Phase Transitions in the Detection of Correlated Databases

Dor Elimelech [* 1] **Wasim Huleihel** [* 2]

Abstract

We study the problem of detecting the correlation between two Gaussian databases $X \in \mathbb{R}^{n \times d}$ and $Y^{n \times d}$, each composed of n users with d features. This problem is relevant in the analysis of social media, computational biology, etc. We formulate this as a hypothesis testing problem: under the null hypothesis, these two databases are statistically independent. Under the alternative, however, there exists an unknown permutation σ over the set of n users (or, row permutation), such that X is ρ-correlated with Y^σ, a permuted version of Y. We determine sharp thresholds at which optimal testing exhibits a phase transition, depending on the asymptotic regime of n and d. Specifically, we prove that if $\rho^2 d \to 0$, as $d \to \infty$, then weak detection (performing slightly better than random guessing) is statistically impossible, *irrespectively* of the value of n. This compliments the performance of a simple test that thresholds the sum all entries of $X^T Y$. Furthermore, when d is fixed, we prove that strong detection (vanishing error probability) is impossible for any $\rho < \rho^\star$, where $\rho^\star$ is an explicit function of d, while weak detection is again impossible as long as $\rho^2 d = o(1)$, as $n \to \infty$. These results close significant gaps in current recent related studies.

1. Introduction

Database alignment and the quantification of the relation between different databases are among the most fundamental tasks in modern applications of statistics. In many cases, the observed databases are high-dimensional, unlabeled, noisy, and scrambled, making the task of inference challenging.

An example of such an inference task between, say, two databases, formulated as an hypothesis testing problem, is the following: under the null hypothesis, the databases are statistically uncorrelated, while under the alternative, there exists a permutation which scrambles one database, and for which the two databases are correlated. Then, observing the databases, under what conditions can we tell if they are correlated or not?

It turns out that the above general inference formulation is relevant in many fields, such as, computational biology (Singh et al., 2008; Kang et al., 2012), social network analysis (Narayanan & Shmatikov, 2008; 2009), computer vision (Berg et al., 2005; Cour et al., 2006), and data anonymization/privacy based systems. A concrete classical example is: consider two data sources (e.g., Netflix and IMDb), each supplying lists of features for a set of entities, say, users. Those features might be various characteristics of those users, such as, names, user identifications, ratings. In many cases, reliable labeling for features is either not available or deleted (so as to hide sensitive unique identifying information) due to privacy concerns. In general, this precludes trivial identification of feature pairs from the two sources that correspond to the same user. Nonetheless, the hope is that if the correlation between the two databases is sufficiently large, then it is possible to identify correspondences between the two databases, and generate an alignment between the feature lists (Narayanan & Shmatikov, 2008; 2009).

Quite recently, the *data alignment problem*, a seemingly simple probabilistic model which captures the scenario above, was introduced and investigated in (Cullina et al., 2018; Dai et al., 2019). Specifically, there are two databases $X \in \mathbb{R}^{n \times d}$ and $Y^{n \times d}$, each composed of n users with d features. There exist an unknown permutation/correspondence which match users in X to users in Y. For a pair of matched database entries, the features are dependent according to a known distribution, and, for unmatched entries, the features are independent. The goal is to *recover* the unknown permutation, and derive statistical guarantees for which recovery is possible and impossible, as a function of the correlation level, n, and d. Roughly speaking, this recovery problem is well-understood for a wide family of probability distributions. For example, in the Gaussian case, denoting the correlation coefficient by ρ, it was shown in (Dai et al., 2019) that if $\rho^2 = 1 - o(n^{-4/d})$ then perfect recovery is possible, while

[*]Equal contribution [1]School of Electrical and Computer Engineering, Ben-Gurion university, Beer Sheva 84105, Israel. [2]Department of Electrical Engineering-Systems, Tel Aviv university, Tel Aviv 6997801, Israel.. Correspondence to: Dor Elimelech <doreli@post.bgu.ac.il>, Wasim Huleihel <wasimh@tauex.tau.ac.il>.

impossible if $\rho^2 = 1 - \omega(n^{-4/d})$, as $n, d \to \infty$.

The *detection* counterpart of the above recovery problem was also investigated in (K & Nazer, 2022a;b). Here, as mentioned above, the underlying question is, given two databases, can we determine whether they are correlated? It was shown that if $\rho^2 d \to \infty$ then efficient detection is possible (with vanishing error probability), simply by thresholding the sum of entries of $X^T Y$. On the other hand, it was also shown that if $\rho^2 d\sqrt{n} \to 0$ and $d = \Omega(\log n)$, then detection is information-theoretically impossible (i.e., lower bound), again, as $n, d \to \infty$. Unfortunately, it is evident that there is a substantial gap between those two bounds. Most notably, the aforementioned upper bound is completely independent of n, implying that it does not play any significant role in the detection problem, while the lower bound depends on n strongly. This sets precisely the main goal of this paper: *we would like to characterize the detection boundaries tightly, as a function of n and d.*

At first glance, one may suspect that the source for this gap is the proposed algorithm. Indeed, note that under the alternative distribution, the latent permutation represents a *hidden* correspondence under which the databases are correlated. Accordingly, the "thresholding the sum" approach ignores this hidden combinatorial structure, and therefore, seemingly suboptimal. However, it turns out that in the regime where $d \to \infty$, *independently* of n, this simple approach is actually surprisingly optimal: we prove that whenever $\rho^2 d \to 0$, then weak detection (performing slightly better than random guessing) is information-theoretically impossible, while if $\rho^2 \lesssim 1/d$ (that is, $\rho^2 \leq (1 - \varepsilon)/d$ for some $\varepsilon > 0$), then strong detection (vanishing error probability) is information-theoretically impossible. This behaviour, however, changes when d is fixed. In this case, we prove that strong detection is impossible for any $\rho < \rho^\star$, where $\rho^\star$ is an explicit function of d, while possible when $\rho = 1 - o(n^{-2/(d-1)})$. The later is achieved by counting the number of empirical pairwise correlations that exceed as certain threshold. Finally, we prove that weak detection (performing slightly better than random guessing) is impossible when $\rho^2 d \to 0$, while possible for any $\rho > \rho^{\star\star}$, where $\rho^{\star\star}$ is again an explicit function of d. The bounds from previous work and our new results described above are summarized in Table 1, shown at the top of the page.

We now mention other related work briefly. In (Dai et al., 2020) the problem of partial recovery of the permutation aligning the databases was analyzed. In (Shirani et al., 2019) necessary and sufficient conditions for successful recovery matching using a typicality-based framework were established. Furthermore, (Bakirtas & Erkip, 2021) and (Bakirtas & Erkip, 2022) proposed and explored the problem of permutation recovery under feature deletions and repetitions, respectively. Recently, the problem of joint correlation de-

tection and alignment of Gaussian database was analyzed in (Tamir, 2022). Finally, we note that the problem of database alignment and detection is closely related to a wide verity of planted matching problems, specifically, the graph alignment problem, with many exciting and interesting results, and useful mathematical techniques, see, e.g., (Moharrami et al., 2021; Ding et al., 2021; Pedarsani & Grossglauser, 2011; Ding et al., 2018; Wu et al., 2020; Mao et al., 2021; Wu et al., 2022; Ganassali, 2020), and many references therein.

Notation. For any $n \in \mathbb{N}$, the set of integers $\{1, 2, \ldots, n\}$ is denoted by $[n]$. Let $\mathbb{S}_n$ denotes the set of all permutations on $[n]$. For a given permutation $\sigma \in \mathbb{S}_n$, let σ_i denote the value to which σ maps $i \in [n]$. We use $\log$ to denote the natural logarithm function (i.e., of base e). Random vectors are denoted by capital letters such as X with transpose X^T. A collection of n random vectors is written as $\mathsf{X} = (X_1, \ldots, X_n)$. The notation $(X_1, \ldots, X_n) \sim P_X^{\otimes n}$ means that the random vectors $(X_1, \ldots, X_n)$ are independent and identically distributed (i.i.d.) according to P_X. We use $\mathcal{N}(\eta, \Sigma)$ to represent the multivariate normal distribution with mean vector η and covariance matrix Σ. Let $\mathsf{Poisson}(\lambda)$ denote the Poisson distribution with parameter λ. The $n \times n$ identity matrix is denoted by $I_{n \times n}$, and 0_d denotes the all zero d-dimensional column vector. Let $\mathcal{L}(Y)$ denote the law, that is, the probability distribution, of a random variable Y. For probability measures $\mathbb{P}$ and $\mathbb{Q}$, let $d_{\mathsf{TV}}(\mathbb{P}, \mathbb{Q}) = \frac{1}{2} \int |d\mathbb{P} - d\mathbb{Q}|$ denote the total variation distance. For a probability measure μ on a space Ω, we use $\mu^{\otimes d}$ for the product measure of μ (d times) on the product space Ω^d. For a measure $\nu \ll \mu$ (that is, a measure absolutely continuous with respect to μ), we denote (by abuse of notation) the Randon-Nikodym derivative ν with respect to μ by $\frac{\nu}{\mu}$. For functions $f, g : \mathbb{N} \to \mathbb{R}$, we say that $f = O(g)$ (and $f = \Omega(g)$) if there exists $c > 0$ such that $f(n) \leq cg(n)$ (and $f(n) \geq cg(n)$) for all n. We say that $f = o(g)$ if $\lim_{n \to \infty} f(n)/g(n) = 0$, and that $f = \omega(g)$ if $g = o(f)$.

2. Problem Formulation and Main Results

Probabilistic Model. Consider the following binary hypothesis testing problem. Under the null hypothesis $\mathcal{H}_0$, the Gaussian databases X and Y are generated independently with $X_1, \ldots, X_n, Y_1, \ldots, Y_n \sim \mathcal{N}(0_d, I_{d \times d})$. Let $\mathbb{P}_0$ denote the resulting distribution over (X, Y). Under the alternate hypothesis $\mathcal{H}_1$, the databases X and Y are correlated with permutation σ for some unknown $\sigma \in \mathbb{S}_n$ and some known correlation coefficient $\rho \in [-1, 1] \setminus \{0\}$. Namely,

$$
\begin{aligned}
\mathcal{H}_0 &: (X_1, Y_1), \ldots, (X_n, Y_n) \overset{\text{i.i.d}}{\sim} \mathcal{N}^{\otimes d}(0_2, I_{2 \times 2}) \\
\mathcal{H}_1 &: (X_1, Y_{\sigma_1}), \ldots, (X_n, Y_{\sigma_n}) \overset{\text{i.i.d}}{\sim} \mathcal{N}^{\otimes d}(0_2, \Sigma_\rho),
\end{aligned} \tag{1}
$$

	Weak Detection		Strong Detection	
Asymptotics	**Possible**	**Impossible**	**Possible**	**Impossible**
$n, d \to \infty$	$\Omega(d^{-1})^*$	$o(d^{-1})$	$\omega(d^{-1})^*$	$(1-\varepsilon)d^{-1}$
$d \to \infty$, n constant	$\Omega(d^{-1})^*$	$o(d^{-1})$	$\omega(d^{-1})^*$	$O(d^{-1})$
$n \to \infty$, d constant	$\rho^2 \geq \frac{60 \log 2}{d}^*$	$o(1)$	$1 - o(n^{-\frac{2}{d-1}})$	$\rho^\star(d)$

Table 1. A summary of our bounds on ρ^2, for weak and strong detection, as a function of the asymptotic regime. Bounds marked with $*$ follows from the upper-bound from (K & Nazer, 2022b).

for some permutation $\sigma \in \mathbb{S}_n$, where

$$\Sigma_\rho \triangleq \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}. \quad (2)$$

For a fixed $\sigma \in \mathbb{S}_n$, we denote the joint distribution measure of (X, Y) under the hypothesis $\mathcal{H}_1$ by $\mathbb{P}_{\mathcal{H}_1|\sigma}$. See Fig. 1 for a visual illustration of our probabilistic model.

Learning Problem. A test function for our problem is a function $\phi : \mathbb{R}^{n \times d} \times \mathbb{R}^{n \times d} \to \{0, 1\}$, designed to determine which of the hypothesis $\mathcal{H}_0, \mathcal{H}_1$ occurred. The risk of a test ϕ is defined as the sum of its Type-I and (worst-case) Type-II error probabilities, i.e.,

$$\mathsf{R}(\phi) \triangleq \mathbb{P}_{\mathcal{H}_0}[\phi(\mathsf{X}, \mathsf{Y}) = 1] + \max_{\sigma \in \mathbb{S}_n} \mathbb{P}_{\mathcal{H}_1|\sigma}[\phi(\mathsf{X}, \mathsf{Y}) = 0]. \quad (3)$$

The *minimax* risk for our hypothesis detection problem is

$$\mathsf{R}^\star \triangleq \inf_{\phi : \mathbb{R}^{n \times d} \times \mathbb{R}^{n \times d} \to \{0,1\}} \mathsf{R}(\phi). \quad (4)$$

We remark that R is a function of ρ, d, and n, however, we omit them from our notation for the benefit of readability.

Definition 2.1. A sequence $(\rho, d, n) = (\rho_k, d_k, n_k)_k$ is:

1. Admissible for strong detection if $\lim_{k \to \infty} \mathsf{R}^\star = 0$.
2. Admissible for weak detection if $\limsup_{k \to \infty} \mathsf{R}^\star < 1$.

Clearly, admissibility of strong detection implies the admissibility of weak detection.

While admissibility of strong detection clearly refers to the existence of algorithms that correctly detects with probability that tends to 1, weak detection implies the the existence of algorithms which are asymptotically better then randomly guessing which of the hypothesis occurred. A useful way to rule out the possibility of weak/strong detection is by considering the relaxed average-case problem, where the permutation is uniformly drawn rather then been arbitrary. In that case, the risk function is characterized by the total

variation distance between the null hypothesis distribution $\mathbb{P}_{\mathcal{H}_0}$ and the distribution under the alternative hypothesis $\mathbb{P}_{\mathcal{H}_1}$. In particular, it can be shown that,

$$d_{\mathsf{TV}}(\mathbb{P}_{\mathcal{H}_0}, \mathbb{P}_{\mathcal{H}_1}) = o(1) \implies \lim_{k \to \infty} \mathsf{R}^\star = 1, \quad (5)$$

and

$$d_{\mathsf{TV}}(\mathbb{P}_{\mathcal{H}_0}, \mathbb{P}_{\mathcal{H}_1}) \leq 1 - \Omega(1) \implies \liminf_{k \to \infty} \mathsf{R}^\star > 0, \quad (6)$$

which correspond to the impossibility of weak and strong detection, respectively. We further discuss the relations specified in (5) and (6) in Appendix B.

