EXACT RECOVERY OF COMMUNITY DETECTION IN K-PARTITE GRAPH MODELS

ZHONGYANG LI

ABSTRACT. We study the vertex classification problem on a graph whose vertices are in k ($k \geq 2$) different communities, edges are only allowed between distinct communities, and the number of vertices in different communities are not necessarily equal. The observation is a weighted adjacency matrix, perturbed by a scalar multiple of the Gaussian Orthogonal Ensemble (GOE), or Gaussian Unitary Ensemble (GUE) matrix. For the exact recovery of the maximum likelihood estimation (MLE) with various weighted adjacency matrices, we prove sharp thresholds of the intensity σ of the Gaussian perturbation. These weighted adjacency matrices may be considered as natural models for the electric network. Surprisingly, these thresholds of σ do not depend on whether the sample space for MLE is restricted to such classifications that the number of vertices in each group is equal to the true value. In contrast to the $\mathbb{Z}_2$-synchronization, a new complex version of the semi-definite programming (SDP) is designed to efficiently implement the community detection problem when the number of communities k is greater than 2, and a common region (independent of k) for σ such that SDP exactly recovers the true classification is obtained.

1. INTRODUCTION

Most graphs of interest display community structure, i.e., their vertices are organized into groups, called communities, clusters or modules. In some cases, edges are concentrated within groups. For example, vertices of a graph may represent scientists, edges join coauthors. Each group consists of vertices representing scientists working on the same research topic, where collaborations are more frequent. Likewise, communities could represent molecules with similar structure in molecule interaction networks among which interactions are more likely, groups of friends in social networks who communicate more often, websites on similar topics in the web graph where there are more hyperlinks in between, and so on. In some other cases, edges may only be possible between vertices in distinct groups. For instance, in an electrical network, electrical current can only be observed between two sites with different electrical potential; commercial trades can only occur when two individuals own different goods. Identifying communities may offer insight on how the network is organized. It allows us to focus on regions having some degree of similarity within the graph. It helps to classify the vertices, based on their role with respect to the communities they belong to. For instance we can distinguish vertices in the interior of their clusters from vertices at the boundary of the clusters, which may act as brokers between the modules and, in that case, could play a crucial role both in holding the modules together and in the dynamics of spreading processes across the network.

Identifying different communities in the stochastic block model is a central topic in many fields of science and technology; see [1] for a summary. A lot of spectacular work has been done when the graph has two equal-sized communities, see, for example, [7, 8, 2] for an incomplete list. Community detection with two equal-sized communities has also been studied on hyper-graphs, see [6].

In this paper, we instead study the community detection on a graph in which there are $k(k \geq 2)$ distinct clusters, not necessarily equal-sized, and edges are only allowed between vertices in different communities. This corresponds to the famous k-partite graph in graph theory. The observation is a weighted adjacency matrix, perturbed by a σ-multiple of the Gaussian Orthogonal Ensemble (GOE), or the Gaussian Unitary Ensemble (GUE) matrix. Here σ is a positive number representing the intensity of the Gaussian perturbation. These weighted adjacency matrices, as will be explained later, may be considered as natural models for the electric network. Given such observations, we apply the maximum likelihood estimation (MLE) to determine which vertex belongs to which group, or community. We obtain a division, or an assignment of vertices of the graph into communities by the MLE, if this assignment of vertices is the same as the true assignment of vertices into communities for every vertex, we say that the MLE exactly recovers the true community structure of the graph, or the exact recovery occurs for the MLE.

The main goal of the paper is to investigate the condition when the MLE exactly recovers the true community structure of the graph. We prove sharp phase transition results with respect to the intensity σ of the Gaussian perturbation for the exact recovery of the MLE. More precisely, we explicitly find the critical value for σ, such that if σ is less than the critical value, then as the size of the graph goes to infinity, with probability tending to 1 exact recovery occurs. On the other hand, if σ is greater than the critical value, then as the size of the graph goes to infinity, with probability tending to 0 exact recovery occurs. Interestingly, the threshold, or critical value, of σ does not depend on whether or not we restrict the sample space for MLE to those classifications in which the number of vertices of each group, or community, is the same as the true value. These results are obtained by analyzing the Gaussian distribution through various inequalities.

Semidefinite programming (SDP) is one of the most exciting developments in mathematical programming in the 1990's. SDP has applications in diverse fields including, but not restricted to, traditional convex constrained optimization, control theory, and combinatorial optimization. A linear programming (LP) problem is one in which we wish to maximize or minimize a linear objective function of real variables over a polytope. In SDP, we instead use real-valued vectors and are allowed to take the dot product of vectors; non-negativity constraints on real variables in LP are replaced by semi-definiteness constraints on matrix variables in SDP. Because SDP is solvable via interior point methods, most of these applications can usually be solved very efficiently in practice as well as in theory.

It is well-known that the community detection problem with $k = 2$ equal-sized communities may be efficiently solved by a semi-definite programming algorithm; see, for instance, [4, 5]. When there are $k \geq 3$ different communities, we can design a "complex version" of the semi-definite programming for efficient recovery. The idea is to relax the constraint

on the rank of the optimal solution, solve the optimization problem on a larger space of the semi-definite matrices, and then achieve efficient recovery. We also obtain an interval of σ to guarantee the exact recovery of the SDP, by applying the celebrated result of the Tracy-Widom fluctuation of the maximal eigenvalue of the GOE matrix; see [9].

2. Main Results.

In this section, we state the main results proved in the paper. We first discuss the basic definition and notation of the k-partite graph, where k is a positive integer with value at least 2.

A k-partite graph $G = (V, E)$ is a graph whose vertices can be colored in k different colors such that any two vertices of the same color cannot be adjacent, or joined by an edge. Assume $V = [n] := \{1, 2, \ldots, n\}$ is the vertex set of G. Define the set of colors

$$R_k := \{c_1, \ldots, c_k\} \subset \mathbb{R}$$

to be a set consisting of k distinct real numbers representing k different colors.

Let $x : [n] \to R_k$ be a mapping from the set of vertices to the set of colors, i.e., it assigns a unique color in R_k to each vertex in $[n]$. Such a mapping x is also called a color assignment mapping of a graph. For $1 \le i \le k$, let

$$x^{-1}(c_i) = \{j \in [n] : y(j) = c_i\};$$

that is, $y^{-1}(c_i)$ is the set consisting of all the vertices with color c_i under the mapping y. For each color $c_i \in R_k$ $(1 \le i \le k)$, let

$$n_i(x) := \left|x^{-1}(c_i)\right|$$

In other words, $n_i(x)$ is the total number of vertices in $[n]$ with color c_i under the mapping x. It is straightforward to see that $n_i(x)$'s are positive integers satisfying

$$(2.1) \qquad \sum_{i=1}^{k} n_i(x) = n.$$

Given these definitions, to determine the colors of all the vertices of a graph is the same as to identify the color assignment mapping of the the graph. In vertex-color-detection problems that will be discussed later, we shall find color assignment mappings of graphs from the following spaces of color assignment mappings:

(1) For each positive integer n, let $n_1 \ge n_2 \ge \ldots \ge n_k$ be fixed and satisfy (2.1). Let $\Omega_{n_1, \ldots, n_k}$ be the set of all the color assignment mappings under which the number of vertices with color c_i is exactly n_i for each $1 \le i \le k$; that is,

$$\Omega_{n_1, \ldots, n_k} = \{x : [n] \to R_k | \, |x^{-1}(c_i)| = n_i, \, \forall \, 1 \le i \le k\}.$$

(2) Let Ω be the set of all the color assignment mappings; that is,

$$\Omega = \{x : [n] \to R_k\}.$$

(3) Let $c > 0$. Let Ω_c be the set of all the color assignment mappings in Ω such that the number of vertices in each color is at least cn, i.e.

$$\Omega_{n_1,\ldots,n_k} = \{\, x : [n] \to R_k | : \min_{j \in [k]} \left| x^{-1}(c_j) \right| \geq cn \}.$$

2.1. **Real Weighted Adjacency Matrix with Gaussian Perturbation.** We first consider the community detection problem when the observation is a real weighted adjacency matrix with Gaussian perturbation.

In Theorems 2.1 and 2.2, we observe different weighted adjacency matrices for a k-partite graph, both of which are perturbed by a σ-multiple of a matrix with i.i.d. standard Gaussian entries. The weighted adjacency matrix in 2.1 can be considered as a natural model for an electrical network, where each one of the n vertices has one of the k distinct electric potentials. The weight of each (oriented) edge is the difference of electric potentials between its initial point and its terminal point. This difference in potentials is proportional to the intensity of electric current on the edge. Given the observation, the goal is to find the difference of electric potentials between each pair of vertices; then we can determine the electrical potential of each vertex up to an additive constant. We consider the probability of the exact recovery of MLE, and find a sharp threshold with respect to the intensity σ of the Gaussian perturbation. More precisely, there is a critical value σ_c depending on n, such that if $\lim_{n \to \infty} \frac{|\sigma|}{\sigma_c} < 1$, the limit of the probabilities for the exact recoveries of MLE as $n \to \infty$ is 1; while if $\lim_{n \to \infty} \frac{|\sigma|}{\sigma_c} > 1$, the limit of the probabilities for the exact recovery of MLE as $n \to \infty$ is 0.

Before stating Theorem 2.1, recall that the Frobenius norm of an $m \times m$ complex matrix $A = \{A_{i,j}\}_{i,j=1}^m \in \mathbb{C}^{m \times m}$ is defined by

$$\|A\|_F := \sqrt{\sum_{i,j=1}^m |A_{ij}|^2}$$

Theorem 2.1. *Let* $n_1 \geq \ldots \geq n_k$ *be the numbers of vertices in* k *different colors* $c_1, \ldots, c_k$, *respectively. For an arbitrary mapping* $x \in \Omega$, *let* $\mathbf{G}(x)$ *be the* $n \times n$ *square matrix whose entries are defined by*

$$(2.2) \qquad\qquad \mathbf{G}_{i,j}(x) = x(i) - x(j);$$

where $1 \leq i, j \leq n$. *Let* $y \in \Omega_{n_1,\ldots,n_k}$ *be the true color assignment function. Assume the observation is given by*

$$(2.3) \qquad\qquad \mathbf{T} = \mathbf{G}(y) + \sigma \mathbf{W}$$

where $\mathbf{W}$ *is a random* $n \times n$ *matrix with i.i.d. standard Gaussian entries, and* $\sigma \in \mathbb{R}$ *is deterministic. Let* k *and* $\{c_1, \ldots, c_k\}$ *be fixed as* $n \to \infty$. *Assume that there exists a constant* $c > 0$ *independent of* n, *such that* $\frac{n_k}{n} \geq c$ *for all* n, *Let*

$$(2.4) \qquad \hat{y} \;\; = \;\; \mathrm{argmin}_{x \in \Omega_{n_1,\ldots,n_k}} \|\mathbf{T} - \mathbf{G}(x)\|_F^2;$$

$$(2.5) \qquad \check{y} \;\; = \;\; \mathrm{argmin}_{x \in \Omega} \|\mathbf{T} - \mathbf{G}(x)\|_F^2.$$

We have

(1) *Assume there exists a constant $\delta > 0$ independent of n, such that*

(2.6)
$$\sigma^2 < \frac{(1-\delta)n \min_{1 \leq i < j \leq k}(c_i - c_j)^2}{4 \log n}.$$

Then

$$\lim_{n \to \infty} \Pr(\hat{y} = y) = 1; \text{ and } \lim_{n \to \infty} \Pr(\check{y} = y) = 1.$$

(2) *Assume there exists a constant $\delta > 0$ independent of n, such that*

(2.7)
$$\sigma^2 > \frac{(1+\delta)n \min_{1 \leq i < j \leq k}(c_i - c_j)^2}{4 \log n}.$$

Then

$$\lim_{n \to \infty} \Pr(\hat{y} = y) = 0; \text{ and } \lim_{n \to \infty} \Pr(\check{y} = y) = 0.$$

Note that $\hat{y}$ (resp. $\check{y}$) are actually the maximum likelihood estimation (MLE) of the true value y in the sample space $\Omega_{n_1,\ldots,n_k}$ (resp. Ω) with respect to the given observation $\mathbf{T}$, since for each given color assignment mapping $x \in \Omega$, the probability density at each $\mathbf{T}$ is proportional to $e^{-\frac{\|\mathbf{T} - \mathbf{G}(x)\|_F^2}{2}}$.

More general versions of Theorem 2.1 are proved in Sections 3 and 4; see Propositions 3.5, 3.7, 4.3, 4.4. In these propositions, we also allow the total number of colors k to change with n. It is straight forward to check that when k and R_k are fixed as $n \to \infty$, Theorem 2.1 is a special case of these propositions.

In Theorem 2.2, we observe the uniformly-weighted adjacency matrix for the undirected k-partite graph, perturbed by a noise which is a σ-multiple of a matrix with i.i.d. standard Gaussian entries. Given the observation, the goal is to determine whether two vertices have the same color or not. For $x, y \in \Omega$, we say x and y are equivalent if for all $i, j \in [n]$, $x(i) = x(j)$ if and only fi $y(i) = y(j)$. We write $x \in C(y)$ if x and y are equivalent. Now we only need the algorithm to find a color assignment mapping which is equivalent to the true color assignment mapping. Again we find a sharp threshold for the probability of the exact recovery of the MLE with respect to the parameter σ.

Theorem 2.2. *Let $n_1 \geq \ldots \geq n_k$ be the numbers of vertices in the k different colors, respectively. For an arbitrary mapping $x \in \Omega$, let $\mathbf{K}(x)$ be the $n \times n$ square matrix with entries defined by*

(2.8)
$$\mathbf{K}_{i,j}(x) = \begin{cases} 1 & \text{if } x(i) \neq x(j) \\ 0 & \text{if } x(i) = x(j) \end{cases};$$

where $1 \leq i, j \leq n$. Let $y \in \Omega_{n_1,\ldots,n_k}$ be the true color assignment function. Assume the observation is given by

(2.9)
$$\mathbf{R} = \mathbf{K}(y) + \sigma \mathbf{W}$$

where $\mathbf{W}$ is a random $n \times n$ matrix with i.i.d. standard Gaussian entries as before. Let the total number of colors k and the set of all colors $\{c_1, \ldots, c_k\}$ be fixed as $n \to \infty$. Assume

there exists a constant $c > 0$, such that $n_k \geq cn$ for all n. Let

$$(2.10) \qquad \tilde{y} = \operatorname{argmin}_{x \in \Omega_{\frac{2c}{3}}} \|\mathbf{R} - \mathbf{K}(x)\|_F^2$$

$$(2.11) \qquad \overline{y} = \operatorname{argmin}_{x \in \Omega_{n_1,\ldots,n_k}} \|\mathbf{R} - \mathbf{K}(x)\|_F^2.$$

We have

(1) Assume there exists $\delta > 0$, such that

$$(2.12) \qquad \sigma^2 < \frac{(1-\delta)(n_k + n_{k-1})}{4 \log n}$$

Then

$$\lim_{n \to \infty} \Pr(\tilde{y} \in C(y)) = 1; \text{ and } \lim_{n \to \infty} \Pr(\overline{y} \in C(y)) = 1.$$

(2) Assume there exists $\delta > 0$, such that

$$(2.13) \qquad \sigma^2 > \frac{(1+\delta)(n_k + n_{k-1})}{4 \log n}$$

then

$$\lim_{n \to \infty} \Pr(\tilde{y} \in C(y)) = 0; \text{ and } \lim_{n \to \infty} \Pr(\overline{y} \in C(y)) = 0.$$

More general versions of Theorem 2.2 are proved in Sections 5 and 6; see Propositions 5.7, 5.8, 6.7, 6.8. In these propositions, we also allow the total number of colors k to change with n. It is straight forward to check that when k and R_k are fixed as $n \to \infty$, Theorem 2.2 is a special case of these propositions. The proofs of Theorems 2.1 and 2.2 are based on various inequalities of Gaussian distributions.

From these two theorems, we can see that we may either choose the sample space for MLE as all the possible assignments of k colors, or potentials, to n vertices in Theorem 2.1 (all the possible classifications of n vertices in k distinct groups in Theorem 2.2), or choose the sample space for MLE to be restricted on all the classifications such that the number of vertices of each type coincides with that of the true value - either way we obtain the same threshold.

2.2. **Complex Unitary Matrix with Gaussian Perturbation.** Now we consider the community detection problem with $k \geq 2$ different communities when the observation is a complex unitary matrix with Gaussian perturbation. Community detection problems with such an observation matrix may be efficiently recovered by the SDP.

Let $d_1, \ldots, d_k \in [0, 2\pi)$ be k distinct real numbers. Let $\mathbf{i}$ satisfy $\mathbf{i}^2 = -1$ be the imaginary unit. Let $x : [n] \to \{e^{\mathbf{i}d_1}, \ldots, e^{\mathbf{i}d_k}\}$ be a mapping which assigns each vertex in $[n]$ a unique color represented by a complex number of modulus 1. Let Θ be the set consisting of all such mappings, that is

$$\Theta := \{x : [n] \to \{e^{\mathbf{i}d_1}, \ldots, e^{\mathbf{i}d_k}\}\}$$

For a mapping $x \in \Theta$, let $\mathbf{P}(x)$ be an $n \times n$ matrix whose entries are given by

$$\mathbf{P}_{a,b}(x) = x(a)\overline{x(b)} = e^{\operatorname{Log}[x(a)] - \operatorname{Log}[x(b)]}, \qquad 1 \leq a, b \leq n,$$

where $\overline{x(b)}$ is the complex conjugate of $x(b)$ and $\mathrm{Log}[\cdot]$ is the principal branch of the complex logarithmic function.

For each $x \in \Theta$, if we consider $\mathbf{x}$ as an $n \times 1$ vector given by

$$\mathbf{x} = (x(1), \ldots, x(n))^t,$$

then

$$\mathbf{P}(x) = \mathbf{x}\overline{\mathbf{x}}^t;$$

which is a rank-1 positive semi-definite Hermitian matrix whose diagonal entries are 1.

We may further restrict the MLE to a subspace of Θ satisfying the following assumptions.

Assumption 2.3.   *(1) The number of vertices in each color is the same. In particular, this implies that the total number of vertices n, is an integer multiple of the total number of colors k.*

(2) $e^{\mathrm{i}d_1}, \ldots, e^{\mathrm{i}d_k}$ are the kth roots of unity. Without loss of generality, assume that $d_l = \frac{2(l-1)\pi}{k}$, for $l = 1, \ldots, k$.

Let Θ_A be the subset of Θ consisting of all the mappings satisfying the above two assumptions. That is,

$$\Theta_A := \left\{ x \in \Theta : \left| x^{-1}\left(e^{\mathrm{i}d_j} \right) \right| = \frac{n}{k}, \, \forall 1 \leq j \leq k \right\}.$$

We define an equivalence class on Θ_A as follows. We say $x, y \in \Theta_A$ are equivalent if there exists a fixed angle α, such that $e^{\mathrm{i}\alpha}\mathbf{x} = \mathbf{y}$. For each $z \in \Theta_A$, let $\theta(z)$ be the equivalence class containing y.

Let $y \in \Theta_A$ be the true color assignment function. Define the observation by

$$\mathbf{U} = \mathbf{P}(y) + \sigma \mathbf{W}_c$$

where $\mathbf{W}_c$ is the standard GUE random matrix. More precisely, $\mathbf{W}_c$ is a random Hermitian matrix whose diagonal entries are i.i.d. standard real Gaussian random variables $(\mathcal{N}(0,1)_{\mathbb{R}})$, and upper triangular entries are i.i.d. standard complex Gaussian random variables $(\mathcal{N}(0,1)_{\mathbb{C}})$.

Given each observation $\mathbf{U}$, the goal is to determine the true color assignment function y, up to a multiplicative constant. Let

$$(2.14) \qquad y^A = \mathrm{argmin}_{x \in \Theta_A}\|\mathbf{U} - \mathbf{P}(x)\|_F^2$$

Note that for any $x \in \Theta$,

$$\|\mathbf{P}(x)\|_F^2 = \sum_{1 \leq a,b \leq n} x(a)\overline{x(b)}x(b)\overline{x(a)} = n^2,$$

which is independent of x. Hence we have

$$y^A = \mathrm{argmax}_{x \in \Theta_A}\Re\langle \mathbf{U}, \mathbf{P}(x) \rangle$$

where $\Re$ denotes the real part of a complex number, and $\langle \cdot, \cdot \rangle$ denotes the inner product of two matrices defined by

$$(2.15) \qquad \langle M_1, M_2 \rangle = \sum_{i,j \in [n]} M_1(i,j) \overline{M_2(i,j)}.$$

where $M_1, M_2 \in \mathbb{C}^{n \times n}$.

It is not hard to see that for any $x, z \in \Theta_A$, if $x \in \theta(z)$, then $\mathbf{P}(x) = \mathbf{P}(z)$. So given any observation $\mathbf{U}$, we cannot distinguish color assignment mappings in the same equivalence class of θ_A. Therefore the best we can do by an MLE algorithm is to recover the equivalence class of the true color assignment function. We have the following theorem.

Theorem 2.4. *Let $y \in \Theta_A$ be the true color assignment function satisfying Assumption 2.3. Let k be fixed and $n \to \infty$.*
(a) If there exists $\delta > 0$, such that

$$(2.16) \qquad \sigma^2 < \frac{(1-\delta)\left[n(1-\cos\frac{2\pi}{k})\right]}{2\log n}$$

then

$$\lim_{n \to \infty} \Pr(y^A \in \theta(y)) = 1$$

(b) If there exists $\delta > 0$, such that

$$\sigma^2 > \frac{(1+\delta)\left[n(1-\cos\frac{2\pi}{k})\right]}{2\log n}$$

then

$$\lim_{n \to \infty} \Pr(y^A \in \theta(y)) = 0.$$

Now we describe a complex semi-definite programming for community detection with multiple (more than 2) communities. For a true color assignment function $y \in \Theta$, let the observation be given by

$$\mathbf{V} = \mathbf{P}(y) + \sigma \mathrm{diag}(\mathbf{y}) \mathbf{W}_s \mathrm{diag}(\overline{\mathbf{y}})$$

where $\mathbf{W}_s$ is the standard GOE random matrix. More precisely, $\mathbf{W}_s$ is a random symmetric matrix whose diagonal entries and upper triangular entries are i.i.d. standard real Gaussian random variables ($\mathcal{N}(0,1)_{\mathbb{R}}$). Note that $\mathbf{V}$ is a Hermitian matrix.

Given each observation $\mathbf{V}$, the goal is to find the true color assignment function y, up to a multiplicative constant. We may consider the following optimization problem

$$(2.17) \qquad\qquad \max \Re \langle \mathbf{V}, X \rangle$$
$$\text{subject to} \quad X_{ii} = 1$$
$$\text{and} \quad X \succeq 0$$

where $X \succeq 0$ means that X is a positive semi-definite Hermitian matrix.

For any mapping $x \in \Theta$, the matrix $P(x) := \mathbf{x}\overline{\mathbf{x}}^t$ is a rank-1 Hermitian matrix. The idea for the MLE is to find the minimizer

$$\text{argmin}_{P(x)} \|P(x) - \mathbf{V}\|^2$$

among all the rank-1 positive semi-definite Hermitian matrices with diagonal entries 1 and the form $\mathbf{x}\mathbf{x}^t$. To achieve efficient recovery, we may relax the rank-1 constraint and consider instead an optimization over all the positive semi-definite Hermitian matrix with diagonal entries 1.

Then we have the following theorem

Theorem 2.5. *Let $p(Y;\sigma)$ be the probability that the solution Y of (2.17) coincides with $\mathbf{y}\overline{\mathbf{y}}^t$, where y is the true color assignment mapping for vertices. If there exists a constant $\delta > 0$ independent of N, such that $\sigma < \frac{(1-\delta)\sqrt{n}}{\sqrt{2\log n}}$, then*

$$\lim_{n\to\infty} p(Y;\sigma) = 1$$

In Theorem 2.5, it turns out that the bound $\frac{(1-\delta)\sqrt{n}}{\sqrt{2\log n}}$ on σ to guarantee the exact recovery of the SDP is independent of k - the total number of communities. However, if we instead use the GUE matrix $\mathbf{W}_c$ instead of $\text{diag}(\mathbf{y})\mathbf{W}_s\text{diag}(\overline{\mathbf{y}})$ to represent the noise, Theorem 2.4 shows that threshold of σ to guarantee the exact recovery of MLE does depend on k. This threshold is of order $O\left(\frac{1}{\sqrt{n\log n}}\right)$ when $k \sim n$. Since the SDP is an algorithm obtained from the relaxation of the rank constraint of the MLE, one may naturally expect a smaller common upper bound for σ to guarantee the exact recovery of the SDP for all k, when the noise is represented by a σ-multiple of $\mathbf{W}_c$.

Theorem 2.4 is proved in Section 7.1; and Theorem 2.5 is proved in Section 7.2.

3. PROOF OF THEOREM 2.1 WHEN THE NUMBER OF VERTICES IN EACH COLOR IS FIXED IN THE SAMPLE SPACE

We first consider the MLE in Theorem 2.1 with sample space $\Omega_{n_1,\ldots,n_k}$, where the number of vertices with color c_i is fixed to be n_i - the same as the true value, for $1 \le i \le k$. For a mapping $x \in \Omega_{n_1,\ldots,n_k}$, let $\mathbf{G}(x)$, $\mathbf{T}$ be defined as in (2.2), (2.3), respectively.

Given a sample $\mathbf{T}$, the goal is to determine the color assignment function y. Let $\hat{y}$ be defined by (2.4). Note that for all $x \in \Omega_{n_1,\ldots,n_k}$,

$$\|\mathbf{G}(x)\|_F^2 = 2 \sum_{1 \le i < j \le k} n_i n_j (c_i - c_j)^2,$$

which depends only on $n_1, \ldots, n_k$, but is independent of x. Then we have

$$\hat{y} = \text{argmax}_{x \in \Omega_{n_1,\ldots,n_k}} \langle \mathbf{G}(x), \mathbf{T} \rangle$$

Let

(3.1) $$p(\hat{y}, \sigma) = \Pr(\hat{y} = y),$$

where $y \in \Omega_{n_1,\ldots,n_k}$ is the true color assignment function.

For each $x \in \Omega_{n_1,\ldots,n_k}$, define

$$f(x) = \langle \mathbf{G}(x), \mathbf{T} \rangle$$

Then

$$p(\hat{y}, \sigma) = \Pr\left(f(y) > \max_{x \in \Omega_{n_1,\ldots,n_k} \setminus \{y\}} f(x) \right)$$

Note that

$$f(x) - f(y) = \langle \mathbf{G}(y), \mathbf{G}(x) - \mathbf{G}(y) \rangle + \sigma \langle \mathbf{W}, \mathbf{G}(x) - \mathbf{G}(y) \rangle.$$

The expression above shows that $f(x) - f(y)$ is a Gaussian random variable with mean $\langle \mathbf{G}(y), \mathbf{G}(x) - \mathbf{G}(y) \rangle$ and variance $\sigma^2 \|\mathbf{G}(x) - \mathbf{G}(y)\|_F^2$.

For $i, j \in [k]$, $x, y \in \Omega_{n_1,\ldots,n_k}$, let

$$(3.2) \qquad S_{i,j}(x, y) = \{l \in [n] : x(l) = c_i, y(l) = c_j\};$$

i.e., $S_{i,j}(x, y)$ consists of all the vertices which have color c_i in x and color c_j in y.

