

**FUNDAMENTAL LIMITS AND ALGORITHMS FOR DATABASE AND GRAPH
ALIGNMENT**

A Dissertation
Presented to
The Academic Faculty

By

Osman Emre Dai

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in
Algorithms, Combination and Optimization

Georgia Institute of Technology

December 2023

© Osman Emre Dai 2023

FUNDAMENTAL LIMITS AND ALGORITHMS FOR DATABASE AND GRAPH ALIGNMENT

Thesis committee:

Dr. Negar Kiyavash
Business Analytics
Ecole Polytechnique Fédérale de Lausanne

Dr. Daniel Cullina
Electrical Engineering and Computer Science
Pennsylvania State University

Dr. Mohit Singh
Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Cheng Mao
Mathematics
Georgia Institute of Technology

Dr. Ashwin Pananjady
Industrial and Systems Engineering and
Electrical and Computer Engineering
Georgia Institute of Technology

Date approved: December 5, 2023

To my mother

ACKNOWLEDGMENTS

I am sincerely grateful to Negar Kiyavash and Daniel Cullina. The guidance and encouragement Negar offered throughout my PhD allowed me to stay motivated and on track and I consider myself fortunate to have embarked on the PhD journey with her as my advisor. Daniel generously acted as a close mentor and co-advisor from early on. Working closely with such a brilliant researcher, while intimidating at the beginning, has been a great pleasure and an invaluable opportunity. I would like to further thank Mohit Singh, who took over the role of official co-advisor at Georgia Tech upon Negar's departure to EPFL, and whose support and guidance has been instrumental. I am also thankful to other members of my dissertation committee, Cheng Mao and Ashwin Pananjady. I owe special thanks to Cheng Mao for serving as thesis reader.

I am grateful to my dear friends Ecem, Ahmet, Hakan, Dilara, Selçuk, Berkay, Can, Çağlar, Aykan, Çağrı, Cemre, Furkan, Utku, Selin, Eldon, Eshan, Alfredo, Stella, Mo, Mitch, Caner and Kantwon, whose support and joyful company, whether in Atlanta or from afar, has meant a lot to me. I want to especially thank my dear friends Cansu, Edanur and Ömer whose support helped me through a particularly challenging period in my life. I owe special thanks to Edanur and Caner who generously offered me their homes, where I was able to focus and write this dissertation.

Finally, I want to thank my mother, Meltem, whose love and support have been my greatest fortune.

TABLE OF CONTENTS

Acknowledgments	iv
List of Tables	x
List of Figures	xi
List of Acronyms	xiii
Summary	xiv
Chapter 1: Introduction and Related Work	1
1.1 Introduction	1
1.2 Outline of thesis	2
1.3 Related work	4
1.3.1 Graph alignment	4
1.3.2 Database alignment	6
1.3.3 Planted matching	6
Chapter 2: Preliminaries for Gaussian planted matching and Gaussian database alignment	8
2.1 Models	8
2.1.1 Notation	8

2.1.2	Planted Matching on Gaussian Bigraphs	9
2.1.3	Correlated Gaussian Databases	9
2.1.4	Particular settings for Gaussian databases	11
2.2	Algorithms for Planted Matching and Database Alignment	11
2.2.1	Information density matrix for database alignment	12
2.2.2	Log-likelihood for planted matching	12
2.2.3	Algorithms	13
2.2.4	Performance metrics for algorithms	15
2.2.5	Computational complexity	16
2.3	Combinatorial Analysis of Misalignment	19
2.3.1	Elementary misalignments between mappings	19
2.3.2	Counting elementary misalignments	22
2.4	Canonical Form of Correlation Statistics for Correlated Gaussian Databases	24
Chapter 3: Gaussian Planted Matching		27
3.1	Sketch of analysis	28
3.2	Concentration Inequalities	29
3.3	Achievability	34
3.3.1	Asymptotic results	34
3.3.2	Finite-size results	37
3.3.3	Proof for Threshold Testing	39
3.3.4	Proofs for Maximum Row Estimation	41
3.3.5	Proof for Maximum likelihood estimation	45

3.4	Converse	55
3.4.1	Converse statement	55
3.5	Interpretation and intuition behind results	56
3.5.1	Merging of boundary for maximum likelihood estimation and maximum row estimation	56
3.5.2	Transition from the elliptic to the quadratic boundary for maximum likelihood estimation	57
3.5.3	Transition between quadratic boundary to linear boundary of maximum likelihood estimation and maximum row estimation	58
3.5.4	Halving of slope of linear boundary of maximum likelihood estimation	58
3.5.5	Vertical segment of boundary of maximum likelihood estimation	59
3.5.6	Gap between maximum row estimation and threshold testing	60
Chapter 4: Gaussian database alignment in the low per-feature correlation limit		61
4.1	Low Per-Feature Correlation Regime	62
4.2	Generating function for information density matrix	64
4.2.1	Main lemmas on the generating function	69
4.2.2	Generating function evaluated for cycles and even paths	73
4.3	Concentration Inequalities	83
4.3.1	Statements of concentration inequalities and their proofs	84
4.3.2	Geometric intuition behind concentration inequalities	87
4.4	Achievability	91
4.4.1	Asymptotic results	92
4.4.2	Finite-size results	94

4.4.3	Proof for Threshold Testing	95
4.4.4	Proofs for Maximum Row Estimation	96
4.4.5	Proof for Maximum likelihood estimation	99
4.5	Exact alignment converse	103
Chapter 5: Gaussian Database Alignment with Features of Arbitrary Dimension		110
5.1	Threshold testing	111
5.1.1	Achievability	111
5.1.2	Converse for almost-exact alignment	123
5.2	Maximum row estimation	125
5.3	Majorization of distributions	129
5.3.1	Majorization	129
5.3.2	Lorenz ordering and majorization for distributions	129
5.3.3	Lorenz ordering of terms that make up the generating function	131
5.4	Maximum likelihood estimation	144
5.4.1	Sketch of proof	145
5.4.2	Convexity of generating function	148
5.4.3	Normalized error exponent	149
5.4.4	Critical point	151
5.4.5	Proof of achievability result	154
Chapter 6: Analysis of a Canonical Labeling Algorithm for the Alignment of Correlated Erdős-Rényi Graphs		156
6.1	Model	157

6.1.1	Notation	157
6.1.2	Problem Definition	157
6.1.3	Alignment by Canonical Labeling	158
6.1.4	Correlated Erdős-Rényi Graphs	161
6.1.5	Outline and Intuition for the Analysis	164
6.2	Achievability result	167
6.3	Anchor alignment	170
6.4	Bipartite graph alignment	178
6.5	General alignment algorithm	183
Appendices		188
	Appendix A: Statistics of the information density matrix for Gaussian database alignment	189
	Appendix B: Other lemmas	204
References		213

LIST OF TABLES

3.1	Sufficient conditions for almost-exact alignment for Gaussian planted matching	35
3.2	Sufficient conditions for exact alignment for Gaussian planted matching	36
3.3	Sufficient conditions for error bounds for Gaussian planted matching in balanced case	37
3.4	Sufficient conditions for error bounds for Gaussian planted matching in unbalanced case	39
3.5	Error bounds for Gaussian planted matching with finite-size	40
3.6	Necessary conditions for error bound for Gaussian planted matching	55
4.1	Sufficient conditions for almost-exact alignment for Gaussian database alignment	92
4.2	Sufficient conditions for exact alignment for Gaussian database alignment	93
4.3	Sufficient conditions for error bounds for Gaussian database alignment in balanced case	94
4.4	Sufficient conditions for error bounds for Gaussian database alignment in unbalanced case	95
4.5	Error bounds for Gaussian database alignment with finite-size	95

LIST OF FIGURES

2.1	Illustration of a pair of databases	9
2.2	Examples of 3 types of elementary misalignments as bigraphs	20
2.3	Examples of 3 types of elementary misalignments as matrices	21
3.1	Illustration of the partition of the error event into atypicality and misalignment-despite-typicality events	30
3.2	Comparison of boundaries for achievability regions of exact alignment and almost-exact alignment	35
3.3	Comparison of the boundaries for the achievability regions in the balanced case	36
3.4	Comparison of boundaries for achievability regions of the maximum likelihood algorithm for various values of α	38
3.5	Comparison of the boundaries for the achievability regions when $ \mathcal{U} = M = n$ and $ \mathcal{V} = n^{1.5}$	38
4.1	2-dimensional projection of probability space. $G_{u,v}$ corresponds to a true pair and has mean ζ , while $G_{u,v'}$ corresponds to a false pair and has mean $-\zeta$	87
4.2	2-dimensional projection of the regions of the probability space corresponding to atypicality events	88
4.3	2-dimensional projection of the regions of the probability space corresponding to typicality	89
4.4	2-dimensional projection of the half-space corresponding to misalignment	89

4.5	Partition of misalignment event	90
4.6	Covering of partition of misalignment event	90
5.1	Comparison of boundaries for achievability regions for threshold testing for various values of the number of dimensions	112
5.2	Comparison of boundaries for achievability regions for maximum row estimation for various values of the number of dimensions	126
5.3	Comparison of non-standardized Lorenz curves for W_2, W_3, W_4, W_5, W_6 and W_∞	133
5.4	Comparison non-standardized Lorenz curves for W_2, Z_ℓ, W_ℓ and W_∞ for $\ell \in \{3, 4, 5, 6\}$	133
5.5	Comparison of boundaries for achievability regions for maximum likelihood estimation for various values of the number of dimensions	145
5.6	Comparison of error bound exponents for elementary misalignments of different sizes in the $\log n$ -dimensional case	146
5.7	Comparison of error bound exponents for elementary misalignments of different sizes in the $1000 \log n$ dimensional case	146
5.8	Comparison of error bound exponents for elementary misalignments of different sizes in the $0.15 \log n$ -dimensional case	147
6.1	Illustration of AnchorSignAlign	160
6.2	Illustration of the degree separation argument for the success of anchor alignment	165
6.3	Illustration for the packing argument that gives intuition on the upper bound on h	166
6.4	Comparison of regions of achievability for symmetric noise	168

SUMMARY

Data alignment refers to a class of problems where given two sets of anonymized data pertaining to overlapping sets of users, the goal is to identify the correspondences between the two sets. If the data of a user is contained in both sets, the correlation between the two data points associated with the user might make it possible to determine that both belong to the same user and hence link the data points. Alignment problems are of practical interest in applications such as privacy and data junction. Data alignment can be used to de-anonymize data, therefore, studying the feasibility of alignment allows for a more reliable understanding of the limitations of anonymization schemes put in place to protect against privacy breaches. Additionally, data alignment can aid in finding the correspondence between data from different sources, e.g. different sensors. The data fusion performed through data alignment in turn can help with variety of inference problems that arise in scientific and engineering applications.

This thesis considers two types of data alignment problems: database and graph alignment. Database alignment refers to the setting where each feature (i.e. data points) in a data set is associated with a single user. Graph alignment refers to the setting where data points in each data set are associated with pairs of users. For both problems, we are particularly interested in the asymptotic case where n , the number of users with data in both sets, goes to infinity. Nevertheless our analyses often yield results applicable to the finite n case. To develop a preliminary understanding of the database alignment problem, we first study the closely related problem of planted matching with Gaussian weights of unit variance, and derive tight achievability bounds that match our converse bounds: Specifically we identify different inequalities between $\log n$ and the signal strength (which corresponds to the square of the difference between the mean weights of planted and non-planted edges) that guarantee upper bounds on the log of the expected number of errors. Then, we study the database alignment problem with Gaussian features in the low per-feature correlation

setting where the number of dimensions of each feature scales as $\omega(\log n)$: We derive inequalities between $\log n$ and signal strength (which, for database alignment, corresponds to the mutual information between correlated features) that guarantee error bounds matching those of the planted matching setting, supporting the claimed connection between the two problems. Then, relaxing the restriction on the number of dimensions of features, we derive conditions on signal strength and dimensionality that guarantee smaller upper bounds on the log of the expected number of errors. The stronger results in the $\mathcal{O}(\log n)$ -dimensional-feature setting for Gaussian databases show how planted matching, while useful, is not a perfect substitute to understand the dynamics of the more complex problem of database alignment. For graph alignment, we focus on the correlated Erdős-Rényi graph model where the data point (i.e. edge) associated with each pair of users in a graph is a Bernoulli random variable that is correlated with the data point associated with the same pair in the other graph. We study a canonical labeling algorithm for alignment and identify conditions on the density of the graphs and correlation between edges across graphs that guarantees the recovery of the true alignment with high probability.

CHAPTER 1

INTRODUCTION AND RELATED WORK

1.1 Introduction

The modern ubiquity of data collection has drawn interest to the problem of data alignment which could be described as follows. Consider two data sets containing information pertaining to various anonymized users. The data associated with a particular user is correlated across the two data sets. Alignment refers to the problem of identifying such correlated pairs. This alignment can enable data fusion (e.g. in applications the field of computational biology [1] or computer vision [2]) or de-anonymization (with several high-profile instances, such as the 2006 Netflix Prize incident [3] or 2016 release of the MBS/PBS healthcare data [4]).

Database alignment and graph alignment are two well studied versions of the alignment problem. In the former setting, the data consists of multi-dimensional features, each associated with an individual user. These features are correlated across the two databases only if they are associated with the same user. In the graph setting, features are associated with pairs of users, and they are correlated across the two graphs only if the pairs match.

Database alignment is closely related to the so-called planted matching problem: The planted matching problem, specifically refers to the setting where we are given a complete bipartite graph with random edge weights and an underlying true matching. Edges that are part of the underlying matching have one probability distribution, while the un-matched edges have another distribution. All edge weights are independently distributed.

For database alignment, the quality of matches between pairs of features across the databases can be represented on a complete bipartite graph with weighted edges: Each user feature corresponds to a vertex and the weight of edges between a pair of vertices is

a measure of the quality of the match of the corresponding pair of features. The weight of the edge between features corresponding to a single user is likely to be high, while the weights of other edges are likely to be low. Then, identifying a good alignment between the databases roughly corresponds to identifying the underlying (or planted) matching on the bipartite graph. When the features consist of a high number of independent dimensions, then, by a central limit theorem phenomenon, the edge weights become close to Gaussian. This phenomenon is independent from the distribution of database features themselves and only depends on them having a large number of independent dimensions. This is the justification behind our interest in planted matching on Gaussian bipartite graphs.

1.2 Outline of thesis

In Chapter 2, we introduce models, algorithms and some preliminary analysis that are useful throughout the following three chapters. We start by an explicit mathematical description of the Gaussian bigraph with planted matching and correlated Gaussian database in Section 2.1. In Section 2.2, we describe how, for both problems, the maximum likelihood estimate for the underlying mapping can be described as an instance of the assignment problem, a particular type of linear optimization program. We also consider the performance of various relaxations of the maximum likelihood estimator to understand the significance of various constraints under different settings. In Section 2.3, we perform a combinatorial analysis of misalignments, which is general and applicable to any problem where we estimate a bijective mapping between a pair of sets. Finally in Section 2.4, we present a canonical representation of the parameters of the correlated Gaussian database model, which serves to significantly simplify notation in our analysis.

In Chapter 3, we study the planted matching problem on Gaussian bipartite graphs. We start by a sketch of our analysis in Section 3.1. After stating the useful concentration inequalities in Section 3.2, we derive conditions on the parameters for the Gaussian bipartite graph that guarantee upper bounds on the log of the expected number of errors in Section

3.3. The tightness of these bounds are shown by the matching the lower bounds given in Section 3.4. A discussion and interpretation of how these bounds vary across different regimes are given in Section 3.5. Gaussian planted matching serves as a model to study and identify fundamental bounds which provide a baseline for our study of Gaussian database alignment. Both the sketch of analysis in Section 3.1 and the discussion and interpretation of results in Section 3.5 are relevant to the database alignment problem studied in the following chapter.

In Chapter 4, we study Gaussian database alignment in the setting where features are high dimensional and each dimension carries little mutual information. A good portion of the analysis in Chapter 3 carries over directly to the analysis for Chapter 4. (This is described in the sketch of analysis in Section 3.1.) We do not repeat portions of the analysis that carry over from the previous chapter and instead simply refer to the corresponding section of the analysis. Instead we dedicate a significant portion of the chapter, specifically Sections 4.2 and 4.3, to the derivation of concentration inequalities, which are essentially trivial in the Gaussian bipartite graph setting, but requires significant work in the correlated Gaussian database setting. We obtain concentration inequalities that mirror those from the previous chapter, however involve some additional terms. These terms are carefully smoothed out in Section 3.3 to obtain achievability results that match those from Chapter 3. This chapter also includes a converse result for exact alignment in Gaussian databases in Section 4.5.

In Chapter 5, we relax the condition from Chapter 4 that requires features to be high dimensional. We derive error bounds optimized for any given scale of dimensions. These bounds are stronger, i.e. they guarantee fewer errors at any given value of mutual information. For threshold testing and maximum row estimation, which are studied in Sections 5.1 and 5.2 respectively, the derivation of the stronger bounds follow from a more careful optimization of the Chernoff bounds studied in the previous chapter. For maximum likelihood estimation, the analysis is less straightforward and relies on the Lorenz ordering between

the distributions of some variables which turn out to be useful. An introduction of Lorenz orderings as well as the derivation of the orderings between different variables are given in Section 5.3, which allow us to derive the achievability result for maximum likelihood estimation in Section 5.4. The achievability results in this chapter for Gaussian database alignment with $\mathcal{O}(\log n)$ -dimensional features violate the converse results in Chapter 3 for Gaussian planted matching. While this is not a real contradiction, since these are two separate problems, this conflict between the results demonstrates that the planted matching problem, while useful, is not a perfect substitute to understand the dynamics of the more complex problem of database alignment.

Finally, in Chapter 6, we study a polynomial time canonical labeling algorithm for graph alignment under the correlated Erdős-Rényi graph model. We identify the conditions on the correlation and density parameters for the graph model that guarantee that the canonical labeling algorithm retrieves the true alignment with high probability.

The analyses and results presented in this thesis have been obtained through collaboration between Daniel Cullina, Negar Kiyavash and the author of the thesis, Osman Emre Dai.

1.3 Related work

1.3.1 Graph alignment

Graph alignment has applications in a variety of domains, including data fusion [5, 6], privacy [7, 8, 9] and in computational biology [10, 11, 12, 13, 14]. For example, in computational biology, a coarse description of the metabolic machinery of a particular species is via a protein-protein interaction (PPI) network, which essentially captures which protein can react with which other protein in that species. Across species, the PPI networks tend to be strongly correlated, because evolution transfers metabolic processes from species to species. Therefore, by identifying correspondences among proteins in different species (so-called *orthologs*), one is able to transfer biological knowledge from one species to the

other. However, crucially, the actual proteins tend to be chemically different across species, because random mutations alter these proteins over time without affecting their function. It is therefore not possible to find correspondences between proteins in different species simply by examining their amino-acid sequences. Graph alignment computes such correspondences by exploiting the correlation across networks in different species.

A similar challenge arises in social networks: suppose a set of users have accounts in several social networks. It is plausible that their links in these networks would be correlated, in the sense that given u and v are linked in the first network, it makes it conditionally more likely that they are connected in the second. This can help network reconciliation (e.g., if one wants to create a single network out of several component networks), and it can hurt privacy (e.g., by exploiting one public network to de-anonymize a private network whose node identities have been obfuscated).

A line of work has studied alignment of correlated Erdős-Rényi graphs: Pedersani and Grossglauser first identified information-theoretic achievability bounds for exact alignment (i.e. finding the true pair for every user) [15], which were then improved and matched with converse bounds by Cullina and Kiyavash. Lyzinski et al. identified bounds for exact alignment for the variant of the problem where the true alignment of number of vertices, referred to as seeds, are known in the beginning [16]. Ganassali et al. and Georgina Hall and Laurent Massoulié introduced results on partial alignment, identifying conditions that guarantee the achievability or impossibility of correctly mapping of a fraction of users with high probability [17, 18]. Several works analyze various efficient and nearly-efficient algorithms [19, 20, 21, 22, 23, 24, 25, 26]. There are other results on variants of this problem, considering alignment with side information [16, 19, 27] or alignment of graphs with Gaussian edge weights [28, 26].

1.3.2 Database alignment

For database alignment, the earliest result by Cullina and Kiyavash identified the sharp information-theoretic condition for the exact alignment of databases with finite-alphabet features [29]. A later study by Shirani et al. identified tight bounds for almost-exact alignment when features are high-dimensional and each dimension is i.i.d with an arbitrary distribution [30]. Zeynep and Nazer identified achievability and converse bounds for the related problem of testing whether Gaussian databases are correlated [31].

1.3.3 Planted matching

In the original formulation of the planted matching problem [32], it is motivated by the problem of tracking moving particles between two images. With particles in the two images forming the vertex set of the bipartite graph, each edge weight is some calculated measure of likelihood of two particles from different images correspond to each other. This can be considered as variant of the database alignment formulation where the features in each database is the information about the particle in each image. One significant difference between the two formulations is that in planted matching, edge weights are typically considered to be independent, while for database alignment, the actual likelihood measures between pairs are not truly independent. This follows from the observation that, in database alignment, the likelihood of the matched pair (u, v) and that of (u, v') both depend on the value of the feature associated with u .

Earliest work studied planted matching in the case where the non-matched edges have uniform distribution while the distribution matched edges is folded Gaussian [32, 33]. Later work studied the case where all edges are exponentially distributed, the matched ones having finite mean while the non-matched edges have mean on the order of the number of vertices in the bipartite graph [34], which was then extended to consider a variant of the problem with side information, where only a subset of the vertex pairs are eligible to be included in the matching, i.e. the matching is planted in a non-complete bipartite graph

[35]. This last study established the sharp information-theoretic conditions for almost-exact alignment in the case where the matched edge has a distribution with fixed density (i.e. not dependent on the size of the vertex set) while that of the non-matched edges is scaled (i.e. stretched) by the average degree in the bipartite graph. The Gaussian case we study in this thesis is distinct from the aforementioned planted matching work: We consider the case where all edges (whether they matched or non-matched) are Gaussian with unit variance, and the only variable for the distribution of edge weights is the distance between the means of matched and non-matched edges. In the regime of interest, the difference in means scales with the square root of log of the number of vertices. We derive sharp thresholds that dictate the order of magnitude of errors at any signal level. Furthermore, we extend our analysis to the unbalanced case where the two vertex sets differ in size and the true matching is injective but not surjective.

CHAPTER 2

PRELIMINARIES FOR GAUSSIAN PLANTED MATCHING AND GAUSSIAN DATABASE ALIGNMENT

In this chapter we introduce notation, models, algorithms and present some preliminary analyses that is used in Chapters 3, 4 and 5.

In Section 2.1 we introduce our notation and the models for the Gaussian planted matching and Gaussian database alignment problems. In Section 2.2 we present a detailed explanation of the three algorithms we consider in our analyses of these two problems. In Section 2.3, we classify different types of elementary misalignments and derive bounds on the number of possible elementary misalignments. Finally, in Section 2.4, we present the canonical form for the correlation statistics for Gaussian databases.

2.1 Models

In this section we introduce notation as well as models and problem settings that we study over multiple chapters. In particular, we introduce the correlated Gaussian database model as well as the closely related planted matching on Gaussian bigraph model, which we study in Chapters 3, 4 and 5.. The correlated Erdős-Rényi graph model, which is studied in Chapter 6 is introduced in the corresponding chapter.

2.1.1 Notation

Random variables are denoted by upper-case Latin letters while their instances are denoted with the corresponding lowercase letter. Vectors and vector-valued functions are expressed by arrows (e.g. \vec{a}) while bold font is used for matrices (e.g. \mathbf{a}). The natural numbers and real numbers are denoted \mathbb{N} and \mathbb{R} respectively while calligraphic notation is used for others sets (e.g. \mathcal{A}).

2.1.2 Planted Matching on Gaussian Bigraphs

Let \mathcal{U} and \mathcal{V} denote sets of users. Let M denote an underlying partial mapping between \mathcal{U} and \mathcal{V} : a bijection between some subset of \mathcal{U} and \mathcal{V} . We write $u \stackrel{M}{\sim} v$ if $u \in \mathcal{U}$ and $v \in \mathcal{V}$ are mapped to each other by M . Any user from one database is mapped to at most one user from the other database. We use $|M|$ to denote the number of pairs mapped by M . Let \mathbf{M} denote the matrix encoding of the mapping in $\{0, 1\}^{\mathcal{U} \times \mathcal{V}}$ such that $M_{u,v} = 1$ if u and v are mapped.

Given parameter $\mu \in \mathbb{R}$, let \mathbf{W} , taking values in $\mathbb{R}^{\mathcal{U} \times \mathcal{V}}$, denote the weight matrix of bipartite graph over \mathcal{U} and \mathcal{V} such that, given M , \mathbf{W} has independent Gaussian entries with unit variance and mean $\mathbb{E}[W_{u,v}] = \mu$ if $u \stackrel{M}{\sim} v$ and $\mathbb{E}[W_{u,v}] = 0$ otherwise. Without loss of generality, we assume $\mu > 0$.

Problem setting: We observe \mathbf{W} and want to identify the proper matching M . The parameter μ is known. We have no prior knowledge of M beyond its size.

2.1.3 Correlated Gaussian Databases

\mathcal{U} , \mathcal{V} and M are defined as in Subsection 2.1.2.

Databases are represented by functions that return feature vectors for each user in the relevant user set. $\vec{A} : \mathcal{U} \rightarrow \mathbb{R}^{\mathcal{D}_a}$ and $\vec{B} : \mathcal{V} \rightarrow \mathbb{R}^{\mathcal{D}_b}$ are the databases associated with the two sets of users. The features in each database are indexed by elements in the sets \mathcal{D}_a and \mathcal{D}_b respectively. A representation of a pair of databases is given in Fig. 2.1.

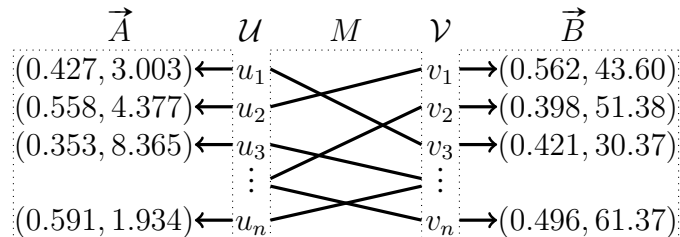


Figure 2.1: Pair of databases with 2-dimensional features and matching M between their user sets \mathcal{U} and \mathcal{V} .

We say \vec{A} and \vec{B} are a pair of correlated Gaussian databases with covariance $\Sigma = \begin{bmatrix} \Sigma_a & \Sigma_{ab} \\ \Sigma_{ab}^\top & \Sigma_b \end{bmatrix}$ if

- All entries in \vec{A} and \vec{B} are, together, jointly Gaussian.
- $\vec{A}(u)$ is independent and identically distributed with variance Σ_a for every $u \in \mathcal{U}$.
Similarly, $\vec{B}(v)$ is independent and identically distributed with variance Σ_b for every $v \in \mathcal{V}$.
- $\text{Cov}(\vec{A}(u), \vec{B}(v)) = \Sigma_{ab}$ if $u \stackrel{M}{\sim} v$ and $\text{Cov}(\vec{A}(u), \vec{B}(v)) = 0$ if $u \not\stackrel{M}{\sim} v$.

Under this model, features in \vec{A} and \vec{B} may have arbitrary dimension $|\mathcal{D}_a|$ and $|\mathcal{D}_b|$ respectively. However, as shown in Section 2.4, knowledge of the statistic Σ can be used to perform linear transformations on features from each database and eliminate degrees of freedom of the features that are not correlated with the other database. Under the canonical form, each dimension from a feature of one database is correlated with exactly one other dimension of the feature from the other database if the pair of features have correlation. Then a single correlation vector $\vec{\rho} \in (-1, 1)^{\mathcal{D}}$ indexed over some set of dimensions \mathcal{D} fully describes the distribution of the pair of databases.

Definition 1. *The number of non-zero entries in $\vec{\rho}$ is referred to as the **effective dimensionality** of the pair of correlated Gaussian databases.*

Problem setting: We consider the scenario where we observe a pair of correlated Gaussian databases \vec{A} and \vec{B} with an unknown partial mapping M between the user sets \mathcal{U} and \mathcal{V} . The statistics of the i.i.d. distribution of correlated feature pairs are known. We have no prior knowledge of the mapping M beyond its size. We say the problem is unbalanced if $|\mathcal{U}| \neq |\mathcal{V}|$.

2.1.4 Particular settings for Gaussian databases

Given a pair of correlated Gaussian databases, suppose the covariance matrix $\Sigma = \begin{bmatrix} \Sigma_a & \Sigma_{ab} \\ \Sigma_{ab}^\top & \Sigma_b \end{bmatrix}$ is such that $\Sigma_a = \Sigma_b$ are both identity and $\Sigma_{a,b}$ is a diagonal matrix with diagonal entries corresponding to the entries in some vector correlation vector $\vec{\rho} \in (-1, 1)^D$. As described later on in Section 2.4, such a representation of correlated Gaussian databases always exists.

We describe two settings of particular interest for database alignment based on this correlation vector $\vec{\rho}$.

Condition (Low per-feature correlation regime). *We say that we are in the low per-feature correlation regime if $\rho^2 \leq o(1)$ for each $i \in \mathcal{D}$.*

The low per-feature correlation regime is of interest as it presents a setting where the dynamics for database alignment become closely comparable to that of planted matching.

The low per-feature correlation regime, as well as the very low per-feature correlation regime are re-defined in Condition 4 and Condition 5 in Section 4.1 based on the original covariance matrix, without referring to the canonical correlation vector.

Condition 2. *We say that mutual information is distributed equally across all dimensions if $-\frac{1}{2} \log(1 - \rho_i^2) = \frac{I_{XY}}{|\mathcal{D}|}$ for each $i \in \mathcal{D}$.*

2.2 Algorithms for Planted Matching and Database Alignment

First we introduce the information density matrix \mathbf{G} in Subsection 2.2.1, which justifies the implementations of the algorithms that we present. Then in Subsection 2.2.3 we formulate these algorithms as linear programs with a clear hierarchy in their constraints. Finally Subsection 2.2.5 presents an analysis of the computational complexities of each algorithm.

The performances of these algorithms for planted matching and database alignment are studied under different settings in Chapters 3, 4 and 5.

2.2.1 Information density matrix for database alignment

Let \vec{A}, \vec{B} denote correlated Gaussian databases as described in Subsection 2.1.3. Let f_{XY} , f_X and f_Y denote the joint and marginal distributions for correlated features in \vec{A} and \vec{B} . Given any partial mapping m , let $\mathcal{U}_m \subseteq \mathcal{U}$ and $\mathcal{V}_m \subseteq \mathcal{V}$ denote the sets of users that have a mapping under m and $\mathcal{W}_m \subseteq \mathcal{U}_m \times \mathcal{V}_m$ denote the set of pairs mapped by m . Then the log-likelihood of observing \vec{A} and \vec{B} under the assumption that $M = m$ is given by

$$\sum_{(u,v) \in \mathcal{W}_m} \log f_{XY}(\vec{A}(u), \vec{B}(v)) + \sum_{u \in \mathcal{U} \setminus \mathcal{U}_m} \log f_X(\vec{A}(u)) + \sum_{v \in \mathcal{V} \setminus \mathcal{V}_m} \log f_Y(\vec{B}(v)). \quad (2.1)$$

Let $\mathbf{G} \in \mathbb{R}^{\mathcal{U} \times \mathcal{V}}$ denote the information density matrix such that $G_{u,v} = \log \frac{f_{XY}(\vec{A}(u), \vec{B}(v))}{f_X(\vec{A}(u))f_Y(\vec{B}(v))}$. In other words, $G_{u,v}$ is the log-likelihood ratio between hypotheses $u \stackrel{M}{\sim} v$ vs. $u \not\stackrel{M}{\sim} v$.

Let $\mathbf{m} \in \{0, 1\}^{\mathcal{U} \times \mathcal{V}}$ denote a matrix encoding of the mapping m such that $m_{u,v} = 1 \iff u \stackrel{m}{\sim} v$. Then, the inner product $\langle \mathbf{G}, \mathbf{m} \rangle$ equals $\sum_{(u,v) \in \mathcal{W}_m} \log f_{XY}(\vec{A}(u), \vec{B}(v)) - \sum_{u \in \mathcal{U}_m} \log f_X(\vec{A}(u)) - \sum_{v \in \mathcal{V}_m} \log f_Y(\vec{B}(v))$. It then follows that

$$\langle \mathbf{G}, \mathbf{m} \rangle + \sum_{u \in \mathcal{U}} \log f_X(\vec{A}(u)) + \sum_{v \in \mathcal{V}} \log f_Y(\vec{B}(v)) \quad (2.2)$$

exactly equals the expression given in (2.1). The terms following $\langle \mathbf{G}, \mathbf{m} \rangle$ in (2.2) do not depend on m . So, the choice of m that maximizes $\langle \mathbf{G}, \mathbf{m} \rangle$ is the same as the maximizer for log-likelihood, as given in (2.2). Then, \mathbf{G} contains all information relevant to identifying the underlying mapping M .

2.2.2 Log-likelihood for planted matching

The probability density function of \mathbf{W} given M under the planted matching modeled described in Subsection 2.1.2 is given by the expression

$$\frac{1}{(2\pi)^{|\mathcal{U} \times \mathcal{V}|/2}} \exp \left(-\frac{1}{2} \langle \mathbf{W} - \mu \mathbf{M}, \mathbf{W} - \mu \mathbf{M} \rangle \right).$$

Then, the log-likelihood is a constant shift away from $-\frac{1}{2} \langle \mathbf{W} - \mu \mathbf{M}, \mathbf{W} - \mu \mathbf{M} \rangle = \langle \mu \mathbf{W}, \mathbf{m} \rangle - \frac{1}{2} \|\mathbf{W}\|_F^2 - \frac{\mu^2}{2} \|\mathbf{M}\|_F^2$ where $\|\cdot\|_F$ denotes the Frobenius norm. $\|\mathbf{W}\|_F^2$ does not depend on \mathbf{M} and $\|\mathbf{M}\|_F^2$ is equal to the size of M . Then, maximizing over mappings with fixed size, \mathbf{m} that maximizes $\langle \mu \mathbf{W}, \mathbf{m} \rangle$ also maximizes the likelihood of \mathbf{W} given $M = m$.

Alternatively, we can optimize over $\langle \mathbf{W}_G, \mathbf{m} \rangle$ where $\mathbf{W}_G = \mu \mathbf{W} - \mu^2/2$ since $\langle \mathbf{W}_G, \mathbf{m} \rangle$ is a constant shift of $|\mathcal{U} \times \mathcal{V}| \mu^2/2$ away from $\langle \mu \mathbf{W}, \mathbf{m} \rangle$.

This \mathbf{W}_G matrix is of interest as it is analogous to the information matrix G for database alignment.

2.2.3 Algorithms

Maximum likelihood estimation

For database alignment, $\langle \mathbf{G}, \mathbf{m} \rangle$ is a constant factor away from the log-likelihood of mapping m , as shown in Subsection 2.2.1. For planted mtaching, given $\mathbf{W}_G \triangleq \mu \mathbf{W} - \mu^2/2$, $\langle \mathbf{W}_G, \mathbf{m} \rangle$ is a constant factor away from the log-likelihood of mapping m , as shown in Subsection 2.2.2. So, optimizing $\langle \mathbf{G}, \mathbf{m} \rangle$ or $\langle \mathbf{W}_G, \mathbf{m} \rangle$ over mapping matrices \mathbf{m} gives us the maximum likelihood estimate for the two problems.

This is an instance of the linear assignment problem, and therefore can be solved by the Hungarian algorithm in polynomial time ([36]).

Alternatively, this can be expressed as a linear problem:

$$\begin{aligned} \text{maximize } \langle \mathbf{G} - \tau, \mathbf{m} \rangle \quad & \text{(a) } \sum_{u \in \mathcal{U}} m_{u,v} \leq 1, \quad \forall v \in \mathcal{V} \\ & \text{(b) } \sum_{v \in \mathcal{V}} m_{u,v} = 1, \quad \forall u \in \mathcal{U} \\ & \text{(c) } m_{u,v} \in [0, 1], \quad \forall (u, v) \in \mathcal{U} \times \mathcal{V} \end{aligned}$$

The value of $\tau \in \mathbb{R}$ is irrelevant under constraint (b) in finding the maximizer \mathbf{m} : Given constraint (b), \mathbf{m} has fixed sum of entries, and therefore $\langle \mathbf{G} - \tau, \mathbf{m} \rangle = \langle \mathbf{G}, \mathbf{m} \rangle - \tau \sum m_{u,v} =$

$\langle \mathbf{G}, \mathbf{m} \rangle - \tau|\mathcal{U}|$, so the objective function is shifted by a constant term that depends on τ but not on \mathbf{m} .

Maximum row estimation

The objective function $\langle \mathbf{G} - \tau, \mathbf{m} \rangle$ can be broken down into its row-wise sums:

$$\langle \mathbf{G} - \tau, \mathbf{m} \rangle = \sum_u (\mathbf{G} - \tau)_{u,*}^\top \mathbf{m}_{u,*}$$

where $(\cdot)_{u,*}$ denotes the row of the matrix corresponding to user $u \in \mathcal{U}$. Then, removing (a), which is the only constraint that takes into account multiple rows at once, breaks down the optimization problem into a collection of row-wise optimization problems, where each row of \mathbf{m} can be optimized independently. That is, alignment is performed independently over each row. This gives us **maximum row estimation**.

Given any u , the algorithm looks at the log likelihood scores of mappings (u, v) for each $v \in \mathcal{V}$ and picks v that has the highest likelihood. Users in \mathcal{V} may be mapped to multiple users if they happen to be the best match for multiple users in \mathcal{U} . The mapping of each user $u \in \mathcal{U}$ under this relaxation would be the maximum likelihood estimate for u if we were blind to the existence of other users in \mathcal{U} .

Once again, the value of $\tau \in \mathbb{R}$ is irrelevant under constraint (b) in finding the maximizer \mathbf{m} .

Threshold testing

The objective function $\langle \mathbf{G} - \tau, \mathbf{m} \rangle$ can be broken down into entry-wise sums $\sum_{u,v} (G_{u,v} - \tau)m_{u,v}$. Removing conditions (a) and (b) breaks down any dependence between entries of \mathbf{m} allows us to optimize all entries in the matrix independently from each other. Then we are left with an algorithm that independently considers each pair of users and makes a decision on whether they are true pairs or not. Specifically, (u, v) is estimated to be a true match if and only if $G_{u,v} - \tau$ is positive. Since $G_{u,v}$ is defined to be the log-likelihood ratio between hypotheses $u \stackrel{M}{\sim} v$ vs. $u \stackrel{\mathcal{L}}{\not\sim} v$, the decision rule $G_{u,v} - \tau \geq 0$ is equivalent to the

likelihood-ratio test at some significance level determined by the log threshold τ . We refer to this relaxation as **threshold testing**.

2.2.4 Performance metrics for algorithms

We evaluate the performance of the algorithms on a given instance of the problem based on the metric $d(\hat{M}, M) \triangleq \frac{1}{2} \|\hat{M} - M\|_F^2$ that defines a distance between the true underlying mapping M and the estimated mapping \hat{M} . For \hat{M}_{ML} and \hat{M}_{MR} this distance gives the number of users in \mathcal{U} mapped to a false match in \mathcal{V} . For \hat{M}_{MR} , the distance gives the arithmetic mean of the number of false positives and false negatives.

Lemma 1. *Given any true mapping M , any threshold τ and any value of \mathbf{G} or \mathbf{W}_G , $d(\hat{M}_{MR}, M) \leq 2d(\hat{M}_{TT}, M)$.*

Proof. It suffices to show that for every error under maximum row estimation, there is at least one error (i.e. a false positive or a false negative) under threshold testing.

Let $(u, v) \in \mathcal{U} \times \mathcal{V}$ such that $u \stackrel{M}{\sim} v$. u is misaligned under maximum row estimation if and only if there exists some $v' \in \mathcal{V}$ such that $G_{u,v'} > G_{u,v}$.

Consider the two cases for the value of τ : If $\tau \geq G_{u,v'}$, then $G_{u,v} < \tau$, i.e. (u, v) fails the threshold test so we have a false negative for (u, v) . On the other hand if $G_{u,v'} > \tau$, then (u, v') beats the threshold so we have a false positive at (u, v') . \square

Remark 1. *There is no direct comparison between $d(\hat{M}_{ML}, M)$ and $d(\hat{M}_{MR}, M)$ or $d(\hat{M}_{ML}, M)$ and $d(\hat{M}_{TT}, M)$ analogous to the statement in Lemma 1. In fact, $d(\hat{M}_{ML}, M)$ can be arbitrarily greater than $d(\hat{M}_{MR}, M)$ for marginal values of \mathbf{G} or \mathbf{W}_G . To show this, we describe an example that holds for any threshold τ where $d(\hat{M}_{ML}, M) = n$ and $d(\hat{M}_{MR}, M) = 2d(\hat{M}_{TT}, M) = 1$.*

Without loss of generality, let $\mathcal{U} = \{u_1, u_2, \dots, u_n\}$, $\mathcal{V} = \{v_1, v_2, \dots, v_n\}$ and $M = m$ such that $u_i \stackrel{m}{\sim} v_i$ for each $i \in \{1, 2, \dots, n\}$. Compare this to m' such that $u_i \stackrel{m'}{\sim} v_{i+1}$ for each $i \in \{1, 2, \dots, n-1\}$ and $u_n \stackrel{m'}{\sim} v_1$, which corresponds to a misalignment of size n .