Main Results. In this section, we present our results concerning the thresholds for admissibility and impossibility of weak and strong detection, in different asymptotic regimes: (a) both n and d tends to infinity, (b) n is a constant and d tends to infinity, and (c) d is a constant and n tends to infinity. We begin with our upper-bounds. As was mentioned in the Introduction, (K & Nazer, 2022b) proposed the following simple test:

$$\phi_{\mathsf{sum}}(\mathsf{X}, \mathsf{Y}) \triangleq \mathbb{1}\left\{ \mathsf{sign}(\rho) \sum_{i,j=1}^{n} X_i^T Y_j > |\rho|\frac{dn}{2} \right\}. \quad (7)$$

We have the following result.

Theorem 2.2. *(K & Nazer, 2022b) Consider the detection problem in* (1). *Then,*

$$\mathsf{R}(\phi_{\mathsf{sum}}) \leq 2 \cdot \exp\left(-\frac{d\rho^2}{60} \right). \quad (8)$$

The implication of Theorem 2.2 is that if $\rho^2 = \omega(d^{-1})$, then ϕ_{sum} achieves strong detection. Furthermore, if $\rho^2 > \frac{60 \log 2}{d}$, then ϕ_{sum} achieves weak detection. Note that the upper bound in (8) is completely independent of n. We observe that in the case where d is constant, the bound (8) can never guarantee strong detection using ϕ_{sum}, not even
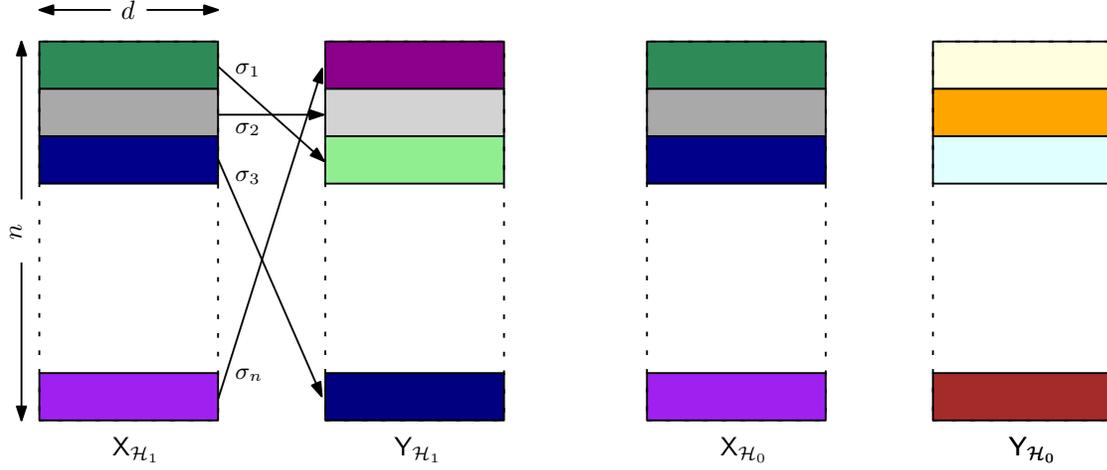
Figure 1. An illustration of the probabilistic model. On the left are the databases X and Y under the hypothesis $\mathcal{H}_1$, where correlated vectors are marked with a similar color. On the right are the uncorrelated databases under the null hypothesis.

in the trivial case where $\rho^2 = 1$ (where detection with zero risk is possible). It should be emphasized that the above phenomenon is inherent; it can be shown that the boundary $\rho^2 = \omega(d^{-1})$ associated with $\mathsf{R}(\phi_{\mathsf{sum}})$ cannot be improved and is not an artifact of the bounding technique used to establish (8).

Aiming for strong detection in the scenario where d is constant (and ρ^2 is smaller than unity), we propose the following alternative detection algorithm. For $i \in [n]$, let us define $\bar{X}_i \triangleq \frac{X_i}{\|X_i\|_2}$ and $\bar{Y}_i \triangleq \frac{Y_i}{\|Y_i\|_2}$. Also, let $\mathsf{W}_{ij}(\mathsf{X}, \mathsf{Y}) \triangleq \mathbb{1}\left\{\mathsf{sign}(\rho)\bar{X}_i^T\bar{Y}_j \geq |\rho|\right\}$, for $i, j \in [n]$. Then, we define the test

$$\phi_{\mathsf{count}}(\mathsf{X}, \mathsf{Y}) \triangleq \mathbb{1}\left\{\sum_{i,j=1}^{n} \mathsf{W}_{ij}(\mathsf{X}, \mathsf{Y}) \geq \frac{1}{2}n\mathcal{P}_{d,\rho}\right\}, \quad (9)$$

where $\mathcal{P}_{d,\rho} \triangleq \mathbb{P}_\rho\left(\bar{X}_1^T\bar{Y}_1 \geq \rho\right)$, and

$$\mathbb{P}_\rho \triangleq \mathcal{N}^{\otimes d}\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1 & \rho\\\rho & 1\end{bmatrix}\right). \quad (10)$$

We study the possibility and impossibility of our detection problem in multiple asymptotic regimes. These regimes are characterized by sequences of the parameters $(\rho, d, n) = (\rho_k, d_k, n_k)_{k \in \mathbb{N}}$. We consider the scenarios where d_k and n_k are either bounded or diverge to infinity. Accordingly, whenever we use asymptotic notations, such as, $\rho^2 = o(\cdot)$, $\rho^2 = \Omega(\cdot)$, etc., it should be understood in the context of the sequences above. For example, the condition $\rho^2 = o(d^{-1})$ means that the sequence (ρ, d, n) satisfies $\rho_k^2 d_k \to 0$, as $k \to \infty$.

The following theorem, shows that as long as ρ tends to one sufficiently fast, strong detection is possible.

Theorem 2.3. *Consider the detection problem in* (1) *and fix* $2 \leq d \in \mathbb{N}$. *Then,* $\mathsf{R}(\phi_{\mathsf{count}}) \to 0$, *as* $n \to \infty$, *if* $\rho^2 = 1 - o(n^{-\frac{2}{d-1}})$.

Prior to our work, for the case where d is fixed and independent of n, it was not clear if strong detection (vanishing error probabilities) can be even achieved. In Theorem, it is shown for the first time that strong detection is possible under non-trivial conditions (i.e, $\rho^2 < 1$). We also point out the fact that our analysis hols only whenever $d \geq 2$. The case where $d = 1$, on the other hand, remains a mystery. It turns out that many other alternative tests fail when $d = 1$ as well. We suspect that this phenomenon might be inherent to the probabilistic structure whenever $d = 1$, rather then an artifact of our algorithms. We leave this intriguing question open for future work.

Remark 2.4 (Recovery vs. detection). As mentioned in the introduction, the recovery problem of the permutation σ was considered in (Dai et al., 2019). It was shown that recovery is possible via the maximum-likelihood estimator if $\rho^2 = 1 - o(n^{-\frac{4}{d}})$, while recovery is impossible if $\rho^2 = 1 - \omega(n^{-\frac{4}{d}})$. Thus, Theorem 2.3 above, shows that the detection problem is statistically easier than recovery even when d is fixed. In fact, denoting the maximum-likelihood estimator by $\hat{\sigma}^{\mathsf{ML}}$, consider the following test

$$\phi_{\mathsf{max}}(\mathsf{X}, \mathsf{Y}) \triangleq \mathbb{1}\left\{\mathsf{sign}(\rho)\sum_{i=1}^{n} X_i^T Y_{\hat{\sigma}_i^{\mathsf{ML}}} > |\rho|\frac{dn}{2}\right\}. \quad (11)$$

It was claimed in (K & Nazer, 2022b) that as a corollary of (Dai et al., 2019), this test achieves strong detection under the same recovery guarantee, namely, if $\rho^2 = 1 - o(n^{-\frac{4}{d}})$. However, we suspect that this claim is in fact imprecise, as the authors overlooked the analysis of the Type-I error probability. Furthermore, it should be mentioned that (Tamir,

2022) proposed and analyzed a similar test as in (9). It was shown that in the case where d is constant, detection is possible if $\frac{\rho^2}{4} = 1 - o(n^{-\frac{4}{d}})$. However, this condition is clearly meaningless since $|\rho| \leq 1$. Finally, we also mention that in other regimes, the threshold for the recovery problem has a significantly different behaviour compared to the detection problem; for example, if $d = \omega(\log n)$, then the recovery barrier is $\frac{\log n}{d}$, i.e., recovery is possible (impossible) if $\rho^2 \gg \frac{\log n}{d}$ ($\rho^2 \ll \frac{\log n}{d}$). For detection, on the other hand, the barrier is $\frac{1}{d}$, independently of n.

In Theorem 2.5, we provide lower bounds, establishing thresholds for which weak detection is impossible.

Theorem 2.5 (Impossibility of weak detection). *Weak detection is impossible as long as $\rho^2 = o(d^{-1})$. That is, for a sequence $(\rho, d, n) = (\rho_k, d_k, n_k)_k$ such that $\rho^2 = o(d^{-1})$:*

- *If d is any function of k, and $n \to \infty$ then $\lim_{k \to \infty} \mathsf{R}^\star = 1$.*
- *If n is constant and $d \to \infty$ then $\lim_{k \to \infty} \mathsf{R}^\star = 1$.*

Namely, (ρ, d, n) is not admissible for weak detection.

Our second lower bound concerns with impossibility of strong detection. In the case where d is constant. In Theorem 2.6 we prove that for any d there exists $\rho^\star = \rho^\star(d)$ such that for $\rho^2 < \rho^\star$, strong detection is impossible. In the remaining cases, we show that the function d^{-1} is a threshold for strong detection.

Theorem 2.6 (Impossibility of strong detection). *A sequence (ρ, d, n) is not admissible for strong detection at either of the following scenarios:*

1. *$d \in \mathbb{N}$ and $\rho \in (-1, 1)$ are constants such that $d < \frac{\log(\rho^2)}{\log(1-\rho^2)}$, and $n \to \infty$.*
2. *$n, d \to \infty$, and $\rho^2 < (1 - \varepsilon)d^{-1}$ for some fixed $\varepsilon > 0$ which does not depend on n and d.*
3. *$d \to \infty$, n is constant, and $\rho^2 = O(d^{-1})$.*

We remark that the function $d^\star(\rho^2) = \frac{\log(\rho^2)}{\log(1-\rho^2)}$ is invertible as a function $(0, 1) \to (0, \infty)$. Denoting the inverse function by $\rho^\star$, the condition established in Theorem 2.6 is equivalent to $\rho^2 < \rho^\star(d)$.

3. Proofs

In this section, we present some of the proofs of our results. Detailed proofs are relegated to the appendix.

3.1. Proof of Theorem 2.3

To prove Theorem 2.3 we use the first and second moment methods, which reduce to the task of bounding the probabilities that the inner product between two random vectors (either independent or ρ-correlated) exceed a certain thresh-

old. To that end, we apply certain powerful integration techniques, to estimate the volumes of high-dimensional spherical caps.

Without loss of generality we assume below that $\rho > 0$. Recall that for $i \in [n]$, we define $\bar{X}_i \triangleq \frac{X_i}{\|X_i\|_2}$ and $\bar{Y}_i \triangleq \frac{Y_i}{\|Y_i\|_2}$. Furthermore, let $\mathcal{Q}_{d,\rho} \triangleq \mathbb{P}_0 \left(\bar{X}_1^T \bar{Y}_1 \geq \rho \right)$, and $\mathcal{P}_{d,\rho} \triangleq \mathbb{P}_\rho \left(\bar{X}_1^T \bar{Y}_1 \geq \rho \right)$, where $\mathbb{P}_\rho$ is defined in (10), and $\mathbb{P}_0 \equiv \mathbb{P}_{\rho=0}$. Consider the test in (9). We start by bounding the Type-I error probability. Markov's inequality implies that

$$\mathbb{P}_{\mathcal{H}_0} \left(\phi_{\mathsf{count}} = 1 \right)$$

$$= \mathbb{P}_{\mathcal{H}_0} \left(\sum_{i,j=1}^n \mathbb{1} \left\{ \bar{X}_i^T \bar{Y}_j \geq \rho \right\} \geq \frac{1}{2} n \mathcal{P}_{d,\rho} \right) \quad (12)$$

$$\leq \frac{2n^2 \mathcal{Q}_{d,\rho}}{n \mathcal{P}_{d,\rho}}. \quad (13)$$

On the other hand, we bound the Type-II error probability as follows. Under $\mathcal{H}_1$, since our proposed test is invariant to reordering of X and Y, we may assume without loss of generality that the latent permutation is the identity one, i.e., $\sigma = \mathsf{Id}$. Then, Chebyshev's inequality implies that

$$\mathbb{P}_{\mathcal{H}_1} \left(\phi_{\mathsf{count}} = 0 \right)$$

$$= \mathbb{P}_{\mathcal{H}_1} \left(\sum_{i,j=1}^n \mathbb{1} \left\{ \bar{X}_i^T \bar{Y}_j \geq \rho \right\} < \frac{1}{2} n \mathcal{P}_{d,\rho} \right) \quad (14)$$

$$\leq \mathbb{P}_{\mathcal{H}_1} \left(\sum_{i=1}^n \mathbb{1} \left\{ \bar{X}_i^T \bar{Y}_i \geq \rho \right\} < \frac{1}{2} n \mathcal{P}_{d,\rho} \right) \quad (15)$$

$$\leq \frac{4 \cdot \mathsf{Var}_\rho \left(\sum_{i=1}^n \mathbb{1} \left\{ \bar{X}_i^T \bar{Y}_i \geq \rho \right\} \right)}{n^2 \mathcal{P}_{d,\rho}^2}, \quad (16)$$

where Var_ρ denotes the variance with respect to $\mathbb{P}_\rho$, which for the random variable $\sum_{i=1}^n \mathbb{1} \left\{ \bar{X}_i^T \bar{Y}_i \right\}$ equals to the variance under the hypothesis $\mathcal{H}_1$. Noticing that

$$\mathsf{Var}_\rho \left(\sum_{i=1}^n \mathbb{1} \left\{ \bar{X}_i^T \bar{Y}_i \geq \rho \right\} \right)$$

$$= \sum_{i=1}^n \mathsf{Var}_\rho \left(\mathbb{1} \left\{ \bar{X}_i^T \bar{Y}_i \geq \rho \right\} \right) \quad (17)$$

$$= n \mathcal{P}_{d,\rho} (1 - \mathcal{P}_{d,\rho}), \quad (18)$$

we finally obtain,

$$\mathbb{P}_{\mathcal{H}_1} \left(\phi_{\mathsf{count}} = 0 \right) \leq \frac{4(1 - \mathcal{P}_{d,\rho})}{n \mathcal{P}_{d,\rho}} \leq \frac{4}{n \mathcal{P}_{d,\rho}}. \quad (19)$$