Let $t_{i,j}(x, y) = |S_{i,j}(x, y)|$, i.e., $t_{i,j}$ is the total number of vertices which have color c_i in x and color c_j in y. We may write $t_{i,j}$ (resp. $S_{i,j}$) instead of $t_{i,j}(x, y)$ (resp. $S_{i,j}(x, y)$) when there is no confusion. Note that since $x, y \in \Omega_{n_1,\ldots,n_k}$, we have

$$(3.3) \qquad \sum_{j \in [k]} t_{i,j} = n_i, \quad \forall i \in [k]$$

$$(3.4) \qquad \sum_{i \in [k]} t_{i,j} = n_j, \quad \forall j \in [k]$$

For each vertex $l \in S_{i,j}$, the inner product of the row in $\mathbf{G}(x)$ corresponding to l and the row in $\mathbf{G}(y)$ corresponding to l is

$$
\begin{aligned}
\langle \mathbf{G}_l(x), \mathbf{G}_l(y) \rangle &= \sum_{r \in [n]} \mathbf{G}_{l,r}(x) \overline{\mathbf{G}_{l,r}(y)} \\
&= \sum_{r \in [n]} [x(l) - x(r)] [y(l) - y(r)] \\
&= \sum_{u=1}^{k} \sum_{v=1}^{k} t_{u,v} (c_i - c_u)(c_j - c_v)
\end{aligned}
$$

Then

$$\langle \mathbf{G}(x), \mathbf{G}(y) \rangle = \sum_{l=1}^{n} \langle \mathbf{G}_l(x), \mathbf{G}_l(y) \rangle$$

(3.5)
$$= \sum_{i,j,u,v \in [k]} t_{i,j} t_{u,v} (c_i - c_u)(c_j - c_v)$$

$$= 2 \left[\sum_{i,j \in [k]} t_{i,j} \cdot c_i \cdot c_j \right] \left[\sum_{u,v \in [k]} t_{u,v} \right] - 2 \left[\sum_{i,j \in [k]} t_{i,j} \cdot c_i \right] \left[\sum_{u,v \in [k]} t_{u,v} \cdot c_v \right]$$

$$= 2n \left[\sum_{i,j \in [k]} t_{i,j} \cdot c_i \cdot c_j \right] - 2 \left[\sum_{i \in [k]} n_i \cdot c_i \right]^2,$$

where the last identity follows from (3.3) and (3.4).

Note that for $i, j \in [k]$ and $x \in \Omega_{n_1,\ldots,n_k}$,

$$t_{i,j}(x,x) = \begin{cases} n_i & \text{if } i = j; \\ 0 & \text{else.} \end{cases}$$

Therefore

$$\langle \mathbf{G}(y), \mathbf{G}(y) \rangle = 2n \sum_{i \in [k]} \left[n_i \cdot c_i^2 \right] - 2 \left[\sum_{i \in [k]} n_i \cdot c_i \right]^2$$

For $x, y \in \Omega_{n_1,\ldots,n_k}$, let

$$M(x,y) : = -\mathbb{E}[f(x) - f(y)] = -\langle \mathbf{G}(y), \mathbf{G}(x) - \mathbf{G}(y) \rangle$$

(3.6)
$$= 2n \sum_{i \in [k]} \left[n_i \cdot c_i^2 \right] - 2n \left[\sum_{i,j \in [k]} t_{i,j} \cdot c_i \cdot c_j \right]$$

Then

$$\mathbf{Var}[f(x) - f(y)] = 2\sigma^2 M(x,y)$$

Therefore we have

$$f(x) - f(y) \sim \mathcal{N}(-M(x,y), 2\sigma^2 M(x,y))$$

and

$$\xi := \frac{f(x,y) + M(x,y)}{\sigma \sqrt{2M(x,y)}} \sim \mathcal{N}(0,1).$$

Then for $x \in \Omega_{n_1,\ldots,n_k} \setminus \{y\}$

$$\Pr\left(f(x) - f(y) > 0 \right) = \Pr_{\xi \sim \mathcal{N}(0,1)} \left(\xi > \frac{\sqrt{M(x,y)}}{\sqrt{2}\sigma} \right)$$

Using the standard Gaussian tail bound $\Pr_{\xi \in \mathcal{N}(0,1)}(\xi > x) < e^{-\frac{1}{2}x^2}$, we obtain

$$\Pr\left(f(x) - f(y) > 0 \right) \le e^{-\frac{M(x,y)}{4\sigma^2}}$$

Then

$$(3.7) \qquad 1 - p(\hat{y}; \sigma) \; \leq \; \sum_{x \in \Omega_{n_1,\dots,n_k} \setminus \{y\}} \Pr(f(x) - f(y) > 0)$$

$$= \sum_{x \in \Omega_{n_1,\dots,n_k} \setminus \{y\}} \Pr_{\xi \sim \mathcal{N}(0,1)} \left(\xi > \frac{\sqrt{M(x,y)}}{\sqrt{2}\sigma} \right)$$

$$\leq \sum_{x \in \Omega_{n_1,\dots,n_k} \setminus \{y\}} e^{-\frac{M(x,y)}{4\sigma^2}}$$

Lemma 3.1. *Let $y \in \Omega_{n_1,\dots,n_k}$ be the fixed color assignment mapping, and let x change over $\Omega_{n_1,\dots,n_k}$. Under the constraints (3.3) and (3.4), $M(x,y)$ achieves its minimum if and only if*

$$t_{i,i} = n_i$$
$$t_{i,j} = 0, \qquad \text{if } i \neq j.$$

and the minimal value of $M(x,y)$ is 0.

Proof. Note that

$$\sum_{i,j \in [k]} t_{i,j} \cdot c_i \cdot c_j \leq \sum_{i,j \in [k]} \frac{(c_i^2 + c_j^2) t_{i,j}}{2} = \sum_{i \in [k]} [n_i \cdot c_i^2];$$

where the identity holds if and only if

- $c_i = c_j$ whenever $t_{i,j} \neq 0$.

Then the lemma follows from the assumption that $c_i \neq c_j$ whenever $i \neq j$. $\qquad \square$

Before proving Theorem 2.1, we introduce a definition.

Definition 3.2. *Let $l \geq 2$ be a positive integer. Let $x, y \in \Omega_{n_1,\dots,n_k}$. We say l distinct colors $(i_1, \dots, i_l) \in [k]^l$ is an l-cycle for (x,y), if $t_{i_{s-1}, i_s}(x,y) > 0$ for all $2 \leq s \leq l+1$, where $i_{l+1} := i_1$.*

Lemma 3.3. *Let $x, y \in \Omega_{n_1,\dots,n_k}$ and $x \neq y$. Then there exists an l-cycle for (x,y) with $2 \leq l \leq k$.*

Proof. Since $x \neq y$ and $x, y \in \Omega_{n_1,\dots,n_k}$, there exists $i_1 \in [k]$, such that $t_{i_1,i_1}(x,y) < n_{i_1}$. Since $\sum_{j \in [k]} t_{i_1,j}(x,y) = n_{i_1}$, there exists $i_2 \in ([k] \setminus \{i_1\})$, such that $t_{i_1,i_2}(x,y) > 0$. The following cases might occur:

- If $t_{i_2,i_1}(x,y) > 0$, then we find a 2-cycle (i_1, i_2) for (x,y).
- If $t_{i_2,i_1}(x,y) = 0$, since $t_{i_1,i_2}(x,y) > 0$ and $\sum_{j \in [k]} t_{j,i_2}(x,y) = n_{i_2}$, we obtain that $t_{i_2,i_2}(x,y) < n_{i_2}$. Moreover, since $\sum_{j \in [k]} t_{i_2,j}(x,y) = n_{i_2}$, there exists $i_3 \in ([k] \setminus \{i_1, i_2\})$, such that $t_{i_2,i_3}(x,y) > 0$.
- If $t_{i_3,i_1}(x,y) > 0$, then we find a 3-cycle (i_1, i_2, i_3) for (x,y).

In general, let $s \geq 2$. Assume we find distinct $i_1, i_2, \dots, i_{s+1} \in [k]$, such that

- for each $1 \leq r \leq s$, we have $t_{i_s, i_{s+1}}(x,y) > 0$; and

- $t_{i_{s+1}, i_1}(x, y) = 0$.

Since $\sum_{j \in [k]} t_{j, i_{s+1}}(x, y) = n_{i_{s+1}}$, we have $t_{i_{s+1}, i_{s+1}}(x, y) < n_{i_{s+1}}$. Since $\sum_{j \in [k]} t_{i_{s+1}, j}(x, y) = n_{i_{s+1}}$, there exists $i_{s+2} \in ([k] \setminus \{i_{s+1}, i_1\})$, such that $t_{i_{s+1}, i_{s+2}}(x, y) > 0$. The following cases might occur

(1) if $i_{s+2} = i_r$, for some $2 \le r \le s$. Then we find a $(s + 2 - r)$-cycle $(i_r, i_{r+1} \dots, i_{s+1})$ for (x, y).

(2) if $i_{s+2} \ne i_r$, for all $1 \le r \le s + 1$, but there exists $1 \le g \le s + 1$, such that

(3.8)
$$t_{i_{s+2}, t_{i_g}} > 0$$

Then let $g_M \in [s + 1]$ be the maximal $g \in [s + 1]$ such that (3.8) holds. Then we find a $(s + 3 - g_M)$-cycle $(i_{g_M}, i_{g_M + 1}, \dots, i_{s+2})$ cycle for (x, y).

(3) if neither (1) or (2) occurs, increase s by 1 and repeat the above process.

Since there are k distinct colors in total, we can always find an l-cycle for (x, y) with $2 \le l \le k$.   □

To measure the difference between two color assignment functions $x, y \in \Omega_{n_1, \dots, n_k}$, we introduce the following definition.

Definition 3.4. *Define the distance function on Ω $D_\Omega : \Omega \times \Omega \to \mathbb{N}$ as follows*
$$D_\Omega(x, y) = n - \sum_{i \in [k]} t_{i,i}(x, y)$$

Proposition 3.5. *Assume $\delta \in \left(0, \frac{1}{2}\right)$, such that*

(3.9)
$$\log k - \frac{\delta \log n}{1 - \delta} \le -B_0 < 0$$

and

(3.10)
$$\sigma^2 < \frac{(1 - \delta) C_0 n}{4 \log n}$$

where $C_0 > 0$ is a constant given by

(3.11)
$$C_0 = \min_{1 \le i < j \le k} (c_i - c_j)^2$$

Then

(3.12)
$$p(\hat{y}; \sigma) \ge 2 - e^{\frac{e^{-2B_0}}{(1 - e^{-B_0})(1 - e^{-2B_0})}};$$

where $p(\hat{y}; \sigma)$ is given by (6.1).

Proof. Let

(3.13)
$$I := \sum_{x \in \Omega_{n_1, \dots, n_k} \setminus \{y\}} e^{-\frac{M(x, y)}{4\sigma^2}}.$$

By (3.7), it suffices to show that $\lim_{n \to \infty} I = 0$.

We shall find an upper bound for I.

Recall that $y \in \Omega_{n_1, \dots, n_k}$ is the true color assignment mapping. Since $x \ne y$, by Lemma 3.3, there exists an l-cycle $(i_1, \dots, i_l)$ for (x, y) with $2 \le l \le k$. Then for each $2 \le s \le (l+1)$,

choose an arbitrary vertex u_s in $S_{i_{s-1},i_s}(x,y)$, and let $y_1(u_s) = c_{i_{s-1}}$, where $i_{l+1} := i_1$. For any vertex $z \in [n] \setminus \{u_2, \ldots, u_{l+1}\}$, let $y_1(z) = y(z)$.

Note that $y_1 \in \Omega_{n_1,\ldots,n_k}$. More precisely, for $1 \leq s \leq l$, we have

$$t_{i_s,i_s}(x,y) + 1 = t_{i_s,i_s}(x,y_1);$$
$$t_{i_s,i_{s+1}}(x,y) - 1 = t_{i_s,i_{s+1}}(x,y_1)$$

and

$$t_{a,b}(x,y) = t_{a,b}(x,y_1), \ \forall (a,b) \notin \{(i_s,i_s),(i_s,i_{s+1})\}_{s=1}^{l}.$$

From (3.6) we obtain

$$
\begin{aligned}
M(x,y_1) - M(x,y) &= -2n \sum_{a,b \in [k]} [t_{a,b}(x,y_1) - t_{a,b}(x,y)]\, c_a c_b \\
&= -2n \left[\sum_{s \in [l]} (c_{i_s}^2 - c_{i_s} c_{i_{s+1}}) \right] \\
&= -n \left[\sum_{s \in [l]} (c_{i_s} - c_{i_{s+1}})^2 \right] \\
&\leq -nl C_0
\end{aligned}
$$

where C_0 is given in (3.11). Therefore

(3.14)
$$e^{-\frac{M(x,y)}{4\sigma^2}} \leq e^{-\frac{M(x,y_1)}{4\sigma^2}} e^{-\frac{nlC_0}{4\sigma^2}}.$$

If $y_1 \neq x$, we find an l_2-cycle $(2 \leq l_2 \leq k)$ for (x,y_1), change colors along the l_2-cycle as above, and obtain another cycle assignment mapping $y_2 \in \Omega_{n_1,\ldots,n_k}$, and so on. Let $y_0 := y$, and note that for each $r \geq 1$, if y_r is obtained from y_{r-1} by changing colors along an l_r cycle, we have

$$D_\Omega(x,y_r) = D_\Omega(x,y_{r-1}) - l_r$$

Therefore finally we can obtain x from y by changing colors along at most $\lfloor \frac{n}{2} \rfloor$ cycles. Using similar arguments as those used to derive (6.18), we obtain that for each r

$$e^{-\frac{M(x,y_{r-1})}{4\sigma^2}} \leq e^{-\frac{M(x,y_r)}{4\sigma^2}} e^{-\frac{nl_r C_0}{4\sigma^2}}.$$

Therefore if $y_s = x$ for some $1 \leq s \leq \lfloor \frac{n}{2} \rfloor$, we have

$$e^{-\frac{M(x,y)}{4\sigma^2}} \leq e^{-\frac{M(x,y_s)}{4\sigma^2}} e^{-\frac{nC_0\left(\sum_{i \in [s]} l_i\right)}{4\sigma^2}}.$$

By (3.6), we have $M(x,y_s) = M(x,x) = 0$, hence

$$e^{-\frac{M(x,y)}{4\sigma^2}} \leq \prod_{i \in [s]} e^{-\frac{nC_0 l_i}{4\sigma^2}}.$$

Note also that for any $r_1 \neq r_2$, in the process of obtaining y_{r_1} from y_{r_1-1} and the process of obtaining y_{r_2} from y_{r_2-1}, we change colors on disjoint sets of vertices. Hence the order

of these steps of changing colors along cycles does not affect the final color assignment mapping we obtain. From (3.13) we obtain

$$(3.15) \qquad I \leq \prod_{l=2}^{k} \left(\sum_{m_l=0}^{\infty} (nk)^{m_l l} e^{-\frac{nm_l l C_0}{4\sigma^2}} \right) - 1$$

On the right hand side of (6.19), when expanding the product, each summand has the form

$$\left[(nk)^{2m_2} e^{-\frac{2m_2 n C_0}{4\sigma^2}} \right] \cdot \left[(nk)^{3m_3} e^{-\frac{3m_3 n C_0}{4\sigma^2}} \right] \cdot \ldots \cdot \left[(nk)^{km_k} e^{-\frac{km_k n C_0}{4\sigma^2}} \right]$$

where the factor $\left[(nk)^{2m_2} e^{-\frac{2m_2 n C_0}{4\sigma^2}} \right]$ represents that we changed along 2-cycles m_2 times, the factor $\left[(nk)^{3m_3} e^{-\frac{3m_3 n C_0}{4\sigma^2}} \right]$ represents that we changed along 3-cycles m_3 times, and so on. Moreover, each time we changed along an l-cycle, we need to first determine the l different colors involved in the l-cycle, and there are at most k^l different l-cycles; we then need to chose l vertices to change colors, and there are at most n^l choices. It is straightforward to check that if σ satisfies (3.10), then

$$nk e^{-\frac{n C_0}{4\sigma^2}} \leq e^{\log k - \frac{\delta \log n}{1-\delta}}$$

as $n \to \infty$. Therefore we have

$$\sum_{m_l=0}^{\infty} (nk)^{m_l l} e^{-\frac{nm_l l C_0}{4\sigma^2}} \leq \frac{1}{1 - e^{l \log k - \frac{\delta l \log n}{1-\delta}}}$$

Let

$$U := \prod_{l=2}^{k} \left(\sum_{m_l=0}^{\infty} (nk)^{m_l l} e^{-\frac{nm_l l C_0}{4\sigma^2}} \right).$$

Since $\log(1+x) \leq x$ for $x \geq 0$, we have

$$\begin{aligned} 0 \leq \log U \;=\; & \sum_{l=2}^{k} \log \left(1 + \sum_{m_l=1}^{\infty} (nk)^{m_l l} e^{-\frac{nm_l l C_0}{4\sigma^2}} \right) \\ \leq\; & \sum_{l=2}^{k} \sum_{m_l=1}^{\infty} (nk)^{m_l l} e^{-\frac{nm_l l C_0}{4\sigma^2}} \\ \leq\; & \sum_{l=2}^{k} \frac{\left(e^{\log k - \frac{\delta \log n}{1-\delta}} \right)^l}{1 - \left(e^{\log k - \frac{\delta \log n}{1-\delta}} \right)^l} \\ \leq\; & \frac{e^{2\log k - \frac{2\delta \log n}{1-\delta}}}{\left(1 - e^{2\log k - \frac{2\delta \log n}{1-\delta}} \right) \left(1 - e^{\log k - \frac{\delta \log n}{1-\delta}} \right)} \end{aligned}$$

Then the proposition follows from (3.9). □

Corollary 3.6. *Assume* $\delta \in \left(0, \frac{1}{2}\right)$, *such that*

$$\lim_{n \to \infty} \log k - \frac{\delta \log n}{1 - \delta} = -\infty.$$

Assume (3.10) holds. Then

$$\lim_{n \to \infty} p(\hat{y}; \sigma) = 1.$$

Proof. The corollary follows from (3.12) by letting $B_0 \to \infty$.   □

Proposition 3.7. *Let* $C_0 > 0$ *be defined by (3.11), and* $y \in \Omega_{n_1,\dots,n_k}$ *be the true color assignment mapping. Assume there exists* $u, v \in [k]$, *such that*

$$(c_u - c_v)^2 = C_0,$$

and

$$(3.16) \qquad\qquad \frac{\log \min\{n_u, n_v\}}{\log n} \geq \beta > 0.$$

Assume

$$(3.17) \qquad\qquad \delta = \left(1 + \frac{4}{\beta}\right) \frac{\log \log n}{\log n},$$

and

$$(3.18) \qquad\qquad \sigma^2 > \frac{(1 + \delta) C_0 n}{4\beta \log n}$$

then

$$\lim_{n \to \infty} p(\hat{y}; \sigma) = 0,$$

where β *is a constant independent of* n *as* $n \to \infty$, *while the total number of colors* k, *the set of colors* R_k, *and* C_0 *may depend on* n.

Proof. For $y \in \Omega_{n_1,\dots,n_k}$, $a, b \in [n]$ such that $c_u = y(a) \neq y(b) = c_v$. Let $y^{(ab)} \in \Omega_{n_1,\dots,n_k}$ be the coloring of vertices defined by

$$(3.19) \qquad\qquad y^{(ab)}(i) = \begin{cases} y(i) & \text{if } i \in [n] \setminus \{a, b\} \\ c_v & \text{if } i = a \\ c_u & \text{if } i = b \end{cases}$$

Then

$$t_{u,v}(y^{(ab)}, y) - 1 = t_{u,v}(y, y) = 0$$
$$t_{u,u}(y^{(ab)}, y) + 1 = t_{u,u}(y, y) = n_u$$
$$t_{v,u}(y^{(ab)}, y) - 1 = t_{v,u}(y, y) = 0$$
$$t_{v,v}(y^{(ab)}, y) + 1 = t_{v,v}(y, y) = n_v.$$

and

$$t_{i,j}(y^{(ab)}, y) = t_{i,j}(y), \ \forall\, (i, j) \in \left([k]^2 \setminus \{(u, u), (u, v), (v, u), (v, v)\}\right)$$

Note that

$$1 - p(\hat{y}; \sigma) \geq \Pr\left(\cup_{a,b \in [n], c_u = y(a) \neq y(b) = c_v} [f(y^{(ab)}) - f(y) > 0]\right),$$

since any of the event $[f(y^{(ab)}) - f(y) > 0]$ implies $\hat{y} \neq y$. By (3.6) we obtain

$$
\begin{aligned}
f(y^{(ab)}) - f(y) &= \langle \mathbf{G}(y), \mathbf{G}(y^{(ab)}) - \mathbf{G}(y) \rangle + \sigma \langle \mathbf{W}, \mathbf{G}(y^{(ab)}) - \mathbf{G}(y) \rangle \\
&= -2n(c_u - c_v)^2 + \sigma \langle \mathbf{W}, \mathbf{G}(y^{(ab)}) - \mathbf{G}(y) \rangle.
\end{aligned}
$$

So $1 - p(\hat{y}; \sigma)$ is at least

$$
\begin{aligned}
&\Pr\left(\cup_{a,b \in [n], c_u = y(a) \neq y(b) = c_v} [f(y^{(ab)}) - f(y) > 0]\right) \\
\geq\ &\Pr\left(\max_{a,b \in [n], c_u = y(a) \neq y(b) = c_v} \sigma \langle \mathbf{W}, \mathbf{G}(y^{(ab)}) - \mathbf{G}(y) \rangle > 2nC_0\right)
\end{aligned}
$$

For $i \in \{u, v\}$, let $H_i \subset y^{-1}(c_i)$ such that

(3.20) $$|H_i| = \frac{\min\{n_u, n_v\}}{\log^2 n} = h.$$

Then

$$1 - p(\hat{y}; \sigma) \geq \Pr\left(\max_{a \in H_u, b \in H_v} \sigma \langle \mathbf{W}, \mathbf{G}(y^{(ab)}) - \mathbf{G}(y) \rangle > 2nC_0\right)$$

Let $(\mathcal{X}, \mathcal{Y}, \mathcal{Z})$ be a partition of $[n]^2$ defined by

$$
\begin{aligned}
\mathcal{X} &= \{\alpha = (\alpha_1, \alpha_2) \in [n]^2, \{\alpha_1, \alpha_2\} \cap [H_u \cup H_v] = \emptyset\} \\
\mathcal{Y} &= \{\alpha = (\alpha_1, \alpha_2) \in [n]^2, |\{\alpha_1, \alpha_2\} \cap [H_u \cup H_v]| = 1\} \\
\mathcal{Z} &= \{\alpha = (\alpha_1, \alpha_2) \in [n]^2, |\{\alpha_1, \alpha_2\} \cap [H_u \cup H_v]| = 2\}
\end{aligned}
$$

For $\eta \in \{\mathcal{X}, \mathcal{Y}, \mathcal{Z}\}$, define the $n \times n$ matrix $\mathbf{W}_\eta$ from the entries of $\mathbf{W}$ as follows

$$\mathbf{W}_\eta(i, j) = \begin{cases} 0 & \text{if } (i, j) \notin \eta \\ \mathbf{W}(i, j), & \text{if } (i, j) \in \eta \end{cases}$$

For each $a \in H_u$ and $b \in H_v$, let

$$
\begin{aligned}
\mathcal{X}_{ab} &= \langle \mathbf{W}_{\mathcal{X}}, \mathbf{G}(y^{(ab)}) - \mathbf{G}(y) \rangle \\
\mathcal{Y}_{ab} &= \langle \mathbf{W}_{\mathcal{Y}}, \mathbf{G}(y^{(ab)}) - \mathbf{G}(y) \rangle \\
\mathcal{Z}_{ab} &= \langle \mathbf{W}_{\mathcal{Z}}, \mathbf{G}(y^{(ab)}) - \mathbf{G}(y) \rangle
\end{aligned}
$$

Lemma 3.8. *The followings are true:*

(1) $\mathcal{X}_{ab} = 0$ for $a \in H_u$ and $b \in H_v$.

(2) For each $a \in H_u$ and $b \in H_v$, the variables $\mathcal{Y}_{ab}$ and $\mathcal{Z}_{ab}$ are independent.

(3) Each $\mathcal{Y}_{ab}$ can be decomposed into $Y_a + Y_b$ where $\{Y_a\}_{a \in H_u} \cup \{Y_b\}_{b \in H_v}$ is a collection of i.i.d. Gaussian random variables.

Proof. Note that for $i, j \in [n]$,

$$(3.21) \qquad \mathbf{G}_{i,j}(y^{(ab)}) - \mathbf{G}_{i,j}(y) = \begin{cases} c_v - c_u & \text{if } i = a, j \notin \{a, b\} \\ c_u - c_v & \text{if } i \notin \{a, b\}, j = a \\ c_u - c_v & \text{if } i = b, j \notin \{a, b\} \\ c_v - c_u & \text{if } i \notin \{a, b\}, j = b \\ 2(c_v - c_u) & \text{if } (i, j) = (a, b) \\ 2(c_u - c_v) & \text{if } (i, j) = (b, a) \\ 0 & \text{otherwise.} \end{cases}$$

It is straightforward to check (1). (2) holds because $\mathcal{Y} \cap \mathcal{Z} = \emptyset$.

For $s \in H_u \cup H_v$, let $\mathcal{Y}_s \subseteq \mathcal{Y}$ be defined by

$$\mathcal{Y}_s = \{\alpha = (\alpha_1, \alpha_2) \in \mathcal{Y} : \alpha_1 = s, \text{ or } \alpha_2 = s\}.$$

Note that for $s_1, s_2 \in H_u \cup H_v$ and $s_1 \neq s_2$, $\mathcal{Y}_{s_1} \cap \mathcal{Y}_{s_2} = \emptyset$. Moreover, $\mathcal{Y} = \cup_{s \in H_u \cup H_v} \mathcal{Y}_s$. Therefore

$$\mathcal{Y}_{ab} = \sum_{s \in H_u \cup H_v} \langle \mathbf{W}_{\mathcal{Y}_s}, \mathbf{G}(y^{(ab)}) - \mathbf{G}(y) \rangle$$

Note also that $\langle \mathbf{W}_{\mathcal{Y}_s}, \mathbf{G}(y^{(ab)}) - \mathbf{G}(y) \rangle = 0$, if $s \notin \{a, b\}$. Hence

$$\mathcal{Y}_{ab} = \sum_{\alpha \in \mathcal{Y}_a \cup \mathcal{Y}_b} [\mathbf{W}(\alpha)] \cdot \{[\mathbf{G}(y^{(ab)}) - \mathbf{G}(y)](\alpha)\}$$

From (3.21) we obtain that for $\alpha = (\alpha_1, \alpha_2) \in \mathcal{Y}_a$ and $\alpha_1 \neq \alpha_2$,

$$[\mathbf{G}(y^{(ab)}) - \mathbf{G}(y)](\alpha) = \begin{cases} c_v - c_u & \text{if } \alpha_1 = a \\ c_u - c_v & \text{if } \alpha_2 = a. \end{cases}$$

So, we can define

$$Y_a := \sum_{\alpha \in \mathcal{Y}_a} [\mathbf{W}(\alpha)] \cdot \{[\mathbf{G}(y^{(ab)}) - \mathbf{G}(y)](\alpha)\}$$

$$= \left\{ \sum_{\alpha \in \mathcal{Y}_a; \alpha_1 = a} [\mathbf{W}(\alpha)] - \sum_{\alpha \in \mathcal{Y}_a; \alpha_2 = a} [\mathbf{W}(\alpha)] \right\} (c_v - c_u)$$

Similarly, define

$$Y_b := \left\{ \sum_{\alpha \in \mathcal{Y}_b; \alpha_2 = b} [\mathbf{W}(\alpha)] - \sum_{\alpha \in \mathcal{Y}_b; \alpha_1 = b} [\mathbf{W}(\alpha)] \right\} (c_v - c_u)$$

Then $\mathcal{Y}_{ab} = Y_a + Y_b$ and $\{Y_s\}_{s \in H_u \cup H_v}$ is a collection of independent Gaussian random variables. Moreover, the variance of Y_s is equal to $(2n - 4h)C_0$ independent of the choice of s. $\square$

By the Lemma 3.8, we obtain

$$\langle \mathbf{W}, \mathbf{G}(y^{(ab)}) - \mathbf{G}(y) \rangle = Y_a + Y_b + \mathcal{Z}_{ab}$$

Moreover,

$$
\begin{aligned}
\max_{a \in H_u, b \in H_v} Y_a + Y_b + \mathcal{Z}_{ab} \;\geq\; & \max_{a \in H_u, b \in H_v} (Y_a + Y_b) - \max_{a \in H_u, b \in H_v} (-\mathcal{Z}_{ab}) \\
= \; & \max_{a \in H_u} Y_a + \max_{b \in H_v} Y_b - \max_{a \in H_u, b \in H_v} (-\mathcal{Z}_{ab})
\end{aligned}
$$

Recall the following tail bound result on the maximum of Gaussian random variables:

Lemma 3.9. *Let $G_1, \ldots, G_N$ be Gaussian random variables with variance 1. Let $\epsilon \in (0,1)$. Then*

$$\Pr\left(\max_{i=1,\ldots,N} G_i > (1 + \epsilon)\sqrt{2 \log N} \right) \leq N^{-\epsilon}$$

and moreover, if G_i's are independent, and ϵ, N satisfy

$$(3.22) \qquad \frac{N^{\epsilon - \epsilon^2}(1 - \epsilon)\sqrt{2 \log N}}{\sqrt{2\pi}(1 + 2(1 - \epsilon)^2 \log N)} > 1$$

Then

$$\Pr\left(\max_{i=1,\ldots,N} G_i < (1 - \epsilon)\sqrt{2 \log N} \right) \leq \exp(-N^{\epsilon})$$

The proof of Lemma 3.9 is in the appendix.

