Random Graph Matching in Geometric Models: the Case of Complete Graphs

Haoyu Wang HAOYU.WANG@YALE.EDU
Department of Mathematics, Yale University, New Haven, USA

Yihong Wu YIHONG.WU@YALE.EDU
Department of Statistics and Data Science, Yale University, New Haven, USA

Jiaming Xu JX77@DUKE.EDU
Fuqua School of Business, Duke University, Durham NC, USA

Israel Yolou ISRAEL.YOLOU@YALE.EDU
Departments of Mathematics and Computer Science, Yale University, New Haven, USA

Editors: Po-Ling Loh and Maxim Raginsky

Abstract

This paper studies the problem of matching two complete graphs with edge weights correlated through latent geometries, extending a recent line of research on random graph matching with independent edge weights to geometric models. Specifically, given a random permutation π^* on $[n]$ and n iid pairs of correlated Gaussian vectors $\{X_{\pi^*(i)}, Y_i\}$ in $\mathbb{R}^d$ with noise parameter σ, the edge weights are given by $A_{ij} = \kappa(X_i, X_j)$ and $B_{ij} = \kappa(Y_i, Y_j)$ for some link function κ. The goal is to recover the hidden vertex correspondence π^* based on the observation of A and B. We focus on the dot-product model with $\kappa(x, y) = \langle x, y\rangle$ and Euclidean distance model with $\kappa(x, y) = \|x - y\|^2$, in the low-dimensional regime of $d = o(\log n)$ wherein the underlying geometric structures are most evident. We derive an approximate maximum likelihood estimator, which provably achieves, with high probability, perfect recovery of π^* when $\sigma = o(n^{-2/d})$ and almost perfect recovery with a vanishing fraction of errors when $\sigma = o(n^{-1/d})$. Furthermore, these conditions are shown to be information-theoretically optimal even when the latent coordinates $\{X_i\}$ and $\{Y_i\}$ are observed, complementing the recent results of Dai et al. (2019) and Kunisky and Niles-Weed (2022) in geometric models of the planted bipartite matching problem. As a side discovery, we show that the celebrated spectral algorithm of Umeyama (1988) emerges as a further approximation to the maximum likelihood in the geometric model.

1. Introduction

Graph matching (or network alignment) refers to finding the best vertex correspondence between two graphs that maximizes the total number of common edges. While this problem, as an instance of quadratic assignment problem, is computationally intractable in the worst case, significant headways, both information-theoretic and algorithmic, have been achieved in the average-case analysis under meaningful statistical models Cullina and Kiyavash (2016, 2017); Ding et al. (2021a); Barak et al. (2019); Fan et al. (2019a,b); Hall and Massoulié (2020); Wu et al. (2021); Ganassali and Massoulié (2020); Ganassali et al. (2022); Mao et al. (2021b,a). One of the most popular models is the *correlated Erdős-Rényi graph* model Pedarsani and Grossglauser (2011), where both observed graphs are Erdős-Rényi graphs with edges correlated through a latent vertex matching; more generally, in the *correlated Wigner* model, the observations are two weighted graph with correlated edge weights (e.g. Gaussians Ding et al. (2021a); Dai et al. (2019); Fan et al. (2019a); Ganassali (2021)).

Despite their simplicity, these models inspired a number of new algorithms that achieve strong performance both theoretically and practically Ding et al. (2021a); Fan et al. (2019a,b); Ganassali and Massoulié (2020); Ganassali et al. (2022); Mao et al. (2021b,a). Nevertheless, one of the major limitations of models with independent edges is that they fail to capture graphs with spatial structure Armiti and Gertz (2014), such as those arising in computer vision datasets (e.g. mesh graphs obtained by triangulating 3D shapes Lähner et al. (2016)). In contrast to Erdős-Rényi-style model with iid edges, *geometric graph models*, such as random dot-product graphs and random geometric graphs, take into account the latent geometry by embedding each node in a Euclidean space and determines edge connection between two nodes by the proximity of their geographical location. While the coordinates are typically assumed to be independent (e.g. Gaussians or uniform over spheres or hypercubes), the edges or edge weights are now dependent. The main objective for the present paper is to study graph matching in correlated geometric graph models, where the network correlation is due to that of the latent coordinates.

1.1. Model

Given two point clouds $\{X_1, \dots, X_n\}$ and $\{Y_1, \dots, Y_n\}$ in $\mathbb{R}^d$, we construct two weighted graphs on the vertex set $[n]$ with weighted adjacency matrices A and B as follows. For each i, j, let $A_{ij} \overset{\text{ind}}{\sim} W(\cdot|X_i, X_j)$ and $B_{ij} \overset{\text{ind}}{\sim} W(\cdot|Y_i, Y_j)$, for some probability transition kernel W. The coordinates are correlated through a latent matching as follows: Consider a Gaussian model

$$Y_i = X_{\pi^*(i)} + \sigma Z_i, \quad i = 1, \dots, n,$$

where X_i, Z_i's are iid $\mathcal{N}(0, I_d)$ vectors and π^* is uniform on S_n, the set of all permutations on $[n]$. In matrix form, we have

$$Y = \Pi^* X + \sigma Z, \tag{1}$$

where $X, Y, Z \in \mathbb{R}^{n\times d}$ are matrices whose rows are X_i's, Y_i's and Z_i's respectively, $\Pi^* \in \mathfrak{S}_n$ denotes the permutation matrix corresponding to π^*, and $\mathfrak{S}_n$ is the collection of all permutation matrices. Given the observation A and B, the goal is to recover the latent correspondence π^*.

Of particular interest are the following special cases:

- *Dot-product model*: The observations are complete graphs with pairwise inner products as edge weights, namely, $A_{ij} = \langle X_i, X_j\rangle$ and $B_{ij} = \langle Y_i, Y_j\rangle$. As such, the weighted adjacency matrices are $A = XX^\top$ and $B = YY^\top$, both Wishart matrices. It is clear that from A and B one can reconstruct X and Y respectively, each up to a global orthogonal transformation on the rows. In this light, the model is also equivalent to the so-called *Procrustes Matching* problem Maron et al. (2016); Dym and Lipman (2017); Grave et al. (2019), where Y in (1) undergoes a further random orthogonal transformation – see Appendix A for a detailed discussion.

- *Distance model*: The edge weights are pairwise squared distances $A_{ij} = \|X_i - X_j\|^2$ and $B_{ij} = \|X_i - X_j\|^2$. This setting corresponds to the classical problem of multi-dimensional scaling (MDS), where the goal is to reconstruct the coordinates (up to global shift and orthogonal transformation) from the distance data (cf. Borg and Groenen (2005)).

- *Random Dot Product Graph (RDPG)*: In this model, the observed data are two graphs with adjacency matrices A and B, where $A_{ij} \overset{\text{ind}}{\sim} \text{Bern}\left(\kappa(\langle X_i, X_j\rangle)\right)$ and $B_{ij} \overset{\text{ind}}{\sim} \text{Bern}\left(\kappa(\langle Y_i, Y_j\rangle)\right)$

conditioned on X and Y, and $\kappa : \mathbb{R} \to [0,1]$ is some link function, e.g. $\kappa(t) = e^{-t^2/2}$. In this way, we observe two instances of RDPG that are correlated through the underlying points and the latent matching. See Athreya et al. (2017) for a recent survey on RDPG.

- *Random Geometric Graph (RGG)*: Similar to RDPG, $A_{ij} \overset{\text{ind}}{\sim} \text{Bern}(\kappa(\|X_i - X_j\|))$ conditioned on $X_1, \dots, X_n$ for some link function $\kappa : \mathbb{R}_+ \to [0,1]$ applied to the pairwise distances. The second RGG instance B is constructed in the same way using $Y_1, \dots, Y_n$. A simple example is $\kappa(t) = \mathbf{1}_{\{t \le r\}}$ for some threshold $r > 0$, where each pair of points within distance r is connected Gilbert (1961); see the monograph Penrose (2003) for a comprehensive discussion on RGG.

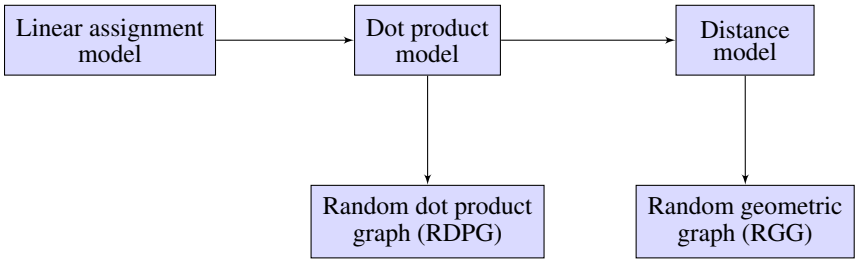


Figure 1: Geometric matching models. Here arrows denote statistical ordering.

Let us mention that the model where the two point clouds are directly observed has been recently studied by Dai et al. (2019, 2020) in the context of feature matching and independently by Kunisky and Niles-Weed (2022) as a geometric model for the planted matching problem, extending the previous work in Chertkov et al. (2010); Moharrami et al. (2021); Ding et al. (2021b) with iid weights to a geometric (low-rank) setting. In this model, X and Y in (1) are observed and the maximum likelihood estimator (MLE) of π^* amounts to solving

$$\max_{\Pi \in \mathfrak{S}_n} \langle Y, \Pi X \rangle. \tag{2}$$

which is a linear assignment (max-weight matching) problem on the weighted complete bipartite graph with weight matrix $YX^\top$. In the sequel we shall refer to this setting as the linear assignment model, which we also study in this paper for the sake of proving impossibility results for the more difficult graph matching problem, as the coordinates are latent and only *pairwise information* are available.

Fig. 1.1 elucidates the logical connections between the aforementioned models. Among these, linear assignment model is the most informative, followed by the dot product model and the distance model, whose further stochastically degraded versions are RDPG and RGG, respectively. As a first step towards understanding graph matching on geometric models, in this paper we study the case of weighted complete graphs in the dot product and distance models.

1.2. Main results

By analyzing the MLE (2) in the stronger linear assignment model (1), Kunisky and Niles-Weed (2022) identified a critical scaling of dimension d at $\log n$:

- In the low-dimensional regime of $d \ll \log n$, accurate reconstruction requires the noise level σ to be vanishingly small. More precisely, with high probability, the MLE (2) recovers the latent π^* perfectly (resp. with a vanishing fraction of errors) provided that $\sigma = o(n^{-2/d})$ (resp. $\sigma = o(n^{-1/d})$).
- In the high-dimensional regime of $d \gg \log n$, it is possible for σ^2 to be as large as $\frac{d}{(4+o(1))\log n}$. Since the dependency between the edges weakens as the latent dimension increases,[1] this is consistent with the known results in the planted matching model with independent Gaussian weights Ding et al. (2021b).[2]

In this paper we mostly focus on the low-dimensional setting as this is the regime where geometric graph ensembles are structurally distinct from Erdős-Rényi graphs. Our main findings are two-fold:

1. The same reconstruction thresholds remain achievable even when the coordinates are latent and only inner-product or distance data are accessible.
2. Furthermore, these thresholds cannot be improved even when the coordinates are observed.

To make these results precise, we start with the dot-product model with $A = XX^\top$ and $B = YY^\top$, and $Y = \Pi^* X + \sigma Z$ according to (1). In this case the MLE turns out to be much more complicated than (2) for the linear assignment model. As shown in Appendix B, the MLE takes the form

$$\widehat{\Pi}_{\mathrm{ML}} = \arg\max_{\Pi \in \mathfrak{S}_n} \int_{O(d)} \mathrm{d}Q \exp\left(\frac{\langle B^{1/2}, \Pi A^{1/2} Q\rangle}{\sigma^2}\right), \tag{3}$$

where the integral is with respect to the Haar measure on the orthogonal group $O(d)$, $A^{1/2} \triangleq U\Lambda^{1/2} \in \mathbb{R}^{n\times d}$ based on the SVD $A = U\Lambda U^\top$, and similarly for $B^{1/2}$. It is unclear whether the above Haar integral has a closed-form solution,[3] let alone how to optimize it over all permutations. Next, we turn to its approximation.

As we will show later, in the low-dimensional case of $d = o(\log n)$, meaningful reconstruction of the latent matching is information-theoretically impossible unless σ vanishes with n at a certain speed. In the regime of small σ, Laplace's method suggests that the predominant contribution to the integral in (3) comes from the maximum $\langle B^{1/2}, \Pi A^{1/2} Q\rangle$ over $Q \in O(d)$. Using the dual form of the nuclear norm $\|X\|_* = \max_{Q\in O(d)} \langle X, Q\rangle$, where $\|X\|_*$ denotes the sum of all singular values of X, we arrive at the following approximate MLE:

$$\widehat{\Pi}_{\mathrm{AML}} = \arg\max_{\Pi \in \mathfrak{S}_n} \|(A^{1/2})^\top \Pi^\top B^{1/2}\|_*. \tag{4}$$

We stress that the above approximation to the MLE (3) is justified for the low-dimensional regime where σ is small. In the high-dimensional (high-noise) case, the approximate MLE actually takes

1. For the Wishart matrix, it is known Jiang and Li (2015); Bubeck et al. (2016); Bubeck and Ganguly (2018) that the total variation between the joint law of the off-diagonals and their iid Gaussian counterpart converges to zero provided that $d = \omega(n^3)$. Based on these results, Bubeck et al. (2016) further shows that high-dimensional RGG is approximately Erdős-Rényi.
2. Analogously, to match two GOE matrices with correlation coefficient ρ, the sharp reconstruction threshold is at $\rho^2 = \frac{(4+o(1))\log n}{n}$ Ganassali (2021); Wu et al. (2021).
3. The integral in (3) can be reduced to computing $\int \mathrm{d}Q \exp(\langle \Lambda, Q\rangle)$ for a diagonal Λ, which, in principle, can be evaluated by Taylor expansion and applying formulas for the joint moments of Q in (Matsumoto, 2013, Theorem 2.2).

on the form of a quadratic assignment problem (QAP), which is the MLE for the well-studied iid model Cullina and Kiyavash (2016); in the special case of the dot-product model, it amounts to replacing the nuclear norm in (4) by the Frobenius norm. We postpone this discussion to Section 4.

To measure the accuracy of a given estimator $\widehat{\pi}$, we define

$$\mathsf{overlap}(\widehat{\pi}, \pi) \triangleq \frac{1}{n} |\{i \in [n] : \widehat{\pi}(i) = \pi(i)\}|$$

as the fraction of nodes whose matching is correctly recovered. The following result identifies the threshold at which the approximate MLE achieves perfect or almost perfect recovery.

Theorem 1 (Recovery guarantee of AML in the dot-product model) *Assume the dot-product model with $d = o(\log n)$. Let $\widehat{\pi}_{\mathrm{AML}}$ be the approximate MLE defined in (4).*

(i) If $\sigma \ll n^{-2/d}$, the estimator $\widehat{\pi}_{\mathrm{AML}}$ achieves perfect recovery with high probability:

$$\mathbb{P}\{\mathsf{overlap}(\widehat{\pi}_{\mathrm{AML}}, \pi^*) = 1\} = 1 - o(1). \tag{5}$$

(ii) If $\sigma \ll n^{-1/d}$, the estimator $\widehat{\pi}_{\mathrm{AML}}$ achieves almost perfect recovery with high probability:

$$\mathbb{P}\{\mathsf{overlap}(\widehat{\pi}_{\mathrm{AML}}, \pi^*) \geq 1 - o(1)\} = 1 - o(1). \tag{6}$$

A few remarks are in order:

- In fact we will show the following nonasymptotic estimate that implies (6): For all sufficiently small ε, if $\sigma^{-d} > 16n2^{2/\varepsilon}$, then $\mathsf{overlap}(\widehat{\pi}_{\mathrm{AML}}, \pi^*) \geq 1 - \varepsilon$ with probability tending to one.
- The estimator (4) has previously appeared in the literature of Procrustes matching Grave et al. (2019), albeit not as an approximation to the MLE in a generative model. See Appendix A for a detailed discussion.
- Unlike linear assignment, it is unclear how to solve the optimization in (4) over permutations efficiently. Nevertheless, for constant d we show that it is possible to find an approximate solution in time that is polynomial in n that achieves the same statistical guarantee as in Theorem 1. Indeed, note that (3) is equivalent to the double maximization

$$\widehat{\Pi}_{\mathrm{AML}} = \arg\max_{\Pi \in \mathfrak{S}_n} \max_{Q \in O(d)} \langle B^{1/2}, \Pi A^{1/2} Q \rangle. \tag{7}$$

  Approximating the inner maximum over a suitable discretization of $O(d)$, each maximization over Π for fixed Q is a linear assignment problem, which can be solved in $O(n^3)$ time. In Section 3, we provide a heuristic that shows (7) can be further approximated by the classical spectral algorithm of Umeyama Umeyama (1988) which is much faster in practice and achieves good empirical performance. For d that grows with n, it is an open question to find a polynomial-time algorithm that attains the (optimal, as we show next) threshold in Theorem 1.

Next, we proceed to the more difficult distance model, where $A_{ij} = \|X_i - X_j\|^2$ and $B_{ij} = \|Y_i - Y_j\|^2$. Deriving the exact MLE in this model appears to be challenging; instead, we apply the

estimator (4) to an appropriately centered version of the data matrices. Let $\mathbf{1} \in \mathbb{R}^n$ denotes the all-one vector and define $\mathbf{F} = \frac{1}{n}\mathbf{1}\mathbf{1}^\top$. Then $A = -2XX^\top + a\mathbf{1}^\top + \mathbf{1}a^\top$ and $B = -2YY^\top + b\mathbf{1}^\top + \mathbf{1}b^\top$, where $a = (\|X_i\|^2)$ and $b = (\|Y_i\|^2)$. Strictly speaking, the vectors a and b are correlated with the ground truth π^*, since b can be viewed as a noisy version of $\Pi^* a$; however, we expect them to inform very little about π^* because such scalar-valued observations are highly sensitive to noise (analogous to degree matching in correlated Erdős-Rényi graphs (Ding et al., 2021a, Section 1.3)). As such, we ignore a and b by projecting A and B to the orthogonal complement of the vector $\mathbf{1}$. Specifically, we compute, as commonly done in the MDS literature (see e.g. Shang et al. (2003); Oh et al. (2010)),

$$\widetilde{A} = -\frac{1}{2}(I - \mathbf{F})A(I - \mathbf{F}), \quad \widetilde{B} = -\frac{1}{2}(I - \mathbf{F})B(I - \mathbf{F}). \tag{8}$$

It is easy to verify that $\widetilde{A} = \widetilde{X}\widetilde{X}^\top$ and $\widetilde{B} = \widetilde{Y}\widetilde{Y}^\top$, where $\widetilde{X} = (I - \mathbf{F})X$ and $\widetilde{Y} = (I - \mathbf{F})Y$ consist of centered coordinates $\widetilde{X}_i = X_i - \bar{X}$ and $\widetilde{Y}_i = Y_i - \bar{Y}$ respectively, with $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$ and $\bar{Y} = \frac{1}{n}\sum_{i=1}^n Y_i$. Overall, we have reduced the distance model to a dot product model where the latent coordinates are now centered.

One can show that the MLE of Π^* given the reduced data $(\widetilde{A}, \widetilde{B})$ is of the same Haar-integral form (3). Using again the small-σ approximation, we arrive at the following estimator by applying (4) to the centered data $\widetilde{A}$ and $\widetilde{B}$:

$$\widetilde{\Pi}_{\mathrm{AML}} = \arg\max_{\Pi \in \mathfrak{S}_n} \|(\widetilde{A}^{1/2})^\top \Pi^\top \widetilde{B}^{1/2}\|_*. \tag{9}$$

Theorem 2 (Recovery guarantee in the distance model) *Assuming the distance model, Theorem 1 holds under the same condition on d and σ, with the estimator $\widetilde{\Pi}_{\mathrm{AML}}$ in (9) replacing $\widehat{\Pi}_{\mathrm{AML}}$ in (4).*

Finally, we state an impossibility result for the linear assignment model, proving that the perfect and almost perfect recovery threshold of $\sigma = o(n^{-2/d})$ and $\sigma = o(n^{-1/d})$ obtained by analyzing the MLE in Kunisky and Niles-Weed (2022) are in fact information-theoretically necessary. Complementing Theorem 1 and Theorem 2, this result also establishes the optimality of the estimator (4) and (9) for their respective model.

Theorem 3 (Impossibility result in the linear assignment model) *Consider the linear assignment model with $d = o(\log n)$.*

(i) If there exists an estimator that achieves perfect recovery with high probability, then $\sigma \le n^{-2/d}$.

(ii) If there exists an estimator that achieves almost perfect recovery with high probability, then $\sigma \le n^{-(1-o(1))/d}$.

Furthermore, in the special case of $d = \Theta(1)$, necessary conditions in (i) and (ii) can be improved to $\sigma \le o(n^{-2/d})$ and $\sigma \le o(n^{-1/d})$, respectively.

Theorem 3(i) slightly improves the necessary condition for perfect recovery in Kunisky and Niles-Weed (2022) from $\sigma = O(n^{-2/d})$ to $\sigma = o(n^{-2/d})$. For almost perfect recovery, the negative result in Kunisky and Niles-Weed (2022) is limited to MLE, while Theorem 3 holds for all algorithms. Moreover, the necessary condition in Theorem 3(ii) was conjectured in (Kunisky and Niles-Weed, 2022, Conjecture 1.4, item 1), which we now resolve in the positive. Finally, while

our focus is in the low-dimensional case of $d = o(\log n)$, we also provide necessary conditions that hold for general d. (See Appendix E for details).

In view of Fig. 1.1, since the negative results in Theorem 3 are proved for the strongest model and the positive results in Theorem 2 are for the weakest model, we conclude that for all three models, namely, linear assignment, dot-product, and distance model, the thresholds for exact and almost perfect reconstruction is given by $n^{-2/d}$ and $n^{-1/d}$, respectively.

