ON THE LANDSCAPE OF SYNCHRONIZATION NETWORKS: A PERSPECTIVE FROM NONCONVEX OPTIMIZATION*

SHUYANG LING†, RUITU XU†, AND AFONSO S. BANDEIRA†

Abstract. Studying the landscape of nonconvex cost functions is key towards a better understanding of optimization algorithms widely used in signal processing, statistics, and machine learning. Meanwhile, the famous Kuramoto model has been an important mathematical model for studying the synchronization phenomena of coupled oscillators over various network topologies. In this paper, we bring together these two seemingly unrelated objects by investigating the optimization landscape of a nonlinear function $E(\boldsymbol{\theta}) = \frac{1}{2}\sum_{1\le i,j\le n} a_{ij}(1 - \cos(\theta_i - \theta_j))$ associated with an underlying network and exploring the relationship between the existence of local minima and the network topology. This function arises naturally in the Burer–Monteiro method applied to $\mathbb{Z}_2$ synchronization as well as matrix completion on the torus. Moreover, it corresponds to the energy function of the homogeneous Kuramoto model on complex networks for coupled oscillators. We prove that the minimizer of the energy function is unique up to a global translation under deterministic dense graphs and Erdős–Rényi random graphs with tools from optimization and random matrix theory. Consequently, the stable equilibrium of the corresponding homogeneous Kuramoto model is unique and the basin of attraction for the synchronous state of these coupled oscillators is the whole phase space minus a set of measure zero. In addition, our results address when the Burer–Monteiro method recovers the ground truth exactly from highly incomplete observations in $\mathbb{Z}_2$ synchronization and shed light on the robustness of nonconvex optimization algorithms with respect to certain types of so-called monotone adversaries. Numerical simulations are performed to illustrate our results.

Key words. nonconvex optimization, optimization landscape, synchronization on networks, dynamical system, Kuramoto model

AMS subject classifications. 90C26, 90C35, 34D06, 35B35

DOI. 10.1137/18M1217644

1. Introduction. Nonconvex optimization plays a crucial role in the mathematics of signal processing [7, 13, 29, 42, 43, 44, 50], statistics [12, 33], and machine learning [14, 20, 21, 39, 45], and has attracted substantial attention in the past few years. Unlike convex optimization, nonconvex objective functions often exhibit complex geometric structures, making it essential to analyze the optimization landscape and to characterize the location and number of local optima and critical points, respectively.

This paper concerns the optimization landscape of the nonconvex function

$$(1.1) \qquad E(\boldsymbol{\theta}) := \frac{1}{2}\sum_{1\le i,j\le n} a_{ij}(1 - \cos(\theta_i - \theta_j)),$$

where $\boldsymbol{A} = (a_{ij})_{1\le i,j\le n}$ with $a_{ij} \ge 0$ is the weight matrix of an undirected graph. As a special case, $\boldsymbol{A}$ can be an adjacency matrix. We want to address the following key question.

> **Key question.** *What is the relationship between the existence of local minima and the topology of the network?*

One may wonder what motivates us to study the optimization landscape of $E(\boldsymbol{\theta})$ and why it is an important and valuable topic. Surprisingly, this seemingly simple function appears in numerous situations in mathematics, physics, engineering, and computer science. Despite tremendous effort, many mysteries behind this simple function still remain unsolved. We briefly introduce how this function emerges in various contexts in the next sections.

Before we move on, we first conduct a preliminary analysis of $E(\boldsymbol{\theta})$. Note that it suffices to consider $E(\boldsymbol{\theta})$ over $[0, 2\pi)^n$ due to the periodicity. A direct computation implies that $\boldsymbol{\theta}_0 := \mathbf{0}$ modulo a global translation, i.e., $\theta_i = \theta_j$ for all $i \neq j$, is a global minimizer of $E(\boldsymbol{\theta})$, independent of the underlying graph $\boldsymbol{A}$ (from now on, we use $\boldsymbol{A}$ to denote the weight (adjacency) matrix as well as the graph it represents). However, it is unclear whether there exist *spurious local minima* besides the global minimum achieved at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$. In particular, we say the optimization landscape of $E(\boldsymbol{\theta})$ is *benign* if there is no spurious local minimum at all.

1.1. Group synchronization, matrix completion, and monotone adversaries. Suppose there are n group elements $\{g_i\}_{i=1}^n$ of a group G. Instead of observing these elements directly, only a subset of their noisy offsets $g_i g_j^{-1}$ is available. The goal is to recover all the group elements up to a global phase from $\{g_i g_j^{-1}\}_{\{(i,j) \in \mathcal{E}\}}$, where $\mathcal{E}$ denotes the index set of available observations. This is called the *group synchronization* problem; see [1, 2, 6, 7, 8, 13, 35, 51] for more details.

The simplest example of group synchronization is $\mathbb{Z}_2$ synchronization [1], where g_i is assumed to belong to $\{\pm 1\}$ and thus the offsets reduce to $g_i g_j^{-1} = g_i g_j$. More precisely, we have $\boldsymbol{Y}$ as the observed data, which is the entrywise product of a binary symmetric matrix $\boldsymbol{A} \in \{0, 1\}^{n \times n}$ and the perturbed ground truth $\boldsymbol{g}\boldsymbol{g}^\top + \boldsymbol{W}$, i.e.,

$$\boldsymbol{Y} = \boldsymbol{A} \circ (\boldsymbol{g}\boldsymbol{g}^\top + \boldsymbol{W}),$$

where $\boldsymbol{W}$ is the noise matrix and "$\circ$" denotes the entrywise multiplication. Without the group constraints on $\{g_i\}_{i=1}^n$, it is exactly a matrix completion problem [11, 20, 36, 45], which originally comes from the *Netflix Prize* problem.

One possible formulation is to maximize the following combinatorial quadratic form

$$\max_{\boldsymbol{z} \in \{\pm 1\}^n} \boldsymbol{z}^\top \boldsymbol{Y} \boldsymbol{z}$$

and then use the maximizer as an estimator of $\boldsymbol{g}$ from its partial noisy measurements. In fact, the program above is equivalent to the least squares loss function after a simple transformation. However, solving this program is NP-hard in general. One may consider semidefinite programming (SDP) relaxation as an alternative by trying to recover the rank-1 matrix $\boldsymbol{g}\boldsymbol{g}^\top$ instead of individual group elements:

$$\max \text{Tr}(\boldsymbol{Y}\boldsymbol{Z}) \quad \text{s.t.} \ \boldsymbol{Z} \succeq 0, \ Z_{ii} = 1 \ \forall \, 1 \leq i \leq n.$$

Note the ground truth $\boldsymbol{g}\boldsymbol{g}^\top$ satisfies the two constraints above. Under mild conditions, this SDP relaxation is proven to recover $\boldsymbol{g}\boldsymbol{g}^\top$ exactly [1]. Despite the effectiveness of SDP relaxation [37], it tends to suffer from a high computational complexity. Hence, numerous efforts have been made in order to find efficient and robust gradient-descent-based methods as alternatives to solving the SDP arising from the relaxation of the low-rank matrix recovery problem.

The approach proposed by Burer and Monteiro in [9, 10] may arguably be one of the most successful methods for tackling these otherwise difficult problems by

conveniently keeping $\boldsymbol{Z}$ in a factorized form and taking full advantage of the low-rank property. By letting $\boldsymbol{Z} = \boldsymbol{Q}\boldsymbol{Q}^\top$, where $\boldsymbol{Q} \in \mathbb{R}^{n \times q}$ with $q < n$, we have

$$(1.2) \qquad \max \ \mathrm{Tr}(\boldsymbol{Y}\boldsymbol{Q}\boldsymbol{Q}^\top) \quad \text{s.t.} \ (\boldsymbol{Q}\boldsymbol{Q}^\top)_{ii} = 1 \ \forall \, 1 \le i \le n.$$

In other words, each row of $\boldsymbol{Q}$ needs to be normalized. These constraints yield a search space consisting of a Riemannian manifold, and Riemannian gradient descent is used to solve this program [3, 7].

In particular, $E(\boldsymbol{\theta})$ is equivalent to the Burer–Monteiro method [7, 9, 10] applied to $\mathbb{Z}_2$ synchronization on networks when $q = 2$. Without loss of generality, we assume $\boldsymbol{g} = \mathbf{1}_n$ by considering $\mathrm{diag}(\boldsymbol{g})\boldsymbol{Y}\,\mathrm{diag}(\boldsymbol{g})$ instead of $\boldsymbol{Y}$. The ith row of $\boldsymbol{Q}$ as $[\cos\theta_i, \ \sin\theta_i]$ gives $(\boldsymbol{Q}\boldsymbol{Q}^\top)_{ij} = \cos(\theta_i - \theta_j)$. Thus we have an equivalent form as

$$(1.3) \qquad \min_{\boldsymbol{\theta} \in \mathbb{R}^n} \quad \sum_{i,j} a_{ij}(1 + w_{ij})(1 - \cos(\theta_i - \theta_j)),$$

which is in the exact form of $E(\boldsymbol{\theta})$.

Note that the Burer–Monteiro method yields a nonconvex cost function. This fact immediately brings up a number of questions: when does this Burer–Monteiro approach recover the ground truth? Does (1.3) have spurious local minima? How does the existence of spurious local minima depend on the topology of the network? All these problems reduce to the analysis of the optimization landscape of $E(\boldsymbol{\theta})$ in (1.1) under different network topologies.

The answers to these aforementioned questions are closely related to the robustness of algorithms with respect to *monotone adversaries*, an increasingly important topic in algorithm design in theoretic computer science. For simplicity, we assume no noise in the measurements, i.e., $w_{ij} = 0$. Suppose we collect more and more noiseless observations. It is not hard to imagine that more noiseless data help us to recover the underlying signal $\boldsymbol{g}$. Moreover, it can be easily proven that the SDP relaxation works perfectly in this scenario. On the other hand, the Burer–Monteiro approach exhibits satisfactory performance in practice and enjoys much higher efficiency than the SDP relaxation does. However, it is unclear whether the Burer–Monteiro approach would inherit the robustness to such seemingly helpful "noise" from the SDP relaxation. This type of "noise" is often referred to as a *monotone adversary*. Note that more observations mean adding more edges to the current network and thus we propose the following question:

> *Does adding more edges to the network improve the optimization landscape or worsen it by creating spurious local optima?*

We defer a more detailed discussion regarding monotone adversaries to section 2.3.

1.2. The Kuramoto model and synchronization on complex networks. The synchronization problem of coupled oscillators has a long history. It was originally proposed and studied by Christiaan Huygens in 1665 when investigating the behavior of two pendulum clocks mounted side by side on the same support. He observed that the two oscillators would always end up swinging in exactly opposite directions, independent of their respective motion [34, 5].

Since then, the synchronization of coupled oscillators has fascinated the scientific community, and remarkable progress has been made regarding this topic. One of the breakthroughs was made by Kuramoto [27, 28] who came up with a mathematically tractable model to describe the synchronization behavior for a large set of coupled oscillators. The general Kuramoto model assumes n oscillators which interact with

one another based on sinusoidal coupling, i.e.,

$$(1.4) \qquad \frac{\mathrm{d}\theta_i}{\mathrm{d}t} = \omega_i - \sum_{j=1}^{n} a_{ij} \sin(\theta_i - \theta_j), \quad 1 \le i \le n,$$

where each $\theta_i \in [0, 2\pi)$ is a function of time t. Here ω_i refers to the natural frequency of the ith oscillator and a_{ij} stands for the strength of the mutual coupling. The original work by Kuramoto [27] required the weights $\{a_{ij}\}$ to be identical and showed the oscillators $\{\theta_i\}_{i=1}^{n}$ would synchronize if the mutual interactions are stronger than the effect of natural frequencies $\{\omega_i\}$, as discussed in [27, 28]. In particular, the local asymptotic stability of (1.4) with general network topologies and natural frequencies was first studied in [24]. It is also worth noting that the Kuramoto model appears in many applications such as electric power networks, neuroscience, chemical oscillations, and spin glasses; see [4, 5, 17, 18, 38] and the references therein for more details.
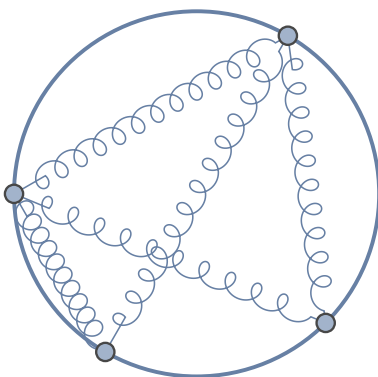

FIG. 1. *Illustration of coupled oscillators.*

The Kuramoto model was later studied on complex networks [24, 18, 4, 38] where $a_{ij} \ge 0$ describes how much the ith and the jth oscillator are coupled. In particular, if $\boldsymbol{A}$ is an adjacency matrix, it means two oscillators i and j are coupled if $a_{ij} = 1$; see Figure 1.

Note that the synchronous state is highly related to the stable equilibrium of continuous time dynamical systems (1.4). On the other hand, [18] points out that the coupled oscillators model is a system of particles which tends to minimize the energy function

$$E_\omega(\boldsymbol{\theta}) = \sum_{i,j} a_{ij}(1 - \cos(\theta_i - \theta_j)) - \sum_{i=1}^{n} \omega_i \cdot \theta_i,$$

where the two terms above model the particle interaction and the external force, respectively. If we consider a standard Cauchy problem associated with $E_\omega(\boldsymbol{\theta})$ by setting the velocity of each oscillator to the gradient of $-E_\omega(\boldsymbol{\theta})$, i.e.,

$$\frac{\mathrm{d}\theta_i(t)}{\mathrm{d}t} = -\frac{\partial E_\omega(\boldsymbol{\theta})}{\partial \theta_i}, \quad 1 \le i \le n, \quad \theta_i = \theta_i(0),$$

or equivalently,

$$(1.5) \qquad \frac{\mathrm{d}\boldsymbol{\theta}(t)}{\mathrm{d}t} = -\nabla E_\omega(\boldsymbol{\theta}), \quad \boldsymbol{\theta} = \boldsymbol{\theta}(0),$$

then the evolution is always moving in the direction such that the energy $E_\omega(\boldsymbol{\theta})$ decreases maximally, which is referred to as a gradient flow. Discretizing (1.5) via an Euler scheme corresponds to gradient descent applied to the energy function $E_\omega(\theta)$.

It is natural to ask what the behavior of $\boldsymbol{\theta}$ is as time evolves. In fact each stable equilibrium corresponds to a local minimizer of the energy function $E_\omega(\boldsymbol{\theta})$ whose existence is the main focus of this work. From this point on, we consider a simplified version, the homogeneous Kuramoto model [27, 40, 46], where the natural frequencies $\{\omega_i\}_{i=1}^n$ are 0. Under this assumption, $E_\omega(\boldsymbol{\theta})$ reduces to $E(\boldsymbol{\theta})$ and thus it suffices to understand the local minima of $E(\boldsymbol{\theta})$ in order to characterize the synchronization of the homogeneous Kuramoto model on networks.