Next we construct \mathbf{G} . Pick some $\varepsilon > 0$. Let $G_{u_i, v_i} = \tau + \varepsilon$ for each $i \in \{1, \dots, n\}$ and let $G_{u_i, v_{i+1}} = \tau - \varepsilon$ for each $i \in \{1, \dots, n-1\}$. Let $G_{u_n, v_1} = \tau + 2n\varepsilon$ and $G_{u, v} = \tau - 2n\varepsilon$ for all other $(u, v) \in \mathcal{U} \times \mathcal{V}$.

Then

- $\langle \mathbf{G}, \mathbf{m} \rangle = n\tau + n\varepsilon$.

The sum is made up of n entries equal to $\tau + \varepsilon$.

- $\langle \mathbf{G}, \mathbf{m}' \rangle = n\tau + (n+1)\varepsilon$.

The sum is made up of $(n-1)$ entries equal to $\tau - \varepsilon$ and a single entry equal to $\tau + 2n\varepsilon$.

- For any other m'' , $\langle \mathbf{G}, \mathbf{m}'' \rangle \leq n\tau + (n-1)\varepsilon$.

This follows from the fact that if a misalignment does not coincide exactly with m' , then it must contain a misaligned pair that is not covered by m' . It then follows that the sum contains some entry equal to $\tau - 2n\varepsilon$. Even if the remaining entries in the sum were the $(n-1)$ largest entries in the \mathbf{G} (i.e. a single $\tau + 2n\varepsilon$ and $(n-2)$ copies of $\tau + \varepsilon$), the sum would not be greater than $n\tau + (n-2)\varepsilon$.

It then follows that m' is the maximum likelihood alignment, which corresponds to an error size of n .

Under maximum likelihood alignment, G_{u_i, v_i} is the largest entry on every row except the row u_n where $G_{u_n, v_1} = \tau + 2n\varepsilon$ beats $G_{u_n, v_n} = \tau + \varepsilon$. Then, there is one misaligned user under maximum row estimation.

Finally, under threshold testing, there are $(n+1)$ entries in \mathbf{G} that beat the threshold: n of these correspond to the true pairs (u_i, v_i) and the last one is G_{u_n, v_1} . Then there is a single false positive and zero false negatives under threshold testing.

2.2.5 Computational complexity

Let $n_M = |M|$ the size of the true mapping, and $n_{\mathcal{U}}, n_{\mathcal{V}}$ denote the sizes of the two user sets \mathcal{U}, \mathcal{V} . Furthermore, in this subsection, let d denote the maximum of the number of features

per user in the two databases. (Not to be confused with the distance metric $d(\hat{M}, M)$, which is not referred to in this subsection.)

Summary

Let $d \leq \mathcal{O}(n_M)$. The computational complexity of the algorithms are given as follows:

- Maximum likelihood estimation: $\mathcal{O}(n_M \cdot n_U \cdot n_V)$ for entire set of users $\mathcal{U} \times \mathcal{V}$
- Maximum row estimation: $\mathcal{O}(d \cdot n_U \cdot n_V)$ for entire set of users $\mathcal{U} \times \mathcal{V}$,
 $\mathcal{O}(d^2 \cdot n_V)$ for a given user $u \in \mathcal{U}$ against entire set \mathcal{V} .
- Threshold testing: $\mathcal{O}(d \cdot n_U \cdot n_V)$ for entire set of users $\mathcal{U} \times \mathcal{V}$,
 $\mathcal{O}(d^2)$ for a given pair of users $(u, v) \in \mathcal{U} \times \mathcal{V}$.

Computing G through the canonical form

Identifying the affine transformations described in Section 2.4 to transform features into canonical form takes a sequence of two Cholesky decompositions ($\Sigma_a = \mathbf{L}_a^\top \mathbf{L}_a$, $\Sigma_b = \mathbf{L}_b^\top \mathbf{L}_b$), two matrix multiplications with inverted triangular matrices ($\mathbf{L}_a^{-1} \Sigma_{ab} (\mathbf{L}_b^\top)^{-1}$), one singular value decomposition ($\mathbf{L}_a^{-1} \Sigma_{ab} (\mathbf{L}_b^\top)^{-1} = \mathbf{U} \mathbf{P} \mathbf{V}^\top$) and two more matrix multiplications with inverted triangular matrices ($\mathbf{U}^\top \mathbf{L}_a^{-1}$ and $\mathbf{V}^\top \mathbf{L}_b^{-1}$). This can be done in $\mathcal{O}(d^3)$ -time.

Performing the affine transformation to transform features into canonical form as described in consists of one vector addition and one matrix-vector multiplication ($(\mathbf{U}^\top \mathbf{L}_a^{-1})(\vec{x} - \vec{\mu}_a)$ or $(\mathbf{V}^\top \mathbf{L}_b^{-1})(\vec{y} - \vec{\mu}_b)$). Then, transforming a single feature vector takes $\mathcal{O}(d^2)$ -time and transforming the entire database takes $\mathcal{O}(d^2 \cdot n_U)$ and $\mathcal{O}(d^2 \cdot n_V)$ for \mathcal{U} and \mathcal{V} respectively.

Given databases in canonical form, a single entry in \mathbf{G} can be computed in $\mathcal{O}(d)$. This follows from the fact that, in canonical form, there is a one-to-one correspondence between feature entries from the two databases and therefore

$$\log \frac{f_{XY}(\vec{A}(u), \vec{B}(v))}{f_X(\vec{A}(u)) f_Y(\vec{B}(v))} = \sum_{i \in \mathcal{D}} \log \frac{f_{X_i Y_i}(A_i(u), B_i(v))}{f_{X_i}(A_i(u)) f_{Y_i}(B_i(v))}.$$

Then it takes $\mathcal{O}(d \cdot n_{\mathcal{U}} \cdot n_{\mathcal{V}})$ to compute the entire matrix \mathbf{G} based on features in canonical form.

Therefore, computing values of \mathbf{G} from the databases has complexity

- $\mathcal{O}(d^3 + d^2 \cdot (n_{\mathcal{U}} + n_{\mathcal{V}}) + d \cdot n_{\mathcal{U}} \cdot n_{\mathcal{V}})$ for the entire matrix \mathbf{G} ,
- $\mathcal{O}(d^3 + d^2 \cdot n_{\mathcal{V}})$ for a row of \mathbf{G} and
- $\mathcal{O}(d^3)$ for a single entry of \mathbf{G} .

Computing \mathbf{G} without going through the canonical form

Given \mathbf{G} in raw form (i.e. not necessarily canonical form), finding the likelihood requires calculating $\det(\boldsymbol{\Sigma})$, $\det(\boldsymbol{\Sigma}_a)$ and $\det(\boldsymbol{\Sigma}_b)$, which takes $\mathcal{O}(d^3)$ -time, as well as $[\vec{x}^\top, \vec{y}^\top] \boldsymbol{\Sigma}^{-1} [\vec{x}^\top, \vec{y}^\top]^\top$, $\vec{x}^\top \boldsymbol{\Sigma}_a^{-1} \vec{x}$ and $\vec{y}^\top \boldsymbol{\Sigma}_b^{-1} \vec{y}$ for each feature pair, which takes $\mathcal{O}(d^2)$ time per feature pair. Then it takes $\mathcal{O}(d^3 + d^2 \cdot n_{\mathcal{U}} \cdot n_{\mathcal{V}})$ to compute the entire matrix \mathbf{G} based on features in raw form. This is less efficient than doing the calculation through the canonical form which takes $\mathcal{O}(d^3 + d^2 \cdot (n_{\mathcal{U}} + n_{\mathcal{V}}))$ to obtain features in canonical form and $\mathcal{O}(d \cdot n_{\mathcal{U}} \cdot n_{\mathcal{V}})$ to get \mathbf{G} based on features in canonical form.

Maximum likelihood estimation

The (unbalanced) linear assignment problem which can be solved by the Hungarian algorithm in $\mathcal{O}(n_M \cdot n_{\mathcal{U}} \cdot n_{\mathcal{V}})$ [36]. Then, the total complexity of maximum likelihood estimation for database alignment, including the computation of \mathbf{G} , is $\mathcal{O}(d^3 + d^2 \cdot (n_{\mathcal{U}} + n_{\mathcal{V}}) + d \cdot n_{\mathcal{U}} \cdot n_{\mathcal{V}} + n_M \cdot n_{\mathcal{U}} \cdot n_{\mathcal{V}})$. The complexity for planted matching is $\mathcal{O}(n_M \cdot n_{\mathcal{U}} \cdot n_{\mathcal{V}})$.

Maximum row estimation

For database alignment, given the corresponding row of \mathbf{G} , identifying the match of a user in \mathcal{U} takes $\mathcal{O}(n_{\mathcal{V}})$ -time. Then the total complexity to align a single user $u \in \mathcal{U}$, including the complexity of calculating row $(\mathbf{G})_{u,*}$, is $\mathcal{O}(d^3 + d^2 n_{\mathcal{V}})$. Consequently, aligning the entire set \mathcal{U} takes $\mathcal{O}(d^3 + d^2 \cdot (n_{\mathcal{U}} + n_{\mathcal{V}}) + d \cdot n_{\mathcal{U}} \cdot n_{\mathcal{V}})$ -time.

For planted matching, identifying the match of a user in \mathcal{U} takes $\mathcal{O}(n_{\mathcal{V}})$ -time, while aligning the entire set takes $\mathcal{O}(n_{\mathcal{U}} \cdot n_{\mathcal{V}})$ -time.

Threshold testing algorithm

For database alignment, given the corresponding entry in \mathbf{G} , performing threshold testing over a pair of users takes $\mathcal{O}(1)$ -time. Then the total complexity to test a single user pair (u, v) , including the complexity of calculating $G_{u,v}$, is $\mathcal{O}(d^3)$. Then performing the test over all pairs would take $\mathcal{O}(d^3 + d^2 \cdot (n_U + n_V) + d \cdot n_U \cdot n_V)$ -time.

For planted matching, performing threshold testing over a pair of users takes $\mathcal{O}(1)$ -time, while performing the test over all pairs would take $\mathcal{O}(n_U \cdot n_V)$ -time.

2.3 Combinatorial Analysis of Misalignment

In this section we study present a combinatorial alignment of misalignments. Specifically, in Subsection 2.3.1 we define 3 types of elementary misalignments that all misalignments can be decomposed into. Then, in Subsection 2.3.2, we present bounds on the number of such elementary misalignments. These bounds are used in Chapters 3, 4 and 5.

2.3.1 Elementary misalignments between mappings

Let $\mathbf{m}, \mathbf{m}' \in \{0, 1\}^{U \times V}$ be the matrix encodings of mappings m, m' and let $\mathbf{G} \in \mathbb{R}^{U \times V}$ denote the score matrix for databases \vec{A}, \vec{B} . As shown in Section 2.2, comparing the likelihoods of \vec{A}, \vec{B} being generated by m versus m' is equivalent to comparing the values of $\langle \mathbf{G}, \mathbf{m} \rangle$ and $\langle \mathbf{G}, \mathbf{m}' \rangle$.

Assume $\mathbf{m} - \mathbf{m}'$ can be written in block diagonal form $\begin{bmatrix} \Delta \mathbf{m}_1 & \mathbf{0} \\ \mathbf{0} & \Delta \mathbf{m}_2 \end{bmatrix}$. Let $\mathbf{m}'_1 \triangleq \mathbf{m} - \begin{bmatrix} \Delta \mathbf{m}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}$ and $\mathbf{m}'_2 \triangleq \mathbf{m} - \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \Delta \mathbf{m}_2 \end{bmatrix}$. m'_1, m'_2 are two valid mappings that in

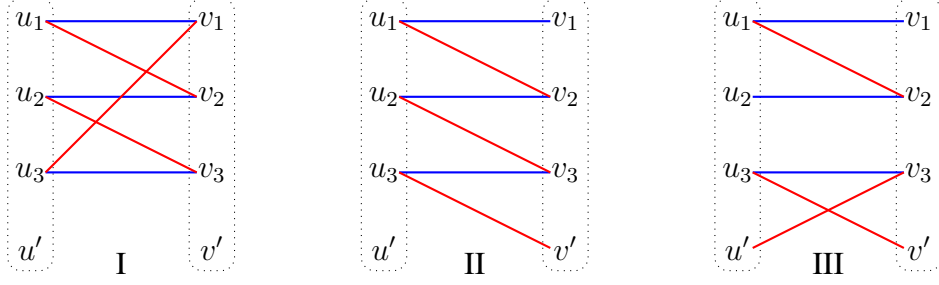


Figure 2.2: Examples of 3 types of elementary misalignments (of size 3) as bigraphs: Cycle (I), even path (II), pair of odd paths (III)

some sense partition the disagreement between m and m' into two.

$$\begin{aligned}
\langle \mathbf{G}, \mathbf{m} \rangle < \langle \mathbf{G}, \mathbf{m}' \rangle &\iff \langle \mathbf{G}, \mathbf{m} - \mathbf{m}' \rangle < 0 \\
\iff \left\langle \mathbf{G}, \begin{bmatrix} \Delta \mathbf{m}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \right\rangle + \left\langle \mathbf{G}, \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \Delta \mathbf{m}_2 \end{bmatrix} \right\rangle < 0 \\
&\iff \langle \mathbf{G}, \mathbf{m} - \mathbf{m}'_2 \rangle + \langle \mathbf{G}, \mathbf{m} - \mathbf{m}'_1 \rangle < 0 \\
&\implies \langle \mathbf{G}, \mathbf{m} - \mathbf{m}'_2 \rangle \leq 0 \text{ or } \langle \mathbf{G}, \mathbf{m} - \mathbf{m}'_1 \rangle \leq 0
\end{aligned}$$

Then m' has higher score than m only if at least one of m'_1, m'_2 also has higher score than m . Furthermore, m' is the minimizer for the inner product $\langle \mathbf{G}, \mathbf{m} - \mathbf{m}' \rangle$ only if both $\langle \mathbf{G}, \mathbf{m} - \mathbf{m}'_1 \rangle$ and $\langle \mathbf{G}, \mathbf{m} - \mathbf{m}'_2 \rangle$ are negative. So m' is the optimal mapping only if each of its ‘submappings’ (i.e. mappings whose mismatch with m are entirely contained in m') have higher score than m .

It is then of interest to define elementary misalignments between mappings.

Definition 3. Let m_1, m_2 be a pair of mappings between \mathcal{U} and \mathcal{V} that are bijective between from their domain to their co-domain. Let $\mathbf{m}_1, \mathbf{m}_2 \in \{0, 1\}^{\mathcal{U} \times \mathcal{V}}$ be binary matrices that encode these mappings. We say the mismatch between the two mappings is elementary if and only if $\mathbf{m}_1 - \mathbf{m}_2$ does not have a block-diagonal representation with multiple zero-sum, non-zero blocks.

$$\begin{array}{ccc}
\begin{bmatrix} + & - & 0 & 0 \\ 0 & + & - & 0 \\ - & 0 & + & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} &
\begin{bmatrix} + & - & 0 & 0 \\ 0 & + & - & 0 \\ 0 & 0 & + & - \\ 0 & 0 & 0 & 0 \end{bmatrix} &
\begin{bmatrix} + & - & 0 & 0 \\ 0 & + & 0 & 0 \\ 0 & 0 & + & - \\ 0 & 0 & - & 0 \end{bmatrix} \\
\text{I} & \text{II} & \text{III}
\end{array}$$

Figure 2.3: Examples of 3 types of elementary misalignments (of size 3) as matrices: Cycle (I), even path (II), pair of odd paths (III)

There are three types of elementary misalignments, as shown in Fig. 2.2 and Fig. 2.3.

These are

- I - Cycles: The two mappings use the same set of users from \mathcal{U} and \mathcal{V} but pair them up differently. This type of mismatch consists of a single cycle.
- II - Even paths: The two mappings use the same set of users from one of the sets (say \mathcal{U}) but differ in the users they map from the other side (\mathcal{V}) by 1 user. This type of mismatch consists of one path and is cycle-free.
- III - Pair of odd paths: The two mappings differ in the users they map on both sides, by 1 user per side. This type of mismatch consists of two paths and is cycle-free.

The bigraph representation in Fig. 2.2 can be used to explain why these three are the only types of elementary misalignments. Since m_1 and m_2 map each user at most once, each vertex can have at most one edge from each mapping and a degree of at most 2. Then the bigraph has alternating edges and maximum degree 2. Graphs of maximum degree 2 decompose into cycles and paths. Each component in the bigraph corresponds to a block in the adjacency matrix.

Since edges are alternating between the two mappings, each cycle has even length and an equal number of edges coming from both graphs. Then each cycle corresponds to a block in $\mathbf{m}_1 - \mathbf{m}_2$ with sum of entries equal to 0. Therefore each cycle is an elementary misalignment. The same holds for even paths.

Odd paths contain one more edge from one mapping than from the other. Therefore these correspond to blocks in $\mathbf{m}_1 - \mathbf{m}_2$ whose sum equal 1 or -1 . Since $\mathbf{m}_1 - \mathbf{m}_2$ has

sum of entries equal to zero, it follows that there must be an equal number of blocks whose entries sum up to 1 and blocks whose entries sum up to -1 . Pairing these up gives us elementary blocks.

2.3.2 Counting elementary misalignments

Lemma 2. *Let \mathcal{U} and \mathcal{V} sets of users of size n and $n + s$ respectively. Let m be the true mapping of size n . Consider the elementary misalignments induced by all mappings m' of size n .*

The number of distinct elementary type-I misalignments of size δ is upper bounded by $\frac{n^\delta}{\delta}$ if $\delta \in \{2, 3, \dots, n\}$ and for other δ . The number of distinct elementary type-II misalignments of size δ is upper bounded by sn^δ . There are no elementary misalignments of type III.

Proof. We count the ways to pick some m' that induces an elementary misalignment with m of size δ .

There are $\binom{n}{\delta}$ ways to pick the δ pairs from m to be misaligned by m' . Let us denote the sets of these users as $\mathcal{U}' \subseteq \mathcal{U}$ and $\mathcal{V}' \subseteq \mathcal{V}$. $|\mathcal{U}'| = |\mathcal{V}'| = \delta$. Assume these sets are fixed.

- If $\delta = 1$, there is no way to obtain a type I mismatch, since the only way to pair the single user in \mathcal{U}' to the single user in \mathcal{V}' is the same as the original mapping in m .

For $\delta \in \{2, 3, \dots\}$, there are $(\delta - 1)!$ ways to pair \mathcal{U}' and \mathcal{V}' to obtain a type I mismatch. (Forming the ‘cycle’ in Fig. 2.2 is simply a matter of arranging the blue edges around the cycle, which results in a unique way to pick the red edges.)

Then, in total, there are $\binom{n}{\delta}(\delta - 1)! = \frac{1}{\delta} \binom{n}{\delta} \delta!$ ways to pick a type I mismatch.

- There are s ways to pick a user w' from \mathcal{V} that is not mapped by m .

Given this user, there are $\delta!$ ways to pair \mathcal{U}' , \mathcal{V}' and w' , leaving one user from either \mathcal{U}' unpaired. We can generate this pairing as follows: Take any of the $(\delta - 1)!$ type I matchings. Break the cycle at any of the δ red edges and connect that edge to w' from the appropriate side. This gives us an even path.

Then, in total, there are $s \binom{n}{\delta} \delta!$ ways to pick a type II mismatch.

- We only consider mappings m' of size n , which is the same as the true mapping m . So, given the representation in Fig. 2.2, there must be an equal number of red and blue edges.

Odd paths have an extra edge of either color. To construct an odd path with more edges belonging to m' , there need to vertices in both \mathcal{U} and \mathcal{V} not covered by m . (These correspond to vertices u' and v' in Fig. 2.2.) Since \mathcal{U} has size equal to that of m , all vertices in \mathcal{U} are covered, and there can be no odd path with more edges from m' .

Since the total number of edges from each mapping needs to be equal, it then follows that there can also be no odd path with more edges from m .

Using the fact that $\binom{n}{\delta} \delta! \leq n^\delta$, we simplify the expression to get the result. \square

Lemma 3. *Let \mathcal{U} and \mathcal{V} sets of users of size n and $n + s$ respectively. Let m be the true mapping of size n . Let $c \in (0, \infty)$ some arbitrary constant. The number of different mappings m' that result in a misalignment of size δ is upper bounded by:*

- $\exp(\delta(1 + \log n + \log(1 + 1/c)))$ if $\delta \geq cs$, and
- $\exp(\delta(1 + \log \frac{ns}{\delta} + \log(1 + c)))$ if $\delta \leq cs$.

Proof. We count the number of different ways to construct m' that results in a misalignment of size δ .

There are $\binom{n}{\delta}$ different ways to pick the set of vertices to be misaligned by m' .

Given the δ pairs of vertices to be misaligned, there are no more than $\delta + s$ ways to misalign each vertex. So, there are no more than $(\delta + s)^\delta$ ways to misalign the set of δ pairs.

$\binom{n}{\delta}$ is strictly less than $(\frac{en}{\delta})^\delta$. If $\delta \geq cs$, then $(\delta + s)^\delta$ is at most $(1 + 1/c)^\delta \delta^\delta$. If $\delta \leq cs$, then $(\delta + s)^\delta$ is at most $(1 + c)^\delta s^\delta$. The products of these terms give us the claimed results. \square

2.4 Canonical Form of Correlation Statistics for Correlated Gaussian Databases

For simplicity of computation as well as analysis, we would like the correlated feature indices in $\vec{A}(u) \in \mathbb{R}^{\mathcal{D}_a}$ and $\vec{B}(v) \in \mathbb{R}^{\mathcal{D}_b}$ to have a one-to-one correspondance. Specifically, we would like the index sets \mathcal{D}_a and \mathcal{D}_b to be identical, and for features across databases to be correlated only if they have the same index. So, given true pair $u \stackrel{M}{\sim} v$, the features $A_i(u)$ and $B_j(v)$ are correlated if and only if $i = j$.

Given the correlation statistics $\vec{\mu}$ and Σ , it is possible to perform affine transformations on features to guarantee this type of correspondance between correlated feature vectors. We say a pair of databases with statistics of this desired form is in canonical form.

The generality of the canonical form is stated in the following lemma while the construction of the transformation that gives features in canonical form is described in the proof of the lemma. The lemma and the proof only consider a pair of correlated features. This directly generalizes to the entire database since pairs of correlated features are independent and identically distributed.

Lemma 4 (Existence of the canonical form). *Let \vec{X} taking values in $\mathbb{R}^{\mathcal{D}_a}$ and \vec{Y} taking values in $\mathbb{R}^{\mathcal{D}_b}$ be a pair of correlated Gaussian features. If the mean and joint variance is known, one can define a pair of affine transformations $\vec{t}_a : \mathbb{R}^{\mathcal{D}_a} \rightarrow \mathbb{R}^{\mathcal{D}}$ and $\vec{t}_b : \mathbb{R}^{\mathcal{D}_b} \rightarrow \mathbb{R}^{\mathcal{D}}$ for some set \mathcal{D} such that the mutual information between $\vec{t}_a(\vec{X})$ and $\vec{t}_b(\vec{Y})$ equals the mutual information between \vec{X} and \vec{Y} and $\vec{t}_a(\vec{X}), \vec{t}_b(\vec{Y})$ taking values in $\mathbb{R}^{\mathcal{D}}$ are a pair of correlated features with mean $\vec{0}$ and joint variance $\begin{bmatrix} \mathbf{I} & \text{diag}(\vec{\rho}) \\ \text{diag}(\vec{\rho}) & \mathbf{I} \end{bmatrix}$ for some correlation vector $\vec{\rho} \in (-1, 1)^{\mathcal{D}}$.*

Proof. Let $\vec{\mu} = \begin{bmatrix} \vec{\mu}_a \\ \vec{\mu}_b \end{bmatrix}$ and $\Sigma = \begin{bmatrix} \Sigma_a & \Sigma_{ab} \\ \Sigma_{ab}^\top & \Sigma_b \end{bmatrix}$ denote the mean and variance of $\begin{bmatrix} \vec{X} \\ \vec{Y} \end{bmatrix}$.

If Σ_a is not full rank, then there is some subset of \mathcal{D}_a that can be discarded without loss of information. That is, we can throw away some indices of \vec{X} to get a shorter vector which allows us to reconstruct the original vector \vec{X} . This follows from the fact that a multivariate

Gaussian vector with covariance Σ can be written as a linear combination of $\text{rank}(\Sigma)$ i.i.d. Gaussian normal random variables. Then, without loss of generality, assume Σ_a and Σ_b are full rank.

Σ_a and Σ_b are covariance matrices, therefore they are positive semi-definite. It then follows that these matrices have Cholesky decompositions: $\Sigma_a = \mathbf{L}_a \mathbf{L}_a^\top$ and $\Sigma_b = \mathbf{L}_b \mathbf{L}_b^\top$ where \mathbf{L}_a and \mathbf{L}_b are lower triangular matrices with non-negative diagonal entries. By the assumption that Σ_a and Σ_b are full rank, it follows that the Cholesky decomposition gives triangular matrix with strictly positive entries. Then \mathbf{L}_a and \mathbf{L}_b are invertible.

Let $d_a \triangleq |\mathcal{D}_a|$ and $d_b \triangleq |\mathcal{D}_b|$. Consider the singular value decomposition of $\mathbf{L}_a^{-1} \Sigma_{ab} (\mathbf{L}_b^\top)^{-1}$: $\mathbf{U} \in \mathbb{R}^{d_a \times \{1,2,\dots,d_a\}}$ and $\mathbf{V} \in \mathbb{R}^{d_b \times \{1,2,\dots,d_b\}}$ orthonormal matrices and $\mathbf{P} \in \mathbb{R}^{d_a \times d_b}$ a diagonal matrix such that $\mathbf{U} \mathbf{P} \mathbf{V}^\top = \mathbf{L}_a^{-1} \Sigma_{ab} (\mathbf{L}_b^\top)^{-1}$.

Let $\vec{f}_a : \mathbb{R}^{d_a} \rightarrow \mathbb{R}^{d_a}$ and $\vec{f}_b : \mathbb{R}^{d_b} \rightarrow \mathbb{R}^{d_b}$ such that

$$\begin{aligned}\vec{f}_a(\vec{x}) &= \mathbf{U}^\top \mathbf{L}_a^{-1} (\vec{x} - \vec{\mu}_a) \\ \vec{f}_b(\vec{y}) &= \mathbf{V}^\top \mathbf{L}_b^{-1} (\vec{y} - \vec{\mu}_b).\end{aligned}$$

Note that both these transformations are invertible.

We can verify that

$$\begin{bmatrix} \mathbf{U}^\top \mathbf{L}_a^{-1} \\ \mathbf{V}^\top \mathbf{L}_b^{-1} \end{bmatrix} \begin{bmatrix} \Sigma_a & \Sigma_{ab} \\ \Sigma_{ab}^\top & \Sigma_b \end{bmatrix} \begin{bmatrix} \mathbf{U}^\top \mathbf{L}_a^{-1} \\ \mathbf{V}^\top \mathbf{L}_b^{-1} \end{bmatrix}^\top = \begin{bmatrix} \mathbf{I} & \mathbf{P} \\ \mathbf{P}^\top & \mathbf{I} \end{bmatrix}.$$

Then

$$(\vec{X}, \vec{Y}) \sim \mathcal{N}(\vec{\mu}, \Sigma) \iff (\vec{f}_a(\vec{X}), \vec{f}_b(\vec{Y})) \sim \mathcal{N}\left(\vec{0}, \begin{bmatrix} \mathbf{I} & \mathbf{P} \\ \mathbf{P}^\top & \mathbf{I} \end{bmatrix}\right).$$

By the invertibility of the transformations, there has been no loss of mutual information.

If \mathbf{P} has no empty rows or columns, then $d_a = d_b$ must hold and we are done.

If the i -th row of \mathbf{P} is all-zero, then the i -th entry of $\vec{f}_a(\vec{X})$ is completely independent from \vec{Y} or $\vec{f}_b(\vec{Y})$. It then follows that we can drop this entry without any loss of mutual information. The same argument applies for columns of \mathbf{P} in relation to entries of $\vec{f}_b(\vec{Y})$.

Let d denote the number of non-zero entries in the diagonal matrix \mathbf{P} , \mathcal{D} be some arbitrary set of size d . Let $\mathbf{E}_a \in \{0, 1\}^{\mathcal{D} \times \{1, 2, \dots, d_a\}}$ such that rows of \mathbf{E}_a are the d standard basis vectors corresponding to the d non-empty rows of \mathbf{P} . Left multiplying a vector by \mathbf{E}_a gives us a shorter vector by ‘throwing away’ all entries corresponding to empty rows of \mathbf{P} . Let $\mathbf{E}_b \in \{0, 1\}^{\mathcal{D} \times \{1, 2, \dots, d_b\}}$ be a matrix of the same kind that ‘throws away’ entries corresponding to empty columns of \mathbf{P} . Then $\mathbf{E}_a \mathbf{P} \mathbf{E}_b^\top$ is a diagonal matrix with no zeros on the diagonal. We use $\vec{\rho} \in \mathbb{R}^{\mathcal{D}}$ to denote the vector formed by the diagonal entries of $\mathbf{E}_a \mathbf{P} \mathbf{E}_b^\top$ (i.e. the non-zero diagonal entries of \mathbf{P} .)

Let $\vec{t}_a : \mathbb{R}^{d_a} \rightarrow \mathbb{R}^{\mathcal{D}}$ and $\vec{t}_b : \mathbb{R}^{d_b} \rightarrow \mathbb{R}^{\mathcal{D}}$ such that

$$\begin{aligned}\vec{t}'_a(\vec{x}) &= \mathbf{E}_a \mathbf{U}^\top \mathbf{L}_a^{-1} (\vec{x} - \vec{\mu}_a) \\ \vec{t}'_b(\vec{y}) &= \mathbf{E}_b \mathbf{V}^\top \mathbf{L}_b^{-1} (\vec{y} - \vec{\mu}_b).\end{aligned}$$

It can be verified that

$$(\vec{X}, \vec{Y}) \sim \mathcal{N}(\vec{\mu}, \Sigma) \iff (\vec{t}_a(\vec{X}), \vec{t}_b(\vec{Y})) \sim \mathcal{N}\left(\vec{0}, \begin{bmatrix} \mathbf{I} & \text{diag}(\vec{\rho}) \\ \text{diag}(\vec{\rho}) & \mathbf{I} \end{bmatrix}\right).$$

□

CHAPTER 3

GAUSSIAN PLANTED MATCHING

We study the Gaussian planted matching problem, which corresponds to identifying the planted matching in a Gaussian bigraph, as described in Subsection 2.1.2. We analyze the performance of the three algorithms presented in Section 2.2.

The Gaussian planted matching problem is closely related to database alignment. Database alignment with any independent and identical feature involves computing the log-likelihood score of all potential matches which are to form the estimated mapping. This can be expressed as a weighted bipartite graph where edge weights corresponding to log-likelihood scores of corresponding node pairs. If the features of the databases are high dimensional and have little dependence between dimensions, then the weights of the edges of the bigraph are nearly independent and Gaussian, and the distribution can be closely approximated by that of a bigraph with a planted matching. This does not depend on the Gaussian feature assumption: it is a central limit theorem phenomenon that will be common to any database model with enough almost independent features. Therefore studying the Gaussian planted matching problem facilitates detailed analysis of the database alignment problem and allows us to identify qualitative phenomena that may apply more generally. The connection between Gaussian planted matching and database alignment with high dimensional independent features can be seen through the parallels between the results in this chapter and Chapter 4, which builds on the analyses presented in this chapter.

In Section 3.1, we present a sketch of our analysis and its connections to the analysis in Chapter 4. In Section 3.2 we derive the concentration inequalities to be used in the proofs of the main results. In Section 3.3, we present asymptotic and finite-size achievability bounds. In Section 3.4, we present asymptotic converse bounds which match the achievability results in the previous section. Finally, in Section 3.5, we present a discussion of

phase transitions in the achievability and matching converse bounds, providing intuition and sketches of technical arguments that explain these transitions.

3.1 Sketch of analysis

Our analyses in Chapters 3 and 4 for each algorithm involves deriving Chernoff bounds for the probability of each error event, then taking the union bound over all of these. The bounds are immediate for Gaussian planted matching (Section 3.2), but for database alignment significant work is required to get useful versions of the bounds: We carefully study the generating function \mathbf{G} in Section 4.2 before we can derive the concentration inequalities given in Section 4.3.

The union bound overestimates the total probability of error: E.g. for maximum row estimation $\{G_{u,v'} \geq G_{u,v}\}$ and $\{G_{u,v''} \geq G_{u,v}\}$ can both occur, yet we will only make one error on the row corresponding to user u . For high correlation, such intersecting error events occurring simultaneously is negligibly unlikely and the union bound gives a good bound. This gives us the linear boundaries in Fig. 3.3.

For the low correlation case, we come up with a way to break up error events into two sub-events: one that is strongly correlated with all error events involving that true pair, and one that is not. E.g. for maximum row estimation, we use $\{G_{u,v} \leq \tau\}$ (atypicality) and $\{G_{u,v'} \geq G_{u,v}\} \cap \{G_{u,v} > \tau\}$ (misalignment-despite-typicality) to cover the error event $\{G_{u,v'} \geq G_{u,v}\}$. The atypicality event has higher likelihood than the original error event, however it only needs to be counted once for each error involving $G_{u,v}$. The misalignment-despite-atypicality event needs to be counted separately for every possible misalignment on a given row, however the correlation between misalignment-despite-atypicality events are much weaker than those of the original error events, allowing us to get a stronger union bound.

Events of the type $\left\{ \sum_{u,v} c_{u,v} G_{u,v} \geq d \right\}$ can be represented as half-spaces in the $|\mathcal{U} \times \mathcal{V}|$ dimensional event space where each dimension corresponds to a different entry of the

matrix \mathbf{G} . The covering of the half-space corresponding to the original error event by two auxiliary half-spaces is illustrated in figures 3.1. A more detailed description of the geometric intuition behind these probability bounds are given in Subsection 4.3.2.

3.2 Concentration Inequalities

\mathbf{W} refers to the edge weight matrix of the bipartite graph under the planted matching setting described in Subsection 2.1.2. The concentration inequalities are stated in two forms, a simpler more straightforward form with \mathbf{W} and a rearranged form involving $\mathbf{W}_G \triangleq \mu\mathbf{W} - \frac{\mu^2}{2}$, which is \mathbf{W} scaled and shifted to match the 1st and 2nd order statistics of \mathbf{G} in the low per-feature correlation limit, as shown in Lemma 50 in Appendix A. The second form with \mathbf{W}_G mirrors the concentration inequalities in Chapter 4. This allows for a comparison between the probability densities in the planted matching setting and the database alignment setting.

Lemma 5 (Atypicality). *Let m a partial mapping fully contained in the true mapping M . Given some $\tau_W \leq \mu$ and $\tau_G = \mu\tau - \frac{\mu^2}{2}$,*

$$\Pr[\tau_W | m| \geq \langle \mathbf{W}, \mathbf{m} \rangle | m \subseteq M] \leq \exp\left(-\frac{(\mu - \tau_W)^2}{2}\right)$$

which is equivalent to

$$\Pr[\tau_G | m| \geq \langle \mathbf{W}_G, \mathbf{m} \rangle | m \subseteq M] \leq \exp\left(-\frac{(\zeta - \tau_G)^2}{4\zeta}\right)$$

where $\zeta = \mu^2/2$.

Proof. Given $u \stackrel{M}{\sim} v$, $W_{u,v}$ is normal with mean μ and unit variance. Then, its moment generating function is given by $\mathbb{E}[e^{\theta W_{u,v}}] = e^{\theta\mu + \frac{\theta^2}{2}}$. By Markov's inequality, for any

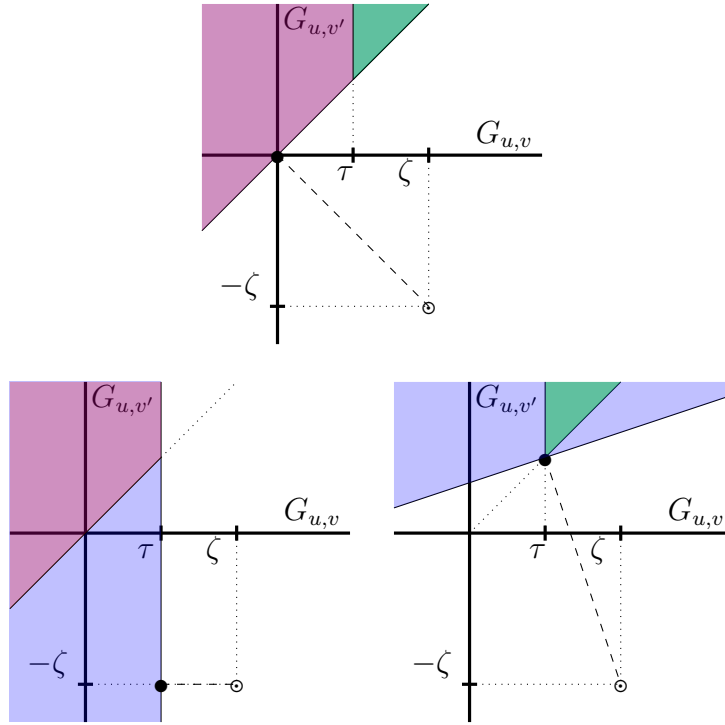


Figure 3.1: Illustration of the partition of the error event into atypicality and misalignment-despite-typicality events. The horizontal axis represents the information density $G_{u,v}$ of true pair (u, v) while the vertical axis correspond to that of a false pair $G_{u,v'}$. The point $(\zeta, -\zeta)$ corresponds to the mean of the distribution. Minimum distance from the mean to the half-space is shown by the bold dashed lines. The top plot shows the partition of the original error event into ‘slices’ while the bottom plots show half-spaces containing those ‘slices’.

Top: Partition of the event half-space corresponding to error event $\{G_{u,v'} \geq G_{u,v}\}$ into two ‘slices’ according to whether information density $G_{u,v}$ of the true pair (u, v) is above or below the threshold τ . Bottom-left: The atypicality ‘slice’ is closer to the mean, however it only needs to be counted once for all error events involving $G_{u,v}$. Bottom-right: The misalignment-despite-typicality ‘slice’ is farther away from the mean than the original half-space and therefore has smaller likelihood.

$\theta < 0$,

$$\Pr \left[W_{u,v} \leq \tau_W | u \stackrel{M}{\sim} v \right] \leq e^{-\theta \tau_W} \mathbb{E} \left[e^{\theta W_{u,v}} \right] = \exp \left(-\theta \tau_W + \theta \mu + \frac{\theta^2}{2} \right)$$

Pick $\theta = -(\mu - \tau_W)$. Then $\Pr \left[W_{u,v} \leq \tau_W | u \stackrel{M}{\sim} v \right] \leq \exp \left(-\frac{(\mu - \tau_W)^2}{2} \right)$. Since all entries in \mathbf{W} are independent, $\Pr \left[\langle \mathbf{W}, \mathbf{m} \rangle \leq \tau_W | m | \mid m \subseteq M \right]$ is the product of all of these terms.

$\frac{(\mu - \tau_W)^2}{2} = \frac{(\mu^2 - \mu \tau_W)^2}{2\mu^2} = \frac{(\mu^2/2 - (\mu \tau_W - \mu^2/2))^2}{2\mu^2} = \frac{(\zeta - \tau_G)^2}{4\zeta}$. Then $\Pr \left[(W_G)_{u,v} \leq \tau_G | u \stackrel{M}{\sim} v \right] = \Pr \left[W_{u,v} \leq \tau_W | u \stackrel{M}{\sim} v \right] \leq \exp \left(-\frac{(\zeta - \tau_G)^2}{4\zeta} \right)$. Once again, taking the product of all of this term over all pairs in m gives us the claimed result. \square

Lemma 6 (False positive probability). *Given some $\tau_W \geq 0$ and $\tau_G = \mu\tau - \frac{\mu^2}{2}$,*

$$\Pr \left[W_{u,v} \geq \tau_W | u \not\stackrel{M}{\sim} v \right] \leq \exp \left(-\frac{\tau_W^2}{2} \right)$$

which is equivalent to

$$\Pr \left[(W_G)_{u,v} \leq \tau_G | u \not\stackrel{M}{\sim} v \right] \leq \exp \left(-\frac{(\zeta + \tau_G)^2}{4\zeta} \right)$$

where $\zeta = \mu^2/2$.