Next, we derive bounds on $\mathcal{Q}_{d,\rho}$ and $\mathcal{P}_{d,\rho}$. We start with a lower bound on $\mathcal{P}_{d,\rho}$. Let $Z_1, \ldots, Z_d$ be i.i.d $\mathcal{N}(0, I_{d \times d})$ random vectors, independent with $X_1, \ldots, X_n$. Let $Y_i' \triangleq$

$\rho X_i + \sqrt{1-\rho^2}Z_i$. We note that under $\mathbb{P}_\rho$, (X, Y) is equally distributed as $(\mathsf{X}, \mathsf{Y}')$. Thus, for our analysis, we shall assume without loss of generality that $\mathsf{Y} = \mathsf{Y}'$. Under this assumption, we show in the appendix that $\bar{X}_1^T \bar{Y}_1 \geq \rho$ holds if

$$\cos(\varphi_{X_1 Z_1}) \geq \frac{\rho}{2\sqrt{1-\rho^2}} \left(\frac{\|Z_1\|_2}{\|X_1\|_2} - \frac{\|X_1\|_2}{\|Z_1\|_2} \right),$$

where $\varphi_{X_1 Z_1}$ denotes the angle between X_1 and Z_1. Since $\varphi_{X_1 Z_1}$ is statistically independent of $\|X_1\|_2$ and $\|Z_1\|_2$, we get that

$$
\begin{aligned}
\mathcal{P}_{d,\rho} &= \mathbb{P}_\rho \left(\bar{X}_1^T \bar{Y}_1 \geq \rho \right) \\
&\geq \mathbb{P}_0 \left[\cos(\varphi_{X_1 Z_1}) \geq \frac{\rho}{2\sqrt{1-\rho^2}} \left(\frac{\|Z_1\|_2}{\|X_1\|_2} - \frac{\|X_1\|_2}{\|Z_1\|_2} \right) \right] \\
&\geq \mathbb{P}_0 \left(\cos(\varphi_{X_1 Z_1}) \geq 0, \|X_1\|_2 \geq \|Z_1\|_2 \right) \\
&= \mathbb{P}_0 \left(\cos(\varphi_{X_1 Z_1}) \geq 0 \right) \cdot \mathbb{P}_0 \left(\|X_1\|_2 \geq \|Z_1\|_2 \right) \\
&= 1/4. \quad\quad (20)
\end{aligned}
$$

We now derive an upper bound on $\mathcal{Q}_{d,\rho}$. For a fixed vector $y_1 \in \mathbb{S}^{d-1}$ on the d-dimensional sphere, let us define $\mathcal{B}_\rho(y_1) \triangleq \{ x_1 \in \mathbb{S}^{d-1} : x_1^T y_1 \geq \rho \}$. Since under $\mathbb{P}_0$, $\bar{X}_1$ and $\bar{Y}_1$ are independently and uniformly distributed on $\mathbb{S}^{d-1}$ we have

$$\mathcal{Q}_{d,\rho} = \int_{\mathbb{S}^{d-1}} \frac{\mathsf{Vol}(\mathcal{B}_\rho(y_1))}{\mathsf{Vol}(\mathbb{S}^{d-1})^2} \mathrm{d}y_1, \quad\quad (21)$$

where the volume of a set $\mathcal{A} \subset \mathbb{R}^d$ is defined as $\mathsf{Vol}(\mathcal{A}) = \int_\mathcal{A} \mathrm{d}x$, where the integration is with respect to Lebesgue's measure on the sphere. Let us bound $\mathsf{Vol}(\mathcal{B}_\rho)$ from above. To that end, define the set

$$\tilde{\mathcal{B}}_\rho(y_1) = \left\{ v \in \mathbb{R}^d : \frac{v}{\|v\|} \in B_\rho(y_1) \right\} \quad\quad (22)$$
$$= \left\{ v \in \mathbb{R}^d : v^T y_1 \geq \rho \|v\| \right\}. \quad\quad (23)$$

Let $f : (v_1, \ldots, v_d) \to (\varphi_1, \ldots, \varphi_{d-1}, r)$ denote the spherical coordinates transformation and let P_{d-1} denote the projection of an element from $\mathbb{R}^d$ onto its first $d-1$ coordinates. We observe that since $\tilde{\mathcal{B}}_\rho(y_1)$ is the cone defined by the spherical cap $\mathcal{B}_\rho(y_1)$, it follows that,

$$f(\tilde{\mathcal{B}}_\rho(y_1)) = P(f(B_\rho(y_1))) \times [0, \infty]. \quad\quad (24)$$

Finally, let μ denote the Gaussian measure of a multivariate random vector with mean y_1, and covariance matrix $\Sigma = \sigma^2 I_{d \times d}$, where $\sigma^2 \geq 0$. We denote the integration with respect to $d\varphi_1 d\varphi_2 \ldots, d\varphi_{d-1} dr$ by $d\overline{\varphi}dr$ an obtain

$$1 = \int_{x \in \mathbb{R}^d} \mu(\mathrm{d}x)$$

$$
\begin{aligned}
&= \int_{\mathbb{R}^d} \frac{1}{(2\pi\sigma^2)^{d/2}} e^{-\frac{1}{2\sigma^2}\|v - y_1\|_2^2} dv \\
&\geq \int_{\tilde{\mathcal{B}}_\rho(y_1)} \frac{1}{(2\pi\sigma^2)^{d/2}} e^{-\frac{1}{2\sigma^2}\|v - y_1\|_2^2} dv \\
&\overset{(a)}{\geq} \int_{\tilde{\mathcal{B}}_\rho(y_1)} \frac{1}{(2\pi\sigma^2)^{d/2}} e^{-\frac{1}{2\sigma^2}(\|v\|^2 - 2\rho\|v\| + 1)} dv \\
&\overset{(b)}{=} \frac{1}{(2\pi\sigma^2)^{d/2}} \int_{f(\tilde{\mathcal{B}}_\rho(y_1))} e^{-\frac{1}{2\sigma^2}(r^2 - 2\rho r + 1)} \cdot
\end{aligned}
$$

$$\left(r^{d-1} \prod_{i=1}^{d-1} \sin^{d-1-i}(\varphi_i) \right) d\overline{\varphi}dr$$

$$\overset{(c)}{=} \frac{1}{(2\pi\sigma^2)^{d/2}} \underbrace{\int_{P(f(\mathcal{B}_\rho(y_1)))} \left(\prod_{i=1}^{d-1} \sin^{d-1-i}(\varphi_i) \right) d\overline{\varphi}}_{\mathsf{Vol}(\mathcal{B}_\rho(y_1))} \cdot$$

$$\int_0^\infty r^{d-1} e^{-\frac{1}{2\sigma^2}(r^2 - 2\rho r + 1)} dr$$

$$\geq \frac{\mathsf{Vol}(\mathcal{B}_\rho(y_1))}{(2\pi\sigma^2)^{d/2}} \int_\rho^\infty \rho^{d-1} e^{-\frac{1}{2\sigma^2}(r^2 - 2\rho r + 1)} dr$$

$$= \frac{\mathsf{Vol}(\mathcal{B}_\rho(y_1))\rho^{d-1}}{(2\pi\sigma^2)^{(d-1)/2}} e^{-\frac{1-\rho^2}{2\sigma^2}} \int_\rho^\infty \frac{1}{(2\pi\sigma^2)^{1/2}} e^{-\frac{(r-\rho)^2}{2\sigma^2}} dr$$

$$= \frac{\rho^{d-1}}{(2\pi\sigma^2)^{(d-1)/2}} e^{-\frac{1-\rho^2}{2\sigma^2}} \mathsf{Vol}(\mathcal{B}_\rho(y_1)) \frac{1}{2},$$

where (a) follows from the definition of $\tilde{\mathcal{B}}_\rho(y_1)$, (b) follows from change of variables and (c) follows from (24). Thus, for every $y_1 \in \mathbb{S}^{d-1}$,

$$\mathsf{Vol}(\mathcal{B}_\rho(y_1)) \leq \min_{\sigma^2 \geq 0} 2 e^{\frac{1-\rho^2}{2\sigma^2} + \frac{d-1}{2}\log(2\pi\sigma^2)} \rho^{1-d} \quad (25)$$

$$= 2 e^{\frac{d-1}{2}\log\left(2\pi e \frac{1-\rho^2}{d-1}\right)} \rho^{1-d}. \quad\quad (26)$$

On the other hand, it is well-known that,

$$\mathsf{Vol}(\mathbb{S}^{d-1}) = \frac{2\pi^{d/2}}{\Gamma(d/2)} \geq \frac{2\pi^{d/2}}{(d/2)^{d/2-1}} = \frac{4}{d} e^{\frac{d}{2}\log(2\pi/d)},$$
$$\quad\quad (27)$$

where we have used the fact that $\Gamma(x) < x^{x-1}$, for $x > 1$ (see, e.g., (Anderson & Qiu, 1997)). Combining (27), (26), and (21), we obtain

$$\mathcal{Q}_{d,\rho} \leq f(d) e^{\frac{d-1}{2}\log(1-\rho^2)} \rho^{1-d}, \quad\quad (28)$$

where $f(d) \triangleq \frac{d}{2} \left(\frac{ed}{d-1} \right)^{d/2} \left(\frac{2\pi e}{d-1} \right)^{1/2}$. Finally, using (20) and (28), we see that (13) can be further upper bounded as

$$\mathbb{P}_{\mathcal{H}_0} \left(\phi_{\mathsf{count}} = 1 \right) \leq 8n f(d) e^{\frac{d-1}{2}\log(1-\rho^2)} \rho^{1-d} \quad (29)$$

$$= 8f(d)n(\rho^{-2} - 1)^{\frac{d-1}{2}}, \qquad (30)$$

while (74) can be upper bounded as,

$$\mathbb{P}_{\mathcal{H}_1}(\phi_{\mathsf{count}} = 0) \leq \frac{16}{n}. \qquad (31)$$

Thus, for a fixed $d \geq 2$, it is clear that $\mathbb{P}_{\mathcal{H}_1}(\phi_{\mathsf{count}} = 0) \to 0$, as $n \to \infty$, and $\mathbb{P}_{\mathcal{H}_0}(\phi_{\mathsf{count}} = 1) \to 0$, if $n(\rho^{-2} - 1)^{\frac{d-1}{2}} = o(1)$. The later holds if $\rho^{-2} = 1 + o(n^{-2/(d-1)})$, which implies that $\rho^2 = 1 - o(n^{-2/(d-1)})$, as stated.

3.2. Proof of Theorem 2.5

We focus on weak detection in the regime where $n, d \to \infty$. As in many detection problems, evaluating the minimax risk function opposes a great challenge due to the error term obtained by maximizing over the error for all permutations in $\mathbb{S}_n$. A well known strategy for overcoming this inherent obstacle is by considering the softer average-case version of the problem. Let π be the uniform measure on $\mathbb{S}_n$, and let us denote by $\mathbb{P}_{\mathcal{H}_1}$ the probability measure obtained by averaging $\mathbb{P}_{\mathcal{H}_1|\sigma}$ with respect to π. For a test ϕ, we consider the Bayesian risk function given by

$$\bar{R}(\phi) \triangleq \mathbb{P}_{\mathcal{H}_0}[\phi(\mathsf{X},\mathsf{Y}) = 1] + \mathbb{E}_{\sigma \sim \pi}\left[\mathbb{P}_{\mathcal{H}_1|\sigma}[\phi(\mathsf{X},\mathsf{Y}) = 0]\right],$$

and the Bayesian risk for our problem:

$$\bar{R}^\star \triangleq \inf_\phi \bar{R}(\phi).$$

Clearly, any test ϕ satisfies $R(\phi) \geq \bar{R}(\phi)$, and therefore $R^\star \geq \bar{R}^\star$. We conclude that in order to prove Theorem 2.5, it is sufficient to show that under the given assumptions, $\bar{R}^\star = 1 + o(1)$. Using a well-known equivalent characterization of the Bayesian risk function by the total variation distance and Cauchy-Schwartz inequality one shows that

$$R^\star \geq \bar{R}^\star = 1 - d_{\mathsf{TV}}(\mathbb{P}_{\mathcal{H}_0}, \mathbb{P}_{\mathcal{H}_1}) \geq 1 - \frac{1}{2}\sqrt{\mathbb{E}_0[L^2] - 1}, \qquad (32)$$

where $L \triangleq \frac{\mathbb{P}_{\mathcal{H}_1}}{\mathbb{P}_{\mathcal{H}_0}}$ is the likelihood ratio, and the expectation is taken with respect to $\mathbb{P}_{\mathcal{H}_0}$. Using the bound given in (32), it is sufficient to show that under the assumptions of Theorem 2.5, $\mathbb{E}_0[L^2] \leq 1 + o(1)$.