2. Outline of proofs

2.1. Positive results

The positive results of Theorem 1 and Theorem 2 are proved in Appendix C and Appendix D. Here we briefly describe the proof strategy in the dot product model. Suppose we want to bound the probability that the approximate MLE $\widehat{\Pi}_{\mathrm{AML}}$ in (4) makes more than t number of errors. Denote by $\mathrm{d}(\pi_1, \pi_2) \triangleq \sum_{i=1}^n \mathbf{1}_{\{\pi_1(i) \neq \pi_2(i)\}}$ the Hamming distance between two permutations $\pi_1, \pi_2 \in S_n$. Without loss of generality, we will assume that $\pi^* = \mathrm{Id}$. By the orthogonal invariance of $\|\cdot\|_*$, we can assume, *for the sake of analysis*, that $A^{1/2} = X$ and $B^{1/2} = Y$. Applying (7),

$$\begin{aligned}\mathbb{P}\left\{\mathrm{d}(\widehat{\Pi}_{\mathrm{AML}}, \mathrm{Id}) > t\right\} \leq & \mathbb{P}\left\{\max_{\pi:\mathrm{d}(\pi,\mathrm{Id})>t} \|X^\top \Pi^\top Y\|_* \geq \|X^\top Y\|_*\right\} \\ \leq & \mathbb{P}\left\{\max_{\pi:\mathrm{d}(\pi,\mathrm{Id})>t} \max_{Q\in O(d)} \langle X^\top \Pi^\top Y, Q\rangle \geq \langle X^\top Y, I_d\rangle\right\}. \end{aligned} \tag{10}$$

For each fixed Π and Q, averaging over the noise yields, for some absolute constant c_0,

$$\mathbb{P}\left\{\langle X^\top \Pi^\top Y, Q\rangle \geq \langle X^\top Y, I_d\rangle\right\} \leq \mathbb{E}\exp\left\{-\frac{c_0}{\sigma^2}\|X - \Pi X Q\|_{\mathrm{F}}^2\right\}. \tag{11}$$

In the remaining argument, there are three places where the structure of the orthogonal group $O(d)$ plays a crucial role:

1. The quantity in (11) turns out to depend on Π through its cycle type and on Q through its eigenvalues. Crucially, the eigenvalues of an orthogonal matrix Q lie on the unit circle, denoted by $(e^{\mathrm{i}\theta_1}, \ldots, e^{\mathrm{i}\theta_d})$, with $|\theta_\ell| \leq \pi$. We then show that the error probability in (11) can be further bounded by, for some absolute constant C_0,

$$(C_0\sigma)^{d(n-\mathfrak{c})}\left(\prod_{\ell=1}^d \frac{C_0\sigma}{\sigma + |\theta_\ell|}\right)^{n_1}, \tag{12}$$

   where n_1 is the number of fixed points in π and $\mathfrak{c}$ is the total number of cycles.

2. In order to bound (10), we take a union bound over π and another union bound over an appropriate discretization of $O(d)$. This turns out to be much subtler than the usual δ-net-based argument, as one needs to implement a localized covering and take into account the local geometry of the orthogonal group. Specifically, note that the error probability in (11) becomes larger when π is near Id and when Q is near I_d (i.e. the phases $|\theta_\ell|$'s are small); fortunately, the entropy (namely, the number of such π and such Q within a certain resolution) also becomes smaller, balancing out the deterioration in the probability bound. This is the second place where the structure of $O(d)$ is used crucially, as the local metric entropy of $O(d)$ in the vicinity of I_d is much lower than that elsewhere.

3. Controlling the approximation error of the nuclear norm is another key step. Note that for any matrix norm of the dual form $\|A\| = \sup_{\|Q\|' \le 1} \langle A, Q \rangle$, where $\|\cdot\|'$ is the dual norm of $\|\cdot\|$, the standard δ-net argument (cf. (Vershynin, 2018, Lemma 4.4.1)) yields a multiplicative approximation $\max_{Q \in N} \langle A, Q \rangle \ge (1-\delta)\|A\|$, where N is any δ-net of the dual norm ball. In general, this result cannot be improved (e.g. for Frobenius norm); nevertheless, for the special case of nuclear norm, this approximation ratio can be improved from $1-\delta$ to $1-\delta^2$, as the following result of independent interest shows. This improvement turns out to be crucial for obtaining the sharp threshold.

   Lemma 4 *Let $N \subset O(d)$ be a δ-net in operator norm of the orthogonal group $O(d)$. For any $A \in \mathbb{R}^{d \times d}$,*

$$\max_{Q \in N} \langle A, Q \rangle \ge \left(1 - \frac{\delta^2}{2}\right) \|A\|_*. \tag{13}$$

The proof of Theorem 1 is completed by combining (12) with a union bound over a specific discretization of $O(d)$, whose cardinality satisfies the desired eigenvalue-based local entropy estimate, followed by a union bound over π which can be controlled using moment generating function of the number of cycles in a random derangement.

2.2. Negative results

The information-theoretic lower bounds in Theorem 3 for the linear assignment model are proved in Appendix E. Here we sketch the main ideas. We first derive a necessary condition for almost perfect recovery that holds for any d via a simple mutual information argument Hajek et al. (2017): On one hand, the mutual information $I(\pi^*; X, Y)$ can be upper bounded by the Gaussian channel capacity as $\frac{nd}{2}\log(1+\sigma^{-2})$. On the other hand, to achieve almost perfect recovery, $I(\pi^*; X, Y)$ needs be asymptotically equal to the full entropy $H(\pi^*)$ which is $(1-o(1))\log n$. These two assertions together immediately imply that $\frac{nd}{2}\log(1+\sigma^{-2}) \ge ((1-o(1))\log n$, which further simplifies to $\sigma = n^{-(1-o(1))/d}$ when $d = o(\log n)$. However, for constant d, this necessary condition turns out to be loose and the main bulk of our proof is to improve it to the optimal condition $\sigma = o(n^{-1/d})$. To this end, we follow the program recently developed in Ding et al. (2021b) in the context of the planted matching model by analyzing the posterior measure of the latent π^* given the data (X, Y).

To start, a simple yet crucial observation in Ding et al. (2021b) is that to prove the impossibility of almost perfect recovery, it suffices to show a random permutation sampled from the posterior distribution is at Hamming distance $\Omega(n)$ away from the ground truth with constant probability. As such, it suffices to show there is more posterior mass over the bad permutations (those far away from the ground truth) than that over the good permutations (those near the ground truth) in the posterior distribution. To proceed, we first bound from above the total posterior mass of good permutations by a truncated first moment calculation applying the large deviation analysis developed in the proof of the positive results. To bound from below the posterior mass of bad permutations, we aim to construct exponentially many bad permutations π whose log likelihood $L(\pi)$ is no smaller than $L(\pi^*)$. A key observation is that $L(\pi) - L(\pi^*)$ can be decomposed according to the orbit decomposition of $(\pi^*)^{-1} \circ \pi$:

$$L(\pi) - L(\pi^*) = \frac{1}{\sigma^2} \langle \Pi X - \Pi^* X, Y \rangle = \frac{1}{\sigma^2} \sum_{O \in \mathcal{O}} \Delta(O), \tag{14}$$

where $\mathcal{O}$ denotes the set of orbits in $(\pi^*)^{-1} \circ \pi$ and for any orbit $O = (i_1, i_2, \ldots, i_t)$,

$$\Delta(O) \triangleq \sum_{k=1}^{t} \left\langle X_{\pi^*(i_{k+1})} - X_{\pi^*(i_k)}, Y_{i_k} \right\rangle . \tag{15}$$

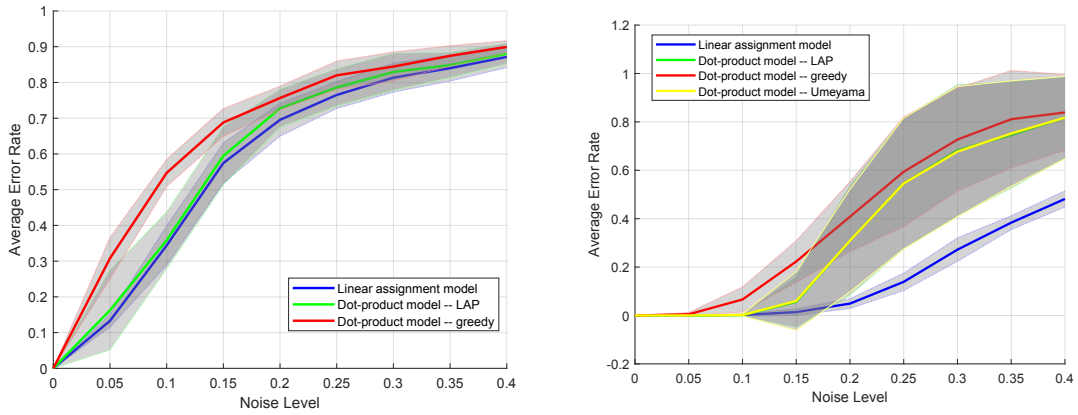
Thus, the goal is to find a collection of vertex-disjoint orbits O whose total lengths add up to $\Omega(n)$ and each of which is *augmenting* in the sense that $\Delta(O) \geq 0$. Here, a key difference to Ding et al. (2021b) is that in the planted matching model with independent edge weights studied there, short augmenting orbits are insufficient to meet the $\Omega(n)$ total length requirement; instead, Ding et al. (2021b) resorts to a sophisticated two-stage process that first finds many augmenting paths then connects then into long cycles. Fortunately, for the linear assignment model in low dimensions of $d = \Theta(1)$, as also observed in Kunisky and Niles-Weed (2022) in their analysis of the MLE, it suffices to look for augmenting 2-orbits and take their disjoint unions. More precisely, we show that there are $\Omega(n)$ many vertex-disjoint augmenting 2-orbits. This has already been done in Kunisky and Niles-Weed (2022) using a second-moment method enhanced by an additional concentration inequality. It turns out that the correlation among the augmenting 2-orbits is mild enough so that a much simpler argument via a basic second-moment calculation followed by an application of Turán's theorem suffices to extract a large vertex-disjoint subcollection. Finally, these vertex-disjoint augmenting 2-orbits give rise to exponentially many permutations that differ from the ground truth by $\Omega(n)$.

Finally, we briefly remark on perfect recovery, for which it suffices to focus on the MLE (2) which minimizes the error probability for uniform π^*. In view of the likelihood decomposition given in (14), it further suffices to prove the existence of *an* augmenting 2-orbit. This can be easily done using the second-moment method. A similar strategy was adopted in Dai et al. (2019), but our first-moment and second-moment estimates are tighter and hence yield nearly optimal conditions.

3. Experiments

In this section we present preliminary numerical results on synthetic data from the dot product model. As observed in Grave et al. (2019), the form of the approximate MLE $\widehat{\Pi}_{\mathrm{AML}}$ in (7) as a double maximization over $\Pi \in \mathfrak{S}_n$ and $Q \in O(d)$ naturally suggests an alternating maximization strategy by iterating between the two steps: (a) For a fixed Q, the Π-maximization is a linear assignment; (b) For a fixed Π, the Q-maximization is the so-called orthogonal Procrustes problem and easily solved via SVD Schönemann (1966). However, with random initialization this method performs rather poorly falling short of the optimal threshold predicted by Theorem 1. While more informative initialization (such as starting from a Π obtained by the doubly-stochastic relaxation of QAP Grave et al. (2019)) can potentially help, in this section we focus on methods that are closer to the original approximate MLE.

As the proof of Theorem 1 shows, as far as achieving the optimal threshold is concerned it suffices to consider a finely discretized $O(d)$. This can be easily implemented in $d = 2$, since any 2×2 orthogonal matrix is either a rotation or reflection of the form: $\left(\begin{smallmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{smallmatrix}\right)$ or $\left(\begin{smallmatrix} \cos(\theta) & \sin(\theta) \\ \sin(\theta) & -\cos(\theta) \end{smallmatrix}\right)$. We then solve (7) on a grid of θ values, by solving the Π-maximization for each such Q and reporting the solution with the highest objective value. As shown in Fig. 2(*a*) for $n = 200$, the performance of the approximate MLE in the dot-product model (green) follows closely that of the MLE in the linear assignment model (blue). Using the greedy matching algorithm (red) in place of the linear assignment solver greatly speeds up the computation at the price of some performance degradation.



(*a*) $n = 200$ and $d = 2$. The green and red curves correspond to (7) on $T_0 = 100$ discretized angles, with exact linear assignment or greedy matching.

(*b*) $n = 200$ and $d = 4$. The green and red curves are based on (17) with exact linear assignment or greedy matching, and the yellow curve corresponds to the Umeyama algorithm (18).

Figure 2: Comparison of dot product model and linear assignment model (averaged over 10 random instances). In the latter, the blue curve corresponds to the MLE (2).

As the dimension increases, it becomes more difficult and computationally more expensive to discretize $O(d)$. Instead, we take a different approach. Note that in the noiseless case ($Y = \Pi^* X$), as long as all singular values have multiplicity one, we have $B^{1/2} = \Pi^* A^{1/2} Q$ for some Q in

$$\mathbb{Z}_2^{\otimes d} = \{\mathsf{diag}(q_i) : q_i \in \{\pm 1\}\}. \tag{16}$$

As such, in the noiseless case it suffices to restrict the inner maximization of (7) to the subgroup $\mathbb{Z}_2^{\otimes d}$ corresponding to coordinate reflections. Since the noise is weak in the low-dimensional setting, we continue to apply this heuristic by computing

$$\widehat{\Pi}_{\mathrm{AML},\mathbb{Z}_2^{\otimes d}} = \arg\max_{\Pi \in \mathfrak{S}_n} \max_{Q \in \mathbb{Z}_2^{\otimes d}} \langle B^{1/2}, \Pi A^{1/2} Q \rangle, \tag{17}$$

which turns out to work very well in practice. Taking this method one step further, notice that in the low-dimensional regime, all non-zero singular values of X and Y are tightly concentrated on the same value $\sqrt{n}$. If we ignore the singular values and simply replace $A^{1/2}$ and $B^{1/2}$ by their left singular vectors $U = [u_1, \ldots, u_d]$ and $V = [v_1, \ldots, v_d]$, (17) can be written more explicitly as

$$\widehat{\Pi}_{\mathrm{Umeyama}} = \arg\max_{\Pi \in \mathfrak{S}_n} \max_{q \in \{\pm 1\}^d} \left\langle \Pi, \sum_{i=1}^{d} q_i v_i u_i^\top \right\rangle, \tag{18}$$

which, somewhat unexpectedly, coincides with the celebrated Umeyama algorithm Umeyama (1988), a specific type of spectral method that is widely used in practice for graph matching. In Fig. 2(*b*) we compare for $n = 200$ and $d = 4$. Consistent with Theorem 1, the error rates in the dot-product

model and the linear assignment model are both near zero until σ exceeds a certain threshold, after which the former departs from the latter. Finally, comparing Fig. 2(*a*) and Fig. 2(*b*) confirms that the reconstruction threshold improves as the latent dimension increases as predicted by Theorem 1.

4. Discussion

In this paper we studied the problem of graph matching in the special case of correlated complete weighted graphs in the dot product and distance model, as a first step towards the more challenging case of random dot-product graphs and random geometric graphs. Within the confines of the present paper, there still remain a number of interesting directions and open problems which we discuss below.

Non-isotropic distribution The present paper assumes the latent coordinates X_i's and Y_i's are isotropic Gaussians. For the linear assignment model, Dai et al. (2019, 2020) has considered a more general setup where $X_i \overset{\text{i.i.d.}}{\sim} N(0, \Sigma)$ for some covariance matrix Σ. As explained in (Dai et al., 2019, Appendix A), it is not hard to see, based on a simple reduction argument (by scaling both X_i's and Y_i's with $\Sigma^{1/2}$ and add noise if needed), that as long as the singular values of Σ are bounded from above and below, the information-theoretic limits in terms of σ remain unchanged. For the dot product or distance model, this is also true but less obvious – see Appendix F for a proof.

While the statistical limits in the nonisotropic case remain the same, potentially it allows more computationally tractable algorithms to succeed. For example, the spectral method recently proposed in Fan et al. (2019a,b) finds a matching by rounding the so-called GRAMPA similarity matrix

$$X = \sum_{i,j=1}^{n} \frac{\langle u_i, \mathbf{1}\rangle \langle v_j, \mathbf{1}\rangle}{(\lambda_i - \mu_j)^2 + \eta^2} u_i v_j^\top. \tag{19}$$

Here $A = \sum \lambda_i u_i u_i^\top$ and $B = \sum \mu_j v_j v_j^\top$ are the SVD of the observed weighted adjacency matrices, and η is a small regularization parameter. In the isotropic case, applying this algorithm to the dot-product model is unlikely to achieve the optimal threshold in Theorem 1. The reason is that in the low-dimensional regime of small d, both A and B and rank-d and all singular values λ_i's and μ_j's are largely concentrated on the same value of $\sqrt{n}$. As such, the similarity matrix (19) degenerates into $X \approx \frac{1}{n\eta^2} \sum_{i,j=1}^n \lambda_i \mu_j \langle u_i, \mathbf{1}\rangle \langle v_j, \mathbf{1}\rangle u_i v_j^\top \propto ab^\top$, where $a = A\mathbf{1}$ and $b = B\mathbf{1}$ are the row-sum vectors. Rounding $ab^\top$ to a permutation matrix is equivalent to "degree-matching", that is, finding the permutation by sorting a and b, which can only tolerate $\sigma = n^{-c}$ type of noise level, for constant c *independent of* the dimension d, due to the small spacing in the order statistics Ding et al. (2021a). However, in the nonisotropic case where Σ has distinct singular values, we expect A and B to have descent spectral gaps and the spectral method (19) may succeed at the dimension-dependent thresholds of Theorem 1. A theoretical justification of this heuristic is outside the scope of this paper.

High-dimensional regime Recall the exact MLE (3), wherein the objective function is an average over the Haar measure on $O(d)$, can be approximated by (4) for small σ. Next, we derive its large-σ approximation. Rewriting the objective function in (3) as $\mathbb{E}[\exp(\frac{1}{\sigma^2}\langle B^{1/2}, \Pi A^{1/2}\mathbf{Q}\rangle)]$ for a random uniform $\mathbf{Q} \in O(d)$ and taking its second-order Taylor expansion for large σ, we get

$$\mathbb{E}\left[\exp\left(\frac{1}{\sigma^2}\langle B^{1/2}, \Pi A^{1/2}\mathbf{Q}\rangle\right)\right] = 1 + \frac{1}{2d\sigma^4}\langle B, \Pi A \Pi^\top\rangle + o(\sigma^{-4}),$$

where we applied $\mathbb{E}[\langle \mathbf{Q}, X\rangle] = 0, \mathbb{E}[\langle \mathbf{Q}, X\rangle^2] = \|X\|_{\mathrm{F}}^2/d$, and $\|(A^{1/2})^\top \Pi^\top B^{1/2}\|_{\mathrm{F}}^2 = \langle B, \Pi A \Pi^\top\rangle$. This expansion suggests that for large σ (which can be afforded in the high-dimensional regime of $d \gg \log n$), the MLE is approximated by the solution to the following QAP:

$$\widehat{\Pi}_{\mathrm{QAP}} = \arg\max_{\Pi \in \mathfrak{S}_n} \langle B, \Pi A \Pi^\top\rangle. \tag{20}$$

This observation aligns with the better studied correlated Erdős-Rényi models or correlated Gaussian Wigner models, where the MLE is exactly given by the QAP (20).

To further compare with the estimator (4) that has been shown optimal in low dimensions, let us rewrite (20) in a form that parallels (7):

$$\widehat{\Pi}_{\mathrm{QAP}} = \arg\max_{\Pi \in \mathfrak{S}_n} \|(A^{1/2})^\top \Pi^\top B^{1/2}\|_{\mathrm{F}} = \arg\max_{\Pi \in \mathfrak{S}_n} \max_{\|Q\|_{\mathrm{F}} \le 1} \langle B^{1/2}, \Pi A^{1/2} Q\rangle. \tag{21}$$

In contrast, the dual variable Q in (7) is constrained to be an orthogonal matrix, which, as discussed in the proof sketch in Section 2.1, is crucial for the proof of Theorem 1. Overall, the above evidence points to the potential suboptimality of QAP in low and moderate-dimensional regime of $d \lesssim \log n$ and its potential optimality in the high-dimensional regime of $d \gg \log n$.

Practical algorithms As demonstrated by extensive numerical experiments in (Fan et al., 2019a, Sec. 4.2), for correlated random graph models with iid pairs of edge weights, the Umeyama algorithm (18) significantly improves over classical "low-rank" spectral methods involving only the top few eigenvectors, but still lags behind the more recent spectral methods such as the GRAMPA algorithm (19) that uses all pairs of eigenvalues and eigenvectors. Surprisingly, in the low-dimensional dot product model with $d = o(\log n)$, while the GRAMPA algorithm is expected to perform poorly, empirical result in Section 3 indicates that the Umeyama method actually works very well in this setting. In fact, it is not hard to show that the Umeyama algorithm returns the true permutation with high probability in the noiseless case of $\sigma = 0$; however, understanding its theoretical performance in the noisy setting remains open.

Random geometric and intersection graphs As mentioned in Section 1.1, one of the motivations for studying complete weighted graphs is to understand random geometric graphs (RGG) which provides more realistic models than Erdős-Rényi graphs for datasets with geometric structures. Note that the necessary conditions in Theorem 3 proved for the linear assignment model also hold for RGG. For the positive side, one conceivable approach is to first estimate the latent positions X, Y up to rotation and translation using multidimensional scaling algorithms (such as MDS-MAP Oh et al. (2010)) based on the two correlated RGGs, and apply the AML estimator in (4) to the estimated latent positions $\widehat{X}$ and $\widehat{Y}$ to reconstruct the latent matching. Establishing performance guarantee of this procedure is an interesting open problem. Another interesting variation is to consider the Bernoulli latent positions, which includes the random intersection graph model Singer (1996) as a special case.

References

Yonathan Aflalo, Alexander Bronstein, and Ron Kimmel. On convex relaxation of graph isomorphism. *Proceedings of the National Academy of Sciences*, 112(10):2942–2947, 2015.

Noga Alon and Joel H. Spencer. *The Probabilistic Method*. Wiley-Interscience Series in Discrete Mathematics and Optimization, 3 edition, 2008.

Ayser Armiti and Michael Gertz. Geometric graph matching and similarity: A probabilistic approach. In *Proceedings of the 26th International Conference on Scientific and Statistical Database Management*, pages 1–12, 2014.

Avanti Athreya, Donniell E Fishkind, Minh Tang, Carey E Priebe, Youngser Park, Joshua T Vogelstein, Keith Levin, Vince Lyzinski, and Yichen Qin. Statistical inference on random dot product graphs: a survey. *The Journal of Machine Learning Research*, 18(1):8393–8484, 2017.

László Babai, Paul Erdös, and Stanley M Selkow. Random graph isomorphism. *SIAM Journal on computing*, 9(3):628–635, 1980.

Boaz Barak, Chi-Ning Chou, Zhixian Lei, Tselil Schramm, and Yueqi Sheng. (nearly) efficient algorithms for the graph matching problem on correlated random graphs. In *Advances in Neural Information Processing Systems*, pages 9186–9194, 2019.

Ingwer Borg and Patrick JF Groenen. *Modern multidimensional scaling: Theory and applications*. Springer Science & Business Media, 2005.

Sébastien Bubeck and Shirshendu Ganguly. Entropic clt and phase transition in high-dimensional wishart matrices. *International Mathematics Research Notices*, 2018(2):588–606, 2018.

Sébastien Bubeck, Jian Ding, Ronen Eldan, and Miklós Z Rácz. Testing for high-dimensional geometry in random graphs. *Random Structures & Algorithms*, 49(3):503–532, 2016.