Proof. Given $u \not\stackrel{M}{\sim} v$, $W_{u,v}$ is normal with zero μ and unit variance. Then, its moment generating function is given by $\mathbb{E} \left[e^{\theta W_{u,v}} \right] = e^{\frac{\theta^2}{2}}$. By Markov's inequality, for any $\theta > 0$,

$$\Pr \left[W_{u,v} \leq \tau_W | u \not\stackrel{M}{\sim} v \right] \leq e^{-\theta \tau_W} \mathbb{E} \left[e^{\theta W_{u,v}} \right] = \exp \left(-\theta \tau_W + \frac{\theta^2}{2} \right)$$

Pick $\theta = \tau_W$. Then $\Pr \left[W_{u,v} \geq \tau_W | u \not\stackrel{M}{\sim} v \right] \leq \exp \left(-\frac{\tau_W^2}{2} \right)$.

$\frac{\tau_W^2}{2} = \frac{(\mu \tau_W)^2}{2\mu^2} = \frac{(\mu^2/2 + (\mu \tau_W - \mu^2/2))^2}{2\mu^2} = \frac{(\zeta + \tau_G)^2}{4\zeta}$. Then $\Pr \left[(W_G)_{u,v} \leq \tau_G | u \not\stackrel{M}{\sim} v \right] = \Pr \left[W_{u,v} \leq \tau_W | u \not\stackrel{M}{\sim} v \right] \leq \exp \left(-\frac{(\zeta + \tau_G)^2}{4\zeta} \right)$. \square

Lemma 7 (Misalignment). *Let m and m' denote two mappings of same size and δ denote the number of pairs mapped by m but not by m' . Then $\Pr[\langle \mathbf{W}, \mathbf{m} \rangle \leq \langle \mathbf{W}, \mathbf{m}' \rangle] \leq \exp\left(-\delta \frac{\mu^2}{4}\right)$.*

$$\Pr[\langle \mathbf{W}, \mathbf{m} \rangle \leq \langle \mathbf{W}, \mathbf{m}' \rangle | M = m] \leq \exp\left(-\delta \frac{\mu^2}{4}\right)$$

which is equivalent to

$$\Pr[\langle \mathbf{W}_G, \mathbf{m} \rangle \leq \langle \mathbf{W}_G, \mathbf{m}' \rangle | M = m] \leq \exp\left(-\delta \frac{\zeta}{2}\right)$$

where $\zeta = \mu^2/2$.

Proof. $\langle \mathbf{W}, \mathbf{m} \rangle - \langle \mathbf{W}, \mathbf{m}' \rangle$ is the linear combination of 2δ independent Gaussian random variables and therefore is Gaussian. ($|m| - \delta$ of the terms in $\langle \mathbf{W}, \mathbf{m} \rangle$ get canceled out by the $|m| - \delta$ common terms in $\langle \mathbf{W}, \mathbf{m}' \rangle$.) The difference has mean $\delta\mu$ and variance 2δ . Then, the moment generating function is given by $\mathbb{E}[\exp(\theta \langle \mathbf{W}, \mathbf{m} \rangle - \theta \langle \mathbf{W}, \mathbf{m}' \rangle)] = \exp(\theta\delta\mu + \theta^2\delta)$. Then, by Markov's inequality, for any $\theta < 0$,

$$\begin{aligned} \Pr[\langle \mathbf{W}, \mathbf{m} \rangle \leq \langle \mathbf{W}, \mathbf{m}' \rangle | M = m] &= \Pr[\langle \mathbf{W}, \mathbf{m} \rangle - \langle \mathbf{W}, \mathbf{m}' \rangle \leq 0 | M = m] \\ &\leq \mathbb{E}[\exp(\theta \langle \mathbf{W}, \mathbf{m} \rangle - \theta \langle \mathbf{W}, \mathbf{m}' \rangle)] \\ &= \exp(\theta\delta\mu + \theta^2\delta). \end{aligned}$$

Picking $\theta = -\mu/2$ gives us the claimed result. □

Lemma 8 (Misalignmen-despite-typicality). *Let m and m' denote two mappings of same size δ such that no pair is mapped under both mappings. Given some $\tau_W \geq \mu/2$ and $\tau_G = \mu\tau - \frac{\mu^2}{2}$,*

$$\Pr[\langle \mathbf{W}, \mathbf{m} \rangle \geq \tau_W |m| \text{ and } \langle \mathbf{W}, \mathbf{m}' \rangle \geq \langle \mathbf{W}, \mathbf{m} \rangle | m \subseteq M] \leq \exp(-\delta \cdot (\tau_W - \mu/2)^2 - \delta\mu^2/4)$$

which is equivalent to

$$\Pr [\langle \mathbf{W}_G, \mathbf{m} \rangle \geq \tau_G |m| \text{ and } \langle \mathbf{W}_G, \mathbf{m}' \rangle \geq \langle \mathbf{W}_G, \mathbf{m} \rangle \mid m \subseteq M] \leq \exp \left(-\delta \cdot \frac{\tau_G^2 + \zeta^2}{2\zeta} \right)$$

where $\zeta = \mu^2/2$.

Proof. If $y \geq x$ and $x \geq t$, then $\theta_1(y - x) + \theta_2(x - t) \geq 0$ for any choice of $\theta_1, \theta_2 > 0$.

Replacing y by $\langle \mathbf{W}, \mathbf{m}' \rangle$, x by $\langle \mathbf{W}, \mathbf{m} \rangle$ and t by $\tau_W |m|$, we get the implication between the events of interest: $\langle \mathbf{W}, \mathbf{m}' \rangle \geq \langle \mathbf{W}, \mathbf{m} \rangle$ and $\langle \mathbf{W}, \mathbf{m} \rangle \geq \tau_W |m|$ implies $\theta_1 (\langle \mathbf{W}, \mathbf{m}' \rangle - \langle \mathbf{W}, \mathbf{m} \rangle) + \theta_2 \langle \mathbf{W}, \mathbf{m} \rangle \geq \theta_2 \tau_W |m|$.

$\langle \mathbf{W}, \mathbf{m} \rangle \geq \tau_W |m|$ implies $\theta_1 (\langle \mathbf{W}, \mathbf{m}' \rangle - \langle \mathbf{W}, \mathbf{m} \rangle) + \theta_2 \langle \mathbf{W}, \mathbf{m} \rangle$ is the linear combination of 2δ independent Gaussian random variables and is therefore Gaussian. It has mean $\delta\mu(\theta_2 - \theta_1)$ and variance $\delta\theta_1^2 + \delta(\theta_2 - \theta_1)^2 = \delta(2\theta_1^2 - 2\theta_1\theta_2 + \theta_2^2)$.

By Markov's inequality

$$\begin{aligned} & \Pr [\theta_1 (\langle \mathbf{W}, \mathbf{m}' \rangle - \langle \mathbf{W}, \mathbf{m} \rangle) + \theta_2 \langle \mathbf{W}, \mathbf{m} \rangle \geq \theta_2 \tau_W |m| \mid m \subseteq M] \\ & \leq e^{-\theta_2 \tau_W |m|} \mathbb{E} \left[e^{\theta_1 (\langle \mathbf{W}, \mathbf{m}' \rangle - \langle \mathbf{W}, \mathbf{m} \rangle) + \theta_2 \langle \mathbf{W}, \mathbf{m} \rangle} \mid m \subseteq M \right] \\ & = \exp \left(-\theta_2 \tau_W \delta + \delta\mu(\theta_2 - \theta_1) + \frac{\delta}{2} (2\theta_1^2 - 2\theta_1\theta_2 + \theta_2^2) \right) \end{aligned}$$

Pick $\theta_1 = \tau_W$ and $\theta_2 = 2(\tau_W - \mu/2)$. Then, the expression in the last line simplifies to

$$\exp \left(-(\tau_W - \mu/2)^2 - \mu^2/4 \right),$$

which matches the first part of the claim.

$-(\tau_W - \mu/2)^2 - \mu^2/4 = -\frac{(\mu\tau - \mu^2/2) + (\mu^2/2)^2}{\mu^2} = \frac{\tau_G^2 + \zeta^2}{2\zeta}$ gives us the second part of the claim. \square

3.3 Achievability

In Subsection 3.3.1, we state asymptotic results which give bounds on the signal strength ζ that guarantee upper bounds on the expected number of errors as n , the size of the matching, goes to infinity. In Subsection 3.3.2, we state finite-size results that bound the expected number of errors as a function of the signal strength ζ for matchings of any finite size. The proofs for these results are given in Subsections 3.3.3 (threshold testing), 3.3.4 (maximum row estimation) and 3.3.5 (maximum likelihood estimation), where we perform different analyses to derive the expressions for the bounds of various forms.

As described in Subsection 2.1.2, the Gaussian bigraph has normally distributed edge weights with unit variance, mean 0 for pairs not included in the planted matching and mean μ for pairs included in it. All results in this section are stated in terms of $\zeta \triangleq \mu^2/2$, which we refer to as signal strength. This re-parametrization is done in order to state results in a form that mirrors the ones in Chapter 4.

Recall that, as defined in Subsection 2.2.4, $d(\hat{M}, M) \triangleq \frac{1}{2} \|\hat{M} - M\|_F^2$. For \hat{M}_{ML} and \hat{M}_{MR} this distance gives the number of users in \mathcal{U} mapped to a false match in \mathcal{V} . For \hat{M}_{MR} , the distance gives the arithmetic mean of the number of false positives and false negatives.

3.3.1 Asymptotic results

In the asymptotic case, as the size of the matching n goes to infinity, we say an algorithm achieves exact alignment if $\mathbb{E}[d(\hat{M}, M)] \leq o(1)$, and almost-exact alignment if $\mathbb{E}[d(\hat{M}, M)] \leq o(n)$.

Theorem 3.3.1 (Exact alignment and almost-exact alignment recovery). *Let $n = |M| = |\mathcal{U}|$. Define $\alpha = \frac{\log(|\mathcal{V}| - n)}{\log n}$ if $|\mathcal{V}| > n$. Consider the Gaussian planted matching problem.*

- *Lower bounds on ζ as given in Table 3.1 are sufficient conditions for almost-exact alignment for different cases.*
- *$\zeta \geq c \log n + \omega(1)$ is a sufficient condition for exact alignment, where the values of*

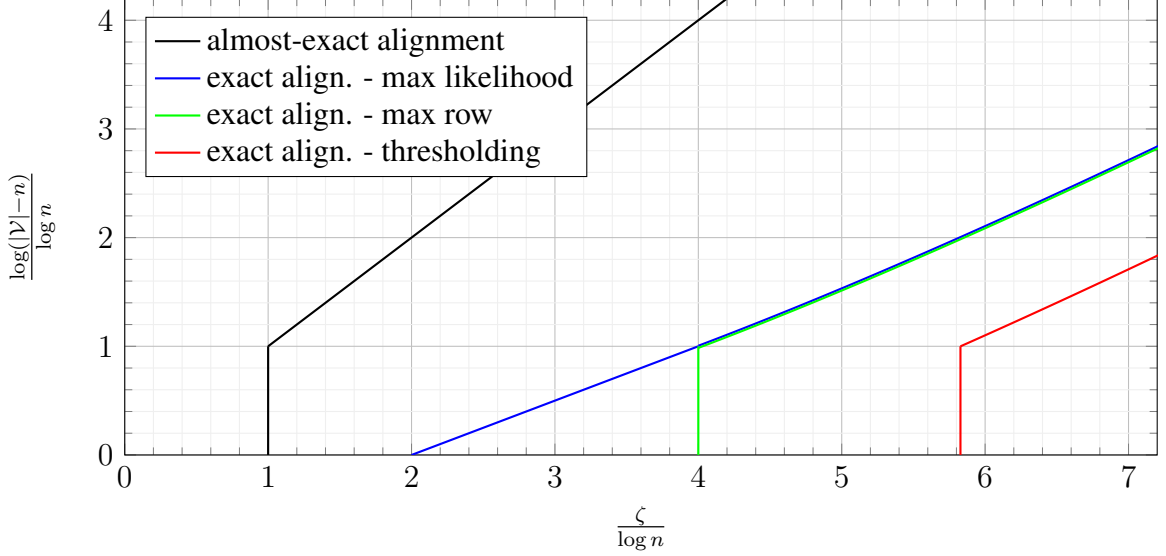


Figure 3.2: Comparison of boundaries for achievability regions of exact alignment and almost-exact alignment as given in Theorems 3.3.1 and 4.4.1. x-axis corresponds to signal strength and y-axis corresponds to order of magnitude of number of unmatched users in \mathcal{V} . Achievability regions are areas below/right of curves.

c for different cases are given in Table 3.2.

Table 3.1: Sufficient conditions for almost-exact alignment for Gaussian planted matching. Values in the table are lower bounds on ζ that guarantee almost-exact alignment for different sizes of $|\mathcal{V}|$.

Algorithm	Threshold.	Max row	Max likelihood
$ \mathcal{V} \leq n + o(n)$		$\log n + \omega(\sqrt{\log n})$	
$ \mathcal{V} \geq n + \Omega(n)$		$\alpha \log n + \omega(\sqrt{\log n})$	

Note that for threshold testing and maximum row estimation, the exact-alignment thresholds does not change with the size of $|\mathcal{V}|$ as long as $|\mathcal{V}|$ is on the order of n . For maximum likelihood estimation, the exact-alignment threshold increases linearly with α in this regime. For $|\mathcal{V}| \geq n + \Omega(n)$, the exact-alignment thresholds for all three algorithms increase quadratically with $\sqrt{\alpha}$. The boundaries for exact and almost-exact alignment for the algorithms are illustrated in Fig. 3.2. Boundaries for maximum likelihood and maximum row algorithms completely overlap for $\zeta/\log n \geq 4$.

Theorem 3.3.2 (Expected number of errors in the balanced case). *Let $n = |M| = |\mathcal{U}| =$*

Table 3.2: Sufficient conditions for exact alignment for Gaussian planted matching. Sufficient conditions are of the form $\zeta \geq c \log n + \omega(1)$, where the value of c is given in the table for different algorithms and for different sizes of $|\mathcal{V}|$.

Size of \mathcal{V}	$ \mathcal{V} = n$	$n < \mathcal{V} \leq n + o(n)$	$ \mathcal{V} \geq n + \Omega(n)$
Thresholding	$(1 + \sqrt{2})^2$	$(1 + \sqrt{2})^2$	$(1 + \sqrt{1 + \alpha})^2$
Max row	4	4	$(1 + \sqrt{\alpha})^2$
Max likelihood	2	$2(\alpha + 1)$	$(1 + \sqrt{\alpha})^2$

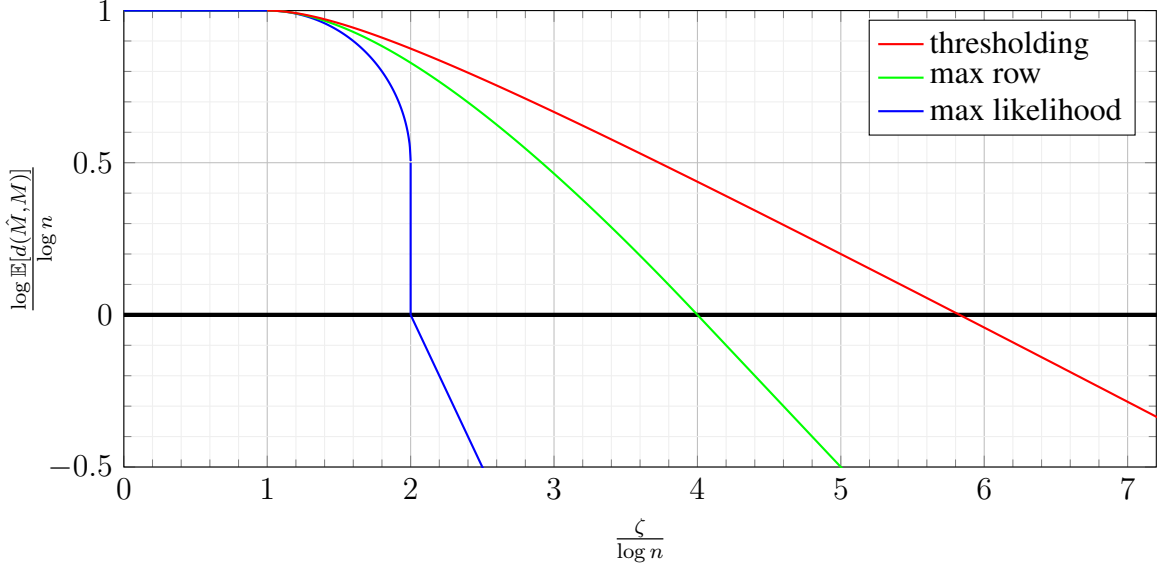


Figure 3.3: Comparison of the boundaries for the achievability regions in the balanced case (i.e. when $|\mathcal{V}| = |\mathcal{U}| = |M| = n$) as given in Theorems 3.3.2 and 4.4.2. The x-axis is to signal strength and the y-axis is the order of magnitude of the number of the expected number of mismatched users. Achievability regions are areas above/right of curves.

$|\mathcal{V}|$. Let $\beta > 0$ some constant. For Gaussian planted matching, as $n \rightarrow \infty$, lower bounds on $\zeta / \log n$ that guarantee error bounds $\mathbb{E}[d(\hat{M}, M)] \leq \mathcal{O}(n^{1-\beta})$ are given in Table 3.3:

The boundaries for the achievability regions of the algorithms are illustrated in Fig. 3.3.

Theorem 3.3.3 (Expected number of errors in unbalanced case). *Let $n = |M| = |\mathcal{U}|$ and $|\mathcal{V}| > n$. Define¹ $\nu \triangleq \frac{\log |\mathcal{V}|}{\log n}$ and $\alpha \triangleq \frac{\log(|\mathcal{V}| - n)}{\log n}$, and let these be constant functions of n . Let $\beta > 0$ some constant. For Gaussian database alignment, as $n \rightarrow \infty$, lower bounds on $\zeta / \log n$ that guarantee error bounds $\mathbb{E}[d(\hat{M}, M)] \leq \mathcal{O}(n^{1-\beta})$ are given in Table 3.4:*

¹Asymptotically, ν can be replaced by $\max\{1, \alpha\}$ as $1 \leq \frac{\nu}{\max\{1, \alpha\}} \leq 1 + \Omega(1/\log n)$.

Table 3.3: Sufficient conditions for error bounds for Gaussian planted matching in the balanced case, i.e. $|\mathcal{V}| = |\mathcal{U}| = |M| = n$. Sufficient conditions are expressed as lower bounds on $\zeta/\log n$. Different conditions are relevant for different ranges of β . The form of the achievability boundary (as illustrated in Fig. 3.3) corresponding to each condition is given in the 3rd column.

Bound on $\zeta/\log n$	Range of β	Form of boundary	Bound on $\mathbb{E}[d(\hat{M}, M)]$
Threshold testing			
$(\sqrt{\beta} + \sqrt{1 + \beta})^2$	$0 < \beta$	parabolic	$n^{1-\beta}$
Maximum row estimation			
$(1 + \sqrt{\beta})^2$	$0 < \beta \leq 1$	parabolic	$2n^{1-\beta}$
$2(1 + \beta)$	$1 < \beta$	linear	$n^{1-\beta}$
Maximum likelihood estimation			
$1 + 2\sqrt{\beta(1 - \beta)}$	$0 < \beta < 1/2$	elliptic	$\mathcal{O}(n^{1-\beta})$
$2 + \frac{2 \log\left(\frac{\sqrt{5}-1}{2}\right)}{\log n}$	$1/2 \leq \beta \leq 1$	vertical	$n^{1-\beta}$
$1 + \beta$	$1 < \beta$	linear	$n^{1-\beta}(1 + o(1))$

The boundaries for the achievability region of the maximum likelihood estimator for various values of α are illustrated in Fig. 3.4. Boundaries for $\alpha = 1.5$ and $\alpha = 1.25$ are tangent to $y = 1$ at $x = 1.5$ and $x = 1.25$ respectively, while all other boundaries are tangent to that line at $x = 1$. These match the almost-exact alignment threshold.

The boundaries for the achievability region of maximum likelihood/maximum row estimation, which coincide, and for thresholding at $\alpha = 1.5$ are illustrated in Fig. 3.5. Both boundaries are tangent to $y = 1$ at $x = 1.5$. This matches the almost-exact alignment threshold.

3.3.2 Finite-size results

Theorem 3.3.4 (Expected number of errors for finite-size case). *Define $n \triangleq |M|$, $x \triangleq \frac{\zeta}{\log n}$, $\nu \triangleq \frac{\log |\mathcal{V}|}{\log n}$ and $\varepsilon \triangleq n^{1-x/2}$. Suppose $|\mathcal{U}| = n$. Then, $\log \mathbb{E} \left[d(\hat{M}, M) \right]$, the log of the expected number of errors, is bounded as given in Table 3.3.4 for the planted matching problem.*

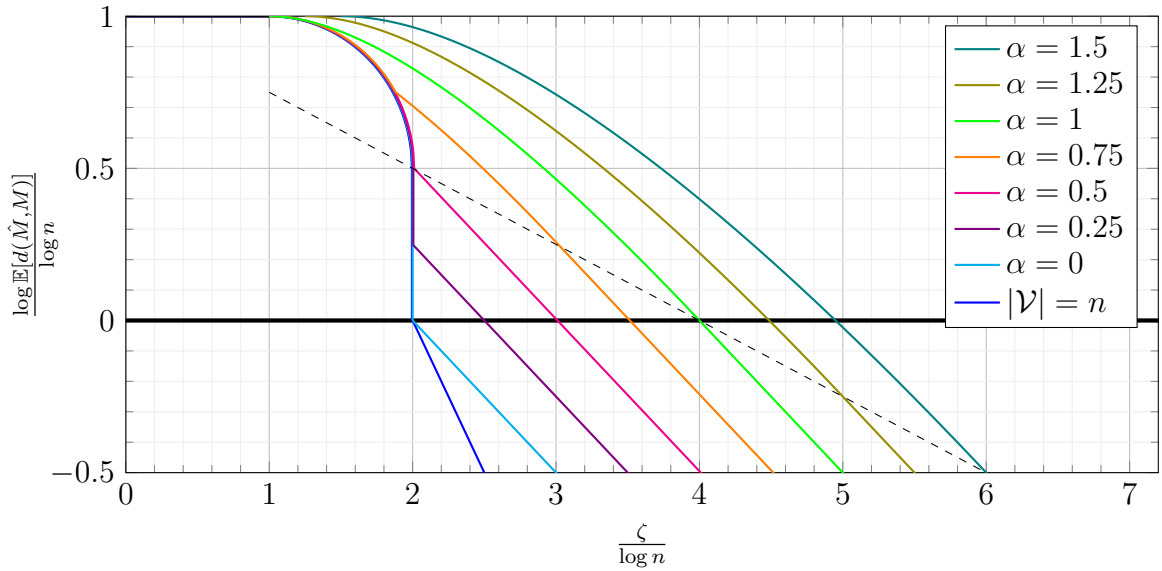


Figure 3.4: Comparison of boundaries for achievability regions of the maximum likelihood algorithm for various values of α as give in Theorems 3.3.3 and 4.4.3. The x-axis is to signal strength and y-axis is the order of magnitude of expected number of mismatched users. Achievability regions are areas above/right of curves. Boundaries are parabolic/elliptic in the region above the dashed black line and consist of linear segments below it.

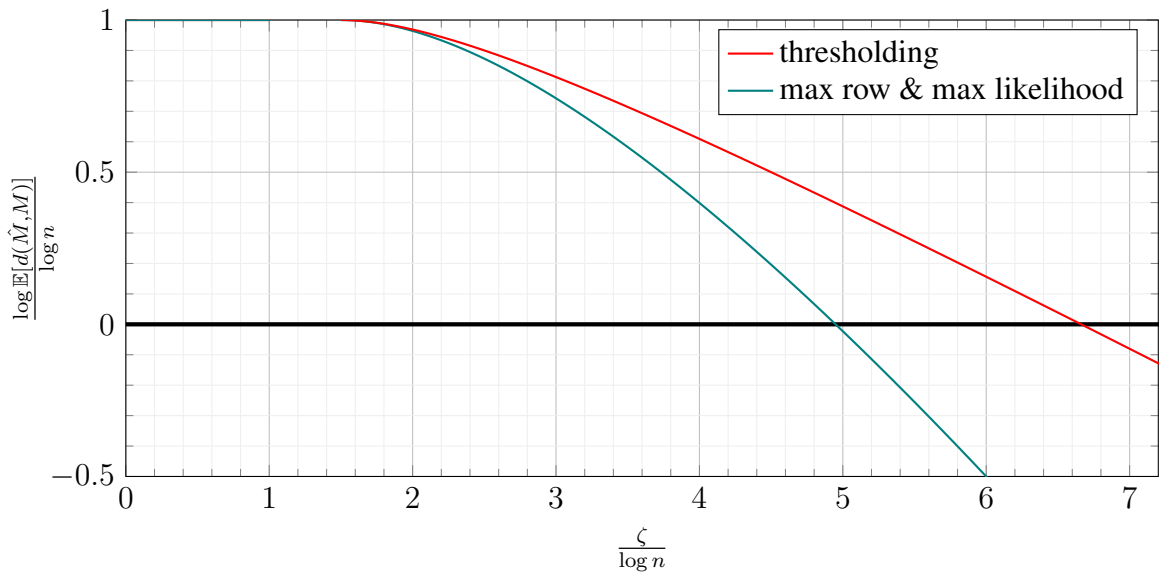


Figure 3.5: Comparison of the boundaries for the achievability regions when $|\mathcal{U}| = |M| = n$ and $|\mathcal{V}| = n^{1.5}$. The x-axis is signal strength and the y-axis is the order of magnitude of the expected number of mismatched users. Achievability regions are areas above/right of curves.

Table 3.4: Sufficient conditions for error bounds for Gaussian planted matching in unbalanced case, i.e. $|\mathcal{V}| > |\mathcal{U}| = |M| = n$. Sufficient conditions are expressed as lower bounds on $\zeta / \log n$. Different conditions are relevant for different ranges of β . The form of the achievability boundary (as illustrated in Fig. 3.4 and Fig. 3.5) corresponding to each condition is given in the 3rd column.

Bound on $\zeta / \log n$	Range of β	Form of boundary	Bound on $\mathbb{E}[d(\hat{M}, M)]$
Threshold testing			
$(\sqrt{\nu} + \beta + \sqrt{\beta})^2$	$0 < \beta$	parabolic	$n^{1-\beta}$
Maximum row estimation			
$(\sqrt{\nu} + \sqrt{\beta})^2$	$0 < \beta \leq \max\{\alpha, 1\}$	parabolic	$2n^{1-\beta}$
$2(\nu + \beta) \log n$	$\max\{\alpha, 1\} < \beta$	linear	$n^{1-\beta}$
Maximum likelihood estimation			
$1 + 2\sqrt{\beta(1-\beta)}$	$0 < \beta < \min\{1-\alpha, 1/2\}$	elliptic	$\mathcal{O}(n^{1-\beta})$
$(\sqrt{\alpha} + \sqrt{\beta})^2$	$1-\alpha < \beta \leq \alpha$	parabolic	$\mathcal{O}(n^{1-\beta})$
$2 + \frac{2 \log\left(\frac{3+\sqrt{5}}{2}\right)}{\log n}$	$1/2 \leq \beta \leq 1-\alpha$	vertical	$n^{1-\beta} + \mathcal{O}(1)$
$2(\alpha + \beta)$	$\max\{\alpha, 1-\alpha\} < \beta$	linear	$n^{1-\beta}(1 + o(1))$

3.3.3 Proof for Threshold Testing

The proofs for threshold testing and maximum row estimation state inequalities use the variable $\nu \triangleq \frac{\log |\mathcal{V}|}{\log n}$ instead of the variable $\alpha \triangleq \frac{\log(|\mathcal{V}|-n)}{\log n}$. $\nu = 1 + \mathcal{O}\left(\frac{1}{\log n}\right)$ and $\alpha \in (0, 1]$ if $n < |\mathcal{V}| \leq 2n$. $\nu = \frac{\log(|\mathcal{V}|-n)}{\log n} \cdot \frac{\log |\mathcal{V}|}{\log(|\mathcal{V}|-n)} = \alpha \cdot \left(1 + \mathcal{O}\left(\frac{1}{\log n}\right)\right)$ if $|\mathcal{V}| \geq n + \Omega(n)$. Then $\max\{1, \alpha\} \leq \nu \leq \max\{1, \alpha\} \left(1 + \mathcal{O}\left(\frac{1}{\log n}\right)\right)$, which allows us to translate asymptotic statements with expressions containing ν to those that only involve α .

Parabolic boundary

We re-state the following results for threshold testing, and introduce their proofs below:

Let $\tau = \log \frac{|\mathcal{U} \cdot |\mathcal{V}|}{|M|}$ be the value of the threshold used for threshold testing.

- Theorem 3.3.1

Almost-exact alignment is achieved if $\zeta \geq \nu \log n + \omega(\sqrt{\log n})$.

By Lemma 1, this condition also guarantees almost-exact alignment for maximum row estimation.

Exact alignment is achieved if $\zeta \geq (1 + \sqrt{1 + \nu})^2 \log n + \omega(1)$.

Table 3.5: Error bounds for Gaussian planted matching with finite-size

Error bound	Condition
Threshold testing $\left(1 - \frac{(x-\nu)^2}{4x}\right) \log n$	$\nu < x$
Maximum row estimation $(1 - (\sqrt{\nu} - \sqrt{x})^2) \log n + \log 2$ $\left(1 + \nu - \frac{x}{2}\right) \log n$	$\nu < x < 4\nu$ $4\nu \leq x$
Maximum likelihood estimation $2 \log \varepsilon - \log(1 - \varepsilon)$ $\log \varepsilon - 2 \log(1 - \varepsilon) + \log(\mathcal{V} - n + 1)$	$x > 2$ and $ \mathcal{V} = n$ $x > 2$

- Theorem 3.3.3

The expected number of false positive and of false negatives are both bounded by $n^{1-\beta}$ if $\zeta \geq (\sqrt{\nu + \beta} + \sqrt{\beta})^2 \log n$.

- Theorem 3.3.2

This is a special case of Theorem 3.3.3 with $\nu = 1$.

- Theorem 3.3.4

Given $x > \nu$, the expected number of errors is bounded by $n^{1 - \frac{(x-\nu)^2}{4x}}$.

Proof. Let τ the threshold such that $|\tau| \leq \zeta$. By Lemma 5, the probability of a true pair failing the test is at most $\exp\left(-\frac{(\zeta-\tau)^2}{4\zeta}\right)$ and by Lemma 6 the probability of a false pair passing the test is at most $\exp\left(-\frac{(\zeta+\tau)^2}{4\zeta}\right)$. The number of true pairs and false pairs are $|M|$ and $|\mathcal{U}| \cdot |\mathcal{V}| - |M|$ respectively. We bound the latter by $|\mathcal{U}| \cdot |\mathcal{V}|$. So the expected number of false negatives is bounded by $|M| \exp\left(-\frac{(\zeta-\tau)^2}{4\zeta}\right)$ and expected number of false positives is bounded by $|\mathcal{U}| \cdot |\mathcal{V}| \exp\left(-\frac{(\zeta+\tau)^2}{4\zeta}\right)$.

The log of the ratio of these two bounds is

$$\log \frac{|M| \exp\left(-\frac{(\zeta-\tau)^2}{4\zeta}\right)}{|\mathcal{U}| \cdot |\mathcal{V}| \exp\left(-\frac{(\zeta+\tau)^2}{4\zeta}\right)} = \tau - \log \frac{|\mathcal{U}| \cdot |\mathcal{V}|}{|M|}.$$

The choice of $\tau = \log \frac{|\mathcal{U}| \cdot |\mathcal{V}|}{|M|}$ makes the log of the ratio zero. Hence the bounds for the

expected numbers of false positives and negatives are equal. Then the bound on the number of errors is twice the bound on the number of false negatives.

Let $n \triangleq |M|$, $|\mathcal{U}| = n$ and $|\mathcal{V}| = n^\nu$ for some $\nu \geq 1$. Then, $\tau = \log \frac{|\mathcal{U}||\mathcal{V}|}{|M|} = \nu \log n$. Since we require $|\tau| \leq \zeta$, then we need $\zeta \geq \nu \log n$.

Let $x \triangleq \frac{\zeta}{\log n}$. Then the number of false negatives is given by

$$|M| \exp\left(-\frac{(\zeta - \tau)^2}{4\zeta}\right) = n^{1 - \frac{(x - \nu)^2}{4x}}.$$

This expression is bounded by $n^{1-\beta}$ if

$$x \geq \left(\sqrt{\nu + \beta} + \sqrt{\beta}\right)^2. \quad (3.1)$$

This gives us the following inequalities that form part of the main results:

- Almost-exact alignment is achieved if (3.1) is satisfied for some β such that $n^{-\beta} \leq o(1)$, which is equivalent to $\beta \geq \omega(1/\log n)$. Such β exists if $\zeta \geq \nu \log n + \omega(\sqrt{\log n})$.

Exact alignment is achieved if (3.1) is satisfied for some β such that $n^{1-\beta} \leq o(1)$, which is equivalent to $\beta - 1 \geq \omega(1/\log n)$. Such β exists if $\zeta \geq (1 + \sqrt{1 + \nu})^2 \log n + \omega(1)$.

- The expected number of false positive and of false negatives are both bounded by $n^{1-\beta}$ if $\zeta \geq (\sqrt{\nu + \beta} + \sqrt{\beta})^2 \log n$.

□

3.3.4 Proofs for Maximum Row Estimation

Consider users $u \in \mathcal{U}$ and $v, v' \in \mathcal{V}$ such that $u \stackrel{M}{\sim} v$. We want to bound the probability of the error event where u is falsely mapped to v' . Under maximum row estimation, this corresponds to the event $G_{u,v} \leq G_{u,v'}$. Without loss of generality, assume the set \mathcal{U} consists of the single user u .

Recall that, as described in Subsection 3.3.3, proofs for threshold testing and maximum row estimation state inequalities use the variable $\nu \triangleq \frac{\log |\mathcal{V}|}{\log n}$ instead of the variable $\alpha \triangleq \frac{\log(|\mathcal{V}|-n)}{\log n}$. The asymptotic statements for threshold testing and maximum row estimation containing ν directly translate to the statements in the main results using the fact that $\max\{1, \alpha\} \leq \nu \leq \max\{1, \alpha\} \left(1 + \mathcal{O}\left(\frac{1}{\log n}\right)\right)$.

Linear boundary - relevant for large β and small α

We re-state the following results for maximum row estimation, and introduce their proofs below:

- Theorem 3.3.1

Exact alignment is achieved if $\zeta \geq 2(1 + \nu) \log n + \omega(1)$.

- Theorem 3.3.3

The number of errors is bounded by $n^{1-\beta}$ if $\zeta \geq (\nu + \beta) \log n$.

- Theorem 3.3.2

This is a special case of Theorem 3.3.3 with $\nu = 1$.

- Theorem 3.3.4

Given $\nu < x$, the expected number of errors is bounded by $n^{1+\nu-\frac{x}{2}}$.

Proof. By Lemma 7, the probability of $\{G_{u,v} \leq G_{u,v'}\}$ is upper bounded by $\exp\left(-\frac{\zeta}{2}\right)$.

There are no more than $|\mathcal{V}|$ vertices $v' \in \mathcal{V} \setminus \{v\}$ that to which u can be falsely mapped. By the union bound, the probability that any of these events happens is upper bounded by $|\mathcal{V}| \exp\left(-\frac{\zeta}{2}\right)$. Then, the expected number of misalignments over all $|\mathcal{U}|$ of rows is upper bounded by $|\mathcal{U}| \cdot |\mathcal{V}| \exp\left(-\frac{\zeta}{2}\right)$.

Let $n \triangleq |M|$, $|\mathcal{U}| = n$, $|\mathcal{V}| = n^\nu$ for some $\nu \geq 1$, and $x \triangleq \frac{\zeta}{\log n}$. Then, the bound on the expected number of misalignments is given by $n^{1+\nu-\frac{x}{2}}$. This expression is bounded by $n^{1-\beta}$ if

$$x \geq 2(\nu + \beta). \tag{3.2}$$

This gives us the following inequalities that form part of the main results:

- Exact alignment is achieved if (3.2) is satisfied for some β such that $n^{1-\beta} \leq o(1)$, which is equivalent to $\beta - 1 \geq \omega(1/\log n)$. Such β exists if $\zeta \geq 2(1+\nu) \log n + \omega(1)$.
- The number of errors is bounded by $n^{1-\beta}$ if $\zeta \geq (\nu + \beta) \log n$.

□

Parabolic boundary - relevant for small β and large α

We re-state the following results for maximum row estimation, and introduce their proofs below:

Suppose $\zeta \leq 4 \log |\mathcal{V}|$.

- Theorem 3.3.1

Exact alignment is achieved if $\zeta \geq (1 + \sqrt{\nu})^2 \log n + \omega(1)$.

(The proof for the condition for almost-exact alignment for maximum row estimation follows the proof for threshold testing as a result of Lemma 1.)

- Theorem 3.3.3

The number of errors is bounded by $2n^{1-\beta}$ if $\beta \leq \nu$ and $\zeta \geq (\sqrt{\nu} + \sqrt{\beta})^2 \log n$.

- Theorem 3.3.2

This is a special case of Theorem 3.3.3 with $\nu = 1$.

- Theorem 3.3.4

Given $\nu < x < 4\nu$, the expected number of errors is bounded by $2n^{1-(\sqrt{\nu}-\sqrt{x})^2}$.

Proof. Let τ a threshold such that $0 \leq \tau \leq \zeta$. For the purpose of analysis, let us consider the alignment of the row corresponding to u a failure if either $\{G_{u,v} < \tau\}$ or $\{G_{u,v} \geq \tau \text{ and } G_{u,v} \leq G_{u,v'}\}$ for some $v' \in \mathcal{V} \setminus \{v\}$. In other words, we say the the algorithm has failed on the given row if either the true pair has score atypically low, in which case many false pairs will beat the true pair, or if a false pair beats the true pair despite the true pair having sufficiently high score. These two events fully cover the actual error event $\{G_{u,v} \leq G_{u,v'}\}$.

By Lemma 5, the probability of $\{G_{u,v} \leq \tau\}$ is bounded by $\exp\left(-\frac{(\zeta-\tau)^2}{4\zeta}\right)$. By Lemma 8, the probability of $\{G_{u,v} \geq \tau \text{ and } G_{u,v} \leq G_{u,v'}\}$ is bounded by $\exp\left(-\frac{\zeta^2+\tau^2}{2\zeta}\right)$. There are no more than $|\mathcal{V}|$ vertices $v' \in \mathcal{V} \setminus \{v\}$ that to which u can be falsely mapped. Then, by the union bound, the probability of $\{G_{u,v} \geq \tau \text{ and } G_{u,v} \leq G_{u,v'}\}$ for some v' is bounded by $|\mathcal{V}| \exp\left(-\frac{\zeta^2+\tau^2}{2\zeta}\right)$.

The log of the ratio of these two bounds is

$$\log \frac{\exp\left(-\frac{(\zeta-\tau)^2}{4\zeta}\right)}{|\mathcal{V}| \exp\left(-\frac{\zeta^2+\tau^2}{2\zeta} + 6\rho_{\max}^2 \tau\right)} = \frac{(\zeta + \tau)^2}{4\zeta} - \log |\mathcal{V}|.$$

The choice of τ as $\tau^* \triangleq 2\sqrt{\zeta \log |\mathcal{V}|} - \zeta$ makes the log of the ratio 0. Then the bound on the failure probability is equal to that of the atypicality bound. Therefore, the bound on the total error probability is twice that of the atypicality bound.

Let $n \triangleq |M|$, $|\mathcal{U}| = n$, $|\mathcal{V}| = n^\nu$ for some $\nu \geq 1$, and $x \triangleq \frac{\zeta}{\log n}$. Then

$$\frac{\tau^*}{\log n} = 2\sqrt{\nu \cdot x} - x = \nu - (\sqrt{x} - \sqrt{\nu})^2.$$

Such τ^* satisfies $0 \leq \tau^* \leq \zeta$ (which is required for Lemma 8) if $\nu \leq x \leq 4\nu$.

There are n rows. Then, the bound on the expected number of atypicality errors is given by

$$n \exp\left(-\frac{(\zeta - \tau)^2}{4\zeta}\right) = n^{1-(\sqrt{\nu}-\sqrt{x})^2}.$$

This expression is bounded by $n^{1-\beta}$ if

$$x \geq \left(\sqrt{\nu} + \sqrt{\beta}\right)^2. \quad (3.3)$$

For such x , we get $\frac{\tau^*}{\log n} = 2\sqrt{\nu x} - x \leq \nu - \beta \leq x$. Then, τ^* satisfies $0 \leq \tau^* \leq \zeta$ if (3.3) is satisfied as well as $x \leq 4\nu$. (At $x \geq 4\nu$, we can pick $\beta \geq \nu$ and get a bound based on

the linear boundary whose proof was given above.)