Inspired by (Wu et al., 2020), K and Nazer (K & Nazer, 2022b) gave an exact description of $\mathbb{E}_0[L^2]$ using the distribution of cycles in a uniformly drawn random permeation. In order to prove Theorem 2.5, we shall carefully analyse $\mathbb{E}_0[L^2]$, and improve the bounds proved (K & Nazer, 2022b). The first step in the calculation calls for a use of Ingster-Suslina method, stating that by Fubini's theorem, $\mathbb{E}_0[L^2]$ may be equivalently written as

$$\mathbb{E}_0[L^2] = \mathbb{E}_{\pi \perp\!\!\!\perp \pi'}\left[\mathbb{E}_0\left[\frac{\mathbb{P}_{\mathcal{H}_1|\pi}}{\mathbb{P}_{\mathcal{H}_0}} \cdot \frac{\mathbb{P}_{\mathcal{H}_1|\pi'}}{\mathbb{P}_{\mathcal{H}_0}}\right]\right], \qquad (33)$$

where the expectation is taken with respect to the independent coupling of π and π', two copies of the uniform measure on $\mathbb{S}_n$. For fixed permutations σ and σ', we note that $\mathbb{P}_{\mathcal{H}_1|\sigma}, \mathbb{P}_{\mathcal{H}_1|\sigma'}$ and $\mathbb{P}_0$ are absolutely continuous with respect to Lebesgue's measure on $\mathbb{R}^{2 \times d \times n}$ and therefore we have

$$\frac{\mathbb{P}_{\mathcal{H}_1|\sigma}}{\mathbb{P}_{\mathcal{H}_0}} = \frac{f_{\mathcal{H}_1|\sigma}}{f_{\mathcal{H}_0}}, \quad \text{and} \quad \frac{\mathbb{P}_{\mathcal{H}_1|\sigma'}}{\mathbb{P}_{\mathcal{H}_0}} = \frac{f_{\mathcal{H}_1|\sigma'}}{f_{\mathcal{H}_0}},$$

where $f_{\mathcal{H}_i}$ denotes the Radon-Nikodym derivative of $\mathbb{P}_{\mathcal{H}_i}$ with respect to Lebesgue's measure, which is the density function (X,Y) under the corresponding hypothesis. Let $\mathcal{N}_\rho : \mathbb{R}^{2 \times d} :\to \mathbb{R}_+$ denote the density function of a pair of random vectors distributed as

$$\mathcal{N}^{\otimes d}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right).$$

We note that

$$\frac{f_{\mathcal{H}_1|\sigma}(\mathsf{X},\mathsf{Y})}{f_{\mathcal{H}_0}(\mathsf{X},\mathsf{Y})} \frac{f_{\mathcal{H}_1|\sigma'}(\mathsf{X},\mathsf{Y})}{f_{\mathcal{H}_0}(\mathsf{X},\mathsf{Y})}$$
$$= \prod_{i=1}^n \frac{\mathcal{N}_\rho(X_i, Y_{\sigma_i})}{\mathcal{N}_0(X_i, Y_{\sigma_i})} \frac{\mathcal{N}_\rho(X_i, Y_{\sigma'_i})}{\mathcal{N}_0(X_i, Y_{\sigma'_i})}. \qquad (34)$$

In order to proceed with the calculation, we make two key observations. First, we note that the distribution under the null hypothesis ($\mathcal{H}_0$) is invariant to reordering the coordinates. In a similar manner, the uniform measure on $\mathbb{S}_n$, is invariant under composition with a fixed permutation. Thus,

$$\mathbb{E}_{\pi \perp\!\!\!\perp \pi'}\left[\mathbb{E}_0\left[\frac{\mathbb{P}_{\mathcal{H}_1|\pi}}{\mathbb{P}_{\mathcal{H}_0}} \cdot \frac{\mathbb{P}_{\mathcal{H}_1|\pi'}}{\mathbb{P}_{\mathcal{H}_0}}\right]\right]$$
$$= \mathbb{E}_\pi\left[\mathbb{E}_0\left[\frac{\mathbb{P}_{\mathcal{H}_1|\pi}}{\mathbb{P}_{\mathcal{H}_0}} \cdot \frac{\mathbb{P}_{\mathcal{H}_1|\mathsf{Id}}}{\mathbb{P}_{\mathcal{H}_0}}\right]\right]. \qquad (35)$$

We consider the product given in (34) for a fixed $\sigma \in \mathbb{S}_n$ and $\sigma' = \mathsf{Id}$, which we denote by Z_σ. The second key observation, is that Z_σ can be decomposed to independent terms, corresponding to the cycles of the permeation σ. We recall that a cycle of a permutation σ is a string $(i_0, i_2, \ldots, i_{|C|-1})$ of elements in $[n]$ such that $\sigma(i_j) = i_{j+1 \mod |C|}$ for all j. If $|C| = k$, we call C a k-cycle. For a fixed cycle C, we denote

$$Z_C \triangleq \prod_{i \in C} \frac{\mathcal{N}_\rho(X_i, Y_{\sigma_i})}{\mathcal{N}_0(X_i, Y_{\sigma_i})} \frac{\mathcal{N}_\rho(X_i, Y_i)}{\mathcal{N}_0(X_i, Y_i)}. \qquad (36)$$

Since the set of cycles of a permutation induce a partition of $[n]$, the random variables $\{Z_C\}_C$, corresponding to all cycles of σ, are independent (with respect to $\mathbb{P}_{\mathcal{H}_0}$) and

$$Z_\sigma = \prod_C Z_C. \qquad (37)$$

The following lemma states that for a fixed cycle C, $\mathbb{E}_0[Z_C]$ depends on ρ and $|C|$. The proof of the lemma is based on the properties of Gaussian random vectors. For further details the reader is referred to Lemma 10 in (K & Nazer, 2022b).

Lemma 3.1. *For a fixed cycle C of a permutation σ,*

$$\mathbb{E}_0[Z_C] = \frac{1}{(1 - \rho^{2|C|})^d}. \tag{38}$$

For a fixed permutation $\sigma \in \mathbb{S}_n$ and $k \in [n]$, let $N_k(\sigma)$ denote the number of k-cycles of σ. Combining (33), (35), (37), and Lemma 3.1 we obtain

$$\mathbb{E}_0[L^2] = \mathbb{E}_\pi\left[\prod_C Z_C\right] = \mathbb{E}_\pi\left[\prod_{k=1}^n \frac{1}{(1 - \rho^{2k})^{dN_k}}\right]. \tag{39}$$

We carefully bound (39) from above, utilizing the statistical properties of k-cycles in a uniformly distributed random permutation. We shall require several technical results. The following lemma concerns the approximation of the joint distribution of k-cycles by independent Poisson random variables.

Lemma 3.2. *(Arratia & Tavaré, 1992)[Theorem 2] Let $1 \leq k \leq n$ be an integer, and let $Z_1 \dots, Z_k$ be independent random variables such that for all $1 \leq i \leq k$, $Z_i \sim \mathsf{Poisson}(i^{-1})$. Then, the total variation between the law of $N_1, \dots, N_k$ and $Z_1, \dots, Z_k$ satisfies*

$$d_{\mathsf{TV}}(\mathcal{L}(N_1, N_2, \dots, N_k), \mathcal{L}(Z_1, Z_2, \dots, Z_k)) \leq F\left(\frac{n}{k}\right),$$

where $F(x)$ is a monotone decreasing function satisfying $\log F(x) = -x \log x(1 + o(1))$ as $x \to \infty$ and we recall that $\mathcal{L}(\cdot)$ denote the law, that is, the probability distribution, of a given random variable.

Lemma 3.3. *Let $1 \leq m \leq n$ be an integer, and let $Z_1 \dots, Z_m$ be independent random variables such that for all $1 \leq i \leq m$, $Z_i \sim \mathsf{Poisson}(i^{-1})$. Then,*

$$\mathbb{E}_\pi\left[\prod_{k=1}^m \left(\frac{1}{1 - \rho^{2k}}\right)^{dZ_k}\right] \leq \exp\left(\frac{d\rho^2}{1 - \rho^2} + \frac{c(d, \rho^2)\rho^4}{1 - \rho^4}\right),$$

where $c(d, \rho^2) = \frac{d(d+1)}{2(1-\rho^2)^{d+2}}$, and therefore if $\rho^2 = o(d^{-1})$,

$$\mathbb{E}_\pi\left[\prod_{k=1}^m \left(\frac{1}{1 - \rho^{2k}}\right)^{dZ_k}\right] \leq 1 + o(1). \tag{40}$$

Proof. By rearranging the expression in the expectation and using independence we have

$$\mathbb{E}_\pi\left[\prod_{k=1}^m \left(\frac{1}{1 - \rho^{2k}}\right)^{dZ_k}\right] = \prod_{k=1}^m \mathbb{E}_\pi\left[\left(\frac{1}{1 - \rho^{2k}}\right)^{dZ_k}\right]$$

$$= \prod_{k=1}^m \mathbb{E}_\pi\left[\exp\left(-dZ_k \log\left(1 - \rho^{2k}\right)\right)\right] \tag{41}$$

$$\overset{(a)}{=} \prod_{k=1}^m \exp\left(\frac{1}{k}\left(e^{-d\log(1-\rho^{2k})} - 1\right)\right) \tag{42}$$

$$= \prod_{k=1}^m \exp\left(\frac{1}{k}\left(\frac{1}{(1 - \rho^{2k})^d} - 1\right)\right). \tag{43}$$

where (a) is followed by the definition of the moment generating function of a Poisson random variable. We prove in the appendix that for any $k > 0$,

$$\frac{1}{(1 - \rho^{2k})^d} \leq 1 + d\rho^{2k} + \frac{d(d+1)}{2(1 - \rho^{2k})^{(d+2)}}\rho^{4k}$$

$$\leq 1 + d\rho^{2k} + c(d, \rho^2)\rho^{4k}, \tag{44}$$

where

$$c(d, \rho^2) \triangleq \frac{d(d+1)}{2(1 - \rho^2)^{(d+2)}}\rho^{4k}. \tag{45}$$

Therefore, we get

$$\prod_{k=1}^m \exp\left(\frac{1}{k}\left(\frac{1}{(1 - \rho^{2k})^d} - 1\right)\right)$$

$$\leq \exp\left(d\sum_{k=1}^m \frac{\rho^{2k}}{k} + c(d, \rho^2)\sum_{k=1}^m \frac{\rho^{4k}}{k}\right) \tag{46}$$

$$\leq \exp\left(d\sum_{k=1}^\infty \frac{\rho^{2k}}{k} + c(d, \rho^2)\sum_{k=1}^\infty \frac{\rho^{4k}}{k}\right) \tag{47}$$

$$= \exp\left(-d\log(1 - \rho^2) - c(d, \rho^2)\log(1 - \rho^4)\right) \tag{48}$$

$$\overset{(a)}{\leq} \exp\left(\frac{d\rho^2}{1 - \rho^2} + \frac{c(d, \rho^2)\rho^4}{1 - \rho^4}\right), \tag{49}$$

where (a) follows from the well known inequality $\log(1 + x) \geq x/(1 + x)$, for $x > -1$. In the case where $\rho = o(d^{-1})$ clearly $d\rho^2 = o(1)$. Furthermore,

$$c(d, \rho^2)\rho^4 = \frac{d(d+1)}{2(1 - \rho^2)^{d+2}}\rho^4$$

$$\leq \frac{1}{2}\exp((d+2)\rho^2)d(d+1)\rho^4$$

$$= \exp(o(1)) \cdot o(1) = o(1). \tag{50}$$

In particular, we get

$$\mathbb{E}_\pi\left[\prod_{k=1}^m \left(\frac{1}{1 - \rho^{2k}}\right)^{dZ_k}\right] \leq 1 + o(1). \tag{51}$$

$\square$

We are now in a position to prove Theorem 2.5, in the case where $n, d \to \infty$. The idea of the proof is as follows: we

consider the expectation of the product given in (39). In the case that $n \to \infty$, we show that the product of the last $n - \log n$ terms is always upper-bounded by $1 + o(1)$. For the expectation of the product of the first $\log n$ terms, use the Poisson approximation of $\{N_k\}_k$ given in Lemma 3.2 and the estimation in that case, given in Lemma 3.3.

We assume that $n, d \to \infty$ and $\rho^2 = o(d^{-1})$ or $\rho^2 < (1-\varepsilon)d^{-1}$. We choose $m = \lceil \log n \rceil$ and we get

$$dn\rho^{2m} \leq (d\rho^2) \cdot n(\rho^2)^{\log(n)-1} \qquad (52)$$

$$= (d\rho^2)n(\rho^2)^{\log\left(\frac{n}{e}\right)} \qquad (53)$$

$$= e(d\rho^2)\left(\frac{n}{e}\right)^{1+\log\rho^2}. \qquad (54)$$

Since $\rho^2 < d^{-1}$ (which is clearly true for sufficiently large d in the particular case where $\rho^2 = o(d^{-1})$), and $d \to \infty$, we have $\log \rho^2 \to -\infty$ as $n \to \infty$. In the appendix we prove that the tail of the product may be simply bounded as follows:

$$\prod_{k=m=\log n}^{n} \left(\frac{1}{1-\rho^{2k}}\right)^{dN_k} \leq \exp\left(\frac{dn\rho^{2m}}{1-\rho^{2m}}\right) \qquad (55)$$

$$= \exp(o(1)) = 1 + o(1). \qquad (56)$$

Next, let $\{Z_k\}_k$ be independent $\mathsf{Poisson}\left(k^{-1}\right)$ random variables as in Lemma 3.2. Since $\sum_{k=1}^{m} N_k \leq n$ with probability 1, using Lemma 3.2 we show in the appendix that

$$\mathbb{E}_\pi \left[\prod_{k=1}^{m} \left(\frac{1}{1-\rho^{2k}}\right)^{dN_k} \right] \leq \mathbb{E}_\pi \left[\prod_{k=1}^{m} \left(\frac{1}{1-\rho^{2k}}\right)^{dZ_k} \right]$$

$$+ d_{\mathsf{TV}}(\mathcal{L}(N_1^m), \mathcal{L}(Z_1^m)) \left(\frac{1}{1-\rho^2}\right)^{dn}$$

$$\overset{(a)}{\leq} \exp\left(\frac{d\rho^2}{1-\rho^2} + \frac{c(d,\rho^2)\rho^4}{1-\rho^4}\right) + F\left(\frac{n}{m}\right) \cdot \left(\frac{1}{1-\rho^2}\right)^{dn},$$

where (a) follows from Lemma 3.3. By Lemma 3.2, we can also show that

$$\log\left(F\left(\frac{n}{\lceil \log n \rceil}\right)\left(\frac{1}{1-\rho^2}\right)^{dn} \right) \leq -n(1 - d\rho^2 + o(1)), \qquad (57)$$

which tends to $-\infty$ as $\rho^2 = o(d^{-1})$. This implies that

$$F\left(\frac{n}{\lceil \log n \rceil}\right)\left(\frac{1}{1-\rho^2}\right)^{dn} = o(1), \qquad (58)$$

and therefore,

$$\mathbb{E}_\pi \left[\prod_{k=1}^{\lceil \log n \rceil} \left(\frac{1}{1-\rho^{2k}}\right)^{dN_k} \right]$$

$$\leq \exp\left(\frac{d\rho^2}{1-\rho^2} + \frac{c(d,\rho^2)\rho^4}{1-\rho^4}\right) + o(1) \qquad (59)$$

$$= 1 + o(1), \qquad (60)$$

where the last equality follows from the assumption that $\rho^2 = o(d^{-1})$. Combining (55) and (60) together we conclude:

$$\mathbb{E}_\pi \left[\prod_{k=1}^{n} \left(\frac{1}{1-\rho^{2k}}\right)^{dN_k} \right]$$

$$\leq \mathbb{E}_\pi \left[\prod_{k=1}^{\lceil \log n \rceil} \left(\frac{1}{1-\rho^{2k}}\right)^{dN_k} \prod_{k=\lceil \log n \rceil}^{n} \left(\frac{1}{1-\rho^{2k}}\right)^{dN_k} \right]$$

$$= (1 + o(1)) \cdot \mathbb{E}_\pi \left[\prod_{k=1}^{\lceil \log n \rceil} \left(\frac{1}{1-\rho^{2k}}\right)^{dN_k} \right] \qquad (61)$$

$$= (1 + o(1)). \qquad (62)$$

Finally, we combine (33), (39) and (62), and obtain:

$$\mathsf{R}^\star \geq 1 - \sqrt{\mathbb{E}_0[L^2] - 1} \qquad (63)$$

$$= 1 - \sqrt{\mathbb{E}_\pi \left[\prod_{k=1}^{n} \frac{1}{(1-\rho^{2k})^{dN_k}} \right] - 1} \qquad (64)$$

$$\geq 1 - \sqrt{1 + o(1) - 1} \qquad (65)$$

$$= 1 - o(1). \qquad (66)$$

4. Conclusions

In this paper, we have studied the asymptotic thresholds for weak and strong sense detection in the Gaussian correlated databases problem. Our results, summarized in Table 1, resolve significant gaps in current literature (where the focus was exclusively on the regime where $n, d \to \infty$). When d is fixed and $n \to \infty$, the picture is incomplete and a gap between our bounds is evident. Closing this gap is an intriguing open problem.