M. Chertkov, L. Kroc, F. Krzakala, M. Vergassola, and L. Zdeborová. Inference in particle tracking experiments by passing messages between images. *PNAS*, 107(17):7663–7668, 2010. ISSN 0027-8424. doi: 10.1073/pnas.0910994107.

Olivier Collier and Arnak S Dalalyan. Minimax rates in permutation estimation for feature matching. *The Journal of Machine Learning Research*, 17(1):162–192, 2016.

Daniel Cullina and Negar Kiyavash. Improved achievability and converse bounds for Erdös-Rényi graph matching. In *Proceedings of the 2016 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Science*, pages 63–72. ACM, 2016.

Daniel Cullina and Negar Kiyavash. Exact alignment recovery for correlated Erdös-Rényi graphs. *arXiv preprint arXiv:1711.06783*, 2017.

Osman E Dai, Daniel Cullina, and Negar Kiyavash. Database alignment with Gaussian features. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3225–3233. PMLR, 2019.

Osman Emre Dai, Daniel Cullina, and Negar Kiyavash. Achievability of nearly-exact alignment for correlated gaussian databases. In *2020 IEEE International Symposium on Information Theory (ISIT)*, pages 1230–1235. IEEE, 2020.

Jian Ding, Zongming Ma, Yihong Wu, and Jiaming Xu. Efficient random graph matching via degree profiles. *Probability Theory and Related Fields*, 179(1):29–115, 2021a.

Jian Ding, Yihong Wu, Jiaming Xu, and Dana Yang. The planted matching problem: Sharp threshold and infinite-order phase transition. *arXiv preprint arXiv:2103.09383*, 2021b.

Nadav Dym and Yaron Lipman. Exact recovery with symmetries for procrustes matching. *SIAM Journal on Optimization*, 27(3):1513–1530, 2017.

Nadav Dym, Haggai Maron, and Yaron Lipman. DS++: a flexible, scalable and provably tight relaxation for matching problems. *ACM Transactions on Graphics (TOG)*, 36(6):184, 2017.

Zhou Fan, Cheng Mao, Yihong Wu, and Jiaming Xu. Spectral graph matching and regularized quadratic relaxations I: The Gaussian model. *to appear in* Foundations of Computational Mathematics, 2019a. arxiv preprint arXiv:1907.08880.

Zhou Fan, Cheng Mao, Yihong Wu, and Jiaming Xu. Spectral graph matching and regularized quadratic relaxations II: Erdős-Rényi graphs and universality. *to appear in* Foundations of Computational Mathematics, 2019b. arxiv preprint arXiv:1907.08883.

Philippe Flajolet and Robert Sedgewick. *Analytic combinatorics*. Cambridge University Press, 2009.

Luca Ganassali. Sharp threshold for alignment of graph databases with gaussian weights. In *MSML21 (Mathematical and Scientific Machine Learning)*, 2021.

Luca Ganassali and Laurent Massoulié. From tree matching to sparse graph alignment. In *Conference on Learning Theory*, pages 1633–1665. PMLR, 2020.

Luca Ganassali, Laurent Massoulié, and Marc Lelarge. Correlation detection in trees for planted graph alignment. In *13th Innovations in Theoretical Computer Science Conference (ITCS 2022)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik, 2022.

Edward N Gilbert. Random plane networks. *Journal of the society for industrial and applied mathematics*, 9(4):533–543, 1961.

Edouard Grave, Armand Joulin, and Quentin Berthet. Unsupervised alignment of embeddings with wasserstein procrustes. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1880–1890. PMLR, 2019.

B. Hajek, Y. Wu, and J. Xu. Information limits for recovering a hidden community. *IEEE Trans. on Information Theory*, 63(8):4729 – 4745, 2017.

Georgina Hall and Laurent Massoulié. Partial recovery in the graph alignment problem. *arXiv preprint arXiv:2007.00533*, 2020.

Tiefeng Jiang and Danning Li. Approximation of rectangular beta-laguerre ensembles and large deviations. *Journal of Theoretical Probability*, 28(3):804–847, 2015.

Dmitriy Kunisky and Jonathan Niles-Weed. Strong recovery of geometric planted matchings. In *Proceedings of the 2022 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 834–876. SIAM, 2022.

Z Lähner, Emanuele Rodolà, MM Bronstein, Daniel Cremers, Oliver Burghard, Luca Cosmo, Andreas Dieckmann, Reinhard Klein, and Y Sahillioglu. SHREC'16: Matching of deformable shapes with topological noise. *Proc. 3DOR*, 2(10.2312), 2016.

Vince Lyzinski, Donniell Fishkind, Marcelo Fiori, Joshua Vogelstein, Carey Priebe, and Guillermo Sapiro. Graph matching: Relax at your own risk. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 38(1):60–73, 2016.

Cheng Mao, Mark Rudelson, and Konstantin Tikhomirov. Exact matching of random graphs with constant correlation. *arXiv preprint arXiv:2110.05000*, 2021a.

Cheng Mao, Mark Rudelson, and Konstantin Tikhomirov. Random graph matching with improved noise robustness. In *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pages 3296–3329, 2021b.

Haggai Maron, Nadav Dym, Itay Kezurer, Shahar Kovalsky, and Yaron Lipman. Point registration via efficient convex relaxation. *ACM Transactions on Graphics (TOG)*, 35(4):1–12, 2016.

Sho Matsumoto. Weingarten calculus for matrix ensembles associated with compact symmetric spaces. *Random Matrices: Theory and Applications*, 2(02):1350001, 2013.

Mehrdad Moharrami, Cristopher Moore, and Jiaming Xu. The planted matching problem: Phase transitions and exact results. *The Annals of Applied Probability*, 31(6):2663–2720, 2021.

Sewoong Oh, Andrea Montanari, and Amin Karbasi. Sensor network localization from local connectivity: Performance analysis for the MDS-MAP algorithm. In *2010 IEEE Information Theory Workshop on Information Theory (ITW 2010, Cairo)*, pages 1–5. IEEE, 2010.

Pedram Pedarsani and Matthias Grossglauser. On the privacy of anonymized networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1235–1243. ACM, 2011.

Mathew Penrose. *Random geometric graphs*, volume 5. OUP Oxford, 2003.

G. Pisier. *The volume of convex bodies and Banach space geometry*. Cambridge University Press, 1999.

Anand Rangarajan, Haili Chui, and Fred L Bookstein. The softassign procrustes matching algorithm. In *Biennial International Conference on Information Processing in Medical Imaging*, pages 29–42. Springer, 1997.

Mark Rudelson and Roman Vershynin. Hanson-Wright inequality and sub-Gaussian concentration. *Electron. Commun. Probab.*, 18:no. 82, 9, 2013. ISSN 1083-589X. doi: 10.1214/ECP.v18-2865. URL https://doi.org/10.1214/ECP.v18-2865.

Peter H Schönemann. A generalized solution of the orthogonal procrustes problem. *Psychometrika*, 31(1):1–10, 1966.

Guilhem Semerjian, Gabriele Sicuro, and Lenka Zdeborová. Recovery thresholds in the sparse planted matching problem. *Physical Review E*, 102(2):022304, 2020.

Yi Shang, Wheeler Ruml, Ying Zhang, and Markus PJ Fromherz. Localization from mere connectivity. In *Proceedings of the 4th ACM international symposium on Mobile ad hoc networking & computing*, pages 201–212, 2003.

Karen B Singer. *Random intersection graphs*. PhD thesis, The Johns Hopkins University, 1996.

Shinji Umeyama. An eigendecomposition approach to weighted graph matching problems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 10(5):695–703, 1988.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018. doi: 10.1017/9781108231596.

Joshua T Vogelstein, John M Conroy, Vince Lyzinski, Louis J Podrazik, Steven G Kratzer, Eric T Harley, Donniell E Fishkind, R Jacob Vogelstein, and Carey E Priebe. Fast approximate quadratic programming for graph matching. *PLOS one*, 10(4):e0121002, 2015.

Yihong Wu, Jiaming Xu, and Sophie H. Yu. Settling the sharp reconstruction thresholds of random graph matching. *arXiv preprint2102.00082*, 2021.

Mikhail Zaslavskiy, Francis Bach, and Jean-Philippe Vert. A path following algorithm for the graph matching problem. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(12): 2227–2242, 2008.

Appendix A. Further related work

The present paper bridges several streams of literature such as planted matching, feature matching, Procrustes matching, and graph matching, which we describe below.

Planted matching and feature matching The planted matching problem aims to recover a perfect matching hidden in a weighted complete $n \times n$ bipartite graph, where the edge weights are independently drawn from either $\mathcal{P}$ or $\mathcal{Q}$ depending on whether edges are on the hidden matching or not. Originally proposed by Chertkov et al. (2010) to model the application of object tracking, a sharp phase transition from almost perfect recovery to partial recovery is conjectured to exist for the special case where $\mathcal{P}$ is a folded Gaussian and $\mathcal{Q}$ is a uniform distribution over $[0, n]$. A recent line of work initiated by Moharrami et al. (2021) and followed by Semerjian et al. (2020); Ding et al. (2021b) has successfully resolved the conjecture and characterized the sharp threshold for general distributions.

Despite these fascinating advances, they crucially rely on the independent weight assumption which does not account for the latent geometry in the object tracking applications. As a remedy, the linear assignment model (1) was proposed and studied by Kunisky and Niles-Weed (2022) as a geometric model for planted matching, where the edge weights are pairwise inner products and no longer independent. In the low-dimensional setting of $d = o(\log n)$, the MLE is shown to achieve perfect recovery when $\sigma = o(n^{-2/d})$ and almost perfect recovery when $\sigma = o(n^{-1/d})$. Further bounds on the number of errors made by MLE and recovery guarantees in the high-dimensional setting are provided. However, the necessary conditions derived in Kunisky and Niles-Weed (2022) only pertain to the MLE, leaving open the possibility that almost perfect recovery might be attained by other algorithms at lower threshold. This is resolved in the negative by the information-theoretic converse in Theorem 3, showing that $\sigma = o(n^{-1/d})$ is necessary for any algorithm to achieve almost perfect recovery. Along the way, we also slightly improve the necessary condition for perfect recovery from $\sigma = O(n^{-2/d})$ to $\sigma = o(n^{-2/d})$.

The linear assignment model (1) was in fact studied earlier in Dai et al. (2019, 2020) in a different context of feature matching, where X_i's and Y_i's are viewed as two correlated Gaussian feature vectors in $\mathbb{R}^d$ and the goal is to find their best alignment. It is shown in Dai et al. (2019) that perfect recovery is possible when $\frac{d}{4}\log\left(1+\sigma^{-2}\right) - \log n \to +\infty$, and impossible when $\frac{d}{4}\log\left(1+\sigma^{-2}\right) \le (1-\Omega(1))\log n$ and $1 \ll d = O(\log n)$.[4] In comparison, the necessary condition in Theorem 19 is tighter and holds for any d, agreeing with their sufficient condition within an additive $\log d$ factor. It is also shown in Dai et al. (2020) that almost perfect recovery is possible when $\frac{d}{2}\log\left(1+\sigma^{-2}\right) \ge (1+\epsilon)\log n$ in the high-dimensional regime $d = \omega(\log n)$ for a small constant $\epsilon > 0$. This matches our necessary condition in Proposition 13 with a sharp constant.

Related problems on feature matching were also studied in the statistics literature. For example, Collier and Dalalyan (2016) studies the model of observing $Y = X + \sigma Z$ and $Y' = \Pi^* X + \sigma Z'$, where Z, Z' are two independent random Gaussian matrices and X is deterministic. The minimum separation (in Euclidean distance) of rows of X needed for perfect recovery, denoted by κ, is shown to be on the order of $\sigma \max\{(\log n)^{1/2}, (d\log n)^{1/4}\}$. Note that in the low-dimensional regime $d = o(\log n)$, this condition is comparable to our threshold for perfect recovery $\sigma = o(n^{-2/d})$, as the typical value of κ scales as $n^{-2/d}$ when X is Gaussian. However, the average-case setup is more challenging as κ can be atypically small due to the stochastic variation of X.

Procrustes matching Our dot-product model is also closely related to the problem of Procrustes matching, which finds numerous applications in natural language processing and computer vision Rangarajan et al. (1997); Maron et al. (2016); Dym and Lipman (2017); Grave et al. (2019). Given two point clouds stacked as rows of X and Y, Procrustes matching aims to find an orthogonal matrix $Q \in O(d)$ and a permutation $\Pi \in \mathfrak{S}_n$ that minimizes the Euclidean distance between the point clouds, i.e., $\min_{\Pi\in\mathfrak{S}_n} \min_{Q\in O(d)} \|YQ - \Pi X\|_{\mathrm{F}}^2$. As observed in Grave et al. (2019), this is equivalent to $\max_{\Pi\in\mathfrak{S}_n} \max_{Q\in O(d)} \langle YQ, \Pi X\rangle$, which further reduces to $\max_{\Pi\in\mathfrak{S}_n} \|X^\top \Pi^\top Y\|_*$. Thus our approximate MLE (4) under the dot-product model is equivalent to Procrustes matching on $A^{1/2}$ and $B^{1/2}$. A semi-definite programming relaxation is proposed in Maron et al. (2016) and further shown to return the optimal solution in the noiseless case when X is generic and asymmetric Maron et al. (2016); Dym and Lipman (2017). In contrast, the more recent work Grave et al. (2019) proposes an iterative algorithm based on the alternating maximization over Π and Q with an initialization provided by solving a doubly-stochastic relaxation of the QAP $\max_{\Pi\in\mathfrak{S}_n} \|X^\top \Pi^\top Y\|_{\mathrm{F}}^2$. Its performance is empirically evaluated on real datasets, but no theoretical performance guarantee is provided. Since the dot-product model is equivalent to the statistical model for Procrustes matching, where $Y = \Pi^* X Q + \sigma Z$ for a random permutation Π^* and orthogonal matrix Q, our results in Theorem 1 and Theorem 3 thus characterize the statistical limits of Procrustes matching.

Graph matching There has been a recent surge of interest in understanding the information-theoretic and algorithmic limits of random graph matching Cullina and Kiyavash (2016, 2017); Hall and Massoulié (2020); Wu et al. (2021); Ding et al. (2021a); Barak et al. (2019); Fan et al. (2019a,b); Ganassali and Massoulié (2020); Ganassali et al. (2022); Mao et al. (2021b,a), which is an average-case model for the QAP and a noisy version of random graph isomorphism Babai et al. (1980). Most of the existing work is restricted to the correlated Erdős-Rényi-type models in

4. While the impossibility result in (Dai et al., 2019, Theorem 2) only states the assumption that $d \gg 1$, its proof, specifically the proof of (Dai et al., 2019, Lemma 4.5), implicitly assumes $\sigma = O(1)$ which further implies $d = O(\log n)$.

which $\left(A_{\pi^*(i)\pi^*(j)}, B_{ij}\right)$ are iid pairs of two correlated Bernoulli or Gaussian random variables. In this case, the maximum likelihood estimator reduces to solving the QAP (20). Sharp information-theoretic limits are derived by analyzing this QAP Cullina and Kiyavash (2016, 2017); Ganassali (2021); Wu et al. (2021) and various efficient algorithms are developed based on its spectral or convex relaxations Umeyama (1988); Zaslavskiy et al. (2008); Aflalo et al. (2015); Vogelstein et al. (2015); Lyzinski et al. (2016); Dym et al. (2017); Fan et al. (2019a,b). However, as discussed in Section 4, for geometric models such as the dot-product model, the QAP is the high-noise approximation of the MLE (3), which differs from the low-noise approximation (3) that is shown to be optimal in the low-dimensional regime of $d = o(\log n)$. This observation suggests that for geometric models one may need to rethink the algorithm design and move beyond the QAP-inspired methods.

Appendix B. Maximal likelihood estimator in the dot-product model

To compute the "likelihood" of the observation (A, B) given the ground truth Π^*, it is useful to keep in mind of the graphical model

$$\begin{array}{ccccc} \Pi^* & \longrightarrow & Y & \longrightarrow & B \\ & & \uparrow & & \\ & & X & \longrightarrow & A \end{array}$$

where X, Y, Π^* is related via (1), $A = XX^\top$, and $B = YY^\top$.

Note that A are B are rank-deficient. To compute the density of (A, B) conditioned on Π^* meaningfully, one needs to choose an appropriate reference measure μ and evaluate the relative density $\frac{\mathrm{d}P_{A,B|\Pi^*}}{\mathrm{d}\mu}$. Let us choose μ to be the product of the marginal distributions of A and B, which does not depend on Π^*. For any rank-d positive semidefinite matrices A_0 and B_0, define $A_0^{1/2} \triangleq U_0\Lambda^{1/2}$ and $B_0^{1/2} \triangleq V_0 D^{1/2}$ based on the SVD $A_0 = U_0\Lambda_0^{1/2}Q_0^\top$ and $B_0 = V_0 D_0 O_0^\top$, where $Q_0, O_0 \in O(d)$ and $U_0, V_0 \in V_{n,d} \triangleq \{U \in \mathbb{R}^{n\times d} : U^\top U = I_d\}$ (the Stiefel manifold). We aim to show

$$\frac{\mathrm{d}P_{A,B|\Pi^*}(A_0, B_0|\Pi)}{\mathrm{d}\mu(A_0, B_0)} = h(A_0, B_0)\int_{O(d)} \mathrm{d}Q \exp\left(\frac{\langle B_0^{1/2}, \Pi A_0^{1/2} Q\rangle}{\sigma^2}\right) \tag{22}$$

for some fixed function h, where the integral is with respect to the Haar measure on $O(d)$. This justifies the MLE in (3) for the dot-product model.

To show (22), denote by $N_\delta(U_0) = \{U \in V_{n,d} : \|U - U_0\|_{\mathrm{F}} \le \delta\}$ and $N_\delta(\Lambda_0) = \{\Lambda \text{ diagonal} : \|\Lambda - \Lambda_0\|_{\ell_\infty} \le \delta\}$ neighborhoods of U_0 and Λ_0 respectively. (Their specific definitions are not crucial.) Consider a δ-neighborhood of A_0 of the following form:

$$N_\delta(A_0) \triangleq \{U\Lambda U^\top : U \in N_\delta(U_0), \Lambda \in N_\delta(\Lambda_0)\}$$

and similarly define $N_\delta(B_0)$. Write the SVD for X as $X = URQ^\top$, where $U \in V_{n,d}, Q \in O(d)$ and the diagonal matrix R are mutually independent; in particular, Q is uniformly distributed over

$O(d)$. Then for constant $C = C(n, d, \sigma)$,

$$\begin{aligned}
&\mathbb{P}[A \in N_\delta(A_0), B \in N_\delta(B_0) | \Pi^* = \Pi] \\
&= \mathbb{E}[\mathbf{1}_{\{XX^\top \in N_\delta(A_0)\}} \mathbf{1}_{\{YY^\top \in N_\delta(B_0)\}} | \Pi^* = \Pi] \\
&= \mathbb{E}[\mathbf{1}_{\{U \in N_\delta(U_0)\}} \mathbf{1}_{\left\{R \in N_\delta(D_0^{1/2})\right\}} \mathbf{1}_{\{YY^\top \in N_\delta(B_0)\}} | \Pi^* = \Pi] \\
&= C \cdot \mathbb{E}\left[\mathbf{1}_{\{U \in N_\delta(U_0)\}} \mathbf{1}_{\left\{R \in N_\delta(D_0^{1/2})\right\}} \int_{\mathbb{R}^{n\times d}} \mathrm{d}y \mathbf{1}_{\{yy^\top \in N_\delta(B_0)\}} \exp\left(-\frac{\|y - \Pi U R Q^\top\|_{\mathrm{F}}^2}{2\sigma^2}\right)\right] \\
&= C \cdot \mathbb{E}\left[\mathbf{1}_{\{U \in N_\delta(U_0)\}} \mathbf{1}_{\left\{R \in N_\delta(D_0^{1/2})\right\}} \int_{\mathbb{R}^{n\times d}} \mathrm{d}y \mathbf{1}_{\{yy^\top \in N_\delta(B_0)\}} \exp\left(-\frac{\|y\|_{\mathrm{F}}^2 + \|R\|_{\mathrm{F}}^2}{2\sigma^2}\right) F(y, \Pi U R)\right],
\end{aligned}$$

where $F : \mathbb{R}^{n\times d} \times \mathbb{R}^{n\times d} \to \mathbb{R}_+$ is defined by

$$F(y, x) \triangleq \mathbb{E}_Q\left[\exp\left(\frac{\langle y, xQ^\top\rangle}{\sigma^2}\right)\right] = \int_{O(d)} \mathrm{d}Q \exp\left(\frac{\langle y, xQ^\top\rangle}{\sigma^2}\right).$$

Note that this function is continuous, strictly positive, and right-invariant, in the sense that $F(YO, XO') = F(Y, X)$ for any $O, O' \in O(d)$. Thus, as $\delta \to 0$, we have for some constant $C' = C'(n, d, \sigma)$,

$$\begin{aligned}
&\mathbb{P}[A \in N_\delta(A_0), B \in N_\delta(B_0) | \Pi^* = \Pi] \\
&= (1 + o(1)) \underbrace{C' \exp\left(\frac{\mathrm{Tr}(A_0)}{2\sigma^2} - \frac{\mathrm{Tr}(B_0)}{2\sigma^2(\sigma^2+1)}\right)}_{\triangleq h(A_0, B_0)} F(B_0^{1/2}, \Pi A_0^{1/2}) \\
&\quad \cdot \underbrace{\mathbb{E}\left[\mathbf{1}_{\{U \in N_\delta(U_0)\}} \mathbf{1}_{\left\{R \in N_\delta(D_0^{1/2})\right\}}\right] \cdot (2\pi(1+\sigma^2))^{-nd/2} \int_{\mathbb{R}^{n\times d}} \mathrm{d}y \mathbf{1}_{\{yy^\top \in N_\delta(B_0)\}} \exp\left(-\frac{\|y\|_{\mathrm{F}}^2}{2(1+\sigma^2)}\right)}_{\mu[A \in N_\delta(A_0), B \in N_\delta(B_0)]},
\end{aligned}$$

proving (22).

Appendix C. Analysis of approximate maximum likelihood

In this section we prove Theorem 1 for the dot product model. The proof of Theorem 2 for the distance model follows the same program and is postponed to Appendix D.

C.1. Discretization of orthogonal group

We first prove Lemma 4 on the approximation of nuclear norm on a discretization of $O(d)$.