By Lemma 3.9 we obtain that when ϵ, h satisfy (3.22) with N replaced by h, each one of the following two events

$$E_1 := \left\{ \max_{a \in H_u} Y_a \geq (1 - \epsilon)\sqrt{2 \log h \cdot 2 C_0 (n - 2h)} \right\}$$

$$E_2 := \left\{ \max_{b \in H_v} Y_b \geq (1 - \epsilon)\sqrt{2 \log h \cdot 2 C_0 (n - 2h)} \right\}$$

has probability at least $1 - e^{-h^\epsilon}$. Moreover, the event

$$E_3 := \left\{ \max_{a \in H_u, b \in H_v} \mathcal{Z}_{ab} \leq (1 + \epsilon)\sqrt{4 \log h \cdot \max \mathrm{Var}(Z_{ab})} \right\}$$

occurs with probability at least $1 - h^{-2\epsilon}$. Then by (3.21) we have

$$
\begin{aligned}
\mathrm{Var}\,\mathcal{Z}_{ab} \;=\; & \|\mathbf{G}(y^{(ab)}) - \mathbf{G}(y)\|_F^2 - \mathrm{Var}(Y_a) - \mathrm{Var}(Y_b) \\
= \; & 4 n C_0 - 4 C_0 (n - 2h) \\
= \; & 8 h C_0
\end{aligned}
$$

Hence the probability of the event

$$E := \left\{ \max_{a \in H_u, b \in H_v} \langle \mathbf{W}, \mathbf{G}(y^{(ab)}) - \mathbf{G}(y) \rangle \geq 4(1 - \epsilon)\sqrt{\log h C_0(n - 2h)} - 4(1 + \epsilon)\sqrt{2 C_0 h \log h} \right\}$$

is at least

$$
\begin{aligned}
\Pr(E_1 \cap E_2 \cap E_3) &= 1 - \Pr(E_1^c \cup E_2^c \cup E_3^c) \\
&\geq 1 - \Pr(E_1^c) - \Pr(E_2^c) - \Pr(E_3^c) \\
&\geq 1 - 2e^{-h^\epsilon} - h^{-2\epsilon}.
\end{aligned}
$$

Moreover, from (6.25) we obtain

$$
\begin{aligned}
&4(1-\epsilon)\sqrt{C_0(n-2h)\log h} - 4(1+\epsilon)\sqrt{2C_0 h \log h} \\
&= 4\sqrt{C_0(n-2h)\log h}\left[1 - \epsilon - (1+\epsilon)\sqrt{\frac{2h}{n-2h}}\right] \\
&\geq 4\sqrt{C_0(n-2h)\log h}\left[1 - \epsilon - \frac{2}{\log n}\right]
\end{aligned}
$$

By (3.16) we have

$$
\begin{aligned}
&4\sqrt{C_0(n-2h)\log h}\left[1 - \epsilon - \frac{2}{\log n}\right] \\
&\geq 4\sqrt{C_0 n\beta \log n \left(1 - \frac{2}{\log^2 n}\right)\left(1 - \frac{2\log\log n}{\beta \log n}\right)}\left[1 - \epsilon - \frac{2}{\log n}\right]
\end{aligned}
$$

Let

(3.23)
$$
\epsilon = \frac{\log\log n}{\beta \log n};
$$

then when n is sufficiently large, (3.22) holds with N replaced by h. Define an event

$$
\begin{aligned}
\tilde{E} : &= \left\{ \max_{a \in H_u, b \in H_v} \langle \mathbf{W}, \mathbf{G}(y^{(ab)}) - \mathbf{G}(y) \rangle \right. \\
&\left. \geq 4\sqrt{C_0 n\beta \log n \left(1 - \frac{2}{\log^2 n}\right)\left(1 - \frac{2\log\log n}{\beta \log n}\right)}\left[1 - \frac{\log\log n}{\beta \log n} - \frac{2}{\log n}\right] \right\}
\end{aligned}
$$

Then $E \subseteq \tilde{E}$

When (3.17) and (3.18) hold, we have

$$
\begin{aligned}
&\Pr\left(\max_{a,b \in \{1,2,\ldots,n\}, y(a) \neq y(b)} \sigma\langle \mathbf{W}, \mathbf{G}(y^{(ab)}) - \mathbf{G}(y) \rangle > 2C_0 n \right) \\
&\geq \Pr(\tilde{E}) \geq \Pr(E) \geq 1 - \frac{1}{\log n},
\end{aligned}
$$

as n is sufficiently large. Then the proposition follows. $\square$

4. Proof of Theorem 2.1 when the number of vertices in each color is arbitrary in the sample space

Now we consider the MLE in Theorem 2.1 whose sample space consists of all the possible mappings from $[n]$ to R_k, with no constraints on the number of vertices in each color. Assume that for each $x \in \Omega$, $n_1(x), \ldots, n_k(x)$ are arbitrary positive integers satisfying (2.1) and denoting the number of vertices in the colors $c_1, \ldots, c_k$ under the mapping x,

respectively. For a mapping $x \in \Omega$, let $\mathbf{G}(x)$ be defined as in (2.2). Let y be the true color assignment mapping, and let $\mathbf{T}$ be defined as in (2.3).

Given a sample $\mathbf{T}$, the goal is to determine the color assignment mapping y. Let $\check{y}$ be defined by (2.5). Then

$$\check{y} = \operatorname{argmin}_{x \in \Omega} \left(\|\mathbf{G}(x)\|_F^2 - 2\langle \mathbf{G}(x), \mathbf{T} \rangle \right)$$

Let

$$p(\check{y}, \sigma) = \Pr(\check{y} = y)$$

For $x \in \Omega$, define

$$d(x) = \|\mathbf{G}(x)\|_F^2 - 2\langle \mathbf{G}(x), \mathbf{T} \rangle$$

Then

$$p(\check{y}, \sigma) = \Pr\left(d(y) < \min_{x \in \Omega \setminus \{y\}} d(x) \right)$$

Note that

(4.1) $\qquad d(x) - d(y) =$
$$\|\mathbf{G}(x)\|_F^2 - \|\mathbf{G}(y)\|_F^2 - 2\langle \mathbf{G}(y), \mathbf{G}(x) - \mathbf{G}(y) \rangle - 2\sigma \langle \mathbf{W}, \mathbf{G}(x) - \mathbf{G}(y) \rangle.$$

The expression above shows that $d(x) - d(y)$ is a Gaussian random variable with mean $\|\mathbf{G}(x)\|_F^2 - \|\mathbf{G}(y)\|_F^2 - 2\langle \mathbf{G}(y), \mathbf{G}(x) - \mathbf{G}(y) \rangle$ and variance $4\sigma^2 \|\mathbf{G}(x) - \mathbf{G}(y)\|_F^2$.

For $i, j \in [k]$, let $S_{i,j}(x, y)$ be defined as in (3.2), and let $t_{i,j}(x, y) = |S_{i,j}(x, y)|$. Then

(4.2) $$\sum_{j \in [k]} t_{i,j}(x, y) = n_i(x), \quad \forall i \in [k]$$

(4.3) $$\sum_{i \in [k]} t_{i,j}(x, y) = n_j(y), \quad \forall j \in [k]$$

Then as in (3.5),

$$
\begin{aligned}
\langle \mathbf{G}(x), \mathbf{G}(y) \rangle &= \sum_{i,j,u,v \in [k]} t_{i,j} t_{u,v} (c_i - c_u)(c_j - c_v) \\
&= 2\left[\sum_{i,j \in [k]} t_{i,j} \cdot c_i \cdot c_j \right]\left[\sum_{u,v \in [k]} t_{u,v} \right] - 2\left[\sum_{i,j \in [k]} t_{i,j} \cdot c_i \right]\left[\sum_{u,v \in [k]} t_{u,v} \cdot c_v \right] \\
&= 2n\left[\sum_{i,j \in [k]} t_{i,j} \cdot c_i \cdot c_j \right] - 2\left[\sum_{i \in [k]} n_i(x) \cdot c_i \right]\left[\sum_{j \in [k]} n_j(y) \cdot c_j \right],
\end{aligned}
$$

where the last identity follows from (4.2) and (4.3).

In particular

$$\langle \mathbf{G}(y), \mathbf{G}(y) \rangle = 2n \sum_{i \in [k]} \left[n_i(y) \cdot c_i^2 \right] - 2\left[\sum_{i \in [k]} n_i(y) \cdot c_i \right]^2$$

Let

$$(4.4) \quad Q(x,y): \quad = \quad \mathbb{E}[d(x) - d(y)] = \|\mathbf{G}(x)\|_F^2 + \|\mathbf{G}(y)\|_F^2 - 2\langle \mathbf{G}(x), \mathbf{G}(y)\rangle$$

$$= \quad 2n \sum_{i,j\in[k]} t_{i,j}(x,y)(c_i - c_j)^2 - 2\left[\sum_{i\in[k]} n_i(x)c_i - \sum_{j\in[k]} n_j(y)c_j\right]^2$$

$$= \quad \sum_{u,v,i,j\in[k]} t_{u,v}(x,y)t_{i,j}(x,y)(c_u - c_v - c_i + c_j)^2$$

Then

$$\mathbf{Var}[d(x) - d(y)] = 4\sigma^2 Q(x,y)$$

For $x \in \Omega \setminus \{y\}$

$$\Pr\left(d(y) - d(x) > 0\right) \quad = \quad \Pr_{\xi \sim \mathcal{N}(0,1)}\left(\xi \geq \frac{\sqrt{Q(x,y)}}{2\sigma}\right)$$

Using the standard Gaussian tail bound $\Pr_{\xi \in \mathcal{N}(0,1)}(\xi > x) < e^{-\frac{1}{2}x^2}$, we obtain

$$\Pr\left(d(y) - d(x) > 0\right) \leq e^{-\frac{Q(x,y)}{8\sigma^2}}$$

Then

$$(4.5) \qquad 1 - p(\check{y};\sigma) \quad \leq \quad \sum_{x\in\Omega\setminus\{y\}} \Pr(d(y) - d(x) > 0)$$

$$= \quad \sum_{x\in\Omega\setminus\{y\}} \Pr_{\xi\sim\mathcal{N}(0,1)}\left(\xi > \frac{\sqrt{Q(x,y)}}{2\sigma}\right)$$

$$\leq \quad \sum_{x\in\Omega\setminus\{y\}} e^{-\frac{Q(x,y)}{8\sigma^2}}$$

Lemma 4.1. *Let $y \in \Omega$ be the fixed true color assignment mapping, and assume that x changes over Ω. Under the constraint (4.3), $Q(x,y)$ achieves its minimum if*

$$t_{i,i} = n_i(y)$$
$$t_{i,j} = 0, \qquad \text{if } i \neq j.$$

and the minimal value of $Q(x,y)$ is 0.

Proof. The lemma follows from (4.4). □

Let $\mathcal{B}$ be the set given by

$$\mathcal{B} = \left\{(t_{1,1}, t_{2,1} \ldots, t_{k,k}) \in \prod_{j\in[k]} \{0, 1, \ldots, n_j(y)\}^k : \sum_{i=1}^k t_{i,j}(x,y) = n_j(y)\right\}.$$

For a small positive number $\epsilon > 0$, let $\mathcal{B}_\epsilon$ be the domain given by

$$\mathcal{B}_\epsilon = \{(t_{1,1}, \ldots, t_{k,k}) \in \mathcal{B} : t_{i,i} \geq n_i(y) - n\epsilon, \ \forall\ i \in [k]\}$$

Then

$$\sum_{x \in \Omega \setminus \{y\}} e^{-\frac{Q(x,y)}{8\sigma^2}} = J_1 + J_2$$

where

$$J_1 = \sum_{x \in \Omega \setminus \{y\}: (t_{1,1},\ldots,t_{k,k}) \in [\mathcal{B} \setminus \mathcal{B}_\epsilon]} e^{-\frac{Q(x,y)}{8\sigma^2}}.$$

and

$$J_2 = \sum_{x \in \Omega \setminus \{y\}: (t_{1,1},\ldots,t_{k,k}) \in \mathcal{B}_\epsilon} e^{-\frac{Q(x,y)}{8\sigma^2}}.$$

Fix a constant $c > 0$. Define a region

(4.6) $$\mathcal{R}_c := \{(v_1, \ldots, v_k) \in \mathbb{R}^k : \sum_{i \in [k]} v_k = 1, \text{ and } \min_{i \in [k]} v_i \geq c\}.$$

Lemma 4.2. *Assume*

(4.7) $$\left(\frac{n_1(y)}{n}, \ldots, \frac{n_k(y)}{n}\right) \in \mathcal{R}_c$$

and

$$(t_{1,1}(x,y), \ldots, t_{k,k}(x,y)) \in \mathcal{B} \setminus \mathcal{B}_\epsilon.$$

Then if $\epsilon > 0$ is small enough,

(4.8) $$Q(x,y) \geq 4C_0 \epsilon^2 n^2$$

where $C_0 > 0$ is given by (3.11).

Proof. When $(t_{1,1}(x,y), \ldots, t_{k,k}(x,y)) \in \mathcal{D} \setminus \mathcal{D}_\epsilon$, we have there exists $1 \leq i \leq k$,

$$\sum_{j \in [k], j \neq i} t_{j,i}(x,y) \geq n\epsilon$$

The following cases might occur

(1) there exists $l \in [k]$, such that

$$t_{l,l}(x,y) > n_l(y) - n\epsilon$$

Then by (4.4)

$$\begin{aligned}
Q(x,y) &= \sum_{u,v,i,j \in [k]} t_{u,v}(x,y) t_{i,j}(x,y)(c_u - c_v - c_i + c_j)^2 \\
&\geq t_{l,l}(x,y) \sum_{j \in [k], j \neq i} t_{j,i}(x,y)(c_i - c_j)^2 \\
&\geq (n_l(y) - n\epsilon)\epsilon C_0 n \\
&\geq (c - \epsilon)\epsilon C_0 n^2,
\end{aligned}$$

where the last inequality follows from (4.7). Then (4.8) holds when $\epsilon \leq \frac{c}{5}$.

(2) For each $b \in [k]$, we have

$$\sum_{a \in [k], a \neq b} u_{a,b} \geq \epsilon$$

Without loss of generality, assume that $c_1 > c_2 > \ldots > c_k$. Again by (4.4)

$$
\begin{aligned}
Q(x,y) &= \sum_{u,v,i,j \in [k]} t_{u,v}(x,y) t_{i,j}(x,y)(c_u - c_v - c_i + c_j)^2 \\
&\geq \sum_{i \in [k], i \neq 1} \sum_{j \in [k], j \neq k} t_{i,1}(x,y) t_{j,k}(x,y)(c_i - c_1 - c_j + c_k)^2 \\
&\geq 4 C_0 \epsilon^2 n^2
\end{aligned}
$$

$\square$

Proposition 4.3. *Let y be the true color assignment mapping. Assume all the following conditions hold:*

- *(4.7) holds;*
- *The total number of colors k and the total number of vertices n satisfy*

(4.9) $$\lim_{n \to \infty} \frac{\log k}{\log n} = 0.$$

- *The quotient of the maximal difference of two colors and the minimal difference of two colors is uniformly bounded for all n; that is,*

(4.10) $$\sup_n \frac{\max_{u,v \in [k]} |c_u - c_v|}{\min_{u,v \in [k], u \neq v} |c_u - c_v|} < \infty$$

- *there exists $\delta > 0$ (independent of n), such that*

(4.11) $$\sigma^2 < \frac{(1 - \delta) C_0 n}{4 \log n}$$

where $C_0 > 0$ is defined by (3.11).

Then

$$\lim_{n \to \infty} p(\check{y}; \sigma) = 1.$$

Proof. First of all, for any fixed $\epsilon > 0$, if (4.11) holds, then $\sigma \sim o(\sqrt{n})$, by (4.7) and Lemma 4.2 we obtain that

$$J_1 \leq \sum_{x \in \Omega \setminus \{y\}} e^{-\frac{C_0 \epsilon^2 n^2}{2\sigma^2}}$$

Note that

$$|\Omega| \leq k^n$$

Hence

$$J_1 \leq e^{n\left(\log k - \frac{2\epsilon^2 \log n}{1 - \delta}\right)} \to 0$$

as $n \to \infty$, for all $\epsilon > 0$ by (4.9).

Now let us consider J_2. Let $x \in \Omega$ be an arbitrary color assignment mapping. Then x can be obtained from y as follows:

(1) If for all $(i,j) \in [k]^2$, $i \neq j$, $t_{i,j}(x,y) = 0$, then $x = y$.
(2) If $x \neq y$, find the least $(a,b) \in [k]^2$ in lexicographic order such that $a \neq b$ and $t_{a,b}(x,y) > 0$. Arbitrarily choose a vertex u in $y^{-1}(c_b)$ and define $y_1 \in \Omega$ by

$$y_1(z) := \begin{cases} c_a & \text{if } z = u \\ y(z) & \text{if } z \in [n] \setminus u \end{cases}$$

Then we obtain

$$t_{a,b}(x,y_1) + 1 = t_{a,b}(x,y)$$
$$t_{a,a}(x,y_1) - 1 = t_{a,a}(x,y)$$
$$t_{i,j}(x,y_1) = t_{i,j}(x,y), \ \forall \ (i,j) \in [k]^2 \setminus \{(a,b),(a,a)\}$$

Recall from (4.4) that

$$Q(x,y) = 2n \sum_{i,j \in [k]} t_{i,j}(x,y)(c_i - c_j)^2 - 2 \left[\sum_{i,j \in [k]} t_{i,j}(x,y)(c_i - c_j) \right]^2;$$

hence we have

$$
\begin{aligned}
& Q(x,y_1) - Q(x,y) \\
= \ & -2n(c_a - c_b)^2 + 2 \left[\sum_{i,j \in [k]} t_{i,j}(x,y)(c_u - c_v) \right]^2 - 2 \left[\sum_{i,j \in [k]} t_{i,j}(x,y)(c_u - c_v) - (c_a - c_b) \right]^2 \\
= \ & -(2n+2)(c_a - c_b)^2 + 4(c_a - c_b) \left[\sum_{i,j \in [k]} t_{i,j}(x,y)(c_i - c_j) \right] \\
\leq \ & -(2n+2)C_0 + C_1 \epsilon n,
\end{aligned}
$$

where $C_1 > 0$ is a constant given by

$$C_1 := 4k\sqrt{C_0} \max_{p,q \in [k]} |c_p - c_q|$$

Therefore

$$e^{-\frac{Q(x,y)}{8\sigma^2}} \leq e^{-\frac{Q(x,y_1)}{8\sigma^2}} e^{-\frac{n}{8\sigma^2} \left[2C_0 + \frac{2}{n}C_0 - C_1\epsilon \right]}$$

Recall the distance in Ω was defined by (3.4). Note that

$$d_\Omega(x,y_1) = d_\Omega(x,y) - 1$$

In general for $r \geq 1$, if we obtained y_r, we can obtain y_{r+1} as follows:

(1) If for all $(i,j) \in [k]^2$, $i \neq j$, $t_{i,j}(x,y) = 0$, then $x = y_r$.

(2) If $x \neq y_r$, find the least $(a, b) \in [k]^2$ in lexicographic order such that $a \neq b$ and $t_{a,b}(x, y_r) > 0$. Arbitrarily choose a vertex u in $y_r^{-1}(c_b)$ and define $y_{r+1} \in \Omega$ by

$$y_{r+1}(z) := \begin{cases} c_a & \text{if } z = u \\ y_r(z) & \text{if } z \in [n] \setminus u \end{cases}$$

Then by the same arguments as before, we have

$$d_\Omega(x, y_{r+1}) = d_\Omega(x, y_r) - 1$$

and

(4.12)
$$e^{-\frac{Q(x,y_r)}{8\sigma^2}} \leq e^{-\frac{Q(x,y_{r+1})}{8\sigma^2}} e^{-\frac{n}{8\sigma^2}\left[2C_0 + \frac{2}{n}C_0 - C_1\epsilon\right]}$$

By (3.4), we have for any $x, y \in \Omega$, $D_\Omega(x, y) \leq n$. Therefore if $x \neq y$, there exists $1 \leq l \leq n$, such that $y_s = x$. By (4.12) we have

$$\begin{aligned} e^{-\frac{Q(x,y)}{8\sigma^2}} &\leq e^{-\frac{Q(x,y_l)}{8\sigma^2}} e^{-\frac{nl}{8\sigma^2}\left[2C_0 + \frac{2}{n}C_0 - C_1\epsilon\right]} \\ &= e^{-\frac{nl}{8\sigma^2}\left[2C_0 + \frac{2}{n}C_0 - C_1\epsilon\right]}; \end{aligned}$$

where the last identity follows from (4.4) and

$$e^{-\frac{Q(x,y_l)}{8\sigma^2}} = e^{-\frac{Q(x,x)}{8\sigma^2}} = 1.$$

Therefore

(4.13)
$$J_2 \leq \sum_{l=1}^{\infty} (nk)^l e^{-\frac{[(2n+2)C_0 - C_1\epsilon n]l}{8\sigma^2}}$$

In the sum, l represents $D_\Omega(x, y)$. If $D_\Omega(x, y) = l$, then x can be obtained from y by changing colors at l vertices, each time there are at most n choices of vertices, and the color of each chosen vertex is changed to one of the k choices of colors. That is where the factor $(nk)^l$ on the right hand side comes from.

When σ satisfies (4.11), we have

$$(nk)e^{-\frac{[(2n+2)C_0 - C_1\epsilon n]}{8\sigma^2}} < e^{\log k - \log n\left(\frac{\delta}{1-\delta} + \frac{1}{n(1-\delta)} - \frac{C_1\epsilon}{2(1-\delta)C_0}\right)}$$

Then for (4.13) we obtain

$$J_2 \leq \frac{e^{\log k - \log n\left(\frac{\delta}{1-\delta} + \frac{1}{n(1-\delta)} - \frac{C_1\epsilon}{2(1-\delta)C_0}\right)}}{1 - e^{\log k - \log n\left(\frac{\delta}{1-\delta} + \frac{1}{n(1-\delta)} - \frac{C_1\epsilon}{2(1-\delta)C_0}\right)}}$$

By (4.9) and (4.10), we can choose $0 < \epsilon < \frac{2C_0\delta}{C_1}$ independent of n, then $\lim_{n\to\infty} J_2 = 0$. Then the proposition follows from (4.5) and the fact that $1 - p(\check{y}; \sigma) \leq J_1 + J_2$. $\square$

Let C_0 be defined as in (3.11). Let

$$A_m := \{c_u \in [k], \exists c_v \in [k], \text{s.t.} (c_u - c_v)^2 = C_0\}.$$

Proposition 4.4. *Assume for all n,*

(4.14)
$$\frac{\log |y^{-1}(A_m)|}{\log n} \geq \beta > 0$$

where β is a constant independent of n.

  Assume

(4.15)
$$\delta = \left(1 + \frac{4}{\beta}\right)\frac{\log\log n}{\log n},$$

and

(4.16)
$$\sigma^2 > \frac{(1+\delta)C_0 n}{4\beta \log n}$$

then

$$\lim_{n\to\infty} p(\hat{y}; \sigma) = 0,$$

Proof. For $y \in \Omega$, $a \in [n]$ such that $y(a) = c_u \in A_m$. Assume $c_v \in [k]$ such that $(c_u - c_v)^2 = C_0$. Define $y^{(a)} \in \Omega$ as follows:

$$y^{(a)}(i) = \begin{cases} y(i) & \text{if } i \in [n] \setminus \{a\} \\ c_v & \text{if } i = a \end{cases}$$

Then

$$
\begin{aligned}
t_{u,u}(y^{(a)}, y) &= n_u(y) - 1; \\
t_{v,u}(y^{(a)}, y) &= 1; \\
t_{i,i}(y^{(a)}, y) &= n_i(y), \text{ if } i \in [k] \setminus \{u\}; \\
t_{i,j}(y^{(a)}, y) &= 0, \text{ if } (i,j) \in [k]^2 \setminus \{(v,u)\} \text{ and } i \neq j;
\end{aligned}
$$

and for $z, w \in [n]$

(4.17)
$$G_{zw}(y^{(a)}) - G_{zw}(y) = \begin{cases} c_v - c_u & \text{if } z = a, \ w \neq a \\ c_u - c_v & \text{if } z \neq a, \ w = a \\ 0 & \text{Otherwise} \end{cases}$$

Note also that

$$1 - p(\check{y}; \sigma) \geq \Pr\left(\cup_{a\in[n], y(a)\in A_m} d(y^{(a)}) - d(y) < 0]\right),$$

since any of the event $[d(y^{(a)}) - d(y) < 0]$ implies $\check{y} \neq y$.

  By (4.1) we have

$$
\begin{aligned}
d(y^{(a)}) - d(y) &= \|\mathbf{G}(y^{(a)}) - \mathbf{G}(y)\|_F^2 - 2\sigma\langle \mathbf{W}, \mathbf{G}(y^{(a)}) - \mathbf{G}(y)\rangle \\
&= (2n-2)(c_u - c_v)^2 - 2\sigma\langle \mathbf{W}, \mathbf{G}(y^{(a)}) - \mathbf{G}(y)\rangle. \\
&= (2n-2)C_0 - 2\sigma\langle \mathbf{W}, \mathbf{G}(y^{(a)}) - \mathbf{G}(y)\rangle
\end{aligned}
$$

So $1 - p(\check{y}; \sigma)$ is at least

$$\Pr\left(\cup_{a, \in [n], y(a) \in A_m}[d(y^{(a)}) - d(y) < 0]\right)$$

$$\geq \quad \Pr\left(\max_{a \in [n], y(a) \in A_m} \sigma\langle \mathbf{W}, \mathbf{G}(y^{(a)}) - \mathbf{G}(y)\rangle > (n-1)C_0\right)$$

Let $H \subset y^{-1}(A_m)$ such that

$$(4.18) \qquad\qquad |H| = \frac{|y^{-1}(A_m)|}{\log^2 n} = h.$$

Then

$$1 - p(\check{y}; \sigma) \geq \Pr\left(\max_{a \in H} \sigma\langle \mathbf{W}, \mathbf{G}(y^{(a)}) - \mathbf{G}(y)\rangle > (n-1)C_0\right)$$

Let $(\mathcal{X}, \mathcal{Y}, \mathcal{Z})$ be a partition of $[n]^2$ defined by

$$\mathcal{X} = \{\alpha = (\alpha_1, \alpha_2) \in [n]^2, \{\alpha_1, \alpha_2\} \cap H| = \emptyset\}$$
$$\mathcal{Y} = \{\alpha = (\alpha_1, \alpha_2) \in [n]^2, |\{\alpha_1, \alpha_2\} \cap H| = 1\}$$
$$\mathcal{Z} = \{\alpha = (\alpha_1, \alpha_2) \in [n]^2, |\{\alpha_1, \alpha_2\} \cap H| = 2\}$$

For $\eta \in \{\mathcal{X}, \mathcal{Y}, \mathcal{Z}\}$, define the $n \times n$ matrix $\mathbf{W}_\eta$ from the entries of $\mathbf{W}$ as follows

$$\mathbf{W}_\eta(i, j) = \begin{cases} 0 & \text{if } (i, j) \notin \eta \\ \mathbf{W}(i, j), & \text{if } (i, j) \in \eta \end{cases}$$

For each $a \in H$, let

$$\mathcal{X}^{(a)} = \langle \mathbf{W}_{\mathcal{X}}, \mathbf{G}(y^{(a)}) - \mathbf{G}(y)\rangle$$
$$\mathcal{Y}^{(a)} = \langle \mathbf{W}_{\mathcal{Y}}, \mathbf{G}(y^{(a)}) - \mathbf{G}(y)\rangle$$
$$\mathcal{Z}^{(a)} = \langle \mathbf{W}_{\mathcal{Z}}, \mathbf{G}(y^{(a)}) - \mathbf{G}(y)\rangle$$

Claim 4.5. *The followings are true:*

(1) $\mathcal{X}^{(a)} = 0$ for $a \in H$.
(2) For each $a \in H$, the variables $\mathcal{Y}^{(a)}$ and $\mathcal{Z}^{(a)}$ are independent.
(3) $\{\mathcal{Y}^{(a)}\}_{a \in H}$ is a collection of i.i.d. Gaussian random variables.