Acknowledgements

References

Anderson, G. and Qiu, S.-L. A monotoneity property of the gamma function. *Proceedings of the American Mathematical Society*, 125(11):3355–3362, 1997.

Arratia, R. and Tavaré, S. The cycle structure of random permutations. *The Annals of Probability*, pp. 1567–1591, 1992.

Bakirtas, S. and Erkip, E. Database matching under column deletions. *2021 IEEE International Symposium on Information Theory (ISIT)*, pp. 2720–2725, 2021.

Bakirtas, S. and Erkip, E. Database matching under column repetitions. *ArXiv*, abs/2202.01730, 2022.

Berg, A., Berg, T., and Malik, J. Shape matching and object recognition using low distortion correspondences. In *Proc. Computer Vision and Pattern Recognition*, 2005.

Cour, T., Srinivasan, P., and Shi, J. Balanced graph matching. In *Proceedings of the 19th International Conference on Neural Information Processing Systems*, NIPS'06, pp. 313–320, Cambridge, MA, USA, 2006. MIT Press.

Cullina, D., Mittal, P., and Kiyavash, N. Fundamental limits of database alignment. In *2018 IEEE International Symposium on Information Theory (ISIT)*, pp. 651–655. IEEE Press, 2018.

Dai, O. E., Cullina, D., and Kiyavash, N. Database alignment with gaussian features. In Chaudhuri, K. and Sugiyama, M. (eds.), *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 3225–3233. PMLR, 16–18 Apr 2019.

Dai, O. E., Cullina, D., and Kiyavash, N. Achievability of nearly-exact alignment for correlated gaussian databases. In *2020 IEEE International Symposium on Information Theory (ISIT)*, pp. 1230–1235, 2020. doi: 10.1109/ISIT44484.2020.9174507.

Ding, J., Ma, Z., Wu, Y., and Xu, J. Efficient random graph matching via degree profiles. *Probability Theory and Related Fields*, 179:29–115, 2018.

Ding, J., Wu, Y., Xu, J., and Yang, D. The planted matching problem: Sharp threshold and infinite-order phase transition. *ArXiv*, abs/2103.09383, 2021.

Ganassali, L. Sharp threshold for alignment of graph databases with gaussian weights. In *MSML*, 2020.

K, Z. and Nazer, B. Detecting correlated gaussian databases. In *2022 IEEE International Symposium on Information Theory (ISIT)*, pp. 2064–2069, 2022a. doi: 10.1109/ISIT50566.2022.9834731.

K, Z. and Nazer, B. Detecting correlated gaussian databases. *arXiv preprint arXiv:2206.12011*, 2022b.

Kang, U., Hebert, M., and Park, S. Fast and scalable approximate spectral graph matching for correspondence problems. *Information Sciences*, 2012.

Mao, C., Wu, Y., Xu, J., and Yu, S. H. Testing network correlation efficiently via counting trees. 2021.

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

Narayanan, A. and Shmatikov, V. Robust de-anonymization of large sparse datasets. In *2008 IEEE Symposium on Security and Privacy (sp 2008)*, pp. 111–125, 2008. doi: 10.1109/SP.2008.33.

Narayanan, A. and Shmatikov, V. De-anonymizing social networks. In *2009 30th IEEE Symposium on Security and Privacy*, pp. 173–187, 2009. doi: 10.1109/SP.2009.22.

Pedarsani, P. and Grossglauser, M. On the privacy of anonymized networks. In *Knowledge Discovery and Data Mining*, 2011.

Shirani, F., Garg, S., and Erkip, E. A concentration of measure approach to database de-anonymization. In *2019 IEEE International Symposium on Information Theory (ISIT)*, pp. 2748–2752. IEEE Press, 2019.

Singh, R., Xu, J., and Berger, B. Global alignment of multiple protein interaction networks with application to functional orthology detection. *Proceedings of the National Academy of Sciences of the United States of America*, 105(35):12763–8, Sep 2008. doi: 10.1073/pnas.0806627105.

Tamir, R. Joint correlation detection and alignment of Gaussian databases. 2022. URL https://arxiv.org/abs/2211.01069.

Tsybakov, A. B. Introduction to nonparametric estimation, 2009. *URL https://doi. org/10.1007/b13794. Revised and extended from the*, 9(10), 2004.

Wu, Y., Xu, J., and Yu, S. H. Testing correlation of unlabeled random graphs. *arXiv preprint arXiv:2008.10097*, 2020.

Wu, Y., Xu, J., and Yu, S. H. Settling the sharp reconstruction thresholds of random graph matching. *IEEE Transactions on Information Theory*, 68(8):5391–5417, 2022. doi: 10.1109/TIT.2022.3169005.

A. Proof of Theorem 2.3

Without loss of generality we assume below that $\rho > 0$. Recall that for $i \in [n]$, we define $\bar{X}_i \triangleq \frac{X_i}{\|X_i\|_2}$ and $\bar{Y}_i \triangleq \frac{Y_i}{\|Y_i\|_2}$. Furthermore, let $\mathcal{Q}_{d,\rho} \triangleq \mathbb{P}_0\left(\bar{X}_1^T \bar{Y}_1 \geq \rho\right)$, and $\mathcal{P}_{d,\rho} \triangleq \mathbb{P}_\rho\left(\bar{X}_1^T \bar{Y}_1 \geq \rho\right)$, where $\mathbb{P}_\rho$ is defined in (10), and $\mathbb{P}_0 \equiv \mathbb{P}_{\rho=0}$. Consider the test in (9). We start by bounding the Type-I error probability. Markov's inequality implies that

$$\mathbb{P}_{\mathcal{H}_0}\left(\phi_{\mathsf{count}} = 1\right) = \mathbb{P}_{\mathcal{H}_0}\left(\sum_{i,j=1}^n \mathbb{1}\left\{\bar{X}_i^T \bar{Y}_j \geq \rho\right\} \geq \frac{1}{2}n\mathcal{P}_{d,\rho}\right) \tag{67}$$

$$\leq \frac{2n^2 \mathcal{Q}_{d,\rho}}{n\mathcal{P}_{d,\rho}}. \tag{68}$$

On the other hand, we bound the Type-II error probability as follows. Under $\mathcal{H}_1$, since our proposed test is invariant to reordering of X and Y, we may assume without loss of generality that the latent permutation is the identity one, i.e., $\sigma = \mathsf{Id}$. Then, Chebyshev's inequality implies that

$$\mathbb{P}_{\mathcal{H}_1}\left(\phi_{\mathsf{count}} = 0\right) = \mathbb{P}_{\mathcal{H}_1}\left(\sum_{i,j=1}^n \mathbb{1}\left\{\bar{X}_i^T \bar{Y}_j \geq \rho\right\} < \frac{1}{2}n\mathcal{P}_{d,\rho}\right) \tag{69}$$

$$\leq \mathbb{P}_{\mathcal{H}_1}\left(\sum_{i=1}^n \mathbb{1}\left\{\bar{X}_i^T \bar{Y}_i \geq \rho\right\} < \frac{1}{2}n\mathcal{P}_{d,\rho}\right) \tag{70}$$

$$\leq \frac{4 \cdot \mathsf{Var}_\rho\left(\sum_{i=1}^n \mathbb{1}\left\{\bar{X}_i^T \bar{Y}_i \geq \rho\right\}\right)}{n^2 \mathcal{P}_{d,\rho}^2}, \tag{71}$$

where Var_ρ denotes the variance with respect to $\mathbb{P}_\rho$, which for the random variable $\sum_{i=1}^n \mathbb{1}\left\{\bar{X}_i^T \bar{Y}_i\right\}$ equals to the variance under the hypothesis $\mathcal{H}_1$. Noticing that

$$\mathsf{Var}_\rho\left(\sum_{i=1}^n \mathbb{1}\left\{\bar{X}_i^T \bar{Y}_i \geq \rho\right\}\right) = \sum_{i=1}^n \mathsf{Var}_\rho\left(\mathbb{1}\left\{\bar{X}_i^T \bar{Y}_i \geq \rho\right\}\right) \tag{72}$$

$$= n\mathcal{P}_{d,\rho}(1 - \mathcal{P}_{d,\rho}), \tag{73}$$

we finally obtain,

$$\mathbb{P}_{\mathcal{H}_1}\left(\phi_{\mathsf{count}} = 0\right) \leq \frac{4(1 - \mathcal{P}_{d,\rho})}{n\mathcal{P}_{d,\rho}} \leq \frac{4}{n\mathcal{P}_{d,\rho}}. \tag{74}$$

Next, we derive bounds on $\mathcal{Q}_{d,\rho}$ and $\mathcal{P}_{d,\rho}$. We start with a lower bound on $\mathcal{P}_{d,\rho}$. Let $Z_1, \ldots, Z_d$ be i.i.d $\mathcal{N}(0, I_{d \times d})$ random vectors, independent with $X_1, \ldots, X_n$. Let $Y_i' \triangleq \rho X_i + \sqrt{1 - \rho^2} Z_i$. We note that under $\mathbb{P}_\rho$, (X, Y) is equally distributed as $(\mathsf{X}, \mathsf{Y}')$. Thus, for our analysis, we shall assume without loss of generality that $\mathsf{Y} = \mathsf{Y}'$. Under this assumption, we have

$$|\bar{X}_1^T \bar{Y}_1|^2 = \frac{\rho^2 \|X_1\|_2^4 + 2\rho\sqrt{1 - \rho^2} \|X_1\|_2^2 X_1^T Z_1 + (1 - \rho^2)|X_1^T Z_1|^2}{\rho^2 \|X_1\|_2^4 + 2\rho\sqrt{1 - \rho^2} \|X_1\|_2^2 X_1^T Z_1 + (1 - \rho^2) \|X_1\|_2^2 \|Z_1\|_2^2} \tag{75}$$

$$\geq \frac{\rho^2 \|X_1\|_2^4 + 2\rho\sqrt{1 - \rho^2} \|X_1\|_2^2 X_1^T Z_1}{\rho^2 \|X_1\|_2^4 + 2\rho\sqrt{1 - \rho^2} \|X_1\|_2^2 X_1^T Z_1 + (1 - \rho^2) \|X_1\|_2^2 \|Z_1\|_2^2} \tag{76}$$

$$= \frac{\rho^2 \|X_1\|_2^2 + 2\rho\sqrt{1 - \rho^2} \|X_1\|_2 \|Z_1\|_2 \cos(\varphi_{X_1 Z_1})}{\rho^2 \|X_1\|_2^2 + 2\rho\sqrt{1 - \rho^2} \|X_1\|_2 \|Z_1\|_2 \cos(\varphi_{X_1 Z_1}) + (1 - \rho^2) \|Z_1\|_2^2}, \tag{77}$$

where we have used the fact that for $X_1, Z_1 \in \mathbb{R}^d$, the inner product $X_1^T Z_1$ can be represented as $X_1^T Z_1 = \|X_1\|_2 \|Z_1\|_2 \cos\varphi_{X_1 Z_1}$, where $\varphi_{X_1 Z_1}$ denotes the angle between X_1 and Z_1. Also, note that $\varphi_{X_1 Z_1}$ is statistically independent of $\|X_1\|_2$ and $\|Z_1\|_2$. Thus, straightforward algebra steps reveal that,

$$\mathcal{P}_{d,\rho} = \mathbb{P}_\rho\left(\bar{X}_1^T \bar{Y}_1 \geq \rho\right) \tag{78}$$

$$\geq \mathbb{P}_0\left[\cos(\varphi_{X_1 Z_1}) \geq \frac{\rho}{2\sqrt{1-\rho^2}}\left(\frac{\|Z_1\|_2}{\|X_1\|_2} - \frac{\|X_1\|_2}{\|Z_1\|_2}\right)\right] \tag{79}$$

$$\geq \mathbb{P}_0\left[\cos(\varphi_{X_1 Z_1}) \geq \frac{\rho}{2\sqrt{1-\rho^2}}\left(\frac{\|Z_1\|_2}{\|X_1\|_2} - \frac{\|X_1\|_2}{\|Z_1\|_2}\right), \|X_1\|_2 \geq \|Z_1\|_2\right] \tag{80}$$

$$\geq \mathbb{P}_0\left(\cos(\varphi_{X_1 Z_1}) \geq 0, \|X_1\|_2 \geq \|Z_1\|_2\right) \tag{81}$$

$$= \mathbb{P}_0\left(\cos(\varphi_{X_1 Z_1}) \geq 0\right) \cdot \mathbb{P}_0\left(\|X_1\|_2 \geq \|Z_1\|_2\right) \tag{82}$$

$$= 1/4. \tag{83}$$

We now derive an upper bound on $\mathcal{Q}_{d,\rho}$. For a fixed vector $y_1 \in \mathbb{S}^{d-1}$ on the d-dimensional sphere, let us define $\mathcal{B}_\rho(y_1) \triangleq \left\{x_1 \in \mathbb{S}^{d-1} : x_1^T y_1 \geq \rho\right\}$. Since under $\mathbb{P}_0$, $\bar{X}_1$ and $\bar{Y}_1$ are independently and uniformly distributed on $\mathbb{S}^{d-1}$ we have

$$\mathcal{Q}_{d,\rho} = \int_{\mathbb{S}^{d-1}} \frac{\mathsf{Vol}(\mathcal{B}_\rho(y_1))}{\mathsf{Vol}(\mathbb{S}^{d-1})^2} dy_1, \tag{84}$$

where the volume of a set $\mathcal{A} \subset \mathbb{R}^d$ is defined as $\mathsf{Vol}(\mathcal{A}) = \int_{\mathcal{A}} dx$, where the integration is with respect to Lebesgue's measure on the sphere. Let us bound $\mathsf{Vol}(\mathcal{B}_\rho)$ from above. To that end, define the set

$$\tilde{\mathcal{B}}_\rho(y_1) = \left\{v \in \mathbb{R}^d : \frac{v}{\|v\|} \in B_\rho(y_1)\right\} = \left\{v \in \mathbb{R}^d : v^T y_1 \geq \rho \|v\|\right\}. \tag{85}$$

