Maximum entropy distributions on graphs

Christopher J. Hillar* Andre Wibisono†

University of California, Berkeley

August 26, 2013

Abstract

Inspired by the problem of sensory coding in neuroscience, we study the maximum entropy distribution on weighted graphs with a given expected degree sequence. This distribution on graphs is characterized by independent edge weights parameterized by vertex potentials at each node. Using the general theory of exponential family distributions, we prove the existence and uniqueness of the maximum likelihood estimator (MLE) of the vertex parameters. We also prove the consistency of the MLE from a single graph sample, extending the results of Chatterjee, Diaconis, and Sly for unweighted (binary) graphs. Interestingly, our extensions require an intricate study of the inverses of diagonally dominant positive matrices. Along the way, we derive analogues of the Erdős-Gallai criterion of graphical sequences for weighted graphs.

1 Introduction

Maximum entropy models are an important class of statistical models for biology. For instance, they have been found to be a good model for protein folding [29, 35], antibody diversity [24], neural population activity [31, 33, 37, 36, 4, 42, 34], and flock behavior [5]. In this paper we develop a general framework for studying maximum entropy distributions on weighted graphs, extending recent work of Chatterjee, Diaconis, and Sly [8]. The development of this theory is partly motivated by the problem of sensory coding in neuroscience.

In the brain, information is represented by discrete electrical pulses, called *action potentials* or *spikes* [28]. This includes neural representations of sensory stimuli which can take on a continuum of values. For instance, large photoreceptor arrays in the retina respond to a range of light intensities in a visual environment, but the brain does not receive information from these photoreceptors directly. Instead, retinal ganglion cells must convey this detailed input to the visual cortex using only a series of binary electrical signals. Continuous stimuli are therefore converted by networks of neurons to sequences of spike times.

An unresolved controversy in neuroscience is whether information is contained in the precise timings of these spikes or only in their "rates" (i.e., counts of spikes in a window of time). Early theoretical studies [22] suggest that information capacities of timing-based codes are superior to those that are rate-based (also see [16] for an implementation in a simple model). Moreover, a number of scientific articles have appeared suggesting that precise spike timing [1, 3, 26, 40, 21, 6, 23, 25, 11, 18] and synchrony [38] are important for various computations in the brain.[1] Here, we briefly explain a possible scheme for encoding continuous vectors with spiking neurons that takes advantage of precise spike timing and the mathematics of maximum entropy distributions.

*Redwood Center for Theoretical Neuroscience, `chillar@msri.org`; partially supported by National Science Foundation Grant IIS-0917342 and a National Science Foundation All-Institutes Postdoctoral Fellowship administered by the Mathematical Sciences Research Institute through its core grant DMS-0441170.

†Department of Electrical Engineering and Computer Science, `wibisono@eecs.berkeley.edu`.

[1] Although it is well-known that precise spike timing is used for time-disparity computation in animals [7], such as when owls track prey with binocular hearing or when electric fish use electric fields around their bodies for locating objects.

Consider a network of n neurons in one region of the brain which transmits a continuous vector $\theta \in \mathbb{R}^n$ using sequences of spikes to a second receiver region. We assume that this second region contains a number of coincidence detectors that measure the absolute difference in spike times between pairs of neurons projecting from the first region. We imagine three scenarios for how information can be obtained by these detectors. In the first, the detector is only measuring for synchrony between spikes; that is, either the detector assigns a 0 to a nonzero timing difference or a 1 to a coincidence of spikes. In another scenario, timing differences between projecting neurons can assume an infinite but countable number of possible values. Finally, in the third scenario, we allow these differences to take on any nonnegative real values. We further assume that neuronal output and thus spike times are stochastic variables. A basic question now arises: How can the first region encode θ so that it can be recovered robustly by the second?

We answer this question by first asking the one symmetric to this: How can the second region recover a real vector transmitted by an unknown sender region from spike timing measurements? We propose the following possible solution to this problem. Fix one of the detector mechanics as described above, and set a_{ij} to be the measurement of the absolute timing difference between spikes from projecting neurons i and j. We assume that the receiver population can compute the (local) sums $d_i = \sum_{j \neq i} a_{ij}$ efficiently. The values $\mathbf{a} = (a_{ij})$ represent a weighted graph G on n vertices, and we assume that a_{ij} is randomly drawn from a distribution on timing measurements (A_{ij}). Making no further assumptions, a principle of Jaynes [17] suggests that the second region propose that the timing differences are drawn from the (unique) distribution over weighted graphs with the highest entropy [32, 10] having the vector $\mathbf{d} = (d_1, \dots, d_n)$ for the expectations of the degree sums $\sum_{j \neq i} A_{ij}$. Depending on which of the three scenarios described above is true for the coincidence detector, this prescription produces one of three different maximum entropy distributions.

Consider the third scenario above (the other cases are also subsumed by our results). As we shall see in Section 3.2, the distribution determined in this case is parameterized by a real vector $\theta = (\theta_1, \dots, \theta_n)$, and finding the maximum likelihood estimator (MLE) for these parameters using $\mathbf{d}$ as sufficient statistics boils down to solving the following set of n algebraic equations in the n unknowns $\hat{\theta}_1, \dots, \hat{\theta}_n$:

$$d_i = \sum_{j \neq i} \frac{1}{\hat{\theta}_i + \hat{\theta}_j} \quad \text{for } i = 1, \dots, n. \tag{1}$$

Given our motivation, we call the system of equations (1) the *retina equations* for theoretical neuroscience, and note that they have been studied in a more general context by Sanyal, Sturmfels, and Vinzant [30] using matroid theory and algebraic geometry. Remarkably, a solution $\hat{\theta}$ to (1) has the property that with high probability, it is arbitrarily close to the original parameters θ for sufficiently large network sizes n (in the scenario of binary measurements, this is a result of [8]). In particular, it is possible for the receiver region to recover reliably a continuous vector θ from a *single* cycle of neuronal firing emanating from the sender region.

We now know how to answer our first question: *The sender region should arrange spike timing differences to come from a maximum entropy distribution.* We remark that this conclusion is consistent with modern paradigms in theoretical neuroscience and artificial intelligence, such as the concept of the Boltzmann machine [2], a stochastic version of its (zero-temperature) deterministic limit, the Little-Hopfield network [20, 15].

Organization. The organization of this paper is as follows. In Section 2, we lay out the general theory of maximum entropy distributions on weighted graphs. In Section 3, we specialize the general theory to three classes of weighted graphs. For each class, we provide an explicit characterization of the maximum entropy distributions and prove a generalization of the Erdős-Gallai criterion for weighted graphical sequences. Furthermore, we also present the consistency property of the MLE of the vertex parameters from one graph sample. Section 4 provides the proofs of the main technical results presented in Section 3. Finally, in Section 5 we discuss the results in this paper and some future research directions.

Notation. In this paper we use the following notation. Let $\mathbb{R}_+ = (0, \infty)$, $\mathbb{R}_0 = [0, \infty)$, $\mathbb{N} = \{1, 2, \dots\}$, and $\mathbb{N}_0 = \{0, 1, 2, \dots\}$. We write $\sum_{\{i,j\}}$ and $\prod_{\{i,j\}}$ for the summation and product, respectively, over all

$\binom{n}{2}$ pairs $\{i,j\}$ with $i \neq j$. For a subset $C \subseteq \mathbb{R}^n$, C° and $\overline{C}$ denote the interior and closure of C in $\mathbb{R}^n$, respectively. For a vector $x = (x_1, \dots, x_n) \in \mathbb{R}^n$, $\|x\|_1 = \sum_{i=1}^n |x_i|$ and $\|x\|_\infty = \max_{1\le i\le n} |x_i|$ denote the ℓ_1 and ℓ_∞ norms of x. For an $n \times n$ matrix $J = (J_{ij})$, $\|J\|_\infty$ denotes the matrix norm induced by the $\|\cdot\|_\infty$-norm on vectors in $\mathbb{R}^n$, that is,

$$\|J\|_\infty = \max_{x\neq 0} \frac{\|Jx\|_\infty}{\|x\|_\infty} = \max_{1\le i\le n} \sum_{j=1}^n |J_{ij}|.$$

2 General theory via exponential family distributions

In this section we develop the general machinery of maximum entropy distributions on graphs via the theory of exponential family distributions [41], and in subsequent sections we specialize our analysis to some particular cases of weighted graphs.

Consider an undirected graph G on $n \ge 3$ vertices with edge (i,j) having weight $a_{ij} \in S$, where $S \subseteq \mathbb{R}$ is the set of possible weight values. We will later consider the following specific cases:

1. *Finite discrete weighted graphs,* with edge weights in $S = \{0, 1, \dots, r-1\}$, $r \ge 2$.
2. *Infinite discrete weighted graphs,* with edge weights in $S = \mathbb{N}_0$.
3. *Continuous weighted graphs,* with edge weights in $S = \mathbb{R}_0$.

A graph G is fully specified by its *adjacency matrix* $\mathbf{a} = (a_{ij})_{i,j=1}^n$, which is an $n \times n$ symmetric matrix with zeros along its diagonal. For fixed n, a probability distribution over graphs G corresponds to a distribution over adjacency matrices $\mathbf{a} = (a_{ij}) \in S^{\binom{n}{2}}$. Given a graph with adjacency matrix $\mathbf{a} = (a_{ij})$, let $\deg_i(\mathbf{a}) = \sum_{j\neq i} a_{ij}$ be the degree of vertex i, and let $\deg(\mathbf{a}) = (\deg_1(\mathbf{a}), \dots, \deg_n(\mathbf{a}))$ be the degree sequence of $\mathbf{a}$.

2.1 Characterization of maximum entropy distribution

Let $\mathcal{S}$ be a σ-algebra over the set of weight values S, and assume there is a canonical σ-finite probability measure ν on $(S, \mathcal{S})$. Let $\nu^{\binom{n}{2}}$ be the product measure on $S^{\binom{n}{2}}$, and let $\mathfrak{P}$ be the set of all probability distributions on $S^{\binom{n}{2}}$ that are absolutely continuous with respect to $\nu^{\binom{n}{2}}$. Since $\nu^{\binom{n}{2}}$ is σ-finite, these probability distributions can be characterized by their density functions, i.e. the Radon-Nikodym derivatives with respect to $\nu^{\binom{n}{2}}$. Given a sequence $\mathbf{d} = (d_1, \dots, d_n) \in \mathbb{R}^n$, let $\mathfrak{P}_{\mathbf{d}}$ be the set of distributions in $\mathfrak{P}$ whose expected degree sequence is equal to $\mathbf{d}$,

$$\mathfrak{P}_{\mathbf{d}} = \{\mathbb{P} \in \mathfrak{P} \colon \mathbb{E}_{\mathbb{P}}[\deg(A)] = \mathbf{d}\},$$

where in the definition above, the random variable $A = (A_{ij}) \in S^{\binom{n}{2}}$ is drawn from the distribution $\mathbb{P}$. Then the distribution $\mathbb{P}^*$ in $\mathfrak{P}_{\mathbf{d}}$ with maximum entropy is precisely the exponential family distribution with the degree sequence as sufficient statistics [41, Chapter 3]. Specifically, the density of $\mathbb{P}^*$ at $\mathbf{a} = (a_{ij}) \in S^{\binom{n}{2}}$ is given by[2]

$$p^*(\mathbf{a}) = \exp\big(-\theta^\top \deg(\mathbf{a}) - Z(\theta)\big), \tag{2}$$

where $Z(\theta)$ is the *log-partition function*,

$$Z(\theta) = \log \int_{S^{\binom{n}{2}}} \exp\big(-\theta^\top \deg(\mathbf{a})\big)\, \nu^{\binom{n}{2}}(d\mathbf{a}),$$

and $\theta = (\theta_1, \dots, \theta_n)$ is a parameter that belongs to the *natural parameter space*

$$\Theta = \{\theta \in \mathbb{R}^n \colon Z(\theta) < \infty\}.$$

[2]We choose to use $-\theta$ in the parameterization (2), instead of the canonical parameterization $p^*(\mathbf{a}) \propto \exp(\theta^\top \deg(\mathbf{a}))$, because it simplifies the notation in our later presentation.

We will also write $\mathbb{P}^*_\theta$ if we need to emphasize the dependence of $\mathbb{P}^*$ on the parameter θ.

Using the definition $\deg_i(\mathbf{a}) = \sum_{j \neq i} a_{ij}$, we can write

$$\exp\big(-\theta^\top \deg(\mathbf{a})\big) = \exp\Big(-\sum_{\{i,j\}} (\theta_i + \theta_j) a_{ij}\Big) = \prod_{\{i,j\}} \exp\big(-(\theta_i + \theta_j) a_{ij}\big).$$

Hence, we can express the log-partition function as

$$Z(\theta) = \log \prod_{\{i,j\}} \int_S \exp\big(-(\theta_i + \theta_j) a_{ij}\big)\, \nu(da_{ij}) = \sum_{\{i,j\}} Z_1(\theta_i + \theta_j), \tag{3}$$

in which $Z_1(t)$ is the marginal log-partition function

$$Z_1(t) = \log \int_S \exp(-ta)\, \nu(da).$$

Consequently, the density in (2) can be written as

$$p^*(\mathbf{a}) = \prod_{\{i,j\}} \exp\big(-(\theta_i + \theta_j) a_{ij} - Z_1(\theta_i + \theta_j)\big).$$

This means the edge weights A_{ij} are independent random variables, with $A_{ij} \in S$ having distribution $\mathbb{P}^*_{ij}$ with density

$$p^*_{ij}(a) = \exp\big(-(\theta_i + \theta_j) a - Z_1(\theta_i + \theta_j)\big).$$

In particular, the edge weights A_{ij} belong to the same exponential family distribution but with different parameters that depend on θ_i and θ_j (or rather, on their sum $\theta_i + \theta_j$). The parameters $\theta_1, \dots, \theta_n$ can be interpreted as the potential at each vertex that determines how strongly the vertices are connected to each other. Furthermore, we can write the natural parameter space Θ as

$$\Theta = \{\theta \in \mathbb{R}^n \colon Z_1(\theta_i + \theta_j) < \infty \;\text{ for all }\; i \neq j\}.$$

2.2 Maximum likelihood estimator and moment-matching equation

Using the characterization of $\mathbb{P}^*$ as the maximum entropy distribution in $\mathfrak{P}_{\mathbf{d}}$, the condition $\mathbb{P}^* \in \mathfrak{P}_{\mathbf{d}}$ means we need to choose the parameter θ for $\mathbb{P}^*_\theta$ such that $\mathbb{E}_\theta[\deg(A)] = \mathbf{d}$.[3] This is an instance of the moment-matching equation, which, in the case of exponential family distributions, is well-known to be equivalent to finding the maximum likelihood estimator (MLE) of θ given an empirical degree sequence $\mathbf{d} \in \mathbb{R}^n$.

Specifically, suppose we draw graph samples $G_1, \dots, G_m$ i.i.d. from the distribution $\mathbb{P}^*$ with parameter θ^*, and we want to find the MLE $\hat{\theta}$ of θ^* based on the observations $G_1, \dots, G_m$. Using the parametric form of the density (2), this is equivalent to solving the maximization problem

$$\max_{\theta \in \Theta} \mathcal{F}(\theta) \equiv -\theta^\top \mathbf{d} - Z(\theta),$$

where $\mathbf{d}$ is the average of the degree sequences of $G_1, \dots, G_m$. Setting the gradient of $\mathcal{F}(\theta)$ to zero reveals that the MLE $\hat{\theta}$ satisfies

$$-\nabla Z(\hat{\theta}) = \mathbf{d}. \tag{4}$$

Recall that the gradient of the log-partition function in an exponential family distribution is equal to the expected sufficient statistics. In our case, we have $-\nabla Z(\hat{\theta}) = \mathbb{E}_{\hat{\theta}}[\deg(A)]$, so the MLE equation (4) recovers the moment-matching equation $\mathbb{E}_{\hat{\theta}}[\deg(A)] = \mathbf{d}$.

[3] Here we write $\mathbb{E}_\theta$ in place of $\mathbb{E}_{\mathbb{P}^*}$ to emphasize the dependence of the expectation on the parameter θ.

In Section 3 we study the properties of the MLE of θ from a *single* sample $G \sim \mathbb{P}^*_\theta$. In the remainder of this section, we address the question of the existence and uniqueness of the MLE with a given empirical degree sequence $\mathbf{d}$.

Define the *mean parameter space* $\mathcal{M}$ to be the set of expected degree sequences from all distributions on $S^{\binom{n}{2}}$ that are absolutely continuous with respect to $\nu^{\binom{n}{2}}$,

$$\mathcal{M} = \{\mathbb{E}_{\mathbb{P}}[\deg(A)] \colon \mathbb{P} \in \mathfrak{P}\}.$$

The set $\mathcal{M}$ is necessarily convex, since a convex combination of probability distributions in $\mathfrak{P}$ is also a probability distribution in $\mathfrak{P}$. Recall that an exponential family distribution is *minimal* if there is no linear combination of the sufficient statistics that is constant almost surely with respect to the base distribution. This minimality property clearly holds for $\mathbb{P}^*$, for which the sufficient statistics are the degree sequence. We say that $\mathbb{P}^*$ is *regular* if the natural parameter space Θ is open. By the general theory of exponential family distributions [41, Theorem 3.3], in a regular and minimal exponential family distribution, the gradient of the log-partition function maps the natural parameter space Θ to the interior of the mean parameter space $\mathcal{M}$, and this mapping[4]

$$-\nabla Z \colon \Theta \to \mathcal{M}^\circ$$

is bijective. We summarize the preceding discussion in the following result.

Proposition 2.1. *Assume Θ is open. Then there exists a solution $\theta \in \Theta$ to the MLE equation $\mathbb{E}_\theta[\deg(A)] = \mathbf{d}$ if and only if $\mathbf{d} \in \mathcal{M}^\circ$, and if such a solution exists then it is unique.*

We now characterize the mean parameter space $\mathcal{M}$. We say that a sequence $\mathbf{d} = (d_1, \dots, d_n)$ is *graphic* (or a *graphical sequence*) if $\mathbf{d}$ is the degree sequence of a graph G with edge weights in S, and in this case we say that G *realizes* $\mathbf{d}$. It is important to note that whether a sequence $\mathbf{d}$ is graphic depends on the weight set S, which we consider fixed for now.

Proposition 2.2. *Let $\mathcal{W}$ be the set of all graphical sequences, and let conv($\mathcal{W}$) be the convex hull of $\mathcal{W}$. Then $\mathcal{M} \subseteq$ conv($\mathcal{W}$). Furthermore, if $\mathfrak{P}$ contains the Dirac delta measures, then $\mathcal{M} =$ conv($\mathcal{W}$).*

Proof. The inclusion $\mathcal{M} \subseteq \mathrm{conv}(\mathcal{W})$ is clear, since any element of $\mathcal{M}$ is of the form $\mathbb{E}_{\mathbb{P}}[\deg(A)]$ for some distribution $\mathbb{P}$ and $\deg(A) \in \mathcal{W}$ for every realization of the random variable A. Now suppose $\mathfrak{P}$ contains the Dirac delta measures δ_B for each $B \in S^{\binom{n}{2}}$. Given $\mathbf{d} \in \mathcal{W}$, let B be the adjacency matrix of the graph that realizes $\mathbf{d}$. Then $\mathbf{d} = \mathbb{E}_{\delta_B}[\deg(A)] \in \mathcal{M}$, which means $\mathcal{W} \subseteq \mathcal{M}$, and hence $\mathrm{conv}(\mathcal{W}) \subseteq \mathcal{M}$ since $\mathcal{M}$ is convex. □

As we shall see in Section 3, the result above allows us to conclude that $\mathcal{M} = \mathrm{conv}(\mathcal{W})$ for the case of discrete weighted graphs. On the other hand, for the case of continuous weighted graphs we need to prove $\mathcal{M} = \mathrm{conv}(\mathcal{W})$ directly since $\mathfrak{P}$ in this case does not contain the Dirac measures.

Remark 2.3. We emphasize the distinction between a *valid* solution $\theta \in \Theta$ and a *general* solution $\theta \in \mathbb{R}^n$ to the MLE equation $\mathbb{E}_\theta[\deg(A)] = \mathbf{d}$. As we saw from Proposition 2.1, we have a precise characterization of the existence and uniqueness of the valid solution $\theta \in \Theta$, but in general, there are multiple solutions θ to the MLE equation. In this paper we shall be concerned only with the valid solution; Sanyal, Sturmfels, and Vinzant study some algebraic properties of general solutions [30].

