*, 2023.
- Bo’az Klartag and Joseph Lehec. Bourgain’s slicing problem and kls isoperimetry up to polylog. *Geometric and functional analysis*, 32(5):1134–1159, 2022.
- Jon Kleinberg and Eva Tardos. *Algorithm design*. Pearson Education India, 2006.
- Frederic Koehler. Fast convergence of belief propagation to global optima: Beyond correlation decay. *Advances in Neural Information Processing Systems*, 32, 2019.
- Frederic Koehler and Elchanan Mossel. Reconstruction on trees and low-degree polynomials. *Advances in Neural Information Processing Systems*, 35:18942–18954, 2022.
- Florent Krzakala, Andrea Montanari, Federico Ricci-Tersenghi, Guilhem Semerjian, and Lenka Zdeborová. Gibbs states and the set of solutions of random constraint satisfaction problems. *Proceedings of the National Academy of Sciences*, 104(25):10318–10323, 2007.
- Steffen L Lauritzen. *Graphical models*, volume 17. Clarendon Press, 1996.
- James R Lee, Prasad Raghavendra, and David Steurer. Lower bounds on the size of semidefinite programming relaxations. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pages 567–576, 2015.
- Jing Lei. Convergence and concentration of empirical measures under Wasserstein distance in unbounded functional spaces. *Bernoulli*, 26(1):767 – 798, 2020. doi: 10.3150/19-BEJ1151. URL <https://doi.org/10.3150/19-BEJ1151>.
- Shuangping Li and Tselil Schramm. Some easy optimization problems have the overlap-gap property. In Nika Haghtalab and Ankur Moitra, editors, *Proceedings of Thirty Eighth Conference on Learning Theory*, volume 291 of *Proceedings of Machine Learning Research*, pages 3582–3622. PMLR, 30 Jun–04 Jul 2025. URL <https://proceedings.mlr.press/v291/li25b.html>.

- László Lovász. Submodular functions and convexity. *Mathematical Programming The State of the Art: Bonn 1982*, pages 235–257, 1983.
- Eric Ma and Tselil Schramm. Polynomial-time sampling despite disorder chaos. *arXiv preprint arXiv:2508.04133*, 2025.
- Saunders Mac Lane. *Categories for the working mathematician*, volume 5. Springer Science & Business Media, 2013.
- Marc Mezard and Andrea Montanari. *Information, physics, and computation*. Oxford University Press, 2009.
- Marc Mézard, Giorgio Parisi, and Miguel Angel Virasoro. *Spin glass theory and beyond: An Introduction to the Replica Method and Its Applications*, volume 9. World Scientific Publishing Company, 1987.
- Marc Mézard, Thierry Mora, and Riccardo Zecchina. Clustering of solutions in the random satisfiability problem. *Physical Review Letters*, 94(19):197205, 2005.
- Andrea Montanari and Emile Richard. A statistical model for tensor pca. *Advances in neural information processing systems*, 27, 2014.
- Rajeev Motwani and Prabhakar Raghavan. *Randomized Algorithms*. Cambridge University Press, 1995.
- Kazuo Murota. *Discrete Convex Analysis*, volume 10. SIAM, 2003.
- Kazuo Murota and Akiyoshi Shioura. Dijkstra’s algorithm and l-concave function maximization. *Mathematical Programming*, 145(1):163–177, 2014.
- James B Orlin. A polynomial time primal network simplex algorithm for minimum cost flows. *Mathematical Programming*, 78(2):109–129, 1997.
- James B Orlin. On the simplex algorithm for networks and generalized networks. In *Mathematical Programming Essays in Honor of George B. Dantzig Part I*, pages 166–178. Springer, 2009.
- Giorgio Parisi. *Statistical field theory*. New York: Addison-Wesley, 1988.
- Will Perkins and Changji Xu. Frozen 1-rsb structure of the symmetric ising perceptron. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 1579–1588, 2021.
- Oliver Riordan and Nicholas Wormald. The diameter of sparse random graphs. *Combinatorics, Probability and Computing*, 19(5-6):835–926, 2010.
- Thomas Rothvoß. The matching polytope has exponential extension complexity. *Journal of the ACM (JACM)*, 64(6):1–19, 2017.
- Nicholas Ruozzi. The bethe partition function of log-supermodular graphical models. *Advances in Neural Information Processing Systems*, 25, 2012.

- YiFan Sun, Andrea Crisanti, Florent Krzakala, Luca Leuzzi, and Lenka Zdeborová. Following states in temperature in the spherical s+ p-spin glass model. *Journal of Statistical Mechanics: Theory and Experiment*, 2012(07):P07002, 2012.
- Kenneth E Train. *Discrete choice methods with simulation*. Cambridge university press, 2009.
- Ramon Van Handel. Probability in high dimension. *Lecture Notes (Princeton University)*, 2(3):2–3, 2014.
- Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018.
- Cédric Villani. *Topics in optimal transportation*, volume 58. American Mathematical Soc., 2021.
- Martin J Wainwright, Michael I Jordan, et al. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008.
- Alexander S Wein. Optimal low-degree hardness of maximum independent set. *Mathematical Statistics and Learning*, 4(3):221–251, 2022.
- Alexander S Wein, Ahmed El Alaoui, and Cristopher Moore. The kikuchi hierarchy and tensor pca. In *2019 IEEE 60th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 1446–1468. IEEE, 2019.
- Elizabeth L Wilmer, David A Levin, and Yuval Peres. Markov chains and mixing times. *American Mathematical Soc., Providence*, 107, 2009a.
- Elizabeth L Wilmer, David A Levin, and Yuval Peres. Markov chains and mixing times. *American Mathematical Soc., Providence*, 2009b.
- Jonathan S Yedidia, William T Freeman, Yair Weiss, et al. Understanding belief propagation and its generalizations. *Exploring artificial intelligence in the new millennium*, 8(236–239):0018–9448, 2003.
- Lenka Zdeborová and Florent Krzakala. Generalization of the cavity method for adiabatic evolution of gibbs states. *Physical Review B—Condensed Matter and Materials Physics*, 81(22):224205, 2010.

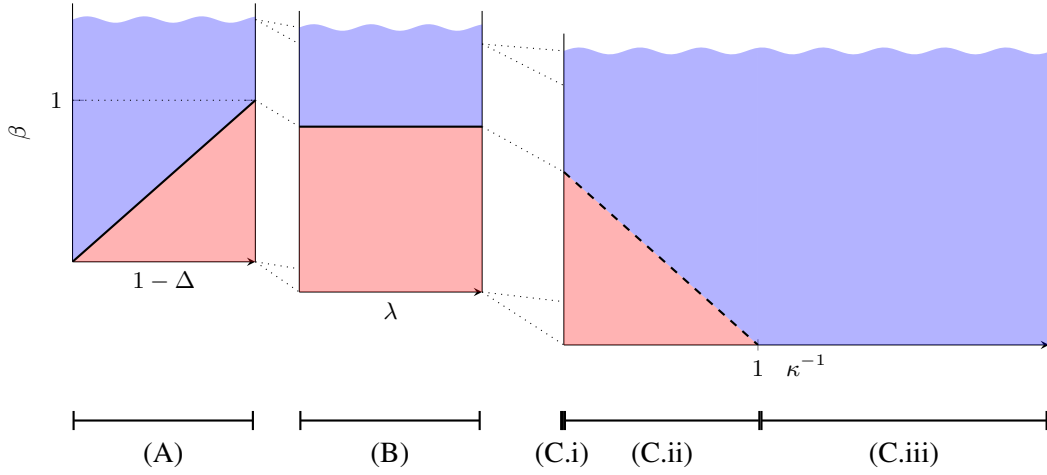


Figure 3: Phase diagram in the limit  $n \rightarrow \infty$ . The second diagram is obtained by “zooming in” the first diagram at  $1 - \Delta = 1$ , and the third diagram is from zooming in the second diagram at  $\lambda = 1$ . The labels for the regions correspond to the regimes (A) through (C.iii). The **blue region** is the low-temperature phase where the Gibbs measure is close to the uniform measure in Wasserstein distance, and the **red region** is the high-temperature phase where they are far apart. Solid black lines indicate a discontinuous (“first-order”) phase transition, and dashed lines indicate a continuous (“second-order”) phase transition. The point at which the line changes from solid to dashed is precisely where the parameter regime transitions from  $\lambda < 1$  to  $\lambda = 1$ .

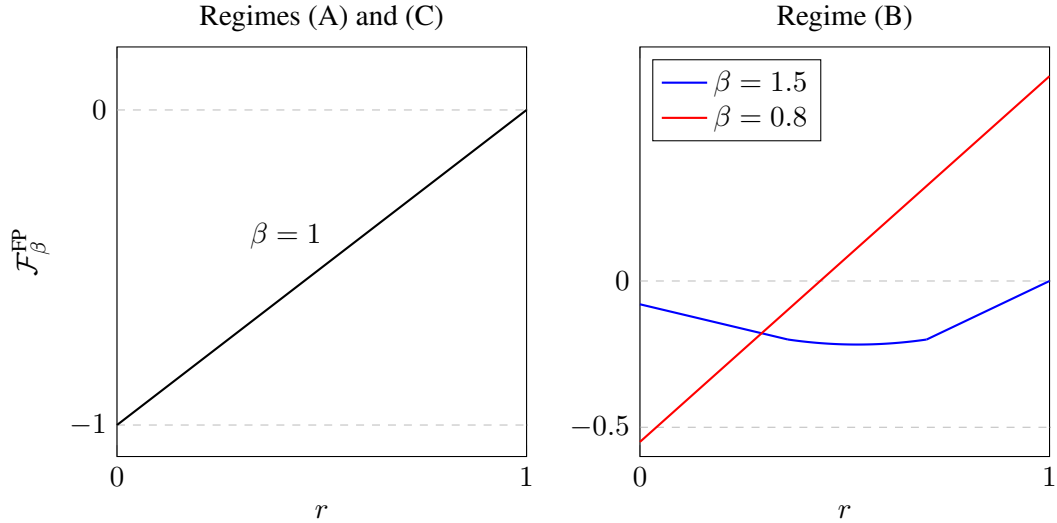


Figure 4: A plot of the Franz–Parisi potential against the correlation  $r \in [0, 1]$  for different regimes. The  $y$ -axes (for  $\mathcal{F}_\beta^{\text{FP}}$ ) are scaled differently to better visualize the landscape. For the right plot we choose  $\lambda = 0.3$  (and  $\Delta = 0$ ). A low temperature phase  $\beta = 1.5$  in the right plot has its unique *quasiconvex* landscape which is quite different from the other cases which simply increase in  $r$ .

## Contents

|          |                                                                                   |           |
|----------|-----------------------------------------------------------------------------------|-----------|
| <b>1</b> | <b>Introduction</b>                                                               | <b>1</b>  |
| 1.1      | Background on the OGP and analogy with Gibbs measures . . . . .                   | 5         |
| 1.2      | Main results . . . . .                                                            | 6         |
| 1.2.1    | Basic models and notations . . . . .                                              | 6         |
| 1.2.2    | Overlap concentration and ensemble-OGP . . . . .                                  | 7         |
| 1.2.3    | Gibbs measures and absence of free energy barrier . . . . .                       | 8         |
| <b>2</b> | <b>Technical ingredients: the shortest path structure of sparse random graphs</b> | <b>9</b>  |
| 2.1      | Asymptotics for a single graph . . . . .                                          | 10        |
| 2.2      | Asymptotics for correlated graphs . . . . .                                       | 10        |
| 2.3      | Gibbs measures, phase transitions, and energy landscape . . . . .                 | 11        |
| 2.3.1    | Free energy density . . . . .                                                     | 12        |
| 2.3.2    | Phase transitions . . . . .                                                       | 13        |
| <b>A</b> | <b>Analysis overview</b>                                                          | <b>28</b> |
| A.1      | Asymptotics of correlated random graphs . . . . .                                 | 28        |
| A.2      | Overlap of shortest path trees for correlated graphs . . . . .                    | 31        |
| A.3      | Gibbs measures over spanning trees . . . . .                                      | 32        |
| <b>B</b> | <b>Preliminaries</b>                                                              | <b>35</b> |
| B.1      | Additional notations . . . . .                                                    | 35        |
| B.2      | Wasserstein distance . . . . .                                                    | 37        |
| B.3      | Shortest path structures . . . . .                                                | 37        |
| B.4      | Gibbs measures . . . . .                                                          | 38        |
| <b>C</b> | <b>Asymptotics for a single graph</b>                                             | <b>39</b> |
| C.1      | The distribution of distances in a single graph . . . . .                         | 39        |
| C.2      | Shortest paths between two fixed vertices . . . . .                               | 45        |
| <b>D</b> | <b>Asymptotics for correlated Erdős–Rényi graphs</b>                              | <b>47</b> |
| D.1      | The overlap of the distance vectors . . . . .                                     | 47        |
| D.2      | The overlap of the shortest path DAGs . . . . .                                   | 53        |
| D.3      | The overlap of random shortest path trees . . . . .                               | 58        |
| D.3.1    | The overlap under optimal coupling . . . . .                                      | 58        |
| D.3.2    | Aside: the overlap under independent coupling . . . . .                           | 63        |
| D.4      | No overlap concentration for random shortest paths . . . . .                      | 64        |
| <b>E</b> | <b>The partition function of Gibbs measures over trees</b>                        | <b>65</b> |
| E.1      | Preliminaries . . . . .                                                           | 66        |
| E.2      | Transformation into a Markov random field . . . . .                               | 67        |
| E.3      | Analysis of the partition function . . . . .                                      | 68        |
| E.3.1    | Main results . . . . .                                                            | 69        |
| E.3.2    | Regularity properties of the underlying graph . . . . .                           | 71        |
| E.3.3    | An informal description of the Gibbs measures . . . . .                           | 72        |

|          |                                                                                        |            |
|----------|----------------------------------------------------------------------------------------|------------|
| E.3.4    | Kernels and conditional Gibbs measures . . . . .                                       | 73         |
| E.3.5    | Proof of Proposition 46 . . . . .                                                      | 78         |
| E.3.6    | Concentration of the log partition function . . . . .                                  | 85         |
| E.3.7    | Proof of Theorem 41 . . . . .                                                          | 91         |
| E.3.8    | Proof of Theorem 42 . . . . .                                                          | 93         |
| <b>F</b> | <b>Analysis of Gibbs measures over trees</b>                                           | <b>98</b>  |
| F.1      | Structure of the Gibbs measures . . . . .                                              | 99         |
| F.1.1    | Uniform measure as ground state . . . . .                                              | 99         |
| F.1.2    | Distribution of distance vectors . . . . .                                             | 100        |
| F.1.3    | Conditional Gibbs measures are almost uniform . . . . .                                | 104        |
| F.1.4    | Gibbs measures in the low temperature phase . . . . .                                  | 108        |
| F.1.5    | Gibbs measures in the high temperature phase . . . . .                                 | 111        |
| F.1.6    | Replica symmetry . . . . .                                                             | 118        |
| F.2      | Phase transitions . . . . .                                                            | 120        |
| F.2.1    | The free energy density . . . . .                                                      | 120        |
| F.2.2    | Phase transitions in Wasserstein space . . . . .                                       | 122        |
| F.3      | The Franz–Parisi potential . . . . .                                                   | 125        |
| F.3.1    | Definitions and main results . . . . .                                                 | 125        |
| F.3.2    | Proof of Proposition 85 . . . . .                                                      | 129        |
| F.3.3    | Proof of Proposition 87 . . . . .                                                      | 132        |
| F.4      | The energy landscape of the Gibbs measures . . . . .                                   | 134        |
| <b>G</b> | <b>Related work</b>                                                                    | <b>136</b> |
| <b>H</b> | <b>Mathematical tools</b>                                                              | <b>138</b> |
| H.1      | Concentration inequalities . . . . .                                                   | 138        |
| H.2      | Logarithm of binomial random variables . . . . .                                       | 142        |
| H.3      | Poisson approximation . . . . .                                                        | 146        |
| H.4      | LogSumExp functions and empirical measures . . . . .                                   | 147        |
| H.5      | Large deviations . . . . .                                                             | 148        |
| H.6      | Useful inequalities . . . . .                                                          | 153        |
| <b>I</b> | <b>Gibbs measures over trees near the critical temperature</b>                         | <b>153</b> |
| I.1      | The mean-field approximation . . . . .                                                 | 153        |
| I.2      | Optimal kernel size at the critical temperature . . . . .                              | 155        |
| I.3      | Internal energy density as a function of temperature . . . . .                         | 157        |
| <b>J</b> | <b>Further discussion</b>                                                              | <b>158</b> |
| J.1      | Additional analogies: geodesics and extension complexity . . . . .                     | 158        |
| J.2      | The physical interpretation of Dijkstra’s algorithm . . . . .                          | 159        |
| J.3      | Dijkstra’s algorithm and belief propagation . . . . .                                  | 161        |
| J.4      | Instability of the projection map from shortest path trees to shortest paths . . . . . | 162        |
| J.5      | Sampling a Gibbs measure is equivalent to a random optimization problem . . . . .      | 163        |
| J.6      | Free energy barriers trap Markov chains . . . . .                                      | 164        |
| J.7      | Comment on replica method approach . . . . .                                           | 166        |

|          |                                                                                                    |            |
|----------|----------------------------------------------------------------------------------------------------|------------|
| J.8      | Simulation details . . . . .                                                                       | 167        |
| <b>K</b> | <b>Relationship to discrete convexity</b>                                                          | <b>167</b> |
| K.1      | L-convexity and M-convexity . . . . .                                                              | 167        |
| K.2      | Graphic matroid and network simplex . . . . .                                                      | 169        |
| K.3      | Dijkstra’s algorithm and L-concave function maximization <a href="#">Murota and Shioura (2014)</a> | 170        |
| K.4      | Shortest path tree is not discrete convex . . . . .                                                | 171        |
| <b>L</b> | <b>Free energy barrier in the space of paths</b>                                                   | <b>171</b> |
| L.1      | Basic properties of Gibbs measures on paths . . . . .                                              | 172        |
| L.2      | Key structural result from Li–Schramm <a href="#">Li and Schramm (2025)</a> . . . . .              | 173        |
| L.3      | Free energy barriers and Franz–Parisi Potential . . . . .                                          | 173        |
| <b>M</b> | <b>An analogous example: local search on multilinear vs Lovász extension</b>                       | <b>174</b> |
| M.1      | Local search in a simple example . . . . .                                                         | 175        |
| M.2      | High-dimensional version . . . . .                                                                 | 176        |
| M.2.1    | Ising Model and Franz–Parisi Potential . . . . .                                                   | 176        |
| M.2.2    | Analytical computation of Franz–Parisi Potential . . . . .                                         | 177        |
| M.2.3    | Metastability and contrast to log-concave setting . . . . .                                        | 178        |

## Notation

|                         |                                                                                             |
|-------------------------|---------------------------------------------------------------------------------------------|
| $n$                     | Number of vertices of a graph                                                               |
| $q_n$                   | Edge probability                                                                            |
| $\mathcal{G}(n, q_n)$   | Erdős–Rényi model                                                                           |
| $\overline{G}$          | Connected component of $G$ containing 1                                                     |
| $d_G(u, v)$             | Distance between $u$ and $v$ in graph $G$                                                   |
| $\text{par}_G(v)$       | Neighbor(s) of $v$ with distance to 1 shorter than $v$                                      |
| $\text{ch}_G(v)$        | Neighbor(s) of $v$ with distance to 1 longer than $v$                                       |
| $\Gamma_d$              | Set of vertices at distance $d$ from 1                                                      |
| $N_d$                   | Number of vertices at distance $d$ from 1, i.e., $N_d =  \Gamma_d $                         |
| $\alpha_n$              | Parameter satisfying $q_n = \alpha_n \log n/n$                                              |
| $\rho_n$                | Graph correlation; see Section 1.2.1                                                        |
| $\ell_n^*$              | Parameter satisfying $(nq_n)^{\ell_n^*} = n$                                                |
| $d_n^*$                 | Smallest integer $d$ satisfying $(nq_n)^d \geq n/(\log \log n)^2$                           |
| $\Delta_n$              | Gap $d_n^* - \ell_n^*$                                                                      |
| $\lambda_n$             | Parameter satisfying $\log(1/\lambda_n) = (nq_n)^{\Delta_n}$                                |
| $\kappa_n$              | Parameter satisfying $\kappa_n = (1 - \lambda_n) \log \log n$                               |
| $\gamma_n$              | Parameter satisfying $\gamma_n = \rho_n^{d_n^*}$                                            |
| $R(T, T')$              | Overlap of two spanning trees; see (3)                                                      |
| $\beta$                 | Inverse temperature in Gibbs measure                                                        |
| $\overline{\beta}$      | Scaled inverse temperature $\overline{\beta} = \beta \log \log n$                           |
| $\mu_{G,\beta}$         | Gibbs measure defined by the sum of distances as the Hamiltonian (energy function); see (5) |
| $Z_{G,\beta}$           | Partition function of the Gibbs measure $\mu_{G,\beta}$                                     |
| $\mathcal{F}_{G,\beta}$ | Scaled free energy density; see (13)                                                        |
| $W_1(\mu, \nu)$         | 1-Wasserstein distance between two measures                                                 |
| $\text{KL}(\mu, \nu)$   | Kullback–Leibler divergence of two measures                                                 |
| $H_\mu(f(X))$           | Shannon entropy of $f(X)$ where $X \sim \mu$                                                |
| $\varphi(T)$            | Kernel of a spanning tree $T$ ; see Definition 44                                           |

## Organization

Our proofs of the main results are very long and consists of many sophisticated arguments, so we break them down into a number of components. As a high-level overview of our proof strategy, Appendix A explains some key ideas that help the interested reader go through the detailed proofs.

In Appendix B, we begin with some preliminaries such as graph-theoretic notation which is used throughout the main parts of the proofs. Appendix C is the beginning of our analysis where we establish some key facts about the distribution of distances in a random graph. Here as in the rest of the paper, we reference mathematical tools such as Chernoff bounds, Poisson approximation, large deviations theory, etc. which are discussed in Appendix H.

Next is Appendix D, where we analyze the overlap between shortest path trees (and other related structures) under partial rerandomization of the graph. These are the results which directly imply that the OGP does *not* hold in the landscape over trees. Note that Appendix D is the *only* section where two correlated random graphs are considered — in the Gibbs measure analysis, a single random graph is sampled once (“quenched”) and the randomness comes only from sampling spanning trees.

In Appendix E we compute asymptotics for the log partition function of the finite temperatures Gibbs measure over random graphs. This is at the heart of the results in the following Appendix F, where we prove our main results about phase transitions, the Franz–Parisi potential, and the approximate structure of the Gibbs measures/energy landscape. The parts of the analysis which are directly used in both the OGP and Franz-Parisi analysis are all contained in Appendix C, so Appendix D can be read before or after Appendices E and F.

The paper concludes with a few additional appendices. Appendix G covers some related work on the overlap gap property and the Franz–Parisi potential. Appendix I discusses the behavior of the Gibbs measures near the critical temperature, complementing Appendix E and Appendix F. Appendix J contains some discussion and mathematical background which is helpful for understanding and interpreting the main results of the paper. Appendix K is a self-contained appendix discussing the shortest path problem from the perspective of matroid theory and discrete convexity. In particular, in Appendix K.2 we elaborate on the connection between the landscape over trees and the *network simplex* algorithm. This discussion explains why as a basic fact we know that local search in the space of trees must always be globally convergent, as well as generalizations of this observation which exist in the literature. In Appendix L we elaborate on the interpretation of the main structural result of Li and Schramm (2025) in terms of the Franz–Parisi potential on the space of paths. In Appendix M, we elaborate on the analogy between our result and local search on different extensions of a submodular function. This includes discussion of the corresponding Franz–Parisi potentials.

## Appendix A. Analysis overview

In this section, we give a high level overview of our analysis along with some informal arguments. Throughout this section, in the interest of simplicity, we ignore any subtleties that might result from the graph  $G_n$  being disconnected.

### A.1. Asymptotics of correlated random graphs

**The distribution of distances and BFS.** Before analyzing the correlated graph setting relevant to the OGP, it helps to carefully understand what happens with a single graph. As illustrated in Figure 6, we sample/reveal the Erdős–Rényi graph  $G = G_n$  progressively via Breadth-First Search (BFS), as in the literature regarding the diameter of random graphs [Chung and Lu \(2001\)](#); [Riordan and Wormald \(2010\)](#). Conditioned on the set of vertices of distance up to  $d$  from a fixed source vertex 1, we next reveal which vertices are distance  $d+1$ . The edges between the remaining vertices and the vertices of distance  $d$  are sampled independently of all the observed edges/non-edges so far. Therefore, the number  $N_d$  of the vertices of distance  $d$  follows the stochastic process

$$\begin{aligned} N_0 &= 1 \\ N_1 &\sim \text{Binom}(n - N_0, q_n) \\ &\vdots \\ N_d &\sim \text{Binom}\left(n - \sum_{i=0}^{d-1} N_i, 1 - (1 - q_n)^{N_{d-1}}\right) \\ &\vdots \end{aligned} \tag{15}$$

which terminates once the BFS has exhausted its connected component.

Based on well-known properties of sparse random graphs, we know that for small depths  $d$  the graph is locally tree-like, so  $N_d$  will concentrate around  $(nq_n)^d$ . Let's review why this occurs, so we can understand what happens at large depths. Suppose that  $N_k$  are given for all  $k < d$  and it is known that  $N_k \approx (nq)^k$ . Then from (15),  $N_d$  would concentrate around

$$N_d \approx \left(n - \sum_{i=0}^{d-1} N_i\right) \left(1 - (1 - q_n)^{N_{d-1}}\right). \tag{16}$$

If  $d$  is small enough, then we should have

$$n - \sum_{i=0}^{d-1} N_i \approx n - \sum_{i=0}^{d-1} (nq_n)^i \approx n \tag{17}$$

and

$$1 - (1 - q_n)^{N_{d-1}} \approx 1 - e^{-q_n(nq_n)^{d-1}} \approx q_n(nq_n)^{d-1} \tag{18}$$

so  $N_d \approx (nq_n)^d$ .

However, this starts to fail when we run low on unexplored vertices, i.e. when  $(nq_n)^d$  is close to  $n$ , equivalently when  $d$  is close to  $\ell_n^* = \frac{\log n}{\log(nq_n)}$ . Suppose now that we have reached  $d = d_n^*$ . By the definition (7) of  $d_n^*$ , we still have  $N_{d_n^*-1} \approx (nq_n)^{d_n^*-1} < \frac{n}{(\log \log n)^2} = o(n)$ , so (17) still

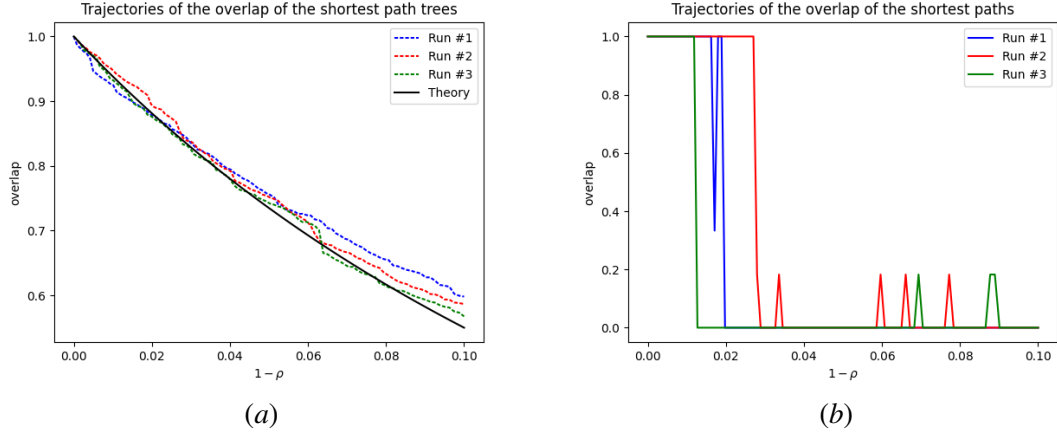


Figure 5: Numerical simulation of the overlap of uniformly random shortest path trees vs. shortest paths. It was conducted independently three times with parameters  $n = 10^5$  and  $q = 10^{-4}$ . The overlap of the uniformly randomly sampled shortest path trees from the optimal coupling (left) and the overlap of the shortest paths from the root to a fixed vertex, where the paths are taken from the sampled shortest path trees (right).

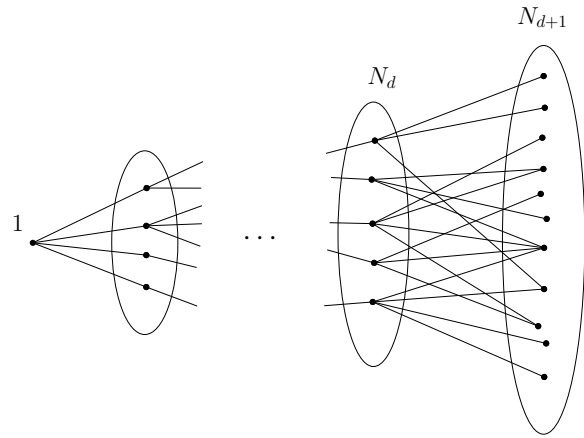


Figure 6: An illustration of the BFS process.

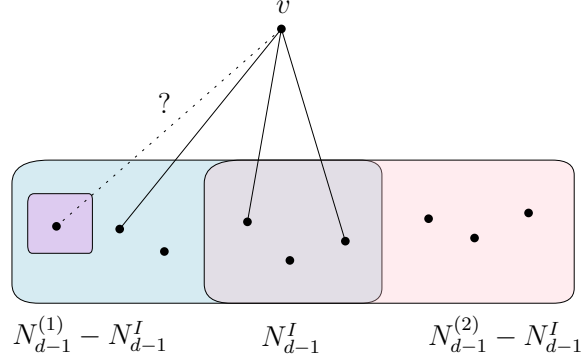


Figure 7: The “adjacency probabilities” problem in correlated graphs. The blue region in the left and the pink region in the right are those explored at step  $d - 1$  for  $G_n^{(1)}$  and  $G_n^{(2)}$ , respectively. If the vertices in the purple box in the blue region is previously explored in  $G_n^{(2)}$ , then it is already known that they are not connected to the new vertex  $v$ .

holds true. On the other hand, the last approximation in (18) fails if  $q_n(nq_n)^{d_n^*-1} = \Omega(1)$ . This is the case where, using the notation of (9), we have that  $\lambda_n = e^{-q_n(nq_n)^{d_n^*-1}} = \Omega(1)$ , and  $N_{d_n^*}$  occupies a nontrivial portion of the vertices. This causes (17) to now fail, and instead we observe that if  $N_{d_n^*}$  does not exhaust most of the remaining vertices, then  $N_{d_n^*+1}$  will. The formal version of this analysis is detailed in Proposition 16; it requires some care since the number of layers  $d_n^*$  is diverging with  $n$ , and we need fairly precise bounds to enable the analysis in what follows.

**Correlated graphs analysis.** We now build upon the ideas above to analyze the more difficult setting of two  $\rho_n$ -correlated graphs  $G_n^{(1)}$  and  $G_n^{(2)}$ . We are now interested in not only  $N_d^{(1)}$  and  $N_d^{(2)}$ , which denote  $N_d$  for  $G_n^{(1)}$  and  $G_n^{(2)}$  respectively, but also  $N_d^I$  — the set of vertices that have distance  $d$  in both. Suppose that we are running a BFS on both graphs simultaneously up to distance  $d - 1$ , and we are currently about to explore the vertices at distance  $d \leq d_n^*$ . By (17), most of the vertices are still unexplored (in both graphs) and among those we are searching for vertices at distance  $d$  in both  $G_n^{(1)}$  and  $G_n^{(2)}$ . We can categorize each vertex  $u$  explored in the previous step  $d - 1$  into the three categories:

- (a)  $u$  has depth  $d - 1$  in  $G_n^{(1)}$  but not in  $G_n^{(2)}$ , counted in  $N_{d-1}^{(1)} - N_{d-1}^I$ ;
- (b)  $u$  has depth  $d - 1$  in  $G_n^{(2)}$  but not in  $G_n^{(1)}$ , counted in  $N_{d-1}^{(2)} - N_{d-1}^I$ ;
- (c)  $u$  has depth  $d - 1$  in both  $G_n^{(1)}$  and  $G_n^{(2)}$ , counted in  $N_{d-1}^I$ .

For each unexplored vertex  $v$ , we would like to express the probability of  $v$  being counted towards  $N_d^I$  in terms of  $(N_{d-1}^{(1)}, N_{d-1}^{(2)}, N_{d-1}^I)$ . Marginally, any two vertices are adjacent only in  $G_n^{(1)}$  with probability  $q_n(1 - \rho_n)$ , only in  $G_n^{(2)}$  with the same probability, and both in  $G_n^{(1)}$  and  $G_n^{(2)}$  with probability  $q_n\rho_n$ . It is tempting to apply these “adjacency probabilities” to  $u$  and  $v$ , but these adjacency probabilities might be *incorrect* if  $u$  is in (a) or (b)! For instance, if  $u$  in (a) has depth  $d - 1$  in  $G_n^{(1)}$  but  $d - 2$  in  $G_n^{(2)}$ , then the fact that  $v$  is unexplored already implies that  $u$  and  $v$  are

not adjacent in  $G_n^{(2)}$ . This in turn implies that, due to correlation,  $u$  and  $v$  are unlikely to be adjacent in  $G_n^{(1)}$  either. See Figure 7 for a visualization of this issue.

To handle this issue, we observe that those satisfying these adjacency probabilities are in the following two subsets of (a) and (b), along with (c):

- (A)  $u$  has depth  $d - 1$  in  $G_n^{(1)}$  and is *unexplored* in  $G_n^{(2)}$ ;
- (B)  $u$  has depth  $d - 1$  in  $G_n^{(2)}$  and is *unexplored* in  $G_n^{(1)}$ .

A key claim is that most of the vertices in (a) and (b) are in fact (A) and (B), i.e., it is very unlikely that a vertex explored in  $G_n^{(1)}$  at step  $d - 1$  happens to have already been explored in  $G_n^{(2)}$  earlier. Since those “bad” vertices not conforming to the adjacency probabilities always negatively impact  $N_d^I$ , this claim allows us to stochastically squeeze  $N_d^I$  between two random variables which are easier to establish concentration bounds. This argument is rigorously given in Proposition 24.

There are two contributions to  $N_d^I$  coming from the previous step. One is *contribution by correlation*, stemming from common edges of  $G_n^{(1)}$  and  $G_n^{(2)}$  connected to  $N_{d-1}^I$ , where each common edge occurs with probability  $\approx q_n \rho_n$ . The other is *contribution by randomness*, which appears when  $v$  happens to be adjacent to depth  $d - 1$  vertices in  $G_n^{(1)}$  and in  $G_n^{(2)}$  through separate edges; such pair of edges occurs with probability  $\approx q_n^2$ . When  $\rho_n$  is large enough and  $N_{d-1}^I$  constitute a non-negligible portion of  $N_{d-1}^{(1)}$  and  $N_{d-1}^{(2)}$ , the contribution by correlation is non-negligible and yields  $N_d^I \approx (n q_n \rho_n)^d$  as long as  $d < d_n^*$  and  $\rho_n^d = \Theta(1)$ . However, if  $\rho_n^d = o(1)$ , then the contribution by correlation diminishes, resulting in *correlation decay*. Hence, the asymptotics of  $\gamma_n = \rho_n^{d_n^*}$  defined in (10) plays an important role in describing what happens at step  $d = d_n^*$ . Indeed, Theorem 6 states that

$$\frac{N_{d_n^*}^I}{n} \xrightarrow{p} 1 - 2\lambda + \lambda^{2-\gamma}.$$

When  $\gamma = 1$ , the RHS is  $1 - \lambda$  which equals the single graph asymptotics in Theorem 5, demonstrating the dominance of the contribution by correlation. On the other hand, when  $\gamma = 0$ , the RHS is precisely  $(1 - \lambda)^2$ , showing that the correlation decays completely and the contribution by randomness dominates.

## A.2. Overlap of shortest path trees for correlated graphs

**From distance asymptotics to shortest path trees.** Once we know the distance  $d_G(1, v)$  for all  $v \in V$ , then every shortest path tree can be constructed by picking for each  $v \in V$  its *parent*  $u$  in the neighborhood of  $v$  satisfying  $d_G(1, u) + 1 = d_G(1, v)$ . We denote the set of parent candidates of  $v$  by

$$\text{par}_G(v) := \{u \in V : u \text{ and } v \text{ are adjacent in } G \text{ and } d_G(1, u) + 1 = d_G(1, v)\}.$$

Then the set of shortest path trees can be identified with the Cartesian product

$$\prod_{v \in V \setminus \{1\}} \text{par}_G(v). \quad (19)$$

As a consequence, the uniform measure over the shortest path trees is simply a vertex-wise product measure. When we have two graphs  $G^{(1)}$  and  $G^{(2)}$  sharing the same vertex set, this means that the

optimal coupling (under the normalized Hamming metric) of the uniform measures is the product of vertex-wise couplings, so the overlap is given by

$$\tilde{R}_n = \frac{1}{n} \sum_{v \in V \setminus \{1\}} \frac{|\text{par}_{G^{(1)}}(v) \cap \text{par}_{G^{(2)}}(v)|}{\max(|\text{par}_{G^{(1)}}(v)|, |\text{par}_{G^{(2)}}(v)|)}.$$

By our distance asymptotics (Theorem 5 and Theorem 6), we only need to include in the sum the vertices at distance  $d_n^*$  or  $d_n^* + 1$  in both graphs. These vertices fall under one of the four subsets depending on the precise pair of distances (e.g., distance  $d_n^*$  in both graphs,  $d_n^*$  in  $G^{(1)}$  and  $d_n^* + 1$  in  $G^{(2)}$ , and so on), and we know the asymptotic proportion of these subsets from our distance asymptotics. Then for each subset we apply concentration inequalities to control the sum. The details can be found in Section D.3.1.

**From overlap concentration to the absence of OGP.** A result of this analysis, Theorem 2, implies that the overlap concentrates for any limit values of the proxies  $\lambda_n \rightarrow \lambda$ ,  $\rho_n \rightarrow \rho$ , and  $\gamma_n \rightarrow \gamma$ . Moreover, the probability limit of the overlap is a continuous function of  $\rho$  or  $\gamma$  onto  $[0, 1]$  for any fixed  $\lambda$ . This means that for any sequence of  $\alpha_n$  and any open interval  $\mathcal{I} \subseteq [0, 1]$ , by choosing a suitable sequence of  $\rho_n$ ,  $\mathcal{I}$  contains the overlap infinitely often with probability 1.

### A.3. Gibbs measures over spanning trees

**Computation of the log partition function.** To analyze the Gibbs measures, we need to obtain a very precise understanding of the log partition function/free energy of the model. A basic principle of statistical physics is that the Gibbs measure is always obtained by maximizing entropy while keeping the average energy fixed. This can be made precise via the Gibbs variational principle (cf. Bogolyubov inequality Ellis (2007)), which tells us that

$$\log Z_{G,\beta} = \sup_{\nu} \left[ H_{\nu}(T) - \beta \log \log(n) \mathbb{E}_{\nu} \left[ \sum_v d_T(1, v) \right] \right] \quad (20)$$

where the supremum ranges over all probability measures; in this formulation, we see that the inverse temperature  $\beta \geq 0$  exactly functions as a Lagrange multiplier. We simplify our system to a mean-field model that is thermodynamically equivalent in the sense that its log-partition function approximates (20), effectively converting it to a one-dimensional optimization. We will shortly demonstrate this by sketching a procedure of solving (20) up to  $o_p(n \log \log n)$  error, which is leading-order in  $n$  and accurate enough for Theorem 7. However, it turns out to really understand the system and prove all the other results, we actually need to build a conceptually similar but more delicate analysis in Section E, where we derive a more precise formula accurate up to  $o_p(n)$  error.

**A variational upper bound.** It is discussed in the previous section that the set of shortest path trees is the Cartesian product of  $\text{par}_G(v)$ . In general, any spanning tree  $T$  can be identified with a map  $\text{par}_T : V \setminus \{1\} \rightarrow V$  where  $u = \text{par}_T(v)$  is in the neighborhood  $N_G(v)$  of  $v$  satisfying  $d_T(1, u) + 1 = d_T(1, v)$ . In other words, a spanning tree can be thought of as an element of the Cartesian product

$$\text{par}_T \in \prod_{v \in V \setminus \{1\}} N_G(v)$$

and the Gibbs measure  $\mu_{G,\beta}$  can also be defined on the corresponding product  $\sigma$ -algebra. It should be noted that, unlike shortest path trees, not every element of the Cartesian product is a valid spanning tree — the parent map  $\text{par}_T$  must be *acyclic*, in the sense that no sequence of vertices  $v_1, \dots, v_k$  satisfies  $v_{i+1} = \text{par}_T(v_i)$  and  $v_1 = \text{par}_T(v_k)$ . This means we cannot apply a vanilla mean-field approximation since arbitrary product measures might place nontrivial mass outside the support of  $\mu_{G,\beta}$ . Nonetheless, this perspective gives an important upper bound to (20) which is useful in understanding the Gibbs measure:

$$\begin{aligned} \log Z_{G,\beta} &= \sup_{\nu} \left[ H_{\nu}(T) - \beta \log \log(n) \mathbb{E}_{\nu} \left[ \sum_v d_T(1, v) \right] \right] \\ &\leq \sup_{\nu} \left[ \sum_v H_{\nu}(\text{par}_T(v)) - \beta \log \log(n) \mathbb{E}_{\nu} \left[ \sum_v d_T(1, v) \right] \right] \\ &= \sup_{\nu} \left[ \sum_v (H_{\nu}(\text{par}_T(v)) - \beta \log \log(n) \mathbb{E}_{\nu}[d_T(1, v)]) \right]. \end{aligned}$$

Making the ansatz that this upper bound is nearly tight, we can start to understand the behavior of the model by optimizing each summand in the RHS. For each  $v \in V \setminus \{1\}$ , it wants to maximize the *entropy*  $H_{\nu}(\text{par}_T(v))$  but at the same time minimize the *energy*  $d_T(1, v)$ , and this tradeoff is balanced by the inverse temperature  $\beta$ . If  $v$  tries to be at the minimum energy state, namely  $d_T(1, v) = d_G(1, v)$ , then the parent choice  $\text{par}_T(v)$  is confined to the set  $\text{par}_G(v)$  (of course, this alone does not guarantee that  $v$  will have minimum energy), which gives the entropy of at most  $\log |\text{par}_G(v)|$ . If this is smaller by at least  $\beta \log \log n$  than the maximum possible entropy  $\log |\text{N}_G(v)|$ , then  $v$  might jump to a higher energy state  $d_T(1, v) \geq d_G(1, v) + 1$  with a non-negligible probability seeking for a larger entropy.

**Remark 9 (Transformed system)** *In our rigorous analysis, we show there is a simple way to rewrite the partition function  $Z_{G,\beta}$  as the partition function of a Markov random field or “spin system” related to the original system. This is explained in Section E.2 and helpful for formalizing many of the ideas we discuss here.*

**Reduction to a one-dimensional system.** Note that since we are working with sparse random graphs, a typical vertex can contribute at most  $H_{\nu}(\text{par}_T(v)) \leq \log |\text{N}_G(v)| = (1 + o(1)) \log \log n$  to the sum, which allows us to ignore  $o(n)$  summands. Here, our results on distance asymptotics come into play: since most of the vertices will be at distance either  $d_n^*$  or  $d_n^* + 1$  with high probability, we only need to consider those vertices in  $V^* = \{v : d_G(1, v) \in \{d_n^*, d_n^* + 1\}\}$ <sup>5</sup>. This gives

$$\log Z_{G,\beta} \leq \sup_{\nu} \left[ \sum_{v \in V^*} (H_{\nu}(\text{par}_T(v)) - \beta \log \log(n) \mathbb{E}_{\nu}[d_T(1, v)]) \right] + o_p(n \log \log n). \quad (21)$$

Recall from the preceding discussion that each vertex decides whether to attempt to be at the minimum energy state by picking its parent from  $\text{par}_G(v)$ , or jump to higher energy state and enjoy a

5. If we want an accuracy of at most  $o(n)$  error, then  $O(\frac{n}{\log \log n})$  number of vertices start to matter. This is the reason why in (7) we defined  $d_n^*$  so that  $N_{d_n^*-1} \approx (nq_n)^{d_n^*-1} < \frac{n}{(\log \log n)^2}$ : this assures that it is still safe to ignore vertices at depth at most  $d_n^* - 1$ . Those at depth at least  $d_n^* + 2$  can also be ignored but needs more care to handle; see Lemma 54.

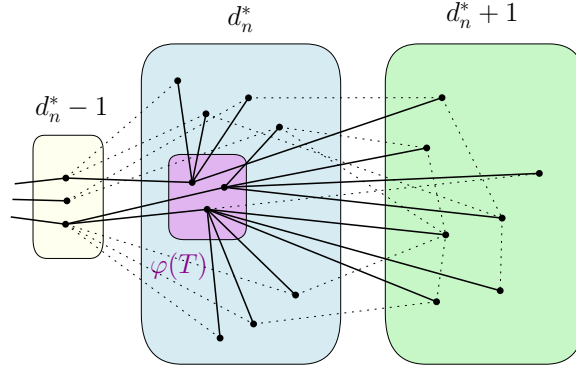


Figure 8: An illustration of the kernel  $\varphi(T)$  of a spanning tree  $T$ . Each line (either dotted or normal) is an edge of the underlying graph  $G_n$ , and the normal lines are the edges of  $T$ . The labels  $d_n^* - 1$ ,  $d_n^*$ , and  $d_n^* + 1$  denote the distances from the root in  $G_n$ , not  $T$ .

possibly larger entropy. In the spirit of mean-field approximation, we make the following mean-field assumptions and observe how the system behaves.

1. Every vertex  $v$  with depth less than  $d_n^*$  always stays at the minimum energy state  $d_T(1, v) = d_G(1, v)$ , in order to “support” the vertices in  $V^*$ .
2. Each of the vertices at depth  $d_n^*$ , (approximately) independently with a common probability  $p$ , decides to pick its parent from  $\text{par}_G(v)$  and gains an energy of  $d_n^*$ . Let  $A$  be the set of vertices that made this decision, i.e.,  $\text{par}_T(v) \in \text{par}_G(v)$ .
3. Every other vertex  $v \in V^* \setminus A$ , hoping that  $A$  supplies enough entropy, always picks its parent from  $A$  and gains an energy of  $d_n^* + 1$ .

On average, a typical vertex at depth  $d_n^*$  would have  $N_{d_n^*-1} q_n$  neighbors in  $\text{par}_G(v)$ ; by our results on distance asymptotics, this is approximately  $(nq_n)^{d_n^*-1} q_n = (nq_n)^{\Delta_n}$ . Hence, each vertex in  $A$  gains a maximum entropy of  $(\Delta_n + o(1)) \log \log n$  on average. Similarly, a typical vertex would have  $q_n |A|$  neighbors in  $A$ , so each vertex in  $V^* \setminus A$  gains a maximum entropy of  $\approx \log(|A| q_n)$ . By our assumptions, the size of  $A$  concentrates around  $N_{d_n^*} p$ , so this is  $\approx \log(N_{d_n^*} p q_n)$ . Thus, the contribution of each vertex at depth  $d_n^*$  to the sum in (21) is approximately

$$H(p) + p \Delta_n \log \log n + (1 - p) \log(N_{d_n^*} p q_n) - \beta \log \log(n) (p d_n^* + (1 - p)(d_n^* + 1))$$

and for each vertex at depth  $d_n^* + 1$

$$\log(N_{d_n^*} p q_n) - \beta \log \log(n) (d_n^* + 1).$$

Collecting terms depending on  $p$ , (21) reduces to optimizing

$$\sup_{0 < p \leq 1} [N_{d_n^*} (H(p) + p(\Delta_n + \beta) \log \log n) + ((1 - p)N_{d_n^*} + N_{d_n^*+1}) \log(N_{d_n^*} p q_n)].$$

Assuming the convergence of the proxy  $\Delta_n \rightarrow \Delta \in [0, 1]$  and observing that  $N_{d_n^*} H(p) = O(n)$  is negligible and  $N_{d_n^*} + N_{d_n^*+1} = |V^*| = (1 + o_p(1))n$ , this can also be stated as optimizing  $m = N_{d_n^*} p$ :

$$\sup_{0 < m \leq N_{d_n^*}} [m(\Delta + \beta) \log \log n + (n - m) \log(m q_n)]. \quad (22)$$

Note that  $m$  represents an approximate size of  $A$ . Thus, we have reduced the problem (20) of optimizing the measure to the problem (22) of optimizing the number of vertices at depth  $d_n^*$  satisfying  $\text{par}_T(v) \in \text{par}_G(v)$ . In Definition 44, we call this set  $A$  a *kernel*, depicted in Figure 8, and is a central object in our rigorous derivation throughout Section E. Once we have completely analyzed the variational upper bound as above, we prove its tightness by constructing measures on trees which approximately satisfy the “mean-field assumptions” above.

**Phase transitions and the Franz–Parisi potential.** We now present an overview of interesting consequences of our analysis of partition function, whose formal proofs are detailed in Section F. By solving (22), one can see that if  $\beta > 1 - \Delta$  (low temperature phase) then  $m = N_{d_n^*}$  is an approximate optimizer. This can be interpreted as the following: in the low temperature phase, almost all vertices at depth  $d_n^*$  decides to be in the minimum energy state  $d_T(1, v) = d_G(1, v) = d_n^*$ . This indicates that a typical tree  $T$  under the Gibbs measure might actually be similar to a shortest path tree, which turns out to be true in Wasserstein sense (cf. Theorem 8). A crucial mathematical tool in proving this is an entropy-transport inequality: we round  $\mu_{G,\beta}$  to the support (19) of the uniform measure by a negligible change in Wasserstein distance, then argue that this rounded measure has near maximum entropy; an entropy-transport inequality then guarantees that this should be close to the uniform measure in Wasserstein distance<sup>6</sup>. In addition, in the low temperature phase, the landscape of the Franz–Parisi potential largely depends on the size of  $\text{par}_G(v)$  for each  $v$  at depth  $d_n^*$ . If  $\Delta > 0$ , then a typical size of  $\text{par}_G(v)$  is unbounded; this means that two independent samples from  $\mu_{G,\beta}$  will have negligible overlap, leading to a strictly increasing landscape of the Franz–Parisi potential (e.g., the left plot of Figure 4). Otherwise, if a typical size of  $\text{par}_G(v)$  is bounded, then two independent samples might have nontrivial overlap with high probability (e.g., the blue curve in Figure 4).

Now we consider the case  $\beta < 1 - \Delta$  (high temperature phase) where  $m = \frac{n}{(1-\Delta-\beta)\log\log n}$  is an approximate optimizer, given that this is smaller than  $N_{d_n^*}$ . In this case, some non-negligible portion of the vertices at depth  $d_n^*$  decide to jump to a higher energy state  $d_T(1, v) = d_G(1, v) + 1 = d_n^* + 1$ . As a result, those staying in the minimum energy state tend to have more children in a typical tree under the Gibbs measure  $\mu_{G,\beta}$ , compared to a uniformly random shortest path tree. Thus, the “children profile” of the vertices at depth  $d_n^*$  can be used to distinguish  $\mu_{G,\beta}$  from the uniform measure; if we look at the empirical distribution of the number of children for vertices at depth  $d_n^*$ , trees sampled from the former would look like a mixture of two distributions, whereas those sampled from the latter would just concentrate around  $N_{d_n^*+1}/N_{d_n^*}$ . This children profile is in fact Lipschitz under the Hamming metric and thus plays a role as a *witness* to show that  $\mu_{G,\beta}$  is not close in Wasserstein metric to the uniform measure. Moreover, the fact that the optimal  $m$  is only  $o(n)$  implies that all but  $o(n)$  vertices enjoy high entropy, i.e., have an unbounded number of parent choices. Thus, two independent samples from  $\mu_{G,\beta}$  are unlikely to have a non-negligible overlap, leading to a strictly increasing landscape of the Franz–Parisi potential in any regime.

## Appendix B. Preliminaries

### B.1. Additional notations

In Section 1.2.1 we introduced our models and notations which are sufficient to present our results. Here, we complement that with additional graph theoretic notations that will be handy throughout

6. As is evident from Corollary 65, this requires that we know the entropy of  $\mu_{G,\beta}$  accurate up to an  $o(n)$  error. See Section F.1.4 for a rigorous argument.

the proofs and intermediate results in the rest of the paper. Besides these, throughout the rest of the paper, especially for the proofs, we will usually omit the subscripts for  $q = q_n$ ,  $G = G_n$ ,  $d^* = d_n^*$ , and  $\ell^* = \ell_n^*$ . On the other hand, we never drop the subscripts for proxy sequences such as  $\lambda_n$ ,  $\Delta_n$ , and  $\rho_n$ , since they can be confused with their limiting values.

**Notations for general graphs.** Here we introduce several notations for common objects in graph theory. For  $u, v \in V(G)$ , we write  $u \sim v$  to mean that  $u$  and  $v$  are adjacent in  $G$ . We denote the neighborhood of  $v$  by

$$N_G(v) := \{u \in V(G) : u \sim v\}.$$

More generally, for any  $A, B \subseteq V(G)$ , we denote the neighborhood of  $A$  in  $B$  by

$$N_B(A) := \bigcup_{v \in A} N_G(v) \cap B \setminus A.$$

We define the *distance function*  $d_G(u, v)$  for  $u, v \in V(G)$  as the length of a shortest path from  $u$  to  $v$  in  $G$ . When there is no such path, the distance is set to be  $\infty$  by convention. More generally, for a set  $A \subseteq V(G)$ ,  $d_G(u, A)$  is the length of a shortest path from  $u$  to any vertex in  $A$ . The function  $d_G(u, -)$  is the *distance vector* of  $G$  starting from  $u$ . Define the distance  $d$  neighborhood of  $v$  by

$$\Gamma_d^G(v) := \{u \in V(G) : d_G(u, v) = d\}$$

and its size

$$N_d^G(v) := |\Gamma_d^G(v)|.$$

In addition, we sometimes use

$$\Gamma_{\leq d}^G(v) := \bigcup_{k=1}^d \Gamma_k^G(v)$$

to denote the set of vertices with distance at most  $d$ . We may sometimes drop the subscript/superscript  $G$  when it is obvious from the context.

**Notations for graphs with fixed source.** As a quick reminder, we mostly consider paths from a fixed starting vertex 1. Thus, we may simply write  $d_G = d_G(1, -)$ ,  $\Gamma_d^G = \Gamma_d^G(1)$ , and  $N_d^G = N_d^G(1)$ . As we study paths from the fixed source 1, we naturally focus on the connected component of  $G$  containing 1. We denote the set of vertices in this component by

$$\overline{V}(G) := \{v \in V(G) : d_G(1, v) < \infty\}$$

and

$$\overline{G} = G[\overline{V}]$$

the induced subgraph of  $G$  by  $\overline{V}$ . On top of the distance structure induced by 1, we define the *parent set* of  $v \in \overline{V} \setminus \{1\}$  by

$$\text{par}_G(v) := \{u \in N_G(v) : d_G(1, u) = d_G(1, v) - 1\}.$$

For any spanning tree  $T$  of  $\overline{G}$ , the parent set of  $v$  in  $T$  is always a singleton and by  $\text{par}_T(v)$  we refer to this unique parent. In particular,  $\text{par}_T(v) \in \text{par}_G(v)$ . As an inverse function of  $\text{par}_G$ , we define the *children set* of  $v \in \overline{V}$  by

$$\text{ch}_G(v) := \{u \in N_G(v) : v \in \text{par}_G(u)\}.$$

## B.2. Wasserstein distance

**Definition 10** (see, e.g., [Villani \(2021\)](#)) *The 1-Wasserstein distance between two probability measures  $\mu, \nu$  over a metric space  $(d, \mathcal{X})$  is given by*

$$W_1(\mu, \nu) = \inf_{\Pi} \mathbb{E}_{(X,Y) \sim \Pi} [d(X, Y)]$$

where  $\Pi$  ranges over all coupling of  $\mu$  and  $\nu$ .

**Proposition 11** *Suppose that  $\mu = \otimes_{i=1}^m \mu_i$  and  $\nu = \otimes_{i=1}^m \nu_i$  are two product measures on  $\mathcal{X}^n$ , let  $d$  be a metric on  $\mathcal{X}$ , and let  $c(x, y) = \sum_{i=1}^n d(x_i, y_i)$  be a metric on  $\mathcal{X}^n$ . Then if we consider the 1-Wasserstein distance induced by these metrics, we have that*

$$W_1(\mu, \nu) = \sum_{i=1}^n W_1(\mu_i, \nu_i),$$

and also the infimum in the definition of  $W_1(\mu, \nu)$  can be obtained by a product measure.

**Proof** We have that

$$W_1(\mu, \nu) = \inf_{\Pi} \mathbb{E}_{(X,Y) \sim \Pi} [c(X, Y)] = \inf_{\Pi} \sum_{i=1}^n \mathbb{E}_{(X,Y) \sim \Pi} [d(X_i, Y_i)] = \sum_{i=1}^n \inf_{\Pi_i} \mathbb{E}_{(X_i, Y_i) \sim \Pi_i} [d(X_i, Y_i)],$$

where we used that  $\sum_{i=1}^n \mathbb{E}_{(X,Y) \sim \Pi} [d(X_i, Y_i)] \leq \sum_{i=1}^n \inf_{\Pi_i} \mathbb{E}_{(X_i, Y_i) \sim \Pi_i} [d(X_i, Y_i)]$  holds for all  $\Pi$ . Also, equality is achieved if we consider a product coupling of the optimal  $\Pi_i$ . ■

## B.3. Shortest path structures

**The shortest path DAG.** The *shortest path DAG* from  $v$  is the union of all shortest paths starting from  $v$ . This is also called the *shortest path subgraph* (see, e.g., [Khuller et al. \(1995\)](#); [Bauer and Wagner \(2009\)](#)); because our original graph is undirected, we use the term “shortest path DAG” to emphasize that the shortest path structure is naturally directed and acyclic.

**Uniformly random shortest path tree.** A shortest path tree from vertex  $v$  is any tree which is formed as the union of shortest paths from  $v$  to each other vertex. Interestingly, given the shortest path DAG, we can easily count and sample a uniformly random element from the set of *shortest path trees* of the graph. All we need to do is, independently for each vertex  $u$  in the shortest path DAG, pick a uniformly incoming edge to  $u$  and drop all of the other incoming edges. In other words, the uniform distribution over shortest path trees naturally factorizes as a *product measure* and the shortest path DAG encodes the information about this product measure. This is closely connected with what happens in the finite-temperature setting, which we study later in [Section E](#).

**Proposition 12** *Fix a graph  $G$ , let  $\mu_{G,v}$  be the uniform distribution over shortest path trees from vertex  $v$  in  $G$ , and let  $F$  be the shortest path DAG of  $G$  also from  $v$ . Then  $\mu_v$  factorizes as a product measure where each vertex  $u \neq v$  picks its parent uniformly at random from its parents in  $F$ .*

The following propositions shows that the process of sampling a spanning tree, given as input the shortest path DAG, is *stable in the worst-case* (i.e. Lipschitz). Here,  $\text{par}_T(v)$  and  $\text{par}_F(v)$  denote the set of parents of  $v$  in the tree  $T$  (in which case is a singleton for non-root vertices) and the DAG  $F$ , respectively.

**Proposition 13** *Given two graphs  $G_1$  and  $G_2$  on the same vertex set  $V$  of size  $n$  and a choice of vertex  $v \in V$ , let  $\mu_1, \mu_2$  be the corresponding uniform measures over shortest path trees from  $v$ . Let  $\Pi$  be the product measure coupling of  $\mu_1$  and  $\mu_2$  achieving  $W_1(\mu_1, \mu_2)$  under the directed<sup>7</sup> Hamming metric (from Proposition 11) and let  $(T_1, T_2) \sim \Pi$  be random shortest path trees. Then with probability at least  $1 - \delta$ ,*

$$\left| \#\{e : e \in T_1 \cap T_2\} - \sum_{u \neq v} \Pr(\text{par}_{T_1}(u) = \text{par}_{T_2}(u)) \right| = O(\sqrt{n \log(2/\delta)}).$$

**Proof** This follows from Hoeffding's inequality. ■

**Proposition 14** *With the notation of the previous proposition, and letting  $F_1$  and  $F_2$  be the corresponding shortest path DAGs starting from vertex  $v$ , we have*

$$\Pr(\text{par}_{T_1}(u) = \text{par}_{T_2}(u)) = \frac{\#\{e : \text{par}_{F_1}(u) \cap \text{par}_{F_2}(u)\}}{\max(\#\text{par}_{F_1}(u), \#\text{par}_{F_2}(u))}.$$

**Proof** This is the probability they are equal in the optimal coupling (Wilmer et al., 2009b, Proposition 4.7). ■

#### B.4. Gibbs measures

**Finite-temperature Gibbs measure.** For the shortest path tree problem, there is a natural way to define a class of finite-temperature *Gibbs measures*. Given a graph  $G = (V, E)$ , its connected component  $\bar{G}$  containing the root vertex 1, and for inverse temperature  $\beta > 0$ , we define a class of Gibbs measures over spanning trees  $T$  of  $\bar{G}$ <sup>8</sup>

$$\mu_{G,\beta}(T) = \frac{1}{Z_{G,\beta}} \exp \left( -\beta \log \log(n) \sum_{v \in \bar{V}} d_T(1, v) \right) \quad (23)$$

where  $Z_{G,\beta}$  is the *partition function* (normalizing constant), and  $d_T$  is the graph metric in the tree  $T$ . The connection to the minimum-cost flow formulation (175) is that for a shortest path tree  $T$ ,  $\sum_v d_T(1, v)$  is exactly the objective value for (175) of the unique unidirectional flow on  $T$  sending 1 unit of flow from the root 1 to each of the other vertices  $v$ . (The entries of the corresponding flow vector  $f$  on an edge  $u \rightarrow v$  correspond to the number of shortest paths which contain this edge.) We

7. The edges of the tree inherit directions from the DAG, and are only considered equal if they point in the same direction. Ignoring direction would make negligible difference in our results since opposite-direction overlaps are not very common in our setting.

8. In our study of sparse Erdős–Rényi graphs,  $\bar{G}$  is typically the giant component of  $G$ .

also show in Appendix J.3 that with this definition of the Gibbs measure, one can naturally recover Dijkstra's algorithm as a special case of belief propagation.

As is customary in statistical physics, as the inverse temperature  $\beta$  increases, the Gibbs measure starts to concentrate more and more on near-minimizers of

$$H(T) = \log \log(n) \sum_v d_T(1, v).$$

In other words, as  $\beta$  increases the Gibbs measure reweighs itself to prefer trees where the sum of the distances of vertices to the root becomes smaller. The scaling factor  $\log \log(n)$  is included so that the regime where  $\beta = \Theta(1)$  is where the model exhibits interesting changes to the Gibbs measure, as will soon become clear from the results and analysis. Note that if we take  $\beta = 0$ , this measure is simply the uniform measure over spanning trees, which itself has been extensively studied over general graphs (see, e.g., Kirchhoff (1958); Anari et al. (2021)).

## Appendix C. Asymptotics for a single graph

In this section, we study the asymptotics of  $G = G_n$ , centered on the set of fixed distance neighborhoods  $\Gamma_d = \Gamma_d^{G_n}(1)$ . These results, especially those developed in Section C.1, are fundamental and will be frequently referred back throughout the paper.

### C.1. The distribution of distances in a single graph

We start with the asymptotic behavior of the distance vector. As mentioned in the introduction, the distance between  $d^* = d_n^*$  and  $\ell^* = \ell_n^*$  will be critically important in classifying the behavior of the random graph. Because of this, even if we consider an arbitrary convergent sequence of relative densities  $\alpha_n \rightarrow \alpha$ , this is not enough to determine the behavior of the graph for large values of  $n$ . (This will be clear from the conclusion of the following theorem.) The key information which is needed to determine the behavior is actually the relative location of  $d^*$  and  $\ell^*$ , which is encoded by the parameter  $\lambda_n$  in the definition below.

**Theorem 15** *Suppose that  $\lambda_n \rightarrow \lambda \in [0, 1]$ . Then we have*

$$N_{d_n^*}/n \xrightarrow{p} 1 - \lambda, \quad \text{and} \quad N_{d_n^*+1}/n \xrightarrow{p} \lambda$$

as  $n \rightarrow \infty$ . In other words, with probability  $1 - o(1)$  all vertices except for at most  $o(n)$  of them have distance either  $d_n^*$  or  $d_n^* + 1$ , and the proportion of vertices with distance  $d^*$  converges to  $1 - \lambda$  in probability.

Recall that  $d^*$  is defined to be the smallest integer such that  $(nq)^{d^*} \geq n/(\log \log n)^2$ . This implies

$$\frac{1}{(\log \log n)^2} \leq (nq)^{d^* - \ell^*} = \log(1/\lambda_n) < \frac{nq}{(\log \log n)^2} = \frac{\alpha_n \log n}{(\log \log n)^2} \quad (24)$$

which gives

$$-\frac{2 \log \log \log n}{\log \alpha_n + \log \log n} \leq d^* - \ell^* = \Delta_n < 1 - \frac{2 \log \log \log n}{\log \alpha_n + \log \log n} < 1 \quad (25)$$

for large enough  $n$ .

The idea of our proof of Theorem 15 is to obtain tight concentration bounds for  $N_d$ . Since the distribution of  $N_d$  depends on  $N_0, \dots, N_{d-1}$ , we need an inductive argument. The following result is much stronger than we actually need here, but this form turns out to be useful in proving Theorem 26 in Section D.2.

**Proposition 16** *Let  $\frac{C}{\log \log n} < \delta < \frac{1}{2}$  for some constant  $C > 0$ . Then with probability  $1 - o(1)$ , we have*

$$(1 - \delta)(nq)^d \leq N_d \leq (1 + \delta)(nq)^d$$

for all  $0 \leq d < d^*$ . In particular, we have

$$\frac{N_d}{(nq)^d} \xrightarrow{p} 1$$

for all  $0 \leq d < d^*$ .