1.3. Related work and our contributions. The landscape of nonconvex loss functions has drawn extensive attention recently. Examples range from signal processing to statistics and machine learning, including phase retrieval [42, 44], matrix completion [20], blind deconvolution [50], community detection and synchronization [7], kernel PCA [12], low-rank matrix sensing problems [29], tensor decomposition [21], dictionary learning [43], and neural networks [39]. Among all those existing works, ours shares more commonalities with group synchronization and matrix completion.

Synchronization problems, which aim to recover group elements g_i from their relative alignment $g_i g_j^{-1}$ (or noisy alignment), are considered on different groups such as phase synchronization [6, 51] (corresponding to the group $U(1)$), $\mathbb{Z}_2$ synchronization [1, 7], joint alignment from pairwise differences [13] (corresponding to the cyclic group, i.e., $\mathbb{Z}/n\mathbb{Z}$), and general group synchronization problems [35, 2]. Many algorithms, some based on convex and nonconvex optimization, have been proposed to tackle these problems and have proven to work well on group elements retrieval. In these aforementioned works, we are mostly influenced by $\mathbb{Z}_2$ and phase synchronization [6, 51, 1, 7, 13], especially [7].

The convex relaxation approach for $\mathbb{Z}_2$ and phase synchronization was initially studied in [6, 1]. The proposed method was shown to be tight, i.e., it recovers group elements exactly under mild conditions. Later, [8, 13, 51] showed that one can retrieve group elements via a two-step nonconvex optimization procedure, i.e., a carefully chosen initialization followed by (projected) power iteration. Compared with convex relaxation, iterative methods based on gradient descent and power methods achieve much higher efficiency while still enjoying rigorous theoretic guarantees. However, discussion on the landscape of the corresponding cost function is not covered [51, 13]. In [7], the authors studied the Burer–Monteiro method applied to $\mathbb{Z}_2$ synchronization and obtained an objective function in the form of $E(\boldsymbol{\theta})$ (modulo a proper transformation). The resulting function is shown to enjoy a benign optimization landscape by taking advantage of the randomness in the weight matrix. The work provides a strategy for analyzing the landscape of the energy function via an effective perturbation argument. Unlike [7], our work concerns the properties of landscapes of the energy function associated with an unweighted network and how the landscape depends on the network topology, which is a vastly different setting.

Our work is also closely related to matrix completion [20, 11, 45], where one wants to recover a low-rank matrix from partial observations. The landscape of the least squares loss function associated with matrix completion was proven to be benign in [20]; fast and efficient nonconvex algorithms were developed in [26, 45] to solve the matrix completion problem. The main distinction of our work lies in the fact

that the ground truth signals have entrywise unit modulus. In fact, the additional unit modulus constraints lead to a more complicated landscape since they potentially create many critical points and local optima, thus making the analysis different.

Another direction of relevant research comes from dynamical systems and synchronization on complex networks [41]. One prominent example is the Kuramoto model [27], which models the collective behavior of a large population of oscillators which interact with one another via sinusoidal coupling [17, 18, 38, 40, 49, 24]. We gladly acknowledge being influenced by the excellent reviews in this field [18, 17, 38]. For the Kuramoto model, the core problem is to find and analyze the equilibria, which means solving a nonlinear system of equations [32] where the stable equilibria correspond to the local minima of the energy function. The work [40] introduces a class of illuminating examples of networks called Wiley–Strogatz–Girvan networks, and the corresponding energy functions of the coupled oscillators yield spurious local minima. These examples indicate that there may exist local minima even if the graph is well connected. The work by Taylor [46] on homogeneous Kuramoto model suggests exploring the existence of local minima for deterministic dense graphs and plays a role in guiding our theoretic analysis. Several recent works [15, 23, 30, 31] have studied the Kuramoto model on random graphs via mean field analysis; they do not overlap with the current work since we consider the landscape of the energy function with finite n instead of its continuum limit and asymptotic behavior when $n \to \infty$. Another distinct feature of our work is the analysis of the global landscape instead of the local optimization landscape restricted on the cohesive phases, i.e., $\{\boldsymbol{\theta} : |\theta_i - \theta_j| < \pi/2\}$, as seen in [17, 18].

The contributions of this work are multifold. We give a precise characterization of the connection between the function landscape and the associated network topology. We consider two important types of graphs, dense deterministic graphs and Erdős–Rényi random graphs, and prove the resulting function landscape enjoys no other local minima except the synchronized state $\boldsymbol{\theta} = \boldsymbol{\theta}_0$. In other words, we prove that the homogeneous Kuramoto model has only one stable synchronized solution and the basin of synchronization is the whole phase space with high probability if the underlying graph is a *finite* Erdős–Rényi graph. We also prove that if each node has at least $(3-\sqrt{2})(n-1)/2 \approx 0.7929(n-1)$ neighbors, no spurious local minima exist, which improves the state-of-the-art bound $\mu \geq 0.9375(n-1)$ in [46] by Taylor. Moreover, the theoretic analysis provides explicit examples under which the Burer–Monteiro method is not robust against monotone adversaries. The techniques involved in this manuscript share very little similarities with those used in the existing literature on the Kuramoto model on complex networks. Thus our work likely offers new inspirations for solving Kuramoto models on other types of random graphs. In conclusion, our work not only provides solutions to a few important questions about synchronization problems on networks but also motivates a series of open problems.

1.4. Organization of our paper. This paper is organized as follows. In section 2 we go through the basics of $E(\boldsymbol{\theta})$ and introduce a few examples. Our main theorems and open problems are presented in section 3. Section 4 is devoted to numerical experiments that complement the theoretical analysis. Finally, the proofs of our results can be found in section 5.

1.5. Notation. We introduce notation which will be used throughout the paper. Matrices are denoted by boldface uppercase letters, e.g., $\boldsymbol{Z}$; vectors are denoted by boldface lowercase letters, e.g., $\boldsymbol{z}$. The individual entries of a matrix or a vector are denoted in normal font such as z_{ij} or z_i. For any matrix $\boldsymbol{Z}$, $\|\boldsymbol{Z}\|$ denotes its operator

norm and $\|\boldsymbol{Z}\|_F$ denotes its the Frobenius norm. For any vector $\boldsymbol{z}$, $\|\boldsymbol{z}\|$ denotes its Euclidean norm; $\|\boldsymbol{z}\|_\infty$ stands for the ℓ_∞ norm. For both matrices and vectors, $\boldsymbol{Z}^\top$ and $\boldsymbol{z}^\top$ stand for the transposes of $\boldsymbol{Z}$ and $\boldsymbol{z}$, respectively. We equip the matrix space $\mathbb{R}^{K \times N}$ with the inner product defined by $\langle \boldsymbol{U}, \boldsymbol{V} \rangle := \mathrm{Tr}(\boldsymbol{U}^\top \boldsymbol{V})$. A special case is the inner product of two vectors, i.e., $\langle \boldsymbol{u}, \boldsymbol{v} \rangle = \mathrm{Tr}(\boldsymbol{u}^\top \boldsymbol{v}) = \boldsymbol{u}^\top \boldsymbol{v}$. For a given vector $\boldsymbol{v}$, $\mathrm{diag}(\boldsymbol{v})$ represents the diagonal matrix whose diagonal entries are given by the vector $\boldsymbol{v}$, and $\mathrm{ddiag}(\boldsymbol{Z})$ denotes a diagonal matrix whose diagonal entries are the same as those of $\boldsymbol{Z}$. $\boldsymbol{I}_n$ and $\boldsymbol{1}_n$ always denote the $n \times n$ identity matrix and a column vector of "1" in $\mathbb{R}^n$, respectively; $\boldsymbol{J}_n$ represents $\boldsymbol{1}_n \boldsymbol{1}_n^\top$; $\boldsymbol{e}_i \in \mathbb{R}^n$ denotes a vector with a 1 in the ith coordinate and 0's elsewhere. For any $\boldsymbol{U}$ and $\boldsymbol{V}$ in $\mathbb{R}^{K \times N}$, their Hadamard product is denoted as $\boldsymbol{U} \circ \boldsymbol{V}$. For any $\boldsymbol{Z} \in \mathbb{R}^{N \times N}$, $\boldsymbol{Z} \succeq 0$ if $\boldsymbol{Z}$ is symmetric and positive semidefinite.

2. Preliminaries. This section is devoted to providing basic properties of $E(\boldsymbol{\theta})$ and some concrete examples. The examples illustrate the connection between the energy landscape and network topology, motivating our theoretical and numerical investigations. Moreover, they address the robustness issue of the Burer–Monteiro approach with respect to monotone adversaries.

2.1. Problem setup. We start with $E(\boldsymbol{\theta})$ and write it in a more compact form. Let

$$(2.1) \qquad \boldsymbol{Q} := [\boldsymbol{x} \ \boldsymbol{y}] \in \mathbb{R}^{n \times 2},$$

where $\boldsymbol{x} := [\cos \theta_1, \ldots, \cos \theta_n]^\top \in \mathbb{R}^n$ and $\boldsymbol{y} := [\sin \theta_1, \ldots, \sin \theta_n]^\top \in \mathbb{R}^n$. Then we have $(\boldsymbol{Q}\boldsymbol{Q}^\top)_{ij} = x_i x_j + y_i y_j = \cos(\theta_i - \theta_j)$, where x_i and y_i are the ith entry of $\boldsymbol{x}$ and $\boldsymbol{y}$, respectively. Hence $E(\boldsymbol{\theta})$ can be recast in terms of $\boldsymbol{Q}$:

$$E(\boldsymbol{\theta}) = \frac{1}{2}\langle \boldsymbol{A}, \boldsymbol{J}_n - \boldsymbol{Q}\boldsymbol{Q}^\top \rangle.$$

To study the behavior of local minima, it is essential to obtain the first- and second-order necessary conditions. Both conditions are easy to obtain via direct computation:

$$(2.2) \qquad (\nabla E(\boldsymbol{\theta}))_i = \sum_{j=1}^{n} a_{ij} \sin(\theta_i - \theta_j) = 0.$$

Thanks to a simple trigonometric identity $\sin(\theta_i - \theta_j) = \sin\theta_i \cos\theta_j - \cos\theta_i \sin\theta_j$, we have

$$\boldsymbol{A}\boldsymbol{x} \circ \boldsymbol{y} = \boldsymbol{A}\boldsymbol{y} \circ \boldsymbol{x}.$$

Now we proceed to compute the Hessian matrix of $E(\boldsymbol{\theta})$:

$$(2.3) \qquad (\nabla^2 E(\boldsymbol{\theta}))_{ij} = \begin{cases} \dfrac{\partial^2 E}{\partial \theta_i \partial \theta_j} = -a_{ij} \cos(\theta_i - \theta_j) & \forall i \neq j, \\ \dfrac{\partial^2 E}{\partial \theta_i^2} = \displaystyle\sum_{j \neq i} a_{ij} \cos(\theta_i - \theta_j) & \forall i = j. \end{cases}$$

In fact, the first- and second-order derivatives above are exactly the Riemannian gradient and Hessian of (1.2) with a proper trigonometric transformation since one can think of (1.2) as an optimization problem on an n-dimensional torus. Interested

readers are referred to [25, 3, 7] for more details about how to solve optimization problems on a manifold.

The Hessian matrix $\nabla^2 E(\boldsymbol{\theta})$ can be regarded as the graph Laplacian associated with the weight matrix $\{a_{ij}(\cos(\theta_i - \theta_j))\}_{1 \leq i,j \leq n}$ if we allow negative weights. As a result, the constant vector $\mathbf{1}_n$ is in the null space of $\nabla^2 E(\boldsymbol{\theta})$. In particular, when $\boldsymbol{\theta}$ is a local minimizer of $E(\boldsymbol{\theta})$, we have $\nabla^2 E(\boldsymbol{\theta}) \succeq 0$. This necessary condition has the following equivalent form:

$$(2.4) \qquad \nabla^2 E(\boldsymbol{\theta}) = \mathrm{ddiag}(\boldsymbol{A}\boldsymbol{Q}\boldsymbol{Q}^\top) - \boldsymbol{A} \circ \boldsymbol{Q}\boldsymbol{Q}^\top \succeq 0.$$

Following from (2.4), we know that if $\{\theta_i\}_{i=1}^n$ simultaneously stay inside a single quadrant, $\nabla^2 E(\boldsymbol{\theta}) \succeq 0$ and $E(\boldsymbol{\theta})$ is convex. This is because $\nabla^2 E(\boldsymbol{\theta})$ is the graph Laplacian associated with the weight given by $a_{ij} \cos(\theta_i - \theta_j) \geq 0$ if $|\theta_i - \theta_j| < \pi/2$. In particular, if the graph is connected, then $E(\boldsymbol{\theta})$ is strictly convex within $\{\boldsymbol{\theta} : |\theta_i - \theta_j| < \pi/2\}$. Therefore, local convexity holds in a neighborhood around the global minimizer $\boldsymbol{\theta}_0 = \mathbf{0}$.

However, the global landscape is more complex. In fact, $E(\boldsymbol{\theta})$ has at least 2^n distinct critical points in the form of $\{\boldsymbol{\theta} \in \mathbb{R}^n : \theta_i = 0 \text{ or } \pi\}$ up to a global rotation. Moreover, other critical points and local optima may coexist while not being obtained analytically. This adds difficulties to the landscape analysis of $E(\boldsymbol{\theta})$.

2.2. Examples. We introduce a few examples to illustrate the complex relationship between the existence of spurious local minima of $E(\boldsymbol{\theta})$ and the network represented by $\boldsymbol{A}$. First, it is natural to ask whether there are no spurious local minima at all if $\boldsymbol{A}$ is well connected. Note that the Hessian matrix at $\boldsymbol{\theta}_0 = \mathbf{0}$ satisfies

$$\nabla^2 E(\boldsymbol{\theta}_0) = \mathrm{ddiag}(\boldsymbol{A}\boldsymbol{J}_n) - \boldsymbol{A} = \mathrm{diag}(\boldsymbol{A}\mathbf{1}_n) - \boldsymbol{A},$$

which corresponds exactly to the graph Laplacian associated with the adjacency matrix $\boldsymbol{A}$. When adding more edges to the graph, i.e., putting more nonzero entries into $\boldsymbol{A}$, the second smallest eigenvalue of $\nabla^2 E(\boldsymbol{\theta})$ would increase and thus strengthen the local convexity at $\boldsymbol{\theta}_0$. However, there are examples indicating that the local curvature at $\boldsymbol{\theta}_0$ *fails* to determine the properties of the global landscape we are interested in. We defer the technical details behind these examples to section 5.3.