Let $f : (v_1, \ldots, v_d) \to (\varphi_1, \ldots, \varphi_{d-1}, r)$ denote the spherical coordinates transformation and let P_{d-1} denote the projection of an element from $\mathbb{R}^d$ onto its first $d-1$ coordinates. We observe that since $\tilde{\mathcal{B}}_\rho(y_1)$ is the cone defined by the spherical cap $\mathcal{B}_\rho(y_1)$, it follows that,

$$f(\tilde{\mathcal{B}}_\rho(y_1)) = P(f(B_\rho(y_1))) \times [0, \infty]. \tag{86}$$

Finally, let μ denote the Gaussian measure of a multivariate random vector with mean y_1, and covariance matrix $\Sigma = \sigma^2 I_{d\times d}$, where $\sigma^2 \geq 0$. Then,

$$1 = \int_{x \in \mathbb{R}^d} \mu(\mathrm{d}x) \tag{87}$$

$$= \int_{\mathbb{R}^d} \frac{1}{(2\pi\sigma^2)^{d/2}} e^{-\frac{1}{2\sigma^2}\|v - y_1\|_2^2} dv \tag{88}$$

$$\geq \int_{\tilde{\mathcal{B}}_\rho(y_1)} \frac{1}{(2\pi\sigma^2)^{d/2}} e^{-\frac{1}{2\sigma^2}\|v - y_1\|_2^2} dv \tag{89}$$

$$\overset{(a)}{\geq} \int_{\tilde{\mathcal{B}}_\rho(y_1)} \frac{1}{(2\pi\sigma^2)^{d/2}} e^{-\frac{1}{2\sigma^2}(\|v\|^2 - 2\rho\|v\| + 1)} dv \tag{90}$$

$$\overset{(b)}{=} \frac{1}{(2\pi\sigma^2)^{d/2}} \int_{f(\tilde{\mathcal{B}}_\rho(y_1))} e^{-\frac{1}{2\sigma^2}(r^2 - 2\rho r + 1)}\left(r^{d-1}\prod_{i=1}^{d-1}\sin^{d-1-i}(\varphi_i)\right) d\varphi_1 \cdots d\varphi_{d-1} dr \tag{91}$$

$$\overset{(c)}{=} \frac{1}{(2\pi\sigma^2)^{d/2}} \underbrace{\int_{P(f(\mathcal{B}_\rho(y_1)))}\left(\prod_{i=1}^{d-1}\sin^{d-1-i}(\varphi_i)\right) d\varphi_1 \cdots d\varphi_{d-1}}_{\mathsf{Vol}(\mathcal{B}_\rho(y_1))} \cdot \int_0^\infty r^{d-1} e^{-\frac{1}{2\sigma^2}(r^2 - 2\rho r + 1)} dr \tag{92}$$

$$\geq \frac{1}{(2\pi\sigma^2)^{d/2}} \mathsf{Vol}(\mathcal{B}_\rho(y_1)) \int_\rho^\infty \rho^{d-1} e^{-\frac{1}{2\sigma^2}(r^2 - 2\rho r + 1)} dr \tag{93}$$

12

$$= \frac{1}{(2\pi\sigma^2)^{(d-1)/2}} \text{Vol}(\mathcal{B}_\rho(y_1)) e^{-\frac{1-\rho^2}{2\sigma^2}} \rho^{d-1} \int_\rho^\infty \frac{1}{(2\pi\sigma^2)^{1/2}} e^{-\frac{(r-\rho)^2}{2\sigma^2}} \, dr \tag{94}$$

$$= \frac{\rho^{d-1}}{(2\pi\sigma^2)^{(d-1)/2}} e^{-\frac{1-\rho^2}{2\sigma^2}} \text{Vol}(\mathcal{B}_\rho(y_1)) \frac{1}{2}, \tag{95}$$

where (a) follows from the definition of $\tilde{\mathcal{B}}_\rho(y_1)$, (b) follows from change of variables and (c) follows from (86).

Thus, for every $y_1 \in \mathbb{S}^{d-1}$,

$$\text{Vol}(\mathcal{B}_\rho(y_1)) \leq \min_{\sigma^2 \geq 0} 2e^{\frac{1-\rho^2}{2\sigma^2} + \frac{d-1}{2}\log(2\pi\sigma^2)} \rho^{1-d} \tag{96}$$

$$= 2e^{\frac{d-1}{2}\log\left(2\pi e \frac{1-\rho^2}{d-1}\right)} \rho^{1-d}. \tag{97}$$

On the other hand, it is well-known that,

$$\text{Vol}(\mathbb{S}^{d-1}) = \frac{2\pi^{d/2}}{\Gamma(d/2)} \geq \frac{2\pi^{d/2}}{(d/2)^{d/2-1}} = \frac{4}{d} e^{\frac{d}{2}\log(2\pi/d)}, \tag{98}$$

where we have used the fact that $\Gamma(x) < x^{x-1}$, for $x > 1$ (see, e.g., (Anderson & Qiu, 1997)). Combining (98), (97), and (84), we obtain

$$\mathcal{Q}_{d,\rho} \leq f(d) e^{\frac{d-1}{2}\log(1-\rho^2)} \rho^{1-d}, \tag{99}$$

where $f(d) \triangleq \frac{d}{2}\left(\frac{ed}{d-1}\right)^{d/2}\left(\frac{2\pi e}{d-1}\right)^{1/2}$. Finally, using (83) and (99), we see that (68) can be further upper bounded as

$$\mathbb{P}_{\mathcal{H}_0}(\phi_{\text{count}} = 1) \leq 8nf(d) e^{\frac{d-1}{2}\log(1-\rho^2)} \rho^{1-d} = 8f(d)n(\rho^{-2} - 1)^{\frac{d-1}{2}}, \tag{100}$$

while (74) can be upper bounded as,

$$\mathbb{P}_{\mathcal{H}_1}(\phi_{\text{count}} = 0) \leq \frac{16}{n}. \tag{101}$$

Thus, for a fixed $d \geq 2$, it is clear that $\mathbb{P}_{\mathcal{H}_1}(\phi_{\text{count}} = 0) \to 0$, as $n \to \infty$, and $\mathbb{P}_{\mathcal{H}_0}(\phi_{\text{count}} = 1) \to 0$, if $n(\rho^{-2} - 1)^{\frac{d-1}{2}} = o(1)$. The later holds if $\rho^{-2} = 1 + o(n^{-2/(d-1)})$, which implies that $\rho^2 = 1 - o(n^{-2/(d-1)})$, as stated.

B. Lower Bounds

As in many detection problems, evaluating the minimax risk function opposes a great challenge due to the error term obtained by maximizing over the error for all permutations in $\mathbb{S}_n$. A well known strategy for overcoming this inherent obstacle is by considering the softer average-case version of the problem. Let π be the uniform measure on $\mathbb{S}_n$, and let us denote by $\mathbb{P}_{\mathcal{H}_1}$ the probability measure obtained by averaging $\mathbb{P}_{\mathcal{H}_1|\sigma}$ with respect to π. For a test ϕ, we consider the Bayesian risk function given by

$$\bar{\mathsf{R}}(\phi) \triangleq \mathbb{P}_{\mathcal{H}_0}[\phi(\mathsf{X}, \mathsf{Y}) = 1] + \mathbb{E}_{\sigma \sim \pi}\left[\mathbb{P}_{\mathcal{H}_1|\sigma}[\phi(\mathsf{X}, \mathsf{Y}) = 0]\right],$$

and the Bayesian risk for our problem:

$$\bar{\mathsf{R}}^\star \triangleq \inf_\phi \bar{\mathsf{R}}(\phi).$$

Clearly, any test ϕ satisfies $\mathsf{R}(\phi) \geq \bar{\mathsf{R}}(\phi)$, and therefore $\mathsf{R}^\star \geq \bar{\mathsf{R}}^\star$. We conclude that in order to prove Theorem 2.5, it is sufficient to show that under the given assumptions, $\bar{\mathsf{R}}^\star = 1 + o(1)$. Using a well-known equivalent characterization of the Bayesian risk function by the total variation distance and Cauchy-Schwartz inequality one shows that

$$\mathsf{R}^\star \geq \bar{\mathsf{R}}^\star = 1 - d_{\text{TV}}(\mathbb{P}_{\mathcal{H}_0}, \mathbb{P}_{\mathcal{H}_1}) \geq 1 - \frac{1}{2}\sqrt{\mathbb{E}_0[L^2] - 1}, \tag{102}$$

where $L \triangleq \frac{\mathbb{P}_{\mathcal{H}_1}}{\mathbb{P}_{\mathcal{H}_0}}$ is the likelihood ratio, and the expectation is taken with respect to $\mathbb{P}_{\mathcal{H}_0}$. Using the bound given in (102), it is sufficient to show that under the assumptions of Theorem 2.5, $\mathbb{E}_0\big[L^2\big] \leq 1 + o(1)$.

Inspired by (Wu et al., 2020), Zeynep and Nazer (K & Nazer, 2022b) gave an exact description of $\mathbb{E}_0\big[L^2\big]$ using the distribution of cycles in a uniformly drawn random permeation. In order to prove Theorem 2.5, we shall carefully analyse $\mathbb{E}_0[L^2]$, and improve the bounds proved (K & Nazer, 2022b). For completeness of the paper, we outline the main ideas behind Nazer and Zeynep's calculation of $\mathbb{E}_0\big[L^2\big]$ before we proceed toward our refined analysis.

The first step in the calculation calls for a use of Ingster-Suslina method, stating that by Fubini's theorem, $\mathbb{E}_0[L^2]$ may be equivalently written as

$$\mathbb{E}_0[L^2] = \mathbb{E}_{\pi \perp\!\!\!\perp \pi'}\left[\mathbb{E}_0\left[\frac{\mathbb{P}_{\mathcal{H}_1|\pi}}{\mathbb{P}_{\mathcal{H}_0}} \cdot \frac{\mathbb{P}_{\mathcal{H}_1|\pi'}}{\mathbb{P}_{\mathcal{H}_0}}\right]\right], \tag{103}$$

where the expectation is taken with respect to the independent coupling of π and π', two copies of the uniform measure on $\mathbb{S}_n$. For fixed permutations σ and σ', we note that $\mathbb{P}_{\mathcal{H}_1|\sigma}, \mathbb{P}_{\mathcal{H}_1|\sigma'}$ and $\mathbb{P}_0$ are absolutely continuous with respect to Lebesgue's measure on $\mathbb{R}^{2 \times d \times n}$ and therefore we have

$$\frac{\mathbb{P}_{\mathcal{H}_1|\sigma}}{\mathbb{P}_{\mathcal{H}_0}} = \frac{f_{\mathcal{H}_1|\sigma}}{f_{\mathcal{H}_0}} \quad \text{and} \quad \frac{\mathbb{P}_{\mathcal{H}_1|\sigma'}}{\mathbb{P}_{\mathcal{H}_0}} = \frac{f_{\mathcal{H}_1|\sigma'}}{f_{\mathcal{H}_0}},$$

where $f_{\mathcal{H}_i}$ denotes the Radon-Nikodym derivative of $\mathbb{P}_{\mathcal{H}_i}$ with respect to Lebesgue's measure, which is the density function (X, Y) under the corresponding hypothesis. Let $\mathcal{N}_\rho : \mathbb{R}^{2 \times d} :\to \mathbb{R}_+$ denote the density function of a pair of random vectors distributed as

$$\mathcal{N}^{\otimes d}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right).$$

We note that

$$\frac{f_{\mathcal{H}_1|\sigma}(\mathsf{X}, \mathsf{Y})}{f_{\mathcal{H}_0}(\mathsf{X}, \mathsf{Y})} \frac{f_{\mathcal{H}_1|\sigma'}(\mathsf{X}, \mathsf{Y})}{f_{\mathcal{H}_0}(\mathsf{X}, \mathsf{Y})} = \prod_{i=1}^n \frac{\mathcal{N}_\rho(X_i, Y_{\sigma_i})}{\mathcal{N}_0(X_i, Y_{\sigma_i})} \frac{\mathcal{N}_\rho(X_i, Y_{\sigma'_i})}{\mathcal{N}_0(X_i, Y_{\sigma'_i})}. \tag{104}$$

In order to proceed with the calculation, we make two key observations. First, we note that the distribution under the null hypothesis $(\mathcal{H}_0)$ is invariant to reordering the coordinates. In a similar manner, the uniform measure on $\mathbb{S}_n$, is invariant under composition with a fixed permutation. Thus,

$$\mathbb{E}_{\pi \perp\!\!\!\perp \pi'}\left[\mathbb{E}_0\left[\frac{\mathbb{P}_{\mathcal{H}_1|\pi}}{\mathbb{P}_{\mathcal{H}_0}} \cdot \frac{\mathbb{P}_{\mathcal{H}_1|\pi'}}{\mathbb{P}_{\mathcal{H}_0}}\right]\right] = \mathbb{E}_\pi\left[\mathbb{E}_0\left[\frac{\mathbb{P}_{\mathcal{H}_1|\pi}}{\mathbb{P}_{\mathcal{H}_0}} \cdot \frac{\mathbb{P}_{\mathcal{H}_1|\mathsf{Id}}}{\mathbb{P}_{\mathcal{H}_0}}\right]\right]. \tag{105}$$

We consider the product given in (104) for a fixed $\sigma \in \mathbb{S}_n$ and $\sigma' = \mathsf{Id}$, which we denote by Z_σ. The second key observation, is that Z_σ can be decomposed to independent terms, corresponding to the cycles of the permeation σ. We recall that a cycle of a permutation σ is a string $(i_0, i_2, \ldots, i_{|C|-1})$ of elements in $[n]$ such that $\sigma(i_j) = i_{j+1 \mod |C|}$ for all j. If $|C| = k$, we call C a k-cycle. For a fixed cycle C, we denote

$$Z_C \triangleq \prod_{i \in C} \frac{\mathcal{N}_\rho(X_i, Y_{\sigma_i})}{\mathcal{N}_0(X_i, Y_{\sigma_i})} \frac{\mathcal{N}_\rho(X_i, Y_i)}{\mathcal{N}_0(X_i, Y_i)}.$$

Since the set of cycles of a permutation induce a partition of $[n]$, the random variables $\{Z_C\}_C$, corresponding to all cycles of σ, are independent (with respect to $\mathbb{P}_{\mathcal{H}_0}$) and

$$Z_\sigma = \prod_C Z_C. \tag{106}$$

The following lemma states that for a fixed cycle C, $\mathbb{E}_0[Z_C]$ depends on ρ and $|C|$. The proof of the lemma is based on the properties of Gaussian random vectors. For further details the reader is referred to (K & Nazer, 2022b)[Lemma 10].