Proof [Proof of Lemma 4] Consider the singular value decomposition $A = UDV^\top$, where $U, V \in O(d)$ and D is diagonal. Then the nuclear norm $\|A\|_* = \max_{Q \in O(d)} \langle A, Q\rangle = \mathrm{Tr}(D)$ is attained at $Q_* = UV^\top$. Pick an element $Q \in N$ with $Q = Q_* + \Delta$, where $\|\Delta\| \le \delta$. By orthogonality of Q and Q_*, we have

$$\Delta Q_*^\top + Q_* \Delta^\top + \Delta\Delta^\top = 0. \tag{23}$$

Note that

$$AQ_*^\top = Q_* A^\top = UDU^\top =: B. \tag{24}$$

Also, we have

$$\langle A, \Delta\rangle = \left\langle AQ_*^\top, \Delta Q_*^\top \right\rangle, \quad \langle A, \Delta\rangle = \left\langle A^\top, \Delta^\top \right\rangle = \left\langle Q_* A^\top, Q_* \Delta^\top \right\rangle.$$

Adding the above equations and applying (23)-(24) yield

$$\langle A, \Delta\rangle = \frac{1}{2}\left\langle B, \Delta Q_*^\top + Q_*\Delta^\top \right\rangle = -\frac{1}{2}\left\langle B, \Delta\Delta^\top \right\rangle.$$

This implies

$$|\langle A, \Delta\rangle| \le \frac{1}{2}\|B\|_*\|\Delta\|^2 = \frac{1}{2}\|A\|_*\|\Delta\|^2,$$

which completes the proof. ■

Next we give a specific construction of a δ-net for $O(d)$ that is suitable for the purpose of proving Theorem 1. Since orthogomal matrices are normal, by the spectral decomposition theorem, each orthogonal matrix $Q \in O(d)$ can be written as $Q = U^*\Lambda U$, where $\Lambda = \mathsf{diag}(e^{\mathrm{i}\theta_1}, \ldots, e^{\mathrm{i}\theta_d})$ with $\theta_j \in [-\pi, \pi]$ for all $j = 1, \ldots, d$ and $U \in U(d)$ is an unitary matrix. To construct a net for $O(d)$, we first discretize the eigenvalues uniformly and then discretize the eigenvectors according to the optimal local entropy of orthogonal matrices with prescribed eigenvalues.

For any fixed $\delta > 0$, let $\Theta \triangleq \{\theta_k = \frac{k\delta}{4} : k = \lfloor -\frac{4\pi}{\delta}\rfloor, \lfloor -\frac{4\pi}{\delta}\rfloor + 1, \ldots, \lceil \frac{4\pi}{\delta}\rceil\}$. Then the set

$$\mathbf{\Lambda} \triangleq \left\{(\lambda_1, \ldots, \lambda_d) \in \mathbb{C}^d : \lambda_j = e^{\mathrm{i}\theta_j}, \theta_j \in \Theta, j = 1, \ldots, d\right\}$$

is a $\frac{\delta}{4}$-net in ℓ_∞ norm for the set of all possible spectrum $\{(\lambda_1, \ldots, \lambda_d) \in \mathbb{C}^d : |\lambda_j| = 1\}$. For each $(\lambda_1, \ldots, \lambda_d) \in \mathbb{C}^d$, let $O(\lambda_1, \ldots, \lambda_d)$ denote the set of orthogonal matrices with a prescribed spectrum $\{\lambda_j\}_{j=1}^d$, i.e.

$$O(\lambda_1, \ldots, \lambda_d) \triangleq \{O \in O(d) : \lambda_i(O) = \lambda_i, i = 1, \ldots, d\},$$

where $\lambda_i(O)$'s are the eigenvalues of O sorted in the counterclockwise way from $-\pi$ to π. Similarly, define $U(\lambda_1, \ldots, \lambda_d)$ to be the set of unitary matrices with a given spectrum

$$U(\lambda_1, \ldots, \lambda_d) \triangleq \{U^*\mathsf{diag}(\lambda_1, \ldots, \lambda_d)U : U \in U(d)\}.$$

Then $O(\lambda_1, \ldots, \lambda_d) \subset U(\lambda_1, \ldots, \lambda_d) \subset U(d)$. Let $N'(\lambda_1, \ldots, \lambda_d)$ be the optimal $\frac{\delta}{4}$-net in operator norm for $U(\lambda_1, \ldots, \lambda_d)$, and let $N(\lambda_1, \ldots, \lambda_d)$ be the projection (with respect to $\|\cdot\|_{\mathrm{op}}$) of $N'(\lambda_1, \ldots, \lambda_d)$ to $O(d)$. Define

$$N \triangleq \bigcup_{(\lambda_1, \ldots, \lambda_d) \in \mathbf{\Lambda}} N(\lambda_1, \ldots, \lambda_d). \tag{25}$$

We claim that N is a δ-net in operator norm for the orthogonal group.

Lemma 5 *The set $N \subset O(d)$ defined in (25) is a δ-net in operator norm for $O(d)$.*

Proof Given $Q \in O(d)$, let its eigenvalue decomposition be $Q = U^*\Lambda U$. where $\Lambda = \mathsf{diag}(\lambda_1, \dots, \lambda_d)$. Then there exists $\widetilde{\Lambda} = \mathsf{diag}(\widetilde{\lambda}_1, \dots, \widetilde{\lambda}_d)$ where $(\widetilde{\lambda}_1, \dots, \widetilde{\lambda}_d) \in \mathbf{\Lambda}$, such that $\|\Lambda - \widetilde{\Lambda}\| \le \frac{\delta}{4}$. By definition, there exists $\widetilde{U} \in U(d)$ such that $\widetilde{U}^*\widetilde{\Lambda}\widetilde{U} \in N'(\widetilde{\lambda}_1, \dots, \widetilde{\lambda}_d)$ and $\|\widetilde{U}^*\widetilde{\Lambda}\widetilde{U} - U^*\widetilde{\Lambda}U\| \le \frac{\delta}{4}$. Let $\widetilde{Q} \in N$ denote the projection of $\widetilde{U}^*\widetilde{\Lambda}\widetilde{U}$. Then

$$\begin{aligned}\|Q - \widetilde{Q}\| &\le \|Q - \widetilde{U}^*\widetilde{\Lambda}\widetilde{U}\| + \|\widetilde{U}^*\widetilde{\Lambda}\widetilde{U} - \widetilde{Q}\| \\ &\le 2\|\widetilde{U}^*\widetilde{\Lambda}\widetilde{U} - Q\| \\ &= 2\|\widetilde{U}^*\widetilde{\Lambda}\widetilde{U} - U^*\Lambda U\| \\ &\le 2(\|\widetilde{U}^*\widetilde{\Lambda}\widetilde{U} - U^*\widetilde{\Lambda}U\| + \|U^*(\widetilde{\Lambda} - \Lambda)U\|) \le \delta,\end{aligned}$$

where the second inequality follows from projection. ∎

The size of this δ-net is estimated in the following lemma.

Lemma 6 (Local entropy of $O(d)$) *For each $(\lambda_1, \dots, \lambda_d)$ where $\lambda_\ell = e^{\mathrm{i}\theta_\ell}$, we have*

$$|N(\lambda_1, \dots, \lambda_d)| \le \left(1 + \frac{2\max|\theta_\ell|}{\delta}\right)^{2d^2} \tag{26}$$

Proof Note that

$$U(\lambda_1, \dots, \lambda_d) = I + \{U^*\mathsf{diag}\,(\lambda_1 - 1, \dots, \lambda_d - 1)\,U : U \in U(d)\} =: I + \widetilde{U}(\lambda_1, \dots, \lambda_d).$$

For any matrix $Q \in \widetilde{U}(\lambda_1, \dots, \lambda_d)$, we have

$$\|Q\|_{\mathrm{op}}^2 = \max\left|e^{\mathrm{i}\theta_\ell} - 1\right|^2 = \max|2 - 2\cos\theta_\ell| \le \max|\theta_\ell|^2.$$

where $\|\cdot\|_{\mathrm{op}}$ is the the operator norm with respect to $\mathbb{C}^d \to \mathbb{C}^d$. This implies

$$U(\lambda_1, \dots, \lambda_d) \subset \mathrm{B}(I, \max|\theta_\ell|),$$

where $\mathrm{B}(I, r)$ is the operator norm ball centered at I_d with radius r. As a normed vector space over $\mathbb{R}$, the space of $d \times d$ complex matrices has dimension $2d^2$ since $\mathbb{C}^{d\times d} \simeq \mathbb{R}^{2d^2}$. Then the desired result follows from a standard volume bound (c.f. e.g. (Pisier, 1999, Lemma 4.10)) for the metric entropy

$$|N(\lambda_1, \dots, \lambda_d)| \le \left|N'(\lambda_1, \dots, \lambda_d)\right| \le \left(1 + \frac{2\max|\theta_\ell|}{\delta}\right)^{2d^2}.$$

∎

C.2. Moment generating functions and cycle decomposition

This subsection is devoted to estimating the following quantity

$$\sum_{\Pi \ne I_n} \sum_{(\lambda_1,\dots,\lambda_d)\in\mathbf{\Lambda}} \sum_{Q\in N(\lambda_1,\dots,\lambda_d)} p(\Pi, Q),$$

where

$$p(\Pi, Q) \triangleq \mathbb{E} \exp\left\{-\frac{1}{32\sigma^2} \|X - \Pi X Q\|_{\mathrm{F}}^2\right\}. \tag{27}$$

As we will see later in (55), this quantity bounds the probability that our AMLE is not the ground truth. The moment generating function (MGF) $p(\Pi, Q)$ in the summation is estimated in the following lemma.

Lemma 7 *For any fixed $\Pi \in \mathfrak{S}_n$, let $\mathcal{O}$ denote the set of orbits of the permutation and n_k be the number of orbits with length k. Let $Q \in O(d)$ and denote by $e^{\mathrm{i}\theta_1}, \ldots, e^{\mathrm{i}\theta_d}$ the eigenvalues of Q, where $\theta_1, \ldots, \theta_d \in [-\pi, \pi]$. Then*

$$p(\Pi, Q) = \prod_{O \in \mathcal{O}, |O| \geq 1} a_{|O|}(Q) = \prod_{k=1}^{n} a_k(Q)^{n_k}, \tag{28}$$

where

$$a_k(Q) \triangleq (4\sigma)^{kd} \prod_{\ell=1}^{d} \left[(\sqrt{1+4\sigma^2} + 2\sigma)^{2k} + (\sqrt{1+4\sigma^2} - 2\sigma)^{2k} - 2\cos(k\theta_\ell)\right]^{-1/2}, \tag{29}$$

satisfying, for all $1 \leq k \leq n$,

$$a_k(Q) \leq a_k(I) \leq (4\sigma)^{(k-1)d}. \tag{30}$$

Furthermore,

$$a_1(Q) \leq (C\sigma)^d \prod_{\ell=1}^{d} \frac{1}{\sigma + |\theta_\ell|}, \tag{31}$$

where $C > 0$ is a universal constant independent of d, n, σ.

Proof For simplicity, denote $t = \frac{1}{32\sigma^2}$. Let $x = \mathsf{vec}(X) \in \mathbb{R}^{nd}$ be the vectorization of X, and note that $x \sim \mathcal{N}(0, I_{nd})$. Through the vectorization, we have

$$\|X - \Pi X Q\|_{\mathrm{F}}^2 = \left\|(I_{nd} - Q^\top \otimes \Pi)x\right\|^2.$$

Let $H \triangleq I_{nd} - Q^\top \otimes \Pi$, then

$$p(\Pi, Q) = \mathbb{E} \exp\left(-tx^\top H^\top H x\right) = \left[\det\left(I + 2tH^\top H\right)\right]^{-\frac{1}{2}}. \tag{32}$$

Note that the eigenvalues of H are

$$\lambda_{ij}(H) = 1 - \lambda_i(Q^\top)\lambda_j(\Pi), \;\; i = 1, \ldots, d, \; j = 1, \ldots, n.$$

This leads to

$$p(\Pi, Q) = \prod_{i=1}^{d} \prod_{j=1}^{n} \left(1 + 2t\left|1 - \lambda_i(Q^\top)\lambda_j(\Pi)\right|^2\right)^{-\frac{1}{2}}. \tag{33}$$

Through a cycle decomposition, the spectrum of Π is the same as a block diagonal matrix $\widetilde{\Pi}$ of the following form

$$\widetilde{\Pi} = \mathsf{diag}\left(P_1^{(1)}, \ldots, P_{n_1}^{(1)}, \ldots, P_1^{(k)}, \ldots, P_{n_k}^{(k)}, \ldots, P_1^{(n)}, \ldots, P_{n_n}^{(n)}\right),$$

where n_k is the number of k-cycles in π, and $P_1^{(k)} = \cdots = P_{n_k}^{(k)} = P^{(k)}$ is a $k \times k$ circulant matrix given by

$$P^{(k)} = \begin{bmatrix} 0 & 1 & \cdots & & 0 \\ 0 & 0 & 1 & & \\ \vdots & 0 & 0 & \ddots & \vdots \\ & & \ddots & \ddots & 1 \\ 1 & & \cdots & 0 & 0 \end{bmatrix}.$$

It is well known that the eigenvalues of $P^{(k)}$ are the k-th roots of unity $\{e^{\mathrm{i}\frac{2\pi}{k}j}\}_{j=0}^{k-1}$. Therefore, the spectrum of Π is the following multiset

$$\mathsf{Spec}(\Pi) = \{e^{\mathrm{i}\frac{2\pi}{k}j_k} \text{ with multiplicity } n_k : 1 \le k \le n, j_k = 0, \ldots, k-1\}. \tag{34}$$

Recall that $e^{\mathrm{i}\theta_1}, \ldots, e^{\mathrm{i}\theta_d}$ are the eigenvalues of Q. Note that the eigenvalues of $Q^\top$ are the complex conjugate of the eigenvalues of Q. Combined with (33) and (34), we have

$$\begin{aligned} p(\Pi, Q) &= \left[\prod_{\ell=1}^{d}\prod_{k=1}^{n}\prod_{j=0}^{k-1}\left(1 + 2t\left|1 - e^{-\mathrm{i}\theta_\ell}e^{\mathrm{i}\frac{2\pi}{k}j}\right|^2\right)^{n_k}\right]^{-1/2} \\ &= \prod_{k=1}^{n}\left[\prod_{\ell=1}^{d}\prod_{j=0}^{k-1}\left(1 + 4t - 4t\cos(-\theta_\ell + \tfrac{2\pi}{k}j)\right)^{-1/2}\right]^{n_k} \triangleq \prod_{k=1}^{n} a_k(Q)^{n_k}. \end{aligned} \tag{35}$$

Define

$$f(\theta) \triangleq \prod_{j=0}^{k-1}\left(1 + 4t - 4t\cos(\theta + \tfrac{2\pi}{k}j)\right),$$

To simplify $f(\theta)$, let $p = \frac{\sqrt{1+8t}+1}{2}$ and $q = \frac{\sqrt{1+8t}-1}{2}$ so that $p^2 + q^2 = 1 + 4t$ and $pq = 2t$. Thus,

$$f(\theta) = \prod_{j=0}^{k-1}\left(p^2 + q^2 - 2pq\cos\left(\tfrac{2\pi}{k}j + \theta\right)\right).$$

Note that

$$p^k - q^k e^{\mathrm{i}k\theta} = \prod_{j=0}^{k-1}\left(p - qe^{\mathrm{i}\frac{2\pi}{k}j + \mathrm{i}\theta}\right), \quad p^k - q^k e^{-\mathrm{i}k\theta} = \prod_{j=0}^{k-1}\left(p - qe^{\mathrm{i}\frac{2\pi}{k}j - \mathrm{i}\theta}\right).$$

Multiplying the above two equations gives us

$$p^{2k} + q^{2k} - 2p^k q^k \cos k\theta = \prod_{j=0}^{k-1}\left(p^2 + q^2 - 2pq\cos\left(\tfrac{2\pi}{k}j + \theta\right)\right) = f(\theta).$$

which implies

$$f(\theta) = \left(\frac{\sqrt{1+8t}+1}{2}\right)^{2k} + \left(\frac{\sqrt{1+8t}-1}{2}\right)^{2k} - 2(2t)^k \cos(k\theta)$$
$$= \left(\frac{1}{4\sigma}\right)^{2k} \left[\left(\sqrt{1+4\sigma^2}+2\sigma\right)^{2k} + \left(\sqrt{1+4\sigma^2}-2\sigma\right)^{2k} - 2\cos k\theta\right].$$

Note that $a_k(Q) = \prod_{\ell=1}^d f(-\theta_\ell)^{-1/2}$, and therefore we have shown (29). In particular,

$$a_1(Q) = (4\sigma)^d \prod_{\ell=1}^d (2 - 2\cos\theta_\ell + 16\sigma^2)^{-\frac{1}{2}}. \tag{36}$$

Since $\sin^2\theta \geq \frac{\theta^2}{4}$ for $\theta \in [-\frac{\pi}{2}, \frac{\pi}{2}]$, we have

$$\sqrt{2-2\cos\theta_\ell + 16\sigma^2} = \sqrt{4\sin^2(\theta_\ell/2) + 16\sigma^2} \geq \sqrt{2\sin^2(\theta_\ell/2)} + \sqrt{8\sigma^2}$$
$$\geq \sqrt{2(\theta_\ell/4)^2} + \sqrt{8\sigma^2} = \sqrt{2}|\theta_\ell|/4 + 2\sqrt{2}\sigma$$

Consequently, this gives us (31). In general, note that

$$(\sqrt{1+4\sigma^2}+2\sigma)^{2k} + (\sqrt{1+4\sigma^2}-2\sigma)^{2k} - 2 \geq (4k\sigma)^2,$$

which completes the proof for (30). To see this, define $g(x) = x^k - x^{-k}$ which is increasing in x. Then

$$(\sqrt{1+4\sigma^2}+2\sigma)^{2k} + (\sqrt{1+4\sigma^2}-2\sigma)^{2k} - 2 = g\left(\sqrt{1+4\sigma^2}+2\sigma\right)^2 \geq g\left(1+2\sigma\right)^2 \geq (4k\sigma)^2,$$

where the last inequality holds because $(1+a)^k - (1-a)^k \geq 2ak$ for $a \geq 0$. Finally, (28) follows from (35). ■

Based on the above representation via cycle decomposition, we have the following estimate for the moment generating function. This estimate is a key result in this paper as it is the basis of both Theorem 1 and Lemma 9.

Lemma 8 *Suppose $d = o(\log n)$. For some $\sigma_0 > 0$, let $\delta = \sigma_0/\sqrt{n}$ and $N \subset O(d)$ be the δ-net defined in (25).*

(i) If $\sigma_0 = o(n^{-2/d})$, then

$$\sum_{\Pi \neq I_n} \sum_{Q \in N} \mathbb{E} \exp\left\{-\frac{1}{32\sigma_0^2}\|X - \Pi X Q\|_{\mathrm{F}}^2\right\} = o(1). \tag{37}$$

(ii) For any $\varepsilon = \varepsilon(n) > 0$, if $\sigma_0^{-d} > 16n2^{2/\varepsilon}$, then the following is true

$$\sum_{\mathrm{d}(\pi,\mathrm{Id}) \geq \varepsilon n} \sum_{Q \in N} \mathbb{E} \exp\left\{-\frac{1}{32\sigma_0^2}\|X - \Pi X Q\|_{\mathrm{F}}^2\right\} = o(1). \tag{38}$$

Proof (i) For any fixed $\Pi \in \mathfrak{S}_n$, combining (30) and (31) yields

$$\begin{aligned}\prod_{k\geq 1} a_k(Q)^{n_k} &\leq (C\sigma_0)^{n_1 d+\sum_{k\geq 2} n_k(k-1)d}\left(\prod_{\ell=1}^{d}\frac{1}{\theta_\ell+\sigma_0}\right)^{n_1}\\ &= (C\sigma_0)^{d(n-\sum_{k\geq 2} n_k)}\left(\prod_{\ell=1}^{d}\frac{1}{\theta_\ell+\sigma_0}\right)^{n_1}\\ &\leq (C\sigma_0)^{\frac{n+n_1}{2}d}\left(\prod_{\ell=1}^{d}\frac{1}{\theta_\ell+\sigma_0}\right)^{n_1}. \end{aligned} \tag{39}$$

Note that by Lemma 6, we have

$$\left|N\left(e^{\mathrm{i}\frac{m_1\delta}{4}},\ldots,e^{\mathrm{i}\frac{m_d\delta}{4}}\right)\right| \leq \left(1+\frac{\max|m_\ell|}{2}\right)^{2d^2} \leq \left(1+\frac{\sum_{\ell=1}^{d}|m_\ell|}{2}\right)^{2d^2} \leq \prod_{\ell=1}^{d}\left(1+\frac{|m_\ell|}{2}\right)^{2d^2}. \tag{40}$$

Using Lemma 7 and (39), this leads to

$$\begin{aligned}&\sum_{\Pi\neq I_n}\sum_{Q\in N} p(\Pi,Q)\\ &\leq \sum_{n_1=0}^{n-2}\sum_{m_1,\ldots,m_d=\lfloor-\frac{4\pi}{\delta}\rfloor}^{\lceil\frac{4\pi}{\delta}\rceil}\left|N\left(e^{\mathrm{i}\frac{m_1\delta}{4}},\ldots,e^{\mathrm{i}\frac{m_d\delta}{4}}\right)\right|(n-n_1)!\binom{n}{n_1}(C\sigma_0)^{\frac{n+n_1}{2}d}\left(\prod_{\ell=1}^{d}\frac{1}{\frac{\delta|m_\ell|}{4}+\sigma_0}\right)^{n_1}\\ &\leq \sum_{n_1=0}^{n-2}(C\sigma_0)^{\frac{n+n_1}{2}d}(n-n_1)!\binom{n}{n_1}\left[\sum_{m=\lfloor-\frac{4\pi}{\delta}\rfloor}^{\lceil\frac{4\pi}{\delta}\rceil}\frac{1}{(\frac{\delta|m|}{4}+\sigma_0)^{n_1}}(1+\tfrac{|m|}{2})^{2d^2}\right]^d\\ &\leq \sum_{n_1=0}^{n-2}(C\sigma_0)^{\frac{n-n_1}{2}d}(n-n_1)!\binom{n}{n_1}\left[\sum_{m=\lfloor-\frac{4\pi}{\delta}\rfloor}^{\lceil\frac{4\pi}{\delta}\rceil}\frac{1}{(1+\frac{\delta}{4\sigma_0}|m|)^{n_1}}(1+\tfrac{|m|}{2})^{2d^2}\right]^d\\ &\leq \sum_{n_1=0}^{n-2}\left((C\sigma_0)^d n^2\right)^{\frac{n-n_1}{2}}\left[\sum_{m=\lfloor-\frac{4\pi}{\delta}\rfloor}^{\lceil\frac{4\pi}{\delta}\rceil}\frac{1}{(1+\frac{\delta}{4\sigma_0}|m|)^{n_1}}(1+\tfrac{|m|}{2})^{2d^2}\right]^d, \end{aligned}$$

where the second line follows from Lemma 6 and the fourth line follows from the fact that the number of permutations with n_1 fixed points is at most $(n-n_1)!\binom{n}{n_1} \leq n^{n-n_1}$.