Proof. It is straightforward to check (1). (2) holds because $\mathcal{Y} \cap \mathcal{Z} = \emptyset$.
For $s \in H$, let $\mathcal{Y}_s \subseteq \mathcal{Y}^{(a)}$ be defined by

$$\mathcal{Y}_s = \{\alpha = (\alpha_1, \alpha_2) \in \mathcal{Y} : \alpha_1 = s, \text{ or } \alpha_2 = s\}.$$

Note that for $s_1, s_2 \in H$ and $s_1 \neq s_2$, $\mathcal{Y}_{s_1} \cap \mathcal{Y}_{s_2} = \emptyset$. Moreover, $\mathcal{Y} = \cup_{s \in H}\mathcal{Y}_s$. Therefore

$$\mathcal{Y}^{(a)} = \sum_{s \in H} \langle \mathbf{W}_{\mathcal{Y}_s}, \mathbf{G}(y^{(a)}) - \mathbf{G}(y)\rangle$$

Note also that $\langle \mathbf{W}_{\mathcal{Y}_s}, \mathbf{G}(y^{(a)}) - \mathbf{G}(y)\rangle = 0$, if $s \neq a$. Hence

$$\mathcal{Y}^{(a)} = \sum_{\alpha \in \mathcal{Y}_a} [\mathbf{W}(\alpha)] \cdot \{[\mathbf{G}(y^{(a)}) - \mathbf{G}(y)](\alpha)\}$$

From (4.17) we obtain that $\alpha = (\alpha_1, \alpha_2) \in \mathcal{Y}_a$ with $\alpha_1 \neq \alpha_2$,

$$[\mathbf{G}(y^{(a)}) - \mathbf{G}(y)](\alpha) = \begin{cases} c_v - c_u & \text{if } \alpha_1 = a \\ c_u - c_v & \text{if } \alpha_2 = a. \end{cases}$$

So,

$$\begin{aligned} \mathcal{Y}^{(a)} &= \sum_{\alpha \in \mathcal{Y}_a} [\mathbf{W}(\alpha)] \cdot \{[\mathbf{G}(y^{(a)}) - \mathbf{G}(y)](\alpha)\} \\ &= \left\{ \sum_{\alpha \in \mathcal{Y}_a; \alpha_1 = a} [\mathbf{W}(\alpha)] - \sum_{\alpha \in \mathcal{Y}_a; \alpha_2 = a} [\mathbf{W}(\alpha)] \right\} (c_v - c_u) \end{aligned}$$

and $\{\mathcal{Y}^{(a)}\}_{a \in H}$ is a collection of independent Gaussian random variables. Moreover, the variance of $\mathcal{Y}^{(a)}$ is equal to $(2n - 2h)C_0$ independent of the choice of a. $\square$

By the claim, we obtain

$$\langle \mathbf{W}, \mathbf{G}(y^{(a)}) - \mathbf{G}(y) \rangle = \mathcal{Y}^{(a)} + \mathcal{Z}^{(a)}$$

Moreover,

$$\max_{a \in H} \mathcal{Y}^{(a)} + \mathcal{Z}^{(a)} \geq \max_{a \in H_u} \left[\mathcal{Y}^{(a)} \right] - \max_{a \in H_u} \left[\mathcal{Z}^{(a)} \right]$$

By the Lemma 3.9 we obtain when ϵ, h satisfy (3.22) with N replaced by h, the event

$$F_1 := \left\{ \max_{a \in H} \mathcal{Y}^{(a)} \geq (1 - \epsilon) \sqrt{2 \log h \cdot 2C_0 (n - h)} \right\}$$

has probability at least $1 - e^{-h^\epsilon}$; and the event

$$F_2 := \left\{ \max_{a \in H} \mathcal{Z}^{(a)} \leq (1 + \epsilon) \sqrt{2 \log h \cdot \max \mathrm{Var}(Z^{(a)})} \right\}$$

with probability at least $1 - h^{-\epsilon}$. Moreover,

$$\begin{aligned} \mathrm{Var} \mathcal{Z}^{(a)} &= \|\mathbf{G}(y^{(a)}) - \mathbf{G}(y)]\|_F^2 - \mathrm{Var}(\mathcal{Y}^{(a)}) \\ &= (2n - 2)C_0 - 2C_0 (n - h) \\ &= C_0(2h - 2) \end{aligned}$$

Hence

$$[F_1 \cap F_1] \subseteq F$$

where the event

$$F := \left\{ \max_{a \in H} \langle \mathbf{W}, \mathbf{G}(y^{(a)}) - \mathbf{G}(y) \rangle \geq \left(1 - \epsilon - (1 + \epsilon) \sqrt{\frac{h - 1}{n - h}} \right) \sqrt{2 \log h \cdot 2C_0(n - h)} \right\}$$

Hence

$$\begin{aligned} \mathrm{Pr}(F) &\geq \mathrm{Pr}(F_1 \cap F_2) = 1 - \mathrm{Pr}(F_1^c \cup F_2^c) \geq 1 - \mathrm{Pr}(F_1^c) - \mathrm{Pr}(F_2^c) \\ &= \mathrm{Pr}(F_1) + \mathrm{Pr}(F_2) - 1 \geq 1 - h^{-\epsilon} - e^{-h^\epsilon} \end{aligned}$$

Moreover, by (4.14) and (4.18) we have

$$\left(1 - \epsilon - (1+\epsilon)\sqrt{\frac{h-1}{n-h}}\right)\sqrt{2\log h \cdot 2C_0(n-h)}$$

$$\geq\ 2\sqrt{C_0\beta n \log n}\sqrt{1 - \frac{2\log\log n}{\beta \log n}}\sqrt{1 - \frac{1}{\log^2 n}}\left(1 - \epsilon - \frac{2}{\log n}\right)$$

Let

$$\tilde{F} :=$$
$$\left\{\max_{a \in H}\langle \mathbf{W}, \mathbf{G}(y^{(a)}) - \mathbf{G}(y)\rangle \geq 2\sqrt{C_0\beta n \log n}\sqrt{1 - \frac{2\log\log n}{\beta \log n}}\sqrt{1 - \frac{1}{\log^2 n}}\left(1 - \epsilon - \frac{2}{\log n}\right)\right\}$$

Then $F \subseteq \tilde{F}$.

Let ϵ be given as in (6.27), then when n is sufficiently large (3.22) holds with N replaced by h. When (4.15) and (4.16) hold,

$$\Pr\left(\max_{a \in y^{-1}(A_m)}\sigma\langle\mathbf{W}, \mathbf{G}(y^{(ab)}) - \mathbf{G}(y)\rangle > 2C_0 n\right)$$

$$\geq\ \Pr(\tilde{F}) \geq \Pr(F) \geq 1 - \frac{1}{\log n},$$

as n is sufficiently large. Then the proposition follows.   □

5. Proof of Theorem 2.2 when the number of vertices in each color is arbitrary in the sample space

Now we prove theorem 2.2 when the sample space for the MLE is Ω. Let $y \in \Omega$ be the true color assignment mapping. Assume that for each $x \in \Omega$,

$$n_i(x) = |x^{-1}(c_i)|, \text{ for } i \in [k]$$

are arbitrary positive integers satisfying (2.1) Let $\mathbf{K}(x)$ be defined as in (2.8), and $\mathbf{R}$ be defined as in (2.9).

Given a sample $\mathbf{R}$, we want to determine the the groups of vertices such that vertices within each group have the same color under the mapping y. Let $\tilde{y}$ be defined as in (2.11).

Again for $i, j \in [k]$, let $S_{i,j}(x, y)$ be defined as in (7.25), and $t_{i,j}(x, y) = |S_{i,j}(x, y)|$. Then

$$(5.1)\quad \sum_{i \in [k]} t_{i,j}(x, y) = n_j(y); \qquad \sum_{j \in [k]} t_{i,j}(x, y) = n_i(x); \qquad \sum_{i,j \in [k]} t_{i,j}(x, y) = n;$$

and

$$\langle \mathbf{K}(x), \mathbf{K}(y)\rangle\ =\ \sum_{a,b \in [n]} \mathbf{1}_{x(a) \neq x(b)}\mathbf{1}_{y(a) \neq y(b)}$$

$$=\ \sum_{i,j \in [k]} t_{i,j}(x, y)\left[n - n_i(x) - n_j(y) + t_{i,j}(x, y)\right]$$

$$(5.2)\qquad =\ n^2 - \sum_{i \in [k]}[n_i(x)]^2 - \sum_{j \in [k]}[n_j(y)]^2 + \sum_{i,j \in [k]}[t_{i,j}(x, y)]^2;$$

where

$$n - n_i(x) - n_j(y) + t_{i,j}(x, y) = \{u \in [n] : x(u) \neq i, \text{ and } y(u) \neq j\};$$

and (5.2) follows from (5.1). In particular,

$$\langle \mathbf{K}(y), \mathbf{K}(y) \rangle = n^2 - \sum_{j \in [k]} [n_j(y)]^2$$

Hence from (2.11) we obtain

$$\tilde{y} = \mathrm{argmax}_{x \in \Omega} \left[\sum_{j=1}^k [n_j(x)]^2 + 2\langle \mathbf{K}(x), \mathbf{R} \rangle \right]$$

Define

(5.3) $$g(x) := \sum_{j=1}^k [n_j(x)]^2 + 2\langle \mathbf{K}(x), \mathbf{R} \rangle;$$

then

(5.4) $$g(x) - g(y)$$

$$= 2\sigma \langle \mathbf{K}(x) - \mathbf{K}(y), \mathbf{W} \rangle + 2 \sum_{i,j \in [k]} [t_{i,j}(x, y)]^2 - \sum_{i=1}^k [n_i(x)]^2 - \sum_{j=1}^k [n_j(y)]^2.$$

Note that $g(x) - g(y)$ is a Gaussian random variable with mean $2\sum_{i,j \in [k]} [t_{i,j}(x, y)]^2 - \sum_{i=1}^k [n_i(x)]^2 - \sum_{j=1}^k [n_j(y)]^2$ and variance $4\sigma^2 \|\mathbf{K}(x) - \mathbf{K}(y)\|_F^2$.

Let

(5.5) $$L(x, y) = \sum_{i=1}^k [n_i(x)]^2 + \sum_{j=1}^k [n_j(y)]^2 - 2 \sum_{i,j \in [k]} [t_{i,j}(x, y)]^2$$

Then it is straightforward to check that

(5.6) $$\|\mathbf{K}(x) - \mathbf{K}(y)\|_F^2 = L(x, y)$$

Therefore

$$\mathrm{Pr}(g(x) - g(y) > 0) = \mathrm{Pr}_{\xi \sim \mathcal{N}(0,1)} \left(\xi > \frac{\sqrt{L(x, y)}}{2\sigma} \right) \leq e^{-\frac{L(x,y)}{8\sigma^2}},$$

where the last inequality follows from the fact that if $\xi \sim \mathcal{N}(0, 1)$, then for $x > 0$, $\mathrm{Pr}(\xi > x) \leq e^{-\frac{x^2}{2}}$.

Definition 5.1. *For $y \in \Omega$, let $C(y)$ consist of all the $x \in \Omega$ such that x can be obtained from y by a permutation of colors. More precisely, $x \in C(y) \subset \Omega$ if and only if the following condition holds*

- *for $i, j \in [n]$, $y(i) = y(j)$ if and only if $x(i) = x(j)$.*

We define an equivalence relation on Ω as follows: we say $x, z \in \Omega$ are equivalent if and only if $x \in C(z)$. Let $\overline{\Omega}$ be the set of all the equivalence classes in Ω.

We have the following elementary lemma:

Lemma 5.2. *If $x, z \in \Omega$ are equivalent, then*

$$\sum_{i,j \in [k]} [t_{i,j}(x,y)]^2 = \sum_{i,j \in [k]} [t_{i,j}(z,y)]^2$$

Proof. By definition if $x, z \in \Omega$ are equivalent, then there exists a permutation ω of $[k]$, such that for all $l \in [n]$, we have

(5.7) $z(l) = \omega(x(l)).$

Therefore for all $i \in [k]$, the following two sets are equal:

(5.8) $x^{-1}(i) = z^{-1}(\omega(i)).$

In other words, for any $u \in [n]$ and $i \in [k]$, $x(u) = i$ if and only if $z(u) = \omega(i)$. Then we have

$$t_{i,j}(x,y) = t_{\omega(i),j}(z,y)$$

by summing over all the i, j's in $[k]$, and using the fact that ω is a bijection from $[k]$ to $[k]$, we obtain the lemma. $\square$

Lemma 5.3. *Let $y \in \Omega$ be the true color assignment mapping. If x and z are equivalent elements in Ω, then for any chosen sample $\mathbf{W}$, we have*

(5.9) $g(x) = g(z).;$

and

(5.10) $L(x,y) = L(z,y).$

Moreover, if $y^* \in C(y)$, then

(5.11) $L(x,y) = L(x,y^*)$

Proof. From the definition (5.3) of $g(x)$ we obtain

$$g(x) = 2\langle \mathbf{K}(x), \mathbf{K}(y) \rangle + 2\sigma \langle \mathbf{K}(x), \mathbf{W} \rangle + \sum_{j=1}^{k} [n_j(x)]^2$$

Recall that $\mathbf{K}_{i,j}(x) = 1$ if and only if $x(i) \neq x(j)$; if $x(i) = x(j)$, $\mathbf{K}_{i,j}(x) = 0$. Since x and z are equivalent $x(i) \neq x(j)$ if and only if $z(i) \neq z(j)$, therefore

(5.12) $\mathbf{K}(x) = \mathbf{K}(z).$

Moreover, if x and z are equivalent and (5.7) holds, by (5.8) we obtain $x \in \Omega_{n_{\omega(1)}(z),\dots,n_{\omega(k)}(z)}$; in particular this implies

$$\sum_{j=1}^{k} [n_j(x)]^2 = \sum_{j=1}^{k} [n_{\omega(j)}(z)]^2 = \sum_{j=1}^{k} [n_j(z)]^2$$

Then we obtain (5.9). The expression (5.10) follows from (5.12) and (5.6). The expression (5.11) follows from (5.10) by observing that the expression (5.5) of $L(x,y)$ is symmetric in x and y. $\square$

Let

$$p(\tilde{y}, \sigma) = \Pr(\tilde{y} \in C(y))$$

Then

$$p(\tilde{y}, \sigma) = \Pr\left(g(y) > \max_{x \in \Omega \setminus C(y)} g(x)\right);$$

hence

$$1 - p(\tilde{y}, \sigma) \leq \sum_{C(x) \in \overline{\Omega} \setminus \{C(y)\}} \Pr(g(x) - g(y) \geq 0) \leq \sum_{C(x) \in \overline{\Omega} \setminus \{C(y)\}} e^{-\frac{L(x,y)}{8\sigma^2}}.$$

Lemma 5.4. *For any $x, y \in \Omega$, $L(x, y) \geq 0$, where the equality holds if and only if $x \in C(y)$.*

Proof. By (5.1) and (5.5) we have

$$(5.13) \quad L(x, y)$$

$$= 2\left[\sum_{i \in [k]} \sum_{1 \leq j_1 < j_2 \leq k} t_{i,j_1}(x, y) t_{i,j_2}(x, y) + \sum_{j \in [k]} \sum_{1 \leq i_1 < i_2 \leq k} t_{i_1,j}(x, y) t_{i_2,j}(x, y)\right] \geq 0$$

We first note that $L(x, y) = 0$ if $x \in C(y)$. Indeed, if $x \in C(y)$, then there exists a permutation $\omega : [k] \to [k]$, such that $x = \omega \circ y$. Then for all $u \in [k]$ satisfying $x(u) = i$, we have $y(u) = \omega^{-1}(i)$. Then

$$t_{i,\omega^{-1}(i)} = n_i(x);$$
$$t_{i,j} = 0, \ \forall \, j \in [k] \setminus \{\omega^{-1}(i)\}.$$

Therefore for any $j_1, j_2 \in [k]$ and $j_1 < j_2$, at least one of $t_{i,j_1}(x, y)$ and $t_{i,j_2}(x, y)$ is 0. Similarly for any $i_1, i_2 \in [k]$ and $i_1 < i_2$, at least one of $t_{i_1,j}(x, y)$ and $t_{i_2,j}(x, y)$ is 0. Then $L(x, y) = 0$ by (5.13).

It remains to show that if $L(x, y) = 0$, then $x \in C(y)$. Note that if $L(x, y) = 0$, then

$$t_{i,j_1}(x, y) t_{i,j_2}(x, y) = 0, \ \forall i \in [k], \ 1 \leq j_1 < j_2 \leq k; \text{ and}$$
$$t_{i_1,j}(x, y) t_{i_2,j}(x, y) = 0, \ \forall j \in [k], \ 1 \leq i_1 < i_2 \leq k$$

Then for any fixed $i \in [k]$, there exists exactly one $j \in [k]$, such that $t_{i,j} \neq 0$; and for each fixed $j \in [k]$, there exists exactly one $i \in [k]$, such that $t_{i,j} \neq 0$. Then the lemma follows. $\qquad\square$

Let $\tilde{\mathcal{B}}$ be the set given by

$$(5.14) \quad \tilde{\mathcal{B}} = \left\{(t_{1,1}, t_{1,2}, \ldots, t_{k,k}) \in \{0, 1, \ldots, n\}^{k^2} : \sum_{i=1}^{k} t_{i,j} = n_j(y)\right\}.$$

For a small positive number $\epsilon > 0$, let $\tilde{\mathcal{B}}_\epsilon$ be the set given by

$$(5.15) \quad \tilde{\mathcal{B}}_\epsilon = \left\{(t_{1,1}, \ldots, t_{k,k}) \in \tilde{\mathcal{B}} : \ \forall \, i \in [k], \exists j \in [k], \text{ s.t. } t_{j,i} \geq n_i(y) - n\epsilon\right\}$$

Lemma 5.5. *Let $\mathcal{R}_c$ be defined in (4.6). When*

(5.16) $$(t_{1,1}(x,y), \ldots, t_{k,k}(x,y)) \in \tilde{\mathcal{B}} \setminus \tilde{\mathcal{B}}_\epsilon,$$

and (4.7) hold, we have

$$L(x,y) \geq \frac{c\epsilon n^2}{k}.$$

Proof. When (5.16) holds, by (5.15) we have there exists $i_0 \in [k]$, such that

$$n_{i_0}(y) - \max_{j \in [k]} t_{j,i_0}(x,y) \geq \epsilon n.$$

By (5.13), we obtain

$$L(x,y) \geq \max_{j \in [k]} t_{j,i_0}(x,y) \left[n_{i_0}(y) - \max_{j \in [k]} t_{j,i_0}(x,y) \right] \geq \frac{n_{i_0}(y)}{k} \epsilon n$$

By (4.7), we have

$$\frac{n_{i_0}(y)}{k} \epsilon n \geq \frac{c\epsilon n^2}{k}$$

$\square$

Lemma 5.6. *Let $\mathcal{R}_c$ be defined in (4.6). Assume (4.7) holds. Let $x \in \Omega$ satisfy*

(5.17) $$\left(\frac{n_1(x)}{n}, \frac{n_2(x)}{n}, \ldots, \frac{n_k(x)}{n} \right) \in \mathcal{R}_{\frac{2c}{3}}$$

For $i \in [k]$, let

(5.18) $$t_{w(i),i}(x,y) = \max_{j \in [k]} t_{j,i}(x,y),$$

where $w(i) \in [k]$. When $\epsilon \in \left(0, \frac{2c}{3k} \right)$ and $(t_{1,1}(x,y), \ldots, t_{k,k}(x,y)) \in \tilde{\mathcal{B}}_\epsilon$, w is a bijection from $[k]$ to $[k]$.

Proof. When $(t_{1,1}(x,y), \ldots, t_{k,k}(x,y)) \in \tilde{\mathcal{B}}_\epsilon$, by (5.15) we have

$$\max_{j \in [k]} t_{j,i}(x,y) > n_i(y) - \epsilon n, \ \forall \ i \in [k].$$

If there exist $j_1, j_2 \in [k]$, such that $j_1 \neq j_2$, and

$$t_{j_1,i}(x,y) = t_{j_2,i}(x,y) = \max_{j \in [k]} t_{j,i}(x,y);$$

then

$$n_i(y) \geq t_{j_1,i} + t_{j_2,i} > 2n_i(y) - 2\epsilon n \geq \left(2 - \frac{2\epsilon}{c} \right) n_i(y);$$

where the last inequality follows from (4.7). But this is impossible when $\epsilon < \frac{c}{2}$. Therefore $w(i)$ satisfying (5.18) is unique for each $i \in [k]$, and w is a mapping from $[k]$ to $[k]$.

If there exist $i, j \in [k]$ such that $i \neq j$ and $w(i) = w(j) = l \in [k]$; then there exists $q \in [k]$, such that $q \neq w(s)$ for all $s \in [k]$. More precisely,

$$t_{q,s}(x,y) \neq \max_{r \in [k]} t_{r,s}(x,y) > n_s(y) - \epsilon n, \ \forall \ s \in [k].$$

By (5.1) we have

(5.19)
$$t_{q,s}(x,y) \le \epsilon n, \ \forall s \in [k].$$

Therefore

$$\frac{2cn}{3} \le \sum_{s \in [k]} t_{q,s}(x,y) = n_q(x) \le k\epsilon n,$$

where the first inequality follows from (5.17), and the last inequality follows from (5.19). But this is impossible when $\epsilon < \frac{2c}{3k}$. The contradiction implies that w must be a bijection from $[k]$, and the lemma follows. $\qquad\square$

We have the following proposition

Proposition 5.7. *Let $\mathcal{R}_c$ be defined as in (4.6). Assume (4.7) and (5.17) holds. Recall that k is the total number of colors. Here c, k may depend on n. Assume the following conditions hold*

(5.20)
$$\lim_{n \to \infty} \frac{\log k}{c^3 \log n} = 0;$$

If there exists $\delta > 0$ independent of n, such that

(5.21)
$$\sigma^2 < \frac{(1-\delta)[n_k(y) + n_{k-1}(y)]}{4 \log n}$$

then

$$\lim_{n \to \infty} p(\tilde{y}; \sigma) = 1$$

Proof. Note that

$$\sum_{C(x) \in \overline{\Omega} \backslash \{C(y)\}} e^{-\frac{L(x,y)}{8\sigma^2}} \le I_3 + I_4$$

where

$$I_3 \quad = \sum_{C(x) \in \overline{\Omega}: \left(t_{1,1}(x,y),\dots,t_{k,k}(x,y)\right) \in [\mathcal{B} \backslash \tilde{\mathcal{B}}_\epsilon], C(x) \ne C(y)} e^{\frac{-L(x,y)}{8\sigma^2}}$$

and

$$I_4 = \sum_{C(x) \in \overline{\Omega}: \left(t_{1,1}(x,y),\dots,t_{k,k}(x,y)\right) \in \tilde{\mathcal{B}}_\epsilon, C(x) \ne C(y)} e^{\frac{-L(x,y)}{8\sigma^2}}.$$

By Lemma 5.5, when (5.21) holds, we have

$$I_3 \quad \le \quad k^n e^{-\frac{c\epsilon n^2 \log n}{2k(1-\delta)[n_k(y)+n_{k-1}(y)]}}$$

Since

(5.22)
$$n_k(y) + n_{k-1}(y) \le \frac{2n}{k},$$

we have

(5.23)
$$I_3 \le e^{n\left(\log k - \frac{c\epsilon \log n}{4(1-\delta)}\right)}.$$

Now let us consider I_4. Assume $\epsilon \in \left(0, \frac{2c}{3k}\right)$. Let w be the bijection from $[k]$ to $[k]$ as defined in (5.18). Let $y^* \in \Omega$ be defined by

$$y^*(z) = w(y(z)), \ \forall z \in [n].$$

Then $y^* \in C(y)$. Moreover, x and y^* satisfies

$$(5.24) \qquad\qquad t_{i,i}(x, y^*) \geq n_i(y^*) - n\epsilon, \ \forall i \in [k].$$

We consider the following color changing process to obtain x from y^*.

(1) If for all $(j, i) \in [k]^2$, and $j \neq i$, $t_{j,i}(x, y^*) = 0$, then $x = y^*$.
(2) If (1) does not hold, find the least $(j, i) \in [k]^2$ in lexicographic order such that $j \neq i$ and $t_{j,i}(x, y^*) > 0$. Choose an arbitrary vertex $u \in S_{j,i}(x, y^*)$. Define $y_1 \in \Omega$ as follows

$$y_1(z) = \begin{cases} c_j & \text{if } z = u \\ y^*(z) & \text{if } z \in [n] \setminus \{u\} \end{cases}$$

Then we have

$$(5.25) \qquad\qquad t_{j,i}(x, y_1) = t_{j,i}(x, y^*) - 1$$
$$(5.26) \qquad\qquad t_{j,j}(x, y_1) = t_{j,j}(x, y^*) + 1$$
$$(5.27) \qquad\qquad t_{a,b}(x, y_1) = t_{a,b}(x, y^*) \ \forall (a, b) \in \left([k]^2 \setminus \{(j, i), (j, j)\}\right).$$

Therefore x and y^* satisfies

$$t_{i,i}(x, y_1) \geq n_i(y_1) - n\epsilon, \ \forall i \in [k].$$

From (5.11) and (5.13) we obtain

$$
\frac{1}{2}\left(L(x, y_1) - L(x, y)\right) = \frac{1}{2}\left(L(x, y_1) - L(x, y^*)\right)
$$

$$
= t_{j,i}(x, y_1)\left(\sum_{l \neq i} t_{j,l}(x, y_1)\right) + t_{j,j}(x, y_1)\left(\sum_{l \neq j} t_{j,l}(x, y_1)\right) - t_{j,i}(x, y_1)t_{j,j}(x, y_1)
$$

$$
+ t_{j,i}(x, y_1)\left(\sum_{l \neq j} t_{l,i}(x, y_1)\right) + t_{j,j}(x, y_1)\left(\sum_{l \neq j} t_{l,j}(x, y_1)\right)
$$

$$
- \Bigg[t_{j,i}(x, y^*)\left(\sum_{l \neq i} t_{j,l}(x, y^*)\right) + t_{j,j}(x, y^*)\left(\sum_{l \neq j} t_{j,l}(x, y^*)\right) - t_{j,i}(x, y^*)t_{j,j}(x, y
$$

$$
+ t_{j,i}(x, y^*)\left(\sum_{l \neq j} t_{l,i}(x, y^*)\right) + t_{j,j}(x, y^*)\left(\sum_{l \neq j} t_{l,j}(x, y^*)\right) \Bigg]
$$

By (5.25) we obtain

$$t_{j,i}(x,y_1)\left(\sum_{l\neq i}t_{j,l}(x,y_1)\right)-t_{j,i}(x,y^*)\left(\sum_{l\neq i}t_{j,l}(x,y^*)\right)$$

$$=\ t_{j,i}(x,y_1)\left[n_j(x)-t_{j,i}(x,y_1)\right]-t_{j,i}(x,y^*)\left[n_j(x)-t_{j,i}(x,y^*)\right]$$

$$=\ t_{j,i}(x,y^*)+t_{j,i}(x,y_1)-n_j(x)$$

Similarly by (5.26) we obtain

$$t_{j,j}(x,y_1)\left(\sum_{l\neq j}t_{j,l}(x,y_1)\right)-t_{j,j}(x,y^*)\left(\sum_{l\neq j}t_{j,l}(x,y^*)\right)$$

$$=\ n_j(x)-t_{j,j}(x,y_1)-t_{j,j}(x,y^*).$$

$$-t_{j,i}(x,y_1)t_{j,j}(x,y_1)+t_{j,i}(x,y^*)t_{j,j}(x,y^*)=-t_{j,i}(x,y^*)+t_{j,j}(x,y^*)+1$$

$$t_{j,i}(x,y_1)\left(\sum_{l\neq j}t_{l,i}(x,y_1)\right)-t_{j,i}(x,y^*)\left(\sum_{l\neq j}t_{l,i}(x,y^*)\right)=t_{j,i}(x,y^*)-n_i(y^*)$$

$$t_{j,j}(x,y_1)\left(\sum_{l\neq j}t_{l,j}(x,y_1)\right)-t_{j,j}(x,y^*)\left(\sum_{l\neq j}t_{l,j}(x,y^*)\right)=n_j(y^*)-t_{j,j}(x,y^*)$$

Therefore we obtain

$$L(x,y_1)-L(x,y)\ =\ 2\left[n_j(y^*)-n_i(y^*)-2t_{j,j}(x,y^*)+2t_{j,i}(x,y^*)-1\right]$$

Then by (6.17),

$$L(x,y_1)-L(x,y)\leq -2n_i(y^*)-2n_j(y^*)+8n\epsilon\leq -2(n_{k-1}(y)+n_k(y)-4n\epsilon)$$

Therefore when (5.21) holds,

$$(5.28)\qquad e^{-\frac{L(x,y)}{8\sigma^2}}\leq e^{-\frac{L(x,y_1)}{8\sigma^2}}e^{-\frac{\log n}{1-\delta}\left(1-\frac{4n\epsilon}{n_{k-1}(y)+n_k(y)}\right)}$$

In general, if we have constructed y_r $(r\geq 1)$, we shall construct y_{r+1} as follows.