If (3.3) is satisfied as well as $x \leq 4\nu$, the expected number of atypicality errors is bounded by $n^{1-\beta}$, so the total number of errors (including misalignment-despite-typicality errors) is bounded by $2n^{1-\beta}$.

This gives us the following inequalities that form part of the main results:

- Exact alignment is achieved if (3.3) is satisfied for some β such that $2n^{1-\beta} \leq o(1)$, which is equivalent to $\beta - 1 \geq \omega(1/\log n)$. (3.3) is satisfied for such β if $\zeta \geq (1 + \sqrt{\nu})^2 \log n + \omega(1)$.
- The number of errors is bounded by $2n^{1-\beta}$ if $\beta \leq \nu$ and $\zeta \geq (\sqrt{\nu} + \sqrt{\beta})^2 \log n$.

□

3.3.5 Proof for Maximum likelihood estimation

Linear boundary - relevant for large β and small α

We re-state the following results for maximum likelihood estimation, and introduce their proofs below:

- Theorem 3.3.1

Exact alignment is achieved if $\zeta \leq 2 \log n + \omega(1)$ for $s = 0$, or if $\zeta \geq 2(1 + \alpha) \log n + \omega(1)$ for $s \geq 1$.

- Theorem 3.3.2

The expected number of errors is bounded by $n^{1-\beta}(1 + o(1))$ if $\zeta \geq (1 + \beta) \log n$ and $\beta > 1 + \omega(1/\log n)$.

If $\zeta \geq 2 \log n + 2 \log \left(\frac{1+\sqrt{5}}{2} \right)$, then the expected number of errors is bounded by 1, which is bounded by $n^{1-\beta}$ for any $\beta \leq 1$.

- Theorem 3.3.3

The expected number of errors is bounded by $n^{1-\beta}(1 + o(1))$ if $\zeta \geq 2(\alpha + \beta) \log n$ and $\beta > 1 + \omega(1/\log n)$.

If $\zeta \geq 2 \log n + 2 \log \left(\frac{3+\sqrt{5}}{2} \right)$, then the expected number of errors is bounded by

$s + \sqrt{5} - 2$, which is bounded by $n^{1-\beta} + \mathcal{O}(1)$ for any $\beta \leq 1 - \alpha$.

• Theorem 3.3.4

Let $\varepsilon \triangleq \exp(\log n - \frac{\zeta}{2})$.

Given $x > \nu$ and $s = 0$, the expected number of errors is bounded by $\frac{\varepsilon^2}{1-\varepsilon}$.

In general, the number of errors is bounded by $\frac{\varepsilon^2}{1-\varepsilon} + \frac{s\varepsilon}{(1-\varepsilon)^2}$, which is bounded by the simpler expression $\frac{(s+1)\varepsilon}{(1-\varepsilon)^2}$.

Proof. Consider some elementary misalignment of size δ . (See Subsection 2.3.1 for elementary misalignments.) By Lemma 7, the probability of the given misalignment is at most $\exp(-\delta \cdot \frac{\zeta}{2})$.

Let $|\mathcal{U}| = n$ and $|\mathcal{V}| = n + s$, where $n \triangleq |M|$ is the size of the matching. By Lemma 2, there are at most $\frac{n^\delta}{\delta}$ different type-I misalignments and at most sn^δ different type-II misalignments of size δ . There are no misalignments of type-III. Furthermore, the number of type-I misalignments is 0 if $\delta = 1$. Define $\varepsilon \triangleq \exp(\log n - \frac{\zeta}{2})$. Then, the expected number of type-I and type-II misalignments of size δ are bounded by $\varepsilon^\delta/\delta$ and $s\varepsilon^\delta$ respectively.

The contribution of a misalignment of size δ is equal to δ . The total total contribution of all elementary misalignments gives us the expected number of errors. If $\varepsilon < 1$, then in expectation, this total contribution is bounded by

$$\sum_{\delta=2}^n \varepsilon^\delta + \sum_{\delta=1}^n s\delta\varepsilon^\delta \leq \frac{\varepsilon^2}{1-\varepsilon} + \frac{s\varepsilon}{(1-\varepsilon)^2}$$

Let us write $x \triangleq \frac{\zeta}{\log n}$ and $\alpha \triangleq \frac{\log s}{\log n}$ if $s \geq 1$. Then the expression above can be written as

$$\frac{n^{2-x}}{1-n^{1-x/2}} + \frac{n^{\alpha+1-x/2}}{(1-n^{1-x/2})^2}.$$

If $x \geq 2 + \frac{2\log(\frac{\sqrt{5}+1}{2})}{\log n}$ and $s = 0$, then this expression is bounded by 1. If $x \geq 2 + \frac{2\log(\frac{3+\sqrt{5}}{2})}{\log n}$ and $s \geq 1$, then the expression is bounded by $s + \sqrt{5} - 2$. In both of these cases, we have $x > 2$, which implies $\varepsilon \triangleq \exp(\log n - \frac{\zeta}{2}) < 1$.

The expected number of errors is bounded by $\frac{n^{1-\beta}}{(1-n^{(1-\beta)/2})^2}$

$$\text{if } s = 0 \text{ and } x \geq 1 + \beta \quad (3.4)$$

$$\text{or if } s \geq 1 \text{ and } x \geq 2(\alpha + \beta) \quad (3.5)$$

Given $\beta > 1$, both conditions imply $x > 2$, which implies $\varepsilon \triangleq \exp(\log n - \frac{\zeta}{2}) < 1$.

- Exact alignment is achieved if (3.4) or (3.5) is satisfied for some β such that $n^{1-\beta} \leq o(1)$, which is equivalent to $\beta - 1 \geq \omega(1/\log n)$. Such β exists if $\zeta \leq 2 \log n + \omega(1)$ for $s = 0$, or if $\zeta \geq 2(1 + \alpha) \log n + \omega(1)$ for $s \geq 1$.

- By (3.4), the expected number of errors is bounded by $n^{1-\beta}(1 + o(1))$ if $\zeta \geq (1 + \beta) \log n$ and $\beta > 1 + \omega(1/\log n)$. The expected number of errors is bounded by $\mathcal{O}(n^{1-\beta})$ if $\zeta \geq (1 + \beta) \log n$ and $\beta > 1 + \Omega(1/\log n)$.

If $\zeta \geq 2 \log n + 2 \log\left(\frac{1+\sqrt{5}}{2}\right)$, then the expected number of errors is bounded by 1, which is bounded by $n^{1-\beta}$ for any $\beta \leq 1$.

- By (3.5), the expected number of errors is bounded by $n^{1-\beta}(1 + o(1))$ if $\zeta \geq 2(\alpha + \beta) \log n$ and $\beta > 1 + \omega(1/\log n)$. The expected number of errors is bounded by $\mathcal{O}(n^{1-\beta})$ if $\zeta \geq 2(\alpha + \beta) \log n$ and $\beta > 1 + \Omega(1/\log n)$.

If $\zeta \geq 2 \log n + 2 \log\left(\frac{3+\sqrt{5}}{2}\right)$, then the expected number of errors is bounded by $s + \sqrt{5} - 2$, which is bounded by $n^{1-\beta} + \mathcal{O}(1)$ for any $\beta \leq 1 - \alpha$.

□

Elliptic boundary - relevant for smallest β and small α

We re-state the following results for maximum likelihood estimation, and introduce their proofs below:

- Theorem 3.3.1

If $\alpha < 1$ some constant bounded away from 1, almost-exact alignment is achieved if $\zeta \geq \log n + \omega(\sqrt{\log n})$.

- Theorem 3.3.2 and Theorem 3.3.3

Let $\alpha \triangleq \frac{\log(|\mathcal{V}|-n)}{\log n}$ and $\beta \in [0, 1/2)$ some constants and suppose $\beta < \min\{1-\alpha, 1/2\}$ and bounded away from both $1-\alpha$ and $1/2$.

The number of errors is bounded by $\exp\left(\frac{2(1+\log 2)+\mathcal{O}(1)}{(1-2\beta)^2}\right) \cdot (n^{1-\beta}) \leq \mathcal{O}(n^{1-\beta})$ if $\zeta \geq \left(1 + 2\sqrt{\beta(1-\beta)}\right) \log n$.

Proof. Let τ a threshold such that $0 \leq \tau \leq \zeta$. Consider a specific misalignment of size δ . Such a misalignment occurs if and only if one of the following is true:

- Atypicality event: the average information density scores of the δ true pairs is below τ , or
- Misalignment-despite-typicality event: the set of δ true pairs have average score greater than τ but nevertheless the set of δ false pairs have greater score than the corresponding set of true pairs.

Lemma 5, the probability of the true pairs having average score below the threshold is bounded by $\exp\left(-\delta \cdot \frac{(\zeta-\tau)^2}{4\zeta}\right)$. By Lemma 8, the probability of that the false pairs have score greater than the true pair despite the true pairs having high score is bounded by $\exp\left(-\delta \cdot \frac{\zeta^2+\tau^2}{2\zeta}\right)$.

Let $|\mathcal{U}| = n$ and $|\mathcal{V}| = n + s$, where $n \triangleq |M|$ is the size of the matching. Let $\alpha \triangleq \frac{\log s}{\log n}$ and let $\beta \in (0, 1/2)$ some number strictly less than $1-\alpha$. Such β exists as long as $\alpha < 1$. Define $\delta^* \triangleq n^{1-\beta}$. Since $\beta < 1-\alpha$, we have $\delta^* = n^{1-\beta} > n^\alpha = s$. Consider some $\delta \geq \delta^*$.

By Lemma 3, there are no more than $\exp(\delta(1 + \log n + \log 2))$ different misalignment-despite-typicality events of size δ . Then, the log of the expected number of such events is no more than $\delta \left((1 + \log n + \log 2) - \frac{\zeta^2+\tau^2}{2\zeta} \right)$. Pick τ to be

$$\tau^* = \sqrt{2\zeta(\log n - \log \varepsilon + 1 + \log 2) - \zeta^2}. \quad (3.6)$$

where ε is some variable to be set later on. Then, the expected misalignment-despite-typicality events of size δ is bounded by ε^δ . If $\varepsilon < 1$, then the bound on the number of such

events of size at least δ^* is bounded by

$$\sum_{\delta=\delta^*}^n \varepsilon^\delta \leq \frac{\varepsilon^{\delta^*}}{1-\varepsilon}.$$

If $\beta \leq 1/2$, then $\delta^* = n^{1-\beta} \geq \sqrt{n}$. If, furthermore, $-\log \varepsilon \geq \Omega(1)$, then $\frac{\varepsilon^{\delta^*}}{1-\varepsilon} = \frac{\exp(-\Omega(\sqrt{n}))}{1-\exp(-\Omega(1))} \leq e^{-\Omega(\sqrt{n})}$. Let ε some constant. A misalignment can result in no more than n errors. Then, the expected number of errors caused by misalignment-despite-typicality errors of size at least δ^* is $ne^{-\Omega(\sqrt{n})} \leq o(1)$.

Next we confirm the number of atypicality errors is small: There are no more than $\binom{n}{\delta}$ different ways to get an atypicality event. This is bounded by $\exp(\delta + \delta \log \frac{n}{\delta})$, which can further be bounded by $\exp(\delta + \delta \log \frac{n}{\delta^*})$. Then, in expectation, there are no more than $\exp\left(\delta + \delta \log \frac{n}{\delta^*} - \delta \frac{(\zeta - \tau^*)^2}{4\zeta}\right)$ atypicality events of size δ .

The log of the ratio of the bound on the expected number of atypicality errors versus the expected number of misalignment-despite-typicality errors is equal to

$$\begin{aligned} & \left[\delta + \delta \log \frac{n}{\delta^*} - \delta \frac{(\zeta - \tau^*)^2}{4\zeta} \right] - \left[\delta \left((1 + \log n + \log 2) - \frac{\zeta^2 + (\tau^*)^2}{2\zeta} \right) \right] \\ &= \delta \left(\frac{(\zeta + \tau^*)^2}{4\zeta} - \log \delta^* - 2 \right). \end{aligned}$$

If

$$\tau^* \leq 2\sqrt{\zeta(\log \delta^* + \log 2)} - \zeta, \quad (3.7)$$

then the log-ratio is greater than 0 and the bound on the expected number of atypicality events of size δ is bounded by that of misalignment-despite-typicality events of size δ .

We identify the smallest δ^* (i.e. the largest β) such that our choice of $\tau = \tau^*$ in (3.6) satisfies the inequality in (3.7): Let us write $x \triangleq \frac{\zeta}{\log n}$. Given some $b \in [0, 1/2)$, let

$x = 1 + 2\sqrt{b(1-b)}$. Define $\eta \triangleq 1 - \log \varepsilon + \log 2$. Then, (3.6) gives us

$$\begin{aligned} \frac{\tau^*}{\log n} &= \sqrt{2x \left(1 + \frac{\eta}{\log n}\right) - x^2} \\ &= \sqrt{1 - 4b + 4b^2 + 2x \cdot \frac{\eta}{\log n}} \\ &= (1 - 2b) \cdot \sqrt{1 + \frac{2x \cdot \eta}{(1 - 2b)^2 \log n}}. \end{aligned}$$

For $x = 1 + 2\sqrt{b(1-b)}$, (3.7) is equivalent to

$$\begin{aligned} \frac{\tau^*}{\log n} &\leq 2\sqrt{x \left(1 - \beta + \frac{\log 2}{\log n}\right) - x} \\ &= 2\sqrt{x(1-b) + x \left(b - \beta + \frac{\log 2}{\log n}\right) - x} \\ &= 2\sqrt{\left(1 - b + \sqrt{b(1-b)}\right)^2 + x \left(b - \beta + \frac{\log 2}{\log n}\right) - x} \\ &= 2 \left(1 - b + \sqrt{b(1-b)}\right) \sqrt{1 + \frac{x(b - \beta) + x \log 2 / \log n}{\left(1 - b + \sqrt{b(1-b)}\right)^2} - 1} - 2\sqrt{b(1-b)} \\ &= 1 - 2b + 2 \left(1 - b + \sqrt{b(1-b)}\right) \left(-1 + \sqrt{1 + \frac{x(b - \beta) + x \log 2 / \log n}{\left(1 - b + \sqrt{b(1-b)}\right)^2}}\right) \end{aligned}$$

We can simplify this last expression to get a more compact inequality that implies (3.7):

$$\frac{\tau^*}{\log n} \leq 1 - 2b + (1 - 2b) \left(-1 + \sqrt{1 + \frac{x(b - \beta)}{\left(1 - b + \sqrt{b(1-b)}\right)^2}}\right)$$

It then follows that, given $x = 1 + 2\sqrt{b(1-b)}$, (3.6) satisfies (3.7) if

$$(1 - 2b) \cdot \sqrt{1 + \frac{2x \cdot \eta}{(1 - 2b)^2 \log n}} \leq 1 - 2b + (1 - 2b) \left(-1 + \sqrt{1 + \frac{x(b - \beta)}{\left(1 - b + \sqrt{b(1-b)}\right)^2}}\right)$$

or equivalently if

$$\begin{aligned} \frac{2\eta/\log n}{(1-2b)^2} &\leq \frac{(b-\beta)}{(1-b+\sqrt{b(1-b)})^2} \\ \iff b-\beta &\geq \frac{2\eta}{\log n} \cdot \frac{(1-b+\sqrt{b(1-b)})^2}{(1-2b)^2} \\ &\geq \frac{2\eta/\log n}{(1-2b)^2}. \end{aligned}$$

Then, we can conclude that (3.6) satisfies (3.7) for $x = 1 + 2\sqrt{b(1-b)}$ if b and β such that $\beta \leq b - \frac{2\eta/\log n}{(1-2b)^2}$. For any finite $b \in [0, 1/2)$ bounded away from $1/2$, this inequality can be simplified to write $\beta \leq b \leq \Omega(1/\log n)$.

We have shown that errors from misalignments of size greater than δ^* is $o(1)$: Those due to misalignment-despite-typicality type errors is $o(1)$ and those due to atypicality type errors is bounded by that of misalignment-despite-typicality type errors, so also $o(1)$. Finally, the expected number of errors due to misalignments smaller than δ^* is at most δ^* . (Given some misalignment m' of size δ , i.e. m' such that $d(m', M) = \delta$, the event $\hat{M} = m'$ results in exactly δ errors. Then, the total contribution of all misalignments of size at most δ^* to the expected total number of errors is no more than δ^* .) This gives the total expected number of errors as $\delta^* + o(1)$.

- Almost-exact alignment is achieved if $x \geq 1 + 2\sqrt{b(1-b)}$ is satisfied for some b such that $\beta = b - \frac{2\eta/\log n}{(1-2b)^2}$ satisfies $n^{-\beta} \leq o(1)$, which is equivalent to $\beta \geq \omega(1/\log n)$. Such β exists if $\zeta \geq \log n + \omega(\sqrt{\log n})$. The condition $\beta \leq 1 - \alpha$ implies $\alpha \leq 1 - \omega(1/\log n)$.
- The number of errors is bounded by $n^{1-b+\frac{2\eta/\log n}{(1-2b)^2}} = n^{1-b} \cdot \exp\left(\frac{2\eta}{(1-2b)^2}\right)$ if $\zeta \geq \left(1 + 2\sqrt{b(1-b)}\right) \log n$. Since $\varepsilon < 1$ some arbitrary constant bounded away from 1, replacing η with the explicit expression $\eta \triangleq 1 + \log 2 - \varepsilon$, we can write the bound as $\exp\left(\frac{2(1+\log 2)+\mathcal{O}(1)}{(1-2b)^2}\right) n^{1-b}$. If $b \in [0, 1/2)$ some constant bounded away from $1/2$, then this bound is $\mathcal{O}(n^{1-\beta})$. We also require $b \leq 1 - \alpha + \frac{2\eta/\log n}{(1-2b)^2} \leq 1 - \alpha + \Omega(1/\log n)$ as per our assumption.

□

Parabolic boundary - relevant for small β and large α

We re-state the following results for maximum likelihood estimation, and introduce their proofs below:

- Theorem 3.3.1

Suppose $\alpha \geq 1$.

Almost-exact alignment is achieved if $\beta \geq \alpha \log n + \omega(\sqrt{\log n})$.

Exact alignment is achieved if $\zeta \geq (1 + \sqrt{\alpha})^2 \log n + \omega(1)$.

- Theorem 3.3.3

Given $1 - \alpha < \beta \leq \alpha$, the number of errors is bounded by $3e \cdot (n^{1-\beta}) \leq \mathcal{O}(n^{1-\beta})$ if $\zeta \geq (\sqrt{\alpha} + \sqrt{\beta})^2 \log n$.

Proof. Let $x \triangleq \frac{\log \zeta}{\log n}$ and $\alpha \triangleq \frac{\log s}{\log n}$ as in the previous. Define $\tilde{\beta} \triangleq 1 - \alpha$. If $\alpha < 1 \iff \tilde{\beta} > 0$, then, by the previous proof, we know that the expected number of errors due to misalignments of size at least s is $o(1)$ if $\zeta \geq \left(1 + 2\sqrt{\tilde{\beta}(1 - \tilde{\beta})}\right) \log n$.

Here we show that, given a stronger bound on ζ , we can also bound the number of errors due to misalignments of size between δ^* and s for some $\delta^* = n^{1-\beta+\eta} \leq 2s$ where $\beta \in [0, 1/2]$ another constant strictly greater than $1 - \alpha$ and η some non-negative function of n which is to be determined later.

Let τ a threshold such that $0 \leq \tau \leq \zeta$. Consider a specific misalignment of size δ . Once again, we cover the misalignment event using two auxiliary events:

- Atypicality event: the average information density scores of the δ true pairs is below τ . By Lemma 5, the probability of this event for size δ is bounded by $\exp\left(-\delta \frac{(\zeta - \tau)^2}{4\zeta}\right)$.
- Misalignment-despite-Typicality event: the set of δ true pairs have average score greater than τ but nevertheless the set of δ false pairs have greater score than the corresponding set of true pairs. By Lemma 8, the probability of that the false pairs have score greater than the true pair despite the true pairs having high score is bounded by

$$\exp\left(-\delta \cdot \frac{\zeta^2 + \tau^2}{2\zeta}\right).$$

Let $|\mathcal{U}| = n$ and $|\mathcal{V}| = n + s$, where $n \triangleq |M|$ is the size of the matching. The number of atypicality events of size δ is bounded by $\exp\left(\delta + \delta \log \frac{n}{\delta}\right)$. By Lemma 3, the number of misalignment-despite-typicality events of size δ (which is between δ^* and $2s$) is bounded by $\exp\left(\delta\left(1 + \log \frac{ns}{\delta} + \log 3\right)\right)$.

The log of the ratio of the bounds on the expected number of atypicality events and the expected number of misalignment-despite-typicality events is equal to $\delta\left(\frac{(\zeta + \tau)^2}{4\zeta} - \log s - \log 3\right)$. If $\tau = \tau^* \triangleq 2\sqrt{\zeta(\log s + \log 3)} - \zeta$, or equivalently if

$$\frac{\tau}{\log n} = 2\sqrt{x(\alpha + \log 3/\log n)} - x, \quad (3.8)$$

then the log-ratio is 0 and the bound on the expected number of atypicality events of size δ is equal to that of misalignment-despite-typicality events of size δ . Such τ^* satisfies $0 \leq \tau \leq \zeta$ if $\alpha + \log 3/\log n \leq x \leq 4(\alpha + \log 3/\log n)$.

Let's plug in the variables defined as $x \triangleq \frac{\zeta}{\log n}$ and $\alpha \triangleq \frac{\log s}{\log n}$. The expected number of misalignment-despite-typicality errors of size δ for some size greater δ^* and less than $2s$ is bounded by

$$\exp\left(\delta(1 + \log 3) + \delta \log \frac{ns}{\delta} - \delta \cdot \frac{\zeta^2 + (\tau^*)^2}{2\zeta}\right),$$

which, given the value of τ in (3.8), is further bounded by

$$\exp\left(\delta(1 + \log 3) + \delta \log \frac{ns}{\delta^*} - \delta \cdot \frac{\zeta^2 + (\tau^*)^2}{2\zeta}\right) = n^{\delta(\beta - \eta - (\sqrt{x} - \sqrt{\alpha})^2)} e^{\delta(1 + \log 3)}.$$

Pick η to be $\frac{1 + \log 3}{\log n}$. Then the expression in the previous bound simplifies as $n^{\delta(\beta - (\sqrt{x} - \sqrt{\alpha})^2)}$. Define $\varepsilon \triangleq n^{(\beta - (\sqrt{x} - \sqrt{\alpha})^2)}$. $\varepsilon < 1$ if $x > (\sqrt{\alpha} + \sqrt{\beta})^2$ and $\alpha \geq \beta$. The contribution of each event to the size of the misalignment is at most δ . (This is because at most one misalignment can occur at a time. So the contribution of a misalignment-despite-typicality

event is either 0 or δ .) Then, the total number of errors due to misalignment-despite-typicality events is bounded as

$$\sum_{\delta=0}^{\infty} \delta \varepsilon^{\delta} = \frac{\varepsilon}{(1-\varepsilon)^2}.$$

The bound above is $o(1)$ if $\varepsilon \leq o(1)$. In that case, the expected number of errors due to all possible misalignments of size $\delta \in [n^{1-\beta+\eta}, 2s]$ is $o(1)$. If $s \geq n \iff \alpha \geq 1$, then this covers all misalignments of size greater than $n^{1-\beta+\eta}$. Otherwise, if $\alpha < 1$, from the previous proof we know that the expected number of errors due to all possible misalignments of size $\delta \in [n^{1-\tilde{\beta}}, n]$ is $o(1)$. The expected number of errors due to misalignments smaller than $n^{1-\beta+\eta}$ is at most $n^{1-\beta+\eta}$. (This follows from the fact that only one of the misalignments can occur.)

$\varepsilon = n^{(\beta - (\sqrt{x} - \sqrt{\alpha})^2)} \leq o(1)$ is equivalent to $x > (\sqrt{\alpha} + \sqrt{\beta + \omega(1/\log n)})^2$, which we can rewrite as

$$x \geq (\sqrt{\alpha} + \sqrt{\beta})^2 + \left(1 + \sqrt{\frac{\alpha}{\beta}}\right) \omega(1/\log n) \quad (3.9)$$

Replacing the last term $\left(1 + \sqrt{\frac{\alpha}{\beta}}\right) \omega(1/\log n)$ by $\left(1 + \sqrt{\frac{\alpha}{\beta}}\right) \Omega(1/\log n)$ changes the bound on the expected number of errors due to all misalignments of size $\delta \in [n^{1-\beta+\eta}, n^{1-\tilde{\beta}}]$ from $o(1)$ to $\mathcal{O}(1)$.

This gives us the following inequality that form part of the main results:

- Suppose $\alpha \geq 1$. Our choice for $\eta = \frac{1+\log 3}{\log n}$ is on the order of $1/\log n$.

Almost-exact alignment is achieved if (3.9) is satisfied for some β such that $n^{1-\beta+\eta} \leq o(n)$, which is equivalent to $\beta \geq \eta + \omega(1/\log n)$. Such β exists if $\zeta \geq \nu \log n + \omega(\sqrt{\log n})$.

Exact alignment is achieved if (3.9) is satisfied for some β such that $n^{1-\beta+\eta} \leq o(1)$, which is equivalent to $\beta \geq 1 + \eta + \omega(1/\log n)$. Such β exists if $\zeta \geq (1 + \sqrt{\alpha})^2 \log n +$

$\omega(1)$.

- The number of errors is bounded by $n^{1-\beta+n} = 3e \cdot (n^{1-\beta})$ if $\zeta \geq (\sqrt{\alpha} + \sqrt{\beta})^2 \log n$.

We also have the initial assumption that $\beta < 1 - \alpha$, as well as the condition $\alpha + \log 3 \leq x \leq 4(\alpha + \log 3)$. $\alpha + \log 3 \leq x$ is weaker than $\zeta \geq (\sqrt{\alpha} + \sqrt{\beta})^2 \log n$ while $x \leq 4(\alpha + \log 3)$ is implied by $\beta \leq \alpha$.

□

3.4 Converse

3.4.1 Converse statement

We present converse results that match the achievability result in Theorem 3.3.3 in multiple regimes. Statements are given in terms of signal strength $\zeta \triangleq \mu^2/2$ to match the form of the expressions in Section 3.3.

Theorem 3.4.1 ([37]). *Let $n = |M| = |\mathcal{U}|$ and $\alpha = \frac{\log(|\mathcal{V}|-n)}{\log n}$ if $|\mathcal{V}| > n$. As $n \rightarrow \infty$, conditions of the form $\zeta \leq c \log n - \omega(\log \log n)$ guarantee that, for any estimator, $d(\hat{M}, M) \geq \Omega(n^{1-\beta})$ with probability $1 - o(1)$. The appropriate values of c for different ranges of β are given in Table 3.6 for different values of β .*

Table 3.6: Necessary conditions for error bound for Gaussian planted matching. $\zeta \leq c \log n - \omega(\log \log n)$ guarantees that $d(\hat{M}, M) \geq \Omega(n^{1-\beta})$ with probability $1 - o(1)$ for any algorithm. The conditions given by these three inequalities match the achievability bounds in Theorem 3.3.3 (Table 3.4) for all $0 < \beta \leq \max\{\alpha, 1 - \alpha\}$.

Value of c in necessary cond.	Range of β where converse holds	Range of β where converse tight
$1 + 2\sqrt{\beta(1-\beta)}$	$0 < \beta \leq \frac{1}{2}$	$0 \leq \beta \leq \min\{1 - \alpha, \frac{1}{2}\}$
$(\sqrt{\alpha} + \sqrt{\beta})^2$	$0 < \beta \leq 1$	$\max\{1 - \alpha, 0\} < \beta \leq \min\{\alpha, 1\}$
2	$\frac{1}{2} \leq \beta \leq 1$	$1/2 < \beta \leq 1 - \alpha$

The maximum likelihood estimator is also the maximum a posteriori estimator and thus is the optimal estimator for exact recovery. It is not necessarily optimal for partial recover: it does not necessarily maximize the probability that it makes at most $n^{1-\beta}$ errors for $\beta < 1$.

However, this converse shows that the maximum likelihood estimator is asymptotically optimal for partial recovery: the conditions needed to ensure partial correctness match the converse conditions in the leading term.

3.5 Interpretation and intuition behind results

We present the intuition behind the various phase transitions of the achievability boundaries given by Theorems 3.3.2 and 3.3.3 as illustrated in Figures 3.3 and 3.4. All explanations in this subsection are also applicable to the matching achievability boundaries given by Theorems 4.4.2 and 4.4.3 for the Gaussian database alignment problem in the low per-feature correlation regime in Chapter 4.

3.5.1 Merging of boundary for maximum likelihood estimation and maximum row estimation

As shown by Theorem 3.3.3 and Fig. 3.4, the boundaries for the achievability regions of maximum likelihood estimation and maximum row estimation fully coincide if the number of unmatched users in \mathcal{V} is on the order of n or greater. As shown by Theorem 3.3.1 and Fig. 3.2, this is also true for the almost-exact alignment threshold.

Maximum row estimation corresponds to a relaxation of the maximum likelihood estimation by removing the constraint that every vertex in \mathcal{V} can have at most one match in \mathcal{U} . That is, when looking for the true mapping matrix $\mathbf{M} \in \{0, 1\}^{\mathcal{U} \times \mathcal{V}}$, maximum likelihood estimation still only accepts a single non-zero entry on each row, but ignores the number of entries on each column.

For $\alpha < 1$ bounded away from 1, the average non-zero entries in each column is $\frac{n}{n+n^\alpha} \geq 1 - o(1)$. So the column constraint of maximum likelihood estimation is tight for almost every column. Therefore, for such α , the column constraint is relevant and its removal results in the introduction of a significant increase in the expected number of errors. This creates a gap between the boundaries of maximum likelihood estimation and maximum row estimation.

On the other hand, for $\alpha > 1$ bounded away from 1, the average non-zero entries in each column is $\frac{n}{n+n^\alpha} \leq o(1)$. So the column constraint is loose for almost every column. Then, for such α , relaxing the column constraint results in no significant loss in performance and the boundaries for the two algorithms coincide.

3.5.2 Transition from the elliptic to the quadratic boundary for maximum likelihood estimation

By Theorem 3.3.3, given $\alpha > 1/2$, there is a phase transition in the achievability boundary for maximum likelihood estimation as the bound on the expected number of errors goes beyond n^α .

Recall that $n = |M| = |\mathcal{U}|$, and $n^\alpha = |\mathcal{V}| - n$. Consider a simplified model to generate an estimated mapping m : In step 1, make an independent decision for every true pair whether to include it in m or not. Each pair is failed to be included with probability $\frac{n^{1-\beta}}{n}$. In step 2, randomly assign each of the users in \mathcal{U} that haven't been mapped in step 1, to a random user in \mathcal{V} that also hasn't been mapped in step 1. In expectation, there are $n^{1-\beta}$ and $n^{1-\beta} + n^\alpha$ users that haven't been mapped in step 1 in \mathcal{U} and \mathcal{V} respectively.

Based on this process, every falsely-paired user in \mathcal{V} will be mapped to about $\frac{n^{1-\beta}}{n^{1-\beta}+n^\alpha} = (1 + n^{\alpha+\beta-1})^{-1}$ users in expectation. If $\alpha + \beta < 1$ bounded away from 1, then this value is $1 - o(1)$. So the restriction on having each user from \mathcal{V} mapped to at most 1 user is tight for almost all users in \mathcal{V} . This implies that, at that point, this constraint is still able to contribute to the elimination of misalignments: An increase in signal strength leads to users previously not paired in step 1 to be correctly paired, which in turn would trigger the mentioned constraint, leading to chain reactions that might fix even more misaligned pairs.

For $\alpha + \beta > 1$ bounded away from 1, the average number of users mapped to the falsely-paired users in \mathcal{V} is $o(1)$. In this regime, the constraint is loose for almost every falsely-paired user in \mathcal{V} , and the restriction does not help in eliminating misalignments.

The fact that this restriction stops being relevant around the point $\beta = 1 - \alpha$ demon-

strates itself on the boundary as an immediate decrease in the absolute value of the slope. At this point, increasing mutual information does not decrease the error exponent as quickly, as the algorithm can no longer leverage the constraint on the number of pairs of users from \mathcal{V} .

Note that, for $\alpha < 1$ bounded away from 1, the gap between maximum likelihood estimation and maximum row estimation still persists whether $\alpha + \beta < 1$ or $\alpha + \beta > 1$. This is due to the fact the correct pairing of the $n - n^{1-\beta}$ users in step 1 does rely on the column constraint.

3.5.3 Transition between quadratic boundary to linear boundary of maximum likelihood estimation and maximum row estimation

When a true pair has a sufficiently low score in \mathbf{G} or \mathbf{W} , the expected number of errors involving that pair is larger than 1. Call this a “bad true pair.” This phase transition between the parabolic boundary and the linear one corresponds to a change in the importance of the bad true pairs. In the parabolic region, most errors involve a bad true pair. In the linear region, most errors involve true pairs that are not bad. See the discussion in Appendix 2.3.1 of the cycle-path decomposition of a pair of matchings for the precise interpretation of “most errors” in these statements.

3.5.4 Halving of slope of linear boundary of maximum likelihood estimation

As shown by Theorem 3.3.3, the slope of the linear boundary for maximum likelihood estimation is halved when going from $|\mathcal{V}| = n$ for $|\mathcal{V}| > n$. This is illustrated by the last two curves in Fig. 3.4.

For both cases, the linear region boundary is only relevant for β large, i.e. when the error exponent is very small. In this regime, the misalignment of any pair is very rare. For $|\mathcal{V}| = n$, a misalignment requires at least two pairs of users. Specifically, we need some pairs (u_0, v_0) and (u_1, v_1) such that, (u_0, v_1) and (u_1, v_0) jointly make a better pairing.

This requires the occurrence of two relatively low score pairs as well as the corresponding misalignment to have two relatively high score false pairs. As such, this error event requires many unlikely things to coincide, making its likelihood very small as we increase signal strength beyond the point $\zeta = 2 \log n$.

On the other hand, when $|\mathcal{V}| \geq n + 1$, there is at least one user, say $v' \in \mathcal{V}$, that already has no pair. Therefore a misalignment may consist of a single misaligned pair, e.g. (u_0, v') instead of (u_0, v_0) . This requires the occurrence of a single relatively low score pair as well as a relatively high score misalignment of the first pair with an with unpaired user, of which there are $|\mathcal{V}| - n$. The inclusion of even a single extra therefore makes the unlikely misalignment event somewhat less exceptional.

3.5.5 Vertical segment of boundary of maximum likelihood estimation

This phase transition involves a shift in the structure of the typical error. A misalignment can be composed of errors that involve any number of users, as described in Appendix 2.3.1. In the balanced case all errors come from cycles. When $\zeta = 2 \log n$, the expected contributions of each cycle length to the number of errors are equal. Each of these contributions has a linear dependence on ζ with a slope proportional to the cycle length. To the right of $\zeta = 2 \log n$, the contribution of the 2-cycle dominates the contributions of others that go down with a steeper descent. To the left of that point, the slopes of long cycles can be arbitrarily steep, dominating the contributions of shorter cycles. This gives us the vertical boundary as we move to the left of $\zeta = 2 \log n$. The top of the vertical boundary occurs due to an effect similar to that described in Subsection 3.5.3. In the elliptical boundary region, most long-cycle errors involve a bad true pair, so the number of bad true pairs controls the overall number of errors.

3.5.6 Gap between maximum row estimation and threshold testing

The difference between the maximum row estimators and the threshold testing estimator is constraint (b) in the linear program, which ensures that the estimated matrix has exactly one 1 in each row. This constraint can be included because of our assumption that $|M| = |\mathcal{U}|$, i.e. that every user in the first database has a match in the second database. The gap between the performance of the maximum row and threshold testing estimators means that this constraint corrects most errors in the threshold testing estimator. If we have a sufficiently large gap between $|M|$ and $|\mathcal{U}|$, we would see the performance of these two estimators converge, just as the ML and maximum row estimators converge in performance when $|\mathcal{V}|$ is sufficiently larger than $|\mathcal{U}|$.

CHAPTER 4

GAUSSIAN DATABASE ALIGNMENT IN THE LOW PER-FEATURE CORRELATION LIMIT

We study the Gaussian database alignment problem under the setting where features consist of a very large number of dimensions, each with with a very small contribution to the mutual information between features. This setting is of particular interest: The entries of the information density matrix \mathbf{G} , as defined in Subsection 2.2.1, become Gaussian, with negligible correlation between pairs of entries, as shown by Lemma 50. This results in a distribution closely related to that of the adjacency matrix \mathbf{W} of the Gaussian bigraph with a planted matching which we studied in Chapter 3.

The exact description of this setting, which we refer to as the low per-feature correlation regime, is given in Section 4.1. In Section 4.2, we study the distribution of \mathbf{G} through its generating function. In Section 4.3, we use the results on the generating function to derive useful concentration inequalities, which closely mirror those for Gaussian planted matching in Section 3.2. In Section 3.3, we present asymptotic and finite-size achievability bounds. Finally, in Section 4.5, we present an asymptotic converse bound for exact alignment that matches the achievability result in the previous section.

More limited versions of our results in Section 3.3 on the conditions for exact alignment by maximum likelihood estimation and almost exact alignment by threshold testing (Theorem 4.4.1) and in Section 4.5 (Theorem 4.5.1) have been presented at AISTATS 2019 [38]. Our result in Section 3.3 for maximum likelihood estimation in the almost-exact alignment regime in the balanced case (Theorem 4.4.2) has been presented at ISIT 2020 [39].

Refer to Section 3.2 for a sketch of our analysis.

4.1 Low Per-Feature Correlation Regime

For database alignment, when per-feature correlation is low, measures of correlation relevant for our analysis can always be expressed in terms of mutual information I_{XY} . Therefore the statements of some of our results are only accurate in this setting. It is also of interest that under the low per-feature correlation regime, the first and second order statistics of the information density matrix \mathbf{G} , as defined in Subsection 2.2.1, converges to that of the Gaussian bigraph with a planted matching: Specifically, as shown by Lemma 50, under the low per-feature correlation regime (as defined in Condition 4 below), entries corresponding to true and false pairs asymptotically have mean I_{XY} and $-I_{XY}$ respectively. All entries asymptotically have variance $2I_{XY}$ and the correlation between any pair of entries is asymptotically 0. Given $\zeta \triangleq I_{XY}$, then \mathbf{G} has entries with mean ζ and $-\zeta$, variance 2ζ and asymptotically zero correlation between entries. This exactly matches the statistics of the Gaussian matrix \mathbf{W}_G as described in Subsection 2.2.2 when we replace ζ by $\mu^2/2$. Under this regime, we are able to derive results for Gaussian database alignment that closely mirror those for Gaussian planted matching.

We formally define this regime, where a large number of dimensions each carry infinitesimally small information:

Condition 4 (Low per-feature correlation in database alignment). *The covariance matrix $\Sigma = \begin{bmatrix} \Sigma_a & \Sigma_{ab} \\ \Sigma_{ab}^\top & \Sigma_b \end{bmatrix}$ is said to satisfy the low per-feature correlation condition if $\rho_{\max} \triangleq \left\| \begin{bmatrix} \Sigma_a^{-1/2} & \Sigma_{ab} \Sigma_b^{-1/2} \end{bmatrix} \right\|_2 \leq o(1)$, where $\|\cdot\|_2$ denotes the ℓ_2 operator norm, i.e. largest singular value.*

By Lemma 50, under Condition 4 (i.e. in the low per-feature correlation regime), the correlation between a pair of entries of \mathbf{G} is $o(I_{XY})$.

Condition 5 (Very low per-feature correlation in database alignment). *The covariance*

matrix $\Sigma = \begin{bmatrix} \Sigma_a & \Sigma_{ab} \\ \Sigma_{ab}^\top & \Sigma_b \end{bmatrix}$ is said to satisfy the low per-feature correlation condition if $\rho_{\max} \triangleq \left\| \Sigma_a^{-1/2} \Sigma_{ab} \Sigma_b^{-1/2} \right\|_2 \leq o(1/\log n)$, where n is the size of the true mapping M and $\|\cdot\|_2$ denotes the ℓ_2 operator norm, i.e. largest singular value.

By Lemma 50, under Condition 4 (i.e. in the low per-feature correlation regime), the correlation between a pair of entries of \mathbf{G} is $o(1)$.

Lemma 9 shows the significance of $\left\| \Sigma_a^{-1/2} \Sigma_{ab} \Sigma_b^{-1/2} \right\|_2$, which is used to characterize Condition 4.