Recall that $\delta = \sigma_0/\sqrt{n}$ and $\sigma_0 = o(n^{-2/d})$. For any fixed $1 \leq n_1 \leq n-2$,

$$T_{n_1} \triangleq \sum_{m=\lfloor-\frac{4\pi}{\delta}\rfloor}^{\lceil\frac{4\pi}{\delta}\rceil}\frac{1}{(1+\frac{\delta}{4\sigma_0}|m|)^{n_1}}(1+\tfrac{|m|}{2})^{2d^2} = \sum_{m=\lfloor-\frac{4\pi}{\delta}\rfloor}^{\lceil\frac{4\pi}{\delta}\rceil}\frac{1}{(1+\frac{|m|}{4\sqrt{n}})^{n_1}}(1+\tfrac{|m|}{2})^{2d^2}.$$

If $n_1 \leq \sqrt{n}$, we have

$$T_{n_1} \leq \sum_{m=\lfloor-\frac{4\pi}{\delta}\rfloor}^{\lceil\frac{4\pi}{\delta}\rceil}\left(1+\frac{|m|}{2}\right)^{2d^2} \leq \frac{8\pi}{\delta}\left(1+\frac{2\pi}{\delta}\right)^{2d^2} \leq 2\left(\frac{4\pi\sqrt{n}}{\sigma_0}\right)^{2d^2+1}. \tag{41}$$

Therefore, let $\sigma_0^{-d} = L$ and $L = n^2 K$ where $K \gg 1$, then

$$\begin{aligned}
\sum_{n_1=0}^{\sqrt{n}} \left((C\sigma_0)^d n^2\right)^{\frac{n-n_1}{2}} T_{n_1}^d &\le \sqrt{n}\left[2\left(\frac{4\pi\sqrt{n}}{\sigma_0}\right)^{2d^2+1}\right]^d \left((C\sigma_0)^d n^2\right)^{\frac{n-\sqrt{n}}{2}} \\
&\le C^{d^3} n^{2d^3} L^{2d^2+1} K^{-\frac{n-\sqrt{n}}{2}} \\
&\le C^{d^3} n^{2d^3} L^{3d^2} K^{-\frac{n}{3}} \\
&\le C^{d^3} n^{2d^3} \exp\left(3d^2 \log(n^2 K)\right) \exp\left(-\frac{n}{3}\log K\right) \\
&\le C^{d^3} \exp\left((6d^2+2d^3)\log n - \left(\frac{n}{3} - 3d^2\right)\log K\right) \\
&= o(1), \qquad (42)
\end{aligned}$$

where the last line follows from $K \gg 1$ and $d = o(\log n)$.

On the other hand, for $\sqrt{n} \le n_1 \le n-2$,we decompose it into two parts $T_{n_1} = J_1 + J_2$, where

$$\begin{aligned}
J_1 &\triangleq \sum_{|m|\le 8\sqrt{n}} \frac{1}{(1+\frac{|m|}{4\sqrt{n}})^{n_1}} (1+\tfrac{|m|}{2})^{2d^2}, \\
J_2 &\triangleq \sum_{8\sqrt{n}<|m|\le \frac{4\pi}{\delta}} \frac{1}{(1+\frac{|m|}{4\sqrt{n}})^{n_1}} (1+\tfrac{|m|}{2})^{2d^2}.
\end{aligned}$$

We first show that the contribution of J_2 is negligible. To see this, note that

$$\begin{aligned}
J_2 &\le C(4\sqrt{n})^{n_1} \sum_{m=1+8\sqrt{n}}^{4\pi/\delta} m^{-n_1+2d^2} \\
&\le C(4\sqrt{n})^{n_1} \int_{8\sqrt{n}}^{4\pi/\delta} x^{-(n_1-2d^2)} \mathrm{d}x \\
&\le C(4\sqrt{n})^{n_1} \frac{1}{n_1-2d^2-1} (8\sqrt{n})^{-n_1+2d^2+1} \\
&\le C 2^{-n_1+6d^2+3} \frac{1}{n_1-2d^2-1} n^{d^2+\frac{1}{2}} \\
&\le C 2^{-n_1/2} n^{d^2}.
\end{aligned}$$

Recall that $n_1 \ge \sqrt{n}$ and $d = o(\log n)$. Therefore we have $J_2 = o(1)$. Moreover, a simple observation is that $T_{n_1} \ge 1$. This concludes that J_2 is negligible and it suffices to bound J_1. Note that for $0 \le x \le 2$ we have $1 + x \ge e^{x/2}$. Therefore, this implies

$$J_1 \le C \sum_{m=0}^{8\sqrt{n}} \exp\left(-\left(\frac{n_1}{8\sqrt{n}} - 2d^2\right) m\right).$$

For $n_1 \ge 32\sqrt{n}(\log n)^2$, we have $\frac{n_1}{8\sqrt{n}} - 2d^2 > \frac{n_1}{16\sqrt{n}}$ since $d = o(\log n)$. Consequently, in this regime we have

$$J_1 \le C \sum_{m=0}^{8\sqrt{n}} \exp\left(-\frac{n_1}{16\sqrt{n}} m\right) \le \frac{C}{1 - e^{-\frac{n_1}{16\sqrt{n}}}} \le \frac{C}{1 - e^{-4(\log n)^2}}.$$

Thus, for $n_1 \geq 32\sqrt{n}(\log n)^2$, we have

$$T_{n_1}^d \leq (2J_1)^d \leq C^d \left(1 - e^{-4(\log n)^2}\right)^{-d} \leq C^d \exp\left(de^{-4(\log n)^2}\right) \leq C^d. \tag{43}$$

For $\sqrt{n} \leq n_1 < 32\sqrt{n}(\log n)^2$, we use a trivial bound

$$J_1 \leq C \sum_{m=0}^{8\sqrt{n}} \left(1 + \frac{m}{2}\right)^{2d^2} \leq C(8\sqrt{n})^{2d^2+1}.$$

In this case,

$$T_{n_1}^d \leq C^d (8\sqrt{n})^{2d^2+1}. \tag{44}$$

Thus, (43) and (44) together imply

$$\begin{aligned}
&\sum_{n_1=\sqrt{n}}^{n-2} \left((C\sigma_0)^d n^2\right)^{\frac{n-n_1}{2}} T_{n_1}^d \\
&\leq \sum_{n_1=\sqrt{n}}^{32\sqrt{n}(\log n)^2} \left((C\sigma_0)^d n^2\right)^{\frac{n-n_1}{2}} T_{n_1}^d + \sum_{n_1=32\sqrt{n}(\log n)^2}^{n-2} \left((C\sigma_0)^d n^2\right)^{\frac{n-n_1}{2}} T_{n_1}^d \\
&\leq 32\sqrt{n}(\log n)^2 C^{d^2} (8\sqrt{n})^{2d^3+d} 2^{-n} + C^d \sigma_0^d n^2 \\
&= o(1)
\end{aligned} \tag{45}$$

Combining (42) and (45) together, we obtain

$$\sum_{\Pi \neq I_n} \sum_{Q \in N} p(\Pi, Q) = o(1),$$

which completes the proof.

(ii) Due to the stronger noise level, we need to be more careful in (39):

$$\begin{aligned}
\prod_{j\geq 1} a_k(Q)^{n_j} &\leq (C\sigma_0)^{n_1 d + \sum_{j\geq 2} n_j (j-1)d} \left(\prod_{\ell=1}^{d} \frac{1}{|\theta_\ell| + \sigma_0}\right)^{n_1} \\
&= (C\sigma_0)^{dn - d\sum_{j=1}^{n} n_j} \prod_{\ell=1}^{d} \frac{1}{(1 + \frac{|\theta_\ell|}{\sigma_0})^{n_1}}.
\end{aligned} \tag{46}$$

For simplicity, denote by $k \triangleq \mathrm{d}(\pi, \mathrm{Id}) = n - n_1$ the number of non-fixed points of π. Let $\widetilde{\pi}$ be the restriction of the permutation $\pi \in S_n$ on its non-fixed points, which by definition is a derangement. Denote the number of cycles of a permutation π by $\mathfrak{c}(\pi)$. An observation is that $\mathfrak{c}(\pi) = \sum_{j=1}^{n} n_j = n_1 + \mathfrak{c}(\widetilde{\pi})$. Then Lemma 7 and (46) yield

$$\begin{aligned}
&\sum_{\mathrm{d}(\pi,\mathrm{Id})\geq \varepsilon n} \sum_{Q\in N} p(\Pi, Q) \\
&\leq \sum_{k=\varepsilon n}^{n} \binom{n}{k} \sum_{\widetilde{\pi} \text{ derangement}} \sum_{m_1,\ldots,m_d=\lfloor -\frac{4\pi}{\delta} \rfloor}^{\lceil \frac{4\pi}{\delta} \rceil} \left| N\left(e^{\mathrm{i}\frac{m_1\delta}{4}}, \ldots, e^{\mathrm{i}\frac{m_d\delta}{4}}\right)\right| (C\sigma_0)^{d(k-\mathfrak{c}(\widetilde{\pi}))} \prod_{\ell=1}^{d} \frac{1}{(1 + \frac{\delta|m_\ell|}{4\sigma_0})^{n-k}}.
\end{aligned}$$

Denote $L = \sigma_0^{-d}$. Using (40) and rearranging the above inequality give us

$$\sum_{\mathrm{d}(\pi,\mathrm{Id})\geq\varepsilon n}\sum_{Q\in N} p(\Pi, Q)$$
$$\leq \sum_{k=\varepsilon n}^{n}\binom{n}{k}L^{-k}\sum_{\widetilde{\pi}\text{ derangement}} L^{\mathfrak{c}(\widetilde{\pi})}\left[\sum_{m=\lfloor-\frac{4\pi}{\delta}\rfloor}^{\lceil\frac{4\pi}{\delta}\rceil}\frac{1}{(1+\frac{\delta}{4\sigma_0}|m|)^{n-k}}(1+\tfrac{|m|}{2})^{2d^2}\right]^d. \quad (47)$$

Note that

$$\sum_{\widetilde{\pi}\text{ derangement}} L^{\mathfrak{c}(\widetilde{\pi})} = k!\,\mathbb{E}_\tau\left[L^{\mathfrak{c}(\tau)}\mathbb{1}_{\{\tau\text{ is a derangement}\}}\right],$$

where the expectation $\mathbb{E}_\tau$ is taken for a uniformly random permutation $\tau \in S_k$. To bound the above truncated generating function, recall that the generating function of $\mathfrak{c}(\tau)$ is given by (see, e.g., (Flajolet and Sedgewick, 2009, Eq. (39)))

$$\mathbb{E}_\tau[L^{\mathfrak{c}(\tau)}] = \binom{L+k-1}{k} = \frac{L(L+1)\cdots(L+k-1)}{k!}. \quad (48)$$

Pick some $\alpha \in (0,1)$ to be determined later and obtain the following

$$\mathbb{E}_\tau\left[L^{\mathfrak{c}(\tau)}\mathbb{1}_{\{\tau\text{ is a derangement}\}}\right] \leq \mathbb{E}_\tau\left[L^{\mathfrak{c}(\tau)}\mathbb{1}_{\{\mathfrak{c}(\tau)\leq k/2\}}\right]$$
$$\leq \mathbb{E}_\tau\left[L^{\alpha\mathfrak{c}(\tau)+(1-\alpha)\frac{k}{2}}\right] = L^{(1-\alpha)\frac{k}{2}}\mathbb{E}_\tau\left[L^{\alpha\mathfrak{c}(\tau)}\right] = L^{(1-\alpha)\frac{k}{2}}\binom{L^\alpha+k-1}{k}.$$

Choosing $\alpha = \frac{\log k}{\log L}$, we have

$$\mathbb{E}_\tau\left[L^{\mathfrak{c}(\tau)}\mathbb{1}_{\{\tau\text{ is a derangement}\}}\right] \leq \binom{2k-1}{k}\left(\frac{L}{k}\right)^{k/2} \leq \left(\frac{16L}{k}\right)^{k/2}. \quad (49)$$

Recall that

$$T_{n-k} = \sum_{m=\lfloor-\frac{4\pi}{\delta}\rfloor}^{\lceil\frac{4\pi}{\delta}\rceil}\frac{1}{(1+\frac{\delta}{4\sigma_0}|m|)^{n-k}}(1+\tfrac{|m|}{2})^{2d^2}.$$

For $k \leq n-\sqrt{n}$, each term T_{n-k} is bounded by (43) and (44). On the other hand, if $k \geq n-\sqrt{n}$, we control T_{n-k} via (41). Here in the case of almost perfect recovery, combined with (49), the assumption on σ_0 yields a superexponentially decaying term in the summation (47). Specifically, combined this with (47) and (49), we obtain

$$\sum_{\mathrm{d}(\pi,\mathrm{Id})\geq\varepsilon n}\sum_{Q\in N} p(\Pi, Q) \leq J_1 + J_2,$$

where

$$J_1 \triangleq C^d \sum_{k=\varepsilon n}^{n-32\sqrt{n}(\log n)^2}\binom{n}{k}L^{-k}k!\left(\frac{16L}{k}\right)^{k/2},$$
$$J_2 \triangleq C^{d^3}n^{2d^3}L^{2d^2+1}\sum_{k=n-32\sqrt{n}(\log n)^2+1}^{n}\binom{n}{k}L^{-k}k!\left(\frac{16L}{k}\right)^{k/2}.$$

Let $L = nK$ where $\frac{\varepsilon}{2}\log\frac{K}{16} > \log 2$. Recall that $d = o(\log n)$. Then applying Stirling's approximation gives us

$$J_1 \le C^d n 2^n \left(\frac{16n}{L}\right)^{\varepsilon n/2} \le C^d n \exp\left(n\log 2 - \frac{\varepsilon n}{2}\log\frac{K}{16}\right) = o(1), \tag{50}$$

and

$$\begin{aligned} J_2 &\le C^{d^3} n^{2d^3+1} L^{2d^2+1} 2^n \left(\frac{16n}{L}\right)^{n/3} \\ &\le C^{d^3} n^{2d^3+1} \exp\left[(2d^2+1)\log n + (2d^2+1)\log K + n\log 2 - \frac{n}{3}\log\frac{K}{16}\right] = o(1). \end{aligned} \tag{51}$$

Combining (50) and (51) implies

$$\sum_{\mathrm{d}(\pi,\mathrm{Id})\ge\varepsilon n}\sum_{Q\in N} p(\Pi, Q) = o(1),$$

which completes the proof. ■

The estimate of the moment generating functions results in the following lemma, which plays a crucial rule in the probability reduction estimate (55).

Lemma 9 *For some $\sigma_0 > 0$, let $\delta = \sigma_0/\sqrt{n}$ and N be the δ-net defined in (25).*

(i) If $\sigma_0 = o(n^{-2/d})$, for any constant $c > 0$, the following inequality is true with high probability

$$\min_{\Pi\neq I_n}\min_{Q\in N} \|X - \Pi X Q\|_{\mathrm{F}} \ge c\sqrt{d}\sigma_0. \tag{52}$$

(ii) For any $\varepsilon = \varepsilon(n) > 0$, if $\sigma_0^{-d} > 16n2^{2/\varepsilon}$, the following is true for any fixed constant $c > 0$ with high probability

$$\min_{\mathrm{d}(\pi,\mathrm{Id})\ge\varepsilon n}\min_{Q\in N} \|X - \Pi X Q\|_{\mathrm{F}} \ge c\sqrt{d}\sigma_0. \tag{53}$$

Proof (i) For fixed $\Pi \neq I_n$ and $Q \in N$, by the Chernoff bound, for every $t \ge 0$ we have

$$\begin{aligned} &\mathbb{P}\left\{\|X - \Pi X Q\|_{\mathrm{F}} < c\sqrt{d}\sigma_0\right\} \\ &\qquad = \mathbb{P}\left\{e^{-t\|X-\Pi XQ\|_{\mathrm{F}}^2} > e^{-tc^2 d\sigma_0^2}\right\} \le e^{tc^2 d\sigma_0^2}\mathbb{E}\exp\left(-t\|X - \Pi X Q\|_{\mathrm{F}}^2\right). \end{aligned}$$

Taking $t = \frac{1}{32\sigma_0^2}$, by the union bound we have

$$\begin{aligned} \mathbb{P}\left\{\min_{\Pi\neq I_n}\min_{Q\in N}\|X - \Pi X Q\|_{\mathrm{F}} \ge c\sqrt{d}\sigma_0\right\} &= 1 - \mathbb{P}\left\{\exists \Pi \neq I_n, \exists Q \in N \ s.t.\ \|X - \Pi X Q\|_{\mathrm{F}} \le c\sigma_0\right\} \\ &\ge 1 - e^{\frac{c^2 d}{32}}\sum_{\Pi\neq I_d}\sum_{Q\in N}\mathbb{E}\exp\left\{-\frac{1}{32\sigma_0^2}\|X - \Pi X Q\|_{\mathrm{F}}^2\right\} \\ &\ge 1 - o(1), \end{aligned}$$

where the last step follows from Lemma 8.

(ii) The arguments are similar with Part (i). Using Chernoff bound and Lemma 8, we have

$$\begin{aligned}
&\mathbb{P}\left\{\min_{\mathrm{d}(\pi,\mathrm{Id})\geq\varepsilon n}\min_{Q\in N}\|X-\Pi XQ\|_{\mathrm{F}}\geq c\sqrt{d}\sigma_0\right\}\\
\geq\ &1-e^{\frac{c^2d}{32}}\sum_{\mathrm{d}(\pi,\mathrm{Id})\geq\varepsilon n}\sum_{Q\in N}\mathbb{E}\exp\left\{-\frac{1}{32\sigma_0^2}\|X-\Pi XQ\|_{\mathrm{F}}^2\right\}\\
\geq\ &1-o(1),
\end{aligned}$$

which completes the proof. ■

C.3. Proof of Theorem 1

Proof (i) For $\sigma\ll n^{-2/d}$, let $\delta=\sigma/\sqrt{n}$ and let N be the δ-net in operator norm for $O(d)$ defined in (25). Applying Lemma 4, we have

$$\begin{aligned}
\mathbb{P}\left\{\|X^\top\Pi^\top Y\|_*\geq\|X^\top Y\|_*\right\}&\leq\mathbb{P}\left\{\max_{Q\in O(d)}\langle X^\top\Pi^\top Y,Q\rangle\geq\langle X^\top Y,I_d\rangle\right\}\\
&\leq\mathbb{P}\left\{\max_{Q\in N}\langle X^\top\Pi^\top Y,Q\rangle\geq(1-\delta^2)\langle X^\top Y,I_d\rangle\right\}.
\end{aligned}$$

For fixed Π and Q, we have

$$\begin{aligned}
&\mathbb{P}\left\{\langle X^\top\Pi^\top Y,Q\rangle\geq(1-\delta^2)\langle X^\top Y,I_d\rangle\right\}\\
&\qquad=\mathbb{P}\left\{\sigma\langle Z,(1-\delta^2)X-\Pi XQ\rangle\geq(1-\delta^2)\|X\|_{\mathrm{F}}^2-\langle X,\Pi XQ\rangle\right\}.
\end{aligned}$$

Note that we have the following observations

$$\|X\|_{\mathrm{F}}^2-\langle X,\Pi XQ\rangle=\frac{1}{2}\|X-\Pi XQ\|_{\mathrm{F}}^2,$$

and

$$\begin{aligned}
\left\|(1-\delta^2)X-\Pi XQ\right\|_{\mathrm{F}}^2&=(1-\delta^2)^2\|X\|_{\mathrm{F}}^2+\|X\|_{\mathrm{F}}^2-2(1-\delta^2)\langle X,\Pi XQ\rangle\\
&=(1-\delta^2)\|X-\Pi XQ\|_{\mathrm{F}}^2-\delta^4\|X\|_{\mathrm{F}}^2.
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\mathbb{P}\left\{\langle X^\top\Pi^\top Y,Q\rangle\geq(1-\delta^2)\langle X^\top Y,I_d\rangle\right\}\\
=&\mathbb{P}\left\{\sigma\mathcal{N}\left(0,(1-\delta^2)\|X-\Pi XQ\|_{\mathrm{F}}^2-\delta^4\|X\|_{\mathrm{F}}^2\right)\geq\frac{1}{2}\|X-\Pi XQ\|_{\mathrm{F}}^2-\delta^2\|X\|_{\mathrm{F}}^2\right\}\\
\leq&\mathbb{P}\left\{\sigma\mathcal{N}\left(0,\|X-\Pi XQ\|_{\mathrm{F}}^2\right)\geq\frac{1}{2}\|X-\Pi XQ\|_{\mathrm{F}}^2-\delta^2\|X\|_{\mathrm{F}}^2\right\}.
\end{aligned}\tag{54}$$

Consider the following events

$$\mathcal{E}_1 \triangleq \left\{cdn \le \|X\|_{\mathrm{F}}^2 \le Cdn\right\}, \quad \mathcal{E}_2 \triangleq \left\{\min_{\Pi \neq I} \min_{Q \in N} \|X - \Pi X Q\|_{\mathrm{F}} \ge C\sqrt{d}\sigma\right\}.$$

It is well known that $\mathbb{P}\{\mathcal{E}_1\} = 1 - o(1)$, and by Lemma 9 we also have $\mathbb{P}\{\mathcal{E}_2\} = 1 - o(1)$. On the events $\mathcal{E}_1$ and $\mathcal{E}_2$, the previous estimate (54) for $\Pi \neq I$ reduces to

$$\begin{aligned}
&\mathbb{P}\left\{\langle X^\top \Pi^\top Y, Q\rangle \ge (1-\delta^2)\langle X^\top Y, I_d\rangle, \mathcal{E}_1, \mathcal{E}_2\right\} \\
&\quad \le \mathbb{P}\left\{\sigma \mathcal{N}\left(0, \|X - \Pi X Q\|_{\mathrm{F}}^2\right) \ge \frac{1}{4}\|X - \Pi X Q\|_{\mathrm{F}}^2\right\} \le \mathbb{E}\exp\left\{-\frac{1}{32\sigma^2}\|X - \Pi X Q\|_{\mathrm{F}}^2\right\}.
\end{aligned} \tag{55}$$

By Lemma 8, the reduction (55) and a union bound, we have

$$\begin{aligned}
&\mathbb{P}\left\{\max_{\Pi \neq I} \|X^\top \Pi^\top Y\|_* \ge \|X^\top Y\|_*\right\} \\
&\le \mathbb{P}\left\{\max_{\Pi \neq I} \|X^\top \Pi^\top Y\|_* \ge \|X^\top Y\|_*, \mathcal{E}_1, \mathcal{E}_2\right\} + \mathbb{P}\{\mathcal{E}_1^c\} + \mathbb{P}\{\mathcal{E}_2^c\} \\
&\le \mathbb{P}\left\{\max_{\Pi \neq I_n} \max_{Q \in N} \langle X^\top \Pi^\top Y, Q\rangle \ge (1-\delta^2)\langle X^\top Y, I_d\rangle, \mathcal{E}_1, \mathcal{E}_2\right\} + o(1) \\
&\le \sum_{\Pi \neq I_n} \sum_{Q \in N} \mathbb{P}\left\{\langle X^\top \Pi^\top Y, Q\rangle \ge (1-\delta^2)\langle X^\top Y, I_d\rangle, \mathcal{E}_1, \mathcal{E}_2\right\} + o(1) \\
&\le \sum_{\Pi \neq I_n} \sum_{Q \in N} \mathbb{E}\exp\left\{-\frac{1}{32\sigma^2}\|X - \Pi X Q\|_{\mathrm{F}}^2\right\} + o(1) \\
&= o(1).
\end{aligned}$$

This implies that the ground truth $\Pi^* = I_n$ is the approximate MLE with probability $1 - o(1)$, i.e.,

$$\mathbb{P}\left\{\mathrm{argmax}_{\Pi \in S_n} \|X^\top \Pi^\top Y\|_* = I_n\right\} = 1 - o(1),$$

which shows the success of perfect recovery with high probability.