Lemma B.1. *For a fixed cycle C of a permutation σ,*

$$\mathbb{E}_0[Z_C] = \frac{1}{(1 - \rho^{2|C|})^d}.$$

For a fixed permutation $\sigma \in \mathbb{S}_n$ and $k \in [n]$, let $N_k(\sigma)$ denote the number of k-cycles of σ. Combining (103), (105), (106), and Lemma B.1 we obtain

$$\mathbb{E}_0[L^2] = \mathbb{E}_\pi\left[\prod_C Z_C\right] = \mathbb{E}_\pi\left[\prod_{k=1}^n \frac{1}{(1-\rho^{2k})^{dN_k}}\right]. \tag{107}$$

By analysis of (107), Zeynep and Nazer showed in (K & Nazer, 2022b)[Lemma 3] that $E_0[L^2] \le (1-\rho^2)^{-dn}$, which equals $1 + o(1)$ if $\rho^2 = o((nd)^{-1})$. Inspired by the calculation performed in (Wu et al., 2020)[Proposition 2], we carefully bound (107) from above, utilizing the statistical properties of k-cycles in a uniformly distributed random permutation. Our refined analysis enables us to prove that $\mathbb{E}_0[L^2] \le 1 + o(1)$ assuming only that $\rho^2 = o(d^{-1})$. The following proposition is makes the main argument for the proof of our lower bounds given in Theorem 2.5 and Theorem 2.6.

Proposition B.2. *Let N_k be the number of k-cycles in a uniformly distributed permutation π, and $1 \le k \le n$. Then:*

1. *For all (ρ^2, d, n) is holds that*

$$\mathbb{E}_\pi\left[\prod_{k=1}^n \left(\frac{1}{1-\rho^{2k}}\right)^{dN_k}\right] \le \exp\left(\frac{nd\rho^2}{1-\rho^2}\right). \tag{108}$$

2. *If at least one of n, d tends to ∞ and $\rho^2 = o(d^{-1})$, then*

$$\mathbb{E}_\pi\left[\prod_{k=1}^n \left(\frac{1}{1-\rho^{2k}}\right)^{dN_k}\right] \le 1 + o(1). \tag{109}$$

3. *If both n, d tends to ∞ and $\rho^2 < (1-\varepsilon)d^{-1}$ for some $\varepsilon > 0$ then*

$$\mathbb{E}_\pi\left[\prod_{k=1}^n \left(\frac{1}{1-\rho^{2k}}\right)^{dN_k}\right] \le (1+o(1))\exp\left(\frac{d\rho^2}{1-\rho^2} + \frac{c(d,\rho^2)\rho^4}{1-\rho^4}\right). \tag{110}$$

4. *If d and ρ^2 are a constant satisfying $d < \frac{\log(\rho^2)}{\log(1-\rho^2)}$ and $n \to \infty$*

$$\mathbb{E}_\pi\left[\prod_{k=1}^n \left(\frac{1}{1-\rho^{2k}}\right)^{dN_k}\right] \le (1+o(1))\exp\left(\frac{d\rho^2}{1-\rho^2} + \frac{c(d,\rho^2)\rho^4}{1-\rho^4}\right), \tag{111}$$

where $c(d,\rho^2) = \frac{d(d+1)}{2(1-\rho^2)^{d+2}}$.

For the proof of this proposition, we shall require several technical results. The following lemma concerns the approximation of the joint distribution of k-cycles by independent Poisson random variables.

Lemma B.3. *(Arratia & Tavaré, 1992)[Theorem 2] Let $1 \le k \le n$ be an integer, and let $Z_1 \ldots, Z_k$ be independent random variables such that for all $1 \le i \le k$, $Z_i \sim \text{Poisson}(i^{-1})$. Then, the total variation between the law of $N_1, \ldots, N_k$ and $Z_1, \ldots, Z_k$ satisfies*

$$d_{\text{TV}}(\mathcal{L}(N_1, N_2, \ldots, N_k), \mathcal{L}(Z_1, Z_2, \ldots, Z_k)) \le F\left(\frac{n}{k}\right),$$

where $F(x)$ is a monotone decreasing function satisfying $\log F(x) = -x \log x(1+o(1))$ as $x \to \infty$ and we recall that $\mathcal{L}(\cdot)$ denote the law, that is, the probability distribution, of a given random variable.

Lemma B.4. *Let $1 \le m \le n$ be an integer, and let $Z_1 \ldots, Z_m$ be independent random variables such that for all $1 \le i \le m$, $Z_i \sim \text{Poisson}(i^{-1})$. Then,*

$$\mathbb{E}_\pi\left[\prod_{k=1}^m \left(\frac{1}{1-\rho^{2k}}\right)^{dZ_k}\right] \le \exp\left(\frac{d\rho^2}{1-\rho^2} + \frac{c(d,\rho^2)\rho^4}{1-\rho^4}\right),$$

where $c(d,\rho^2) = \frac{d(d+1)}{2(1-\rho^2)^{d+2}}$, and therefore if $\rho^2 = o(d^{-1})$,

$$\mathbb{E}_\pi\left[\prod_{k=1}^m \left(\frac{1}{1-\rho^{2k}}\right)^{dZ_k}\right] \le 1 + o(1).$$

Proof. The proof of this lemma is elementary, only requires the moment generating function of Poisson random variables, and some linear approximations of elementary functions. By rearranging the expression in the expectation and using independence we have

$$
\mathbb{E}_\pi\left[\prod_{k=1}^m\left(\frac{1}{1-\rho^{2k}}\right)^{dZ_k}\right] = \prod_{k=1}^m\mathbb{E}_\pi\left[\left(\frac{1}{1-\rho^{2k}}\right)^{dZ_k}\right]
$$

$$
= \prod_{k=1}^m\mathbb{E}_\pi\left[\exp\left(-dZ_k\log\left(1-\rho^{2k}\right)\right)\right]
$$

$$
\overset{(a)}{=} \prod_{k=1}^m\exp\left(\frac{1}{k}\left(e^{-d\log(1-\rho^{2k})}-1\right)\right)
$$

$$
= \prod_{k=1}^m\exp\left(\frac{1}{k}\left(\frac{1}{(1-\rho^{2k})^d}-1\right)\right).
$$

where (a) is followed by the definition of the moment generating function of a Poisson random variable.

We shall now bound the term $(1-\rho^{2k})^{-d}-1$ from above. A straight forward calculation of the Taylor expansion of the function $f(x) = (1-x)^{-d}$ show that for $x \in (0,1)$ we have

$$
\frac{1}{(1-x)^d} = \sum_{m=0}^\infty \binom{m+d-1}{d-1}x^m.
$$

Using Lagrange's remainder theorem and obtain that for all $x > (0,1)$,

$$
\frac{1}{(1-x)^d} = 1 + dx + \frac{d(d+1)}{2}\cdot\frac{1}{(1-c)^{(d+2)}}x^2
$$

$$
\leq 1 + dx + \frac{d(d+1)}{2(1-x)^{(d+2)}}x^2.
$$

where c is a point in $[0,x]$. Choosing $x = \rho^{2k}$ we get that for any $k > 0$

$$
\frac{1}{(1-\rho^{2k})^d} \leq 1 + d\rho^{2k} + \frac{d(d+1)}{2(1-\rho^{2k})^{(d+2)}}\rho^{4k}
$$

$$
\leq 1 + d\rho^{2k} + \frac{d(d+1)}{2(1-\rho^2)^{(d+2)}}\rho^{4k}.
$$

For the rest of our analysis we denote

$$
c(d,\rho^2) \triangleq \frac{d(d+1)}{2(1-\rho^2)^{(d+2)}}\rho^{4k},
$$

and we get

$$
\mathbb{E}_\pi\left[\prod_{k=1}^m\left(\frac{1}{1-\rho^{2k}}\right)^{dZ_k}\right] = \prod_{k=1}^m\mathbb{E}_\pi\left[\left(\frac{1}{1-\rho^{2k}}\right)^{dZ_k}\right]
$$

$$
= \prod_{k=1}^m\exp\left(\frac{1}{k}\left(\frac{1}{(1-\rho^{2k})^d}-1\right)\right)
$$

$$
\leq \prod_{k=1}^m\exp\left(\frac{1}{k}\left(1+\left(d\rho^{2k}+c(d,\rho^2)\rho^{4k}\right)-1\right)\right)
$$

$$
= \exp\left(d\sum_{k=1}^m\frac{\rho^{2k}}{k}+c(d,\rho^2)\sum_{k=1}^m\frac{\rho^{4k}}{k}\right)
$$

$$
\leq \exp\left(d\sum_{k=1}^\infty\frac{\rho^{2k}}{k}+c(d,\rho^2)\sum_{k=1}^\infty\frac{\rho^{4k}}{k}\right)
$$

$$= \exp\left(-d\log(1-\rho^2) - c(d,\rho^2)\log(1-\rho^4)\right)$$

$$\overset{(a)}{\leq} \exp\left(\frac{d\rho^2}{1-\rho^2} + \frac{c(d,\rho^2)\rho^4}{1-\rho^4}\right),$$

where (a) follows from the well known inequality $\log(1+x) \geq x/(1+x)$, for $x > -1$. In the case where $\rho^2 = o(d^{-1})$ clearly $d\rho^2 = o(1)$. Furthermore,

$$c(d,\rho^2)\rho^4 = \frac{d(d+1)}{2(1-\rho^2)^{d+2}}\rho^4 \leq \frac{1}{2}\exp((d+2)\rho^2)d(d+1)\rho^4 = \exp(o(1)) \cdot o(1) = o(1).$$

In particular, we get

$$\mathbb{E}_\pi\left[\prod_{k=1}^m \left(\frac{1}{1-\rho^{2k}}\right)^{dZ_k}\right] \leq \exp\left(\frac{d\rho^2}{1-\rho^2} + \frac{c(d,\rho^2)\rho^4}{1-\rho^4}\right) \tag{112}$$

$$= \exp\left(\frac{o(1)}{1-o(1)} + \frac{o(1)}{1-o(1)}\right) = \exp(o(1)) = 1 + o(1). \tag{113}$$

$$\square$$

We are now ready to prove Proposition B.2. The idea of the proof is as follows: we consider the expectation of the product given in (109). In the case that $n \to \infty$, we show that the product of the last $n - \alpha\log n$ terms is always upper-bounded by $1 + o(1)$ for an appropriate choice of α. For the expectation of the product of the first $\alpha\log n$ terms, use the Poisson approximation of $\{N_k\}_k$ given in Lemma B.3 and the estimation in that case, given in Lemma B.4. The other case, where n is constant, is solvable using elementary arguments.

Proof of Proposition B.2. We divide our proof into three parts, with respect to the asymptotic regimes of n and d. We start by a simple observation - whenever n is fixed we have $\sum_{k=1}^n kN_K = n$, which implies that for any $1 \leq m \leq n$ we have

$$\prod_{k=m}^n \left(\frac{1}{1-\rho^{2k}}\right)^{dN_k} \leq \prod_{k=m}^n \left(\frac{1}{1-\rho^{2m}}\right)^{dN_k} = \left(\frac{1}{1-\rho^{2m}}\right)^{d\sum_{k=m}^n N_k} \tag{114}$$

$$\leq \left(\frac{1}{1-\rho^{2m}}\right)^{dn} = \left(1 + \frac{\rho^{2m}}{1-\rho^{2m}}\right)^{dn} \leq \exp\left(\frac{dn\rho^{2m}}{1-\rho^{2m}}\right). \tag{115}$$

This immediately proves (108).