Lemma 9. Let $\Sigma = \begin{bmatrix} \Sigma_a & \Sigma_{ab} \\ \Sigma_{ab}^\top & \Sigma_b \end{bmatrix}$ be the covariance matrix between pairs of correlated feature vectors and let and the $\vec{\rho} \in (-1, 1)^{\mathcal{D}}$ correlation vector that characterizes the correlation in canonical form. Then

$$\left\| \Sigma_a^{-1/2} \Sigma_{ab} \Sigma_b^{-1/2} \right\|_2 = \max_{i \in \mathcal{D}} |\rho_i|,$$

where $\|\cdot\|_2$ denotes the ℓ_2 operator norm, i.e. largest singular value.

Proof. Let $\mathbf{P}, \mathbf{L}_a, \mathbf{L}_b, \mathbf{U}, \mathbf{V}$ be as defined in the proof of Lemma 4. Specifically, let \mathbf{L}_a and \mathbf{L}_b be triangular matrices such that $\mathbf{L}_a \mathbf{L}_a^\top = \Sigma_a$ and $\mathbf{L}_b \mathbf{L}_b^\top = \Sigma_b$, and \mathbf{P} diagonal and \mathbf{U}, \mathbf{V} orthonormal matrices such that $\mathbf{U} \mathbf{P} \mathbf{V}^\top = \mathbf{L}_a^{-1} \Sigma_{ab} (\mathbf{L}_b^\top)^{-1}$.

Define $\mathbf{A}_0 \triangleq \Sigma_a^{-1/2} \Sigma_{ab} \Sigma_b^{-1/2}$, $\mathbf{A}_1 \triangleq \Sigma_a^{-1/2} \Sigma_{ab} (\mathbf{L}_b^\top)^{-1} \mathbf{V}$ and $\mathbf{A}_2 \triangleq \mathbf{U}^\top \mathbf{L}_a^{-1} \Sigma_{ab} (\mathbf{L}_b^\top)^{-1} \mathbf{V}$.

First we show that \mathbf{A}_2 has the same singular values as \mathbf{A}_0 : The singular values of some matrix \mathbf{A} can be found by finding the eigenvalues of $\mathbf{A}^\top \mathbf{A}$, or those of $\mathbf{A} \mathbf{A}^\top$.

- Since $\mathbf{L}_b \mathbf{L}_b^\top = \Sigma_b$ and $\mathbf{V} \mathbf{V}^\top = \mathbf{I}$, it follows that $\mathbf{A}_0 \mathbf{A}_0^\top = \mathbf{A}_1 \mathbf{A}_1^\top$.

Then, \mathbf{A}_0 and \mathbf{A}_1 must have the same singular values.

- Since $\mathbf{L}_a \mathbf{L}_a^\top = \Sigma_a$ and $\mathbf{U} \mathbf{U}^\top = \mathbf{I}$, it follows that $\mathbf{A}_1^\top \mathbf{A}_1 = \mathbf{A}_2^\top \mathbf{A}_2$.

Then, \mathbf{A}_1 and \mathbf{A}_2 must have the same singular values.

Then \mathbf{A}_0 and \mathbf{A}_2 have the same singular values.

In the proof of Lemma 4, we are given that $\mathbf{U}\mathbf{P}\mathbf{V}^\top = \mathbf{L}_a^{-1}\boldsymbol{\Sigma}_{ab}(\mathbf{L}_b^\top)^{-1}$. Then, by the orthonormality of \mathbf{U} and \mathbf{V} , we have $\mathbf{P} = \mathbf{U}^\top \mathbf{L}_a^{-1}\boldsymbol{\Sigma}_{ab}(\mathbf{L}_b^\top)^{-1}\mathbf{V} = \mathbf{A}_2$.

Since $\mathbf{P} = \text{diag}(\vec{\rho})$ is a diagonal matrix, its singular values are simply its diagonal entries in absolute value. Then the largest singular value is $\max |\rho_i|$ \square

Mutual information $I_{XY} \triangleq I(\vec{A}(u); \vec{B}(v) | u \stackrel{M}{\sim} v)$ between pairs of correlated features is equal to $I_{XY} = -\frac{1}{2} \sum_{i \in \mathcal{D}} \log(1 - \rho_i^2)$. Lemma 9 shows that, under Condition 4, $\rho_i^2 \leq o(1)$ for all $i \in \mathcal{D}$. It then follows that $I_{XY} = \frac{1}{2} \sum \rho_i^2(1 + o(1)) \leq \frac{1}{2} |\mathcal{D}| \rho_{max}^2$. $I_{XY} \geq \Omega(\log n)$ in the regime where alignment is feasible. (n is the size of the matching.) Then Condition 4 implies $|\mathcal{D}| \geq \omega(\log n)$, i.e. in the low per-feature correlation limit alignment is feasible only if the number of dimensions is $\omega(\log n)$. Similarly, Then Condition 4 implies $|\mathcal{D}| \geq \omega(\log^2 n)$.

4.2 Generating function for information density matrix

In this section we study the distribution of \mathbf{G} through its generating function, as defined in Definition 6. This function allows us to directly obtain concentration inequalities, as shown in Lemma 11.

In Subsection 4.2.1, we present a more general study of the generating function, while in Subsection 4.2.2 we focus specifically on elementary misalignments as defined in Subsection 2.3.1.

Definition 6 (Generating function for information density matrix). *The generating function $R = R^{\mathcal{U}, \mathcal{V}} : \mathbb{R}^{\mathcal{U} \times \mathcal{V}} \rightarrow \mathbb{R}$ is defined such that*

$$R(\boldsymbol{\Theta}) \triangleq \int \int \exp(\langle \mathbf{G}, \boldsymbol{\Theta} \rangle) f_{\vec{A}}(\vec{a}) f_{\vec{B}}(\vec{b}) d\vec{a} d\vec{b}$$

where $f_{\vec{A}}, f_{\vec{B}}$ denote the marginal probabilities for the two databases and $\mathbf{G} \in \mathbb{R}^{\mathcal{A} \times \mathcal{B}}$ denotes the information density matrix for \vec{a}, \vec{b} as defined in Section 2.2.1.

Lemma 10. $\mathbb{E}[\exp(\langle \mathbf{G}, \Theta \rangle) | M = m_1] = R(\Theta + \mathbf{m}_1)$

Proof. The key equality of the proof is that $\langle \mathbf{G}, \mathbf{M} \rangle = \log f_{\vec{A}, \vec{B}|M}(\vec{a}, \vec{b}) - \log f_{\vec{A}}(\vec{a}) - \log f_{\vec{B}}(\vec{b})$, where $f_{\vec{A}, \vec{B}|M}$ is the joint distribution between databases given M . We show as follows:

Let $\mathcal{U}_M \subseteq \mathcal{U}$ and $\mathcal{V}_M \subseteq \mathcal{V}$ denote the set of users that have a mapping under M and $\mathcal{W}_M \subset \mathcal{U}_M \times \mathcal{V}_M$ denote the set of pairs mapped by M . By the model for Gaussian data structures (as given in Subsection 2.1.3), all matched feature pairs and unmatched features are mutually independent. It then follows that

$$\begin{aligned} \log f_{\vec{A}, \vec{B}|M}(\vec{a}, \vec{b}) &= \sum_{(u,v) \in \mathcal{W}_M} \log f_{XY}(\vec{A}(u), \vec{B}(v)) \\ &\quad + \sum_{u' \in \mathcal{U} \setminus \mathcal{U}_M} \log f_X(\vec{A}(u')) + \sum_{v' \in \mathcal{V} \setminus \mathcal{V}_M} \log f_Y(\vec{A}(v')) \end{aligned}$$

where f_{XY} denotes the joint distribution of correlated features while f_X, f_Y denote the marginals.

As defined in Section 2.2.1, $G_{u,v} = \log \frac{f_{XY}(\vec{A}(u), \vec{B}(v))}{f_X(\vec{A}(u))f_Y(\vec{B}(v))} = \log f_{XY}(\vec{A}(u), \vec{B}(v)) - \log f_X(\vec{A}(u)) - \log f_Y(\vec{B}(v))$ for any $u \in \mathcal{U}$ and $v \in \mathcal{V}$. Then,

$$\begin{aligned} \log f_{\vec{A}, \vec{B}|M}(\vec{a}, \vec{b}) - \langle \mathbf{G}, \mathbf{M} \rangle &= \sum_{(u,v) \in \mathcal{W}_M} (\log f_{XY}(a_u, b_v) - G_{u,v}) \\ &\quad + \sum_{u \in \mathcal{U} \setminus \mathcal{U}_M} \log f_X(a_u) + \sum_{v \in \mathcal{V} \setminus \mathcal{V}_M} \log f_Y(a_v) \\ &= \sum_{(u,v) \in \mathcal{W}_M} (\log f_X(a_u) + \log f_Y(b_v)) \\ &\quad + \sum_{u \in \mathcal{U} \setminus \mathcal{U}_M} \log f_X(a_u) + \sum_{v \in \mathcal{V} \setminus \mathcal{V}_M} \log f_Y(a_v) \\ &= \sum_{u \in \mathcal{U}} \log f_X(a_u) + \sum_{v \in \mathcal{V}} \log f_Y(b_v) \end{aligned}$$

which shows that $\langle \mathbf{G}, \mathbf{M} \rangle = \log f_{\vec{A}, \vec{B}|M}(\vec{a}, \vec{b}) - \log f_{\vec{A}}(\vec{a}) - \log f_{\vec{B}}(\vec{b})$.

Then

$$\begin{aligned}
& \mathbb{E} [\exp (\langle \mathbf{G}, \boldsymbol{\Theta} - \mathbf{M} \rangle) | M] \\
&= \int \int \exp (\langle \mathbf{G}, \boldsymbol{\Theta} - \mathbf{M} \rangle) f_{\vec{A}, \vec{B} | M}(\vec{a}, \vec{b}) d\vec{a} d\vec{b} \\
&= \int \int \exp (\langle \mathbf{G}, \boldsymbol{\Theta} \rangle + \log f_{\vec{A}, \vec{B} | M}(\vec{a}, \vec{b}) - \langle \mathbf{G}, \mathbf{M} \rangle) d\vec{a} d\vec{b} \\
&= \int \int \exp (\langle \mathbf{G}, \boldsymbol{\Theta} \rangle + \log f_{\vec{A}}(\vec{a}) + \log f_{\vec{B}}(\vec{b})) d\vec{a} d\vec{b} \\
&= \int \int \exp (\langle \mathbf{G}, \boldsymbol{\Theta} \rangle) f_{\vec{A}}(\vec{a}) f_{\vec{B}}(\vec{b}) d\vec{a} d\vec{b} = R(\boldsymbol{\Theta})
\end{aligned}$$

This completes the proof. □

Corollary 11. *Let $\mathbf{m}_1, \mathbf{m}_2 \in \{0, 1\}^{\mathcal{U} \times \mathcal{V}}$ be the matrix encodings of the mappings m_1 and m_2 respectively. We have the following Chernoff bounds:*

a) *Probability of atypicality:*

$$\Pr (\tau |m_1| \geq \langle \mathbf{G}, \mathbf{m}_1 \rangle | M = m_1) \leq \exp (\theta \tau |m_1|) R ((1 - \theta) \mathbf{m}_1)$$

for any $\theta > 0$ such that the function R is non-negative.

b) *False positive probability:*

$$\Pr \left(G_{u,v} \geq \tau |u \not\stackrel{M}{\sim} v \right) \leq \exp (-\theta \tau) R ([\theta])$$

for any $\theta > 0$ such that the function R is non-negative.

c) *Probability of misalignment:*

$$\Pr (\langle \mathbf{G}, \mathbf{m}_2 \rangle \geq \langle \mathbf{G}, \mathbf{m}_1 \rangle | M = m_1) \leq R ((1 - \theta) \mathbf{m}_1 + \theta \mathbf{m}_2)$$

for any $\theta > 0$ such that the function R is non-negative.

d) *Probability of misalignment despite typicality:*

$$\begin{aligned} & \Pr (\langle \mathbf{G}, \mathbf{m}_1 \rangle \geq \tau |m_1| \wedge \langle \mathbf{G}, \mathbf{m}_2 \rangle \geq \langle \mathbf{G}, \mathbf{m}_1 \rangle \mid M = m_1) \\ & \leq e^{-\tau |m_1|(\nu-1)} R(\nu(1-\theta)\mathbf{m}_1 + \nu\theta\mathbf{m}_2) \end{aligned}$$

for any $\theta > 0$ and $\nu > 1$ such that the function R is non-negative.

Proof. a) **Probability of atypicality:**

$$\begin{aligned} & \Pr (\tau |m_1| \geq \langle \mathbf{G}, \mathbf{m}_1 \rangle \mid M = m_1) \\ & \Pr (-\theta\tau |m_1| \leq -\theta \langle \mathbf{G}, \mathbf{m}_1 \rangle \mid M = m_1) \\ & = \Pr (e^{-\theta\tau |m_1|} \leq \exp(\langle \mathbf{G}, -\theta\mathbf{m}_1 \rangle) \mid M = m_1) \\ & \leq \exp(\theta\tau |m_1|) \mathbb{E} [\exp(\langle \mathbf{G}, -\theta\mathbf{m}_1 \rangle) \mid M = m_1] \\ & = \exp(\theta\tau |m_1|) \mathbb{E} [\exp(\langle \mathbf{G}, -\theta\mathbf{m}_1 \rangle) \mid M = m_1] \\ & = \exp(\theta\tau |m_1|) R((1-\theta)\mathbf{m}_1). \end{aligned}$$

b) **False positive probability**

Ignore all vertices except a given pair (u, v) , i.e. suppose $\mathcal{U} = \{u\}$ and $\mathcal{V} = \{v\}$. Let m_1 denote an empty mapping, i.e. no pair is mapped under m_1 . Then $\mathbf{m}_1 = [0]$. Let m_2

denote the mapping that maps (u, v) . Then $\mathbf{m}_2 = [1]$.

$$\begin{aligned}
& \Pr \left(G_{u,v} \geq \tau \mid u \stackrel{M}{\not\sim} v \right) \\
&= \Pr \left(\langle \mathbf{G}, \mathbf{m}_2 - \mathbf{m}_1 \rangle \geq \tau \mid M = m_1 \right) \\
&= \Pr \left(\theta \langle \mathbf{G}, \mathbf{m}_2 - \mathbf{m}_1 \rangle \geq \theta \tau \mid M = m_1 \right) \\
&= \Pr \left(\exp \left(\langle \mathbf{G}, \theta(\mathbf{m}_2 - \mathbf{m}_1) \rangle \right) \geq e^{\theta \tau} \mid M = m_1 \right) \\
&\leq \exp(-\theta \tau) \mathbb{E} \left[\exp \left(\langle \mathbf{G}, \theta(\mathbf{m}_2 - \mathbf{m}_1) \rangle \right) \mid M = m_1 \right] \\
&= \exp(-\theta \tau) \mathbb{E} \left[\exp \left(\langle \mathbf{G}, \theta(\mathbf{m}_2 - \mathbf{m}_1) \rangle \right) \mid M = m_1 \right] \\
&= \exp(-\theta \tau) R(\theta \mathbf{m}_2 + (1 - \theta) \mathbf{m}_1) \\
&= \exp(-\theta \tau) R([\theta]).
\end{aligned}$$

c) Probability of misalignment:

$$\begin{aligned}
& \Pr \left(\langle \mathbf{G}, \mathbf{m}_2 \rangle \geq \langle \mathbf{G}, \mathbf{m}_1 \rangle \mid M = m_1 \right) \\
&= \Pr \left(\langle \mathbf{G}, \mathbf{m}_2 - \mathbf{m}_1 \rangle \geq 0 \mid M = m_1 \right) \\
&= \Pr \left(\exp \left(\langle \mathbf{G}, \theta(\mathbf{m}_2 - \mathbf{m}_1) \rangle \right) \geq 1 \mid M = m_1 \right) \\
&\leq \mathbb{E} \left[\exp \left(\langle \mathbf{G}, \theta(\mathbf{m}_2 - \mathbf{m}_1) \rangle \right) \mid M = m_1 \right] \\
&= R((1 - \theta) \mathbf{m}_1 + \theta \mathbf{m}_2).
\end{aligned}$$

d) Probability of misalignment despite typicality:

If $y \geq x$ and $x \geq t$, then $\theta_1(y - x) + \theta_2(x - t) \geq 0$ for any choice of $\theta_1, \theta_2 > 0$. Replace θ_1 by $\nu \theta$ and θ_2 by $\nu - 1$. $\theta_1, \theta_2 > 0$ holds for any $\theta > 0$ and $\nu > 1$. It then follows that,

if $y \geq x$ and $x \geq t$, then $\nu\theta(y - x) - (1 - \nu)x \geq (\nu - 1)t$. Then

$$\begin{aligned}
& \Pr(\langle \mathbf{G}, \mathbf{m}_2 \rangle \geq \langle \mathbf{G}, \mathbf{m}_1 \rangle \wedge \langle \mathbf{G}, \mathbf{m}_1 \rangle \geq \tau |m_1| \mid M = m_1) \\
& \leq \Pr(\langle \mathbf{G}, \nu\theta(\mathbf{m}_2 - \mathbf{m}_1) - (1 - \nu)\mathbf{m}_1 \rangle \geq \tau |m_1| (\nu - 1) \mid M = m_1) \\
& = \Pr(e^{\langle \mathbf{G}, \nu\theta(\mathbf{m}_2 - \mathbf{m}_1) - (1 - \nu)\mathbf{m}_1 \rangle} \geq e^{\tau |m_1| (\nu - 1)} \mid M = m_1) \\
& \leq e^{-\tau |m_1| (\nu - 1)} \mathbb{E} [e^{\langle \mathbf{G}, \nu\theta(\mathbf{m}_2 - \mathbf{m}_1) - (1 - \nu)\mathbf{m}_1 \rangle} \mid M = m_1] \\
& = e^{-\tau |m_1| (\nu - 1)} R(\nu\theta(\mathbf{m}_2 - \mathbf{m}_1) + \nu\mathbf{m}_1) \\
& = e^{-\tau |m_1| (\nu - 1)} R(\nu(1 - \theta)\mathbf{m}_1 + \nu\theta\mathbf{m}_2)
\end{aligned}$$

□

4.2.1 Main lemmas on the generating function

Lemma 12. Define $\mathbf{P} = \mathbf{P}^{\mathcal{U}, \mathcal{V}} : \mathbb{R}^{\mathcal{U} \times \mathcal{V}} \times (-1, 1) \rightarrow \mathbb{R}^{(\mathcal{U} \sqcup \mathcal{V}) \times (\mathcal{U} \sqcup \mathcal{V})}$ such that

$$\mathbf{P}(\Theta, \rho) \triangleq (1 - \rho^2) \mathbf{I} + \begin{bmatrix} \rho^2 \cdot \text{diag}(\Theta \vec{\mathbf{1}}) & -\rho\Theta \\ -\rho\Theta^\top & \rho^2 \cdot \text{diag}(\Theta^\top \vec{\mathbf{1}}) \end{bmatrix}$$

where $\vec{\mathbf{1}}$ represents appropriately indexed vectors of all ones.

If $\mathbf{P}(\Theta, \rho_i)$ positive definite for each $i \in \mathcal{D}$, then evaluating the generating function R at Θ gives the expression

$$R(\Theta) = \prod_{i \in \mathcal{D}} \left[\frac{(1 - \rho_i^2)^{|\mathcal{U}| + |\mathcal{V}| - \sum_{(u,v) \in \Theta_{u,v}} \Theta_{u,v}}}{\det \mathbf{P}(\Theta, \rho_i)} \right]^{\frac{1}{2}}$$

where $\vec{\rho} \in (-1, 1)^{\mathcal{D}}$ is the correlation vector in canonical form.

Proof. Let \vec{A}, \vec{B} databases in canonical form with statistics $\vec{\mu} = \vec{0}$ and

$$\Sigma = \begin{bmatrix} \mathbf{I} & \text{diag}(\vec{\rho}) \\ \text{diag}(\vec{\rho}) & \mathbf{I} \end{bmatrix} \text{ and } \mathbf{G} \in \mathbb{R}^{\mathcal{U} \times \mathcal{V}} \text{ their information density matrix.}$$

$$\begin{aligned} g_{XY}(\vec{x}, \vec{y}) &= \log \frac{f_{\vec{X}, \vec{Y}}(\vec{x}, \vec{y})}{f_{\vec{X}}(\vec{x}) f_{\vec{Y}}(\vec{y})} \\ &= \sum_{i \in \mathcal{D}} \left[-\frac{1}{2} \log(1 - \rho_i^2) - \frac{\rho_i^2 (x_i^2 + y_i^2) - 2\rho_i x_i y_i}{2(1 - \rho_i^2)} \right] \end{aligned}$$

Define $\vec{a} \in \mathbb{R}^{\mathcal{U}}$ and $\vec{b} \in \mathbb{R}^{\mathcal{V}}$ such that a_u and b_v denote the features associated with users u and v respectively. As shown in the proof for Lemma 10, $\log f_{\vec{A}, \vec{B}|M}(\vec{a}, \vec{b}) - \langle \mathbf{G}, \mathbf{M} \rangle = \sum \log f_X(a_u) + \sum \log f_Y(b_v)$. Without loss of generality, assume $\vec{\rho} = [\rho]$. Then,

$$\begin{aligned} &\langle \mathbf{G}, \Theta \rangle + \log f_{\vec{A}, \vec{B}|M}(\vec{a}, \vec{b}) - \langle \mathbf{G}, \mathbf{M} \rangle \\ &= \langle \mathbf{G}, \Theta \rangle + \sum_{u \in \mathcal{U}} \log f_X(a_u) + \sum_{v \in \mathcal{V}} \log f_Y(b_v) \\ &= -\frac{1}{2} \log(1 - \rho^2) \sum_{(u,v)} \Theta_{u,v} - \frac{|\mathcal{U}| + |\mathcal{V}|}{2} \log(2\pi) \\ &\quad - \frac{1}{2(1 - \rho^2)} \sum_{u \in \mathcal{U}} a_u^2 \left(1 - \rho^2 + \rho^2 \sum_{v \in \mathcal{V}} \Theta_{u,v} \right) \\ &\quad - \frac{1}{2(1 - \rho^2)} \sum_{v \in \mathcal{V}} b_v^2 \left(1 - \rho^2 + \rho^2 \sum_{u \in \mathcal{U}} \Theta_{u,v} \right) \\ &\quad - \frac{1}{2(1 - \rho^2)} \sum_{u \in \mathcal{V}} \sum_{v \in \mathcal{V}} \rho^{\Theta_{u,v}} a_u b_v + \rho^{\Theta_{u,v}} b_v a_u \\ &= -\frac{1}{2} \log(1 - \rho^2) \sum_{(u,v)} \Theta_{u,v} - \frac{|\mathcal{U}| + |\mathcal{V}|}{2} \log(2\pi) \\ &\quad - \frac{1}{2(1 - \rho^2)} \begin{bmatrix} \vec{a} \\ \vec{b} \end{bmatrix}^\top \mathbf{P}(\Theta, \rho) \begin{bmatrix} \vec{a} \\ \vec{b} \end{bmatrix} \end{aligned}$$

Then we can write

$$\begin{aligned} & \exp(\langle \mathbf{G}, \Theta - \mathbf{M} \rangle) f_{\vec{A}, \vec{B} | M}(\vec{a}, \vec{b}) \\ &= \frac{(1 - \rho^2)^{-\frac{1}{2} \sum_{(u,v)} \Theta_{u,v}}}{(2\pi)^{\frac{|U|+|V|}{2}}} \cdot \exp\left(-\frac{1}{2(1 - \rho^2)} \begin{bmatrix} \vec{\alpha}_i \\ \vec{\beta}_i \end{bmatrix}^\top \mathbf{P}(\Theta, \rho) \begin{bmatrix} \vec{\alpha}_i \\ \vec{\beta}_i \end{bmatrix}\right) \end{aligned}$$

Taking the integral of this expression gives us the claimed result. For the case where $\vec{\rho}$ is multi-dimensional, taking the product of this expression over each $i \in \mathcal{D}$ gives us the proper expression. \square

Lemma 13 (Decomposition for block diagonal matrices). *If $\Theta \in \mathbb{R}^{U \times V}$ can be written in block diagonal form, i.e. if U and V can be partitioned into U_1, U_2 and V_1, V_2 such that Θ can be written as $\Theta = \begin{bmatrix} \Theta_1 & 0 \\ 0 & \Theta_2 \end{bmatrix} \in \mathbb{R}^{(U_1 \sqcup U_2) \times (V_1 \sqcup V_2)}$, then*

$$R^{U,V}(\Theta) = R^{U_1, V_1}(\Theta_1) \cdot R^{U_2, V_2}(\Theta_2)$$

Proof. This follows from the fact that $\mathbf{P}^{U,V}(\Theta, \rho) \in \mathbb{R}^{U \times V}$ can be transformed into block matrix form as $\begin{bmatrix} \mathbf{P}^{U_1, V_1}(\Theta_1, \rho) & 0 \\ 0 & \mathbf{P}^{U_2, V_2}(\Theta_2, \rho) \end{bmatrix}$ by simultaneously permuting rows and columns.

Then $\det \mathbf{P}^{U,V}(\Theta) = \det \mathbf{P}^{U_1, V_1}(\Theta_1, \rho) \cdot \det \mathbf{P}^{U_2, V_2}(\Theta_2, \rho)$ and we get the claimed result. \square

Lemma 14. *If $\theta \in \mathbb{R}$ such that $|\theta| < |1/\rho_i|$ for any $i \in \mathcal{D}$, then*

$$R([1 - \theta]) = \prod_{i \in \mathcal{D}} \left[\frac{(1 - \rho_i^2)^\theta}{1 - \rho_i^2 \theta^2} \right]^{\frac{1}{2}}$$

Furthermore, if $\theta \in [-1, 1]$ then

$$R([1 - \theta]) \leq \exp(-\theta(1 - \theta)I_{XY}),$$

where $I_{XY} = -\frac{1}{2} \sum_{i \in \mathcal{D}} \log(1 - \rho_i^2)$.

Proof.

$$\mathbf{P}([1 - \theta], \rho) = \begin{bmatrix} 1 - \rho^2\theta & -\rho(1 - \theta) \\ -\rho(1 - \theta) & 1 - \rho^2\theta \end{bmatrix}$$

The eigenvalues of this matrix are $(1 - \rho^2\theta) - \rho(1 - \theta)$ and $(1 - \rho^2\theta) + \rho(1 - \theta)$. They are both strictly positive if and only if $|\theta| < |1/\rho|$. Their product equals $(1 - \rho^2)(1 - \rho^2\theta^2)$. Plugging in this value in the denominator of $\frac{(1 - \rho_i^2)^{|\mathcal{U}| + |\mathcal{V}| - \sum_{(u,v)} \Theta_{u,v}}}{\det \mathbf{P}(\Theta, \rho_i)}$ as given in Lemma 12 gives us the exact expression for $R([1 - \theta])$.

By Lemma 51, $(1 - \rho^2\theta^2) \geq (1 - \rho^2)^{\theta^2}$ if $\theta \in [-1, 1]$, which gives us

$$R([1 - \theta]) \leq \prod_{i \in \mathcal{D}} (1 - \rho_i^2)^{\frac{\theta(1 - \theta)}{2}}$$

This gives us the upper bound for $R([1 - \theta])$. □

Corollary 15. *Let $\mathbf{m} \in \{0, 1\}^{\mathcal{U} \times \mathcal{V}}$ some binary matrix with row sums and columns sums at most 1. Let n denote the sum of its entries. Then*

$$R((1 - \theta)\mathbf{m}) = \prod_{i \in \mathcal{D}} \left[\frac{(1 - \rho_i^2)^\theta}{1 - \rho_i^2\theta^2} \right]^{\frac{n}{2}}$$

Furthermore, if $\theta \in [-1, 1]$ then

$$R([1 - \theta]\mathbf{m}) \leq \exp(-n\theta(1 - \theta)I_{XY}),$$

where $I_{XY} = -\frac{1}{2} \sum_{i \in \mathcal{D}} \log(1 - \rho_i^2)$.

Proof. $(1 - \theta)\mathbf{m}$ has at most 1 non-zero entry in each row and in each column. Then it can be arranged to have block diagonal form where each non-zero block on the diagonal has size 1 and is equal to $[1 - \theta]$. By Lemma 13, $R((1 - \theta)\mathbf{m})$ decomposes into the product of $[R([1 - \theta])]^n$ and $R(\mathbf{0})$ where $\mathbf{0}$ is the zero block of the block diagonal decomposition. By Lemma 12, $R(\mathbf{0}) = 1$. The expression for $R([1 - \theta])$ is given by Lemma 14. \square

4.2.2 Generating function evaluated for cycles and even paths

Lemma 16 (Exact Chernoff bound for a cycle). *Let $\theta \in \mathbb{R}$ and $\nu \in \mathbb{R}$ such that $|\nu| \leq \min_i 1/|\rho_i|$. Let m_1 and m_2 be two mappings of size n such that $\Theta \triangleq \nu((1 - \theta)\mathbf{m}_1 + \theta\mathbf{m}_2)$ is a matrix block corresponding to a cycle of the type given in Fig. 2.3-I. Then*

$$\begin{aligned} \log R(\Theta) &\geq \log R(\Theta^*) \\ &= -n(2 - \nu) \cdot I_{XY} \\ &\quad - \frac{1}{2} \sum_{i \in \mathcal{D}} \sum_{k=0}^{n-1} \log \left((1 - \rho_i^2(1 - \nu))^2 - \frac{\rho_i^2 \nu^2}{2} \left[1 + \cos \left(\frac{2k\pi}{n} \right) \right] \right), \end{aligned}$$

where $I_{XY} = -\frac{1}{2} \sum \log(1 - \rho_i^2)$, and $\Theta^* = \frac{\nu(\mathbf{m}_1 + \mathbf{m}_2)}{2}$ is the value for Θ optimized over $\theta \in \mathbb{R}$, which corresponds to $\theta = 1/2$.

In particular, for $n = 2$ and $\nu = 1$,

$$\log R(\Theta) \geq \log R(\Theta^*) = -I_{XY}$$

Proof. Without loss of generality, suppose the canonical correlation vector $\vec{\rho} \in (-1, 1)^{\mathcal{D}}$ is actually one-dimensional: $\vec{\rho} = [\rho]$. Our final result generalizes to the multi-dimensional case by taking the product over all $i \in \mathcal{D}$.

Without loss of generality, let Θ have its rows and columns arranged such that

$$\Theta = \nu \begin{bmatrix} \theta & 1-\theta & 0 & \cdots & 0 & 0 \\ 0 & \theta & 1-\theta & \cdots & 0 & 0 \\ 0 & 0 & \theta & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \theta & 1-\theta \\ 1-\theta & 0 & 0 & \cdots & 0 & \theta \end{bmatrix}.$$

Θ has $|\mathcal{U}| = |m_1| = n$ rows and $|\mathcal{V}| = |m_1| = n$ columns.

Recall the definition of $\mathbf{P}(\Theta, \rho)$ as given in Lemma 12:

$$\mathbf{P}(\Theta, \rho) \triangleq (1 - \rho^2) \mathbf{I} + \begin{bmatrix} \rho^2 \cdot \text{diag}(\Theta \vec{\mathbf{1}}) & -\rho \Theta \\ -\rho \Theta^\top & \rho^2 \cdot \text{diag}(\Theta^\top \vec{\mathbf{1}}) \end{bmatrix}$$

Define the \mathbf{P}_1 and \mathbf{P}_2 to be the matrices that form the diagonal blocks of $\mathbf{P}(\Theta, \rho) =$

$$\begin{bmatrix} \mathbf{P}_1 & -\rho \Theta \\ -\rho \Theta^\top & \mathbf{P}_2 \end{bmatrix}:$$

$$\mathbf{P}_1 \triangleq (1 - \rho^2) \mathbf{I} + \rho^2 \text{diag}(\Theta \vec{\mathbf{1}})$$

$$= (1 - \rho^2(1 - \nu)) \mathbf{I}$$

$$\mathbf{P}_2 \triangleq (1 - \rho^2) \mathbf{I} + \rho^2 \text{diag}(\Theta^\top \vec{\mathbf{1}})$$

$$= (1 - \rho^2(1 - \nu)) \mathbf{I}$$

Then

$$\begin{aligned}
\det \mathbf{P}(\boldsymbol{\Theta}, \rho) &= [\det \mathbf{P}_1] [\det (\mathbf{P}_2 - \rho^2 \boldsymbol{\Theta}^\top \mathbf{P}_1^{-1} \boldsymbol{\Theta})] \\
&= [1 - \rho^2(1 - \nu)]^n \det \left[(1 - \rho^2(1 - \nu)) \mathbf{I} - \frac{\rho^2}{1 - \rho^2(1 - \nu)} \boldsymbol{\Theta}^\top \boldsymbol{\Theta} \right] \\
&= \det \left[(1 - \rho^2(1 - \nu))^2 \mathbf{I} - \rho^2 \boldsymbol{\Theta}^\top \boldsymbol{\Theta} \right]
\end{aligned}$$

where $\mathbf{P}_2 - \rho^2 \boldsymbol{\Theta}^\top \mathbf{P}_1^{-1} \boldsymbol{\Theta} = (1 - \rho^2(1 - \nu)) \mathbf{I} - \frac{\rho^2}{1 - \rho^2(1 - \nu)} \boldsymbol{\Theta}^\top \boldsymbol{\Theta}$ is the Schur complement of \mathbf{P}_1 .

We can write

$$\boldsymbol{\Theta}^\top \boldsymbol{\Theta} = \nu^2 (1 - 2\theta(1 - \theta)) \mathbf{I} + 2\nu^2 \theta(1 - \theta) \mathbf{Q}$$

where

$$\mathbf{Q} \triangleq \frac{1}{2} \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 & 0 & 1 \\ 1 & 0 & 1 & \cdots & 0 & 0 & 0 \\ 0 & 1 & 0 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 1 & 0 \\ 0 & 0 & 0 & \cdots & 1 & 0 & 1 \\ 0 & 0 & 0 & \cdots & 0 & 1 & 0 \end{bmatrix}$$

and \mathbf{Q} has eigenvalues $\cos\left(\frac{2k\pi}{n}\right)$, $k \in \{0, 1, 2, \dots, n-1\}$. These are all between -1 and 1 . Furthermore, the sum of the eigenvalues of \mathbf{Q} is $\text{tr}(\mathbf{Q}) = 0$ Thus the n eigenvalues of the matrix

$$\begin{aligned}
&(1 - \rho^2(1 - \nu))^2 \mathbf{I} - \rho^2 \boldsymbol{\Theta}^\top \boldsymbol{\Theta} \\
&= (1 - \rho^2(1 - \nu))^2 \mathbf{I} - \rho^2 \nu^2 [(1 - 2\theta(1 - \theta)) \mathbf{I} + 2\theta(1 - \theta) \mathbf{Q}]
\end{aligned}$$

are given by the expression

$$\begin{aligned} & (1 - \rho^2(1 - \nu))^2 - \rho^2\nu^2 \left((1 - 2\theta(1 - \theta)) + 2\theta(1 - \theta) \cos\left(\frac{2k\pi}{n}\right) \right) \\ &= (1 - \rho^2(1 - \nu))^2 - \rho^2\nu^2 + 2\rho^2\nu^2\theta(1 - \theta) \left[1 - \cos\left(\frac{2k\pi}{n}\right) \right] \end{aligned}$$

for $k \in \{0, 1, 2, \dots, n-1\}$.

For any value of k , $\theta = 1/2$ gives us the maximum value for this expression over all $\theta \in \mathbb{R}$. The value for $\det \mathbf{P}(\Theta, \rho) = \det \left[(1 - \rho^2(1 - \nu))^2 \mathbf{I} - \rho^2 \Theta^\top \Theta \right]$ is the product of all eigenvalues. Then, $\det \mathbf{P}(\Theta, \rho)$ is maximized over all $\theta \in \mathbb{R}$ at $\theta = 1/2$. Consequently, $R(\Theta)$ is minimized over all $\theta \in \mathbb{R}$ at $\theta = 1/2$.

Henceforth assume $\theta = 1/2$. This gives us $2\theta(1 - \theta) = 1/2$ and

$$\begin{aligned} \det \mathbf{P}(\Theta, \rho) &= \prod_{k=0}^{n-1} \left((1 - \rho^2(1 - \nu))^2 - \rho^2\nu^2 + \frac{\rho^2\nu^2}{2} \left[1 - \cos\left(\frac{2k\pi}{n}\right) \right] \right) \\ &= \prod_{k=0}^{n-1} \left((1 - \rho^2(1 - \nu))^2 - \frac{\rho^2\nu^2}{2} \left[1 + \cos\left(\frac{2k\pi}{n}\right) \right] \right). \end{aligned}$$

$R(\Theta)$ has $\frac{1}{2} \left(|\mathcal{U}| + |\mathcal{V}| - \sum_{(u,v)} \Theta_{u,v} \right) = \frac{n}{2}(1 + 1 - \nu) = \frac{n}{2}(2 - \nu)$ multiplicative terms of $(1 - \rho^2)$ in the numerator. This gives us

$$R(\Theta) = \left[\frac{(1 - \rho^2)^{n(2-\nu)}}{\prod_k \left((1 - \rho^2(1 - \nu))^2 - \frac{\rho^2\nu^2}{2} \left[1 + \cos\left(\frac{2k\pi}{n}\right) \right] \right)} \right]^{1/2}.$$

$I_{XY} = -\frac{1}{2} \sum \log(1 - \rho_i^2)$, which allows us to convert the expression above to the claimed result. \square

Lemma 17 (Chernoff bound for a cycle in the low per-feature correlation regime). *Given $\nu \in [0, 2]$, let m_1 and m_2 be two mappings of size n such that $\Theta = \frac{\nu}{2}(\mathbf{m}_1 + \mathbf{m}_2)$ is a matrix*

block corresponding to a cycle of the type given in Fig. 2.3-I.

$$\frac{\log R(\Theta)}{I_{XY}} \leq -\frac{n}{2}\nu(2-\nu) + n \cdot \rho_{\max}^2(\nu-1)$$

where $I_{XY} = -\frac{1}{2}\sum \log(1-\rho_i^2)$, mutual information between correlated features, and $\rho_{\max} = \max |\rho_i|$, the largest correlation coefficient under the canonical form.

Proof. Without loss of generality, suppose the canonical correlation vector $\vec{\rho} \in (-1, 1)^{\mathcal{D}}$ is actually one-dimensional: $\vec{\rho} = [\rho]$. Our final result generalizes to the multi-dimensional case by taking the product over all $i \in \mathcal{D}$.

By Lemma 16,

$$\log R(\Theta) = -n(2-\nu) \cdot I_{XY} - \frac{1}{2} \sum_{k=0}^{n-1} \log \left((1-\rho^2(1-\nu))^2 - \frac{\rho^2\nu^2}{2} \left[1 + \cos \left(\frac{2k\pi}{n} \right) \right] \right).$$

By Lemma 52 with $\tau = (1-\rho^2(1-\nu))^2 - \rho^2\nu^2/2$ and $\sigma = -\rho^2\nu^2/2$,

$$\begin{aligned} \det \mathbf{P}(\Theta, \rho) &= \prod_{k=0}^{n-1} \left((1-\rho^2(1-\nu))^2 - \frac{\rho^2\nu^2}{2} \left[1 + \cos \left(\frac{2k\pi}{n} \right) \right] \right) \\ &\geq \left[(1-\rho^2(1-\nu))^2 \right]^{\frac{n}{2}} \left[(1-\rho^2(1-\nu))^2 - \rho^2\nu^2 \right]^{\frac{n}{2}} \\ &= [1-\rho^2(1-\nu)]^n [1-\rho^2]^{\frac{n}{2}} [1-\rho^2(1-\nu)^2]^{\frac{n}{2}} \end{aligned}$$

Since $\nu \in [0, 2] \iff (1-\nu)^2 \in [0, 1]$ by Lemma 51, we have the bound $1-\rho^2(1-\nu)^2 \geq (1-\rho^2)^{(1-\nu)^2}$. Furthermore, since $-(1-\nu) \in [0, 1]$, we have the bound $1-\rho^2(1-\nu) \geq (1+\rho^2)^{-(1-\nu)}$. Then,

$$\log \det \mathbf{P}(\Theta, \rho) \geq \frac{n}{2} (1 + (1-\nu)^2) \log(1-\rho^2) - n(1-\nu) \log(1+\rho^2)$$

$R(\Theta)$ has $\frac{1}{2} \left(|\mathcal{U}| + |\mathcal{V}| - \sum_{(u,v)} \Theta_{u,v} \right) = \frac{n}{2}(1+1-\nu) = \frac{n}{2}(2-\nu)$ multiplicative terms of

$(1 - \rho^2)$ in the numerator and our bound has $\frac{n}{4}(1 + (1 - \nu)^2)$ of them in the denominator, which gives, in total, $\frac{n}{4}\nu(2 - \nu) + \frac{n}{2}(1 - \nu)$ such terms. Then

$$\log R(\Theta) = \frac{n}{4}\nu(2 - \nu) \log(1 - \rho^2) - \frac{n}{2}(\nu - 1) [\log(1 + \rho^2) + \log(1 - \rho^2)]$$

By Lemma 53, $-\frac{1}{2} \sum \log(1 - \rho_i^2) + \log(1 + \rho_i^2)$ is upper bounded by $-\frac{1}{2} \sum \rho_i^2 \log(1 - \rho_i^2)$, which itself is upper bounded by $\rho_{\max}^2 I_{XY}$ where $\rho_{\max} = \max |\rho_i|$ and $I_{XY} = -\frac{1}{2} \sum \log(1 - \rho_i^2)$.