(ii) The arguments are essentially the same as Part (i). For a sufficiently small $\varepsilon = \varepsilon(n) > 0$, take $\sigma^{-d} > 16n2^{2/\varepsilon}$ and consider the event

$$\mathcal{E}_2' \triangleq \left\{\min_{\mathrm{d}(\pi, \mathrm{Id}) \ge \varepsilon n} \min_{Q \in N} \|X - \Pi X Q\|_{\mathrm{F}} \ge C\sqrt{d}\sigma\right\}.$$

Then Lemma 9 implies $\mathbb{P}\{\mathcal{E}_2'\} = 1 - o(1)$. On the event $\mathcal{E}_1$ and $\mathcal{E}_2'$, the reduction estimate for Π with $\mathrm{d}(\pi, \mathrm{Id}) \ge \varepsilon n$ still holds

$$\mathbb{P}\left\{\langle X^\top \Pi^\top Y, Q\rangle \ge (1-\delta^2)\langle X^\top Y, I_d\rangle, \mathcal{E}_1, \mathcal{E}_2'\right\} \le \mathbb{E}\exp\left\{-\frac{1}{32\sigma^2}\|X - \Pi X Q\|_{\mathrm{F}}^2\right\}.$$

Combining this with Lemma 8, we have

$$\mathbb{P}\left\{\max_{\mathrm{d}(\pi,\mathrm{Id})\geq\varepsilon n}\|X^\top\Pi^\top Y\|_*\geq\|X^\top Y\|_*\right\}$$
$$\leq\sum_{\mathrm{d}(\pi,\mathrm{Id})\geq\varepsilon n}\sum_{Q\in N}\mathbb{E}\exp\left\{-\frac{1}{32\sigma^2}\|X-\Pi XQ\|_{\mathrm{F}}^2\right\}+o(1)=o(1).$$

Thus,

$$\mathbb{P}\left\{\mathsf{overlap}(\widehat{\pi}_{\mathrm{AML}},\pi^*)\geq 1-\varepsilon\right\}=1-o(1).$$

Taking $\sigma\ll n^{-1/d}$ so that $\epsilon=o(1)$, this implies the desired (6). ■

Appendix D. Proof for the distance model

In this section, we prove Theorem 2. Let $\widetilde{X}\triangleq(I-\mathbf{F})X$, $\widetilde{Y}\triangleq(I-\mathbf{F})Y$ and $\widetilde{Z}\triangleq(I-\mathbf{F})Z$. Recall that the approximate MLE for the distance model is given by (9). As in the proof of Theorem 1, thanks to the orthogonal invariance of the nuclear norm $\|\cdot\|_*$, we may assume $\widetilde{A}^{1/2}=\widetilde{X}$ and $\widetilde{B}^{1/2}=\widetilde{Y}$ without loss of generality, so that

$$\widetilde{\Pi}_{\mathrm{AML}}=\arg\max_{\Pi\in\mathfrak{S}(n)}\|\widetilde{X}^\top\Pi^\top\widetilde{Y}\|_*.$$

Following the arguments for the dot-product model in Appendix C, a key step is to extend the estimate for $p(\Pi,Q)$ in (27) to the following MGF:

$$\widetilde{p}(\Pi,Q)\triangleq\mathbb{E}\exp\left\{-\frac{1}{32\sigma^2}\left\|\widetilde{X}-\Pi\widetilde{X}Q\right\|_{\mathrm{F}}^2\right\},\tag{56}$$

where $\Pi\in\mathfrak{S}_n$ and $Q\in O(d)$. The following lemma gives a comparison between the MGF for the distance model and that for the dot-product model defined in (27), the latter of which was previous estimated in Lemma 7.

Lemma 10 *Fix a permutation matrix $\Pi\in\mathfrak{S}_n$. For $Q\in O(d)$, denote by $e^{\mathrm{i}\theta_1},\ldots,e^{\mathrm{i}\theta_d}$ the eigenvalues of Q, where $\theta_1,\ldots,\theta_d\in[-\pi,\pi]$. Then*

$$\widetilde{p}(\Pi,Q)\leq p(\Pi,Q)\prod_{\ell=1}^{d}\left(1+\frac{\theta_l^2}{16\sigma^2}\right)^{1/2}.\tag{57}$$

Proof Let $t=\frac{1}{32\sigma^2}$. Denote by $\widetilde{x}=\mathsf{vec}(\widetilde{X})\in\mathbb{R}^{nd}$ the vectorization of $\widetilde{X}$ and recall that $x=\mathsf{vec}(X)\in\mathbb{R}^{nd}$ satisfies $x\sim\mathcal{N}(0,I_{nd})$. Then

$$\left\|\widetilde{X}-\Pi\widetilde{X}Q\right\|_{\mathrm{F}}^2=\left\|(I_{nd}-Q^\top\otimes\Pi)\widetilde{x}\right\|^2=\left\|(I_{nd}-Q^\top\otimes\Pi)(I_d\otimes(I_n-\mathbf{F}))x\right\|^2.$$

Denote $\widetilde{H}\triangleq(I_{nd}-Q^\top\otimes\Pi)(I_d\otimes(I_n-\mathbf{F}))$, then

$$\widetilde{p}(\Pi,Q)=\mathbb{E}\exp\left(-tx^\top\widetilde{H}^\top\widetilde{H}x\right)=\left[\det\left(I+2t\widetilde{H}^\top\widetilde{H}\right)\right]^{-\frac{1}{2}}.$$

It suffices to compute the eigenvalues of $\widetilde{H}$. Recall that the spectrum of Π is given by (34). We claim that the spectrum of $\widetilde{H}$ is the following multiset

$$\mathsf{Spec}(\widetilde{H}) = \Big(\mathsf{Spec}(H)\backslash\{1 - e^{-\mathrm{i}\theta_\ell} : \ell = 1, \dots, d\}\Big) \cup \{0 \text{ with multiplicity } d\}\,, \tag{58}$$

where $\mathsf{Spec}(H)$ is the spectrum of H defined in Lemma 7, given by

$$\mathsf{Spec}(H) = \Big\{1 - e^{-\mathrm{i}\theta_\ell}\lambda_j : \lambda_j \in \mathsf{Spec}(\Pi),\, j = 1, \dots, n,\, \ell = 1, \dots, d\Big\}\,.$$

Now we prove (58). As shown in (34), Π has eigenvalue 1 with multiplicity $\mathfrak{c}(\Pi)$, where $\mathfrak{c}(\Pi)$ denote the number of cycles. We denote these by $\lambda_1 = \cdots = \lambda_{\mathfrak{c}(\Pi)} = 1$. Using the cycle decomposition and the block diagonal structure as in Lemma 7, we know that the eigenvectors corresponding to $\lambda_1, \dots, \lambda_{\mathfrak{c}(\Pi)}$ are of the following form

$$v_i = (0, \dots, 0, 1, \dots, 1, 0, \dots, 0)^\top, \quad i = 1, \dots, \mathfrak{c}(\Pi)$$

where the number of 1's equals the length of the corresponding cycle. In particular, due to the block diagonal structure, the 1 blocks in v_i's do not overlap. Therefore, we know that the vector $\widetilde{v}_1 = \frac{1}{\sqrt{n}}\sum_{i=1}^{\mathfrak{c}(\Pi)} v_i = \frac{1}{\sqrt{n}}\mathbf{1} = \frac{1}{\sqrt{n}}(1, \dots, 1)^\top \in \mathbb{R}^n$ is in the eigenspace of 1. Using the Gram-Schmidt process, we can construct vectors $\widetilde{v}_2, \dots, \widetilde{v}_{\mathfrak{c}(\Pi)}$ such that $\{\widetilde{v}_i\}_{i=1}^n$ is a orthonormal basis of the eigenspace, i.e.

$$\langle \widetilde{v}_i, \widetilde{v}_j\rangle = \delta_{ij}, \;\; \mathsf{span}(\widetilde{v}_1, \dots, \widetilde{v}_{\mathfrak{c}(\Pi)}) = \mathsf{span}(v_1, \dots, v_{\mathfrak{c}(\Pi)}).$$

Pick an arbitrary eigenvalue μ of $Q^\top$ with eigenvector $w \in \mathbb{R}^d$, and also pick an arbitrary eigenvalue λ of Π with eigenvector $v \in \mathbb{R}^n$. Based on the arguments above, if $\lambda \neq \lambda_1$, then $v \perp \widetilde{v}_1$, and therefore

$$\widetilde{H}(w \otimes v) = w \otimes (I - \mathbf{F})v - (Q^\top w) \otimes \Pi(I - \mathbf{F})v = w \otimes v - \mu w \otimes \lambda v = (1 - \mu\lambda)(w \otimes v). \tag{59}$$

For the eigenpair $(\lambda_1, \widetilde{v}_1)$, we have

$$\widetilde{H}(w \otimes \widetilde{v}_1) = w \otimes (I - \mathbf{F})\widetilde{v}_1 - (Q^\top w) \otimes \Pi(I - \mathbf{F})\widetilde{v}_1 = w \otimes 0 - \mu w \otimes 0 = 0. \tag{60}$$

Combining (59) and (60), we conclude that for $\ell = 1, \dots, d$ and $j = 2, \dots, n$, the eigenvalue $1 - e^{-\mathrm{i}\theta_\ell}\lambda_j$ of H remains to be an eigenvalue of $\widetilde{H}$, while the eigenvalues $1 - e^{-\mathrm{i}\theta_\ell}\lambda_1 = 1 - e^{-\mathrm{i}\theta_\ell}$ of H are replaced by 0 in the spectrum of $\widetilde{H}$. Hence we have shown (58) is true.

Using (58) and (33), we obtain

$$\begin{aligned}
\widetilde{p}(\Pi, Q) &= \prod_{j=2}^{n}\prod_{\ell=1}^{d}\Big(1 + 2t|1 - e^{-\mathrm{i}\theta_\ell}\lambda_j|^2\Big)^{-1/2} \\
&= p(\Pi, Q)\prod_{\ell=1}^{d}\Big(1 + 2t|1 - e^{-\mathrm{i}\theta_\ell}|^2\Big)^{1/2} \\
&= p(\Pi, Q)\prod_{\ell=1}^{d}\left(1 + 2t(2 - 2\cos\theta_\ell)\right)^{1/2} \\
&\leq p(\Pi, Q)\prod_{\ell=1}^{d}\left(1 + \frac{\theta_\ell^2}{16\sigma^2}\right)^{1/2},
\end{aligned}$$

which completes the proof. ■

Applying Lemma 10, the following lemma is the counterpart of Lemma 8.

Lemma 11 *Suppose $d = o(\log n)$. For some $\sigma_0 > 0$, let $\delta = \sigma_0/\sqrt{n}$ and $N \subset O(d)$ be the δ-net defined in (25).*

(i) If $\sigma_0 = o(n^{-2/d})$, then

$$\sum_{\Pi \neq I_n} \sum_{Q \in N} \mathbb{E} \exp\left\{ -\frac{1}{32\sigma_0^2} \left\| \widetilde{X} - \Pi \widetilde{X} Q \right\|_{\mathrm{F}}^2 \right\} = o(1). \tag{61}$$

(ii) For any $\varepsilon = \varepsilon(n) > 0$, if $\sigma_0^{-d} > 16n2^{2/\varepsilon}$, then the following is true

$$\sum_{\mathrm{d}(\pi,\mathrm{Id}) \geq \varepsilon n} \sum_{Q \in N} \mathbb{E} \exp\left\{ -\frac{1}{32\sigma_0^2} \left\| \widetilde{X} - \Pi \widetilde{X} Q \right\|_{\mathrm{F}}^2 \right\} = o(1). \tag{62}$$

Proof (i) Similarly as in Lemma 8 Part (i), using (57) we have

$$\begin{aligned}
&\sum_{\Pi \neq I_n} \sum_{Q \in N} \widetilde{p}(\Pi, Q) \\
&\leq \sum_{n_1=0}^{n-2} \sum_{m_1,\ldots,m_d = \lfloor -\frac{4\pi}{\delta} \rfloor}^{\lceil \frac{4\pi}{\delta} \rceil} \left\{ \left| N\left(e^{\mathrm{i}\frac{m_1\delta}{4}}, \ldots, e^{\mathrm{i}\frac{m_d\delta}{4}} \right) \right| (n-n_1)! \binom{n}{n_1} (C\sigma_0)^{\frac{n+n_1}{2}} d \right. \\
&\qquad \left. \times \left[\prod_{\ell=1}^{d} \frac{1}{(\frac{\delta|m_\ell|}{4} + \sigma_0)^{n_1}} \left(1 + \frac{\delta^2 m_\ell^2}{256\sigma_0^2} \right)^{\frac{1}{2}} \right] \right\} \\
&\leq \sum_{n_1=0}^{n-2} \left((C\sigma_0)^d n^2 \right)^{\frac{n-n_1}{2}} \left[\sum_{m=\lfloor -\frac{4\pi}{\delta} \rfloor}^{\lceil \frac{4\pi}{\delta} \rceil} \frac{1}{(1 + \frac{\delta}{4\sigma_0}|m|)^{n_1}} \left(1 + \frac{\delta^2 m^2}{256\sigma_0^2} \right)^{\frac{1}{2}} \left(1 + \frac{|m|}{2} \right)^{2d^2} \right]^d \\
&= \sum_{n_1=0}^{n-2} \left((C\sigma_0)^d n^2 \right)^{\frac{n-n_1}{2}} \left[\sum_{m=\lfloor -\frac{4\pi}{\delta} \rfloor}^{\lceil \frac{4\pi}{\delta} \rceil} \frac{1}{(1 + \frac{|m|}{4\sqrt{n}})^{n_1}} \left(1 + \frac{m^2}{256n} \right)^{\frac{1}{2}} \left(1 + \frac{|m|}{2} \right)^{2d^2} \right]^d \\
&\leq \sum_{n_1=0}^{n-2} \left((C\sigma_0)^d n^2 \right)^{\frac{n-n_1}{2}} \left[\sum_{m=\lfloor -\frac{4\pi}{\delta} \rfloor}^{\lceil \frac{4\pi}{\delta} \rceil} \frac{1}{(1 + \frac{|m|}{4\sqrt{n}})^{n_1}} \left(1 + \frac{|m|}{2} \right)^{2d^2+1} \right]^d .
\end{aligned}$$

Let

$$\widetilde{T}_{n_1} \triangleq \sum_{m=\lfloor -\frac{4\pi}{\delta} \rfloor}^{\lceil \frac{4\pi}{\delta} \rceil} \frac{1}{(1 + \frac{|m|}{4\sqrt{n}})^{n_1}} \left(1 + \frac{|m|}{2} \right)^{2d^2+1} .$$

Using the same arguments as in (41), (43) and (44), $\widetilde{T}_{n_1}$ can be bounded by

$$\widetilde{T}_{n_1}^d \le \begin{cases} C^{d^3} n^{d^3+d} L^{2d^2+2} & \text{if } n_1 \le \sqrt{n}, \\ C^d (8\sqrt{n})^{2d^2+2} & \text{if } \sqrt{n} < n_1 < 32\sqrt{n}(\log n)^2, \\ C^d & \text{if } n_1 \ge 32\sqrt{n}(\log n)^2, \end{cases} \tag{63}$$

where $L = \sigma_0^{-d}$. Consequently, following the similar estimates in (42) and (45),

$$\sum_{\Pi \neq I_n} \sum_{Q \in N} \widetilde{p}(\Pi, Q) = o(1),$$

which completes the proof.

(ii) Combined with (57), using the same arguments as in Lemma 8 Part (ii) yields

$$\sum_{\mathrm{d}(\pi,\mathrm{Id}) \ge \varepsilon n} \sum_{Q \in N} \widetilde{p}(\Pi, Q) \le \sum_{k=\varepsilon n}^{n} \binom{n}{k} L^{-k} k! \left(\frac{16L}{k}\right)^{k/2} \widetilde{T}_{n-k}^d = \widetilde{J}_1 + \widetilde{J}_2$$

where

$$\widetilde{J}_1 \triangleq \sum_{k=\varepsilon n}^{n-32\sqrt{n}(\log n)^2} \binom{n}{k} L^{-k} k! \left(\frac{16L}{k}\right)^{k/2} \widetilde{T}_{n-k}^d,$$

$$\widetilde{J}_2 \triangleq \sum_{k=n-32\sqrt{n}(\log n)^2+1}^{n} \binom{n}{k} L^{-k} k! \left(\frac{16L}{k}\right)^{k/2} \widetilde{T}_{n-k}^d.$$

By (63), these two term can be bounded in the same way as in (50) and (51). Thus,

$$\sum_{\mathrm{d}(\pi,\mathrm{Id}) \ge \varepsilon n} \sum_{Q \in N} \widetilde{p}(\Pi, Q) = o(1),$$

which completes the proof. ∎

Lemma 11 implies the following high probability estimates. The proof is the same as in Lemma 9 via Chernoff bound and therefore we omit it here.

Lemma 12 *Suppose $d = o(\log n)$. For some $\sigma_0 > 0$, let $\delta = \sigma_0/\sqrt{n}$ and $N \subset O(d)$ be the δ-net defined in (25).*

(i) If $\sigma_0 = o(n^{-2/d})$, for any constant $c > 0$, the following inequality is true with high probability

$$\min_{\Pi \neq I_n} \min_{Q \in N} \left\| \widetilde{X} - \Pi \widetilde{X} Q \right\|_{\mathrm{F}} \ge c\sqrt{d}\sigma_0. \tag{64}$$

(ii) For any $\varepsilon = \varepsilon(n) > 0$, if $\sigma_0^{-d} > 16n2^{2/\varepsilon}$, the following is true for any fixed constant $c > 0$ with high probability

$$\min_{\mathrm{d}(\pi,\mathrm{Id}) \ge \varepsilon n} \min_{Q \in N} \left\| \widetilde{X} - \Pi \widetilde{X} Q \right\|_{\mathrm{F}} \ge c\sqrt{d}\sigma_0. \tag{65}$$

Now we are ready to prove Theorem 2 . Similarly as in the dot-product model (see the remark following Theorem 1), for almost perfect recovery, we actually prove a stronger nonasymptotic bound: For all sufficiently small ε, if $\sigma^{-d} > 16n2^{2/\varepsilon}$, then $\mathsf{overlap}(\widetilde{\pi}_{\mathrm{AML}}, \pi^*) \geq 1-\varepsilon$ with high probability, which clearly implies Theorem 2 by taking $\sigma \ll n^{-1/d}$.