We close this section by discussing the symmetry of the valid solution to the MLE equation. Recall the decomposition (3) of the log-partition function $Z(\theta)$ into the marginal log-partition functions $Z_1(\theta_i + \theta_j)$. Let $\mathrm{Dom}(Z_1) = \{t \in \mathbb{R} \colon Z_1(t) < \infty\}$, and let $\mu \colon \mathrm{Dom}(Z_1) \to \mathbb{R}$ denote the (marginal) *mean function*

$$\mu(t) = \int_S a \, \exp\big(-ta - Z_1(t)\big) \, \nu(da).$$

[4]The mapping is $-\nabla Z$, instead of ∇Z, because of our choice of the parameterization in (2) using $-\theta$.

Observing that we can write

$$\mathbb{E}_\theta[A_{ij}] = \int_S a \, \exp\big(-(\theta_i+\theta_j)a - Z_1(\theta_i+\theta_j)\big)\, \nu(da) = \mu(\theta_i+\theta_j),$$

the MLE equation $\mathbb{E}_\theta[\deg(A)] = \mathbf{d}$ then becomes

$$d_i = \sum_{j\neq i} \mu(\theta_i+\theta_j) \quad \text{for } i = 1,\dots,n. \tag{5}$$

In the statement below, sgn denotes the sign function: $\text{sgn}(t) = t/|t|$ if $t \neq 0$, and $\text{sgn}(0) = 0$.

Proposition 2.4. *Let $\mathbf{d} \in \mathcal{M}^\circ$, and let $\theta \in \Theta$ be the unique solution to the system of equations* (5)*. If μ is strictly increasing, then*

$$sgn(d_i - d_j) = sgn(\theta_i - \theta_j) \quad \textit{for all } i \neq j,$$

and similarly, if μ is strictly decreasing, then

$$sgn(d_i - d_j) = sgn(\theta_j - \theta_i) \quad \textit{for all } i \neq j.$$

Proof. Given $i \neq j$,

$$\begin{aligned} d_i - d_j &= \Big(\mu(\theta_i+\theta_j) + \sum_{k\neq i,j} \mu(\theta_i+\theta_k)\Big) - \Big(\mu(\theta_j+\theta_i) + \sum_{k\neq i,j} \mu(\theta_j+\theta_k)\Big) \\ &= \sum_{k\neq i,j} \big(\mu(\theta_i+\theta_k) - \mu(\theta_j+\theta_k)\big). \end{aligned}$$

If μ is strictly increasing, then $\mu(\theta_i+\theta_k) - \mu(\theta_j+\theta_k)$ has the same sign as $\theta_i - \theta_j$ for each $k \neq i,j$, and thus $d_i - d_j$ also has the same sign as $\theta_i - \theta_j$. Similarly, if μ is strictly decreasing, then $\mu(\theta_i+\theta_k) - \mu(\theta_j+\theta_k)$ has the opposite sign of $\theta_i - \theta_j$, and thus $d_i - d_j$ also has the opposite sign of $\theta_i - \theta_j$. □

3 Analysis for specific edge weights

In this section we analyze the maximum entropy random graph distributions for several specific choices of the weight set S. For each case, we specify the distribution of the edge weights A_{ij}, the mean function μ, the natural parameter space Θ, and characterize the mean parameter space $\mathcal{M}$. We also study the problem of finding the MLE $\hat{\theta}$ of θ from one graph sample $G \sim \mathbb{P}^*_\theta$ and prove the existence, uniqueness, and consistency of the MLE. Along the way, we derive analogues of the Erdős-Gallai criterion of graphical sequences for weighted graphs. We defer the proofs of the results presented here to Section 4.

3.1 Finite discrete weighted graphs

We first study weighted graphs with edge weights in the finite discrete set $S = \{0,1,\dots,r-1\}$, where $r \geq 2$. The case $r = 2$ corresponds to unweighted graphs, and our analysis in this section recovers the results of Chatterjee, Diaconis, and Sly [8]. The proofs of the results in this section are provided in Section 4.2.

3.1.1 Characterization of the distribution

We take ν to be the counting measure on S. Following the development in Section 2, the edge weights $A_{ij} \in S$ are independent random variables with density

$$p^*_{ij}(a) = \exp\big(-(\theta_i+\theta_j)a - Z_1(\theta_i+\theta_j)\big), \quad 0 \leq a \leq r-1,$$

where the marginal log-partition function Z_1 is given by

$$Z_1(t) = \log \sum_{a=0}^{r-1} \exp(-at) = \begin{cases} \log \frac{1-\exp(-rt)}{1-\exp(-t)} & \text{if } t \neq 0, \\ \log r & \text{if } t = 0. \end{cases}$$

Since $Z_1(t) < \infty$ for all $t \in \mathbb{R}$, the natural parameter space $\Theta = \{\theta \in \mathbb{R}^n \colon Z_1(\theta_i + \theta_j) < \infty, i \neq j\}$ is given by $\Theta = \mathbb{R}^n$. The mean function is given by

$$\mu(t) = \sum_{a=0}^{r-1} a \exp(-at - Z_1(t)) = \frac{\sum_{a=0}^{r-1} a \ \exp(-at)}{\sum_{a=0}^{r-1} \exp(-at)}. \tag{6}$$

At $t = 0$ the mean function takes the value

$$\mu(0) = \frac{\sum_{a=0}^{r-1} a}{r} = \frac{r-1}{2},$$

while for $t \neq 0$, the mean function simplifies to

$$\mu(t) = -\left(\frac{1-\exp(-t)}{1-\exp(-rt)}\right) \cdot \frac{d}{dt} \sum_{a=0}^{r-1} \exp(-at) = \frac{1}{\exp(t)-1} - \frac{r}{\exp(rt)-1}. \tag{7}$$

Figure 1 shows the behavior of the mean function $\mu(t)$ and its derivative $\mu'(t)$ as r varies.

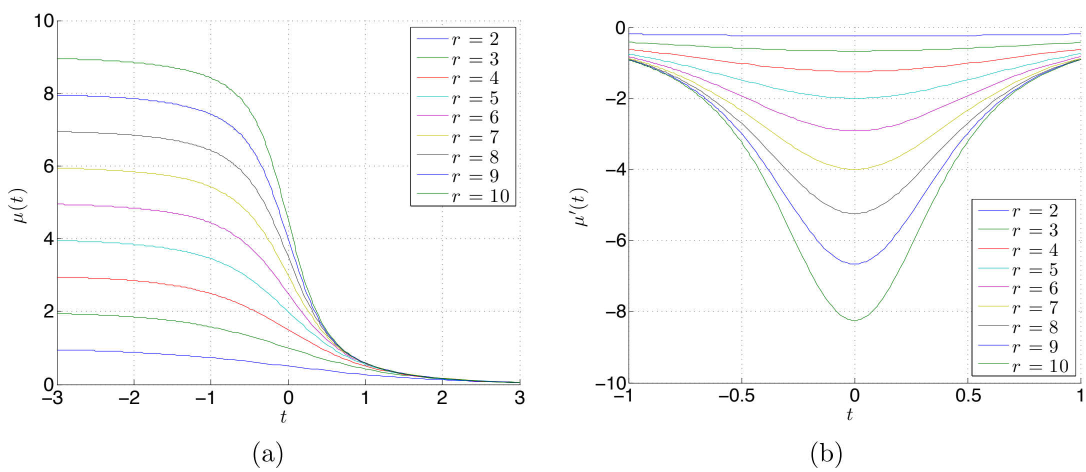

Figure 1: Plot of the mean function $\mu(t)$ (left) and its derivative $\mu'(t)$ (right) as r varies.

Remark 3.1. For $r = 2$, the edge weights A_{ij} are independent Bernoulli random variables with

$$\mathbb{P}^*(A_{ij} = 1) = \mu(\theta_i + \theta_j) = \frac{\exp(-\theta_i - \theta_j)}{1 + \exp(-\theta_i - \theta_j)} = \frac{1}{1 + \exp(\theta_i + \theta_j)}.$$

As noted above, this is the model recently studied by Chatterjee, Diaconis, and Sly [8] in the context of graph limits. When $\theta_1 = \theta_2 = \cdots = \theta_n = t$, we recover the classical Erdős-Rényi model with edge emission probability $p = 1/(1 + \exp(2t))$.

3.1.2 Existence, uniqueness, and consistency of the MLE

Consider the problem of finding the MLE of θ from one graph sample. Specifically, let $\theta \in \Theta$ and suppose we draw a sample $G \sim \mathbb{P}^*_\theta$. Then, as we saw in Section 2, the MLE $\hat{\theta}$ of θ is a solution to the moment-matching equation $\mathbb{E}_{\hat{\theta}}[\deg(A)] = \mathbf{d}$, where $\mathbf{d}$ is the degree sequence of the sample graph G. As in (5), the moment-matching equation is equivalent to the following system of equations:

$$d_i = \sum_{j \neq i} \mu(\hat{\theta}_i + \hat{\theta}_j), \quad i = 1, \ldots, n. \tag{8}$$

Since the natural parameter space $\Theta = \mathbb{R}^n$ is open, Proposition 2.1 tells us that the MLE $\hat{\theta}$ exists and is unique if and only if the empirical degree sequence $\mathbf{d}$ belongs to the interior $\mathcal{M}^\circ$ of the mean parameter space $\mathcal{M}$.

We also note that since $\nu^{\binom{n}{2}}$ is the counting measure on $S^{\binom{n}{2}}$, all distributions on $S^{\binom{n}{2}}$ are absolutely continuous with respect to $\nu^{\binom{n}{2}}$, so $\mathfrak{P}$ contains all probability distributions on $S^{\binom{n}{2}}$. In particular, $\mathfrak{P}$ contains the Dirac measures, and by Proposition 2.2, this implies $\mathcal{M} = \mathrm{conv}(\mathcal{W})$, where $\mathcal{W}$ is the set of all graphical sequences.

The following result characterizes when $\mathbf{d}$ is a degree sequence of a weighted graph with edge weights in S; we also refer to such $\mathbf{d}$ as a *(finite discrete) graphical sequence*. The case $r = 2$ recovers the classical Erdős-Gallai criterion [12].

Theorem 3.2. *A sequence $(d_1, \ldots, d_n) \in \mathbb{N}_0^n$ with $d_1 \geq d_2 \geq \cdots \geq d_n$ is the degree sequence of a graph G with edge weights in the set $S = \{0, 1, \ldots, r-1\}$, if and only if $\sum_{i=1}^n d_i$ is even and*

$$\sum_{i=1}^k d_i \leq (r-1)k(k-1) + \sum_{j=k+1}^n \min\{d_j, (r-1)k\} \quad \text{for } k = 1, \ldots, n. \tag{9}$$

Although the result above provides a precise characterization of the set of graphical sequences $\mathcal{W}$, it is not immediately clear how to characterize the convex hull $\mathrm{conv}(\mathcal{W})$, or how to decide whether a given $\mathbf{d}$ belongs to $\mathcal{M}^\circ = \mathrm{conv}(\mathcal{W})^\circ$. Fortunately, in practice we can circumvent this issue by employing the following algorithm to compute the MLE. The case $r = 2$ recovers the iterative algorithm proposed by Chatterjee et al. [8] in the case of unweighted graphs.

Theorem 3.3. *Given $\mathbf{d} = (d_1, \ldots, d_n) \in \mathbb{R}^n_+$, define the function $\varphi \colon \mathbb{R}^n \to \mathbb{R}^n$ by $\varphi(\mathbf{x}) = (\varphi_1(\mathbf{x}), \ldots, \varphi_n(\mathbf{x}))$, where*

$$\varphi_i(\mathbf{x}) = x_i + \frac{1}{r-1}\left(\log \sum_{j \neq i} \mu(x_i + x_j) - \log d_i\right). \tag{10}$$

Starting from any $\theta^{(0)} \in \mathbb{R}^n$, define

$$\theta^{(k+1)} = \varphi(\theta^{(k)}), \quad k \in \mathbb{N}_0. \tag{11}$$

Suppose $\mathbf{d} \in \mathrm{conv}(\mathcal{W})^\circ$, so the MLE equation (8) has a unique solution $\hat{\theta}$. Then $\hat{\theta}$ is a fixed point of the function φ, and the iterates (11) converge to $\hat{\theta}$ geometrically fast: there exists a constant $\beta \in (0, 1)$ that only depends on $(\|\hat{\theta}\|_\infty, \|\theta^{(0)}\|_\infty)$, such that

$$\|\theta^{(k)} - \hat{\theta}\|_\infty \leq \beta^{k-1}\, \|\theta^{(0)} - \hat{\theta}\|_\infty, \quad k \in \mathbb{N}_0. \tag{12}$$

Conversely, if $\mathbf{d} \notin \mathrm{conv}(\mathcal{W})^\circ$, then the sequence $\{\theta^{(k)}\}$ has a divergent subsequence.

Figure 2 demonstrates the performance of the algorithm presented above. We set $n = 200$ and sample $\theta \in [-1, 1]^n$ uniformly at random. Then for each $2 \leq r \leq 10$, we sample a graph from the distribution $\mathbb{P}^*_\theta$, compute the empirical degree sequence $\mathbf{d}$, and run the iterative algorithm starting with $\theta^{(0)} = 0$ until

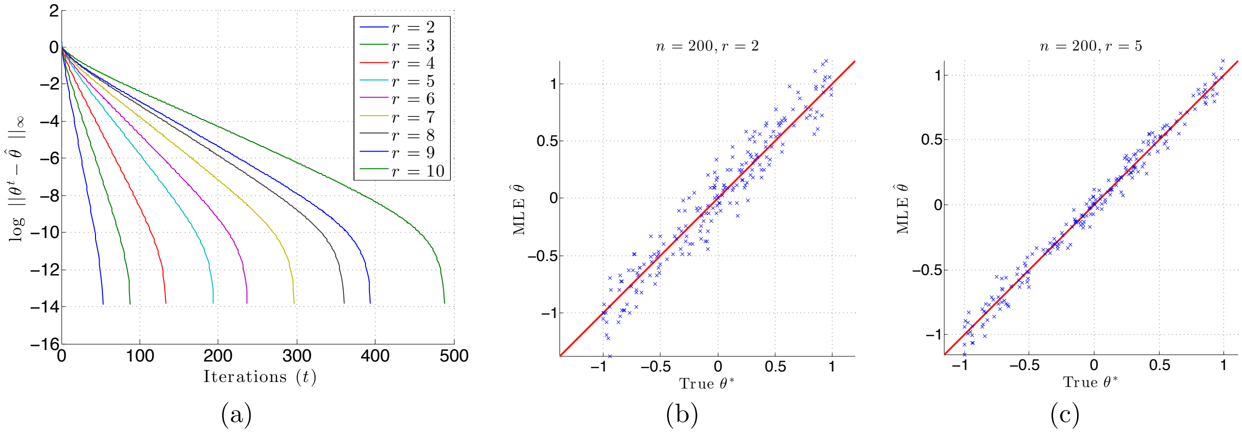

Figure 2: (a) Plot of $\log \|\theta^{(t)} - \hat{\theta}\|_\infty$ for various values of r, where $\hat{\theta}$ is the final value of $\theta^{(t)}$ when the algorithm converges; (b) Scatter plot of the estimate $\hat{\theta}$ vs. the true parameter θ for $r = 2$; (c) Scatter plot for $r = 5$.

convergence. The left panel (Figure 2(a)) shows the rate of convergence (on a logarithmic scale) of the algorithm for various values of r. We observe that the iterates $\{\theta^{(t)}\}$ indeed converge geometrically fast to the MLE $\hat{\theta}$, but the rate of convergence decreases as r increases. By examining the proof of Theorem 3.3 in Section 4.2, we see that the term β has the expression

$$\beta^2 = 1 - \frac{1}{(r-1)^2} \left(\min \left\{ \frac{\exp(2K) - 1}{\exp(2rK) - 1}, \, -\frac{\mu'(2K)}{\mu(-2K)} \right\} \right)^2,$$

where $K = 2\|\hat{\theta}\|_\infty + \|\theta^{(0)}\|_\infty$. This shows that β is an increasing function of r, which explains the empirical decrease in the rate of convergence as r increases.

Figures 2(b) and (c) show the plots of the estimate $\hat{\theta}$ versus the true θ. Notice that the points lie close to the diagonal line, which suggests that the MLE $\hat{\theta}$ is very close to the true parameter θ. Indeed, the following result shows that $\hat{\theta}$ is a consistent estimator of θ. Recall that $\hat{\theta}$ is *consistent* if $\hat{\theta}$ converges in probability to θ as $n \to \infty$.

Theorem 3.4. *Let $M > 0$ and $k > 1$ be fixed. Given $\theta \in \mathbb{R}^n$ with $\|\theta\|_\infty \leq M$, consider the problem of finding the MLE $\hat{\theta}$ of θ based on one graph sample $G \sim \mathbb{P}^*_\theta$. Then for sufficiently large n, with probability at least $1 - 2n^{-(k-1)}$ the MLE $\hat{\theta}$ exists and satisfies*

$$\|\hat{\theta} - \theta\|_\infty \leq C \sqrt{\frac{k \log n}{n}},$$

where C is a constant that only depends on M.

3.2 Continuous weighted graphs

In this section we study weighted graphs with edge weights in $\mathbb{R}_0$. The proofs of the results presented here are provided in Section 4.3.

3.2.1 Characterization of the distribution

We take ν to be the Lebesgue measure on $\mathbb{R}_0$. The marginal log-partition function is

$$Z_1(t) = \log \int_{\mathbb{R}_0} \exp(-ta)\, da = \begin{cases} \log(1/t) & \text{if } t > 0 \\ \infty & \text{if } t \leq 0. \end{cases}$$

Thus $\mathrm{Dom}(Z_1) = \mathbb{R}_+$, and the natural parameter space is

$$\Theta = \{(\theta_1, \ldots, \theta_n) \in \mathbb{R}^n : \theta_i + \theta_j > 0 \text{ for } i \neq j\}.$$

For $\theta \in \Theta$, the edge weights A_{ij} are independent exponential random variables with density

$$p^*_{ij}(a) = (\theta_i + \theta_j) \, \exp\big(-(\theta_i + \theta_j)\, a \big) \quad \text{for } a \in \mathbb{R}_0$$

and mean parameter $\mathbb{E}_\theta[A_{ij}] = 1/(\theta_i + \theta_j)$. The corresponding mean function is given by

$$\mu(t) = \frac{1}{t}, \quad t > 0.$$

3.2.2 Existence, uniqueness, and consistency of the MLE

We now consider the problem of finding the MLE of θ from one graph sample $G \sim \mathbb{P}^*_\theta$. As we saw previously, the MLE $\hat{\theta} \in \Theta$ satisfies the moment-matching equation $\mathbb{E}_{\hat{\theta}}[\deg(A)] = \mathbf{d}$, where $\mathbf{d}$ is the degree sequence of the sample graph G. Equivalently, $\hat{\theta} \in \Theta$ is a solution to the system of equations

$$d_i = \sum_{j \neq i} \frac{1}{\hat{\theta}_i + \hat{\theta}_j}, \quad i = 1, \ldots, n. \tag{13}$$

Remark 3.5. The system (13) is a special case of a general class that Sanyal, Sturmfels, and Vinzant [30] study using algebraic geometry and matroid theory (extending the work of Proudfoot and Speyer [27]). Define

$$\chi(t) = \sum_{k=0}^{n} \left(\left\{ {n \atop k} \right\} + n \left\{ {n-1 \atop k} \right\} \right) (t-1)^{(2)}_k,$$

in which $\left\{ {n \atop k} \right\}$ is the Stirling number of the second kind and $(x)^{(2)}_{k+1} = x(x-2)\cdots(x-2k)$ is a generalized falling factorial. Then, there is a polynomial $H(\mathbf{d})$ in the d_i such that for $\mathbf{d} \in \mathbb{R}^n$ with $H(\mathbf{d}) \neq 0$, the number of solutions $\theta \in \mathbb{R}^n$ to (13) is $(-1)^n \chi(0)$. Moreover, the polynomial $H(\mathbf{d})$ has degree $2(-1)^n(n\chi(0) + \chi'(0))$ and characterizes those $\mathbf{d}$ for which the equations above have multiple roots. We refer to [30] for more details.