**Proof** Let  $L = (1 - \delta)^{1/3}$  and  $R = (1 + \delta)^{1/3}$ . Note that for large enough  $n$ ,

$$L^{1/k} = (1 - \delta)^{1/(3k)} \leq \left(1 - \frac{C}{\log \log n}\right)^{1/(3k)} \leq 1 - \frac{C}{3k \log \log n} \quad (26)$$

and

$$R^{1/k} = (1 + \delta)^{1/(3k)} \geq \left(1 + \frac{C}{\log \log n}\right)^{1/(3k)} \geq 1 + \frac{C}{6k \log \log n} \quad (27)$$

where (26) is from Lemma 115 and (27) is from the fact that  $1 + x \leq e^x \leq 1 + 2x$  for small enough  $x > 0$ . Based on (15), our proof strategy is to apply concentration of measure and the union bound to establish that the collection of events  $E_d$  (parameterized by depth  $d$ ) hold simultaneously with high probability, where

$$E_d = \begin{cases} L^{1/2}nq \leq N_1 \leq R^{1/2}nq & \text{if } d = 1, \\ L(nq)^2 \leq N_2 \leq R(nq)^2 & \text{if } d = 2, \\ L^{1+(d-2)/\ell^*}(nq)^d \leq N_d \leq R^{1+(d-2)/\ell^*}(nq)^d & \text{if } 3 \leq d \leq d^* - 2, \\ L^3(nq)^d \leq N_d \leq R^3(nq)^d & \text{if } d = d^* - 1. \end{cases}$$

The proof consists of three steps; from  $d = 0$  to 2, from  $d = 3$  to  $d^* - 2$ , and  $d = d^* - 1$ . For convenience, we define

$$n_d := n - \sum_{i=0}^{d-1} N_i$$

and

$$q_d := 1 - (1 - q)^{N_{d-1}}.$$

**Step 1.** We prove  $\Pr(E_1) \geq 1 - o(1)$  and  $\Pr(E_2 \mid E_1) \geq 1 - o(1)$ . Since  $N_1 \sim \text{Binom}(n-1, q)$ , Lemma 99 together with (26) and (27) gives that

$$L^{1/2}nq \leq N_1 \leq R^{1/2}nq$$

with probability  $1 - o(1)$ .

Under this event, we have

$$n_2 = n - N_0 - N_1 \geq n - 1 - R^{1/2}nq \geq n - 2\alpha_n \log n \geq L^{1/6}n$$

and

$$qN_1 \leq 2R^{1/2}nq^2 \leq 2(1 - L^{1/6})$$

for large  $n$ , which gives by Lemma 115 and Lemma 116

$$L^{2/3}nq^2 \leq L^{1/6}qN_1 \leq q_2 = 1 - (1 - q)^{N_1} \leq qN_1 \leq R^{1/2}nq^2.$$

By Lemma 99, we obtain

$$\Pr(N_2 \leq R^{1/2}n_2q_2 \leq R(nq)^2 \mid E_1) \geq 1 - o(1)$$

and

$$\Pr(N_2 \geq L^{1/6}n_2q_2 \geq L(nq)^2 \mid E_1) \geq 1 - o(1).$$

**Step 2.** We prove  $\Pr(E_d \mid E_1, \dots, E_{d-1}) \geq 1 - o(\frac{1}{\log n})$  for  $3 \leq d \leq d^* - 2$  by generalizing Step 1. Suppose that under  $E_1, \dots, E_{d-1}$ , we condition on  $N_0, \dots, N_{d-1}$ . We first obtain a lower bound on  $n_d$  and  $q_d$ .

**Claim 1** Under  $E_{d-1}$ , we have  $n_d \geq L^{1/(3\ell^*)}n$  and  $q_d \geq L^{1/(3\ell^*)}qN_{d-1}$ .

To prove the claim, we need an upper bound of  $L^{1/(3\ell^*)}$  to make it easier to work with. By (26) we get

$$L^{1/(3\ell^*)} \leq 1 - \frac{C}{9\ell^* \log \log n} = 1 - \frac{C \log(nq)}{9 \log n \log \log n} \leq 1 - \frac{C}{18 \log n} \quad (28)$$

for large enough  $n$ , since  $nq = \Omega(\log n)$ .

Now we bound  $n_d$ . Under  $E_{d-1}$ , we have that

$$\sum_{k=0}^{d-1} N_k \leq 1 + Rnq + \sum_{k=2}^{d-1} R(R^{1/d^*}nq)^k \leq R^2 \left( \sum_{k=0}^{d-1} (nq)^k \right) \leq 2R^2(nq)^{d-1}.$$

From (25) we have  $d \leq d^* - 2 < \ell^* - 1$ , so we can further bound this as

$$2R^2(nq)^{d-1} \leq 2R^2(nq)^{\ell^*-2} = \frac{2R^2}{(nq)^2} \cdot n.$$

We want to show  $\frac{2R^2}{(nq)^2} \leq 1 - L^{1/(3\ell^*)}$ , and by (28) it is enough to show

$$\frac{2R^2}{(nq)^2} \leq \frac{C}{18 \log n}$$

for sufficiently large  $n$ , which easily follows from  $nq = \Omega(\log n)$ .

To lower bound  $q_d$ , we use (24) to observe that

$$qN_{d-1} \leq qR^2(nq)^{d-1} \leq qR^2(nq)^{d^*-3} = R^2(nq)^{d^*-\ell^*-2} \leq \frac{R^2}{nq(\log \log n)^2}.$$

Similar to the preceding paragraph, we have for sufficiently large  $n$  that

$$\frac{R^2}{nq(\log \log n)^2} \leq \frac{C}{9 \log n} \leq 2(1 - L^{1/(3\ell^*)}).$$

Hence, by Lemma 115 and Lemma 116, we obtain

$$q_d = 1 - (1 - q)^{N_{d-1}} \geq 1 - e^{-qN_{d-1}} \geq L^{1/(3\ell^*)} qN_{d-1}$$

which proves the claim.

Combining the two lower bounds, we have under  $E_{d-1}$  that

$$\mathbb{E}[N_d \mid N_0, \dots, N_{d-1}] = n_d q_d \geq L^{2/(3\ell^*)} nqN_{d-1} \geq L^{2/(3\ell^*)} \cdot L^{1+(d-3)/\ell^*} (nq)^d. \quad (29)$$

Now we are ready to achieve bounds for the probability  $\Pr(E_d \mid E_1, \dots, E_{d-1})$ . Since  $n_d \leq n$  and  $q_d \leq qN_{d-1}$  by Lemma 115, Lemma 99 gives

$$\Pr(N_d \leq R^{1/\ell^*} nqN_{d-1} \leq R^{1+(d-2)/\ell^*} (nq)^d \mid E_1, \dots, E_{d-1}) \geq 1 - \exp\left(-\frac{(R^{1/\ell^*} - 1)^2 n_d q_d}{3}\right).$$

Note that by (27)

$$R^{1/\ell^*} - 1 \geq \frac{C}{6\ell^* \log \log n} \geq \frac{C}{12 \log n}$$

and by (29)

$$n_d q_d \geq L^{2/(3\ell^*)} \cdot L^{1+(d-3)/\ell^*} (nq)^d \geq L^2 (nq)^3 = \Omega((\log n)^3).$$

Thus, we arrive at

$$\Pr(N_d \leq R^{1+(d-2)/\ell^*} (nq)^d \mid E_1, \dots, E_{d-1}) \geq 1 - e^{-\Omega(\log n)} \geq 1 - o\left(\frac{1}{\log n}\right).$$

Similarly, from (29),

$$\Pr(N_d \geq L^{1/(3\ell^*)} n_d q_d \geq L^{1+(d-2)/\ell^*} (nq)^d \mid E_1, \dots, E_{d-1}) \geq 1 - \exp\left(-\frac{(1 - L^{1/(3\ell^*)})^2 n_d q_d}{2}\right).$$

Using (28), we get a similar bound

$$\Pr(N_d \geq L^{1+(d-2)/\ell^*} (nq)^d \mid E_1, \dots, E_{d-1}) \geq 1 - e^{-\Omega(\log n)} \geq 1 - o\left(\frac{1}{\log n}\right).$$

**Step 3.** For  $d = d^* - 1$  we prove  $\Pr(E_d \mid E_1, \dots, E_{d-1}) \geq 1 - o(1)$ , and to this end we need  $n_d > L^{1/3}n$  and  $q_d > L^{1/3}qN_{d-1}$ . The proof is almost identical, but since it is important to see what could possibly break down if  $d = d^*$ , we include a proof.

Bounding  $n_d$  is nearly identical;  $\sum_{k=0}^{d-1} N_k \leq 2R^2(nq)^{d-1}$  still holds. By (25) we have

$$2R^2(nq)^{d-1} \leq 2R^2(nq)^{\ell^*-1} = \frac{2R^2}{nq} \cdot n.$$

By (26) it follows that  $\frac{2R^2}{nq} \leq 1 - L^{1/3}$  for large  $n$ . This gives  $n_d \geq L^{1/3}n$ .

For  $q_d$ , note that by (24)

$$qN_{d-1} \leq qR^2(nq)^{d^*-2} = R^2(nq)^{d^*-\ell^*-1} \leq \frac{R^2}{(\log \log n)^2}.$$

We have  $\frac{R^2}{(\log \log n)^2} \leq 2(1 - L^{1/3})$ , which implies

$$q_d \geq L^{1/3}qN_{d-1}.$$

Now by the same argument with Step 2 it is easy to see that

$$\Pr(E_d \mid E_1, \dots, E_{d-1}) \geq 1 - o(1)$$

for  $d = d^* - 1$ .

Combining the three steps gives the desired result. ■

We are now ready to prove Theorem 15. In fact, we prove a slightly stronger result which turns out to be useful in Section E.

**Theorem 17** *For any constant  $C > 0$ , we have*

$$\left| \frac{N_{d^*}}{(1 - \lambda_n)n} - 1 \right| < \frac{C}{\log \log n}$$

and

$$\left| \frac{N_{d^*+1}}{n} - \lambda_n \right| < \frac{C}{\log \log n}$$

with probability at least  $1 - o(1)$ .

**Proof** For  $0 < \delta_1 < 1$  and  $0 < \delta_2 < 1$  that may depend on  $n$ , consider an event

$$E = \{|\Gamma_{\leq d^*-1}| \leq \delta_1 n\} \cap \left\{ (1 - \delta_2)(nq)^{d^*-1} \leq N_{d^*-1} \leq (1 + \delta_2)(nq)^{d^*-1} \right\}.$$

Now we condition on  $N_{d^*-1}$ ,  $|\Gamma_{\leq d^*-1}|$ , and the event  $E$ . Recall that

$$N_{d^*} \sim \text{Binom}(n_{d^*}, q_{d^*})$$

where  $n_{d^*} = n - |\Gamma_{\leq d^*-1}|$  and  $q_{d^*} = 1 - (1 - q)^{N_{d^*-1}}$ . First, note that by Lemma 115

$$q_{d^*} \geq 1 - e^{-qN_{d^*-1}} \geq 1 - e^{-(1-\delta_2)q(nq)^{d^*-1}} = 1 - e^{-(1-\delta_2)(nq)^{d^*-\ell^*}} = 1 - \lambda_n^{1-\delta_2}.$$

By Lemma 115 we have

$$q_{d^*} \geq 1 - \lambda_n^{1-\delta_2} = 1 - (1 - (1 - \lambda_n))^{1-\delta_2} \geq 1 - (1 - (1 - \delta_2)(1 - \lambda_n)) = (1 - \lambda_n)(1 - \delta_2). \quad (30)$$

For an upper bound on  $q_{d^*}$ , we use the fact  $1 - q \geq e^{-q}(1 - q^2)$  and again apply Lemma 115 to get

$$q_{d^*} \leq 1 - e^{-qN_{d^*-1}}(1 - q^2)^{N_{d^*-1}} \leq 1 - e^{-qN_{d^*-1}}(1 - q^2N_{d^*-1}) \leq 1 - \lambda_n^{1+\delta_2} + nq^2.$$

Similar to the lower bound, Lemma 115 gives

$$q_{d^*} \leq 1 - \lambda_n^{1+\delta_2} + nq^2 = 1 - (1 - (1 - \lambda_n))^{1+\delta_2} + nq^2 \leq (1 + \delta_2)(1 - \lambda_n) + nq^2$$

Thus, we have

$$n_{d^*}q_{d^*} \geq n(1 - \lambda_n)(1 - \delta_1)(1 - \delta_2) \geq n(1 - \lambda_n)(1 - (\delta_1 + \delta_2))$$

and

$$n_{d^*}q_{d^*} \leq n(1 - \lambda_n)(1 + \delta_2) + n^2q^2 = n(1 - \lambda_n) \left( 1 + \delta_2 + \frac{nq^2}{1 - \lambda_n} \right).$$

Note that for large enough  $n$ ,

$$1 - \lambda_n \geq 1 - e^{-1/(\log \log n)^2} \geq \frac{1}{2(\log \log n)^2}$$

by, e.g., Lemma 116. Thus,

$$\frac{nq^2}{1 - \lambda_n} \leq \frac{\alpha_n^2(\log n)^2(\log \log n)^2}{n}.$$

We may now invoke Proposition 16 with  $\delta_1 = C_1/\log \log n$  and  $\delta_2 = C_2/\log \log n$  which gives that for any constant  $C_3 > 0$

$$|n_{d^*}q_{d^*} - (1 - \lambda_n)n| < \frac{C_3(1 - \lambda_n)n}{\log \log n}$$

with probability  $1 - o(1)$ . Using Lemma 99 we can argue that the concentration of  $\text{Binom}(n_{d^*}, q_{d^*})$  is much stronger than the RHS, leading to

$$\left| \frac{N_{d^*}}{(1 - \lambda_n)n} - 1 \right| < \frac{C}{\log \log n}$$

with probability  $1 - o(1)$ .

To prove the second part of the theorem, we note that with high probability

$$N_{d^*} \geq \frac{1 - \lambda_n}{2}n \geq \frac{1}{2}(1 - e^{-1/(\log \log n)^2})n \geq \frac{n}{4(\log \log n)^2}.$$

Conditioned on  $\Gamma_{\leq d^*}$ , the probability that a vertex in  $V \setminus \Gamma_{\leq d^*}$  is not connected to any of  $\Gamma_{d^*}$  is at most

$$(1 - q)^{N_{d^*}} \leq e^{-qN_{d^*}} \leq e^{-\frac{nq}{4(\log \log n)^2}} \quad (31)$$

which is far less than  $o(1/\log \log n)$ . Thus, the second part follows from the first part.  $\blacksquare$

We note that the asymptotics for  $N_{d^*}$  is somewhat different from  $N_d$  for  $d \leq d^* - 1$ . Nevertheless, if we only consider a subsequence where  $\lambda_n \rightarrow \lambda = 1$ , then  $N_{d^*}$  follows a similar asymptotics, though slightly weaker than Proposition 16. We state this as the following lemma, which turns out to be useful in a number of proofs.

**Lemma 18** *Consider a subsequence of  $n \in \mathbb{Z}^+$  such that  $\lambda_n \rightarrow \lambda = 1$ . Then we have*

$$\frac{N_{d^*}}{(nq)^{d^*}} \xrightarrow{p} 1.$$

**Proof** We have

$$\frac{N_{d^*}}{(nq)^{d^*}} = \frac{N_{d^*}}{n \log(1/\lambda_n)} = \frac{N_{d^*}}{n(1 - \lambda_n)} \cdot \frac{1 - \lambda_n}{\log(1/\lambda_n)}.$$

Since  $\frac{1 - \lambda_n}{\log(1/\lambda_n)} \rightarrow 1$  if  $\lambda_n \rightarrow 1$ , the conclusion follows from Theorem 17.  $\blacksquare$

## C.2. Shortest paths between two fixed vertices

We develop some basic properties of sparse random graphs, which are more or less standard. These are useful in Section D.4 where we prove that the overlap of uniformly random shortest paths does not concentrate.

**Proposition 19** *Let  $d \leq \frac{d^* - 1}{2}$ . Then with probability  $1 - o(1)$ , the shortest path DAG is locally a tree up to distance  $d$ . In other words, with probability  $1 - o(1)$ , there is a unique shortest path from 1 to any vertex with distance at most  $d$ .*

**Proof** We prove that for all  $1 \leq k \leq d$  every vertex in  $\Gamma_k$  is connected to at most one vertex in  $\Gamma_{k-1}$ , simultaneously with high probability. Conditioned on  $\Gamma_1, \dots, \Gamma_{k-1}$ , the number of edges between  $v \in \Gamma_{\geq k}$  and  $\Gamma_{k-1}$  follows  $\text{Binom}(N_{k-1}, q)$ , so the probability that this number is greater than 1 is

$$1 - (1 - q)^{N_{k-1}} - qN_{k-1}(1 - q)^{N_{k-1}-1}.$$

Using Lemma 116, this can be bounded by

$$\begin{aligned} 1 - (1 - q)^{N_{k-1}} - qN_{k-1}(1 - q)^{N_{k-1}-1} &\leq qN_{k-1} - qN_{k-1}(1 - q)^{N_{k-1}-1} \\ &\leq q^2 N_{k-1}^2. \end{aligned}$$

Hence, the probability that none of the vertices in  $\Gamma_{\geq k}$  have more than one connection to  $\Gamma_{k-1}$  is at least

$$(1 - q^2 N_{k-1}^2)^{|\Gamma_{\geq k}|} \geq (1 - q^2 N_{k-1}^2)^n \geq 1 - nq^2 N_{k-1}^2.$$

Thus, the probability that the shortest path DAG is locally a tree up to distance  $d$  is at least

$$\prod_{k=1}^d (1 - nq^2 N_{k-1}^2) \geq 1 - nq^2 \sum_{k=1}^d N_{k-1}^2.$$

By Proposition 16, with overwhelming probability this is further bounded by

$$1 - nq^2 \sum_{k=1}^d N_{k-1}^2 \geq 1 - nq^2 \sum_{k=1}^d 2(nq)^{2(k-1)} \geq 1 - 4nq^2(nq)^{2(d-1)} = 1 - 4q(nq)^{2d-1}.$$

By (24) we have

$$q(nq)^{2d-1} \leq q(nq)^{d^*-2} = (nq)^{d^*-\ell^*-1} \leq \frac{1}{(\log \log n)^2} = o(1)$$

which proves the result. ■

**Corollary 20** *With probability  $1 - o(1)$ , there is no cycle of length smaller than  $d^*$  that contains a fixed vertex 1.*

**Proof** First, suppose that  $d^*$  is odd. Then by Proposition 19 the shortest path DAG is locally a tree up to distance  $d = \frac{d^*-1}{2}$ . In addition to this, we prove that the whole graph  $G$  is locally a tree up to distance  $d - 1$  with high probability, which would follow by showing that

$$\sum_{k=1}^{d-1} \binom{N_k}{2} q \xrightarrow{p} 0.$$

Indeed, by (24) and Proposition 16, we get

$$\sum_{k=1}^{d-1} \binom{N_k}{2} q \leq \binom{2N_{d-1}}{2} q \leq 2qN_{d-1}^2 \leq 8q(nq)^{2d-2} \leq 8(nq)^{d^*-\ell^*-2} \leq \frac{8}{nq(\log \log n)^2}$$

with probability  $1 - o(1)$ , as desired. Under these events, the shortest possible cycle containing 1 has length at least  $2d + 1 = d^*$ .

A similar argument works when  $d^*$  is even. In this case, we set  $d = \frac{d^*}{2} - 1$  and deduce that  $G$  is locally a tree up to distance  $d$ , so under the good event the shortest possible cycle has length at least  $2(d + 1) = d^*$ . The details are omitted. ■

Proposition 19 gives a nearly complete picture of the union of the shortest paths from fixed vertices 1 to 2. Note that by Proposition 16 the distance between 1 and 2 is either  $d^*$  or  $d^* + 1$  with overwhelming probability. When  $d^*$  is odd,  $d = \frac{d^*-1}{2}$  is an integer, so if we run Dijkstra's algorithm (or simply BFS) then the shortest path DAGs from 1 and from 2 will be disjoint trees up to distance  $d$  with high probability, so what happens between  $\Gamma_d(1)$  and  $\Gamma_d(2)$  will determine the distance between 1 and 2 and the set of shortest paths between them. If there exists an edge between them, the distance will be  $d^*$  and otherwise  $d^* + 1$ .

Now the most interesting case is when  $0 < \lambda < 1$ , i.e. the distance can be either  $d^*$  or  $d^* + 1$ , both with non-negligible probability. Then the number of edges between  $\Gamma_d(1)$  and  $\Gamma_d(2)$  will follow the distribution

$$\text{Binom}(N_d(1) \cdot N_d(2), q). \tag{32}$$

By Proposition 16, note that

$$\frac{N_d(1) \cdot N_d(2) \cdot q}{(nq)^{2d}q} \xrightarrow{p} 1.$$

Since  $(nq)^{2d}q = (nq)^{d^*-1}q = \log(1/\lambda_n) \rightarrow \log(1/\lambda)$ , it follows that the limiting distribution of (32) is in fact  $\text{Poisson}(\log(1/\lambda))$ . Hence, the probability that there exists a unique edge between  $\Gamma_d(1)$  and  $\Gamma_d(2)$  converges in probability to  $\lambda \log(1/\lambda)$ . Since the shortest path DAGs up to distance  $d$  were trees, this actually implies there is a unique path between 1 and 2.

We summarize our discussion to the following theorem.

**Theorem 21** *Suppose  $\{\alpha_n\}_{n \in \mathbb{Z}^+}$  is chosen so that  $d^*$  is odd and  $\lambda_n \rightarrow \lambda \in (0, 1)$ . Then the probability that there is a unique shortest path of length  $d^*$  from 1 to 2 converges to  $\lambda \log(1/\lambda)$ .*

## Appendix D. Asymptotics for correlated Erdős–Rényi graphs

In this section, we analyze the setting of *correlated* random graphs relevant to OGP theory. The main result in this section is the computation of the overlap of random shortest path trees in Section D.3. Before we get there, it is helpful to first analyze the behavior of the correlated *distance vectors* and *shortest path DAGs* which we do first.

**Additional notation.** As usual, we drop the subscripts  $G^{(1)} = G_n^{(1)}$ ,  $G^{(2)} = G_n^{(2)}$ . We write  $\Gamma_d^{(1)} = \Gamma_d^{G^{(1)}}$  and  $\Gamma_d^{(2)} = \Gamma_d^{G^{(2)}}$  to denote the set of vertices with distance  $d$  within the graphs  $G^{(1)}$  and  $G^{(2)}$ , respectively. Similarly, we define  $N_d^{(1)}$  and  $N_d^{(2)}$ . Also, we denote by  $\Gamma_d^I$  the intersection  $\Gamma_d^{(1)} \cap \Gamma_d^{(2)}$ , and  $N_d^I = |\Gamma_d^I|$  the cardinality of the intersection.

### D.1. The overlap of the distance vectors

Recall that as in Theorem 15, the shortest distance distribution heavily depends on the limit of the difference between  $d_n^*$  and  $\ell_n^*$ . For cases 1 and 3, i.e.  $\lambda_n = o(1)$  or  $\lambda_n = 1 - o(1)$ , then the overlap is quite trivial, because there is one fixed distance  $d_n^*$  or  $d_n^* + 1$  which dominates almost the entire graph. Hence, the most interesting case to study the overlap is case 2, where  $\lambda_n \rightarrow \lambda \in (0, 1)$ .

As a reminder, we restate the definition (10) of  $\gamma_n$ :

$$\gamma_n := \rho_n^{d_n^*} \in (0, 1).$$

**Theorem 22** *Suppose that  $\lambda_n \rightarrow \lambda$  and  $\gamma_n \rightarrow \gamma$  with  $\lambda, \gamma \in [0, 1]$ . Then as  $n \rightarrow \infty$ , we have that*

$$\frac{N_{d_n^*}^I}{n} \xrightarrow{p} 1 - 2\lambda + \lambda^{2-\gamma},$$

*i.e. the proportion of the vertices that have distance  $d^*$  for both  $G_n^{(1)}$  and  $G_n^{(2)}$  converges to  $1 - 2\lambda + \lambda^{2-\gamma}$  in probability.*

It is helpful to consider the two extreme cases. If  $\gamma = 1$ , then  $N_{d^*}^I/n \rightarrow 1 - \lambda$  which is the same with the single graph case;  $G^{(1)}$  and  $G^{(2)}$  are nearly identical. On the other hand,  $\gamma = 0$ , we get  $N_{d^*}^I/n \rightarrow (1 - \lambda)^2$  which can be interpreted as  $(N_{d^*}^{(1)}/n) \cdot (N_{d^*}^{(2)}/n)$ ; due to correlation decay, the events of a vertex contained in  $\Gamma_{d^*}^{(1)}$  and in  $\Gamma_{d^*}^{(2)}$  are almost independent.

**Corollary 23** *Under the same setting with Theorem 22, the proportion of the vertices that have the same distance for  $G_n^{(1)}$  and  $G_n^{(2)}$  converges to  $1 - 2\lambda + 2\lambda^{2-\gamma}$  in probability.*

**Proof** The remaining vertices not contained in  $\bigcup_{d=0}^{d^*} \Gamma_d^{(1)}$  and  $\bigcup_{d=0}^{d^*} \Gamma_d^{(2)}$  have proportion  $\lambda^{2-\gamma}$  in probability limit and both  $\Gamma_{d^*+1}^{(1)}$  and  $\Gamma_{d^*+1}^{(2)}$  occupy them by Theorem 15.  $\blacksquare$

The proof of Theorem 22 is similar to the proof of Theorem 15, except that we need to obtain concentration bounds for  $N_d^I$  as well. Thus, we will prove a result similar to Proposition 16. Recall the  $\gamma_n$  defined in (10).

**Proposition 24** *Let  $\frac{C}{\log \log n} < \delta < \frac{1}{2}$  for some constant  $C$ . Suppose that  $\gamma_n \rightarrow \gamma > 0$ . Then with probability  $1 - o(1)$ , we have*

$$(1 - \delta)(nq\rho_n)^d \leq N_d^I \leq (1 + \delta)(nq\rho_n)^d$$

for all  $0 \leq d < d^*$ . Moreover, if  $\gamma = 0$ , then

$$\frac{N_{d^*-1}^I}{(nq)^{d^*-1}} \xrightarrow{p} 0.$$

**Proof** The cardinality of the intersection  $N_d^I = |\Gamma_d^I| = |\Gamma_d^{(1)} \cap \Gamma_d^{(2)}|$  unfortunately does not have a nice exact hierarchical structure like (15). Nevertheless, we will obtain bounds using a similar strategy.

Suppose that we are already given (conditioned on)  $\Gamma_i^{(1)}$  and  $\Gamma_i^{(2)}$  for all  $0 \leq i \leq d-1$ , and we are at the step of constructing  $\Gamma_d^{(1)}$  and  $\Gamma_d^{(2)}$ . We define the set of “fresh” vertices  $V_d := V \setminus \Gamma_{\leq d-1}^{(1)} \setminus \Gamma_{\leq d-1}^{(2)}$  and its cardinality  $n_d := |V_d|$ . Each vertex  $v \in V_d$  is either:

- (a) connected to both  $\Gamma_{d-1}^{(1)}$  and  $\Gamma_{d-1}^{(2)}$ ,
- (b) connected to  $\Gamma_{d-1}^{(1)}$  only or  $\Gamma_{d-1}^{(2)}$  only, or
- (c) connected to neither  $\Gamma_{d-1}^{(1)}$  nor  $\Gamma_{d-1}^{(2)}$ .

Then for (a)  $v$  will be included in  $\Gamma_d^I$  and for (c)  $v$  will be included in  $V_{d+1}$ . Let  $\sim_i$  and  $\approx_i$  denote the connectivity and disconnectivity in graph  $G^{(i)}$ , and when there is no subscript we mean for both. We are interested in the probability of (a), which is calculated as

$$\begin{aligned} q_d &:= \Pr(v \sim_1 \Gamma_{d-1}^{(1)} \wedge v \sim_2 \Gamma_{d-1}^{(2)}) \\ &= 1 - \Pr(v \approx_1 \Gamma_{d-1}^{(1)}) - \Pr(v \approx_2 \Gamma_{d-1}^{(2)}) + \Pr(v \approx_1 \Gamma_{d-1}^{(1)} \wedge v \approx_2 \Gamma_{d-1}^{(2)}). \end{aligned}$$

Here,  $\Pr(v \approx_1 \Gamma_{d-1}^{(1)} \wedge v \approx_2 \Gamma_{d-1}^{(2)})$  equals

$$\Pr(v \approx_1 \Gamma_{d-1}^{(1)} \wedge v \approx_2 \Gamma_{d-1}^{(2)}) = \Pr(v \approx_1 \Gamma_{d-1}^{(1)} \setminus \Gamma_{d-1}^I) \cdot \Pr(v \approx_2 \Gamma_{d-1}^{(2)} \setminus \Gamma_{d-1}^I) \cdot \Pr(v \approx \Gamma_{d-1}^I). \quad (33)$$

The probability  $\Pr(v \approx \Gamma_{d-1}^I)$  is easy to compute. The probability that a pair of vertices is not connected by an edge in both  $G^{(1)}$  and  $G^{(2)}$  is  $(1-q)(\rho_n + (1-\rho_n)(1-q)) = (1-q)(1-q(1-\rho_n))$ , so

$$\Pr(v \approx \Gamma_{d-1}^I) = (1-q)^{N_{d-1}^I} (1-q(1-\rho_n))^{N_{d-1}^I}.$$

However, the calculation of  $\Pr(v \approx_1 \Gamma_{d-1}^{(1)} \setminus \Gamma_{d-1}^I)$  is more complicated: the edge probability in  $G^{(1)}$  between  $v$  and a vertex in  $\Gamma_{d-1}^{(1)} \setminus \Gamma_{d-1}^I$  is not exactly  $q$ , because some of the vertices in  $\Gamma_{d-1}^{(1)} \setminus \Gamma_{d-1}^I$  might be contained in  $\Gamma_{\leq d-2}^{(2)}$ . For these vertices, it is already conditioned that they are not connected to  $v$  in  $G^{(2)}$ . Fortunately, this means that for these vertices the edge probability in  $G^{(1)}$  is *less* than  $q$ , which allows us to obtain bounds using “truly fresh” pair of vertices.

Put it formally, let  $\tilde{\Gamma}_d^{(1)} := \Gamma_d^{(1)} \setminus \Gamma_{\leq d-1}^{(2)}$  and  $\tilde{\Gamma}_d^{(2)} := \Gamma_d^{(2)} \setminus \Gamma_{\leq d-1}^{(1)}$  (also set  $\tilde{N}_d^{(1)} := |\tilde{\Gamma}_d^{(1)}|$  and  $\tilde{N}_d^{(2)} := |\tilde{\Gamma}_d^{(2)}|$ ). Then when we set all edge probabilities to  $q$ ,  $\Pr(v \sim_1 \Gamma_{d-1}^{(1)} \wedge v \sim_2 \Gamma_{d-1}^{(2)})$  only gets bigger, and when we replace  $\Gamma_{d-1}^{(1)}$  and  $\Gamma_{d-1}^{(2)}$  with  $\tilde{\Gamma}_{d-1}^{(1)}$  and  $\tilde{\Gamma}_{d-1}^{(2)}$ , respectively, the probability becomes smaller. This gives bounds

$$q_d \leq 1 - (1 - q)^{N_{d-1}^{(1)}} - (1 - q)^{N_{d-1}^{(2)}} + (1 - q)^{N_{d-1}^{(1)} + N_{d-1}^{(2)} - N_{d-1}^I} (1 - q(1 - \rho_n))^{N_{d-1}^I} \quad (34)$$

and

$$q_d \geq 1 - (1 - q)^{\tilde{N}_{d-1}^{(1)}} - (1 - q)^{\tilde{N}_{d-1}^{(2)}} + (1 - q)^{\tilde{N}_{d-1}^{(1)} + \tilde{N}_{d-1}^{(2)} - N_{d-1}^I} (1 - q(1 - \rho_n))^{N_{d-1}^I}. \quad (35)$$

The main part of the proof is similar to Proposition 16. First, we consider the case  $\gamma > 0$ . Let  $L = (1 - \delta)^{1/3}$  and  $R = (1 + \delta)^{1/3}$ . We define the collection of events

$$E_d = \begin{cases} L^{1/2} n q \rho_n \leq N_1^I \leq R^{1/2} n q \rho_n & \text{if } d = 1, \\ L(n q \rho_n)^2 \leq N_2^I \leq R(n q \rho_n)^2 & \text{if } d = 2, \\ L^{1+(d-2)/\ell^*} (n q \rho_n)^d \leq N_d^I \leq R^{1+(d-2)/\ell^*} (n q \rho_n)^d & \text{if } 3 \leq d \leq d^* - 2, \\ L^3 (n q \rho_n)^d \leq N_d^I \leq R^3 (n q \rho_n)^d & \text{if } d = d^* - 1. \end{cases}$$

Again, the proof consists of three steps, namely

$$\Pr(E_d \mid E_1, \dots, E_{d-1}) \geq \begin{cases} 1 - o(1) & \text{if } 1 \leq d \leq 2, \\ 1 - o\left(\frac{1}{\log n}\right) & \text{if } 3 \leq d \leq d^* - 2, \\ 1 - o(1) & \text{if } d = d^* - 1. \end{cases}$$

We only prove for  $3 \leq d \leq d^* - 2$  since the other cases are similar (see the proof of Proposition 16). In order to apply Lemma 99, we need to derive bounds for  $n_d$  and  $q_d$ .

We first handle  $n_d$ . Note that

$$n_d \geq n - \sum_{k=0}^{d-1} N_k^{(1)} - \sum_{k=0}^{d-1} N_k^{(2)}.$$

Thus, by applying the same argument as in the proof of Proposition 16, it is easy to see that

$$n_d \geq L^{1/(3\ell^*)} n. \quad (36)$$

For  $q_d$ , we establish a lower bound first. By Proposition 16, we may safely assume  $N_k^{(i)} \leq 2(nq)^k$  for all  $k \leq d^* - 1$ . For some  $\epsilon > 0$  (not necessarily constant) to be chosen later, assume

$$q(nq)^{d-1} \leq \epsilon. \quad (37)$$

Then

$$qN_{d-1}^{(i)} \leq 2q(nq)^{d-1} \leq 2\epsilon.$$

Hence, by Lemma 115 and Lemma 116, we have

$$(1-q)^{\tilde{N}_d^{(i)}} \leq 1 - (1-\epsilon)q\tilde{N}_{d-1}^{(i)}.$$

Also by Lemma 115, we get

$$(1-q)^{\tilde{N}_{d-1}^{(1)} + \tilde{N}_{d-1}^{(2)} - N_{d-1}^I} \geq 1 - q(\tilde{N}_{d-1}^{(1)} + \tilde{N}_{d-1}^{(2)} - N_{d-1}^I)$$

and

$$(1 - q(1 - \rho_n))^{N_d^I} \geq 1 - q(1 - \rho_n)N_d^I.$$

By plugging in these bounds to (35), we have

$$\begin{aligned} q_d &\geq q\rho N_{d-1}^I - \epsilon q(\tilde{N}_{d-1}^{(1)} + \tilde{N}_{d-1}^{(2)}) + q^2(1-\rho)(\tilde{N}_{d-1}^{(1)} + \tilde{N}_{d-1}^{(2)} - N_{d-1}^I)N_{d-1}^I \\ &\geq q\rho N_{d-1}^I - \epsilon q(N_{d-1}^{(1)} + N_{d-1}^{(2)}) \\ &\geq q\rho N_{d-1}^I - 4\epsilon q(nq)^{d-1}. \end{aligned}$$

Under  $E_{d-1}$ , we have  $N_{d-1}^I \geq \frac{1}{2}(nq\rho_n)^{d-1}$ . Also note  $\rho_n^d \geq \rho_n^{d*} \geq \gamma/2$  for large enough  $n$ . This gives

$$q\rho_n N_{d-1}^I \geq \frac{1}{2}q\rho_n(nq\rho_n)^{d-1} \geq \frac{1}{4}\gamma q(nq)^{d-1}. \quad (38)$$

Hence,

$$q_d \geq q\rho_n N_{d-1}^I - 4\epsilon q(nq)^{d-1} \geq \left(1 - \frac{16\epsilon}{\gamma}\right) q\rho_n N_{d-1}^I.$$

Recall from (28) that we have for large enough  $n$

$$L^{1/(3\ell^*)} \leq 1 - \frac{C}{18 \log n}.$$

Thus, we set (note that the assumption  $\gamma > 0$  is used here)

$$\epsilon = \frac{C\gamma}{288 \log n} = \Theta\left(\frac{1}{\log n}\right)$$

which satisfies (37) since

$$q(nq)^{d-1} \leq (nq)^{d^* - \ell^* - 2} \leq \frac{1}{nq(\log \log n)^2} \leq O\left(\frac{1}{(\log n)(\log \log n)^2}\right).$$

This gives a lower bound

$$q_d \geq L^{1/(3\ell^*)} q\rho_n N_{d-1}^I.$$

Combined with (36) and (38), we have

$$n_d q_d \geq L^{2/(3\ell^*)} nq\rho_n N_{d-1}^I \geq \frac{1}{8}\gamma(nq)^d = \Omega((\log n)^3). \quad (39)$$

Therefore, applying Lemma 99 and (28) gives

$$\begin{aligned} \Pr(N_d \geq L^{1/(3\ell^*)} n_d q_d \geq L^{1+(d-2)/\ell^*} (nq)^d \mid E_1, \dots, E_{d-1}) &\geq 1 - \exp\left(-\frac{(1 - L^{1/(3\ell^*)})^2 n_d q_d}{2}\right) \\ &\geq 1 - e^{-\Omega(\log n)} \\ &\geq 1 - o\left(\frac{1}{\log n}\right). \end{aligned}$$

Deriving an upper bound is similar. We abuse notation and reuse  $\epsilon$  but possibly with a different value. Since

$$q(N_{d-1}^{(1)} + N_{d-1}^{(2)}) \leq 4q(nq)^{d-1} \leq 4\epsilon$$

by Lemma 115 and Lemma 116 we have

$$(1 - q)^{N_{d-1}^{(1)} + N_{d-1}^{(2)} - N_{d-1}^I} \leq 1 - (1 - 2\epsilon)q(N_{d-1}^{(1)} + N_{d-1}^{(2)} - N_{d-1}^I)$$

and

$$(1 - q(1 - \rho))^{N_d^I} \geq 1 - (1 - 2\epsilon)q(1 - \rho)n N_d^I.$$

Also, note that  $(1 - q)^{N_{d-1}^{(i)}} \leq 1 - qN_{d-1}^{(i)}$ . Applying these inequalities to (34), one can see that

$$\begin{aligned} q_d &\leq 1 - (1 - q)^{N_{d-1}^{(1)}} - (1 - q)^{N_{d-1}^{(2)}} + (1 - q)^{N_{d-1}^{(1)} + N_{d-1}^{(2)} - N_{d-1}^I} (1 - q(1 - \rho))^{N_d^I} \\ &\leq q\rho N_{d-1}^I + 2\epsilon q(N_{d-1}^{(1)} + N_{d-1}^{(2)} - N_{d-1}^I) + (1 - 2\epsilon)^2 q^2 (1 - \rho) N_{d-1}^I (N_{d-1}^{(1)} + N_{d-1}^{(2)} - N_{d-1}^I) \\ &\leq q\rho N_{d-1}^I + 2\epsilon q(N_{d-1}^{(1)} + N_{d-1}^{(2)}) + q^2 (N_{d-1}^{(1)} + N_{d-1}^{(2)})^2 \\ &\leq q\rho N_{d-1}^I + 6\epsilon q(N_{d-1}^{(1)} + N_{d-1}^{(2)}) \\ &\leq q\rho N_{d-1}^I + 24\epsilon q(nq)^{d-1} \\ &\leq \left(1 + \frac{96\epsilon}{\gamma}\right) q\rho N_{d-1}^I. \end{aligned}$$

Note that by (27) we have

$$R^{1/(2\ell^*)} \geq 1 + \frac{C}{12\ell^* \log \log n} = 1 + \frac{C \log(nq)}{12 \log n \log \log n} \geq 1 + \frac{C}{24 \log n}. \quad (40)$$

for large  $n$ . Thus, this time we set

$$\epsilon = \frac{C\gamma}{2304 \log n} = \Theta\left(\frac{1}{\log n}\right).$$

which satisfies (37) as well. This gives an upper bound

$$q_d \leq R^{1/(2\ell^*)} q\rho n N_{d-1}^I$$

and thus by (39) and (40)

$$\begin{aligned} \Pr(N_d \leq R^{1/(2\ell^*)} n_d q_d \leq R^{1+(d-2)/\ell^*} (nq)^d \mid E_1, \dots, E_{d-1}) &\geq 1 - \exp\left(-\frac{(R^{1/(2\ell^*)} - 1)^2 n_d q_d}{2}\right) \\ &\geq 1 - e^{-\Omega(\log n)} \\ &\geq 1 - o\left(\frac{1}{\log n}\right) \end{aligned}$$

which concludes the proof.

The proof for the case  $\gamma = 0$  is mostly similar. Note that the above proof step works as long as  $\rho^d$  is bounded away from zero. Intuitively, since  $\rho^{d^*} \rightarrow 0$ ,  $\rho^d$  will eventually be  $o(1)$  as  $d$  increases. After this step, what the inequality

$$q_d \leq q\rho_n N_{d-1}^I + 24\epsilon q(nq)^{d-1}$$

tells us is that  $\rho_n N_{d-1}^I = q(nq)^{d-1} \cdot \rho_n^d = o(q(nq)^{d-1})$ , so  $n_d q_d \leq o((nq)^d)$  as desired. More formally,  $\gamma = 0$  implies that for any  $\delta' > 0$  we can find  $d < d^*$  such that  $\rho_n^d < \delta'$ . This implies that

$$q_d \leq q\rho_n N_{d-1}^I + 24\epsilon q(nq)^{d-1} \leq 2\delta' q(nq)^{d-1} + 24\epsilon q(nq)^{d-1} = (2\delta' + 24\epsilon)q(nq)^{d-1}.$$

Thus, it easily follows that  $N_d^I/(nq)^d \xrightarrow{d} 0$ . It is now straightforward to prove that this continues to hold for larger  $d$ . ■

The following lemma complements Proposition 24, similar to Lemma 18.

**Lemma 25** *Suppose that  $\lambda_n \rightarrow \lambda = 1$  and  $\gamma_n \rightarrow \gamma$ . Then*

$$\frac{N_{d^*}^I}{(nq)^{d^*}} \xrightarrow{p} \gamma.$$

**Proof** The proof is nearly identical to Proposition 24, thus omitted. ■

**Proof** [Proof of Theorem 22] Using Propositions 16 and 24, we prove that each of the bounds from (34) and (35) converges in probability to the desired bound. The result is trivial if  $\lambda = 0$  or  $\lambda = 1$  by Theorem 15, so we assume  $0 < \lambda < 1$ . Note that for any  $1 \leq d \leq d^*$  and a constant  $\epsilon > 0$  we have

$$\tilde{N}_d^{(1)} = |\Gamma_d^{(1)} \setminus \Gamma_{\leq d-1}^{(2)}| \geq |\Gamma_d^{(1)}| - |\Gamma_{\leq d-1}^{(2)}| \geq (1 - \epsilon)(nq)^d - 2(nq)^{d-1} \geq (1 - 2\epsilon)(nq)^d \quad (41)$$

with high probability. Hence, as in the proof of Theorem 15, the terms  $(1 - q)^{N_{d^*-1}^{(i)}}$  and  $(1 - q)^{\tilde{N}_{d^*-1}^{(i)}}$  all converge in probability to  $\lambda$ .

First, consider the case  $\gamma > 0$ . Then  $\rho_n = 1 - o(1)$  so  $q(1 - \rho_n)N_{d^*-1}^I \xrightarrow{p} 0$ , which implies

$$(1 - q(1 - \rho_n))^{N_{d^*-1}^I} \xrightarrow{p} 0.$$

Also, since  $\rho_n^{d^*} \rightarrow \gamma$ , Proposition 24 gives  $qN_{d^*-1}^I \xrightarrow{p} \gamma \log(1/\lambda)$ , so that

$$(1 - q)^{N_{d^*-1}^{(1)} + N_{d^*-1}^{(2)} - N_{d^*-1}^I} \xrightarrow{p} \lambda^{2-\gamma}.$$

Due to (41) the same convergence holds when  $\tilde{N}_{d^*-1}^{(i)}$  are used in place of  $N_{d^*-1}^{(i)}$ . Combining these facts gives the desired result.

If  $\gamma = 0$ , then by the second statement of Proposition 24 we have  $qN_{d^*-1}^I \rightarrow 0$ , so the same result holds. ■

## D.2. The overlap of the shortest path DAGs

We define the *shortest path DAG* of a graph  $G$ , denoted by  $\text{SPD}_G(v)$ , as the union of the shortest paths starting from  $v \in V(G)$  where the orientation agrees with the direction of the outgoing paths from  $v$ . We also study a natural measure of overlap for shortest path DAGs which we now formally present. This result does not have direct implications for shortest path trees, but shortest path DAGs themselves are natural objects to consider from the perspective of, e.g., linear programming duality — see the discussion in Preliminaries. When  $\lambda = 0$  and  $\rho = 1$ , the overlap of shortest path DAGs exhibits a subtle behavior not seen for shortest path trees.

Since 1 is the source, we simply write  $\text{SPD}_G = \text{SPD}_G(1)$ , and the overlap is defined by

$$S_n := \frac{|\text{SPD}_{G_n^{(1)}} \cap \text{SPD}_{G_n^{(2)}}|}{\sqrt{|\text{SPD}_{G_n^{(1)}}| \cdot |\text{SPD}_{G_n^{(2)}}|}}. \quad (42)$$

As with trees, we require that the edges in the overlap have the same orientation for  $G^{(1)}$  and  $G^{(2)}$ .

For the distance vectors discussed in Section D.1, the cases  $\lambda = 0$  and  $\lambda = 1$  were trivial and the interesting regime was  $\lambda \in (0, 1)$ . For the shortest path DAGs, the situation is the opposite: the case  $\lambda \in (0, 1)$  is easy and the case  $\lambda = 0$  will be the trickiest. This is because when  $\lambda \in (0, 1)$ , it is obvious that most of the edges in the shortest path DAG will connect  $\Gamma_{d^*}$  and  $\Gamma_{d^*+1}$ , but when  $\lambda = 0$ , it is not clear which of  $\Gamma_{d^*-1}$  and  $\Gamma_{d^*+1}$  is larger<sup>9</sup>.

We first discuss intuitively what is happening here. By Proposition 16,  $N_{d^*-1} = |\Gamma_{d^*-1}|$  concentrates around  $(nq)^{d^*-1} = \frac{\log(1/\lambda_n)}{q}$ . On the other hand, by Theorem 17,  $N_{d^*+1} = |\Gamma_{d^*+1}|$  concentrates around  $\lambda_n n$ . Thus, we can compare the two sets by comparing  $\frac{\log(1/\lambda_n)}{\lambda_n}$  with  $nq = \alpha_n \log n$ . But since  $\lambda_n$  does not converge to a particular value as discussed in Section C.1, the result of the comparison is of course not consistent. Hence, just as we fixed a subsequence so that we could say  $\lambda_n \rightarrow \lambda$ , we will consider a finer subsequence such that

$$\eta_n := \frac{\log(1/\lambda_n)}{\lambda_n nq} = \frac{\log(1/\lambda_n)}{\alpha_n \lambda_n \log n} \rightarrow \eta \in [0, \infty].$$

For instance,  $\eta = 0$  would mean  $N_{d^*-1} \ll N_{d^*+1}$  so most of the edges in  $\text{SPD}_G$  connect  $\Gamma_{d^*}$  and  $\Gamma_{d^*+1}$ .

To compute the overlap, we need to compute the proportions of  $N_{d^*-1}^I$  and  $N_{d^*+1}^I$  relative to  $N_{d^*-1}^{(i)}$  and  $N_{d^*+1}^{(i)}$ , respectively. The former is well understood through Proposition 24. For the latter, we go back to the proof of Proposition 24 where we were constructing  $\Gamma_d^{(i)}$  and  $\Gamma_d^I$ . At step  $d = d^*$ , one can realize those vertices  $v \in V_{d^*}$  that are not connected to  $\Gamma_{d^*-1}^{(1)}$  or  $\Gamma_{d^*-1}^{(2)}$  comprise  $\Gamma_{d^*+1}^I$ . Thus, our probability of interest will be

$$\Pr(v \approx_1 \Gamma_{d^*-1}^{(1)} \wedge v \approx_2 \Gamma_{d^*-1}^{(2)})$$

which can be calculated through (33). From the proof of Proposition 24, one may expect this to be around  $\lambda^{2-\rho_{d^*}^{d^*}}$ . Since  $|N_{d^*+1}^{(i)}| \approx \lambda n$  the proportion is roughly  $\lambda^{1-\gamma}$ . If  $\gamma = 1$  and  $\lambda = 0$ , this is not well defined. Hence, we need to consider another proxy sequence

$$\xi_n := \lambda_n^{1-\gamma_n} \rightarrow \xi \in [0, 1]$$

9. At first glance, one might expect  $\lambda = 1$  would suffer from the same issue, but fortunately it turns out that we always have  $|\Gamma_{d^*}| \gg |\Gamma_{d^*+2}|$ , so it can be handled in the same way as the case  $\lambda \in (0, 1)$ .

converges. The overlap of the shortest path DAGs will be described in terms of these quantities.

We formalize our discussion to the following theorem.

**Theorem 26** *Suppose that:*

- $\lambda_n \rightarrow \lambda \in [0, 1]$ ,
- $\gamma_n \rightarrow \gamma \in [0, 1]$ ,
- $\frac{\log(1/\lambda_n)}{\alpha_n \lambda_n \log n} \rightarrow \eta \in [0, \infty]$ ,
- $\lambda_n^{1-\gamma_n} \rightarrow \xi \in [0, 1]$ , and
- $\rho_n \rightarrow \rho \in [0, 1]$ .

Then we have

$$S_n \xrightarrow{p} \begin{cases} \gamma & \text{if } \lambda = 1 \text{ or } \eta = \infty, \\ \rho \cdot \frac{(1 - (2 - \xi)\lambda)\xi + \eta\gamma}{1 - \lambda + \eta} & \text{otherwise.} \end{cases}$$

**Proof** We split the proof into several cases, each of which will require different concentration results. The fully rigorous proofs for these concentrations are a bit tedious, so we only detail the proof for the first case and afterwards only present informally. Let  $q_I = q(q(1 - \rho_n) + \rho_n)$  be the edge probability of  $G^{(1)} \cap G^{(2)}$ .

**Case 1:**  $\lambda \in (0, 1)$ . This is the easiest case, which intuitively follows from our earlier results. Combining Lemma 99 and Theorem 15, it is easy to see that

$$\frac{|\text{SPD}_{G^{(i)}}|}{\lambda(1 - \lambda)n^2q} \xrightarrow{p} 1. \quad (43)$$

Now we need to obtain a concentration result for  $\text{SPD}_{G^{(1)}} \cap \text{SPD}_{G^{(2)}}$ . It is intuitively clear that most of the edges are between  $N_{d^*}^I$  and  $N_{d^*+1}^I$ . To prove it, we follow the procedure of the proof of Proposition 24, and after each step  $d$ , obtain a concentration bound on the number of edges between  $N_{d-1}^I$  and  $N_d^I$ . Conditioned on  $\Gamma_1^{(i)}, \dots, \Gamma_{d-1}^{(i)}$ , the number of edges in  $\text{SPD}_{G^{(1)}} \cap \text{SPD}_{G^{(2)}}$  that connects  $N_{d-1}^I$  and  $N_d^I$  follows

$$\text{Binom} \left( N_{d-1}^I \cdot \left| V \setminus \Gamma_{\leq d-1}^{(1)} \setminus \Gamma_{\leq d-1}^{(2)} \right|, q_I \right).$$

We may assume the conclusion of Proposition 24 holds. Then for  $d \leq d^*$ , we have that  $N_{d-1}^I \leq 2(nq)^{d-1}$  and  $|V \setminus \Gamma_{\leq d-1}^{(1)} \setminus \Gamma_{\leq d-1}^{(2)}| \leq n$ . Thus, it easily follows from the Chernoff bound that the number of edges connecting  $N_{d-1}^I$  and  $N_d^I$  is bounded by

$$4(nq)^{d-1} \cdot n \cdot q_I.$$

with probability  $1 - o(1)$ , simultaneously for all  $d \leq d^*$ . Hence, the number of edges connecting  $N_{d-1}^I$  and  $N_d^I$  for all  $d \leq d^*$  is bounded by

$$\sum_{d=1}^{d^*} 4(nq)^{d-1} \cdot n \cdot q_I \leq 8(nq)^{d^*-1} \cdot n \cdot q_I \leq \frac{8 \log(1/\lambda_n) n^2 q_I}{nq}.$$

This is negligible compared to  $|\text{SPD}_{G^{(i)}}|$  as written in (43). For  $d \geq d^* + 2$ , we observe that the number of edges conditioned on the previous steps is stochastically dominated by

$$\text{Binom} \left( N_{d^*+1}^I \left| V \setminus \Gamma_{\leq d^*+1}^{(1)} \setminus \Gamma_{\leq d^*+1}^{(2)} \right| + \binom{|V \setminus \Gamma_{\leq d^*+1}^{(1)} \setminus \Gamma_{\leq d^*+1}^{(2)}|}{2}, q_I \right).$$

Since  $|V \setminus \Gamma_{\leq d^*+1}^{(1)} \setminus \Gamma_{\leq d^*+1}^{(2)}| = o(n)$ , it is clear that this is also negligible compared to  $|\text{SPD}_{G^{(i)}}|$ . Thus, we only need to consider the edges between  $N_{d^*}^I$  and  $N_{d^*+1}^I$ , whose cardinality follows

$$\text{Binom}(N_{d^*}^I | V \setminus \Gamma_{\leq d^*}^{(1)} \setminus \Gamma_{\leq d^*}^{(2)}, q_I)$$

conditioned on  $\Gamma_1^{(i)}, \dots, \Gamma_{d^*}^{(i)}$ . Since we know that  $N_{d^*}^I$  concentrates around

$$(1 - 2\lambda + \lambda^{2-\gamma})n = (1 - (2 - \xi)\lambda)n$$

and  $|V \setminus \Gamma_{\leq d^*}^{(1)} \setminus \Gamma_{\leq d^*}^{(2)}|$  concentrates around

$$\lambda^{2-\gamma}n = \lambda\xi n$$

we see that

$$\frac{|\text{SPD}_{G^{(1)}} \cap \text{SPD}_{G^{(2)}}|}{\lambda(1 - (2 - \xi)\lambda)\xi n^2 q_I} \xrightarrow{p} 1.$$

This completes the proof when  $\lambda \in (0, 1)$ .

**Case 2:**  $\lambda = 1$ . From Lemma 18, we know that  $N_{d^*}^{(i)}$  concentrates around  $(nq)^{d^*} = \log(1/\lambda_n)n$ . Conditioned on  $N_1^{(i)}, \dots, N_{d^*+1}^{(i)}$ , the number of vertices with distance at least  $d^* + 2$  follows

$$\text{Binom} \left( n - \sum_{d=0}^{d^*} N_d, (1 - q)^{N_{d^*}} \right).$$

Note that  $(1 - q)^{N_{d^*}} \leq e^{-qN_{d^*}} \leq e^{-\log(1/\lambda_n)nq/2} = e^{-\alpha_n \log(1/\lambda_n) \log n/2}$  with high probability. From (24) we have  $\log(1/\lambda_n) \geq 1/(\log \log n)^2$ , so for large enough  $n$

$$(1 - q)^{N_{d^*}} \leq e^{-\alpha_n \log(1/\lambda_n) \log n/2} \leq e^{-\log n/(\log \log n)^2} \ll \frac{1}{(\log \log n)^2} \leq \log(1/\lambda_n)$$

with high probability. Hence, it follows that  $N_{d^*}^{(i)} \gg N_{d^*+2}^{(i)}$  in probability limit, which means we only need to consider the edges connecting  $N_{d^*}^{(i)}$  and  $N_{d^*+1}^{(i)}$ . Applying Lemma 25 gives  $N_{d^*}^I/N_{d^*}^{(i)} \xrightarrow{p} \gamma$ , which implies the desired result.

**Case 3:**  $\lambda = 0$  and  $\eta = \infty$ . This is similar to Case 2 since we can prove that  $N_{d^*-1}^{(i)} \gg N_{d^*+1}^{(i)}$ . Indeed,  $N_{d^*-1}^{(i)}$  concentrates around  $(nq)^{d^*-1} = \log(1/\lambda_n)/(nq)$  and  $N_{d^*+1}$  is bounded by

$$n(1 - q)^{N_{d^*-1}} \leq ne^{-qN_{d^*-1}} = \lambda_n n$$

with high probability. Thus,  $\eta = \infty$  implies  $N_{d^*-1}^{(i)} \gg N_{d^*+1}^{(i)}$  and the concentration follows.

**Case 4:**  $\lambda = 0$  and  $\eta < \infty$ . This case is the trickiest, since  $N_{d^*+1}^{(i)}$  is not negligible anymore. Note that  $\eta < \infty$  implies

$$\lambda_n \geq \frac{1}{\log n} \quad (44)$$

for large  $n$ . This case is the trickiest, since  $N_{d^*+1}^{(i)}$  is not negligible anymore. Our goal is obtain concentration bounds for the quantities

$$N_{\geq d^*+1}^{(i)} := \left| V \setminus \Gamma_{\leq d^*}^{(i)} \right|, \quad (45)$$

$$N_{\geq d^*+2}^{(i)} := \left| V \setminus \Gamma_{\leq d^*+1}^{(i)} \right|, \quad (46)$$

$$N_{d^*+1}^I := \left| \Gamma_{d^*+1}^{(1)} \cap \Gamma_{d^*+1}^{(2)} \right|. \quad (47)$$

We note that (the second part of) Theorem 17 is not strong enough to handle our needs here. We must leverage (44) and extend the proof of Theorem 17 to obtain a tighter result.

Let  $\epsilon > 0$  be an arbitrary constant. Conditioned on  $\Gamma_1^{(i)}, \dots, \Gamma_{d^*-1}^{(i)}, N_{\geq d^*+1}^{(i)}$  follows

$$\text{Binom} \left( n - \sum_{d=0}^{d^*-1} N_d^{(i)}, (1-q)^{N_{d^*-1}^{(i)}} \right).$$

We handle  $(1-q)^{N_{d^*-1}^{(i)}}$ , then the concentration for  $N_{\geq d^*+1}^{(i)}$  will readily follow. For the upper bound, we apply Proposition 16 with  $\delta_1 = \frac{\log(1+\epsilon)}{\log \log n}$ , which gives

$$\begin{aligned} (1-q)^{N_{d^*-1}^{(i)}} &\leq e^{-qN_{d^*-1}^{(i)}} \\ &\leq e^{-(1-\delta_1)q(nq)^{d^*-1}} \\ &= \lambda_n \cdot e^{\delta_1 \log(1/\lambda_n)} \\ &\leq \lambda_n \cdot e^{\delta_1 \log \log n} \\ &\leq (1+\epsilon)\lambda_n \end{aligned}$$

with probability  $1 - o(1)$ . To get a lower bound, we use

$$q = \frac{\alpha_n \log n}{n} \leq \frac{2 \log \left( \frac{2}{2-\epsilon} \right)}{\alpha_n \log n + \log \left( \frac{2}{2-\epsilon} \right)}$$

to apply Lemma 115 and Lemma 116, which implies

$$\begin{aligned} (1-q)^{N_{d^*-1}^{(i)}} &\geq \exp \left( - \left( 1 + \frac{\log \left( \frac{2}{2-\epsilon} \right)}{\alpha_n \log n} \right) q N_{d^*-1}^{(i)} \right) \\ &\geq e^{-qN_{d^*-1}^{(i)}} \cdot \exp \left( - \left( \frac{\log \left( \frac{2}{2-\epsilon} \right)}{\alpha_n \log n} \right) nq \right) \\ &\geq \left( 1 - \frac{\epsilon}{2} \right) e^{-qN_{d^*-1}^{(i)}} \end{aligned} \quad (48)$$

Now in the same manner as the derivation of the upper bound, we have

$$e^{-qN_{d^*-1}^{(i)}} \geq \left(1 - \frac{\epsilon}{2}\right) \lambda_n$$

with probability  $1 - o(1)$ . Thus, we obtain

$$(1 - q)^{N_{d^*-1}^{(i)}} \geq (1 - \epsilon) \lambda_n.$$

Therefore, we have proved that

$$\frac{(1 - q)^{N_{d^*-1}^{(i)}}}{\lambda_n} \xrightarrow{p} 1$$

which in turn implies

$$\frac{N_{\geq d^*+1}^{(i)}}{\lambda_n n} \xrightarrow{p} 1.$$

Next, we bound (46). In fact, we only need a crude upper bound, which can be achieved by

$$(1 - q)^{N_{d^*}^{(i)}} \leq e^{-qN_{d^*}^{(i)}} \leq e^{-\frac{1}{2}nq} \leq n^{-\frac{\alpha n}{2}}.$$

Since  $\lambda \geq \frac{1}{\log n}$ , this shows that  $N_{\geq d^*+1}^{(i)} \gg N_{\geq d^*+2}^{(i)}$ , so we can safely neglect  $N_{\geq d^*+2}^{(i)}$ . In particular, we have proved that

$$\frac{N_{d^*+1}^{(i)}}{\lambda_n n} \xrightarrow{p} 1. \quad (49)$$

Finally, we handle (47). As noted in the discussion preceding the statement of Theorem 26, we actually want to prove concentration bounds for (33) when  $d = d^*$ , which we temporarily denote by  $\bar{q}$ . Following the argument in the proof of Proposition 24 that led to the bounds (34) and (35), we have

$$\bar{q} \leq (1 - q)^{\tilde{N}_{d^*-1}^{(1)} + \tilde{N}_{d^*-1}^{(2)} - N_{d^*-1}^I} (1 - q(1 - \rho_n))^{N_{d^*-1}^I} \quad (50)$$

and

$$\bar{q} \geq (1 - q)^{N_{d^*-1}^{(1)} + N_{d^*-1}^{(2)} - N_{d^*-1}^I} (1 - q(1 - \rho_n))^{N_{d^*-1}^I}. \quad (51)$$

Starting from (50),

$$\begin{aligned} \bar{q} &\leq (1 - q)^{\tilde{N}_{d^*-1}^{(1)} + \tilde{N}_{d^*-1}^{(2)} - N_{d^*-1}^I} (1 - q(1 - \rho_n))^{N_{d^*-1}^I} \\ &\leq e^{-q(\tilde{N}_{d^*-1}^{(1)} + \tilde{N}_{d^*-1}^{(2)} - \rho_n N_{d^*-1}^I)}. \end{aligned}$$

Since Proposition 16 yields

$$\tilde{N}_{d^*-1}^{(i)} \geq N_{d^*-1}^{(i)} - N_{\leq d^*-2}^{(i)} \geq N_{d^*-1}^{(i)} - 2(nq)^{d^*-2}$$

with high probability, we obtain

$$\begin{aligned} \bar{q} &\leq e^{-q(\tilde{N}_{d^*-1}^{(1)} + \tilde{N}_{d^*-1}^{(2)} - \rho_n N_{d^*-1}^I)} \\ &\leq e^{-q(N_{d^*-1}^{(1)} + N_{d^*-1}^{(2)} - \rho_n N_{d^*-1}^I)} \cdot e^{4q(nq)^{d^*-2}} \\ &\leq (1 + \epsilon) e^{-q(N_{d^*-1}^{(1)} + N_{d^*-1}^{(2)} - \rho_n N_{d^*-1}^I)} \end{aligned}$$

for large  $n$ . On the other hand, applying (48) to (51) gives

$$\bar{q} \geq (1 - \epsilon)e^{-q(N_{d^*-1}^{(1)} + N_{d^*-1}^{(2)} - \rho_n N_{d^*-1}^I)}.$$

Hence, it suffices to derive concentration bounds for  $e^{-q(N_{d^*-1}^{(1)} + N_{d^*-1}^{(2)} - \rho_n N_{d^*-1}^I)}$ . This can be done similarly as bounding  $(1 - q)^{N_{d^*-1}^{(i)}}$ , by combining together Proposition 24, which we omit the details. Namely, one can show that

$$(1 - 2\epsilon)\lambda_n^{2 - \rho_n^{d^*}} \leq \bar{q} \leq (1 + 2\epsilon)\lambda_n^{2 - \rho_n^{d^*}}.$$

Recall that  $\rho_n^{d^*} = \gamma_n$ . Due to the fact that  $N_{\geq d^*+2}^{(i)}$  are negligible, we arrive at

$$\frac{N_{d^*+1}^I}{\lambda_n^{2 - \gamma_n} \cdot n} \xrightarrow{p} 1. \quad (52)$$

With these results, we can prove concentrations for the shortest path DAGs. Combining (49) and (52) yields  $N_{d^*+1}^I / N_{d^*+1}^{(i)} \xrightarrow{p} \xi$ , the same argument with Case 2 leads to  $N_{d^*}^I / N_{d^*}^{(i)} \xrightarrow{p} \gamma$ , and  $N_{d^*}^{(i)} / N_{d^*+1}^{(i)} \rightarrow \eta$  by the assumption; altogether completes the proof. ■

### D.3. The overlap of random shortest path trees

As noted in Section B.3, the shortest path DAG of a graph naturally encodes the information about the distribution of a uniformly random shortest path tree. The goal of this section is to analyze

$$R_n := \frac{1}{n} |T_n^{(1)} \cap T_n^{(2)}|.$$

As briefly mentioned in the introduction,  $R_n$  is a still random variable after conditioned on the graphs  $G^{(1)}$  and  $G^{(2)}$ , and the distribution depends on how we sample a pair of shortest path trees.

We will focus on the setting of Proposition 13, where we sample a pair of shortest path trees  $(T_n^{(1)}, T_n^{(2)})$  from the optimal coupling with respect to the Hamming metric. Another choice is to sample each tree independently — the analysis is much simpler, and will result in the usual notion of disorder chaos.

#### D.3.1. THE OVERLAP UNDER OPTIMAL COUPLING

By Proposition 13, it suffices to prove that

$$\tilde{R}_n := \mathbb{E}[R_n \mid G_n^{(1)}, G_n^{(2)}] = \frac{1}{n} \sum_{v \in \bar{V}_n \setminus \{1\}} \frac{|\text{par}_{G_n^{(1)}}(v) \cap \text{par}_{G_n^{(2)}}(v)|}{\max(|\text{par}_{G_n^{(1)}}(v)|, |\text{par}_{G_n^{(2)}}(v)|)} \quad (53)$$

concentrates, where  $\bar{V}_n$  is the set of vertices reachable from 1. An equivalent way of formulating (53) is to consider a weighted version of shortest path DAG, where to each incoming edge  $e$  to  $v$  we give weight  $1/\text{indeg}(v)$ , and compute the  $\ell_1$  distance between two weighted shortest path DAGs.

**Theorem 27** Suppose that  $\lambda_n \rightarrow \lambda$  and  $\gamma_n \rightarrow \gamma$  with  $\lambda, \gamma \in [0, 1]$ , and  $\rho_n \rightarrow \rho$ . Then as  $n \rightarrow \infty$ , we have that

$$\tilde{R}_n \xrightarrow{p} \begin{cases} \frac{1 - 2\lambda + \lambda^{2-\gamma}}{1 - \lambda} \cdot \lambda^{2-\gamma} \rho + g((1 - \gamma) \log(1/\lambda), \gamma \log(1/\lambda))(1 - \lambda^\gamma) & \text{if } 0 < \lambda < 1, \\ \gamma & \text{otherwise} \end{cases}$$

where  $g(a, b)$  is defined by

$$g(a, b) = \begin{cases} \mathbb{E} \left[ \frac{Z}{\max(X_1, X_2) + Z} \mid Z > 0 \right] & \text{if } a, b > 0, \\ 1 & \text{if } a = 0, \\ 0 & \text{if } b = 0 \end{cases}$$

where the expectation is over independent  $X_1, X_2 \sim \text{Poisson}(a)$  and  $Z \sim \text{Poisson}(b)$ .