The path graph of n nodes. We start with a simple example where the graph is a path of n vertices; see Figure 2. For such a path graph, we claim there are exactly 2^n critical points represented by $\{\boldsymbol{\theta} \in \mathbb{R}^n : \theta_i = 0 \text{ or } \pi\}$ but the only local minimizer is $\boldsymbol{\theta} = \boldsymbol{\theta}_0$. More precisely, the adjacency matrix of a path satisfies

$$a_{ij} = \begin{cases} 1 & \text{if } |j - i| = 1, \\ 0 & \text{otherwise.} \end{cases}$$

By substituting a_{ij} into (2.2), solving for critical points, and checking the second-order necessary condition, we arrive at our claim. In fact, if the graph is a connected tree (i.e., a connected graph without loop), one can draw a similar conclusion.

Wiley–Strogatz–Girvan networks. Wiley–Strogatz–Girvan networks [49] are constructed by linking each node on a cycle with its k nearest neighbors; see Figure 3. Obviously, the larger k is, the better the graph is connected. On the other hand, for such a graph, at least one local minimum besides $\boldsymbol{\theta}_0$ can be constructed explicitly if $k \leq 0.34n$, i.e., each node has at most $0.68n$ neighbors. More precisely, we can verify that $\boldsymbol{\theta} = 2\pi n^{-1}[0, 1, \ldots, n - 1]^\top$ is a local minimizer (also known as a uniformly
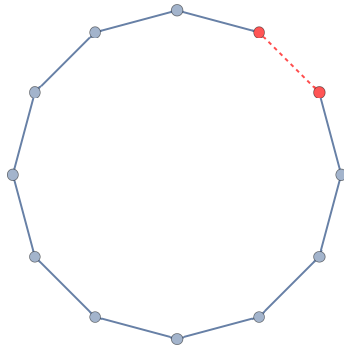
FIG. 2. *A path graph with n nodes and an n-cycle. While the energy function associated with a path graph has the unique minimum $\boldsymbol{\theta}_0$, it has at least one spurious local minimum if the underlying network is an n-cycle.*
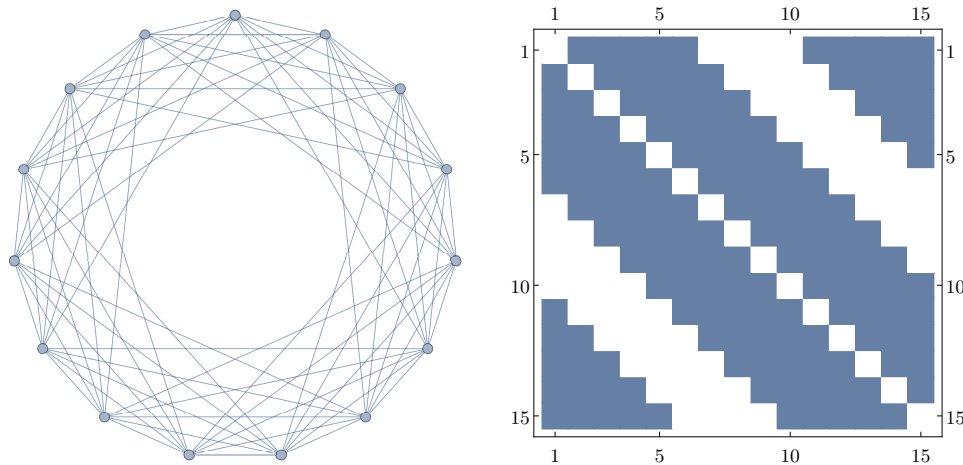

FIG. 3. *A Wiley–Strogatz–Girvan network* [49] *and its adjacency matrix.*

twisted state; see [49]), as shown in Figure 4, and we provide a short derivation via the discrete Fourier transform in section 5.3.

In conclusion, one can always find an explicit graph whose nodes are of degree as large as $0.68n$ such that at least one spurious local minimum exists. It is worth noting that if the underlying graph is complete, there is no local minimum [46]. Therefore, we want to find out if there exists a critical value for the degree which prevents the existence of spurious local minima.

2.3. Monotone adversary. We now return to the Burer–Monteiro approach when applied to $\mathbb{Z}_2$ synchronization and look into its robustness with respect to the *monotone adversary*. Here the monotone adversary means adding noise that seemingly helps us to solve the problem by providing more information. The motivation behind this is to design efficient and reliable algorithms to tackle difficult optimization problems *without* implicitly taking advantage of the statistical properties of noise. Therefore, the study of monotone adversaries becomes an active research topic in algorithm designs in theoretic computer science [19] and has been discussed in several contexts such as community detection [33] and matrix completion [14].
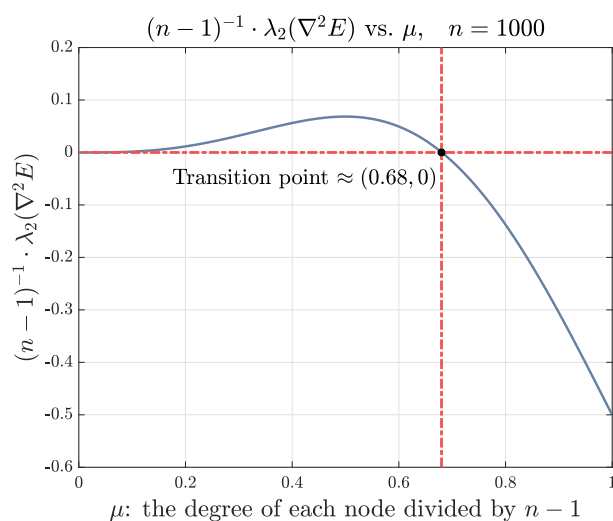
FIG. 4. *The rescaled second smallest eigenvalue of the Hessian matrix* $(n-1)^{-1}\lambda_2(\nabla^2 E(\boldsymbol{\theta}))$ *with respect to* μ *at the twisted state for a Wiley–Strogatz–Girvan network.*

For $\mathbb{Z}_2$ synchronization, let us investigate the monotone adversary for a simplified model in (1.3) with $\boldsymbol{W} = 0$ and study the landscape of the corresponding objective function associated with a given graph $\boldsymbol{A}$. In this case, the objective function is exactly equal to $E(\boldsymbol{\theta})$ and the monotone adversary is equivalent to adding more edges to the network. In other words, we have more observed data by doing that, and it should be helpful for the retrieval of underlying information. Note that SDP relaxation is robust to this type of adversary as more information is provided. However, it is unclear whether nonconvex optimization methods such as the Burer–Monteiro approach would enjoy this type of robustness property.

Unfortunately, this is not the case in general as we have already seen explicit examples where adding more edges may *worsen* the landscape of $E(\boldsymbol{\theta})$ by introducing undesired spurious local minima. Let $\boldsymbol{A}$ be the adjacency matrix of the n-path graph, and there is no local minimum at all. Suppose we add one more edge to make it into an n-cycle graph. The resulting graph becomes an example of the Wiley–Strogatz–Girvan networks and it is known that the twisted state leads to a spurious local minimum.

3. Main results and open problems. The connection between the optimization landscape and general network topology is far from being well understood. From now on, we consider two important types of graphs and the optimization landscape of their corresponding energy function $E(\boldsymbol{\theta})$.

1. *Deterministic dense graphs.* Concrete examples from the Wiley–Strogatz–Girvan networks are constructed showing the existence of spurious local minima even if the graph is highly connected. On the other hand, we note that if the underlying graph is complete, i.e., $a_{ij} = 1$, no spurious local minimum exists. Therefore, one may immediately wonder how the optimization landscape looks like if the graph is sufficiently dense, i.e., each vertex has at least $\mu(n-1)$ neighbors for μ close to 1. We refer to this as the worst-case scenario.

2. *Erdős–Rényi random graphs.* While the deterministic dense graphs correspond to the worst-case analysis, is it possible that the landscape of $E(\boldsymbol{\theta})$ is benign

on average? For this purpose, we consider the underlying graph as Erdős–Rényi random graphs $G(n,p)$ whose adjacency matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ satisfies $a_{ij} = a_{ji}$ and

$$(3.1) \qquad a_{ij} = \begin{cases} 1 & \text{with probability } p, \\ 0 & \text{with probability } 1 - p \end{cases}$$

for certain constant $p \in [0, 1]$. This is called the average-case scenario.

Our first theorem states that if the graph is sufficiently dense, then the corresponding energy function has a benign optimization landscape, i.e., all local minima are global and attained at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$.

THEOREM 3.1. *Given a graph with n vertices, suppose the degree of each vertex is at least $\mu(n-1)$ with*

$$\mu \geq \frac{3 - \sqrt{2}}{2}.$$

Then the only local minimum of the energy function $E(\boldsymbol{\theta})$ is attained when

$$\boldsymbol{\theta} = \boldsymbol{\theta}_0,$$

which is unique modulo a global translation.

The result above significantly improves on a previous result by Taylor [46] where the optimization landscape is proven benign if $\mu \geq 0.9395$. Combined with the existing facts from the Wiley–Strogatz–Girvan networks [49], we have a more complete characterization regarding the existence of local minima with respect to the degree of each vertex: the smallest possible μ is between 0.6809 and $(3 - \sqrt{2})/2 \approx 0.7929$ such that no spurious local minima exists if all nodes have at least $\mu(n-1)$ neighbors. Based on all the results obtained so far, we propose the following open problem.

Open Problem 3.2. Does there exist a critical constant μ_c such that, where each node has at least $\mu(n-1)$ neighbors,
- if $\mu < \mu_c$, there are other local minima besides $\boldsymbol{\theta} = \boldsymbol{\theta}_0$;
- if $\mu > \mu_c$, the only local minimum is achieved at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ (modulo a global translation)?

While the deterministic dense graphs yield the worst-case analysis, we present an average-case result with the help of randomness.

THEOREM 3.3. *Given an Erdős–Rényi graph $G(n,p)$, suppose $p = 32\gamma n^{-1/3} \log n$ with $\gamma \geq 1$. Then the only local minimum of the energy function $E(\boldsymbol{\theta})$ is attained at*

$$\boldsymbol{\theta} = \boldsymbol{\theta}_0$$

with probability at least

$$1 - 4 \exp\left(n \left(\log(100 n^{\frac{1}{3}}) - 2\gamma \log n\right)\right) - 10 n^{-\gamma + 1}.$$

This theorem says that for an Erdős–Rényi graph $G(n,p)$ with $p = \Omega(n^{-1/3} \log n)$, i.e., each node is of degree about $\Omega(n^{2/3} \log n)$, the optimization landscape has no spurious local minima with high probability. In other words, for the homogeneous Kuramoto model whose underlying network is an Erdős–Rényi graph, all the oscillators will synchronize and the basin of attraction is the whole phase space minus a set of measure zero with high probability. Note that our analysis does not rely on

$n \to \infty$, and thus provides a nonasymptotic characterization of the landscape unlike the existing works [15, 23, 30, 31] on the continuum limit of Kuramoto model.

As pointed out previously, the energy function can also be derived from the least squares loss for matrix completion restricted to the torus. Without this additional constraint, [20] showed that the loss function has no spurious local minima with high probability if $p = \Omega(n^{-1}\log n)$. Our theorem loses a factor of $n^{2/3}$ and requires $p = \mathcal{O}(n^{-1/3}\log n)$, mainly due to the constraint of entrywise unit modulus. However, the numerical simulations indicate that the loss of the $n^{2/3}$ factor may well be an artifact of proof, as shown in section 4. We will point out in the proof where our analysis leads to this potentially suboptimal bound. As a result, we also propose a second open problem.

Open Problem 3.4. Is it true that no spurious local minimum exists with high probability if $p = \Omega(n^{-1}\log n)$?

Note that for an Erdős–Rényi graph, the graph is connected with high probability if $p > n^{-1}\log n$. Therefore, our conjecture claims that as long as the Erdős–Rényi graph is connected, the corresponding optimization landscape should be benign with high probability.

Besides the two aforementioned open problems, our work also motivates a series of mathematical problems, the solutions to which would significantly contribute to the development of nonconvex optimization as well as the understanding of the synchronization phenomena of coupled oscillators on complex networks. While our work concerns the homogeneous Kuramoto model, many mysteries behind the inhomogeneous Kuramoto models remain to be solved where the natural frequencies $\{\omega_i\}_{i=1}^n$ in (1.4) are nonzero. It is also interesting to look into the Kuramoto model over other random graphs and graphs with special structures with tools from the recent progress in the nonconvex optimization. The study of the optimization landscape under the semirandom model (cf. [14]) is particularly relevant to the robustness of algorithms against the monotone adversary. Other future directions include finding out what kinds of graph properties will determine the optimization landscape of the energy function. As discussed already, the second smallest eigenvalue of the associated graph Laplacian may not be a perfect candidate because the spurious local minima may still exist even if the graph is highly connected. Thus it is especially appealing to find a single quantity associated with the underlying network which is able to characterize the optimization landscape.

4. Numerics. In this section, we present some experimental results on the landscape of $E(\boldsymbol{\theta})$ with the Erdős–Rényi graph as the underlying network. We consider systems with small and large size, i.e., $n = 5, 10, \ldots, 100$ and $n = 100, 150, \ldots, 1500$, respectively. For each system, we run gradient descent on the energy function $E(\boldsymbol{\theta})$ with random initialization, which is equivalent to a discrete approximation of the Cauchy problem in (1.5). For the systems with smaller size, we vary p between 0 and 1, and for each pair of (n, p) we take 50 independent random instances; for the systems with larger size, we vary p between 0 and $4n^{-1}\log n$ for each n and also perform 50 experiments for each set of parameters. The gradient descent method is implemented with a fixed step size of 0.005 and the maximum number of iterations of 1000 per round; in each round the local search stops when $\|\nabla E(\boldsymbol{\theta})\| \leq 10^{-8}$. The shade of grey in the square represents the fraction of instances for which gradient descent reaches its global minimum. The red line represents the phase transition threshold we conjecture in Open Problem 3.4, i.e., $p = n^{-1}\log n$.

In Figure 5 (see online version for color figures) we observe that the gradient descent with random initialization finds the global minimum with high probability as soon as p becomes slightly greater than $n^{-1} \log n$, i.e., the graph is connected. Figure 6 delivers a similar message that the gradient descent does not converge to global minimum with high probability when $p = n^{-1} \log n$. On the other hand, when p is slightly larger than $n^{-1} \log n$, i.e., $p = 2n^{-1} \log n$, the iterates from gradient descent converge to $\boldsymbol{\theta}_0$ successfully with high probability.