(a) If for all $(j,i)\in[k]^2$, and $j\neq i$, $t_{j,i}(x,y_r)=0$, then $x=y_r$.
(b) If (a) does not hold, find the least $(j,i)\in[k]^2$ in lexicographic order such that $j\neq i$ and $t_{j,i}(x,y_r)>0$. Choose an arbitrary vertex $u\in S_{j,i}(x,y_r)$. Define $y_{r+1}\in\Omega$ as follows

$$y_{r+1}(z)=\begin{cases}c_j & \text{if } z=u\\ y_r(z) & \text{if } z\in[n]\setminus\{u\}\end{cases}$$

Then if (6.17) holds with y^* replaced by y_r, then (6.17) holds with y^* replaced by y_{r+1}. By similar computations as above we obtain

$$(5.29)\qquad e^{-\frac{L(x,y_r)}{8\sigma^2}}\leq e^{-\frac{L(x,y_{r+1})}{8\sigma^2}}e^{-\frac{\log n}{1-\delta}\left(1-\frac{4n\epsilon}{n_{k-1}(y)+n_k(y)}\right)}.$$

Recall that the distance D_Ω in Ω is defined in (3.4). From the constructions of y_{r+1} we have

$$D_\Omega(x, y_{r+1}) = D_\Omega(x, y) - 1.$$

Therefore there exists $s \in [n]$, such that $y_s = x$. By (5.28) and (5.29) and the fact that $L(x, x) = 0$ we obtain

$$e^{-\frac{L(x,y)}{8\sigma^2}} \leq e^{-\frac{s \log n}{1-\delta}\left(1 - \frac{4n\epsilon}{n_{k-1}(y) + n_k(y)}\right)}$$

Since any x in $\tilde{\mathcal{B}}_\epsilon$ can be obtained by y be the color changing process described above, we have

$$(5.30) \qquad I_4 \leq \sum_{l=1}^{\infty} (nk)^l e^{-\frac{l \log n}{1-\delta}\left(1 - \frac{4n\epsilon}{n_{k-1}(y) + n_k(y)}\right)}$$

The right hand side of (5.30) is the sum of geometric series with both initial term and common ratio equal to

$$(5.31) \qquad V := e^{\log k - \log n\left(\frac{\delta}{1-\delta} - \frac{4n\epsilon}{(1-\delta)(n_{k-1}(y) + n_k(y))}\right)}$$

By (4.7) we have

$$n_{k-1}(y) + n_k(y) \geq 2cn$$

$$(5.32) \qquad V \leq e^{\log k - \log n\left(\frac{\delta}{1-\delta} - \frac{2\epsilon}{c(1-\delta)}\right)}$$

Let

$$(5.33) \qquad \epsilon = \min\left\{\frac{c}{2k}, \frac{\delta c}{4}\right\} < \frac{c}{k};$$

then from (5.23) and using the fact that $c \leq \frac{1}{k}$, we obtain

$$0 \leq I_3 \leq e^{n\left(\log k - \log n \min\left(\frac{c^3}{8}, \frac{\delta c^2}{16}\right)\right)}$$

Then from (5.20), when n is sufficiently large

$$\log k \leq \frac{c^3 \log n}{16};$$

Since $k \geq 2$ we have

$$n\left(\log k - \frac{c^3 \log n}{8}\right) \leq -n \log k \to -\infty,$$

as $n \to \infty$. Moreover, since $c \leq 1$ from (5.20) we obtain

$$\lim_{n \to \infty} \frac{\log k}{c^2 \log n} = 0$$

Hence when n is sufficiently large, we obtain

$$n\left(\log k - \frac{\delta c^2 \log n}{16}\right) \leq -n \log k \to -\infty,$$

as $n \to \infty$.

Therefore we obtain

$$\lim_{n\to\infty} I_3 = 0. \tag{5.34}$$

Moreover, when ϵ, V are given by (5.33) and (5.31), from (5.32) we obtain

$$0 \le V \le e^{\log k - \frac{\delta \log n}{2(1-\delta)}}$$

By (5.20), for any constant $\delta > 0$ independent of n, $\lim_{n\to\infty} V = 0$, therefore

$$\lim_{n\to\infty} I_4 = \lim_{n\to\infty} \frac{V}{1-V} = 0. \tag{5.35}$$

Then the proposition follows from (5.34), (5.35) and the fact that $1 - p(\tilde{y}; \sigma) \le I_3 + I_4$. $\quad\square$

Proposition 5.8. *Let $\mathcal{R}_c$ be defined as in (4.6). Assume (4.7) and (5.17) holds. Here k and c may depend on n. Assume $n_1(y) \ge n_2(y) \ge \ldots \ge n_k(y)$ and*

$$\frac{\log n_k(y)}{\log n} \ge 1 + \frac{\log c}{\log n} \ge \beta > 0, \ \forall \ n, \tag{5.36}$$

where β is a constant independent of n.

    Assume

$$\delta = \left(1 + \frac{4}{\beta}\right) \frac{\log \log n}{\log n}, \tag{5.37}$$

and

$$\sigma^2 > \frac{(1+\delta)[n_k(y) + n_{k-1}(y)]}{4\beta \log n}, \tag{5.38}$$

then

$$\lim_{n\to\infty} p(\tilde{y}; \sigma) = 0.$$

Proof. For $y \in \Omega$, $a \in [n]$ such that $c_{k-1} = y(a)$. Let $y^{(a)} \in \Omega$ be defined by

$$y^{(a)}(i) = \begin{cases} y(i) & \text{if } i \in [n], \text{ and } i \ne a \\ c_k & \text{if } i = a. \end{cases}$$

Then

$$t_{k,k-1}(y^{(a)}, y) = 1;$$
$$t_{k-1,k-1}(y^{(a)}, y) = n_{k-1}(y) - 1;$$
$$t_{i,i}(y^{(a)}, y) = n_i(y); \ \forall \ i \in [k] \setminus \{k-1\};$$
$$t_{i,j}(y^{(a)}, y) = 0; \ \forall (i,j) \in [k]^2 \setminus \{(k, k-1)\}, \text{ and } i \ne j.$$

and

$$n_k(y^{(a)}) = n_k(y) + 1;$$
$$n_{k-1}(y^{(a)}) = n_{k-1}(y) - 1;$$
$$n_i(y^{(a)}) = n_i(y); \ \forall \ i \in [k] \setminus \{k, k-1\}.$$

Moreover,

$$1 - p(\tilde{y}; \sigma) \geq \Pr\left(\cup_{a \in [n] \cap y^{-1}(c_{k-1})}[g(y^{(a)}) - g(y) > 0]\right)$$

Since any of the event $[g(y^{(a)}) - g(y) > 0]$ implies $\tilde{y} \neq y$.

From (5.2) we obtain

$$(5.39) \quad \langle \mathbf{K}(y), \mathbf{K}(y^{(a)}) - \mathbf{K}(y) \rangle = \sum_{i,j \in [k]} [t_{i,j}(y^{(a)}, y)]^2 - \sum_{i \in [k]} [n_i(y^{(a)})]^2 = -2n_k(y)$$

Then from (5.4) we have

$$g(y^{(a)}) - g(y) = \sum_{i \in [k]} [n_i(y^{(a)})]^2 - \sum_{i \in [k]} [n_i(y)]^2 + 2\langle \mathbf{K}(y), \mathbf{K}(y^{(a)}) - \mathbf{K}(y) \rangle + 2\sigma\langle \mathbf{W}, \mathbf{K}(y^{(a)}) - \mathbf{K}(y) \rangle$$

$$= -2n_{k-1}(y) - 2n_k(y) + 2 + 2\sigma\langle \mathbf{W}, \mathbf{K}(y^{(a)}) - \mathbf{K}(y) \rangle.$$

Let $H_{k-1} \subset y^{-1}(c_{k-1})$, such that

$$(5.40) \qquad\qquad |H_{k-1}| = \frac{n_{k-1}(y)}{\log^2 n} = h.$$

Then $1 - p(\tilde{y}; \sigma)$ is at least

$$\Pr\left(\cup_{a \in [n] \cap y^{-1}(c_{k-1})}[g(y^{(a)}) - g(y) > 0]\right)$$

$$\geq \Pr\left(\max_{a \in [n] \cap y^{-1}(c_{k-1})} \sigma\langle \mathbf{W}, \mathbf{K}(y^{(a)}) - \mathbf{K}(y) \rangle > n_k(y) + n_{k-1}(y) - 1\right)$$

$$\geq \Pr\left(\max_{a \in H_{k-1}} \sigma\langle \mathbf{W}, \mathbf{K}(y^{(a)}) - \mathbf{K}(y) \rangle > n_k(y) + n_{k-1}(y) - 1\right)$$

Let $(\mathcal{X}, \mathcal{Y}, \mathcal{Z})$ be a partition of $[n]^2$ defined by

$$\mathcal{X} = \{\alpha = (\alpha_1, \alpha_2) \in [n]^2, \{\alpha_1, \alpha_2\} \cap [H_{k-1}] = \emptyset\}$$
$$\mathcal{Y} = \{\alpha = (\alpha_1, \alpha_2) \in [n]^2, |\{\alpha_1, \alpha_2\} \cap [H_{k-1}]| = 1\}$$
$$\mathcal{Z} = \{\alpha = (\alpha_1, \alpha_2) \in [n]^2, |\{\alpha_1, \alpha_2\} \cap [H_{k-1}]| = 2\}$$

For $\eta \in \{\mathcal{X}, \mathcal{Y}, \mathcal{Z}\}$, define the $n \times n$ matrix $\mathbf{W}_\eta$ from the entries of $\mathbf{W}$ as follows

$$\mathbf{W}_\eta(i, j) = \begin{cases} 0 & \text{if } (i, j) \notin \eta \\ \mathbf{W}(i, j), & \text{if } (i, j) \in \eta \end{cases}$$

For each $a \in H_{k-1}$, let

$$\mathcal{X}_a = \langle \mathbf{W}_\mathcal{X}, \mathbf{K}(y^{(a)}) - \mathbf{K}(y) \rangle$$
$$\mathcal{Y}_a = \langle \mathbf{W}_\mathcal{Y}, \mathbf{K}(y^{(a)}) - \mathbf{K}(y) \rangle$$
$$\mathcal{Z}_a = \langle \mathbf{W}_\mathcal{Z}, \mathbf{K}(y^{(a)}) - \mathbf{K}(y) \rangle$$

Explicit computations show that

$$\mathbf{K}_{i,j}(y^{(a)}) - \mathbf{K}_{i,j}(y) = \begin{cases} -1 & \text{if } i = a, \text{ and } j \in y^{-1}(c_k) \\ 1 & \text{if } i = a, \text{ and } j \in y^{-1}(c_{k-1}), j \neq a \\ -1 & \text{if } j = a, \text{ and } i \in y^{-1}(c_k) \\ 1 & \text{if } j = a, \text{ and } i \in y^{-1}(c_{k-1}), \ i \neq a \\ 0 & \text{otherwise.} \end{cases}$$

Claim 5.9. *The followings are true:*

   (1) $\mathcal{X}_a = 0$ for $a \in H_{k-1}$.
   (2) For each $a \in H_{k-1}$, the variables $\mathcal{Y}_a$ and $\mathcal{Z}_a$ are independent.

Proof. It is straightforward to check (1). (2) holds because $\mathcal{Y} \cap \mathcal{Z} = \emptyset$. □

For $s \in H_{k-1}$, let $\mathcal{Y}^s \subseteq \mathcal{Y}$ be defined by

$$\mathcal{Y}^s = \{\alpha = (\alpha_1, \alpha_2) \in \mathcal{Y} : \alpha_1 = s, \text{ or } \alpha_2 = s\}.$$

Note that for $s_1, s_2 \in H_{k-1}$ and $s_1 \neq s_2$, $\mathcal{Y}^{s_1} \cap \mathcal{Y}^{s_2} = \emptyset$. Moreover, $\mathcal{Y} = \cup_{s \in H_{k-1}} \mathcal{Y}^s$. Therefore

$$\mathcal{Y}_a = \sum_{s \in H_{k-1}} \langle \mathbf{W}_{\mathcal{Y}^s}, \mathbf{K}(y^{(a)}) - \mathbf{K}(y) \rangle$$

Note also that $\langle \mathbf{W}_{\mathcal{Y}^s}, \mathbf{K}(y^{(a)}) - \mathbf{K}(y) \rangle = 0$, if $s \neq a$. Hence

$$\mathcal{Y}_a = \sum_{\alpha \in \mathcal{Y}^a} [\mathbf{W}(\alpha)] \cdot \{[\mathbf{K}(y^{(a)}) - \mathbf{K}(y)](\alpha)\}$$

Note that for $\alpha \in \mathcal{Y}^a$,

$$[\mathbf{K}(y^{(a)}) - \mathbf{K}(y)](\alpha) = \begin{cases} -1 & \text{if } \{\alpha_1, \alpha_2\} \cap [y^{-1}(c_k)] = 1 \\ 1 & \text{if } \{\alpha_1, \alpha_2\} \cap [y^{-1}(c_{k-1})] = 2 \\ 0 & \text{else.} \end{cases}$$

So,

$$\sum_{\alpha \in \mathcal{Y}^a} [\mathbf{W}(\alpha)] \cdot \{[\mathbf{K}(y^{(a)}) - \mathbf{K}(y)](\alpha)\} = \sum_{\alpha \in \mathcal{Y}^a; \{\alpha_1, \alpha_2\} \cap [y^{-1}(c_k)] = 2} [\mathbf{W}(\alpha)] - \sum_{\alpha \in \mathcal{Y}^a; \{\alpha_1, \alpha_2\} \cap [y^{-1}(c_k)] = 1} [\mathbf{W}(\alpha)]$$

$\{\mathcal{Y}_s\}_{s \in H_{k-1}}$ is a collection of independent centered Gaussian random variables. Moreover, the variance of $\mathcal{Y}_s$ is equal to $2(n_k(y) + n_{k-1}(y) - h)$ independent of the choice of s.

By the claim, we obtain

$$\langle \mathbf{W}, \mathbf{K}(y^{(a)}) - \mathbf{K}(y) \rangle = \mathcal{Y}_a + \mathcal{Z}_a$$

Moreover,

$$\max_{a \in H_{k-1}} \mathcal{Y}_a + \mathcal{Z}_a \ \geq \ \max_{a \in H_{k-1}} \mathcal{Y}_a - \max_{a \in H_{k-1}} (-\mathcal{Z}_a)$$

By Lemma 3.9 about the tail bound result of the maximum of Gaussian random variables, if (3.22) holds with N replaced by h, the event

$$E_1^* := \left\{ \max_{a \in H_{k-1}} \mathcal{Y}_a \geq (1 - \epsilon)\sqrt{2\log h \cdot 2\left(n_k(y) + n_{k-1}(y) - h\right)} \right\}$$

has probability at least $1 - e^{-h^\epsilon}$; and the event

$$E_2^* := \left\{ \max_{a \in H_{k-1}} \mathcal{Z}_a \leq (1 + \epsilon)\sqrt{2\log h \cdot \max_{a \in H_{k-1}} \mathrm{Var}(\mathcal{Z}_a)} \right\}$$

has probability $1 - h^{-\epsilon}$.

Note that

$$(5.41) \qquad \langle \mathbf{K}(y^{(a)}), \mathbf{K}(y^{(a)}) - \mathbf{K}(y) \rangle \;=\; -\sum_{i,j \in [k]} [t_{i,j}(y, y^{(a)})]^2 + \sum_{i \in [k]} [n_i(y)]^2$$

$$= \; 2n_{k-1}(y) - 2$$

Moreover, by (5.39) and (5.41)

$$\mathrm{Var}\mathcal{Z}_a \;=\; \|\mathbf{K}(y^{(a)}) - \mathbf{K}(y)\|_F^2 - \mathrm{Var}(\mathcal{Y}_a)$$

$$= \; 2\left[n_{k-1}(y) + n_k(y)\right] - 2 - 2\left(n_{k-1}(y) + n_k(y) - h\right)$$

$$= \; 2h - 2$$

Let

$$E^* := \left\{ \max_{a \in H_{k-1}} \mathcal{Y}_a + \mathcal{Z}_a \geq \left(1 - \epsilon - (1 + \epsilon)\sqrt{\frac{h-1}{n_k(y) + n_{k-1}(y) - h}}\right)\sqrt{4\log h(n_k(y) + n_{k-1}(y) - h)} \right\}$$

Then $E_1^* \cap E_2^* \subseteq E^*$. Moreover, by (5.40) and (5.36) we have

$$\left(1 - \epsilon - (1 + \epsilon)\sqrt{\frac{h-1}{n_k(y) + n_{k-1}(y) - h}}\right)\sqrt{4\log h(n_k(y) + n_{k-1}(y) - h)}$$

$$\geq \; 2\left(1 - \epsilon - \frac{1+\epsilon}{\log n}\right)\sqrt{\beta\log n\,(n_k(y) + n_{k-1}(y))}\sqrt{\left(1 - \frac{1}{\log^2 n}\right)\left(1 - \frac{2\log\log n}{\beta\log n}\right)}$$

Let $\tilde{E}^*$ be the event defined by

$$\tilde{E}^* := \left\{ \max_{a \in H_{k-1}} \mathcal{Y}_a + \mathcal{Z}_a \geq 2\left(1 - \epsilon - \frac{1+\epsilon}{\log n}\right)\sqrt{\beta\log n\,(n_k(y) + n_{k-1}(y))}\sqrt{\left(1 - \frac{1}{\log^2 n}\right)\left(1 - \frac{2\log\log n}{\beta\log n}\right)} \right\}$$

Then $E^* \subseteq \tilde{E}^*$.

Let ϵ be given as in (6.27), then when n is sufficiently large (3.22) holds with N replaced by h. When (5.37) and (5.38) holds,

$$\mathrm{Pr}\left(\max_{a \in H_{k-1}} \sigma\langle \mathbf{W}, \mathbf{K}(y^{(a)}) - \mathbf{K}(y)\rangle > n_k(y) + n_{k-1}(y) - 1\right) \geq \mathrm{Pr}(\tilde{E}^*) \geq \mathrm{Pr}(E^*),$$

as n is sufficiently large. Moreover, we have

$$
\begin{aligned}
\Pr(E^*) &\geq \Pr(E_1^* \cap E_2^*) \\
&\geq 1 - \Pr(E_1^c) - \Pr(E_2^c) \\
&\geq 1 - \log n,
\end{aligned}
$$

when n is sufficiently large. Then the proposition follows.

$\square$

6. Proof of Theorem 2.2 when the number of vertices in each color is fixed in the sample space

We now discuss the MLE in Theorem 2.2 when the number of vertices of each color in the sample space is fixed to be the same as the true value. Let $y \in \Omega_{n_1,\ldots,n_k}$ be the unknown true color assignment mapping. Let $\mathbf{K}(y)$ be defined as in (2.8), and let $\mathbf{R}$ be defined as in (2.9). For given sample $\mathbf{R}$, let $\overline{y}$ be defined as in (2.11).

Given a sample $\mathbf{R}$, the goal is to find the exact division of $[n]$ into groups of vertices such that all the vertices in the same group have the same color under the true color assignment mapping y. Note that for all $x \in \Omega_{n_1,\ldots,n_k}$,

$$
\|\mathbf{K}(x)\|_F^2 = 2 \sum_{1 \leq i < j \leq k} n_i n_j,
$$

which depends only on $n_1, \ldots, n_k$, but is independent of x. Then we have

$$
\overline{y} = \mathrm{argmax}_{x \in \Omega_{n_1,\ldots,n_k}} \langle \mathbf{K}(x), \mathbf{R} \rangle
$$

We define an equivalence relation on $\Omega_{n_1,\ldots,n_k}$ as follows:

Definition 6.1. *For $y \in \Omega_{n_1,\ldots,n_k}$, let $C^*(y)$ consist of all the $x \in \Omega_{n_1,\ldots,n_k}$ such that x can be obtained from y by a permutation of colors. More precisely, $x \in C^*(y) \subset \Omega_{n_1,\ldots,n_k}$ if and only if the following condition holds*

- *for $i, j \in [n]$, $y(i) = y(j)$ if and only if $x(i) = x(j)$.*

We define an equivalence relation on $\Omega_{n_1,\ldots,n_k}$ as follows: we say $x, z \in \Omega_{n_1,\ldots,n_k}$ are equivalent if and only if $x \in C^*(z)$. Let $\overline{\Omega}_{n_1,\ldots,n_k}$ be the set of all the equivalence classes in $\Omega_{n_1,\ldots,n_k}$.

It is not hard to check that for each $x \in \Omega_{n_1,\ldots,n_k}$, $C^*(x) = C(x) \cap \Omega_{n_1,\ldots,n_k}$, where $C(x)$ is defined as in Definition 5.1. Moreover, for $x, z \in \Omega_{n_1,\ldots,n_k}$, $x \in C^*(z)$ if and only if there exists a permutation ω of $[k]$, such that

$$
x = \omega(z)
$$

and for any $i \in [k]$,

$$
|z^{-1}(i)| = |z^{-1}(\omega(i))|.
$$

Let

(6.1) $$p(\overline{y}, \sigma) = \Pr(\overline{y} \in C^*(y)),$$

where $y \in \Omega_{n_1,\dots,n_k}$ is the true color assignment function.

For each $x \in \Omega_{n_1,\dots,n_k}$, define

$$(6.2) \qquad\qquad h(x) = \langle \mathbf{K}(x), \mathbf{R} \rangle$$

Lemma 6.2. *Let $y \in \Omega_{n_1,\dots,n_k}$ be the true color assignment mapping. If $x, z \in \Omega_{n_1,\dots,n_k}$ is such that $x \in C^*(z)$, then*

$$(6.3) \qquad\qquad \mathbf{K}(x) = \mathbf{K}(z);$$

and

$$(6.4) \qquad\qquad h(x) = h(z).$$

Proof. From (6.2) we see that (6.4) follows from (6.3). Recall that $\mathbf{K}_{i,j}(x) = 1$ if and only if $x(i) = x(j)$; if $x(i) \neq x(j)$, $\mathbf{K}_{i,j}(x) = 0$. Since x and z are equivalent $x(i) = x(j)$ if and only if $z(i) = z(j)$, therefore $\mathbf{K}(x) = \mathbf{K}(z)$. $\qquad\square$

Then

$$p(\overline{y}, \sigma) = \Pr\left(h(y) > \max_{x \in \Omega_{n_1,\dots,n_k} \setminus C^*(y)} h(x) \right)$$

Note that

$$h(x) - h(y) = \langle \mathbf{K}(y), \mathbf{K}(x) - \mathbf{K}(y) \rangle + \sigma \langle \mathbf{W}, \mathbf{K}(x) - \mathbf{K}(y) \rangle.$$

The expression above shows that $h(x) - h(y)$ is a Gaussian random variable with mean $\langle \mathbf{K}(y), \mathbf{K}(x) - \mathbf{K}(y) \rangle$ and variance $\sigma^2 \|\mathbf{K}(x) - \mathbf{K}(y)\|_F^2$.

For $i, j \in [k]$, $x, y \in \Omega_{n_1,\dots,n_k}$, let $S_{i,j}(x, y)$ be defined as in (3.2). Let $t_{i,j}(x, y) = |S_{i,j}(x, y)|$. Let

$$U(x, y) := -\mathbb{E}[h(x) - h(y)] = \langle \mathbf{K}(y), \mathbf{K}(y) - \mathbf{K}(x) \rangle;$$

Then by (5.2) we have

$$(6.5) \qquad\qquad U(x, y) = \sum_{i \in [k]} [n_i]^2 - \sum_{i,j \in [k]} [t_{i,j}(x, y)]^2$$

and

$$\|\mathbf{K}(x) - \mathbf{K}(y)\|_F^2 = 2U(x, y).$$

For $x \in \Omega_{n_1,\dots,n_k} \setminus C^*(y)$

$$\Pr\left(h(x) - h(y) > 0 \right) \quad = \quad \Pr_{\xi \sim \mathcal{N}(0,1)}\left(\xi \geq \frac{\sqrt{U(x,y)}}{\sqrt{2}\sigma} \right)$$

Using the standard Gaussian tail bound $\Pr_{\xi \in \mathcal{N}(0,1)}(\xi > x) < e^{-\frac{1}{2}x^2}$, we obtain

$$\Pr\left(h(x) - h(y) > 0 \right) \leq e^{-\frac{U(x,y)}{4\sigma^2}}$$

Then

(6.6)
$$1 - p(\overline{y}; \sigma) \leq \sum_{x \in \Omega \setminus C^*(y)} \Pr(h(x) - h(y) > 0)$$

$$\leq \sum_{x \in \Omega \setminus C^*(y)} e^{-\frac{U(x,y)}{4\sigma^2}}$$

Lemma 6.3. *For any $x, y \in \Omega_{n_1,\ldots,n_k}$, $U(x,y) \geq 0$, where the equality holds if and only if $x \in C^*(y)$.*

Proof. Note that

(6.7)
$$U(x,y) = \frac{1}{2} L(x,y)\bigg|_{x,y \in \Omega_{n_1,\ldots,n_k}}.$$

Then the lemma follows from Lemma 5.4. $\square$

Lemma 6.4. *Assume $x, z_1, z_2 \in \Omega_{n_1,\ldots,n_k}$. Assume $z_1 \in C^*(z_2)$. Then*

$$U(x, z_1) = U(x, z_2).$$

Proof. The lemma follows from (5.11) and (6.7). $\square$

Lemma 6.5. *Let $\tilde{\mathcal{B}}$ and $\tilde{\mathcal{B}}_\epsilon$ be defined as in (5.14) and (5.15), respectively. Let $\mathcal{R}_c$ be defined as in (4.6). Let $x, y \in \Omega_{n_1,\ldots,n_k}$. Assume (4.7) and (5.16) hold. Then*

$$U(x,y) \geq \frac{c\epsilon n^2}{2k}$$

Proof. The lemma follows from (6.7) and Lemma 5.5. $\square$

Lemma 6.6. *Let $y \in \Omega_{n_1,\ldots,n_k}$ be the true color assignment mapping. Let $\mathcal{R}_c$ be defined in (4.6). Assume (4.7) holds. Let $x \in \Omega_{n_1,\ldots,n_k}$ For $i \in [k]$, let $w(i) \in [k]$ be defined as in (5.18). Then*

   (1) when $\epsilon \in \left(0, \frac{c}{k}\right)$ and $(t_{1,1}(x,y), \ldots, t_{k,k}(x,y)) \in \tilde{\mathcal{B}}_\epsilon$, w is a bijection from $[k]$ to $[k]$.
   (2) Assume there exist $i, j \in [k]$, such that $n_i \neq n_j$. If

(6.8)
$$\epsilon < \min_{i,j \in [k]: n_i \neq n_j} \left| \frac{n_i - n_j}{n} \right|$$

   Then for any $i \in [k]$,

(6.9)
$$n_i = |y^{-1}(i)| = |y^{-1}(w(i))| = n_{w(i)}.$$

Proof. Part (1) of the lemma is a special case of Lemma 5.6 when $x \in \Omega_{n_1,\ldots,n_k}$. Now we prove Part (2) of the lemma. We shall prove by contradiction. Assume (6.9) does not hold, then there exists $i_0 \in [k]$, such that

(6.10)
$$n_i > n_{w(i)}.$$

By (5.18), we obtain

$$n_{w_i} \geq t_{w(i),i}(x,y) > n_i - \epsilon n$$

Hence we have

$$\epsilon > \frac{n_i - n_w(i)}{n}$$

But this is a contradiction to (6.8) and (6.10). The contradiction implies Part (2) of the lemma. $\square$

We have the following proposition

Proposition 6.7. *Let $\mathcal{R}_c$ be defined as in (4.6). Let $y \in \Omega_{n_1,\dots,n_k}$ be the true color assignment mapping satisfying (4.7), where $n_1 \geq n_2 \geq \dots \geq n_k$. Here c, k may depend on n. Assume*

$$(6.11) \qquad \lim_{n\to\infty} \frac{c}{k(n_k + n_{k-1})} = 0;$$

$$(6.12) \qquad \lim_{k\to\infty} \frac{\log k}{\log n} = 0;$$

and

$$(6.13) \qquad \lim_{n\to\infty} \frac{cn \min\left\{\frac{c}{k}, \min_{i,j\in[k],n_i\neq n_j} \left|\frac{n_i - n_j}{n}\right|\right\}}{(n_{k-1} + n_k)} \geq \alpha > 0,$$

where $\alpha > 0$ is a constant independent of n; and $\min_{i,j\in[k],n_i\neq n_j} \left|\frac{n_i - n_j}{n}\right| = \infty$ if $n_i = n_j$ for all $i, j \in [k]$. If there exists a constant $\delta > 0$ independent of n, such that

$$(6.14) \qquad \sigma^2 < \frac{(1-\delta)[n_{k-1} + n_k]}{4\log n}$$

then

$$\lim_{n\to\infty} p(\overline{y}; \sigma) = 1.$$

Proof. Let

$$\Gamma := \sum_{x\in\overline{\Omega}_{n_1,\dots,n_k}\backslash C^*(y)} e^{-\frac{U(x,y)}{4\sigma^2}}.$$

By (6.6), it suffices to show that $\lim_{n\to\infty} \Gamma = 0$.