In particular, Theorem 27 implies that the overlap concentrates around a continuous function of  $\rho_n$ .

**Corollary 28 (Overlap concentration)** For a given sequence  $\{\alpha_n\}$ , there is a sequence of automorphisms (i.e., strictly increasing continuous bijections)  $f_n : [0, 1] \rightarrow [0, 1]$  with the following property. For any sequence of correlations  $\{\rho_n\}$ , independently sample for each  $n$  a pair of shortest path trees  $(T_n^{(1)}, T_n^{(2)})$  from the optimal coupling  $\Pi_n$ . Then

$$\left| R(T_n^{(1)}, T_n^{(2)}) - f_n(\rho_n) \right| \xrightarrow{p} 0.$$

**Proof** Note that given  $\{\alpha_n\}$  the sequence  $\{\lambda_n\}$  is fixed. We set

$$f_n(\rho_n) = \frac{1 - 2\lambda_n + \lambda_n^{2-\gamma_n}}{1 - \lambda_n} \cdot \lambda_n^{2-\gamma_n} \rho_n + g((1 - \gamma_n) \log(1/\lambda_n), \gamma_n \log(1/\lambda_n))(1 - \lambda_n^{\gamma_n}). \quad (54)$$

The rest of the proof follows by Theorem 27 and a compactness argument. Namely, suppose for contradiction that we have an  $\epsilon > 0$  and a sequence of indices  $\{n_k\}_{k \in \mathbb{Z}^+}$  such that

$$\Pr \left( \left| R(T_{n_k}^{(1)}, T_{n_k}^{(2)}) - f_{n_k}(\rho_{n_k}) \right| > \epsilon \right) > \delta$$

for some constant  $\delta > 0$ . From that we can find a further subsequence  $\{n_{k_i}\}_{i \in \mathbb{Z}^+}$  such that the corresponding subsequences of  $\lambda_n$ ,  $\rho_n$ , and  $\gamma_n$  all converge, contradicting Theorem 27 and Proposition 13.  $\blacksquare$

Another consequence of Theorem 27 is that the uniform measure over shortest path trees satisfies a type of Wasserstein disorder chaos [El Alaoui et al. \(2022\)](#).

**Corollary 29 (Wasserstein disorder chaos)** Suppose  $\{\rho_n\}$  is a constant sequence  $\rho$ . Then

$$\liminf_{\rho \rightarrow 1^-} \lim_{n \rightarrow \infty} \tilde{R}_n \leq 4/27 < 1.$$

**Proof** We use Theorem 27 and the fact that  $\max_{\lambda \in [0,1]} (1-\lambda)\lambda^2 = 4/27$ . ■

Now we prove Theorem 27. In contrast to Theorem 26 where we had to deal with minority vertices,  $\tilde{R}_n$  is an average over all *vertices*, so by Theorem 15 we can safely disregard all vertices in  $\Gamma_{\leq d^*-1}$  and  $\Gamma_{\geq d^*+1}$ . The proof of Theorem 27 is slightly more sophisticated since we have two graphs, so we first state two crucial observations as a separate lemma.

**Lemma 30** *For each  $d \geq 1$  and  $v \in V$ , we define random variables*

$$Y_{d,v} = \begin{cases} \frac{|\text{par}_{G^{(1)}}(v) \cap \text{par}_{G^{(2)}}(v)|}{\max(|\text{par}_{G^{(1)}}(v)|, |\text{par}_{G^{(2)}}(v)|)} & \text{if } v \in \Gamma_d^I \text{ and } v \sim \Gamma_{d-1}^I, \\ 0 & \text{otherwise} \end{cases}$$

where  $\sim$  means connectivity in both graphs. Then the following hold.

- (a) *Conditioned on  $\Gamma_1^{(i)}, \dots, \Gamma_{d-1}^{(i)}$ , random variables in  $\{Y_{d,v}\}_{v \in V \setminus \Gamma_{\leq d-1}^{(1)} \setminus \Gamma_{\leq d-1}^{(2)}}$  are independent and identically distributed.*
- (b) *Let  $\bar{Y} = \frac{1}{n} \sum_{v \in V} (Y_{d^*,v} + Y_{d^*+1,v})$ . Then for any  $\epsilon > 0$  we have  $\Pr(|\tilde{R}_n - \bar{Y}| > \epsilon) \rightarrow 0$  as  $n \rightarrow \infty$ .*

**Proof** The first part follows by observing that  $Y_{d,v}$  are nonzero if and only if  $v \sim \Gamma_{d-1}^I$ , so  $|\text{par}_{G^{(i)}}(v)|$  are functions of the edge connectivity between  $v$  and  $\Gamma_{d-1}^{(i)}$ .

To prove the second part, we first note that  $\bar{Y} \leq \tilde{R}_n$ ; intuitively,  $(n-1)\bar{Y}$  is the summation as in (53), but only over vertices contained in  $\Gamma_{d^*}^I$  and  $\Gamma_{d^*+1}^I$ . It remains to prove that the other vertices are either contained in a small set, or the overlap of its parents is negligible. Indeed, by Proposition 16, the sets  $\Gamma_d^{(i)}$  where  $d < d^*$  or  $d > d^* + 1$  have cardinality  $o(n)$ , which does not contribute to the sum. Thus, we only need to deal with the case  $\Gamma_{d^*}^{(i)} \setminus \Gamma_{d^*}^I$ , and more specifically  $0 < \lambda < 1$ .

Without loss of generality, we consider  $\Gamma_{d^*}^{(1)} \cap \Gamma_{d^*+1}^{(2)}$ , which occupies most of  $\Gamma_{d^*}^{(1)} \setminus \Gamma_{d^*}^I$  by Theorem 15. We show that most vertices here are connected to many vertices in  $\Gamma_{d^*}^{(2)}$  but few vertices in  $\Gamma_{d^*-1}^{(1)}$ . Conditioned on  $\Gamma_{\leq d^*}^{(2)}$ , each vertex  $v \in V \setminus \Gamma_{\leq d^*}^{(2)}$  satisfies

$$|\{w \in \Gamma_{d^*}^{(2)} : v \sim_2 w\}| \sim \text{Binom}(N_{d^*}^{(2)}, q).$$

Let  $E_v$  be the event such that  $v$  is connected to at most  $\frac{1}{2}(1-\lambda)nq$  vertices in  $\Gamma_{\leq d^*-1}^{(2)}$ . Since  $N_{d^*}^{(2)}$  concentrates around  $(1-\lambda)n$  by Proposition 16, Lemma 99 gives

$$\Pr(E_v \mid \Gamma_{\leq d^*-1}^{(2)}) \leq o(1).$$

Again, by Lemma 99, this implies that at most  $o(n)$  vertices in  $V \setminus \Gamma_{\leq d^*-1}^{(2)}$  are connected to at most  $\frac{1}{2}(1-\lambda)nq$  vertices in  $\Gamma_{\leq d^*}^{(2)}$  with high probability.

Next, conditioned on  $\Gamma_{\leq d^*-1}^{(1)}$ , each vertex  $v \in V \setminus \Gamma_{\leq d^*-1}^{(1)}$  satisfies

$$|\{w \in \Gamma_{d^*-1}^{(1)} : v \sim_2 w\}| \sim \text{Binom}(N_{d^*-1}^{(1)}, q).$$

Since  $qN_{d^*-1}^{(1)}$  concentrates around  $\log(1/\lambda)$  by Proposition 16, we have

$$\text{Binom}(N_{d^*-1}^{(1)}, q) \xrightarrow{d} \text{Poisson}(\log(1/\lambda))$$

so for any  $\epsilon > 0$  there exists  $K$  such that at most  $\epsilon n$  vertices in  $V \setminus \Gamma_{\leq d^*-1}^{(1)}$  are connected to at least  $K$  vertices in  $\Gamma_{\leq d^*-1}^{(1)}$  with probability  $1 - o(1)$ .

Therefore, all but at most  $o(n)$  vertices in  $\Gamma_{d^*}^{(1)} \cap \Gamma_{d^*+1}^{(2)}$  have at least  $\frac{1}{2}(1 - \lambda)nq = \Omega(\log n)$  connections to  $\Gamma_{d^*}^{(2)}$  and at most  $K$  connections to  $\Gamma_{d^*}^{(1)}$ , so their contributions to (53) are  $o(1)$ . ■

**Lemma 31** *Conditioned on  $\Gamma_1^{(i)}, \dots, \Gamma_{d-1}^{(i)}$  and  $v \in V \setminus \Gamma_{\leq d-1}^{(1)} \setminus \Gamma_{\leq d-1}^{(2)}$ , define independent random variables  $X_d^{(i)}$ ,  $Z_d$ , and  $W_d^{(i)}$  such that*

$$X_d^{(i)} \sim \text{Binom}(N_{d-1}^{(i)} - N_{d-1}^I, q)$$

and

$$(Z_d, W_d^{(1)}, W_d^{(2)}, -) \sim \text{Multinomial}(N_{d-1}^I; q_I, q - q_I, q - q_I, -)$$

where  $q_I = q(q(1 - \rho_n) + \rho_n)$  is the edge probability of  $G^{(1)} \cap G^{(2)}$ . Then we have that

$$\begin{bmatrix} |\text{par}_{G^{(1)}}(v)| \\ |\text{par}_{G^{(2)}}(v)| \\ |\text{par}_{G^{(1)}}(v) \cap \text{par}_{G^{(2)}}(v)| \end{bmatrix} \mathbf{1}_{v \sim \Gamma_{d-1}^I} \stackrel{d}{=} \begin{bmatrix} X_d^{(1)} + Z_d + W_d^{(1)} \\ X_d^{(2)} + Z_d + W_d^{(2)} \\ Z_d \end{bmatrix} \mathbf{1}_{Z_d > 0}.$$

In particular, conditioned on  $\Gamma_1^{(i)}, \dots, \Gamma_{d-1}^{(i)}$ , we have

$$Y_{d,v} \stackrel{d}{=} \begin{cases} \frac{Z_d}{\max(X_d^{(1)} + W_d^{(1)}, X_d^{(2)} + W_d^{(2)}) + Z_d} \cdot \mathbf{1}_{Z_d > 0} & \text{if } v \in V \setminus \Gamma_{\leq d-1}^{(1)} \setminus \Gamma_{\leq d-1}^{(2)}, \\ 0 & \text{otherwise.} \end{cases}$$

**Proof** This is clear once we realize that  $X_d^{(i)}$  represents edges in  $G^{(i)}$  connected to  $\Gamma_{d-1}^{(i)} \setminus \Gamma_{d-1}^I$ ,  $Z_d$  represents edges in  $G^{(1)} \cap G^{(2)}$  connected to  $\Gamma_{d-1}^I$ , and  $W_d^{(i)}$  represents edges only in  $G^{(i)}$  connected to  $\Gamma_{d-1}^I$ . These are precisely the parents of  $v$  in the corresponding shortest path DAGs if  $v \sim \Gamma_{d-1}^I$ , i.e.  $Z_d > 0$ . ■

Combining the above lemmata, we obtain the following lemma which is extremely useful in analyzing  $\tilde{R}_n$ .

**Lemma 32** *Consider a random variable  $\Pi_d$  which is a function of  $\Gamma_1^{(i)}, \dots, \Gamma_{d-1}^{(i)}$  defined by*

$$\Pi_d := \mathbb{E} \left[ \frac{Z_d}{\max(X_d^{(1)} + W_d^{(1)}, X_d^{(2)} + W_d^{(2)}) + Z_d} \cdot \mathbf{1}_{Z_d > 0} \middle| \Gamma_1^{(i)}, \dots, \Gamma_{d-1}^{(i)} \right].$$

Then the random variable

$$\bar{\Pi} := \Pi_{d^*} + \lambda^{2-\gamma} \Pi_{d^*+1}$$

satisfies  $|\tilde{R}_n - \bar{\Pi}| \xrightarrow{p} 0$ .

**Proof** We build on Lemma 31. For convenience, we denote  $V_d := V \setminus \Gamma_{\leq d-1}^{(1)} \setminus \Gamma_{\leq d-1}^{(2)}$  and define

$$\bar{Y}_d := \frac{1}{|V_d|} \sum_{v \in V_d} Y_{d,v}.$$

Note that  $\mathbb{E}[\bar{Y}_d \mid \Gamma_1^{(i)}, \dots, \Gamma_{d-1}^{(i)}] = \Pi_d$ . By Lemma 30,  $\bar{Y}_d$  conditioned on  $\Gamma_1^{(i)}, \dots, \Gamma_{d-1}^{(i)}$  is a sample mean of i.i.d. random variables. Since each  $Y_{d,v}$  can have value in  $[0, 1]$ , we have

$$\text{Var}(\bar{Y}_d \mid \Gamma_1^{(i)}, \dots, \Gamma_{d-1}^{(i)}) = \frac{1}{|V_d|} \text{Var}(Y_{d,v} \mid \Gamma_1^{(i)}, \dots, \Gamma_{d-1}^{(i)}) \leq \frac{1}{4|V_d|}.$$

Hence, Chebyshev's inequality gives

$$\Pr(|\bar{Y}_d - \Pi_d| \geq \delta \mid \Gamma_1^{(i)}, \dots, \Gamma_{d-1}^{(i)}) \leq \frac{1}{4\delta^2|V_d|} \quad (55)$$

Now we write

$$\begin{aligned} \bar{Y} &= \frac{1}{n} \sum_{v \in V} (Y_{d^*,v} + Y_{d^*+1,v}) \\ &= \frac{1}{n} \sum_{v \in V} Y_{d^*,v} + \frac{1}{n} \sum_{v \in V} Y_{d^*+1,v} \\ &= \frac{|V_{d^*}|}{n} \bar{Y}_{d^*} + \frac{|V_{d^*+1}|}{n} \bar{Y}_{d^*+1}. \end{aligned}$$

By Theorems 15 and 22, we have  $|V_{d^*}|/n \xrightarrow{p} 1$  and  $|V_{d^*+1}|/n \xrightarrow{p} \lambda^{2-\gamma}$ . For  $\bar{Y}_{d^*}$ , (55) easily gives  $|\bar{Y}_{d^*} - \Pi_{d^*}| \xrightarrow{p} 0$ . For  $\bar{Y}_{d^*+1}$ , (55) gives a similar result if  $\lambda > 0$ ; otherwise,  $|V_{d^*+1}|/n \xrightarrow{p} 0$  so the result trivially holds. ■

**Proof** [Proof of Theorem 27] By Lemma 32, it is enough to prove that  $\bar{\Pi}$  has the given limit in probability.

First, assume  $\lambda = 1$ . Then by Theorem 15 (or in the language of Lemma 32,  $\Pr(Z_{d^*} > 0) \rightarrow 0$ ) it suffices to prove that  $\Pi_{d^*+1}$  concentrates around the given limit. Since  $q(nq)^{d^*} = \log(1/\lambda_n)nq = \Omega(\log n / (\log \log n)^2)$ , Lemma 99 and Proposition 16 give

$$\frac{X_{d^*+1}^{(i)} + W_{d^*+1}^{(i)} + Z_{d^*+1}}{q(nq)^{d^*}} \xrightarrow{p} 1.$$

If  $\gamma = 0$ , then by Proposition 24 we have  $N_{d^*}^{(i)} / (nq)^{d^*} \rightarrow 0$  so  $Z_{d^*+1} / X_{d^*+1}^{(i)} \rightarrow 0$ , thus the expectation tends to zero and the result is trivial. Assuming  $\gamma > 0$ , we get

$$\frac{Z_{d^*+1}}{q(nq)^{d^*}\gamma} \xrightarrow{p} 1.$$

Also,  $\gamma > 0$  implies  $\rho_n \rightarrow 1$ , so  $W_{d^*+1}^{(i)} / Z_{d^*+1} \rightarrow 0$ . Hence, the expectation converges in probability to  $\gamma$  as desired. The proof for  $\lambda = 0$  is almost the same by replacing  $\Pi_{d^*+1}$  with  $\Pi_{d^*}$ .

Now assume  $0 < \lambda < 1$ . Again, we appeal to Lemma 32 and establish probability limits of  $\Pi_{d^*}$  and  $\Pi_{d^*+1}$ . We first prove that

$$\Pi_{d^*} \xrightarrow{p} g((1 - \gamma) \log(1/\lambda), \gamma \log(1/\lambda))(1 - \lambda^\gamma). \quad (56)$$

We may assume  $\gamma > 0$ , otherwise  $Z_{d^*} \rightarrow 0$  by Proposition 24 so the result is trivial. Since  $q(nq)^{d^*-1} \xrightarrow{p} \log(1/\lambda)$ , Propositions 16 and 24 yield  $qN_{d^*-1} \xrightarrow{p} \log(1/\lambda)$  and  $qN_{d^*-1} \xrightarrow{p} \gamma \log(1/\lambda)$  so

$$X_{d^*}^{(i)} \xrightarrow{d} \text{Poisson}((1 - \gamma) \log(1/\lambda)).$$

and

$$Z_{d^*} \xrightarrow{d} \text{Poisson}(\gamma \log(1/\lambda)). \quad (57)$$

Since  $\gamma > 0$  implies  $\rho_n = 1 - o(1)$ , we have  $(q - q_I)/q = o(1)$  so  $W_{d^*}^{(i)} \xrightarrow{p} 0$ . Note that we can write  $\Pi_{d^*}$  as

$$\mathbb{E} \left[ \frac{Z_{d^*}}{\max(X_{d^*}^{(1)} + W_{d^*}^{(1)}, X_{d^*}^{(2)} + W_{d^*}^{(2)}) + Z_{d^*}} \middle| Z_{d^*} > 0, \Gamma_1^{(i)}, \dots, \Gamma_{d-1}^{(i)} \right] \cdot \Pr(Z_{d^*} > 0 \mid \Gamma_1^{(i)}, \dots, \Gamma_{d-1}^{(i)}).$$

Since  $\Pr(Z_{d^*} > 0 \mid \Gamma_1^{(i)}, \dots, \Gamma_{d-1}^{(i)}) \rightarrow 1 - \lambda^\gamma$  by (57), we have proved (56).

Next, we show

$$\Pi_{d^*+1} \xrightarrow{p} \frac{1 - 2\lambda + \lambda^{2-\gamma}}{1 - \lambda} \cdot \rho. \quad (58)$$

By Theorems 15 and 22, we have  $N_{d^*}^{(i)}/n \xrightarrow{p} 1 - \lambda$  and  $N_d^I/n \xrightarrow{p} 1 - 2\lambda + \lambda^{2-\gamma}$ . This gives

$$\frac{X_{d^*+1}^{(i)} + W_{d^*+1}^{(i)} + Z_{d^*+1}}{nq} \xrightarrow{p} 1 - \lambda$$

and

$$\frac{Z_{d^*+1}}{nq_I} \xrightarrow{p} 1 - 2\lambda + \lambda^{2-\gamma}$$

where  $q_I = q(q(1 - \rho_n) + \rho_n)$ . This proves (58). Combining (56) and (58) proves the desired limit in probability.  $\blacksquare$

#### D.3.2. ASIDE: THE OVERLAP UNDER INDEPENDENT COUPLING

When shortest path trees are independently sampled, the conditional expectation of  $R_n$  is given by

$$Q_n := \frac{1}{n-1} \sum_{v \in \bar{V}_n \setminus \{1\}} \frac{|\text{par}_{G^{(1)}}(v) \cap \text{par}_{G^{(2)}}(v)|}{|\text{par}_{G^{(1)}}(v)| \cdot |\text{par}_{G^{(2)}}(v)|}. \quad (59)$$

From this, it is obvious to see that when a vertex has an asymptotically unbounded number of parent choices, then regardless of  $|\text{par}_{G^{(1)}}(v) \cap \text{par}_{G^{(2)}}(v)|$  the probability of choosing the same parent is  $o(1)$ .

Therefore, the only nontrivial contribution to (59) comes from  $\Gamma_{d^*}$  and only when  $0 < \lambda < 1$  and  $\gamma > 0$  (and thus  $\rho = 1$ ), since this is the only case where  $\Theta(n)$  vertices asymptotically have only

a finite number of parents and a nonzero intersection. In this case, the expectation of the probability of choosing the same parent conditioned on  $\Gamma_1^{(i)}, \dots, \Gamma_{d^*-1}^{(i)}$  can be described as

$$\mathbb{E} \left[ \frac{Z_{d^*}}{(X_{d^*}^{(1)} + Z_{d^*})(X_{d^*}^{(1)} + Z_{d^*})} \middle| Z_{d^*} > 0, \Gamma_1^{(i)}, \dots, \Gamma_{d^*-1}^{(i)} \right] \cdot \Pr(Z_{d^*} > 0 \mid \Gamma_1^{(i)}, \dots, \Gamma_{d-1}^{(i)}).$$

The reader may want to compare this with the proof of (56).

This sketches the proof of the following theorem. We omit the rigorous proof since it is essentially the same with the proof of Theorem 27.

**Theorem 33** *Suppose that  $\lambda_n \rightarrow \lambda$ ,  $\gamma_n \rightarrow \gamma$ , and  $\rho_n \rightarrow \rho$  with  $\lambda, \gamma, \rho \in [0, 1]$ . Then as  $n \rightarrow \infty$ , we have that*

$$Q_n \xrightarrow{p} \begin{cases} h((1-\gamma)\log(1/\lambda), \gamma\log(1/\lambda))(1-\lambda^\gamma) & \text{if } 0 < \lambda < 1 \text{ and } \gamma > 0, \\ 0 & \text{otherwise} \end{cases}$$

where  $h(a, b)$  is defined by

$$h(a, b) = \begin{cases} \mathbb{E} \left[ \frac{Z}{(X_1 + Z)(X_2 + Z)} \middle| Z > 0 \right] & \text{if } a, b > 0, \\ \mathbb{E} [Z^{-1} \mid Z > 0] & \text{if } a = 0 \end{cases}$$

for independent  $X_1, X_2 \sim \text{Poisson}(a)$  and  $Z \sim \text{Poisson}(b)$ .

In particular, when  $\rho_n = \rho < 1$  is a fixed constant then the overlap of the shortest path trees will be asymptotically zero, implying the usual notion of disorder chaos (see, e.g., Chatterjee (2009)).

**Remark 34 (Different notions of disorder chaos)** *This notion of disorder chaos is incomparable with Wasserstein disorder chaos, see discussion in El Alaoui et al. (2022). In El Alaoui et al. (2022), they define Wasserstein disorder chaos using  $W_2$  distance instead of  $W_1$ , but this is equivalent for the purpose of determining if Wasserstein disorder chaos is present or not. Explicitly, this is because their underlying distance metric (normalized Hamming distance) is bounded, and for bounded random variables we can compare  $L_1$  and  $L_2$  norms via Jensen and Hölder's inequality.*

#### D.4. No overlap concentration for random shortest paths

In this section, we show that for fixed vertices 1 and 2, there is a parameter regime where the overlap of uniformly random shortest paths provably fails to concentrate. This lies in stark contrast to the case of shortest path trees we studied above. For correlated instances, we sample a pair of shortest paths from the optimal coupling, similar to shortest path trees in the previous section.

This is conceptually related to the Overlap Gap Property established in Li and Schramm (2025), which proves that shortest path cannot be computed by a sufficiently stable algorithm, but not a formal consequence<sup>10</sup>.

**Theorem 35** *There exist sequences  $\alpha_n$  and  $\rho_n$  such that the overlap of uniformly random shortest paths between fixed vertices 1 and 2 has variance bounded away from zero.*

10. As a reminder, the OGP implies that the shortest path is not “stable” under edge-by-edge resampling of the underlying graph.

The idea to construct such parameters is to find a regime where there is non-negligible probability of having a unique path between 1 and 2. Then we argue that the path may either be destroyed or be preserved with non-negligible probability if we choose the right  $\rho_n$ . Lastly, by the absence of short cycles in a typical sparse random graph, another shortest path can't have large overlap should the original path gets destroyed.

**Proof** [Proof of Theorem 35] We choose  $\alpha_n$  that meets the assumption of Theorem 21 and  $\rho_n = 1 - \mu/\ell_n^*$  for any constant  $\mu > 0$ . Then conditioned on the event of Theorem 21, the number of edges along the shortest path  $P$  that gets resampled converges in distribution to  $\text{Poisson}(\mu)$ . Thus, the probability that the path is being preserved converges to  $e^{-\mu}$ . From the discussion above, it is easy to see that the probability that this path is still unique (i.e. no additional edge generated between  $\Gamma_d^{(2)}(1)$  and  $\Gamma_d^{(2)}(2)$ ) in  $G^{(2)}$  converges to  $\lambda$ , in which case the overlap is 1.

On the other hand, with probability converging to  $1 - e^{-\mu}$ , the path gets destroyed. In this case, let  $P'$  be any shortest path from the resampled graph  $G^{(2)}$ . By Theorem 15, the length of  $P'$  is either  $d^*$  or  $d^* + 1$  with high probability, so  $P \cap P' \neq \emptyset$  guarantees the existence of a cycle of length smaller than  $d^*$  in the graph  $G^{(1)} \cup G^{(2)}$ . Since  $G^{(1)} \cup G^{(2)}$  is stochastically dominated by a graph  $G' \sim \mathcal{G}(n, q(2 - \rho_n))$ , we define a parameter

$$\ell^\# := \frac{\log n}{\log(nq(2 - \rho_n))} = \frac{\log n}{\log(nq) + \log(2 - \rho_n)}$$

for  $G'$ . Note that

$$\begin{aligned} \ell^* - \ell^\# &= \frac{\log n}{\log(nq)} - \frac{\log n}{\log(nq) + \log(2 - \rho_n)} \\ &= \frac{(\log n) \log(2 - \rho_n)}{\log(nq)(\log(nq) + \log(2 - \rho_n))} \\ &\leq \frac{(1 - \rho_n) \ell^*}{\log(nq) + \log(2 - \rho_n)} \\ &\leq \frac{\mu}{\log(nq)} \end{aligned}$$

which implies  $|\ell^* - \ell^\#| = o(1)$ . Since  $0 < \lambda < 1$ , we actually have  $|\ell^* - d^*| = o(1)$ , so it follows that the closest integer to  $\ell^\#$  is also  $d^*$ . By applying Corollary 20 to  $G'$ , we have that there is no cycle of length smaller than  $d^*$  with high probability in  $G'$  and thus in  $G^{(1)} \cup G^{(2)}$  as well. Hence, it follows that  $P \cap P' = \emptyset$  for any  $P'$  with probability  $1 - o(1)$ .

In conclusion, we have proved that, for any  $\epsilon > 0$ , the probability that the overlap is 1 is at least  $e^{-\mu} \lambda^2 \log(1/\lambda) - \epsilon$  and the probability that the overlap is zero is at least  $(1 - e^{-\mu}) \lambda \log(1/\lambda) - \epsilon$  for sufficiently large  $n$ . The theorem is proved.  $\blacksquare$

## Appendix E. The partition function of Gibbs measures over trees

In this section, we compute the limiting asymptotics for the log partition function  $\log Z$  of the Gibbs measure over trees. This is the technical heart of the analysis. We leave statements and proofs of most structural properties to Section F afterwards.

One interesting observation in our analysis is that we can compute the partition function of our model in terms of a “spin system” or *Markov random field* (MRF), by relating the distribution over

trees to the corresponding distribution over distance vectors. This is explained in Section E.2 and is helpful for conceptualizing understanding our analysis, so we make use of it throughout the rest of the section.

### E.1. Preliminaries

Recall the definition from (23) of the Gibbs measure over spanning trees  $T$  at inverse temperature  $\beta$ :

$$\mu_{G,\beta}(T) = \frac{1}{Z_{G,\beta}} \exp \left( -\beta \log \log(n) \sum_{v \in \bar{V}} d_T(1, v) \right) \quad (60)$$

where  $\bar{V}$  is the set of vertices in the connected component  $\bar{G}$  of  $G$  containing the source vertex 1.

**Definition 36 (Entropy and Relative Entropy)** Given a discrete probability measure  $\nu$ , we define

$$H_\nu(X) = -\mathbb{E}_{X \sim \nu}[\log(\nu(X))]$$

to be the Shannon entropy. Similarly, for any probability measure  $P$  which is absolute continuous with respect to  $Q$ ,

$$\text{KL}(P, Q) = \mathbb{E}_P \left[ \log \frac{dP}{dQ} \right] = \mathbb{E}_Q \left[ \frac{dP}{dQ} \log \frac{dP}{dQ} \right]$$

is the relative entropy or Kullback–Leibler divergence (see, e.g., Dembo et al. (1991); Van Handel (2014)). When  $\mu$  is the uniform measure on a finite set  $S$ , then for any probability measure  $\nu$  on  $S$ , we have the relation

$$\text{KL}(\nu, \mu) = \mathbb{E}_{X \sim \nu}[\log \nu(X) + \log |S|] = \log |S| - H_\nu(X).$$

**Proposition 37 (Gibbs variational principle, see, e.g., Ellis (2007); Van Handel (2014); Wainwright et al. (2008))**

$$\log Z_{G,\beta} = \sup_{\nu} \left[ H_\nu(T) - \beta \log \log(n) \mathbb{E}_\nu \left[ \sum_v d_T(1, v) \right] \right]$$

where the supremum ranges over all probability measures on the set of spanning trees of the giant component  $\bar{G}$  of  $G$ . Furthermore, the supremum is uniquely attained at the Gibbs measure  $\mu_{G,\beta}$ .

**Proof** This follows by rearranging

$$0 \leq \text{KL}(\nu, \mu_{G,\beta}) = \mathbb{E}_\nu \left[ \log \frac{\nu(T)}{\mu_{G,\beta}(T)} \right] = \log Z_{G,\beta} - H_\nu(T) + \beta \log \log(n) \mathbb{E}_\nu \left[ \sum_v d_T(1, v) \right] \quad (61)$$

where we used the fact that the KL is zero iff  $\nu = \mu_{G,\beta}$ . ■

**Definition 38** Define the naïve mean-field approximation to  $\log Z$  by the right hand side of the following inequality:

$$\log Z_{G,\beta} \geq \sup_{\nu \in \mathcal{M}} \left[ H_\nu(T) - \beta \log \log(n) \mathbb{E}_\nu \left[ \sum_v d_T(1, v) \right] \right]$$

where  $\nu$  ranges over the set  $\mathcal{M}$  of vertex-wise product measures over spanning trees in the following sense: under  $\nu$ , the parent  $\text{par}_T(v)$  of every vertex  $v$  is chosen independently. The inequality is just a special case of the Gibbs variational principle.

Informally, we will say that the naïve mean-field approximation is accurate if it gets the value of  $\log Z$  correct to leading order. Although we do not apply these results, deterministic conditions for guaranteeing the accuracy of the approximation have been extensively studied (Eldan, 2020; Chatterjee and Dembo, 2016; Jain et al., 2019; Austin, 2019; Eldan, 2018; Basak and Mukherjee, 2017; Augeri, 2021, e.g.). We note that accuracy of the mean-field approximation should not always be interpreted as implying that the Gibbs measure is close to a product measure — this is true in a sense for one direction of the KL divergence, but false for the Wasserstein metric in some simple examples. Instead, it generally means that the measure is well-approximated by a small mixture of product measures (see Eldan (2018); Austin (2019)). We will explain and confirm this part of the picture later in Section F with more detail.

**Diameter bound.** We use a fundamental result about the diameter of random graphs, which tells us the diameter of the giant component is not too much bigger than  $d^*$ . The references prove tighter concentration bounds than what is stated here, but the below statement is sufficient for our purposes.

**Theorem 39 (Chung and Lu (2001); Riordan and Wormald (2010))** *Let  $p_n$  be a sequence of elements of  $(0, 1)$ . Asymptotically almost surely, the diameter of the giant component of  $G(n, p_n)$  is*

$$(1 + o(1)) \frac{\log n}{\log(np_n)}$$

provided that  $np_n \rightarrow \infty$ .

## E.2. Transformation into a Markov random field

We next describe a transformation of our model to a Markov random field (MRF). This transformation works on *arbitrary graphs* and is not specific to the random graph setting. Recall that we consider the Gibbs measure on spanning trees of a weighted undirected graph  $G$  given by

$$\mu_{G, \bar{\beta}}(T) = \frac{1}{Z_{G, \bar{\beta}}} \exp \left( -\bar{\beta} \sum_v d_T(1, v) \right)$$

where<sup>11</sup>  $\bar{\beta} \geq 0$ . Note that the solution to the single-source shortest path problem itself is the mode of the distribution for any  $\bar{\beta} > 0$ .

If we think of  $\mu_{G, \bar{\beta}}$  as inducing a probability measure over distance vectors  $D_v = d_T(1, v)$  and let  $\ell_{uv}$  denote the length of the edge between each pair of neighbors  $u$  and  $v$ , then we find  $D$  is a Markov random field<sup>12</sup> on the underlying graph  $G$ . More precisely, we find:

- 
11. We use the notation  $\bar{\beta}$ , because as mentioned above on random graphs we will ultimately let  $\bar{\beta} = \beta \log \log n$  and view  $\beta$  as the actual inverse temperature. This scaling is needed because the degree of vertices in our graph diverges as  $n \rightarrow \infty$ .
  12. I.e., a random vector which satisfies the Markov property that  $D_v$  is conditionally independent of the rest of  $D$  given  $(D_w)_{w \sim v}$ , see e.g., Mezard and Montanari (2009); Lauritzen (1996); Wainwright et al. (2008). This always holds if the PMF is a product over factors on cliques of the graph  $G$ , c.f. the Hammersley-Clifford theorem Lauritzen (1996)

**Proposition 40** *The marginal law of  $D$  given  $T \sim \mu$  is*

$$\Pr[D = d] = \frac{1(d_1 = 0)}{Z_{G,\bar{\beta}}} \prod_{v \neq 1} \exp [H(\text{par}_T(v) \mid D = d) - \bar{\beta} d_v] 1(\exists w \sim v, d_v = d_w + \ell_{vw}). \quad (62)$$

Furthermore, the conditional entropy respects the graph structure, more precisely

$$H(\text{par}_T(v) \mid D = d) = H(\text{par}_T(v) \mid D_v = d_v, \forall w \sim v, D_w = d_w),$$

so  $D$  is a Markov random field on  $G$ .

**Proof** Observe that its marginal law under  $\mu = \mu_{G,\bar{\beta}}$  is

$$\Pr_\mu[D = d] = \Pr_\mu[\forall v : d_T(1, v) = d_v] = \frac{1}{Z_{G,\bar{\beta}}} \sum_T \exp \left( -\bar{\beta} \sum_v d_v \right) 1(\forall v : d_T(1, v) = d_v).$$

Observe

$$\exp \left( -\bar{\beta} \sum_v d_v \right) 1(\forall v : d_T(1, v) = d_v) = 1(d_1 = 0) \prod_{v \neq 1} \exp(-\bar{\beta} d_v) 1(\exists w \sim v, d_v = d_w + \ell_{vw}).$$

Furthermore, given that  $\Pr(D = d) > 0$  we know that  $T$  is a node-wise product measure in the sense that

$$\Pr[\forall v \neq 1, \text{par}(v)_v = p_v \mid D] = \prod_{v \neq 1} \Pr[\text{par}_T(v)_v = p_v \mid D]$$

since, once the distances to the root are known, the choice of parent of each vertex is an independent process which picks for each vertex  $v$  uniformly at random among its neighbors which are at the correct distance, so we have that

$$\log \#\{T : \forall v, d_T(1, v) = d_v\} = H(T \mid D = d) = \sum_v H(\text{par}_T(v) \mid D = d).$$

Combining the above equations gives the stated conclusion. ■

### E.3. Analysis of the partition function

We analyze the partition function of the Gibbs measure defined in (60). For notational convenience, we write

$$\mu_{G,\beta}(T) = \frac{1}{Z_{G,\beta}} \exp \left( -\bar{\beta} \sum_v d_T(1, v) \right) \quad (63)$$

where  $\bar{\beta} = \beta \log \log n$ . From the Gibbs variational principle (Proposition 37), the log partition function can be written as

$$\Phi_{G,\beta} := \log Z_{G,\beta} = \sup_\nu \left[ H_\nu(T) - \bar{\beta} \mathbb{E}_\nu \left[ \sum_v d_T(1, v) \right] \right]. \quad (64)$$

Here,  $\nu$  is a measure on the set of spanning trees  $T$  of the connected component  $\bar{G}$  containing 1 of  $G$ . Our goal of this section is to find a formula for (64) that is accurate up to an  $o(n)$  (absolute) error asymptotically almost surely. More formally, we find a quantity  $\tilde{\Phi}_{G,\beta}^*$  such that

$$\Phi_{G,\beta} = \tilde{\Phi}_{G,\beta}^* + o_p(n).$$

**Organization of the analysis.** The intuition behind our analysis is simple, but the mathematical rigor to justify the idea and formalize the intuition might seem rather long and sophisticated. To present the formal results early but also to guide the readers through the proof idea and an intuitive picture, we organize this section as follows. In Section E.3.1, we formally state core results including the formula for  $\log Z_{G,\beta}$  with brief explanations and remarks, but we do not yet attempt to explain in detail what the terms in the formula mean. Section E.3.2 serves as a quick review of the single graph asymptotics developed in Section C which the rest of the sections rely on. Section E.3.3 then gives a non-rigorous analysis based on the Markov random field representation. Section E.3.4 through Section E.3.8 utilize a handful of mathematical tools to give a formal proof to the main results.

### E.3.1. MAIN RESULTS

**Reduction to a one-dimensional optimization.** Our first result shows that we can solve the variational optimization (64) by instead solving a one-dimensional optimization. Namely, given the graph  $G = G_n$ , we construct a set of measures

$$\hat{\mu}_{G,1}, \hat{\mu}_{G,2}, \dots, \hat{\mu}_{G,N_{d^*}} \quad (65)$$

indexed by the set of integers at most  $N_{d^*}$ , the number of vertices at distance  $d^*$  in  $G$ . The index  $m$  of each measure  $\hat{\mu}_{G,m}$  represents the number of vertices at depth  $d^*$  satisfying  $\text{par}_T(v) \in \text{par}_G(v)$  (i.e., the size of the *kernel* of  $T$  which we study in Section E.3.4), and it maps to a value  $\tilde{\Phi}_{G,\beta}(m)$  which we will define shortly. A measure  $\hat{\mu}_{G,m}$  with the largest value of  $\tilde{\Phi}_{G,\beta}(m)$  is supposed to be a good approximation of  $\mu_{G,\beta}$ . An explicit construction of these measures is delayed to (82) in Section E.3.4.

To define  $\tilde{\Phi}_{G,\beta}(m)$ , we let  $\Psi(m; n, \lambda)$  be a function of integers  $1 \leq m \leq n$  and a real number  $0 < \lambda < 1$

$$\Psi(m; n, \lambda) := \mathbb{E} \left[ \log \left( \sum_{I \in \binom{[n]}{m}} \prod_{i \in I} X_i \right) \right], \quad X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} \text{ZTP}(\log(1/\lambda)) \quad (66)$$

where ZTP is the *zero-truncated Poisson* distribution defined in Definition 107. Now  $\tilde{\Phi}_{G,\beta}(m)$  is defined by

$$\tilde{\Phi}_{G,\beta}(m) := \Psi(m; N_{d^*}, \lambda_n) + (n - m) \log(mq) + \bar{\beta}m - \bar{\beta}(d^* + 1)n. \quad (67)$$

Note that (67) is formulated only in terms of  $n$ ,  $\alpha_n$ , and  $N_{d^*}$ . The following theorem shows that we can approximate  $\Phi_{G,\beta} = \log Z_{G,\beta}$  by solving a one-dimensional optimization

$$\tilde{\Phi}_{G,\beta}^* := \max_{1 \leq m \leq N_{d^*}} \tilde{\Phi}_{G,\beta}(m). \quad (68)$$

**Theorem 41** *There exists a collection of measures (65) with the following property. For any fixed constant  $\beta > 0$ , let  $m^*$  be any function of  $n$ ,  $\alpha_n$ , and  $N_{d_n^*}$  (and  $\beta$ ) such that  $m^* \in \{1, \dots, N_{d_n^*}\}$  almost surely and it approximately solves (68) in a sense that*

$$\tilde{\Phi}_{G_n,\beta}^* = \tilde{\Phi}_{G_n,\beta}(m^*) + o_p(n). \quad (69)$$

*Then we have the following.*

- $\hat{\mu}_{G_n, m^*}$  approximates  $\mu_{G_n, \beta}$  in KL divergence:

$$\text{KL}(\hat{\mu}_{G_n, m^*}, \mu_{G_n, \beta}) = o_p(n).$$

- The log partition function  $\Phi_{G_n, \beta}$  is well approximated by  $\tilde{\Phi}_{G_n, \beta}(m^*)$ :

$$\Phi_{G_n, \beta} = \tilde{\Phi}_{G_n, \beta}(m^*) + o_p(n).$$

**Proof** See Section E.3.7. ■

An equivalent but slightly different viewpoint to this reduction is to consider the following mean-field spin system with a random external field. Suppose that the nature samples  $A_1, \dots, A_n \stackrel{\text{i.i.d.}}{\sim} \text{Poisson}(\log(1/\lambda_n))$ . We consider a (random) measure defined on  $\mathbf{x} = (x_1, \dots, x_n) \in \{0, 1\}^n$

$$\frac{\mathbf{1}_{A_i=0 \implies x_i=0}}{Z_{\mathbf{A}}} \exp \left( \sum_{i: A_i > 0} (\log A_i - \log(nq) + \bar{\beta})x_i + n \cdot g(\bar{\mathbf{x}}) \right) \quad (70)$$

where  $\bar{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^n x_i$  and  $g(x) = (1-x) \log x$ . Then the second part of Theorem 41 can be stated in the following way:

$$\log Z_{G, \beta} = \log Z_{\mathbf{A}} + n \log(nq) - \bar{\beta}(d^* + 1)n + o_p(n).$$

This fact is not needed nor is explicitly proved in our analysis. However, this formulation may provide further insights to our model and is especially useful when we discuss the behavior near the critical temperature. See Appendix I for details.

**Solving the optimization.** The next step is to find a suitable  $m^*$  that satisfies (69). As we will see shortly, depending on the constant  $\beta > 0$  and the asymptotics of the proxy sequences, there are two regimes where the solutions are far apart. To state the result, we define a new proxy

$$\kappa_n := (1 - \lambda_n) \log \log n.$$

Since  $N_{d^*} \approx (1 - \lambda_n)n$  by Theorem 17,  $\kappa_n$  expresses the size of  $N_{d^*}$  at the order of  $n / \log \log n$ .

Now we precisely define the two regimes where we state our result. The first is the low temperature phase, which is the union of two regimes:

$$\begin{cases} \beta > \limsup_{n \rightarrow \infty} (1 - \Delta_n - \kappa_n^{-1}) \\ \beta = \limsup_{n \rightarrow \infty} (1 - \Delta_n - \kappa_n^{-1}) \end{cases} \quad \text{and} \quad \limsup_{n \rightarrow \infty} \kappa_n < \infty. \quad (\text{LT})$$

Next is the high temperature phase

$$\beta < \liminf_{n \rightarrow \infty} (1 - \Delta_n - \kappa_n^{-1}). \quad (\text{HT})$$

Note that these do not cover all possible regimes, especially when  $\beta$  is near the limits of  $1 - \Delta_n - \kappa_n^{-1}$ . Nevertheless, we present a formula of  $\log Z_{G, \beta}$  for the two regimes.

**Theorem 42 (Approximate log partition function)** *In each phase, we have the following formula for the log partition function  $\Phi_{G_n, \beta} = \log Z_{G_n, \beta}$ .*

- In the low temperature phase (LT),  $m^* = N_{d_n^*}$  satisfies (69) and we have

$$\Phi_{G_n, \beta} = n\Psi_0(\lambda_n) + \lambda_n n \log(\alpha_n(1 - \lambda_n) \log n) - \bar{\beta}(d_n^* + \lambda_n)n + o_p(n)$$

where

$$\Psi_0(\lambda) := \mathbb{E}[\log(X \vee 1)], \quad X \sim \text{Poisson}(\log(1/\lambda)).$$

- In the high temperature phase (HT),  $m^* = \lfloor \frac{N_{d_n^*}}{(1 - \Delta_n - \beta)\kappa_n} \rfloor$  satisfies (69) and we have

$$\Phi_{G_n, \beta} = n \log \left( \frac{\alpha_n \log n}{(1 - \Delta_n - \beta) \log \log n} \right) - n - \bar{\beta}(d_n^* + 1)n + o_p(n).$$

**Proof** See Section E.3.8. ■

**Remark 43** In Section E.3.4, we will construct the measures  $\hat{\mu}_{G, m}$  in a way that the last measure  $\hat{\mu}_{G, N_{d^*}}$  corresponding to  $m = N_{d^*}$  is precisely the uniform measure  $\mu_{G, \infty}$  over the shortest path trees. Then Theorem 42 indicates that the Gibbs measure  $\mu_{G, \beta}$  is close in KL divergence to the uniform measure  $\mu_{G, \infty}$  in the low temperature phase (LT).

### E.3.2. REGULARITY PROPERTIES OF THE UNDERLYING GRAPH

Before we proceed on studying the Gibbs measures  $\mu_{G, \beta}$ , we remind the reader that  $\mu_{G, \beta}$  themselves are *random* measures, since the graph  $G$  is drawn from the Erdős–Rényi model  $\mathcal{G}(n, q)$ . Thus, when we say we sample a random spanning tree  $T$  from  $\mu_{G, \beta}$ , there are in fact two sources of randomness: one from  $G$ , and one from  $\mu_{G, \beta}$  conditioned on  $G$ . Thus, for most of the results in this section, we actually prove that  $\mu_{G, \beta}$  enjoys certain properties for “nice” graphs  $G$ , where  $G \sim \mathcal{G}(n, q)$  is “nice” with probability tending to 1 as  $n \rightarrow \infty$ . Many of such “nice” features of a random graph come from what we have developed in Section C. In this section, we collect these “regularity properties” on  $G$  we will frequently refer back to throughout the proofs in Section E. Some of the proofs need additional regularity properties specific to them, and these will usually be stated as a separate lemma.

As for notations, we try to distinguish the randomness from  $G$  and from  $\mu_{G, \beta}$  to minimize confusion. Whenever we study the probability of a certain event  $E$  of spanning trees under the Gibbs measure  $\mu_{G, \beta}$  conditioned on  $G$ , we write

$$\mu_{G, \beta}(E)$$

instead of the usual probability symbol, e.g.,  $\Pr_{T \sim \mu_{G, \beta} | G}(E)$ . For the expectation, we always include a subscript to indicate under what measure we are taking the expectation over, such as

$$\mathbb{E}_{\mu_{G, \beta}}[f(T)].$$

Before we proceed, we briefly recap some relevant facts regarding the shortest path structure of  $G$  we developed in Section C. The purpose of this section is to provide an intuitive picture, so the results will be stated rather informally.

- The single most important groundwork for our analysis is perhaps the stochastic process (15). A useful corollary of this is that conditioned on  $\Gamma_{\leq d-1}$ , we have

$$|\text{par}_G(v)| \sim \text{ZTB}(N_{d-1}, q)$$

for all  $v \in \Gamma_d$ , where  $\text{ZTB}(n, p)$  is the *zero-truncated binomial* distribution, the binomial. Essential properties of zero-truncated distributions can be found in Appendix H.2.

- Proposition 16 tells us that

$$|\Gamma_{\leq d^*-1}| = O((nq)^{d^*-1}) = O_p\left(\frac{n}{(\log \log n)^2}\right). \quad (71)$$

Since a typical vertex has degree  $O(\log n)$ , the entropy gained from picking its parent  $\text{par}_T(v)$  is at most  $O(\log \log n)$ . This implies that the entropy coming from vertices of distance at most  $d^* - 1$  is only  $o_p(n)$  and thus can be safely ignored.

- By Theorem 17, most of the vertices are in  $\Gamma_{d^*} \cup \Gamma_{d^*+1}$ , i.e., have distance either  $d^*$  or  $d^* + 1$  to the root 1 in  $G$ . In addition,  $N_{d^*}$  is well approximated by  $(1 - \lambda_n)n$ :

$$N_{d^*} = \left(1 + o_p\left(\frac{1}{\log \log n}\right)\right) (1 - \lambda_n)n. \quad (72)$$

- Conditioned on  $\Gamma_{\leq d^*-1}$ , each  $v \in \Gamma_{d^*}$  has  $\text{ZTB}(N_{d^*-1}, q)$  parents. Since  $qN_{d^*-1}$  concentrates around  $q(nq)^{d^*-1} = \log(1/\lambda_n)$  by Proposition 16, this is approximately  $\text{ZTP}(\log(1/\lambda_n))$ , the *zero-truncated Poisson* distribution which is briefly discussed in Appendix H.3. If, for instance,  $\Delta_n \rightarrow \Delta \in (0, 1]$  (and thus  $\lambda_n \rightarrow \lambda = 0$ ), this concentrates around  $(nq)^\Delta$ .
- Conditioned on  $\Gamma_{\leq d^*}$ ,  $G[\Gamma_{d^*}]$  is marginally an Erdős–Rényi graph; consequently, a typical vertex  $v \in \Gamma_{d^*}$  has approximately  $qN_{d^*}$  neighbors in  $\Gamma_{d^*}$ . By Lemma 18 this concentrates around  $q(nq)^{d^*}$ , which is  $\Omega(\log n / (\log \log n)^2)$ .
- Conditioned on  $\Gamma_{\leq d^*}$ , each  $v \in \Gamma_{d^*+1}$  has  $\text{ZTB}(N_{d^*}, q)$  parents, which concentrates around  $qN_{d^*}$ . Again, this is at least  $\Omega(\log n / (\log \log n)^2)$  asymptotically almost surely.

### E.3.3. AN INFORMAL DESCRIPTION OF THE GIBBS MEASURES

To get a sense of the behavior of  $\mu_{G,\beta}$ , we revisit the informal analysis in Section A.3 from a slightly different viewpoint — a Markov Random Field perspective from Section E.2. Namely, we examine the (marginal) law of the *distance vector*  $\mathbf{d}_T$  under  $\mu_{G,\beta}$ . Imagine that we are sampling a distance vector using a Gibbs sampler. Namely, we sample a value of  $\mathbf{d}_T(1, v)$  for each non-root vertex  $1 \neq v \in V$ , conditioned on the distance values of all the other vertices. The (unnormalized) log probability of a particular distance  $d$  is the sum of the *parent selection entropy*

$$H(\text{par}_T(v)) := \log |\{u \in N_G(v) : \mathbf{d}_T(1, u) = d - 1\}| = \log |N_G(v) \cap \Gamma_{d-1}^T|$$

and the negative *potential energy* times the inverse temperature

$$-\bar{\beta}d = -\beta d \log \log n.$$

Now we run an “informal” Gibbs sampling, starting with the initial vector  $d_T = d_G$ . In terms of the distribution over spanning trees, we are starting with the uniform measure over the shortest path trees. For each vertex  $v$ , we consider candidate distances that  $v$  can choose, and roughly estimate how much parent selection entropy and potential energy  $v$  might gain.

- Suppose that  $v \in \Gamma_{d^*}$ . If we choose to leave  $v$  at the same level, i.e.,  $d_T(1, v) = d^*$ , then the parent selection entropy is approximately the log of  $ZTP(\log(1/\lambda_n))$ . For  $\Delta > 0$ , this would concentrate around  $\Delta \log(nq) \approx \Delta \log \log n$ . On the other hand, if we choose  $d^* + 1$ , the parent selection entropy is at least  $\log(\log n / (\log \log n)^2) \approx \log \log n$ , but  $v$  also gains  $\beta \log \log n$  potential energy. Since  $v$  gains almost the maximum parent selection entropy  $\log \log n$  at level  $d^* + 1$ , it will almost never move to level  $d^* + 2$  or greater. From this observation, we make the following guess:

- if  $\beta > 1 - \Delta$ , then  $v$  will most likely stay at distance  $d^*$  and
- if  $\beta < 1 - \Delta$ , then  $v$  will most likely jump to distance  $d^* + 1$ .

This is an example of a *phase transition*. See Section F.2 for more details including formal statements.

- Suppose that  $v \in \Gamma_{d^*+1}$ . Then the parent selection entropy is at least  $\log(\log / (\log \log n)^2) \approx \log \log n$ , which is already nearly the maximum. Thus, as long as  $\beta$  is a positive constant,  $v$  will almost never move.
- The behavior of the vertices at level  $d^* - 1$  or below might be a little more complicated. For now, we will just accept that they will almost never move. Roughly speaking, this is because a typical vertex in  $\Gamma_{\leq d^*-1}$  has a lot of descendants at level  $d^*$  and  $d^* + 1$ ; if too many of  $\Gamma_{\leq d^*-1}$  leave their level, then a lot of vertices might gain a lot of potential energy which wouldn’t justify the entropy gain.

This suggests that the law of a spanning tree  $T \sim \mu_{G,\beta}$  of  $\bar{G}$  largely depends on the dynamics of the vertices in  $\Gamma_{d^*}$ , which is captured by the notion of *kernels* defined in the following section.

#### E.3.4. KERNELS AND CONDITIONAL GIBBS MEASURES

As mentioned in the preceding section, a key notion that lies in the heart of our analysis is the *kernel* of a spanning tree, which we define as follows.

**Definition 44** Let  $T$  be a spanning tree of  $\bar{G}$ . The kernel of  $T$ , denoted  $\varphi(T)$ , is defined by

$$\varphi(T) := \Gamma_{d^*} \cap \text{ch}_T(\Gamma_{d^*-1}).$$

In other words,  $\varphi(T)$  is the set of vertices which are children of  $\Gamma_{d^*-1}$ , both for  $G$  and  $T$ . Intuitively, from the view of our imaginary Gibbs sampler from Section E.3.3,  $\varphi(T)$  is related to the set of vertices in  $\Gamma_{d^*}$  that “stay”, though we don’t require vertices in the kernel to have distance  $d^*$  in  $T$  (in fact, they do most of the time; see Section F.1). The following proposition characterizes all possible kernels.

**Proposition 45** Any kernel intersects every connected component of  $\bar{G} \setminus \Gamma_{\leq d^*-1}$ . Conversely, any subset of  $\Gamma_{d^*}$  intersecting every component of  $\bar{G} \setminus \Gamma_{\leq d^*-1}$  is the kernel of a spanning tree of  $\bar{G}$ .

**Proof** Assume  $A = \varphi(T)$ . For any vertex  $v \in \bar{V} \setminus \Gamma_{\leq d^*-1}$ , a (unique) path in  $T$  from  $v$  to the root must contain an edge  $u_1 \rightarrow u_2$  where  $u_1 \in \Gamma_{d^*}$  and  $u_2 \in \Gamma_{d^*-1}$ . This implies  $u_1 \in \text{ch}_T(\Gamma_{d^*-1})$  and thus  $u_1 \in A$ . Hence, there is a path in  $T$  from  $v$  to  $A$ .

Conversely, let  $A \subseteq \Gamma_{d^*}$  be a set that intersects every component of  $\bar{G} \setminus \Gamma_{\leq d^*-1}$ . We construct a spanning tree  $T$  as follows:

1. choose a spanning tree  $T_0$  of  $G[\Gamma_{\leq d^*-1}]$ ;
2. choose edges  $e_1, \dots, e_{|A|}$  between the vertices in  $A$  and  $\Gamma_{d^*-1}$ , one for each vertex;
3. choose spanning trees  $T_1, \dots, T_k$  of the all  $k$  components of  $\bar{G} \setminus \Gamma_{\leq d^*-1}$ ;
4. set  $T = T_0 \cup T_1 \cup \dots \cup T_k \cup \{e_1, \dots, e_{|A|}\}$ .

It is straightforward to verify that  $T$  is indeed a spanning tree of  $\bar{G}$  and that  $A = \varphi(T)$ . ■

We denote by

$$\mathcal{K}(G) := \{\varphi(T) : T \text{ is a spanning tree of } \bar{G}\}$$

the set of all possible kernels as characterized in the proposition.

**Conditioning on the kernel.** One of the reasons why kernels are useful is because the Gibbs measures  $\mu_{G,\beta}$  are easy to describe when conditioned on the kernel. Consider a decomposition of the partition function  $Z_{G,\beta}$  by grouping together the spanning trees with the same kernel, i.e.,

$$Z_{G,\beta} = \sum_{A \in \mathcal{K}(G)} Z_{G,\beta}|_A \quad (73)$$

where

$$Z_{G,\beta}|_A := \sum_{T: \varphi(T)=A} \exp \left( -\bar{\beta} \sum_v d_T(1, v) \right).$$

This corresponds to decomposing a measure  $\nu$  on the spanning trees of  $\bar{G}$  by

$$\nu = \sum_{A \in \mathcal{K}(G)} \varphi_*(\nu)(A) \cdot \nu|_A$$

where

$$\varphi_*(\nu)(A) := \nu(\varphi(T) = A)$$

are the mixture weights and

$$\nu|_A(T) = \nu(T \mid \varphi(T) = A)$$

are the conditional measures. Here,  $\varphi_*(\nu)$  can equivalently be understood as the pushforward of  $\nu$  by  $\varphi$ . Our first result is that whenever  $|A|$  is reasonably large, regardless of the constant  $\beta > 0$ ,  $\mu_{G,\beta}|_A$  behaves like its zero-temperature limit  $\mu_{G,\infty}|_A$  which can be described by the following sampling procedure:

1. for each  $v \in \Gamma_{\leq d^*-1}$ , sample  $\text{par}_T(v)$  from  $\text{par}_G(v)$  uniformly at random;
2. for each  $v \in A$ , sample  $\text{par}_T(v)$  from  $\text{par}_G(v)$  uniformly at random;

3. for each  $v \in \bar{V} \setminus \Gamma_{\leq d^*-1} \setminus A$ , sample  $\text{par}_T(v)$  uniformly at random from the set of vertices  $u \in \mathcal{N}_G(v)$  such that  $d_G(v, A) = d_G(u, A) + 1$ .

In other words,  $\mu_{G,\infty}|_A$  is a product measure that always samples a minimum energy configuration, conditioned on  $\varphi(T) = A$ . To state the formal result, we define

$$m_\ell := m_0 \vee \left\lceil \frac{n}{3(\log \log n)^2} \right\rceil \quad (74)$$

where

$$m_0 := \min_{A \in \mathcal{K}(G)} |A| = \#\{\text{connected components of } \bar{G} \setminus \Gamma_{\leq d^*-1}\} \quad (75)$$

is the minimum size of a kernel.

**Proposition 46** *Let  $\beta > 0$  be a constant. Then with probability at least  $1 - o(1)$ , we have*

$$\text{KL}(\mu_{G,\infty}|_A, \mu_{G,\beta}|_A) = o(n)$$

and

$$\log(Z_{G,\beta}|_A) = \sum_{v \in A} \log |\text{par}_G(v)| + (n - |A|) \log(|A|q) + \bar{\beta}|A| - \bar{\beta}(d^* + 1)n + o(n) \quad (76)$$

simultaneously for all  $A \in \mathcal{K}(G)$  with size at least

$$|A| \geq m_\ell. \quad (77)$$

We postpone the proof of Proposition 46 to Section E.3.5. Once we understand  $\log(Z_{G,\beta}|_A)$ , (73) gives

$$\begin{aligned} \Phi_{G,\beta} &= \log Z_{G,\beta} \\ &\leq \max_{A \in \mathcal{K}(G)} \log(Z_{G,\beta}|_A) + \log |\mathcal{K}(G)| \\ &\leq \max_{A \in \mathcal{K}(G)} \log(Z_{G,\beta}|_A) + O(n). \end{aligned}$$

We clearly have  $\max_{A \in \mathcal{K}(G)} \log(Z_{G,\beta}|_A) \leq \Phi_{G,\beta}$ , so this in fact implies

$$\Phi_{G,\beta} = \max_{A \in \mathcal{K}(G)} \log(Z_{G,\beta}|_A) + O(n). \quad (78)$$

Thus, if we know that the maximum is attained where (77) holds, then we obtain an approximate value of  $\Phi_{G,\beta}$  by optimizing (76) over  $A$ , and the corresponding measure  $\mu_{G,\infty}|_A$  will be close to  $\mu_{G,\beta}$  in KL distance. However, this approximation has error up to  $O(n)$ , due to  $\mathcal{K}(G)$  being too large.

**Conditioning on the kernel size.** If we aim at a more delicate analysis, we should use a coarser decomposition of  $Z_{G,\beta}$ . Namely, instead of grouping by the kernel, we group by the kernel size. We define

$$\mathcal{K}_m(G) := \{A \in \mathcal{K}(G) : |A| = m\}$$

as the collection of kernels of size  $m$ . Consider a decomposition of  $Z_{G,\beta}$  by

$$Z_{G,\beta} = \sum_{m=m_0}^{N_{d^*}} Z_{G,\beta}|_m$$

where

$$Z_{G,\beta}|_m := \sum_{A \in \mathcal{K}_m(G)} Z_{G,\beta}|_A. \quad (79)$$

The corresponding decomposition of a measure  $\nu$  can be written as

$$\nu = \sum_{m=m_0}^{N_{d^*}} |\varphi|_*(\nu)(m) \cdot \nu|_m$$

where

$$|\varphi|_*(\nu)(m) := \nu(|\varphi(T)| = m)$$

and

$$\nu|_m(T) := \nu(T \mid |\varphi(T)| = m).$$

Here, in contrast to (78), we have

$$\begin{aligned} \Phi_{G,\beta} &\leq \max_{m_0 \leq m \leq N_{d^*}} \log(Z_{G,\beta}|_m) + \log(N_{d^*} - m_0 + 1) \\ &\leq \max_{m_0 \leq m \leq N_{d^*}} \log(Z_{G,\beta}|_m) + O(\log n) \end{aligned}$$

so

$$\Phi_{G,\beta} = \max_{m_0 \leq m \leq N_{d^*}} \log(Z_{G,\beta}|_m) + O(\log n) \quad (80)$$

which is tight enough. Our goal now is to analyze (79). Plugging in (76) from Proposition 46, we get

$$\begin{aligned} \log(Z_{G,\beta}|_m) &= \log \left( \sum_{A \in \mathcal{K}_m(G)} Z_{G,\beta}|_A \right) \\ &= \log \left( \sum_{A \in \mathcal{K}_m(G)} \prod_{v \in A} |\text{par}_G(v)| \right) + (n - m) \log(mq) + \bar{\beta}m - \bar{\beta}(d^* + 1)n + o_p(n). \end{aligned}$$

Here, we define

$$\mathcal{H}_G(m) := \log \left( \sum_{A \in \mathcal{K}_m(G)} \prod_{v \in A} |\text{par}_G(v)| \right) \quad (81)$$

which is the log partition function of the measure  $\tau_{G,m}$  on  $\mathcal{K}_m(G)$  defined by

$$\tau_{G,m}(A) \propto \prod_{v \in A} |\text{par}_G(A)|.$$

Thus, to approximate the measures  $\mu_{G,\beta} \|_m$ , we build measures  $\hat{\mu}_{G,m}$  by mixing  $\mu_{G,\infty}|_A$  with mixture weight  $\tau_{G,m}$ :

$$\hat{\mu}_{G,m} := \sum_{A \in \mathcal{K}_m(G)} \tau_{G,m}(A) \cdot \mu_{G,\infty}|_A. \quad (82)$$

Then we arrive at the following result.

**Proposition 47** *Let  $\beta > 0$  be a constant. Then with probability at least  $1 - o(1)$ , we have*

$$\text{KL}(\hat{\mu}_{G,m}, \mu_{G,\beta} \|_m) = o(n)$$

and

$$\log(Z_{G,\beta} \|_m) = \mathcal{H}_G(m) + (n - m) \log(mq) + \bar{\beta}m - \bar{\beta}(d^* + 1)n + o(n)$$

simultaneously for all  $m_\ell \leq m \leq N_{d^*}$ .

**Proof** We already proved the second part. For the first part, for the sake of convenience, we define

$$f(\nu) := H_\nu(T) - \bar{\beta} \mathbb{E}_\nu \left[ \sum_v d_T(1, v) \right].$$

We observe that by Proposition 46 and the Gibbs variational principle (Proposition 37)

$$\begin{aligned} \text{KL}(\hat{\mu}_{G,m}, \mu_{G,\beta} \|_m) &= \log(Z_{G,\beta} \|_m) - f(\hat{\mu}_{G,m}) \\ &= \log(Z_{G,\beta} \|_m) - (H_{\tau_{G,m}}(\mathbf{A}) + \mathbb{E}_{\tau_{G,m}} [f(\mu_{G,\infty}|\mathbf{A})]) \\ &= \log(Z_{G,\beta} \|_m) - (H_{\tau_{G,m}}(\mathbf{A}) + \mathbb{E}_{\tau_{G,m}} [\log(Z_{G,\beta}|\mathbf{A})]) + o_p(n) \\ &= \mathcal{H}_G(m) - \left( H_{\tau_{G,m}}(\mathbf{A}) + \mathbb{E}_{\tau_{G,m}} \left[ \sum_{v \in \mathbf{A}} \log |\text{par}_G(v)| \right] \right) + o_p(n) \\ &= o_p(n) \end{aligned}$$

as desired. ■

**Remark 48** *The measures  $\hat{\mu}_{G,m}$  defined in (82) will be our desired set of measures (65) that satisfies Theorem 41. Of course we only have  $\hat{\mu}_{G,m}$  for  $m \geq m_0$  because otherwise there is no available kernel. For  $m < m_0$  we can just set  $\hat{\mu}_{G,m}$  arbitrarily since these values of  $m$  can never approximately solve (68); this will be proved in Section E.3.7. Note that when  $m = N_{d^*}$ ,  $\tau_{G,m}$  simply puts all mass to  $A = \Gamma_{d^*}$  so  $\hat{\mu}_{G,m} = \mu_{G,\infty}$  as we discussed in the remark below Theorem 42.*

## E.3.5. PROOF OF PROPOSITION 46

**An upper bound.** We start by deriving an upper bound of  $\log(Z_{G,\beta}|_A)$  which leverages the Gibbs variational principle and a nice property of kernels.

**Lemma 49** *Let  $M$  be a constant greater than the maximum degree of  $G$ . Then with probability at least  $1 - o(1)$ , we have*

$$\log(Z_{G,\beta}|_A) \leq \sum_{v \in A} \log |\text{par}_G(v)| + \sum_{v \in \bar{V} \setminus \Gamma_{\leq d^*-1} \setminus A} \log(|N_A(v)| + M e^{-\bar{\beta}}) + \bar{\beta}|A| - \bar{\beta}(d^* + 1)n + o(n)$$

simultaneously for all  $A \in \mathcal{K}(G)$ .