Thus the numerical experiments provide some evidence for Open Problem 3.4 that the phase transition might exist at $p = \Omega(n^{-1} \log n)$, i.e., no spurious local minima exist if $p \geq cn^{-1} \log n$ with high probability. It is important to note that the global convergence of plain gradient descent with random initialization does not necessarily rule out the possible existence of spurious minima. Therefore, the experiments serve to support our belief that a near-optimal bound of p may be obtained by using new techniques.
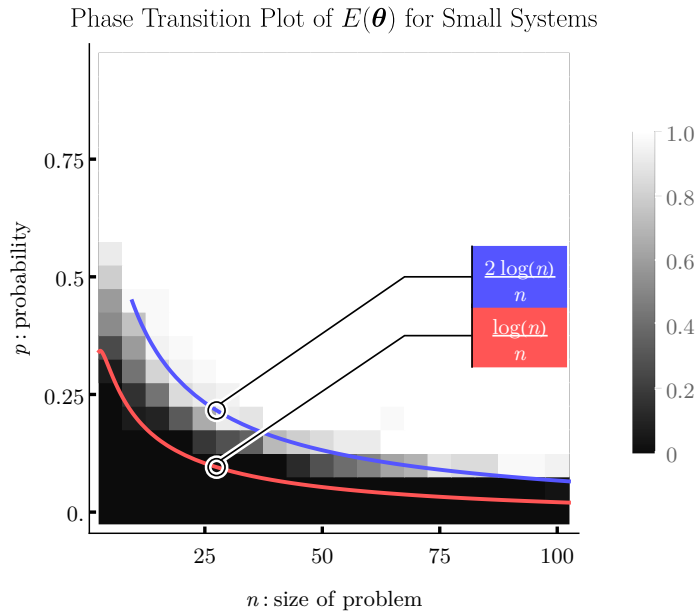
FIG. 5. *Phase transition plot on $E(\boldsymbol{\theta})$ with smaller systems. Here n is chosen between 5 and 100 and p varies from 0 to 1. For each pair of (n, p), we run 50 independent experiments. The shade of grey in each square denotes the fraction of trials for which gradient descent reaches its global minimum; the square is lighter if there are more successes. The red line represents the phase transition threshold we conjecture in Open Problem 3.4, i.e., $p = n^{-1} \log n$; the blue line represents $p = 2n^{-1} \log n$, which achieves a high probability of successful recovery.*

5. Proofs. The proofs of main theorems include a key ingredient, which is summarized in the following proposition. This important proposition is inspired by the work of Taylor [46].

PROPOSITION 5.1. *Let $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_n]^\top$ be a local minimizer of $E(\boldsymbol{\theta})$ with $\boldsymbol{A}$ as the adjacency matrix of a connected graph. Suppose there exists some $r = \|r\| e^{i\theta_r} \in \mathbb{C}$ such that all $\{\theta_i\}_{i=1}^n$ satisfy*

$$|\sin(\theta_i - \theta_r)| < \frac{1}{\sqrt{2}} \quad \forall 1 \leq i \leq n.$$

Then $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ is the only local minimizer of $E(\boldsymbol{\theta})$ modulo a global translation.

Phase Transition Plot of $E(\boldsymbol{\theta})$ for Large Systems

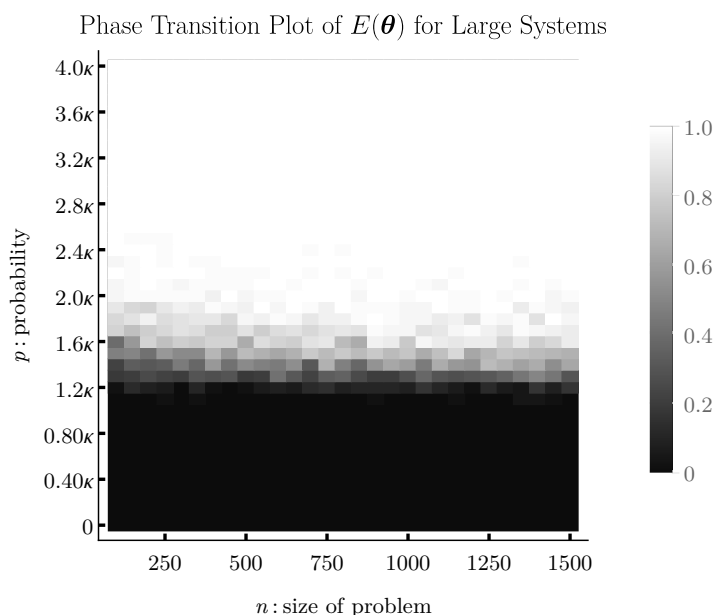Fig. 6. *Phase transition plot on $E(\boldsymbol{\theta})$ with larger systems. We let n range from 100 to 1500 and $p = c\kappa$ where $\kappa = \frac{\log n}{n}$ and $0 \le c \le 4$. For each pair of (n, p), we run 50 independent trials and compute the fraction of successful trials among all 50 instances. The black region stands for failure and the white region means success. A clear phase transition occurs around $p = \frac{1.5 \log n}{n}$ and the probability of success is overwhelmingly high when $p \ge \frac{2 \log n}{n}$.*

Remark 5.2. In the proof of Theorems 3.1 and 3.3, we pick

$$r(\boldsymbol{\theta}) := \sum_{j=1}^{n} e^{\mathrm{i}\theta_j},$$

where $n^{-1} r(\boldsymbol{\theta})$ is called the order parameter introduced by Kuramoto in [27]. The absolute value of the order parameter measures how much the current state is correlated with the global minimum.

Proof. Since $\boldsymbol{\theta}$ is a local minimizer, it must satisfy the first and second necessary conditions (2.2) and (2.4), respectively.

The assumption implies that

$$\{\theta_i\}_{i=1}^{n} \subseteq \underbrace{\left\{\theta : -\frac{\pi}{4} < \theta - \theta_r < \frac{\pi}{4}\right\}}_{\Gamma} \bigcup \underbrace{\left\{\theta : \frac{3\pi}{4} < \theta - \theta_r < \frac{5\pi}{4}\right\}}_{\Gamma^c},$$

which is the union of two disjoint quadrants.

We claim that all $\{\theta_i\}_{i=1}^{n}$ must be inside one of these two quadrants, and we will prove it by contradiction. Suppose both quadrants contain at least one element in $\{\theta_i\}_{i=1}^{n}$. Since the graph is connected, the induced partitions Γ and Γ^c yield

$$\sum_{i \in \Gamma} \sum_{j \in \Gamma^c} a_{ij} \cos(\theta_i - \theta_j) < 0,$$

which follows from $\cos(\theta_i - \theta_j) < 0$ for all $\theta_i \in \Gamma$ and $\theta_j \in \Gamma^c$ and at least one element

a_{ij} is nonzero for $i \in \Gamma$ and $j \in \Gamma^c$ due to the graph connectivity. However, this contradicts the fact that $\mathrm{ddiag}(\boldsymbol{A} \circ \boldsymbol{Q}\boldsymbol{Q}^\top) - \boldsymbol{A} \circ \boldsymbol{Q}\boldsymbol{Q}^\top \succeq 0$.

By letting $\boldsymbol{w} \in \mathbb{R}^n$, whose ith entry w_i satisfies

$$w_i = \begin{cases} 1 & \text{if } \theta_i \in \Gamma, \\ -1 & \text{if } \theta_i \in \Gamma^c, \end{cases}$$

it holds that

$$\sum_{i \in \Gamma} \sum_{j \in \Gamma^c} a_{ij} \cos(\theta_i - \theta_j) = \frac{1}{4} \sum_{i \in \Gamma} \sum_{j \in \Gamma^c} a_{ij} \cos(\theta_i - \theta_j)(w_i - w_j)^2$$
$$= \frac{1}{4}\boldsymbol{w}^\top (\mathrm{ddiag}(\boldsymbol{A} \circ \boldsymbol{Q}\boldsymbol{Q}^\top) - \boldsymbol{A} \circ \boldsymbol{Q}\boldsymbol{Q}^\top)\boldsymbol{w} \geq 0,$$

which leads to a contradiction.

Therefore, all $\{\theta_i\}_{i=1}^n$ must stay in the interior of a single quadrant. Now we show that if this is the case, the only critical point of $E(\boldsymbol{\theta})$ within this quadrant must be $\boldsymbol{\theta} = \boldsymbol{0}$.

Define an auxiliary function as

$$E(t) := \frac{1}{2} \sum_{1 \leq i,j \leq n} a_{ij}(1 - \cos((\theta_i - \theta_j)t)),$$

where $t \in [0, 1]$. Then

$$E'(t) = \frac{1}{2} \sum_{1 \leq i,j \leq n} a_{ij} \cdot (\theta_i - \theta_j)\sin((\theta_i - \theta_j)t) \geq \frac{t}{\pi} \sum_{1 \leq i,j \leq n} a_{ij}(\theta_i - \theta_j)^2,$$

which follows from $x\sin(x) \geq 2\pi^{-1}x^2$ for $-\pi/2 \leq x \leq \pi/2$ and $\theta_i - \theta_j \in (-\pi/2, \pi/2)$.

From spectral graph theory [16, Chapter 1], we know that the quadratic form $\sum_{i,j} a_{ij}(\theta_i - \theta_j)^2$ must be strictly positive for a connected graph unless $\theta_i = \theta_j$ for $i \neq j$. In other words, if $\boldsymbol{\theta} \neq \boldsymbol{0}$ (modulo a global translation), $E'(t) > 0$ implies that $\boldsymbol{\theta}$ cannot be a critical point of $E(\boldsymbol{\theta})$ since the energy function strictly decreases along the straight line pointing from $\boldsymbol{\theta}$ to the origin. Thus $\boldsymbol{\theta}$ is not a local minimizer unless $\boldsymbol{\theta} = \boldsymbol{0}$. This guarantees that $\boldsymbol{\theta} = \boldsymbol{0}$ holds and it is the only local minimizer within the quadrant. □

5.1. Proof of Theorem 3.1. Consider a graph with adjacency matrix $\boldsymbol{A}$ and with each vertex of degree at least $\mu(n-1)$, i.e.,

$$\min_{1 \leq i \leq n} \sum_{j:j \neq i} a_{ij} \geq \mu(n-1).$$

Here we assume the diagonal entries of $\boldsymbol{A}$ are zero, i.e., $a_{ii} = 0$, since they do not contribute to $E(\boldsymbol{\theta})$.

Proposition 5.1 indicates that it suffices to show that $\{\theta_i\}_{i=1}^n$ be within two disjoint quadrants for any given local minimizer $\boldsymbol{\theta}$ under $\mu > (3 - \sqrt{2})/2$. For any given local minimizer $\boldsymbol{\theta}$, we let $\boldsymbol{Q} \in \mathbb{R}^{n \times 2}$ be of the form in (2.1) and we have $(\boldsymbol{Q}\boldsymbol{Q}^\top)_{ij} = \cos(\theta_i - \theta_j)$. The second-order necessary condition in (2.3) implies that

$$\langle \mathrm{diag}(\boldsymbol{A}\boldsymbol{Q}\boldsymbol{Q}^\top) - \boldsymbol{A} \circ \boldsymbol{Q}\boldsymbol{Q}^\top, \boldsymbol{Q}\boldsymbol{Q}^\top \rangle = \langle \boldsymbol{A}, \boldsymbol{Q}\boldsymbol{Q}^\top \rangle - \langle \boldsymbol{A}, \boldsymbol{Q}\boldsymbol{Q}^\top \circ \boldsymbol{Q}\boldsymbol{Q}^\top \rangle \geq 0,$$

where $(\boldsymbol{Q}\boldsymbol{Q}^\top)_{ii} = 1$, $1 \leq i \leq n$, and $\boldsymbol{Q}\boldsymbol{Q}^\top$ is positive semidefinite. The inequality above is equivalent to

$$\langle \boldsymbol{J}_n, \boldsymbol{Q}\boldsymbol{Q}^\top \rangle \geq \langle \boldsymbol{J}_n, \boldsymbol{Q}\boldsymbol{Q}^\top \circ \boldsymbol{Q}\boldsymbol{Q}^\top \rangle + \langle \boldsymbol{J}_n - \boldsymbol{A}, \boldsymbol{Q}\boldsymbol{Q}^\top - \boldsymbol{Q}\boldsymbol{Q}^\top \circ \boldsymbol{Q}\boldsymbol{Q}^\top \rangle$$

$$(5.1) \qquad \geq \langle \boldsymbol{J}_n, \boldsymbol{Q}\boldsymbol{Q}^\top \circ \boldsymbol{Q}\boldsymbol{Q}^\top \rangle + \langle \boldsymbol{J}_n - \boldsymbol{I}_n - \boldsymbol{A}, \boldsymbol{Q}\boldsymbol{Q}^\top - \boldsymbol{Q}\boldsymbol{Q}^\top \circ \boldsymbol{Q}\boldsymbol{Q}^\top \rangle.$$

Our first goal is to get a lower bound of $\langle \boldsymbol{J}_n, \boldsymbol{Q}\boldsymbol{Q}^\top \rangle$ by estimating each term on the right-hand side of (5.1).

Note that $\boldsymbol{Q}\boldsymbol{Q}^\top$ is rank 2 and its trace is n. Thus its Frobenius norm has a lower bound as

$$(5.2) \quad \langle \boldsymbol{J}_n, \boldsymbol{Q}\boldsymbol{Q}^\top \circ \boldsymbol{Q}\boldsymbol{Q}^\top \rangle = \|\boldsymbol{Q}\boldsymbol{Q}^\top\|_F^2 = \lambda_1^2 + \lambda_2^2 \geq \frac{(\lambda_1 + \lambda_2)^2}{2} = \frac{1}{2}(\mathrm{Tr}(\boldsymbol{Q}\boldsymbol{Q}^\top))^2 = \frac{n^2}{2},$$

where λ_1 and λ_2 are the eigenvalues of $\boldsymbol{Q}\boldsymbol{Q}^\top$ and $\mathrm{Tr}(\boldsymbol{Q}\boldsymbol{Q}^\top) = n$.