The case where both n and d tends to ∞: we assume that $n, d \to \infty$ and $\rho^2 = o(d^{-1})$ or $\rho^2 < (1-\varepsilon)d^{-1}$. We choose $m = \lceil \log n \rceil$ and we get

$$dn\rho^{2m} \leq (d\rho^2) \cdot n(\rho^2)^{\log(n)-1} = (d\rho^2)n(\rho^2)^{\log\left(\frac{n}{e}\right)} = e(d\rho^2)\left(\frac{n}{e}\right)^{1+\log\rho^2}.$$

Since $\rho^2 < d^{-1}$ (which is clearly true for sufficiently large d in the particular case where $\rho^2 = o(d^{-1})$), and $d \to \infty$, we have $\log\rho^2 \to -\infty$ as $n \to \infty$. Plugging our chosen m in (115) obtain:

$$\prod_{k=\log n}^n \left(\frac{1}{1-\rho^{2k}}\right)^{dN_k} \leq \exp\left(\frac{dn\rho^{2m}}{1-\rho^{2m}}\right) = \exp(o(1)) = 1 + o(1). \tag{116}$$

For a fixed integer m, we consider the set $S_{n,m} \subseteq \mathbb{N}^m$ given by

$$S_{n,m} = \left\{(n_1,\ldots,n_m) \in \mathbb{N}^d \,\middle|\, \sum_{k=1}^m n_k \leq n\right\},$$

and a function $f_{n,m} : \mathbb{N}^m \to [0, \infty]$ given by

$$f_{n,m}(n_1, n_2, \ldots, n_m) = \prod_{k=1}^{m} \left(\frac{1}{1 - \rho^{2k}} \right)^{dn_k} \cdot \mathbb{1}_{S_{n,m}}(n_1, \ldots, n_m).$$

We note that for all $n_1, \ldots, n_m \in \mathbb{N}$

$$f_{n,m}(n_1, n_2, \ldots, n_m) \le \prod_{k=1}^{m} \left(\frac{1}{1 - \rho^2} \right)^{dn_k} \cdot \mathbb{1}_{S_{n,m}}(n_1, \ldots, n_m) \le \left(\frac{1}{1 - \rho^2} \right)^{dn}. \tag{117}$$

We set $m = \lceil \log n \rceil$, and let $\{Z_k\}_k$ be independent $\mathsf{Poisson}(k^{-1})$ random variables as in Lemma B.3. Since $\sum_{k=1}^{m} N_k \le n$ with probability 1, we have

$$\mathbb{E}_\pi \left[\prod_{k=1}^{m} \left(\frac{1}{1 - \rho^{2k}} \right)^{dN_k} \right] = \mathbb{E}_\pi [f_{n,m}(N_1, \ldots, N_m)]$$

$$\le \mathbb{E}[f_{n,m}(Z_1, \ldots, Z_m)] + d_{\mathsf{TV}}(\mathcal{L}(N_1^m), \mathcal{L}(Z_1^m)) \cdot \|f_{n,m}\|_\infty$$

$$\le \mathbb{E}_\pi \left[\prod_{k=1}^{m} \left(\frac{1}{1 - \rho^{2k}} \right)^{dZ_k} \right] + d_{\mathsf{TV}}(\mathcal{L}(N_1^m), \mathcal{L}(Z_1^m)) \cdot \|f_{n,m}\|_\infty$$

$$\overset{(a)}{\le} \mathbb{E}_\pi \left[\prod_{k=1}^{m} \left(\frac{1}{1 - \rho^{2k}} \right)^{dZ_k} \right] + F\left(\frac{n}{m} \right) \cdot \left(\frac{1}{1 - \rho^2} \right)^{dn}$$

$$\overset{(b)}{\le} \exp\left(\frac{d\rho^2}{1 - \rho^2} + \frac{c(d, \rho^2)\rho^4}{1 - \rho^4} \right) + F\left(\frac{n}{m} \right) \cdot \left(\frac{1}{1 - \rho^2} \right)^{dn}$$

$$= \exp\left(\frac{d\rho^2}{1 - \rho^2} + \frac{c(d, \rho^2)\rho^4}{1 - \rho^4} \right) + F\left(\frac{n}{\lceil \log n \rceil} \right) \cdot \left(\frac{1}{1 - \rho^2} \right)^{dn},$$

where (a) follows from (117) and Lemma B.3, (b) follows from Lemma B.4. By Lemma B.3, we also have

$$\log\left(F\left(\frac{n}{\lceil \log n \rceil} \right) \left(\frac{1}{1 - \rho^2} \right)^{dn} \right) \le \log\left(F\left(\frac{n}{\log n} \right) \left(\frac{1}{1 - \rho^2} \right)^{dn} \right)$$

$$= \log\left(F\left(\frac{n}{\log n} \right) \right) - nd \log(1 - \rho^2)$$

$$\le -(1 + o(1)) \frac{n}{\log n} \log\left(\frac{n}{\log n} \right) - nd \log(1 - \rho^2)$$

$$\overset{(a)}{\le} -(1 + o(1)) \frac{n}{\log n} \log\left(\frac{n}{\log n} \right) + nd\rho^2(1 + o(1))$$

$$= n\left(-(1 + o(1)) \left(1 - \frac{\log \log n}{\log n} \right) + d\rho^2(1 + o(1)) \right)$$

$$= -n(1 - d\rho^2 + o(1)) \xrightarrow[n \to \infty]{(b)} -\infty,$$

where (a) follows from the Taylor expansion of the function $\log(1 - x)$ and $\rho^2 = o(1)$, and (b) follows from the assumption that $\rho^2 < (1 - \varepsilon)d^{-1}$. This implies that

$$F\left(\frac{n}{\lceil \log n \rceil} \right) \left(\frac{1}{1 - \rho^2} \right)^{dn} = o(1),$$

and therefore,

$$\mathbb{E}_\pi \left[\prod_{k=1}^{\lceil \log n \rceil} \left(\frac{1}{1 - \rho^{2k}} \right)^{dN_k} \right] \le \exp\left(\frac{d\rho^2}{1 - \rho^2} + \frac{c(d, \rho^2)\rho^4}{1 - \rho^4} \right) + o(1). \tag{118}$$

Combining (116) and (118) together we conclude:

$$\mathbb{E}_\pi \left[\prod_{k=1}^n \left(\frac{1}{1-\rho^{2k}} \right)^{dN_k} \right] \leq \mathbb{E}_\pi \left[\prod_{k=1}^{\lceil \log n \rceil} \left(\frac{1}{1-\rho^{2k}} \right)^{dN_k} \prod_{k=\lceil \log n \rceil}^n \left(\frac{1}{1-\rho^{2k}} \right)^{dN_k} \right]$$

$$= (1+o(1)) \cdot \mathbb{E}_\pi \left[\prod_{k=1}^{\lceil \log n \rceil} \left(\frac{1}{1-\rho^{2k}} \right)^{dN_k} \right]$$

$$= (1+o(1)) \cdot \exp\left(\frac{d\rho^2}{1-\rho^2} + \frac{c(d,\rho^2)\rho^4}{1-\rho^4} \right) + o(1).$$

We have now proved (110). We note that by assuming furthermore that $\rho^2 = o(d^{-1})$, by the second part of Lemma B.4 we get

$$\mathbb{E}_\pi \left[\prod_{k=1}^n \left(\frac{1}{1-\rho^{2k}} \right)^{dN_k} \right] = (1+o(1)) \exp\left(\frac{d\rho^2}{1-\rho^2} + \frac{c(d,\rho^2)\rho^4}{1-\rho^4} \right) + o(1) = 1 + o(1).$$

We have now proved (109).

The case where n is constant and d tends to ∞: We assume that $\rho^2 = o(d^{-1})$. Since n is constant, we have $dn\rho^2 = o(1)$ and therefore by (115) we have

$$E_\pi \left[\prod_{k=1}^n \left(\frac{1}{1-\rho^{2k}} \right)^{dN_k} \right] \leq \exp\left(\frac{dn\rho^2}{1-\rho^2} \right) = 1 + o(1).$$

The case where d is constant and n tends to ∞: we also assume that ρ^2 is a constant such that $d < \frac{\log(\rho^2)}{\log(1-\rho^2)}$. We repeat the same steps as in the case where $n \to \infty$ and $\rho^2 = o(d^{-1})$ with a minor change. Instead of approximating the product of the first $m = \lceil \log n \rceil$ terms in the product, we take $m' = \lceil \alpha \log(n) \rceil$, where $\alpha = -\frac{1}{\log(\rho^2)} + \varepsilon$, where ε sufficiently small so that

$$d \frac{\log(1-\rho^2)}{\log(\rho^2)} < \frac{1}{1-\varepsilon \log(\rho^2)}, \tag{119}$$

(such exists by the assumption $d < d^\star(\rho^2) = \frac{\log(\rho^2)}{\log(1-\rho^2)}$). Repeating the same steps as in the previous part, we have

$$\prod_{k=m'}^n \left(\frac{1}{1-\rho^{2k}} \right)^{dN_k} \leq \exp\left(\frac{dn\rho^{2m'}}{1-\rho^{2m'}} \right).$$

We observe that

$$nd\rho^{2m'} \leq nd(\rho^2)^{\alpha \log n} = ndn^{\alpha \log(\rho^2)} = dn^{1+\alpha \log(\rho^2)} = dn^{\varepsilon \log(\rho^2)} = o(1),$$

which implies that

$$\prod_{k=m'}^n \left(\frac{1}{1-\rho^{2k}} \right)^{dN_k} \leq \exp\left(\frac{dn\rho^{2m'}}{1-\rho^{2m'}} \right) = 1 + o(1).$$

We now evaluate the product of the first m' terms. In a similar fashion to the first part, by Lemma B.3 and Lemma B.4,

$$\mathbb{E}_\pi \left[\prod_{k=1}^{m'} \left(\frac{1}{1-\rho^{2k}} \right)^{dN_k} \right] \leq \mathbb{E}_\pi \left[\prod_{k=1}^{m'} \left(\frac{1}{1-\rho^{2k}} \right)^{dZ_k} \right] + F\left(\frac{n}{m'} \right) \cdot \left(\frac{1}{1-\rho^2} \right)^{dn}$$

$$\leq \exp\left(\frac{d\rho^2}{1-\rho^2} + \frac{c(d,\rho^2)\rho^4}{1-\rho^4} \right) + F\left(\frac{n}{m'} \right) \cdot \left(\frac{1}{1-\rho^2} \right)^{dn}$$

$$= \exp\left(\frac{d\rho^2}{1-\rho^2} + \frac{c(d,\rho^2)\rho^4}{1-\rho^4} \right) + F\left(\frac{n}{\lceil \alpha \log n \rceil} \right) \cdot \left(\frac{1}{1-\rho^2} \right)^{dn}.$$

Using Lemma B.3 once again, we obtain

$$\log\left(F\left(\frac{n}{\lceil \alpha \log n \rceil} \right) \left(\frac{1}{1-\rho^2} \right)^{dn} \right) \leq -(1+o(1)) \frac{n}{\alpha \log n} \log\left(\frac{n}{\alpha \log n} \right) - nd \log(1-\rho^2)$$

$$= -n\left((1+o(1)) \frac{1}{\alpha} \left(1 - \frac{\log(\alpha \log n)}{\log n} \right) + d\log(1-\rho^2) \right)$$

$$= -n(1+o(1))\left(\frac{1}{\alpha} - d\log(1-\rho^2) \right)$$

$$= -n(1+o(1))\left(\frac{\log(\rho^2)}{1-\varepsilon\log(\rho^2)} - d\log(1-\rho^2) \right).$$

By (119) we have

$$\frac{\log(\rho^2)}{1-\varepsilon\log(\rho^2)} - d\log(1-\rho^2) > 0,$$

which implies that

$$F\left(\frac{n}{m'} \right) \cdot \left(\frac{1}{1-\rho^2} \right)^{dn} = o(1).$$

We now conclude

$$\mathbb{E}_\pi\left[\prod_{k=1}^{n} \left(\frac{1}{1-\rho^{2k}} \right)^{dN_k} \right] = \mathbb{E}_\pi\left[\prod_{k=1}^{m'} \left(\frac{1}{1-\rho^{2k}} \right)^{dN_k} \prod_{k=m'}^{n} \left(\frac{1}{1-\rho^{2k}} \right)^{dN_k} \right]$$

$$= \mathbb{E}_\pi\left[\prod_{k=1}^{m'} \left(\frac{1}{1-\rho^{2k}} \right)^{dN_k} (1+o(1)) \right]$$

$$\leq (1+o(1))\left(\exp\left(\frac{d\rho^2}{1-\rho^2} + \frac{c(d,\rho^2)\rho^4}{1-\rho^4} \right) + o(1) \right)$$

$$= (1+o(1)) \exp\left(\frac{d\rho^2}{1-\rho^2} + \frac{c(d,\rho^2)\rho^4}{1-\rho^4} \right).$$

$\square$

The proofs of Theorem 2.5 and Theorem 2.6 now follows:

Proof of Theorem 2.5. We combine (103), (107) and Proposition B.2 and obtain:

$$\mathsf{R}^\star \geq 1 - \sqrt{\mathbb{E}_0[L^2] - 1}$$

$$= 1 - \sqrt{\mathbb{E}_\pi\left[\prod_{k=1}^{n} \frac{1}{(1-\rho^{2k})^{dN_k}} \right] - 1}$$

$$\geq 1 - \sqrt{1+o(1) - 1}$$

$$= 1 - o(1).$$

$\square$

Proof of Theorem 2.6. Let (ρ, d, n) be a sequence satisfying the assumptions of Theorem 2.6. We start by recalling an important well-known fact (see, for example, (Tsybakov, 2004)[Lemma 2.6 and 2.7]): for any sequence of measures

$(\mathbb{P}_{0,k})_k$, $(\mathbb{P}_{1,k})_k$,

$$\mathbb{E}_{\mathbb{P}_{0,k}}\left[\left(\frac{\mathbb{P}_{1,k}}{\mathbb{P}_{0,k}}\right)^2\right] = O(1) \implies d_{\mathsf{TV}}(\mathbb{P}_{0,k}, \mathbb{P}_{1,k}) = 1 - \Omega(1).$$

Thus, by (102), if $\mathbb{E}_0[L^2] = O(1)$ we have

$$\mathsf{R}^\star \geq 1 - d_{\mathsf{TV}}(\mathbb{P}_{\mathcal{H}_0}, \mathbb{P}_{\mathcal{H}_1}) = \Omega(1), \tag{120}$$

which implies that (ρ, d, n) is not admissible for strong detection (as (120) implies $\limsup \mathsf{R}^\star > 0$).

Indeed, by Proposition B.2 and (107): if d and ρ are constants such that $d < d^\star(\rho^2)$, and $n \to \infty$, then

$$\mathbb{E}_0[L^2] \leq (1 + o(1)) \exp\left(\frac{d\rho^2}{1 - \rho^2} + \frac{c(d, \rho^2)\rho^4}{1 - \rho^4}\right) = O(1).$$

On the other hand, if $d, n \to \infty$ and $\rho^2 < (1 - \varepsilon)d^{-1}$, we have that $\rho^2 = o(1)$. Thus, it also follows from Proposition B.2 that

$$\begin{aligned}
\mathbb{E}_0[L^2] &\leq (1 + o(1)) \exp\left(\frac{d\rho^2}{1 - \rho^2} + \frac{c(d, \rho^2)\rho^4}{1 - \rho^4}\right) \\
&\leq (1 + o(1)) \exp\left((1 + o(1))\left(d\rho^2 + \frac{d(d+1)}{2(1 - \rho^2)^{d+2}}\rho^4\right)\right) \\
&\leq (1 + o(1)) \exp\left((1 + o(1))\left(1 + \frac{1}{2e^{-2}(1 + o(1))}\right)\right) = O(1).
\end{aligned}$$

The remaining case is where $d \to \infty$, n is constant and $\rho^2 = O(d^{-1})$. By (108), which is true without any assumptions on (ρ, d, n), we have

$$\mathbb{E}_0[L^2] \leq \exp\left(\frac{nd\rho^2}{1 - \rho^2}\right) = O(1).$$

That concludes the proof. $\qquad\square$