**Proof** First, let  $A \in \mathcal{K}(G)$  be fixed. By the Gibbs variational principle (Proposition 37), we have

$$\log(Z_{G,\beta}|_A) = \sup_{\nu} \left[ H_{\nu}(T) - \bar{\beta} \mathbb{E}_{\nu} \left[ \sum_{v \in \bar{V}} d_T(1, v) \right] \right]$$

where  $\nu$  ranges over all measures over the spanning trees  $T$  of  $\bar{G}$  such that  $\varphi(T) = A$ . Using the sub-additivity of entropy (see, e.g., Dembo et al. (1991)), we write

$$\begin{aligned} H_{\nu}(T) &\leq \sum_{v \in \Gamma_{\leq d^*-1} \setminus \{1\}} H_{\nu}(\text{par}_T(v)) + \sum_{v \in A} H_{\nu}(\text{par}_T(v)) + \sum_{v \in \bar{V} \setminus \Gamma_{\leq d^*-1} \setminus A} H_{\nu}(\text{par}_T(v)) \\ &\leq |\Gamma_{\leq d^*-1}| \log M + \sum_{v \in A} \log |\text{par}_G(v)| + \sum_{v \in \bar{V} \setminus \Gamma_{\leq d^*-1} \setminus A} H_{\nu}(\text{par}_T(v)) \end{aligned}$$

where we have used the fact that since  $A$  is the kernel of  $T$ , the choice of parents for each  $v \in A$  is limited to its parents in  $G$ . For the energy part, we have

$$\begin{aligned} \sum_{v \in \bar{V}} d_T(1, v) &= \sum_{v \in A} d_T(1, v) + \sum_{v \in \bar{V} \setminus A} d_T(1, v) \\ &\geq \sum_{v \in A} d_T(1, v) + \sum_{v \in \Gamma_{d^*-1}} d_T(1, v) + \sum_{v \in \bar{V} \setminus \Gamma_{\leq d^*-1} \setminus A} d_T(1, v) \\ &\geq N_{d^*-1}(d^* - 1) + |A|d^* + \sum_{v \in \bar{V} \setminus \Gamma_{\leq d^*-1} \setminus A} d_T(1, v). \end{aligned}$$

A vertex  $v \in \bar{V} \setminus \Gamma_{\leq d^*-1} \setminus A$  is not in the kernel, so any path to the root must pass through  $A$ . This implies that  $d_T(1, v) \geq d^* + d_T(v, A)$ . Thus,

$$\begin{aligned} \sum_{v \in \bar{V}} d_T(1, v) &\geq N_{d^*-1}(d^* - 1) + |A|d^* + \sum_{v \in \bar{V} \setminus \Gamma_{\leq d^*-1} \setminus A} (d^* + d_T(v, A)) \\ &= |\bar{V} \setminus \Gamma_{\leq d^*-1}|d^* + \sum_{v \in \bar{V} \setminus \Gamma_{\leq d^*-1} \setminus A} d_T(v, A) - N_{d^*-1}. \end{aligned}$$

This gives

$$\begin{aligned} \log(Z_{G,\beta}|_A) &\leq |\Gamma_{\leq d^*-1}| \log M + \sum_{v \in A} \log |\text{par}_G(v)| - \bar{\beta} |\bar{V} \setminus \Gamma_{\leq d^*-2}| d^* + \bar{\beta} N_{d^*-1} \\ &\quad + \sup_{\nu} \left[ \sum_{v \in \bar{V} \setminus \Gamma_{\leq d^*-1} \setminus A} (H_{\nu}(\text{par}_T(v)) - \bar{\beta} \mathbb{E}_{\nu} [\text{d}_T(v, A)]) \right]. \end{aligned} \quad (83)$$

Thus supremum in this expression is bounded by

$$\sum_{v \in \bar{V} \setminus \Gamma_{\leq d^*-1} \setminus A} \sup_{\nu} [H_{\nu}(\text{par}_T(v)) - \bar{\beta} \mathbb{E}_{\nu} [\text{d}_T(v, A)]] .$$

A trick here is to use a simple bound  $\text{d}_T(v, A) \geq 1 + \mathbf{1}_{\text{par}_T(v) \notin A}$ . Again using the Gibbs variational principle, we have

$$\begin{aligned} \sup_{\nu} [H_{\nu}(\text{par}_T(v)) - \bar{\beta} \mathbb{E}_{\nu} [\mathbf{1}_{\text{par}_T(v) \notin A}]] &\leq \log \left( \sum_{u \in \mathbf{N}_A(v)} e^0 + \sum_{u \in \mathbf{N}_G(v) \setminus \mathbf{N}_A(v)} e^{-\bar{\beta}} \right) \\ &\leq \log(|\mathbf{N}_A(v)| + M e^{-\bar{\beta}}) . \end{aligned}$$

Plugging into (83), we get

$$\begin{aligned} \log(Z_{G,\beta}|_A) &\leq |\Gamma_{\leq d^*-1}| \log M + \sum_{v \in A} \log |\text{par}_G(v)| + \sum_{v \in \bar{V} \setminus \Gamma_{\leq d^*-1} \setminus A} \log(|\mathbf{N}_A(v)| + M e^{-\bar{\beta}}) + \bar{\beta} |A| \\ &\quad - \bar{\beta} |\bar{V} \setminus \Gamma_{\leq d^*-2}| (d^* + 1) + \bar{\beta} N_{d^*-1} . \end{aligned}$$

Now by Proposition 16 and the definition on  $d^*$ , we have

$$|\Gamma_{\leq d^*-1}| = O_p((nq)^{d^*-1}) = O_p \left( \frac{n}{(\log \log n)^2} \right)$$

which implies  $|\Gamma_{\leq d^*-1}| \log M = o_p(n)$ . Similarly,  $\bar{\beta} N_{d^*-1} = o_p(n)$ . Also, note that

$$|\bar{V} \setminus \Gamma_{\leq d^*-2}| = n - O_p((nq)^{d^*-2}) = n - O_p \left( \frac{n}{\log n (\log \log n)^2} \right) \quad (84)$$

so that

$$\bar{\beta} |\bar{V} \setminus \Gamma_{\leq d^*-1}| (d^* + 1) = \bar{\beta} n (d^* + 1) - o_p(n) .$$

This gives the desired bound. ■

Our goal here is to control the sum

$$\sum_{v \in} \log(\mathbf{N}_A(v) + M e^{-\bar{\beta}}) .$$

A key observation is that conditioned on  $\Gamma_{\leq d^*-1}$  and  $A$ , we have that  $|\mathbf{N}_A(v)|$  for  $v \in V \setminus \Gamma_{\leq d^*-1} \setminus A$  are i.i.d. with distribution  $\text{Binom}(|A|, q)$ . Thus, we hope to have an approximation

$$\sum_{v \in \bar{V} \setminus \Gamma_{\leq d^*-1} \setminus A} \log(Me^{-\bar{\beta}} + |\mathbf{N}_A(v)|) \approx |\bar{V} \setminus \Gamma_{\leq d^*-1} \setminus A| \mathbb{E}[\log(Me^{-\bar{\beta}} + \text{Binom}(|A|, q))]$$

which are accurate up to, hopefully,  $o(n)$ . We would like to apply Lemma 100 to  $\log(Me^{-\bar{\beta}} + |\mathbf{N}_A(v)|)$ . An issue is that if  $\bar{\beta}$  is very large then  $Me^{-\bar{\beta}}$  gets close to zero, which may cause its subgaussian norm to blow up. Fortunately, we know intuitively that the Gibbs measure prefers  $A$  to be large enough to support  $V \setminus \Gamma_{\leq d^*-1} \setminus A$ , i.e.,  $\log |\mathbf{N}_A(v)|$  is likely to be of order  $\Omega(\log \log n)$ . Thus, our trick is to replace  $Me^{-\bar{\beta}}$  with larger value when it is too small, large enough to achieve strong concentration but strictly smaller in order than  $\log n$ .

**Lemma 50** *Let  $L = Me^{-\bar{\beta}} \vee \sqrt{\log n}$ . Then with probability at least  $1 - o(1)$ ,*

$$\sum_{v \in \bar{V} \setminus \Gamma_{\leq d^*-1} \setminus A} \log(Me^{-\bar{\beta}} + |\mathbf{N}_A(v)|) \leq (n - |A|) \log(L + |A|q) + o(n)$$

*simultaneously for all  $A \in \mathcal{K}(G)$ .*

**Proof** We first condition on  $\Gamma_{\leq d^*-1}$  and  $A$ . We first slightly extend the sum to

$$\sum_{v \in \bar{V} \setminus \Gamma_{\leq d^*-1} \setminus A} \log(Me^{-\bar{\beta}} + |\mathbf{N}_A(v)|) \leq \sum_{v \in V \setminus \Gamma_{\leq d^*-1} \setminus A} \log(L + |\mathbf{N}_A(v)|)$$

so that we can say  $|\mathbf{N}_A(v)|$  are (conditionally) i.i.d. binomial random variables. Letting  $Y = L + |\mathbf{N}_A(v)|$ , by Lemma 100 we have

$$\|\log(L + |\mathbf{N}_A(v)|) - \log(L + |A|q)\|_{\psi_2}^2 \leq \frac{6}{L} \leq \frac{6}{\sqrt{\log n}}.$$

Hence, by Corollary 97, the probability that the sum  $\sum \log(L + |\mathbf{N}_A(v)|)$  is away from its mean by at least  $t$  is bounded by

$$2 \exp\left(-\frac{t^2 \sqrt{\log n}}{432 |V \setminus \Gamma_{\leq d^*-1} \setminus A|}\right) \leq 2 \exp\left(-\frac{t^2 \sqrt{\log n}}{432 n}\right).$$

By setting  $t = n / \log \log n$ , we get

$$2 \exp\left(-\frac{n \sqrt{\log n}}{432 (\log \log n)^2}\right).$$

Now we take the union bound over all possible  $A$ , which gives

$$\begin{aligned} \sum_{v \in V \setminus \Gamma_{\leq d^*-1} \setminus A} \log(L + |\mathbf{N}_A(v)|) &\leq |V \setminus \Gamma_{\leq d^*-1} \setminus A| \mathbb{E}[\log(L + |\mathbf{N}_A(v)|) \mid \Gamma_{\leq d^*-1}, A] + \frac{n}{\log \log n} \\ &\leq (n - |A|) \mathbb{E}[\log(L + |\mathbf{N}_A(v)|) \mid \Gamma_{\leq d^*-1}, A] + \frac{n}{\log \log n} \end{aligned}$$

for all  $A \subseteq \mathcal{K}(G)$  with probability at least

$$1 - 2 \exp \left( -\frac{n\sqrt{\log n}}{432(\log \log n)^2} + n \right) = 1 - o(1).$$

Finally, Jensen's inequality gives

$$\mathbb{E}[\log(L + |\mathbf{N}_A(v)|) \mid \Gamma_{\leq d^*-1}, A] \leq \log(L + \mathbb{E}[|\mathbf{N}_A(v)| \mid \Gamma_{\leq d^*-1}, A]) = \log(L + |A|q)$$

completing the proof. ■

Combining Lemma 49 with Lemma 50 gives the following result.

**Lemma 51** *Let  $L = Me^{-\bar{\beta}} \vee \sqrt{\log n}$ . Then with probability at least  $1 - o(1)$ ,*

$$\log(Z_{G,\beta}|_A) \leq \sum_{v \in A} \log |\text{par}_G(v)| + (n - |A|) \log(L + |A|q) + \bar{\beta}|A| - \bar{\beta}(d^* + 1)n + o(n)$$

*simultaneously for all  $A \in \mathcal{K}(G)$ .*

This also immediately yields a bound for  $Z_{G,\beta}\|_m$  by following an argument similar to the proof of Proposition 47. This is not needed for the proof of Proposition 46 but is useful for our analysis in Section F.1.

**Lemma 52** *Let  $L = Me^{-\bar{\beta}} \vee \sqrt{\log n}$ . Then with probability at least  $1 - o(1)$ ,*

$$\log(Z_{G,\beta}\|_m) \leq \mathcal{H}_G(m) + (n - m) \log(L + mq) + \bar{\beta}m - \bar{\beta}(d^* + 1)n + o(n)$$

*simultaneously for all  $m_0 \leq m \leq N_{d^*}$ .*

**A lower bound.** By the Gibbs variational principle,  $\log(Z_{G,\beta}|_A)$  is lower bounded by

$$\log(Z_{G,\beta}|_A) \geq H_{\mu_{G,\infty}|_A}(T) - \bar{\beta} \mathbb{E}_{\mu_{G,\infty}|_A} \left[ \sum_v \mathbf{d}_T(1, v) \right]. \quad (85)$$

Since  $\mu_{G,\infty}|_A$  is a product measure, we have

$$\begin{aligned} H_{\mu_{G,\infty}|_A}(T) &= \sum_{v \in \bar{V} \setminus \{1\}} H(\text{par}_T(v)) \\ &\geq \sum_{v \in A} \log |\text{par}_G(v)| + \sum_{v \in V \setminus \Gamma_{\leq d^*-1} \setminus A} \log(|\mathbf{N}_A(v)| \vee 1) \end{aligned} \quad (86)$$

The first thing we need is the concentration of

$$\sum_{v \in V \setminus \Gamma_{d^*-1} \setminus A} \log(|\mathbf{N}_A(v)| \vee 1). \quad (87)$$

simultaneously for all  $A \subseteq \Gamma_{d^*}$ .

**Lemma 53** *There is a universal constant  $C > 0$  such that with probability at least  $1 - o(1)$ ,*

$$\sum_{v \in V \setminus \Gamma_{\leq d^*-1} \setminus A} \log(|N_A(v)| \vee 1) \geq |V \setminus \Gamma_{\leq d^*-1} \setminus A| \log(|A|q) - \frac{Cn}{\sqrt{|A|q}}$$

*simultaneously for all  $A \subseteq \Gamma_{d^*}$ .*

**Proof** Condition on  $\Gamma_{d^*-1}$  and  $A$ . By Lemma 101 we have

$$\|\log(|N_A(v)| \vee 1) - \log(|A|q \vee 1)\|_{\psi_2} \leq \frac{C_1}{\sqrt{|A|q}}$$

for some constant  $C_1 > 0$ . By Corollary 97 the probability that the sum  $\sum \log(|N_A(v)| \vee 1)$  is away from its mean by at least  $t > 0$  is bounded by

$$2 \exp\left(-\frac{2t^2|A|q}{C_2^2|V \setminus \Gamma_{\leq d^*-1} \setminus A|}\right) \leq 2 \exp\left(-\frac{2t^2|A|q}{C_2^2n}\right)$$

for some constant  $C_2$ . By setting  $t = C_2n/\sqrt{|A|q}$  the probability above is bounded by  $2e^{-2n}$ . Taking the union bound over all  $A \subseteq \Gamma_{d^*}$  we get

$$\sum_{v \in V \setminus \Gamma_{d^*-1} \setminus A} \log(|N_A(v)| \vee 1) \geq |V \setminus \Gamma_{d^*-1} \setminus A| \mathbb{E}[\log(|\text{par}_A(v)| \vee 1) \mid \Gamma_{\leq d^*-1}, A] - \frac{C_3n}{\sqrt{|A|q}}$$

simultaneously for all  $A$ , with probability at least  $1 - o(1)$ . Now using Lemma 96 we have

$$\begin{aligned} \sum_{v \in V \setminus \Gamma_{d^*-1} \setminus A} \log(|N_A(v)| \vee 1) &\geq |V \setminus \Gamma_{d^*-1} \setminus A| \left( \log(|A|q \vee 1) - \frac{C_1}{\sqrt{|A|q}} \right) - \frac{C_3n}{\sqrt{|A|q}} \\ &\geq |V \setminus \Gamma_{d^*-1} \setminus A| \log(|A|q) - \frac{Cn}{\sqrt{|A|q}} \end{aligned}$$

for some constant  $C > 0$ , which is as desired. ■

Applying Lemma 53 to (86), we get

$$H_{\mu_{G,\infty|A}}(T) \geq \sum_{v \in A} |\text{par}_G(v)| + |V \setminus \Gamma_{\leq d^*-1} \setminus A| \log(|A|q) - \frac{Cn}{\sqrt{|A|q}}.$$

By Proposition 16 we have  $|\Gamma_{\leq d^*-1}| = O_p((nq)^{d^*-1}) = O_p(n/(\log \log n)^2)$  and clearly  $\log(|A|q) = O(\log \log n)$ , so

$$H_{\mu_{G,\infty|A}}(T) \geq \sum_{v \in A} |\text{par}_G(v)| + (n - |A|) \log(|A|q) - \frac{Cn}{\sqrt{|A|q}} - O_p\left(\frac{n}{\log \log n}\right) \quad (88)$$

simultaneously for all  $A \in \mathcal{K}(G)$ .

Now we bound the energy term of (85). We have

$$\begin{aligned} \sum_v d_T(1, v) &\leq d^*|A| + (d^* + 1)(n - |A|) + \sum_{v \in \bar{V} \setminus \Gamma_{\leq d^*+1}^T} d_G(v, A) \\ &\leq d^*|A| + (d^* + 1)(n - |A|) + |V \setminus \Gamma_{\leq d^*+1}^T| + |V \setminus \Gamma_{\leq d^*+2}^T| \text{diam}(\bar{G} \setminus \Gamma_{\leq d^*-1}) \end{aligned} \quad (89)$$

where  $\text{diam}(\bar{G} \setminus \Gamma_{\leq d^*-1})$  is the graph diameter of  $\bar{G} \setminus \Gamma_{\leq d^*-1}$ . Since  $G \setminus \Gamma_{\leq d^*-1}$  is an Erdős–Rényi graph conditioned on  $\Gamma_{\leq d^*-1}$ , Theorem 39 implies that

$$\text{diam}(\bar{G} \setminus \Gamma_{\leq d^*-1}) = O\left(\frac{\log n}{\log \log n}\right) \quad (90)$$

asymptotically almost surely. Now we control  $|V \setminus \Gamma_{\leq d^*+1}^T|$  and  $|V \setminus \Gamma_{\leq d^*+2}^T|$  given that  $|A|$  is reasonably large. Note that by the definition of  $\mu_{G,\infty}|_A$ ,  $\Gamma_{\leq d^*+1}^T$  and  $\Gamma_{\leq d^*+2}^T$  are fixed for any  $T \in \text{supp}(\mu_{G,\infty}|_A)$  as long as  $A$  is fixed. As such, we abuse notation and write  $\Gamma_{\leq d^*+1}^A$  and  $\Gamma_{\leq d^*+2}^A$  to emphasize that they are functions of  $A$  (and  $G$ ). Also, recall the definition of  $m_\ell$  in (74).

**Lemma 54** *With probability at least  $1 - o(1)$ , we have*

$$|V \setminus \Gamma_{\leq d^*+1}^A| = o\left(\frac{n}{\log \log n}\right)$$

and

$$|V \setminus \Gamma_{\leq d^*+2}^A| = o\left(\frac{n}{\log n}\right)$$

simultaneously for all  $A \in \mathcal{K}(G)$  with  $|A| \geq m_\ell$ .

**Proof** It suffices to prove that with high probability the bound simultaneously holds for all  $A$  of the same size  $|A| = \lceil \frac{n}{3(\log \log n)^2} \rceil$ . Conditioned on  $\Gamma_{d^*-1}$  and  $A$ , we have

$$|V \setminus \Gamma_{\leq d^*+1}^A| = \sum_{v \in V \setminus \Gamma_{\leq d^*-1} \setminus A} \mathbf{1}\{\mathbf{N}_A(v) = \emptyset\} \sim \text{Binom}(|V \setminus \Gamma_{d^*-1} \setminus A|, (1 - q)^{|A|}).$$

This is stochastically dominated by  $\text{Binom}(n, e^{-q|A|})$ . For simplicity write  $\mu = ne^{-q|A|}$ . Using Lemma 98, we have

$$\Pr(|V \setminus \Gamma_{\leq d^*+1}^A| \geq \mu + t) \leq \exp\left(t - (\mu + t) \log\left(1 + \frac{t}{\mu}\right)\right).$$

Setting  $\mu + t = \frac{n}{(\log \log n)^2}$  which is greater than  $\mu$  for large enough  $n$ ,

$$\begin{aligned} (\mu + t) \log\left(1 + \frac{t}{\mu}\right) &= \frac{n}{(\log \log n)^2} \log\left(\frac{e^{q|A|}}{(\log \log n)^2}\right) \\ &= \frac{\alpha_n |A| \log n}{(\log \log n)^2} - \frac{2n \log \log \log n}{(\log \log n)^2} \\ &\geq \frac{\alpha_n n \log n}{3(\log \log n)^4} - o(n). \end{aligned}$$

This gives

$$\Pr(|V \setminus \Gamma_{\leq d^*+1}^A| \geq \mu + t) \leq \exp\left(-\frac{\alpha_n n \log n}{3(\log \log n)^4} + o(n)\right). \quad (91)$$

For  $|V \setminus \Gamma_{\leq d^*+2}^A|$ , we proceed similarly. Under the good event  $|V \setminus \Gamma_{\leq d^*+1}^A| < \mu + t$ ,  $\Gamma_{d^*+1}^A$  has size  $n - o(n)$ , so we may assume  $N_{d^*+1}^A \geq n/2$ . Thus,  $|V \setminus \Gamma_{\leq d^*+2}^A|$  is stochastically dominated by  $\text{Binom}(n, e^{-nq/2})$ . Applying the same Chernoff bound, this time with  $\mu + t = \frac{n}{(\log n)(\log \log n)}$ , it is easy to see

$$\Pr\left(|V \setminus \Gamma_{\leq d^*+2}^A| \geq \frac{n}{(\log n)(\log \log n)}\right) \leq \exp\left(-\Omega\left(\frac{n}{(\log \log n)^2}\right)\right). \quad (92)$$

Now we take the union bound over all  $A$ . Note that

$$\begin{aligned} \binom{N_{d^*}}{\lceil \frac{n}{3(\log \log n)^2} \rceil} &\leq \binom{n}{\lceil \frac{n}{3(\log \log n)^2} \rceil} \\ &\leq \exp\left(nH\left(\frac{1}{(\log \log n)^2}\right)\right) \\ &\leq \exp\left(\frac{4n \log \log \log n}{(\log \log n)^2}\right) \end{aligned} \quad (93)$$

where for the last inequality we used the fact that  $H(p) \leq -2p \log p$  for all  $0 < p < 1/2$ . Hence, taking the union bound gives  $o(1)$  bounds for the tail probabilities (91) and (92).  $\blacksquare$

Now we have a bound

$$\sum_v d_T(1, v) \leq (d^* + 1)n - |A| + o_p\left(\frac{n}{\log \log n}\right) \quad (94)$$

for the energy term (89), simultaneously for all  $A$  with  $|A| \geq m_\ell$ . Combining with (88) yields the following result.

**Lemma 55** *With probability at least  $1 - o(1)$ , we have*

$$\log(Z_{G,\beta}|_A) \geq \sum_{v \in A} \log |\text{par}_G(v)| + (n - |A|) \log(|A|q) + \bar{\beta}|A| - \bar{\beta}(d^* + 1)n - o(n)$$

*simultaneously for all  $A \in \mathcal{K}(G)$  with  $|A| \geq m_\ell$ .*

**Putting the bounds together.** Lemma 51 and Lemma 55 are more than sufficient to prove Proposition 46. For  $L = Me^{-\bar{\beta}} \vee \sqrt{\log n}$  in Lemma 51, we have

$$\log(L + |A|q) = \log(|A|q) + \log\left(1 + \frac{L}{|A|q}\right) \leq \log(|A|q) + \frac{L}{|A|q}.$$

Since  $|A| \geq m_\ell$ , we have

$$|A|q \geq \frac{nq}{3(\log \log n)^2} = \Omega\left(\frac{\log n}{(\log \log n)^2}\right)$$

while, since  $M = O_p(\log n)$ ,

$$L = O_p((\log n)^{(1-\beta)\vee\frac{1}{2}}).$$

Hence, we have

$$\frac{L}{|A|q} = o_p(n).$$

Now applying (61), Lemma 51, and Lemma 55 proves Proposition 46.

### E.3.6. CONCENTRATION OF THE LOG PARTITION FUNCTION

The formulae in Proposition 46 and Proposition 47 involve  $\mathcal{H}_G(m)$  which heavily depends on  $G$ . We prove that this concentrates uniformly over all  $m_0 \leq m \leq N_{d^*}$ , which allows us to replace it with  $\Psi(m; N_{d^*}, \lambda_n)$  which depends on  $G$  only through  $N_{d^*}$ . The goal of this section is to prove the following lemma.

**Lemma 56** *With probability at least  $1 - o(1)$ , we have*

$$\mathcal{H}_G(m) = \Psi(m; N_{d^*}, \lambda_n) + o(n)$$

*simultaneously for all  $m_0 \leq m \leq N_{d^*}$ .*

Revisit (81) and (66) for definitions. We repeat the definition of  $\mathcal{H}_G(m)$ :

$$\mathcal{H}_G(m) := \log \left( \sum_{A \in \mathcal{K}_m(G)} \prod_{v \in A} |\text{par}_G(v)| \right).$$

Here,  $A$  ranges over all kernels in  $\mathcal{K}_m(G)$ , but the collection  $\mathcal{K}_m(G)$  is a bit unwieldy. Our first task is to replace this with the collection of all  $m$ -elements subsets which we temporarily write

$$\tilde{\mathcal{H}}_G(m) := \log \left( \sum_{A \in \binom{\Gamma_{d^*}}{m}} \prod_{v \in A} |\text{par}_G(v)| \right).$$

**Lemma 57** *With probability at least  $1 - o(1)$ , we have*

$$\tilde{\mathcal{H}}_G(m) = \mathcal{H}_G(m) + o(n)$$

*simultaneously for all  $m_0 \leq m \leq N_{d^*}$ .*

**Proof** We let  $J \subseteq \bar{V} \setminus \Gamma_{\leq d^*-1}$  be the set of vertices outside the giant component of  $\bar{G} \setminus \Gamma_{\leq d^*-1}$ , which has cardinality  $|J| = n^{1-\Theta(1)}$  with high probability. Also, let  $M$  be the maximum degree of  $G$ .

If  $m \leq |J| = n^{1-\Theta(1)}$ , then  $\tilde{\mathcal{H}}_G(m) = o_p(n)$  so the result is trivial. Otherwise, if  $m > |J|$ , then any subset of  $\Gamma_{d^*}$  containing  $J$  is a kernel, so we have

$$\exp(\tilde{\mathcal{H}}_G(m) - \mathcal{H}_G(m)) \leq \frac{\sum_A \prod_{v \in A} |\text{par}_G(v)|}{\sum_{A \supseteq J} \prod_{v \in A} |\text{par}_G(v)|}.$$

For the denominator, we use a simple bound  $|\text{par}_G(v)| \geq 1$  to obtain

$$\sum_{\substack{J \subseteq A \subseteq \Gamma_{d^*} \\ |A|=m}} \prod_{v \in A} |\text{par}_G(v)| \geq \sum_{\substack{B \subseteq \Gamma_{d^*} \setminus J \\ |B|=m-|J|}} \prod_{v \in B} |\text{par}_G(v)|.$$

Now we handle the numerator. The idea is to factor out  $(m - |J|)$  sized subsets of  $A$  to cancel the numerator, and bound the remaining terms by  $M$ . For this, we bound each product as

$$\begin{aligned} \prod_{v \in A} |\text{par}_G(v)| &\leq M^{|A \cap J|} \prod_{v \in A \setminus J} |\text{par}_G(v)| \\ &\leq M^{|A \cap J|} \left( \frac{1}{\binom{|A \setminus J|}{m-|J|}} \sum_{\substack{B \subseteq A \setminus J \\ |B|=m-|J|}} \prod_{v \in B} |\text{par}_G(v)| \cdot M^{|J|-|A \cap J|} \right) \\ &= \frac{M^{|J|}}{\binom{|A \setminus J|}{m-|J|}} \sum_{\substack{B \subseteq A \setminus J \\ |B|=m-|J|}} \prod_{v \in B} |\text{par}_G(v)|. \end{aligned}$$

Now we take the sum over all  $A \subseteq \Gamma_{d^*}$ . By the double counting argument the multiplicative factor of each  $B \subseteq \Gamma_{d^*} \setminus J$  with size  $|B| = m - |J|$  is

$$M^{|J|} \sum_{k=0}^{|J|} \frac{\binom{|J|}{k} \binom{n-m}{|J|-k}}{\binom{m-k}{m-|J|}}$$

where  $k$  encodes  $|A \cap J|$ . Combining with the numerator bound, we have

$$\exp(\tilde{\mathcal{H}}_G(m) - \mathcal{H}_G(m)) \leq M^{|J|} \sum_{k=0}^{|J|} \frac{\binom{|J|}{k} \binom{n-m}{|J|-k}}{\binom{m-k}{m-|J|}} \leq M^{|J|} |J| n^{|J|}.$$

Since  $|J| = n^{1-\Theta(1)}$  the result follows. ■

Next, we show that  $\tilde{\mathcal{H}}_G(m)$  concentrates around  $\Psi(m; N_{d^*}, \lambda_n)$ . Throughout this part, we treat  $\Gamma_{\leq d^*-1}$  as fixed unless otherwise stated. We first construct random variables that have the same distribution as  $|\text{par}_G(v)|$  and set up notations to better describe  $\tilde{\mathcal{H}}_G(m)$  in terms of these random variables. Let  $n^* = |V \setminus \Gamma_{\leq d^*-1}|$  and consider i.i.d. random variables

$$X_1, \dots, X_{n^*} \sim \text{Binom}(N_{d^*-1}, q).$$

We also define an index set  $I_{\mathbf{X} > 0}$  as a function of  $\mathbf{X} = (X_1, \dots, X_{n^*})$  by

$$I_{\mathbf{X} > 0} := \{i \in [n^*] : X_i > 0\}$$

and its size

$$N_{\mathbf{X} > 0} := |I_{\mathbf{X} > 0}|.$$

Note that

$$(|\text{par}_G(v)|)_{v \in \Gamma_{d^*}} \stackrel{d}{=} (X_i)_{i \in I_{\mathbf{X} > 0}}.$$

We also define  $Y_i = \log(X_i \vee 1)$ . For fixed  $N_{d^*} = N_{\mathbf{X}>0}$ , we have

$$\begin{aligned}\tilde{\mathcal{H}}_G(m) &\stackrel{d}{=} \log \left( \sum_{J \in \binom{I_{\mathbf{X}>0}}{m}} \prod_{i \in J} e^{Y_i} \right) \\ &= \log \left( \sum_{J \in \binom{I_{\mathbf{X}>0}}{m}} \exp \left( \sum_{i \in J} Y_i \right) \right) \\ &= \text{LSE}_m(\mathbf{Y}_{\mathbf{X}>0})\end{aligned}\tag{95}$$

where  $\text{LSE}_m$  is the LogSumExp function applied over subset sums of  $m$  elements, and  $\mathbf{Y}_{\mathbf{X}>0} := (Y_i)_{i \in I_{\mathbf{X}>0}}$ . Note that (95) holds jointly, i.e.,

$$\left( \tilde{\mathcal{H}}_G(m) \right)_{m=1}^{N_{d^*}} \stackrel{d}{=} (\text{LSE}_m(\mathbf{Y}_{\mathbf{X}>0}))_{m=1}^{N_{\mathbf{X}>0}}\tag{96}$$

where  $\mathbf{Y}_{\mathbf{X}>0} := (Y_i)_{i \in I_{\mathbf{X}>0}}$ . Necessary properties of the  $m$ -wise LogSumExp function can be found in Appendix H.4.

For the sake of convenience, we introduce several shorthand notations. First, we define

$$\phi_{n,p} := \text{law}(\log(X \vee 1)), \quad X \sim \text{Binom}(n, p)$$

which will be used to describe the law of  $Y_i$ . Also, we define

$$\pi_\lambda := \text{law}(\log(X \vee 1)), \quad X \sim \text{Poisson}(\lambda)$$

and

$$\tilde{\pi}_{p,\lambda} := \text{law}(\log(X \vee 1)), \quad X \sim p\delta_0 + (1-p)\text{ZTP}(\lambda)$$

which are useful for  $\Psi(m; N_{d^*}, \lambda_n)$ . Here,  $\tilde{\pi}_{p,\lambda}$  can be understood as a “patched” version of  $\pi_\lambda$ , where the mass at zero has been adjusted. In particular,  $\pi_\lambda = \tilde{\pi}_{p,\lambda}$  if  $p = e^{-\lambda}$ .

By Lemma 108,  $\tilde{\pi}_{0,\lambda}$  has a bounded subgaussian norm regardless of  $\lambda$ , so Lemma 112 implies that there is a universal constant  $C$  such that

$$\begin{aligned}|\text{LSE}_m(\mathbf{Y}_{\mathbf{X}>0}) - \Psi(m; N_{\mathbf{X}>0}, \lambda_n)| &\leq N_{\mathbf{X}>0} W_1(\tilde{\pi}_{0,\log(1/\lambda_n)}, \mu_{\mathbf{Y}_{\mathbf{X}>0}}) + C\sqrt{N_{\mathbf{X}>0}} \\ &= n^* W_1(\tilde{\pi}_{p_{\mathbf{X}=0}, \log(1/\lambda_n)}, \mu_{\mathbf{Y}}) + C\sqrt{N_{\mathbf{X}>0}}\end{aligned}\tag{97}$$

where

$$p_{\mathbf{X}=0} := \frac{N_{\mathbf{X}=0}}{n^*} = 1 - \frac{N_{\mathbf{X}>0}}{n^*}.$$

We can decompose the RHS as

$$n^* W_1(\tilde{\pi}_{p_{\mathbf{X}=0}, \log(1/\lambda_n)}, \mu_{\mathbf{Y}}) = \underbrace{n^* W_1(\tilde{\pi}_{p_{\mathbf{X}=0}, \log(1/\lambda_n)}, \pi_{\log(1/\lambda_n)})}_{(A)} + n^* W_1(\pi_{\log(1/\lambda_n)}, \mu_{\mathbf{Y}}).\tag{98}$$

The second term is further decomposed into the following.

$$n^* W_1(\pi_{\log(1/\lambda_n)}, \mu_{\mathbf{Y}}) \leq \underbrace{n^* W_1(\pi_{\log(1/\lambda_n)}, \pi_{qN_{d^*-1}})}_{(B)} + \underbrace{n^* W_1(\pi_{qN_{d^*-1}}, \phi_{N_{d^*-1}, q})}_{(C)} + \underbrace{n^* W_1(\phi_{N_{d^*-1}, q}, \mu_{\mathbf{Y}})}_{(D)}.$$

There are four terms we need to bound. We proceed in reverse order. For (D), Lemma 101 and Lemma 111 suffice:

$$(D) \leq O_p(\sqrt{n^*}) = O_p(\sqrt{n}).$$

For (C) we use Theorem 109 and the fact that  $x \mapsto \log(x \vee 1)$  is 1-Lipschitz:

$$\begin{aligned} (C) &\leq n^* W_1(\text{Binom}(N_{d^*-1}, q), \text{Poisson}(qN_{d^*-1})) \\ &= O\left(n^* \sqrt{N_{d^*-1}} q^{3/2}\right) \\ &= O\left((nq)^{3/2}\right) \\ &= O((\log n)^{3/2}). \end{aligned}$$

To deal with (B) we use the following lemma.

**Lemma 58** *For  $0 < \lambda_1 < \lambda_2$  we have*

$$W_1(\pi_{\lambda_1}, \pi_{\lambda_2}) \leq \frac{2(\lambda_2 - \lambda_1)}{\lambda_1} \cdot (1 - e^{-\lambda_1}).$$

**Proof** For  $X \sim \text{Poisson}(\lambda_1)$  and  $Y \sim \text{Poisson}(\lambda_2)$ , we consider  $Z \sim \text{Poisson}(\lambda_2 - \lambda_1)$  independent of  $X$  and construct a coupling of  $X$  and  $Y$  by  $Y = X + Z$ . Then

$$\begin{aligned} W_1(\pi_{\lambda_1}, \pi_{\lambda_2}) &\leq \mathbb{E}[\log(Y \vee 1) - \log(X \vee 1)] \\ &= \mathbb{E}\left[\log\left(\frac{(X + Z) \vee 1}{X \vee 1}\right)\right] \\ &\leq \mathbb{E}\left[\log\left(1 + \frac{Z}{X \vee 1}\right)\right] \\ &\leq \mathbb{E}\left[\frac{Z}{X \vee 1}\right] \\ &\leq 2(\lambda_2 - \lambda_1) \mathbb{E}\left[\frac{1}{X + 1}\right] \\ &= \frac{2(\lambda_2 - \lambda_1)}{\lambda_1} \cdot (1 - e^{-\lambda_1}). \end{aligned}$$

■

Using this we have

$$\begin{aligned} (B) &\leq n^* \cdot \frac{2|qN_{d^*-1} - \log(1/\lambda_n)|}{qN_{d^*-1} \wedge \log(1/\lambda_n)} \\ &= n^* \cdot \frac{2|N_{d^*-1} - (nq)^{d^*-1}|}{N_{d^*-1} \wedge (nq)^{d^*-1}} \\ &= O_p\left(\frac{n}{\log \log n}\right) \end{aligned}$$

where the last line is due to Proposition 16. Now we are left with (A). We first establish the following lemma.

**Lemma 59** *We have*

$$W_1(p\delta_0 + (1-p) \text{ZTP}(\lambda), \text{Poisson}(\lambda)) = \frac{\lambda|p - e^{-\lambda}|}{1 - e^{-\lambda}}.$$

**Proof** Let  $X \sim p\delta_0 + (1-p) \text{ZTP}(\lambda)$  and  $Y \sim \text{Poisson}(\lambda)$ . For an integer  $x \geq 0$ , we have

$$\begin{aligned} F_X(x) &:= \Pr(X \leq x) = p + \sum_{i=1}^x \Pr(X = i) \\ &= p + \frac{1-p}{1-e^{-\lambda}} \sum_{i=1}^x \Pr(Y = i) \\ &= p + \frac{1-p}{1-e^{-\lambda}} (F_Y(x) - e^{-\lambda}) \\ &= \frac{1-p}{1-e^{-\lambda}} F_Y(x) + \frac{p-e^{-\lambda}}{1-e^{-\lambda}} \end{aligned}$$

This gives

$$\begin{aligned} W_1(p\delta_0 + (1-p) \text{ZTP}(\lambda), \text{Poisson}(\lambda)) &= \sum_{x=0}^{\infty} |F_X(x) - F_Y(x)| \\ &= \sum_{x=0}^{\infty} \frac{|p - e^{-\lambda}|}{1 - e^{-\lambda}} (1 - F_Y(x)) \\ &= \frac{|p - e^{-\lambda}|}{1 - e^{-\lambda}} \mathbb{E}[Y] \\ &= \frac{\lambda|p - e^{-\lambda}|}{1 - e^{-\lambda}}. \end{aligned}$$

■

Since  $x \mapsto \log(x \vee 1)$  is 1-Lipschitz, Lemma 59 implies

$$\begin{aligned} (\text{A}) &\leq n^* \cdot \frac{\log(1/\lambda_n)}{1 - \lambda_n} |p_{\mathbf{X}=0} - \lambda_n| \\ &\leq n^* \cdot \frac{\log(1/\lambda_n)}{1 - \lambda_n} (|p_{\mathbf{X}=0} - (1-q)^{N_{d^*-1}}| + |(1-q)^{N_{d^*-1}} - e^{-qN_{d^*-1}}| + |e^{-qN_{d^*-1}} - \lambda_n|) \\ &= \underbrace{\frac{\log(1/\lambda_n)}{1 - \lambda_n} |N_{\mathbf{X}=0} - n^*(1-q)^{N_{d^*-1}}|}_{(\text{A1})} + \underbrace{\frac{n^* \log(1/\lambda_n)}{1 - \lambda_n} |(1-q)^{N_{d^*-1}} - e^{-qN_{d^*-1}}|}_{(\text{A2})} \\ &\quad + \underbrace{\frac{n^* \log(1/\lambda_n)}{1 - \lambda_n} |e^{-qN_{d^*-1}} - \lambda_n|}_{(\text{A3})}. \end{aligned}$$

This again consists of three terms which we bound separately. Before that, we note that

$$\frac{\log(1/\lambda_n)}{1 - \lambda_n} = O\left(\frac{\log n}{(\log \log n)^2}\right)$$

which can be derived from (24). For (A1), recall that

$$N_{\mathbf{X}=0} \sim \text{Binom}(n^*, (1-q)^{N_{d^*-1}})$$

which gives

$$(A1) = \frac{\log(1/\lambda_n)}{1-\lambda_n} \cdot O(\sqrt{n}) = O\left(\frac{\sqrt{n} \log n}{(\log \log n)^2}\right).$$

For (A2), we utilize an inequality

$$e^{-np} - (1-p)^n \leq \frac{1}{2}np^2$$

which holds for all  $n \geq 0$  and  $0 \leq p \leq 1$ . This can be verified by taking derivative of  $e^{-np} - (1-p)^n - \frac{1}{2}np^2$  by  $p$  and showing that it is decreasing. We have

$$\begin{aligned} (A2) &\leq \frac{n^* \log(1/\lambda_n)}{1-\lambda_n} \cdot \frac{1}{2} N_{d^*-1} q^2 \\ &\leq \frac{1}{2} \cdot \frac{\log(1/\lambda_n)}{1-\lambda_n} \cdot (nq)^2 \\ &= O\left(\frac{(\log n)^3}{(\log \log n)^2}\right). \end{aligned}$$

Finally, for (A3), we use

$$\frac{e^x - 1}{x} \geq e^{x/2}$$

which holds for all  $x > 0$ . Thus,

$$\frac{\log(1/\lambda_n)}{1-\lambda_n} \geq \lambda_n^{-1/2}.$$

This gives

$$\begin{aligned} (A3) &\leq n^* \left| \lambda_n^{-1/2} e^{-qN_{d^*-1}} - \lambda_n^{1/2} \right| \\ &= n^* \left| \lambda_n^{\frac{qN_{d^*-1}}{\log(1/\lambda_n)} - \frac{1}{2}} - \lambda_n^{1/2} \right|. \end{aligned}$$

To further upper bound this, we note that for  $t > 0$  and  $0 \leq x \leq 1$  the function  $|x^{t-1/2} - x^{1/2}|$  attains its maximum at  $x^* = (2t-1)^{-1/(t-1)}$  (other than the trivial edge case  $t = 1$ ) which implies

$$|x^{t-1/2} - x^{1/2}| \leq \frac{2|t-1|}{2t-1} (x^*)^{1/2} \leq \frac{2|t-1|}{2t-1}.$$

In our case,

$$\frac{qN_{d^*-1}}{\log(1/\lambda_n)} = \frac{N_{d^*-1}}{(nq)^{d^*-1}} = 1 \pm O_p\left(\frac{1}{\log \log n}\right)$$

by Proposition 16, which gives

$$(A3) = O_p\left(\frac{n}{\log \log n}\right).$$

As a result, we have proved that (97) is  $o_p(n)$  simultaneously for all  $m$ . Going back to (96), this is equivalent to the following result.

**Lemma 60** *With probability at least  $1 - o(1)$ , we have*

$$\tilde{\mathcal{H}}_G(m) = \Psi(m; N_{d^*}, \lambda_n) + o(n)$$

*simultaneously for all  $1 \leq m \leq N_{d^*}$ .*

Combining this with Lemma 57 completes the proof of Lemma 56. Note that Lemma 56 also yields the following version of Proposition 47.

**Corollary 61** *Let  $\beta > 0$  be a constant. Then with probability at least  $1 - o(1)$ , we have*

$$\log(Z_{G,\beta} \| m) = \tilde{\Phi}_{G,\beta}(m) + o(n)$$

*simultaneously for all  $m_\ell \leq m \leq N_{d^*}$ .*

### E.3.7. PROOF OF THEOREM 41

We show that the measures (82) (for  $m < m_0$ , set them arbitrary) satisfy the desired properties. Recall that by Lemma 52 we have an upper bound of  $\log(Z_{G,\beta} \| m)$  for  $m_0 \leq m \leq N_{d^*}$ :

$$\Psi(m; N_{d^*}, \lambda_n) + (n - m) \log(L + mq) + \bar{\beta}m. \quad (99)$$

We claim that it suffices to prove that:

- (a)  $\Psi(m; N_{d^*}, \lambda_n)$  is an increasing function of  $m$  for  $m \leq m_\ell$  with high probability,
- (b)  $(n - m) \log(L + mq)$  is also an increasing function of  $m$  for  $m \leq m_\ell$  with high probability, and
- (c)  $f(m) = (n - m) \log(mq)$  is also an increasing function of  $m$  for  $m \leq m_\ell$  with high probability and there exists an integer  $m^\# \in [m_\ell, N_{d^*}]$  as a function of  $n$  such that  $f(m^\#) - f(m_\ell) = \Omega_p(n)$ .

If we have these, then we get the following facts:

- By (a) and (b), the maximum value of (99) is attained at  $m \geq m_\ell$  with high probability, at which the bound is tight by Corollary 61. This implies that

$$\log Z_{G,\beta} = \max_{m_\ell \leq m \leq N_{d^*}} \log(Z_{G,\beta} \| m) + o_p(n). \quad (100)$$

- By (a) and (c), any  $m^*$  satisfying the assumption of the theorem has  $m^* \geq m_\ell$  with high probability, in which case the objective (67) agrees with (99) ignoring  $o_p(n)$  terms. Thus,  $m^*$  is an approximate solution to the maximization in (100), yielding

$$\log Z_{G,\beta} = \log(Z_{G,\beta} \| m^*)$$

which is enough to prove the theorem, similar to Proposition 47.

Note that we may assume

$$m_\ell = \left\lceil \frac{n}{3(\log \log n)^2} \right\rceil$$

since  $m_0$  is by definition the number of connected components of  $\overline{G} \setminus \Gamma_{\leq d^*-1}$  which is  $n^{1-\Theta(1)}$  with high probability. We first prove (a) that  $\Psi(m; N_{d^*}, \lambda_n)$  is non-decreasing if  $m \leq m_\ell$ . By Lemma 18 and the definition of  $d^*$ ,  $N_{d^*}$  concentrates around  $(nq)^{d^*} \geq \frac{n}{(\log \log n)^2}$  when it is the smallest. Thus, we have that

$$m_\ell = \left\lceil \frac{n}{3(\log \log n)^2} \right\rceil < \frac{N_{d^*}}{2}$$

asymptotically almost surely. As a consequence, for any  $m < m_\ell$  and  $x_1, \dots, x_{N_{d^*}} \geq 1$ , we have

$$\begin{aligned} \sum_{I \in \binom{[N_{d^*}]}{m}} \prod_{i \in I} x_i &= \frac{1}{N_{d^*} - m} \sum_{I \in \binom{[N_{d^*}]}{m+1}} \sum_{J \in \binom{I}{m}} \prod_{i \in J} x_i \\ &\leq \frac{m+1}{N_{d^*} - m} \sum_{I \in \binom{[N_{d^*}]}{m+1}} \prod_{i \in I} x_i \end{aligned}$$

which implies that  $\Psi(m; N_{d^*}, \lambda_n) \leq \Psi(m+1; N_{d^*}, \lambda_n)$  simultaneously for all  $m < m_\ell$ , asymptotically almost surely.

Now we prove (b). To see that  $g(x) = (n-x) \log(L+qx)$  is non-decreasing, observe that

$$g'(x) = -\log(L+qx) + \frac{q(n-x)}{L+qx}.$$

Using that  $L+qx \leq L+m_\ell q = O_p(\log n / (\log \log n)^2)$ , we get

$$\log(L+qx) = O_p(\log \log n)$$

and

$$\frac{q(n-x)}{L+qx} = \Omega_p((\log \log n)^2)$$

which imply  $g'(x) > 0$  uniformly over all  $x < m_\ell$  for sufficiently large  $n$ , with high probability.

The first part of (c) is similar to (b). For the second part, we may take

$$m^\# = \left\lceil \frac{n}{2(\log \log n)^2} \right\rceil \wedge N_{d^*}.$$

By the same argument with the previous paragraph, we may assume  $m^\# = \left\lceil \frac{n}{2(\log \log n)^2} \right\rceil$ . Then it is straightforward to see that

$$f(m^\#) - f(m_\ell) \geq \log(3/2)n - o_p(n)$$

as desired. This concludes the proof of Theorem 41.

## E.3.8. PROOF OF THEOREM 42

In this section, we prove Theorem 42 by solving the optimization (68), which we re-state here for the reader's convenience:

$$\tilde{\Phi}_{G,\beta}^* := \max_{m_0 \leq m \leq N_{d^*}} \left\{ \Psi(m; N_{d^*}, \lambda_n) + (n - m) \log(mq) + \bar{\beta}m \right\} - \bar{\beta}(d^* + 1)n. \quad (101)$$

The first thing we need to handle is the mysterious function  $\Psi(m; N_{d^*}, \lambda_n)$ . Conditioned on  $N_{d^*}$ , consider i.i.d. random variables  $X_1, \dots, X_{N_{d^*}} \sim \text{ZTP}(\log(1/\lambda_n))$ . Recall that by definition

$$\begin{aligned} \Psi(m; N_{d^*}, \lambda_n) &= \mathbb{E} \left[ \text{LSE}_m(\log X_1, \dots, \log X_{N_{d^*}}) \mid N_{d^*} \right] \\ &= \mathbb{E} \left[ \log \left( \sum_{I \in \binom{[N_{d^*}]}{m}} \exp \left( \sum_{i \in I} \log X_i \right) \right) \mid N_{d^*} \right]. \end{aligned}$$

If the sum  $\sum \log X_i$  enjoys strong concentration, we might be able to replace each of these with  $m\Psi_1(\lambda_n)$  where we define

$$\Psi_1(\lambda_n) := \mathbb{E}[\log X_1], \quad X_1 \sim \text{ZTP}(\log(1/\lambda_n)).$$

This might introduce a large error (up to  $O(n)$ ) if  $\log(1/\lambda_n) = \Theta(1)$  and  $m = \Theta(n)$ . In other cases, namely if  $m$  is either small or close to  $N_{d^*}$ , it is possible to obtain more explicit results by concentration of measure.

**Lemma 62** *There is a universal constant  $C > 0$  such that with probability at least  $1 - o(1)$ ,*

$$\left| \sum_{i \in A} \log X_i - |A| \Psi_1(\lambda_n) \right| \leq C \sqrt{(|A| \wedge (N_{d^*} - |A| + \sqrt{N_{d^*}})) N_{d^*}}$$

*simultaneously for all  $A \subseteq [N_{d^*}]$ .*

**Proof** Define

$$K = \|\log X_1 - \log(\log(1/\lambda_n) \vee 1)\|_{\psi_2}.$$

By Corollary 97 we have for any  $t > 0$  that

$$\Pr \left( \left| \sum_{i=1}^{N_{d^*}} \log X_i - N_{d^*} \Psi_1(\lambda_n) \right| \geq t \right) \leq 2 \exp \left( -\frac{C_1 t^2}{N_{d^*} K^2} \right)$$

for some constant  $C_1 > 0$ . By setting  $t = K N_{d^*}^{3/4}$ , we have

$$\left| \sum_{i=1}^{N_{d^*}} \log X_i - N_{d^*} \Psi_1(\lambda_n) \right| \leq K N_{d^*}^{3/4} \quad (102)$$

with probability at least  $1 - 2e^{-C_1 \sqrt{N_{d^*}}}$ . Similarly, for any  $t > 0$ ,

$$\Pr \left( \left| \sum_{i \in A} \log X_i - |A| \Psi_1(\lambda_n) \right| \geq t \right) \leq 2 \exp \left( -\frac{C_1 t^2}{|A| K^2} \right).$$

By setting  $t = 2K\sqrt{|A|N_{d^*}/C_1}$ , we have

$$\Pr\left(\left|\sum_{i \in A} \log X_i - |A|\Psi_1(\lambda_n)\right| \geq C_2K\sqrt{|A|N_{d^*}}\right) \leq 2e^{-4N_{d^*}}$$

for some constant  $C_2 > 0$ . Taking the union bound over all  $A \subseteq [N_{d^*}]$ , we have that

$$\left|\sum_{i \in A} \log X_i - |A|\Psi_1(\lambda_n)\right| \leq C_2K\sqrt{|A|N_{d^*}} \quad (103)$$

for all  $A \subseteq [N_{d^*}]$ , with probability at least  $1 - 2e^{-3N_{d^*}}$ , conditioned on  $\Gamma_{\leq d^*-1}$ . Now we write

$$\sum_{i \in A} \log X_i = \sum_{i=1}^{N_{d^*}} \log X_i - \sum_{i \notin A} \log X_i.$$

Under the events (102) and (103), we get

$$\left|\sum_{i \in A} \log X_i - |A|\Psi_1(\lambda_n)\right| \leq KN_{d^*}^{3/4} + C_2K\sqrt{(N_{d^*} - |A|)N_{d^*}}.$$

Thus, we conclude that

$$\left|\sum_{i \in A} \log X_i - |A|\Psi_1(\lambda_n)\right| \leq C_3K\sqrt{(|A| \wedge (N_{d^*} - |A| + \sqrt{N_{d^*}}))N_{d^*}}$$

for some constant  $C_3 > 0$ , with probability at least  $1 - 2e^{-C_1\sqrt{N_{d^*}}} - 2e^{-3N_{d^*}} = 1 - o(1)$ . The conclusion now follows since  $K$  is bounded by a universal constant by Lemma 108. ■

**Lemma 63** *There is a universal constant  $C > 0$  such that with probability at least  $1 - o(1)$ ,*

$$\Psi(m; N_{d^*}, \lambda_n) \leq N_{d^*}H(m/N_{d^*}) + m\Psi_1(\lambda_n) + C\sqrt{(m \wedge (N_{d^*} - m + \sqrt{N_{d^*}}))N_{d^*}} + o(n)$$

and

$$\Psi(m; N_{d^*}, \lambda_n) \geq m\Psi_1(\lambda_n) - o(n)$$

simultaneously for all  $1 \leq m \leq N_{d^*}$ .

**Proof** Lemma 62 implies that

$$\begin{aligned} \text{LSE}_m(\log X_1, \dots, \log X_{N_{d^*}}) &\leq \log \binom{N_{d^*}}{m} + m\Psi_1(\lambda_n) + C\sqrt{(m \wedge (N_{d^*} - m + \sqrt{N_{d^*}}))N_{d^*}} \\ &\leq N_{d^*}H(m/N_{d^*}) + m\Psi_1(\lambda_n) + C\sqrt{(m \wedge (N_{d^*} - m + \sqrt{N_{d^*}}))N_{d^*}}. \end{aligned}$$

simultaneously for all  $m$ , with probability  $1 - o(1)$ . By Lemma 112  $\text{LSE}_m$  concentrates around its mean given that the distribution is subgaussian, so this in turn implies

$$\Psi(m; N_{d^*}, \lambda_n) \leq N_{d^*} H(m/N_{d^*}) + m\Psi_1(\lambda_n) + C\sqrt{(m \wedge (N_{d^*} - m + \sqrt{N_{d^*}}))N_{d^*}} + o_p(n)$$

simultaneously for all  $m$ . On the other hand, Jensen's inequality gives

$$\begin{aligned} \text{LSE}_m(\log X_1, \dots, \log X_{N_{d^*}}) &= \log \left( \frac{1}{\binom{N_{d^*}}{m}} \sum_{I \in \binom{[N_{d^*}]}{m}} \exp \left( \sum_{i \in I} \log X_i \right) \right) + \log \binom{N_{d^*}}{m} \\ &\geq \frac{\binom{N_{d^*}-1}{m-1}}{\binom{N_{d^*}}{m}} \sum_{i=1}^{N_{d^*}} \log X_i \\ &= \frac{m}{N_{d^*}} \sum_{i=1}^{N_{d^*}} \log X_i \end{aligned}$$

which implies

$$\Psi(m; N_{d^*}, \lambda_n) \geq m\Psi_1(\lambda_n) - o_p(n).$$

■

Now we prove Theorem 42. Our objective function is

$$f(x) = \Psi(x; N_{d^*}, \lambda_n) + (n - x) \log(qx) + \bar{\beta}x - \bar{\beta}(d^* + 1)n \quad (104)$$

with domain  $\mathcal{X} = \{1, \dots, N_{d^*}\}$ . We also define

$$f_u(x) := g(x) + h(x) \quad (105)$$

where

$$g(x) = \Psi_1(\lambda_n)x + (n - x) \log(qx) + \bar{\beta}x - \bar{\beta}(d^* + 1)n$$

and

$$h(x) = N_{d^*} H(x/N_{d^*}) + C\sqrt{\left(x \wedge (N_{d^*} - x + \sqrt{N_{d^*}})\right) N_{d^*}}.$$

Note that  $f_u(x)$  is a concave function. By Lemma 63 we have with high probability

$$g(x) - o(n) \leq f(x) \leq f_u(x) + o(n) \quad (106)$$

for all  $x \in \mathcal{X}$ . Our proof strategy is as follows: instead of directly optimizing  $f(x)$ , we find a set  $I \subseteq \mathcal{X}$  where  $f_u(x)$  attains its maximum  $f_u(x_u^*)$ , and prove that  $h(x_u^*) = o(n)$ . This implies  $|f(x_u^*) - g(x_u^*)| = o(n)$  so  $g(x_u^*)$  can be a good approximation for the maximum of  $f(x)$  asymptotically almost surely.

Before we start the main proof, we note that Lemma 108 gives

$$|\Psi_1(\lambda_n) - (\Delta_n \vee 0) \log(nq)| = O(1) \quad (107)$$

with high probability. This further implies

$$|\Psi_1(\lambda_n) - \Delta_n \log(nq)| = O(\log \log \log n). \quad (108)$$

**Case 1: low temperature phase.** To deal with the phase (LT), we consider  $x = (1 - \epsilon\kappa_n^{-1})N_{d^*}$  where  $0 \leq \epsilon \leq (\kappa_n \wedge 1)/2$ . Note that

$$\begin{aligned} (1 - \epsilon\kappa_n^{-1})N_{d^*} &= N_{d^*} - \frac{\epsilon N_{d^*}}{(1 - \lambda_n) \log \log n} \\ &= N_{d^*} - \frac{\epsilon n}{\log \log n} - o\left(\frac{n}{\log \log n}\right) \\ &\geq N_{d^*} - \frac{n}{\log \log n} - o\left(\frac{n}{\log \log n}\right). \end{aligned} \quad (109)$$

This readily implies  $h((1 - \epsilon\kappa_n^{-1})N_{d^*}) = o(n)$ . For  $g((1 - \epsilon\kappa_n^{-1})N_{d^*})$  we have

$$\begin{aligned} g(N_{d^*}) - g((1 - \epsilon\kappa_n^{-1})N_{d^*}) &= \epsilon\kappa_n^{-1}N_{d^*}(\Psi_1(\lambda_n) - \log(qN_{d^*}) + \bar{\beta}) - (n - (1 - \epsilon\kappa_n^{-1})N_{d^*}) \log(1 - \epsilon\kappa_n^{-1}) \\ &\geq \epsilon\kappa_n^{-1}N_{d^*}(\Psi_1(\lambda_n) - \log(qN_{d^*}) + \bar{\beta}) - n \log(1 - \epsilon\kappa_n^{-1}) \\ &= \epsilon(\beta - 1 + \Delta_n + \kappa_n^{-1})n - (\epsilon\kappa_n^{-1} + \log(1 - \epsilon\kappa_n^{-1}))n - o(n) \end{aligned} \quad (110)$$

where the  $o(n)$  term does not depend on  $\epsilon$ . Since  $\beta \geq \limsup_{n \rightarrow \infty} (1 - \Delta_n - \kappa_n^{-1})$  implies  $\beta - 1 + \Delta_n + \kappa_n^{-1} \geq -o(1)$ , this gives

$$f_u(N_{d^*}) \geq f_u((1 - \epsilon\kappa_n^{-1})N_{d^*}) - o(n)$$

for all  $\epsilon$ . Now we argue that the maximum of  $f_u(x)$  is attained at some  $\epsilon$ . Since  $f_u(x)$  is concave, it suffices to show that there is some  $0 < \epsilon_0 < \kappa_n \wedge 1$  such that  $f_u(N_{d^*}) > f_u((1 - \epsilon_0\kappa_n^{-1})N_{d^*})$ . We prove that  $\epsilon_0 = (\kappa_n \wedge 1)/2$  satisfies our desired property by further splitting the low temperature phase into the following two regimes.

- Suppose that  $\beta > \limsup_{n \rightarrow \infty} (1 - \Delta_n - \kappa_n^{-1})$ . This implies that there exists a  $\delta > 0$  such that  $\beta - 1 + \Delta_n + \kappa_n^{-1} > \delta$  for all large  $n$ . If  $\kappa_n \geq 1$ , then  $\epsilon_0 = 1/2$  so

$$\epsilon(\beta - 1 + \Delta_n + \kappa_n^{-1})n \geq \frac{\delta}{2}n.$$

Otherwise, if  $\kappa_n < 1$ , then  $\epsilon_0 = \kappa_n/2$  so

$$\begin{aligned} -(\epsilon_0\kappa_n^{-1} + \log(1 - \epsilon_0\kappa_n^{-1}))n &= \left(-\frac{1}{2} + \log 2\right)n \\ &\geq \frac{1}{6}n. \end{aligned}$$

This implies that

$$g(N_{d^*}) - g((1 - \epsilon_0\kappa_n^{-1})N_{d^*}) \geq \left(\frac{\delta}{2} \wedge \frac{1}{6}\right)n - o(n).$$

In particular,  $f_u(N_{d^*}) > f_u((1 - \epsilon_0\kappa_n^{-1})N_{d^*})$  for all sufficiently large  $n$ .

- Otherwise, we have  $\beta = \limsup_{n \rightarrow \infty} (1 - \Delta_n - \kappa_n^{-1})$  and  $\limsup_{n \rightarrow \infty} \kappa_n = K < \infty$  for a constant  $K$ . Then  $\liminf_{n \rightarrow \infty} \kappa_n^{-1} = 1/K$ , so by substituting  $\epsilon$  with  $\epsilon_0 = (\kappa_n \wedge 1)/2$  we have  $\epsilon_0 \kappa_n^{-1} \geq \frac{1}{2} \wedge \frac{1}{4K}$  which implies that  $-\epsilon_0 \kappa_n^{-1} - \log(1 - \epsilon_0 \kappa_n^{-1})$  is bounded away from zero. Thus, similar to the preceding regime, we get

$$g(N_{d^*}) - g((1 - \epsilon_0 \kappa_n^{-1})N_{d^*}) = \Omega(n)$$

and  $f_u(N_{d^*}) > f_u((1 - \epsilon_0 \kappa_n^{-1})N_{d^*})$  for all large  $n$ .

Therefore, we have proved that  $\max_{x \in \mathcal{X}} f(x) = f_u(N_{d^*}) + o_p(n)$ . Since  $N_{d^*} \Psi_1(\lambda_n) = n \Psi_0(\lambda_n) + o_p(n)$ , this gives the formula

$$\Phi_{G,\beta} = n \Psi_0(\lambda_n) + \lambda_n n \log(\alpha_n(1 - \lambda_n) \log n) - \bar{\beta}(d^* + \lambda_n)n \pm o_p(n).$$

**Case 2: high temperature phase.** In the high temperature phase (HT), we can pick a constant  $\delta > 0$  such that  $\beta < 1 - \Delta_n - \kappa_n^{-1} - \delta$  for all sufficiently large  $n$ . This also implies  $\kappa_n > 1$ . Here, we consider  $x = (1 + \epsilon) \frac{\kappa_n^{-1} N_{d^*}}{1 - \Delta_n - \beta}$  for  $-1/2 \leq \epsilon \leq \delta$ . Note that

$$(1 + \delta) \frac{\kappa_n^{-1} N_{d^*}}{1 - \Delta_n - \beta} < \frac{(\kappa_n^{-1} + \delta \kappa_n^{-1}) N_{d^*}}{\kappa_n^{-1} + \delta} < N_{d^*}.$$

Also, we have

$$\begin{aligned} \frac{\kappa_n^{-1} N_{d^*}}{1 - \Delta_n - \beta} &= \frac{N_{d^*}}{(1 - \Delta_n - \beta)(1 - \lambda_n) \log \log n} \\ &= \frac{n}{(1 - \Delta_n - \beta) \log \log n} + o\left(\frac{n}{\log \log n}\right). \end{aligned} \tag{111}$$

This gives  $h\left((1 + \epsilon) \frac{\kappa_n^{-1} N_{d^*}}{1 - \Delta_n - \beta}\right) = o(n)$ . Moreover,

$$\begin{aligned} &g\left(\frac{\kappa_n^{-1} N_{d^*}}{1 - \Delta_n - \beta}\right) - g\left((1 + \epsilon) \frac{\kappa_n^{-1} N_{d^*}}{1 - \Delta_n - \beta}\right) \\ &= \frac{\epsilon \kappa_n^{-1} N_{d^*}}{1 - \Delta_n - \beta} \left( \Psi_1(\lambda_n) - \log\left(\frac{\kappa_n^{-1} N_{d^*} q}{1 - \Delta_n - \beta}\right) + \bar{\beta} \right) - \left( n - \frac{(1 + \epsilon) \kappa_n^{-1} N_{d^*}}{1 - \Delta_n - \beta} \right) \log(1 + \epsilon) \\ &\geq \frac{\epsilon \kappa_n^{-1} N_{d^*}}{1 - \Delta_n - \beta} \left( \Psi_1(\lambda_n) - \log\left(\frac{\kappa_n^{-1} N_{d^*} q}{1 - \Delta_n - \beta}\right) + \bar{\beta} \right) - n \log(1 + \epsilon) \\ &= (\epsilon - \log(1 + \epsilon))n - o(n). \end{aligned} \tag{112}$$

This implies

$$f_u\left(\frac{\kappa_n^{-1} N_{d^*}}{1 - \Delta_n - \beta}\right) \geq f_u\left((1 + \epsilon) \frac{\kappa_n^{-1} N_{d^*}}{1 - \Delta_n - \beta}\right) - o(n)$$

for all  $\epsilon$ . Also, by plugging in  $\epsilon = \epsilon_1 = -1/2$  or  $\epsilon = \epsilon_2 = \delta$ , we have for sufficiently large  $n$

$$\begin{aligned} f_u\left(\frac{\kappa_n^{-1} N_{d^*}}{1 - \Delta_n - \beta}\right) &\geq f_u\left((1 + \epsilon_i) \frac{\kappa_n^{-1} N_{d^*}}{1 - \Delta_n - \beta}\right) + \Omega(n) \\ &> f_u\left((1 + \epsilon_i) \frac{\kappa_n^{-1} N_{d^*}}{1 - \Delta_n - \beta}\right) \end{aligned}$$

for either  $\epsilon_i$ , so by the concavity of  $f_u$  it follows that the maximum is attained in  $(\epsilon_1, \epsilon_2)$ . Therefore, an approximate maximum can be obtained by  $\epsilon = 0$  leading to

$$\Phi_{G,\beta} = n \log \left( \frac{\alpha_n \log n}{(1 - \Delta_n - \beta) \log \log n} \right) - n - \bar{\beta}(d^* + 1)n \pm o_p(n).$$

## Appendix F. Analysis of Gibbs measures over trees

Building on top of our analysis of the log partition function  $\log Z_{G,\beta}$ , we continue on analyzing the Gibbs measures  $\mu_{G,\beta}$ . We show the existence of different forms of phase transitions in the system by analyzing the trajectory of  $\mu_{G,\beta}$  in the space of probability measures equipped with 1-Wasserstein metric. We also compute several thermodynamic quantities such as free energy density and Franz–Parisi potential, which let us give a picture of the optimization landscape from a statistical physics perspective.