Since the natural parameter space Θ is open, Proposition 2.1 tells us that the MLE $\hat{\theta}$ exists and is unique if and only if the empirical degree sequence $\mathbf{d}$ belongs to the interior $\mathcal{M}^\circ$ of the mean parameter space $\mathcal{M}$. We characterize the set of graphical sequences $\mathcal{W}$ and determine its relation to the mean parameter space $\mathcal{M}$.

We say $\mathbf{d} = (d_1, \ldots, d_n)$ is a *(continuous) graphical sequence* if there is a graph G with edge weights in $\mathbb{R}_0$ that realizes $\mathbf{d}$. The finite discrete graphical sequences from Section 3.1 have combinatorial constraints because there are only finitely many possible edge weights between any pair of vertices, and these constraints translate into a set of inequalities in the generalized Erdős-Gallai criterion in Theorem 3.2. In the case of continuous weighted graphs, however, we do not have these constraints because every edge can have as much weight as possible. Therefore, the criterion for a continuous graphical sequence should be simpler than in Theorem 3.2, as the following result shows.

Theorem 3.6. *A sequence* $(d_1, \ldots, d_n) \in \mathbb{R}_0^n$ *is graphic if and only if*

$$\max_{1 \le i \le n} d_i \le \frac{1}{2} \sum_{i=1}^{n} d_i. \tag{14}$$

We note that condition (14) is implied by the case $k = 1$ in the conditions (9). This is to be expected, since any finite discrete weighted graph is also a continuous weighted graph, so finite discrete graphical sequences are also continuous graphical sequences.

Given the criterion in Theorem 3.6, we can write the set $\mathcal{W}$ of graphical sequences as

$$\mathcal{W} = \Big\{(d_1, \dots, d_n) \in \mathbb{R}_0^n\colon\ \max_{1\le i\le n} d_i \le \frac{1}{2}\sum_{i=1}^n d_i\Big\}.$$

Moreover, we can also show that the set of graphical sequences coincide with the mean parameter space.

Lemma 3.7. *The set $\mathcal{W}$ is convex, and $\mathcal{M} = \mathcal{W}$.*

The result above, together with the result of Proposition 2.1, implies that the MLE $\hat{\theta}$ exists and is unique if and only if the empirical degree sequence $\mathbf{d}$ belongs to the interior of the mean parameter space, which can be written explicitly as

$$\mathcal{M}^\circ = \Big\{(d_1', \dots, d_n') \in \mathbb{R}_+^n\colon\ \max_{1\le i\le n} d_i' < \frac{1}{2}\sum_{i=1}^n d_i'\Big\}.$$

Example 3.8. Let $n = 3$ and $\mathbf{d} = (d_1, d_2, d_3) \in \mathbb{R}^n$ with $d_1 \ge d_2 \ge d_3$. It is easy to see that the system of equations (13) gives us

$$\begin{aligned}
\frac{1}{\hat{\theta}_1 + \hat{\theta}_2} &= \frac{1}{2}(d_1 + d_2 - d_3),\\
\frac{1}{\hat{\theta}_1 + \hat{\theta}_3} &= \frac{1}{2}(d_1 - d_2 + d_3),\\
\frac{1}{\hat{\theta}_2 + \hat{\theta}_3} &= \frac{1}{2}(-d_1 + d_2 + d_3),
\end{aligned}$$

from which we obtain a unique solution $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3)$. Recall that $\hat{\theta} \in \Theta$ means $\hat{\theta}_1 + \hat{\theta}_2 > 0$, $\hat{\theta}_1 + \hat{\theta}_3 > 0$, and $\hat{\theta}_2 + \hat{\theta}_3 > 0$, so the equations above tell us that $\hat{\theta} \in \Theta$ if and only if $d_1 < d_2 + d_3$. In particular, this also implies $d_3 > d_1 - d_2 \ge 0$, so $\mathbf{d} \in \mathbb{R}_+^3$. Hence, there is a unique solution $\hat{\theta} \in \Theta$ to the system of equations (13) if and only if $\mathbf{d} \in \mathcal{M}^\circ$, as claimed above.

Finally, we prove that the MLE $\hat{\theta}$ is a consistent estimator of θ.

Theorem 3.9. *Let $M \ge L > 0$ and $k \ge 1$ be fixed. Given $\theta \in \Theta$ with $L \le \theta_i + \theta_j \le M$, $i \ne j$, consider the problem of finding the MLE $\hat{\theta} \in \Theta$ of θ from one graph sample $G \sim \mathbb{P}_\theta^*$. Then for sufficiently large n, with probability at least $1 - 2n^{-(k-1)}$ the MLE $\hat{\theta} \in \Theta$ exists and satisfies*

$$\|\hat{\theta} - \theta\|_\infty \le \frac{100M^2}{L}\sqrt{\frac{k\log n}{\gamma n}},$$

where $\gamma > 0$ is a universal constant.

3.3 Infinite discrete weighted graphs

We now turn our focus to weighted graphs with edge weights in $\mathbb{N}_0$. The proofs of the results presented here can be found in Section 4.4.

3.3.1 Characterization of the distribution

We take ν to be the counting measure on $\mathbb{N}_0$. In this case the marginal log-partition function is given by

$$Z_1(t) = \log\sum_{a=0}^\infty \exp(-at) = \begin{cases} -\log\big(1 - \exp(-t)\big) & \text{if } t > 0,\\ \infty & \text{if } t \le 0.\end{cases}$$

Thus, the domain of Z_1 is $\text{Dom}(Z_1) = (0, \infty)$, and the natural parameter space is

$$\Theta = \{(\theta_1, \ldots, \theta_n) \in \mathbb{R}^n : \theta_i + \theta_j > 0 \text{ for } i \neq j\},$$

which is the same natural parameter space as in the case of continuous weighted graphs in the preceding section. Given $\theta \in \Theta$, the edge weights A_{ij} are independent geometric random variables with probability mass function

$$\mathbb{P}^*(A_{ij} = a) = \big(1 - \exp(-\theta_i - \theta_j)\big)\ \exp\big(-(\theta_i + \theta_j)\, a\big), \quad a \in \mathbb{N}_0.$$

The mean parameters are

$$\mathbb{E}_{\mathbb{P}^*}[A_{ij}] = \frac{\exp(-\theta_i - \theta_j)}{1 - \exp(-\theta_i - \theta_j)} = \frac{1}{\exp(\theta_i + \theta_j) - 1},$$

induced by the mean function

$$\mu(t) = \frac{1}{\exp(t) - 1}, \quad t > 0.$$

3.3.2 Existence, uniqueness, and consistency of the MLE

Consider the problem of finding the MLE of θ from one graph sample $G \sim \mathbb{P}^*_\theta$. Let $\mathbf{d}$ denote the degree sequence of G. Then the MLE $\hat{\theta} \in \Theta$, which satisfies the moment-matching equation $\mathbb{E}_{\hat{\theta}}[\deg(A)] = \mathbf{d}$, is a solution to the system of equations

$$d_i = \sum_{j \neq i} \frac{1}{\exp(\hat{\theta}_i + \hat{\theta}_j) - 1}, \quad i = 1, \ldots, n. \tag{15}$$

We note that the natural parameter space Θ is open, so by Proposition 2.1, the MLE $\hat{\theta}$ exists and is unique if and only if $\mathbf{d} \in \mathcal{M}^\circ$, where $\mathcal{M}$ is the mean parameter space. Since $\nu^{\binom{n}{2}}$ is the counting measure on $\mathbb{N}_0^{\binom{n}{2}}$, the set $\mathfrak{P}$ contains all the Dirac measures, so we know $\mathcal{M} = \text{conv}(\mathcal{W})$ from Proposition 2.2. Here $\mathcal{W}$ is the set of all *(infinite discrete) graphical sequences*, namely, the set of degree sequences of weighted graphs with edge weights in $\mathbb{N}_0$. The following result provides a precise criterion for such graphical sequences. Note that condition (16) below is implied by the limit $r \to \infty$ in Theorem 3.2.

Theorem 3.10. *A sequence $(d_1, \ldots, d_n) \in \mathbb{N}_0^n$ is graphic if and only if $\sum_{i=1}^n d_i$ is even and*

$$\max_{1 \le i \le n} d_i \le \frac{1}{2} \sum_{i=1}^n d_i. \tag{16}$$

The criterion in Theorem 3.10 allows us to write an explicit form for the set of graphical sequences $\mathcal{W}$,

$$\mathcal{W} = \Big\{(d_1, \ldots, d_n) \in \mathbb{N}_0^n : \sum_{i=1}^n d_i \text{ is even and } \max_{1 \le i \le n} d_i \le \frac{1}{2} \sum_{i=1}^n d_i\Big\}.$$

Now we need to characterize $\text{conv}(\mathcal{W})$. Let $\mathcal{W}_1$ denote the set of all continuous graphical sequences from Theorem 3.6, when the edge weights are in $\mathbb{R}_0$,

$$\mathcal{W}_1 = \Big\{(d_1, \ldots, d_n) \in \mathbb{R}_0^n : \max_{1 \le i \le n} d_i \le \frac{1}{2} \sum_{i=1}^n d_i\Big\}.$$

It turns out that when we take the convex hull of $\mathcal{W}$, we essentially recover $\mathcal{W}_1$.

Lemma 3.11. $\overline{\textit{conv}(\mathcal{W})} = \mathcal{W}_1$.

Recalling that a convex set and its closure have the same interior points, the result above gives us

$$\mathcal{M}^\circ = \text{conv}(\mathcal{W})^\circ = \left(\overline{\text{conv}(\mathcal{W})}\right)^\circ = \mathcal{W}_1^\circ = \Big\{(d_1, \dots, d_n) \in \mathbb{R}_+^n : \max_{1 \le i \le n} d_i < \frac{1}{2}\sum_{i=1}^n d_i\Big\}.$$

Example 3.12. Let $n = 3$ and $\mathbf{d} = (d_1, d_2, d_3) \in \mathbb{R}^n$ with $d_1 \ge d_2 \ge d_3$. It can be easily verified that the system of equations (15) gives us

$$\hat{\theta}_1 + \hat{\theta}_2 = \log\left(1 + \frac{2}{d_1 + d_2 - d_3}\right),$$
$$\hat{\theta}_1 + \hat{\theta}_3 = \log\left(1 + \frac{2}{d_1 - d_2 + d_3}\right),$$
$$\hat{\theta}_2 + \hat{\theta}_3 = \log\left(1 + \frac{2}{-d_1 + d_2 + d_3}\right),$$

from which we can obtain a unique solution $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3)$. Recall that $\hat{\theta} \in \Theta$ means $\hat{\theta}_1 + \hat{\theta}_2 > 0$, $\hat{\theta}_1 + \hat{\theta}_3 > 0$, and $\hat{\theta}_2 + \hat{\theta}_3 > 0$, so the equations above tell us that $\hat{\theta} \in \Theta$ if and only if $2/(-d_1 + d_2 + d_3) > 0$, or equivalently, $d_1 < d_2 + d_3$. This also implies $d_3 > d_1 - d_2 \ge 0$, so $\mathbf{d} \in \mathbb{R}_+^3$. Thus, the system of equations (15) has a unique solution $\hat{\theta} \in \Theta$ if and only if $\mathbf{d} \in \mathcal{M}^\circ$, as claimed above.

Finally, we prove that with high probability the MLE $\hat{\theta}$ exists and converges to θ.

Theorem 3.13. *Let $M \ge L > 0$ and $k \ge 1$ be fixed. Given $\theta \in \Theta$ with $L \le \theta_i + \theta_j \le M$, $i \ne j$, consider the problem of finding the MLE $\hat{\theta} \in \Theta$ of θ from one graph sample $G \sim \mathbb{P}_\theta^*$. Then for sufficiently large n, with probability at least $1 - 3n^{-(k-1)}$ the MLE $\hat{\theta} \in \Theta$ exists and satisfies*

$$\|\hat{\theta} - \theta\|_\infty \le \frac{8\ \exp(5M)}{L}\sqrt{\frac{k \log n}{\gamma n}},$$

where $\gamma > 0$ is a universal constant.

4 Proofs of main results

In this section we provide proofs for the technical results presented in Section 3. The proofs of the characterization of weighted graphical sequences (Theorems 3.2, 3.6, and 3.10) are inspired by the constructive proof of the classical Erdős-Gallai criterion by Choudum [9].

4.1 Preliminaries

We begin by presenting several results that we will use in this section. We use the definition of sub-exponential random variables and the concentration inequality presented in [39].

4.1.1 Concentration inequality for sub-exponential random variables

We say that a real-valued random variable X is *sub-exponential* with parameter $\kappa > 0$ if

$$\mathbb{E}[|X|^p]^{1/p} \le \kappa p \quad \text{for all } p \ge 1.$$

Note that if X is a κ-sub-exponential random variable with finite first moment, then the centered random variable $X - \mathbb{E}[X]$ is also sub-exponential with parameter 2κ. This follows from the triangle inequality applied to the p-norm, followed by Jensen's inequality for $p \ge 1$:

$$\mathbb{E}\big[\big|X - \mathbb{E}[X]\big|^p\big]^{1/p} \le \mathbb{E}[|X|^p]^{1/p} + \big|\mathbb{E}[X]\big| \le 2\mathbb{E}[|X|^p]^{1/p}.$$

Sub-exponential random variables satisfy the following concentration inequality.

Theorem 4.1 ([39, Corollary 5.17])**.** *Let $X_1, \dots, X_n$ be independent centered random variables, and suppose each X_i is sub-exponential with parameter κ_i. Let $\kappa = \max_{1 \le i \le n} \kappa_i$. Then for every $\epsilon \ge 0$,*

$$\mathbb{P}\left(\left| \frac{1}{n} \sum_{i=1}^{n} X_i \right| \ge \epsilon \right) \le 2 \exp\left[-\gamma\, n \cdot \min\left(\frac{\epsilon^2}{\kappa^2},\, \frac{\epsilon}{\kappa} \right) \right],$$

where $\gamma > 0$ is an absolute constant.

We will apply the concentration inequality above to exponential and geometric random variables, which are the distributions of the edge weights of continuous weighted graphs (from Section 3.2) and infinite discrete weighted graphs (from Section 3.3).

Lemma 4.2. *Let X be an exponential random variable with $\mathbb{E}[X] = 1/\lambda$. Then X is sub-exponential with parameter $1/\lambda$, and the centered random variable $X - 1/\lambda$ is sub-exponential with parameter $2/\lambda$.*

Proof. For any $p \ge 1$, we can evaluate the moment of X directly:

$$\mathbb{E}[|X|^p] = \int_0^\infty x^p \cdot \lambda \exp(-\lambda x)\, dx = \frac{1}{\lambda^p} \int_0^\infty y^p \exp(-y)\, dy = \frac{\Gamma(p+1)}{\lambda^p},$$

where Γ is the gamma function, and in the computation above we have used the substitution $y = \lambda x$. It can be easily verified that $\Gamma(p+1) \le p^p$ for $p \ge 1$, so

$$\mathbb{E}[|X|^p]^{1/p} = \frac{\big(\Gamma(p+1)\big)^{1/p}}{\lambda} \le \frac{p}{\lambda}.$$

This shows that X is sub-exponential with parameter $1/\lambda$. □

Lemma 4.3. *Let X be a geometric random variable with parameter $q \in (0,1)$, so*

$$\mathbb{P}(X = a) = (1-q)^a\, q, \quad a \in \mathbb{N}_0.$$

Then X is sub-exponential with parameter $-2/\log(1-q)$, and the centered random variable $X - (1-q)/q$ is sub-exponential with parameter $-4/\log(1-q)$.

Proof. Fix $p \ge 1$, and consider the function $f \colon \mathbb{R}_0 \to \mathbb{R}_0$, $f(x) = x^p (1-q)^x$. One can easily verify that f is increasing for $0 \le x \le \lambda$ and decreasing on $x \ge \lambda$, where $\lambda = -p/\log(1-q)$. In particular, for all $x \in \mathbb{R}_0$ we have $f(x) \le f(\lambda)$, and

$$f(\lambda) = \lambda^p (1-q)^\lambda = \left(\frac{p}{-\log(1-q)} \cdot (1-q)^{-1/\log(1-q)} \right)^p = \left(\frac{p}{-e \cdot \log(1-q)} \right)^p.$$

Now note that for $0 \le a \le \lfloor \lambda \rfloor - 1$ we have $f(a) \le \int_a^{a+1} f(x)\, dx$, and for $a \ge \lceil \lambda \rceil + 1$ we have $f(a) \le \int_{a-1}^{a} f(x)\, dx$. Thus, we can bound

$$\begin{aligned} \sum_{a=0}^{\infty} f(a) &= \sum_{a=0}^{\lfloor \lambda \rfloor - 1} f(a) + \sum_{a=\lfloor \lambda \rfloor}^{\lceil \lambda \rceil} f(a) + \sum_{a=\lceil \lambda \rceil + 1}^{\infty} f(a) \\ &\le \int_0^{\lfloor \lambda \rfloor} f(x)\, dx + 2 f(\lambda) + \int_{\lceil \lambda \rceil}^{\infty} f(x)\, dx \\ &\le \int_0^\infty f(x)\, dx + 2 f(\lambda). \end{aligned}$$

Using the substitution $y = -x\log(1-q)$, we can evaluate the integral to be

$$\int_0^\infty f(x)\,dx = \int_0^\infty x^p \exp\left(x \cdot \log(1-q)\right)\,dx = \frac{1}{(-\log(1-q))^{p+1}} \int_0^\infty y^p\,\exp(-y)\,dy$$
$$= \frac{\Gamma(p+1)}{(-\log(1-q))^{p+1}} \le \frac{p^p}{(-\log(1-q))^{p+1}},$$

where in the last step we have again used the relation $\Gamma(p+1) \le p^p$. We use the result above, along with the expression of $f(\lambda)$, to bound the moment of X:

$$\begin{aligned}
\mathbb{E}[|X|^p] &= \sum_{a=0}^{\infty} a^p \cdot (1-q)^a\, q = q \sum_{a=0}^{\infty} f(a) \\
&\le q \int_0^\infty f(x)\,dx + 2q\, f(\lambda) \\
&\le \left(\frac{q^{1/p}\, p}{(-\log(1-q))^{1+1/p}}\right)^p + \left(\frac{(2q)^{1/p}\, p}{-e\cdot\log(1-q)}\right)^p \\
&\le \left(\frac{q^{1/p}\, p}{(-\log(1-q))^{1+1/p}} + \frac{(2q)^{1/p}\, p}{-e\cdot\log(1-q)}\right)^p,
\end{aligned}$$

where in the last step we have used the fact that $x^p + y^p \le (x+y)^p$ for $x, y \ge 0$ and $p \ge 1$. This gives us

$$\begin{aligned}
\frac{1}{p}\,\mathbb{E}[|X|^p]^{1/p} &\le \frac{q^{1/p}}{(-\log(1-q))^{1+1/p}} + \frac{2^{1/p}\, q^{1/p}}{-e\cdot\log(1-q)} \\
&= \frac{1}{-\log(1-q)} \left(\left(\frac{q}{-\log(1-q)}\right)^{1/p} + \frac{2^{1/p}\, q^{1/p}}{e}\right).
\end{aligned}$$

Now note that $q \le -\log(1-q)$ for $0 < q < 1$, so $(-q/\log(1-q))^{1/p} \le 1$. Moreover, $(2q)^{1/p} \le 2^{1/p} \le 2$. Therefore, for any $p \ge 1$, we have

$$\frac{1}{p}\,\mathbb{E}[|X|^p]^{1/p} \le \frac{1}{-\log(1-q)}\left(1 + \frac{2}{e}\right) < \frac{2}{-\log(1-q)}.$$

Thus, we conclude that X is sub-exponential with parameter $-2/\log(1-q)$. □

4.1.2 Bound on the inverses of diagonally-dominant matrices

An $n \times n$ real matrix J is *diagonally dominant* if

$$\Delta_i(J) := |J_{ii}| - \sum_{j \ne i} |J_{ij}| \ge 0, \quad \text{for } i = 1, \ldots, n.$$

We say that J is *diagonally balanced* if $\Delta_i(J) = 0$ for $i = 1, \ldots, n$. We have the following bound from [13] on the inverses of diagonally dominant matrices. This bound is independent of Δ_i, so it is also applicable to diagonally balanced matrices. We will use this result in the proofs of Theorems 3.9 and 3.13.