Proof [Proof of Theorem 2] (i) Let N be the δ-net for $O(d)$ defined in (25). Following the same argument as in Theorem 1

$$\mathbb{P}\left\{\|\widetilde{X}^\top \Pi^\top \widetilde{Y}\|_* \geq \|\widetilde{X}^\top \widetilde{Y}\|_*\right\} \leq \mathbb{P}\left\{\max_{Q\in N} \langle \widetilde{X}^\top \Pi^\top \widetilde{Y}, Q\rangle \geq (1-\delta^2)\langle \widetilde{X}^\top \widetilde{Y}, I_d\rangle\right\}.$$

For fixed Π and Q, we have

$$\begin{aligned}&\mathbb{P}\left\{\langle \widetilde{X}^\top \Pi^\top \widetilde{Y}, Q\rangle \geq (1-\delta^2)\langle \widetilde{X}^\top \widetilde{Y}, I_d\rangle\right\}\\ &\qquad = \mathbb{P}\left\{\sigma\langle \widetilde{Z}, (1-\delta^2)\widetilde{X} - \Pi\widetilde{X}Q\rangle \geq (1-\delta^2)\|\widetilde{X}\|_{\mathrm{F}}^2 - \langle \widetilde{X}, \Pi\widetilde{X}Q\rangle\right\}.\end{aligned}$$

Since the entries of $\widetilde{Z}$ are not independent, we need to be more careful:

$$\begin{aligned}\langle \widetilde{Z}, (1-\delta^2)\widetilde{X} - \Pi\widetilde{X}Q\rangle &= \langle (I-\mathbf{F})Z, (1-\delta^2)\widetilde{X} - \Pi\widetilde{X}Q\rangle\\ &= \langle Z, (I-\mathbf{F})((1-\delta^2)\widetilde{X} - \Pi\widetilde{X}Q)\rangle\\ &= \langle Z, (1-\delta^2)\widetilde{X} - \Pi\widetilde{X}Q\rangle,\end{aligned}$$

because $(I-\mathbf{F})\widetilde{X} = \widetilde{X}$ and $I-\mathbf{F}$ commutes with any permutation matrix Π. Therefore, similarly as in (54),

$$\begin{aligned}&\mathbb{P}\left\{\langle \widetilde{X}^\top \Pi^\top \widetilde{Y}, Q\rangle \geq (1-\delta^2)\langle \widetilde{X}^\top \widetilde{Y}, I_d\rangle\right\}\\ =&\mathbb{P}\left\{\sigma\mathcal{N}\left(0, (1-\delta^2)\left\|\widetilde{X} - \Pi\widetilde{X}Q\right\|_{\mathrm{F}}^2 - \delta^4\left\|\widetilde{X}\right\|_{\mathrm{F}}^2\right) \geq \frac{1}{2}\left\|\widetilde{X} - \Pi\widetilde{X}Q\right\|_{\mathrm{F}}^2 - \delta^2\left\|\widetilde{X}\right\|_{\mathrm{F}}^2\right\}\\ \leq&\mathbb{P}\left\{\sigma\mathcal{N}\left(0, \left\|\widetilde{X} - \Pi\widetilde{X}Q\right\|_{\mathrm{F}}^2\right) \geq \frac{1}{2}\left\|\widetilde{X} - \Pi\widetilde{X}Q\right\|_{\mathrm{F}}^2 - \delta^2\left\|\widetilde{X}\right\|_{\mathrm{F}}^2\right\}. \qquad (66)\end{aligned}$$

Consider the events

$$\mathcal{E}_1 \triangleq \left\{cdn \leq \left\|\widetilde{X}\right\|_{\mathrm{F}}^2 \leq Cdn\right\},\ \ \mathcal{E}_2 \triangleq \left\{\min_{\Pi\neq I}\min_{Q\in N}\left\|\widetilde{X} - \Pi\widetilde{X}Q\right\|_{\mathrm{F}} \geq C\sqrt{d}\sigma\right\}.$$

We claim that $\mathbb{P}\{\mathcal{E}_1\} = 1 - o(1)$. To see this, note that

$$\begin{aligned}\|\widetilde{X}\|_{\mathrm{F}}^2 = \langle (I-\mathbf{F})X, (I-\mathbf{F})X\rangle &= \langle X, (I-\mathbf{F})X\rangle\\ &= \mathrm{Tr}(X^\top(I-\mathbf{F})X) = \sum_{i=1}^d\sum_{\alpha,\beta=1}^n X_{\alpha i}X_{\beta i}(I-F)_{\alpha\beta}.\end{aligned}$$

For each $i = 1, \ldots, d$, we have $\sum_{\alpha,\beta=1}^n X_{\alpha i}X_{\beta i}(I-\mathbf{F})_{\alpha\beta} = \mathsf{Col}_i(X)^\top(I-\mathbf{F})\mathsf{Col}_i(X)$, where $\mathsf{Col}_i(X) \sim \mathcal{N}(0, I_n)$ is the i-th column of X. By Hanson-Wright inequality (see e.g. (Rudelson

and Vershynin, 2013, Theorem 1.1)), for each $t \geq 0$,

$$\mathbb{P}\left\{|\mathsf{Col}_i(X)^\top(I-\mathbf{F})\mathsf{Col}_i(X) - \mathbb{E}\mathsf{Col}_i(X)^\top(I-\mathbf{F})\mathsf{Col}_i(X)| > t\right\} \leq 2\exp\left[-c\min\left(\frac{t^2}{\|I-\mathbf{F}\|_{\mathrm{F}}^2}, \frac{t}{\|I-\mathbf{F}\|}\right)\right].$$

Taking $t = n^{3/4}$ and simplifying the above inequality yield

$$\mathbb{P}\left\{|\mathsf{Col}_i(X)^\top(I-\mathbf{F})\mathsf{Col}_i(X) - (n-1)| > n^{3/4}\right\} \leq 2\exp\left(-cn^{1/2}\right). \tag{67}$$

Note that (67) is true for every $i = 1, \ldots, d$, and the columns $\mathsf{Col}_i(X)$'s are independent. This immediately gives us $\mathbb{P}\{\mathcal{E}_1\} = 1 - o(1)$. Moreover, by Lemma 12 we also have $\mathbb{P}\{\mathcal{E}_2\} = 1 - o(1)$. On the events $\mathcal{E}_1$ and $\mathcal{E}_2$, the estimate (66) reduces to

$$\begin{aligned}
&\mathbb{P}\left\{\langle\widetilde{X}^\top\Pi^\top\widetilde{Y}, Q\rangle \geq (1-\delta^2)\langle\widetilde{X}^\top\widetilde{Y}, I_d\rangle, \mathcal{E}_1, \mathcal{E}_2\right\} \\
&\leq \mathbb{P}\left\{\sigma\mathcal{N}\left(0, \left\|\widetilde{X} - \Pi\widetilde{X}Q\right\|_{\mathrm{F}}^2\right) \geq \frac{1}{4}\left\|\widetilde{X} - \Pi\widetilde{X}Q\right\|_{\mathrm{F}}^2\right\} \leq \mathbb{E}\exp\left\{-\frac{1}{32\sigma^2}\left\|\widetilde{X} - \Pi\widetilde{X}Q\right\|_{\mathrm{F}}^2\right\}.
\end{aligned} \tag{68}$$

Combining this with Lemma 11 and applying a union bound, we have

$$\begin{aligned}
&\mathbb{P}\left\{\max_{\Pi\neq I}\|\widetilde{X}^\top\Pi^\top\widetilde{Y}\|_* \geq \|\widetilde{X}^\top\widetilde{Y}\|_*\right\} \\
&\leq \mathbb{P}\left\{\max_{\Pi\neq I_n}\max_{Q\in N}\langle\widetilde{X}^\top\Pi^\top\widetilde{Y}, Q\rangle \geq (1-\delta^2)\langle\widetilde{X}^\top\widetilde{Y}, I_d\rangle, \mathcal{E}_1, \mathcal{E}_2\right\} + \mathbb{P}\{\mathcal{E}_1^c\} + \mathbb{P}\{\mathcal{E}_2^c\} \\
&\leq \sum_{\Pi\neq I_n}\sum_{Q\in N}\mathbb{P}\left\{\langle\widetilde{X}^\top\Pi^\top\widetilde{Y}, Q\rangle \geq (1-\delta^2)\langle\widetilde{X}^\top\widetilde{Y}, I_d\rangle, \mathcal{E}_1, \mathcal{E}_2\right\} + o(1) \\
&\leq \sum_{\Pi\neq I_n}\sum_{Q\in N}\mathbb{E}\exp\left\{-\frac{1}{32\sigma^2}\left\|\widetilde{X} - \Pi\widetilde{X}Q\right\|_{\mathrm{F}}^2\right\} + o(1) \\
&= o(1).
\end{aligned}$$

This implies $\widetilde{\pi}_{\mathrm{AML}} = \mathrm{Id}$ with high probability, which completes the proof.

(ii) The idea is the same as Theorem 1 Part (ii). For a sufficiently small $\varepsilon = \varepsilon(n) > 0$, take $\sigma^{-d} > 16n2^{2/\varepsilon}$ and consider the event

$$\mathcal{E}_2' \triangleq \left\{\min_{\mathrm{d}(\pi,\mathrm{Id})\geq\varepsilon n}\min_{Q\in N}\left\|\widetilde{X} - \Pi\widetilde{X}Q\right\|_{\mathrm{F}} \geq C\sqrt{d}\sigma\right\}.$$

Then Lemma 12 implies $\mathbb{P}\{\mathcal{E}_2'\} = 1 - o(1)$. On the event $\mathcal{E}_1$ and $\mathcal{E}_2'$, the reduction estimate (66) for Π with $\mathrm{d}(\pi, \mathrm{Id}) \geq \varepsilon n$ still holds

$$\mathbb{P}\left\{\langle\widetilde{X}^\top\Pi^\top\widetilde{Y}, Q\rangle \geq (1-\delta^2)\langle\widetilde{X}^\top\widetilde{Y}, I_d\rangle, \mathcal{E}_1, \mathcal{E}_2'\right\} \leq \mathbb{E}\exp\left\{-\frac{1}{32\sigma^2}\left\|\widetilde{X} - \Pi\widetilde{X}Q\right\|_{\mathrm{F}}^2\right\}.$$

Combined with Lemma 11, we have

$$\mathbb{P}\left\{\max_{\mathrm{d}(\pi,\mathrm{Id})\geq\varepsilon n}\|\widetilde{X}^\top\Pi^\top\widetilde{Y}\|_*\geq\|\widetilde{X}^\top\widetilde{Y}\|_*\right\}$$
$$\leq\sum_{\mathrm{d}(\pi,\mathrm{Id})\geq\varepsilon n}\sum_{Q\in N}\mathbb{E}\exp\left\{-\frac{1}{32\sigma^2}\left\|\widetilde{X}-\Pi\widetilde{X}Q\right\|_{\mathrm{F}}^2\right\}+o(1)=o(1).$$

Thus,

$$\mathbb{P}\left\{\mathsf{overlap}(\widetilde{\pi}_{\mathrm{AML}},\pi^*)\geq 1-\varepsilon\right\}=1-o(1),$$

which completes the proof. ■

Appendix E. Information-theoretic necessary conditions

In this section, we derive necessary conditions for both almost perfect recovery and perfect recovery for the linear assignment model (1). These conditions also hold for the weaker dot-product and distance models.

E.1. Impossibility of almost perfect recovery

We first derive a necessary condition for almost perfect recovery that holds for any d via a simple mutual information argument. Then we focus on the special case where d is a constant and give a much sharper analysis, improving the necessary condition from $\sigma\leq n^{-(1-o(1))/d}$ to $\sigma=o(n^{-1/d})$. Note that achieving a vanishing recovery error in expectation is equivalent to that with high probability (see e.g. (Hajek et al., 2017, Appendix A)). Thus without loss of generality, we focus on the expected number of errors $\mathbb{E}\mathrm{d}\left(\pi^*,\widehat{\pi}\right)$ in this subsection.

Proposition 13 *For any $\epsilon\in(0,1)$, if there exists an estimator $\widehat{\pi}\equiv\widehat{\pi}(X,Y)$ such that $\mathbb{E}\mathrm{d}\left(\pi^*,\widehat{\pi}\right)\leq\epsilon n$, then*

$$\frac{d}{2}\log\left(1+\frac{1}{\sigma^2}\right)-(1-\epsilon)\log n+1+\frac{\log(n+1)}{n}\geq 0. \tag{69}$$

The necessary condition (69) further specializes to:

- $d=o(\log n)$:

$$\sigma=O\left(n^{-(1-\epsilon)/d}\right). \tag{70}$$

  This yields Theorem 3(ii) and resolves (Kunisky and Niles-Weed, 2022, Conjecture 1.4, item 1) in the positive;

- $d=\Theta(\log n)$:

$$\sigma\leq\frac{1-\epsilon+o(1)}{\sqrt{n^{2/d}-1}};$$

- $d = \omega(\log n)$:

$$\sigma \leq \sqrt{\frac{d}{2(1-\epsilon-o(1))\log n}}.$$

  In this case, this necessary condition matches the sufficient condition of almost perfect recovery in (Dai et al., 2020, Theorem 1) and (Kunisky and Niles-Weed, 2022, Section A.2) up to $1 + o(1)$ factor, thereby determining the sharp information-theoretic limit for the linear assignment model in high dimensions.

Proof Since $\pi^* \to (X, Y) \to \widehat{\pi}$ form a Markov chain, by the data processing inequality of mutual information, we have

$$I(\pi^*; X, Y) \geq I(\pi^*; \widehat{\pi}) = H(\pi^*) - H(\pi^*|\widehat{\pi}). \tag{71}$$

On the one hand, note that $H(\pi^*) = \log(n!) \geq n \log n - n$. Moreover, for any fixed realization of $\widehat{\pi}$, the number of π^* such that $\mathrm{d}(\pi^*, \widehat{\pi}) = \ell$ is $\binom{n}{\ell}!\ell \leq n^\ell$, where $!\ell$ denotes the number of derangements of ℓ elements, given by

$$!\ell = \ell! \sum_{i=0}^{\ell} \frac{(-1)^i}{i!} = \left[\frac{\ell!}{e}\right],$$

and $[\cdot]$ denotes rounding to the nearest integer. Therefore,

$$H(\pi^*|\widehat{\pi}, \mathrm{d}(\pi^*, \widehat{\pi})) \leq \mathbb{E}\mathrm{d}(\pi^*, \widehat{\pi}) \log n \leq \epsilon n \log n.$$

Furthermore, $\mathrm{d}(\pi^*, \widehat{\pi})$ takes values in $\{0, 1, \ldots, n\}$. Thus from the chain rule,

$$H(\pi^*|\widehat{\pi}) = H(\mathrm{d}(\pi^*, \widehat{\pi})|\widehat{\pi}) + H(\pi^*|\widehat{\pi}, \mathrm{d}(\pi^*, \widehat{\pi})) \leq \log(n+1) + \epsilon n \log n. \tag{72}$$

On the other hand, the information provided by the observation (X, Y) about π^* satisfies

$$\begin{aligned} I(\pi^*; X, Y) &= I(\Pi^* X; \Pi^* X + \sigma Z|X) \\ &\overset{(a)}{\leq} \frac{nd}{2} \log\left(1 + \frac{\mathbb{E}[\|X\|^2]}{nd\sigma^2}\right) \\ &= \frac{nd}{2} \log\left(1 + \frac{1}{\sigma^2}\right), \end{aligned} \tag{73}$$

where (a) follows from the Gaussian channel capacity formula and the fact that the mutual information in the Gaussian channel under a second moment constraint is maximized by the Gaussian input distribution. Combining (71)–(73), we get that

$$\frac{nd}{2} \log\left(1 + \frac{1}{\sigma^2}\right) \geq (1-\epsilon) n \log n - n - \log(n+1),$$

arriving at the desired necessary condition (69). ■

While the negative result in Proposition 13 holds for any d, the necessary condition (69) turns out to be loose for bounded d. The following result gives the optimal condition in this case.

Theorem 14 *Assume $\sigma = \sigma_0 n^{-1/d}$ for any constant $\sigma_0 \in (0, 1/2)$. There exists a constant $\delta_0(\sigma_0, d)$ that only depends on σ_0, d such that for any estimator $\widehat{\Pi}$ and all sufficiently large n,*

$$\mathbb{E}\mathrm{d}\left(\Pi^*, \widehat{\Pi}\right) \geq \delta_0 n.$$

Theorem 14 readily implies that for constant d, $\sigma = o(n^{-1/d})$ is necessary for achieving the almost perfect recovery, i.e., $\mathbb{E}\mathrm{d}(\Pi^*, \widehat{\Pi}) = o(n)$. To prove Theorem 14, we follow the program in Ding et al. (2021b) of analyzing the posterior distribution. The likelihood function of (X, Y) given $\Pi^* = \Pi$ is proportional to $\exp(-\frac{1}{2\sigma^2}\|Y - \Pi X\|_{\mathrm{F}}^2)$. Therefore, conditional on (X, Y), the posterior distribution of Π^* is a Gibbs measure, given by

$$\mu_{X,Y}(\Pi) = \frac{1}{Z(X,Y)} \exp\left(L(\Pi)\right), \quad \text{where } L(\Pi) = \frac{1}{\sigma^2}\left\langle \Pi X, Y\right\rangle,$$

and $Z(X, Y)$ is the normalization factor.

As observed in (Ding et al., 2021b, Section 3.1), in order to prove the impossibility of almost perfect recovery, it suffices to consider the estimator $\widetilde{\Pi}$ which is sampled from the posterior distribution $\mu_{X,Y}(\Pi)$. To see this, given any estimator $\widehat{\Pi} \equiv \widehat{\Pi}(X, Y)$, $(\widehat{\Pi}, \Pi^*)$ and $(\widehat{\Pi}, \widetilde{\Pi})$ are equal in law, and hence

$$\mathbb{E}[\mathrm{d}(\widetilde{\Pi}, \Pi^*)] \leq \mathbb{E}[\mathrm{d}(\widetilde{\Pi}, \widehat{\Pi})] + \mathbb{E}[\mathrm{d}(\Pi^*, \widehat{\Pi})] = 2\mathbb{E}[\mathrm{d}(\Pi^*, \widehat{\Pi})],$$

which shows that $\widetilde{\Pi}$ is optimal within a factor of two. Thus it suffices to bound $\mathbb{E}[\mathrm{d}(\widetilde{\Pi}, \Pi^*)]$ from below.

To this end, fix some δ to be specified later and define the sets of good and bad solutions respectively as

$$\begin{aligned}\mathbf{\Pi}_{\mathsf{good}} &= \{\Pi \in \mathfrak{S}_n : \mathrm{d}(\Pi, \Pi^*) < \delta n\},\\ \mathbf{\Pi}_{\mathsf{bad}} &= \{\Pi \in \mathfrak{S}_n : \mathrm{d}(\Pi, \Pi^*) \geq \delta n\}.\end{aligned}$$

By the definition of $\widetilde{\Pi}$, we have

$$\mathbb{E}[\mathrm{d}(\widetilde{\Pi}, \Pi^*)] \geq \delta n \cdot \mathbb{E}[\mu_{X,Y}(\mathbf{\Pi}_{\mathsf{bad}})].$$

Next we show two key lemmas, which bound the posterior mass of $\mathbf{\Pi}_{\mathsf{good}}$ and $\mathbf{\Pi}_{\mathsf{good}}$ from above and below, respectively.

Lemma 15 *Assume $\sigma = \sigma_0 n^{-1/d}$ for any constant $\sigma_0 \in (0, 1/2)$. For any constant δ such that $\delta \leq 16(2\sigma_0)^d$, with probability at least $1 - 4\delta n e^{-\delta n/\log n}$,*

$$\frac{\mu_{X,Y}(\mathbf{\Pi}_{\mathsf{good}})}{\mu_{X,Y}(\Pi^*)} \leq 2\left(\frac{16e^2(2\sigma_0)^d}{\delta}\right)^{\delta n}. \tag{74}$$

Lemma 16 *Assume $\sigma = \sigma_0 n^{-1/d}$ for some constant σ_0. There exist constants $\delta_0(\sigma_0, d)$ and $c(\sigma_0, d)$ that only depend on σ_0, d such that for all $\delta \leq \delta_0$ and sufficiently large n, with probability at least $1/2 - c/n$,*

$$\frac{\mu_{X,Y}(\mathbf{\Pi}_{\mathsf{bad}})}{\mu_{X,Y}(\Pi^*)} \geq e^{\delta_0 n/2}. \tag{75}$$

Given the above two lemmas, Theorem 14 readily follows. Indeed, combining Lemma 15 and Lemma 16 and choosing δ such that $\delta \log(16e^2(2\sigma_0)^d/\delta) = \delta_0/4$ we get $\mu_{X,Y}(\mathbf{\Pi}_{\mathsf{bad}}) \geq \frac{e^{\delta_0 n/4}}{2+e^{\delta_0 n/4}}$ with probability at least $1/2 - c/n - 4\delta n e^{-\delta n/\log n}$, which shows that $\mathbb{E}[\mathrm{d}(\widetilde{\Pi}, \Pi^*)] \gtrsim \delta n$ as desired.

E.2. Upper bounding the posterior mass of good permutations

In this section, we prove Lemma 15 by a truncated first moment calculation. We need the following key auxiliary result.

Lemma 17 *Assume that* $n(2\sigma)^d \le 1$. *Then for any* $\ell \in [0, n]$,

$$\sum_{\Pi:\mathrm{d}(\Pi,\Pi^*)=\ell} \mathbb{E}\exp\left(-\frac{1}{8\sigma^2}\|\Pi X - \Pi^* X\|_{\mathrm{F}}^2\right) \le \left(\frac{16n^2(2\sigma)^d}{\ell}\right)^{\ell/2}.$$

Proof It follows from (30) in Lemma 7 that

$$\mathbb{E}\exp\left(-\frac{1}{8\sigma^2}\|\Pi X - \Pi^* X\|_{\mathrm{F}}^2\right) \le \prod_{k=1}^{n}\left[(2\sigma)^{k-1}\right]^{dn_k} \le (2\sigma)^{d(\ell-\mathfrak{c}(\widetilde{\pi}))},$$

where $\ell = n - n_1$ is the number of non-fixed points, $\widetilde{\pi}$ is the restriction of the permutation π on its non-fixed points, and $\mathfrak{c}(\widetilde{\pi})$ denotes the number of cycles of $\widetilde{\pi}$. It follows that

$$\begin{aligned}
\sum_{\Pi:\mathrm{d}(\Pi,\Pi^*)=\ell} \mathbb{E}\exp\left(-\frac{1}{8\sigma^2}\|\Pi X - \Pi^* X\|_{\mathrm{F}}^2\right) &\le \binom{n}{\ell}\frac{\ell!}{L^\ell}\mathbb{E}_\tau\left[L^{\mathfrak{c}(\tau)}\mathbb{1}_{\{\tau \text{ is a derangement}\}}\right] \\
&\le \left(\frac{n}{L}\right)^\ell \mathbb{E}_\tau\left[L^{\mathfrak{c}(\tau)}\mathbb{1}_{\{\tau \text{ is a derangement}\}}\right] \\
&\le \left(\frac{16n^2}{\ell L}\right)^{\ell/2},
\end{aligned}$$

where $L = (2\sigma)^{-d}$, the expectation $\mathbb{E}_\tau$ is taken for a uniformly random permutation $\tau \in S_\ell$, and the last inequality follows from (49). ∎

Proof [Proof of Lemma 15] Note that

$$\frac{\mu_{X,Y}(\mathbf{\Pi}_{\mathsf{good}})}{\mu_{X,Y}(\Pi^*)} = \sum_{\Pi\in\mathbf{\Pi}_{\mathsf{good}}} e^{L(\Pi)-L(\Pi^*)} = R_1 + R_2,$$

where

$$\begin{aligned}
R_1 &\triangleq \sum_{\Pi:\mathrm{d}(\Pi,\Pi^*)<\beta n/\log n} e^{L(\Pi)-L(\Pi^*)} \\
R_2 &\triangleq \sum_{\Pi:\frac{\beta n}{\log n}\le\mathrm{d}(\Pi,\Pi^*)<\delta n} e^{L(\Pi)-L(\Pi^*)}
\end{aligned}$$

for some β to be specified. Next we bound R_1 and R_2 separately.

First, the number of permutations Π such that $\Pi^{-1}\circ\Pi^*$ has ℓ non-fixed points is

$$|\{\Pi \in \mathfrak{S}_n : \mathrm{d}(\Pi,\Pi^*) = \ell\}| = !\ell \cdot \binom{n}{\ell}, \tag{76}$$

where $!\ell = \left[\frac{\ell!}{e}\right]$. Thus

$$\frac{1}{2e}n(n-1)\cdots(n-\ell+1) \le |\{\Pi \in \mathfrak{S}_n : \mathrm{d}(\Pi,\Pi^*) = \ell\}| \le \frac{2}{e}n(n-1)\cdots(n-\ell+1). \quad (77)$$

Furthermore, for any Π,

$$\begin{aligned}\mathbb{E}e^{L(\Pi)-L(\Pi^*)} &= \mathbb{E}\exp\left(\frac{1}{\sigma^2}\langle \Pi X - \Pi^* X, Y\rangle\right)\\ &= \mathbb{E}\exp\left(\frac{1}{\sigma^2}\langle \Pi X - \Pi^* X, \Pi^* X\rangle + \frac{1}{2\sigma^2}\|\Pi X - \Pi^* X\|_{\mathrm{F}}^2\right)\\ &= 1, \end{aligned} \quad (78)$$

where the first equality holds due to $Y = \Pi^* X + \sigma^2 Z$ and $\mathbb{E}\exp(\langle A, Z\rangle) = \exp(\|A\|_{\mathrm{F}}^2/2)$ and the second equality follows from $\langle \Pi X - \Pi^* X, \Pi^* X\rangle = -\frac{1}{2}\|\Pi X - \Pi^* X\|_{\mathrm{F}}^2$.