**Entropy and transportation.** We will make use of *entropy-transport* inequalities, specifically:

**Theorem 64 (Corollary 4.16, Example 4.17 of [Van Handel \(2014\)](#))** *Suppose that  $\mu_i$  is a probability measure on metric space  $(\mathcal{X}_i, d_i)$  for  $i = 1$  to  $n$  and that for some  $\sigma > 0$ , each satisfies the  $W_1$ -entropy transport inequality  $W_1(\nu, \mu_i) \leq \sqrt{2\sigma^2 \text{KL}(\nu, \mu_i)}$  for all  $\nu$ . Then for all probability measures  $\nu$  on  $\bigotimes_i \mathcal{X}_i$  and letting  $\mu = \bigotimes_i \mu_i$ , we have<sup>13</sup>*

$$W_1(\nu, \mu) \leq \sqrt{2\sigma^2 n \text{KL}(\nu, \mu)}$$

where we equip  $\bigotimes_i \mathcal{X}_i$  with the metric  $d(x, y) = \sum_i d_i(x_i, y_i)$ .

In particular, by combining the above with Pinsker’s inequality, this holds with  $\sigma^2 = 1$  if  $d_i(x, y) = 1(x \neq y)$ , in which case  $d$  is the Hamming metric.

**Corollary 65 (High entropy  $\implies$  Wasserstein closeness to the uniform measure)** *Let  $\mathcal{X}_1, \dots, \mathcal{X}_n$  be finite sets and write*

$$S = \prod_{i=1}^n \mathcal{X}_i, \quad \mu = \bigotimes_{i=1}^n \mu_i,$$

where each  $\mu_i$  is the uniform measure on  $\mathcal{X}_i$ . Equip  $S$  with the (unnormalised) Hamming metric

$$d(x, y) = \sum_{i=1}^n \mathbf{1}(x_i \neq y_i), \quad x = (x_1, \dots, x_n), y = (y_1, \dots, y_n) \in S.$$

For any probability measure  $\nu$  on  $S$ , such that  $\nu$  has  $\varepsilon$ -near-maximal entropy, i.e.,

$$H_\nu(X) \geq \log |S| - \varepsilon \quad (\varepsilon > 0),$$

we have

$$W_1(\nu, \mu) \leq \sqrt{2n\varepsilon}.$$

13. We use a different normalization convention for the Hamming metric compared to [Van Handel \(2014\)](#).

**Proof** Because each  $\mu_i$  is uniform on the finite set  $\mathcal{X}_i$ , it satisfies the  $W_1$ -entropy transport inequality with parameter  $\sigma^2 = 1$  under the 0–1 metric  $d_i(x, y) = \mathbf{1}(x \neq y)$  (cf. the remark following Theorem 64). Applying Theorem 64 to the product space  $(S, d)$  therefore gives

$$W_1(\nu, \mu) \leq \sqrt{2n \text{KL}(\nu, \mu)}.$$

For the uniform measure  $\mu$  on  $S$ , we recall the standard identity  $\text{KL}(\nu, \mu) = \mathbb{E}_\nu \left[ \log \frac{\nu}{\mu} \right] = \log |S| - H_\nu(X)$ . Substituting this into the previous inequality yields the result.  $\blacksquare$

### F.1. Structure of the Gibbs measures

In Section E.3, we have constructed measures  $\hat{\mu}_{G,m}$  that approximates  $\mu_{G,\beta}$  in KL divergence. Built on these results, we analyze the geometric structure of a typical spanning tree under the Gibbs measures, with respect to the 1-Wasserstein metric. Our main goal is to prove Theorem 73 and Theorem 75, which state that the Gibbs measures are Wasserstein close to the uniform measure over the set of shortest path trees in the low temperature phase, and far from the uniform measure in the high temperature phase.

#### F.1.1. UNIFORM MEASURE AS GROUND STATE

As a warm-up, we analyze the uniform measure  $\mu_{G,\infty}$  over the shortest path trees of  $\overline{G}$ . In statistical physics terminology, this is referred to as the *ground state*. In this section, we compute the ground state energy and entropy of our system, which is quite similar to the computation in the proof of Proposition 46 in Section E.3.5.

The ground state energy of the system is equivalent to the sum of  $d(1, v)$  over all  $v$  in the giant component of  $G$ .

**Theorem 66 (Ground state energy)** *We have*

$$\sum_{v \in \overline{V}} d_{G_n}(1, v) = (d_n^* + \lambda_n)n + o_p \left( \frac{n}{\log \log n} \right)$$

**Proof** An upper bound is already proved in (94):

$$\sum_v d_G(1, v) \leq (d^* + 1)n - N_{d^*} + o_p \left( \frac{n}{\log \log n} \right).$$

For the lower bound, we have

$$\begin{aligned} \sum_v d_G(1, v) &\geq N_{d^*-1}(d^* - 1) + N_{d^*}d^* + |\overline{V} \setminus \Gamma_{\leq d^*}|(d^* + 1) \\ &= |\overline{V} \setminus \Gamma_{\leq d^*-2}|(d^* + 1) - 2N_{d^*-1} - N_{d^*}. \end{aligned}$$

By Proposition 16,  $N_{d^*-1}$  concentrates around  $(nq)^{d^*-1} = O_p(n/(\log \log n)^2)$ . Also applying (84), we get the desired lower bound

$$\sum_v d_G(1, v) \geq (d^* + 1)n - N_{d^*} - o_p \left( \frac{n}{\log \log n} \right).$$

Now applying Theorem 17, we obtain

$$\begin{aligned} \sum_v d_G(1, v) &= (d^* + 1)n - (1 - \lambda_n)n + o_p\left(\frac{n}{\log \log n}\right) \\ &= (d^* + \lambda_n)n + o_p\left(\frac{n}{\log \log n}\right). \end{aligned}$$

■

The ground state entropy is just the log of the number of shortest path trees. In terms of the conditional Gibbs measure (79), we have

$$Z_{G,\beta} \|_{N_{d^*}} = \#\{\text{shortest path trees of } \overline{G}\} \cdot \exp\left(-\bar{\beta} \sum_v d_G(1, v)\right)$$

so

$$\log(Z_{G,\beta} \|_{N_{d^*}}) = \log \#\{\text{shortest path trees of } \overline{G}\} - \beta \log \log(n) \sum_v d_G(1, v).$$

Since we know the ground state energy, the entropy can be easily obtained from  $\log(Z_{G,\beta} \|_{N_{d^*}})$ . By Corollary 61 this is approximately  $\Phi_{G,\beta}(N_{d^*})$  which is calculated in the low temperature part of Theorem 42. This immediately yields the following theorem.

**Theorem 67 (Entropy of the ground state)** *We have*

$$\log \#\{\text{shortest path trees of } \overline{G}_n\} = n\Psi_0(\lambda_n) + \lambda_n n \log(\alpha_n(1 - \lambda_n) \log n) + o_p(n).$$

*In particular, if  $\lambda_n \rightarrow \lambda \in [0, 1]$  and  $\Delta_n \rightarrow \Delta \in [0, 1]$ , then*

$$\frac{\log \#\{\text{shortest path trees of } \overline{G}\}}{n \log \log n} \xrightarrow{p} \lambda + \Delta = \begin{cases} \Delta & \text{if } 0 < \Delta \leq 1, \\ \lambda & \text{if } \Delta = 0. \end{cases}$$

#### F.1.2. DISTRIBUTION OF DISTANCE VECTORS

In Section E.3.4, we studied the conditional Gibbs measures  $\mu_{G,\beta}|_A$  for a fixed kernel  $A$ , which are building blocks for the Gibbs measures and are easier to analyze. Similarly, we start by analyzing  $\mu_{G,\beta}|_A$ , extending the results from the previous section where  $A = N_{d^*}$ . The first step is to prove that a typical spanning tree under  $\mu_{G,\beta}|_A$  has near-minimum energy.

**Proposition 68** *For any constant  $C > 0$ , consider a set of spanning trees of  $\overline{G}$  defined by*

$$E_{G,m} = \left\{ T : \sum_v d_T(1, v) \leq (d^* + 1)n - m + \frac{Cn}{\log \log n} \right\}. \quad (113)$$

*Let  $\beta > 0$  be a fixed constant. Then with probability at least  $1 - o(1)$ , we have*

$$\mu_{G,\beta}|_A(E_{G,|A|}) = 1 - \exp(-(C\beta - o(1))n)$$

*simultaneously for all  $A \in \mathcal{K}(G)$  of size  $|A| \geq m_\ell$ .*

To prove Proposition 68, we observe that the log partition function  $\log(Z_{G,\beta}|_A)$  established in Proposition 46 can be decomposed into the entropy part

$$\sum_{v \in A} \log |\text{par}_G(v)| + (n - |A|) \log(|A|q)$$

and the energy part

$$\bar{\beta}|A| - \bar{\beta}(d^* + 1)n.$$

The entropy part is in fact nearly the largest possible. This is because by the definition of kernel, any vertex  $v \in A$  has at most  $\log |\text{par}_G(v)|$  parent selection entropy, and for all other vertices  $\log(|A|q) \approx \log \log n$  is the maximum possible for sparse graphs. Hence, a typical tree under  $\mu_{G,\beta}|_A$  cannot have energy too greater than  $\bar{\beta}|A| - \bar{\beta}(d^* + 1)n$ .

Before we prove the proposition, we first establish the following lemma, which allows us to replace  $\log(|A|q)$  with a more convenient form.

**Lemma 69** *There is a universal constant  $C > 0$  such that with probability at least  $1 - o(1)$ ,*

$$\sum_{v \in B} \log |\mathbf{N}_A(v)| \leq |B| \log(|A|q \vee 1) + \frac{Cn}{\sqrt{|A|q}}$$

*simultaneously for all  $A \subseteq \Gamma_{d^*}$  and  $B \subseteq \mathbf{N}_{G \setminus \Gamma_{\leq d^*-1}}(A)$ .*

**Proof** The proof is similar to that of Lemma 53. Treating  $\Gamma_{\leq d^*-1}$  and  $A$  as fixed, by Lemma 103

$$\|\log |\mathbf{N}_A(v)| - \log(|A|q \vee 1)\|_{\psi_2} \leq \frac{C_1}{\sqrt{|A|q}}$$

for some constant  $C_1 > 0$ . By Corollary 97 the probability that  $\sum \log |\mathbf{N}_A(v)|$  is away from its mean by at least  $t > 0$  is bounded by

$$2 \exp \left( -\frac{2t^2|A|q}{C_2^2|B|} \right) \leq 2 \exp \left( -\frac{2t^2|A|q}{C_2^2n} \right)$$

for some constant  $C_2$ . By setting  $t = C_2n/\sqrt{|A|q}$  this is bounded by  $2e^{-2n}$ , so taking the union bound over all  $A$  and  $B$  gives

$$\sum_{v \in B} \log |\mathbf{N}_A(v)| \leq |B| \mathbb{E}[\log |\mathbf{N}_A(v)| \mid \Gamma_{\leq d^*-1}, A] + \frac{C_3n}{\sqrt{|A|q}}$$

simultaneously for all  $A$  and  $B$  asymptotically almost surely. Finally, Lemma 96 gives

$$\begin{aligned} \sum_{v \in B} \log |\mathbf{N}_A(v)| &\leq |B| \left( \log(|A|q \vee 1) + \frac{C_1}{\sqrt{|A|q}} \right) + \frac{C_3n}{\sqrt{|A|q}} \\ &\leq |B| \log(|A|q \vee 1) + \frac{Cn}{\sqrt{|A|q}} \end{aligned}$$

for some constant  $C > 0$ . ■

**Proof** [Proof of Proposition 68] By Proposition 46 we may assume

$$\begin{aligned} \log(Z_{G,\beta}|_A) &= \log \left( \sum_{T \in \varphi^{-1}(A)} \exp \left( -\bar{\beta} \sum_v d_T(1, v) \right) \right) \\ &= \sum_{v \in A} \log |\text{par}_G(v)| + (n - |A|) \log(|A|q) + \bar{\beta}|A| - \bar{\beta}(d^* + 1)n + o(n). \end{aligned}$$

Define

$$\bar{E}_{G,A} := \varphi^{-1}(A) \cap E_{G,|A|} \quad (114)$$

and

$$\bar{E}_{G,A}^c := \varphi^{-1}(A) \setminus E_{G,|A|}.$$

Note that we have

$$\log \left( \mu_{G,\beta}|_A(E_{G,|A|}^c) \right) = \log \left( \sum_{T \in \bar{E}_{G,A}^c} \exp \left( -\bar{\beta} \sum_v d_T(1, v) \right) \right) - \log(Z_{G,\beta}|_A)$$

so our goal is to analyze

$$\sum_{T \in \bar{E}_{G,A}^c} \exp \left( -\bar{\beta} \sum_v d_T(1, v) \right).$$

We decompose

$$\bar{E}_{G,A}^c = \bigsqcup_B \bar{E}_{G,A,B}^c$$

where

$$\bar{E}_{G,A,B}^c := \left\{ T \in \bar{E}_{G,A}^c : \Gamma_{d^*+1}^T = B \right\}.$$

If  $T \in \bar{E}_{G,A,B}^c$  we have

$$\begin{aligned} \sum_v d_T(1, v) &= d^*|A| + (d^* + 1)|B| + (d^* + 2)(n - |A| - |B|) \pm o_p \left( \frac{\log \log n}{n} \right) \\ &= (d^* + 1)n - |A| + (n - |A| - |B|) \pm o_p \left( \frac{\log \log n}{n} \right). \end{aligned}$$

Thus,  $\bar{E}_{G,A,B}^c \neq \emptyset$  implies that

$$n - |A| - |B| \geq \frac{(C - o_p(1))n}{\log \log n}. \quad (115)$$

To bound the size of  $\bar{E}_{G,A,B}^c$  given that it is nonempty, we observe that  $v \in A$  has  $|\text{par}_G(v)|$  parent choices,  $v \in B$  has  $|\text{N}_A(v)|$  parent choices, and the other vertices have  $O_p(\log n)$  parent choices

which is simply the bound for the maximum degree of a typical sparse graph. Using Lemma 69 this gives

$$\begin{aligned}
 \log \left| \overline{E}_{G,A,B}^c \right| &\leq \sum_{v \in A} \log |\text{par}_G(v)| + \sum_{v \in B} \log |\mathbf{N}_A(v)| + (n - |A| - |B|) \log \log n + o_p(n) \\
 &\leq \sum_{v \in A} \log |\text{par}_G(v)| + |B| \log(|A|q) + (n - |A| - |B|) \log \log n + o_p(n) \\
 &\leq \sum_{v \in A} \log |\text{par}_G(v)| + (n - |A|) \log(|A|q) + (n - |A| - |B|) \log(3(\log \log n)^2) + o_p(n)
 \end{aligned}$$

which holds simultaneously for all  $A$  and  $B$  with  $|A| \geq m_\ell$ . Hence, it follows that

$$\begin{aligned}
 &\log \left( \sum_{T \in \overline{E}_{G,A,B}^c} \exp \left( -\bar{\beta} \sum_v \mathbf{d}_T(1, v) \right) \right) \\
 &\leq \log(Z_{G,\beta}|_A) + (n - |A| - |B|)(\log(3(\log \log n)^2) - \bar{\beta}) + o_p(n) \\
 &= \log(Z_{G,\beta}|_A) - (\beta - o(1))(n - |A| - |B|) \log \log n + o_p(n).
 \end{aligned}$$

This gives

$$\begin{aligned}
 &\log \left( \sum_{|B|=n-|A|-k} \sum_{T \in \overline{E}_{G,A,B}^c} \exp \left( -\bar{\beta} \sum_v \mathbf{d}_T(1, v) \right) \right) \\
 &\leq \log \binom{n}{k} + \log(Z_{G,\beta}|_A) - (\beta - o(1))k \log \log n + o_p(n) \\
 &\leq k \log \left( \frac{en}{k} \right) + \log(Z_{G,\beta}|_A) - (\beta - o(1))k \log \log n + o_p(n) \\
 &\leq k \log \left( \frac{e \log \log n}{C - o_p(1)} \right) + \log(Z_{G,\beta}|_A) - (\beta - o(1))k \log \log n + o_p(n) \\
 &\leq \log(Z_{G,\beta}|_A) - (\beta - o_p(1))k \log \log n + o_p(n) \\
 &\leq \log(Z_{G,\beta}|_A) - (C\beta - o_p(1))n + o_p(n)
 \end{aligned}$$

where we have used (115). This holds simultaneously for all  $k$  such that the sum is nonzero. Hence, we have

$$\begin{aligned}
 &\log \left( \sum_{T \in \overline{E}_{G,A}^c} \exp \left( -\bar{\beta} \sum_v \mathbf{d}_T(1, v) \right) \right) \\
 &\leq \max_k \left\{ \log \left( \sum_{|B|=n-|A|-k} \sum_{T \in \overline{E}_{G,A,B}^c} \exp \left( -\bar{\beta} \sum_v \mathbf{d}_T(1, v) \right) \right) \right\} + O(\log n) \\
 &\leq \log(Z_{G,\beta}|_A) - (C\beta - o_p(1))n + o_p(n).
 \end{aligned}$$

The proposition is proved. ■

### F.1.3. CONDITIONAL GIBBS MEASURES ARE ALMOST UNIFORM

Now we show that  $\mu_{G,\beta}|_A$  is close in Wasserstein distance to  $\mu_{G,\infty}|_A$ , the uniform measure over the spanning trees with kernel  $A$  and minimum energy. We again use the fact that the entropy is nearly maximum and apply the entropy-transport inequality through Corollary 65. Since some vertices might have parents not in the support of  $\mu_{G,\infty}|_A$ , we need to round them, for which we appeal to Proposition 68.

**Proposition 70** *Let  $\beta > 0$  be a fixed constant. Then with probability at least  $1 - o(1)$ , we have*

$$W_1(\mu_{G,\infty}|_A, \mu_{G,\beta}|_A) = o(n)$$

*simultaneously for all  $A \in \mathcal{K}(G)$  of size  $|A| \geq m_\ell$ .*

**Proof** For an arbitrary constant  $C > 0$ , let  $E_{G,m}$  and  $\bar{E}_{G,A}$  be as defined in (113) and (114). Let  $\mu_{G,\beta}|\bar{E}_{G,A}$  be the measure  $\mu_{G,\beta}|_A$  conditioned on  $E_{G,|A|}$  (or  $\bar{E}_{G,A}$ ). From Proposition 68, it is clear that

$$W_1(\mu_{G,\beta}|\bar{E}_{G,A}, \mu_{G,\beta}|_A) = o_p(n) \quad (116)$$

and

$$\log(Z_{G,\beta}|\bar{E}_{G,A}) := \log \left( \sum_{T \in \bar{E}_{G,A}} \exp \left( -\bar{\beta} \sum_v d_T(1, v) \right) \right) = \log(Z_{G,\beta}|_A) - o_p(n)$$

simultaneously for all  $|A| \geq m_\ell$ . By Proposition 46 and the Gibbs variational principle (Proposition 37), this implies that

$$\begin{aligned} H_{\mu_{G,\beta}|\bar{E}_{G,A}}(T) &= \log(Z_{G,\beta}|\bar{E}_{G,A}) + \bar{\beta} \mathbb{E}_{\mu_{G,\beta}|\bar{E}_{G,A}} \left[ \sum_v d_T(1, v) \right] \\ &\geq \log(Z_{G,\beta}|_A) + \bar{\beta}(d^* + 1)n - \bar{\beta}|A| - o_p(n) \\ &= \sum_{v \in A} \log |\text{par}_G(v)| + (n - |A|) \log(|A|q) - o_p(n) \end{aligned}$$

and

$$\mathbb{E}_{\mu_{G,\beta}|\bar{E}_{G,A}} \left[ \sum_v d_T(1, v) \right] \leq (d^* + 1)n - |A| + \frac{Cn}{\log \log n}. \quad (117)$$

Now we consider the distance vector  $D = d_T : v \mapsto d_T(1, v)$ . Since  $D$  is a function of  $T$  we have

$$H_{\mu_{G,\beta}|\bar{E}_{G,A}}(T) = H_{\mu_{G,\beta}|\bar{E}_{G,A}}(D) + H_{\mu_{G,\beta}|\bar{E}_{G,A}}(T | D).$$

By the stars and bars method, we see that the number of possible distance vectors of a spanning tree in  $E_{G,m}$  is bounded by

$$\binom{n + \frac{Cn}{\log \log n} + o(n) - 1}{\frac{Cn}{\log \log n} + o(n)} = \binom{n + o(n)}{o(n)}$$

which implies that

$$H_{\mu_{G,\beta}|\bar{E}_{G,A}}(D) \leq \log \binom{n + o(n)}{o(n)} = o(n).$$

As a consequence, we get

$$\begin{aligned} H_{\mu_{G,\beta}|\bar{E}_{G,A}}(T | D) &= H_{\mu_{G,\beta}|\bar{E}_{G,A}}(T) - o(n) \\ &\geq \sum_{v \in A} \log |\text{par}_G(v)| + (n - |A|) \log(|A|q) - o_p(n). \end{aligned} \quad (118)$$

To apply the entropy-transport inequality (Corollary 65), we round  $\mu_{G,\beta}|\bar{E}_{G,A}$  to the support of  $\mu_{G,\infty}|_A$ . This is equivalent to saying that we are forcing every vertex to have minimum distance to the root, provided that the kernel is  $A$ . We denote by  $d_A$  the unique distance vector in the support of  $\mu_{G,\infty}|_A$ , i.e.,

$$d_A(v) := \begin{cases} d_G(1, v) & \text{if } d_G(1, v) \leq d^* - 1, \\ d^* + d_G(v, A) & \text{if } v \in \bar{V} \setminus \Gamma_{\leq d^*-1}. \end{cases}$$

Now we define the rounding procedure from  $\mu_{G,\beta}|\bar{E}_{G,A}$  to  $\nu$  as follows. For a spanning tree  $T \sim \mu_{G,\beta}|\bar{E}_{G,A}$ , for vertices  $v$  such that  $d_A(1, v) \neq 1 + d_A(1, \text{par}_T(v))$  (which implies  $d_T(1, v) \neq d_A(v)$ ), we replace each  $\text{par}_T(v)$  with an arbitrary vertex  $u \in \text{par}_G(v)$  satisfying  $d_A(1, v) = 1 + d_A(1, u)$ . It is easy to see that the resulting graph  $S$  is a valid spanning tree in the support of  $\mu_{G,\infty}|_A$  (i.e.,  $d_T = d_A$ ). Also, since we do not modify the parents of more than  $\frac{Cn}{\log \log n}$  vertices,  $\nu$  is not too far from the original measure in terms of Wasserstein distance:

$$W_1(\nu, \mu_{G,\beta}|\bar{E}_{G,A}) = o(n). \quad (119)$$

We claim that this procedure does not lose entropy by more than  $o_p(n)$ , by analyzing the entropy on a per distance vector basis. Since the law of  $T$  conditioned on distance vector is a product measure, we have

$$\begin{aligned} H_{\mu_{G,\beta}|\bar{E}_{G,A}}(T | D = d) &= \sum_{v: d(v) \neq d_A(v)} H_{\mu_{G,\beta}|\bar{E}_{G,A}}(\text{par}_T(v) | D = d) \\ &\quad + \sum_{v: d(v) = d_A(v)} H_{\mu_{G,\beta}|\bar{E}_{G,A}}(\text{par}_T(v) | D = d) \\ &\leq |\{v : d(v) \neq d_A(v)\}| \log \log n + o_p(n) \\ &\quad + \sum_{v: d(v) = d_A(v)} H_{\mu_{G,\beta}|\bar{E}_{G,A}}(\text{par}_T(v) | D = d) \end{aligned}$$

where we have used the fact that the maximum degree is  $O_p(\log n)$ . By our construction of  $S \sim \nu$ , we have

$$\begin{aligned} H_\nu(S | D = d) &\geq \sum_{v: d(v) = d_A(v)} H_{\mu_{G,\beta}|\bar{E}_{G,A}}(\text{par}_T(v) | D = d) \\ &\geq H_{\mu_{G,\beta}|\bar{E}_{G,A}}(T | D = d) - |\{v : d(v) \neq d_A(v)\}| \log \log n - o_p(n). \end{aligned}$$

This gives

$$\begin{aligned} H_\nu(S) &\geq H_\nu(S \mid D) \\ &\geq H_{\nu_{G,\beta}|\bar{E}_{G,A}}(T \mid D) - (\log \log n) \mathbb{E}_{\mu_{G,\beta}|\bar{E}_{G,A}} [|\{v : d_T(1, v) \neq d_A(v)\}|] - o_p(n). \end{aligned}$$

Now (117) implies that

$$(\log \log n) \mathbb{E}_{\mu_{G,\beta}|\bar{E}_{G,A}} [|\{v : d_T(1, v) \neq d_A(v)\}|] \leq Cn$$

which by (118) gives

$$\begin{aligned} H_\nu(S) &\geq H_{\nu_{G,\beta}|\bar{E}_{G,A}}(T \mid D) - Cn - o_p(n) \\ &\geq \sum_{v \in A} \log |\text{par}_G(v)| + (n - |A|) \log(|A|q) - Cn - o_p(n). \end{aligned}$$

This is at most  $Cn + o_p(n)$  smaller than the maximum entropy attained by  $\mu_{G,\infty}|_A$ . Therefore, applying Corollary 65 with  $\nu$  and  $\mu = \mu_{G,\infty}|_A$ , we arrive at

$$W_1(\nu, \mu_{G,\infty}|_A) \leq \sqrt{2Cn} + o_p(n).$$

Since  $C$  is arbitrary, this implies  $W_1(\nu, \mu_{G,\infty}|_A) = o_p(n)$ . Together with (116) and (119), we complete the proof.  $\blacksquare$

Now that we better understand the structure of  $\mu_{G,\beta}|_A$ , we mix these measures to build  $\mu_{G,\beta}$ . Since our results so far only apply to the case where the kernel size is at least  $m_\ell$ , we prove that the kernel is indeed large enough asymptotically almost surely.

**Lemma 71** *Let  $\beta > 0$  be a fixed constant. Then we have*

$$\mu_{G,\beta}(\{T : |\varphi(T)| \geq m_\ell\}) \geq 1 - \exp((1/3 - o(1))n)$$

with probability at least  $1 - o(1)$ .

**Proof** This is trivial if  $m_\ell = m_0$  so we assume

$$m_\ell = \frac{n}{3(\log \log n)^2}.$$

We begin with

$$\begin{aligned} \log(\mu_{G,\beta}(\{T : |\varphi(T)| \leq m_\ell\})) &= \log\left(\sum_{m=m_0}^{m_\ell} Z_{G,\beta}\|_m\right) - \log Z_{G,\beta} \\ &\leq \max_{m_0 \leq m \leq m_\ell} \log(Z_{G,\beta}\|_m) - \log Z_{G,\beta} + \log n. \end{aligned} \tag{120}$$

By Lemma 52, we have

$$\log(Z_{G,\beta}\|_m) \leq \mathcal{H}_G(m) + (n - m) \log(L + mq) + \bar{\beta}m - \bar{\beta}(d^* + 1)n + o_p(n).$$

By Lemma 56 and Lemma 63, we see that  $\mathcal{H}_G(m) = o_p(n)$  for  $m \leq m_\ell$ , which gives

$$\begin{aligned} \log(Z_{G,\beta}\|_m) &\leq (n-m) \log(L+mq) + \bar{\beta}m - \bar{\beta}(d^*+1)n + o_p(n) \\ &\leq n \log(L+mq) - \bar{\beta}(d^*+1)n + o_p(n) \\ &\leq n \log\left(\frac{nq}{(\log \log n)^2}\right) - n \log 3 - \bar{\beta}(d^*+1)n + o_p(n). \end{aligned}$$

Now we lower bound  $\log Z_{G,\beta}$ . Let

$$m_1 := \left\lfloor \frac{n}{2(\log \log n)^2} \right\rfloor$$

which is strictly smaller than  $N_{d^*}$  asymptotically almost surely (by Lemma 18). By Proposition 47 we have

$$\begin{aligned} \log Z_{G,\beta} &\geq \log(Z_{G,\beta}\|_{m_1}) \\ &= \mathcal{H}_G(m_1) + (n-m_1) \log(m_1q) + \bar{\beta}m_1 - \bar{\beta}(d^*+1)n - o_p(n) \\ &\geq (n-m_1) \log(m_1q) - \bar{\beta}(d^*+1)n - o_p(n) \\ &\geq \left(1 - \frac{1}{2(\log \log n)^2} - \frac{1}{n}\right) n \log\left(\frac{nq}{2(\log \log n)^2}\right) - \bar{\beta}(d^*+1)n - o_p(n) \\ &= n \log\left(\frac{nq}{(\log \log n)^2}\right) - n \log 2 - \bar{\beta}(d^*+1)n - o_p(n). \end{aligned}$$

Continuing from (120), we have

$$\begin{aligned} \log(\mu_{G,\beta}(\{T : |\varphi(T)| \leq m_\ell\})) &\leq -n \log(3/2) + o_p(n) \\ &= -\frac{1}{3}n + o_p(n). \end{aligned}$$

■

Since the measures  $\mu_{G,\beta}|_A$  can be approximated by  $\mu_{G,\infty}$ , we mix these using the mixture weights of  $\mu_{G,\beta}$ . Namely, we define

$$\tilde{\mu}_{G,\beta}(T) := \sum_{A \in \mathcal{K}(G)} \mu_{G,\beta}(\varphi(T) = A) \cdot \mu_{G,\infty}|_A(T).$$

An equivalent way of defining  $\tilde{\mu}_{G,\beta}$  is through the following sampling algorithm:

1. for each  $v \in \Gamma_{\leq d^*-1}$ , sample  $\text{par}_T(v)$  from  $\text{par}_G(v)$  uniformly at random;
2. sample  $A \subseteq \Gamma_{d^*}$  according to the marginal law of the kernel  $\varphi(T)$  under  $\mu_{G,\beta}$ ;
3. for each  $v \in A$ , sample  $\text{par}_T(v)$  from  $\text{par}_G(v)$  uniformly at random;
4. for each  $v \in \bar{V} \setminus \Gamma_{d^*-1} \setminus A$ , sample  $\text{par}_T(v)$  uniformly at random from the set of vertices  $u \in \mathbb{N}_G(v)$  such that  $d_G(v, A) = d_G(u, A) + 1$ .

Then our results immediately yield the following corollary.

**Corollary 72** *Let  $\beta > 0$  be a fixed constant. Then we have*

$$\text{KL}(\tilde{\mu}_{G,\beta}, \mu_{G,\beta}) = o_p(n)$$

and

$$W_1(\tilde{\mu}_{G,\beta}, \mu_{G,\beta}) = o_p(n).$$

**Proof** The first result follows by Lemma 71, Proposition 46, and the convexity of the KL divergence. The second result follows by Lemma 71 and Proposition 70.  $\blacksquare$

#### F.1.4. GIBBS MEASURES IN THE LOW TEMPERATURE PHASE

Now we complement Theorem 42 by studying how the Gibbs measures are compared to the uniform measure  $\mu_{G,\infty}$  in 1-Wasserstein metric. We begin with the low temperature phase (LT), in which case the Gibbs measures are close to  $\mu_{G,\infty}$ .

**Theorem 73** *In the low temperature phase (LT) we have*

$$W_1(\mu_{G_n,\infty}, \mu_{G_n,\beta}) = o_p(n).$$

The idea of our proof is to show that a typical tree in the low temperature phase has a nearly maximum sized kernel. Similar to the our proof method in Section F.1.3, we observe that the entropy part of the formula in Theorem 42 is almost the maximum possible, and then apply the entropy-transport inequality.

**Lemma 74** *For any constant  $C > 0$ , let*

$$m_1 := m_\ell \vee \lfloor (1 - C\kappa_n^{-1})N_{d^*} \rfloor. \quad (121)$$

*Then in the low temperature phase (LT), we have with probability at least  $1 - o(1)$  that*

$$\mu_{G,\beta}(\{T : |\varphi(T)| \leq m_1\}) \leq \exp(-\Omega(n)).$$

**Proof** The statement follows by Lemma 71 if  $m_1 = m_\ell$  so we assume  $m_1 = \lfloor (1 - C\kappa_n^{-1})N_{d^*} \rfloor$ . We have

$$\begin{aligned} \log(\mu_{G,\beta}(\{T : |\varphi(T)| \leq m_1\})) &= \log\left(\sum_{m=m_0}^{m_1} Z_{G,\beta}\|_m\right) - \log Z_{G,\beta} \\ &\leq \max_{m_0 \leq m \leq m_1} \log(Z_{G,\beta}\|_m) - \log Z_{G,\beta} + \log n. \end{aligned}$$

We will use  $f_u(x)$  defined in (105) to upper bound  $\tilde{\Phi}_{G,\beta}(m)$ . Namely, by Corollary 61, we have

$$\begin{aligned} \max_{m_0 \leq m \leq m_1} \log(Z_{G,\beta}\|_m) &= \max_{m_0 \leq m \leq m_1} \tilde{\Phi}_{G,\beta}(m) \\ &\leq \max_{m_0 \leq x \leq m_1} f_u(x) + o_p(n) \end{aligned}$$

and

$$\log Z_{G,\beta} \geq f_u(N_{d^*}).$$

We now take a look at  $f_u((1 - C\kappa_n^{-1})N_{d^*})$ . Similar to (109), we have

$$(1 - C\kappa_n^{-1})N_{d^*} \geq N_{d^*} - \frac{Cn}{\log \log n} - o_p\left(\frac{n}{\log \log n}\right) \quad (122)$$

which implies  $h((1 - C\kappa_n^{-1})N_{d^*}) = o_p(n)$ . Also similar to (110),

$$g(N_{d^*}) - g((1 - C\kappa_n^{-1})N_{d^*}) = C(\beta - 1 + \Delta_n + \kappa_n^{-1})n - (C\kappa_n^{-1} + \log(1 - C\kappa_n^{-1}))n - o_p(n).$$

We consider two regimes for the low temperature phase. If  $\beta > \limsup_{n \rightarrow \infty} (1 - \Delta_n - \kappa_n^{-1})$ , then the RHS is clearly  $\Omega_p(n)$ . Otherwise, if  $\beta = \limsup_{n \rightarrow \infty} (1 - \Delta_n - \kappa_n^{-1})$  and  $\limsup_{n \rightarrow \infty} \kappa_n = K < \infty$  for some constant  $K$ , then  $C\kappa_n^{-1} > \frac{C}{2K}$  for all sufficiently large  $n$ , which again implies that the RHS is  $\Omega(n)$ . In either case, we arrive at  $g(N_{d^*}) - g((1 - C\kappa_n^{-1})N_{d^*}) = \Omega_p(n)$  which gives

$$f_u(N_{d^*}) - f_u((1 - C\kappa_n^{-1})N_{d^*}) = \Omega_p(n).$$

By the concavity of  $f_u$  this implies that  $f_u(x)$  is increasing in  $x \leq (1 - C\kappa_n^{-1})N_{d^*}$ . As a result, we conclude that

$$\begin{aligned} \log(\mu_{G,\beta}(\{T : |\varphi(T)| \leq m_1\})) &\leq \max_{m_0 \leq m \leq m_1} \log(Z_{G,\beta}|_m) - \log Z_{G,\beta} + o_p(n) \\ &\leq f_u((1 - C\kappa_n^{-1})N_{d^*}) - f_u(N_{d^*}) + o_p(n) \\ &\leq -\Omega_p(n). \end{aligned}$$

proving the lemma. ■

**Proof** [Proof of Theorem 73] Let  $C > 0$  be an arbitrary constant and define  $m_1$  as in (121). Consider the measure  $\tilde{\mu}_{G,\beta}|_{\geq m_1}$  defined by  $\tilde{\mu}_{G,\beta}$  conditioned on  $|\varphi(T)| \geq m_1$ :

$$\tilde{\mu}_{G,\beta}|_{\geq m_1}(T) := \sum_{A \in \mathcal{K}(G)} \mu_{G,\beta}(\varphi(T) = A \mid |\varphi(T)| \geq m_1) \cdot \mu_{G,\infty}|_A(T).$$

By Lemma 74, Proposition 46, and the convexity of the KL divergence we have

$$\text{KL}(\tilde{\mu}_{G,\beta}|_{\geq m_1}, \mu_{G,\beta}) = o_p(n) \quad (123)$$

and by Proposition 70 and Lemma 74

$$W_1(\tilde{\mu}_{G,\beta}|_{\geq m_1}, \mu_{G,\beta}) = o_p(n). \quad (124)$$

By Proposition 46 and the Gibbs variational principle (Proposition 37) we have

$$\begin{aligned} H_{\tilde{\mu}_{G,\beta}|_{\geq m_1}}(T) &\geq \log Z_{G,\beta} + \bar{\beta} \mathbb{E}_{\tilde{\mu}_{G,\beta}|_{\geq m_1}} \left[ \sum_v d_T(1, v) \right] - o_p(n) \\ &\geq \log(Z_{G,\beta}|_{\Gamma_{d^*}}) + \bar{\beta} \sum_v d_G(1, v) - o_p(n) \\ &= \sum_{v \in V \setminus \{1\}} \log |\text{par}_G(v)| - o_p(n) \end{aligned} \quad (125)$$

and by (122)

$$\mathbb{E}_{\tilde{\mu}_{G,\beta} \parallel \geq m_1} \left[ \sum_v d_T(1, v) \right] \leq \sum_v d_G(1, v) + \frac{Cn}{\log \log n} + o_p \left( \frac{n}{\log \log n} \right). \quad (126)$$

Now we round  $\tilde{\mu}_{G,\beta} \parallel \geq m_1$  to a measure  $\nu$  supported on the set of shortest path trees using the following procedure. For a spanning tree  $T \sim \tilde{\mu}_{G,\beta} \parallel \geq m_1$ , for each vertex  $v \in \Gamma_{d^*} \setminus \varphi(T)$ , replace  $\text{par}_T(v)$  with an arbitrary vertex in  $\Gamma_{d^*-1}$ . Since  $\tilde{\mu}_{G,\beta} \parallel \geq m_1$  is a mixture of measures of the form  $\mu_{G,\infty}|_A$ , it is not difficult to see that the resulting graph  $S$  is a shortest path tree and that

$$W_1(\nu, \tilde{\mu}_{G,\beta} \parallel \geq m_1) = o_p(n). \quad (127)$$

For  $\mathbf{A} = \varphi(T)$  note that

$$H_{\tilde{\mu}_{G,\beta} \parallel \geq m_1}(T) = H_{\tilde{\mu}_{G,\beta} \parallel \geq m_1}(\mathbf{A}) + H_{\tilde{\mu}_{G,\beta} \parallel \geq m_1}(T \mid \mathbf{A})$$

where

$$\begin{aligned} H_{\tilde{\mu}_{G,\beta} \parallel \geq m_1}(\mathbf{A}) &\leq \log \left( \sum_{m=m_1}^{N_{d^*}} \binom{N_{d^*}}{m} \right) \\ &\leq \log \binom{n}{m_1} + \log n \\ &= o_p(n). \end{aligned}$$

Thus,

$$H_{\tilde{\mu}_{G,\beta} \parallel \geq m_1}(T \mid \mathbf{A}) \geq H_{\tilde{\mu}_{G,\beta} \parallel \geq m_1}(T) - o_p(n). \quad (128)$$

Also,

$$\begin{aligned} H_{\tilde{\mu}_{G,\beta} \parallel \geq m_1}(T \mid \mathbf{A} = A) &= H_{\mu_{G,\infty}|_A}(T) \\ &= \sum_{v \in \Gamma_{d^*} \setminus A} H_{\mu_{G,\infty}|_A}(\text{par}_T(v)) + \sum_{v \in V \setminus \{1\} \setminus (\Gamma_{d^*} \setminus A)} H_{\mu_{G,\infty}|_A}(\text{par}_T(v)) \\ &\leq |\Gamma_{d^*} \setminus A| \log \log n + \sum_{v \in V \setminus \{1\} \setminus (\Gamma_{d^*} \setminus A)} H_{\mu_{G,\infty}|_A}(\text{par}_T(v)) + o_p(n) \end{aligned}$$

which gives

$$\begin{aligned} H_\nu(S \mid \mathbf{A} = A) &\geq \sum_{v \in V \setminus \{1\} \setminus (\Gamma_{d^*} \setminus A)} H_{\mu_{G,\infty}|_A}(\text{par}_T(v)) \\ &\geq H_{\tilde{\mu}_{G,\beta} \parallel \geq m_1}(T \mid \mathbf{A} = A) - |\Gamma_{d^*} \setminus A| \log \log n - o_p(n). \end{aligned}$$

Thus,

$$\begin{aligned} H_\nu(S) &\geq H_\nu(S \mid \mathbf{A}) \\ &\geq H_{\tilde{\mu}_{G,\beta} \parallel \geq m_1}(T \mid \mathbf{A}) - (\log \log n) \mathbb{E}_{\tilde{\mu}_{G,\beta} \parallel \geq m_1}[|\Gamma_{d^*} \setminus \mathbf{A}|] - o_p(n). \end{aligned}$$

Here, (126) gives

$$\begin{aligned} (\log \log n) \mathbb{E}_{\tilde{\mu}_{G,\beta} \| \geq m_1} [|\Gamma_{d^*} \setminus \mathbf{A}|] &\leq (\log \log n) \mathbb{E}_{\tilde{\mu}_{G,\beta} \| \geq m_1} [|v \in V : d_T(1, v) \neq d_G(1, v)|] \\ &\leq Cn + o_p(n). \end{aligned}$$

Combined with (125) and (128), we get

$$H_\nu(S) \geq \sum_{v \in V \setminus \{1\}} \log |\text{par}_G(v)| - Cn - o_p(n).$$

Applying Corollary 65 to  $\nu$  and  $\mu = \mu_{G,\infty}$  gives

$$W_1(\nu, \mu_{G,\infty}) \leq \sqrt{2C}n + o_p(n). \quad (129)$$

Since  $C$  is arbitrary, this implies  $W_1(\nu, \mu_{G,\infty}) = o_p(n)$ . Together with (124) and (127) the proposition is proved.  $\blacksquare$

#### F.1.5. GIBBS MEASURES IN THE HIGH TEMPERATURE PHASE

Now we move to the high temperature phase. Here, we wish to prove that  $\mu_{G,\beta}$  is far from the uniform measure over the shortest path trees.

**Theorem 75** *In the high temperature phase (HT) we have*

$$W_1(\mu_{G_n,\infty}, \mu_{G_n,\beta}) = \Theta_p(n).$$

To show that the Wasserstein distance is large, we use the dual formulation of Wasserstein metric by explicitly constructing a 1-Lipschitz function of spanning trees and compare the expectations under  $\mu_{G,\beta}$  and  $\mu_{G,\infty}$ . The key object that distinguishes the high temperature phase from the low temperature phase is again kernels. Recall from Theorem 42 that the optimal kernel size that contributes the most to the log partition function is

$$m^* = \left\lfloor \frac{N_{d^*}}{(1 - \Delta_n - \beta)\kappa_n} \right\rfloor.$$

When there is an asymptotic gap between  $\beta$  and  $1 - \Delta_n - \kappa_n^{-1}$  as in (HT), this optimal kernel size is strictly smaller than  $N_{d^*}$ . We first prove that a typical tree in the high temperature phase has kernel size very close to this  $m^*$ , thereby is strictly smaller than  $N_{d^*}$ .

**Lemma 76** *Suppose that  $\beta < 1 - \Delta_n - \kappa_n^{-1} - \delta$  for a constant  $\delta > 0$ . Let*

$$m_2 := \left\lceil \left(1 + \frac{\delta}{7}\right) \cdot \frac{\kappa_n^{-1} N_{d^*}}{1 - \Delta_n - \beta} \right\rceil. \quad (130)$$

*Then we have asymptotically almost surely that*

$$\mu_{G,\beta}(\{T : |\varphi(T)| \geq m_2\}) \leq \exp(-\Omega(n)).$$

**Proof** We proceed similar to the proof of Lemma 74. We have

$$\begin{aligned} \log(\mu_{G,\beta}(\{T : |\varphi(T)| \geq m_2\})) &= \log\left(\sum_{m=m_2}^{N_{d^*}} Z_{G,\beta}\|_m\right) - \log Z_{G,\beta} \\ &\leq \max_{m_0 \vee m_2 \leq m \leq N_{d^*}} \log(Z_{G,\beta}\|_m) - \log Z_{G,\beta} + \log n \\ &\leq \max_{m_2 \leq m \leq N_{d^*}} f_u(x) - f_u\left(\frac{\kappa_n^{-1} N_{d^*}}{1 - \Delta_n - \beta}\right) + o_p(n). \end{aligned}$$

From (111), we have

$$h\left(\left(1 + \frac{\delta}{7}\right) \cdot \frac{\kappa_n^{-1} N_{d^*}}{1 - \Delta_n - \beta}\right) = o_p(n)$$

and by (112)

$$\begin{aligned} g\left(\frac{\kappa_n^{-1} N_{d^*}}{1 - \Delta_n - \beta}\right) - g\left(\left(1 + \frac{\delta}{7}\right) \cdot \frac{\kappa_n^{-1} N_{d^*}}{1 - \Delta_n - \beta}\right) &= \left(\frac{\delta}{7} - \log\left(1 + \frac{\delta}{7}\right)\right) n - o_p(n) \\ &= \Omega_p(n). \end{aligned}$$

This in particular gives

$$f_u\left(\frac{\kappa_n^{-1} N_{d^*}}{1 - \Delta_n - \beta}\right) - f_u\left(\left(1 + \frac{\delta}{7}\right) \cdot \frac{\kappa_n^{-1} N_{d^*}}{1 - \Delta_n - \beta}\right) = \Omega_p(n).$$

Since  $f_u$  is concave, for all large enough  $n$ ,  $f_u(x)$  is decreasing in  $x \geq \left(1 + \frac{\delta}{7}\right) \cdot \frac{\kappa_n^{-1} N_{d^*}}{1 - \Delta_n - \beta}$ . Therefore,

$$\begin{aligned} \log(\mu_{G,\beta}(\{T : |\varphi(T)| \geq m_2\})) &\leq f_u\left(\left(1 + \frac{\delta}{7}\right) \cdot \frac{\kappa_n^{-1} N_{d^*}}{1 - \Delta_n - \beta}\right) - f_u\left(\frac{\kappa_n^{-1} N_{d^*}}{1 - \Delta_n - \beta}\right) + o_p(n) \\ &\leq -\Omega_p(n). \end{aligned}$$

■

One might be tempted to try proving that  $W_1(\mu_{G,\beta}, \mu_{G,\infty})$  is large using the kernel size  $|\varphi(T)|$  as the Lipschitz function that witnesses a large shift in Wasserstein metric, since it clearly “jumps” when moving between the two phases. However,  $|\varphi(T)|$  might not be big enough if  $N_{d^*} = o_p(n)$  (i.e.,  $\lambda_n \rightarrow 1$ ), in which case it is not possible to witness a  $\Theta(n)$  shift.

Instead, we notice that the kernel must “support” most of the remaining vertices. More precisely, from Proposition 68, we know that most of the vertices not in the kernel must have distance  $d^* + 1$ , which means that the parents of those vertices must be in the kernel. As a consequence, vertices in the kernel must have large numbers of children, which is the fact we leverage to construct a Lipschitz function. To be specific, we make the following (informal) hypotheses.

- In the low temperature phase, the kernel of a typical tree is nearly the entire  $\Gamma_{d^*}$ . Hence, a typical vertex in  $\Gamma_{d^*}$  would have the number of children around  $N_{d^*+1}/N_{d^*}$ .
- In the high temperature phase, a nontrivial portion of  $\Gamma_{d^*}$  is not in the kernel. Thus, those in the kernel  $A$  would have the number of children around  $(n - |A|)/|A|$ , and those not in the kernel would mostly have no child.

Based on these observations, the Lipschitz function we construct is the sum of absolute deviation from  $N_{d^*+1}/N_{d^*}$ , given by

$$\sum_{v \in \Gamma_{d^*}} \left| \text{ch}_T(v) - \frac{N_{d^*+1}}{N_{d^*}} \right|. \quad (131)$$

Thus, we first need to understand how the absolute deviation behaves. For the sake of convenience, we introduce a few notations here. For a fixed kernel  $A \in \mathcal{K}(G)$ , we define

$$\mathbf{d}_A(v) := \begin{cases} \mathbf{d}_G(1, v) & \text{if } \mathbf{d}_G(1, v) \leq d^* - 1, \\ d^* + \mathbf{d}_G(v, A) & \text{if } v \in \bar{V} \setminus \Gamma_{\leq d^*-1}. \end{cases} \quad (132)$$

as the distance vector of any  $T$  in the support of  $\mu_{G,\infty}|_A$ . Also, we write

$$N_d^A := |\{v : \mathbf{d}_A(v) = d\}|. \quad (133)$$

**Lemma 77** *Conditioned on  $A \subseteq \Gamma_{d^*}$ , for any  $t, \epsilon > 0$ , we have*

$$\mathbb{E}_{\mu_{G,\infty}|_A} \left[ \sum_{v \in A} \left| \text{ch}_T(v) - \frac{N_{d^*+1}^A}{|A|} \right| \right] \leq t + 2n\epsilon$$

with probability at least

$$1 - 2^{|A|} \epsilon^{-1} \exp \left( -\frac{t^2}{2(n - |A|)} \right).$$

**Proof** For notational convenience we define

$$f_{G,A}(T) := \sum_{v \in A} \left| \text{ch}_T(v) - \frac{N_{d^*+1}^A}{|A|} \right|.$$

We first note that

$$\sum_{v \in A} |\text{ch}_T(v)| = N_{d^*+1}^A.$$

Hence,

$$\begin{aligned} f_{G,A}(T) &= \sum_{v \in A} \left| \text{ch}_T(v) - \frac{N_{d^*+1}^A}{|A|} \right| \\ &= 2 \max_{B \subseteq A} \sum_{v \in B} \left( \left| \text{ch}_T(v) - \frac{N_{d^*+1}^A}{|A|} \right| \right) \\ &= 2 \max_{B \subseteq A} \left\{ \sum_{v \in B} |\text{ch}_T(v)| - \frac{|B| N_{d^*+1}^A}{|A|} \right\} \\ &= 2 \max_{B \subseteq A} \left\{ \sum_{u \in \Gamma_{d^*+1}^A} \mathbf{1}_{\text{par}_T(u) \in B} - \frac{|B| N_{d^*+1}^A}{|A|} \right\} \end{aligned}$$

Conditioned on  $\Gamma_{d^*}$ ,  $A \subseteq \Gamma_{d^*}$ , and  $\Gamma_{d^*+1}^A$ , the events  $\text{par}_T(u) \in B$  are independent over all  $u \in \Gamma_{d^*+1}^A$  and have probability  $|B|/|A|$ . Hence,  $\sum_{u \in \Gamma_{d^*+1}^A} \mathbf{1}_{\text{par}_T(u) \in B}$  is conditionally  $\text{Binom}(N_{d^*+1}^A, |B|/|A|)$ . Thus, Hoeffding's inequality gives

$$\mathbb{E} \left[ \mu_{G,\infty|A} \left( \left\{ T : \sum_{u \in \Gamma_{d^*+1}^A} \mathbf{1}_{\text{par}_T(u) \in B} - \frac{|B|N_{d^*+1}^A}{|A|} \geq t \right\} \right) \middle| A \subseteq \Gamma_{d^*}, N_{d^*+1}^A \right] \leq \exp \left( -\frac{2t^2}{N_{d^*+1}^A} \right).$$

Taking the union bound, we get

$$\begin{aligned} \mathbb{E} [\mu_{G,\infty|A}(\{T : f_{G,A}(T) \geq t\}) \mid A \subseteq \Gamma_{d^*}, N_{d^*+1}^A] &\leq 2^{|A|} \exp \left( -\frac{t^2}{2N_{d^*+1}^A} \right) \\ &\leq 2^{|A|} \exp \left( -\frac{t^2}{2(n-|A|)} \right) \end{aligned}$$

By Markov's inequality, we have

$$\Pr(\mu_{G,\infty|A}(\{T : f_{G,A}(T) \geq t\}) \geq \epsilon \mid A \subseteq \Gamma_{d^*}) \leq 2^{|A|} \epsilon^{-1} \exp \left( -\frac{t^2}{2(n-|A|)} \right).$$

Note that

$$\begin{aligned} f_{G,A}(T) &\leq \sum_{v \in A} |\text{ch}_T(v)| + \sum_{v \in A} \frac{N_{d^*+1}^A}{|A|} \\ &= 2N_{d^*+1}^A \\ &\leq 2n. \end{aligned}$$

Thus,  $\mu_{G,\infty|A}(\{T : f_{G,A}(T) \geq t\}) < \epsilon$  implies

$$\mathbb{E}_{\mu_{G,\infty|A}}[f_{G,A}(T)] < t + 2n\epsilon.$$

This gives

$$\Pr \left( \mathbb{E}_{\mu_{G,\infty|A}}[f_{G,A}(T)] \geq t + 2n\epsilon \mid A \subseteq \Gamma_{d^*} \right) \leq 2^{|A|} \epsilon^{-1} \exp \left( -\frac{t^2}{2(n-|A|)} \right)$$

which is as desired. ■

With this lemma, we prove our main result by analyzing (131).

**Proof** [Proof of Theorem 75] Let  $m_2$  be defined in (130). By (111), we may assume  $m_\ell \leq m_2$  which holds asymptotically almost surely. We consider a measure  $\tilde{\mu}_{G,\beta} \ll_{[m_\ell, m_2]}$  defined by  $\tilde{\mu}_{G,\beta}$  conditioned on  $m_\ell \leq |\varphi(T)| \leq m_2$ . In other words,

$$\tilde{\mu}_{G,\beta} \ll_{[m_\ell, m_2]}(T) := \sum_{A \in \mathcal{K}(G)} \mu_{G,\beta}(\varphi(T) = A \mid m_\ell \leq |\varphi(T)| \leq m_2) \cdot \mu_{G,\infty|A}(T).$$

Lemma 71, Lemma 76, Proposition 46, and the convexity of the KL divergence give

$$\text{KL}(\tilde{\mu}_{G,\beta} \ll_{[m_\ell, m_2]}, \mu_{G,\beta}) = o_p(n)$$

and Proposition 70, Lemma 71, and Lemma 76 give

$$W_1(\tilde{\mu}_{G,\beta}|_{[m_\ell, m_2]}, \mu_{G,\beta}) = o_p(n).$$

Now our goal is to prove that  $\mu_{G,\infty}$  and  $\tilde{\mu}_{G,\beta}|_{[m_\ell, m_2]}$  are far apart in Wasserstein distance. Thus, it suffices to find a 1-Lipschitz function  $f$  such that

$$\mathbb{E}_{\mu_{G,\infty}|_A}[f(T)] - \mathbb{E}_{\mu_{G,\infty}}[f(T)] = \Theta_p(n)$$

simultaneously for all  $A \in \mathcal{K}(G)$  with  $m_\ell \leq |A| \leq m_2$ .

We define a function  $f$  on the set of spanning trees of  $\bar{G}$  by

$$f(T) := \sum_{v \in \Gamma_{d^*}} \left| |\text{ch}_T(v)| - \frac{N_{d^*+1}}{N_{d^*}} \right|.$$

Since removing or adding an edge in  $T$  changes one of  $\text{ch}_T(v)$  by at most 1,  $f$  is 1-Lipschitz. Using Lemma 77 with  $t = \sqrt{2N_{d^*}(n - N_{d^*})}$  and  $\epsilon = e^{-N_{d^*}/8}$ , we get by Theorem 17

$$\begin{aligned} \mathbb{E}_{\mu_{G,\infty}}[f(T)] &\leq \sqrt{2N_{d^*}(n - N_{d^*})} + o_p(n) \\ &= \sqrt{2 \cdot \frac{N_{d^*}}{n} \cdot \left(1 - \frac{N_{d^*}}{n}\right) \cdot n} + o_p(n) \\ &= \sqrt{2\lambda_n(1 - \lambda_n)} \cdot n + o_p(n) \\ &= \sqrt{2 \cdot \frac{\kappa_n}{\log \log n} \left(1 - \frac{\kappa_n}{\log \log n}\right) \cdot n} + o_p(n) \end{aligned} \tag{134}$$

conditioned on  $N_{d^*}$  with probability at least

$$\begin{aligned} 1 - 2^{N_{d^*}} \epsilon^{-1} \exp\left(-\frac{t^2}{2(n - N_{d^*})}\right) &= 1 - 2^{N_{d^*}} e^{N_{d^*}/8} e^{-N_{d^*}} \\ &\geq 1 - e^{-N_{d^*}/8} \\ &= 1 - o_p(1). \end{aligned}$$

Next, we analyze  $\mathbb{E}_{\mu_{G,\infty}|_A}[f(T)]$ . We write

$$\begin{aligned} \mathbb{E}_{\mu_{G,\infty}|_A}[f(T)] &= \mathbb{E}_{\mu_{G,\infty}|_A} \left[ \sum_{v \in \Gamma_{d^*}} \left| |\text{ch}_T(v)| - \frac{N_{d^*+1}}{N_{d^*}} \right| \right] \\ &\geq \mathbb{E}_{\mu_{G,\infty}|_A} \left[ \sum_{v \in A} \left| |\text{ch}_T(v)| - \frac{N_{d^*+1}}{N_{d^*}} \right| \right] \\ &\geq |A| \left| \frac{N_{d^*+1}^A}{|A|} - \frac{N_{d^*+1}}{N_{d^*}} \right| - \mathbb{E}_{\mu_{G,\infty}|_A} \left[ \sum_{v \in A} \left| |\text{ch}_T(v)| - \frac{N_{d^*+1}^A}{|A|} \right| \right] \\ &= \left| N_{d^*+1}^A - |A| \cdot \frac{N_{d^*+1}}{N_{d^*}} \right| - \mathbb{E}_{\mu_{G,\infty}|_A} \left[ \sum_{v \in A} \left| |\text{ch}_T(v)| - \frac{N_{d^*+1}^A}{|A|} \right| \right]. \end{aligned}$$

For the first term in the RHS, Lemma 54 gives

$$\begin{aligned}
 \left| N_{d^*+1}^A - |A| \cdot \frac{N_{d^*+1}}{N_{d^*}} \right| &\geq N_{d^*+1}^A - |A| \cdot \frac{N_{d^*+1}}{N_{d^*}} \\
 &\geq n - |A| - |A| \cdot \frac{N_{d^*+1}}{N_{d^*}} - o_p\left(\frac{n}{\log \log n}\right) \\
 &\geq n \left(1 - \frac{|A|}{N_{d^*}}\right) - o_p\left(\frac{n}{\log \log n}\right) \\
 &\geq \frac{6}{7} \cdot \frac{\delta \kappa_n}{1 + \delta \kappa_n} \cdot n - o_p\left(\frac{n}{\log \log n}\right)
 \end{aligned} \tag{135}$$

where the last inequality is due to  $|A| \leq m_2$ . Also, applying Lemma 77 with  $t = n/(\log \log n)^{1/4}$  and  $\epsilon = 1/n$  gives

$$\begin{aligned}
 \mathbb{E}_{\mu_{G,\infty}|A} \left[ \sum_{v \in A} \left| \text{ch}_T(v) - \frac{N_{d^*+1}^A}{|A|} \right| \right] &\leq \frac{n}{(\log \log n)^{1/4}} + 2 \\
 &= o(n)
 \end{aligned} \tag{136}$$

with probability at least

$$\begin{aligned}
 1 - 2^{|A|} \epsilon^{-1} \exp\left(-\frac{t^2}{2(n-|A|)}\right) &\geq 1 - \exp\left(-\frac{t^2}{2n} + |A| + \log(1/\epsilon)\right) \\
 &\geq 1 - \exp\left(-\frac{n}{2\sqrt{\log \log n}} + m_2 + \log n\right).
 \end{aligned}$$

Note that (135) and (136) imply

$$\mathbb{E}_{\mu_{G,\infty}|A} [f(T)] \geq \frac{\delta \kappa_n / 2}{1 + \delta \kappa_n / 2} \cdot n - o_p(n). \tag{137}$$

Now we will apply the union bound for all  $A \in \mathcal{K}(G)$  with  $m_\ell \leq |A| \leq m_2$ . Note that

$$\begin{aligned}
 |\{A \in \mathcal{K}(G) : m_\ell \leq |A| \leq m_2\}| &\leq \sum_{m=m_\ell}^{m_2} \binom{n}{m} \\
 &\leq \sum_{m=m_\ell}^{m_2} \left(\frac{en}{m}\right)^m \\
 &\leq n \left(\frac{en}{m_\ell}\right)^{m_2} \\
 &\leq n (3e(\log \log n)^2)^{m_2}.
 \end{aligned}$$

Hence, it follows that (137) holds simultaneously for all  $A \in \mathcal{K}(G)$  with  $m_\ell \leq |A| \leq m_2$  with probability at least

$$1 - \exp\left(-\frac{n}{2\sqrt{\log \log n}} + m_2 + 2 \log n + m_2 \log (3e(\log \log n)^2)\right).$$

By (111), this is

$$1 - \exp\left(-\frac{n}{2\sqrt{\log \log n}} + o_p\left(\frac{n}{\sqrt{\log \log n}}\right)\right) = 1 - o_p(1).$$

Hence, asymptotically almost surely, we have both (134) and (137). This implies

$$\mathbb{E}_{\mu_{G,\infty}|A}[f(T)] - \mathbb{E}_{\mu_{G,\infty}}[f(T)] \geq \left(\frac{6}{7} \cdot \frac{\delta \kappa_n}{1 + \delta \kappa_n} - \sqrt{2 \cdot \frac{\kappa_n}{\log \log n} \left(1 - \frac{\kappa_n}{\log \log n}\right)}\right) n - o_p(n) \quad (138)$$

simultaneously for all  $A \in \mathcal{K}(G)$  with  $m_\ell \leq |A| \leq m_2$ .

Now it remains to prove that

$$\frac{6}{7} \cdot \frac{\delta \kappa_n}{1 + \delta \kappa_n} - \sqrt{2 \cdot \frac{\kappa_n}{\log \log n} \left(1 - \frac{\kappa_n}{\log \log n}\right)}$$

is bounded below by a constant. First, recall our assumption  $\beta < 1 - \Delta_n - \kappa_n^{-1} - \delta$  which implies  $\kappa_n > 1$ . This gives

$$\sqrt{2 \cdot \frac{\kappa_n}{\log \log n} \left(1 - \frac{\kappa_n}{\log \log n}\right)} \leq \sqrt{\frac{2}{\log \log n}} \cdot \kappa_n. \quad (139)$$

On the other hand, the AM–GM inequality gives

$$\sqrt{2 \cdot \frac{\kappa_n}{\log \log n} \left(1 - \frac{\kappa_n}{\log \log n}\right)} \leq \frac{1}{\sqrt{2}}. \quad (140)$$

Now we consider two cases. If  $\delta \kappa_n < 7$ , then for all sufficiently large  $n$  such that  $\sqrt{2/\log \log n} \leq \delta/28$ , (139) gives

$$\begin{aligned} \frac{6}{7} \cdot \frac{\delta \kappa_n}{1 + \delta \kappa_n} - \sqrt{2 \cdot \frac{\kappa_n}{\log \log n} \left(1 - \frac{\kappa_n}{\log \log n}\right)} &\geq \frac{3\delta \kappa_n}{28} - \sqrt{\frac{2}{\log \log n}} \cdot \kappa_n \\ &\geq \frac{\delta \kappa_n}{14} \\ &\geq \frac{\delta}{14}. \end{aligned}$$

Otherwise, if  $\delta \kappa_n \geq 7$ , then (140) gives

$$\frac{6}{7} \cdot \frac{\delta \kappa_n}{1 + \delta \kappa_n} - \sqrt{2 \cdot \frac{\kappa_n}{\log \log n} \left(1 - \frac{\kappa_n}{\log \log n}\right)} \geq \frac{3}{4} - \frac{1}{\sqrt{2}}. \quad (141)$$

This completes the proof. ■

### F.1.6. REPLICA SYMMETRY

As a consequence of our analysis, we show that our system  $\mu_{G,\beta}$  is replica symmetric in the sense of the following theorem. Namely, this states that whenever  $\lambda \in \{0, 1\}$  or the temperature is bounded away from the critical temperature, then the overlap of two independently sampled trees concentrates.