Since each row sum of $\boldsymbol{A}$ is at least $\mu(n-1)$, we know that $\boldsymbol{J}_n - \boldsymbol{I}_n - \boldsymbol{A}$ has at most $(1-\mu)(n-1)n$ nonzero entries. Moreover, the absolute values of each entry in $\boldsymbol{Q}\boldsymbol{Q}^\top$ and $\boldsymbol{Q}\boldsymbol{Q}^\top \circ \boldsymbol{Q}\boldsymbol{Q}^\top$ are bounded by 1. Hence,

$$|\langle \boldsymbol{J}_n - \boldsymbol{I}_n - \boldsymbol{A}, \boldsymbol{Q}\boldsymbol{Q}^\top - \boldsymbol{Q}\boldsymbol{Q}^\top \circ \boldsymbol{Q}\boldsymbol{Q}^\top \rangle| \leq \sum_{i \neq j} (1 - a_{ij}) |\cos(\theta_i - \theta_j) - \cos^2(\theta_i - \theta_j)|$$

$$(5.3) \qquad\qquad\qquad\qquad \leq 2 \sum_{i \neq j} (1 - a_{ij}) \leq 2(1 - \mu)(n - 1)n.$$

Combining (5.2) and (5.3) with (5.1), we have a lower bound for the correlation between $\boldsymbol{Q}$ and the constant vector $\mathbf{1}_n$ as follows:

$$\langle \boldsymbol{J}_n, \boldsymbol{Q}\boldsymbol{Q}^\top \rangle \geq \frac{n^2}{2} - 2(1 - \mu)(n - 1)n = \left(2\mu - \frac{3}{2}\right)n^2 + 2(1 - \mu)n.$$

Now we consider the rescaled order parameter r as

$$r = \sum_{j=1}^n e^{\mathrm{i}\theta_j} = \|r\|e^{\mathrm{i}\theta_r},$$

and its ℓ_2-norm is bounded below by

$$\|r\|^2 = \left(\sum_{j=1}^n \cos\theta_j\right)^2 + \left(\sum_{j=1}^n \sin\theta_j\right)^2 = \sum_{1 \leq i,j \leq n} (\cos\theta_i \cos\theta_j + \sin\theta_i \sin\theta_j)$$

$$(5.4) \qquad = \langle \boldsymbol{J}_n, \boldsymbol{Q}\boldsymbol{Q}^\top \rangle \geq \left(2\mu - \frac{3}{2}\right)n^2 + 2(1 - \mu)n.$$

For every $i \in [n]$, it holds that

$$\|r\|e^{-\mathrm{i}(\theta_i - \theta_r)} = re^{-\mathrm{i}\theta_i} = \sum_{j=1}^n e^{-\mathrm{i}(\theta_i - \theta_j)}.$$

By taking the imaginary parts, we have

$$(5.5) \quad \|r\| \cdot \sin(\theta_i - \theta_r) = \sum_{j=1}^n \sin(\theta_i - \theta_j) = \sum_{j=1}^n (1 - a_{ij}) \sin(\theta_i - \theta_j) \quad \forall 1 \leq i \leq n,$$

where the second equality follows from the first-order necessary condition in (2.2), i.e.,

$$\sum_{j=1}^{n} a_{ij} \sin(\theta_i - \theta_j) = 0, \quad 1 \le i \le n.$$

By combining (5.3) with (5.5), we have

$$\|r\| \cdot |\sin(\theta_i - \theta_r)| = \left| \sum_{j=1}^{n} (1 - a_{ij}) \sin(\theta_i - \theta_j) \right| \le \left| \sum_{j=1}^{n} (1 - a_{ij}) \right| \le (1 - \mu)(n - 1).$$

Note that $\|r\|^2 = \langle \boldsymbol{J}_n, \boldsymbol{Q}\boldsymbol{Q}^{\top} \rangle \ge (2\mu - \frac{3}{2})n^2 + 2(1 - \mu)n$ in (5.4). Therefore,

$$|\sin(\theta_i - \theta_r)|^2 \le \frac{(1 - \mu)^2 (n - 1)^2}{\|r\|^2} \le \frac{2(1 - \mu)^2}{4\mu - 3 + 4(1 - \mu)n^{-1}} < \frac{1}{2}$$

for any $1 \le i \le n$ provided that $(3 - \sqrt{2})/2 \le \mu \le 1$, which immediately implies

$$|\sin(\theta_i - \theta_r)| < \frac{1}{\sqrt{2}}, \quad 1 \le i \le n.$$

Here r and $\boldsymbol{\theta}$ satisfy the assumption in Proposition 5.1, and moreover the underlying graph is connected given that $\mu \ge (3 - \sqrt{2})/2$. Now applying Proposition 5.1 yields the desired result. □

5.2. Proof of Theorem 3.3.

5.2.1. The road map for the proof of Theorem 3.3. To prove Theorem 3.3, besides Proposition 5.1, we will also use the toolbox of random matrices and concentration of measure, which are widely used in many signal processing and statistical inference problems. Simply speaking, the main strategy follows from finding a lower bound of the rescaled order parameter $\|r\| = \|\sum_{j=1}^{n} e^{\mathrm{i}\theta_j}\|$, which is exactly equal to $\|\boldsymbol{Q}^{\top}\boldsymbol{1}_n\|$; obtaining an upper bound of $|\mathrm{Im}(re^{-\mathrm{i}\theta_i})|$ for a local minimizer $\boldsymbol{\theta} = \{\theta_i\}_{i=1}^{n}$; and finally applying Proposition 5.1 to finish the proof. Note that by applying Theorem 15 of [7] with a proper transformation, the same performance bound with respect to p can be obtained to ensure a benign optimization landscape of $E(\boldsymbol{\theta})$. We provide an alternative proof based on the "two-disjoint-quadrant" argument (Proposition 5.1) and get a tighter bound of the lower bound of order parameter, which are of independent interest.

From now on, we set

(5.6)
$$p := \frac{32\gamma \log n}{n^{1/3}}$$

in (3.1) under the assumptions of Theorem 3.3 and also define

$$\boldsymbol{\Delta} := \boldsymbol{A} - p\boldsymbol{J}_n$$

as the centered random adjacency matrix. Here, for simplicity of presentation, we set the diagonal entries of $\boldsymbol{A}$ equal to p since they do not appear in $E(\boldsymbol{\theta})$.

Now we present two supporting lemmata whose proofs we defer to section 5.2.2 and show how these two results lead to the final proof of Theorem 3.3.

LEMMA 5.3. *For $\boldsymbol{A}$ an Erdős–Rényi random graph $G(n, p)$, we have*

$$(5.7) \qquad \|\boldsymbol{A} - \mathbb{E}(\boldsymbol{A})\| \leq \sqrt{2\gamma np(1-p)\log n} + \frac{2\gamma \log n}{3}$$

with probability at least $1 - 2n^{-\gamma+1}$. In particular, it holds that

$$\max_{1 \leq i \leq n} \|\boldsymbol{e}_i^\top(\boldsymbol{A} - \mathbb{E}(\boldsymbol{A}))\| \leq \sqrt{2\gamma np(1-p)\log n} + \frac{2\gamma \log n}{3}.$$

Moreover, the infinity norm bound of $\boldsymbol{\Delta}\boldsymbol{1}_n$ obeys

$$\|\boldsymbol{\Delta}\boldsymbol{1}_n\|_\infty \leq \sqrt{2\gamma np(1-p)\log n} + \frac{2\gamma \log n}{3}$$

with probability at least $1 - 2n^{-\gamma+1}$.

The proof of this lemma is standard; it follows from direct application of the matrix Bernstein inequality. The purpose of this lemma is to show how close $\boldsymbol{A}$ is to its expectation; moreover the estimation of the upper bound of $|\mathrm{Im}(re^{-\mathrm{i}\theta_i})|$ is based on this fact. The next lemma demonstrates that with high probability, the rescaled order parameter of a local minimizer, i.e., $r = \sum_{j=1}^n e^{\mathrm{i}\theta_j}$, and the global minimizer $\boldsymbol{1}_n$ are highly correlated.

LEMMA 5.4. *With probability at least*

$$1 - 4\exp\left(n\left(\log(100n^{1/3}) - 2\gamma \log n\right)\right), \quad \gamma \geq 1,$$

it holds that

$$\|\boldsymbol{Q}^\top\boldsymbol{1}_n\|^2 \geq n^2(1 - 3n^{-\frac{1}{3}})$$

uniformly for all $\boldsymbol{Q} \in \mathbb{R}^{n \times 2}$ satisfying the first and second necessary conditions.

Proof of Theorem 3.3. Let $r(\boldsymbol{\theta}) := \sum_{j=1}^n e^{\mathrm{i}\theta_j}$, where $\{\theta_j\}_{j=1}^n$ is a local minimizer and for simplicity we use "r" to represent $r(\boldsymbol{\theta})$. By using the first-order necessary condition $\sum_{j=1}^n a_{ij}\sin(\theta_i - \theta_j) = 0$ and the Cauchy–Schwarz inequality, we have

$$
\begin{aligned}
p|\mathrm{Im}(re^{-\mathrm{i}\theta_i})| &= p\left|\sum_{j=1}^n \sin(\theta_i - \theta_j)\right| \\
&= \left|\sum_{j=1}^n (a_{ij} - p)\sin(\theta_i - \theta_j)\right| \\
&\leq |\sin\theta_i| \cdot \left|\sum_{j=1}^n (a_{ij} - p)\cos\theta_j\right| + |\cos\theta_i| \cdot \left|\sum_{j=1}^n (a_{ij} - p)\sin\theta_j\right| \\
&\leq \sqrt{\left(\sum_{j=1}^n (a_{ij} - p)\cos\theta_j\right)^2 + \left(\sum_{j=1}^n (a_{ij} - p)\sin\theta_j\right)^2} \\
&= \|\boldsymbol{e}_i^\top(\boldsymbol{A} - p\boldsymbol{J}_n)\boldsymbol{Q}\| = \|\boldsymbol{e}_i^\top\boldsymbol{\Delta}\boldsymbol{Q}\|.
\end{aligned}
$$

Now we give an upper bound of $\|\boldsymbol{e}_i^\top(\boldsymbol{A} - p\boldsymbol{J}_n)\boldsymbol{Q}\|$:

$$\|\boldsymbol{e}_i^\top\boldsymbol{\Delta}\boldsymbol{Q}\| \leq \frac{1}{n}|\boldsymbol{e}_i^\top\boldsymbol{\Delta}\mathbf{1}_n| \cdot \|\boldsymbol{Q}^\top\mathbf{1}_n\| + \left\|\boldsymbol{e}_i^\top\boldsymbol{\Delta}\left(\boldsymbol{I}_n - \frac{\mathbf{1}_n\mathbf{1}_n^\top}{n}\right)\boldsymbol{Q}\right\|$$

$$\leq \frac{1}{n}\|\boldsymbol{\Delta}\mathbf{1}_n\|_\infty \cdot \|\boldsymbol{Q}^\top\mathbf{1}_n\| + \|\boldsymbol{e}_i^\top\boldsymbol{\Delta}\| \cdot \left\|\left(\boldsymbol{I}_n - \frac{\mathbf{1}_n\mathbf{1}_n^\top}{n}\right)\boldsymbol{Q}\right\|_F.$$

In fact, the suboptimal bound of our theorem results from the estimation of the second term above. The possible solution may require a sharper upper bound of the correlation between the local minimizers $\boldsymbol{Q}$ and each row of $\boldsymbol{\Delta}$.

It is straightforward to get an upper bound for each term in the expression above:

$$\|\boldsymbol{\Delta}\mathbf{1}_n\|_\infty \leq \sqrt{2\gamma np(1-p)\log n} + \frac{2\gamma\log n}{3},$$

$$\|\boldsymbol{e}_i^\top\boldsymbol{\Delta}\| \leq \sqrt{2\gamma np(1-p)\log n} + \frac{2\gamma\log n}{3},$$

$$\left\|\left(\boldsymbol{I}_n - \frac{\mathbf{1}_n\mathbf{1}_n^\top}{n}\right)\boldsymbol{Q}\right\|_F^2 = n - \frac{1}{n}\|\boldsymbol{Q}^\top\mathbf{1}_n\|^2 \leq n - n\left(1 - 3n^{-\frac{1}{3}}\right) \leq 3n^{\frac{2}{3}},$$

$$\|\boldsymbol{Q}^\top\mathbf{1}_n\| \leq n,$$

where the first two inequalities are based on Lemma 5.3 and the last two are implied by Lemma 5.4.

Therefore, each $\|\boldsymbol{e}_i^\top\boldsymbol{\Delta}\boldsymbol{Q}\|$ is uniformly bounded by

$$\|\boldsymbol{e}_i^\top\boldsymbol{\Delta}\boldsymbol{Q}\| \leq \left(1 + \sqrt{3}n^{\frac{1}{3}}\right) \cdot \left(\sqrt{2\gamma np(1-p)\log n} + \frac{2\gamma\log n}{3}\right)$$

for all $\boldsymbol{Q}$.

For reasonably large n, we consider $p^{-1}\|\boldsymbol{e}_i^\top\boldsymbol{\Delta}\boldsymbol{Q}\|$, which satisfies

$$|\mathrm{Im}(re^{-\mathrm{i}\theta_i})| \leq \frac{\|\boldsymbol{e}_i^\top\boldsymbol{\Delta}\boldsymbol{Q}\|}{p} \leq 2n^{\frac{1}{3}}\left(\sqrt{\frac{2\gamma n(1-p)\log n}{p}} + \frac{2\gamma\log n}{3p}\right)$$

$$= \frac{n}{2}\left(\sqrt{1-p} + \frac{n^{-\frac{1}{3}}}{12}\right)$$

$$(5.8) \qquad\qquad < n\sqrt{\frac{1 - 3n^{-\frac{1}{3}}}{2}}$$

when $p = \frac{32\gamma\log n}{n^{1/3}} > \frac{3}{n^{1/3}}$ for $\gamma \geq 1$. On the other hand, we have

$$(5.9) \qquad\qquad \|r\|^2 = \|\boldsymbol{Q}^\top\mathbf{1}_n\|^2 \geq n^2(1 - 3n^{-\frac{1}{3}}),$$

where the lower bound is given by Lemma 5.4.

For every $i \in [n]$ we consider an upper bound of $|\sin(\theta_i - \theta_r)|$ by using

$$\left|\sum_{j=1}^n \sin(\theta_i - \theta_j)\right| = \left|\mathrm{Im}(re^{-\mathrm{i}\theta_i})\right| = \|r\| \cdot |\sin(\theta_i - \theta_r)| \leq \frac{\|\boldsymbol{e}_i^\top\boldsymbol{\Delta}\boldsymbol{Q}\|}{p},$$

where $r = \sum_{j=1}^n e^{i\theta_j} = \|r\|e^{i\theta_r}$. Combining (5.8) with (5.9) yields
(5.10)

$$|\text{Im}(re^{-i\theta_i})| = \|r\| \cdot |\sin(\theta_i - \theta_r)| \leq \frac{\|e_i^\top \Delta Q\|}{p} < n\sqrt{\frac{1 - 3n^{-\frac{1}{3}}}{2}} \leq \frac{\|Q^\top \mathbf{1}_n\|}{\sqrt{2}} = \frac{\|r\|}{\sqrt{2}}.$$

Therefore, we have

$$|\sin(\theta_i - \theta_r)| < \frac{1}{\sqrt{2}} \quad \forall\, 1 \leq i \leq n.$$

On the other hand, we assume $p = 32\gamma n^{-\frac{1}{3}}\log n \gg n^{-1}\log n$ and thus by Theorem 5.8 of [48, Chapter 5], the Erdős–Rényi random graph is connected with overwhelmingly high probability. Now all the assumptions of Proposition 5.1 are fulfilled and we have $\theta = 0$ as the only local minimizer. $\qquad\square$

5.2.2. Proof of the supporting lemmata in Theorem 3.3.

THEOREM 5.5 (see [47, matrix Bernstein, Theorem 1.4]). *Consider a finite sequence* $\{Z_k\}$ *of independent random self-adjoint matrices with dimension* $n \times n$. *Assume that each random matrix satisfies*

$$\mathbb{E}(Z_k) = 0, \qquad \|Z_k\| \leq R.$$

Then, for all $t \geq 0$,

$$\mathbb{P}\left(\left\|\sum_k Z_k\right\| \geq t\right) \leq n \cdot \exp\left(-\frac{t^2/2}{\sigma^2 + Rt/3}\right),$$

where $\sigma^2 = \|\sum_k \mathbb{E}(Z_k^2)\|$. *In other words,*

$$\left\|\sum_k Z_k\right\| \leq \frac{2\gamma R \log n}{3} + \sqrt{2\gamma\sigma^2 \log n}$$

holds with probability at least $1 - 2n^{-\gamma+1}$.