This gives us the claimed result. □

Lemma 18 (Chernoff bound for the minimal even path with $\nu = 1$ and $\theta = 1/2$). *Given*

$$\Theta = [1/2, 1/2] \in \mathbb{R}^{1 \times 2}$$

$$R(\Theta) = \prod_{i \in \mathcal{D}} \left[\frac{1 - \rho_i^2}{1 - \rho_i^2/2} \right]^{\frac{1}{2}}$$

Proof. Define the \mathbf{P}_1 and \mathbf{P}_2 to be the matrices that form the diagonal blocks of $\mathbf{P}(\Theta, \rho) =$

$$\begin{bmatrix} \mathbf{P}_1 & -\rho\Theta \\ -\rho\Theta^\top & \mathbf{P}_2 \end{bmatrix}:$$

$$\mathbf{P}_1 \triangleq (1 - \rho^2) \mathbf{I} + \rho^2 \text{diag}(\Theta \vec{1}) = \mathbf{I}$$

$$\mathbf{P}_2 \triangleq (1 - \rho^2) \mathbf{I} + \rho^2 \text{diag}(\Theta^\top \vec{1}) = \left(1 - \frac{\rho^2}{2}\right) \mathbf{I}$$

Then

$$\begin{aligned} \det \mathbf{P}(\Theta, \rho) &= \det \mathbf{P}_1 \det [\mathbf{P}_2 - \rho^2 \Theta^\top \mathbf{P}_1^{-1} \Theta] \\ &= \det [\mathbf{P}_2 - \rho^2 \Theta^\top \Theta] \\ &= \det \left[\left(1 - \frac{\rho^2}{2}\right) \mathbf{I} - \frac{\rho^2}{4} \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \right] \end{aligned}$$

$\frac{1}{4} \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$ has eigenvalues $1/2$ and 0 . Then $\left(1 - \frac{\rho^2}{2}\right) \mathbf{I} - \frac{\rho^2}{4} \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$ has eigenvalues $1 - \rho^2$ and $1 - \frac{\rho^2}{2}$. It then follows that

$$\begin{aligned} \det \mathbf{P}(\Theta, \rho) &= \det \left[\left(1 - \frac{\rho^2}{2}\right) \mathbf{I} - \frac{\rho^2}{4} \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \right] \\ &= (1 - \rho^2) (1 - \rho^2/2) \end{aligned}$$

and

$$\begin{aligned} R(\Theta) &= \prod_{i \in \mathcal{D}} \left[\frac{(1 - \rho_i^2)^{|\mathcal{U}|+|\mathcal{V}|-\sum_{(u,v)} \Theta_{u,v}}}{\det \mathbf{P}(\Theta, \rho_i)} \right]^{\frac{1}{2}} \\ &= \prod_{i \in \mathcal{D}} \left[\frac{(1 - \rho_i^2)^{1+2-1}}{(1 - \rho_i^2) (1 - \rho_i^2/2)} \right]^{\frac{1}{2}} \\ &= \prod_{i \in \mathcal{D}} \left[\frac{1 - \rho_i^2}{1 - \rho_i^2/2} \right]^{\frac{1}{2}} \end{aligned}$$

□

Lemma 19 (Chernoff bound for an even path in the low per-feature correlation regime).

Given $\nu \in [1, 2]$, let m_1 and m_2 be two mappings of size n such that $\Theta = \frac{\nu}{2}(\mathbf{m}_1 + \mathbf{m}_2)$ is a matrix block corresponding to a cycle of the type given in Fig. 2.3-II.

$$\frac{\log R(\Theta)}{I_{XY}} \leq -\frac{n}{2} \cdot \nu(2 - \nu) - (\nu - 1)^2 \nu^2 + 6n \cdot \rho_{\max}^2 (\nu - 1)$$

where $I_{XY} = -\frac{1}{2} \sum \log(1 - \rho_i^2)$, mutual information between correlated features, and $\rho_{\max} = \max |\rho_i|$, the largest correlation coefficient under the canonical form.

Proof. Without loss of generality, let Θ have its rows and columns arranged such that

$$\Theta = \nu \begin{bmatrix} \theta & 1-\theta & 0 & \cdots & 0 & 0 & 0 \\ 0 & \theta & 1-\theta & \cdots & 0 & 0 & 0 \\ 0 & 0 & \theta & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \theta & 1-\theta & 0 \\ 0 & 0 & 0 & \cdots & 0 & \theta & 1-\theta \end{bmatrix}.$$

Θ has $|\mathcal{U}| = |m_1| = n$ rows and $|\mathcal{V}| = |m_1| + 1 = n + 1$ columns.

Define the \mathbf{P}_1 and \mathbf{P}_2 to be the matrices that form the diagonal blocks of $\mathbf{P}(\Theta, \rho) = \begin{bmatrix} \mathbf{P}_1 & -\rho\Theta \\ -\rho\Theta^\top & \mathbf{P}_2 \end{bmatrix}$:

$$\begin{aligned} \mathbf{P}_1 &\triangleq (1 - \rho^2) \mathbf{I} + \rho^2 \text{diag}(\Theta \vec{\mathbf{1}}) \\ &= (1 - \rho^2(1 - \nu)) \mathbf{I} \\ \mathbf{P}_2 &\triangleq (1 - \rho^2) \mathbf{I} + \rho^2 \text{diag}(\Theta^\top \vec{\mathbf{1}}) \\ &= (1 - \rho^2(1 - \nu)) \mathbf{I} - \rho^2 \nu \text{diag}(1 - \theta, 0, 0, \dots, 0, \theta) \end{aligned}$$

Then $\det \mathbf{P}(\Theta, \rho)$ is equal to $\det \mathbf{P}_1 = (1 - \rho^2)^n$ times the determinant of the Schur complement of \mathbf{P}_1 , i.e. $\mathbf{P}_2 - \rho^2 \Theta^\top \mathbf{P}_1^{-1}$, which is equal to

$$(1 - \rho^2(1 - \nu)) \mathbf{I} - \rho^2 \nu \text{diag}(1 - \theta, 0, 0, \dots, 0, \theta) - \frac{\rho^2}{1 - \rho^2(1 - \nu)} \Theta^\top \Theta$$

Define another auxilliary matrix

$$\mathbf{Q} \triangleq \frac{1}{2\nu^2\theta(1-\theta)} [\Theta^\top \Theta - [1 - 2\theta(1 - \theta)] \nu^2 \mathbf{I} + \nu (1 - \rho^2(1 - \nu)) \text{diag}(1 - \theta, 0, 0, \dots, 0, \theta)]$$

such that the Schur complement of \mathbf{P}_1 can be expressed as

$$\begin{aligned} & (1 - \rho^2(1 - \nu)) \mathbf{I} - \rho^2 \nu \operatorname{diag}(1 - \theta, 0, 0, \dots, 0, \theta) - \frac{\rho^2}{1 - \rho^2(1 - \nu)} \Theta^\top \Theta \\ &= \left[1 - \rho^2(1 - \nu) - \frac{\rho^2 \nu^2 (1 - 2\theta(1 - \theta))}{1 - \rho^2(1 - \nu)} \right] \mathbf{I} - \frac{2\rho^2 \nu^2 \theta (1 - \theta)}{1 - \rho^2(1 - \nu)} \mathbf{Q} \end{aligned}$$

It can be shown that

$$\mathbf{Q} \triangleq \frac{1}{2} \begin{bmatrix} 1 & 1 & 0 & \cdots & 0 & 0 & 0 \\ 1 & 0 & 1 & \cdots & 0 & 0 & 0 \\ 0 & 1 & 0 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 1 & 0 \\ 0 & 0 & 0 & \cdots & 1 & 0 & 1 \\ 0 & 0 & 0 & \cdots & 0 & 1 & 1 \end{bmatrix} + \nu(1 - \nu)(1 - \rho^2) \operatorname{diag}(1 - \theta, 0, 0, \dots, 0, \theta).$$

We have $\nu(1 - \nu) \max\{\theta, 1 - \theta\} \in [-1, 0]$. Then, \mathbf{Q} is an irreducible non-negative square matrix with row sums at most 1. Consequently, by the Perron-Frobenius theorem, its eigenvalues are all between -1 and 1 . Then the eigenvalues of the Schur complement of \mathbf{P}_1 are between $1 - \rho^2(1 - \nu) - \frac{\rho^2 \nu^2}{1 - \rho^2(1 - \nu)}$ and $1 - \rho^2(1 - \nu) - \frac{\rho^2 \nu^2 (1 - 2\theta)^2}{1 - \rho^2(1 - \nu)}$. Furthermore, the sum of the eigenvalues of \mathbf{Q} is $\operatorname{tr}(\mathbf{Q}) = 1 + \nu(1 - \nu)(1 - \rho^2)$. Multiplying the $n + 1$ eigenvalues of \mathbf{Q} by $-\frac{2\rho^2 \nu^2 \theta (1 - \theta)}{1 - \rho^2(1 - \nu)}$ and adding $1 - \rho^2(1 - \nu) - \frac{\rho^2 \nu^2 (1 - 2\theta(1 - \theta))}{1 - \rho^2(1 - \nu)}$ gives us the $n + 1$ eigenvalues of the Schur complement of \mathbf{P}_1 . Now we plug in $\theta = 1/2$. By Lemma 52, the determinant of the Schur complement of \mathbf{P}_1 is lower bounded by

$$\left[1 - \rho^2(1 - \nu) \right]^{\frac{n}{2}} \left[1 - \rho^2(1 - \nu) - \frac{\rho^2 \nu^2}{1 - \rho^2(1 - \nu)} \right]^{\frac{n}{2} + 1 + \nu(1 - \nu)(1 - \rho^2)}.$$

Then, multiplying the determinant of the Schur complement of \mathbf{P}_1 by $\det \mathbf{P}_1 = [1 - \rho^2(1 - \nu)]^n$

gives us

$$\begin{aligned} \det \mathbf{P}(\Theta, \rho) &\geq [1 - \rho^2(1 - \nu)]^{n-1} \left[(1 - \rho^2(1 - \nu))^2 - \rho^2 \nu^2 \right]^{\frac{n}{2} + 1 + \nu(1 - \nu)(1 - \rho^2)} \\ &= [1 - \rho^2(1 - \nu)]^{n-1 - \nu(1 - \nu)(1 - \rho^2)} \left[(1 - \rho^2) (1 - \rho^2(1 - \nu)^2) \right]^{\frac{n}{2} + 1 + \nu(1 - \nu)(1 - \rho^2)}. \end{aligned}$$

Since $\nu \in [0, 2] \iff (1 - \nu)^2 \in [0, 1]$ by Lemma 51, we have the bound $1 - \rho^2(1 - \nu)^2 \geq (1 - \rho^2)^{(1 - \nu)^2}$. Furthermore, since $-(1 - \nu) \in [0, 1]$, we have the bound $1 - \rho^2(1 - \nu) \geq (1 + \rho^2)^{-(1 - \nu)}$. Then,

$$\begin{aligned} \log \det \mathbf{P}(\Theta, \rho) &\geq \left(\frac{n}{2} + 1 + \nu(1 - \nu) (1 - \rho^2) \right) (1 + (1 - \nu)^2) \log (1 - \rho^2) \\ &\quad - (n - 1 - \nu(1 - \nu) (1 - \rho^2))(1 - \nu) \log (1 + \rho^2). \end{aligned}$$

$R(\Theta)$ has $\frac{1}{2} \left(|\mathcal{U}| + |\mathcal{V}| - \sum_{(u,v)} \Theta_{u,v} \right) = \frac{1}{2}(n + n + 1 - \nu n) = \frac{n}{2}(2 - \nu) + \frac{1}{2}$ multiplicative terms of $(1 - \rho^2)$ in the numerator and our bound has $\frac{1}{2} \left(\frac{n}{2} + 1 + \nu(1 - \nu) (1 - \rho^2) \right) (1 + (1 - \nu)^2)$ of them in the denominator, which gives, in total, $\frac{n}{4}\nu(2 - \nu) - \frac{n-1}{2}(\nu - 1) + \frac{1}{2}\nu(\nu - 1)^3 (1 - \rho^2) - \frac{1}{2}\nu(\nu - 1)\rho^2$ such terms. Then

$$\begin{aligned} \log R(\Theta) &\leq \frac{1}{2} \left[\frac{n}{2}\nu(2 - \nu) + \nu^2(\nu - 1)^2 - \nu^2(\nu - 1)^2\rho^2 - \nu(\nu - 1)\rho^2 \right] \log (1 - \rho^2) \\ &\quad - \frac{\nu - 1}{2} \left[(n - 1) + \nu(\nu - 1) (1 - \rho^2) \right] \left[\log (1 - \rho^2) + \log (1 + \rho^2) \right] \end{aligned}$$

For the multi-dimensional case, since under the canonical form all dimensions are mutually

independent, we get $\log R(\Theta)$ by simply summing this expression for all ρ_i . This gives us

$$\begin{aligned} \log R(\Theta) &\leq \left[\frac{n}{2}\nu(2-\nu) + \nu^2(\nu-1)^2 \right] \sum_{i \in \mathcal{D}} \frac{1}{2} \log(1-\rho_i^2) \\ &\quad - (\nu-1) \left[\nu^2(\nu-1) + \nu \right] \sum_{i \in \mathcal{D}} \frac{1}{2} \rho_i^2 \log(1-\rho_i^2) \\ &\quad - (\nu-1)(n-1) \sum_{i \in \mathcal{D}} \frac{1}{2} \left[\log(1-\rho_i^2) + \log(1+\rho_i^2) \right] \\ &\quad + (\nu-1)^2 \nu \sum_{i \in \mathcal{D}} \frac{1}{2} \rho_i^2 \left[\log(1-\rho_i^2) + \log(1+\rho_i^2) \right] \end{aligned}$$

Given $\nu \geq 1$, the last term is non-positive, and therefore we can drop it while maintaining the inequality. By Lemma 53, $-\frac{1}{2} \sum \log(1-\rho_i^2) + \log(1+\rho_i^2)$ is upper bounded by $-\frac{1}{2} \sum \rho_i^2 \log(1-\rho_i^2)$, which itself is upper bounded by $\rho_{\max}^2 I_{XY}$ where $\rho_{\max} = \max |\rho_i|$ and $I_{XY} = -\frac{1}{2} \sum \log(1-\rho_i^2)$. This gives us

$$\frac{\log R(\Theta)}{I_{XY}} \leq -\frac{n}{2} \cdot \nu(2-\nu) - (\nu-1)^2 \nu^2 + n \cdot \rho_{\max}^2 (\nu-1) + \rho_{\max}^2 (\nu-1)^2 (\nu^2 + 1)$$

Since $n \geq 1$ and $\nu \in [1, 2]$, we have $(\nu-1)^2 (\nu^2 + 1) \leq 5n(\nu-1)$. This gives us the claimed result. □

4.3 Concentration Inequalities

In Subsection 4.3.1, we state concentration inequalities and give their proofs based on our results on the generating function in Section 4.2, through Lemma 11. In Subsection 4.3.2, we provide some geometric intuition behind the expressions for the concentration inequalities.

4.3.1 Statements of concentration inequalities and their proofs

\mathbf{G} refers to the information density matrix under the database alignment setting as described in Subsection 2.1.3. Specifically, \mathbf{G} is the matrix such that $G_{u,v}$ is the log-likelihood ratio of hypotheses $u \stackrel{M}{\sim} v$ vs. $u \not\stackrel{M}{\sim} v$ for any $(u, v) \in \mathcal{U} \times \mathcal{V}$.

Lemma 20 (Atypicality). *Given some $|\tau| \leq I_{XY}$ and m a partial mapping fully contained in the true mapping M , $\Pr [\langle \mathbf{G} - \tau, \mathbf{m} \rangle \leq 0 | m \subseteq M] \leq \exp \left(-|m| \cdot \frac{(I_{XY} - \tau)^2}{4I_{XY}} \right)$.*

Proof. The atypicality event is completely independent from users that are not contained in m . Then, without loss of generality, we can assume $m = M$ instead of $m \subseteq M$.

By Corollary 11,

$$\Pr [\tau | m| \geq \langle \mathbf{G}, \mathbf{m} \rangle | M = m] \leq \exp(\theta \tau | m|) R((1 - \theta)\mathbf{m}).$$

By Corollary 15, this last expression is upper bounded by $\exp(\theta \tau | m| - \theta(1 - \theta)I_{XY}| m|)$.

Let $\theta = \frac{I_{XY} - \tau}{2I_{XY}}$. Then $\tau \theta = \frac{I_{XY}\tau - \tau^2}{2I_{XY}}$ and $\theta(1 - \theta)I_{XY} = \frac{I_{XY}^2 - \tau^2}{4I_{XY}}$, which gives us

$$\begin{aligned} & \Pr [\tau | m| \geq \langle \mathbf{G}, \mathbf{m} \rangle | M = m] \\ & \leq \exp(\tau \theta | m| - \theta(1 - \theta)I_{XY}| m|) \\ & = \exp\left(-|m| \cdot \frac{I_{XY}^2 - 2I_{XY}\tau + \tau^2}{4I_{XY}}\right) \end{aligned}$$

which matches the claimed result. \square

Lemma 21 (False positive probability). *Given some $|\tau| \leq I_{XY}$, $\Pr \left[G_{u,v} \geq \tau | u \not\stackrel{M}{\sim} v \right] \leq \exp \left(-\frac{(I_{XY} + \tau)^2}{4I_{XY}} \right)$.*

Proof. By Corollary 11,

$$\Pr \left[G_{u,v} \geq \tau | u \not\stackrel{M}{\sim} v \right] \leq e^{\theta \tau} R([\theta]).$$

By Corollary 15, $R([\theta]) \leq \exp(-\theta(1-\theta)I_{XY})$.

$$\begin{aligned} & \Pr \left[G_{u,v} \geq \tau | u \stackrel{M}{\not\sim} v \right] \\ & \leq \exp(-\tau\theta - \theta(1-\theta)I_{XY}) \\ & = \exp\left(-\frac{I_{XY}^2 + 2I_{XY}\tau + \tau^2}{4I_{XY}}\right) \end{aligned}$$

which matches the claimed result. \square

Lemma 22 (Misalignment). *Let m and m' denote two mappings of same size and δ denote the number of pairs mapped by m but not by m' . Then $\Pr[\langle \mathbf{G}, \mathbf{m} \rangle \leq \langle \mathbf{G}, \mathbf{m}' \rangle | M = m] \leq \exp(-\frac{\delta}{2}I_{XY})$.*

Proof. By Corollary 11, $\Pr[\langle \mathbf{G}, \mathbf{m} \rangle \leq \langle \mathbf{G}, \mathbf{m}' \rangle | M = m]$ is upper bounded by $R(\Theta)$ where $\Theta = \frac{1}{2}\mathbf{m} + \frac{1}{2}\mathbf{m}'$. Θ can be represented in block-diagonal form: For each $n - \delta$ pair of users (u, v) that is mapped both by \mathbf{m} and \mathbf{m}' , we get a 1×1 block $[1]$. The remaining δ pairs whose mapping is not the same between \mathbf{m} and \mathbf{m}' , we get blocks Θ_j that correspond to cycles or even paths as described in Definition 3.

Given this block-diagonal form, by Lemma 13, $R(\Theta)$ is equal to the product $[R([1])]^{n-\delta} \cdot \prod_j R(\Theta_j)$. By Lemma 14, $R([1]) = 1$. By Lemmas 17 and 19, plugging in $\nu = 1$, we have $R(\Theta_j) \leq \exp\left(-\frac{\delta_j I_{XY}}{2}\right)$ where δ_j is the total length (i.e. number of blue edges or number of red edges) of the corresponding cycle or even paths. The total number of user pairs that whose mapping differs between \mathbf{m} and \mathbf{m}' is δ . Then $\prod_j R(\Theta_j) \leq \exp\left(\sum_j -\frac{\delta_j I_{XY}}{2}\right) = \exp\left(-\frac{\delta I_{XY}}{2}\right)$.

It then follows that $\Pr[\langle \mathbf{G}, \mathbf{m} \rangle \leq \langle \mathbf{G}, \mathbf{m}' \rangle | M = m] \leq R(\Theta) \leq \exp\left(-\frac{\delta I_{XY}}{2}\right)$. \square

Lemma 23 (Misalignment-despite-typicality). *Let m and m' denote two mappings of same*

size δ such that no pair is mapped under both mappings. Given $0 \leq \tau \leq I_{XY}$,

$$\Pr [\langle \mathbf{G}, \mathbf{m} \rangle \geq \tau |m| \text{ and } \langle \mathbf{G}, \mathbf{m}' \rangle \geq \langle \mathbf{G}, \mathbf{m} \rangle \mid m \subseteq M] \leq \exp \left(-\delta \cdot \frac{I_{XY}^2 + \tau^2}{2I_{XY}} + \delta \cdot 6\rho_{\max}^2 \tau \right)$$

where $\rho_{\max} = \max |\rho_i|$, the largest correlation coefficient under the canonical form.

By Lemma 9, under Condition 4, $\rho_{\max}^2 \leq o(1)$, so the bound can be simplified as $\exp \left(-\delta \cdot \frac{I_{XY}^2 + \tau^2}{2I_{XY}} (1 - o(1)) \right)$.

Proof. By Corollary 11, $\Pr [\langle \mathbf{G}, \mathbf{m} \rangle \geq \tau |m| \text{ and } \langle \mathbf{G}, \mathbf{m}' \rangle \geq \langle \mathbf{G}, \mathbf{m} \rangle \mid m \subseteq M]$ is upper bounded by $e^{-\tau |m|(\nu-1)} R(\Theta)$ where $\Theta = \nu(1-\theta)\mathbf{m} + \nu\theta\mathbf{m}'$. Consider the decomposition of Θ into blocks: we get cycles and even paths as described in Definition 3. (There are no one-by-one blocks since m and m' have no intersection.) This decomposition gives us a block-diagonal representation of Θ . By Lemma 13, $R(\Theta) = \prod R(\Theta_j)$, where Θ_j denotes the block corresponding to an elementary misalignment (i.e. cycle or even path).

Let m_j and m'_j denote the partial misalignments that correspond to the intersection of block Θ_j with m and m' . Let $\delta_j = |m_j| = |m'_j|$ denote their size. (Cycle and even path type misalignment consist of mappings of equal size.) Under Condition 4, Lemmas 17 and 19 give us

$$\begin{aligned} R(\nu(1-\theta)\mathbf{m}_j + \nu\theta\mathbf{m}'_j) &\leq \exp \left(-\delta_j \cdot \frac{I_{XY}}{2} \nu(2-\nu) + 6\delta_j \rho_{\max}^2 I_{XY} \right) \\ \implies R(\nu(1-\theta)\mathbf{m} + \nu\theta\mathbf{m}') &\leq \exp \left(-\delta \cdot \frac{I_{XY}}{2} \nu(2-\nu) + 6\delta \rho_{\max}^2 I_{XY} \right) \end{aligned}$$

Pick $\nu = 1 + \frac{\tau}{I_{XY}}$. Then

$$R(\nu(1-\theta)\mathbf{m} + \nu\theta\mathbf{m}') \leq \exp \left(-\delta \cdot \frac{I_{XY}^2 - \tau^2}{2I_{XY}} + \delta \cdot 6\rho_{\max}^2 \tau \right)$$

and $e^{-\tau\delta(\nu-1)} = \exp \left(-\delta \cdot \frac{\tau^2}{I_{XY}} \right)$. Then $e^{-\tau |m|(\nu-1)} R(\Theta) \leq \exp \left(-\delta \cdot \frac{I_{XY}^2 + \tau^2}{2I_{XY}} + \delta \cdot 6\rho_{\max}^2 \tau \right)$

and we have the claimed result. \square

4.3.2 Geometric intuition behind concentration inequalities

For Gaussian database alignment in the low per-feature correlation regime, by Lemma 50, entries corresponding to true pairs in \mathbf{G} asymptotically have mean I_{XY} and false pairs have mean $-I_{XY}$. All entries have variance $2I_{XY}$.

Compare this to Gaussian planted matching, where given $\mathbf{W}_G = \mu\mathbf{W} - \mu^2/2$ a scaled and shifted version of the original edge weight matrix \mathbf{W} , entries corresponding to true pairs in \mathbf{W}_G have mean $\mu^2/2$ and false pairs have mean $-\mu^2/2$. All entries have variance μ^2 .

For the rest of the section, we use ζ to refer to I_{XY} in the context of Gaussian database alignment, which is analogous to $\mu^2/2$ in the context of planted matching. Then true pairs have mean ζ , false pair have mean $-\zeta$, and all pairs have variance 2ζ in both \mathbf{G} and \mathbf{W}_G .

We want to bound the measure of the probability spaces that correspond to each type of error event. Consider the probability space $\mathbb{R}^{\mathcal{U}\times\mathcal{V}}$. A two-dimensional projection of this space is given in Fig. 4.1. Note the mean point $(\zeta, -\zeta)$.

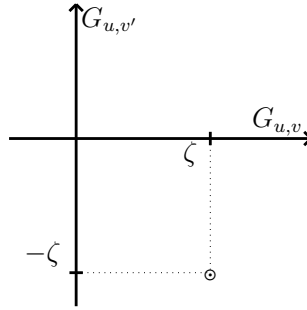


Figure 4.1: 2-dimensional projection of probability space. $G_{u,v}$ corresponds to a true pair and has mean ζ , while $G_{u,v'}$ corresponds to a false pair and has mean $-\zeta$.

As shown in Section 2.2, the objective function for all three algorithms is to maximize a linear combination of a shifted version of \mathbf{G} or \mathbf{W}_G . All concentration equalities we use are bounds on the measures of half-spaces in the probability space $\mathbb{R}^{\mathcal{U}\times\mathcal{V}}$. Approximating the entries of \mathbf{G} as independent normal random variables with appropriate statistics (i.e. $(\mu, \sigma^2) = (\zeta, 2\zeta)$ for true pairs and $(\mu, \sigma^2) = (-\zeta, 2\zeta)$ for false pairs), we are able to get

quick approximations for the bounds on half-spaces using the Chernoff bound. Specifically, the probability of a half-space is bounded by $\exp\left(-\frac{\ell^2}{2\sigma^2}\right)$ where ℓ denotes the separation between the half-space and the mean point and $\sigma^2 = 2\zeta$ is the variance of the terms. These bounds hold exactly in the planted matching case since entries of \mathbf{W}_G are indeed independent normal random variables.

True pair failing threshold testing

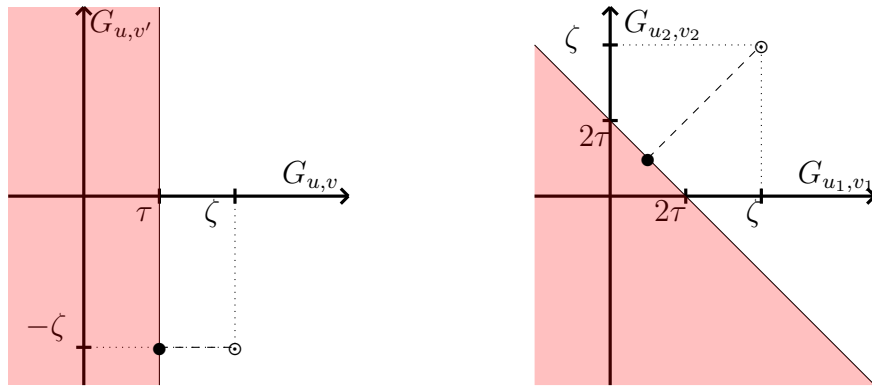


Figure 4.2: 2-dimensional projection of the regions of the probability space corresponding to the event $\{G_{u,v} \leq \tau\}$ (left-hand side) and the event $\{G_{u_1,v_1} + G_{u_2,v_2} \leq 2\tau\}$ (right-hand side).

The left-hand side of Fig. 4.2 illustrates the half-space corresponding to $\{G_{u,v} \leq \tau\}$. The separation between the half-space and the mean point is equal to $\ell = \zeta - \tau$. Then, the Chernoff bound gives us $\exp\left(-\frac{\ell^2}{2\sigma^2}\right) = \exp\left(-\frac{(\zeta - \tau)^2}{4\zeta}\right)$, which exactly matches the statement in Lemma 20.

Similarly, the right-hand side of Fig. 4.2 illustrates the half-space corresponding to the case with 2 true pairs: $\{G_{u_1,v_1} + G_{u_2,v_2} \leq 2\tau\}$. The separation between the half-space and the mean point is equal to $\ell = (\zeta - \tau)\sqrt{2}$. Then, the Chernoff bound gives us $\exp\left(-\frac{\ell^2}{2\sigma^2}\right) = \exp\left(-2 \cdot \frac{(\zeta - \tau)^2}{4\zeta}\right)$, which exactly matches the statement in Lemma 20.

This argument can be generalized to an arbitrary number of true pairs.

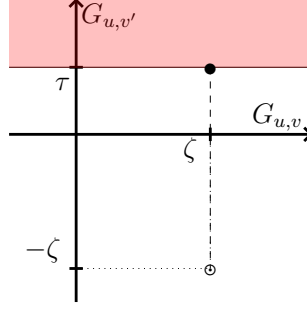


Figure 4.3: 2-dimensional projection of the regions of the probability space corresponding to the event $\{G_{u,v'} \geq \tau\}$

False pair passing threshold testing

Fig. 4.3 illustrates the half-space corresponding to $\{G_{u,v'} \geq \tau\}$. The separation between the half-space and the mean point is equal to $\ell = \zeta + \tau$. Then, the Chernoff bound gives us $\exp\left(-\frac{\ell^2}{2\sigma^2}\right) = \exp\left(-\frac{(\zeta+\tau)^2}{4\zeta}\right)$, which exactly matches the statement in Lemma 21.

Misalignment

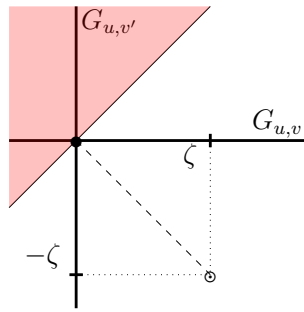


Figure 4.4: 2-dimensional projection of the half-space corresponding to the event $\{G_{u,v} \leq G_{u,v'}\}$.

Fig. 4.4 illustrates the half-space corresponding to $\{G_{u,v} \leq G_{u,v'}\}$. The separation between the half-space and the mean point is equal to $\ell = \zeta\sqrt{2}$. Then, the Chernoff bound gives us $\exp\left(-\frac{\ell^2}{2\sigma^2}\right) = \exp\left(-\frac{\zeta}{4}\right)$, which exactly matches the statement in Lemma 22.

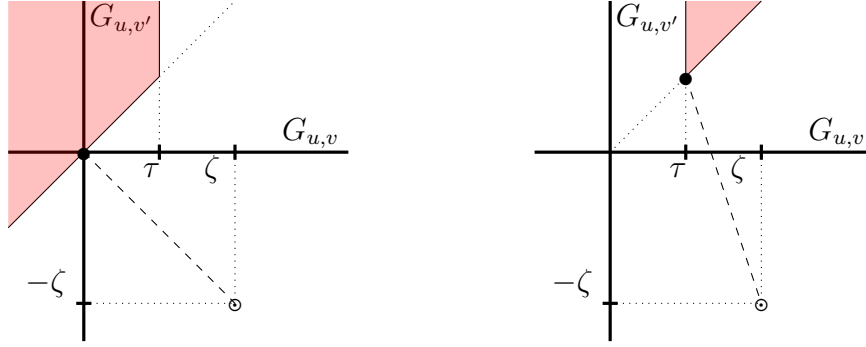


Figure 4.5: Partition of the half-space in Fig. 4.4 according to whether information density $G_{u,v}$ of the true pair (u, v) is above or below the threshold τ .

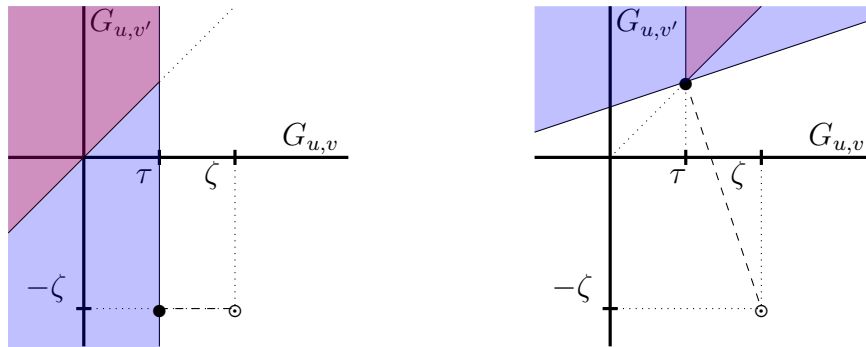


Figure 4.6: Blue-purple half-spaces containing the purple ‘slices’ from Fig. 4.5.

Misalignment despite typicality

The misalignment half-space shown in Fig. 4.4 can be broken down into two cases based on whether or not the true pairs have high enough average score. This is illustrated in Fig. 4.5. These ‘slices’ of a half-space can then be covered by another set of half-spaces, as illustrated in Fig. 4.6. In both figures, the left-hand side corresponds to the atypicality event and the right-hand side corresponds to the misalignment-despite-typicality event.

It can be shown that, the half-space in the right-hand side of Fig. 4.6 is at distance $(\sqrt{2\zeta^2 + 2\eta^2})$ to the mean point. Then, the Chernoff bound gives us $\exp\left(-\frac{\ell^2}{2\sigma^2}\right) = \exp\left(-\frac{\zeta^2 + \tau^2}{2\zeta}\right)$, which exactly matches the statement in Lemma 8, and asymptotically matches that in Lemma 23.

These figures also help demonstrate the contribution of this approach in analysis: Both the original misalignment half-space in Fig. 4.4 as well as the misalignment-despite-

typicality half-space in the right-hand side of Fig. 4.6 change based on the choice of (u, v') . The separation between half-spaces is greater with the misalignment-despite-typicality event, which gives us some improvement in the error bound at the cost of having to consider the atypicality error shown in the left-hand side of Fig. 4.6. This last half-space however, does not depend on (u, v') . The atypicality half-space is fixed once we pick (u, v) . Therefore this term does not require taking a union bound.

For an appropriate choice of τ , the gains made by the improvement from the misalignment half-space to the misalignment-despite-typicality half-space can compensate for the cost of having to consider the atypicality half-space. This is thanks to the fact that, unlike misalignment and misalignment-despite-typicality, we need not need to reconsider atypicality for every choice of (u, v') .

4.4 Achievability

We substitute I_{XY} with ζ in our statements of results to have statements that explicitly show the parallels between Gaussian database alignment in the low per-feature correlation regime and Gaussian planted matching.

In Subsection 4.4.1, we state asymptotic results which give bounds on the signal strength ζ that guarantee upper bounds on the expected number of errors as n , the size of the matching, goes to infinity. In Subsection 4.4.2, we state finite-size results that bound the expected number of errors as a function of the signal strength ζ for matchings of any finite size. The proofs for these results are described in Subsections 3.3.3 (threshold testing), 3.3.4 (maximum row estimation) and 3.3.5 (maximum likelihood estimation). The analyses for Gaussian planted matching in the low per-feature correlation regime very closely follow the analyses for Gaussian planted matching, with some adjustments due to differences between the concentration inequalities for Gaussian database alignment in the low per-feature correlation regime given in Section 4.3 and those for Section 3.2.

where we perform different analyses to derive the expressions for the bounds of various

forms.

Recall that, as defined in Subsection 2.2.4, $d(\hat{M}, M) \triangleq \frac{1}{2} \|\hat{M} - M\|_F^2$. For \hat{M}_{ML} and \hat{M}_{MR} this distance gives the number of users in \mathcal{U} mapped to a false match in \mathcal{V} . For \hat{M}_{MR} , the distance gives the arithmetic mean of the number of false positives and false negatives.

4.4.1 Asymptotic results

In the asymptotic case, as the size of the matching n goes to infinity, we say an algorithm achieves exact alignment if $\mathbb{E}[d(\hat{M}, M)] \leq o(1)$, and almost-exact alignment if $\mathbb{E}[d(\hat{M}, M)] \leq o(n)$.

Theorem 4.4.1 (Exact alignment and almost-exact alignment recovery). *Let $n = |M| = |\mathcal{U}|$. Define $\alpha = \frac{\log(|\mathcal{V}| - n)}{\log n}$ if $|\mathcal{V}| > n$ and $\alpha = 0$ if $|\mathcal{V}| = n$. Consider the Gaussian database alignment problem.*

- Lower bounds on ζ as given in Table 4.1 are sufficient conditions for almost-exact alignment for different cases¹.
- $\zeta \geq c \log n + \omega(1)$ is a sufficient condition for exact alignment, where the values of c for different cases are given in Table 4.2.

Table 4.1: Sufficient conditions for almost-exact alignment for Gaussian database alignment. Values in the table are lower bounds on ζ that guarantee almost-exact alignment for different algorithms, for different sizes of $|\mathcal{V}|$ and different conditions on ρ_{max}^2 .

Algorithm	Threshold.	Max row	Max likelihood	
Cond. on ρ_{max}^2	No cond.	No cond.	Cond. 5	Condition 4
$ \mathcal{V} \leq n + o(n)$	$\log n + \omega(\sqrt{\log n})$		$(1 + \Omega(1)) \log n$	
$ \mathcal{V} \geq n + \Omega(n)$	$\alpha \log n + \omega(\sqrt{\log n})$		$(\alpha + \Omega(1)) \log n$	

The boundaries for exact and almost-exact alignment for the algorithms are illustrated in Fig. 3.2. Boundaries for maximum likelihood and maximum row algorithms completely overlap for $\zeta / \log n \geq 4$.

¹No condition on ρ_{max}^2 is necessary for any of the conditions for threshold testing given in Table 4.1, including the condition $\zeta \geq (1 + \sqrt{1 + \alpha})^2 \log n + \omega(1)$ for the case where $|\mathcal{V}| \geq n + \Omega(n)$.

Table 4.2: Sufficient conditions for exact alignment for Gaussian database alignment. Sufficient conditions are of the form $\zeta \geq c \log n + \omega(1)$, where the value of c is given in the table for different algorithms, for different sizes of $|\mathcal{V}|$ and different conditions on ρ_{\max}^2 .

Size of \mathcal{V}	$ \mathcal{V} = n$	$ \mathcal{V} \leq n + o(n)$	$ \mathcal{V} \geq n + \Omega(n)$	
Cond. on ρ_{\max}^2	No cond.	No condition	Cond. 5	Condition 4
Threshold.	$(1 + \sqrt{2})^2$	$(1 + \sqrt{2})^2$	$(1 + \sqrt{1 + \alpha})^2$	
Max row	4	4	$(1 + \sqrt{\alpha})^2$	$(1 + \sqrt{\alpha})^2 + \Omega(1)$
Max likelihood	2	$2(\alpha + 1)$	$(1 + \sqrt{\alpha})^2$	$(1 + \sqrt{\alpha})^2 + \Omega(1)$

The sufficient conditions in Theorem 4.4.1 (Table 4.2) for database alignment fully match those in Theorem 3.3.1 (Table 3.2), except for the case where we consider . The conditions fully match for threshold testing for the case where $|\mathcal{V}| \leq n + o(n)$ and for the case where Condition 5 holds, i.e. $\rho_{\max}^2 \leq o(1/\log n)$. (We only fail to show matching conditions when Condition 5 does not hold and $|\mathcal{V}| \geq n + \Omega(n)$, and only for maximum row estimation and maximum likelihood estimation.)

Theorem 4.4.2 (Expected number of errors in the balanced case). *Let $n = |M| = |\mathcal{U}| = |\mathcal{V}|$. Let $\beta > 0$ some constant. For Gaussian database alignment, as $n \rightarrow \infty$, lower bounds on $\zeta/\log n$ that guarantee error bounds $\mathbb{E}[d(\hat{M}, M)] \leq n^{1-\beta+o(1)}$, and whether or not they require conditions on ρ_{\max}^2 , are given in Table 4.3:*

From top to bottom, the forms of the boundaries given by the thresholds in Theorem 4.4.2: (Threshold testing) parabolic, (maximum row estimation) parabolic, linear, (maximum likelihood estimation) elliptic, vertical, linear. The boundaries for the achievability regions of the algorithms are illustrated in Fig. 3.3.