Theorem 4.4 ([13, Theorem 1.1])**.** *Let $n \ge 3$. For any symmetric diagonally dominant matrix J with $J_{ij} \ge \ell > 0$, we have*

$$\|J^{-1}\|_\infty \le \frac{3n-4}{2\ell(n-2)(n-1)}.$$

4.2 Proofs for the finite discrete weighted graphs

In this section we present the proofs of the results presented in Section 3.1.

4.2.1 Proof of Theorem 3.2

We first prove the necessity of (9). Suppose $\mathbf{d} = (d_1, \dots, d_n)$ is the degree sequence of a graph G with edge weights $a_{ij} \in S$. Then $\sum_{i=1}^n d_i = 2\sum_{(i,j)} a_{ij}$ is even. Moreover, for each $1 \le k \le n$, $\sum_{i=1}^k d_i$ counts the total edge weights coming out from the vertices $1, \dots, k$. The total edge weights from these k vertices to themselves is at most $(r-1)k(k-1)$, and for each vertex $j \notin \{1, \dots, k\}$, the total edge weights from these k vertices to vertex j is at most $\min\{d_j, (r-1)k\}$, so by summing over $j \notin \{1, \dots, k\}$ we get (9).

To prove the sufficiency of (9) we use induction on $s := \sum_{i=1}^n d_i$. The base case $s = 0$ is trivial. Assume the statement holds for $s-2$, and suppose we have a sequence $\mathbf{d}$ with $d_1 \ge d_2 \ge \cdots \ge d_n$ satisfying (9) with $\sum_{i=1}^n d_i = s$. Without loss of generality we may assume $d_n \ge 1$, for otherwise we can proceed with only the nonzero elements of $\mathbf{d}$. Let $1 \le t \le n-1$ be the smallest index such that $d_t > d_{t+1}$, with $t = n-1$ if $d_1 = \cdots = d_n$. Define $\mathbf{d}' = (d_1, \dots, d_{t-1}, d_t - 1, d_{t+1}, \dots, d_{n-1}, d_n - 1)$, so we have $d_1' = \cdots = d_{t-1}' > d_t' \ge d_{t+1}' \ge \cdots \ge d_{n-1}' > d_n'$ and $\sum_{i=1}^n d_i' = s-2$.

We will show that $\mathbf{d}'$ satisfies (9). By the inductive hypothesis, this means $\mathbf{d}'$ is the degree sequence of a graph G' with edge weights $a_{ij}' \in \{0, 1, \dots, r-1\}$. We now attempt to modify G' to obtain a graph G whose degree sequence is equal to $\mathbf{d}$. If the weight a_{tn}' of the edge (t, n) is less than $r-1$, then we can obtain G by increasing a_{tn}' by 1, since the degree of vertex t is now $d_t' + 1 = d_t$, and the degree of vertex n is now $d_n' + 1 = d_n$. Otherwise, suppose $a_{tn}' = r-1$. Since $d_t' = d_1' - 1$, there exists a vertex $u \ne n$ such that $a_{tu}' < r-1$. Since $d_u' > d_n'$, there exists another vertex v such that $a_{uv}' > a_{vn}'$. Then we can obtain the graph G by increasing a_{tu}' and a_{vn}' by 1 and reducing a_{uv}' by 1, so that now the degrees of vertices t and n are each increased by 1, and the degrees of vertices u and v are preserved.

It now remains to show that $\mathbf{d}'$ satisfies (9). We divide the proof into several cases for different values of k. We will repeatedly use the fact that $\mathbf{d}$ satisfies (9), as well as the inequality $\min\{a, b\} - 1 \le \min\{a-1, b\}$.

1. For $k = n$:
$$\sum_{i=1}^n d_i' = \sum_{i=1}^n d_i - 2 \le (r-1)n(n-1) - 2 < (r-1)n(n-1).$$

2. For $t \le k \le n-1$:
$$\begin{aligned}\sum_{i=1}^k d_i' = \sum_{i=1}^k d_i - 1 &\le (r-1)k(k-1) + \sum_{j=k+1}^n \min\{d_j, (r-1)k\} - 1 \\ &\le (r-1)k(k-1) + \sum_{j=k+1}^{n-1} \min\{d_j, (r-1)k\} + \min\{d_n - 1, (r-1)k\} \\ &= (r-1)k(k-1) + \sum_{j=k+1}^n \min\{d_j', (r-1)k\}.\end{aligned}$$

3. For $1 \le k \le t-1$: first suppose $d_n \ge 1 + (r-1)k$. Then for all j we have
$$\min\{d_j', (r-1)k\} = \min\{d_j, (r-1)k\} = (r-1)k,$$
so
$$\begin{aligned}\sum_{i=1}^k d_i' = \sum_{i=1}^k d_i &\le (r-1)k(k-1) + \sum_{j=k+1}^n \min\{d_j, (r-1)k\} \\ &= (r-1)k(k-1) + \sum_{j=k+1}^n \min\{d_j', (r-1)k\}.\end{aligned}$$

4. For $1 \leq k \leq t-1$: suppose $d_1 \geq 1+(r-1)k$, and $d_n \leq (r-1)k$. We claim that $\mathbf{d}$ satisfies (9) at k with a strict inequality. If this claim is true, then, since $d_t = d_1$ and $\min\{d'_t, (r-1)k\} = \min\{d_t, (r-1)k\} = (r-1)k$,

$$\begin{aligned}\sum_{i=1}^{k} d'_i = \sum_{i=1}^{k} d_i &\leq (r-1)k(k-1) + \sum_{j=k+1}^{n} \min\{d_j, (r-1)k\} - 1 \\ &= (r-1)k(k-1) + \sum_{j=k+1}^{n-1} \min\{d'_j, (r-1)k\} + \min\{d_n, (r-1)k\} - 1 \\ &\leq (r-1)k(k-1) + \sum_{j=k+1}^{n-1} \min\{d'_j, (r-1)k\} + \min\{d_n - 1, (r-1)k\} \\ &= (r-1)k(k-1) + \sum_{j=k+1}^{n} \min\{d'_j, (r-1)k\}.\end{aligned}$$

Now to prove the claim, suppose the contrary that $\mathbf{d}$ satisfies (9) at k with equality. Let $t+1 \leq u \leq n$ be the smallest integer such that $d_u \leq (r-1)k$. Then, from our assumption,

$$\begin{aligned}kd_k = \sum_{i=1}^{k} d_i &= (r-1)k(k-1) + \sum_{j=k+1}^{n} \min\{d_j, (r-1)k\} \\ &\geq (r-1)k(k-1) + (u-k-1)(r-1)k + \sum_{j=u}^{n} d_j \\ &= (r-1)k(u-2) + \sum_{j=u}^{n} d_j.\end{aligned}$$

Therefore, since $d_{k+1} = d_k = d_1$,

$$\begin{aligned}\sum_{i=1}^{k+1} d_i = (k+1)d_k &\geq (r-1)(k+1)(u-2) + \frac{k+1}{k}\sum_{j=u}^{n} d_j \\ &> (r-1)(k+1)k + (r-1)(k+1)(u-k-2) + \sum_{j=u}^{n} d_j \\ &\geq (r-1)(k+1)k + \sum_{j=k+2}^{n} \min\{d_j, (r-1)(k+1)\},\end{aligned}$$

which contradicts the fact that $\mathbf{d}$ satisfies (9) at $k+1$. Thus, we have proved that $\mathbf{d}$ satisfies (9) at k with a strict inequality.

5. For $1 \leq k \leq t-1$: suppose $d_1 \leq (r-1)k$. In particular, we have $\min\{d_j, (r-1)k\} = d_j$ and $\min\{d'_j, (r-1)k\} = d'_j$ for all j. First, if we have

$$d_{k+2} + \cdots + d_n \geq 2, \tag{17}$$

then we are done, since

$$\begin{aligned}\sum_{i=1}^{k} d'_i = \sum_{i=1}^{k} d_i &= (k-1)d_1 + d_{k+1} \\ &\le (r-1)k(k-1) + d_{k+1} + d_{k+2} + \cdots + d_n - 2 \\ &= (r-1)k(k-1) + \sum_{j=k+1}^{n} d'_j \\ &= (r-1)k(k-1) + \sum_{j=k+1}^{n} \min\{d'_j, (r-1)k\}.\end{aligned}$$

Condition (17) is obvious if $d_n \ge 2$ or $k+2 \le n-1$ (since there are $n-k-1$ terms in the summation and each term is at least 1). Otherwise, assume $k+2 \ge n$ and $d_n = 1$, so in particular, we have $k = n-2$ (since $k \le t-1 \le n-2$), $t = n-1$, and $d_1 \le (r-1)(n-2)$. Note that we cannot have $d_1 = (r-1)(n-2)$, for then $\sum_{i=1}^{n} d_i = (n-1)d_1 + d_n = (r-1)(n-1)(n-2)+1$ would be odd, so we must have $d_1 < (r-1)(n-2)$. Similarly, n must be even, for otherwise $\sum_{i=1}^{n} d_i = (n-1)d_1 + 1$ would be odd. Thus, since $1 \le d_1 < (r-1)(n-2)$ we must have $n \ge 4$. Therefore,

$$\begin{aligned}\sum_{i=1}^{k} d'_i = (n-2)d_1 &= (n-3)d_1 + d_{n-1} \\ &\le (r-1)(n-2)(n-3) - (n-3) + d_{n-1} \\ &\le (r-1)(n-2)(n-3) + (d_{n-1}-1) + (d_n - 1) \\ &= (r-1)(n-2)(n-3) + \sum_{j=k+1}^{n} \min\{d'_j, (r-1)k\}.\end{aligned}$$

This shows that $\mathbf{d}'$ satisfies (9) and finishes the proof of Theorem 3.2.

4.2.2 Proof of Theorem 3.3

We follow the outline of the proof of [8, Theorem 1.5]. We first present the following properties of the mean function $\mu(t)$ and the Jacobian matrix of the function φ (10). We then combine these results at the end of this section into a proof of Theorem 3.3.

Lemma 4.5. *The mean function $\mu(t)$ is positive and strictly decreasing, with $\mu(-t)+\mu(t) = r-1$ for all $t \in \mathbb{R}$, and $\mu(t) \to 0$ as $t \to \infty$. Its derivative $\mu'(t)$ is increasing for $t \ge 0$, with the properties that $\mu'(t) < 0$, $\mu'(t) = \mu'(-t)$ for all $t \in \mathbb{R}$, and $\mu'(0) = -(r^2-1)/12$.*

Proof. It is clear from (6) that $\mu(t)$ is positive. From the alternative representation (7) it is easy to see that $\mu(-t)+\mu(t) = r-1$, and $\mu(t) \to 0$ as $t \to \infty$. Differentiating expression (6) yields the formula

$$\mu'(t) = \frac{-(\sum_{a=0}^{r-1} a^2\, \exp(-at))(\sum_{a=0}^{r-1} \exp(-at)) + (\sum_{a=0}^{r-1} a\, \exp(-at))^2}{(\sum_{a=0}^{r-1} \exp(-at))^2},$$

and substituting $t = 0$ gives us $\mu'(0) = -(r^2-1)/12$. The Cauchy-Schwarz inequality applied to the expression above tells us that $\mu'(t) < 0$, where the inequality is strict because the vectors $(a^2 \exp(-at))_{a=0}^{r-1}$ and $(\exp(-at))_{a=0}^{r-1}$ are not linearly dependent. Thus, $\mu(t)$ is strictly decreasing for all $t \in \mathbb{R}$.

The relation $\mu(-t)+\mu(t) = r-1$ gives us $\mu'(-t) = \mu'(t)$. Furthermore, by differentiating the expression (7) twice, one can verify that $\mu''(t) \ge 0$ for $t \ge 0$, which means $\mu'(t)$ is increasing for $t \ge 0$. See also Figure 1 for the behavior of $\mu(t)$ and $\mu'(t)$ for different values of r. □

Lemma 4.6. *For all $t \in \mathbb{R}$, we have*

$$\frac{\mu'(t)}{\mu(t)} \geq -r+1+\frac{1}{\sum_{a=0}^{r-1}\exp(at)} > -r+1.$$

Proof. Multiplying the numerator and denominator of (6) by $\exp((r-1)t)$, we can write

$$\mu(t) = \frac{\sum_{a=0}^{r-1} a\ \exp((r-1-a)t)}{\sum_{a=0}^{r-1}\exp((r-1-a)t)} = \frac{\sum_{a=0}^{r-1}(r-1-a)\exp(at)}{\sum_{a=0}^{r-1}\exp(at)}.$$

Therefore,

$$\begin{aligned}
\frac{\mu'(t)}{\mu(t)} &= \frac{d}{dt}\log\mu(t) = \frac{d}{dt}\left(\log\sum_{a=0}^{r-1}(r-1-a)\exp(at) - \log\sum_{a=0}^{r-1}\exp(at)\right) \\
&= \frac{\sum_{a=0}^{r-1} a(r-1-a)\exp(at)}{\sum_{a=0}^{r-1}(r-1-a)\exp(at)} - \frac{\sum_{a=0}^{r-1} a\exp(at)}{\sum_{a=0}^{r-1}\exp(at)} \\
&\geq -\frac{\sum_{a=0}^{r-1} a\exp(at)}{\sum_{a=0}^{r-1}\exp(at)} \\
&= -\frac{\sum_{a=0}^{r-1}(r-1)\exp(at) - \sum_{a=0}^{r-1}(r-1-a)\exp(at)}{\sum_{a=0}^{r-1}\exp(at)} \\
&= -r+1+\frac{\sum_{a=0}^{r-1}(r-1-a)\exp(at)}{\sum_{a=0}^{r-1}\exp(at)} \\
&\geq -r+1+\frac{1}{\sum_{a=0}^{r-1}\exp(at)}.
\end{aligned}$$

□

We recall the following definition and result from [8]. Given $\delta > 0$, let $\mathcal{L}_n(\delta)$ denote the set of $n \times n$ matrices $A = (a_{ij})$ with $\|A\|_\infty \leq 1$, $a_{ii} \geq \delta$, and $a_{ij} \leq -\delta/(n-1)$, for each $1 \leq i \neq j \leq n$.

Lemma 4.7 ([8, Lemma 2.1])**.** *If $A, B \in \mathcal{L}_n(\delta)$, then*

$$\|AB\|_\infty \leq 1 - \frac{2(n-2)\delta^2}{(n-1)}.$$

In particular, for $n \geq 3$,

$$\|AB\|_\infty \leq 1-\delta^2.$$

Given $\theta, \theta' \in \mathbb{R}^n$, let $J(\theta,\theta')$ denote the $n \times n$ matrix whose (i,j)-entry is

$$J_{ij}(\theta,\theta') = \int_0^1 \frac{\partial\varphi_i}{\partial\theta_j}(t\theta+(1-t)\theta')\,dt. \tag{18}$$

Lemma 4.8. *For all $\theta, \theta' \in \mathbb{R}^n$, we have $\|J(\theta,\theta')\|_\infty = 1$.*

Proof. The partial derivatives of φ (10) are

$$\frac{\partial\varphi_i(\mathbf{x})}{\partial x_i} = 1 + \frac{1}{(r-1)}\frac{\sum_{j\neq i}\mu'(x_i+x_j)}{\sum_{j\neq i}\mu(x_i+x_j)}, \tag{19}$$

and for $i \neq j$,

$$\frac{\partial\varphi_i(\mathbf{x})}{\partial x_j} = \frac{1}{(r-1)}\frac{\mu'(x_i+x_j)}{\sum_{k\neq i}\mu(x_i+x_k)} < 0, \tag{20}$$

where the last inequality follows from $\mu'(x_i + x_j) < 0$. Using the result of Lemma 4.6 and the fact that μ is positive, we also see that

$$\frac{\partial \varphi_i(\mathbf{x})}{\partial x_i} = 1 + \frac{1}{(r-1)} \frac{\sum_{j \neq i} \mu'(x_i + x_j)}{\sum_{j \neq i} \mu(x_i + x_j)} > 1 + \frac{1}{(r-1)} \frac{\sum_{j \neq i} (-r+1)\mu(x_i + x_j)}{\sum_{j \neq i} \mu(x_i + x_j)} = 0.$$

Setting $\mathbf{x} = t\theta + (1-t)\theta'$ and integrating over $0 \le t \le 1$, we also get that $J_{ij}(\theta, \theta') < 0$ for $i \neq j$, and $J_{ii}(\theta, \theta') = 1 + \sum_{j \neq i} J_{ij}(\theta, \theta') > 0$. This implies $\|J(\theta, \theta')\|_\infty = 1$, as desired. □

Lemma 4.9. *Let $\theta, \theta' \in \mathbb{R}^n$ with $\|\theta\|_\infty \le K$ and $\|\theta'\|_\infty \le K$ for some $K > 0$. Then $J(\theta, \theta') \in \mathcal{L}_n(\delta)$, where*

$$\delta = \frac{1}{(r-1)} \min \left\{ \frac{\exp(2K) - 1}{\exp(2rK) - 1}, \; -\frac{\mu'(2K)}{\mu(-2K)} \right\}. \tag{21}$$

Proof. From Lemma 4.8 we already know that $J \equiv J(\theta, \theta')$ satisfies $\|J\|_\infty = 1$, so to show that $J \in \mathcal{L}_n(\delta)$ it remains to show that $J_{ii} \ge \delta$ and $J_{ij} \le -\delta/(n-1)$ for $i \neq j$. In particular, it suffices to show that for each $0 \le t \le 1$ we have $\partial \varphi_i(\mathbf{x})/\partial x_i \ge \delta$ and $\partial \varphi_i(\mathbf{x})/\partial x_j \le -\delta/(n-1)$, where $\mathbf{x} \equiv \mathbf{x}(t) = t\theta + (1-t)\theta'$.

Fix $0 \le t \le 1$. Since $\|\theta\|_\infty \le K$ and $\|\theta'\|_\infty \le K$, we also know that $\|\mathbf{x}\|_\infty \le K$, so $-2K \le x_i + x_j \le 2K$ for all $1 \le i, j \le n$. Using the properties of μ and μ' from Lemma 4.5, we have

$$0 < \mu(2K) \le \mu(x_i + x_j) \le \mu(-2K)$$

and

$$\mu'(0) \le \mu'(x_i + x_j) \le \mu'(2K) < 0.$$

Then from (20) and using the definition of δ,

$$\frac{\partial \varphi_i(\mathbf{x})}{\partial x_j} \le \frac{\mu'(2K)}{(n-1)(r-1)\mu(-2K)} \le -\frac{\delta}{n-1}.$$

Furthermore, by Lemma 4.6 we have

$$\frac{\mu'(x_i + x_j)}{\mu(x_i + x_j)} \ge -r + 1 + \frac{\exp(x_i + x_j) - 1}{\exp(r(x_i + x_j)) - 1} \ge -r + 1 + \frac{\exp(2K) - 1}{\exp(2rK) - 1}.$$

So from (19), we also get

$$\frac{\partial \varphi_i(\mathbf{x})}{\partial x_i} \ge 1 + \frac{1}{(r-1)} \frac{\sum_{j \neq i} (-r + 1 + \frac{\exp(2K)-1}{\exp(2rK)-1}) \mu(x_i + x_j)}{\sum_{j \neq i} \mu(x_i + x_j)} = \frac{1}{(r-1)} \left(\frac{\exp(2K) - 1}{\exp(2rK) - 1} \right) \ge \delta,$$

as required. □

We are now ready to prove Theorem 3.3.