Proof of Lemma 5.3. The difference between A and $\mathbb{E}(A)$ is a sum of rank 1 matrices:

$$A - \mathbb{E}(A) = \sum_{i<j}(a_{ij} - p)(E_{ij} + E_{ji}),$$

where $\mathbb{E}(a_{ij}) = p$ and $E_{ij} = e_i e_j^\top$. Note that each component is bounded by $\max\{p, 1-p\}$ in operator norm since $\|E_{ij} + E_{ji}\| = 1$. The variance is

$$\mathbb{E}(A - \mathbb{E}(A))^2 = p(1-p)\sum_{i<j}(E_{ii} + E_{jj}) \preceq np(1-p)I_n,$$

which follows from the independence among all entries $\{a_{ij}\}$. Applying the matrix Bernstein inequality, we have

$$\mathbb{P}(\|A - \mathbb{E}(A)\| \geq t) \leq 2n \cdot \exp\left(-\frac{t^2/2}{np(1-p) + t/3}\right) \leq 2n^{-\gamma+1}$$

with $t = \sqrt{2\gamma np(1-p)\log n} + 2\gamma \log n/3$. The upper bound of $\|e_i^\top(A - \mathbb{E}(A))\|$ follows directly from $\|e_i^\top(A - \mathbb{E}(A))\| \leq \|A - \mathbb{E}(A)\|$.

For the infinity norm bound of $\boldsymbol{\Delta}\mathbf{1}_n$, we apply the Bernstein inequality to $(\boldsymbol{\Delta}\mathbf{1}_n)_i$ for each fixed i and then take the union bound over $1 \le i \le n$, and then have

$$\mathbb{P}\left(\|\boldsymbol{\Delta}\mathbf{1}_n\|_\infty \ge t\right) = \mathbb{P}\left(\max_{1 \le i \le n}\left|\sum_{j=1}^n (a_{ij} - p)\right| \ge t\right) \le 2n \exp\left(-\frac{t^2/2}{np(1-p) + t/3}\right)$$

$$\le 2n^{-\gamma+1},$$

where $t = \sqrt{2\gamma np(1-p)\log n} + 2\gamma \log n/3$. $\qquad\square$

In order to prove Lemma 5.4, we first introduce the restricted isometry property of $\boldsymbol{A} - p\boldsymbol{J}_n$ as follows.

LEMMA 5.6 (restricted isometry property on a manifold). *Let $\boldsymbol{A}$ be an Erdős–Rényi graph $G(n,p)$. Then the restricted isometry property*

(5.11)
$$|\langle \boldsymbol{A} - p\boldsymbol{J}_n, \boldsymbol{Q}\boldsymbol{Q}^\top \circ \boldsymbol{Q}\boldsymbol{Q}^\top\rangle| \le \delta p\|\boldsymbol{Q}\boldsymbol{Q}^\top\|_F^2,$$
$$|\langle \boldsymbol{A} - p\boldsymbol{J}_n, \boldsymbol{Q}\boldsymbol{Q}^\top\rangle| \le \delta p\|\boldsymbol{Q}\boldsymbol{Q}^\top\|_F^2$$

holds uniformly for all $\boldsymbol{Q}$ with probability at least

$$1 - 4\exp\left(n\log\left(\frac{100}{\delta}\right) - \frac{n^2\delta^2 p}{64(1-p) + 6\delta}\right),$$

where $0 < \delta < 1$.

Proof. Let $\boldsymbol{X} := \boldsymbol{Q}\boldsymbol{Q}^\top$ and $X_{ij} = \cos(\theta_i - \theta_j)$. We expect that $\langle \boldsymbol{A}, \boldsymbol{X} \circ \boldsymbol{X}\rangle$ and $\langle \boldsymbol{A}, \boldsymbol{X}\rangle$ will not deviate too much from their expectations *uniformly* for all $\boldsymbol{X} = \boldsymbol{Q}\boldsymbol{Q}^\top$, where

$$\langle \boldsymbol{A}, \boldsymbol{X} \circ \boldsymbol{X}\rangle = \sum_{i,j} a_{ij}\cos^2(\theta_i - \theta_j), \qquad \langle \boldsymbol{A}, \boldsymbol{X}\rangle = \sum_{i,j} a_{ij}\cos(\theta_i - \theta_j)$$

are sub-Gaussian random variables. It suffices to only look at $\langle \boldsymbol{A}, \boldsymbol{X} \circ \boldsymbol{X}\rangle$ due to their similarities.

Note that the centered variable satisfies

$$\langle \boldsymbol{A}, \boldsymbol{X} \circ \boldsymbol{X}\rangle - p\|\boldsymbol{X}\|_F^2 = \sum_{i \ne j}(a_{ij} - p)\cos^2(\theta_i - \theta_j),$$

which follows from $\mathbb{E}(a_{ij}) = p$. Each term $(a_{ij} - p)\cos^2(\theta_i - \theta_j)$ is contained in $[-p, 1-p]$. Moreover, the variance of $\sum_{i,j}(a_{ij} - p)\cos^2(\theta_i - \theta_j)$ satisfies

$$\sigma^2 = \text{Var}\left(2\sum_{i<j}(a_{ij} - p)\cos^2(\theta_i - \theta_j)\right) = 4p(1-p)\sum_{i<j}\cos^4(\theta_i - \theta_j) \le 2n^2 p(1-p).$$

By Bernstein's inequality, we have

$$\mathbb{P}\left(\left|\langle \boldsymbol{A}, \boldsymbol{X} \circ \boldsymbol{X}\rangle - p\|\boldsymbol{X}\|_F^2\right| \ge t\right) \le 2\exp\left(-\frac{t^2/2}{2n^2 p(1-p) + t/3}\right)$$

(5.12)
$$\le 2\exp\left(-\frac{n^2\delta^2 p}{64(1-p) + 6\delta}\right),$$

where we use

$$t = \frac{\delta p \|\boldsymbol{X}\|_F^2}{2}, \qquad \frac{n^2}{2} \le \|\boldsymbol{X}\|_F^2 \le n^2.$$

Thus it holds that

$$(5.13) \qquad \left| \langle \boldsymbol{A}, \boldsymbol{X} \circ \boldsymbol{X} \rangle - p\|\boldsymbol{X}\|_F^2 \right| \le \frac{\delta p}{2} \|\boldsymbol{X}\|_F^2$$

with probability at least

$$1 - 2\exp\left(-\frac{n^2 \delta^2 p}{64(1-p) + 6\delta} \right).$$

So far we have established a concentration inequality for a fixed set of $\{\theta_i\}_{i=1}^n$. Now we need to extend inequality (5.13) to arbitrary $\theta \in [0, 2\pi]^n$ uniformly.

Define $m := \lfloor \epsilon^{-1} + 1 \rfloor$ and construct an ϵ-net over $[0, 2\pi]^n$ as

$$\mathcal{S}_{n,m} := \left\{ \widetilde{\boldsymbol{\theta}} = \{\widetilde{\theta}_i\}_{i=1}^n \in [0, 2\pi]^n : \widetilde{\theta}_i = \frac{2l\pi}{m} \text{ for some } l : 1 \le l \le m \right\}$$

with $|\mathcal{S}_{n,m}| = m^n$. Formula (5.12) holds for all $\widetilde{\boldsymbol{\theta}}$ on the net $\mathcal{S}_{n,m}$ with probability at least

$$1 - 2|\mathcal{S}_{n,m}| \cdot \exp\left(-\frac{n^2 \delta^2 p}{64(1-p) + 6\delta} \right) = 1 - 2\exp\left(n \log m - \frac{n^2 \delta^2 p}{64(1-p) + 6\delta} \right)$$

by taking the union bound for all $\widetilde{\boldsymbol{\theta}} \in \mathcal{S}_{n,m}$.

For arbitrary $\{\theta_i\}_{i=1}^n$, there always exists a point $\widetilde{\boldsymbol{\theta}}$ on the net $\mathcal{S}_{n,m}$ such that $\|\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}\|_\infty \le \pi/m$, i.e., $|\theta_i - \widetilde{\theta}_i| \le \pi/m$ for $1 \le i \le n$. By letting

$$f(\boldsymbol{\theta}) := \langle \boldsymbol{A} - p\boldsymbol{J}_n, \boldsymbol{Q}\boldsymbol{Q}^\top \circ \boldsymbol{Q}\boldsymbol{Q}^\top \rangle,$$

we have that $f(\boldsymbol{\theta})$ is a Lipschitz continuous function and satisfies

$$
\begin{aligned}
|f(\boldsymbol{\theta}) - f(\widetilde{\boldsymbol{\theta}})| &= \left| \sum_{i \ne j} (a_{ij} - p)(\cos^2(\theta_i - \theta_j) - \cos^2(\widetilde{\theta}_i - \widetilde{\theta}_j)) \right| \\
&\le 2\sum_{i \ne j} |a_{ij} - p| \cdot |(\theta_i - \theta_j) - (\widetilde{\theta}_i - \widetilde{\theta}_j)| \\
&\le 4\|\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}\|_\infty \cdot \sum_{i \ne j} |a_{ij} - p| \\
&\le 4\|\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}\|_\infty \cdot \sum_{i \ne j} (a_{ij} + p) \\
(5.14) \qquad &\le \frac{4\pi}{m} \left(n\|\boldsymbol{A}\| + n^2 p \right),
\end{aligned}
$$

where $|\theta_i - \widetilde{\theta}_i| \le \pi/m$ and $\|\boldsymbol{A}\| \le np + \sqrt{2\gamma np(1-p)\log n} + 2\gamma \log n/3$.

Now we consider the difference between $\|\boldsymbol{X}\|_F^2$ and $\|\widetilde{\boldsymbol{X}}\|_F^2$, where $\widetilde{\boldsymbol{X}} = \widetilde{\boldsymbol{Q}}\widetilde{\boldsymbol{Q}}^\top$ and $\widetilde{X}_{ij} = \cos(\widetilde{\theta}_i - \widetilde{\theta}_j)$, and it holds that

$$\left| \|\boldsymbol{X}\|_F^2 - \|\widetilde{\boldsymbol{X}}\|_F^2 \right| \leq \sum_{i \neq j} |\cos^2(\theta_i - \theta_j) - \cos^2(\widetilde{\theta}_i - \widetilde{\theta}_j)|$$

$$\leq 2 \sum_{i \neq j} |(\theta_i - \theta_j) - (\widetilde{\theta}_i - \widetilde{\theta}_j)|$$

$$(5.15) \qquad\qquad\qquad \leq \frac{4\pi n^2}{m}.$$

Since $|f(\widetilde{\boldsymbol{\theta}})| \leq \frac{\delta p \|\widetilde{\boldsymbol{X}}\|_F^2}{2}$, $f(\boldsymbol{\theta})$ is bounded by

$$|f(\boldsymbol{\theta})| \leq |f(\widetilde{\boldsymbol{\theta}})| + |f(\boldsymbol{\theta}) - f(\widetilde{\boldsymbol{\theta}})|$$

$$\leq \frac{\delta p}{2} \|\widetilde{\boldsymbol{X}}\|_F^2 + \frac{4\pi n}{m}(\|\boldsymbol{A}\| + np)$$

$$\leq \frac{\delta p}{2} \left(\|\boldsymbol{X}\|_F^2 + \frac{4\pi n^2}{m} \right) + \frac{4\pi n}{m}(\|\boldsymbol{A}\| + np)$$

$$\leq \frac{\delta p}{2} \|\boldsymbol{X}\|_F^2 + \frac{2\pi n^2 p}{m} \left(2 + \delta + \sqrt{\frac{2\gamma(1-p)\log n}{np}} + \frac{2\gamma \log n}{3np} \right)$$

for any $\boldsymbol{\theta}$, where (5.7), (5.14), and (5.15) have been used. As a result, we have $|f(\boldsymbol{\theta})| \leq \delta p \|\boldsymbol{X}\|_F^2$ provided that the second term above is bounded by $\frac{n^2 \delta p}{4}$, i.e.,

$$(5.16) \qquad m \geq \frac{8\pi}{\delta} \left(2 + \delta + \sqrt{\frac{2\gamma(1-p)\log n}{np}} + \frac{2\gamma \log n}{3np} \right).$$

In conclusion, $|\langle \boldsymbol{A} - p\boldsymbol{J}_n, \boldsymbol{Q}\boldsymbol{Q}^\top \circ \boldsymbol{Q}\boldsymbol{Q}^\top \rangle| \leq \delta p \|\boldsymbol{X}\|_F^2$ holds with probability at least

$$1 - 2\exp\left(n\log\left(\frac{100}{\delta}\right) - \frac{n^2 \delta^2 p}{64(1-p) + 6\delta} \right).$$

Following a similar procedure, we are also able to prove that

$$|\langle \boldsymbol{A} - p\boldsymbol{J}_n, \boldsymbol{Q}\boldsymbol{Q}^\top \rangle| \leq \delta p \|\boldsymbol{Q}\boldsymbol{Q}^\top\|_F^2$$

holds for all $\boldsymbol{Q}$ uniformly with probability at least

$$1 - 2\exp\left(n\log\left(\frac{100}{\delta}\right) - \frac{n^2 \delta^2 p}{64(1-p) + 6\delta} \right).$$

For completeness, we give a sketch of the proof below.