Theorem 4.4.3 (Expected number of errors in unbalanced case). *Let $n = |M| = |\mathcal{U}|$ and $|\mathcal{V}| > n$. Define² $\nu \triangleq \frac{\log |\mathcal{V}|}{\log n}$ and $\alpha \triangleq \frac{\log(|\mathcal{V}| - n)}{\log n}$, and let these be constant functions of n . Let $\beta > 0$ some constant. For Gaussian database alignment, as $n \rightarrow \infty$, lower bounds on $\zeta/\log n$ that guarantee error bounds $\mathbb{E}[d(\hat{M}, M)] \leq n^{1-\beta+o(1)}$, and whether or not they require conditions on ρ_{\max}^2 , are given in Table 4.3:*

²Asymptotically, ν can be replaced by $\max\{1, \alpha\}$ as $1 \leq \frac{\nu}{\max\{1, \alpha\}} \leq 1 + \Omega(1/\log n)$.

Table 4.3: Sufficient conditions for error bounds for Gaussian database alignment in the balanced case, i.e. $|\mathcal{V}| = |\mathcal{U}| = |M| = n$. Sufficient conditions are expressed as lower bounds on $\zeta/\log n$. Different conditions are relevant for different ranges of β . If it is noted ‘no’ on the 3rd column, then the error bound on the 4th column does not require any condition on ρ_{\max}^2 . Otherwise, the bounds on the 4th and 5th columns correspond to bounds under Cond. 5 and Cond. 4 respectively.

Bound on $\zeta/\log n$	Range of β	Require condition?	Bound on $\mathbb{E}[d(\hat{M}, M)]$	
			Cond. 5	Cond. 4
Threshold testing				
$(\sqrt{\beta} + \sqrt{1 + \beta})^2$	$0 < \beta$	no	$n^{1-\beta}$	
Maximum row estimation				
$(1 + \sqrt{\beta})^2$	$0 < \beta \leq 1$	yes	$2n^{1-\beta}(1 + o(1))$	$\mathcal{O}(n^{1-\beta})$
$2(1 + \beta)$	$1 < \beta$	no	$n^{1-\beta}$	
Maximum likelihood estimation				
$1 + 2\sqrt{\beta(1 - \beta)}$	$0 < \beta < 1/2$	yes	$\mathcal{O}(n^{1-\beta})$	$n^{1-\beta+o(1)}$
$2 + \frac{2 \log(\frac{\sqrt{5}-1}{2})}{\log n}$	$1/2 < \beta \leq 1$	no	$n^{1-\beta}$	
$1 + \beta$	$1 < \beta$	no	$n^{1-\beta}(1 + o(1))$	

From top to bottom, the forms of the boundaries given by the thresholds in Theorem 4.4.2: (Threshold testing) parabolic, (maximum row estimation) parabolic, linear, (maximum likelihood estimation) elliptic, parabolic, vertical, linear. The boundaries for the achievability region of the maximum likelihood estimator for various values of α are illustrated in Fig. 3.4. Boundaries for $\alpha = 1.5$ and $\alpha = 1.25$ are tangent to $y = 1$ at $x = 1.5$ and $x = 1.25$ respectively, while all other boundaries are tangent to that line at $x = 1$. These match the almost-exact alignment threshold.

The boundaries for the achievability region of maximum likelihood/maximum row estimation, which coincide, and for thresholding at $\alpha = 1.5$ are illustrated in Fig. 3.5. Both boundaries are tangent to $y = 1$ at $x = 1.5$. This matches the almost-exact alignment threshold.

4.4.2 Finite-size results

Theorem 4.4.4 (Expected number of errors for finite-size case). *Define $n \triangleq |M|$, $x \triangleq \frac{\zeta}{\log n}$, $\nu \triangleq \frac{\log |\mathcal{V}|}{\log n}$ and $\varepsilon \triangleq n^{1-x/2}$. Suppose $|\mathcal{U}| = n$. Then, $\log \mathbb{E} [d(\hat{M}, M)]$, the log of the*

Table 4.4: Sufficient conditions for error bounds for Gaussian database alignment in the unbalanced case, i.e. $|\mathcal{V}| > |\mathcal{U}| = |M| = n$. Sufficient conditions are expressed as lower bounds on $\zeta/\log n$. Different conditions are relevant for different ranges of β . If it is noted ‘no’ on the 3rd column, then the error bound on the 4th column does not require any condition on ρ_{\max}^2 . Otherwise, the bounds on the 4th and 5th columns correspond to bounds under Cond. 5 and Cond. 4 respectively.

Bound on $\zeta/\log n$	Range of β	Require condition?	Bound on $\mathbb{E}[d(\hat{M}, M)]$	
			Cond. 5	Cond. 4
Threshold testing				
$(\sqrt{\nu + \beta} + \sqrt{\beta})^2$	$0 < \beta$	no	$n^{1-\beta}$	
Maximum row estimation				
$(\sqrt{\nu} + \sqrt{\beta})^2$	$0 < \beta \leq \max\{\alpha, 1\}$	yes	$2n^{1-\beta}(1 + o(1))$	$\mathcal{O}(n^{1-\beta})$
$2(\nu + \beta)$	$\max\{\alpha, 1\} < \beta$	no	$n^{1-\beta}$	
Maximum likelihood estimation				
$1 + 2\sqrt{\beta(1-\beta)}$	$0 < \beta \leq \min\{1-\alpha, 1/2\}$	yes	$\mathcal{O}(n^{1-\beta})$	$n^{1-\beta+o(1)}$
$(\sqrt{\alpha} + \sqrt{\beta})^2$	$1-\alpha < \beta \leq \alpha$	yes	$\mathcal{O}(n^{1-\beta})$	$n^{1-\beta+o(1)}$
$2 + \frac{2\log\left(\frac{3+\sqrt{5}}{2}\right)}{\log n}$	$1/2 < \beta \leq 1-\alpha$	no	$n^{1-\beta} + \mathcal{O}(1)$	
$2(\alpha + \beta)$	$\max\{\alpha, 1-\alpha\} < \beta$	no	$n^{1-\beta}(1 + o(1))$	

expected number of errors is bounded as given in Table 4.4.4

Table 4.5: Error bounds for Gaussian database alignment with finite-size

Error bound	Condition
Threshold testing	
$\left(1 - \frac{(x-\nu)^2}{4x}\right) \log n$	$\nu < x$
Maximum row estimation	
$(1 - (\sqrt{\nu} - \sqrt{x})^2) \log n + \log 2$	$\nu < x < 4\nu$
$(1 + \nu - \frac{x}{2}) \log n$	$4\nu \leq x$
Maximum likelihood estimation	
$2 \log \varepsilon - \log(1 - \varepsilon)$	$x > 2$ and $ \mathcal{V} = n$
$\log \varepsilon - 2 \log(1 - \varepsilon) + \log(\mathcal{V} - n + 1)$	$x > 2$

4.4.3 Proof for Threshold Testing

Recall that, as described in Subsection 3.3.3, proofs for threshold testing and maximum row estimation state inequalities use the variable $\nu \triangleq \frac{\log |\mathcal{V}|}{\log n}$ instead of the variable $\alpha \triangleq$

$\frac{\log(|\mathcal{V}|-n)}{\log n}$. The asymptotic statements for threshold testing and maximum row estimation containing ν directly translate to the statements in the main results using the fact that $\max\{1, \alpha\} \leq \nu \leq \max\{1, \alpha\} \left(1 + \mathcal{O}\left(\frac{1}{\log n}\right)\right)$.

Parabolic boundary

Consider the following results for threshold testing, and introduce their proofs below:

Let $\tau = \log \frac{|\mathcal{U} \cdot |\mathcal{V}|}{|M|}$ be the value of the threshold used for threshold testing.

- Theorem 4.4.1

Almost-exact alignment is achieved if $\zeta \geq \nu \log n + \omega(\sqrt{\log n})$.

By Lemma 1, this condition also guarantees almost-exact alignment for maximum row estimation.

Exact alignment is achieved if $\zeta \geq (1 + \sqrt{1 + \nu})^2 \log n + \omega(1)$.

- Theorem 4.4.3

The expected number of false positive and of false negatives are both bounded by $n^{1-\beta}$ if $\zeta \geq (\sqrt{\nu + \beta} + \sqrt{\beta})^2 \log n$.

- Theorem 4.4.2

This is a special case of Theorem 4.4.3 with $\nu = 1$.

- Theorem 4.4.4

Given $x > \nu$, the expected number of errors is bounded by $n^{1 - \frac{(x-\nu)^2}{4x}}$.

Proof. The proof for database alignment is identical to the corresponding proof in Subsection 3.3.3 for planted matching when we replace the concentration inequality lemmas 5 and 6 for \mathbf{W}_G with the corresponding lemmas 20 and 21 for \mathbf{G} . □

4.4.4 Proofs for Maximum Row Estimation

Consider users $u \in \mathcal{U}$ and $v, v' \in \mathcal{V}$ such that $u \stackrel{M}{\sim} v$. We want to bound the probability of the error event where u is falsely mapped to v' . Under maximum row estimation, this corresponds to the event $G_{u,v} \leq G_{u,v'}$. Without loss of generality, assume the set \mathcal{U} consists of the single user u .

Linear boundary - relevant for large β and small α

Consider the following results for maximum row estimation, and introduce their proofs below:

- Theorem 4.4.1

Exact alignment is achieved if $\zeta \geq 2(1 + \nu) \log n + \omega(1)$.

- Theorem 4.4.3

The number of errors is bounded by $n^{1-\beta}$ if $\zeta \geq (\nu + \beta) \log n$.

- Theorem 4.4.2

This is a special case of Theorem 4.4.3 with $\nu = 1$.

- Theorem 4.4.4

Given $\nu < x$, the expected number of errors is bounded by $n^{1+\nu-\frac{x}{2}}$.

Proof. The proof for database alignment is identical to the corresponding proof in Subsection 3.3.4 for planted matching when we replace the concentration inequality Lemma 7 for \mathbf{W}_G with the corresponding Lemma 22 for \mathbf{G} . □

Parabolic boundary - relevant for small β and large α

Consider the following results for maximum row estimation, and introduce their proofs below:

Suppose $\zeta \leq 4 \log |\mathcal{V}|$.

- Theorem 4.4.1

Under Condition 4 exact alignment is achieved if $\zeta \geq \left[(1 + \sqrt{\nu})^2 + \Omega(1) \right] \log n$.

Under Condition 5 exact alignment is achieved if $\zeta \geq (1 + \sqrt{\nu})^2 \log n + \omega(1)$.

(The proof for the condition for almost-exact alignment for maximum row estimation follows the proof for threshold testing as a result of Lemma 1.)

- Theorem 4.4.3

The number of errors is bounded by $n^{1-\beta} \left(1 + n^{6\rho_{\max}^2(\nu-\beta)} \right)$ if $\beta \leq \nu$ and $\zeta \geq (\sqrt{\nu} + \sqrt{\beta})^2 \log n$. This bound is $n^{1-\beta+o(1)}$ under Condition 4 and $2n^{1-\beta}(1+o(1))$.

- Theorem 3.3.2

This is a special case of Theorem 3.3.3 with $\nu = 1$.

- Theorem 3.3.4

Given $\nu < x < 4\nu$, the expected number of errors is bounded by

$$n^{1-(\sqrt{\nu}-\sqrt{x})^2} \left(1 + n^{6\rho_{\max}^2(2\sqrt{\nu x}-x)}\right).$$

Proof. The proof for database alignment closely follows the the corresponding proof in Subsection 3.3.4 for planted matching when we replace the concentration inequality Lemma 5 for \mathbf{W}_G with the corresponding Lemma 20 for \mathbf{G} . We describe the other adjustments needed to make to the proof to get the claimed result.

The probability bound in Lemma 20 for database alignment differs from the probability bound in Lemma 5 for planted matching by an exponential factor of $6\delta\rho_{\max}^2\tau$. ($\delta = 1$ for this analysis.) Consequently we need to introduce this extra term to the log-ratio between the probability of the events $\{G_{u,v} \leq \tau\}$ (i.e. atypicality for u) and $\{\exists v' \in \mathcal{V} \setminus \{v\} : G_{u,v} \geq \tau \text{ and } G_{u,v} \leq G_{u,v'}\}$ (i.e. misalignment despite typicality for u). Then, picking $\tau = \tau^* \triangleq 2\sqrt{\zeta \log |\mathcal{V}|} - \zeta$ gives us a misalignment-despite-typicality probability bound that is $\exp(6\rho_{\max}^2\tau^*) = \exp(6\rho_{\max}^2(2\sqrt{\nu x} - x) \log n)$ times that of the atypicality bound, which can be further bounded by $\exp(6\rho_{\max}^2(\nu - \beta) \log n)$. Then, the bound on the total number of errors is $1 + n^{6\rho_{\max}^2(\nu - \beta)}$ times the atypicality bound, which is shown to be $n^{1-(\sqrt{\nu}-\sqrt{x})^2}$. By Lemma 9, under Condition 4, $\rho_{\max}^2 \leq o(1)$ so $1 + n^{6\rho_{\max}^2(\nu - \beta)} \leq 1 + n^{o(1)}$.

Then $x \geq (\sqrt{\nu} + \sqrt{\beta})^2$ achieves $n^{1-\beta} \left(1 + n^{6\rho_{\max}^2(\nu - \beta)}\right) \leq n^{1-\beta+o(1)}$ errors. To achieve exact alignment, we need $\beta - 1 - 6\rho_{\max}^2(\nu - \beta) \geq \omega(1/\log n)$, which is equivalent to $\beta \geq 1 + 6\rho_{\max}^2\nu + \omega(1/\log n)$.

Under Condition 4, $\beta \geq 1 + 6\rho_{\max}^2\nu + \omega(1/\log n)$ is dominated by the stronger bound $\beta \geq 1 + \Omega(1)$ since $\rho_{\max}^2 \leq o(1)$. Then, $x \geq (\sqrt{\nu} + \sqrt{\beta})^2$ is satisfied for such β if $\zeta \geq \left[(1 + \sqrt{\nu})^2 + \Omega(1)\right] \log n$.

Under Condition 5, $\beta \geq 1 + 6\rho_{\max}^2\nu + \omega(1/\log n)$ is equivalent to $\beta \geq \omega(1/\log n)$ since $\rho_{\max}^2 \leq o(1/\log n)$. Then, $x \geq (\sqrt{\nu} + \sqrt{\beta})^2$ is satisfied for such β if $\zeta \geq (1 + \sqrt{\nu})^2 \log n + \omega(1)$. \square

4.4.5 Proof for Maximum likelihood estimation

Linear boundary - relevant for large β and small α

Consider the following results for maximum likelihood estimation, and introduce their proofs below:

- Theorem 3.3.1

Exact alignment is achieved if $\zeta \leq 2 \log n + \omega(1)$ for $s = 0$, or if $\zeta \geq 2(1 + \alpha) \log n + \omega(1)$ for $s \geq 1$.

- Theorem 3.3.2

The expected number of errors is bounded by $n^{1-\beta}(1 + o(1))$ if $\zeta \geq (1 + \beta) \log n$ and $\beta > 1 + \omega(1/\log n)$.

If $\zeta \geq 2 \log n + 2 \log \left(\frac{1+\sqrt{5}}{2} \right)$, then the expected number of errors is bounded by 1, which is bounded by $n^{1-\beta}$ for any $\beta \leq 1$.

- Theorem 3.3.3

The expected number of errors is bounded by $n^{1-\beta}(1 + o(1))$ if $\zeta \geq 2(\alpha + \beta) \log n$ and $\beta > 1 + \omega(1/\log n)$.

If $\zeta \geq 2 \log n + 2 \log \left(\frac{3+\sqrt{5}}{2} \right)$, then the expected number of errors is bounded by $s + \sqrt{5} - 2$, which is bounded by $n^{1-\beta} + \mathcal{O}(1)$ for any $\beta \leq 1 - \alpha$.

- Theorem 3.3.4

Let $\varepsilon \triangleq \exp \left(\log n - \frac{\zeta}{2} \right)$.

Given $x > \nu$ and $s = 0$, the expected number of errors is bounded by $\frac{\varepsilon^2}{1-\varepsilon}$.

In general, the number of errors is bounded by $\frac{\varepsilon^2}{1-\varepsilon} + \frac{s\varepsilon}{(1-\varepsilon)^2}$, which is bounded by the simpler expression $\frac{(s+1)\varepsilon}{(1-\varepsilon)^2}$.

Proof. The proof for database alignment is identical to the corresponding proof in Subsec-

tion 3.3.5 for planted matching when we replace the concentration inequality Lemma 7 for \mathbf{W}_G with the corresponding Lemma 22 for \mathbf{G} . \square

Elliptic boundary - relevant for smallest β and small α

Consider the following results for maximum likelihood estimation, and introduce their proofs below:

- Theorem 3.3.1

Let $\alpha < 1$ some constant bounded away from 1.

Under Condition 4, then almost-exact alignment is achieved if $\zeta \geq (1 + \Omega(1)) \log n$.

Under Condition 4, then almost-exact alignment is achieved if $\zeta \geq \log n + \omega(1)$.

- Theorem 3.3.2 and Theorem 3.3.3

Let $\alpha \in (0, \infty)$ and $\beta \in [0, 1/2)$ some constants and suppose $\beta < \min\{1 - \alpha, 1/2\}$ and bounded away from both $1 - \alpha$ and $1/2$.

If $\zeta \geq \left(1 + 2\sqrt{\beta(1 - \beta)}\right) \log n$, then the number of errors is bounded by $\exp\left(\frac{1 + \log 2 - \log \varepsilon + 12\left(1 + 2\sqrt{\beta(1 - \beta)}\right)\rho_{\max}^2 \log n}{(1 - 2\beta)^2}\right) \cdot (n^{1 - \beta})$, where $\varepsilon < 1$ some constant bounded away from 1. This error bound is $n^{1 - \beta + o(1)}$ under Condition 4. It is $\exp\left(\frac{2(1 + \log 2) + \mathcal{O}(1)}{(1 - 2\beta)^2}\right) n^{1 - \beta}$ under Condition 5, which is $\mathcal{O}(n^{1 - \beta})$.

Proof. The proof for database alignment closely follows the the corresponding proof in Subsection 3.3.5 for planted matching when we replace the concentration inequality Lemma 5 for \mathbf{W}_G with the corresponding Lemma 20 for \mathbf{G} . We describe the other adjustments needed to make to the proof to get the claimed result.

The probability bound in Lemma 20 for database alignment differs from the probability bound in Lemma 5 for planted matching by an exponential factor of $6\delta\rho_{\max}^2\tau$. Then, the expression for τ^* needs to be adjusted to include the extra term $6\delta\rho_{\max}^2\tau$, which is bounded by $6\delta\rho_{\max}^2\zeta$: $\tau^* = \sqrt{2\zeta(\log n - \log \varepsilon + 6\rho_{\max}^2\zeta + 1 + \log 2) - \zeta^2}$. Then, the definition of η is also to be adjusted to be $\eta \triangleq 1 + \log 2 - \log \varepsilon + 6\rho_{\max}^2\zeta$. Then $x = 1 + 2\sqrt{b(1 - b)}$

guarantees an error bound of $n^{1-b} \cdot \exp\left(\frac{2\eta}{(1-2b)^2}\right)$. This makes the error bound equal to

$$n^{1-b} \cdot \exp\left(\frac{1 + \log 2 - \log \varepsilon + 12 \left(1 + 2\sqrt{b(1-b)}\right) \rho_{\max}^2 \log n}{(1-2b)^2}\right).$$

By Lemma 9, under Condition 4, $\rho_{\max}^2 \leq o(1)$, so $\rho_{\max}^2 \log n \leq o(\log n)$. Under Condition 5, $\rho_{\max}^2 \leq o(1/\log n)$, so $\rho_{\max}^2 \log n \leq o(1)$.

For almost-exact alignment, we need $x = 1 + 2\sqrt{b(1-b)}$ to be satisfied for some b such that $\beta = b - \frac{2\eta/\log n}{(1-2b)^2}$ satisfies $\beta \geq \omega(1/\log n)$. This is equivalent to $b \geq \frac{2\eta/\log n}{(1-2b)^2} + \omega(1/\log n)$.

Under Condition 4 $b \geq \frac{2\eta/\log n}{(1-2b)^2} + \omega(1/\log n)$ is dominated by $b \geq \Omega(1)$, since $\rho_{\max}^2 \leq o(1)$. Such b exists if $\zeta \geq (1 + \Omega(1)) \log n$.

Under Condition 5 $b \geq \frac{2\eta/\log n}{(1-2b)^2} + \omega(1/\log n)$ is equivalent to $b \geq \omega(1/\log n)$, since $\rho_{\max}^2 \leq o(1/\log n)$ and therefore $\eta \leq \mathcal{O}(1)$. Such b exists if $\zeta \geq \log n + \omega(1)$. \square

Parabolic boundary - relevant for small β and large α

Consider the following results for maximum likelihood estimation, and introduce their proofs below:

- Theorem 3.3.1

Suppose $\alpha \geq 1$. Under Condition 4, almost-exact alignment is achieved if $\beta \geq (\alpha + \Omega(1)) \log n$ and exact alignment is achieved if $\zeta \geq [(1 + \sqrt{\alpha})^2 + \Omega(1)] \log n$.

Under Condition 5, almost-exact alignment is achieved if $\beta \geq \alpha \log n + \omega(\sqrt{\log n})$ and exact alignment is achieved if $\zeta \geq (1 + \sqrt{\alpha})^2 \log n + \omega(1)$.

- Theorem 3.3.3

Given $1 - \alpha < \beta \leq \alpha$, the number of errors is bounded by $3e \cdot n^{1-\beta+7\rho_{\max}^2(\alpha-\beta)}$ if $\zeta \geq (\sqrt{\alpha} + \sqrt{\beta})^2 \log n$. This error bound is $n^{1-\beta+o(1)}$ under Condition 4 and $3e \cdot n^{1-\beta}(1 + o(1))$ under Condition 5, which is $\mathcal{O}(n^{1-\beta})$.

Proof. The proof for database alignment closely follows the the corresponding proof in

Subsection 3.3.5 for planted matching when we replace the concentration inequality Lemma 5 for \mathbf{W}_G with the corresponding Lemma 20 for \mathbf{G} . We describe the other adjustments needed to make to the proof to get the claimed result.

The probability bound in Lemma 20 for database alignment differs from the probability bound in Lemma 5 for planted matching by an exponential factor of $6\delta\rho_{\max}^2\tau$. Given the choice $\tau = \tau^* \triangleq 2\sqrt{\zeta(\log s + \log 3)} - \zeta$, the log-ratio between the bound on the expected number of atypicality events of size δ versus the bound on the expected number of misalignment-despite-typicality events of size δ is $-6\delta\rho_{\max}^2\tau^*$, which is negative. Then the expected number of atypicality events of size δ is bounded by the bound for the expected number of misalignment-despite-typicality events of size δ .

We adjust the choice of η to be $\eta = \frac{1+\log 3+6\rho_{\max}^2\tau^*}{\log n}$. It can be shown that $\tau^* = 2\sqrt{\zeta(\log s + \log 3)} - \zeta \leq \frac{7}{6}(\alpha - \beta) \log n$. Then η is on the order of $\frac{1}{\log n} + \rho_{\max}^2(\alpha - \beta)$.

Consequently, our error bound $n^{1-\beta+\eta}$ guaranteed by

$$x \geq \left(\sqrt{\alpha} + \sqrt{\beta}\right)^2 + \left(1 + \sqrt{\alpha/\beta}\right) \omega(1/\log n)$$

is itself bounded by $3e \cdot n^{1-\beta+7\rho_{\max}^2(\alpha-\beta)}$.

By Lemma 9, under Condition 4, $\rho_{\max}^2 \leq o(1)$, $n^{1-\beta+7\rho_{\max}^2(\alpha-\beta)} \leq n^{1-\beta+o(1)}$. Under Condition 5, $\rho_{\max}^2 \leq o(1/\log n)$, and therefore $n^{1-\beta+7\rho_{\max}^2(\alpha-\beta)} \leq \mathcal{O}(n^{1-\beta})$.

To achieve exact alignment, we need $\beta \geq 1 + \eta + \omega(1/\log n)$.

Under Condition 4, $\beta \geq 1 + \eta + \omega(1/\log n)$ is dominated by the stronger bound $\beta \geq 1 + \Omega(1)$ since η is on the order of $\frac{1}{\log n} + \rho_{\max}^2(\alpha - \beta)$ and $\rho_{\max}^2 \leq o(1)$. To achieve almost-exact alignment, we need $\beta \geq \eta + \omega(1/\log n)$, which is dominated by $\beta \geq \Omega(1)$ by the same reason. $x \geq \left(\sqrt{\alpha} + \sqrt{\beta}\right)^2 + \left(1 + \sqrt{\alpha/\beta}\right) \omega(1/\log n)$ is satisfied for some $\beta \geq 1 + \Omega(1)$ if $\zeta \geq \left[(1 + \sqrt{\alpha})^2 + \Omega(1)\right] \log n$. That condition on x is satisfied for some $\beta \geq \Omega(1)$ if $\zeta \geq (\alpha + \Omega(1)) \log n$.

Under Condition 5, $\beta \geq 1 + \eta + \omega(1/\log n)$ is equivalent to $\beta \geq 1 + \omega(1/\log n)$

since η is on the order of $\frac{1}{\log n} + \rho_{\max}^2(\alpha - \beta)$ and $\rho_{\max}^2 \leq o(1/\log n)$. To achieve almost-exact alignment, we need $\beta \geq \eta + \omega(1/\log n)$, which is equivalent to $\beta \geq \omega(1/\log n)$ by the same reason. $x \geq (\sqrt{\alpha} + \sqrt{\beta})^2 + \left(1 + \sqrt{\alpha/\beta}\right) \omega(1/\log n)$ is satisfied for some $\beta \geq 1 + \omega(1/\log n)$ if $\zeta \geq (1 + \sqrt{\alpha})^2 \log n + \omega(1)$. That condition on x is satisfied for some $\beta \geq \omega(1/\log n)$ if $\zeta \geq \alpha \log n + \omega(\sqrt{\log n})$. \square

4.5 Exact alignment converse

We have a converse for exact alignment that exactly matches the achievability result in Theorem 4.4.1.

Theorem 4.5.1. *Consider a pair of correlated Gaussian databases where mutual information is distributed equally across each dimension as described in Condition 2. Under Condition 4, if $I_{XY} \leq (2 - \Omega(1)) \log n$, then any for algorithm, the probability of returning the proper alignment is $o(1)$.*

Lemma 24. *Suppose $|\mathcal{U}| = |\mathcal{V}| = |M|$. Let m, m' denote two mappings of size n that agree everywhere except on two mapped pairs. Let $\vec{\rho} \in (-1, 1)^{\mathcal{D}}$ denote the correlation vector under canonical form (as described in Section 2.4) and suppose the value of $|\rho_i|$ is the same for all i . Under Condition 4, if $I_{XY} \geq \Omega(\log n)$,*

$$\Pr(\langle \mathbf{G}, \mathbf{m} \rangle \leq \langle \mathbf{G}, \mathbf{m}' \rangle) \geq \exp(-I_{XY}(1 + o(1))).$$

Proof. Recall that, as defined in Subsection 2.2.1 $G_{u,v} \triangleq \log \frac{f_{XY}(\vec{A}(u), \vec{B}(v))}{f_X(\vec{A}(u))f_Y(\vec{B}(v))}$, where f_X and f_Y denote the marginal probability density functions of features in \vec{A} and \vec{B} respectively, and f_{XY} denotes the probability density function of the joint distribution of a pair of correlated features. Assume features are given in canonical form as described in . Then f_X, f_Y and f_{XY} can each be decomposed into the product of distributions marginalized over each dimension $i \in \mathcal{D}$. Specifically, we can define dimension-wise distribution functions

f_{X_i} , f_{Y_i} and $f_{X_i Y_i}$ such that

$$\begin{aligned} f_X(\vec{x}) &= \prod_{i \in \mathcal{D}} f_{X_i}(x_i) = \prod_{i \in \mathcal{D}} \frac{1}{\sqrt{2\pi}} e^{-x_i^2/2} \\ f_Y(\vec{y}) &= \prod_{i \in \mathcal{D}} f_{Y_i}(y_i) = \prod_{i \in \mathcal{D}} \frac{1}{\sqrt{2\pi}} e^{-y_i^2/2} \\ f_{XY}(\vec{x}, \vec{y}) &= \prod_{i \in \mathcal{D}} f_{X_i Y_i}(x_i, y_i) = \prod_{i \in \mathcal{D}} \frac{1}{2\pi\sqrt{1-\rho_i^2}} \exp\left(-\frac{x_i^2 + y_i^2 - 2\rho_i x_i y_i}{2(1-\rho_i^2)}\right) \end{aligned}$$

It then follows that we can decompose $G_{u,v}$ into the sum of terms corresponding to each dimension:

$$G_{u,v} = \sum_{i \in \mathcal{D}} (G_i)_{u,v} = \sum_{i \in \mathcal{D}} \frac{f_{X_i Y_i}(A_i(u), B_i(v))}{f_X(A_i(u)) f_Y(B_i(v))},$$

which allows us to define matrices \mathbf{G}_i taking values in $\mathbb{R}^{\mathcal{U} \times \mathcal{V}}$ for all $i \in \mathcal{D}$. This allows for the decomposition of the matrix $\mathbf{G} = \sum_i \mathbf{G}_i$.

Fix $\rho_i = \rho$. Let $z \triangleq |\mathcal{D}|$.

be a function of I_{XY} and ρ : $z = -\frac{2I_{XY}}{\log(1-\rho^2)}$.

Let R_i denote the generating function (as given in Definition 6) for the single dimensional case. (The function is the same for all $i \in \mathcal{D}$ since the value of ρ_i^2 does not depend on i .) By Cramér's Theorem on the asymptotic tightness of the Chernoff bound (see for example [40]),

$$\begin{aligned} \lim_{z \rightarrow \infty} \frac{1}{z} \log \Pr \left(\sum_{i \in \mathcal{D}} \langle \mathbf{G}_i, \mathbf{m}' - \mathbf{m} \rangle \geq 0 \right) &= - \inf_{\theta > 0} \log \mathbb{E} \left[e^{\langle \mathbf{G}_i, \theta(\mathbf{m}' - \mathbf{m}) \rangle} \right] \\ &= - \inf_{\theta > 0} \log R_i((1-\theta)\mathbf{m} + \theta\mathbf{m}') \end{aligned}$$

$(1-\theta)\mathbf{m} + \theta\mathbf{m}'$ can be written in block diagonal form with $n-2$ 1×1 blocks [1] and a single block $\begin{bmatrix} 1-\theta & \theta \\ \theta & 1-\theta \end{bmatrix}$. By Lemma 13, the value of $R_i(\frac{\mathbf{m}+\mathbf{m}'}{2})$ can be written as the product of the R value for each block. By Lemma 14, $R_i([1]) = 1$ and by Lemma 16,

$$\inf \log R_i \left(\begin{bmatrix} 1 - \theta & \theta \\ \theta & 1 - \theta \end{bmatrix} \right) = \sqrt{1 - \rho_i^2}. \text{ Then}$$

$$\lim_{z \rightarrow \infty} \frac{1}{z} \log \Pr \left(\sum_{i \in \mathcal{D}} \langle \mathbf{G}_i, \mathbf{m}' - \mathbf{m} \rangle \geq 0 \right) = \frac{1}{2} \log (1 - \rho^2).$$

Then,

$$\log \Pr (\langle \mathbf{G}, \mathbf{m}' \rangle \geq \langle \mathbf{G}, \mathbf{m} \rangle) \geq \frac{z}{2} \log (1 - \rho^2) (1 - o(1)).$$

The $\frac{z}{2} \log (1 - \rho^2)$ is by definition equal to I_{XY} .

□

Lemma 25. *Suppose $|\mathcal{U}| = |\mathcal{V}| = |M|$. Let m, m', m'' denote three mappings of size n such that m' and m'' agree everywhere except on δ mapped pairs. Then*

$$\Pr (\langle \mathbf{G}, \mathbf{m}' \rangle \geq \langle \mathbf{G}, \mathbf{m} \rangle \text{ and } \langle \mathbf{G}, \mathbf{m}'' \rangle \geq \langle \mathbf{G}, \mathbf{m} \rangle | M = m) \leq \exp \left(-\frac{\delta}{2} \cdot I_{XY} \right).$$

Proof.

$$\begin{aligned} & \Pr (\langle \mathbf{G}, \mathbf{m}' \rangle \geq \langle \mathbf{G}, \mathbf{m} \rangle \text{ and } \langle \mathbf{G}, \mathbf{m}'' \rangle \geq \langle \mathbf{G}, \mathbf{m} \rangle | M = m) \\ & \leq \Pr \left(\left\langle \mathbf{G}, \frac{\mathbf{m}' + \mathbf{m}'' - 2\mathbf{m}}{2} \right\rangle \geq 0 | M = m \right) \\ & \leq \mathbb{E} \left[e^{\langle \mathbf{G}, \frac{\mathbf{m}' + \mathbf{m}'' - 2\mathbf{m}}{2} \rangle} | M = m \right] \\ & = R \left(\frac{\mathbf{m}' + \mathbf{m}''}{2} \right) \end{aligned}$$

$\frac{\mathbf{m}' + \mathbf{m}''}{2}$ can be written in block diagonal form: We get $n - \delta$ 1×1 blocks [1] for all mapped pairs that agree between the two mappings. For the remaining δ pairs whose mapping disagrees between \mathbf{m}' and \mathbf{m}'' , we get blocks Θ_j that correspond to cycles or even paths as described in Definition 3.

Given this block-diagonal form, by Lemma 13, $R\left(\frac{\mathbf{m}' + \mathbf{m}''}{2}\right)$ is equal to the product $[R([1])]^{n-\delta}$. $\prod_j R(\Theta_j)$. By Lemma 14, $R([1]) = 1$. By Lemmas 17 and 19, plugging in $\nu = 1$ and $\theta = 1/2$ we get $\prod_j R(\Theta_j) = e^{-\frac{\delta}{2} \cdot I_{XY}}$. \square

Proof of Theorem 4.5.1. Suppose $|\mathcal{V}| = |\mathcal{U}| = |M|$ and $I_{XY} \geq \Omega(\log n)$ so that Lemma 24 is applicable.

Let $\mathcal{M}(\mathbf{G}; m) \triangleq \{m' \mid \langle \mathbf{G}, \mathbf{m}' \rangle \geq \langle \mathbf{G}, \mathbf{m} \rangle, m' \neq m\}$ denote the set of matches that are at least as likely as m given the information density matrix \mathbf{G} . Maximum likelihood estimation perfectly estimates the mapping if and only if $\mathcal{M}(\mathbf{G}; M) = \emptyset$.

Also define $\mathcal{M}_2(m) \triangleq \{m' \mid \text{The disagreement between } m \text{ and } m' \text{ corresponds to a 2-cycle.}\}$. For compactness, let $X \triangleq |\mathcal{M}(\mathbf{G}; m) \cap \mathcal{M}_2(m)|$. Clearly $0 \leq X \leq |\mathcal{M}(\mathbf{G}; m)|$.

We apply Chebyshev's inequality:

$$\Pr(|\mathcal{M}(\mathbf{G}; m)| = 0) \leq \Pr(X = 0) \leq \Pr((X - \mathbb{E}[X])^2 \geq \mathbb{E}^2[X]) \leq \frac{\text{Var}(X)}{\mathbb{E}^2[X]}$$

All matchings are equally likely. Therefore, given any bijective matching m ,

$$\begin{aligned} \mathbb{E}[X] &= \mathbb{E}[|\mathcal{M}(\mathbf{G}; m) \cap \mathcal{M}_2(m)|] = \sum_{m' \in \mathcal{M}_2(m)} \Pr(m' \in \mathcal{M}(\mathbf{G}; m) \mid M = m) \\ &= \sum_{m' \in \mathcal{M}_2(m)} \Pr(\langle \mathbf{G}, \mathbf{m}' \rangle \geq \langle \mathbf{G}, \mathbf{m} \rangle \mid M = m) \end{aligned}$$

Let $\varepsilon_1 \triangleq \Pr(\langle \mathbf{G}, \mathbf{m}' \rangle \geq \langle \mathbf{G}, \mathbf{m} \rangle \mid M = m)$ for some m' in $\mathcal{M}_2(m)$. Then $\mathbb{E}[X] = |\mathcal{M}_2(m)| \cdot \varepsilon_1$.

$$\varepsilon_1 = \binom{n}{2} \cdot \varepsilon_1.$$

$$\begin{aligned} |\mathcal{M}(\mathbf{G}; m) \cap \mathcal{M}_2(m)|^2 &= \left[\sum_{m' \in \mathcal{M}_2(m)} \mathbb{1} \{m' \in \mathcal{M}(\mathbf{G}; m)\} \right]^2 \\ &= \sum_{m' \in \mathcal{M}_2(m)} \mathbb{1} \{m' \in \mathcal{M}(\mathbf{G}; m)\} \\ &\quad + 2 \sum_{\{m', m''\} \subset \mathcal{M}_2(m)} \mathbb{1} \{m' \in \mathcal{M}(\mathbf{G}; m) \text{ and } m'' \in \mathcal{M}(\mathbf{G}; m)\} \\ &= \sum_{m' \in \mathcal{M}_2(m)} \mathbb{1} \{\langle \mathbf{G}, \mathbf{m}' - \mathbf{m} \rangle \geq 0\} \\ &\quad + 2 \sum_{\{m', m''\} \subset \mathcal{M}_2(m)} \mathbb{1} \{\langle \mathbf{G}, \mathbf{m}' - \mathbf{m} \rangle \geq 0 \text{ and } \langle \mathbf{G}, \mathbf{m}'' - \mathbf{m} \rangle \geq 0\} \end{aligned}$$

Any pair of mappings $m', m'' \in \mathcal{M}_2(m)$ disagree on either 3 or 4 pairs. There are $\frac{1}{2} \binom{n}{4} \binom{4}{2} = 3 \binom{n}{4}$ different ways to choose to matchings $\{m', m''\} \subset \mathcal{M}_2(m)$ such that m' and m'' disagree on 4 pairs, and $3 \binom{n}{3}$ ways to choose them such that they disagree on 3 pairs. Notice that $3 \binom{n}{4} + 3 \binom{n}{3} = \binom{|\mathcal{M}_2(m)|}{2}$ and these partition all the choices for $\{m', m''\} \subset \mathcal{M}_2(m)$.

When m' and m'' disagree on 4 pairs, the error events become independent and we get

$$\Pr(\langle \mathbf{G}, \mathbf{m}' - \mathbf{m} \rangle \geq 0 \text{ and } \langle \mathbf{G}, \mathbf{m}'' - \mathbf{m} \rangle \geq 0 | M = m) = \varepsilon_1^2.$$

Let $\varepsilon_2 \triangleq \Pr(\langle \mathbf{G}, \mathbf{m}' - \mathbf{m} \rangle \geq 0 \text{ and } \langle \mathbf{G}, \mathbf{m}'' - \mathbf{m} \rangle \geq 0 | M = m)$ given that m' and m'' disagree on 4 pairs.

By the relation $k + 2 \binom{k}{2} = k^2$. For $k = |\mathcal{M}_2(m)| = \binom{n}{2}$ and $\binom{k}{2} = 3 \binom{n}{3} + 3 \binom{n}{4}$ we can

write:

$$\begin{aligned}
\mathbb{E} [|\mathcal{M}(\mathbf{G}; m) \cap \mathcal{M}_2(m)|] &= |\mathcal{M}_2(m)|^2 \varepsilon_1^2 \\
&= \binom{n}{2} \varepsilon_1^2 + \left[6 \binom{n}{3} + 6 \binom{n}{4} \right] \varepsilon_1^2 \\
\mathbb{E} [|\mathcal{M}(\mathbf{G}; m) \cap \mathcal{M}_2(m)|^2] &= \binom{n}{2} \varepsilon_1 + 6 \binom{n}{3} \varepsilon_2 + 6 \binom{n}{4} \varepsilon_1^2 \\
\text{Var} \|\mathcal{M}_2^{\mathcal{E}}(F, M)\| &= \binom{n}{2} (\varepsilon_1 - \varepsilon_1^2) + 6 \binom{n}{3} (\varepsilon_2 - \varepsilon_1^2) \\
&\leq \binom{n}{2} \varepsilon_1 + 6 \binom{n}{3} \varepsilon_2
\end{aligned}$$

Plugging these values into the Chernoff bound we get

$$\Pr (|\mathcal{M}(\mathbf{G}; M)| = 0) \leq \frac{\binom{n}{2} \varepsilon_1 + 6 \binom{n}{3} \varepsilon_2}{\binom{n}{2}^2 \varepsilon_1^2} \leq \mathcal{O} \left(\frac{1}{n^2 \varepsilon_1} + \frac{\varepsilon_2}{n \varepsilon_1^2} \right)$$

By lemmas 24 and 25 we have $\varepsilon_1 \geq e^{-I_{XY}(1-o(1))}$ and $\varepsilon_2 \leq e^{-\frac{3I_{XY}}{2}}$. Thus $\varepsilon_1^2/\varepsilon_2 \geq e^{-\frac{I_{XY}}{2}(1-o(1))}$.