Proof of Theorem 3.3: By the mean-value theorem for vector-valued functions [19, p. 341], for any $\theta, \theta' \in \mathbb{R}^n$ we can write

$$\varphi(\theta) - \varphi(\theta') = J(\theta, \theta')(\theta - \theta'),$$

where $J(\theta, \theta')$ is the Jacobian matrix defined in (18). Since $\|J(\theta, \theta')\|_\infty = 1$ (Lemma 4.8), this gives us

$$\|\varphi(\theta) - \varphi(\theta')\|_\infty \le \|\theta - \theta'\|_\infty. \tag{22}$$

First suppose there is a solution $\hat{\theta}$ to the system of equations (8), so $\hat{\theta}$ is a fixed point of φ. Then by setting $\theta = \theta^{(k)}$ and $\theta' = \hat{\theta}$ to the inequality above, we obtain

$$\|\theta^{(k+1)} - \hat{\theta}\|_\infty \le \|\theta^{(k)} - \hat{\theta}\|_\infty. \tag{23}$$

In particular, this shows that $\|\theta^{(k)}\|_\infty \le K$ for all $k \in \mathbb{N}_0$, where $K := 2\|\hat{\theta}\|_\infty + \|\theta^{(0)}\|_\infty$. By Lemma 4.9, this implies $J(\theta^{(k)}, \hat{\theta}) \in \mathcal{L}_n(\delta)$ for all $k \in \mathbb{N}_0$, where δ is given by (21). Another application of the mean-value theorem gives us

$$\theta^{(k+2)} - \hat{\theta} = J(\theta^{(k+1)}, \hat{\theta})\, J(\theta^{(k)}, \hat{\theta})\, (\theta^{(k)} - \hat{\theta}),$$

so by Lemma 4.7,

$$\|\theta^{(k+2)} - \hat{\theta}\|_\infty \le \|J(\theta^{(k+1)}, \hat{\theta})\, J(\theta^{(k)}, \hat{\theta})\|_\infty\, \|\theta^{(k)} - \hat{\theta}\|_\infty \le \left(1 - \delta^2\right) \|\theta^{(k)} - \hat{\theta}\|_\infty.$$

Unrolling the recursive bound above and using (23) gives us

$$\|\theta^{(k)} - \hat{\theta}\|_\infty \le (1 - \delta^2)^{\lfloor k/2 \rfloor} \|\theta^{(0)} - \hat{\theta}\|_\infty \le (1 - \delta^2)^{(k-1)/2} \|\theta^{(0)} - \hat{\theta}\|_\infty,$$

which proves (12) with $\tau = \sqrt{1 - \delta^2}$.

Now suppose the system of equations (8) does not have a solution, and suppose the contrary that the sequence $\{\theta^{(k)}\}$ does not have a divergent subsequence. This means $\{\theta^{(k)}\}$ is a bounded sequence, so there exists $K > 0$ such that $\|\theta^{(k)}\|_\infty \le K$ for all $k \in \mathbb{N}_0$. Then by Lemma 4.9, $J(\theta^{(k)}, \theta^{(k+1)}) \in \mathcal{L}_n(\delta)$ for all $k \in \mathbb{N}_0$, where δ is given by (21). In particular, by the mean value theorem and Lemma 4.8, we get for all $k \in \mathbb{N}_0$,

$$\|\theta^{(k+3)} - \theta^{(k+2)}\|_\infty \le (1 - \delta^2) \|\theta^{(k+1)} - \theta^{(k)}\|_\infty.$$

This implies $\sum_{k=0}^{\infty} \|\theta^{(k+1)} - \theta^{(k)}\|_\infty < \infty$, which means $\{\theta^{(k)}\}$ is a Cauchy sequence. Thus, the sequence $\{\theta^{(k)}\}$ converges to a limit, say $\hat{\theta}$, as $k \to \infty$. This limit $\hat{\theta}$ is necessarily a fixed point of φ, as well as a solution to the system of equations (8), contradicting our assumption. Hence we conclude that $\{\theta^{(k)}\}$ must have a divergent subsequence. □

A little computation based on the proof above gives us the following result, which will be useful in the proof of Theorem 3.4.

Proposition 4.10. *Assume the same setting as in Theorem 3.3, and assume the MLE equation 8 has a unique solution $\hat{\theta}$. Then*

$$\|\theta^{(0)} - \hat{\theta}\|_\infty \le \frac{2}{\delta^2} \|\theta^{(0)} - \theta^{(1)}\|_\infty,$$

where δ is given by (21) *with $K = 2\|\hat{\theta}\|_\infty + \|\theta^{(0)}\|_\infty$.*

Proof. With the same notation as in the proof of Theorem 3.3, by applying the mean-value theorem twice and using the bound in Lemma 4.8, for each $k \ge 0$ we have

$$\|\theta^{(k+3)} - \theta^{(k+2)}\|_\infty \le (1 - \delta^2) \|\theta^{(k+1)} - \theta^{(k)}\|_\infty.$$

Therefore, since $\{\theta^{(k)}\}$ converges to $\hat{\theta}$,

$$\|\theta^{(0)} - \hat{\theta}\|_\infty \le \sum_{k=0}^{\infty} \|\theta^{(k)} - \theta^{(k+1)}\|_\infty \le \frac{1}{\delta^2} \left(\|\theta^{(0)} - \theta^{(1)}\|_\infty + \|\theta^{(1)} - \theta^{(2)}\|_\infty \right) \le \frac{2}{\delta^2} \|\theta^{(0)} - \theta^{(1)}\|_\infty,$$

where the last inequality follows from (22). □

4.2.3 Proof of Theorem 3.4

Our proof of Theorem 3.4 follows the outline of the proof of Theorem 1.3 in [8]. Recall that $\mathcal{W}$ is the set of graphical sequences, and the MLE equation (8) has a unique solution $\hat{\theta} \in \mathbb{R}^n$ if and only if $\mathbf{d} \in \text{conv}(\mathcal{W})^\circ$. We first present a few preliminary results. We will also use the properties of the mean function μ as described in Lemma 4.5.

The following property is based on [8, Lemma 4.1].

Lemma 4.11. *Let* $\mathbf{d} \in conv(\mathcal{W})$ *with the properties that*

$$c_2(r-1)(n-1) \le d_i \le c_1(r-1)(n-1), \quad i = 1, \dots, n, \tag{24}$$

and

$$\min_{\substack{B \subseteq \{1,\dots,n\}, \\ |B| \ge c_2^2(n-1)}} \left\{ \sum_{j \notin B} \min\{d_j, (r-1)|B|\} + (r-1)|B|(|B|-1) - \sum_{i \in B} d_i \right\} \ge c_3 n^2, \tag{25}$$

where $c_1, c_2 \in (0,1)$ *and* $c_3 > 0$ *are constants. Then the MLE equation* (8) *has a solution* $\hat{\theta}$ *with the property that* $\|\hat{\theta}\|_\infty \le C$*, where* $C \equiv C(c_1, c_2, c_3)$ *is a constant that only depends on* c_1, c_2, c_3*.*

Proof. First assume $\hat{\theta}$ exists, so $\hat{\theta}$ and $\mathbf{d}$ satisfy

$$d_i = \sum_{j \ne i} \mu(\hat{\theta}_i + \hat{\theta}_j), \quad i = 1, \dots, n.$$

Let

$$d_{\max} = \max_{1 \le i \le n} d_i, \quad d_{\min} = \min_{1 \le i \le n} d_i, \quad \hat{\theta}_{\max} = \max_{1 \le i \le n} \hat{\theta}_i, \quad \hat{\theta}_{\min} = \min_{1 \le i \le n} \hat{\theta}_i,$$

and let $i^*, j^* \in \{1, \dots, n\}$ be such that $\hat{\theta}_{i^*} = \hat{\theta}_{\max}$ and $\hat{\theta}_{j^*} = \hat{\theta}_{\min}$.

We begin by observing that since μ is a decreasing function and we have the assumption (24),

$$c_2(r-1) \le \frac{d_{\min}}{n-1} \le \frac{d_{i^*}}{n-1} = \frac{1}{(n-1)} \sum_{j \ne i^*} \mu(\hat{\theta}_{\max} + \hat{\theta}_j) \le \mu(\hat{\theta}_{\max} + \hat{\theta}_{\min}),$$

so

$$\hat{\theta}_{\max} + \hat{\theta}_{\min} \le \mu^{-1}\big(c_2(r-1)\big).$$

Thus, if we have a lower bound on $\hat{\theta}_{\min}$ by a constant, then we also get a constant upper bound on $\hat{\theta}_{\max}$ and we are done.

We now proceed to prove the lower bound $\hat{\theta}_{\min} \ge -C$. If $\theta_{\min} \ge 0$, then there is nothing to prove, so let us assume that $\hat{\theta}_{\min} < 0$. We claim the following property.

Claim. *If* $\hat{\theta}_{\min}$ *satisfies* $\mu(\hat{\theta}_{\min}/2) \ge c_1(r-1)$ *and* $\mu(\hat{\theta}_{\min}/4) \ge (r-1)/(1+c_2)$*, then the set* $A = \{i \colon \hat{\theta}_i \le \hat{\theta}_{\min}/4\}$ *has* $|A| \ge c_2^2(n-1)$*.*

Proof of claim: Let $S = \{i \colon \hat{\theta}_i < -\hat{\theta}_{\min}/2\}$ and $m = |S|$. Note that $j^* \in S$ since $\hat{\theta}_{j^*} = \hat{\theta}_{\min} < 0$, so $|m| \ge 1$. Then using the property that μ is a decreasing function and the assumption on $\mu(\hat{\theta}_{\min}/2)$, we obtain

$$\begin{aligned} c_1(r-1)(n-1) \ge d_{\max} \ge d_{j^*} &= \sum_{i \ne j^*} \mu(\hat{\theta}_{\min} + \hat{\theta}_i) \\ &\ge \sum_{i \in S \setminus \{j^*\}} \mu(\hat{\theta}_{\min} + \hat{\theta}_i) > (m-1)\, \mu\left(\frac{\hat{\theta}_{\min}}{2}\right) \ge c_1(r-1)(m-1). \end{aligned}$$

This implies $m < n$, which means there exists $i \notin S$, so $\hat{\theta}_i \ge -\hat{\theta}_{\min}/2 > 0$.

Let $S_i = \{j \colon j \ne i, \hat{\theta}_j > -\hat{\theta}_i/2\}$, and let $m_i = |S_i|$. Then, using the properties that μ is decreasing and bounded above by $r-1$, and using the assumption on $\mu(\hat{\theta}_{\min}/4)$, we get

$$\begin{aligned} c_2(r-1)(n-1) \le d_{\min} \le d_i &= \sum_{j \in S_i} \mu(\hat{\theta}_i + \hat{\theta}_j) + \sum_{j \notin S_i, j \neq i} \mu(\hat{\theta}_i + \hat{\theta}_j) \\ &< m_i \, \mu\left(\frac{\hat{\theta}_i}{2}\right) + (n-1-m_i)(r-1) \\ &= (n-1)(r-1) - m_i \left(r - 1 - \mu\left(\frac{\hat{\theta}_i}{2}\right)\right) \\ &= (n-1)(r-1) - m_i \, \mu\left(-\frac{\hat{\theta}_i}{2}\right) \\ &\le (n-1)(r-1) - m_i \, \mu\left(\frac{\hat{\theta}_{\min}}{4}\right) \\ &\le (n-1)(r-1) - \frac{m_i(r-1)}{1+c_2}. \end{aligned}$$

Rearranging the last inequality above gives us $m_i \le (1-c_2^2)(n-1)$.

Note that for every $j \neq S_i$, $j \neq i$, we have $\hat{\theta}_j \le -\hat{\theta}_i/2 \le \hat{\theta}_{\min}/4$. Therefore, if $A = \{j : \hat{\theta}_j \le \hat{\theta}_{\min}/4\}$, then we see that $S_i^c \setminus \{i\} \subseteq A$, so

$$|A| \ge |S_i^c \setminus \{i\}| = n - m_i - 1 \ge c_2^2(n-1),$$

as desired. □

Now assume

$$\hat{\theta}_{\min} \le \min\left\{2\mu^{-1}(c_1(r-1)),\ 4\mu^{-1}\left(\frac{r-1}{1+c_2}\right),\ -16\right\},$$

for otherwise we are done. Then $\mu(\hat{\theta}_{\min}/2) \ge c_1(r-1)$ and $\mu(\hat{\theta}_{\min}/4) \ge (r-1)/(1+c_2)$, so by the claim above, the size of the set $A = \{i : \hat{\theta}_i \le \hat{\theta}_{\min}/4\}$ is at least $c_2^2(n-1)$. Let

$$h = \sqrt{-\hat{\theta}_{\min}} > 0,$$

and for integers $0 \le k \le \lceil h/16 \rceil - 1$, define the set

$$D_k = \left\{i :\ -\frac{1}{8}\hat{\theta}_{\min} + kh \le \hat{\theta}_i < -\frac{1}{8}\hat{\theta}_{\min} + (k+1)h\right\}.$$

Since the sets $\{D_k\}$ are disjoint, by the pigeonhole principle we can find an index $0 \le k^* \le \lceil h/16 \rceil - 1$ such that

$$|D_{k^*}| \le \frac{n}{\lceil h/16 \rceil} \le \frac{16n}{h}.$$

Fix k^*, and consider the set

$$B = \left\{i : \hat{\theta}_i \le \frac{1}{8}\hat{\theta}_{\min} - \left(k^* + \frac{1}{2}\right)h\right\}.$$

Note that $\hat{\theta}_{\min}/4 \le \hat{\theta}_{\min}/8 - (k^* + 1/2)h$, which implies $A \subseteq B$, so $|B| \ge |A| \ge c_2^2(n-1)$. For $1 \le i \le n$, define

$$d_i^B = \sum_{j \in B \setminus \{i\}} \mu(\hat{\theta}_i + \hat{\theta}_j),$$

and observe that

$$\sum_{j\notin B} d_j^B = \sum_{j\notin B}\sum_{i\in B} \mu(\hat{\theta}_i + \hat{\theta}_j) = \sum_{i\in B}(d_i - d_i^B). \tag{26}$$

We note that for $i \in B$ we have $\hat{\theta}_i \le \hat{\theta}_{\min}/8$, so

$$\begin{aligned}(r-1)|B|(|B|-1) - \sum_{i\in B} d_i^B &= \sum_{i\in B}\sum_{j\in B\setminus\{i\}} \Big(r-1-\mu(\hat{\theta}_i+\hat{\theta}_j)\Big) \\ &\le |B|(|B|-1)\left(r-1-\mu\left(\frac{\hat{\theta}_{\min}}{4}\right)\right) \qquad (27) \\ &= |B|(|B|-1)\,\mu\left(-\frac{\hat{\theta}_{\min}}{4}\right) \le n^2\mu\left(\frac{h^2}{4}\right),\end{aligned}$$

where in the last inequality we have used the definition $h^2 = -\hat{\theta}_{\min} > 0$. Now take $j \notin B$, so $\hat{\theta}_j > \hat{\theta}_{\min}/8 - (k^* + 1/2)h$. We consider three cases:

1. If $\hat{\theta}_j \ge -\hat{\theta}_{\min}/8 + (k^*+1)h$, then for every $i \notin B$, we have $\hat{\theta}_i + \hat{\theta}_j \ge h/2$, so

$$\min\{d_j,\ (r-1)|B|\} - d_j^B \le d_j - d_j^B = \sum_{i\notin B, i\neq j} \mu(\hat{\theta}_j + \hat{\theta}_i) \le n\mu\left(\frac{h}{2}\right).$$

2. If $\hat{\theta}_j \le -\hat{\theta}_{\min}/8 + k^*h$, then for every $i \in B$, we have $\hat{\theta}_i + \hat{\theta}_j \le -h/2$, so

$$\begin{aligned}\min\{d_j,\ (r-1)|B|\} - d_j^B &\le (r-1)|B| - \sum_{i\in B}\mu(\hat{\theta}_j+\hat{\theta}_i) \\ &\le (r-1)|B| - |B|\,\mu\left(-\frac{h}{2}\right) = |B|\,\mu\left(\frac{h}{2}\right) \le n\mu\left(\frac{h}{2}\right).\end{aligned}$$

3. If $-\hat{\theta}_{\min}/8 + k^*h \le \hat{\theta}_j \le -\hat{\theta}_{\min}/8 + (k^*+1)h$, then $j \in D_{k^*}$, and in this case

$$\min\{d_j,\ (r-1)|B|\} - d_j^B \le (r-1)|B| \le n(r-1).$$

There are at most n such indices j in both the first and second cases above, and there are at most $|D_k^*| \le 16n/h$ such indices j in the third case. Therefore,

$$\sum_{j\notin B}\big(\min\{d_j,\ (r-1)|B|\} - d_j^B\big) \le n^2\mu\left(\frac{h}{2}\right) + \frac{16n^2(r-1)}{h}.$$

Combining this bound with (27) and using (26) give us

$$\sum_{j\notin B}\min\{d_j, (r-1)|B|\} + (r-1)|B|(|B|-1) - \sum_{i\in B} d_i \le n^2\mu\left(\frac{h}{2}\right) + \frac{16n^2(r-1)}{h} + n^2\mu\left(\frac{h^2}{4}\right).$$

Assumption (25) tells us that the left hand side of the inequality above is bounded below by c_3n^2, so we obtain

$$\mu\left(\frac{h}{2}\right) + \frac{16(r-1)}{h} + \mu\left(\frac{h^2}{4}\right) \ge c_3.$$

The left hand side is a decreasing function of $h > 0$, so the bound above tells us that $h \le C(c_3)$ for a constant $C(c_3)$ that only depends on c_3 (and r), and so $\hat{\theta}_{\min} = -h^2 \ge -C(c_3)^2$, as desired.

Showing existence of $\hat{\theta}$. Now let $\mathbf{d} \in \text{conv}(\mathcal{W})$ satisfy (24) and (25). Let $\{\mathbf{d}^{(k)}\}_{k\geq 0}$ be a sequence of points in $\text{conv}(\mathcal{W})^\circ$ converging to $\mathbf{d}$, so by Proposition 2.1, for each $k \geq 0$ there exists a solution $\hat{\theta}^{(k)} \in \mathbb{R}^n$ to the MLE equation (8) with $\mathbf{d}^{(k)}$ in place of $\mathbf{d}$. Since $\mathbf{d}$ satisfy (24), (25), and $\mathbf{d}^{(k)} \to \mathbf{d}$, for all sufficiently large k, $\mathbf{d}^{(k)}$ also satisfy (24) and (25) with some constants c_1', c_2', c_3' depending on c_1, c_2, c_3. The preceding analysis then shows that $\|\hat{\theta}^{(k)}\|_\infty \leq C$ for all sufficiently large k, where $C \equiv C(c_1', c_2', c_3') = C(c_1, c_2, c_3)$ is a constant depending on c_1, c_2, c_3. This means $\{\hat{\theta}^{(k)}\}_{k\geq 0}$ is a bounded sequence, so it contains a convergent subsequence $\{\hat{\theta}^{(k_i)}\}_{k_i\geq 0}$, say $\hat{\theta}^{(k_i)} \to \hat{\theta}$. Then $\|\hat{\theta}\|_\infty \leq C$, and since $\hat{\theta}^{(k_i)}$ is a solution to the MLE equation (8) for $\mathbf{d}^{(k_i)}$, $\hat{\theta}$ is necessarily a solution to (8) for $\mathbf{d}$, and we are done. □

We are now ready to prove Theorem 3.4.