Let $0 < \epsilon < \min\left\{\frac{c}{k}, \min_{i,j\in[k],n_i\neq n_j} \left|\frac{n_i - n_j}{n}\right|\right\}$. Then

$$\Gamma \leq \Gamma_1 + \Gamma_2;$$

where

$$\Gamma_1 \quad = \sum_{C^*(x)\in\overline{\Omega}_{n_1,\dots,n_k}:\left(t_{1,1}(x,y),\dots,t_{k,k}(x,y)\right)\in[\tilde{\mathcal{B}}\backslash\tilde{\mathcal{B}}_\epsilon],C(x)\neq C(y)} e^{\frac{-U(x,y)}{4\sigma^2}}$$

and

$$(6.15) \qquad \Gamma_2 = \sum_{C^*(x)\in\overline{\Omega}_{n_1,\dots,n_k}:\left(t_{1,1}(x,y),\dots,t_{k,k}(x,y)\right)\in\tilde{\mathcal{B}}_\epsilon,C(x)\neq C(y)} e^{\frac{-U(x,y)}{4\sigma^2}}.$$

By Lemma 6.5, when (6.14) holds, we have

$$\Gamma_1 \leq k^n e^{-\frac{c \epsilon n^2 \log n}{2(1-\delta)k(n_{k-1}+n_k)}}$$

When

$$\epsilon = \frac{1}{2} \min \left\{ \frac{c}{k}, \min_{i,j \in [k], n_i \neq n_j} \left| \frac{n_i - n_j}{n} \right| \right\},$$

we have

$$\Gamma_1 \leq e^{n \left(\log k - \frac{cn \log n \min \left\{ \frac{c}{k}, \min_{i,j \in [k], n_i \neq n_j} \left| \frac{n_i - n_j}{n} \right| \right\}}{4(1-\delta)(n_{k-1}+n_k)} \right)}$$

Then by (6.12) and (6.13) we obtain that

(6.16)
$$0 \leq \lim_{n \to \infty} \Gamma_1 \leq \lim_{n \to \infty} e^{-n \log k} = 0.$$

since $k \geq 2$.

Now let us consider Γ_2. Recall that $y \in \Omega_{n_1,\ldots,n_k}$ is the true color assignment mapping. Let w be the bijection from $[k]$ to $[k]$ as defined in (5.18). Let $y^* \in \Omega$ be defined by

$$y^*(z) = w(y(z)), \ \forall z \in [n].$$

Then $y^* \in C(y)$. By Part (2) of Lemma 6.6, we obtain that for $i \in [k]$

$$\left| [y^*]^{-1}(i) \right| = \left| y^{-1}(w^{-1}(i)) \right| = \left| y^{-1}(i) \right|;$$

therefore $y^* \in \Omega_{n_1,\ldots,n_k}$. Moreover, x and y^* satisfies

(6.17)
$$t_{i,i}(x,y^*) \geq n_i(y^*) - n\epsilon, \ \forall i \in [k].$$

If $x \neq y^*$, by Lemma 3.3, there exists an l-cycle $(i_1, \ldots, i_l)$ for (x, y^*) with $2 \leq l \leq k$. Then for each $2 \leq s \leq (l+1)$, choose an arbitrary vertex u_s in $S_{i_{s-1},i_s}(x,y)$, and let $y_1(u_s) = c_{i_{s-1}}$, where $i_{l+1} := i_1$. For any vertex $z \in [n] \setminus \{u_2, \ldots, u_{l+1}\}$, let $y_1(z) = y^*(z)$.

Note that $y_1 \in \Omega_{n_1,\ldots,n_k}$. More precisely, for $1 \leq s \leq l$, we have

$$t_{i_s,i_s}(x,y^*) + 1 = t_{i_s,i_s}(x,y_1);$$
$$t_{i_s,i_{s+1}}(x,y^*) - 1 = t_{i_s,i_{s+1}}(x,y_1)$$

and

$$t_{a,b}(x,y^*) = t_{a,b}(x,y_1), \ \forall (a,b) \notin \{(i_s,i_s),(i_s,i_{s+1})\}_{s=1}^l.$$

From (6.5) and Lemma 6.4 we obtain

$$\begin{aligned}
U(x,y_1) - U(x,y) &= U(x,y_1) - U(x,y^*) \\
&= -\sum_{i,j \in [k]} [t_{i,j}(x,y_1)]^2 + \sum_{i,j \in [k]} [t_{i,j}(x,y^*)]^2 \\
&= 2 \sum_{s=1}^l [t_{i_s,i_{s+1}}(x,y^*) - t_{i_s,i_s}(x,y^*)].
\end{aligned}$$

When $(t_{1,1}(x,y), \ldots, t_{k,k}(x,y)) \in \tilde{\mathcal{B}}_\epsilon$, we obtain

$$U(x, y_1) - U(x, y) \ \leq \ 2 \sum_{s=1}^{l} [2n\epsilon - n_{i_s}(y^*)] \leq -l\,(n_k + n_{k-1} - 4n\epsilon)$$

Therefore

(6.18) $$e^{-\frac{U(x,y)}{4\sigma^2}} \leq e^{-\frac{U(x,y_1)}{4\sigma^2}} e^{-\frac{l(n_k+n_{k-1}-4n\epsilon)}{4\sigma^2}}.$$

If $y_1 \neq x$, we find an l_2-cycle $(2 \leq l_2 \leq k)$ for (x, y_1), change colors along the l_2-cycle as above, and obtain another cycle assignment mapping $y_2 \in \Omega_{n_1,\ldots,n_k}$, and so on. Let $y_0 := y$, and note that for each $r \geq 1$, if y_r is obtained from y_{r-1} by changing colors along an l_r cycle, we have

$$D_\Omega(x, y_r) = D_\Omega(x, y_{r-1}) - l_r$$

Therefore finally we can obtain x from y by changing colors along at most $\lfloor \frac{n}{2} \rfloor$ cycles. Using similar arguments as those used to derive (6.18), we obtain that for each r

$$e^{-\frac{U(x,y_{r-1})}{4\sigma^2}} \leq e^{-\frac{U(x,y_r)}{4\sigma^2}} e^{-\frac{l_r(n_k+n_{k-1}-4n\epsilon)}{4\sigma^2}}.$$

Therefore if $y_s = x$ for some $1 \leq s \leq \lfloor \frac{n}{2} \rfloor$, we have

$$e^{-\frac{U(x,y)}{4\sigma^2}} \leq e^{-\frac{U(x,y_s)}{4\sigma^2}} e^{-\frac{(n_k+n_{k-1}-4n\epsilon)\left(\sum_{i\in[s]} l_i\right)}{4\sigma^2}}.$$

By Lemma 6.3, we have $U(x, y_s) = U(x, x) = 0$, hence

$$e^{-\frac{U(x,y)}{4\sigma^2}} \leq \prod_{i\in[s]} e^{-\frac{(n_k+n_{k-1}-4n\epsilon)l_i}{4\sigma^2}}.$$

Note also that for any $r_1 \neq r_2$, in the process of obtaining y_{r_1} from y_{r_1-1} and the process of obtaining y_{r_2} from y_{r_2-1}, we change colors on disjoint sets of vertices. Hence the order of these steps of changing colors along cycles does not affect the final color assignment mapping we obtain. From (6.15) we obtain

(6.19) $$\Gamma_2 \leq \prod_{l=2}^{k} \left(\sum_{m_l=0}^{\infty} (nk)^{m_l l} e^{-\frac{m_l l(n_k+n_{k-1}-4n\epsilon)}{4\sigma^2}} \right) - 1$$

On the right hand side of (6.19), when expanding the product, each summand has the form

$$\left[(nk)^{2m_2} e^{-\frac{2m_2(n_k+n_{k-1}-4n\epsilon)}{4\sigma^2}} \right] \cdot \left[(nk)^{3m_3} e^{-\frac{3m_3(n_k+n_{k-1}-4n\epsilon)}{4\sigma^2}} \right] \cdot \ldots \cdot \left[(nk)^{km_k} e^{-\frac{km_k(n_k+n_{k-1}-4n\epsilon)}{4\sigma^2}} \right]$$

where the factor $\left[(nk)^{2m_2} e^{-\frac{2m_2(n_k+n_{k-1}-4n\epsilon)}{4\sigma^2}} \right]$ represents that we changed along 2-cycles m_2 times, the factor $\left[(nk)^{3m_3} e^{-\frac{3m_3(n_k+n_{k-1}-4n\epsilon)}{4\sigma^2}} \right]$ represents that we changed along 3-cycles m_3 times, and so on. Moreover, each time we changed along an l-cycle, we need to first determine the l different colors involved in the l-cycle, and there are at most k^l different

l-cycles; we then need to chose l vertices to change colors, and there are at most n^l choices. It is straightforward to check that if σ satisfies (6.14), then

$$nke^{-\frac{n_k+n_{k-1}-4\epsilon}{4\sigma^2}} \leq e^{\log k - \frac{\log n}{1-\delta}\left(\delta - \frac{4\epsilon}{n_k+n_{k-1}}\right)}.$$

When $\epsilon < \frac{c}{k}$ and (6.11) holds, we obtain

$$e^{\log k - \frac{\log n}{1-\delta}\left(\delta - \frac{4\epsilon}{n_k+n_{k-1}}\right)} \leq e^{\log k - \frac{\log n}{1-\delta}\left(\delta - \frac{4c}{k(n_k+n_{k-1})}\right)}$$
$$\leq e^{\log k - \frac{\delta \log n}{2(1-\delta)}}$$

When (6.12) holds, we obtain

$$e^{\log k - \frac{\delta \log n}{2(1-\delta)}} \leq e^{-\frac{\delta \log n}{4(1-\delta)}} \to 0$$

as $n \to \infty$.

Therefore we have

$$\sum_{m_l=0}^{\infty} (nk)^{m_l l} e^{-\frac{m_l l(n_k+n_{k-1}-4n\epsilon)}{4\sigma^2}} \leq \frac{1}{1 - e^{l\log k - \frac{l\log n}{1-\delta}\left(\delta - \frac{4\epsilon}{n_k+n_{k-1}}\right)}} \leq \frac{1}{1 - e^{-\frac{\delta ln}{4(1-\delta)}}}$$

Let

$$\Sigma := \prod_{l=2}^{k} \left(\sum_{m_l=0}^{\infty} (nk)^{m_l l} e^{-\frac{m_l l(n_k+n_{k-1}-4\epsilon)}{4\sigma^2}} \right).$$

Since $\log(1+x) \leq x$ for $x \geq 0$, we have

$$
\begin{aligned}
0 \leq \log \Sigma &= \sum_{l=2}^{k} \log \left(1 + \sum_{m_l=1}^{\infty} (nk)^{m_l l} e^{-\frac{m_l l(n_k+n_{k-1}-4n\epsilon)}{4\sigma^2}} \right) \\
&\leq \sum_{l=2}^{k} \sum_{m_l=1}^{\infty} (nk)^{m_l l} e^{-\frac{m_l l(n_k+n_{k-1}-4n\epsilon)}{4\sigma^2}} \\
&\leq \sum_{l=2}^{k} \frac{\left(e^{-\frac{\delta \log n}{4(1-\delta)}} \right)^l}{1 - \left(e^{-\frac{\delta \log n}{4(1-\delta)}} \right)^l} \\
&\leq \frac{\left(e^{-\frac{\delta \log n}{4(1-\delta)}} \right)^2}{\left[1 - \left(e^{-\frac{\delta \log n}{4(1-\delta)}} \right)^2 \right] \left(1 - e^{-\frac{\delta \log n}{4(1-\delta)}} \right)} \to 0,
\end{aligned}
$$

as $n \to \infty$. Then

(6.20) $$0 \leq \lim_{n\to\infty} \Gamma_2 \leq \lim_{n\to\infty} e^{\log \Sigma} - 1 = 0.$$

Then the proposition follows from (6.16) and (6.20). $\qquad\square$

Proposition 6.8. *Assume $n_1 \geq n_2 \geq n_k$, and $y \in \Omega_{n_1,\dots,n_k}$ is the true color assignment mapping. Suppose that*

$$(6.21) \qquad\qquad\qquad\qquad \frac{\log n_k}{\log n} \geq \beta > 0, \ \forall \ n,$$

where β is a constant independent of n.

  Assume

$$(6.22) \qquad\qquad\qquad\qquad \delta = \left(1 + \frac{4}{\beta}\right)\frac{\log\log n}{\log n},$$

and

$$(6.23) \qquad\qquad\qquad\qquad \sigma^2 > \frac{(1+\delta)[n_k(y) + n_{k-1}(y)]}{4\beta\log n},$$

then

$$\lim_{n\to\infty} p(\overline{y}; \sigma) = 0.$$

Proof. For $y \in \Omega_{n_1,\dots,n_k}$, $a, b \in [n]$ such that $c_k = y(a) \neq y(b) = c_{k-1}$. Let $y^{(ab)} \in \Omega_{n_1,\dots,n_k}$ be the coloring of vertices defined by

$$(6.24) \qquad\qquad\qquad y^{(ab)}(i) = \begin{cases} y(i) & \text{if } i \in [n] \setminus \{a, b\} \\ c_{k-1} & \text{if } i = a \\ c_k & \text{if } i = b \end{cases}$$

Then

$$t_{k-1,k}(y^{(ab)}, y) - 1 = t_{k-1,k}(y, y) = 0$$
$$t_{k-1,k-1}(y^{(ab)}, y) + 1 = t_{k-1,k-1}(y, y) = n_{k-1}$$
$$t_{k,k-1}(y^{(ab)}, y) - 1 = t_{k,k-1}(y, y) = 0$$
$$t_{k,k}(y^{(ab)}, y) + 1 = t_{k,k}(y, y) = n_k.$$

and

$$t_{i,j}(y^{(ab)}, y) = t_{i,j}(y), \ \forall \ (i,j) \in \left([k]^2 \setminus \{(k,k), (k, k-1), (k-1, k), (k-1, k-1)\}\right)$$

Note that

$$1 - p(\overline{y}; \sigma) \geq \Pr\left(\cup_{a,b\in[n], c_k=y(a)\neq y(b)=c_{k-1}}[h(y^{(ab)}) - h(y) > 0]\right),$$

since any of the event $[h(y^{(ab)}) - h(y) > 0]$ implies $\overline{y} \neq y$. By (5.2) we obtain

$$\begin{aligned}
h(y^{(ab)}) - h(y) &= \langle \mathbf{K}(y), \mathbf{K}(y^{(ab)}) - \mathbf{K}(y) \rangle + \sigma\langle \mathbf{W}, \mathbf{K}(y^{(ab)}) - \mathbf{K}(y) \rangle \\
&= \sum_{i,j\in[k]} [t_{i,j}(y^{(ab)}, y)]^2 - \sum_{i\in[k]} [n_i]^2 + \sigma\langle \mathbf{W}, \mathbf{K}(y^{(ab)}) - \mathbf{K}(y) \rangle \\
&= 4 - 2n_k - 2n_{k-1} + \sigma\langle \mathbf{W}, \mathbf{K}(y^{(ab)}) - \mathbf{K}(y) \rangle.
\end{aligned}$$

So $1 - p(\overline{y}; \sigma)$ is at least

$$\Pr\left(\cup_{a,b\in[n],c_k=y(a)\neq y(b)=c_{k-1}}[h(y^{(ab)}) - h(y) > 0]\right)$$

$$\geq \quad \Pr\left(\max_{a,b\in[n],c_k=y(a)\neq y(b)=c_{k-1}}\sigma\langle\mathbf{W},\mathbf{K}(y^{(ab)}) - \mathbf{K}(y)\rangle > 2(n_k + n_{k-1} - 2)\right)$$

For $i \in \{k - 1, k\}$, let $H_i \subset y^{-1}(c_i)$ such that

(6.25) $$|H_i| = \frac{n_k}{\log^2 n} = h.$$

Then

$$1 - p(\overline{y}; \sigma) \geq \Pr\left(\max_{a\in H_{k-1},b\in H_k}\sigma\langle\mathbf{W},\mathbf{K}(y^{(ab)}) - \mathbf{K}(y)\rangle > 2(n_k + n_{k-1} - 2)\right)$$

Let $(\mathcal{X}, \mathcal{Y}, \mathcal{Z})$ be a partition of $[n]^2$ defined by

$$\mathcal{X} = \{\alpha = (\alpha_1, \alpha_2) \in [n]^2, \{\alpha_1, \alpha_2\} \cap [H_{k-1} \cup H_k] = \emptyset\}$$
$$\mathcal{Y} = \{\alpha = (\alpha_1, \alpha_2) \in [n]^2, |\{\alpha_1, \alpha_2\} \cap [H_{k-1} \cup H_k]| = 1\}$$
$$\mathcal{Z} = \{\alpha = (\alpha_1, \alpha_2) \in [n]^2, |\{\alpha_1, \alpha_2\} \cap [H_{k-1} \cup H_k]| = 2\}$$

For $\eta \in \{\mathcal{X}, \mathcal{Y}, \mathcal{Z}\}$, define the $n \times n$ matrix $\mathbf{W}_\eta$ from the entries of $\mathbf{W}$ as follows

$$\mathbf{W}_\eta(i, j) = \begin{cases} 0 & \text{if } (i, j) \notin \eta \\ \mathbf{W}(i, j), & \text{if } (i, j) \in \eta \end{cases}$$

For each $a \in H_k$ and $b \in H_{k-1}$, let

$$\mathcal{X}_{ab} = \langle\mathbf{W}_\mathcal{X}, \mathbf{K}(y^{(ab)}) - \mathbf{K}(y)\rangle$$
$$\mathcal{Y}_{ab} = \langle\mathbf{W}_\mathcal{Y}, \mathbf{K}(y^{(ab)}) - \mathbf{K}(y)\rangle$$
$$\mathcal{Z}_{ab} = \langle\mathbf{W}_\mathcal{Z}, \mathbf{K}(y^{(ab)}) - \mathbf{K}(y)\rangle$$

Lemma 6.9. *The followings are true:*

(1) $\mathcal{X}_{ab} = 0$ for $a \in H_k$ and $b \in H_{k-1}$.
(2) For each $a \in H_k$ and $b \in H_{k-1}$, the variables $\mathcal{Y}_{ab}$ and $\mathcal{Z}_{ab}$ are independent.
(3) Each $\mathcal{Y}_{ab}$ can be decomposed into $Y_a + Y_b$ where $\{Y_a\}_{a\in H_k}\cup\{Y_b\}_{b\in H_{k-1}}$ is a collection of i.i.d. Gaussian random variables.

Proof. Note that for $i, j \in [n]$,

$$(6.26) \qquad \mathbf{K}_{i,j}(y^{(ab)}) - \mathbf{K}_{i,j}(y) = \begin{cases} 1 & \text{if } i = a, j \in y^{-1}(c_k), \text{ and } j \neq a. \\ -1 & \text{if } i = a, j \in y^{-1}(c_{k-1}) \\ -1 & \text{if } i = b, j \in y^{-1}(c_k) \\ 1 & \text{if } i = b, j \in y^{-1}(c_{k-1}) \text{ and } j \neq b \\ 1 & \text{if } j = a, i \in y^{-1}(c_k), \text{ and } i \neq a. \\ -1 & \text{if } j = a, i \in y^{-1}(c_{k-1}) \\ -1 & \text{if } j = b, i \in y^{-1}(c_k) \\ 1 & \text{if } j = b, i \in y^{-1}(c_{k-1}) \text{ and } i \neq b \\ 0 & \text{otherwise.} \end{cases}$$

It is straightforward to check (1). (2) holds because $\mathcal{Y} \cap \mathcal{Z} = \emptyset$.

For $s \in H_{k-1} \cup H_k$, let $\mathcal{Y}_s \subseteq \mathcal{Y}$ be defined by

$$\mathcal{Y}_s = \{\alpha = (\alpha_1, \alpha_2) \in \mathcal{Y} : \alpha_1 = s, \text{ or } \alpha_2 = s\}.$$

Note that for $s_1, s_2 \in H_{k-1} \cup H_k$ and $s_1 \neq s_2$, $\mathcal{Y}_{s_1} \cap \mathcal{Y}_{s_2} = \emptyset$. Moreover, $\mathcal{Y} = \cup_{s \in H_{k-1} \cup H_k} \mathcal{Y}_s$. Therefore

$$\mathcal{Y}_{ab} = \sum_{s \in H_{k-1} \cup H_k} \langle \mathbf{W}_{\mathcal{Y}_s}, \mathbf{K}(y^{(ab)}) - \mathbf{K}(y) \rangle$$

Note also that $\langle \mathbf{W}_{\mathcal{Y}_s}, \mathbf{K}(y^{(ab)}) - \mathbf{K}(y) \rangle = 0$, if $s \notin \{a, b\}$. Hence

$$\mathcal{Y}_{ab} = \sum_{\alpha \in \mathcal{Y}_a \cup \mathcal{Y}_b} [\mathbf{W}(\alpha)] \cdot \{[\mathbf{K}(y^{(ab)}) - \mathbf{K}(y)](\alpha)\}$$

From (6.26) we obtain that for $\alpha = (\alpha_1, \alpha_2) \in \mathcal{Y}_a$ and $\alpha_1 \neq \alpha_2$,

$$[\mathbf{K}(y^{(ab)}) - \mathbf{K}(y)](\alpha) = \begin{cases} 1 & \text{if } |\{\alpha_1, \alpha_2\} \cap y^{-1}(c_k)| = 2. \\ -1 & \text{if } |\{\alpha_1, \alpha_2\} \cap y^{-1}(c_{k-1})| = 1. \end{cases}$$

So, we can define

$$Y_a := \sum_{\alpha \in \mathcal{Y}_a} [\mathbf{W}(\alpha)] \cdot \{[\mathbf{K}(y^{(ab)}) - \mathbf{K}(y)](\alpha)\}$$

$$= \sum_{\alpha \in \mathcal{Y}_a; |\{\alpha_1, \alpha_2\} \cap y^{-1}(c_k)| = 2} [\mathbf{W}(\alpha)] - \sum_{\alpha \in \mathcal{Y}_a; |\{\alpha_1, \alpha_2\} \cap y^{-1}(c_{k-1})| = 1} [\mathbf{W}(\alpha)]$$

Similarly, define

$$Y_b := \sum_{\alpha \in \mathcal{Y}_b; |\{\alpha_1, \alpha_2\} \cap y^{-1}(c_{k-1})| = 2} [\mathbf{W}(\alpha)] - \sum_{\alpha \in \mathcal{Y}_b; |\{\alpha_1, \alpha_2\} \cap y^{-1}(c_k)| = 1} [\mathbf{W}(\alpha)]$$

Then $\mathcal{Y}_{ab} = Y_a + Y_b$ and $\{Y_s\}_{s \in H_{k-1} \cup H_k}$ is a collection of independent Gaussian random variables. Moreover, the variance of Y_s is $2(n_k + n_{k-1} - 2h)$ independent of the choice of s. $\qquad \square$

By the Lemma 6.9, we obtain

$$\langle \mathbf{W}, \mathbf{K}(y^{(ab)}) - \mathbf{K}(y) \rangle = Y_a + Y_b + \mathcal{Z}_{ab}$$

Moreover,

$$
\begin{aligned}
\max_{a \in H_k, b \in H_{k-1}} Y_a + Y_b + \mathcal{Z}_{ab} &\geq \max_{a \in H_k, b \in H_{k-1}} (Y_a + Y_b) - \max_{a \in H_k, b \in H_{k-1}} (-\mathcal{Z}_{ab}) \\
&= \max_{a \in H_k} Y_a + \max_{b \in H_{k-1}} Y_b - \max_{a \in H_k, b \in H_{k-1}} (-\mathcal{Z}_{ab})
\end{aligned}
$$

By Lemma 3.9 we obtain that when ϵ, h satisfy (3.22) with N replaced by h, each one of the following two events

$$F_1^* := \left\{ \max_{a \in H_k} Y_a \geq (1 - \epsilon) \sqrt{2 \log h \cdot 2(n_k + n_{k-1} - 2h)} \right\}$$

$$F_2^* := \left\{ \max_{b \in H_{k-1}} Y_b \geq (1 - \epsilon) \sqrt{2 \log h \cdot 2(n_k + n_{k-1} - 2h)} \right\}$$

has probability at least $1 - e^{-h^\epsilon}$. Moreover, the event

$$F_3^* := \left\{ \max_{a \in H_k, b \in H_{k-1}} \mathcal{Z}_{ab} \leq (1 + \epsilon) \sqrt{4 \log h \cdot \max \mathrm{Var}(Z_{ab})} \right\}$$

occurs with probability at least $1 - h^{-2\epsilon}$. Then by (6.26) we have

$$
\begin{aligned}
\mathrm{Var} \mathcal{Z}_{ab} &= \| \mathbf{K}(y^{(ab)}) - \mathbf{K}(y) \|_F^2 - \mathrm{Var}(Y_a) - \mathrm{Var}(Y_b) \\
&= 4(n_k + n_{k-1}) - 6 - 4(n_k + n_{k-1} - 2h) \\
&= 8h - 6
\end{aligned}
$$

Hence the probability of the event

$$F^* :=$$
$$\left\{ \max_{a \in H_k, b \in H_{k-1}} \langle \mathbf{W}, \mathbf{K}(y^{(ab)}) - \mathbf{K}(y) \rangle \geq 4(1 - \epsilon) \sqrt{\log h (n_k + n_{k-1} - 2h)} - 4(1 + \epsilon) \sqrt{\left(2h - \frac{3}{2} \right) \log h} \right\}$$

is at least

$$
\begin{aligned}
\Pr(F_1^* \cap F_2^* \cap F_3^*) &= 1 - \Pr([F_1^*]^c \cup [F_2^*]^c \cup [F_3^*]^c) \\
&\geq 1 - \Pr([F_1^*]^c) - \Pr([F_2^*]^c) - \Pr([F_3^*]^c) \\
&\geq 1 - 2e^{-h^\epsilon} - h^{-2\epsilon}.
\end{aligned}
$$

Moreover, from (6.25) we obtain

$$
\begin{aligned}
&4(1 - \epsilon) \sqrt{\log h (n_k + n_{k-1} - 2h)} - 4(1 + \epsilon) \sqrt{\left(2h - \frac{3}{2} \right) \log h} \\
&= 4 \sqrt{(n_k + n_{k-1} - 2h) \log h} \left[1 - \epsilon - (1 + \epsilon) \sqrt{\frac{2h - \frac{3}{2}}{n_k + n_{k-1} - 2h}} \right] \\
&\geq 4 \sqrt{(n_k + n_{k-1} - 2h) \log h} \left[1 - \epsilon - \frac{2}{\log n} \right]
\end{aligned}
$$

By (6.21) we have

$$4\sqrt{(n_k + n_{k-1} - 2h)\log h}\left[1 - \epsilon - \frac{2}{\log n}\right]$$

$$\geq 4\sqrt{\beta \log n(n_k + n_{k-1} - 2h)\left(1 - \frac{2\log\log n}{\beta\log n}\right)}\left[1 - \epsilon - \frac{2}{\log n}\right]$$

Let

(6.27) $$\epsilon = \frac{\log\log n}{\beta\log n};$$

then when n is sufficiently large, (3.22) holds with N replaced by h. Define an event

$$\tilde{F}^* := \left\{\max_{a \in H_k, b \in H_{k-1}}\langle\mathbf{W}, \mathbf{K}(y^{(ab)}) - \mathbf{K}(y)\rangle\right.$$

$$\left.\geq 4\sqrt{\beta\log n(n_k + n_{k-1} - 2h)\left(1 - \frac{2\log\log n}{\beta\log n}\right)}\left[1 - \frac{\log\log n}{\beta\log n} - \frac{2}{\log n}\right]\right\}$$

Then $F^* \subseteq \tilde{F}^*$

When (6.22) and (6.23) hold, we have

$$\Pr\left(\max_{a,b\in[n],y(a)\neq y(b)}\sigma\langle\mathbf{W}, \mathbf{G}(y^{(ab)}) - \mathbf{G}(y)\rangle > 2(n_k + n_{k-1} - 2)\right)$$

$$\geq \Pr(\tilde{F}^*) \geq \Pr(F) \geq 1 - \frac{1}{\log n},$$

as n is sufficiently large. Then the proposition follows.