Define $g(\boldsymbol{\theta}) := \langle \boldsymbol{A} - p\boldsymbol{J}_n, \boldsymbol{Q}\boldsymbol{Q}^\top \rangle$ and it is straightforward to verify that

$$\mathbb{P}\left(|g(\boldsymbol{\theta})| \geq \frac{\delta p \|\boldsymbol{Q}\boldsymbol{Q}^\top\|_F^2}{2} \right) = \mathbb{P}\left(|\langle \boldsymbol{A}, \boldsymbol{X} \rangle - p\langle \boldsymbol{J}, \boldsymbol{X} \rangle| \geq \frac{\delta p \|\boldsymbol{Q}\boldsymbol{Q}^\top\|_F^2}{2} \right)$$

$$\leq 2\exp\left(-\frac{n^2 \delta^2 p}{64(1-p) + 6\delta} \right).$$

By using the same ϵ-net, $\mathcal{S}_{n,m}$, and taking the union bound over all elements in $\mathcal{S}_{n,m}$, we have

$$\mathbb{P}\left(\sup_{\widetilde{\boldsymbol{\theta}} \in \mathcal{S}_{n,m}} |g(\widetilde{\boldsymbol{\theta}})| \geq \frac{\delta p \|\boldsymbol{Q}\boldsymbol{Q}^\top\|_F^2}{2}\right) \leq 2\exp\left(n\log m - \frac{n^2\delta^2 p}{64(1-p)+6\delta}\right).$$

Moreover, it holds that

$$|g(\boldsymbol{\theta}) - g(\widetilde{\boldsymbol{\theta}})| \leq \frac{4\pi}{m}(n\|\boldsymbol{A}\| + n^2 p)$$

for any $\boldsymbol{\theta} \in [0,2\pi]^n$ and a point $\widetilde{\boldsymbol{\theta}}$ on the net $\mathcal{S}_{n,m}$ such that $\|\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}\|_\infty \leq \frac{\pi}{m}$. By using (5.15) and following the exact same steps as below (5.15), we can show that

$$|g(\boldsymbol{\theta})| \leq \delta p \|\boldsymbol{Q}\boldsymbol{Q}^\top\|_F^2$$

uniformly for all $\boldsymbol{\theta} \in [0,2\pi]^n$ for m chosen as in (5.16). This finishes the proof of the restricted isometry property. $\qquad\square$

Lemma 5.6 directly implies the following corollary.

COROLLARY 5.7 (restricted isometry property on manifold). *Let $\boldsymbol{A}$ be an Erdős–Rényi graph $G(n,p)$ with*

$$(5.17) \qquad p = \frac{32\gamma\log n}{n^{1/3}}, \qquad \delta = \frac{1}{n^{1/3}}.$$

Then the restricted isometry property

$$(5.18) \qquad \begin{aligned} -\delta p\|\boldsymbol{Q}\boldsymbol{Q}^\top\|_F^2 &\leq \langle \boldsymbol{A} - p\boldsymbol{J}_n, \boldsymbol{Q}\boldsymbol{Q}^\top \circ \boldsymbol{Q}\boldsymbol{Q}^\top\rangle \leq \delta p\|\boldsymbol{Q}\boldsymbol{Q}^\top\|_F^2, \\ -\delta p\|\boldsymbol{Q}\boldsymbol{Q}^\top\|_F^2 &\leq \langle \boldsymbol{A} - p\boldsymbol{J}_n, \boldsymbol{Q}\boldsymbol{Q}^\top\rangle \leq \delta p\|\boldsymbol{Q}\boldsymbol{Q}^\top\|_F^2 \end{aligned}$$

holds uniformly for all $\boldsymbol{Q}$ with probability at least

$$1 - 4\exp\left(n\left(\log(100n^{\frac{1}{3}}) - \frac{\gamma\log n}{2}\right)\right), \quad \gamma \geq 1.$$

From now on, we set p and δ as in (5.17).

Proof of Lemma 5.4. The second-order necessary condition implies

$$\langle \mathrm{ddiag}(\boldsymbol{A}\boldsymbol{Q}\boldsymbol{Q}^\top) - \boldsymbol{A}\circ\boldsymbol{Q}\boldsymbol{Q}^\top, \boldsymbol{Q}\boldsymbol{Q}^\top\rangle = \langle \boldsymbol{A}, \boldsymbol{Q}\boldsymbol{Q}^\top\rangle - \langle \boldsymbol{A}, \boldsymbol{Q}\boldsymbol{Q}^\top \circ \boldsymbol{Q}\boldsymbol{Q}^\top\rangle \geq 0,$$

where $(\boldsymbol{Q}\boldsymbol{Q}^\top)_{ii} = 1$ for $i \in [n]$.

Now we decompose the above inequality into

$$p\langle \boldsymbol{J}_n, \boldsymbol{Q}\boldsymbol{Q}^\top\rangle \geq p\langle \boldsymbol{J}_n, \boldsymbol{Q}\boldsymbol{Q}^\top \circ \boldsymbol{Q}\boldsymbol{Q}^\top\rangle + \langle p\boldsymbol{J}_n - \boldsymbol{A}, \boldsymbol{Q}\boldsymbol{Q}^\top\rangle + \langle \boldsymbol{A} - p\boldsymbol{J}_n, \boldsymbol{Q}\boldsymbol{Q}^\top \circ \boldsymbol{Q}\boldsymbol{Q}^\top\rangle.$$

From Lemma 5.6, we have a lower bound for the last two terms above, i.e.,

$$\begin{aligned} \langle p\boldsymbol{J}_n - \boldsymbol{A}, \boldsymbol{Q}\boldsymbol{Q}^\top\rangle &\geq -\delta p\|\boldsymbol{Q}\boldsymbol{Q}^\top\|_F^2, \\ \langle \boldsymbol{A} - p\boldsymbol{J}_n, \boldsymbol{Q}\boldsymbol{Q}^\top \circ \boldsymbol{Q}\boldsymbol{Q}^\top\rangle &\geq -\delta p\|\boldsymbol{Q}\boldsymbol{Q}^\top\|_F^2 \end{aligned}$$

for all $\boldsymbol{Q}$. Hence the correlation between $\boldsymbol{Q}^\top$ and $\mathbf{1}_n$ is bounded below by

$$\|\boldsymbol{Q}^\top\mathbf{1}_n\|^2 \geq (1-2\delta)\|\boldsymbol{Q}\boldsymbol{Q}^\top\|_F^2 \geq \frac{n^2(1-2\delta)}{2}.$$

Note $\boldsymbol{Q} \in \mathbb{R}^{n \times 2}$ is rank 2, and we let $\boldsymbol{Q} = [\boldsymbol{x} \ \boldsymbol{y}]$ with $\boldsymbol{x} \perp \boldsymbol{y}$. This is made possible by multiplying the 2×2 orthogonal matrix $\boldsymbol{R}$ from the right of $\boldsymbol{Q}$, where $\boldsymbol{R}$ diagonalizes $\boldsymbol{Q}^\top \boldsymbol{Q}$. Recall the first-order necessary condition (2.2) can be written as

$$\boldsymbol{A}\boldsymbol{x} \circ \boldsymbol{y} = \boldsymbol{A}\boldsymbol{y} \circ \boldsymbol{x}.$$

By letting $\boldsymbol{A} = p\boldsymbol{J}_n + \boldsymbol{\Delta}$, where $\boldsymbol{\Delta} = \boldsymbol{A} - p\boldsymbol{J}_n$, we have

$$p\mathbf{1}_n^\top \boldsymbol{x} \cdot \boldsymbol{y} - p\mathbf{1}_n^\top \boldsymbol{y} \cdot \boldsymbol{x} = (\boldsymbol{\Delta}\boldsymbol{y}) \circ \boldsymbol{x} - (\boldsymbol{\Delta}\boldsymbol{x}) \circ \boldsymbol{y}.$$

By taking the squared magnitude of the equation above, we arrive at

$$p^2(|\mathbf{1}_n^\top \boldsymbol{x}|^2 \|\boldsymbol{y}\|^2 + |\mathbf{1}_n^\top \boldsymbol{y}|^2 \|\boldsymbol{x}\|^2) \leq 2\|\boldsymbol{\Delta}\|^2(\|\boldsymbol{x}\|^2 + \|\boldsymbol{y}\|^2) = 2n\|\boldsymbol{\Delta}\|^2,$$

where $\boldsymbol{x} \perp \boldsymbol{y}$, $\|\boldsymbol{x}\|^2 + \|\boldsymbol{y}\|^2 = n$, and

$$\|\boldsymbol{\Delta}\boldsymbol{y} \circ \boldsymbol{x}\| = \|\operatorname{diag}(\boldsymbol{x})\boldsymbol{\Delta}\boldsymbol{y}\| \leq \|\boldsymbol{\Delta}\| \cdot \|\boldsymbol{y}\|.$$

Therefore, we have

$$|\mathbf{1}_n^\top \boldsymbol{x}|^2 \|\boldsymbol{y}\|^2 + |\mathbf{1}_n^\top \boldsymbol{y}|^2 \|\boldsymbol{x}\|^2 \leq \frac{2n\|\boldsymbol{\Delta}\|^2}{p^2}.$$

Without loss of generality, we assume $|\mathbf{1}_n^\top \boldsymbol{x}|^2 \geq n^2(1 - 2\delta)/4$, which follows from

$$|\mathbf{1}_n^\top \boldsymbol{x}|^2 + |\mathbf{1}_n^\top \boldsymbol{y}|^2 = \|\boldsymbol{Q}^\top \mathbf{1}_n\|^2 \geq \frac{n^2(1 - 2\delta)}{2}.$$

Therefore, $\|\boldsymbol{y}\|^2$ is bounded by

$$\|\boldsymbol{y}\|^2 \leq \frac{2n\|\boldsymbol{\Delta}\|^2}{p^2 |\mathbf{1}_n^\top \boldsymbol{x}|^2} \leq \frac{8\|\boldsymbol{\Delta}\|^2}{np^2(1 - 2\delta)},$$

and combined with $\|\boldsymbol{x}\|^2 + \|\boldsymbol{y}\|^2 = n$, we get

$$\|\boldsymbol{x}\|^2 \geq n\left(1 - \frac{8\|\boldsymbol{\Delta}\|^2}{n^2 p^2(1 - 2\delta)}\right).$$

Hence, these local minimizers $\boldsymbol{Q}$ would satisfy

$$\|\boldsymbol{Q}\boldsymbol{Q}^\top\|_F^2 = \|\boldsymbol{x}\|^4 + \|\boldsymbol{y}\|^4 \geq \|\boldsymbol{x}\|^4 \geq n^2\left(1 - \frac{8\|\boldsymbol{\Delta}\|^2}{n^2 p^2(1 - 2\delta)}\right)^2,$$

which yields a refined bound of $\|\boldsymbol{Q}^\top \mathbf{1}_n\|^2$, i.e.,

$$\|\boldsymbol{Q}^\top \mathbf{1}_n\|^2 \geq (1 - 2\delta)\|\boldsymbol{Q}\boldsymbol{Q}^\top\|_F^2 \geq (1 - 2\delta)n^2\left(1 - \frac{8\|\boldsymbol{\Delta}\|^2}{n^2 p^2(1 - 2\delta)}\right)^2$$

$$\geq (1 - 2\delta)n^2\left(1 - \frac{16\|\boldsymbol{\Delta}\|^2}{n^2 p^2(1 - 2\delta)}\right)$$

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

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

where $\delta = n^{-1/3}$, $p = 32\gamma n^{-1/3}\log n$, and

$$\frac{4\|\boldsymbol{\Delta}\|}{np} \leq 4\left(\sqrt{\frac{2\gamma(1 - p)\log n}{np}} + \frac{2\gamma\log n}{3np}\right)$$

$$\leq n^{-\frac{1}{3}}\sqrt{1 - p} + \frac{1}{12}n^{-\frac{2}{3}} \leq n^{-\frac{1}{3}}. \qquad \square$$

5.3. Supplementary technical details for the supporting examples.

The n-path graph. For the n-path graph, its adjacency matrix $\boldsymbol{A}$ has a simple form: $a_{ij} = 1$ if $|i - j| = 1$ and 0 otherwise. If we look at the first-order necessary condition, we have

$$0 = \sum_{j=1}^{n} a_{ij} \sin(\theta_i - \theta_j) = \begin{cases} \sin(\theta_i - \theta_{i-1}) + \sin(\theta_i - \theta_{i+1}) & \text{if } 2 \le i \le n - 1, \\ \sin(\theta_1 - \theta_2) & \text{if } i = 1, \\ \sin(\theta_{n-1} - \theta_n) & \text{if } i = n, \end{cases}$$

which implies that all critical points must belong to

$$\{\boldsymbol{\theta} : \theta_i = 0 \text{ or } \pi \mod 2\pi, \ 1 \le i \le n\}$$

up to a global shift. However, the only global minimum of $E(\boldsymbol{\theta})$ w.r.t. the n-path is $\boldsymbol{\theta}_0$. Note that for these critical points, the corresponding $\boldsymbol{Q}$ is rank 1 and has entries ± 1. Now let $\boldsymbol{Q} = \mathbf{1}_\Gamma - \mathbf{1}_{\Gamma^c}$, where $\Gamma = \{i : \theta_i = 0\}$ and $\Gamma^c = \{i : \theta_i = \pi\}$, and $\mathbf{1}_\Gamma \in \mathbb{R}^n$ is the corresponding indicator function of Γ. Consider the inner product between the Hessian matrix and $\boldsymbol{Q}\boldsymbol{Q}^\top$, and it holds that

$$\begin{aligned} \langle \mathrm{ddiag}(\boldsymbol{A}\boldsymbol{Q}\boldsymbol{Q}^\top) - \boldsymbol{A} \circ \boldsymbol{Q}\boldsymbol{Q}^\top, \boldsymbol{Q}\boldsymbol{Q}^\top \rangle &= \mathrm{Tr}(\boldsymbol{A}\boldsymbol{Q}\boldsymbol{Q}^\top) - \langle \boldsymbol{A}, \boldsymbol{J}_n \rangle \\ &= \mathbf{1}_\Gamma^\top \boldsymbol{A} \mathbf{1}_\Gamma + \mathbf{1}_{\Gamma^c}^\top \boldsymbol{A} \mathbf{1}_{\Gamma^c} - 2 \mathbf{1}_{\Gamma^c}^\top \boldsymbol{A} \mathbf{1}_\Gamma - \langle \boldsymbol{A}, \boldsymbol{J}_n \rangle \\ &= -4 \mathbf{1}_{\Gamma^c}^\top \boldsymbol{A} \mathbf{1}_\Gamma \le 0 \end{aligned}$$

since $\boldsymbol{Q}\boldsymbol{Q}^\top \circ \boldsymbol{Q}\boldsymbol{Q}^\top = \boldsymbol{J}_n$. The inequality above is strict unless either Γ or Γ^c is empty due to the connectivity of the n-path. Therefore, for any critical point $\boldsymbol{\theta}$, where $\boldsymbol{\theta} \ne \boldsymbol{\theta}_0$, it cannot be a local minimum because the Hessian matrix is not positive semidefinite.