Proof of Theorem 3.4: Let $\mathbf{d}^* = (d_1^*, \dots, d_n^*)$ denote the expected degree sequence under $\mathbb{P}_\theta^*$, so $d_i^* = \sum_{j\neq i} \mu(\theta_i + \theta_j)$. Since $-2M \leq \theta_i + \theta_j \leq 2M$ and μ is a decreasing function, we see that

$$(n-1)\,\mu(2M) \leq d_i^* \leq (n-1)\,\mu(-2M), \quad i = 1, \dots, n. \tag{28}$$

For $B \subseteq \{1, \dots, n\}$, let

$$g(\mathbf{d}^*, B) = \sum_{j\notin B} \min\{d_j^*,\ (r-1)|B|\} + (r-1)|B|(|B|-1) - \sum_{i\in B} d_i^*,$$

and similarly for $g(\mathbf{d}, B)$. Using the notation $(d_j^*)^B$ as introduced in the proof of Lemma 4.11, we notice that for $j \notin B$,

$$d_j^* = \sum_{i\neq j} \mu(\theta_j + \theta_i) \geq \sum_{i\in B} \mu(\theta_j + \theta_i) = (d_j^*)^B,$$

and similarly,

$$(r-1)|B| \geq \sum_{i\in B} \mu(\theta_j + \theta_i) = (d_j^*)^B.$$

Therefore, using the relation (26), we see that

$$\begin{aligned}
g(\mathbf{d}^*, B) &\geq \sum_{j\notin B} (d_j^*)^B + (r-1)|B|(|B|-1) - \sum_{i\in B} d_i^* \\
&= (r-1)|B|(|B|-1) - \sum_{i\in B} (d_i^*)^B \\
&= \sum_{i\in B} \sum_{j\in B\setminus\{i\}} \big(r - 1 - \mu(\theta_i + \theta_j)\big) \\
&\geq |B|(|B|-1)\,\big(r - 1 - \mu(-2M)\big) \\
&= |B|(|B|-1)\,\mu(2M).
\end{aligned}$$

We now recall that the edge weights (A_{ij}) are independent random variables taking values in $\{0, 1, \dots, r-1\}$, with $\mathbb{E}_\theta[A_{ij}] = \mu(\theta_i + \theta_j)$. By Hoeffding's inequality [14], for each $i = 1, \dots, n$ we have

$$\begin{aligned}
\mathbb{P}\left(|d_i - d_i^*| \geq (r-1)\sqrt{\frac{kn\log n}{2}}\right) &\leq \mathbb{P}\left(|d_i - d_i^*| \geq (r-1)\sqrt{\frac{k(n-1)\log n}{2}}\right) \\
&= \mathbb{P}\left(\left|\frac{1}{n-1}\sum_{j\neq i}(A_{ij} - \mu(\theta_i + \theta_j))\right| \geq (r-1)\sqrt{\frac{k\log n}{2(n-1)}}\right) \\
&\leq 2\exp\left(-\frac{2(n-1)}{(r-1)^2} \cdot \frac{(r-1)^2 k \log n}{2(n-1)}\right) \\
&= \frac{2}{n^k}.
\end{aligned}$$

Therefore, by union bound, with probability at least $1 - 2/n^{k-1}$ we have $\|\mathbf{d} - \mathbf{d}^*\|_\infty \le (r-1)\sqrt{kn \log n/2}$. Assume we are in this situation. Then from (28) we see that for all $i = 1, \dots, n$,

$$(n-1)\,\mu(2M) - (r-1)\sqrt{\frac{kn\log n}{2}} \le d_i \le (n-1)\,\mu(-2M) + (r-1)\sqrt{\frac{kn\log n}{2}}.$$

Thus, for sufficiently large n, we have

$$c_2(r-1)(n-1) \le d_i \le c_1(r-1)(n-1), \quad i = 1, \dots, n.$$

with

$$c_1 = \frac{3\mu(-2M)}{2(r-1)}, \qquad c_2 = \frac{\mu(2M)}{2(r-1)}.$$

Moreover, it is easy to see that for every $B \subseteq \{1, \dots, n\}$ we have $|g(\mathbf{d}, B) - g(\mathbf{d}^*, B)| \le \sum_{i=1}^n |d_i - d_i^*| \le n\|\mathbf{d} - \mathbf{d}^*\|_\infty$. Since we already know that $g(\mathbf{d}^*, B) \ge |B|(|B|-1)\,\mu(2M)$, this gives us

$$g(\mathbf{d}, B) \ge g(\mathbf{d}^*, B) - n\|\mathbf{d} - \mathbf{d}^*\|_\infty \ge |B|(|B|-1)\mu(2M) - (r-1)\sqrt{\frac{kn^3\log n}{2}}.$$

Thus, for $|B| \ge c_2^2(n-1)$ and for sufficiently large n, we have $g(\mathbf{d}, B) \ge c_3 n^2$ with $c_3 = \frac{1}{2}c_2^4\,\mu(2M)$.

We have shown that $\mathbf{d}$ satisfies the properties (24) and (25), so by Lemma 4.11, the MLE $\hat{\theta}$ exists and satisfies $\|\hat{\theta}\|_\infty \le C$, where the constant C only depends on M (and r). Assume further that $C \ge M$, so $\|\theta\|_\infty \le C$ as well.

To bound the deviation of $\hat{\theta}$ from θ, we use the convergence rate in the iterative algorithm to compute $\hat{\theta}$. Set $\hat{\theta}^{(0)} = \theta$ in the algorithm in Theorem 3.3, so by Proposition 4.10, we have

$$\|\hat{\theta} - \theta\|_\infty \le \frac{2}{\delta^2}\|\theta - \varphi(\theta)\|_\infty, \tag{29}$$

where δ is given by (21) with $K = 2\|\hat{\theta}\|_\infty + \|\theta\|_\infty \le 3C$. From the definition of φ (10), we see that for each $1 \le i \le n$,

$$\theta_i - \varphi_i(\theta) = \frac{1}{r-1}\left(\log d_i - \log \sum_{j \ne i} \mu(\theta_i + \theta_j)\right) = \frac{1}{r-1}\log\frac{d_i}{d_i^*}.$$

Noting that $(y-1)/y \le \log y \le y - 1$ for $y > 0$, we have $|\log(d_i/d_i^*)| \le |d_i - d_i^*|/\min\{d_i, d_i^*\}$. Using the bounds on $\|\mathbf{d} - \mathbf{d}^*\|_\infty$ and d_i, d_i^* that we have developed above, we get

$$\|\theta - \varphi(\theta)\|_\infty \le \frac{\|\mathbf{d} - \mathbf{d}^*\|_\infty}{\min\{\min_i d_i,\ \min_i d_i^*\}} \le (r-1)\sqrt{\frac{kn\log n}{2}} \cdot \frac{2}{\mu(2M)\,(n-1)} \le \frac{2(r-1)}{\mu(2M)}\sqrt{\frac{k\log n}{n}}.$$

Plugging this bound to (29) gives us the desired result. □

4.3 Proofs for the continuous weighted graphs

In this section we present the proofs of the results presented in Section 3.2.

4.3.1 Proof of Theorem 3.6

Clearly if $(d_1, \dots, d_n) \in \mathbb{R}_0^n$ is a graphical sequence, then so is $(d_{\pi(1)}, \dots, d_{\pi(n)})$, for any permutation π of $\{1, \dots, n\}$. Thus, without loss of generality we can assume $d_1 \ge d_2 \ge \cdots \ge d_n$, and in this case condition (14) reduces to

$$d_1 \le \sum_{i=2}^n d_i. \tag{30}$$

First suppose $(d_1, \dots, d_n) \in \mathbb{R}_0^n$ is graphic, so it is the degree sequence of a graph with adjacency matrix $\mathbf{a} = (a_{ij})$. Then condition (30) is satisfied since

$$d_1 = \sum_{i=2}^{n} a_{1i} \leq \sum_{i=2}^{n} \sum_{j \neq i} a_{ij} = \sum_{i=2}^{n} d_i.$$

For the converse direction, we first note the following easy properties of weighted graphical sequences:

(i) The sequence $(c, c, \dots, c) \in \mathbb{R}_0^n$ is graphic for any $c \in \mathbb{R}_0$, realized by the "cycle graph" with weights $a_{i,i+1} = c/2$ for $1 \leq i \leq n-1$, $a_{1n} = c/2$, and $a_{ij} = 0$ otherwise.

(ii) A sequence $\mathbf{d} = (d_1, \dots, d_n) \in \mathbb{R}_0^n$ satisfying (30) with an equality is graphic, realized by the "star graph" with weights $a_{1i} = d_i$ for $2 \leq i \leq n$ and $a_{ij} = 0$ otherwise.

(iii) If $\mathbf{d} = (d_1, \dots, d_n) \in \mathbb{R}_0^n$ is graphic, then so is $\overline{\mathbf{d}} = (d_1, \dots, d_n, 0, \dots, 0) \in \mathbb{R}_0^{n'}$ for any $n' \geq n$, realized by inserting $n' - n$ isolated vertices to the graph that realizes $\mathbf{d}$.

(iv) If $\mathbf{d}^{(1)}, \mathbf{d}^{(2)} \in \mathbb{R}_0^n$ are graphic, then so is $\mathbf{d}^{(1)} + \mathbf{d}^{(2)}$, realized by the graph whose edge weights are the sum of the edge weights of the graphs realizing $\mathbf{d}^{(1)}$ and $\mathbf{d}^{(2)}$.

We now prove the converse direction by induction on n. For the base case $n = 3$, it is easy to verify that (d_1, d_2, d_3) with $d_1 \geq d_2 \geq d_3 \geq 0$ and $d_1 \leq d_2 + d_3$ is the degree sequence of the graph G with edge weights

$$a_{12} = \frac{1}{2}(d_1 + d_2 - d_3) \geq 0, \qquad a_{13} = \frac{1}{2}(d_1 - d_2 + d_3) \geq 0, \qquad a_{23} = \frac{1}{2}(-d_1 + d_2 + d_3) \geq 0.$$

Assume that the claim holds for $n - 1$; we will prove it also holds for n. So suppose we have a sequence $\mathbf{d} = (d_1, \dots, d_n)$ with $d_1 \geq d_2 \geq \dots \geq d_n \geq 0$ satisfying (30), and let

$$K = \frac{1}{n-2} \left(\sum_{i=2}^{n} d_i - d_1 \right) \geq 0$$

If $K = 0$ then (30) is satisfied with an equality, and by property (ii) we know that $\mathbf{d}$ is graphic. Now assume $K > 0$. We consider two possibilities.

1. Suppose $K \geq d_n$. Then we can write $\mathbf{d} = \mathbf{d}^{(1)} + \mathbf{d}^{(2)}$, where

$$\mathbf{d}^{(1)} = (d_1 - d_n,\ d_2 - d_n,\ \dots,\ d_{n-1} - d_n,\ 0) \in \mathbb{R}_0^n$$

and

$$\mathbf{d}^{(2)} = (d_n, d_n, \dots, d_n) \in \mathbb{R}_0^n.$$

The assumption $K \geq d_n$ implies $d_1 - d_n \leq \sum_{i=2}^{n-1}(d_i - d_n)$, so $(d_1 - d_n, d_2 - d_n, \dots, d_{n-1} - d_n) \in \mathbb{R}_0^{n-1}$ is a graphical sequence by induction hypothesis. Thus, $\mathbf{d}^{(1)}$ is also graphic by property (iii). Furthermore, $\mathbf{d}^{(2)}$ is graphic by property (i), so $\mathbf{d} = \mathbf{d}^{(1)} + \mathbf{d}^{(2)}$ is also a graphical sequence by property (iv).

2. Suppose $K < d_n$. Then write $\mathbf{d} = \mathbf{d}^{(3)} + \mathbf{d}^{(4)}$, where

$$\mathbf{d}^{(3)} = (d_1 - K,\ d_2 - K,\ \dots,\ d_n - K) \in \mathbb{R}_0^n$$

and

$$\mathbf{d}^{(4)} = (K, K, \dots, K) \in \mathbb{R}_0^n.$$

By construction, $\mathbf{d}^{(3)}$ satisfies $d_1 - K = \sum_{i=2}^{n}(d_i - K)$, so $\mathbf{d}^{(3)}$ is a graphical sequence by property (ii). Since $\mathbf{d}^{(4)}$ is also graphic by property (i), we conclude that $\mathbf{d} = \mathbf{d}^{(3)} + \mathbf{d}^{(4)}$ is graphic by property (iv).

This completes the induction step and finishes the proof of Theorem 3.6.

4.3.2 Proof of Lemma 3.7

We first prove that $\mathcal{W}$ is convex. Given $\mathbf{d} = (d_1, \dots, d_n)$ and $\mathbf{d}' = (d'_1, \dots, d'_n)$ in $\mathcal{W}$, and given $0 \le t \le 1$, we note that

$$\begin{aligned}\max_{1\le i\le n} \big(td_i + (1-t)d'_i\big) &\le t \max_{1\le i\le n} d_i + (1-t) \max_{1\le i\le n} d'_i \\ &\le \frac{1}{2} t \sum_{i=1}^n d_i + \frac{1}{2}(1-t) \sum_{i=1}^n d'_i \\ &= \frac{1}{2} \sum_{i=1}^n \big(td_i + (1-t)d'_i\big),\end{aligned}$$

which means $t\mathbf{d} + (1-t)\mathbf{d}' \in \mathcal{W}$.

Next, recall that we already have $\mathcal{M} \subseteq \mathrm{conv}(\mathcal{W}) = \mathcal{W}$ from Proposition 2.2, so to conclude $\mathcal{M} = \mathcal{W}$ it remains to show that $\mathcal{W} \subseteq \mathcal{M}$. Given $\mathbf{d} \in \mathcal{W}$, let G be a graph that realizes $\mathbf{d}$ and let $\mathbf{w} = (w_{ij})$ be the edge weights of G, so that $d_i = \sum_{j \ne i} w_{ij}$ for all $i = 1, \dots, n$. Consider a distribution $\mathbb{P}$ on $\mathbb{R}_0^{\binom{n}{2}}$ that assigns each edge weight A_{ij} to be an independent exponential random variable with mean parameter w_{ij}, so $\mathbb{P}$ has density

$$p(\mathbf{a}) = \prod_{\{i,j\}} \frac{1}{w_{ij}} \exp\left(-\frac{a_{ij}}{w_{ij}}\right), \quad \mathbf{a} = (a_{ij}) \in \mathbb{R}_0^{\binom{n}{2}}.$$

Then by construction, we have $\mathbb{E}_{\mathbb{P}}[A_{ij}] = w_{ij}$ and

$$\mathbb{E}_{\mathbb{P}}[\deg_i(A)] = \sum_{j \ne i} \mathbb{E}_{\mathbb{P}}[A_{ij}] = \sum_{j\ne i} w_{ij} = d_i, \quad i = 1, \dots, n.$$

This shows that $\mathbf{d} \in \mathcal{M}$, as desired.

4.3.3 Proof of Theorem 3.9

We first prove that the MLE $\hat{\theta}$ exists almost surely. Recall from the discussion in Section 3.2 that $\hat{\theta}$ exists if and only if $\mathbf{d} \in \mathcal{M}^\circ$. Clearly $\mathbf{d} \in \mathcal{W}$ since $\mathbf{d}$ is the degree sequence of the sampled graph G. Since $\mathcal{M} = \mathcal{W}$ (Lemma 3.7), we see that the MLE $\hat{\theta}$ does not exist if and only if $\mathbf{d} \in \partial\mathcal{M} = \mathcal{M} \setminus \mathcal{M}^\circ$, where

$$\partial\mathcal{M} = \left\{\mathbf{d}' \in \mathbb{R}_0^n \colon \min_{1\le i\le n} d'_i = 0 \;\text{ or }\; \max_{1\le i\le n} d'_i = \frac{1}{2}\sum_{i=1}^n d'_i\right\}.$$

In particular, note that $\partial\mathcal{M}$ has Lebesgue measure 0. Since the distribution $\mathbb{P}^*$ on the edge weights $A = (A_{ij})$ is continuous (being a product of exponential distributions) and $\mathbf{d}$ is a continuous function of A, we conclude that $\mathbb{P}^*(\mathbf{d} \in \partial\mathcal{M}) = 0$, as desired.

We now prove the consistency of $\hat{\theta}$. Recall that θ is the true parameter that we wish to estimate, and that the MLE $\hat{\theta}$ satisfies $-Z(\hat{\theta}) = \mathbf{d}$. Let $\mathbf{d}^* = -\nabla Z(\theta)$ denote the expected degree sequence of the maximum entropy distribution $\mathbb{P}^*_\theta$. By the mean value theorem for vector-valued functions [19, p. 341], we can write

$$\mathbf{d} - \mathbf{d}^* = \nabla Z(\theta) - \nabla Z(\hat{\theta}) = J(\theta - \hat{\theta}). \tag{31}$$

Here J is a matrix obtained by integrating (element-wise) the Hessian $\nabla^2 Z$ of the log-partition function on intermediate points between θ and $\hat{\theta}$:

$$J = \int_0^1 \nabla^2 Z(t\theta + (1-t)\hat{\theta})\, dt.$$

Recalling that $-\nabla Z(\theta) = \mathbb{E}_\theta[\deg(A)]$, at any intermediate point $\xi \equiv \xi(t) = t\theta + (1-t)\hat{\theta}$, we have

$$\big(\nabla Z(\xi)\big)_i = -\sum_{j \neq i} \mu(\xi_i + \xi_j) = -\sum_{j\neq i} \frac{1}{\xi_i + \xi_j}.$$

Therefore, the Hessian $\nabla^2 Z$ is given by

$$\big(\nabla^2 Z(\xi)\big)_{ij} = \frac{1}{(\xi_i + \xi_j)^2} \quad i \neq j,$$

and

$$\big(\nabla^2 Z(\xi)\big)_{ii} = \sum_{j \neq i} \frac{1}{(\xi_i+\xi_j)^2} = \sum_{j\neq i} \big(\nabla^2 Z(\xi)\big)_{ij}.$$

Since $\theta, \theta' \in \Theta$ and we assume $\theta_i + \theta_j \leq M$, it follows that for $i \neq j$,

$$0 < \xi_i + \xi_j \leq \max\{\theta_i + \theta_j,\, \hat{\theta}_i + \hat{\theta}_j\} \leq \max\{M, 2\|\hat{\theta}\|_\infty\} \leq M + 2\|\hat{\theta}\|_\infty.$$

Therefore, the Hessian $\nabla^2 Z$ is a *diagonally balanced* matrix with off-diagonal entries bounded below by $1/(M+2\|\hat{\theta}\|_\infty)^2$. In particular, J is also a symmetric, diagonally balanced matrix with off-diagonal entries bounded below by $1/(M+2\|\hat{\theta}\|_\infty)^2$, being an average of such matrices. By Theorem 4.4, J is invertible and its inverse satisfies the bound

$$\|J^{-1}\|_\infty \leq \frac{(M+2\|\hat{\theta}\|_\infty)^2(3n-4)}{2(n-1)(n-2)} \leq \frac{2}{n}\,(M+2\|\hat{\theta}\|_\infty)^2,$$

where the last inequality holds for $n \geq 7$. Inverting J in (31) and applying the bound on $\|J^{-1}\|_\infty$ gives

$$\|\theta - \hat{\theta}\|_\infty \leq \|J^{-1}\|_\infty \,\|\mathbf{d} - \mathbf{d}^*\|_\infty \leq \frac{2}{n}\,(M+2\|\hat{\theta}\|_\infty)^2\,\|\mathbf{d}-\mathbf{d}^*\|_\infty. \tag{32}$$

Let $A = (A_{ij})$ denote the edge weights of the sampled graph $G \sim \mathbb{P}^*_\theta$, so $d_i = \sum_{j\neq i} A_{ij}$ for $i = 1, \dots, n$. Moreover, since $\mathbf{d}^*$ is the expected degree sequence from the distribution $\mathbb{P}^*_\theta$, we also have $d_i^* = \sum_{j\neq i} 1/(\theta_i + \theta_j)$. Recall that A_{ij} is an exponential random variable with rate $\lambda = \theta_i + \theta_j \geq L$, so by Lemma 4.2, $A_{ij} - 1/(\theta_i+\theta_j)$ is sub-exponential with parameter $2/(\theta_i+\theta_j) \leq 2/L$. For each $i = 1,\dots,n$, the random variables $(A_{ij} - 1/(\theta_i+\theta_j), j \neq i)$ are independent sub-exponential random variables, so we can apply the concentration inequality in Theorem 4.1 with $\kappa = 2/L$ and

$$\epsilon = \left(\frac{4k\log n}{\gamma(n-1)L^2}\right)^{1/2}.$$

Assume n is sufficiently large such that $\epsilon/\kappa = \sqrt{k\log n/\gamma(n-1)} \leq 1$. Then by Theorem 4.1, for each $i = 1,\dots,n$ we have

$$\begin{aligned}
\mathbb{P}\left(|d_i - d_i^*| \geq \sqrt{\frac{4kn\log n}{\gamma L^2}}\right) &\leq \mathbb{P}\left(|d_i - d_i^*| \geq \sqrt{\frac{4k(n-1)\log n}{\gamma L^2}}\right) \\
&= \mathbb{P}\left(\left|\frac{1}{n-1}\sum_{j\neq i}\left(A_{ij} - \frac{1}{\theta_i+\theta_j}\right)\right| \geq \sqrt{\frac{4k\log n}{\gamma(n-1)L^2}}\right) \\
&\leq 2\exp\left(-\gamma\,(n-1)\cdot\frac{L^2}{4}\cdot\frac{4k\log n}{\gamma(n-1)L^2}\right) \\
&= \frac{2}{n^k}.
\end{aligned}$$