**Theorem 78 (Replica symmetry)** *Let  $\beta > 0$  be fixed and assume  $\lambda_n \rightarrow \lambda \in [0, 1]$ . Suppose that for given  $\{G_n\}$  we independently sample  $T_n, T'_n \sim \mu_{G_n,\beta}$ . Then*

$$R(T_n, T'_n) \xrightarrow{p} \begin{cases} 0 & \text{if } \lambda \in \{0, 1\} \text{ or } \beta < 1, \\ f(\lambda) & \text{if } \lambda \in (0, 1) \text{ and } \beta > 1 \end{cases}$$

where

$$f(\lambda) = (1 - \lambda)\mathbb{E}[1/X \mid X > 0], \quad X \sim \text{Poisson}(\log(1/\lambda)).$$

**Proof** First, consider the case  $\lambda \in \{0, 1\}$  or  $\beta < 1$ . By Corollary 72, we may assume  $T_n, T'_n \sim \tilde{\mu}_{G,\beta}$ . In addition, by Lemma 71, it suffices to show that for any  $\epsilon > 0$

$$\Pr(R(T, T') > \epsilon \mid G, T', |\varphi(T)| \geq m_\ell) = o_p(1)$$

for any spanning tree  $T'$  of  $\overline{G}$ . Note that

$$\Pr(R(T, T') > \epsilon \mid G, T', |\varphi(T)| \geq m_\ell) \leq \max_{A \in \mathcal{K}(G): |A| \geq m_\ell} \Pr(R(T, T') > \epsilon \mid G, T', \varphi(T) = A). \quad (142)$$

Thus, our task is to bound  $\Pr(R(T, T') > \epsilon \mid G, T', \varphi(T) = A)$  where it is guaranteed that  $|A| \geq m_\ell$ . By our assumption, the law of  $T$  conditioned on  $\varphi(T) = A$  is  $\mu_{G,\infty}|_A$  which is a product measure. Define

$$\text{par}_A(v) := \{u \in \overline{V} : d_A(1, u) = d_A(1, v) + 1\}$$

where  $d_A$  is defined in (132). Then  $R(T, T')$  can be written as the mean of (conditionally) independent random variables

$$R(T, T') = \frac{1}{|\overline{V}| - 1} \sum_{v \in \overline{V} \setminus \{1\}} \mathbf{1}_{\text{par}_T(v) = \text{par}_{T'}(v)} \quad (143)$$

where

$$\mathbf{1}_{\text{par}_T(v) = \text{par}_{T'}(v)} \sim \text{Bernoulli} \left( \frac{\mathbf{1}_{\text{par}_{T'}(v) \in \text{par}_A(v)}}{|\text{par}_A(v)|} \right).$$

Since we need an upper bound, we can upper bound this using the sum of Bernoulli( $1/|\text{par}_A(v)|$ ). Thus, the quantity of interest here is the mean

$$p_{G,A} := \frac{1}{|\overline{V}| - 1} \sum_{v \in \overline{V} \setminus \{1\}} \frac{1}{|\text{par}_A(v)|}.$$

In light of (142), it suffices to show that with high probability,  $p_{G,A} = o(1)$  simultaneously for all  $A \in \mathcal{K}(G)$  with  $|A| \geq m_\ell$ . By Theorem 17, with high probability,

$$\begin{aligned} (|\bar{V}| - 1)p_{G,A} &= \sum_{v \in \bar{V} \setminus \{1\}} \frac{1}{|\text{par}_A(v)|} \\ &\leq \sum_{v \in A} \frac{1}{|\text{par}_G(v)|} + \sum_{v \in \bar{V} \setminus \Gamma_{\leq d^*-1} \setminus A} \frac{1}{|N_A(v)| \vee 1} + o(n). \end{aligned}$$

We bound the two sums in the RHS separately, starting with the first sum. If  $\lambda \in (0, 1)$  and  $\beta < 1$ , then we are in the high temperature phase (HT) so Lemma 76 we may assume  $|A| = o(n)$  for all  $A$ , which immediately yields  $\sum_{v \in A} 1/|\text{par}_G(v)| \leq |A| = o_p(n)$ . otherwise, if  $\lambda \in \{0, 1\}$ , then we use

$$\sum_{v \in A} \frac{1}{|\text{par}_G(v)|} \leq \sum_{v \in \Gamma_{d^*}} \frac{1}{|\text{par}_G(v)|}.$$

Recall that conditioned on  $\Gamma_{\leq d^*-1}$ ,  $|\text{par}_G(v)|$  are (conditionally) i.i.d. following  $\text{Binom}(N_{d^*-1}, q)$ . If  $\lambda = 0$ , then  $qN_{d^*-1} \xrightarrow{p} \infty$  so it is straightforward to see  $\sum_{v \in \Gamma_{d^*}} 1/|\text{par}_G(v)| = o_p(n)$ , and if  $\lambda = 1$ , then  $N_{d^*} = o_p(n)$  so the same bound obviously holds. Now it remains to deal with the term

$$\sum_{v \in \bar{V} \setminus \Gamma_{\leq d^*-1} \setminus A} \frac{1}{|N_A(v)| \vee 1}. \quad (144)$$

Replacing  $A$  with any of its superset decreases the sum, so it suffices to show that this is  $o(n)$  simultaneously for all  $A$  of size exactly  $\lceil \frac{n}{3(\log \log n)^2} \rceil$ . Since  $|N_A(v)|$  are i.i.d.  $\text{Binom}(|A|, q)$  conditioned on  $\Gamma_{\leq d^*-1}$  and  $A$ , we define

$$g(m, p) = \mathbb{E} \left[ \frac{1}{X \vee 1} \right], \quad X \sim \text{Binom}(m, p)$$

so that by Hoeffding's inequality

$$\Pr \left( \sum_{v \in \bar{V} \setminus \Gamma_{\leq d^*-1} \setminus A} \frac{1}{|N_A(v)| \vee 1} \geq g(|A|, q) + t \mid \Gamma_{\leq d^*-1}, A \right) \leq e^{-\frac{2t^2}{n}}.$$

Since  $|A| \geq m_\ell$  we have  $|A|q \rightarrow \infty$  so  $g(|A|, q) = o(n)$ , and by plugging in  $t = n/\log \log n$ , (93) gives that (144) is  $o_p(n)$  as desired.

Now assume  $\lambda \in (0, 1)$  and  $\beta > 1$ . Then we are in the low temperature phase (LT), so by Theorem 73 we may assume  $T, T' \sim \mu_{G,\infty}$ . Then (143) is the sum of independent random variables  $\mathbf{1}_{\text{par}_T(v)=\text{par}_{T'}(v)} \sim \text{Bernoulli}(1/|\text{par}_G(v)|)$ . By the same argument, it is straightforward to see that this can be written as

$$\frac{1}{|\bar{V}| - 1} \sum_{v \in \Gamma_{d^*}} \frac{1}{|\text{par}_G(v)|} = \frac{N_{d^*}}{|\bar{V}| - 1} \cdot \frac{1}{N_{d^*}} \sum_{v \in \Gamma_{d^*}} \frac{1}{|\text{par}_G(v)|} + o_p(1).$$

From Theorem 17 we have  $N_{d^*}/(|\bar{V}| - 1) \xrightarrow{p} 1 - \lambda$  and  $\frac{1}{N_{d^*}} \sum_{v \in \Gamma_{d^*}} \frac{1}{|\text{par}_G(v)|}$  concentrates around

$$\mathbb{E} \left[ \sum_{v \in \Gamma_{d^*}} \frac{1}{|\text{par}_G(v)|} \middle| \Gamma_{\leq d^*-1}, \Gamma_{d^*} \right]$$

which again converges to  $f(\lambda)$ , since,  $\text{Binom}(N_{d^*-1}, q) \xrightarrow{p} \text{Poisson}(\log(1/\lambda))$ . ■

## F.2. Phase transitions

Our results have shown that, both in KL divergence and Wasserstein distance, the Gibbs measures behave quite differently in the low temperature phase and the high temperature phase. This is fundamentally a phase transition widely studied in statistical physics. In this section, we further investigate the nature of phase transitions occurring in our model of shortest paths in sparse random graphs.

### F.2.1. THE FREE ENERGY DENSITY

The usual notion of *free energy* in statistical physics is the negative log partition function multiplied by the temperature. In our context, the free energy can be expressed as

$$-\bar{\beta}^{-1} \Phi_{G,\beta} = -\frac{1}{\beta \log \log n} \log Z_{G,\beta}.$$

Recall that by Theorem 66 the ground state energy is approximately  $(d^* + \lambda_n)n$  so it is reasonable to subtract off this term to see more precisely how the free energy changes relative to the ground state. As such, we define the relative free energy density as

$$\mathcal{F}_{G,\beta} := -\frac{1}{\beta n \log \log n} \log Z_{G,\beta} - (d_n^* + \lambda_n).$$

We first establish a formula for the limiting free energy density.

**Corollary 79** *Let  $\beta > 0$ ,  $\Delta \in [0, 1]$ , and suppose that  $\Delta_n \rightarrow \Delta$  and  $\lambda_n \rightarrow \lambda$ . Then the free energy density converges in probability to*

$$\mathcal{F}_{G_n,\beta} \xrightarrow{p} \mathcal{F}_{\lambda,\Delta}(\beta)$$

where  $\mathcal{F}_{\lambda,\Delta}$  is a continuous function of  $\beta$  and is one of the following.

(A) If  $\Delta = 1$ , then  $\lambda = 0$  and

$$\mathcal{F}_{\lambda,\Delta}(\beta) = -\frac{1}{\beta}.$$

(B) If  $\Delta \in [0, 1)$  and  $\lambda \in [0, 1)$ , then

$$\mathcal{F}_{\lambda,\Delta}(\beta) = \begin{cases} -\frac{\lambda + \Delta}{\beta} & \text{if } \beta \geq 1 - \Delta, \\ \frac{1 - (\lambda + \Delta)}{1 - \Delta} - \frac{1}{\beta} & \text{if } \beta \leq 1 - \Delta. \end{cases}$$

(C) If  $\lambda = 1$ , then  $\Delta = 0$  and

$$\mathcal{F}_{\lambda,\Delta}(\beta) = -\frac{1}{\beta}.$$

If we directly use Theorem 42 to prove this, we miss the case  $\beta = 1 - \Delta$ . Instead, it is easier and perhaps more illuminating to start from more primitive results, namely Lemma 49 and Lemma 55. Together with (107), it is not difficult to deduce the following corollaries.

**Lemma 80** *With probability at least  $1 - o(1)$ , we have*

$$\frac{1}{\log \log n} \log(Z_{G,\beta}|_A) \leq n - (1 - \Delta - \beta)|A| - \beta(d^* + 1)n + o(n)$$

*simultaneously for all  $A \in \mathcal{K}(G)$ .*

**Lemma 81** *With probability at least  $1 - o(1)$ , we have*

$$\frac{1}{\log \log n} \log(Z_{G,\beta}|_A) \geq n - (1 - \Delta - \beta)|A| - \beta(d^* + 1)n - o(n)$$

*for all  $A \in \mathcal{K}(G)$  with  $|A| \geq m_\ell$ .*

From Section E.3.7, we already know that we only need to consider kernels of size at least  $m_\ell$  to approximate  $\log Z_{G,\beta}$ . Thus,

$$\frac{1}{\log \log n} \log Z_{G,\beta} = \max_{m_\ell \leq m \leq N_{d^*}} \{n - (1 - \Delta - \beta)m - \beta(d^* + 1)n\} + o_p(n).$$

The objective we are maximizing over is simply a linear function, so this is easy to solve and yields Corollary 79.

For the case  $\Delta \in [0, 1)$  and  $\lambda \in [0, 1)$ , the free energy density as a function of  $\beta$  has a non-differentiable point  $\beta = 1 - \Delta$ , indicating the existence of *first-order phase transition*. Interestingly, zooming in the free energy by  $\log \log n$  reveals another form of phase transition. This is easily obtained as a direct consequence of Theorem 42, simply by splitting the formula into different parameter regimes based on the asymptotics of  $\Delta_n$ ,  $\lambda_n$ , and  $\kappa_n$ . Note that the ground state energy is still  $(d^* + \lambda_n)n$  even after zooming in, since the error term in Theorem 66 is  $o_p(n/\log \log n)$ .

**Corollary 82** *Let  $\beta > 0$ ,  $\Delta \in [0, 1]$ ,  $\lambda \in [0, 1]$ , and  $\kappa \geq 0$  be constants. Suppose that  $\Delta_n \rightarrow \Delta$  and  $\lambda_n \rightarrow \lambda$ . Then*

$$\mathcal{F}_{G_n,\beta} \log \log n = -\frac{1}{\beta} \zeta_{G_n,\beta} + o_p(1)$$

*where  $\zeta_{G_n,\beta}$  can be one of the following.*

(A) If  $\Delta = 1$ , then  $\lambda = 0$  and

$$\zeta_{G_n,\beta} = (\lambda_n + \Delta_n) \log \log n + \log \alpha_n.$$

(B) We further split the case  $\Delta \in [0, 1)$  and  $\lambda \in [0, 1)$  into the following.

(B.i) If  $\Delta \in [0, 1)$  and  $\lambda = 0$ , then

$$\zeta_{G_n, \beta} = \begin{cases} (\lambda_n + \Delta_n) \log \log n + \Delta \log \alpha_n & \text{if } \beta > 1 - \Delta, \\ (1 - \beta + \beta \lambda_n) \log \log n - \log \log \log n + \log \left( \frac{\alpha_n}{1 - \Delta - \beta} \right) - 1 & \text{if } \beta < 1 - \Delta. \end{cases}$$

(B.ii) If  $\Delta = 0$ ,  $\lambda \in (0, 1)$ , then

$$\zeta_{G_n, \beta} = \begin{cases} \lambda_n \log \log n + \lambda \log((1 - \lambda)\alpha_n) + \Psi_0(\lambda) & \text{if } \beta > 1, \\ (1 - \beta + \beta \lambda_n) \log \log n - \log \log \log n + \log \left( \frac{\alpha_n}{1 - \beta} \right) - 1 & \text{if } \beta < 1. \end{cases}$$

(C) If  $\lambda = 1$ , then  $\Delta = 0$  and

$$\zeta_{G_n, \beta} = \begin{cases} \log \log n - \log \log \log n + \log(\kappa_n \alpha_n) - \kappa_n & \text{if } \beta > 1 - \kappa_n^{-1}, \\ \log \log n - \log \log \log n + \log \left( \frac{\alpha_n}{1 - \beta} \right) - \kappa_n \beta - 1 & \text{if } \beta < 1 - \kappa_n^{-1} \end{cases}$$

where  $\kappa_n \in [0, \infty]$ .

(C.i) If  $\kappa = \infty$ , then  $\kappa^{-1}$  is replaced with zero.

(C.ii) If  $\kappa \in (1, \infty)$ , then  $\kappa_n$  in the formula can be replaced with  $\kappa$ .

(C.iii) If  $\kappa \in [0, 1]$ , then  $\zeta_{G_n, \beta}$  can be set to either formula for any value of  $\beta > 0$ , and can equivalently be written as

$$\zeta_{G_n, \beta} = (\lambda_n + \Delta_n) \log \log n + \log \alpha_n.$$

The regimes (A), (B), and (C) are set to match the corresponding cases in Corollary 79. The most interesting regimes are (C.i) and (C.ii), which are something new that is not visible only from Corollary 79. An especially intriguing regime is (C.ii), which turns out to exhibit a *second-order phase transition* which we will further discuss in the following section.

### F.2.2. PHASE TRANSITIONS IN WASSERSTEIN SPACE

An important feature of the Gibbs measures in the regimes (B), (C.i), and (C.ii) is that as temperature increases (i.e.,  $\beta$  decreases), the distance between  $\mu_{G, \beta}$  and the uniform measure  $\mu_{G, \infty}$  suddenly changes. As such, we can examine the phase transition phenomena by viewing the Gibbs measures  $\mu_{G, \beta}$  as a function  $\beta \mapsto \mu_{G, \beta}$  to the space of measures on spanning trees.

Formally, for a fixed graph  $G$ , we let  $\mathcal{P}_1(\mathcal{T}_G)$  to be the space of probability measures on the spanning trees of the component  $\overline{G}$  of  $G$  containing a fixed vertex 1, equipped with the normalized 1-Wasserstein metric

$$w_1(\mu, \nu) := \frac{1}{n} W_1(\mu, \nu), \quad \mu, \nu \in \mathcal{P}_1(\mathcal{T}_G).$$

Also, we define the *critical temperature*  $\beta_c$  as

$$\beta_c := \lim_{n \rightarrow \infty} (1 - \Delta_n - \kappa_n^{-1})$$

whenever the limit exists and is positive. In particular,

$$\beta_c = \begin{cases} 1 - \Delta & \text{if } \Delta_n \rightarrow \Delta \in [0, 1) \text{ and } \lambda_n \rightarrow 0, \\ 1 & \text{if } \lambda_n \rightarrow \lambda \in (0, 1] \text{ and } \kappa_n \rightarrow \infty, \\ 1 - 1/\kappa & \text{if } \kappa_n \rightarrow \kappa \in (1, \infty). \end{cases}$$

Our system associated with  $G$  can be completely described by the function  $f_G : \mathbb{R}_{>0} \rightarrow \mathcal{P}_1(\mathcal{T}_G)$ ,  $\beta \mapsto \mu_{G,\beta}$ . Then Corollary 82 implies that

$$w_1(f_G(\beta), \mu_{G,\infty}) = \frac{1}{n} W_1(\mu_{G,\beta}, \mu_{G,\infty}) \xrightarrow{p} 0$$

if  $\beta > \beta_c$ ; in other words,  $f_G$  is asymptotically almost surely constant in  $\beta > \beta_c$ . However, the distance to  $\mu_{G,\infty}$  is bounded away from zero when  $\beta < \beta_c$  — a non-analyticity in  $f_G$  occurs at  $\beta_c$  in the thermodynamic limit, indicating a phase transition.

In fact, the phase transition in the regime (C.ii) is somewhat different from that in the regimes (B) and (C.i). One way to see this is to consider the *optimal kernel size* as denoted by  $m^*$  in Theorem 42. In other words, we look at the kernel size that minimizes the free energy. Since there is no significant difference between (B) and (C.i), for simplicity, we shall focus on comparing (C.i) and (C.ii).

- In the regime (C.i), we have  $\Delta_n \rightarrow 0$  and  $\kappa_n \rightarrow \infty$ . In the low temperature phase, the optimal kernel proportion  $m^*/N_{d^*} = 1$ , but in the high temperature phase, the optimal kernel proportion is approximately

$$\frac{m^*}{N_{d^*}} \approx \frac{1}{(1 - \Delta_n - \beta)\kappa_n} \rightarrow 0.$$

Hence, as  $\beta$  decreases from above  $\beta_c = 1$  to below  $\beta_c = 1$ , the optimal kernel proportion suddenly drops from 1 to 0 *discontinuously*.

- In the regime (C.ii), we also have  $\Delta_n \rightarrow 0$  but  $\kappa_n \rightarrow \kappa \in (1, \infty)$ . The optimal kernel proportion in the low temperature phase is the same, but in the high temperature phase we have

$$\frac{m^*}{N_{d^*}} \approx \frac{1}{(1 - \Delta_n - \beta)\kappa_n} \rightarrow \frac{1}{(1 - \beta)\kappa}.$$

Note that  $\beta_c = 1 - 1/\kappa$  and thus its left limit  $\beta \rightarrow \beta_c - 0$  is 1. Hence, the optimal kernel proportion changes *continuously*.

This observation can be adapted to our Wasserstein framework in the following way. For a sequence of graphs  $\{G_n\}_{n \in \mathbb{Z}^+}$  with  $|V(G_n)| = n$ , we get a sequence of functions  $\{f_{G_n}\}_{n \in \mathbb{Z}^+}$  where  $f_{G_n} : \mathbb{R}_{>0} \rightarrow \mathcal{P}(\mathcal{T}_{G_n})$  and we check if it is *equicontinuous* at  $\beta_c$ . This is formally stated and proved in the following results.

**Proposition 83 (Discontinuous phase transition)** *In regimes (B) and (C.i), there exists a universal constant  $C > 0$  such that for all constant  $\delta > 0$ , we can find a constant  $\beta > 0$  with  $|\beta - \beta_c| < \delta$  satisfying*

$$\frac{1}{n} W_1(\mu_{G_n,\beta}, \mu_{G_n,\beta_c}) \geq C$$

*asymptotically almost surely.*

**Proof** Choose  $C = 1/50$ . Set  $\beta_1 = \beta_c + \delta/2$  and  $\beta_2 = \beta_c - \delta/2$  which are in (LT) and (HT), respectively. By Theorem 73, we have

$$\frac{1}{n}W_1(\mu_{G,\beta_1}, \mu_{G,\infty}) = o_p(1).$$

For  $\beta_2$ , we note that in the proof of Theorem 75, (138) and (141) imply that

$$\frac{1}{n}W_1(\mu_{G,\beta_2}, \mu_{G,\infty}) \geq \frac{3}{4} - \frac{1}{\sqrt{2}} - o_p(1)$$

if  $(\delta/2)\kappa_n \geq 7$ . Since  $\kappa_n \rightarrow \infty$  in regimes (B) and (C.i), for any  $\delta > 0$  this holds for sufficiently large  $n$ . Thus, we have that

$$\begin{aligned} \frac{1}{n}W_1(\mu_{G,\beta_1}, \mu_{G,\beta_2}) &\geq \frac{3}{4} - \frac{1}{\sqrt{2}} - o_p(1) \\ &\geq 2C \end{aligned}$$

asymptotically almost surely. Hence, at least one of  $\beta_1$  or  $\beta_2$  satisfies the desired condition.  $\blacksquare$

**Proposition 84 (Continuous phase transition)** *In regime (C.ii), for any constant  $\delta > 0$ , any  $\beta > 0$  with  $|\beta - \beta_c| < \delta$  satisfies*

$$\frac{1}{n}W_1(\mu_{G_n,\beta}, \mu_{G_n,\beta_c}) \leq \kappa\sqrt{12\delta}$$

*asymptotically almost surely.*

**Proof** We set  $\delta = \frac{\epsilon^2}{12\kappa^2}$ . The case  $\beta_c \leq \beta < \beta_c + \delta$  is contained in (LT), so by Theorem 73

$$\frac{1}{n}W_1(\mu_{G,\beta_c}, \mu_{G,\infty}) = o_p(1)$$

and the conclusion follows. Thus, it remains to handle the case  $\beta_c - \delta < \beta < \beta_c$ . Define

$$m_3 := m_\ell \vee \left\lfloor \left(1 - \frac{\epsilon^2}{3}\kappa_n^{-1}\right) N_{d^*} \right\rfloor.$$

We claim that

$$\mu_{G,\beta}(\{T : |\varphi(T)| \leq m_3\}) \leq \exp(-\Omega(n)). \quad (145)$$

Before we show this, we first see how (145) yields the conclusion. By following the exact same argument with the proof of Theorem 73, (129) gives

$$\frac{1}{n}W_1(\mu_{G,\beta}, \mu_{G,\infty}) \leq \sqrt{\frac{2}{3}\epsilon^2} + o_p(1) < \epsilon$$

asymptotically almost surely. Thus, it suffices to prove (145). Let

$$\delta' := \frac{\epsilon^2}{12\kappa} \wedge \frac{1}{2}.$$

This time, we follow the argument in the proof of Lemma 76. From (112) we have

$$\begin{aligned} g\left(\frac{\kappa_n^{-1}N_{d^*}}{1-\Delta_n-\beta}\right) - g\left(\frac{(1-\delta')\kappa_n^{-1}N_{d^*}}{1-\Delta_n-\beta}\right) &\geq (-\delta' - \log(1-\delta'))n - o_p(n) \\ &= \Omega_p(n) \end{aligned}$$

which yields

$$\log\left(\mu_{G,\beta}\left(\left\{T : |\varphi(T)| \geq \frac{(1-\delta')\kappa_n^{-1}N_{d^*}}{1-\Delta_n-\beta}\right\}\right)\right) \leq -\Omega_p(n).$$

Since  $\beta > \beta_c - \delta$  and  $\kappa_n \rightarrow \kappa$ , we have for sufficiently large  $n$  that  $\beta > 1 - \Delta_n - \kappa_n^{-1} - \delta$  and  $\kappa_n < 2\kappa$ . Thus,

$$\begin{aligned} \frac{(1-\delta')\kappa_n^{-1}}{1-\Delta_n-\beta} &\geq \frac{1-\delta'}{1+\delta\kappa_n} \\ &\geq (1-\delta')(1-\delta\kappa_n) \\ &\geq 1-\delta'-\delta\kappa_n \\ &\geq 1-\frac{\epsilon^2}{3\kappa_n}. \end{aligned}$$

This, together with Lemma 71, proves (145). ■

### F.3. The Franz–Parisi potential

#### F.3.1. DEFINITIONS AND MAIN RESULTS

Fix a constant  $\beta > 0$ . Consider a sequence of graphs  $G_1, G_2, \dots$  independently sampled from  $G_n \sim \mathcal{G}(n, q_n)$  and a sequence of spanning trees  $\{T_n\}$ . We consider the non-asymptotic Franz–Parisi potential centered at  $T_n$ , defined by

$$\mathcal{F}_{G_n, T_n, \beta}^{\text{FP}}(r, \epsilon) := -\frac{1}{\beta n \log \log n} \log Z_{r, \epsilon} - (d_n^* + \lambda_n)$$

where

$$Z_{r, \epsilon} := \sum_{T': \mathbf{R}(T_n, T') \in [r-\epsilon, r+\epsilon]} \exp\left(-\bar{\beta} \sum_v \mathbf{d}_{T'}(1, v)\right)$$

and

$$\mathbf{R}(T, T') = \frac{1}{|\bar{V}| - 1} \sum_{v \in \bar{V} \setminus \{1\}} \mathbf{1}_{\text{par}_T(v) \neq \text{par}_{T'}(v)}$$

is the normalized Hamming distance between the two (rooted) spanning trees. Similar to the free energy density, we subtract off the ground state energy. Note that  $Z_{r, \epsilon}$  clearly depends on  $G_n$ ,  $T_n$ , and  $\beta$ , but we drop the subscripts in the interest of readability.

Now assume further that  $\{T_n\}$  are independently drawn from the Gibbs measures  $T_n \sim \mu_{G_n, \beta}$ . We define the (asymptotic) two-temperature Franz–Parisi potential by

$$\mathcal{F}_{\beta, \beta'}^{\text{FP}}(r) := \lim_{\epsilon \rightarrow 0^+} \text{plim}_{n \rightarrow \infty} \mathcal{F}_{G_n, T_n, \beta'}^{\text{FP}}(r, \epsilon)$$

and the one-temperature version

$$\mathcal{F}_\beta^{\text{FP}}(r) := \mathcal{F}_{\beta,\beta}^{\text{FP}}(r)$$

whenever the limits exist. To compute the Franz–Parisi potential, we derive a slightly more general formula which holds for any sequence of trees  $\{T_n\}$  with mild asymptotic conditions. Recall the definition of kernels in Definition 44.

**Proposition 85** *Suppose that  $\Delta_n \rightarrow \Delta \in [0, 1]$  and  $\lambda_n \rightarrow \lambda \in [0, 1]$ . Suppose that  $\{T_n\}$  is a sequence of random spanning trees such that*

$$\frac{|\varphi(T_n)|}{n} \xrightarrow{p} r_1, \quad \frac{|\text{ch}_{T_n}(\varphi(T_n))|}{n} \xrightarrow{p} 1 - r_1$$

and

$$\frac{|\{v \in \varphi(T_n) : |\text{par}_{G_n}(v)| = 1\}|}{n} \xrightarrow{p} r_2.$$

Then the limit

$$\lim_{\epsilon \rightarrow 0} \text{plim}_{n \rightarrow \infty} \mathcal{F}_{G_n, T_n, \beta}^{\text{FP}}(r, \epsilon) = f_{\lambda, \Delta}(r)$$

exists and is a continuous function of  $r$ . Moreover, if  $\beta \geq 1 - \Delta$ , we have

$$f_{\lambda, \Delta}(r) = \begin{cases} -r_1 - \lambda - \frac{\Delta}{\beta} + \left(1 + \frac{\Delta}{\beta}\right)r & \text{if } r_1 + \lambda \leq r \leq 1, \\ -\frac{1 - (1 - \lambda - r_1)(1 - \Delta)}{\beta} + \frac{1}{\beta}r & \text{if } r_1 \leq r \leq r_1 + \lambda, \\ -\frac{1 - (1 - \lambda)(1 - \Delta)}{\beta} + \frac{\Delta}{\beta}r & \text{if } r_2 \leq r \leq r_1, \\ r_2 - \frac{1 - (1 - \lambda - r_2)(1 - \Delta)}{\beta} - \left(1 - \frac{1}{\beta}\right)r & \text{if } 0 \leq r \leq r_2. \end{cases}$$

and if  $\beta \leq 1 - \Delta$ , we have

$$f_{\lambda, \Delta}(r) = \begin{cases} 1 - r_1 - \lambda - \frac{1}{\beta} + \frac{1}{\beta}r & \text{if } r_1 \leq r \leq 1, \\ 1 - \lambda - \frac{1}{\beta} + \left(\frac{1}{\beta} - 1\right)r & \text{if } 0 \leq r \leq r_1, \end{cases}$$

We postpone the proof to Section F.3.2 and first specialize this result to the asymptotic case where  $\{T_n\}$  are drawn from the Gibbs measures. Note that the cases are set to match Corollary 79 and Corollary 82.

**Corollary 86** *Suppose that  $\Delta_n \rightarrow \Delta \in [0, 1]$  and  $\lambda_n \rightarrow \lambda \in [0, 1]$ .*

(A) *If  $\Delta = 1$ , then  $\lambda = 0$  and*

$$\mathcal{F}_\beta^{\text{FP}}(r) = -\frac{1}{\beta} + \frac{1}{\beta}r$$

*for all  $\beta > 0$ .*

(B) We further split the case  $\Delta \in [0, 1)$  and  $\lambda \in [0, 1)$  into the following.

(B.i) If  $\Delta \in [0, 1)$  and  $\lambda = 0$ , then one of the following holds.

- If  $\beta > 1 - \Delta$ , then

$$\mathcal{F}_\beta^{FP}(r) = -\frac{\Delta}{\beta} + \frac{\Delta}{\beta}r.$$

- If  $\beta = 1 - \Delta$ , then  $\mathcal{F}_\beta(r)$  may not exist. Instead, we have

$$\mathcal{F}_{\beta+\delta,\beta}^{FP}(r) = -\frac{\Delta}{\beta} + \frac{\Delta}{\beta}r = -\frac{\Delta}{1-\Delta} + \frac{\Delta}{1-\Delta}r$$

for all  $\delta > 0$  and

$$\mathcal{F}_{\beta-\delta,\beta}^{FP}(r) = 1 - \frac{1}{\beta} + \frac{1}{\beta}r = \frac{\Delta}{1-\Delta} + \frac{1}{1-\Delta}r$$

for all  $0 < \delta < 1 - \Delta$ .

- If  $\beta < 1 - \Delta$ , then

$$\mathcal{F}_\beta^{FP}(r) = 1 - \frac{1}{\beta} + \frac{1}{\beta}r.$$

(B.ii) If  $\Delta = 0$  and  $\lambda \in (0, 1)$ , then one of the following holds.

- If  $\beta > 1$ , then

$$\mathcal{F}_\beta^{FP}(r) = \begin{cases} -\frac{1}{\beta} + \frac{1}{\beta}r & \text{if } r \geq 1 - \lambda, \\ -\frac{\lambda}{\beta} & \text{if } \lambda \log(1/\lambda) \leq r \leq 1 - \lambda, \\ \frac{-\lambda + (\beta - 1)\lambda \log(1/\lambda)}{\beta} - \frac{\beta - 1}{\beta}r & \text{if } r \leq \lambda \log(1/\lambda). \end{cases}$$

- If  $\beta = 1$ , then  $\mathcal{F}_\beta(r)$  may not exist but we have

$$\mathcal{F}_{\beta+\delta,\beta}^{FP}(r) = \begin{cases} -\frac{1}{\beta} + \frac{1}{\beta}r = -1 + r & \text{if } r \geq 1 - \lambda, \\ -\frac{\lambda}{\beta} = -\lambda & \text{if } r \leq 1 - \lambda \end{cases}$$

for all  $\delta > 0$  and

$$\mathcal{F}_{\beta-\delta,\beta}^{FP}(r) = 1 - \lambda - \frac{1}{\beta} + \frac{1}{\beta}r = -\lambda + r$$

for all  $0 < \delta < 1$ .

- If  $\beta < 1$ , then

$$\mathcal{F}_\beta^{FP}(r) = 1 - \lambda - \frac{1}{\beta} + \frac{1}{\beta}r.$$

(C) If  $\lambda = 1$ , then  $\Delta = 0$  and

$$\mathcal{F}_\beta^{FP}(r) = -\frac{1}{\beta} + \frac{1}{\beta}r$$

for all  $\beta > 0$ .

**Proof** This is immediate from Proposition 85 by observing the following.

- Recall that by Theorem 17  $N_{d^*}/n \xrightarrow{p} 1 - \lambda$ .
- If  $\{T_n\}$  are sampled in the low temperature phase, Lemma 74 and (122) imply  $r_1 = 1 - \lambda$ . Also conditioned on  $\Gamma_{\leq d^*-1}$ , for each  $v \in V \setminus \Gamma_{\leq d^*-1}$ , we have  $|\text{par}_G(v)| \sim \text{Binom}(N_{d^*-1}, q)$ , so  $\Pr(|\text{par}_G(v)| = 1 \mid \Gamma_{\leq d^*-1}) = qN_{d^*-1}(1-q)^{N_{d^*-1}-1} \xrightarrow{p} \lambda \log(1/\lambda)$  by  $N_{d^*-1}/(nq)^{d^*-1} \xrightarrow{p} 1$  from Proposition 16. This gives  $r_2 = \lambda \log(1/\lambda)$  (which is zero when  $\lambda = 0$ ).
- If  $\{T_n\}$  are sampled in the high temperature phase, Lemma 76 implies that  $r_1 = 0$ , which automatically gives  $r_2 = 0$ .

■

We complement the case (B.ii) by further investigating the “flat” region for  $\lambda \log(1/\lambda) < r < 1 - \lambda$ . As it turns out, if we zoom in this region, it is a smooth convex basin. This is proved in the following proposition. It might be helpful to revisit the formula in Corollary 82.

**Proposition 87** Suppose that  $\Delta_n \rightarrow \Delta = 0$ ,  $\lambda_n \rightarrow \lambda \in (0, 1)$ ,  $\beta > 1$ , and  $\lambda \log(1/\lambda) < r < 1 - \lambda$ . Letting  $\bar{r} = \frac{r - \lambda \log(1/\lambda)}{1 - \lambda - \lambda \log(1/\lambda)} \in (0, 1)$  be a normalized value of  $r$ , we have

$$\lim_{\epsilon \rightarrow 0} \lim_{n \rightarrow \infty} \left( \left( \mathcal{F}_{G_n, T_n \beta}(r, \epsilon) + \frac{\lambda_n}{\beta} \right) \log \log n + \frac{\lambda \log((1 - \lambda)\alpha_n) + \Psi_0(\lambda)}{\beta} \right) = \frac{1}{\beta} I_\lambda(\bar{r})$$

where

$$I_\lambda(x) := \sup_{t \in \mathbb{R}} (tx - \Lambda(t; \lambda)), \quad 0 \leq x \leq 1$$

is a convex function defined by the Legendre transform of

$$\Lambda(t; \lambda) := \mathbb{E} \left[ \log \left( 1 - \frac{1}{X} + \frac{e^t}{X} \right) \mid X \geq 2 \right], \quad X \sim \text{Poisson}(\log(1/\lambda)).$$

The proof of Proposition 87, outlined in Section F.3.3, borrows techniques from large deviations theory (see, e.g., Dembo and Zeitouni (2009)). The function  $I_\lambda$  is called a *rate function*, which encodes how likely the overlap  $r$  will occur if we sample two independent spanning trees from  $\mu_{G, \beta}$  in the low temperature phase. This can also be understood as a relative entropy of  $\mu_{G, \beta}$  conditioned on the overlap, with respect to  $\mu_{G, \beta}$ . For instance, for each vertex  $v \in \Gamma_{d^*}$ , if we independently sample two parents  $\text{par}_T(v)$  from  $\text{par}_G(v)$ , then they overlap with probability  $1/|\text{par}_G(v)|$ . Since the law of  $|\text{par}_G(v)|$  is approximately  $\text{Poisson}(\log(1/\lambda))$ , among the vertices with  $|\text{par}_G(v)| \geq 2$ , we expect that the typical overlap will be approximately

$$\bar{r}^* = \mathbb{E}[1/X \mid X \geq 2], \quad X \sim \text{Poisson}(\log(1/\lambda)).$$

With simple calculus, one can verify that this is where the line  $y = \bar{r}^* t$  with slope  $\bar{r}^*$  is tangent to the convex function  $y = \Lambda(t; \lambda)$ , hence  $I_\lambda(\bar{r}^*) = 0$ . This corresponds to the bottom of the zoomed-in region, in which case the Franz–Parisi potential agrees with the free energy density (cf. Corollary 82).

One may notice that  $\mathcal{F}_\beta(r)$  in the regime (B.i) is also flat if  $\Delta = 0$  and  $\beta > 1$ . This is fundamentally the same phenomenon with (B.ii), and a similar analysis can be used to show that in this case the Franz–Parisi potential is increasing in  $r$  in the entire regime if we “zoom in”. Intuitively, (B.i) corresponds to the case where a typical vertex has distance  $d^*$  and has an unbounded number of parent choices, so any two independent samples are unlikely to have nontrivial overlap. We do not provide detailed analysis for this case.

### F.3.2. PROOF OF PROPOSITION 85

As usual, we drop the subscripts and write  $T = T_n$  and  $G = G_n$  for conciseness. Define

$$Q := \{v \in \varphi(T) : |\text{par}_G(v)| = 1\}.$$

and recall that

$$Z_{r,\epsilon} := \sum_{T': R(T, T') \in [r-\epsilon, r+\epsilon]} \exp \left( -\bar{\beta} \sum_v d_{T'}(1, v) \right).$$

Also, for a set  $R \subseteq V \setminus \{1\}$ , define

$$\mathcal{D}(T, R) = \{T' : \text{par}_T(v) = \text{par}_{T'}(v) \iff v \in R\}$$

which is the collection of spanning trees whose overlap with  $T$  is exactly  $R$ .

Our goal is to analyze

$$\mathcal{F}_{G,T,\beta}(r, \epsilon) = -\frac{1}{\beta n \log \log n} \log Z_{r,\epsilon} - (d_n^* + \lambda_n).$$

We first write

$$Z_{r,\epsilon} = \sum_{R: |R|/(\bar{n}-1) \in [r-\epsilon, r+\epsilon]} \sum_{T' \in \mathcal{D}(T, R)} \exp \left( -\bar{\beta} \sum_v d_{T'}(1, v) \right) \quad (146)$$

where for convenience we denote  $\bar{n} = |\bar{V}|$ . Recall that the key notion in analyzing  $\log Z_{G,\beta}$  is the kernel of a spanning tree, because the sum over a fixed kernel is easy to approximate (cf. Proposition 46). We apply the same idea to analyze  $\log Z_{r,\epsilon}$ . However, if the overlap of  $T'$  and  $T$  is fixed to be  $R$ , not all sets in  $\mathcal{K}(G)$  can be the kernel of  $T'$ .

- If  $v \in R \cap \varphi(T)$ , then  $\text{par}_{T'}(v) = \text{par}_T(v) \in \Gamma_{d^*-1}$ , so  $v$  must be in the kernel of  $T'$ .
- If  $v \in R \cap (\Gamma_{d^*} \setminus \varphi(T))$ , then  $\text{par}_{T'}(v) = \text{par}_T(v) \notin \Gamma_{d^*-1}$ , so  $v$  must not be in the kernel of  $T'$ .
- If  $v \in Q \setminus R$ , then  $\text{par}_{T'}(v) \neq \text{par}_T(v) \in \Gamma_{d^*-1}$  but  $|\text{par}_G(v)| = 1$  so  $\text{par}_T(v)$  is the only neighbor of  $v$  in  $\Gamma_{d^*-1}$ . Hence,  $v$  must not be in the kernel of  $T'$ .

- For all the other vertices in  $\Gamma_{d^*}$ ,  $\text{par}_{T'}(v) \neq \text{par}_T(v)$  but either  $v \notin \varphi(T)$  or  $|\text{par}_G(v)| > 1$ , so  $v$  may or may not be in the kernel of  $T'$ .

In fact, the kernels of trees in  $\mathcal{D}(T, R)$  can be characterized by these properties.

**Lemma 88** *The kernel  $A = \varphi(T')$  of any  $T' \in \mathcal{D}(T, R)$  satisfies*

$$R \cap \varphi(T) \subseteq A \subseteq (R \cap \varphi(T)) \cup (\Gamma_{d^*} \setminus R \setminus Q). \quad (147)$$

*Conversely, for any  $A \in \mathcal{K}(G)$  satisfying (147) there is a  $T' \in \mathcal{D}(T, R)$  such that  $A = \varphi(T')$ .*

We omit the proof of this fact. Let  $\mathcal{K}(T, R)$  be this set of all possible kernels. Then (146) can be further decomposed as

$$Z_{r,\epsilon} = \sum_{R: |R|/(\bar{n}-1) \in [r-\epsilon, r+\epsilon]} Z_R$$

where

$$Z_R := \sum_{A \in \mathcal{K}(T, R)} Z_R|_A$$

and

$$Z_R|_A := \sum_{\substack{T' \in \mathcal{D}(T, R) \\ \varphi(T') = A}} \exp \left( -\bar{\beta} \sum_v d_{T'}(1, v) \right).$$

Similar to what we did with  $\log Z_{G,\beta}$  and  $\log(Z_{G,\beta}|_A)$  (cf. (78)), we have that

$$\begin{aligned} \log Z_{r,\epsilon} &= \max_{R: |R|/(\bar{n}-1) \in [r-\epsilon, r+\epsilon]} \log Z_R + O(n) \\ &= \max_{R: |R|/(\bar{n}-1) \in [r-\epsilon, r+\epsilon]} \max_{A \in \mathcal{K}(T, R)} \log(Z_R|_A) + O(n). \end{aligned}$$

Since we only need  $\log Z_{r,\epsilon}$  up to  $O(n \log \log n)$  error, it suffices to analyze and maximize  $\log(Z_R|_A)$ . Throughout, we assume that  $A \in \mathcal{K}(T, R)$ , i.e.,  $A \in \mathcal{K}(G)$  and satisfies (147). Note that letting

$$B := A \setminus (R \cap \varphi(T))$$

(147) is equivalent to

$$B \subseteq \Gamma_{d^*} \setminus R \setminus Q. \quad (148)$$

**An upper bound.** Now we upper bound  $\log(Z_R|_A)$  analogous to Lemma 51 and Lemma 80. The difference from these formulae is that here vertices in  $R$  have a fixed parent and thus do not contribute any entropy.

**Lemma 89** *Let  $L = Me^{-\bar{\beta}} \vee \sqrt{\log \bar{n}}$ . With probability at least  $1 - o(1)$ , we have*

$$\log(Z_R|_A) \leq \sum_{v \in B} \log(|\text{par}_G(v)| - \mathbf{1}_{v \in \varphi(T)}) + (n - |R| - |B|) \log(L + |A|q) + \bar{\beta}|A| - \bar{\beta}(d^* + 1)n + o(n)$$

*simultaneously for all  $R \subseteq \bar{V} \setminus \{1\}$  and  $A \in \mathcal{K}(T, R)$ .*

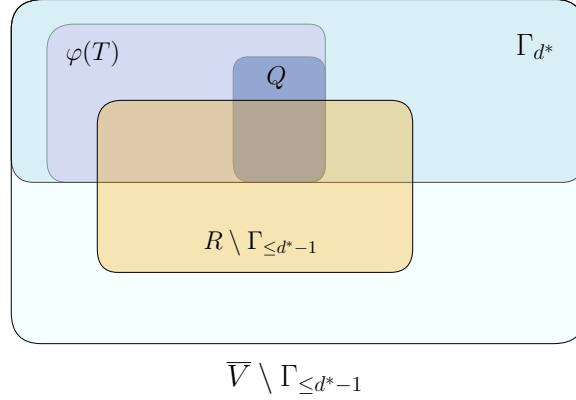


Figure 9: A Venn diagram of  $\Gamma_{d^*}$ ,  $\varphi(T)$ ,  $Q$ , and  $R$  intersecting them.

The proof is nearly identical to Lemma 51 so we omit it. Similar to Lemma 80, this yields

$$\frac{1}{\log \log n} \log(Z_R|_A) \leq |B|\Delta + (n - |R| - |B|) + \beta|A| - \beta(d^* + 1)n + o(n).$$

Rewriting this, we have that

$$\frac{1}{\log \log n} \log(Z_R|_A) \leq n - |R| - |B|(1 - \Delta - \beta) + \beta|R \cap \varphi(T)| - \beta(d^* + 1)n + o(n). \quad (149)$$

Now we consider different phases.

- In the low to moderate temperature phase  $\beta \geq 1 - \Delta$ , maximizing  $\log(Z_R|_A)$  for a fixed  $R$  is done by maximizing  $|B|$ . As seen in (147),  $|B|$  is at most  $|\Gamma_{d^*} \setminus R \setminus Q|$ . Thus, we get an upper bound of  $\log Z_R$

$$\frac{1}{\log \log n} \log Z_R \leq n - |R| + |\Gamma_{d^*} \setminus R \setminus Q|(\beta - 1 + \Delta) + \beta|R \cap \varphi(T)| - \beta(d^* + 1)n + o(n).$$

Thus, our goal is to optimize

$$|\Gamma_{d^*} \setminus R \setminus Q|(\beta - 1 + \Delta) + \beta|R \cap \varphi(T)| \quad (150)$$

over  $R$ . We partition  $R \setminus \Gamma_{\leq d^*-1}$  it into four sets

$$\begin{aligned} U_1 &= R \cap Q \\ U_2 &= R \cap (\varphi(T) \setminus Q) \\ U_3 &= R \setminus \Gamma_{\leq d^*} \\ U_4 &= R \cap (\Gamma_{d^*} \setminus \varphi(T)). \end{aligned} \quad (151)$$

See Figure 9 for an illustration. Letting  $u_i = |U_i|$  the sizes, (150) is written as

$$(N_{d^*} - |Q| - u_2 - u_4)(\beta - 1 + \Delta) + \beta(u_1 + u_2).$$

This is a linear combination of  $u_i$ , and the coefficients of  $u_i$  are  $\beta$ ,  $1 - \Delta$ , 0, and  $1 - \Delta - \beta$ , respectively. Since  $\beta \geq 1 - \Delta \geq 0 \geq 1 - \Delta - \beta$ , we increase  $u_1$ ,  $u_2$ ,  $u_3$ , and  $u_4$  in this order.

In other words, the optimal strategy to construct  $R$  is to pick vertices from  $Q$ ,  $\varphi(T)$ ,  $\bar{V} \setminus \Gamma_{d^*}$ , and then  $\Gamma_{d^*} \setminus \varphi(T)$  until the target size  $|R|$  is reached. Now obvious calculations yield

$$\mathcal{F}_{G,T,\beta}(r, \epsilon) \geq \inf_{r' \in [r-\epsilon, r+\epsilon]} f_{\lambda,\Delta}(r') - o_p(1). \quad (152)$$

- In the moderate to high temperature phase  $\beta \leq 1 - \Delta$ , an upper bound is obtained simply by plugging in  $|B| = 0$ , which gives

$$\max_R \{n - |R| + \beta |R \cap \varphi(T)|\} - \beta(d^* + 1)n + o_p(n).$$

This time, an optimal way to construct  $R$  is simply exhausting  $\varphi(T)$  and then picking vertices outside  $\varphi(T)$  arbitrarily. This also leads to (152).

**A lower bound.** Similar to the upper bound, we have the following analogue of Lemma 55.

**Lemma 90** *With probability at least  $1 - o(1)$ , we have*

$$\log(Z_R|_A) \geq \sum_{v \in B} \log(|\text{par}_G(v)| - \mathbf{1}_{v \in \varphi(T)}) + (n - |R| - |B|) \log(|A|q + \bar{\beta}|A| - \bar{\beta}(d^* + 1)n - o(n))$$

*simultaneously for all  $R \subseteq \bar{V} \setminus \{1\}$  and  $A \in \mathcal{K}(T, R)$  with  $|A| \geq m_\ell$ .*

The proof is simply repeating the argument in the proof of Lemma 55 so we omit it. This also gives

$$\frac{1}{\log \log n} \log(Z_R|_A) \geq n - |R| - |B|(1 - \Delta - \beta) + \beta |R \cap \varphi(T)| - \beta(d^* + 1)n - o(n)$$

which matches (149). Thus, it remains to choose  $R$  and  $A \in \mathcal{K}(T, R)$  such that  $|A| \geq m_\ell$  and the optimal value of (150) is achieved. This can be done by constructing  $R$  (with size within  $(1 \pm \epsilon)n$ ) in the same way we obtained (152) and, if such  $A$  does not exist, add any  $m_\ell$  vertices from  $\Gamma_{d^*}$  intersecting every connected component of  $\bar{G} \setminus \Gamma_{\leq d^*-1}$ . Since  $m_\ell/n \rightarrow 0$  this has little influence on the size constraint on  $R$ . This leads to an upper bound of  $\mathcal{F}_{G,T,\beta}(r, \epsilon)$ :

$$\mathcal{F}_{G,T,\beta}(r, \epsilon) \leq \sup_{r' \in [r-\epsilon, r+\epsilon]} f_{\lambda,\Delta}(r') + o_p(1). \quad (153)$$

Since  $f_{\lambda,\Delta}$  is continuous on  $[0, 1]$ , (152) and (153) prove the conclusion.

### F.3.3. PROOF OF PROPOSITION 87

Since  $T$  is sampled from the low temperature phase, by Theorem 73 we may assume that  $T$  is a uniformly random shortest path tree. In other words, throughout this section, we assume that

$$\varphi(T) = \Gamma_{d^*}$$

and

$$Q = \{v \in \Gamma_{d^*} : |\text{par}_G(v)| = 1\}.$$

Similar to our analysis of  $\log Z_{G,\beta}$ , to get a more accurate estimate of  $\log Z_{r,\epsilon}$ , we partition  $Z_R$  with respect to the kernel size  $|A|$ . In other words,

$$Z_R = \sum_{m=0}^{|\Gamma_{d^*} \setminus R \setminus Q|} \sum_{|A|=m+|R \cap \varphi(T)|} Z_R|_A$$

where  $m$  encodes the size of  $B$  satisfying (148). Note that

$$\log Z_R = \max_{m=0}^{|\Gamma_{d^*} \setminus R \setminus Q|} \log \left( \sum_{|A|=m+|R \cap \varphi(T)|} Z_R|_A \right) + o(n).$$

Since it is assumed that  $\beta > 1 = 1 - \Delta$ , we are in the low temperature phase. Similar to the proof of Theorem 42, it is not difficult to see that an approximate maximum is achieved when  $m$  is the largest uniformly over  $R$ , namely,  $m = |\Gamma_{d^*} \setminus R \setminus Q|$ . This is equivalent to saying that

$$A = (\Gamma_{d^*} \setminus Q) \cup U_1$$

where  $U_i$  are defined in (151). Thus,

$$\log Z_R \leq \log(Z_R|_{(\Gamma_{d^*} \setminus Q) \cup U_1}) + o(n)$$

simultaneously for all  $R$ . Now we apply Lemma 89 and rewrite the bound in terms of  $U_i$  and  $u_i = |U_i|$ . Since we have assumed  $\varphi(T) = \Gamma_{d^*}$ , we have  $U_4 = \emptyset$ . Thus,

$$\begin{aligned} \log Z_R &\leq \sum_{v \in \varphi(T) \setminus Q \setminus U_2} \log(|\text{par}_G(v)| - 1) + (|\bar{V} \setminus \Gamma_{\leq d^*}| + |Q| - u_1 - u_3) \log(qN_{d^*}) \\ &\quad + \bar{\beta}(N_{d^*} - |Q| + u_1) - \bar{\beta}(d^* + 1)n + o(n) \\ &= \sum_{v \in \varphi(T) \setminus Q \setminus U_2} \log(|\text{par}_G(v)| - 1) + (\lambda_n n + |Q| - u_1 - u_3) \log((1 - \lambda_n)nq) \\ &\quad + \bar{\beta}(-|Q| + u_1) - \bar{\beta}(d^* + \lambda_n)n + o(n) \end{aligned} \tag{154}$$

where we also used the fact that  $L + |A|q \leq qN_{d^*}$  asymptotically almost surely. Now note that

$$\log Z_{r,\epsilon} = \max_{u_1, u_3} \log \left( \sum_{\substack{R: |U_i|=u_i \\ i=1,3}} Z_R \right) + o(n). \tag{155}$$

Recall our optimal way of constructing  $R$  in the previous section, where we picked vertices from  $Q$ ,  $\varphi(T) \setminus Q$ ,  $\bar{V} \setminus \Gamma_{d^*}$ , and  $\Gamma_{d^*} \setminus \varphi(T)$ . In the case of Proposition 87, we have  $Q/n \xrightarrow{p} \lambda \log(1/\lambda)$  and  $|\varphi(T)|/n = N_{d^*}/n \xrightarrow{p} 1 - \lambda$  by our assumptions, so an optimal choice of  $R$  satisfies  $u_1 = |Q| - o(n)$  and  $u_3 = o(n)$ . These values are very close to extreme and thus do not contribute to any entropy of choosing  $U_1$  and  $U_3$  for those sizes. Thus, we may simply use  $u_1 = |Q|$  and  $u_3 = 0$  as an optimal choice. In other words, an approximate maximum of (155) is achieved at these values of  $u_1$  and  $u_3$ . Thus, we can write

$$\log Z_{r,\epsilon} = \log \left( \sum_{\substack{Q \subseteq R \subseteq \Gamma_{d^*} \\ |R|/(\bar{n}-1) \in [r-\epsilon, r+\epsilon]}} Z_R \right) + o(n).$$

For  $R$  with  $Q \subseteq R \subseteq \Gamma_{d^*}$ , (154) is written as

$$\log Z_R \leq \sum_{v \in \Gamma_{d^*} \setminus Q \setminus U_2} \log(|\text{par}_G(v)| - 1) + \lambda_n n \log((1 - \lambda_n) n q) - \bar{\beta}(d^* + \lambda_n) n + o(n).$$

Using this, it is straightforward to see that it suffices to show

$$\lim_{\epsilon \rightarrow 0} \text{plim}_{n \rightarrow \infty} \left( \frac{1}{n} \log \left( \sum_{|U_2| \in [\bar{r} - \epsilon, \bar{r} + \epsilon]} \prod_{v \in \Gamma_{d^*} \setminus Q \setminus U_2} (|\text{par}_G(v)| - 1) \right) - \Psi_0(\lambda) \right) = -I_\lambda(\bar{r}).$$

Note that the summand can be written as

$$\begin{aligned} \prod_{v \in \Gamma_{d^*} \setminus Q \setminus U_2} (|\text{par}_G(v)| - 1) &= \prod_{v \in \Gamma_{d^*} \setminus Q \setminus U_2} (|\text{par}_G(v)| - 1) \prod_{v \in U_2} 1 \\ &= \prod_{v \in \Gamma_{d^*} \setminus Q} |\text{par}_G(v)| \prod_{v \in \Gamma_{d^*} \setminus Q \setminus U_2} \frac{|\text{par}_G(v)| - 1}{|\text{par}_G(v)|} \prod_{v \in U_2} \frac{1}{|\text{par}_G(v)|}. \end{aligned}$$

Here, we have

$$\begin{aligned} \sum_{v \in \Gamma_{d^*} \setminus Q} \log |\text{par}_G(v)| &= \sum_{v \in \bar{V} \setminus \Gamma_{\leq d^*-1}} \log(|\mathbf{N}_{\Gamma_{d^*-1}}(v)| \vee 1) \\ &= n \Psi_0(\lambda) + o_p(n) \end{aligned}$$

where the last line is due to Corollary 97 and Lemma 101 (a similar argument was used in Section E.3.7). On the other hand, the product

$$\prod_{v \in \Gamma_{d^*} \setminus Q \setminus U_2} \frac{|\text{par}_G(v)| - 1}{|\text{par}_G(v)|} \prod_{v \in U_2} \frac{1}{|\text{par}_G(v)|}$$

is in fact the joint probability mass function of independent Bernoulli( $1/|\text{par}_G(v)|$ ) random variables evaluated at  $\mathbf{1}_{v \in U_2}$ . Thus, we seek to apply Proposition 113 to conclude that

$$\frac{1}{n} \log \left( \sum_{|U_2| \in [\bar{r} - \epsilon, \bar{r} + \epsilon]} \prod_{v \in \Gamma_{d^*} \setminus Q \setminus U_2} \frac{|\text{par}_G(v)| - 1}{|\text{par}_G(v)|} \prod_{v \in U_2} \frac{1}{|\text{par}_G(v)|} \right) \xrightarrow{p} - \inf_{x \in [\bar{r} - \epsilon, \bar{r} + \epsilon]} I_\lambda(x)$$

which is enough to prove Proposition 87. It only remains to verify that the empirical measure of  $1/|\text{par}_G(v)|$  for  $v \in \Gamma_{d^*} \setminus Q$  is close in 1-Wasserstein metric to the distribution of  $1/X$  conditioned on  $X \geq 2$  where  $X \sim \text{Poisson}(\log(1/\lambda))$ . This can easily be verified using Lemma 111 and Theorem 109; we omit the details.

#### E.4. The energy landscape of the Gibbs measures

We conclude Section E with a simple and intuitive description of the energy landscape depicted in Figure 2. Note that the figure only aims to visualize the relationship between energy and entropy, and does not reflect the actual geometry of the space of spanning trees. Specifically, the depth and width of the basin represent the energy and the entropy of a particular state, respectively. Thus,

the slope represents the derivative of energy with respect to the entropy, which is the temperature needed to overcome that slope. It might be helpful to think that there is a moving particle inside the basin which moves faster as temperature increases.

From Proposition 68, we know that for a fixed kernel size  $m \geq m_\ell$ , the energy of a typical tree under any Gibbs measure is  $(d^* + 1)n - m + o_p(n/\log \log n)$ . Hence, loosely speaking, we can think of  $-m$  as the energy by subtracting off the common term  $(d^* + 1)n$ . In other words, the kernel size  $m$  can be thought of as representing the *depth* of a particular state in the basin. From Corollary 61, we see that the entropy of the state where  $|\varphi(T)| = m$  is approximately

$$\Psi(m; N_{d^*}, \lambda_n) + (n - m) \log(mq) + o(n). \quad (156)$$

**The low temperature state (blue dotted line).** This is the ground state of the system, corresponding to the shortest path trees. Unless the particle is moving fast enough to escape, it stays in the ground state. This agrees with our theoretical results that the Gibbs measure  $\mu_{G,\beta}$  is close to the uniform measure  $\mu_{G,\infty}$  if  $\beta > \beta_c$ .

**The intermediate state (gray region).** Now the particle is energetic enough to climb up the hill. This is the state that only exists in the regimes (B) and (C.i), i.e.,  $\kappa_n \rightarrow \infty$ . This implies that  $N_{d^*} = \omega_p(n/\log \log n)$ . Suppose that  $m = CN_{d^*}$  for some constant  $0 < C \leq 1$ . Then using Lemma 63 and (108) the entropy (156) can be written as

$$\begin{aligned} \Psi(m; N_{d^*}, \lambda_n) + (n - m) \log(mq) &= (\Delta + o_p(1))m \log \log n + n \log(mq) - (1 + o_p(1))m \log \log n \\ &= -(1 - \Delta + o_p(1))m \log \log n + n \log(mq). \end{aligned}$$

Since  $m \log \log n = \omega_p(n)$ ,  $n \log(mq)$  part is negligible so this is simply  $-(1 - \Delta + o_p(1))m \log \log n$ . Hence, the entropy is asymptotically a linear function of the energy  $-m$ , which explains the (approximately) constant slope shown in Figure 2. Also, the approximate slope  $(1 - \Delta) \log \log n$  matches our critical temperature  $\beta_c \log \log n$ . Note that in the regime (C.ii) where  $\kappa_n$  is asymptotically bounded,  $N_{d^*} \log \log n = O_p(n)$  so  $n \log(mq)$  is no longer negligible and this argument does not apply.

**The high temperature state (red region).** This is the state where  $m = O(n/\log \log n)$ . Thus, we let  $m = \frac{Cn}{\log \log n}$  for a constant  $0 < C < \liminf_{n \rightarrow \infty} \kappa_n$  (where the upper bound condition is to ensure  $m < N_{d^*}$  asymptotically almost surely). In this part of the basin, the entropy (156) can be expressed as

$$\begin{aligned} \Psi(m; N_{d^*}, \lambda_n) + (n - m) \log(mq) &= \Delta m \log \log n + n \log(mq) - m \log \log n + o_p(n) \\ &= -(1 - \Delta)m \log \log n + n \log(mq) + o_p(n). \end{aligned}$$

As noted above,  $m \log \log n = Cn$  is a linear term, so  $n \log(mq) = n \log(Cnq/\log \log n)$  is now a non-negligible term. The slope can be estimated by taking derivative with respect to  $m$ , which gives

$$(1 - \Delta) \log \log n + \frac{n}{m} = (1 - \Delta + C^{-1}) \log \log n.$$

Hence, the slope gets steeper as  $C \rightarrow 0$ , i.e., as the depth decreases. This explains a curvy shape of the edges in this red region. In the regimes (B) and (C),  $C$  is unbounded in the right, and as  $C \rightarrow \infty$  the slope approaches  $\beta_c = 1 - \Delta$ , connecting to the intermediate state. In the regime (C.ii),  $C$  is

bounded by  $\kappa \in (1, \infty)$  so the slope approaches  $\beta_c = 1 - \Delta - \kappa^{-1}$ . Since  $C = \kappa$  corresponds to the ground state  $m = N_{d^*}$ , the intermediate state does not exist and thus the state changes continuously from low temperature to high temperature.

In either case, as we go beyond the high temperature state, we have  $1 - \Delta + C^{-1} \rightarrow \infty$ , meaning that the system requires an unbounded temperature to escape the basin. Recall that  $\beta = 0$  gives the uniform measure over the spanning trees.

## Appendix G. Related work

**Related work on tensor PCA.** One average case problem which has been deeply investigated is the problem of *tensor principle component analysis (PCA)* [Montanari and Richard \(2014\)](#) for tensors of order  $p \geq 3$ . Notably, in this algorithm the best guarantees for natural local search algorithms with random initialization like tensor power iteration, approximate message passing (AMP), and gradient ascent or low-temperature Langevin dynamics on the log-likelihood are statistically suboptimal among polynomial time algorithms [Montanari and Richard \(2014\)](#); [Arous et al. \(2020\)](#). Instead, what are believed to be the best possible statistical guarantees for polynomial time algorithms are obtained by algorithms such as higher levels of the Sum-of-Squares (SoS) semidefinite programming hierarchy [Hopkins et al. \(2015\)](#) and more sophisticated spectral methods related to SoS [Hopkins et al. \(2015, 2017\)](#). This example also raised questions about how the predictions from statistical physics style landscape analysis should be interpreted, since unlike in many other problems AMP does not predict the computational threshold. More generally, “local” algorithms with random initialization seem to be suboptimal based on existing analyses [Wein et al. \(2019\)](#).

In a notable work, Wein, El Alaoui and Moore showed that ideas related to statistical physics and landscape analysis *can* predict the same threshold as SoS by analyzing the Hessian of the *Kikuchi free energy*. This is a higher-order analogue of the *Bethe free energy* which is related to belief propagation and approximate message passing (see, e.g., [Yedidia et al. \(2003\)](#)); essentially the Bethe free energy is based on optimization over pairwise marginals, whereas the level  $\ell$  of the Kikuchi free energy is an optimization problem over  $\ell$ -wise marginals. (The levels of the SoS hierarchy have a similar interpretation.) They showed that a linearization analysis of the Kikuchi free energy leads to a spectral algorithm on  $n^{O(p)} \times n^{O(p)}$  matrices which succeeds up to the conjectured optimal threshold.

There are several parallels between these findings for tensor PCA and our results for shortest path on random graphs. In both cases the *choice of optimization landscape/geometry* has an important effect on the performance of local search algorithms like power iteration. In tensor PCA, naively “local” algorithms related to, e.g., optimization the Bethe free energy have suboptimal guarantees; instead, the best algorithms use other landscapes (e.g., corresponding to optimization of the Kikuchi free energy or higher levels of SoS, or the “ironed out” landscape of [Biroli et al. \(2020\)](#)). This mirrors how direct local search on the landscape of the shortest path problem fails whereas lifting local search to the tree landscape succeeds. Relatedly, in both cases ideas related to statistical physics and spin glass theory (e.g., various variational approximations to the free energy) predict the “correct” computational-statistical tradeoff for polynomial time algorithms when we allow for a better choice of optimization landscape.

**Planted vs null problems.** One major difference between our problem and tensor PCA is that tensor PCA is a “planted” problem — the goal for the algorithm is to estimate a planted signal buried in noise — whereas shortest path in random graphs is a “null” or “unplanted” problem because

there is no underlying signal or concept of signal-to-noise ratio. This is why the OGP is relevant to our setting (it is generally used for “null”/pure random optimization problems) whereas AMP and related methods for making predictions in planted problems are used in tensor PCA.

Recently there have also been studies of planted analogues of the OGP in order to study problems like the *planted clique* problem. This had led to findings which appear to have a parallel moral to the case of tensor PCA — naive local search can fail to achieve the optimal computational-statistical tradeoff [Gamarnik and Zadik \(2024\)](#); [Chen et al. \(2024\)](#); [Gheissari et al. \(2025\)](#) and natural changes to the landscape like over-parameterization lead to closing the gap. See also [Deshpande and Montanari \(2015\)](#) for a perspective on planted clique related to AMP.

**Some pointers regarding shattering, OGP, etc.** Some examples of very well-studied problems where there has been a lot of deep work on understanding the solution geometry (through Gibbs measures, OGP, and other means) the  $p$ -spin model from spin glass theory, independent sets in random graphs, and different versions of random  $k$ -SAT, see ([Krzakala et al., 2007](#); [Alaoui, 2024](#); [Kızıldağ, 2023](#); [Gamarnik and Sudan, 2017](#); [Bresler and Huang, 2022](#); [Ding et al., 2015](#); [Coja-Oghlan, 2014](#); [Coja-Oghlan and Panagiotou, 2012](#); [Gamarnik, 2021](#), e.g.). For instance, the works [Achlioptas and Coja-Oghlan \(2008\)](#); [Mézard et al. \(2005\)](#); [Achlioptas and Ricci-Tersenghi \(2006\)](#); [Coja-Oghlan and Efthymiou \(2015\)](#) rigorously established shattering results for the solution space of several optimization problems such as  $k$ -SAT and observed that this implies lower bounds for MCMC algorithms. Here shattering means that the solution space is largely made up of many small clusters which are pairwise distant from each other; this is one example of the type of free energy barrier which the Franz–Parisi potential can pick up on (see, e.g., [El Alaoui et al. \(2025\)](#) for more discussion). These shattering properties also motivated the introduction of the OGP [Gamarnik and Sudan \(2014\)](#).

As previously mentioned, the OGP framework and other analyses of the solution landscape have been applied to a variety of different random optimization problems. We will not attempt to survey the OGP literature, the related literature on spin glass theory etc (see e.g. [Mézard et al. \(1987\)](#)), but just mention a couple of relevant applications where the stability of algorithms have been explicitly studied. One such example is the extensive literature on the symmetric and asymmetric binary perceptron ([Abbe et al., 2022b](#); [Kim and Roche, 1998](#); [Perkins and Xu, 2021](#); [Baldassi et al., 2020](#); [Abbe et al., 2022a](#); [Gamarnik et al., 2022](#), e.g.). Here it was found that although stable algorithms exist in certain regimes, they succeed by finding solutions only in a relatively small region of the solution space; around typical solutions free energy barriers actually exist. This illustrates one potential subtlety involved in making predictions based on landscape geometry. See also [Chou et al. \(2022\)](#) for another work proving the stability of a natural (quantum) algorithm, shown in order to conclude that it cannot overcome the OGP.