To bound R_1, using (77) and (78) we have

$$\mathbb{E}R_1 = \sum_{\mathrm{d}(\Pi,\Pi^*)<\frac{\beta n}{\log n}} \mathbb{E}e^{L(\Pi)-L(\Pi^*)} \le \sum_{\ell<\frac{\beta n}{\log n}} \frac{2}{e}n^\ell \le \frac{2\beta n}{e\log n}\exp(\beta n).$$

By Markov's inequality,

$$\mathbb{P}\left\{R_1 \ge e^{2\beta n}\right\} \le \frac{2n}{e}\exp(-\beta n). \quad (79)$$

To bound R_2, the calculation above shows that directly applying the Markov inequality is too crude since $\mathbb{E}[R_2] = e^{\Theta(n\log n)}$. Note that although $L(\Pi) - L(\Pi^*)$ is negatively biased, when $L(\Pi) - L(\Pi^*)$ is atypically large it results in an excessive contribution to the exponential moments. Thus we truncate on the following event:

$$\mathcal{E} \triangleq \bigcap_{\Pi:\frac{\beta n}{\log n}\le \mathrm{d}(\Pi,\Pi^*)<\delta n} \{L(\Pi) - L(\Pi^*) \le \tau(\mathrm{d}(\Pi,\Pi^*))\}$$

for some threshold $\tau(\ell)$ to be chosen.

Then for any $c' > 0$,

$$\begin{aligned}&\mathbb{P}\left\{R_2 \ge e^{c'n}\right\}\\ &\le \mathbb{P}\{\mathcal{E}^c\} + \mathbb{P}\left\{\{R_2 \ge e^{c'n}\}\cap\mathcal{E}\right\}\\ &\le \mathbb{P}\{\mathcal{E}^c\} + \mathbb{P}\left\{\sum_{\frac{\beta n}{\log n}\le \mathrm{d}(\Pi,\Pi^*)<\delta n} e^{L(\Pi)-L(\Pi^*)}\mathbf{1}_{\{L(\Pi)-L(\Pi^*)\le\tau(\mathrm{d}(\Pi,\Pi^*))\}} \ge e^{c'n}\right\}\\ &\le \mathbb{P}\{\mathcal{E}^c\} + e^{-c'n}\sum_{\frac{\beta n}{\log n}\le \mathrm{d}(\Pi,\Pi^*)<\delta n} \mathbb{E}e^{L(\Pi)-L(\Pi^*)}\mathbf{1}_{\{L(\Pi)-L(\Pi^*)\le\tau(\mathrm{d}(\Pi,\Pi^*))\}}. \end{aligned} \quad (80)$$

To bound the first term, note that for any $t > 0$,

$$\begin{aligned}&\mathbb{P}\{L(\Pi) - L(\Pi^*) \ge \tau\}\\ &\qquad \le e^{-t\tau}\mathbb{E}\exp\left(\frac{t}{\sigma^2}\langle \Pi X - \Pi^* X, Y\rangle\right) = e^{-t\tau}\mathbb{E}\exp\left(\frac{t^2-t}{2\sigma^2}\|\Pi X - \Pi^* X\|_{\mathrm{F}}^2\right).\end{aligned}$$

By choosing $t = 1/2$, we get that

$$\mathbb{P}\left\{L(\Pi) - L(\Pi^*) \geq \tau\right\} \leq e^{-\tau/2}\mathbb{E}\exp\left(-\frac{1}{8\sigma^2}\|\Pi X - \Pi^* X\|_{\mathrm{F}}^2\right).$$

Recall from Lemma 17, we have that

$$\sum_{\Pi:\mathrm{d}(\Pi,\Pi^*)=\ell} \mathbb{E}\exp\left(-\frac{1}{8\sigma^2}\|\Pi X - \Pi^* X\|_{\mathrm{F}}^2\right) \leq \left(\frac{16n^2(2\sigma)^d}{\ell}\right)^{\ell/2} = \left(\frac{16n(2\sigma_0)^d}{\ell}\right)^{\ell/2}.$$

Therefore, it follows from a union bound that

$$\begin{aligned}
\mathbb{P}\left\{\mathcal{E}^c\right\} &= \sum_{\frac{\beta n}{\log n} \leq \mathrm{d}(\Pi,\Pi^*) < \delta n} \mathbb{P}\left\{L(\Pi) - L(\Pi^*) \geq \tau\left(\mathrm{d}(\Pi,\Pi^*)\right)\right\} \\
&\leq \sum_{\frac{\beta n}{\log n} \leq \ell < \delta n} e^{-\tau(\ell)/2}\left(\frac{16n(2\sigma_0)^d}{\ell}\right)^{\ell/2} \\
&= \sum_{\frac{\beta n}{\log n} \leq \ell < \delta n} e^{-\ell} \leq \delta n e^{-\frac{\beta n}{\log n}}, \qquad (81)
\end{aligned}$$

where the last equality holds by choosing $\tau(\ell) = \ell\log(16e^2 n(2\sigma_0)^d/\ell)$.

For the second term in (81), we bound the truncated MGF as follows:

$$\begin{aligned}
&\sum_{\Pi:\mathrm{d}(\Pi,\Pi^*)=\ell} \mathbb{E}e^{L(\Pi)-L(\Pi^*)}\mathbf{1}_{\{L(\Pi)-L(\Pi^*)\leq\tau(\mathrm{d}(\Pi,\Pi^*))\}} \\
&\leq \sum_{\Pi:\mathrm{d}(\Pi,\Pi^*)=\ell} \mathbb{E}\exp\left(\frac{1}{2}\left(L(\Pi) - L(\Pi^*) + \tau(\ell)\right)\right) \\
&\leq \sum_{\Pi:\mathrm{d}(\Pi,\Pi^*)=\ell} \mathbb{E}\exp\left(-\frac{1}{8\sigma^2}\|\Pi X - \Pi^* X\|_{\mathrm{F}}^2\right)e^{\tau(\ell)/2} \\
&\leq \left(\frac{16n(2\sigma_0)^d}{\ell}\right)^{\ell/2} e^{\tau(\ell)/2} \\
&\leq \left(\frac{16en(2\sigma_0)^d}{\ell}\right)^{\ell}.
\end{aligned}$$

It follows that

$$\begin{aligned}
\sum_{\frac{\beta n}{\log n} \leq \mathrm{d}(\Pi,\Pi^*) < \delta n} \mathbb{E}e^{L(\Pi)-L(\Pi^*)}\mathbf{1}_{\{L(\Pi)-L(\Pi^*)\leq r\mathrm{d}(\Pi,\Pi^*)\}} &\leq \sum_{\frac{\beta n}{\log n} \leq \ell < \delta n}\left(\frac{16en(2\sigma_0)^d}{\ell}\right)^{\ell} \\
&\leq \delta n\left(\frac{16e(2\sigma_0)^d}{\delta}\right)^{\delta n},
\end{aligned}$$

where the last inequality holds for all $\delta \le 16(2\sigma_0)^d$. Choosing $c' = \delta \log(16e^2(2\sigma_0)^d/\delta)$, we get that

$$e^{-c'n} \sum_{\frac{\beta n}{\log n} \le \mathrm{d}(\Pi,\Pi^*) < \delta n} \mathbb{E} e^{L(\Pi)-L(\Pi^*)} \mathbf{1}_{\{L(\Pi)-L(\Pi^*) \le r\mathrm{d}(\Pi,\Pi^*)\}} \le \delta n e^{-\delta n} \tag{82}$$

Substituting (81) and (82) into (80), we get

$$\mathbb{P}\left\{ R_2 \ge \left(\frac{16e^2(2\sigma_0)^d}{\delta} \right)^{\delta n} \right\} \le 2\delta n e^{-\beta n/\log n}.$$

Combining this with (79) and upon choosing $\beta = \delta$, we have

$$\mathbb{P}\left\{ R_1 + R_2 \ge 2\left(\frac{16e^2(2\sigma_0)^d}{\delta} \right)^{\delta n} \right\} \le 4\delta n e^{-\delta n/\log n},$$

concluding the proof. ■

E.3. Lower bounding the posterior mass of bad permutations

In this section, we prove Lemma 16. We aim to construct exponentially many bad permutations π whose log likelihood $L(\pi)$ is no smaller than $L(\pi^*)$. It turns out that $L(\pi) - L(\pi^*)$ can be decomposed according to the orbit decomposition of $(\pi^*)^{-1} \circ \pi$ as per (14). Thus, following Ding et al. (2021b), we look for vertex-disjoint orbits O whose total lengths add up to $\Omega(n)$ and each of them is *augmenting* in the sense that $\Delta(O) \ge 0$.

In the planted matching model with independent weights Ding et al. (2021b), a great challenge lies in the fact that short augmenting orbits (even after taking their disjoint unions) are insufficient to meet the $\Omega(n)$ total length requirement. As a result, one has to search for long augmenting orbits of length $\Omega(n)$. However, due to the excessive correlations among long augmenting orbits, the second-moment calculation fundamentally fails. To overcome this challenge, Ding et al. (2021b) invents a two-stage finding scheme which first finds many but short augmenting paths and then patches them together to form a long augmenting orbit using the so-called sprinkling idea. Fortunately, in our low-dimensional case of $d = \Theta(1)$, as also observed in Kunisky and Niles-Weed (2022), it suffices to look for augmenting 2-orbits and take their disjoint unions. More precisely, the following lemma shows that there are $\Omega(n)$ vertex-disjoint augmenting 2-orbits, from which we can easily extract exponentially many different unions of total length $\Omega(n)$. In contrast, to prove the failure of the MLE for almost perfect recovery in Kunisky and Niles-Weed (2022), a single union of $\Omega(n)$ vertex-disjoint augmenting 2-orbits is sufficient.

Lemma 18 *If $\sigma = \sigma_0 n^{-1/d}$, then there exist constants $c(\sigma_0, d)$, $\delta_0(\sigma_0, d)$, and $n_0(\sigma_0, d)$ that only depend on σ_0 and d such that for all $n \ge n_0$, with probability at least $1/2 - c/n$, there are at least $\delta_0 n$ many vertex-disjoint augmenting 2-orbits.*

This lemma is proved in (Kunisky and Niles-Weed, 2022, Section 4) using the so-called concentration-enhanced second-moment method. For completeness, here we provide a much simpler proof via the vanilla second-moment method combined with Turán's theorem.

Proof Let I_{ij} denote the indicator that (i,j) is an augmenting 2-orbit and $I = \sum_{i<j} I_{ij}$. To extract a collection of vertex-disjoint augmenting 2-orbits, we construct a graph $G = (V, E)$, where the vertices correspond to (i,j) for which $I_{ij} = 1$, and (i,j) and (k,ℓ) are connected if (i,j) and (k,ℓ) share a common vertex. By construction, any collection of vertex-disjoint 2-orbits corresponds to an independent set in G. By Turán's theorem (see e.g. (Alon and Spencer, 2008, Theorem 1, p. 95)), there exists an independent set S in G of size at least $|V|^2/(2|E| + |V|)$. It remains to bound $|V|$ from below and $|E|$ from above.

Note that $|V| = I = \sum_{i<j} I_{ij}$. For all n sufficiently large, $\sigma^2 \le d/40$ and it follows from (Kunisky and Niles-Weed, 2022, Prop. 4.3) that

$$p \triangleq \mathbb{P}\left\{I_{ij} = 1\right\} \ge \frac{1}{1000\sqrt{d}} \left(1 + \frac{1}{\sigma^2}\right)^{-d/2}.$$

Therefore,

$$\mathbb{E}I = \sum_{i<j} \mathbb{P}\left\{I_{ij} = 1\right\} \ge \binom{n}{2} \frac{1}{1000\sqrt{d}} \left(1 + \frac{1}{\sigma^2}\right)^{-d/2}. \tag{83}$$

Under the assumption that $\sigma = \sigma_0 n^{-1/d}$, it follows that $\mathbb{E}I \ge c_0(d,\sigma_0)n$ for some constant $c_0(d,\sigma_0)$ that only depends on d and σ_0. Moreover,

$$\begin{aligned}
\mathsf{var}(I) &= \sum_{i<j,k<\ell} \mathrm{Cov}\left(I_{ij}, I_{k\ell}\right) \\
&= \sum_{i<j} \mathsf{var}(I_{ij}) + \sum_{i<j} \sum_{k:k\neq i,j} \left(\mathrm{Cov}\left(I_{ij}, I_{ik}\right) + \mathrm{Cov}\left(I_{ij}, I_{jk}\right)\right) \\
&\le \sum_{i<j} \mathbb{E}I_{ij}^2 + \sum_{i<j} \sum_{k:k\neq i,j} \left(\mathbb{E}I_{ij}I_{ik} + \mathbb{E}I_{ij}I_{jk}\right),
\end{aligned}$$

where the second equality holds because I_{ij} and $I_{k\ell}$ are independent when $\{i,j\} \cap \{k,\ell\} = \emptyset$. Recall that $\mathbb{E}I_{ij}^2 = \mathbb{E}I_{ij} = p$. Moreover, it follows from (Kunisky and Niles-Weed, 2022, Prop. 4.5) that

$$\mathbb{E}I_{ij}I_{ik} \le \left(1 + \frac{3}{4\sigma^2}\right)^{-d}.$$

Combining the last three displayed equation yields that

$$\mathsf{var}(I) \le \mathbb{E}I + n^3 \left(1 + \frac{3}{4\sigma^2}\right)^{-d}. \tag{84}$$

Under the assumption that $\sigma = \sigma_0 n^{-1/d}$, it follows that $\mathsf{var}(I) \le \mathbb{E}I + c_1(d,\sigma_0)n$ for some $c_1(d,\sigma_0)$ that only depends on d and σ_0. By Chebyshev's inequality,

$$\mathbb{P}\left\{I \le \frac{1}{2}\mathbb{E}I\right\} \le \frac{4\mathsf{var}(I)}{\left(\mathbb{E}I\right)^2} \le \frac{4(c_0 + c_1)}{c_0^2 n}.$$

Moreover,

$$|E| = \sum_{i<j} \sum_{k:k\neq i,j} \left(I_{ij}I_{ik} + I_{ij}I_{jk}\right)$$

and hence

$$\mathbb{E}|E| = \sum_{i<j} \sum_{k:k\neq i,j} (\mathbb{E}I_{ij}I_{ik} + \mathbb{E}I_{ij}I_{jk}) \leq n^3 \left(1 + \frac{3}{4\sigma^2}\right)^{-d} \leq c_1(d,\sigma_0)n.$$

By Markov's inequality, $|E| \leq 2\mathbb{E}|E|$ with probability at least $1/2$. Therefore, with probability at least $1/2 - 4c_1/(c_0^2 n)$,

$$|S| \geq \frac{|V|^2}{2|E| + |V|} \geq \frac{(\mathbb{E}I)^2/4}{4\mathbb{E}|E| + \mathbb{E}I/2} \geq \frac{c_0^2 n^2/4}{4c_1 n + c_0 n/2} \geq \delta_0 n,$$

for some constant $\delta_0(d,\sigma_0)$ that only depends on d and σ_0. ■

Proof [Proof of Lemma 16] By Lemma 18, from $\delta_0 n$ such vertex-disjoint augmenting 2-orbits, we choose $\delta_0 n/2$ many of them and form a union of augmenting 2-orbits with the total length $\delta_0 n/2 \times 2 = \delta_0 n$. There are $\binom{\delta_0 n}{\delta_0 n/2}$ many different unions, and each of such union corresponds to a permutation Π with $\mathrm{d}(\Pi, \Pi^*) = \delta_0 n$ and $L(\Pi) \geq L(\Pi^*)$ in view of (14). Therefore, for any $\delta \leq \delta_0$,

$$\frac{\mu_{X,Y}(\mathbf{\Pi}_{\mathsf{bad}})}{\mu_{X,Y}(\Pi^*)} \geq \binom{\delta_0 n}{\delta_0 n/2} \geq 2^{\delta_0 n/2}.$$

■

E.4. Impossibility of perfect recovery

In this section, we prove an impossibility condition of perfect recovery.

Theorem 19 *Suppose that* $\sigma^2 \leq d/40$ *and*

$$\frac{d}{4}\log\left(1 + \frac{1}{\sigma^2}\right) - \log n + \log d \leq C, \tag{85}$$

for a constant $C > 0$. *Then there exists a constant* c *that only depends on* C *such that for any estimator* $\widehat{\pi}$, $\mathbb{P}\{\widehat{\pi} \neq \pi^*\} \geq c$.

Theorem 19 immediately implies that if there exists an estimator that achieves perfect recovery with high probability, then

$$\frac{d}{4}\log\left(1 + \frac{1}{\sigma^2}\right) - \log n + \log d \to +\infty. \tag{86}$$

In comparison, it is shown in (Dai et al., 2019, Theorem 1) that perfect recovery is possible if $\frac{d}{4}\log\left(1 + \frac{1}{\sigma^2}\right) - \log n \to +\infty$. Thus our necessary condition agrees with their sufficient condition up to an additive $\log d$ factor. Our necessary condition (86) further specializes to

- $d \ll \log n$:

$$\sigma \leq \begin{cases} o(n^{-2/d}) & \text{if } d = O(1) \\ n^{-2/d} & \text{if } d \gg 1 \end{cases}.$$

  This yields Theorem 3(i) and slightly improves over the necessary condition of MLE in (Kunisky and Niles-Weed, 2022, Theorem 1.1), that is, $\sigma = O(n^{-2/d})$.

- $d = \Theta(\log n)$:
$$\sigma \le \frac{1}{\sqrt{n^{4/d} - 1}};$$
- $d \gg \log n$:
$$\sigma \le \sqrt{\frac{d}{4\log(n/d) + \omega(1)}}.$$

Note that the previous work Dai et al. (2019) shows that $\frac{d}{4}\log\left(1 + \frac{1}{\sigma^2}\right) \ge (1 - \Omega(1))\log n$ is necessary for perfect recovery, under the additional assumption that $1 \ll d = O(\log n)$. The analysis therein is based on showing the existence of an augmenting 2-orbit via the second-moment method. We follow a similar strategy, but our first and second moment estimates are sharper and thus yield a tighter condition.

Proof Recall that I_{ij} denote the indicator that (i, j) is an augmenting 2-orbit and $I = \sum_{i<j} I_{ij}$. For the purpose of lower bound, consider the Bayesian setting where π^* is drawn uniformly at random. Then the MLE $\widehat{\pi}_{\mathrm{ML}}$ given in (2) minimizes the probability of error. Hence, it suffices to bound from below $\mathbb{P}\{\widehat{\pi}_{\mathrm{ML}} \neq \pi^*\}$. Note that on the event $\{I > 0\}$, there exists at least one permutation $\pi \neq \pi^*$ whose likelihood is at least as large as that of π^* and hence the error probability of MLE is at least $1/2$. Therefore,
$$\mathbb{P}\{\widehat{\pi}_{\mathrm{ML}} \neq \pi^*\} \ge \frac{1}{2}\mathbb{P}\{I > 0\}.$$
It remains to bound $\mathbb{P}\{I > 0\}$ from below. To this end, we first bound $\mathsf{var}(I)/(\mathbb{E}I)^2$. In view of (84),
$$\frac{\mathsf{var}(I)}{(\mathbb{E}I)^2} \le \frac{1}{\mathbb{E}I} + \frac{1}{(\mathbb{E}I)^2} n^3 \left(1 + \frac{3}{4\sigma^2}\right)^{-d}.$$
By assumption $\sigma^2 \le d/40$ and (85), it follows from (83) that
$$\mathbb{E}I \gtrsim \frac{n^2}{\sqrt{d}}\left(1 + \frac{1}{\sigma^2}\right)^{-d/2} \ge \exp\left(\frac{3}{2}\log d - 2C\right) \ge \exp(-2C).$$
Moreover,
$$\frac{1}{(\mathbb{E}I)^2} n^3 \left(1 + \frac{3}{4\sigma^2}\right)^{-d} \lesssim \frac{d}{n}\left(\frac{1 + 1/\sigma^2}{1 + 3/(4\sigma^2)}\right)^d \overset{(a)}{\le} \frac{d}{n}\left(1 + \frac{1}{\sigma^2}\right)^{d/4} \overset{(b)}{\le} e^C,$$
where (a) holds because $1 + 3x/4 \ge (1 + x)^{3/4}$ for all $x \ge 0$ and (b) holds due to assumption (85),

Combining the last three displayed equation yields that $\mathsf{var}(I)/(\mathbb{E}I)^2 \le c_0$ for some constant c_0 that only depends on C. By the Paley-Zygmund inequality,
$$\mathbb{P}\{I > 0\} \ge \mathbb{P}\left\{I \ge \frac{1}{2}\mathbb{E}I\right\} \ge \frac{(\mathbb{E}I)^2}{4\left(\mathsf{var}(I) + (\mathbb{E}I)^2\right)} \ge \frac{1}{4c_0 + 1}.$$
■

Appendix F. Recovery thresholds in the nonisotropic case

In this section we argue that Theorem 1 continues to hold under the same conditions in the nonisotropic case of $X_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Sigma)$, provided that $\Sigma \succ cI$ for some absolute constant $c > 0$. In the general nonisotropic case, we denote by $p(\Sigma, \sigma, \Pi, Q)$ the moment generating function given by (27) to highlight the dependency on the covariance matrix Σ and the noise level σ. As in the proof of Lemma 8, recall $x = \mathsf{vec}(X)$ denotes the vectorization of X. Since $X_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Sigma)$, the vector $x \in \mathbb{R}^{nd}$ has distribution $x \sim \mathcal{N}(0, I_n \otimes \Sigma)$. Note that $I_n \otimes \Sigma \succ cI_{nd}$. Modifying (32) accordingly, we have

$$p(\Sigma, \sigma, \Pi, Q) = \mathbb{E} \exp\left(-\frac{1}{32\sigma^2} x^\top H^\top H x\right) = \left[\det\left(I + \frac{1}{16\sigma^2} H^\top H (I_n \otimes \Sigma)\right)\right]^{-\frac{1}{2}}$$
$$\leq \left[\det\left(I + \frac{c}{16\sigma^2} H^\top H\right)\right]^{-\frac{1}{2}} = p(I, \sigma', \Pi, Q),$$

where $H = I_{nd} - Q^\top \otimes \Pi$ and $\sigma' = \sigma/\sqrt{c}$. This shows that the MGF $p(\Sigma, \sigma, \Pi, Q)$ satisfies the same estimates (30), (31) and Lemma 8 for the isotropic case with the original noise σ replaced by a constant multiple of it σ'. This constant multiplicative factor keeps σ' satisfying the same noise threshold in Theorem 1, which implies both prefect recovery and almost perfect recovery can still be achieved for the nonisotropic case under the same conditions, hence confirming our claim in Section 4.