By the union bound,

$$\mathbb{P}\left(\|\mathbf{d} - \mathbf{d}^*\|_\infty \geq \sqrt{\frac{4kn \log n}{\gamma L^2}} \right) \leq \sum_{i=1}^{n} \mathbb{P}\left(|d_i - d_i^*| \geq \sqrt{\frac{4kn \log n}{\gamma L^2}} \right) \leq \frac{2}{n^{k-1}}.$$

Assume for the rest of this proof that $\|\mathbf{d} - \mathbf{d}^*\|_\infty \leq \sqrt{4kn \log n/(\gamma L^2)}$, which happens with probability at least $1 - 2/n^{k-1}$. From (32) and using the triangle inequality, we get

$$\|\hat{\theta}\|_\infty \leq \|\theta - \hat{\theta}\|_\infty + \|\theta\|_\infty \leq \frac{4}{L} \sqrt{\frac{k \log n}{\gamma n}} \, (M + 2\|\hat{\theta}\|_\infty)^2 + M.$$

What we have shown is that for sufficiently large n, $\|\hat{\theta}\|_\infty$ satisfies the inequality $G_n(\|\hat{\theta}\|_\infty) \geq 0$, where $G_n(x)$ is the quadratic function

$$G_n(x) = \frac{4}{L} \sqrt{\frac{k \log n}{\gamma n}} \, (M + 2x)^2 - x + M.$$

It is easy to see that for sufficiently large n we have $G_n(2M) < 0$ and $G_n(\log n) < 0$. Thus, $G_n(\|\hat{\theta}\|_\infty) \geq 0$ means either $\|\hat{\theta}\|_\infty < 2M$ or $\|\hat{\theta}\|_\infty > \log n$. We claim that for sufficiently large n we always have $\|\hat{\theta}\|_\infty < 2M$. Suppose the contrary that there are infinitely many n for which $\|\hat{\theta}\|_\infty > \log n$, and consider one such n. Since $\hat{\theta} \in \Theta$ we know that $\hat{\theta}_i + \hat{\theta}_j > 0$ for each $i \neq j$, so there can be at most one index i with $\hat{\theta}_i < 0$. We consider the following two cases:

1. **Case 1:** suppose $\hat{\theta}_i \geq 0$ for all $i = 1, \ldots, n$. Let i^* be an index with $\hat{\theta}_{i^*} = \|\hat{\theta}\|_\infty > \log n$. Then, using the fact that $\hat{\theta}$ satisfies the system of equations (13) and $\hat{\theta}_{i^*} + \hat{\theta}_j \geq \hat{\theta}_{i^*}$ for $j \neq i^*$, we see that

$$\begin{aligned}
\frac{1}{M} &\leq \frac{1}{n-1} \sum_{j \neq i^*} \frac{1}{\theta_{i^*} + \theta_j} \\
&\leq \frac{1}{n-1} \left| \sum_{j \neq i^*} \frac{1}{\theta_{i^*} + \theta_j} - \sum_{j \neq i^*} \frac{1}{\hat{\theta}_{i^*} + \hat{\theta}_j} \right| + \frac{1}{n-1} \sum_{j \neq i^*} \frac{1}{\hat{\theta}_{i^*} + \hat{\theta}_j} \\
&= \frac{1}{n-1} \, |d_i^* - d_i| + \frac{1}{n-1} \sum_{j \neq i^*} \frac{1}{\hat{\theta}_{i^*} + \hat{\theta}_j} \\
&\leq \frac{1}{n-1} \|\mathbf{d}^* - \mathbf{d}\|_\infty + \frac{1}{\|\hat{\theta}\|_\infty} \\
&\leq \frac{1}{n-1} \sqrt{\frac{4kn \log n}{\gamma L^2}} + \frac{1}{\log n},
\end{aligned}$$

   which cannot hold for sufficiently large n, as the last expression tends to 0 as $n \to \infty$.

2. **Case 2:** suppose $\hat{\theta}_i < 0$ for some $i = 1, \ldots, n$, so $\hat{\theta}_j > 0$ for $j \neq i$ since $\hat{\theta} \in \Theta$. Without loss of generality assume $\hat{\theta}_1 < 0 < \hat{\theta}_2 \leq \cdots \leq \hat{\theta}_n$, so $\hat{\theta}_n = \|\hat{\theta}\|_\infty > \log n$. Following the same chain of inequalities as in the previous case (with $i^* = n$), we obtain

$$\begin{aligned}
\frac{1}{M} &\leq \frac{1}{n-1} \|\mathbf{d}^* - \mathbf{d}\|_\infty + \frac{1}{n-1} \left(\frac{1}{\hat{\theta}_n + \hat{\theta}_1} + \sum_{j=2}^{n-1} \frac{1}{\hat{\theta}_j + \hat{\theta}_n} \right) \\
&\leq \frac{1}{n-1} \sqrt{\frac{4kn \log n}{\gamma L^2}} + \frac{1}{(n-1)(\hat{\theta}_n + \hat{\theta}_1)} + \frac{n-2}{(n-1)\|\hat{\theta}\|_\infty} \\
&\leq \frac{1}{n-1} \sqrt{\frac{4kn \log n}{\gamma L^2}} + \frac{1}{(n-1)(\hat{\theta}_n + \hat{\theta}_1)} + \frac{1}{\log n}.
\end{aligned}$$

So for sufficiently large n,

$$\frac{1}{\hat{\theta}_1 + \hat{\theta}_n} \geq (n-1)\left(\frac{1}{M} - \frac{1}{n-1}\sqrt{\frac{4kn\log n}{\gamma L^2}} - \frac{1}{\log n}\right) \geq \frac{n}{2M},$$

and thus $\hat{\theta}_1 + \hat{\theta}_i \leq \hat{\theta}_1 + \hat{\theta}_n \leq 2M/n$ for each $i = 2, \ldots, n$. However, then

$$\sqrt{\frac{4kn\log n}{\gamma L^2}} \geq \|\mathbf{d}^* - \mathbf{d}\|_\infty \geq |d_1^* - d_1| \geq -\sum_{j=2}^{n} \frac{1}{\theta_1 + \theta_j} + \sum_{j=2}^{n} \frac{1}{\hat{\theta}_1 + \hat{\theta}_j} \geq -\frac{(n-1)}{L} + \frac{n(n-1)}{2M},$$

which cannot hold for sufficiently large n, as the right hand side of the last expression tends to ∞ faster than the left hand side.

The analysis above shows that $\|\hat{\theta}\|_\infty < 2M$ for all sufficiently large n. Plugging in this result to (32), we conclude that for sufficiently large n, with probability at least $1 - 2n^{-(k-1)}$ we have the bound

$$\|\theta - \hat{\theta}\|_\infty \leq \frac{2}{n}\,(5M)^2\sqrt{\frac{4kn\log n}{\gamma L^2}} = \frac{100M^2}{L}\sqrt{\frac{k\log n}{\gamma n}},$$

as desired.

4.4 Proofs for the infinite discrete weighted graphs

In this section we prove the results presented in Section 3.3.

4.4.1 Proof of Theorem 3.10

Without loss of generality we may assume $d_1 \geq d_2 \geq \cdots \geq d_n$, so condition (16) becomes $d_1 \leq \sum_{i=2}^n d_i$. The necessary part is easy: if $(d_1, \ldots, d_n)$ is a degree sequence of a graph G with edge weights $a_{ij} \in \mathbb{N}_0$, then $\sum_{i=1}^n d_i = 2\sum_{\{i,j\}} a_{ij}$ is even, and the total weight coming out of vertex 1 is at most $\sum_{i=2}^n d_i$. For the converse direction, we proceed by induction on $s = \sum_{i=1}^n d_i$. The statement is clearly true for $s = 0$ and $s = 2$. Assume the statement is true for some even $s \in \mathbb{N}$, and suppose we are given $\mathbf{d} = (d_1, \ldots, d_n) \in \mathbb{N}_0^n$ with $d_1 \geq \cdots \geq d_n$, $\sum_{i=1}^n d_i = s+2$, and $d_1 \leq \sum_{i=2}^n d_i$. Without loss of generality we may assume $d_n \geq 1$, for otherwise we can proceed with only the nonzero elements of $\mathbf{d}$. Let $1 \leq t \leq n-1$ be the smallest index such that $d_t > d_{t+1}$, with $t = n-1$ if $d_1 = \cdots = d_n$, and let $\mathbf{d}' = (d_1, \ldots, d_{t-1}, d_t - 1, d_{t+1}, \ldots, d_n - 1)$. We will show that $\mathbf{d}'$ is graphic. This will imply that $\mathbf{d}$ is graphic, because if $\mathbf{d}'$ is realized by the graph G' with edge weights a'_{ij}, then $\mathbf{d}$ is realized by the graph G with edge weights $a_{tn} = a'_{tn} + 1$ and $a_{ij} = a'_{ij}$ otherwise.

Now for $\mathbf{d}' = (d'_1, \ldots, d'_n)$ given above, we have $d'_1 \geq \cdots \geq d'_n$ and $\sum_{i=1}^n d'_i = \sum_{i=1}^n d_i - 2 = s$ is even. So it suffices to show that $d'_1 \leq \sum_{i=2}^n d'_i$, for then we can apply the induction hypothesis to conclude that $\mathbf{d}'$ is graphic. If $t = 1$, then $d'_1 = d_1 - 1 \leq \sum_{i=2}^n d_i - 1 = \sum_{i=2}^n d'_i$. If $t > 1$ then $d_1 = d_2$, so $d_1 < \sum_{i=2}^n d_i$ since $d_n \geq 1$. In particular, since $\sum_{i=1}^n d_i$ is even, $\sum_{i=2}^n d_i - d_1 = \sum_{i=1}^n d_i - 2d_1$ is also even, hence $\sum_{i=2}^n d_i - d_1 \geq 2$. Therefore, $d'_1 = d_1 \leq \sum_{i=2}^n d_i - 2 = \sum_{i=2}^n d'_i$. This finishes the proof of Theorem 3.10.

4.4.2 Proof of Lemma 3.11

Clearly $\mathcal{W} \subseteq \mathcal{W}_1$, so $\overline{\text{conv}(\mathcal{W})} \subseteq \mathcal{W}_1$ since $\mathcal{W}_1$ is closed and convex, by Lemma 3.7. Conversely, let $\mathbb{Q}$ denote the set of rational numbers. We will first show that $\mathcal{W}_1 \cap \mathbb{Q}^n \subseteq \text{conv}(\mathcal{W})$ and then proceed by a limit argument. Let $\mathbf{d} \in \mathcal{W}_1 \cap \mathbb{Q}^n$, so $\mathbf{d} = (d_1, \ldots, d_n) \in \mathbb{Q}^n$ with $d_i \geq 0$ and $\max_{1 \leq i \leq n} d_i \leq \frac{1}{2}\sum_{i=1}^n d_i$. Choose $K \in \mathbb{N}$ large enough such that $Kd_i \in \mathbb{N}_0$ for all $i = 1, \ldots, n$. Observe that $2K\mathbf{d} = (2Kd_1, \ldots, 2Kd_n) \in \mathbb{N}_0^n$ has the property that $\sum_{i=1}^n 2Kd_i \in \mathbb{N}_0$ is even and $\max_{1 \leq i \leq n} 2Kd_i \leq \frac{1}{2}\sum_{i=1}^n 2Kd_i$, so $2K\mathbf{d} \in \mathcal{W}$ by definition. Since $0 = (0, \ldots, 0) \in \mathcal{W}$ as well, all elements along the segment joining 0 and $2K\mathbf{d}$ lie in

$\mathrm{conv}(\mathcal{W})$, so in particular, $\mathbf{d} = (2K\mathbf{d})/(2K) \in \mathrm{conv}(\mathcal{W})$. This shows that $\mathcal{W}_1 \cap \mathbb{Q}^n \subseteq \mathrm{conv}(\mathcal{W})$, and hence $\overline{\mathcal{W}_1 \cap \mathbb{Q}^n} \subseteq \overline{\mathrm{conv}(\mathcal{W})}$.

To finish the proof it remains to show that $\overline{\mathcal{W}_1 \cap \mathbb{Q}^n} = \mathcal{W}_1$. On the one hand we have

$$\overline{\mathcal{W}_1 \cap \mathbb{Q}^n} \subseteq \overline{\mathcal{W}_1} \cap \overline{\mathbb{Q}^n} = \mathcal{W}_1 \cap \mathbb{R}_0^n = \mathcal{W}_1.$$

For the other direction, given $\mathbf{d} \in \mathcal{W}_1$, choose $\mathbf{d}_1, \dots, \mathbf{d}_n \in W_1$ such that $\mathbf{d}, \mathbf{d}_1, \dots, \mathbf{d}_n$ are in general position, so that the convex hull C of $\{\mathbf{d}, \mathbf{d}_1, \dots, \mathbf{d}_n\}$ is full dimensional. This can be done, for instance, by noting that the following $n+1$ points in $\mathcal{W}_1$ are in general position:

$$\{0,\ \mathbf{e}_1 + \mathbf{e}_2,\ \mathbf{e}_1 + \mathbf{e}_3,\ \cdots,\ \mathbf{e}_1 + \mathbf{e}_n,\ \mathbf{e}_1 + \mathbf{e}_2 + \cdots + \mathbf{e}_n\},$$

where $\mathbf{e}_1, \dots, \mathbf{e}_n$ are the standard basis of $\mathbb{R}^n$. For each $m \in \mathbb{N}$ and $i = 1, \dots, n$, choose $\mathbf{d}_i^{(m)}$ on the line segment between $\mathbf{d}$ and $\mathbf{d}_i$ such that the convex hull C_m of $\{\mathbf{d}, \mathbf{d}_1^{(m)}, \dots, \mathbf{d}_n^{(m)}\}$ is full dimensional and has diameter at most $1/m$. Since C_m is full dimensional we can choose a rational point $\mathbf{r}_m \in C_m \subseteq C \subseteq \mathcal{W}_1$. Thus we have constructed a sequence of rational points $(\mathbf{r}_m)$ in $\mathcal{W}_1$ converging to $\mathbf{d}$, which shows that $\mathcal{W}_1 \subseteq \overline{\mathcal{W}_1 \cap \mathbb{Q}^n}$.

4.4.3 Proof of Theorem 3.13

We first address the issue of the existence of $\hat{\theta}$. Recall from the discussion in Section 3.3 that the MLE $\hat{\theta} \in \Theta$ exists if and only if $\mathbf{d} \in \mathcal{M}^\circ$. Clearly $\mathbf{d} \in \mathcal{W}$ since $\mathbf{d}$ is the degree sequence of the sampled graph G, and $\mathcal{W} \subseteq \mathrm{conv}(\mathcal{W}) = \mathcal{M}$ from Proposition 2.2. Therefore, the MLE $\hat{\theta}$ does not exist if and only if $\mathbf{d} \in \partial\mathcal{M} = \mathcal{M} \setminus \mathcal{M}^\circ$, where the boundary $\partial\mathcal{M}$ is explicitly given by

$$\partial\mathcal{M} = \left\{ \mathbf{d}' \in \mathbb{R}_0^n \colon \min_{1 \le i \le n} d'_i = 0 \ \text{ or } \ \max_{1 \le i \le n} d'_i = \frac{1}{2} \sum_{i=1}^n d'_i \right\}.$$

Using union bound and the fact that the edge weights A_{ij} are independent geometric random variables, we have

$$\begin{aligned} \mathbb{P}(d_i = 0 \text{ for some } i) &\le \sum_{i=1}^n \mathbb{P}(d_i = 0) = \sum_{i=1}^n \mathbb{P}(A_{ij} = 0 \text{ for all } j \neq i) \\ &= \sum_{i=1}^n \prod_{j \neq i} (1 - \exp(-\theta_i - \theta_j)) \le n\,(1 - \exp(-M))^{n-1}. \end{aligned}$$

Furthermore, again by union bound,

$$\mathbb{P}\left(\max_{1 \le i \le n} d_i = \frac{1}{2} \sum_{i=1}^n d_i \right) = \mathbb{P}\left(d_i = \sum_{j \neq i} d_j \text{ for some } i \right) \le \sum_{i=1}^n \mathbb{P}\left(d_i = \sum_{j \neq i} d_j \right).$$

Note that we have $d_i = \sum_{j \neq i} d_j$ for some i if and only if the edge weights $A_{jk} = 0$ for all $j, k \neq i$. This occurs with probability

$$\mathbb{P}\left(A_{jk} = 0 \text{ for } j, k \neq i\right) = \prod_{\substack{j,k \neq i \\ j \neq k}} (1 - \exp(-\theta_j - \theta_k)) \le (1 - \exp(-M))^{\binom{n-1}{2}}.$$

Therefore,

$$\begin{aligned} \mathbb{P}(\mathbf{d} \in \partial\mathcal{M}) &\le \mathbb{P}(d_i = 0 \text{ for some } i) + \mathbb{P}\left(\max_{1 \le i \le n} d_i = \frac{1}{2} \sum_{i=1}^n d_i \right) \\ &\le n\,(1 - \exp(-M))^{n-1} + n\,(1 - \exp(-M))^{\binom{n-1}{2}} \\ &\le \frac{1}{n^{k-1}}, \end{aligned}$$

where the last inequality holds for sufficiently large n. This shows that for sufficiently large n, the MLE $\hat{\theta}$ exists with probability at least $1 - 1/n^{k-1}$.