**Some related work using the Franz–Parisi potential.** There is a massive physics literature around the Franz–Parisi potential and its application to predicting metastability and hardness which we will again not attempt to survey. See, e.g., [Sun et al. \(2012\)](#); [Franz and Parisi \(1995\)](#); [Zdeborová and Krzakala \(2010\)](#) for relevant work including the “state following” perspective. Some recent applications in mathematics include its use in the solution of the spiked Wigner model [El Alaoui and Krzakala \(2018\)](#) and in studying shattering in spherical spin glasses [El Alaoui et al. \(2025\)](#). Another recent work [Bandeira et al. \(2022\)](#) has shown that in planted models with Gaussian noise, a version of the *annealed* approximation to the Franz–Parisi potential is (1) related/equivalent to predictions based on low-degree polynomials, and (2) yields rigorous lower bounds against local

Markov chains. The latter point is nonobvious and its proof relies heavily on the noise being Gaussian, unlike the quenched Franz–Parisi potential which always implies hardness for MCMC via bottleneck theory (see, e.g., Appendix J.6). The shortest path problem example of [Li and Schramm \(2025\)](#) gives an example where the low-degree heuristic and direct application of the Franz–Parisi potential make essentially different conclusions.

**Overlap gap and low-degree polynomials.** Another interesting aspect of the OGP for shortest path [Li and Schramm \(2025\)](#) is the separation it leads to between the predictions of overlap gap and the low-degree polynomial method for predicting average-case hardness. One reason this is interesting is that some recent works have also established connections between these two in the positive direction, in the sense that OGP-based techniques can sometimes rule out the success of low-degree polynomials [Wein \(2022\)](#); [Gamarnik et al. \(2024\)](#); [Huang and Sellke \(2025\)](#) — see [Li and Schramm \(2025\)](#) for a much more extensive discussion of this topic, and [Huang and Sellke \(2025\)](#) for connections between the low degree heuristic and the probability of the OGP occurring. Intuitively, this is because low-degree polynomials of independent variables are in some sense average-case Lipschitz due to concentration of measure. See, e.g., [Brennan et al. \(2020\)](#); [Holmgren and Wein \(2021\)](#); [Koehler and Mossel \(2022\)](#); [Huang and Mossel \(2024, 2025\)](#); [Buhai et al. \(2025\)](#) for some other works discussing connections between low-degree polynomials and the Statistical Query model as well as limitations of the low-degree polynomial method for predicting computational tractability.

**The structure of low-complexity Gibbs measures.** The rigorous analysis of our model fits into a general paradigm known as *mean-field approximation* in statistical physics and probability [Parisi \(1988\)](#). In particular, rigorous work on the naïve mean-field approximation (e.g. [Austin \(2019\)](#); [Eldan and Gross \(2018\)](#); [Eldan \(2018, 2020\)](#); [Jain et al. \(2019\)](#); [Augeri \(2021\)](#)) shows that for some Gibbs measures of the form  $p(x) \propto e^{f(x)}$  for  $x \in \{\pm 1\}^n$ , the log partition function  $\log Z$  can be well-approximated by an optimization over the space of product measures, and in Wasserstein space the Gibbs measure can be approximated by a “small” mixture of product measures (pure states). Our model does not satisfy the assumptions needed to apply these general results (enforcing the tree structure requires hard constraints, and it is also unclear whether our model satisfies something like the “low gradient complexity” conditions used in [Eldan \(2018\)](#)). The results we obtain do informally correspond to an approximation by (mixtures of) “almost” product measures, and we do build on related information-theoretic techniques like entropy-transportation inequalities. Related to the literature on mean-field approximation, it is also interesting that we need very precise asymptotics (beyond leading-order) for  $\log Z$  in order to understand the structure of the Gibbs measure — a major focus in the mean-field approximation literature is determining the actual size of the error term for naïve mean-field bounds (see [Eldan \(2020\)](#); [Augeri \(2021\)](#) for the state of the art in the context of Ising models).

## Appendix H. Mathematical tools

Here we collect mathematical results used in our main proofs.

### H.1. Concentration inequalities

**Subgaussian random variables.** We start with basic facts about subgaussian random variables; here we match the conventions in [Vershynin \(2018\)](#) where many further details can be found. We include standard proofs of several results in order to make the constants explicit.

**Definition 91** For a random variable  $X$ , the subgaussian norm of  $X$  is defined by

$$\|X\|_{\psi_2} := \inf\{K > 0 : \mathbb{E}[e^{X^2/K^2}] \leq 2\}.$$

We say  $X$  is subgaussian if it has finite subgaussian norm.

Note that we do not require subgaussian random variables to have mean zero.

**Proposition 92** Let  $X$  be a random variable. If  $X$  is subgaussian then it satisfies the tail bound

$$\Pr(|X| \geq t) \leq 2e^{-t^2/\|X\|_{\psi_2}^2}$$

for all  $t \geq 0$ . Conversely, if for some constant  $K > 0$ ,  $X$  satisfies a tail bound

$$\Pr(|X| \geq t) \leq 2e^{-t^2/K^2}$$

for all  $t \geq 0$ , then we have  $\|X\|_{\psi_2} \leq \sqrt{3}K$ .

**Proof** If  $X$  is subgaussian then the Markov inequality gives

$$\Pr(|X| \geq t) = \Pr(e^{X^2/K^2} \geq e^{t^2/K^2}) \leq e^{-t^2/K^2} \mathbb{E}[e^{X^2/K^2}] \leq 2e^{-t^2/K^2}.$$

For the converse, note that

$$\begin{aligned} \mathbb{E}[e^{X^2/C^2}] &= \int_0^\infty \Pr(e^{X^2/C^2} \geq x) dx \\ &= 1 + \int_1^\infty \Pr(|X| \geq C\sqrt{\log x}) dx \\ &\leq 1 + 2 \int_1^\infty x^{-C^2/K^2} dx. \end{aligned}$$

Setting  $C = \sqrt{3}K$  the RHS is at most 2, so  $\|X\|_{\psi_2} \leq C = \sqrt{3}K$ . ■

**Proposition 93** Let  $X$  be a random variable with mean zero. If  $X$  is subgaussian then its MGF satisfies

$$\mathbb{E}[e^{\lambda X}] \leq e^{3\lambda^2\|X\|_{\psi_2}^2/2}.$$

for all  $\lambda \in \mathbb{R}$ . Conversely, if for some constant  $\sigma > 0$  the MGF of  $X$  satisfies

$$\mathbb{E}[e^{\lambda X}] \leq e^{\lambda^2\sigma^2/2}$$

for all  $\lambda \in \mathbb{R}$ , then  $\|X\|_{\psi_2} \leq \sqrt{6}\sigma$ .

**Proof** First, assume that  $X$  is subgaussian with  $\|X\|_{\psi_2} = K$ . We use the inequality

$$e^x \leq 1 + x + \frac{x^2}{2}e^{|x|}$$

which can be easily verified with the Taylor expansion. Then we have

$$\mathbb{E}[e^{\lambda X}] \leq 1 + \mathbb{E}[\lambda X] + \frac{\lambda^2}{2} \mathbb{E}[X^2 e^{|\lambda X|}] = 1 + \frac{\lambda^2 K^2}{2} \mathbb{E} \left[ \frac{X^2}{K^2} e^{|\lambda X|} \right].$$

Again we use the inequalities  $x^2 \leq e^{x^2/2}$  and  $|\lambda X| \leq \frac{X^2}{2K^2} + \frac{\lambda^2 K^2}{2}$  which give

$$\mathbb{E} \left[ \frac{X^2}{K^2} e^{|\lambda X|} \right] \leq e^{\lambda^2 K^2/2} \mathbb{E}[e^{X^2/K^2}] \leq 2e^{\lambda^2 K^2/2}.$$

Hence,

$$\mathbb{E}[e^{\lambda X}] \leq 1 + \lambda^2 K^2 e^{\lambda^2 K^2/2} \leq (1 + \lambda^2 K^2) e^{\lambda^2 K^2/2} \leq e^{3\lambda^2 K^2/2}.$$

For the converse, we have for any  $\lambda > 0$  that

$$\Pr(|X| \geq t) = \Pr(e^{\lambda X} \geq e^{\lambda t}) + \Pr(e^{-\lambda X} \geq e^{\lambda t}) \leq 2e^{-\lambda t + \lambda^2 \sigma^2/2}.$$

Setting  $\lambda = t/\sigma^2$  we obtain

$$\Pr(|X| \geq t) \leq 2e^{-\frac{t^2}{2\sigma^2}}.$$

Applying Proposition 92 completes the proof. ■

**Proposition 94** *If  $X$  is subgaussian, then  $X - \mathbb{E}[X]$  is also subgaussian with*

$$\|X - \mathbb{E}[X]\|_{\psi_2} \leq 2\|X\|_{\psi_2}.$$

**Proof** Let  $\mu = \mathbb{E}[X]$  and  $K = \|X\|_{\psi_2}$ . Then

$$\mathbb{E} \left[ e^{\frac{(X-\mu)^2}{4K^2}} \right] \leq \mathbb{E} \left[ e^{\frac{X^2 + \mu^2}{2K^2}} \right] = e^{\frac{(\mathbb{E}[X])^2}{2K^2}} \mathbb{E} \left[ e^{\frac{X^2}{2K^2}} \right] \leq \left( \mathbb{E} \left[ e^{\frac{X^2}{2K^2}} \right] \right)^2 \leq \mathbb{E}[e^{X^2/K^2}]$$

where the last two inequalities are due to Jensen's inequality. ■

**Proposition 95** *Let  $X_1, \dots, X_n$  be independent subgaussian random variables with mean zero. Then  $X_1 + \dots + X_n$  has subgaussian norm*

$$\|X_1 + \dots + X_n\|_{\psi_2}^2 \leq 18 \sum_{i=1}^n \|X_i\|_{\psi_2}^2.$$

**Proof** Since the MGF of  $X_1 + \dots + X_n$  is the product of the MGFs of  $X_i$ , Proposition 93 easily gives the result. ■

**Lemma 96** *If  $X$  is subgaussian then  $|\mathbb{E}[X]| \leq \|X\|_{\psi_2}$ .*

**Proof** By Proposition 92

$$\begin{aligned}\mathbb{E}[|X|] &= \int_0^\infty \Pr(|X| \geq t) dt \\ &\leq \int_0^\infty 2e^{-t^2/\|X\|_{\psi_2}^2} dt \\ &\leq \sqrt{\pi}\|X\|_{\psi_2}.\end{aligned}$$

This also implies the existence of  $\mathbb{E}[X]$ . To get a sharper result, we can appeal to Jensen's inequality:

$$e^{(\mathbb{E}[X])^2/\|X\|_{\psi_2}^2} \leq \mathbb{E}\left[e^{X^2/\|X\|_{\psi_2}^2}\right] \leq 2.$$

Hence,  $|\mathbb{E}[X]| \leq \sqrt{\log 2}\|X\|_{\psi_2} \leq \|X\|_{\psi_2}$ . ■

This leads to the following subgaussian version of Hoeffding's inequality.

**Corollary 97** *Suppose that  $X_1, \dots, X_n$  are i.i.d. random variables with mean  $\mu$  and subgaussian norm at most  $K$ . Then we have*

$$\Pr\left(\left|\sum_{i=1}^n X_i - n\mu\right| \geq t\right) \leq 2e^{-\frac{t^2}{72nK^2}}.$$

**Proof** This follows by Proposition 94 and Proposition 95. ■

**Chernoff bound.** We use the well-known Chernoff bound for binomials throughout this paper.

**Lemma 98** *Let  $Y \sim \text{Binom}(n, p)$ . Then for any  $-1 < \delta < 1$  we have*

$$\Pr(Y \geq (1 + \delta)np) \leq \exp(np(\delta - (1 + \delta)\log(1 + \delta))).$$

This yields the following bounds which are usually more convenient.

**Lemma 99 (Chernoff bound, see, e.g. Motwani and Raghavan (1995))** *Let  $Y \sim \text{Binom}(n, p)$ . Then for any  $\delta \geq 0$  we have*

$$\Pr(Y \geq (1 + \delta)np) \leq e^{-\frac{\delta^2 np}{2 + \delta}}$$

and

$$\Pr(Y \leq (1 - \delta)np) \leq e^{-\frac{\delta^2 np}{2}}.$$

*In particular, we have*

$$\Pr(|Y - np| \geq \delta np) \leq 2e^{-\frac{\delta^2 np}{3}}.$$

## H.2. Logarithm of binomial random variables

In Section E when we study Gibbs measures and random spanning trees, we are interested in the entropy of choosing a parent for each vertex, which is naturally related to the sum of the *logarithm* of binomial random variables. In this section, we study the concentration of such random variables. The binomial distribution has a nonzero mass at 0, so its logarithm is of course not well defined. Instead, we consider transformed versions of binomial distributions which put no mass at zero. The first family we consider are *shifted* binomial distributions.

**Lemma 100** *Let  $X \sim \text{Binom}(n, p)$ . Then for  $a > 0$  we have*

$$\|\log(X + a) - \log(np + a)\|_{\psi_2} \leq \sqrt{\frac{6}{a}}.$$

**Proof** Let  $Y = \log(X + a)$  and  $\mu = np$ . By Lemma 99 we have

$$\begin{aligned} \Pr(\log(X + a) - \log(\mu + a) \geq t) &= \Pr(X - \mu \geq (e^t - 1)(\mu + a)) \\ &\leq \exp\left(-\frac{(e^t - 1)^2(\mu + a)^2}{2\mu + (e^t - 1)(\mu + a)}\right) \\ &\leq \exp\left(-\frac{(e^t - 1)^2(\mu + a)^2}{(e^t + 1)(\mu + a)}\right) \\ &= \exp\left(-(\mu + a) \cdot \frac{(e^t - 1)^2}{e^t + 1}\right) \end{aligned}$$

for  $t > 0$ . Note that

$$\frac{(e^t - 1)^2}{e^t + 1} \geq \frac{(e^t - 1)^2}{2e^t} = \cosh t - 1 \geq \frac{t^2}{2}. \quad (157)$$

Thus, we have

$$\Pr(\log(X + a) - \log(\mu + a) \geq t) \leq \exp\left(-\frac{\mu + a}{2} \cdot t^2\right).$$

For the lower bound, we begin with

$$\begin{aligned} \Pr(\log(X + a) - \log(\mu + a) \leq -t) &= \Pr(X - \mu \leq -(1 - e^{-t})(\mu + a)) \\ &\leq \exp\left(-\frac{(1 - e^{-t})^2(\mu + a)^2}{2\mu}\right) \end{aligned}$$

for  $t > 0$ . We only need to care about the case  $t \leq \log(\mu + a) - \log a$ , since otherwise the LHS is zero. This implies that

$$1 - e^{-t} \geq \frac{1 - \frac{a}{\mu + a}}{\log(\mu + a) - \log a} \cdot t = \frac{1}{\mu + a} \cdot \frac{\mu}{\log(\mu + a) - \log a} \cdot t.$$

We use the inequality (verifiable with, e.g., Taylor expansion)

$$\frac{e^x - 1}{x} \geq e^{x/2}$$

with  $x = \log(\mu + a) - \log a$ , giving

$$1 - e^{-t} \geq \sqrt{\frac{a}{\mu + a}} \cdot t.$$

Hence,

$$\Pr(\log(X + a) - \log(\mu + a) \leq -t) \leq \exp\left(-\frac{a(\mu + a)}{2\mu} \cdot t^2\right)$$

for all  $t > 0$ .

By Proposition 92, we have

$$\|\log(X + a) - \log(np + a)\|_{\psi_2}^2 \leq \frac{6}{\mu + a} + \frac{6\mu}{a(\mu + a)} \leq \frac{6}{a}.$$

■

Next, we analyze *rectified* binomial distributions.

**Lemma 101** *Let  $X \sim \text{Binom}(n, p)$ . Then we have*

$$\|\log(X \vee 1) - \log(np \vee 1)\|_{\psi_2} \leq \frac{8\sqrt{3}}{\sqrt{np \vee 1}}.$$

**Proof** Let  $\mu = np$ . First, assume  $\mu > 1$ . Then by Lemma 99

$$\begin{aligned} \Pr(\log(X \vee 1) - \log \mu \geq t) &= \Pr(X \vee 1 \geq e^t \mu) \\ &= \Pr(X \geq e^t \mu) \\ &= \Pr(X - \mu \geq (e^t - 1)\mu) \\ &\leq \exp\left(-\mu \cdot \frac{(e^t - 1)^2}{e^t + 1}\right) \\ &\leq \exp\left(-\frac{\mu}{2} \cdot t^2\right) \end{aligned}$$

for  $t > 0$  where the last line follows from (157). Also,

$$\Pr(\log(X \vee 1) - \log \mu \leq -t) \leq \Pr(X \leq e^{-t} \mu) \leq \exp\left(-\frac{\mu(1 - e^{-t})^2}{2}\right)$$

for  $t > 0$ . The LHS is zero if  $t > \log \mu$ , otherwise  $t \leq \log \mu$  which implies

$$1 - e^{-t} \geq \frac{\mu - 1}{\mu \log \mu} \cdot t.$$

Here we use an inequality (verifiable with simple calculus)

$$\frac{e^x - 1}{x} \geq \frac{1}{2} e^{\frac{3}{4}x}$$

and take  $x = \log \mu$ . Then

$$1 - e^{-t} \geq \frac{1}{2\mu^{1/4}} t$$

which gives

$$\Pr(\log(X \vee 1) - \log \mu \leq -t) \leq \exp\left(-\frac{\sqrt{\mu}}{8} \cdot t^2\right)$$

for all  $t > 0$ . Since

$$\frac{\sqrt{\mu}}{8} \leq \frac{\mu}{2}$$

for all  $\mu > 1$ , we have by Proposition 92

$$\|\log(X \vee 1) - \log(np \vee 1)\|_{\psi_2} \leq \frac{8\sqrt{3}}{\sqrt{\mu}}.$$

Next, we assume  $\mu \leq 1$ . Then

$$\begin{aligned} \Pr(\log(X \vee 1) \geq t) &= \Pr(X \geq e^t) \\ &= \Pr(X - \mu \geq e^t - \mu) \\ &\leq \exp\left(-\frac{(e^t - \mu)^2}{e^t + \mu}\right) \end{aligned}$$

for  $t > 0$ . Since  $(e^t - \mu)^2/(e^t + \mu) \geq (e^t - 1)^2/(e^t + 1)$  we get

$$\Pr(\log(X \vee 1) \geq t) \leq e^{-t^2/2}.$$

Since the lower tail is empty, we simply have

$$\|\log(X \vee 1) - \log(np \vee 1)\|_{\psi_2} \leq \sqrt{6}.$$

■

The last family we consider is a family of *truncated* binomial distributions, which is simply conditioning on positiveness. This occurs the most often in our analysis, so we introduce a definition.

**Definition 102** *The zero-truncated binomial distribution with parameters  $n \in \mathbb{Z}^+$  and  $0 < p \leq 1$ , denoted by  $\text{ZTB}(n, p)$ , is a distribution with probability mass function*

$$\frac{1}{1 - (1 - p)^n} \cdot \binom{n}{x} p^x (1 - p)^{n-x}$$

for  $x = 1, \dots, n$ .

We prove a similar result for the zero-truncated distributions.

**Lemma 103** *Let  $X \sim \text{ZTB}(n, p)$ . Then we have*

$$\|\log X - \log(np \vee 1)\|_{\psi_2} \leq \frac{8\sqrt{6}}{\sqrt{np \vee 1}}.$$

*In particular,  $\|\log X - \log(np \vee 1)\|_{\psi_2}$  is bounded by a universal constant.*

**Proof** Let  $\mu = np$  and first assume  $\mu \geq 1$ . Consider  $Y \sim \text{Binom}(n, p)$ . For  $K = \|\log(Y \vee 1) - \log \mu\|_{\psi_2}$  we have

$$\begin{aligned} 2 &\geq \mathbb{E}[e^{(\log(Y \vee 1) - \log \mu)^2 / K^2}] \\ &= e^{(\log \mu)^2 / K^2} \Pr(Y = 0) + \mathbb{E}[e^{(\log(Y \vee 1) - \log \mu)^2 / K^2} \mid Y > 0] \Pr(Y > 0) \\ &\geq (1 - p)^n + \mathbb{E}[e^{(\log X - \log \mu)^2 / K^2}] (1 - (1 - p)^n) \end{aligned}$$

which gives

$$\begin{aligned} \mathbb{E}[e^{(\log X - \log(\mu \vee 1))^2 / K^2}] &\leq \frac{2 - (1 - p)^n}{1 - (1 - p)^n} \\ &= 1 + \frac{1}{1 - (1 - p)^n} \\ &\leq 1 + \frac{1}{1 - e^{-\mu}} \\ &< 4. \end{aligned}$$

By Jensen's inequality,

$$\mathbb{E}[e^{(\log X - \log(\mu \vee 1))^2 / (2K^2)}] \leq (\mathbb{E}[e^{(\log X - \log(\mu \vee 1))^2 / K^2}])^{1/2} \leq 2$$

which implies  $\|\log X - \log(np \vee 1)\|_{\psi_2} \leq \sqrt{2}K$ . Now we can invoke Lemma 101 to conclude

$$\|\log X - \log(\mu \vee 1)\|_{\psi_2} \leq \frac{8\sqrt{6}}{\sqrt{\mu}}.$$

Now assume  $\mu < 1$ . Here, we use Lemma 98 which gives

$$\begin{aligned} \Pr(\log X \geq t) &= \frac{\Pr(X - \mu \geq e^t - \mu)}{1 - (1 - p)^n} \\ &\leq \frac{1}{1 - e^{-\mu}} \exp\left(e^t - \mu - e^t \log\left(1 + \frac{e^t - \mu}{\mu}\right)\right) \\ &= \exp\left(e^t - \mu - te^t - e^t \log(1/\mu) + \log\left(\frac{1}{1 - e^{-\mu}}\right)\right) \\ &= \exp\left(-(t - 1)e^t - (e^t - 1) \log(1/\mu) - \log\left(\frac{e^\mu - 1}{\mu}\right)\right) \\ &\leq \exp(-(t - 1)e^t) \end{aligned}$$

for  $t > 0$ . If we only consider  $t \geq 2$ , we get

$$\Pr(\log X \geq t) \leq \exp(-(t - 1)e^t) \leq e^{-t^2/2}.$$

Then we can say  $\Pr(|\log X| \geq t) \leq 2e^{-t^2/2}$  for all  $t > 0$  since the lower tail does not exist and the case  $t < 2$  is trivial. ■

### H.3. Poisson approximation

In many cases, approximating binomial distributions with Poisson distributions simplifies the results and often makes them more interpretable. We first state the following results which can easily be translated from those for binomial distributions in the previous section. We omit the proofs since they are nearly identical.

**Lemma 104** *Let  $Y \sim \text{Poisson}(\mu)$ . Then for any  $0 < \delta < 1$  we have*

$$\Pr(Y \geq (1 + \delta)\mu) \leq e^{-\frac{\delta^2 \mu}{2 + \delta}}$$

and

$$\Pr(Y \leq (1 - \delta)\mu) \leq e^{-\frac{\delta^2 \mu}{2}}.$$

**Lemma 105** *Let  $X \sim \text{Poisson}(\mu)$ . Then for  $a > 0$  we have*

$$\|\log(X + a) - \log(\mu + a)\|_{\psi_2} \leq \sqrt{\frac{6}{a}}.$$

**Lemma 106** *Let  $X \sim \text{Poisson}(\mu)$ . Then we have*

$$\|\log(X \vee 1) - \log(\mu \vee 1)\|_{\psi_2} \leq \frac{8\sqrt{3}}{\sqrt{\mu \vee 1}}.$$

The zero-truncated Poisson distributions are also similarly defined.

**Definition 107** *The zero-truncated Poisson distribution with parameter  $\mu > 0$ , denoted by  $\text{ZTP}(\mu)$ , is a distribution with probability mass function*

$$\frac{1}{1 - e^{-\mu}} \cdot \frac{\mu^x e^{-\mu}}{x!}$$

for  $x = 1, 2, \dots$ .

**Lemma 108** *Let  $X \sim \text{ZTP}(\mu)$ . Then we have*

$$\|\log X - \log(\mu \vee 1)\|_{\psi_2} \leq \frac{8\sqrt{6}}{\sqrt{\mu \vee 1}}.$$

We also extract a powerful result from the literature on Poisson approximations [Chen \(1975\)](#); [Barbour et al. \(1992\)](#); [Barbour and Xia \(2006\)](#).

**Theorem 109** ([Barbour and Xia \(2006\)](#)) *There is a universal constant  $C > 0$  such that*

$$W_1(\text{Binom}(n, p), \text{Poisson}(np)) \leq C\sqrt{np}^{3/2}.$$

#### H.4. LogSumExp functions and empirical measures

For a vector  $\mathbf{x} = (x_1, \dots, x_n) \in \mathbb{R}^n$ , we define the  $m$ -wise LogSumExp of  $\mathbf{x}$  by

$$\text{LSE}_m(\mathbf{x}) := \log \left( \sum_{I \in \binom{[n]}{m}} \exp \left( \sum_{i \in I} x_i \right) \right).$$

We establish some useful properties of  $\text{LSE}_m$  which is used in Section E.3.6 to prove uniform concentration of certain log partition functions.

**Lemma 110** *For a vector  $\mathbf{x} = (x_1, \dots, x_n) \in \mathbb{R}^n$ , let  $\mu_{\mathbf{x}}$  be its empirical measure on  $\mathbb{R}$ . Then for two vectors  $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ , we have*

$$\frac{1}{n} |\text{LSE}_m(\mathbf{x}) - \text{LSE}_m(\mathbf{y})| \leq W_1(\mu_{\mathbf{x}}, \mu_{\mathbf{y}}) \quad (158)$$

for all  $1 \leq m \leq n$ .

**Proof** Since  $\text{LSE}_m$  is invariant under permutation, we may assume  $x_1 \leq \dots \leq x_n$  and  $y_1 \leq \dots \leq y_n$  without loss of generality. We use the following simple inequality that holds for any  $a_1, a_2, b_1, b_2 > 0$ :

$$\frac{b_1}{a_1} \wedge \frac{b_2}{a_2} \leq \frac{b_1 + b_2}{a_1 + a_2} \leq \frac{b_1}{a_1} \vee \frac{b_2}{a_2}$$

which implies

$$\left| \log \left( \frac{b_1 + b_2}{a_1 + a_2} \right) \right| \leq \left| \log \left( \frac{b_1}{a_1} \right) \right| \vee \left| \log \left( \frac{b_2}{a_2} \right) \right|.$$

Applying this repeatedly, we get

$$\begin{aligned} |\text{LSE}_m(\mathbf{x}) - \text{LSE}_m(\mathbf{y})| &\leq \max_{I \in \binom{[n]}{m}} \left| \log \left( \frac{\exp(\sum_{i \in I} x_i)}{\exp(\sum_{i \in I} y_i)} \right) \right| \\ &\leq \max_{I \in \binom{[n]}{m}} \left| \sum_{i \in I} x_i - \sum_{i \in I} y_i \right| \\ &\leq \sum_{i=1}^n |x_i - y_i| \\ &= n W_1(\mu_{\mathbf{x}}, \mu_{\mathbf{y}}). \end{aligned}$$

■

Lemma 110 indicates that  $\text{LSE}_m(\mathbf{x})$  is related to the empirical distribution of  $\mathbf{x}$ . As such, we take the following result from Lei (2020) (see, e.g., Proposition 3.1 and Corollary 5.2). This also holds for subexponential distributions, but subgaussian assumption suffices for our application.

**Lemma 111 (Lei (2020))** *Let  $\mu$  be a probability measure on  $\mathbb{R}$  with subgaussian norm at most  $\sigma$ . Then there is a constant  $C_\sigma$  and a sequence  $p_{\sigma,n} = 1 - o(1)$  such that for an empirical measure  $\hat{\mu}_n$  we have*

$$\mathbb{E}[W_1(\hat{\mu}_n, \mu)] \leq C_\sigma n^{-1/2}$$

and

$$\Pr \left( W_1(\hat{\mu}_n, \mu) \leq C_\sigma n^{-1/2} \right) \geq p_{\sigma, n}.$$

In other words, we have  $\mathbb{E}[W_1(\hat{\mu}_n, \mu)] = O(n^{-1/2})$  and  $W_1(\hat{\mu}_n, \mu) = O_p(n^{-1/2})$  uniformly over all  $\mu$  with subgaussian norm at most  $\sigma$ .

This enhances Lemma 110, leading to the following key lemma.

**Lemma 112** *Let  $\mu$  be a probability measure on  $\mathbb{R}$  with subgaussian norm at most  $\sigma$ . Then there is a constant  $C_\sigma$  such that for any vector  $\mathbf{y} = (y_1, \dots, y_n) \in \mathbb{R}^n$  and  $1 \leq m \leq n$ , we have*

$$|\text{LSE}_m(\mathbf{y}) - \mathbb{E}[\text{LSE}_m(\mathbf{X})]| \leq nW_1(\mu, \mu_{\mathbf{y}}) + C_\sigma \sqrt{n}$$

where the expectation is over i.i.d. samples  $X_1, \dots, X_n \sim \mu$ .

**Proof** By Lemma 110 we have

$$\begin{aligned} \frac{1}{n} |\text{LSE}_m(\mathbf{X}) - \text{LSE}_m(\mathbf{y})| &\leq W_1(\mu_{\mathbf{X}}, \mu_{\mathbf{y}}) \\ &\leq W_1(\mu_{\mathbf{X}}, \mu) + W_1(\mu, \mu_{\mathbf{y}}). \end{aligned}$$

Now Jensen's inequality gives

$$\begin{aligned} \frac{1}{n} |\text{LSE}_m(\mathbf{y}) - \mathbb{E}[\text{LSE}_m(\mathbf{X})]| &\leq \frac{1}{n} \mathbb{E} [|\text{LSE}_m(\mathbf{X}) - \text{LSE}_m(\mathbf{y})|] \\ &\leq \mathbb{E}[W_1(\mu_{\mathbf{X}}, \mu)] + W_1(\mu, \mu_{\mathbf{y}}). \end{aligned}$$

Applying Lemma 111 completes the proof. ■

## H.5. Large deviations

In this section, we prove a slight extension of the Cramér's theorem in large deviations theory specific to independent not necessarily identically distributed Bernoulli random variables. We consider the case where the parameters for those Bernoulli distributions are also random, written in terms of a triangular array  $\mathbf{p}$  of random variables

$$\begin{array}{cccc} \mathbf{p}_{11} & & & \\ \mathbf{p}_{21} & \mathbf{p}_{22} & & \\ \mathbf{p}_{31} & \mathbf{p}_{32} & \mathbf{p}_{33} & \\ \vdots & \vdots & \vdots & \ddots \end{array}$$

where  $\mathbf{p}_{ni} \in (0, 1)$  for all  $n$  and  $1 \leq i \leq n$ . We also assume that the empirical measures (which are also random)

$$\hat{\mu}_{\mathbf{p}_n} := \frac{1}{n} \sum_{i=1}^n \delta_{\mathbf{p}_{ni}}$$

asymptotically behaves like a fixed probability distribution  $P$  supported on  $(0, 1)$ , in the sense that

$$W_1(\hat{\mu}_{\mathbf{p}_n}, P) \xrightarrow{P} 0 \tag{159}$$

where  $W_1$  is the usual 1-Wasserstein metric. Conditioned on  $\mathbf{p}$ , we consider a triangular array of conditionally i.i.d. random variables  $X_{ni} \sim \text{Bernoulli}(\mathbf{p}_{ni})$ . We establish a large deviations behavior of the random variables

$$S_n := \frac{1}{n} \sum_{i=1}^n X_{ni}$$

conditioned on  $\mathbf{p}_n$ . To describe the behavior, we define the expected cumulant generating function

$$\Lambda(t) := \mathbb{E}_{\mathbf{q} \sim P}[\log(1 - \mathbf{q} + \mathbf{q}e^t)]$$

and its Legendre transform

$$\Lambda^*(x) := \sup_{t \in \mathbb{R}} (tx - \Lambda(t)).$$

Then we have the following result, which can be viewed as a “random environment” or a “quenched” version of Cramér’s theorem.

**Proposition 113** *Under the assumptions above, we have*

$$\frac{1}{n} \log(\Pr(S_n \in \mathcal{I} \mid \mathbf{p}_n)) \xrightarrow{P} - \inf_{x \in \mathcal{I}} \Lambda^*(x)$$

for any compact interval  $\mathcal{I} \subseteq (0, 1)$  with nonempty interior.

The proof follows very closely along the lines of the proofs of Cramér’s theorem or the Gärtner–Ellis theorem<sup>14</sup> (see, e.g., [Dembo and Zeitouni \(2009\)](#) for more on large deviations). A main technical difficulty is that of course the distribution of  $S_n \mid \mathbf{p}_n$  itself is random. To deal with this, we begin with a simple but powerful lemma for working with convex functions in random environment.

**Lemma 114** *Let  $\{f_n\}$  be a sequence of random strictly convex functions  $f_n : X \rightarrow \mathbb{R}$  defined on a convex open set  $X \subseteq \mathbb{R}$  converging pointwise to a strictly convex function  $f : X \rightarrow \mathbb{R}$  in probability, i.e., for each  $x \in X$  we have  $f_n(x) \xrightarrow{P} f(x)$ . The the following hold.*

(a) *The convergence  $f_n \xrightarrow{P} f$  is uniform over any compact set  $\mathcal{K} \subseteq X$ :*

$$\sup_{x \in \mathcal{K}} |f_n(x) - f(x)| \xrightarrow{P} 0.$$

(b) *Suppose that  $x_n^* \in X$  is the unique minimizer of  $f_n$  for each  $n$  and  $x^* \in X$  is the unique minimizer of  $f$ . Then we have*

$$x_n^* \xrightarrow{P} x^*$$

and

$$f_n(x_n^*) \xrightarrow{P} f(x^*).$$

## Proof

---

14. Though we only consider intervals in the interest of simplicity, Proposition 113 easily extends to a statement about any closed and open sets, just like the standard Cramér’s theorem.

(a) Let  $\epsilon > 0$ . We use the fact that any convex function  $g$  satisfies

$$|g(x) - g(a)| \leq |g(a + \delta) - g(a)| \vee |g(a - \delta) - g(a)|$$

for all  $x \in [a - \delta, a + \delta]$ . For each  $a \in \mathcal{K}$ , since  $f_n(a + \delta) \xrightarrow{p} f(a + \delta)$ ,  $f_n(a) \xrightarrow{p} f(a)$ , and  $f_n(a - \delta) \xrightarrow{p} f(a - \delta)$ , and since convex functions are locally Lipschitz, we can find a small enough neighborhood  $(a - \delta, a + \delta) \subseteq X$  such that

$$\sup_{x \in (a - \delta, a + \delta)} |f_n(x) - f(x)| < \epsilon$$

with probability at least  $1 - o(1)$ . Since  $\mathcal{K}$  is compact, we can find finite values of  $a$  such that those neighborhoods cover  $\mathcal{K}$ .

(b) For any  $\epsilon > 0$  we have  $f_n(x^* - \epsilon) \xrightarrow{p} f(x^* - \epsilon) > f(x^*)$ ,  $f_n(x^* + \epsilon) \xrightarrow{p} f(x^* + \epsilon) > f(x^*)$ , and  $f_n(x^*) \xrightarrow{p} f(x^*)$ . Thus, we have  $x_n^* \in (x^* - \epsilon, x^* + \epsilon)$  with high probability, proving the first statement. The second statement follows since convex functions are locally Lipschitz. ■

**Proof** [Proof of Proposition 113] Let  $\mathcal{I} = [a, b]$  with  $0 < a < b < 1$ . We also define the mean cumulant generating function

$$\Lambda(t; \mathbf{p}_n) := \frac{1}{n} \sum_{i=1}^n \log(1 - \mathbf{p}_{ni} + \mathbf{p}_{ni}e^t)$$

and its Legendre transform

$$\Lambda^*(x; \mathbf{p}_n) := \sup_{t \in \mathbb{R}} (tx - \Lambda(t; \mathbf{p}_n)).$$

Note that by Lemma 114 and the assumption (159) we have

$$\Lambda(t; \mathbf{p}_n) \xrightarrow{p} \Lambda(t)$$

pointwise for all  $t \in \mathbb{R}$  and also

$$\Lambda^*(x; \mathbf{p}_n) \xrightarrow{p} \Lambda^*(x) \tag{160}$$

pointwise for all  $x \in (0, 1)$ .

We first prove an upper bound

$$\frac{1}{n} \log(\Pr(S_n \in [a, b] \mid \mathbf{p}_n)) \leq - \inf_{x \in [a, b]} \Lambda^*(x) + o_p(1). \tag{161}$$

Note that if  $x = \mathbb{E}[S_n \mid \mathbf{p}_n] = \frac{1}{n} \sum_{i=1}^n \mathbf{p}_{ni}$ , then the line  $tx$  as a function of  $t$  is tangent to  $\Lambda(t; \mathbf{p}_n)$ , so  $\Lambda^*(x) = 0$ . Thus, if  $[a, b]$  contains  $\mathbb{E}[S_n \mid \mathbf{p}_n]$  then (161) is trivial. Thus, we assume that  $a > \mathbb{E}[S_n \mid \mathbf{p}_n]$  or  $b < \mathbb{E}[S_n \mid \mathbf{p}_n]$ . We only consider the former, as the latter can be handled in the same way. Using the Chernoff bound, we have

$$\begin{aligned} \Pr(S_n \in [a, b] \mid \mathbf{p}_n) &\leq \Pr(S_n \geq a \mid \mathbf{p}_n) \\ &\leq \exp \left( - \sup_{t \geq 0} (nta - n\Lambda_n(t; \mathbf{p}_n)) \right) \\ &= \exp(-n\Lambda^*(a; \mathbf{p}_n)) \end{aligned}$$

where in the last equality we have used  $a > \mathbb{E}[S_n \mid \mathbf{p}_n]$ . This condition also implies that  $\Lambda^*(a) = \inf_{x \in [a,b]} \Lambda^*(x)$ , so we have

$$\frac{1}{n} \log(\Pr(S_n \in [a, b] \mid \mathbf{p}_n)) \leq - \inf_{x \in [a,b]} \Lambda^*(x; \mathbf{p}).$$

Applying Lemma 114 and (160), (161) is proved.

Next, we prove a lower bound

$$\frac{1}{n} \log(\Pr(S_n \in (a, b) \mid \mathbf{p}_n)) \geq - \inf_{x \in (a,b)} \Lambda^*(x) - o_p(1).$$

Fix any  $y \in (a, b)$  and  $\epsilon > 0$  such that  $a < y - \epsilon < y + \epsilon < b$ . Let  $t_{n,y} \in \mathbb{R}$  satisfy  $\Lambda^*(y; \mathbf{p}_n) = t_{n,y}y - \Lambda(t_{n,y}; \mathbf{p}_n)$  and similarly let  $t_y \in \mathbb{R}$  satisfy  $\Lambda^*(y) = t_y y - \Lambda(t_y)$ . Note that

$$|t_{n,y}| \leq |t_{n,a}| \vee |t_{n,b}|$$

which allows us to bound  $|t_{n,y}|$  uniformly. We denote the RHS by

$$t_{n,[a,b]} = |t_{n,a}| \vee |t_{n,b}|.$$

By Lemma 114,  $t_{n,[a,b]}$  converges in probability, and we denote by  $t_{[a,b]}$  its probability limit. Now we consider random variables

$$\tilde{X}_{ni} \sim \text{Bernoulli} \left( \frac{\mathbf{p}_{ni} e^{t_{n,y}}}{1 - \mathbf{p}_{ni} + \mathbf{p}_{ni} e^{t_{n,y}}} \right)$$

which are i.i.d. conditioned on  $\mathbf{p}$ , and also

$$\tilde{S}_n := \frac{1}{n} \sum_{i=1}^n \tilde{X}_{ni}.$$

It is easy to see that  $\tilde{S}_n$  follows an exponentially tilted distribution of  $S_n$  which satisfies

$$\Pr(\tilde{S}_n = x) = \exp(n(t_{n,y}x - \Lambda_n(t_{n,y}; \mathbf{p}_n))) \cdot \Pr(S_n = x)$$

and has the mean cumulant generating function

$$\tilde{\Lambda}(t; \mathbf{p}_n) = \Lambda(t + t_{n,y}; \mathbf{p}_n) - \Lambda(t_{n,y}; \mathbf{p}_n).$$

Then we have

$$\begin{aligned} \Pr(S_n \in (y - \epsilon, y + \epsilon)) &= \sum_{x \in (y - \epsilon, y + \epsilon)} \exp(-n(t_{n,y}x - \Lambda_n(t_{n,y}; \mathbf{p}_n))) \cdot \Pr(\tilde{S}_n = x) \\ &\geq \exp(-n(t_{n,y}y - \Lambda_n(t_{n,y}; \mathbf{p}_n))) \exp(-n\epsilon|t_{n,y}|) \Pr(\tilde{S}_n \in (y - \epsilon, y + \epsilon)) \\ &= \exp(-n\Lambda^*(y; \mathbf{p}_n) - n\epsilon|t_{n,y}|) \Pr(\tilde{S}_n \in (y - \epsilon, y + \epsilon)). \end{aligned}$$

Put another way,

$$\begin{aligned} \frac{1}{n} \log(\Pr(S_n \in (y - \epsilon, y + \epsilon) \mid \mathbf{p}_n)) &\geq -\Lambda^*(y; \mathbf{p}_n) - \epsilon|t_{n,y}| + \frac{1}{n} \log(\Pr(\tilde{S}_n \in (y - \epsilon, y + \epsilon))) \\ &\geq -\Lambda^*(y; \mathbf{p}_n) - \epsilon t_{n,[a,b]} + \frac{1}{n} \log(\Pr(\tilde{S}_n \in (y - \epsilon, y + \epsilon))) \end{aligned}$$

Due to Lemma 114 the convergences  $\Lambda^*(y; \mathbf{p}_n) \xrightarrow{p} \Lambda^*(y)$  and  $t_{n,[a,b]} \xrightarrow{p} t_{[a,b]}$  are uniform over  $y$ , so we have

$$\frac{1}{n} \log(\Pr(S_n \in (y - \epsilon, y + \epsilon) \mid \mathbf{p}_n)) \geq -\Lambda^*(y) - \epsilon t_{[a,b]} + \frac{1}{n} \log(\Pr(\tilde{S}_n \in (y - \epsilon, y + \epsilon))) - o_p(1)$$

where the  $o_p(1)$  term does not depend on  $y$  or  $\epsilon$ .

Now it remains to show

$$\frac{1}{n} \log(\Pr(\tilde{S}_n \in (y - \epsilon, y + \epsilon) \mid \mathbf{p}_n)) \xrightarrow{p} 0. \quad (162)$$

To see this, we observe that  $y = \mathbb{E}[\tilde{S}_n \mid \mathbf{p}_n]$  since

$$\tilde{\Lambda}'(0; \mathbf{p}_n) = \Lambda'(t_{n,y}; \mathbf{p}_n) = y$$

where the second equality follows from the fact that  $t_{n,y}$  maximizes  $ty - \Lambda(t; \mathbf{p}_n)$ . Now we apply the Chernoff bound to see that

$$\Pr(\tilde{S}_n \geq y + \epsilon \mid \mathbf{p}_n) \leq \exp \left( -n \cdot \sup_{t \in \mathbb{R}} (t(y + \epsilon) - \tilde{\Lambda}(t; \mathbf{p}_n)) \right).$$

Here,

$$\begin{aligned} \sup_{t \in \mathbb{R}} (t(y + \epsilon) - \tilde{\Lambda}(t; \mathbf{p}_n)) &= \sup_{t \in \mathbb{R}} (t(y + \epsilon) - \Lambda(t + t_{n,y}; \mathbf{p}_n) + \Lambda(t_{n,y}; \mathbf{p}_n)) \\ &= \sup_{t \in \mathbb{R}} ((t + t_{n,y})(y + \epsilon) - \Lambda(t + t_{n,y}; \mathbf{p}_n)) + \epsilon t_{n,y} - \Lambda^*(y; \mathbf{p}_n) \\ &= \Lambda^*(y + \epsilon; \mathbf{p}_n) - \Lambda^*(y; \mathbf{p}_n) + \epsilon t_{n,y}. \end{aligned}$$

Hence,

$$\Pr(\tilde{S}_n \geq y + \epsilon \mid \mathbf{p}_n) \leq \exp(-n(\Lambda^*(y + \epsilon; \mathbf{p}_n) - \Lambda^*(y; \mathbf{p}_n) + \epsilon t_{n,y})).$$

Similar to the above argument, by Lemma 114, we have that

$$\Lambda^*(y + \epsilon; \mathbf{p}_n) - \Lambda^*(y; \mathbf{p}_n) + \epsilon t_{n,y} = \Lambda^*(y + \epsilon) - \Lambda^*(y) + \epsilon t_y + o_p(1)$$

where the  $o_p(1)$  term does not depend on  $y$  or  $\epsilon$ . Since  $\Lambda^*(y + \epsilon) - \Lambda^*(y) + \epsilon t_y > 0$ , this implies

$$\Pr(\tilde{S}_n \geq y + \epsilon \mid \mathbf{p}_n) = o_p(1)$$

where the  $o_p(1)$  term is still unaffected by  $y$  or  $\epsilon$ . Similarly, we have

$$\Pr(\tilde{S}_n \leq y - \epsilon \mid \mathbf{p}_n) = o_p(1)$$

and this proves (162) as desired. ■

## H.6. Useful inequalities

**Lemma 115** For  $0 \leq p \leq 1$  and  $t \geq 1$ , we have

$$1 - e^{-pt} \leq 1 - (1 - p)^t \leq pt.$$

**Proof** The left inequality is due to  $1 - p \leq e^{-p}$ . To prove the right inequality, we can set  $f(x) = (1 - x)^t - 1 + tx$  and observe that  $f(0) = 0$  and  $f'(x) = -t(1 - x)^{t-1} + t \geq 0$ . ■

**Lemma 116** Let  $0 \leq c \leq 1$ . If  $0 \leq x \leq 2(1 - c)$ , then  $1 - e^{-x} \geq cx$ .

**Proof** From  $0 \leq x \leq 2(1 - c)$  we have  $x - \frac{1}{2}x^2 \geq cx$ . Then  $e^{-x} \leq 1 - x + \frac{1}{2}x^2 \leq 1 - cx$ . ■

## Appendix I. Gibbs measures over trees near the critical temperature

Recall the Gibbs measure studied in Section E

$$\mu_{G_n, \beta}(T) = \frac{1}{Z_{G_n, \beta}} \exp \left( -\beta \log \log(n) \sum_{v \in \bar{V}_n} d_T(1, v) \right)$$

over the set of spanning trees of the component  $\bar{G}_n$  of  $G_n$  containing 1 with  $\bar{V}_n = V(\bar{G}_n)$ . We analyzed this model and its phase transition phenomena when  $\beta$  crosses the critical temperature  $\beta_c = 1 - \Delta - \kappa^{-1}$ . As discussed in Section F.2, the model exhibits distinct behaviors depending on the limits of the proxies  $\Delta_n \rightarrow \Delta \in [0, 1]$ ,  $\lambda_n \rightarrow \lambda \in [0, 1]$ , and  $\kappa_n \rightarrow \kappa \in [0, \infty]$ :

- if  $\Delta < 1$  and  $\kappa = \infty$ , then the system goes through a *discontinuous phase transition*;
- if  $\lambda = 1$  (and thus  $\Delta = 0$ ) and  $\kappa \in (1, \infty)$ , then the system goes through a *continuous phase transition*.

In terms of kernel size (cf. Section F.4), this means that the typical kernel size changes discontinuously in the former, and it changes continuously in the latter. In the discontinuous case, it is not immediately clear what happens when  $\beta$  is extremely close to the critical temperature. The goal of this appendix is to better understand the picture near  $\beta_c$ . We will focus on one particular regime

$$\Delta_n \rightarrow \Delta = 0, \quad \lambda_n \rightarrow \lambda \in (0, 1) \tag{163}$$

which has the critical temperature  $\beta_c = 1$ . This corresponds to (B.ii) in Corollary 82. The behavior in the other regimes with first-order phase transitions is qualitatively the same.

### I.1. The mean-field approximation

In this section, we revisit the mean-field spin system (70) as a simplification of our model. To look at the near-critical behavior/“finite-size effects” we slightly modify the definition of our model by introducing a sequence of inverse temperatures  $\{\beta_n\}_{n \in \mathbb{Z}_+}$  satisfying

$$\liminf_{n \rightarrow \infty} \beta_n > 0 \tag{164}$$

and define the Gibbs measure

$$\mu_{G_n, \beta_n}(T) = \frac{1}{Z_{G_n, \beta_n}} \exp \left( -\beta_n \log \log(n) \sum_{v \in \bar{V}_n} d_T(1, v) \right).$$

Note that with the assumption (164), our arguments in Section E hold for  $\beta_n$  without modification. In particular, an approximate maximum value of the optimization

$$\max_{1 \leq m \leq N_{d^*}} \{ \Psi(m; N_{d^*}, \lambda_n) + (n - m) \log(mq) + \beta_n m \log \log n - \beta_n (d^* + 1)n \log \log n \} \quad (165)$$

gets the value of  $\log Z_{G_n, \beta_n}$  correct up  $o_p(n)$  error, by Theorem 41. Here,  $m$  represents the kernel size  $|\varphi(T)|$  and an approximate optimizer corresponds to an *optimal kernel size* discussed in Section F.4. Also, recall that

$$\Psi(m; N_{d^*}, \lambda_n) := \mathbb{E}[\text{LSE}_m(\log A_1, \dots, \log A_{N_{d^*}}) \mid N_{d^*}], \quad A_1, \dots, A_{N_{d^*}} \sim \text{ZTP}(\log(1/\lambda_n)).$$

Similar to our argument in Section E.3.6, by the uniform concentration of  $\text{LSE}_m$  established in Lemma 112, the log partition function of the following mean-field spin system<sup>15</sup> on  $\mathbf{x} \in \{0, 1\}^{N_{d^*}} \times \{0\}^{n-N_{d^*}}$  given  $N_{d^*}$  and  $\mathbf{A} = (A_1, \dots, A_{N_{d^*}})$  leads to a variational optimization equivalent to (165):

$$\mu_{\mathbf{A}, N_{d^*}}(\mathbf{x}) = \frac{1}{Z_{\mathbf{A}, N_{d^*}}} \exp \left( \sum_{i=1}^{N_{d^*}} (\log A_i - \log(nq) + \beta_n \log \log n) x_i + n \cdot g(\bar{\mathbf{x}}) \right) \quad (166)$$

where  $\bar{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^n x_i$  and  $g(x) = (1 - x) \log x$ . Indeed, we have

$$\begin{aligned} Z_{\mathbf{A}, N_{d^*}} &= \sum_{m=1}^{N_{d^*}} \sum_{\mathbf{x}: \bar{\mathbf{x}}=m/n} \exp \left( \sum_{i=1}^{N_{d^*}} (\log A_i - \log(nq) + \beta_n \log \log n) x_i + n \cdot g(m/n) \right) \\ &= \sum_{m=1}^{N_{d^*}} \exp \left( \text{LSE}_m(\log A_1, \dots, \log A_{N_{d^*}}) + (n - m) \log(mq) + \beta_n m \log \log n - n \log(nq) \right) \end{aligned}$$

so by the uniform concentration of  $\text{LSE}_m$

$$\log Z_{G_n, \beta_n} = \log Z_{\mathbf{A}, N_{d^*}} + n \log(nq) - \beta_n (d^* + 1)n \log \log n + o_p(n).$$

In our new model (166),  $\mathbf{x}$  and  $\mathbf{A}$  have the following interpretations. The first  $N_{d^*}$  components  $x_1, \dots, x_{N_{d^*}}$  of  $\mathbf{x}$  represent the vertices  $v_1, \dots, v_{N_{d^*}}$  in the set  $\Gamma_{d^*}$  of vertices at depth  $d^*$ , and  $x_i = 1$  if and only if the corresponding vertex  $v_i$  is in the kernel  $\varphi(T)$ . Thus,  $\bar{\mathbf{x}}$  represents  $|\varphi(T)|/n$ , the ratio of the kernel size to  $n$ . Each  $A_i$  represents the number of parent choices of  $v_i$ , namely,  $\text{par}_G(v_i)$ .

15. This is slightly different from (70) in that  $N_{d^*}$  — the number of indices with nonzero value — is also quenched (fixed). They are also asymptotically equivalent, but we deliberately chose this quenched version to have better interpretability in terms of the graph  $G_n$  in our original system.

**Phase transitions revisited.** The phase transitions we analyzed in Section F.2 can also be intuitively seen from (166). In the main regime of interest (163), the coefficient of  $x_i$  can be approximated by  $\log A_i - \log(nq) + \beta_n \log \log n = (-1 + \beta_n + o_p(1)) \log \log n$ . Thus if  $\liminf \beta_n > 1$ , then the coefficients are mostly positive and each  $x_i$  roughly contributes  $\Omega(\log \log n)$  to the log probability, so every  $x_i$  wants to be at 1. Otherwise, if  $\limsup \beta_n < 1$ , then the coefficients of  $x_i$  are negative, but the term  $n \cdot g(\bar{x})$  prevents too many of them from being zero, and the equilibrium is achieved at  $\bar{x} = \Theta(1/\log \log n)$ .

## I.2. Optimal kernel size at the critical temperature

Now we describe the behavior of our system in terms of the kernel  $\varphi(T)$  of  $T$  for a typical tree under  $\mu_{G_n, \beta_n}$  when  $\beta_n \rightarrow 1$ . We further assume that

$$\alpha_n \rightarrow \alpha \in \mathbb{R}_+ \quad (167)$$

and

$$\beta_n = 1 + \frac{b}{\log \log n}, \quad b \in \mathbb{R}. \quad (168)$$

Then the optimization (165), ignoring  $o(n)$  terms, becomes

$$\max_{1 \leq m \leq N_{d^*}} \{ \Psi(m; N_{d^*}, \lambda_n) + (n-m) \log(m/n) + (b - \log \alpha)m \} + n \log(nq) - \beta_n(d^* + 1)n \log \log n$$

and the corresponding mean-field model analogous to (166) can be written as

$$\mu'_{\mathbf{A}, N_{d^*}}(\mathbf{x}) = \frac{1}{Z'_{\mathbf{A}, N_{d^*}}} \exp \left( \sum_{i=1}^{N_{d^*}} \log(A_i) x_i + n \cdot g_{\alpha, b}(\bar{\mathbf{x}}) \right), \quad \mathbf{x} \in \{0, 1\}^{N_{d^*}} \times \{0\}^{n-N_{d^*}} \quad (169)$$

where  $g_{\alpha, b}(x) = (b - \log \alpha)x + (1 - x) \log x$ . In the regime (163),  $\log Z'_{\mathbf{A}, N_{d^*}}$  can be accurately described with our tools in large deviations (Section H.5).

**Proposition 117** *We have*

$$\frac{1}{n} \log Z'_{\mathbf{A}, N_{d^*}} \xrightarrow{p} \sup_{0 < y \leq 1} \{ g_{\alpha, b}((1 - \lambda)y) - I(y) \} + \mathbb{E}[\log(A_i + 1)] \quad (170)$$

where

$$I(y) := \sup_{t \in \mathbb{R}} \left\{ ty - \mathbb{E} \left[ \log \left( \frac{A_i e^t + 1}{A_i + 1} \right) \right] \right\}.$$

**Proof** For a constant  $K \in \mathbb{Z}_+$ , consider the intervals  $\mathcal{I}_k = ((k-1)/K, k/K]$  and let

$$\mathcal{B}_k = \left\{ \mathbf{x} : \frac{1}{N_{d^*}} \sum_{i=1}^{N_{d^*}} x_i \in \mathcal{I}_k \right\}.$$

We have

$$\begin{aligned} Z'_{\mathbf{A}, N_{d^*}} &= \sum_{k=1}^K \sum_{\mathbf{x} \in \mathcal{B}_k} \exp \left( \sum_{i=1}^{N_{d^*}} \log(A_i) x_i + n \cdot g_{\alpha, b}(\bar{\mathbf{x}}) \right) \\ &\leq K \cdot \max_{1 \leq k \leq K} \left( \exp \left( n \cdot \sup_{\mathbf{x} \in \mathcal{B}_k} g_{\alpha, b}(\bar{\mathbf{x}}) \right) \sum_{\mathbf{x} \in \mathcal{B}_k} \exp \left( \sum_{i=1}^{N_{d^*}} \log(A_i) x_i \right) \right) \end{aligned}$$

so

$$\log Z'_{\mathbf{A}, N_{d^*}} \leq \log K + \max_{1 \leq k \leq K} \left\{ n \cdot \sup_{\mathbf{x} \in \mathcal{B}_k} g_{\alpha, b}(\bar{\mathbf{x}}) + \log \left( \sum_{\mathbf{x} \in \mathcal{B}_k} \exp \left( \sum_{i=1}^{N_{d^*}} \log(A_i) x_i \right) \right) \right\}.$$

Similarly, we have a lower bound

$$\log Z'_{\mathbf{A}, N_{d^*}} \geq \max_{1 \leq k \leq K} \left\{ n \cdot \inf_{\mathbf{x} \in \mathcal{B}_k} g_{\alpha, b}(\bar{\mathbf{x}}) + \log \left( \sum_{\mathbf{x} \in \mathcal{B}_k} \exp \left( \sum_{i=1}^{N_{d^*}} \log(A_i) x_i \right) \right) \right\}.$$

Applying Proposition 113 to Bernoulli( $\frac{A_i}{A_i+1}$ ) random variables, we have

$$\begin{aligned} \frac{1}{n} \log \left( \sum_{\mathbf{x} \in \mathcal{B}_k} \exp \left( \sum_{i=1}^{N_{d^*}} \log(A_i) x_i \right) \right) &= \frac{1}{n} \sum_{i=1}^{N_{d^*}} \log(A_i + 1) - \inf_{y \in \mathcal{I}_k} I(y) + o_K(n) \\ &= \mathbb{E}[\log(A_i + 1)] - \inf_{y \in \mathcal{I}_k} I(y) + o_K(n). \end{aligned}$$

Using the fact that  $g_{\alpha, b}$  is continuous and  $N_{d^*} = (1 - \lambda)n + o_p(n)$  by Theorem 17, we arrive at the conclusion by taking  $K$  large enough.  $\blacksquare$

**Corollary 118** *The optimization in the RHS of (170) has a unique maximizer  $y^*$  and for any constant  $\epsilon > 0$  we have*

$$\mu'_{\mathbf{A}, N_{d^*}}(|\bar{\mathbf{x}} - (1 - \lambda)y^*| > \epsilon) \xrightarrow{p} 0.$$

**Proof** The function  $g_{\alpha, b}((1 - \lambda)y) - I(y)$  is strictly concave in  $y$  and tends to  $-\infty$  as  $y \rightarrow 0$ , so it has a unique maximizer in  $(0, 1]$ . The uniqueness in turn implies that the log probability is smaller by  $\Omega(n)$  outside a neighborhood of  $y^*$ , yielding the second conclusion.  $\blacksquare$

Recall from the previous section that  $\mathbf{x}$  in the mean-field model (169) corresponds to the proportion of the kernel  $|\varphi(T)|/n$ . Hence, Corollary 118 in fact implies that if we independently sample  $T_n$  from  $\mu_{G_n, \beta_n}$  for each  $n$ , then the kernel proportion concentrates around a nontrivial value:

$$\frac{|\varphi(T_n)|}{n} \xrightarrow{p} (1 - \lambda)y^*.$$

Note that  $y^*$  depends on  $\alpha$ ,  $b$ , and  $\lambda$ . This is to some extent complementary to Lemma 74 and Lemma 76 where it was proved (in a slightly stronger form) that the kernel proportion concentrates

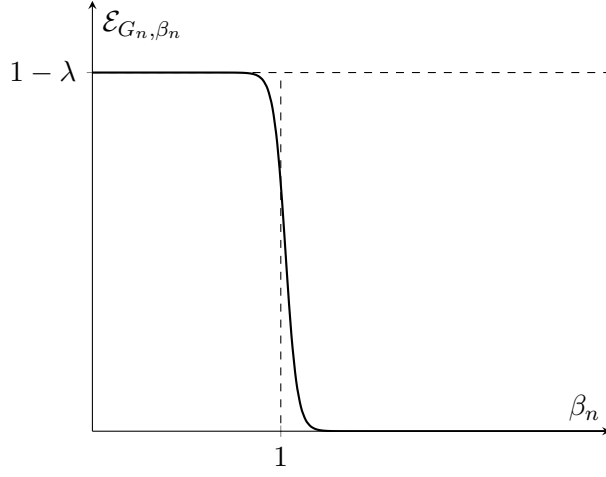


Figure 10: A plot of the internal energy density  $\mathcal{E}_{G_n, \beta_n}$  against the inverse temperature  $\beta_n$ , for a fixed  $n$  (i.e., non-asymptotic). The slope near the critical temperature 1 is at the order of  $-\log \log n$ .

around  $1 - \lambda$  and 0 in the low and high temperature regimes, respectively. We also note that (169), being the exact mean-field model, explains the behavior of each vertex at depth  $d_n^*$  in the following way. Each vertex  $v$  has  $p_v$  chance of being included in the kernel, and those probabilities can approximately be described as the maximizer of

$$\sum_{v \in \Gamma_{d^*}} (\log |\text{par}_G(v)| + H(p_v)) + n \cdot g_{\alpha, b} \left( \frac{1}{n} \sum_{v \in \Gamma_{d^*}} p_v \right)$$

and Corollary 118 implies that  $\frac{1}{n} \sum_v p_v$  can be approximated by  $(1 - \lambda)y^*$ .

**Remark 119 (Accuracy of naive mean-field approximation)** *In the low and high temperature regimes we studied in Section E, a naive mean-field approximation is accurate in the sense that we can find a vertex-wise product measure that approximates the Gibbs measure within  $o_p(n)$  KL divergence. At the critical temperature, the “mean-field approximation” is at least accurate in a looser sense that we have a mean-field model (169) that has asymptotically equal log partition function up to  $o_p(n)$  error. This does not directly give a product measure approximation to the origin model on trees, because if each vertex is independently included in the kernel as we described above, then there is a small chance the kernel is invalid (does not correspond to a valid tree).*

### I.3. Internal energy density as a function of temperature

We conclude this appendix with a non-asymptotic landscape of the (relative) *internal energy density* which we define by

$$\mathcal{E}_{G_n, \beta_n} := \frac{1}{n} \mathbb{E}_{\mu_{G_n, \beta_n}} \left[ \sum_v (d_T(1, v) - d_G(1, v)) \right].$$

Theorem 66 and Proposition 68 are enough to analyze  $\mathcal{E}_{G_n, \beta_n}$  if the kernel proportion  $|\varphi(T)|/n$  concentrates. Specifically, if  $|\varphi(T)|/n \xrightarrow{p} r$ , then Proposition 68 gives

$$\sum_v d_T(1, v) = (d^* + 1 - r)n + o_p(n)$$

with  $\mu_{G_n, \beta_n}$ -exponentially high probability, which implies

$$\mathcal{E}_{G_n, \beta_n} \xrightarrow{p} 1 - \lambda - r.$$

Along with the results in Section F (e.g., Lemma 74 and Lemma 76), we have the following behavior of  $\mathcal{E}_{G_n, \beta_n}$  depending on the temperature of the system.

- In the low temperature phase  $\liminf \beta_n > 1$ , the kernel is almost the entire set  $\Gamma_{d^*}$  most of the time, so  $|\varphi(T)|/n \xrightarrow{p} 1 - \lambda$ . This implies  $\mathcal{E}_{G_n, \beta_n} \xrightarrow{p} 0$ , i.e., the system is in the ground state.
- Near the critical temperature, the limiting kernel proportion depends on the asymptotics of  $\alpha_n$  and  $\beta_n$  as in (167) and (168). For a fixed  $\alpha$ , as  $b$  from (168) moves from  $\infty$  to  $-\infty$ , the limiting kernel proportion  $(1 - \lambda)y^*$  jumps smoothly from 0 to  $1 - \lambda$ . As a consequence, the internal energy density  $\mathcal{E}_{G_n, \beta_n}$  smoothly increases from 0 to  $1 - \lambda$ .
- In the high temperature phase  $\limsup \beta_n < 1$ , the kernel proportion vanishes, i.e.,  $|\varphi(T)|/n \xrightarrow{p} 0$ . This gives  $\mathcal{E}_{G_n, \beta_n} \xrightarrow{p} 1 - \lambda$ .

This is summarized in Figure 10. The change in internal energy at the phase transition corresponds to what is called the “latent heat” for a first-order phase transition (see, e.g., Chaikin and Lubensky (1995)).

## Appendix J. Further discussion

### J.1. Additional analogies: geodesics and extension complexity

**A motivating analogy to geodesics.** In order to better understand the distinction between the OGP in the space of paths vs trees, it is interesting to think about the shortest path problem on other metric spaces, such as Riemannian manifolds. In the Riemannian case, a path between two points  $a$  and  $b$  is a *geodesic* if it is a critical point of the length functional (i.e., the first variation vanishes) and a shortest path, or minimal geodesic, is a global minimum of the length functional Burago et al. (2001). Unlike in Euclidean space, on general manifolds the length functional often has multiple geodesics — or example, on the sphere the geodesics are arcs of great circles, and for generic points  $a$  and  $b$  there will be two geodesics, with only one of them being the shortest path.

A natural optimization procedure in the space of paths would be the *curve-shortening flow* (with fixed end points), which would be the gradient flow of the length functional Grayson (1989); Allen et al. (2012). Clearly, for appropriate initialization the curve shortening flow can converge to any geodesic, so it may fail to find the shortest path between two points. It is also straightforward to define analogues of geodesics in general metric spaces Burago et al. (2001) such as graphs, and we can similarly consider a generalization of the curve shortening flow by iteratively modifying a path via local search — e.g., in the case of graphs, by iteratively moving to the shortest path constrained

to a small Hamming ball around the current iterate. For the same reasons as in the continuous case, this procedure can easily converge to a local instead of global optimum. Intuitively, we can view the overlap-gap result for shortest paths [Li and Schramm \(2025\)](#) as showing that in random graphs, the optimization landscape of paths is particularly adversarial and local search procedures have no hope of succeeding. On the other hand, if we change to a different optimization landscape, where we are allowed to store more information than just a single candidate path, finding shortest path trees is computationally tractable via “local” procedures (in a sense made precise by our results).

**Related analogy with extension complexity.** As we just discussed, while the result of [Li and Schramm \(2025\)](#) shows that shortest paths in a random graph exhibits overlap gap, our results show that shortest path *trees*, which encode more information, are nicer. This has a rough analogy with extended formulations. As a reminder, some high-complexity polytopes in low dimensions are actually *projections* of much “simpler” polytopes in a higher dimensional space [Conforti et al. \(2010\)](#), which has important consequences for, e.g., the tractability of convex optimization over these sets. One notable example is the case of the *permutahedron*, the convex hull of  $\{(\pi(1), \dots, \pi(n)) : \pi \in S_n\}$  where  $S_n$  is the set of permutations of  $[n]$ , has  $2^n - 2$  many facets. However, it can be written as a projection of the set of doubly stochastic  $n \times n$  matrices (Birkhoff polytope), which has only  $O(n^2)$  facets, or as the projection of a different polytope with  $O(n \log n)$  many facets [Goemans \(2015\)](#).