$\square$

7. Complex Unitary Matrix with Gaussian Perturbation

Now we consider the community detection problem when the observation is a complex unitary matrix plus a multiple of a GUE or GOE matrix. In the former case, we prove a threshold with respect the intensity σ of the GUE perturbation for the exact recovery of the MLE. In the latter case, we develop a "complex version" of SDP algorithm for efficient recovery, and explicitly prove the region of the intensity of the GOE perturbation for the exact recovery of the SDP.

7.1. **GUE perturbation.** In this section, we prove Theorem 2.4. Recall that $y \in \Theta_A$ is the true color assignment function satisfying Assumption 2.3. For $x \in \Theta_A$, define

$$r(x) = \Re\langle\mathbf{U}, \mathbf{P}(x)\rangle.$$

Then

(7.1) $$r(x) - r(y) = \Re\left[\langle\mathbf{P}(y), \mathbf{P}(x) - \mathbf{P}(y)\rangle + \sigma\langle\mathbf{W}_c, \mathbf{P}(x) - \mathbf{P}(y)\rangle\right]$$

which is a real Gaussian random variable with mean $\Re[\langle\mathbf{P}(y), \mathbf{P}(x) - \mathbf{P}(y)\rangle]$, and variance

$$4\sigma^2\sum_{i<j}\{1 - \Re[\overline{x(i)}x(j)y(i)\overline{y(j)}]\}.$$

Moreover,

$$\Re\langle \mathbf{P}(x), \mathbf{P}(y)\rangle = n + 2\sum_{i<j} \Re[\overline{x(i)}x(j)y(i)\overline{y(j)}]$$

Hence

$$\Re[\langle \mathbf{P}(y), \mathbf{P}(x) - \mathbf{P}(y)\rangle] = -2\sum_{i<j}[1 - \Re[\overline{x(i)}x(j)y(i)\overline{y(j)}]]$$

Let

$$(7.2) \qquad J(x,y) = -\mathbb{E}[r(x) - r(y)] = 2\sum_{i<j}[1 - \Re[\overline{x(i)}x(j)y(i)\overline{y(j)}]]$$

Therefore for $x \in \Theta_A$

$$\Pr(r(x) - r(y) > 0) = \Pr_{\xi \in \mathcal{N}(0,1)}\left(\xi > \frac{\sqrt{J(x,y)}}{\sqrt{2}\sigma}\right) \le e^{-\frac{J(x,y)}{4\sigma^2}}$$

Lemma 7.1. *For any $x, y \in \Theta_A$, $J(x,y) \ge 0$; $J(x,y) = 0$ if and only if there exists a fixed angle α, such that $e^{\mathbf{i}\alpha}\mathbf{x} = \mathbf{y}$.*

Proof. First of all, $J(x,y) \ge 0$ follows from the fact that $\left|\overline{x(i)}x(j)y(i)\overline{y(j)}\right| = 1$. Moreover, $J(x,y) = 0$ if and only if for any $1 \le i < j \le n$, $\overline{x(i)}x(j)y(i)\overline{y(j)} = 1$. Then the lemma follows. $\square$

Then

Lemma 7.2. *If $x, z \in \Theta_A$ such that $x \in \theta(z)$, then*

    (1) $\mathbf{P}(x) = \mathbf{P}(z)$.
    (2) $r(x) = r(z)$.

Proof. Note that if $x \in \theta(z)$, then $\mathbf{P}(x) = \mathbf{P}(z)$. Then $r(x) = r(z)$ follows from (7.1). $\square$

Lemma 7.3. *If $x, y, y' \in \Theta_A$ such that $y \in \theta(y')$, then*

$$J(x,y) = J(x,y').$$

Proof. The lemma follows from (7.2) and Part (2) of Lemma 7.2. $\square$

By Lemmas 7.1 and 7.2, we obtain

$$p(y^A; \sigma) = \Pr\left[r(y^A) > \max_{x \in \Theta_A, x \notin \theta(y)} r(x)\right]$$

Hence

$$(7.3) \qquad 1 - p(y^A; \sigma) \le \sum_{\theta(x) \subseteq [\Theta_A \backslash \theta(y)]} e^{-\frac{J(x,y)}{4\sigma^2}}$$

For $i, j \in [k]$, let

$$(7.4) \qquad S_{i,j}(x,y) = \{1 \le l \le n : x(l) = e^{\mathbf{i}d_i}, y(l) = e^{\mathbf{i}d_j}\};$$

i.e., $S_{i,j}(x,y)$ consists of all the vertices which have color $e^{\mathbf{i}d_i}$ in x and color $e^{\mathbf{i}d_j}$ in y.

Let $t_{i,j}(x,y) = |S_{i,j}(x,y)|$. Again $t_{i,j}(x,y)$ satisfies (4.2) and (4.3). We shall now define a distance function on Θ.

Definition 7.4. *Let* $D_\Theta : \Theta \times \Theta \to \mathbb{N}$ *be the distance function on* Θ *defined as follows: for* $x, z \in \Theta$,

$$D_\Theta(x,z) = \sum_{i,j\in[k],i\neq j} t_{i,j}(x,z).$$

From (7.2), we obtain

$$
\begin{aligned}
J(x,y) &= \sum_{i,j,p,q\in[k]} t_{i,j}(x,y)t_{p,q}(x,y)[1 - \cos(-d_i + d_p + d_j - d_q)] \\
&= \sum_{i,j,p,q\in[k]} t_{i,j}(x,y)t_{p,q}(x,y)\left[1 - \cos\left(\frac{2\pi(p+j-q-i)}{k}\right)\right]
\end{aligned}
$$

When $p,q,i,j \in [k]$, $p+j-q-i \in [-(2k-2), 2k-2]$; hence $\cos\left(\frac{2\pi(p+j-q-i)}{k}\right) = 1$ if and only if $p+j-q-i \in \{-k,0,k\}$. Therefore,

(7.5) $J(x,y)$

$$
= \sum_{i,j,p,q\in[k],p+j-q-i\notin\{-k,0,k\}} t_{i,j}(x,y)t_{p,q}(x,y)\left[1 - \cos\left(\frac{2\pi(p+j-q-i)}{k}\right)\right]
$$

By Assumption 2.3,

$$n_1(x) = n_2(x) = \ldots = n_k(x) = n_1(y) = n_2(y) = \ldots = n_k(y) = \frac{n}{k}$$

Let $\hat{D}$ be the set defined by

$$\hat{D} = \left\{(t_{1,1},\ldots,t_{k,k}) \in \left\{0,1,\ldots\frac{n}{k}\right\}^{k^2} : \sum_{j\in[k]} t_{i,j} = \frac{n}{k}, \sum_{i\in[k]} t_{i,j} = \frac{n}{k}\right\}$$

For $\epsilon > 0$, let

(7.6) $$\hat{D}_\epsilon = \left\{(t_{1,1},\ldots,t_{k,k}) \in \hat{D} : \sum_{i,j,p,q\in[k],p+j-q-i\notin\{-k,0,k\}} t_{i,j}t_{p,q} \leq \epsilon n^2\right\}$$

Then we have

(7.7) $$\sum_{\theta(x)\subseteq[\Theta_A\backslash\theta(y)]} e^{-\frac{J(x,y)}{4\sigma^2}} \leq \hat{I}_1 + \hat{I}_2$$

where

$$\hat{I}_1 = \sum_{\theta(x)\subseteq\Theta_A\backslash\theta(y):\left(t_{1,1},\ldots,t_{k,k}\right)\in D\backslash\hat{D}_\epsilon} e^{\frac{-J(x,y)}{4\sigma^2}}.$$

and

(7.8) $$\hat{I}_2 = \sum_{\theta(x)\subseteq\Theta_A\backslash\theta(y):\left(t_{1,1},\ldots,t_{k,k}\right)\in\hat{D}_\epsilon} e^{\frac{-J(x,y)}{4\sigma^2}}.$$

Lemma 7.5. *Assume*

$$\lim_{n\to\infty} \frac{\epsilon n^2}{k^2} = \infty.$$

Then we have

$$\lim_{n\to\infty} \hat{I}_1 = 0.$$

Proof. From the definition of the domains $\hat{D}$ and $\hat{D}_\epsilon$, as well as the expression (7.5), we obtain that when $(u_{1,1,},\ldots,u_{k,k}) \in \hat{\mathcal{D}} \setminus \hat{\mathcal{D}}_\epsilon$

$$J(x,y) \geq \epsilon n^2 \left[1 - \cos\left(\frac{2\pi}{k}\right)\right].$$

Then the lemma follows.   □

From (7.5), we obtain

$$J(x,y) = \sum_{u=1}^{\lfloor \frac{k}{2}\rfloor} \sum_{\{i,j,p,q\in[k],\ ((p+j-q-i)\mod k)\in\{u,k-u\}\}} t_{i,j}(x,y)t_{p,q}(x,y)\left[1 - \cos\left(\frac{2u\pi}{k}\right)\right]$$

Lemma 7.6. *For each $x \in \Theta_A$ satisfying $(t_{1,1}(x,y),\ldots,t_{k,k}(x,y)) \in \hat{D}_\epsilon$, there exists $y' \in \theta(y)$ such that*

$$(7.9) \qquad \frac{1}{n}\sum_{i,j\in[k],i\neq j} t_{i,j}(x,y') < (k-1)\sqrt{2\epsilon}$$

In particular, this implies that

$$(7.10) \qquad \frac{1}{n}\sum_{i\in[k]} t_{i,i}(x,y') > 1 - (k-1)\sqrt{2\epsilon};$$

and for each $i \in [k]$,

$$(7.11) \qquad \frac{1}{n}t_{i,i}(x,y') > \frac{1}{k} - (k-1)\sqrt{2\epsilon}.$$

Proof. From the definition (7.6) and the fact that $\sum_{i,j\in[k]} t_{i,j}(x,y) = n$, we obtain

$$\sum_{i,j,p,q\in[k],p+j-q-i\in\{-k,0,k\}} t_{i,j}(x,y)t_{p,q}(x,y) \geq (1-\epsilon)n^2;$$

which is the same as the following inequality

$$(7.12) \qquad \sum_{l=0}^{k-1}\left[\sum_{[i,j\in[k],(i-j)\mod k=l]} t_{i,j}(x,y)\right]^2 \geq (1-\epsilon)n^2$$

Let $a = \sqrt{2\epsilon}$. Again using the fact that $\sum_{i,j\in[k]} t_{i,j}(x,y) = n$; if there exists $l_1, l_2 \in \{0,\ldots,k-1\}$ and $l_1 \neq l_2$ and

$$\min\left\{\sum_{[i,j\in[k],(i-j)\mod k=l_1]} t_{i,j}(x,y), \sum_{[i,j\in[k],(i-j)\mod k=l_2]} t_{i,j}(x,y)\right\} \geq an;$$

then

$$\sum_{l=0}^{k-1}\left[\sum_{[i,j\in[k],(i-j)\mod k=l]}t_{i,j}(x,y)\right]^2$$

$$\leq \left[\sum_{l=0}^{k-1}\sum_{[i,j\in[k],(i-j)\mod k=l]}t_{i,j}(x,y)\right]^2 - \left[\sum_{[i,j\in[k],(i-j)\mod k=l_1]}t_{i,j}(x,y)\right]\cdot\left[\sum_{[i,j\in[k],(i-j)\mod k=l_2]}t_{i,j}(x$$

$$\leq n^2 - a^2 n^2 = (1-2\epsilon)n^2$$

which is a contradiction to (7.12). Hence there exists at most one $l_0 \in \{0, 1, \ldots, k-1\}$, such that

$$\sum_{[i,j\in[k],(i-j)\mod k=l_0]}t_{i,j}(x,y)\geq an.$$

Then for and $l \in \{0, 1, \ldots, k-1\} \setminus \{l_0\}$

$$\sum_{[i,j\in[k],(i-j)\mod k=l]}t_{i,j}(x,y) < an.$$

Since $\sum_{i,j\in[k]}t_{i,j}(x,y) = n$, we have

$$\sum_{[i,j\in[k],(i-j)\mod k=l_0]}t_{i,j}(x,y) = \sum_{i,j\in[k]}t_{i,j}(x,y) - \sum_{[l\in(\{0,1,\ldots,k-1\}\setminus\{l_0\})]}\sum_{[i,j\in[k],(i-j)\mod k=l]}t_{i,j}(x,y)$$

$$\geq [1-(k-1)a]n$$

Let $y' = e^{\frac{2l_0\pi\mathbf{i}}{k}}y$, then (7.9) and (7.10) follows. To obtain (7.11), note that

$$\frac{1}{n}t_{i,i}(x,y') = \frac{1}{n}\sum_{j\in[k]}t_{i,j}(x,y') - \frac{1}{n}\sum_{j\in[k],j\neq i}t_{i,j}(x,y') \geq \frac{n}{k} - (k-1)\sqrt{2\epsilon}.$$

Then the lemma follows. $\square$

Proposition 7.7. *Let $y \in \Theta_A$ be the true color assignment function satisfying Assumption 2.3. Let k be the total number of colors and n be the total number of vertices. Assume*

(7.13) $$\lim_{n\to\infty}\frac{\log k}{\log n} = 0.$$

If there exists $\delta > 0$, such that

(7.14) $$\sigma^2 < \frac{(1-\delta)\left[n(1-\cos\frac{2\pi}{k})\right]}{2\log n}$$

then

$$\lim_{n\to\infty}\Pr(y^A \in \theta(y)) = 1$$

Proof. For given $x \in \Theta_A$, such that $(t_{1,1}(x,y), \ldots, t_{k,k}(x,y)) \in \hat{D}_\epsilon$, and $x \notin \theta(y)$, by Lemma 7.6, let $y' \in \theta(y)$ such that (7.9) holds. By Lemma 3.3, there exists an l-cycle for (x, y') with $2 \leq l \leq k$.

Then for each $2 \le s \le (l+1)$, choose an arbitrary vertex u_s in $S_{i_{s-1},i_s}(x,y)$, and let $y_1(u_s) = e^{\mathbf{i}d_{i_{s-1}}}$, where $i_{l+1} := i_1$. For any vertex $z \in [n] \setminus \{u_2,\ldots,u_{l+1}\}$, let $y_1(z) = y'(z)$.

Note that $y_1 \in \Theta_A$. More precisely, for $1 \le s \le l$, we have

(7.15) $$t_{i_s,i_s}(x,y') + 1 = t_{i_s,i_s}(x,y_1);$$

(7.16) $$t_{i_s,i_{s+1}}(x,y') - 1 = t_{i_s,i_{s+1}}(x,y_1)$$

and

$$t_{a,b}(x,y') = t_{a,b}(x,y_1), \ \forall (a,b) \notin \{(i_s,i_s),(i_s,i_{s+1})\}_{s=1}^{l}.$$

Let

$$\Delta := \{(i_s,i_s),(i_s,i_{s+1})\}_{s=1}^{l}.$$

From (7.5) and Lemma 7.3 we obtain

$$J(x,y_1) - J(x,y) = J(x,y_1) - J(x,y')$$

$$= \sum_{\substack{i,j,p,q\in[k],p+j-q-i\notin\{-k,0,k\}}} \left[t_{i,j}(x,y_1)t_{p,q}(x,y_1) - t_{i,j}(x,y')t_{p,q}(x,y')\right]\left(1 - \cos\left(\frac{2\pi(p+j-q-i)}{k}\right)\right)$$

$$= B_1 + B_2 + B_3 + B_4 + B_5 + B_6 + B_7 + B_8.$$

Here

$$B_1 = \sum_{s=1}^{l} \sum_{p,q\in[k],(p,q)\notin\Delta} \left[t_{i_s,i_s}(x,y_1)t_{p,q}(x,y_1) - t_{i_s,i_s}(x,y')t_{p,q}(x,y')\right]\left(1 - \cos\left(\frac{2\pi(p-q)}{k}\right)\right)$$

$$B_2 = \sum_{s=1}^{l} \sum_{p,q\in[k],(p,q)\notin\Delta} \left[t_{i_s,i_{s+1}}(x,y_1)t_{p,q}(x,y_1) - t_{i_s,i_{s+1}}(x,y')t_{p,q}(x,y')\right]\left(1 - \cos\left(\frac{2\pi(p+i_{s+1}-q-i_s)}{k}\right)\right)$$

$$B_3 = \sum_{i,j\in[k],i,j\notin\Delta} \sum_{s=1}^{l} \left[t_{i,j}(x,y_1)t_{i_s,i_s}(x,y_1) - t_{i,j}(x,y')t_{i_s,i_s}(x,y')\right]\left(1 - \cos\left(\frac{2\pi(j-i)}{k}\right)\right) = B_1$$

$$B_4 = \sum_{i,j\in[k],i,j\notin\Delta} \sum_{s=1}^{l} \left[t_{i,j}(x,y_1)t_{i_s,i_{s+1}}(x,y_1) - t_{i,j}(x,y')t_{i_s,i_{s+1}}(x,y')\right]\left(1 - \cos\left(\frac{2\pi(i_s+j-i_{s+1}-i)}{k}\right)\right) = B_2$$

$$B_5 = \sum_{s=1}^{l}\sum_{r=1}^{l} \left[t_{i_s,i_s}(x,y_1)t_{i_r,i_r}(x,y_1) - t_{i_s,i_s}(x,y')t_{i_r,i_r}(x,y')\right](1 - \cos 0) = 0$$

$$B_6 = \sum_{s=1}^{l}\sum_{r=1}^{l} \left[t_{i_s,i_s}(x,y_1)t_{i_r,i_{r+1}}(x,y_1) - t_{i_s,i_s}(x,y')t_{i_r,i_{r+1}}(x,y')\right]\left(1 - \cos\left(\frac{2\pi(i_r-i_{r+1})}{k}\right)\right)$$

$$B_7 = \sum_{s=1}^{l}\sum_{r=1}^{l} \left[t_{i_s,i_{s+1}}(x,y_1)t_{i_r,i_r}(x,y_1) - t_{i_s,i_{s+1}}(x,y')t_{i_r,i_r}(x,y')\right]\left(1 - \cos\left(\frac{2\pi(i_{s+1}-i_s)}{k}\right)\right) = B_6$$

$$B_8 = \sum_{s=1}^{l}\sum_{r=1}^{l} \left[t_{i_s,i_{s+1}}(x,y_1)t_{i_r,i_{r+1}}(x,y_1) - t_{i_s,i_{s+1}}(x,y')t_{i_r,i_{r+1}}(x,y')\right]\left(1 - \cos\left(\frac{2\pi(i_r-i_{r+1}+i_{s+1}-i_s)}{k}\right)\right)$$

By (7.15) and (7.16) we have

$$B_1 = l \sum_{p,q\in[k],(p,q)\notin\Delta} t_{p,q}(x,y')\left(1-\cos\left(\frac{2\pi(p-q)}{k}\right)\right)$$

When $(t_{1,1}(x,y),\ldots,t_{k,k}(x,y))\in\hat{D}_\epsilon$, by Lemma 7.6 we obtain

$$0\leq B_1 \leq 2l(k-1)\sqrt{2\epsilon}n.$$

Similarly,

$$B_2 = -\sum_{s=1}^{l}\sum_{p,q\in[k],(p,q)\notin\Delta} t_{p,q}(x,y')\left(1-\cos\left(\frac{2\pi(p+i_{s+1}-q-i_s)}{k}\right)\right)$$

Then

$$B_2 \leq -\sum_{s=1}^{l}\sum_{p\in([k]\setminus\{i_1,\ldots,i_l\})} t_{p,p}(x,y')\left(1-\cos\left(\frac{2\pi(i_{s+1}-i_s)}{k}\right)\right)$$

$$\leq -\frac{(k-l)n}{k}\sum_{s=1}^{l}\left(1-\cos\left(\frac{2\pi(i_{s+1}-i_s)}{k}\right)\right)+2l(k-1)\sqrt{2\epsilon}n$$

$$B_6 = \sum_{s=1}^{l}\sum_{r=1}^{l}\left[-t_{i_s,i_s}(x,y')+t_{i_r,i_{r+1}}(x,y')-1\right]\left(1-\cos\left(\frac{2\pi(i_r-i_{r+1})}{k}\right)\right)$$

$$= -l\sum_{r=1}^{l}\left(1-\cos\left(\frac{2\pi(i_r-i_{r+1})}{k}\right)\right)+l\sum_{r=1}^{l}t_{i_r,i_{r+1}}(x,y')\left(1-\cos\left(\frac{2\pi(i_r-i_{r+1})}{k}\right)\right)$$

$$-\sum_{s=1}^{l}t_{i_s,i_s}(x,y')\sum_{r=1}^{l}\left(1-\cos\left(\frac{2\pi(i_r-i_{r+1})}{k}\right)\right)$$

By Lemma 7.6 we obtain

$$B_6 \leq -l\sum_{r=1}^{l}\left(1-\cos\left(\frac{2\pi(i_r-i_{r+1})}{k}\right)\right)+2l(k-1)\sqrt{2\epsilon}n$$

$$-\frac{ln}{k}\sum_{r=1}^{l}\left(1-\cos\left(\frac{2\pi(i_r-i_{r+1})}{k}\right)\right)+2l(k-1)\sqrt{2\epsilon}n$$

$$B_8 = \sum_{s=1}^{l}\sum_{r=1}^{l}\left[-t_{i_s,i_{s+1}}(x,y')-t_{i_r,i_{r+1}}(x,y')-1\right]\left(1-\cos\left(\frac{2\pi(i_r-i_{r+1}+i_{s+1}-i_s)}{k}\right)\right)$$

By Lemma 7.6 we obtain

$$0\geq B_8\geq -4l(k-1)n\sqrt{2\epsilon}-2l^2$$

Then we have

$$J(x,y_1)-J(x,y)\leq\left[16l(k-1)\sqrt{2\epsilon}-2l\left(1-\cos\frac{2\pi}{k}\right)\right]n.$$

Therefore

$$(7.17) \qquad e^{-\frac{J(x,y)}{4\sigma^2}} \le e^{-\frac{J(x,y_1)}{4\sigma^2}} e^{-\frac{n}{4\sigma^2}\left[2l\left(1-\cos\frac{2\pi}{k}\right)-16l(k-1)\sqrt{2\epsilon}\right]}.$$

If $y_1 \ne x$, we find an l_2-cycle $(2 \le l_2 \le k)$ for (x, y_1), change colors along the l_2-cycle as above, and obtain another cycle assignment mapping $y_2 \in \Theta_A$, and so on. Let $y_0 := y$, and note that for each $r \ge 1$, if y_r is obtained from y_{r-1} by changing colors along an l_r cycle, we have

$$D_\Theta(x, y_r) = D_\Theta(x, y_{r-1}) - l_r$$

Therefore finally we can obtain x from y by changing colors along at most $\lfloor\frac{n}{2}\rfloor$ cycles. Using similar arguments as those used to derive (7.17), we obtain that for each r

$$e^{-\frac{J(x,y_{r-1})}{4\sigma^2}} \le e^{-\frac{J(x,y_r)}{4\sigma^2}} e^{-\frac{n}{4\sigma^2}\left[2l_r\left(1-\cos\frac{2\pi}{k}\right)-16l_r(k-1)\sqrt{2\epsilon}\right]}.$$

Therefore if $y_s = x$ for some $1 \le s \le \lfloor\frac{n}{2}\rfloor$, we have

$$e^{-\frac{J(x,y)}{4\sigma^2}} \le e^{-\frac{J(x,y_s)}{4\sigma^2}} e^{-\frac{n}{4\sigma^2}\left[2\left(1-\cos\frac{2\pi}{k}\right)-16(k-1)\sqrt{2\epsilon}\right]\left[\sum_{i=1}^s l_i\right]}.$$

By (7.2), we have $J(x, y_s) = J(x, x) = 0$, hence

$$e^{-\frac{J(x,y)}{4\sigma^2}} \le \prod_{i\in[s]} e^{-\frac{nl_i}{4\sigma^2}\left[2\left(1-\cos\frac{2\pi}{k}\right)-16(k-1)\sqrt{2\epsilon}\right]}.$$

Let

$$\gamma_\epsilon := 2\left(1 - \cos\frac{2\pi}{k}\right) - 16(k-1)\sqrt{2\epsilon}.$$

Note also that for any $r_1 \ne r_2$, in the process of obtaining y_{r_1} from y_{r_1-1} and the process of obtaining y_{r_2} from y_{r_2-1}, we change colors on disjoint sets of vertices. Hence the order of these steps of changing colors along cycles does not affect the final color assignment mapping we obtain. From (7.8) we obtain

$$(7.18) \qquad \hat{I}_2 \le \prod_{l=2}^k \left(\sum_{m_l=0}^\infty (nk)^{m_l l} e^{-\frac{nm_l l\gamma_\epsilon}{4\sigma^2}}\right) - 1$$

On the right hand side of (7.18), when expanding the product, each summand has the form

$$\left[(nk)^{2m_2} e^{-\frac{2m_2 n\gamma_\epsilon}{4\sigma^2}}\right] \cdot \left[(nk)^{3m_3} e^{-\frac{3m_3 n\gamma_\epsilon}{4\sigma^2}}\right] \cdot \ldots \cdot \left[(nk)^{km_k} e^{-\frac{km_k n\gamma_\epsilon}{4\sigma^2}}\right]$$

where the factor $\left[(nk)^{2m_2} e^{-\frac{2m_2 n\gamma_\epsilon}{4\sigma^2}}\right]$ represents that we changed along 2-cycles m_2 times, the factor $\left[(nk)^{3m_3} e^{-\frac{3m_3 n\gamma_\epsilon}{4\sigma^2}}\right]$ represents that we changed along 3-cycles m_3 times, and so on. Moreover, each time we changed along an l-cycle, we need to first determine the l different colors involved in the l-cycle, and there are at most k^l different l-cycles; we then need to chose l vertices to change colors, and there are at most n^l choices. It is straightforward to check that if σ satisfies (7.14), then

$$nke^{-\frac{n\gamma_\epsilon}{4\sigma^2}} \le e^{\log k - \frac{\delta\log n}{1-\delta} + \frac{8(k-1)\sqrt{2\epsilon}\log n}{(1-\delta)\left(1-\cos\frac{2\pi}{k}\right)}}.$$

as $n \to \infty$. Therefore we have

$$\sum_{m_l=0}^{\infty} (nk)^{m_l l} e^{-\frac{n m_l l \gamma_\epsilon}{4\sigma^2}} \leq \frac{1}{1 - e^{l\left(\log k - \frac{\delta \log n}{1-\delta} + \frac{8(k-1)\sqrt{2\epsilon}\log n}{(1-\delta)\left(1-\cos\frac{2\pi}{k}\right)}\right)}}$$

Let

$$\hat{U} := \prod_{l=2}^{k} \left(\sum_{m_l=0}^{\infty} (nk)^{m_l l} e^{-\frac{n m_l l \gamma_\epsilon}{4\sigma^2}} \right).$$

When (7.13) holds, let

$$\epsilon := \frac{1}{\sqrt{n}}$$

Then when σ satisfies (7.14)

$$\frac{\gamma_\epsilon}{4\sigma^2} \leq \frac{\delta \log n}{2(1-\delta)}$$

when n is sufficiently large.