We now turn to proving the consistency of $\hat{\theta}$. For the rest of this proof, assume that the MLE $\hat{\theta} \in \Theta$ exists, which occurs with probability at least $1 - 1/n^{k-1}$. The proof of the consistency of $\hat{\theta}$ follows the same outline as in the proof of Theorem 3.9. Let $\mathbf{d}^* = -\nabla Z(\theta)$ denote the expected degree sequence of the distribution $\mathbb{P}^*_\theta$, and recall that the MLE $\hat{\theta}$ satisfies $\mathbf{d} = -\nabla Z(\hat{\theta})$. By the mean value theorem [19, p. 341], we can write

$$\mathbf{d} - \mathbf{d}^* = \nabla Z(\theta) - \nabla Z(\hat{\theta}) = J(\theta - \hat{\theta}), \tag{33}$$

where J is the matrix obtained by integrating the Hessian of Z between θ and $\hat{\theta}$,

$$J = \int_0^1 \nabla^2 Z(t\theta + (1-t)\hat{\theta})\, dt.$$

Let $0 \le t \le 1$, and note that at the point $\xi = t\theta + (1-t)\hat{\theta}$ the gradient ∇Z is given by

$$\big(\nabla Z(\xi)\big)_i = -\sum_{j \neq i} \frac{1}{\exp(\xi_i + \xi_j) - 1}.$$

Thus, the Hessian $\nabla^2 Z$ is

$$\big(\nabla^2 Z(\xi)\big)_{ij} = \frac{\exp(\xi_i + \xi_j)}{(\exp(\xi_i + \xi_j) - 1)^2} \quad i \neq j,$$

and

$$\big(\nabla^2 Z(\xi)\big)_{ii} = \sum_{j \neq i} \frac{\exp(\xi_i + \xi_j)}{(\exp(\xi_i + \xi_j) - 1)^2} = \sum_{j \neq i} \big(\nabla^2 Z(\xi)\big)_{ij}.$$

Since $\theta, \hat{\theta} \in \Theta$ and we assume $\theta_i + \theta_j \le M$, for $i \neq j$ we have

$$0 < \xi_i + \xi_j \le \max\{\theta_i + \theta_j,\ \hat{\theta}_i + \hat{\theta}_j\} \le \max\{M,\ 2\|\hat{\theta}\|_\infty\} \le M + 2\|\hat{\theta}\|_\infty.$$

This means J is a symmetric, diagonally dominant matrix with off-diagonal entries bounded below by $\exp(M + 2\|\hat{\theta}\|_\infty)/(\exp(M+2\|\hat{\theta}\|_\infty)-1)^2$, being an average of such matrices. Then by Theorem 4.4, we have the bound

$$\|J^{-1}\|_\infty \le \frac{(3n-4)}{2(n-2)(n-1)} \, \frac{(\exp(M + 2\|\hat{\theta}\|_\infty) - 1)^2}{\exp(M + 2\|\hat{\theta}\|_\infty)} \le \frac{2}{n} \, \frac{(\exp(M + 2\|\hat{\theta}\|_\infty) - 1)^2}{\exp(M + 2\|\hat{\theta}\|_\infty)},$$

where the second inequality holds for $n \ge 7$. By inverting J in (33) and applying the bound on J^{-1} above, we obtain

$$\|\theta - \hat{\theta}\|_\infty \le \|J^{-1}\|_\infty \, \|\mathbf{d} - \mathbf{d}^*\|_\infty \le \frac{2}{n} \, \frac{(\exp(M + 2\|\hat{\theta}\|_\infty) - 1)^2}{\exp(M + 2\|\hat{\theta}\|_\infty)} \, \|\mathbf{d} - \mathbf{d}^*\|_\infty. \tag{34}$$

Let $A = (A_{ij})$ denote the edge weights of the sampled graph $G \sim \mathbb{P}^*_\theta$, so $d_i = \sum_{j \neq i} A_{ij}$ for $i = 1, \ldots, n$. Since $\mathbf{d}^*$ is the expected degree sequence from the distribution $\mathbb{P}^*_\theta$, we also have $d^*_i = \sum_{j \neq i} 1/(\exp(\theta_i + \theta_j) - 1)$. Recall that A_{ij} is a geometric random variable with emission probability

$$q = 1 - \exp(-\theta_i - \theta_j) \ge 1 - \exp(-L),$$

so by Lemma 4.3, $A_{ij} - 1/(\exp(\theta_i + \theta_j) - 1)$ is sub-exponential with parameter $-4/\log(1-q) \le 4/L$. For each $i = 1, \ldots, n$, the random variables $(A_{ij} - 1/(\exp(\theta_i + \theta_j) - 1), j \neq i)$ are independent sub-exponential random variables, so we can apply the concentration inequality in Theorem 4.1 with $\kappa = 4/L$ and

$$\epsilon = \left(\frac{16k \log n}{\gamma(n-1)L^2} \right)^{1/2}.$$

Assume n is sufficiently large such that $\epsilon/\kappa = \sqrt{k\log n/\gamma(n-1)} \le 1$. Then by Theorem 4.1, for each $i = 1, \dots, n$ we have

$$\begin{aligned}
\mathbb{P}\left(|d_i - d_i^*| \ge \sqrt{\frac{16kn\log n}{\gamma L^2}}\right) &\le \mathbb{P}\left(|d_i - d_i^*| \ge \sqrt{\frac{16k(n-1)\log n}{\gamma L^2}}\right) \\
&= \mathbb{P}\left(\left|\frac{1}{n-1}\sum_{j\ne i}\left(A_{ij} - \frac{1}{\exp(\theta_i+\theta_j)-1}\right)\right| \ge \sqrt{\frac{16k\log n}{\gamma(n-1)L^2}}\right) \\
&\le 2\exp\left(-\gamma\,(n-1)\cdot\frac{L^2}{16}\cdot\frac{16k\log n}{\gamma(n-1)L^2}\right) \\
&= \frac{2}{n^k}.
\end{aligned}$$

The union bound then gives us

$$\mathbb{P}\left(\|\mathbf{d} - \mathbf{d}^*\|_\infty \ge \sqrt{\frac{16kn\log n}{\gamma L^2}}\right) \le \sum_{i=1}^n \mathbb{P}\left(|d_i - d_i^*| \ge \sqrt{\frac{16kn\log n}{\gamma L^2}}\right) \le \frac{2}{n^{k-1}}.$$

Assume now that $\|\mathbf{d} - \hat{\mathbf{d}}\|_\infty \le \sqrt{16kn\log n/(\gamma L^2)}$, which happens with probability at least $1 - 2/n^{k-1}$. From (34) and using the triangle inequality, we get

$$\|\hat{\theta}\|_\infty \le \|\theta - \hat{\theta}\|_\infty + \|\theta\|_\infty \le \frac{8}{L}\sqrt{\frac{k\log n}{\gamma n}}\,\frac{(\exp(M + 2\|\hat{\theta}\|_\infty) - 1)^2}{\exp(M + 2\|\hat{\theta}\|_\infty)} + M.$$

This means $\|\hat{\theta}\|_\infty$ satisfies the inequality $H_n(\|\hat{\theta}\|_\infty) \ge 0$, where $H_n(x)$ is the function

$$H_n(x) = \frac{8}{L}\sqrt{\frac{k\log n}{\gamma n}}\,\frac{(\exp(M+2x)-1)^2}{\exp(M+2x)} - x + M.$$

One can easily verify that H_n is a convex function, so H_n assumes the value 0 at most twice, and moreover, $H_n(x) \to \infty$ as $x \to \infty$. It is also easy to see that for all sufficiently large n, we have $H_n(2M) < 0$ and $H_n(\frac{1}{4}\log n) < 0$. Therefore, $H_n(\|\hat{\theta}\|_\infty) \ge 0$ implies either $\|\hat{\theta}\|_\infty < 2M$ or $\|\hat{\theta}\|_\infty > \frac{1}{4}\log n$. We claim that for sufficiently large n we always have $\|\hat{\theta}\|_\infty < 2M$. Suppose the contrary that there are infinitely many n for which $\|\hat{\theta}\|_\infty > \frac{1}{4}\log n$, and consider one such n. Since $\hat{\theta}_i + \hat{\theta}_j > 0$ for each $i \ne j$, there can be at most one index i with $\hat{\theta}_i < 0$. We consider the following two cases:

1. **Case 1:** suppose $\hat{\theta}_i \ge 0$ for all $i = 1, \dots, n$. Let i^* be an index with $\hat{\theta}_{i^*} = \|\hat{\theta}\|_\infty > \frac{1}{4}\log n$. Then, since $\hat{\theta}_{i^*} + \hat{\theta}_j \ge \hat{\theta}_{i^*}$ for $j \ne i^*$,

$$\begin{aligned}
\frac{1}{\exp(M)-1} &\le \frac{1}{n-1}\sum_{j\ne i^*}\frac{1}{\exp(\theta_{i^*}+\theta_j)-1} \\
&\le \frac{1}{n-1}\left|\sum_{j\ne i^*}\frac{1}{\exp(\theta_{i^*}+\theta_j)-1} - \sum_{j\ne i^*}\frac{1}{\exp(\hat{\theta}_{i^*}+\hat{\theta}_j)-1}\right| + \frac{1}{n-1}\sum_{j\ne i^*}\frac{1}{\exp(\hat{\theta}_{i^*}+\hat{\theta}_j)-1} \\
&\le \frac{1}{n-1}\|\mathbf{d} - \mathbf{d}^*\|_\infty + \frac{1}{\exp(\|\hat{\theta}\|_\infty)-1} \\
&\le \frac{1}{n-1}\sqrt{\frac{16kn\log n}{\gamma L^2}} + \frac{1}{n^{1/4}-1},
\end{aligned}$$

which cannot hold for sufficiently large n, as the last expression tends to 0 as $n \to \infty$.

2. **Case 2:** suppose $\hat{\theta}_i < 0$ for some $i = 1, \dots, n$, so $\hat{\theta}_j > 0$ for $j \neq i$. Without loss of generality assume $\hat{\theta}_1 < 0 < \hat{\theta}_2 \leq \dots \leq \hat{\theta}_n$, so $\hat{\theta}_n = \|\hat{\theta}\|_\infty > \frac{1}{4}\log n$. Following the same chain of inequalities as in the previous case (with $i^* = n$), we obtain

$$\begin{aligned}\frac{1}{\exp(M)-1} &\leq \frac{1}{n-1}\|\mathbf{d}-\mathbf{d}^*\|_\infty + \frac{1}{n-1}\left(\frac{1}{\exp(\hat{\theta}_n+\hat{\theta}_1)-1} + \sum_{j=2}^{n-1}\frac{1}{\exp(\hat{\theta}_j+\hat{\theta}_n)-1}\right) \\ &\leq \frac{1}{n-1}\sqrt{\frac{16kn\log n}{\gamma L^2}} + \frac{1}{(n-1)(\exp(\hat{\theta}_n+\hat{\theta}_1)-1)} + \frac{n-2}{(n-1)(\exp(\|\hat{\theta}\|_\infty)-1)} \\ &\leq \frac{1}{n-1}\sqrt{\frac{16kn\log n}{\gamma L^2}} + \frac{1}{(n-1)(\exp(\hat{\theta}_n+\hat{\theta}_1)-1)} + \frac{1}{n^{1/4}-1}.\end{aligned}$$

This implies

$$\frac{1}{\exp(\hat{\theta}_1+\hat{\theta}_n)-1} \geq (n-1)\left(\frac{1}{\exp(M)-1} - \frac{1}{n-1}\sqrt{\frac{16kn\log n}{\gamma L^2}} - \frac{1}{n^{1/4}-1}\right) \geq \frac{n}{2(\exp(M)-1)},$$

where the last inequality assumes n is sufficiently large. Therefore, for $i = 2, \dots, n$,

$$\frac{1}{\exp(\hat{\theta}_1+\hat{\theta}_i)-1} \geq \frac{1}{\exp(\hat{\theta}_1+\hat{\theta}_n)-1} \geq \frac{n}{2(\exp(M)-1)}.$$

However, this implies

$$\begin{aligned}\sqrt{\frac{16kn\log n}{\gamma L^2}} \geq \|\mathbf{d}-\mathbf{d}^*\|_\infty \geq |d_1 - d_1^*| &\geq -\sum_{j=2}^{n}\frac{1}{\exp(\theta_1+\theta_j)-1} + \sum_{j=2}^{n}\frac{1}{\exp(\hat{\theta}_1+\hat{\theta}_n)-1} \\ &\geq -\frac{(n-1)}{\exp(L)-1} + \frac{n(n-1)}{2(\exp(M)-1)},\end{aligned}$$

which cannot hold for sufficiently large n, as the right hand side in the last expression grows faster than the left hand side on the first line.

The analysis above shows that we have $\|\hat{\theta}\|_\infty < 2M$ for all sufficiently large n. Plugging in this result to (34) gives us

$$\|\theta - \hat{\theta}\|_\infty \leq \frac{2}{n}\,\frac{(\exp(5M)-1)^2}{\exp(5M)}\sqrt{\frac{16kn\log n}{\gamma L^2}} \leq \frac{8\ \exp(5M)}{L}\sqrt{\frac{k\log n}{\gamma n}}.$$

Finally, taking into account the issue of the existence of the MLE, we conclude that for sufficiently large n, with probability at least

$$\left(1-\frac{1}{n^{k-1}}\right)\left(1-\frac{2}{n^{k-1}}\right) \geq 1 - \frac{3}{n^{k-1}},$$

the MLE $\hat{\theta} \in \Theta$ exists and satisfies

$$\|\theta - \hat{\theta}\|_\infty \leq \frac{8\ \exp(5M)}{L}\sqrt{\frac{k\log n}{\gamma n}},$$

as desired. This finishes the proof of Theorem 3.13.

5 Discussion and future work

In this paper, we have studied the maximum entropy distribution on weighted graphs with a given expected degree sequence. In particular, we focused our study on three classes of weighted graphs: the finite discrete weighted graphs (with edge weights in the set $\{0, 1, \dots, r-1\}$, $r \geq 2$), the infinite discrete weighted graphs (with edge weights in the set $\mathbb{N}_0$), and the continuous weighted graphs (with edge weights in the set $\mathbb{R}_0$). We have shown that the maximum entropy distributions are characterized by the edge weights being independent random variables having exponential family distributions parameterized by the vertex potentials. We also studied the problem of finding the MLE of the vertex potentials, and we proved the remarkable consistency property of the MLE from only one graph sample.

In the case of finite discrete weighted graphs, we also provided a fast, iterative algorithm for finding the MLE with a geometric rate of convergence. Finding the MLE in the case of continuous or infinite discrete weighted graphs can be performed via standard gradient-based methods, and the bounds that we proved on the inverse Hessian of the log-partition function can also be used to provide a rate of convergence for these methods. However, it would be interesting if we can develop an efficient iterative algorithm for computing the MLE, similar to the case of finite discrete weighted graphs.

Another interesting research direction is to explore the theory of maximum entropy distributions when we impose additional structures on the underlying graph. We can start with an arbitrary graph G_0 on n vertices, for instance a lattice graph or a sparse graph, and consider the maximum entropy distributions on the subgraphs G of G_0. By choosing different types of the underlying graphs G_0, we can incorporate additional prior information from the specific applications we are considering.

Finally, given our initial motivation for this project, we would also like to apply the theory that we developed in this paper to applications in neuroscience, in particular, in modeling the early-stage computations that occur in the retina. There are also other problem domains where our theory are potentially useful, including applications in clustering, image segmentation, and modularity analysis.

References

[1] M. Abeles. *Local Cortical Circuits: An Electrophysiological Study.* Springer, 1982.

[2] D. H. Ackley, G. E. Hinton, and T. J. Sejnowski. *A learning algorithm for Boltzmann machines.* Cognitive Science, **9** (1985), 147–169.

[3] W. Bair and C. Koch. *Temporal precision of spike trains in extrastriate cortex of the behaving macaque monkey.* Neural Computation, **8**(6):1185–202, August 1996.

[4] M. Bethge and P. Berens. *Near-maximum entropy models for binary neural representations of natural images.* Advances in Neural Information Processing Systems **20**, 2008.

[5] W. Bialek, A. Cavagna, I. Giardina, T. Mora, E. Silvestri, M. Viale, and A. Walczak. *Statistical mechanics for natural flocks of birds.* Proceedings of the National Academy of Sciences, **109** (2012), 4786–4791.

[6] D. A. Butts, C. Weng, J. Jin, C. I. Yeh, N. A. Lesica, J. M. Alonso, and G. B. Stanley. *Temporal precision in the neural code and the timescales of natural vision.* Nature, **449**(7158):92–5, 2007.

[7] C. E. Carr. *Processing of temporal information in the brain.* Annual Review of Neuroscience, **16** (1993), 223–243.

[8] S. Chatterjee, P. Diaconis, and A. Sly. *Random graphs with a given degree sequence.* Annals of Applied Probability, **21** (2011), 1400–1435.

[9] S. A. Choudum. *A simple proof of the Erdős-Gallai theorem on graph sequences.* Bull. Austr. Math. Soc. 33:67–70, 1986.

[10] T. M. Cover and J. A. Thomas. *Elements of information theory.* Wiley-Interscience, 2006.

[11] G. Desbordes, J. Jin, C. Weng, N. A. Lesica, G. B. Stanley, and J. M. Alonso. *Timing precision in population coding of natural scenes in the early visual system.* PLoS biology, **6**(12), December 2008.

[12] P. Erdős and T. Gallai. *Graphs with prescribed degrees of vertices.* Mat. Lapok, **11** (1960), 264–274.

[13] C. Hillar, S. Lin, and A. Wibisono. *Inverses of symmetric, diagonally dominant positive matrices and applications.* http://arxiv.org/abs/1203.6812, 2012.

[14] W. Hoeffding. *Probability inequalities for sums of bounded random variables.* Journal of the American Statistical Association, **58** (1963), 13–30.

[15] J. J. Hopfield, *Neural networks and physical systems with emergent collective computational abilities.* Proceedings of the National Academy of Sciences, **79** (1982).

[16] J. J. Hopfield. *Pattern recognition computation using action potential timing for stimulus representation.* Nature, **376** (1995), 33–36.

[17] E. T. Jaynes. *Information theory and statistical mechanics.* Physical Review, **106** (1957).

[18] B. E. Kilavik, S. Roux, A. Ponce-Alvarez, J. Confais, S. Grün, and A. Riehle. *Long-term modifications in motor cortical dynamics induced by intensive practice.* The Journal of Neuroscience, **29** (2009), 12653–12663.

[19] S. Lang. *Real and Functional Analysis.* Springer, 1993.

[20] W. A. Little, *The existence of persistent states in the brain.* Mathematical Biosciences, **19** (1974), 101–120.

[21] R. C. Liu, S. Tzonev, S. Rebrik, and K. D. Miller. *Variability and information in a neural code of the cat lateral geniculate nucleus.* Journal of Neurophysiology, **86** (2001), 2789–2806.

[22] D. M. MacKay and W. S. McCulloch. *The limiting information capacity of a neuronal link.* Bulletin of Mathematical Biology, **14** (1952), 127–135.

[23] P. Maldonado, C. Babul, W. Singer, E. Rodriguez, D. Berger, and S. Grün. *Synchronization of neuronal responses in primary visual cortex of monkeys viewing natural images.* Journal of Neurophysiology, **100** (2008), 1523–1532.

[24] T. Mora, A. M. Walczak, W. Bialek, and C. G. Callan. *Maximum entropy models for antibody diversity.* Proceedings of the National Academy of Sciences, **107** (2010), 5405–5410.

[25] I. Nemenman, G. D. Lewen, W. Bialek, and R. R. R. van Steveninck. *Neural coding of natural stimuli: information at sub-millisecond resolution.* PLoS Computational Biology, **4** (2008).

[26] S. Neuenschwander and W. Singer. *Long-range synchronization of oscillatory light responses in the cat retina and lateral geniculate nucleus.* Nature, **379** (1996), 728–733.

[27] N. Proudfoot and D. Speyer. *A broken circuit ring.* Beiträge zur Algebra and Geometrie, **47** (2006), 161–166.

[28] F. Rieke, D. Warland, R. R. van Steveninck, and W. Bialek. *Spikes: Exploring the Neural Code.* MIT Press, 1999.

[29] W. Russ, D. Lowery, P. Mishra, M. Yae, and R. Ranganathan. *Natural-like function in artificial WW domains.* Nature, **437** (2005), 579–583.

[30] R. Sanyal, B. Sturmfels, and C. Vinzant. *The entropic discriminant.* Advances in Mathematics, 2013.

[31] E. Schneidman, M. J. Berry, R. Segev, and W. Bialek. *Weak pairwise correlations imply strongly correlated network states in a neural population.* Nature, **440** (2006), 1007–1012.

[32] C. Shannon, *The mathematical theory of communication*, Bell Syst. Tech. J., **27** (1948).

[33] J. Shlens, G. D. Field, J. L. Gauthier, M. I. Grivich, D. Petrusca, A. Sher, A. M. Litke, and E. J. Chichilnisky. *The structure of multi-neuron firing patterns in primate retina.* Journal of Neuroscience, **26**(32):8254–8266, 2006.

[34] J. Shlens, G. D. Field, J. L. Gauthier, M. Greschner, A. Sher, A. M. Litke, and E. J. Chichilnisky. *The structure of large-scale synchronized firing in primate retina.* Journal of Neuroscience, **29**(15):5022–5031, 2009.

[35] M. Socolich, S. Lockless, W. Russ, H. Lee, K. Gardner, and R. Ranganathan. *Evolutionary information for specifying a protein fold.* Nature, **437** (2005), 512–518.

[36] A. Tang, D. Jackson, J. Hobbs, W. Chen, J. Smith, and et al. *A maximum entropy model applied to spatial and temporal correlations from cortical networks in vitro.* Journal of Neuroscience, **28** (2008), 505–518.

[37] G. Tkacik, E. Schneidman, M. Berry, and W. Bialek. *Ising models for networks of real neurons.* http://arxiv.org/abs/q-bio/0611072, 2006.

[38] P. J. Uhlhaas, G. Pipa, G. B. Lima, L. Melloni, S. Neuenschwander, D. Nikolić, and W. Singer. *Neural synchrony in cortical networks: history, concept and current status.* Frontiers in Integrative Neuroscience, **3** (2009), 1–19.

[39] R. Vershynin. *Compressed sensing, theory and applications.* Cambridge University Press, 2012.

[40] J. D. Victor and K. P. Purpura. *Nature and precision of temporal coding in visual cortex: a metric-space analysis.* Journal of Neurophysiology, **76**(2):1310–1326, 1996.

[41] M. Wainwright and M. I. Jordan. *Graphical models, exponential families, and variational inference.* Foundations and Trends in Machine Learning, **1**(1–2):1–305, January 2008.

[42] S. Yu, D. Huang, W. Singer, and D. Nikolic. *A small world of neuronal synchrony.* Cerebral Cortex, **18** (2008), 2891–2901.