The map from shortest path trees to shortest path is induced by a type of projection map (see Definition 121 below). So as in the case of extended formulations, our result shows that a poorly behaved object (a shortest path in a random graph) can be recovered as the projection of a well-behaved one (a shortest path tree in a random graph). The field of *extension complexity* includes techniques dedicated to establishing the *non-existence* of small extended formulations of polytopes — see e.g. [Rothvoß \(2017\)](#); [Goos et al. \(2018\)](#); [Braverman and Moitra \(2013\)](#); [Lee et al. \(2015\)](#). It would be interesting if any related ideas could be applied in the context of overlap gaps and/or MCMC.

## J.2. The physical interpretation of Dijkstra’s algorithm

The content of this section is mostly classical, see textbooks such as [Bazaraa et al. \(2011\)](#); [Ahuja et al. \(1993\)](#). See also Appendix K for some related content.

**Physical “derivation” of Dijkstra’s algorithm.** Dijkstra’s algorithm is often motivated or explained through the following well-known physical interpretation<sup>16</sup>. We imagine the graph as a collection of physical nodes of equal weights, and an edge  $(v, w)$  is interpreted as an (infinitely strong) physical string of length  $d_G(v, w)$  between the vertices. Then we consider holding fixing the source vertex  $s$  to some fixed height, and then letting all of the other vertices fall down as they are pulled by the force of gravity. If the physical width of the string and nodes are neglected/infinitesimally small, then at equilibrium the *drop in height* between vertex  $s$  and vertex  $v$  will be exactly the distance  $d_G(s, v)$ . In general, the set of strings experiencing tension at equilibrium will contain shortest paths to from the source to every vertex, and if the loads are balanced symmetrically then the set of strings in tension will be exactly the shortest path DAG.

---

16. There is also a related physical interpretation of Dijkstra’s algorithm in terms of a wave propagating through a system of pipes of different lengths (see Chapter 4.4 of [Kleinberg and Tardos \(2006\)](#)).

**Linear programming, network flows, and duality.** Modeling the above physical interpretation of Dijkstra’s algorithm naturally leads to a network flow/linear programming<sup>17</sup> interpretation. Here the variable  $h_v$  corresponds to the *negative* height of the node:

$$\max \sum_v h_v \quad (171)$$

$$\text{s.t. } h_s = 0 \quad (172)$$

$$|h_v - h_w| \leq d_G(v, w) \quad \text{for all } v \sim w \quad (173)$$

$$h_v \in \mathbb{R} \quad \text{for all } v \quad (174)$$

The objective of maximizing the total negative height exactly corresponds to minimizing the *potential energy* of the collection of nodes in the physical analogy, which is exactly where the nodes will come to rest if their positions evolve according to the laws of classical mechanics. (See, e.g., [Goldstein \(1980\)](#).)

We can also naturally interpret the above linear program as the *dual* of a minimum-cost flow problem. Recall that a shortest path between vertices  $s$  and  $t$  can be interpreted as an integral *minimum-cost* flow where the cost of pushing one unit of flow through an edge is proportional to the distance traveled, and edges have infinite capacities. Similarly, we can reinterpret the problem of computing shortest paths from vertex  $s$  to all of the other vertices in the graph minimizing the cost of sending a unit of flow from source  $s$  to every other vertex  $v$  in the graph. This corresponds to the following program:

$$\min \sum_{v \sim w} |f(v, w)| d_G(v, w) \quad (175)$$

$$\text{s.t. } f(v, w) = -f(w, v) \quad \forall v \sim w \quad (176)$$

$$\sum_{w: v \sim w} f(w, v) = 1 \quad \forall v \neq s \quad (177)$$

$$f(v, w) \in \mathbb{R} \quad \forall v \sim w \quad (178)$$

where the second constraint is the conservation of flow equation (recalling that one unit of flow will flow out at each vertex  $v \neq s$ ). This flow problem is a special case of the uncapacitated transportation problem and is somewhat related to another physical system: differential equation models of slime mold algorithms for shortest paths [Bonifaci et al. \(2012\)](#); [Bonifaci \(2013\)](#).

The two linear programs are naturally related by convex duality: the flows  $f(v, w)$  in the latter program are simply the dual variables to the constraints in the former, and similarly the heights  $h_v$  are the dual variables (potentials) for the conservation of flow equations. Concretely, weak duality is the following consequence of Holder’s inequality, which holds for any feasible points of the LPs:

$$\sum_v h_v = \sum_v h_v \sum_w f(w, v) = \frac{1}{2} \sum_{v, w} f(w, v)(h_v - h_w) \leq \sum_{v \sim w} |f(v, w)| d_G(v, w). \quad (179)$$

Strong duality corresponds to the fact that at a pair of optimizing solutions, this becomes an equality.

This duality relationship can also be naturally explained in terms of the physical interpretation. The flow  $f(v, w)$  corresponds to the net force in the up direction exerted upon vertex  $w$  by vertex

---

17. As written, these are convex programs — they are not literally linear programs since we use the absolute value function — but they can straightforwardly be transformed into linear programs via a standard procedure.

$v$  (via the tension of the connecting string), and the flow conservation equation (177) says that at equilibrium, the total upward force exhibit by tension must exactly cancel the downward pull of gravity (so that the net force is zero).

In terms of the two linear programs, we can understand Dijkstra’s algorithm as placing the nodes at their correct heights in a downward scan starting from the origin, which determines the distances from the root, and tracking the saturated constraints in this process corresponds to determining the the union of the supports of all minimum cost flows, or equivalently maintaining the shortest path DAG. This roughly matches the dynamics one would get by taking the physical analogue, setting all of the vertices to the same initial height, and then letting them fall by the force of gravity.

### J.3. Dijkstra’s algorithm and belief propagation

There are many connections between different versions of belief propagation and dynamic programming methods like Viterbi/min-sum decoding in the context of combinatorial optimization. For example, a simple variant of belief propagation can be used to for computing the maximum a posteriori estimate in various settings — see Chapter 14.3 of [Mezard and Montanari \(2009\)](#). Below, we highlight a direct connection in the special case we are interested in, which is computing shortest paths. See [Gamarnik et al. \(2012\)](#) for a related work in the context of general max-flow problems as well as for more references into the literature.

We have already argued that (23) is a natural finite temperature analogue of the shortest path problem. In what follows, we show that starting with this distribution, we can naturally derive Dijkstra’s algorithm from belief propagation. Note that for this section, the argument holds for a *general* graph  $G$  and does not rely upon properties of random graphs. The key to the derivation is the previously described reinterpretation of the random tree  $T$  as a Markov random field (MRF, a.k.a. undirected graphical model or factor model [Mezard and Montanari \(2009\)](#); [Wainwright et al. \(2008\)](#); [Lauritzen \(1996\)](#)) on the graph.

**Belief propagation on the Markov Random Field.** From the factor model, we have the following (loopy) Belief Propagation update (see, e.g., [Mezard and Montanari \(2009\)](#) for the general BP formula): letting 1 be the root node, for every  $u \neq 1$  we have

$$\nu_{u \rightarrow v}^{(t)} = \mathbb{E}_{u \rightarrow v}^{(t-1)} \left[ \frac{\sum_{w:w \in N(u) \setminus \{v\}} e^{-\bar{\beta}(D_w + \ell_{wu})} \delta_{D_w + \ell_{wu}}}{\sum_{w:w \in N(u) \setminus \{v\}} e^{-\bar{\beta}(D_w + \ell_{wu})}} \right] \quad (180)$$

where  $\delta_a$  is the Dirac delta measure at  $a$ ,  $N(u)$  denotes the neighborhood of vertex  $u$ , and  $\mathbb{E}_{u \rightarrow v}^{(t-1)}$  denotes the expectation where independently  $D_w \sim \nu_{w \rightarrow u}^{(t-1)}$  for each  $w \in N(u) \setminus \{v\}$ . and for  $u = 1$  we have

$$\nu_{1 \rightarrow v}^{(0)}(0) = 1$$

since  $d(1, 1) = 0$ . For the initial messages, we can take  $\mu_{u \rightarrow v}^{(0)}(\infty) = 1$  for all  $u \neq 1$  and all neighbors  $v$  of  $u$ . Loopy belief propagation then corresponds to running (180) for  $t$  steps.

**BP converges to Dijkstra in the limit.** The dynamics of loopy belief propagation are nontrivial to analyze. However, if we consider the zero-temperature limit where  $\bar{\beta} \rightarrow \infty$ , then softmax converges to max so the belief propagation converges to

$$\eta_{u \rightarrow v}^{(t)} = \mathbb{E}_{u \rightarrow v}^{(t-1)} \left[ \delta_{\min_{w \in N(u) \setminus \{v\}} (D_w + \ell_{wu})} \right]$$

where  $\mathbb{E}_{u \rightarrow v}^{(t-1)}$  now denotes the expectation where independently  $D_w \sim \eta_{w \rightarrow u}^{(t-1)}$ . By applying induction on time  $t$ , we see that this algorithm only sends delta measures as messages; furthermore, if we track the propagation of messages outward from the root, we see that the message passing is exactly implementing Dijkstra’s algorithm.

**Remark 120** *The belief propagation algorithm above is conceptually related to the success of the “mean-field approximation” for the free energy which shows up in Section E. More specifically, the fixed points of belief propagation correspond to the critical points of a related free energy functional called the Bethe free energy [Mezard and Montanari \(2009\)](#); [Yedidia et al. \(2003\)](#), and in the reverse direction the critical points of the naïve mean-field free energy correspond to fixed points of a similar message-passing algorithm [Wainwright et al. \(2008\)](#). Because the random graphs we consider have  $\omega(1)$  degree and (informally speaking) our system does not exhibit “frustration”, it is reasonable to guess in this setting that the Bethe and naïve mean-field free energy behave very similarly (see, e.g., [Yedidia et al. \(2003\)](#); [Rozzi \(2012\)](#); [Koehler \(2019\)](#) for some related context). Furthermore, essentially the same derivation as above can be used to show that the naïve message-passing algorithm related to mean-field approximation (see, e.g., [Wainwright et al. \(2008\)](#)) also converges to Dijkstra.*

#### J.4. Instability of the projection map from shortest path trees to shortest paths

Given that shortest path trees are “stable” objects which change in a predictable way under resampling of the underlying graph, one might consider trying to compute a shortest path from  $u$  to  $v$  by first computing a shortest path tree rooted at  $u$ , and then outputting the root-to-leaf path from  $u$  to  $v$ . Nevertheless, this procedure cannot be a stable way to construct a shortest path because the OGP rules any such procedure out [Li and Schramm \(2025\)](#). The reason that it fails to be stable is the following: the projection map  $\text{path}_{u,v}$  which takes a shortest path tree from vertex  $u$  to the shortest path from  $u$  to a vertex  $v$  fails badly to be Lipschitz in the Hamming metric.

**Definition 121 (tree-to-path projection map)** *Given any two vertices  $u, v$ , we can define the map  $\text{path}_{u,v}$  which takes as input a tree  $T$  containing  $u$  and  $v$  and outputs the unique shortest path  $P$  between  $u$  and  $v$  in  $T$ . Note that  $\text{path}_{u,v} \circ \text{path}_{u,v} = \text{path}_{u,v}$  so this map is a projector (a.k.a. idempotent ([Mac Lane, 2013](#), Table of Terminology)) in the usual sense. It can be extended to a linear map from the free vector space generated by trees containing  $u$  and  $v$  onto the subspace generated by paths, in which case it would be a projection in the usual sense for linear operators.*

**Example 1** *For an example in a non-random graph, consider a cycle of length  $n$  with  $n$  even and label the vertices by elements of  $\mathbb{Z}/n\mathbb{Z}$ . Then there are two shortest path trees from the vertex 0: one (call it  $T_1$ ) which is a path formed by removing the edge from  $n/2$  to  $n/2 + 1$ , and a symmetrical tree where the path is given by removing the edge from  $n/2 - 1$  to  $n/2$ . While these trees differ only in 2 out of  $n$  many edges, the corresponding shortest paths from 0 to  $n/2$  are disjoint! [Figure 1](#) illustrates this example.*

Our results directly imply that the projection is also not average-case stable in the setting of correlated random graphs. This is because we prove the shortest path tree changes in a smooth and predictable way whereas the prior work [Li and Schramm \(2025\)](#) showed that its projection, the shortest path, will typically change drastically as we resample edges of the graph (see [Section D.4](#) as well).

**Example 2 (Uniformly random shortest path tree is not worst-case stable)** *It is natural to wonder if sampling a uniformly random shortest path tree, or computing the shortest path DAG, could also be stable in the worst case (i.e. Lipschitz). This is not the case. For example, consider a  $d$ -ary tree  $T$  (e.g., with  $d = 2$ ) and depth  $\Theta(\log n)$ , and form a graph  $G$  by identifying the root vertex  $v$  and the leaves of  $T$  with those of an identical tree  $T'$  (so the “interior” vertices of  $T$  and  $T'$  are still disjoint). The initial shortest path DAG is the entire graph  $G$ , oriented away from the root  $v$  and towards the leaves. If we add an edge from  $v$  to any of the vertices at distance 2 from it, the resulting shortcuts leads to a constant proportion of the edges of the graph (those from the “descendants” of the symmetrical copy of the vertex in the tree) being removed from the shortest path DAG, and also leads to a similar change in the uniform distribution over spanning trees — the set of edges which are lost from the DAG go from appearing with probability  $1/(2d)$  to 0, and their symmetrical pairs go from probability  $1/(2d)$  to  $1/d$ , so the mean of the shortest path tree distribution changes by  $\Theta(n)$  in  $\ell_1$  distance.*

*It also follows from our analysis that in Erdos–Rényi graphs, a straightforward analogue of this construction leads to similar unstable (adding a shortcut from the original root vertex to a vertex at distance 2 from  $v$  will significantly change the shortest path DAG and uniformly random shortest path tree). This shows that if insertions/deletions are worst-case, then even in relatively well connected graphs the shortest path DAG and uniformly random shortest path tree no longer behave in a Lipschitz manner.*

### J.5. Sampling a Gibbs measure is equivalent to a random optimization problem

Here we recall the classical connection between sampling Gibbs measures and random optimization problems with Gumbel noise (Train, 2009, Chapter 3). The contents of this section are not needed in the proof of our main results, but are helpful for understanding some of the parallels between the OGP and Gibbs measure frameworks.

Let  $\mathcal{X}$  be a finite set and  $f : \mathcal{X} \rightarrow \mathbb{R}$  a fixed function. Fix an inverse temperature  $\beta > 0$ . The associated *Gibbs distribution* on  $\mathcal{X}$  is

$$p_\beta(x) = \frac{\exp(\beta f(x))}{\sum_{y \in \mathcal{X}} \exp(\beta f(y))} = \frac{\exp(\beta f(x))}{Z_\beta},$$

where  $Z_\beta = \sum_{y \in \mathcal{X}} \exp(\beta f(y))$  is the normalizing constant.

**Gumbel distribution.** A random variable  $G$  is said to have the *standard Gumbel distribution* if its cumulative distribution function is

$$F_G(g) = \Pr[G \leq g] = \exp(-e^{-g}), \quad g \in \mathbb{R}.$$

Equivalently, if  $U \sim \text{Uniform}(0, 1)$  then

$$G = -\log(-\log U)$$

is standard Gumbel (with mean  $\gamma \approx 0.5772$ , the Euler–Mascheroni constant).

**Proposition 122 (Gumbel–Max trick)** *Let  $\{G(x)\}_{x \in \mathcal{X}}$  be i.i.d. standard Gumbel random variables. Then*

$$\Pr\left(\arg \max_{x \in \mathcal{X}} \{\beta f(x) + G(x)\} = x\right) = \frac{\exp(\beta f(x))}{\sum_{y \in \mathcal{X}} \exp(\beta f(y))} = p_\beta(x).$$

*In other words, sampling from  $p_\beta$  is equivalent to solving the random optimization*

$$\hat{x} = \arg \max_{x \in \mathcal{X}} \{\beta f(x) + G(x)\}.$$

**Proof** [Proof sketch.] By independence,

$$\Pr\left(\arg \max_x [\beta f(x) + G(x)] = x\right) = \int_{-\infty}^{\infty} \Pr(G(x) \in dt) \prod_{y \neq x} \Pr(G(y) \leq t + \beta[f(x) - f(y)]).$$

Writing each CDF as  $\exp(-e^{-x})$  and simplifying yields exactly  $\exp(\beta f(x)) / \sum_y \exp(\beta f(y))$ . ■

**Proposition 123 (Log partition function via Gumbel max)** *With the same setup, one has*

$$\mathbb{E}\left[\max_{x \in \mathcal{X}} \{\beta f(x) + G(x)\}\right] = \log\left(\sum_{y \in \mathcal{X}} e^{\beta f(y)}\right) + \gamma = \log Z_\beta + \gamma,$$

where  $\gamma \approx 0.5772$  is the Euler–Mascheroni constant.

**Proof** [Proof sketch.] Compute the law of  $M = \max_x (\beta f(x) + G(x))$  via its CDF

$$\Pr(M \leq t) = \prod_x \exp(-e^{-(t - \beta f(x))})$$

to show that it is a Gumbel distribution with a correspondingly shifted mean. ■

## J.6. Free energy barriers trap Markov chains

It is well known that “free energy barriers” as described by the (quenched) Franz–Parisi potential are rigorous obstructions to the Glauber dynamics and related Markov chains. See Theorem 13.7 of the textbook [Mezard and Montanari \(2009\)](#) as well as, e.g., [Coja-Oghlan and Efthymiou \(2015\)](#); [Bandeira et al. \(2022\)](#); [Arous et al. \(2023\)](#); [Bandeira et al. \(2023\)](#). As discussed in the references, this is a direct implication of classical bottleneck theory for MCMC, but for expositional purposes we recall some of the details here.

Recall that for a Markov chain with transition kernel  $P$  and stationary measure  $\pi$  the bottleneck ratio of a set  $S$  is

$$\Phi(S) = \frac{\sum_{x \in S, y \in S^C} \pi(x) P(x, y)}{\pi(S)}.$$

**Lemma 124 (Proof of Theorem 7.4 of [Wilmer et al. \(2009a\)](#))** *If  $X_0 \sim \pi$  and  $X_0, X_1, \dots$  is a Markov chain evolving according to  $P$ , then*

$$\Pr(\exists s \leq t, X_s \in S^C \mid X_0 \in S) \leq t\Phi(S).$$

**Lemma 125** *For a reversible Markov chain, for any subset  $S$  of the state space and letting  $\partial S = \{y : \exists x \in S, P(x, y) > 0\}$ , we have*

$$\Phi(S) \leq \frac{\pi(\partial S \cap S^C)}{\pi(S)}$$

**Proof** By reversibility, we have

$$\Phi(S) = \frac{\sum_{x \in S, y \in S^C} \pi(x)P(x, y)}{\pi(S)} = \frac{\sum_{x \in S, y \in \partial S \cap S^C} \pi(x)P(x, y)}{\pi(S)} = \frac{\sum_{x \in S, y \in \partial S \cap S^C} \pi(y)P(y, x)}{\pi(S)} \leq \frac{\pi(\partial S \cap S^C)}{\pi(S)}$$

where the last inequality follows since  $\sum_x P(y, x) \leq 1$  for any  $y$ . ■

If the Franz-Parisi potential fails to be *quasi-convex*, i.e. if the potential has multiple disconnected local minima, then the bottleneck theory tells us that Markov chains which are local and reversible will get trapped from suitable initializations.

To illustrate the basic idea, we explicitly write out statements for a special case which often arises. If the Franz-Parisi potential (or its nonasymptotic analogue about point  $x$ ) is not monotonically increasing in the overlap at some point  $r \in (r^*, 1]$ , where  $r^*$  is the location of the global minimum, then for  $\tau = rm$  the above theorem directly implies metastability of states with inner product at least  $\tau$  is metastable, i.e. it takes a long time for the dynamics to escape this set. Symmetrical results and arguments cover the case where the potential is not monotonically decreasing from 0 to  $r^*$ .

**Theorem 126 (Example metastability from FPP, Homogeneous case)** *Let  $P$  be a reversible Markov chain on state space  $\mathcal{X} \subset \{x \in \{0, 1\}^n : \sum_i x_i = m\}$  for some  $m \geq 1$  with stationary measure  $\pi$ , and suppose  $P$  is  $k$ -local for some integer  $k \geq 0$ , in the sense that if  $\|x - x'\|_1 > k$  then  $P(x, x') = 0$ . For any  $\tau \geq 0$  and  $x \in \mathcal{X}$ , define*

$$S_{x, \tau} = \{x' : \langle x, x' \rangle \geq \tau\}.$$

*Then for any  $t \geq 0$ ,*

$$\Pr_{\pi}(\exists s \leq t, X_s \in S_{x, \tau}^C \mid X_0 \in S) \leq t \frac{\pi(\{x' : \langle x, x' \rangle \in (\tau, \tau + k)\})}{\pi(S_{x, \tau})}.$$

**Proof** This follows by combining the previous two lemmas. ■

**Theorem 127 (Example metastability from FPP, Inhomogeneous case)** *Let  $P$  be a reversible Markov chain on state space  $\mathcal{X} \subset \{0, 1\}^n$  with stationary measure  $\pi$ . Suppose that  $P$  is  $\epsilon > 0$  stable with respect to overlap with some fixed  $x \in \mathcal{X}$ , in the sense that if  $P(y, z) > 0$  then*

$$\frac{\langle z, x \rangle}{\|z\|_2 \|x\|_2} \geq \frac{\langle y, x \rangle}{\|y\|_2 \|x\|_2} - \epsilon.$$

*For any  $\tau \in \mathbb{R}$  and  $x \in \mathcal{X}$ , define*

$$S_{x, \tau} = \{x' : \frac{\langle x, x' \rangle}{\|x\|_2 \|x'\|_2} \geq \tau\}.$$

Then for any  $t \geq 0$ ,

$$\Pr_{\pi}(\exists s \leq t, X_s \in S_{\tau}^C \mid X_0 \in S) \leq t \frac{\pi(\{x' : \frac{\langle x, x' \rangle}{\|x\|_2 \|x'\|_2} \in (\tau, \tau + \epsilon)\})}{\pi(S_{x, \tau})}.$$

Note that by the following basic lemma, it is straightforward to prove that Markov chains which are  $k$ -local in the sense of Hamming distance are also local in the sense of normalized overlap, provided that  $\|x\|_2 \gg \sqrt{k}$  for all  $x \in \mathcal{X}$ .

**Lemma 128** For any  $\Delta \in \{-1, 0, 1\}^k$

$$\frac{\langle x, x' + \Delta \rangle}{\|x\|_2 \|x' + \Delta\|_2} \geq \frac{\langle x, x' \rangle - k}{\|x\|_2 (\|x'\|_2 + \sqrt{k})}.$$

### J.7. Comment on replica method approach

In this appendix, we mention the alternative replica method approach to solving the model, which could be interesting to investigate further. The replica method is a non-rigorous but highly reliable way to estimate  $\mathbb{E} \log Z$  to leading order in many problems. See [Mezard and Montanari \(2009\)](#) for more about the replica method. In some cases, the replica method can be much easier to use than rigorous methods, but it does not seem so helpful in our model. This is unsurprising since, while the replica method can be applied to sparse problems, it is usually not the preferred method (see [Biroli et al. \(2000\)](#); [Braunstein and Semerjian \(2023\)](#) for much more context). To illustrate, we show the basic computations which arise.

In the replica trick we start with the identity

$$\mathbb{E}[\log Z] = \lim_{k \rightarrow 0} \frac{1}{k} \log \mathbb{E}[Z^k]$$

which is straightforward to justify using that  $\log(1+x) \approx x$  for small  $x$ . (Note that for fixed  $z > 0$   $\lim_{k \rightarrow 0} z^k = 1$ .) We then apply this formula and guess that we can swap the order of limits to take the high-dimensional limit  $n \rightarrow \infty$  before  $k \rightarrow 0$ :

$$\lim_{n \rightarrow \infty} \frac{1}{n \log \log n} \mathbb{E} \log Z = \lim_{k \rightarrow 0} \lim_{n \rightarrow \infty} \frac{1}{kn \log \log n} \log \mathbb{E}[Z^k].$$

It is difficult to compute  $\mathbb{E}[Z^k]$  for  $k \approx 0$ , so we instead attempt to guess it from the formula of  $\mathbb{E}[Z^k]$  for integer  $k$ , which is more tractable to compute. By expanding the definition of  $Z$ , we know that

$$Z^k = \sum_{T_1, \dots, T_k} \exp \left( -\beta \log \log(n) \sum_{j=1}^k \sum_v d_{T_j}(1, v) \right) \prod_{j=1}^k 1(T_j \subset G)$$

where  $G$  is the random graph and the summation ranges over all spanning trees of the complete graph.

By the definition of the Erdos-Reyni graph  $G(n, \frac{\alpha_n \log n}{n})$  we can compute that

$$\mathbb{E} Z^k = \sum_{T_1, \dots, T_k} \exp \left( -\beta \log \log(n) \sum_{j=1}^k \sum_v d_{T_j}(1, v) \right) (\alpha_n \log(n)/n)^{|T_1 \cup \dots \cup T_k|}$$

where  $|T_1 \cup \dots \cup T_k|$  counts the number of unique edges among the  $k$  spanning trees. In dense spin glass models like the SK model or REM model, at this point the summand would simplify to an expression only depending on the overlap matrix  $Q_{ij} = |T_i \cap T_j|/(n-1)$ . In our case, the summand depends only on  $E_j = \sum_v d_{T_j}(1, v)$  and  $|H| = |T_1 \cup \dots \cup T_k|$ . Computing the formula in the cases  $k = 1$  and  $k = 2$  would correspond to the computations for the first and second moment method, which is already nontrivial. We did not pursue this direction further.

### J.8. Simulation details

We elaborate further on the implementation of the simulation behind Figure 5, since it is slightly nontrivial. The theoretical curve is computed using the formula (54) from Corollary 28 which is the most accurate when  $\rho_n$  is close to 1. To compute  $g(a, b)$ , we used a Monte Carlo approximation by generating i.i.d. Poisson samples, since we are not aware of a closed form. To generate the plot 100 values of  $\rho_n$  were sampled from 1 to 0.9. To simulate progressive resampling, for each pair of vertices  $u, v \in V$  with  $u \neq v$ , we attach a random variable  $T_{u,v} \sim \text{Uniform}(0, 1)$  which models the “time” the pair gets resampled<sup>18</sup>. Namely, for given time indices

$$0 = t_0 < t_1 < \dots < t_m < 1$$

we sample a trajectory of Erdős–Rényi graphs  $(G_0, \dots, G_m)$  as follows. We first generate  $G_0 \sim \mathcal{G}(n, q)$ , and each  $G_k$  for  $k \geq 1$  is inductively generated from  $G_{k-1}$  by resampling  $(u, v)$  with  $t_{k-1} < T_{u,v} \leq t_k$ . Then for each  $k$ , we compute the overlaps for  $G_0$  and  $G_k$ : one for uniformly random shortest path trees (42) and one for the shortest paths (53). Note that each of the shortest paths is determined by the corresponding shortest path tree.

## Appendix K. Relationship to discrete convexity

Discrete convex analysis primarily concerns two natural discrete analogues of convexity, called  $L$  and  $M$ -convexity. Interestingly, while these definitions can be stated in terms of purely geometric properties of a function/set, they have remarkably strong connections to the uncapacitated min-cost flow problem and the shortest path problem.

### K.1. L-convexity and M-convexity

This part is essentially from chapter 1 of Murota’s book Murota (2003).

**Definition 129** For any  $n \geq 1$ , we say that a nonempty set  $D \subset \mathbb{Z}^n$  is  $L^h$ -convex if it is closed under discrete midpoints: for all  $x, y \in D$ ,

$$\lceil \frac{x+y}{2} \rceil, \lfloor \frac{x+y}{2} \rfloor \in D$$

where the ceiling and floor are applied componentwise. We say that such a set  $D$  is  $L$ -convex if it additionally satisfies translation invariance along the all-ones direction, i.e.

$$x \in D \rightarrow x \pm \mathbf{1} \in D.$$

18. In practice, maintaining  $\binom{n}{2}$  random variables for large  $n$  is infeasible. Instead we leverage sparsity: we first sample the edges of  $G_0 \cup \dots \cup G_m$  and assign  $T_{u,v}$  only to these edges, and then generate the graphs using sampled subsets of the edges. It is relatively straightforward to achieve the correct distribution for the trajectory this way.

The two types of  $L$ -convexity are almost equivalent in the sense that any  $L^h$ -convex set in  $\mathbb{Z}^n$  can be directly represented as the intersection of an  $L$ -convex set in  $\mathbb{Z}^{n+1}$  with the hyperplane  $x_{n+1} = 0$ . There are other natural characterizations of these definitions in terms of submodularity and related properties, see [Murota \(2003\)](#).

For our context, it is useful to know the following alternative representation of  $L$ -natural convex sets:

**Theorem 130 (Chapter 1.4.1, Equation 1.34, of [Murota \(2003\)](#))** *A nonempty set  $D \subset \mathbb{Z}^n$  is  $L$ -natural convex if and only if there exists a matrix  $C \in (\mathbb{R} \cup \{\infty\})^{n \times n}$  with zero-diagonal such that*

$$D = \{p : p_i - p_j \leq C_{ij}\}.$$

Given an  $L$ -natural convex set  $D$ , we can define its indicator

$$\delta_D(p) = \begin{cases} 0 & \text{if } p \in D \\ \infty & \text{otherwise} \end{cases}$$

and its *discrete convex/Fenchel conjugate* by

$$\delta_D^*(p) = \sup_x [\langle x, p \rangle - \delta_D(p)] = \sup_{x \in D} \langle x, p \rangle,$$

which is a discrete analogue of the support function of a convex body. It is also a prototypical example of an  $M$ -convex function, as we will see more precisely next.

**Definition 131** *Given a function  $f : \mathbb{Z}^n \rightarrow \mathbb{R} \cup \{\infty\}$ , we define*

$$\text{dom}(f) = \{p : f(p) < \infty\}.$$

**Definition 132** *Let  $f : \mathbb{Z}^n \rightarrow \mathbb{R} \cup \{\infty\}$  be a function with nonempty domain. We say that  $f$  is  $M$ -convex if it satisfies the following exchange axiom: for any  $x, y \in \text{dom}(f)$  and  $i$  such that  $x_i > y_i$ , there exists  $j$  such that  $x_j < y_j$  such that*

$$f(x) + f(y) \geq f(x - e_i + e_j) + f(y + e_i - e_j).$$

**Remark 133** *In many cases in discrete convex analysis, one is interested in integer-valued functions which means that  $f$  will be valued in  $\mathbb{Z} \cup \{\infty\}$ .*

The domain of such a function  $f$  is equivalently an  $M$ -convex set. In the special case that  $f : \{0, 1\}^n \rightarrow \mathbb{R}$ , the support of  $f$  is exactly the set of *bases* of a *matroid*.

**Definition 134** *We say a function  $f : \mathbb{Z}^n \rightarrow \mathbb{R}$  is positively homogeneous if for all  $x, y \in \text{dom}(f)$  such that  $x = \lambda y$  for some  $\lambda > 0$ ,*

$$f(x) = f(\lambda y) = \lambda f(y).$$

**Theorem 135 (Theorem 1.14 of [Murota \(2003\)](#))** *Let  $f : \mathbb{Z}^n \rightarrow \mathbb{Z}$  be an arbitrary integer-valued function with nonempty domain. Then  $f$  is a positively homogeneous  $M$ -convex function if and only if there exists an  $L$ -convex set  $D$  such that  $f = \delta_D^*$ , i.e.  $f$  is the discrete convex conjugate of the indicator of  $D$ . Furthermore, in this case  $(\delta_D^*)^* = f^* = \delta_D$ , i.e. discrete convex conjugation acts as an involution.*

## K.2. Graphic matroid and network simplex

**The support function computes min-cost flows.** Recall that every  $L$ -convex set corresponds to an integer zero-diagonal matrix  $C$  via the representation

$$D = \{p \in \mathbb{Z}^n : p_i - p_j \leq C_{ij}\}$$

and the discrete convex conjugate is

$$\delta_D^*(x) = \sup_{p \in D} \langle p, x \rangle.$$

The convex hull of  $D$  will be an integral polytope, i.e. a convex set with integral vertices, so  $\delta_D^*$  agrees on  $\mathbb{Z}^n$  with the restriction of the support function of  $\text{conv}(D)$ , i.e. for all  $x \in \mathbb{Z}^n$  we have the equivalent formula

$$\delta_D^*(x) = \sup_{p \in \overline{D}} \langle p, x \rangle.$$

where

$$\overline{D} = \{p \in \mathbb{R}^n : p_i - p_j \leq C_{ij}\}$$

is the convex hull of  $D$ . By strong LP duality, this can in turn be rewritten as an optimization problem over flows:

$$\delta_D^*(x) = \inf_{f \in \overline{F}_x} \sum_{i,j} C_{ij} f_{ij}$$

where

$$\overline{F}_x = \{f \in \mathbb{R}_{\geq 0}^{n \times n} : \sum_i C_{ij} f_{ij} - \sum_i C_{ji} f_{ji} = x_i, \quad f_{jj} = 0 \forall j\}$$

is the set of real-valued flows satisfying demand vector  $x$ . By complementary slackness from linear programming theory, any vertex of  $\overline{F}_x$  has at least  $(n-1)(n-1) - (n-1)$  zero entries; equivalently, any vertex of  $\overline{F}_x$  has at most  $n-1$  nonzero entries. In fact, any vertex of  $\overline{F}_x$  is supported on a *forest*: this follows by considering connected components of the graph and applying complementary slackness on each component.

The fundamental *simplex algorithm* from linear programming, in this setting, recasts the linear program as an optimization over the elements of the graphic matroid. Each vertex of  $\overline{F}_x$  correspond to one or more bases in the sense of simplex, which is the same as the sense of basis for the graphic matroid, i.e. a spanning tree. In general, degeneracy can occur when many different spanning trees correspond to the same flow (i.e. the same vertex): then pivoting may not strictly improve objective until we find the “right” basis for this point. This is called *stalling* in the linear programming literature.

For tree-like demands  $x$ , the flow polytope has no nondegenerate vertices<sup>19</sup> so simplex is guaranteed to make progress in every step. Therefore, the network simplex algorithm will necessarily converge in polynomial time (in the size of the numbers in its input, i.e. in pseudopolynomial time) for any standard pivot rule. As a very special case, this tells us that network simplex will converge in polynomial time in the shortest path setting on unweighted graphs considered in our paper.

19. In other words, given a integer-valued flow from the source to all other vertices, we can read off a unique choice of basis (spanning tree).

**Proposition 136** *1-local search on (P2) converges in polynomial time for any connected graph  $G$ .*

**Proof** As discussed above, given the interpretation of local search as the simplex algorithm, this is a direct consequence of basic properties of the simplex algorithm, see e.g. [Bazaraa et al. \(2011\)](#); [Ahuja et al. \(1993\)](#). Explicitly, the objective function is integer-valued and can be at most polynomially large, and there are no *degeneracies* in the corresponding linear program — i.e., every vertex of the linear program is in one-to-one correspondence with a basis (which in turn corresponds to a spanning tree). So the simplex algorithm with any standard pivot rules will always decrease the objective value at every step until it finds the global minimum. ■

Recall that (P2) is

$$\min_{T \in \mathcal{T}} \sum_{v \in V} d_T(s, v)$$

where  $\mathcal{T}$  is the set of spanning trees of  $G$ . By 1-local search we mean any algorithm which, started from a spanning tree  $T$ , iteratively moves to a tree  $T'$  which differs in the addition and removal of one edge, as long as the objective value for  $T'$  is smaller. It is also possible to prove Proposition 136 directly by considering breath-first-search from the root  $s$  in the underlying graph.

**Remark 137 (Simplex is strongly polynomial time for shortest paths)** *Although for worst-case linear programs, all known versions of the simplex algorithm currently require exponential time, it was proved by Orlin [Orlin \(2009\)](#) that the simplex algorithm with Dantzig’s pivot rule is strongly polynomial time for shortest path problems. See also the improved result of Hansen et al [Hansen et al. \(2014\)](#). Orlin [Orlin \(1997\)](#) also showed existence of a pivoting approach for general flow problems with polynomial time guarantees.*

### K.3. Dijkstra’s algorithm and L-concave function maximization [Murota and Shioura \(2014\)](#)

From the discussion in the previous section, we can understand how the *network simplex* algorithm is a natural method to optimize linear functions over  $L$ -convex sets. It was observed by Murota and Shioura [Murota and Shioura \(2014\)](#) that Dijkstra’s algorithm also naturally arises from considerations in discrete convex analysis.

Let  $D$  be as defined in the previous section and interpret  $C_{ij}$  as the edge length from vertex  $i$  to vertex  $j$  in the graph of interest. Then *shortest path problem* from source vertex 1 corresponds to computation of

$$\delta_D^*(-(n-1), 1, \dots, 1) = \max_{p \in D} \sum_v (p(v) - p(1)).$$

This is optimization of a linear function over an  $L$ -convex set, which is a special case of the more general problem of  $L$ -concave function maximization (the larger class consists of functions with obey a midpoint inequality similar to the continuous case, and discretized similarly to the definition of  $L$ -convex set). Any  $L$ -concave maximization problem can be solved in polynomial time by the *steepest ascent algorithm* from discrete convex analysis [Murota \(2003\)](#). Consist with the name, the steepest ascent algorithm at iteration  $t$  applied to  $L$ -concave function  $g$  starts from a point  $p_t \in \mathbb{R}^n$  and moves to  $p_{t+1} = \arg \max_{y \in \{0,1\}^n, \epsilon \in \{\pm 1\}} g(p_t + \epsilon y)$ .  $L$ -concave functions have a “local to global” property similar to continuous concave functions, i.e., local optima of such functions are always global optima. So steepest ascent is globally convergent.

In general, steepest ascent can always be implemented by appealing to an oracle for submodular minimization to solve the optimization over  $y$ . However, in specific cases the steepest ascent algorithm can be simplified. In the case of the shortest path problem, it turns out that the steepest ascent algorithm is *Dijkstra's algorithm* [Murota \(2003\)](#); one always has  $\epsilon = 1$  and steepest ascent is increasing the node potentials  $p$  just as in Dijkstra's algorithm (the  $p$  correspond to the location of the vertices in the “physical” interpretation of Dijkstra's algorithm mention before). See the paper for full details.

#### K.4. Shortest path tree is not discrete convex

The support of the shortest path tree objective is on the bases of the graphic matroid, which is an M-convex set, so it is natural to ask if the function itself is M-convex (equivalently, to ask if it is a valuated matroid). The following example shows that this is not the case. (It is direct to see that shortest path tree objective is not  $L$ -convex as it does not have the needed discrete midpoint property.)

**Example 3** *For notational convenience, we will root the tree in this example at vertex 0 instead of vertex 1. Let  $m \geq 1$ , and consider a graph  $G$  which is given by starting with a cycle on  $2m + 1$  nodes labeled  $0, \dots, 2m$ , and adjoining a separate path of length  $m + 3$  from node 0 to node  $m$  with intermediate vertices labeled  $1', 2', \dots, (m+1)', (m+2)'$ . Let  $T_1$  be the optimal shortest path tree, which is formed by dropping the edges from  $m$  to  $m+1$  and from  $(m+1)'$  to  $(m+2)'$  from the graph. Let  $T_2$  be the spanning tree formed from the original graph by dropping the edge from  $m-1$  to  $m$  and from  $m$  to  $(m+2)'$ . Let  $OPT$  be the objective value achieved by  $T_1$ , and observe that the objective value achieved by  $T_2$  is  $OPT + 2$  since vertices  $m$  and  $(m+2)'$  are now further from the root.*

*If we consider an exchange between the two trees by dropping the edge from  $(m+1)'$  to  $(m+2)'$  from  $T_2$  and adding it to  $T_1$ , we must remove the edge from  $m$  to  $(m+2)'$  from  $T_1$  and add it to  $T_2$ . This makes the objective value of  $T_1$  worsen to  $OPT + 1$  and the objective value of  $T_2$  stays the same, which violates the exchange axiom.*

See Figure 11 for a diagram illustrating the example.

## Appendix L. Free energy barrier in the space of paths

Fix a target vertex  $t$ , and consider for each  $n$  the Gibbs measure

$$\nu_{G,\beta}(P) = \frac{1}{Z_{G,\beta}} \exp(-\beta \log \log(n)|P|) = \frac{1}{Z_{G,\beta}} \log(n)^{-\beta|P|} \quad (181)$$

over paths  $P \in \mathcal{P}_{s,t}$  from 1 to  $t$ . To be more precise, we let paths in  $\mathcal{P}_{s,t}$  repeat vertices (i.e. they are not required to be simple, sometimes called “walks”), and we also restrict paths in  $\mathcal{P}_{s,t}$  to have length at most  $n - 1$  where  $n$  is the number of vertices. We make the latter restriction since the shortest path has at most  $n - 1$  edges in any graph, and we can avoid unnecessary technical details by defining the Gibbs measure over a finite space, but the results below do not depend on this choice. For a path  $P$  we will let  $|P|$  denote its length, i.e. the total number of edges it contains.

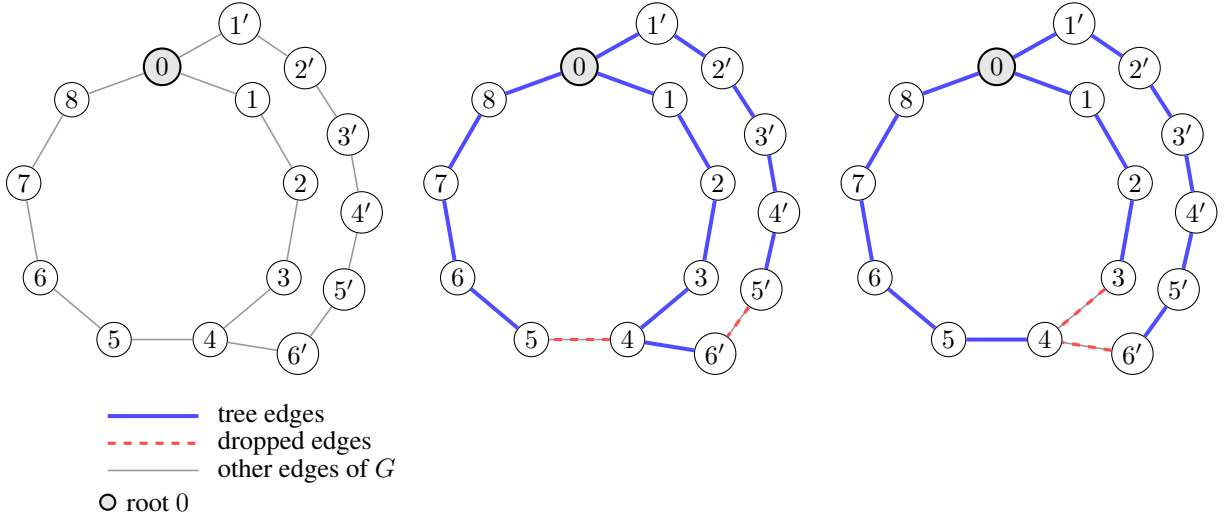


Figure 11: Graph  $G$  and the spanning trees  $T_1$  and  $T_2$  from Example 3 with  $m=4$ .

### L.1. Basic properties of Gibbs measures on paths

Below we show that the model is in a low temperature phase when  $\beta > 1$  and describe its structure.

**Remark 138** *It is not hard to show with similar ideas that Gibbs measure on paths exhibits a phase transition at  $\beta = 1$ . Informally, at low temperature ( $\beta > 1$ ) the measure is almost completely supported on the true shortest paths, whereas at high temperature ( $\beta < 1$ ) the measure is close to uniform and typical sample paths are much longer than the shortest path. Since the low temperature regime is most relevant to the results of [Li and Schramm \(2025\)](#), we omit the details.*

**Lemma 139** *Fix  $\beta > 1$ . With probability  $1 - o(1)$ , the following holds true uniformly over all  $r \geq 1$ . Let  $\ell^*$  be the length of the shortest path from 1 to 2 in the random graph, then*

$$\nu_{G,\beta}(\{P' : |P'| \geq \ell^* + r\}) \leq 2/\log(n)^{(\beta-1)r}.$$

Furthermore,

$$\sum_{\ell \geq \ell^* + r} |\{P : |P| = \ell\}| \log(n)^{-\beta\ell} \leq 2 \log(n)^{-\beta\ell^*} \log(n)^{-(\beta-1)r}.$$

**Proof** Since  $\ell^*$  be the length of the shortest path from 1 to 2,  $Z_{G,\beta} \geq \exp(-\beta\ell^* \log \log(n))$ . Observe that

$$\nu_{G,\beta}(\mathcal{P} \setminus \arg \min |P|) = \frac{1}{Z_{G,\beta}} \sum_{\ell \geq \ell^* + r} |\{P : |P| = \ell\}| \log(n)^{-\beta\ell} \leq \frac{1}{Z_{G,\beta}} \sum_{\ell \geq \ell^* + r} (\Delta/\log(n)^\beta)^\ell \leq \sum_{s \geq r} (\Delta/\log(n)^\beta)^s.$$

where  $\Delta$  is the maximum degree in the graph. Since  $\Delta = O(\log n)$  asymptotically almost surely and  $\beta > 1$ , the geometric series converges to at most

$$2(\Delta/\log(n)^\beta)^r \leq 2/\log(n)^{(\beta-1)r}.$$

The second inequality follows from essentially the same argument. ■

**Lemma 140** Fix  $\beta > 1$ . There exists  $\epsilon = o(1)$  as  $n \rightarrow \infty$  such that

$$\nu_{G,\beta} = (1 - \epsilon) \text{Uni}(\arg \min |P|) + \epsilon$$

with probability  $1 - o(1)$ . Equivalently,  $\nu_{G,\beta}$  is equal to the uniform distribution over the set of shortest paths from 1 to 2 up to  $o(1)$  Huber contamination [Huber \(1964\)](#).

**Proof** Since the restriction of  $\nu$  to  $P \in \arg \min |P|$  is the uniform measure, the conclusion follows from the previous lemma applied with  $r = 1$ .  $\blacksquare$

## L.2. Key structural result from Li–Schramm [Li and Schramm \(2025\)](#)

Li and Schramm showed that the set of approximate optima for the shortest path problem satisfy a very strong structural condition [Li and Schramm \(2025\)](#).

**Definition 141** For  $\epsilon > 0$ , let  $\mathcal{P}_\epsilon(G)$  be the set of paths from vertex 1 to 2 whose length is at most  $(1 + \epsilon)$  times longer than the shortest path from 1 to 2.

**Theorem 142 (Special case of Theorem 2.2 of [Li and Schramm \(2025\)](#))** There exists a constant  $C > 0$  such that for any  $\epsilon > 0$  sufficiently small, the following is true. With probability at least  $1 - O\left(\frac{\log \log n}{\log n}\right)$ , all pairs  $p, p' \in \mathcal{P}_\epsilon$  are nearly disjoint in the following sense:

$$\frac{|p \cap p'|}{\sqrt{|p||p'|}} \leq C\epsilon.$$

They also obtained an estimate of the number of  $\epsilon$ -shortest paths:

**Lemma 143 (Lemma 2.1 of [Li and Schramm \(2025\)](#))** For all  $\epsilon > 0$  sufficiently small, it holds that with probability  $1 - O\left(\frac{\log n}{\log \log n}\right)$  that

$$|\mathcal{P}_\epsilon| = (1 \pm o(1))n^\epsilon.$$

## L.3. Free energy barriers and Franz–Parisi Potential

Recall that the (asymptotic) Franz–Parisi potential [Franz and Parisi \(1995\)](#) (see, e.g., [El Alaoui et al. \(2025\)](#) for the formal mathematical definition below, up to the conventional factor of  $-1/\beta$ ) is given, assuming the limit exists, by

$$\mathcal{F}_\beta(q) = \lim_{\epsilon \rightarrow 0} \text{p-lim}_{n \rightarrow \infty} \frac{-1}{\beta \log n} \log \sum_{P': |P \cap P'| / \sqrt{|P||P'|} \in (q - \epsilon, q + \epsilon)} e^{-\beta \log \log(n) |P'|}$$

where  $P$  is a sample from  $\nu_{G,\beta}$ , the Gibbs measure over paths at inverse temperature  $\beta$ . With this sign convention, if the FPP is not *quasiconvex*, i.e. if it has multiple local minima then we say there is a free energy barrier. In many cases, including this example, quasiconvexity fails because the potential is not monotonically increasing from the global minimum. Glauber dynamics and similar processes can only “roll downward” from  $q = 1$ , so they need the monotone increasing property to be able to escape from their initialization in a small amount of time. See [Appendix J.6](#) for more.

For our purposes, we will work with the corresponding nonasymptotic version

$$\mathcal{F}_\beta(q, n, \epsilon, P) = \frac{-1}{\beta \log n} \log \sum_{P': |P \cap P'| / \sqrt{|P||P'|} \in (q-\epsilon, q+\epsilon)} e^{-\beta \log \log(n) |P'|}$$

which has the same consequences for bottlenecks, dynamics, etc. This is for convenience since the existence of the limit is irrelevant here.

**Proposition 144** *There exists  $c > 0$  such that the following is true for all sufficiently small  $\epsilon > 0$  and sufficiently large  $n$ . For any  $\beta > 1$ , for all  $q \in (0.01, 0.99)$ ,*

$$\mathcal{F}_{\beta, n, \epsilon, P}(q) \geq \mathcal{F}_{\beta, n, \epsilon, P}(1) + c(\beta - 1)$$

*asymptotically almost surely for  $P \sim \nu_{G, \beta}$ .*

**Proof** By Lemma 140,  $P$  sampled from the Gibbs measure will be a true shortest path from 1 to 2 asymptotically almost surely. By Theorem 142, any path  $P' \neq P$  from 1 to 2 with length at most  $|P|(1 + \epsilon)$  will have

$$\frac{|P \cap P'|}{\sqrt{|P||P'|}} \leq C\epsilon.$$

Therefore by Lemma 139,

$$\frac{1}{\log(n)^{\beta \ell^*}} \sum_{P' \neq P: |P \cap P'| / \sqrt{|P||P'|} > C\epsilon} \leq 2 / \log(n)^{(\beta-1)\epsilon |P|}.$$

Since  $|P|$  is asymptotically almost surely of order  $\log(n) / \log \log(n)$  (e.g. by Theorem 15), and  $\mathcal{F}_\beta(1) \geq \log(n)^{-\beta \ell^*}$ , the conclusion follows. ■

**Remark 145 (Alternative free energy barrier)** *We can also show a free energy barrier from the structural result in a slightly different way, by arguing there are no paths (sufficiently) close to an optimal  $P$ . Because for them to have large overlap, they need to have similar length  $|P \cap P'| / \sqrt{|P||P'|} \leq \sqrt{|P|/|P'|}$  so if  $|P'| > (1+\epsilon)|P|$  then the overlap is at most  $\sqrt{1/(1+\epsilon)} \leq 1 - c\epsilon$ . In terms of the Franz–Parisi potential, this alternative argument gives a taller but less “wide” gap than the proposition, so it is slightly weaker in terms of the locality of chains it applies to.*

## Appendix M. An analogous example: local search on multilinear vs Lovász extension

In this section, we consider the consequences of the fact that minimizing the *multilinear* and *Lovász* extensions of a submodular function  $f$  are obviously polynomial time equivalent — the global minimum of either is simply the global minimum of  $f$  on sets — but their optimization landscapes are very different. This illustrates the point that (1) the landscape of two polynomial time equivalent optimization problems can be very different, and (2) this difference is meaningful/useful from an algorithm design perspective. To rephrase point (2), it is helpful to know that the optimization landscape of the Lovász extension is benign whereas that of the multilinear extension is not, so we can know to perform local search on the former rather than the latter.

We also illustrate how the landscape of the multilinear extension is reflected in the Franz–Parisi potential for a related high-dimensional Ising model, which implies rigorous metastability results for the Glauber dynamics at low temperature.

**Optimization of quadratic polynomials, supermodularity, and Lovász extension.** Any quadratic polynomial of the form  $f(x) = \langle x, Mx \rangle / 2 + \langle x, h \rangle$  where  $M_{ij}$  is an  $n \times n$  matrix with zero diagonal and nonnegative entries can in fact be maximized over  $[-1, 1]^n$  in polynomial time. This is because the maximum is achieved at the vertices, and the maximization over the vertices is a supermodular maximization problem<sup>20</sup>, equivalently a submodular minimization of  $-f$ , so it can be solved in polynomial time. Explicitly, we can define the Lovász extension [Lovász \(1983\)](#)

$$\hat{f}(x) = \frac{1}{2} \int_{-1}^1 f((-1)^{1[x_1 < t]}, \dots, (-1)^{1[x_n < t]}) dt.$$

which agrees with  $f$  on  $\{\pm 1\}^n$ . The fundamental property of the Lovász extension is that because  $-f$  is supermodular,  $-\hat{f}$  is convex, i.e.  $\hat{f}$  is a concave function on  $[-1, 1]^n$ .

### M.1. Local search in a simple example

**An example.** Define

$$f(x, y, z, w) = x + y - z - w + xy + zw + \frac{3}{4}(x + y)(z + w), \quad (x, y, z, w) \in [-1, 1]^4.$$

We consider minimizing  $-f$ , equivalently maximizing  $f$ . A direct check over vertices shows  $f(1, 1, 1, 1) = f(-1, -1, -1, -1) = 5$  are the global maxima, while  $f(1, 1, -1, -1) = 3$ . Gradient flow on  $-f$  goes to the suboptimal point  $(1, 1, -1, -1)$  rather than the global optima. This happens because  $f$  is the *multilinear* extension of its restriction to  $\{\pm 1\}^n$ , which is not suitable for direct optimization.

On the other hand, as we will discuss afterwards, optimizing  $f$  is equivalent to maximizing the Lovász extension  $\hat{f}$  of  $f$  restricted to  $\{\pm 1\}^4 \rightarrow \mathbb{R}$ , which is a concave optimization problem and thus can successfully be solved via local search *on the correct choice of objective/landscape*.

**Failure of gradient flow on multilinear extension.** Then

$$\nabla f(x, y, z, w) = (1 + y + \frac{3}{4}(z + w), 1 + x + \frac{3}{4}(z + w), -1 + w + \frac{3}{4}(x + y), -1 + z + \frac{3}{4}(x + y)).$$

The unconstrained gradient (ascent) flow is

$$\frac{d}{dt}(x, y, z, w) = \nabla f(x, y, z, w) \tag{182}$$

On the box  $[-1, 1]^4$  we consider the *projected* gradient flow, i.e. we follow (182) while in the interior and, upon hitting the boundary, we project the velocity onto the tangent cone of  $[-1, 1]^4$  at the current point.

**Proposition 146** *For the projected ascent flow with initial condition  $(x(0), y(0), z(0), w(0)) = (0, 0, 0, 0)$ , the trajectory hits the point  $(1, 1, -1, -1)$  at time  $t^* = \ln 4$  and remains there for all  $t \geq t^*$ . In particular, the flow started at the origin converges to  $(1, 1, -1, -1)$  even though the global maximizers of  $f$  on  $[-1, 1]^4$  are  $(1, 1, 1, 1)$  and  $(-1, -1, -1, -1)$ .*

20. Here we consider  $\{\pm 1\}^n$  in bijective correspondence with  $2^{[n]}$  by viewing  $-1$  in coordinate  $i$  as indicating that element  $i$  is not present in the set, and  $+1$  as indicating that the element is included.

**Proof** The ODE (182) is invariant under the symmetry  $x \leftrightarrow y$ ,  $z \leftrightarrow w$ , and  $(x, y, z, w) \mapsto (y, x, w, z)$ . With the symmetric initial condition, the solution satisfies

$$x(t) = y(t) =: s(t), \quad z(t) = w(t) = -s(t)$$

for as long as it remains in the interior. Substituting into (182) gives a single scalar ODE

$$\dot{s} = 1 + s + \frac{3}{2}(-s) = 1 - \frac{1}{2}s, \quad s(0) = 0,$$

whose solution is

$$s(t) = 2(1 - e^{-t/2}).$$

The first contact with the box occurs when  $s(t) = 1$ , i.e.  $2(1 - e^{-t/2}) = 1$ , which yields  $t^* = \ln 4$ . At  $t = t^*$  we reach  $(x, y, z, w) = (1, 1, -1, -1)$ .

At this point,

$$\nabla f(1, 1, -1, -1) = (1.5, 1.5, -0.5, -0.5).$$

The active constraints are  $x = 1$ ,  $y = 1$ ,  $z = -1$ ,  $w = -1$ . The tangent cone therefore consists of velocities with  $\dot{x} \leq 0$ ,  $\dot{y} \leq 0$ ,  $\dot{z} \geq 0$ ,  $\dot{w} \geq 0$ . Projecting the ascent direction  $(1.5, 1.5, -0.5, -0.5)$  onto this cone gives the zero vector, so the projected flow has  $\dot{x} = \dot{y} = \dot{z} = \dot{w} = 0$  there. Thus the trajectory stays at  $(1, 1, -1, -1)$  for all  $t \geq t^*$ .  $\blacksquare$

**Success of local search on Lovász extension in the same example.** As a consequence of the general theory, we know that  $\hat{f}$  is concave and that projected gradient ascent (i.e., subgradient descent on  $-\hat{f}$ ), as well as other convex optimization procedures like the ellipsoid method, will converge to a global optimum. Explicitly, in the example of zero initialization, we have  $\hat{f}(0, 0, 0, 0) = (1/2)f(1, 1, 1, 1) + (1/2)f(-1, -1, -1, -1) = 5$  so the all-zeros initialization is already a global optimum. From an algorithmic perspective, the fact that  $\hat{f}$  is concave and locally maximized at  $(0, 0, 0, 0)$  tells us that it is a global optima of  $\hat{f}$ , and by the definition of the Lovász extension this exactly tells us that both  $f(1, 1, 1, 1)$  and  $f(-1, -1, -1, -1)$  are global maxima of  $f$ .

## M.2. High-dimensional version

We can define a high-dimensional version of the previous example by, for any  $m \geq 1$ , defining

$$f_m(x, y, z, w) = f(\bar{x}, \bar{y}, \bar{z}, \bar{w}) = f\left(\frac{1}{m} \sum_{i=1}^m x_i, \frac{1}{m} \sum_{i=1}^m y_i, \frac{1}{m} \sum_{i=1}^m z_i, \frac{1}{m} \sum_{i=1}^m w_i\right).$$

Here  $\bar{x} = \frac{1}{m} \sum_{i=1}^m x_i$  is a shorthand for the mean of the  $x$  vector.

### M.2.1. ISING MODEL AND FRANZ-PARISI POTENTIAL

We can define a corresponding mean-field Ising model on  $\{\pm 1\}^{4m}$  at inverse temperature  $\beta > J_0$  by its Gibbs measure

$$p_{\beta, m}(x, y, z, w) = \frac{1}{Z_{\beta, m}} \exp(\beta m f_m(x, y, z, w)).$$

Let  $Z = Z_{\beta,m}$  be the normalizing constant. From the definition, we can compute<sup>21</sup> that

$$Z = \sum_{\bar{x}, \bar{y}, \bar{z}, \bar{w}} e^{\beta m f(\bar{x}, \bar{y}, \bar{z}, \bar{w}) + S_m(x, y, z, w)}$$

where

$$\begin{aligned} S_m(\bar{x}, \bar{y}, \bar{z}, \bar{w}) &= \#\{x : \frac{1}{m} \sum_i x_i = \bar{x}, \dots, \frac{1}{m} \sum_i w_i = \bar{w}\} \\ &= m \left( h\left(\frac{1+\bar{x}}{2}\right) + h\left(\frac{1+\bar{y}}{2}\right) + h\left(\frac{1+\bar{z}}{2}\right) + h\left(\frac{1+\bar{w}}{2}\right) \right) + O(\log m) \end{aligned}$$

where  $h(p) = H(\text{Ber}(p)) = p \log(1/p) + (1-p) \log(1/(1-p))$  is the binary entropy function and the second equality follows by applying Stirling's approximation to the binomial coefficients. Defining the Franz–Parisi potential as the limit

$$\mathcal{F}_\beta(r) = -\frac{1}{\beta} \lim_{\epsilon \rightarrow 0} \lim_{m \rightarrow \infty} \frac{1}{m} \log \sum_{x', y', z', w'} e^{\beta m f_m(x, y, z, w)} \mathbf{1}(\langle (x, y, z, w), (x', y', z', w') \rangle / 4m \in [r - \epsilon, r + \epsilon])$$

where  $(x, y, z, w)$  are sampled from the Gibbs measure at inverse temperature  $\beta$ , we can therefore compute that in the zero-temperature limit  $\beta \rightarrow \infty$ ,

$$\mathcal{F}_\infty(r) := \lim_{\beta \rightarrow \infty} \mathcal{F}_\beta(r) = (-1) \sup_{\bar{x}, \bar{y}, \bar{z}, \bar{w} \in K(r)} f(\bar{x}, \bar{y}, \bar{z}, \bar{w})$$

where

$$K(r) = \{(\bar{x}, \bar{y}, \bar{z}, \bar{w}) \in [-1, 1]^4 : (\bar{x} + \bar{y} + \bar{z} + \bar{w})/4 = r\}$$

because the global maxima of  $f$  are at  $(-1, -1, -1, -1)$  and  $(1, 1, 1, 1)$ .

### M.2.2. ANALYTICAL COMPUTATION OF FRANZ–PARISI POTENTIAL

We analytically compute  $\mathcal{F}_\infty(r)$ , which corresponds to solving a constrained maximization of  $f(x, y, z, w)$  with a linear constraint. Since  $\mathcal{F}_\beta(r) = \mathcal{F}_\infty(r) + o(1)$  as  $\beta \rightarrow \infty$ , we know that the finite but low-temperature Franz–Parisi potential will be a perturbation of the zero-temperature one.

**Symmetrization.** By symmetry within the pairs  $(x, y)$  and  $(z, w)$ , we may assume the maximizer satisfies

$$x = y =: a, \quad z = w =: b.$$

Indeed, for fixed  $x + y$  the product  $xy$  is maximized at  $x = y$ , and similarly  $zw$  at  $z = w$ . Hence it suffices to optimize over

$$a, b \in [-1, 1], \quad a + b = 2r.$$

21. The computations here are similar to those in the Curie-Weiss/mean-field Ising model, so we omit unnecessary details. See, e.g., [Ellis \(2007\)](#) for the details of the calculation in the case of the Curie-Weiss model.

**Reduction to one variable.** With  $x = y = a, z = w = b$ ,

$$f(a, b) = 2a - 2b + a^2 + b^2 + 3ab.$$

Let  $s := a + b = 2r$  and  $d := a - b$ . Then  $ab = \frac{s^2 - d^2}{4}$ , and the box constraints imply

$$a, b \in [-1, 1] \iff |d| \leq 2 - |s| = 2 - 2|r|.$$

A short calculation gives

$$f = 2d - \frac{1}{4}d^2 + \frac{5}{4}s^2 = 2d - \frac{1}{4}d^2 + 5r^2.$$

The function  $d \mapsto 2d - \frac{1}{4}d^2$  is concave with unconstrained maximizer at  $d = 4$ , so under the constraint  $|d| \leq 2 - 2|r|$  the maximum is attained at

$$d^* = 2 - 2|r|.$$

**Value and maximizers.** Plugging  $s = 2r$  and  $d^*$  back,

$$\sup_{K(r)} f = 2(2 - 2|r|) - \frac{1}{4}(2 - 2|r|)^2 + 5r^2 = 4r^2 - 2|r| + 3.$$

Thus

$$\mathcal{F}_\infty(r) = -(4r^2 - 2|r| + 3), \quad r \in [-1, 1].$$

One maximizing configuration is

$$(r \geq 0) : (x, y, z, w) = (1, 1, 2r-1, 2r-1), \quad (r \leq 0) : (x, y, z, w) = (1+2r, 1+2r, -1, -1),$$

which satisfies  $(x + y + z + w)/4 = r$  and  $[-1, 1]$ -bounds.

### M.2.3. METASTABILITY AND CONTRAST TO LOG-CONCAVE SETTING

From the analytical solution at zero temperature, we see that the Franz–Parisi potential is not quasi-convex but rather has 2 global minima and 1 sub-optimal local minima at all sufficiently large  $\beta > 0$ . So we conclude that this model does have exponentially large free energy barriers.

This rigorously rules out the possibility that low-temperature Glauber dynamics can escape from its initialization in  $e^{o(n)}$  time, and we also see visually the fact that the optimization initialized at some states can converge to the suboptimal local minima, as in the explicit analysis of gradient flow in the low-dimensional example.

On the other hand, while local search fails we again remark that this is not a hard optimization problem, since its global minimum of  $f_m$  can be found by minimizing the Lovász extension of the restriction of  $-f$  to  $\{\pm 1\}^{4m}$ . Because the Lovász extension of  $-f_m$  is convex, there will not be any free energy barriers in the corresponding landscape, and in fact we can sample from the corresponding log-concave Gibbs measures via MCMC. For example, the continuous-time Langevin diffusion in  $n$  dimensions is known to have relaxation time  $\text{polylog}(n)$  for log-concave measures in isotropic position due to recent progress on the KLS conjecture [Klartag and Lehec \(2022\)](#). This also implies Lipschitz concentration which shows that the overlap distribution is highly concentrated. See [Bandeira et al. \(2023\)](#) for more references and discussion about the Franz–Parisi potential in the log-concave case.

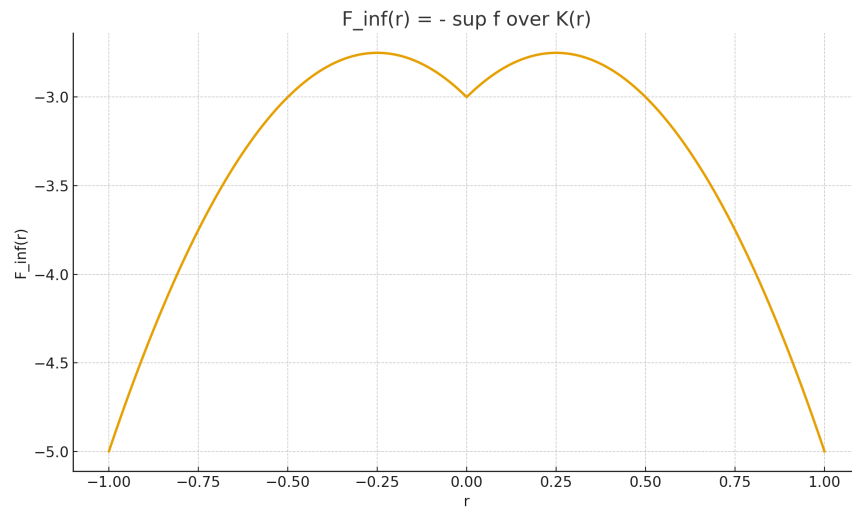


Figure 12: Plot of Franz–Parisi potential  $\mathcal{F}_{\infty}(r)$  in the Ising model example. There are three local minima, so there is an (exponentially large) free energy barrier.