Since $\log(1 + x) \leq x$ for $x \geq 0$, we have

$$
\begin{aligned}
0 \leq \log \hat{U} &= \sum_{l=2}^{k} \log \left(1 + \sum_{m_l=1}^{\infty} (nk)^{m_l l} e^{-\frac{n m_l l \gamma_\epsilon}{4\sigma^2}} \right) \\
&\leq \sum_{l=2}^{k} \sum_{m_l=1}^{\infty} (nk)^{m_l l} e^{-\frac{n m_l l \gamma_\epsilon}{4\sigma^2}} \\
&\leq \sum_{l=2}^{k} \frac{\left(e^{-\frac{\delta \log n}{2(1-\delta)}} \right)^l}{1 - \left(e^{-\frac{\delta \log n}{2(1-\delta)}} \right)^l} \\
&\leq \frac{e^{-\frac{\delta \log n}{1-\delta}}}{\left(1 - e^{-\frac{\delta \log n}{1-\delta}} \right) \left(1 - e^{-\frac{\delta \log n}{2(1-\delta)}} \right)} \to 0
\end{aligned}
$$

as $n \to \infty$. Hence we have $\lim_{n\to\infty} \hat{I}_2 = 0$. By Lemma 7.5, $\lim_{n\to\infty} \hat{I}_1 = 0$. Then the proposition follows.

$\square$

Proof of Theorem 2.4(b). For $y \in \Theta_A$, $a, b \in [n]$ such that $1 = y(a) \neq y(b) = e^{\frac{2\pi i}{k}}$. Let $y^{(ab)}$ be the coloring of vertices defined by

$$(7.19) \qquad y^{(ab)}(i) = \begin{cases} y(i) & \text{if } i \in \{1, 2, \ldots, n\} \setminus \{a, b\} \\ e^{\frac{2\pi i}{k}} & \text{if } i = a \\ 1 & \text{if } i = b \end{cases}$$

Then

$$1 - p(y^A; \sigma) \geq \Pr\left(\cup_{a,b\in[n], 1=y(a)\neq y(b)=e^{\frac{2\pi i}{k}}}[r(y^{(ab)}) - r(y) > 0]\right),$$

since any of the event $[r(y^{(ab)}) - r(y) > 0]$ implies $y^A \neq y$. Recall that

$$
\begin{aligned}
r(y^{(ab)}) - r(y) &= \langle \mathbf{P}(y), \mathbf{P}(y^{(ab)}) - \mathbf{P}(y)\rangle + \sigma\langle \mathbf{W}_c, \mathbf{P}(y^{(ab)}) - \mathbf{P}(y)\rangle \\
&= -4(n-2)\left(1 - \cos\frac{2\pi}{k}\right) - 2\left(1 - \cos\frac{4\pi}{k}\right) + \sigma\langle \mathbf{W}_c, \mathbf{P}(y^{(ab)}) - \mathbf{P}(y)\rangle.
\end{aligned}
$$

Let

$$E(n) := 4(n-2)\left(1 - \cos\frac{2\pi}{k}\right) + 2\left(1 - \cos\frac{4\pi}{k}\right)$$

For $l \in \{0, 1\}$, let $H_l \subset y^{-1}\left(e^{\frac{2l\pi i}{k}}\right)$ such that $|H_l| = \frac{n}{\log^2 n} = h$. Then

$$1 - p(y^A; \sigma) \geq \Pr\left(\max_{a\in H_0, b\in H_1}\sigma\langle \mathbf{W}, \mathbf{P}(y^{(ab)}) - \mathbf{P}(y)\rangle > E(n)\right)$$

Let $(\mathcal{X}, \mathcal{Y}, \mathcal{Z})$ be a partition of $[n]^2$ defined by

$$
\begin{aligned}
\mathcal{X} &= \{\alpha = (\alpha_1, \alpha_2) \in [n]^2, \{\alpha_1, \alpha_2\} \cap [H_0 \cup H_1] = \emptyset\} \\
\mathcal{Y} &= \{\alpha = (\alpha_1, \alpha_2) \in [n]^2, |\{\alpha_1, \alpha_2\} \cap [H_0 \cup H_1]| = 1\} \\
\mathcal{Z} &= \{\alpha = (\alpha_1, \alpha_2) \in [n]^2, |\{\alpha_1, \alpha_2\} \cap [H_0 \cup H_1]| = 2\}
\end{aligned}
$$

For $\eta \in \{\mathcal{X}, \mathcal{Y}, \mathcal{Z}\}$, define the $n \times n$ matrix $\mathbf{W}_\eta$ from the entries of $\mathbf{W}$ as follows

$$\mathbf{W}_{\eta,c}(i,j) = \begin{cases} 0 & \text{if } (i,j) \notin \eta \\ \mathbf{W}_c(i,j), & \text{if } (i,j) \in \eta \end{cases}$$

For each $a \in H_0$ and $b \in H_1$, let

$$
\begin{aligned}
\mathcal{X}_{ab} &= \langle \mathbf{W}_{\mathcal{X},c}, \mathbf{P}(y^{(ab)}) - \mathbf{P}(y)\rangle \\
\mathcal{Y}_{ab} &= \langle \mathbf{W}_{\mathcal{Y},c}, \mathbf{P}(y^{(ab)}) - \mathbf{P}(y)\rangle \\
\mathcal{Z}_{ab} &= \langle \mathbf{W}_{\mathcal{Z},c}, \mathbf{P}(y^{(ab)}) - \mathbf{P}(y)\rangle
\end{aligned}
$$

Claim 7.8. *The followings are true:*

(1) $\mathcal{X}_{ab} = 0$ for $a \in H_0$ and $b \in H_1$.

(2) For each $a \in H_0$ and $b \in H_1$, the variables $\mathcal{Y}_{ab}$ and $\mathcal{Z}_{ab}$ are independent.

(3) Each $\mathcal{Y}_{ab}$ can be decomposed into $Y_a + Y_b$ where $\{Y_a\}_{a\in H_0} \cup \{Y_b\}_{b\in H_1}$ is a collection of i.i.d. Gaussian random variables.

Proof. It is straightforward to check (1). (2) holds because $\mathcal{Y} \cap \mathcal{Z} = \emptyset$.

For $s \in H_0 \cup H_1$, let $\mathcal{Y}_s \subseteq \mathcal{Y}$ be defined by

$$\mathcal{Y}_s = \{\alpha = (\alpha_1, \alpha_2) \in \mathcal{Y} : \alpha_1 = s, \text{ or } \alpha_2 = s\}.$$

Note that for $s_1, s_2 \in H_0 \cup H_1$ and $s_1 \neq s_2$, $\mathcal{Y}_{s_1} \cap \mathcal{Y}_{s_2} = \emptyset$. Moreover, $\mathcal{Y} = \cup_{s\in H_0\cup H_1}\mathcal{Y}_s$. Therefore

$$\mathcal{Y}_{ab} = \sum_{s\in H_0\cup H_1} \langle \mathbf{W}_{\mathcal{Y}_s,c}, \mathbf{P}(y^{(ab)}) - \mathbf{P}(y)\rangle$$

Note also that $\langle \mathbf{W}_{\mathcal{Y}_{s,c}}, \mathbf{P}(y^{(ab)}) - \mathbf{P}(y) \rangle = 0$, if $s \notin \{a, b\}$. Hence

$$\mathcal{Y}_{ab} = \sum_{\alpha \in \mathcal{Y}_a \cup \mathcal{Y}_b} [\mathbf{W}_c(\alpha)] \cdot \{[\mathbf{P}(y^{(ab)}) - \mathbf{P}(y)](\alpha)\}$$

Note that for $\alpha \in \mathcal{Y}_a$,

$$[\mathbf{P}(y^{(ab)}) - \mathbf{P}(y)](\alpha) = \begin{cases} (e^{\frac{2\pi \mathbf{i}}{k}} - 1)\overline{y(\alpha_2)} & \text{if } \alpha_1 = a \\ y(\alpha_1)(e^{-\frac{2\pi \mathbf{i}}{k}} - 1) & \text{if } \alpha_2 = a. \end{cases}$$

So,

$$Y_a := \sum_{\alpha \in \mathcal{Y}_a} [\mathbf{W}_c(\alpha)] \cdot \{[\mathbf{P}(y^{(ab)}) - \mathbf{P}(y)](\alpha)\}$$

$$= \left\{ \sum_{\alpha \in \mathcal{Y}_a; \alpha_1 = a} [\overline{y(\alpha_2)}\mathbf{W}_c(\alpha)] + \sum_{\alpha \in \mathcal{Y}_a; \alpha_2 = a} [y(\alpha_1)\mathbf{W}_c(\alpha)] \right\} \left(\cos \frac{2\pi}{k} - 1 \right)$$

$$+ \mathbf{i} \left\{ \sum_{\alpha \in \mathcal{Y}_a; \alpha_1 = a} [\overline{y(\alpha_2)}\mathbf{W}_c(\alpha)] - \sum_{\alpha \in \mathcal{Y}_a; \alpha_2 = a} [y(\alpha_1)\mathbf{W}_c(\alpha)] \right\} \sin \frac{2\pi}{k}$$

Similarly, define

$$Y_b : = \left\{ \sum_{\alpha \in \mathcal{Y}_b; \alpha_2 = b} [\overline{y(\alpha_2)}\mathbf{W}_c(\alpha)] - \sum_{\alpha \in \mathcal{Y}_b; \alpha_1 = b} [y(\alpha_1)\mathbf{W}_c(\alpha)] \right\} \left(1 - \cos \frac{2\pi}{k} \right)$$

$$- \mathbf{i} \left\{ \sum_{\alpha \in \mathcal{Y}_a; \alpha_1 = b} [\overline{y(\alpha_2)}\mathbf{W}_c(\alpha)] - \sum_{\alpha \in \mathcal{Y}_a; \alpha_2 = b} [y(\alpha_1)\mathbf{W}_c(\alpha)] \right\} \sin \frac{2\pi}{k}$$

Then $\mathcal{Y}_{ab} = Y_a + Y_b$ and $\{Y_s\}_{s \in H_0 \cup H_1}$ is a collection of independent Gaussian random variables. Moreover, the variance of Y_s is equal to $8(n - 2h)\left(1 - \cos \frac{2\pi}{k}\right)$ independent of the choice of s. □

By the claim, we obtain

$$\langle \mathbf{W}_c, \mathbf{P}(y^{(ab)}) - \mathbf{P}(y) \rangle = Y_a + Y_b + \mathcal{Z}_{ab}$$

Moreover,

$$\max_{a \in H_0, b \in H_1} Y_a + Y_b + \mathcal{Z}_{ab} \geq \max_{a \in H_0, b \in H_1} (Y_a + Y_b) - \max_{a \in H_0, b \in H_1} (-\mathcal{Z}_{ab})$$

$$= \max_{a \in H_0} Y_a + \max_{b \in H_1} Y_b - \max_{a \in H_0, b \in H_1} (-\mathcal{Z}_{ab})$$

By Lemma 3.9 we obtain

$$\max_{a \in H_0} Y_a \geq (1 - 0.01\epsilon)\sqrt{2 \log h \cdot 8(n - 2h)\left(1 - \cos \frac{2\pi}{k}\right)}$$

$$\max_{b \in H_1} Y_b \geq (1 - 0.01\epsilon)\sqrt{2 \log h \cdot 8(n - 2h)\left(1 - \cos \frac{2\pi}{k}\right)}$$

$$\max_{a \in H_0, b \in H_1} \mathcal{Z}_{ab} \leq (1 + \epsilon)\sqrt{4 \log h \cdot \max \text{Var}(Z_{ab})}$$

with probability $1 - o_n(1)$.(Here $o_n(1) \to 0$ as $n \to \infty$.) Moreover,

$$\mathrm{Var}\,\mathcal{Z}_{ab} \leq 32h^2$$

which is $o(n)$. Hence

$$\max_{a \in H_1, b \in H_2} \langle \mathbf{W}_c, \mathbf{P}(y^{(ab)}) - \mathbf{P}(y) \rangle \geq 2(1 - 0.01\epsilon - o(1))\sqrt{2 \log n \cdot 8(n - 2h)\left(1 - \cos\frac{2\pi}{k}\right)}$$

$$\geq 8(1 - 0.01\epsilon - o(1))\sqrt{\left(1 - \cos\frac{2\pi}{k}\right)n \log n}$$

with probability $1 - o_n(1)$. Since $\sigma^2 > \frac{(1+\delta)\left[n(1 - \cos\frac{2\pi}{k})\right]}{2 \log n}$, we have

$$\Pr\left(\max_{a \in H_0, b \in H_1} \sigma \langle \mathbf{W}, \mathbf{P}(y^{(ab)}) - \mathbf{P}(y)\rangle > E(n)\right) \geq 1 - o_n(1)$$

Then the lemma follows. $\square$

7.2. **Algorithm: Complex Semi-Definite Programming and GOE perturbation.**
In this section, we prove Theorem 2.5.

Assume

$$\mathbf{V} = V_1 + \mathbf{i}V_2; \qquad \mathbf{X} = X_1 + \mathbf{i}X_2$$

where U_1 and X_1 are $n \times n$ real symmetric matrices, and U_2 and X_2 are $n \times n$ real anti-symmetric matrices.

Let

$$\tilde{X} = \begin{pmatrix} X_1 & -X_2 \\ X_2 & X_1 \end{pmatrix} \qquad \tilde{V} = \begin{pmatrix} V_1 & V_2 \\ -V_2 & V_1 \end{pmatrix},$$

then the complex optimization problem (2.17) is equivalent to the following real optimization problem

(7.20) $$\max \langle \tilde{V}, \tilde{X} \rangle$$

     subject to $\tilde{X}_{ii} = 1$, for $1 \leq i \leq 2n$

     $\tilde{X}_{p,q} = \tilde{X}_{p+n,q+n}$, for $1 \leq p \leq n$, $1 \leq q \leq n$

     $\tilde{X}_{p,q+n} = -\tilde{X}_{p+n,q}$, for $1 \leq p \leq n$, $1 \leq q \leq n$.

     and $\tilde{X} \succeq 0$

The dual program of (7.20) is

(7.21) $$\min \mathrm{tr}(Z)$$

     subject to $$Z - \tilde{V} + \begin{pmatrix} A & 0 \\ 0 & -A \end{pmatrix} + \begin{pmatrix} 0 & B \\ B & 0 \end{pmatrix} \succeq 0$$

     Z is diagonal

By complementary slackness

$$X = X_1 + iX_2 = \mathbf{y}\bar{\mathbf{y}}^t = (\mathbf{y}_1 + i\mathbf{y}_2)(\mathbf{y}_1 - i\mathbf{y}_2)^t$$

is the unique optimum solution of (2.17) if and only if there exists a dual feasible solution (Z, A, B), such that

$$\left\langle Z - \tilde{V} + \begin{pmatrix} A & 0 \\ 0 & -A \end{pmatrix} + \begin{pmatrix} 0 & B \\ B & 0 \end{pmatrix}, \begin{pmatrix} \mathbf{y}_1\mathbf{y}_1^t + \mathbf{y}_2\mathbf{y}_2^t & \mathbf{y}_1\mathbf{y}_2^t - \mathbf{y}_2\mathbf{y}_1^t \\ \mathbf{y}_2\mathbf{y}_1^t - \mathbf{y}_1\mathbf{y}_2^t & \mathbf{y}_1\mathbf{y}_1^t + \mathbf{y}_2\mathbf{y}_2^t \end{pmatrix} \right\rangle = 0;$$

which is equivalent to

(7.22)
$$\left\langle Z - \tilde{V}, \begin{pmatrix} \mathbf{y}_1\mathbf{y}_1^t + \mathbf{y}_2\mathbf{y}_2^t & \mathbf{y}_1\mathbf{y}_2^t - \mathbf{y}_2\mathbf{y}_1^t \\ \mathbf{y}_2\mathbf{y}_1^t - \mathbf{y}_1\mathbf{y}_2^t & \mathbf{y}_1\mathbf{y}_1^t + \mathbf{y}_2\mathbf{y}_2^t \end{pmatrix} \right\rangle = 0.$$

Assume

$$Z = \begin{pmatrix} Z_1 & 0 \\ 0 & Z_2 \end{pmatrix},$$

where Z_1, Z_2 are $n \times n$ real diagonal matrices. Then (7.22) is equivalent to

(7.23)
$$\Re\left[\mathbf{y}^t(Z_1 + Z_2 - 2\mathbf{V})\overline{\mathbf{y}}\right] = 0$$

Note that

$$\Im\left[\mathbf{y}^t(Z_1 + Z_2 - 2\mathbf{V})\overline{\mathbf{y}}\right] = \langle Z_1 + Z_2 - 2V_1, \mathbf{y}_2\mathbf{y}_1^t - \mathbf{y}_1\mathbf{y}_2^t \rangle - \langle 2V_2, \mathbf{y}_1\mathbf{y}_1^t + \mathbf{y}_2\mathbf{y}_2^t \rangle$$

which is identically zero since each term on the right hand side of (7.24) is the inner product of a symmetric matrix and an anti-symmetric matrix. Moreover, if the minimizer of (7.21) is unique, then as a Hermitian matrix, the second smallest eigenvalue of

$$S := Z_1 + Z_2 - 2\mathbf{V}$$

is strictly positive.

From (7.23) and the fact that S is positive semi-definite, we obtain $S\mathbf{y} = 0$. Hence

$$Z_1(i,i) + Z_2(i,i) = 2\sum_{j=1}^{n} \overline{\mathbf{y}(i)}[\mathbf{V}(i,j)]\mathbf{y}(j);$$

and therefore,

(7.24)
$$S(i,i) = 2\sum_{j \in [n], j \neq i} \overline{\mathbf{y}(i)}\mathbf{V}(i,j)\mathbf{y}(j) - 2\mathbf{V}(i,i)$$

For $i \neq j$,

(7.25)
$$S(i,j) = -2\mathbf{V}(i,j)$$

For a Hermitian matrix M, we define the Laplacian $\Delta(M)$ of M by

$$\Delta(M) := \text{diag}(M\mathbf{1}) - M$$

where $\mathbf{1}$ is the column vector all of whose entries are 1. Then from (7.24), (7.25), explicit computations show that

$$S = 2\text{diag}(\mathbf{y})[\Delta(\text{diag}(\overline{\mathbf{y}})\mathbf{V}\text{diag}(\mathbf{y}))]\text{diag}(\overline{\mathbf{y}})$$

Moreover,

$$\text{diag}(\overline{\mathbf{y}})\mathbf{V}\text{diag}(\mathbf{y}) = \text{diag}(\overline{\mathbf{y}})[\mathbf{y}\overline{\mathbf{y}}^t + \sigma\text{diag}(\mathbf{y})\mathbf{W}_s\text{diag}(\overline{\mathbf{y}})]\text{diag}(\mathbf{y})$$
$$= \mathbf{1}\mathbf{1}^t + \sigma\mathbf{W}_s$$

Therefore

$$(7.26) \qquad \Delta[\text{diag}(\overline{\mathbf{y}})\mathbf{U}\text{diag}(\mathbf{y})] = n\left(I_{n\times n} - \frac{1}{n}\mathbf{1}\mathbf{1}^t\right) + \sigma\Delta[\mathbf{W}_s].$$

The matrix $n\left(I_{n\times n} - \frac{1}{n}\mathbf{1}\mathbf{1}^t\right)$ has rank $(n-1)$ and two distinct eigenvalues: 0 and n. The eigenvalue 0 has multiplicity 1 and n has multiplicity $(n-1)$. Note that 0 is also an eigenvalue of the matrix $\Delta[\mathbf{W}_s]$, and the n-dimensional vector $\mathbf{1}$ is an eigenvector with respect to the eigenvalue 0 for both the matrix $n\left(I_{n\times n} - \frac{1}{n}\mathbf{1}\mathbf{1}^t\right)$ and the matrix $\Delta[\mathbf{W}_s]$. Therefore the matrix (7.26) is positive definite if

$$(7.27) \qquad \sigma\|\Delta[\mathbf{W}_s]\| \leq n,$$

where $\|\cdot\|$ is the spectral norm of a matrix defined to be the largest modulus of its eigenvalues. We have, by the triangle inequality,

$$(7.28) \qquad \|\Delta\mathbf{W}_s\| \leq \max_{i\in[n]}\left|\sum_{j=1}^{n}\mathbf{W}_s(i,j)\right| + \|\mathbf{W}_s\|.$$

The matrix $\mathbf{W}_s$ is a standard GOE matrix. Recall the following proposition about the largest eigenvalue of the standard GOE matrix.

Proposition 7.9. *(Tracy-Widom [9]) Let $\lambda_{\max}^{(n)}$ be the largest eigenvalue of an $n \times n$ GOE matrix, then when n is large,*

$$\lambda_{\max}^{(n)} \sim \sqrt{2n} + \frac{n^{-\frac{1}{6}}\xi_1}{\sqrt{2}}$$

where ξ_1 is a random variable independent of n with the GOE Tracy-Widom distribution.

By Proposition 7.9, for any fixed $\delta > 0$,

$$(7.29) \qquad \lim_{n\to\infty}\mathbb{P}(\|\mathbf{W}_s\| \geq (1+\delta)\sqrt{2n}) = 0.$$

Now we consider the distribution of $\max_{i\in[n]}\left|\sum_{j=1}^{n}\mathbf{W}_s(i,j)\right|$. Since we require that $\mathbf{W}_s$ is a symmetric matrix, the identically distributed Gaussion random variables $\{\sum_{j=1}^{n}\mathbf{W}_s(i,j)\}_{i\in[n]}$ are no longer independent. In our case the random vector $(\sum_{j=1}^{n}\mathbf{W}_s(1,j),\ldots,\sum_{j=1}^{n}\mathbf{W}_s(n,j))$ is a Gaussian random vector with mean 0 and covariance matrix given by

$$\Sigma = \begin{pmatrix} n & 1 & \ldots & 1 \\ 1 & n & \ldots & 1 \\ & \ldots & & \\ 1 & \ldots & 1 & n \end{pmatrix}$$

The following proposition was proved in [3].

Proposition 7.10. *(Theorem 2.2 of [3])Consider a triangular array of normal random variables $\xi_{n,i}$, $i = 1, 2, \ldots,$, and $n = 1, 2, \ldots$, such that for each n, $\{\xi_{n,i}, i \geq 0\}$ is a stationary normal sequence. Assume $\xi_{n,i} \sim \mathcal{N}(0, 1)$. Let $\rho_{n,j} := \mathbb{E}(\xi_{n,i}, \xi_{n,i+j})$ and assuming that*

$$(1 - \rho_{n,j}) \log n \to \delta_j \in (0, \infty], \text{ for all } j \geq 1 \text{ as } n \to \infty$$

Assume that there exist positive integers l_n satisfying $l_n = o(n)$ and for which

$$\lim_{n \to \infty} \sup_{j \geq l_n} |\rho_{n,j}| \log n = 0.$$

and

$$\lim_{m \to \infty} \limsup_{n \to \infty} \sum_{j=m}^{l_n} n^{-\frac{1-\rho_{n,j}}{1+\rho_{n,j}}} \frac{(\log n)^{-\frac{\rho_{n,j}}{1+\rho_{n,j}}}}{(1 - \rho_{n,j}^2)^{\frac{1}{2}}} = 0$$

Then

$$\lim_{n \to \infty} \Pr(\max_{1 \leq i \leq n} \xi_{n,i} \leq u_n(x)) = \exp[-\theta \exp(-x)]$$

where

$$u_n(x) = \frac{x}{a_n} + b_n;$$

and

$$
\begin{aligned}
a_n &= \sqrt{2 \log n} \\
b_n &= \sqrt{2 \log n} + \frac{\log \log n + \log 4\pi}{2\sqrt{2 \log n}}
\end{aligned}
$$

and $\theta \in [0, 1]$ is a constant. In particular $\theta = 1$ if $\delta_j = \infty$ for all $j \geq 1$.

By Proposition 7.10,

$$
\begin{aligned}
&\max_{i \in [n]} \left| \sum_{j=1}^{n} \mathbf{W}_s(i, j) \right| \\
&\geq (1 - \delta) \sqrt{2 \log n \left(\max_{i \in [n]} \left[\mathrm{Var} \left(\sum_{j \in [n]} \mathbf{W}_s(i, j) \right) \right] \right)} \\
&= (1 - \delta) \sqrt{2n \log n}
\end{aligned}
$$

with probability $1 - o_n(1)$. Moreover, by Lemma 3.9, for any fixed $\epsilon > 0$

$$(7.30) \qquad \max_{i \in [n]} \left| \sum_{j=1}^{n} \mathbf{W}_s(i, j) \right| \leq (1 + \epsilon) \sqrt{2n \log n}$$

with probability $1 - o(1)$.

Then Theorem 2.5 follows from (7.27), (7.28), (7.29) and (7.30).

Appendix A. Proof of Lemma 3.9

It is known that for a standard Gaussian random variable G_i and $x > 0$,

$$(A.1) \qquad \frac{xe^{-\frac{x^2}{2}}}{\sqrt{2\pi}(1+x^2)} \leq \mathbf{Pr}(G_i > x) \leq \frac{e^{-\frac{x^2}{2}}}{x\sqrt{2\pi}}$$

Let $G_1, \ldots, G_N$ be N standard Gaussian random variables. Then by (A.1) we have

$$\begin{aligned}
\Pr\left(\max_{i\in[N]} G_i \geq (1+\epsilon)\sqrt{2\log N}\right) &\leq \sum_{i\in[N]} \Pr\left(G_i \geq (1+\epsilon)\sqrt{2\log N}\right) \\
&\leq \frac{Ne^{-(1+\epsilon)^2\log N}}{2(1+\epsilon)\sqrt{\pi\log N}} \\
&\leq N^{-\epsilon}
\end{aligned}$$

If we further assume that G_i's are independent, then

$$\begin{aligned}
\Pr\left(\max_{i\in[N]} G_i < (1-\epsilon)\sqrt{2\log N}\right) &= \prod_{i\in[N]} \Pr\left(G_i < (1-\epsilon)\sqrt{2\log N}\right) \\
&= \prod_{i\in[N]} \left[1 - \Pr\left(G_i > (1-\epsilon)\sqrt{2\log N}\right)\right]
\end{aligned}$$

By (A.1) we obtain

$$\Pr\left(\max_{i\in[N]} G_i < (1-\epsilon)\sqrt{2\log N}\right) \leq \left(1 - \frac{(1-\epsilon)\sqrt{2\log N}}{\sqrt{2\pi}(1+2(1-\epsilon)^2\log N)}\frac{1}{N^{(1-\epsilon)^2}}\right)^N$$

When (3.22) holds, we have

$$\Pr\left(\max_{i\in[N]} G_i < (1-\epsilon)\sqrt{2\log N}\right) \leq \left(1 - \frac{1}{N^{1-\epsilon}}\right)^{N^{1-\epsilon}\cdot N^\epsilon} \leq e^{-N^\epsilon}$$

Then the lemma follows.

References

[1] Emmanuel Abbe, *Community detection and stochastic block models: Recent developments*, Journal of Machine Learning Research **18** (2018), 1–86.

[2] Emmanuel Abbe, Afonso S. Bandeira, and Georgina Hall, *Exact recovery in the stochastic block model*, IEEE Transactions on Information Theory **62** (2016), 471–487.

[3] Tailen Hsing, Jürg Hüsler, and Rolf-Dieter Reiss, *The extremes of a triangular array of normal random variables*, The Annals of Applied Probability **6** (1996), 671–686.

[4] A. Javanmard, A. Montanari, and F. Ricci-Tersenghi, *Performance of a community detection algorithm based on semidefinite programming*, J. Phys.: Conf. Ser. 699 (2016), 012015.

[5] _____, *Phase transitions in semidefinite relaxations*, Proceedings of the National Academy of Sciences **113(16)** (2016), E2218–2223.

[6] C. Kim, A. Bandeira, and M. Goemans, *Community detection in hypergraphs, spikedtensor models, and sum-of-squares*, 2017 12th International Conference on Sampling Theory and Applications (2017), 124–128.

[7] L. Massoulié, *Community detection thresholds and the weak Ramanujan property*, Proceedings of the 46th Annual ACM Symposium on Theory of Computing (2014), 694–703.

[8] Elchanan Mossel, Joe Neeman, and Allan Sly, *A proof of the blockmodel threshold conjecture*, Combinatorica **38** (2018), 665–708.

[9] Craig A. Tracy and Harold Widom, *Distribution functions for largest eigenvalues and their applications*, Proceedings of the International Congress of Mathematicians **I** (2002), 587–596.

Department of Mathematics, University of Connecticut, Storrs, Connecticut 06269-3009, USA

Email address: zhongyang.li@uconn.edu

URL: https://mathzhongyangli.wordpress.com