Wiley–Strogatz–Girvan networks. To make the presentation more self-contained, we briefly discuss the network in [49] in the language of linear algebra and discrete harmonic analysis, and show that the twisted state is a local minimizer when μ is approximately smaller than 0.68. If each node is only connected to its k nearest neighbors, the corresponding adjacency matrix, denoted by $\boldsymbol{A}_k$, is a circulant matrix generated by

$$[0, \underbrace{1, 1, \ldots, 1}_{k \text{ ones}}, 0, \ldots, 0, \underbrace{1, \ldots, 1}_{k \text{ ones}}]^\top \in \mathbb{R}^n,$$

as shown in Figure 3. In particular, if $k = 1$, then

$$\boldsymbol{A} = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 & 1 \\ 1 & 0 & 1 & \cdots & 0 & 0 \\ 0 & 1 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 1 \\ 1 & 0 & 0 & \cdots & 1 & 0 \end{bmatrix}.$$

An important fact is that all circulant matrices are diagonalizable under a Fourier basis; see [22, Chapter 4].

Let $\omega = e^{2\pi \mathrm{i}/n}$ be the root of unity and $\boldsymbol{z} := [1, \omega, \ldots, \omega^{n-1}]^\top$. Since $\boldsymbol{z}$ is an eigenvector of $\boldsymbol{A}_k$ and $\boldsymbol{A}_k$ has real eigenvalues due to its symmetry, we have

$$\mathrm{Im}(\mathrm{diag}(\overline{\boldsymbol{z}}) \boldsymbol{A}_k \boldsymbol{z}) = \lambda \mathrm{Im}(\mathrm{diag}(\overline{\boldsymbol{z}}) \boldsymbol{z}) = 0.$$

By letting $\boldsymbol{x} = \mathrm{Re}(\boldsymbol{z})$ and $\boldsymbol{y} = \mathrm{Im}(\boldsymbol{z})$ and substituting $\boldsymbol{z} = \boldsymbol{x} + \mathrm{i}\boldsymbol{y}$ into the equation above, it holds that

$$\mathrm{Im}(\mathrm{diag}(\boldsymbol{x} - \mathrm{i}\boldsymbol{y})\boldsymbol{A}_k(\boldsymbol{x} + \mathrm{i}\boldsymbol{y})) = \mathrm{diag}(\boldsymbol{x})\boldsymbol{A}_k\boldsymbol{y} - \mathrm{diag}(\boldsymbol{y})\boldsymbol{A}_k\boldsymbol{x} = 0,$$

which indicates that $\widetilde{\boldsymbol{\theta}} \coloneqq \{2\pi l/n\}_{l=1}^{n}$ is a critical point of $E(\boldsymbol{\theta})$.

It suffices to compute the Hessian matrix of $E(\boldsymbol{\theta})$ at $\widetilde{\boldsymbol{\theta}}$. Note that $\boldsymbol{Q}\boldsymbol{Q}^{\top} = \mathrm{Re}(\boldsymbol{z}\boldsymbol{z}^{*})$, and hence $\boldsymbol{A}_k \circ \boldsymbol{Q}\boldsymbol{Q}^{\top} = \mathrm{Re}(\mathrm{diag}(\boldsymbol{z})\boldsymbol{A}_k\,\mathrm{diag}(\overline{\boldsymbol{z}}))$ is a symmetric circulant matrix and so is

$$\boldsymbol{L}_k \coloneqq \mathrm{ddiag}(\boldsymbol{A}_k\boldsymbol{Q}\boldsymbol{Q}^{\top}) - \boldsymbol{A}_k \circ \boldsymbol{Q}\boldsymbol{Q}^{\top}.$$

Therefore, $\boldsymbol{L}_k$ is also diagonalizable under the discrete Fourier transform, and all of its eigenvalues are

$$\lambda_l(\boldsymbol{L}_k) = 2\sum_{j=1}^{k}\cos\left(\frac{2\pi j}{n}\right) - 2\sum_{j=1}^{k}\cos\left(\frac{2\pi j}{n}\right)\cos\left(\frac{2\pi(l-1)j}{n}\right), \quad 1 \le l \le n.$$

In particular, if $l = 2$, the second smallest eigenvalue of the Hessian is

$$\lambda_2(\boldsymbol{L}_k) = 2\sum_{j=1}^{k}\cos\left(\frac{2\pi j}{n}\right) - 2\sum_{j=1}^{k}\cos^2\left(\frac{2\pi j}{n}\right) = \frac{2}{n}\langle \boldsymbol{A}_k, \boldsymbol{Q}\boldsymbol{Q}^{\top} - \boldsymbol{Q}\boldsymbol{Q}^{\top} \circ \boldsymbol{Q}\boldsymbol{Q}^{\top}\rangle.$$

Therefore, $\widetilde{\boldsymbol{\theta}}$ is a local minimizer since $E(\boldsymbol{\theta})$ if $\lambda_2(\boldsymbol{L}_k) > 0$ and this is guaranteed by $k \le 0.34n$ approximately, as suggested in Figure 4.

Acknowledgment. S.L. and R.X. thank Tianqi Wu for a very fruitful and inspiring discussion and for bringing to their attention reference [46].

REFERENCES

[1] E. Abbe, A. S. Bandeira, A. Bracher, and A. Singer, *Decoding binary node labels from censored edge measurements: Phase transition and efficient recovery*, IEEE Trans. Network Sci. Eng., 1 (2014), pp. 10–22.

[2] E. Abbe, L. Massoulié, A. Montanari, A. Sly, and N. Srivastava, *Group synchronization on grids*, Math. Stat. Learn. (2018), DOI: 10.4171/MSL/6.

[3] P.-A. Absil, R. Mahony, and R. Sepulchre, *Optimization Algorithms on Matrix Manifolds*, Princeton University Press, Princeton, NJ, 2009.

[4] J. A. Acebrón, L. L. Bonilla, C. J. P. Vicente, F. Ritort, and R. Spigler, *The Kuramoto model: A simple paradigm for synchronization phenomena*, Rev. Modern Phys., 77 (2005), 137.

[5] A. Arenas, A. Díaz-Guilera, J. Kurths, Y. Moreno, and C. Zhou, *Synchronization in complex networks*, Phys. Rep., 469 (2008), pp. 93–153.

[6] A. S. Bandeira, N. Boumal, and A. Singer, *Tightness of the maximum likelihood semidefinite relaxation for angular synchronization*, Math. Program., 163 (2017), pp. 145–167.

[7] A. S. Bandeira, N. Boumal, and V. Voroninski, *On the low-rank approach for semidefinite programs arising in synchronization and community detection*, in 29th Annual Conference on Learning Theory, Proc. Mach. Learn. Res. 49, 2016, pp. 361–382; available at http://proceedings.mlr.press/v49/.

[8] N. Boumal, *Nonconvex phase synchronization*, SIAM J. Optim., 26 (2016), pp. 2355–2377.

[9] S. Burer and R. D. Monteiro, *A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization*, Math. Program., 95 (2003), pp. 329–357.

[10] S. Burer and R. D. Monteiro, *Local minima and convergence in low-rank semidefinite programming*, Math. Program., 103 (2005), pp. 427–444.

[11] E. J. Candès and B. Recht, *Exact matrix completion via convex optimization*, Found. Comput. Math., 9 (2009), pp. 717–772.

[12] J. Chen and X. Li, *Memory-efficient Kernel PCA via Partial Matrix Sampling and Nonconvex Optimization: A Model-free Analysis of Local Minima*, preprint, https://arxiv.org/abs/1711.01742, 2017.

[13] Y. Chen and E. Candès, *The projected power method: An efficient algorithm for joint alignment from pairwise differences*, Comm. Pure Appl. Math., 71 (2018), pp. 1648–171.

[14] Y. Cheng and R. Ge, *Non-convex matrix completion against a semi-random adversary*, in Proceedings of the 31st Conference On Learning Theory, Proc. Mach. Learn. Res. 75, pp. 1362–1394; available at http://proceedings.mlr.press/v75/.

[15] H. Chiba, G. S. Medvedev, and M. S. Mizuhara, *Bifurcations in the Kuramoto model on graphs*, Chaos, 28 (2018), 073109.

[16] F. R. Chung, *Spectral Graph Theory*, CBMS Reg. Conf. Ser. Math. 92, American Mathematical Society, Providence, RI, 1997.

[17] F. Dörfler and F. Bullo, *Synchronization in complex networks of phase oscillators: A survey*, Automatica, 50 (2014), pp. 1539–1564.

[18] F. Dörfler, M. Chertkov, and F. Bullo, *Synchronization in complex oscillator networks and smart grids*, Proc. Natl. Acad. Sci., 110 (2013), pp. 2005–2010.

[19] U. Feige and J. Kilian, *Heuristics for semirandom graph problems*, J. Comput. System Sci., 63 (2001), pp. 639–671.

[20] R. Ge, J. D. Lee, and T. Ma, *Matrix completion has no spurious local minimum*, in Adv. Neural Inf. Process. Syst. 29, Curran Associates, Red Hook, NY, 2016, pp. 2973–2981.

[21] R. Ge and T. Ma, *On the optimization landscape of tensor decompositions*, in Adv. Neural Inf. Process. Syst. 30, Curran Associates, Red Hook, NY, 2017, pp. 3653–3663.

[22] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 3rd ed., The Johns Hopkins University Press, Baltimore, MD, 1996.

[23] T. Ichinomiya, *Frequency synchronization in a random oscillator network*, Phys. Rev. E, 70 (2004), 026116.

[24] A. Jadbabaie, N. Motee, and M. Barahona, *On the stability of the Kuramoto model of coupled nonlinear oscillators*, in Proceedings of the 2004 American Control Conference, Vol. 5, IEEE Press, Piscataway, NJ, 2004, pp. 4296–4301.

[25] M. Journée, F. Bach, P.-A. Absil, and R. Sepulchre, *Low-rank optimization on the cone of positive semidefinite matrices*, SIAM J. Optim., 20 (2010), pp. 2327–2351.

[26] R. H. Keshavan, S. Oh, and A. Montanari, *Matrix completion from a few entries*, in Proceedings of the 2009 IEEE International Symposium on Information Theory, IEEE Press, Piscataway, NJ, 2009, pp. 324–328.

[27] Y. Kuramoto, *Self-entrainment of a population of coupled non-linear oscillators*, in International Symposium on Mathematical Problems in Theoretical Physics, Lecture Notes in Phys. 39, Springer, Berlin, 1975, pp. 420–422.

[28] Y. Kuramoto, *Chemical Oscillations, Waves, and Turbulence*, Springer Ser. Synergetics 19, Springer, Berlin, 1984.

[29] Q. Li, Z. Zhu, and G. Tang, *The non-convex geometry of low-rank matrix optimization*, Inf. Inference, 8 (2018), pp. 51–96.

[30] M. Lopes, E. Lopes, S. Yoon, J. Mendes, and A. Goltsev, *Synchronization in the random-field Kuramoto model on complex networks*, Phys. Rev. E, 94 (2016), 012308.

[31] G. S. Medvedev and X. Tang, *Stability of twisted states in the Kuramoto model on Cayley and random graphs*, J. Nonlinear Sci., 25 (2015), pp. 1169–1208.

[32] D. Mehta, N. S. Daleo, F. Dörfler, and J. D. Hauenstein, *Algebraic geometrization of the Kuramoto model: Equilibria and stability analysis*, Chaos, 25 (2015), 053103.

[33] A. Moitra, W. Perry, and A. S. Wein, *How robust are reconstruction thresholds for community detection?*, in Proceedings of the 48th Annual ACM Symposium on Theory of Computing, Association for Computing Machinery, New York, 2016, pp. 828–841.

[34] J. Peña Ramirez, L. A. Olvera, H. Nijmeijer, and J. Alvarez, *The sympathy of two pendulum clocks: Beyond Huygens' observations*, Sci. Rep., 6 (2016), http://dx.doi.org/10.1038/srep23580.

[35] A. Perry, A. S. Wein, A. S. Bandeira, and A. Moitra, *Message-passing algorithms for synchronization problems over compact groups*, Comm. Pure Appl. Math. (2016), https://doi.org/10.1002/cpa.21750.

[36] B. Recht, *A simpler approach to matrix completion*, J. Mach. Learn. Res., 12 (2011), pp. 3413–3430.

[37] B. Recht, M. Fazel, and P. Parrilo, *Guaranteed minimum rank solutions of matrix equations via nuclear norm minimization*, SIAM Rev., 52 (2010), pp. 471–501.

[38] F. A. Rodrigues, T. K. D. Peron, P. Ji, and J. Kurths, *The Kuramoto model in complex networks*, Phys. Rep., 610 (2016), pp. 1–98.

[39] M. Soltanolkotabi, A. Javanmard, and J. D. Lee, *Theoretical insights into the optimization landscape of over-parameterized shallow neural networks*, IEEE Trans. Inform. Theory, 65 (2019), pp. 742–769.

[40] S. H. Strogatz, *From Kuramoto to Crawford: Exploring the onset of synchronization in populations of coupled oscillators*, Phys. D, 143 (2000), pp. 1–20.

[41] S. H. Strogatz, *Exploring complex networks*, Nature, 410 (2001), pp. 268–276.

[42] J. Sun, Q. Qu, and J. Wright, *When Are Nonconvex Problems Not Scary?*, preprint, https://arxiv.org/abs/1510.06096, 2015.

[43] J. Sun, Q. Qu, and J. Wright, *Complete dictionary recovery over the sphere* I: *Overview and the geometric picture*, IEEE Trans. Inform. Theory, 63 (2017), pp. 853–884.

[44] J. Sun, Q. Qu, and J. Wright, *A geometric analysis of phase retrieval*, Found. Comput. Math., 18 (2018), pp. 1131–1198.

[45] R. Sun and Z.-Q. Luo, *Guaranteed matrix completion via non-convex factorization*, IEEE Trans. Inform. Theory, 62 (2016), pp. 6535–6579.

[46] R. Taylor, *There is no non-zero stable fixed point for dense networks in the homogeneous Kuramoto model*, J. Phys. A, 45 (2012), 055102.

[47] J. A. Tropp, *User-friendly tail bounds for sums of random matrices*, Found. Comput. Math., 12 (2012), pp. 389–434.

[48] R. van der Hofstad, *Random Graphs and Complex Networks*, Vol. 1, Camb. Ser. Stat. Probab. Math., Cambridge University Press, Cambridge, 2016, https://doi.org/10.1017/9781316779422.

[49] D. A. Wiley, S. H. Strogatz, and M. Girvan, *The size of the sync basin*, Chaos, 16 (2006), 015103.

[50] Y. Zhang, H.-W. Kuo, and J. Wright, *Structured local minima in sparse blind deconvolution*, in Adv. Neural Inf. Process. Syst. 31, Curran Associates, Red Hook, NY, 2018, pp. 2322–2331.

[51] Y. Zhong and N. Boumal, *Near-optimal bounds for phase synchronization*, SIAM J. Optim., 28 (2018), pp. 989–1016.