If $I_{XY} \leq (2 - \Omega(1)) \log n$, then

$$\begin{aligned}
n \varepsilon_1^2 / \varepsilon_2 &\geq n e^{-\frac{I_{XY}}{2}(1-o(1))} \geq n^{1-\frac{1}{2}(2-\Omega(1))-o(1)} \geq n^{\Omega(1)} \\
\text{and } n^2 \varepsilon_1 &= n^2 e^{-I_{XY}(1-o(1))} \geq n^{2-(2-\Omega(1))-o(1)} \geq n^{\Omega(1)}
\end{aligned}$$

and therefore $\Pr (|\mathcal{M}(\mathbf{G}; M)| = 0) \leq \mathcal{O}(n^{-\Omega(1)}) \leq o(1)$, i.e. the probability of the maximum likelihood estimator perfectly estimating the mapping is $o(1)$.

The maximum likelihood estimator is also the maximum a posteriori estimator and thus is the optimal estimator for exact recover. Then, the probability of perfectly estimating the mapping is $o(1)$ for any algorithm.

Adding extra users to \mathcal{V} only makes it less likely that the true estimator is perfectly retrieved. Then, our result extends to the case where $|\mathcal{V}| > |\mathcal{U}|$. Similarly, making I_{XY} smaller only makes it less likely to retrieve the true estimator. So our result trivially extends

to the case where $I_{XY} \leq o(\log n)$.

□

CHAPTER 5

GAUSSIAN DATABASE ALIGNMENT WITH FEATURES OF ARBITRARY DIMENSION

In Chapter 4, we derived achievability results for the Gaussian database alignment in the low per-feature correlation regime that matched the achievability (and converse) results for Gaussian planted matching from Chapter 3, which established the connection between the two problems. However, that is not to say that Gaussian planted matching is an exact substitute to study and understand the dynamics of the database alignment problem.

In this chapter, we study the Gaussian database alignment problem with no constraints on the number of dimensions, unlike the previous chapter where we required features to be $\omega(\log n)$ dimensional. We show that the converse result for Gaussian planted matching in Theorem 3.4.1 does not always hold for Gaussian database alignment, which illustrates the more complex nature of the database alignment problem.

In Section 5.1, we analyze threshold testing: Subsection 5.1.1 presents a bound on the expected number of errors and Subsection 5.1.2 presents a converse for almost-exact alignment that matches the achievability result in Subsection 5.1.1. Unlike our other results on database alignment, neither of these assume that the underlying mapping M fully covers either of the user sets \mathcal{U} and \mathcal{V} .

In Section 5.2 we present a bound on the expected number of errors for maximum row estimation.

In Section 5.3, we do some majorization analysis over some random variables which happen to be important in the following section, Section 5.4, where we present achievability results for maximum likelihood alignment.

Our results for both threshold testing and for maximum likelihood alignment, the bounds vary as a function of the ratio $|\mathcal{D}|/\log n$, i.e. the ratio between the number of dimensions

and the log of the size of the underlying mapping, which suggests that $\log n$ is the scaling of interest for the number of dimensions. Furthermore, the results for threshold testing and for maximum likelihood alignment match each other in the $o(\log n)$ -dimensional case, and they match the corresponding results in Chapter 4 in the $\omega(\log n)$ -dimensional case.

Our result in Subsection 5.1.2 (Theorem 5.1.2) has been presented at AISTATS 2019 [38].

5.1 Threshold testing

5.1.1 Achievability

Our achievability result depends on mutual information I_{XY} as well as the effective dimensionality $|\mathcal{D}|$ as defined in Definition 1.

Theorem 5.1.1. *Define $\tau^* \triangleq \log \left(\frac{|\mathcal{U} \times \mathcal{V}|}{n} \right)$. If $|\tau^*| < I_{XY}$, then the expected number of errors by threshold testing with threshold τ is bounded as follows:*

- For any value of the effective dimensionality $|\mathcal{D}|$:

$$\log \mathbb{E} \left[d \left(\hat{M}_{TT}, M \right) \right] \leq \log n - \frac{(I_{XY} - \tau^*)^2}{4I_{XY}} = \log |\mathcal{U} \times \mathcal{V}| - \frac{(I_{XY} + \tau^*)^2}{4I_{XY}}.$$

- For a given value of the effective dimensionality $|\mathcal{D}|$:

$$\log \mathbb{E} \left[d \left(\hat{M}_{TT}, M \right) \right] \leq \log n - \frac{|\mathcal{D}|}{2} \left[\sqrt{1 + \xi^2} - 1 + \log \left(\frac{\sqrt{1 + \xi^2} - 1}{\xi^2} \right) + \log 2 \right],$$

$$\text{where } \xi \triangleq \frac{2(I_{XY} - \tau^*)}{|\mathcal{D}| \sqrt{1 - \exp\left(-\frac{2I_{XY}}{|\mathcal{D}|}\right)}}.$$

- For $|\mathcal{D}|$ such that $|\mathcal{D}| \leq o(\log n)$:

$$\log \mathbb{E} \left[d \left(\hat{M}_{TT}, M \right) \right] \leq \log n - (I_{XY} - \tau^*)(1 - o(1)) \leq \log |\mathcal{U} \times \mathcal{V}| - I_{XY}(1 - o(1)).$$

Trivially, all bounds are decreasing functions of I_{XY} . The second bound is an increas-

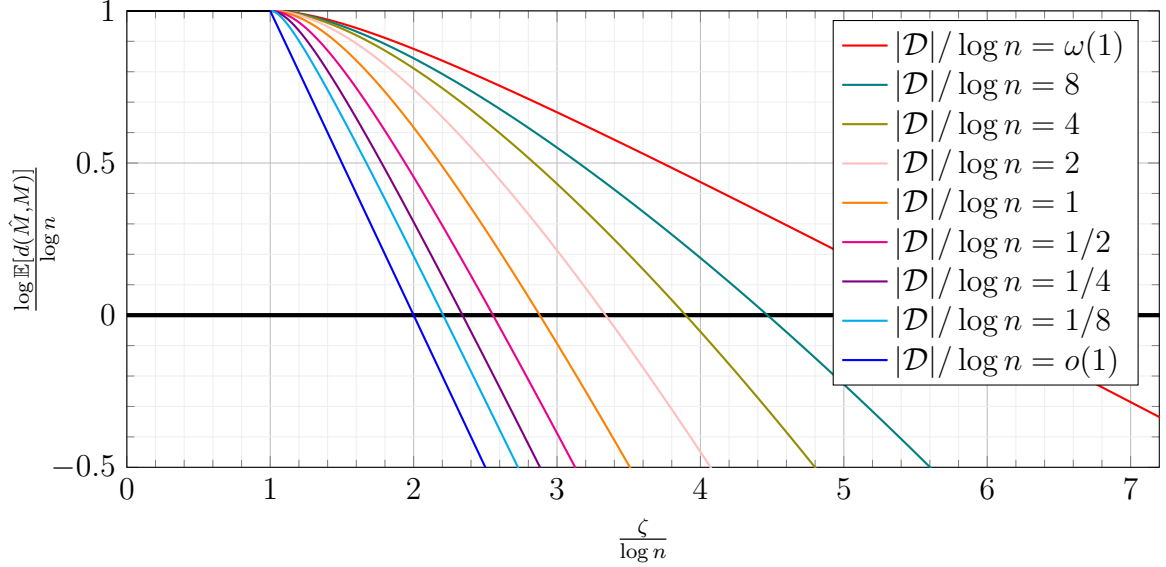


Figure 5.1: Comparison of boundaries for achievability regions for threshold testing under the fully covered case (i.e. $|\mathcal{U}| = |\mathcal{V}| = |M|$) for various values of the number of dimensions $|\mathcal{D}|$ as given in Theorem 5.1.1. The x-axis is to signal strength $\frac{\zeta}{\log n} = \frac{I_{XY}}{\log n}$ and y-axis is the order of magnitude of expected number of mismatched users. Achievability regions are areas above/right of curves.

ing function of $|\mathcal{D}|$ and its limits as $|\mathcal{D}|/\log n$ goes to infinity and to zero give us the first and third bounds respectively.

The achievability regions for various values of the number of dimensions $|\mathcal{D}|$ are illustrated in Fig. 5.1.

In Subsection 5.1.1, we present a lemma that shows that any bound applicable for the case where mutual information is distributed equally among all dimensions is applicable in general with no such constraint. In Subsection 5.1.1, we derive concentration inequalities that are needed to obtain the main result in this subsection. Finally, in Subsection 5.1.1, we give the proof for Theorem 5.1.1.

Convexity of generating function

Lemma 26 shows that, given the number of dimensions and the mutual information I_{XY} , the generating function and consequently the Chernoff bound for the false negative event is maximized, i.e. is worst, when mutual information is distributed equally among all

dimensions as described in Condition 2. Then, we can limit ourselves to the case where mutual information is distributed equally among all dimensions to get a bound that applies generally.

Lemma 26. *Consider two pairs of correlated databases (\vec{A}_0, \vec{B}_0) and (\vec{A}_1, \vec{B}_1) between the same user sets \mathcal{U} and \mathcal{V} and with the same underlying mapping M . Suppose $|\mathcal{U}| = |\mathcal{V}| = 1$ and suppose the two pairs of databases have the same number of effective features, as defined in Definition 1. Let $R_0, R_1 : \mathbb{R}^{1 \times 1} \rightarrow \mathbb{R}$ denote the generating functions for the information density matrices corresponding to the two pairs of correlated databases as defined in Section 4.2.*

If R_0 corresponds to a pair of databases with mutual information distributed equally among all dimensions (Condition 2), then $R_0([\theta]) \geq R_1([\theta])$ for any value of $\theta \in [0, 1]$.

Proof. Let $\vec{\rho} \in (-1, 1)^{\mathcal{D}}$ some arbitrary correlation vector and $I_{XY} = -\frac{1}{2} \sum_i \log(1 - \rho_i^2)$.

For any $i \in \mathcal{D}$, define

$$\lambda_i \triangleq \frac{-\frac{1}{2} \log(1 - \rho_i^2)}{I_{XY}}$$

and express ρ_i as a function of λ_i and I_{XY} :

$$\rho_i = \rho(\lambda_i)$$

where $\rho(x) = \sqrt{1 - \exp(-2I_{XY}\lambda_i)}$.

By definition, $\sum \lambda_i = 1$.

By Lemma 14, the log of the generating function is given by

$$\log R([1 - \theta]) = -I_{XY}\theta - \frac{1}{2} \sum_{i \in \mathcal{D}} \log(1 - \rho_i^2 \theta^2). \quad (5.1)$$

This expression can be written as a function of λ_i :

$$-I_{XY}\theta - \frac{1}{2} \sum_{i \in \mathcal{D}} \log(1 - \rho^2(\lambda_i)\theta^2).$$

Given Λ is a uniform random variable with support over $\{\lambda_i\}_{i \in \mathcal{D}}$, the above expression can also be written as

$$|\mathcal{D}| \mathbb{E} \left[-\frac{I_{XY}}{|\mathcal{D}|} \theta - \frac{1}{2} \log(1 - \rho^2(\Lambda)\theta^2) \right].$$

The expected value of Λ is $\mathbb{E}[\Lambda] = \frac{1}{|\mathcal{D}|} \sum \lambda_i = 1/|\mathcal{D}|$.

The second derivative of the function inside the expectation is equal to

$$\frac{d^2}{d\lambda^2} \left(-I_{XY}\theta - \frac{1}{2} \log(1 - \rho^2(\lambda)\theta^2) \right) = -\frac{2I_{XY}^2 e^{2I_{XY}\lambda\theta^2} (1 - \theta^2)}{(e^{2I_{XY}\lambda} (1 - \theta^2) + \theta^2)^2}$$

which is strictly negative if and only if $\theta \in (0, 1)$. Then, for $\theta \in (0, 1)$, the function inside the expectation is strictly concave. (If $\theta \in \{0, 1\}$, then the second derivative is zero and the function is linear.) It then follows that, by Jensen's inequality,

$$|\mathcal{D}| \mathbb{E} \left[-\frac{I_{XY}}{|\mathcal{D}|} \theta - \frac{1}{2} \log(1 - \rho^2(\Lambda)\theta^2) \right] \leq |\mathcal{D}| \left(-\frac{I_{XY}}{|\mathcal{D}|} \theta - \frac{1}{2} \log(1 - \rho^2(\mathbb{E}[\Lambda])\theta^2) \right)$$

or

$$-I_{XY}\theta - \frac{1}{2} \sum_{i \in \mathcal{D}} \log(1 - \rho^2(\lambda_i)\theta^2) \leq -I_{XY}\theta - \frac{|\mathcal{D}|}{2} \log(1 - \rho^2(1/|\mathcal{D}|)\theta^2).$$

Then, for any value of $\theta \in (0, 1)$, the generating function for an arbitrary partitioning of mutual information across \mathcal{D} is upper bounded by the generating function for the case where mutual information is partitioned equally across \mathcal{D} . \square

Concentration inequalities

Lemma 27 identifies the conditions on τ that result in the optimal Chernoff parameter being within a certain range, which is required for the convexity argument in Lemma 26 and to obtain concentration inequalities through Corollary 11. Lemma 28 gives the concentration inequalities in their raw form, while Lemma 29 evaluates them in the extreme cases where $|\mathcal{D}| \leq o(\log n)$ and $|\mathcal{D}| \geq \omega(\log n)$.

Lemma 27. Define $\xi \triangleq \frac{2(I_{XY} - \tau)}{|\mathcal{D}|\rho^2}$ and $\theta^* = \frac{\sqrt{1+\xi^2}-1}{\xi\sqrt{\rho^2}}$ where $I_{XY} \triangleq -\frac{|\mathcal{D}|}{2} \log(1 - \rho^2)$.

- If $\tau < I_{XY}$, then $0 < \theta^* < \frac{1}{\sqrt{\rho^2}}$.
- If $\tau > I_{XY}$, then $-\frac{1}{\sqrt{\rho^2}} < \theta^* < 0$.
- If $|\tau| < I_{XY}$, then $0 < \theta^* < 1$.

Proof. • $0 < \theta^* < 1/\sqrt{\rho^2}$ if $\tau < I_{XY}$:

Since $|\mathcal{D}| \geq 1$, $\sqrt{\rho^2} > 0$ and $I_{XY} > \tau$, we have $\xi > 0$, which implies $\theta^* > 0$.

$$\begin{aligned}
 & 0 < \xi \\
 \iff & 1 + \xi^2 < 1 + \xi^2 + 2\xi \\
 \iff & \sqrt{1 + \xi^2} < \xi + 1 \\
 \iff & \sqrt{1 + \xi^2} - 1 < \xi \\
 \iff & \theta^* \sqrt{\rho^2} = \frac{\sqrt{1 + \xi^2} - 1}{\xi} < 1
 \end{aligned}$$

So $\theta^* < 1/\sqrt{\rho^2}$.

- $-1/\sqrt{\rho^2} < \theta^* < 0$ if $\tau > I_{XY}$:

Since $|\mathcal{D}| \geq 1$, $\sqrt{\rho^2} > 0$ and $I_{XY} < \tau$, we have $\xi < 0$, which implies $\theta^* < 0$.

$$\begin{aligned}
& 0 < -\xi \\
\iff & 1 + \xi^2 < 1 + \xi^2 - 2\xi \\
\iff & \sqrt{1 + \xi^2} < 1 - \xi \\
\iff & \sqrt{1 + \xi^2} - 1 < -\xi \\
\iff & \theta^* \sqrt{\rho^2} = \frac{\sqrt{1 + \xi^2} - 1}{\xi} > -1
\end{aligned}$$

So $\theta^* < 1/\sqrt{\rho^2}$.

- $0 < \theta^* < 1$ if $|\tau| < I_{XY}$:

$\tau < I_{XY}$ implies $\theta^* > 0$.

The Taylor series expansion of $-(1-x)\log(1-x)$ is given by

$$\begin{aligned}
-(1-x)\log(1-x) &= -(1-x) \sum_{k=1}^{\infty} \frac{x^k}{k} \\
&= x - \sum_{k=2}^{\infty} \frac{x^k}{k(k-1)} \leq x
\end{aligned}$$

It then follows that, given $|\tau| < -I_{XY}$

$$\begin{aligned}
& - (1 - \rho^2) \log (1 - \rho^2) \leq \rho^2 \\
\iff & 2 (1 - \rho^2) I_{XY} \leq |\mathcal{D}| \rho^2 \\
\iff & \frac{2I_{XY}}{|\mathcal{D}| \rho^2} \leq \frac{1}{1 - \rho^2} \\
\implies & \frac{I_{XY} - \tau}{|\mathcal{D}| \rho^2} < \frac{1}{1 - \rho^2} \\
\iff & \frac{\xi}{2\sqrt{\rho^2}} < \frac{1}{1 - \rho^2} \\
\iff & \xi^2 (1 - \rho^2) < 2\xi \sqrt{\rho^2} \\
\iff & \xi^2 + 1 < \rho^2 \xi^2 + 2\xi \sqrt{\rho^2} + 1 \\
\iff & \sqrt{1 + \xi^2} < \xi \sqrt{\rho^2} + 1 \\
\iff & \sqrt{1 + \xi^2} - 1 < \xi \sqrt{\rho^2} \\
\iff & \frac{\sqrt{1 + \xi^2} - 1}{\xi \sqrt{\rho^2}} < 1.
\end{aligned}$$

□

Lemma 28. Define $\xi \triangleq \frac{2(I_{XY} - \tau)}{|\mathcal{D}| \sqrt{1 - \exp(-\frac{2I_{XY}}{|\mathcal{D}|})}}$. Let $\vec{\rho} \in (-1, 1)^{\mathcal{D}}$ the correlation vector under canonical form.

- Let $\tau < I_{XY}$.

If either $\tau > -I_{XY}$ or if mutual information is distributed equally across all dimensions (Condition 2), then

$$\log \Pr \left[\tau \geq G_{u,v} | u \stackrel{M}{\sim} v \right] \leq -\frac{|\mathcal{D}|}{2} \left[\sqrt{1 + \xi^2} - 1 + \log \left(\frac{\sqrt{1 + \xi^2} - 1}{\xi^2} \right) + \log 2 \right]$$

- Let $\tau > -I_{XY}$.

If either $\tau < I_{XY}$ or if mutual information is distributed equally across all dimen-

sions (Condition 2), then

$$\log \Pr \left[\tau \leq G_{u,v}|u \stackrel{M}{\not\sim} v \right] \leq -\frac{|\mathcal{D}|}{2} \left[\sqrt{1 + \xi^2} - 1 + \log \left(\frac{\sqrt{1 + \xi^2} - 1}{\xi^2} \right) + \log 2 \right] - \tau$$

These bounds are decreasing functions of I_{XY} and increasing functions of $|\mathcal{D}|$.

Proof. Assume that mutual information is distributed equally across each dimension as described in Condition 2, i.e. under the canonical form, the correlation vector $\vec{\rho} \in (-1, 1)^{\mathcal{D}}$ is such that

$$\rho_i = \pm \sqrt{1 - \exp \left(-\frac{2I_{XY}}{|\mathcal{D}|} \right)}$$

for each $i \in \mathcal{D}$. Then $\rho_i \neq 0$ as long as $I_{XY} > 0$. Henceforth, in this proof, we refer to this value simply as ρ .

By Corollary 11, for any value of $\theta_{\text{FN}} \in (0, 1/\sqrt{\rho^2})$ and $\theta_{\text{FP}} \in (0, 1 + 1/\sqrt{\rho^2})$, the log-probability of the false negative and false positive events are bounded as follows:

$$\begin{aligned} \log \Pr \left[\tau \geq G_{u,v}|u \stackrel{M}{\sim} v \right] &\leq \tau \theta_{\text{FN}} + \log R([1 - \theta_{\text{FN}}]) \\ \log \Pr \left[\tau \leq G_{u,v}|u \stackrel{M}{\not\sim} v \right] &\leq -\tau \theta_{\text{FP}} + \log R([\theta_{\text{FP}}]) \\ &= (1 - \theta_{\text{FP}})\tau + \log R([1 - (1 - \theta_{\text{FP}})]) - \tau \end{aligned}$$

It then follows that, given any value of τ , the values for θ_{FN} and for $1 - \theta_{\text{FP}}$ within any range that minimize the expressions above satisfy $\theta_{\text{FN}}^* = 1 - \theta_{\text{FP}}^*$. (These optimal θ^* values may or may not be strictly positive. We are not able to get a useful Chernoff bound if the minimizing θ^* is negative.) Plugging in the optimizing values for θ_{FN}^* and θ_{FP}^* gives us an expression for false negatives which is e^τ times that of false positives.

Next we optimize $\theta = \theta_{\text{FN}}$ as a function of τ :

$$\begin{aligned}
\log \Pr \left[\tau \geq G_{u,v} | u \stackrel{M}{\sim} v \right] &\leq \tau\theta + \log R([1 - \theta]) \\
&= \theta\tau - \theta I_{XY} - \frac{|\mathcal{D}|}{2} \log(1 - \rho^2\theta^2) \\
&= -\theta(I_{XY} - \tau) - \frac{|\mathcal{D}|}{2} \log(1 - \rho^2\theta^2)
\end{aligned}$$

The derivative of this last expression is $-(I_{XY} - \tau) + \frac{|\mathcal{D}|\rho^2\theta}{1 - \rho^2\theta^2}$, and it is an increasing function of θ within the ranges $\theta \in (-\infty, -1/\sqrt{\rho^2})$, $\theta \in (-1/\sqrt{\rho^2}, 1/\sqrt{\rho^2})$ and $\theta \in (1/\sqrt{\rho^2}, \infty)$. Then the optimal value of θ that minimizes the Chernoff bound is the only root to $I_{XY} - \tau = \frac{|\mathcal{D}|\rho^2\theta}{1 - \rho^2\theta^2}$ within the range $\theta \in (-1/\sqrt{\rho^2}, 1/\sqrt{\rho^2})$, which is equal to

$$\begin{aligned}
\theta^* &= \frac{\sqrt{|\mathcal{D}|^2\rho^4 + 4(I_{XY} - \tau)^2\rho^2} - |\mathcal{D}|\rho^2}{2\rho^2(I_{XY} - \tau)} \\
&= \frac{\sqrt{1 + \frac{4(I_{XY} - \tau)^2}{|\mathcal{D}|^2\rho^2}} - 1}{2(I_{XY} - \tau)/|\mathcal{D}|} \\
&= \frac{\sqrt{1 + \xi^2} - 1}{\xi\sqrt{\rho^2}},
\end{aligned}$$

where $\xi \triangleq \frac{2(I_{XY} - \tau)}{|\mathcal{D}|\sqrt{\rho^2}}$.

- By Lemma 27, if $\tau < I_{XY}$, then $\theta^* \in (0, 1/\sqrt{\rho^2})$ and therefore we have a probability bound for the false negative event.
- By Lemma 27, if $\tau > -I_{XY}$, then $1 - \theta^* > 0$ and $|\theta^*| < 1/\sqrt{\rho^2}$ and therefore we have a probability bound for the false positive event.

Plugging in the value of θ^* ,

$$\begin{aligned}
1 - \rho^2(\theta^*)^2 &= 1 - \frac{(\sqrt{1 + \xi^2} - 1)^2}{\xi^2} = \frac{2(\sqrt{1 + \xi^2} - 1)}{\xi^2} \\
\theta^*(I_{XY} - \tau) &= \frac{|\mathcal{D}|}{2} (\sqrt{1 + \xi^2} - 1)
\end{aligned}$$

which gives us the expression for the Chernoff bound as

$$-\theta(I_{XY} - \tau) - \frac{|\mathcal{D}|}{2} \log(1 - \rho^2 \theta^2) \leq -\frac{|\mathcal{D}|}{2} \left[\sqrt{1 + \xi^2} - 1 + \log \left(\frac{\sqrt{1 + \xi^2} - 1}{\xi^2} \right) + \log 2 \right]$$

By Lemma 27, $\theta^* \in (0, 1)$ if $|\tau| \leq I_{XY}$. Then, by Lemma 26, the Chernoff bound can only get smaller when we remove the assumption that mutual information is distributed equally across each dimension.

ξ is an increasing function of I_{XY} , and the general bound is a decreasing function of ξ . Then the general bound is a decreasing function of I_{XY} .

The general bound applies in the case where mutual information is equally distributed across $|\mathcal{D}|$ dimensions as defined by Condition 2. The equally distributed case across $|\mathcal{D}|-1$ is equivalent to a non-equally distributed case across $|\mathcal{D}|$ dimensions. By Lemma 26, the Chernoff bound can only get smaller when we violate the equally distributed condition across a given number of dimensions. Then the bound is an increasing function of $|\mathcal{D}|$. \square

We work out the form of the concentration inequalities in Lemma 28 under the low-dimensional case (i.e. $|\mathcal{D}| \leq o(\log n)$) and the high-dimensional case (i.e. $|\mathcal{D}| \geq \omega(\log n)$). This bound in the high-dimensional case is slightly weaker than, yet asymptotically matches, the bound in Lemma 20, which applies without any restriction on the number of dimensions. Therefore we use Lemma 20 instead of the following lemma when working in the high-dimensional case.

Lemma 29. Define $\xi \triangleq \frac{2(I_{XY} - \tau)}{|\mathcal{D}| \sqrt{1 - \exp(-\frac{2I_{XY}}{|\mathcal{D}|})}}$.

- *Low-dimensional case:* $|\mathcal{D}| \leq o(\log n)$

$$-\frac{|\mathcal{D}|}{2} \left[\sqrt{1 + \xi^2} - 1 + \log \left(\frac{\sqrt{1 + \xi^2} - 1}{\xi^2} \right) + \log 2 \right] \leq -|I_{XY} - \tau|(1 - o(1))$$

- *High-dimensional case:* $|\mathcal{D}| \geq \omega(\log n)$

$$-\frac{|\mathcal{D}|}{2} \left[\sqrt{1 + \xi^2} - 1 + \log \left(\frac{\sqrt{1 + \xi^2} - 1}{\xi^2} \right) + \log 2 \right] \leq -\frac{(I_{XY} - \tau)^2}{4I_{XY}}(1 - o(1))$$

Proof. Recall that $\xi \triangleq \frac{2(I_{XY} - \tau)}{|\mathcal{D}| \sqrt{1 - \exp\left(-\frac{2I_{XY}}{|\mathcal{D}|}\right)}}$. Suppose I_{XY} and $I_{XY} - \tau$ are on the order of $\log n$.

- *Low-dimensional case:*

If $|\mathcal{D}| \leq o(\log n)$, then $1 - o(1) \leq \sqrt{1 - \exp\left(-\frac{2I_{XY}}{|\mathcal{D}|}\right)} \leq 1$ and therefore $\frac{2|I_{XY} - \tau|}{|\mathcal{D}|} \leq |\xi| \leq \frac{2|I_{XY} - \tau|}{|\mathcal{D}|}(1 + o(1))$, which is $\omega(1)$.

$\sqrt{1 + x^2} - 1 \geq x(1 - o(1))$ for $x \geq \omega(1)$. Then

$$\begin{aligned} & -\frac{|\mathcal{D}|}{2} \left[\sqrt{1 + \xi^2} - 1 + \log \left(\frac{\sqrt{1 + \xi^2} - 1}{\xi^2} \right) + \log 2 \right] \\ & \leq -\frac{|\mathcal{D}|}{2} \left[|\xi|(1 - o(1)) + \log \left(\frac{|\xi|(1 - o(1))}{\xi^2} \right) + \log 2 \right] \\ & \leq -\frac{|\mathcal{D}|}{2} [|\xi|(1 - o(1)) - \log |\xi| + \log 2 - o(1)] \\ & \leq -\frac{|\mathcal{D}|}{2} \cdot |\xi|(1 - o(1)) \\ & = -\frac{|I_{XY} - \tau|}{\sqrt{1 - \exp\left(-\frac{2I_{XY}}{|\mathcal{D}|}\right)}}(1 - o(1)) \\ & \leq -|I_{XY} - \tau|(1 - o(1)) \end{aligned}$$

- *High-dimensional case:*

$x(1 - o(1)) \leq 1 - e^{-x} \leq x$ for $x \leq o(1)$. Then, given $|\mathcal{D}| \geq \omega(I_{XY})$,

$$\begin{aligned} (1 - o(1)) \sqrt{\frac{2I_{XY}}{|\mathcal{D}|}} & \leq \sqrt{1 - \exp\left(-\frac{2I_{XY}}{|\mathcal{D}|}\right)} \leq \sqrt{\frac{2I_{XY}}{|\mathcal{D}|}} \\ \frac{|I_{XY} - \tau|}{\sqrt{2I_{XY}|\mathcal{D}|/2}} & \leq |\xi| \leq \frac{|I_{XY} - \tau|}{\sqrt{I_{XY}|\mathcal{D}|/2}}(1 + o(1)). \end{aligned}$$

It then follows that, if $|\mathcal{D}| \geq \omega(\log n)$, then $|\xi| \leq o(1)$.

$\sqrt{1+x} - 1 \geq \frac{x}{2} \left(1 - \frac{x}{4}\right)$ and $\log(1-x) \geq -x(1+o(1))$ for $x \leq o(1)$. Then,

$$\begin{aligned}
\log \Pr \left[\tau \geq G_{u,v} | u \stackrel{M}{\sim} v \right] &\leq -\frac{|\mathcal{D}|}{2} \left[\sqrt{1+\xi^2} - 1 + \log \left(\frac{\sqrt{1+\xi^2} - 1}{\xi^2} \right) + \log 2 \right] \\
&\leq -\frac{|\mathcal{D}|}{2} \left[\frac{\xi^2}{2} \left(1 - \frac{\xi^2}{4}\right) + \log \left(\frac{\frac{\xi^2}{2} \left(1 - \frac{\xi^2}{4}\right)}{\xi^2} \right) + \log 2 \right] \\
&= -\frac{|\mathcal{D}|}{2} \left[\frac{\xi^2}{2} \left(1 - \frac{\xi^2}{4}\right) + \log \left(1 - \frac{\xi^2}{4}\right) \right] \\
&\leq -\frac{|\mathcal{D}|}{2} \left[\frac{\xi^2}{2} (1 - o(1)) - \frac{\xi^2}{4} (1 + o(1)) \right] \\
&\leq -\frac{|\mathcal{D}|}{2} \cdot \frac{\xi^2}{4} (1 - o(1)) \\
&= -\frac{(I_{XY} - \tau)^2}{2|\mathcal{D}| \left(1 - \exp\left(-\frac{2I_{XY}}{|\mathcal{D}|}\right)\right)} (1 - o(1)) \\
&\leq -\frac{(I_{XY} - \tau)^2}{2|\mathcal{D}| \left(\frac{2I_{XY}}{|\mathcal{D}|}\right)} (1 - o(1)) \\
&= -\frac{(I_{XY} - \tau)^2}{4I_{XY}} (1 - o(1))
\end{aligned}$$

We get the bound for the false positive event by multiplying the bound for the false negative event by $e^{-\tau}$. □

Proof for achievability result

Proof for Theorem 5.1.1. There are n true pairs and $|\mathcal{U} \times \mathcal{V}| - n$ false pairs, which are the number of possible false negative and false positive events respectively. The Chernoff bound for false negatives is e^τ times that of false positives. Then, picking $\tau = \log \left(\frac{|\mathcal{U} \times \mathcal{V}|}{n} - 1 \right)$ results in an equal number of false negatives and false positives in expectation. Instead, we pick a slightly larger value $\tau = \log \frac{|\mathcal{U} \times \mathcal{V}|}{n}$ with a simpler expression. This means the expected number of false negatives is slightly larger than that of false positives. $\tau^* \geq \log n$ since $|\mathcal{U}| \geq n$ and $|\mathcal{V}| \geq n$.

Summing up the bound in Lemma 28 for all n possible false negative events, or for all

$|\mathcal{U} \times \mathcal{V}|$ possible false positive events, gives us the claimed general bound.

Plugging in the simplified expression in Lemma 29 gives us the expression for the low-dimensional cases. We use Lemma 20 for the high-dimensional case, which is the most general bound, since the bound is an increasing function of $|\mathcal{D}|$. \square

5.1.2 Converse for almost-exact alignment

By Theorem 5.1.1, $I_{XY} \geq \tau^* + \omega(\sqrt{n})$ is a sufficient condition for threshold testing to achieve almost-exact alignment (i.e. alignment with $o(n)$ errors). Next, we show that $I_{XY} \geq \tau^*(1 - \Omega(1))$ is a necessary condition for almost-exact alignment by proving the matching converse.

Theorem 5.1.2. *For threshold testing,*

$$\mathbb{E} \left[d(\hat{M}_{TT}, M) \right] \geq \frac{n}{2} \left(\frac{\tau^* - I_{XY}}{\log |\mathcal{U} \times \mathcal{V}| + 1} \right).$$

where $\tau^* \triangleq \log \frac{|\mathcal{U} \times \mathcal{V}|}{n}$.

Consequently, if $I_{XY} \leq \tau^*(1 - \mathcal{O}(1))$, then $\mathbb{E} \left[d(\hat{M}_{TT}, M) \right] \geq \Omega(n)$ and almost-exact alignment is cannot be achieved by threshold testing.

Lemma 30. *For $u \in \mathcal{U}$ and $v \in \mathcal{V}$, $H(M_{u,v} | \vec{A}_u, \vec{B}_v) \geq \frac{n}{|\mathcal{U} \times \mathcal{V}|} \left(\log \frac{|\mathcal{U} \times \mathcal{V}|}{n} - I_{XY} \right)$.*

Proof. We have

$$H(M_{u,v} | \vec{A}_u, \vec{B}_v) = H(M_{u,v}) + I(\vec{A}_u; \vec{B}_v) - I(M_{u,v}; \vec{A}_u) - I(M_{u,v}; \vec{B}_v) - I(\vec{A}_u; \vec{B}_v | M_{u,v}).$$

$I(\vec{A}_u; \vec{B}_v) \geq 0$ since mutual information is non-negative.

$M_{u,v}$ is a Bernoulli random variable $\Pr(M_{u,v} = 1) = \frac{n}{|\mathcal{U} \times \mathcal{V}|}$. Then

$$\begin{aligned} H(M_{u,v}) &= \frac{n}{|\mathcal{U} \times \mathcal{V}|} \log \frac{|\mathcal{U} \times \mathcal{V}|}{n} + \frac{|\mathcal{U} \times \mathcal{V}| - n}{|\mathcal{U} \times \mathcal{V}|} \log \frac{|\mathcal{U} \times \mathcal{V}|}{|\mathcal{U} \times \mathcal{V}| - n} \\ &\geq \frac{n}{|\mathcal{U} \times \mathcal{V}|} \log \frac{|\mathcal{U} \times \mathcal{V}|}{n} \end{aligned}$$

Furthermore, $I(M_{u,v}; \vec{A}_u) = I(M_{u,v}; \vec{B}_v) = 0$ since, as long as we don't observe the other database, all features in a given database are i.i.d. no matter what the underlying mapping.

Finally,

$$\begin{aligned} I(\vec{A}_u; \vec{B}_v | M_{u,v}) &= \frac{|\mathcal{U} \times \mathcal{V}| - n}{|\mathcal{U} \times \mathcal{V}|} I(\vec{A}_u | (M_{u,v} = 0); \vec{B}_v | (M_{u,v} = 0)) \\ &\quad + \frac{n}{|\mathcal{U} \times \mathcal{V}|} I(\vec{A}_u | (M_{u,v} = 1); \vec{B}_v | (M_{u,v} = 1)) \\ &= \frac{|\mathcal{U} \times \mathcal{V}| - n}{|\mathcal{U} \times \mathcal{V}|} \cdot 0 + \frac{n}{|\mathcal{U} \times \mathcal{V}|} I_{\vec{X}\vec{Y}}. \end{aligned}$$

$$\text{Then } H(M_{u,v} | \vec{A}_u, \vec{B}_v) \geq \frac{n}{|\mathcal{U} \times \mathcal{V}|} \left(\log \frac{|\mathcal{U} \times \mathcal{V}|}{n} - I_{XY} \right) \quad \square$$

Proof of Theorem 5.1.2. Let $\hat{M}_{u,v} \triangleq \mathbb{1} \left\{ (\vec{A}_u, \vec{B}_v) \in H_\tau \right\}$ denote the estimation on the relation between identifiers u and v . We have a correct estimation if $\hat{M}_{u,v} = M_{u,v}$. Define $E \triangleq \mathbb{1} \left\{ \hat{M}_{u,v} \neq M_{u,v} \right\}$. Then by Fano's inequality,

$$H(M_{u,v} | \vec{A}_u, \vec{B}_v) \leq H(E) + \Pr[E = 1],$$

which gives the upper bound as $H(E)$.

Let $\epsilon \triangleq \Pr[E = 1]$. This value can also be expressed as the expected frequency of false matches, i.e. given ε_{FN} and ε_{FP} the expected number of false negatives and false positives, $\epsilon = \frac{\varepsilon_{FN} + \varepsilon_{FP}}{|\mathcal{U} \times \mathcal{V}|}$.

Let H_b denote the binary entropy function. By Fano's inequality, using Lemma 30, we have

$$H_b(\epsilon) \geq H(M_{u,v} | \vec{A}_u, \vec{B}_v) \geq \frac{n}{|\mathcal{U} \times \mathcal{V}|} \left(\log \frac{|\mathcal{U} \times \mathcal{V}|}{n} - I_{XY} \right) \quad (5.2)$$

$-(1-x) \ln(1-x) \leq x$, so

$$\begin{aligned} H_b(\epsilon) &= -\epsilon \log \epsilon - (1-\epsilon) \log(1-\epsilon) \\ &\leq -\epsilon \log \epsilon + \epsilon = \frac{\varepsilon_{FN} + \varepsilon_{FP}}{|\mathcal{U} \times \mathcal{V}|} (\log |\mathcal{U} \times \mathcal{V}| - \log(\varepsilon_{FN} + \varepsilon_{FP}) + 1). \end{aligned}$$

Combining this with (5.2) gives us

$$\frac{\varepsilon_{FN} + \varepsilon_{FP}}{n} \geq \frac{\log \frac{|\mathcal{U} \times \mathcal{V}|}{n} - I_{XY}}{\log |\mathcal{U} \times \mathcal{V}| + 1 - \log(\varepsilon_{FN} + \varepsilon_{FP})}.$$

If $\frac{\log \frac{|\mathcal{U} \times \mathcal{V}|}{n} - I_{XY}}{\log |\mathcal{U} \times \mathcal{V}| + 1 - \log(\varepsilon_{FN} + \varepsilon_{FP})} \geq \frac{1}{n}$, then the inequality above implies

$$\frac{\varepsilon_{FN} + \varepsilon_{FP}}{n} \geq \frac{\log \frac{|\mathcal{U} \times \mathcal{V}|}{n} - I_{XY}}{\log |\mathcal{U} \times \mathcal{V}| + 1}.$$

□

5.2 Maximum row estimation

Theorem 5.2.1. *Suppose $|\mathcal{U}| = |M| = n$. If $I_{XY} \geq \log |\mathcal{V}|$, then the expected number of errors by maximum row estimation is bounded as follows:*

- For a given value of the effective dimensionality $|\mathcal{D}|$:

$$\log \mathbb{E} \left[d \left(\hat{M}_{MR}, M \right) \right] \leq \log |\mathcal{U} \times \mathcal{V}| - I_{XY} - \frac{|\mathcal{D}|}{2} \log \left(\frac{1 + \exp \left(-\frac{2I_{XY}}{|\mathcal{D}|} \right)}{2} \right)$$

and

$$\log \mathbb{E} \left[d \left(\hat{M}_{MR}, M \right) \right] \leq \log n - \frac{|\mathcal{D}|}{2} \left[\sqrt{1 + \xi^2} - 1 + \log \left(\frac{\sqrt{1 + \xi^2} - 1}{\xi^2} \right) + \log 2 \right] + \log 2,$$

where $\xi \triangleq \frac{2(I_{XY} - \log |\mathcal{V}|)}{|\mathcal{D}| \sqrt{1 - \exp \left(-\frac{2I_{XY}}{|\mathcal{D}|} \right)}}$.

- For $|\mathcal{D}|$ such that $|\mathcal{D}| \leq o(\log n)$:

$$\log \mathbb{E} \left[d \left(\hat{M}_{TT}, M \right) \right] \leq \log |\mathcal{U} \times \mathcal{V}| - I_{XY} + \mathcal{O}(|\mathcal{D}|).$$

Trivially, all bounds are decreasing functions of I_{XY} . The second bound is an increasing function of $|\mathcal{D}|$ and its limits as $|\mathcal{D}|/\log n$ goes to infinity and to zero give us the first

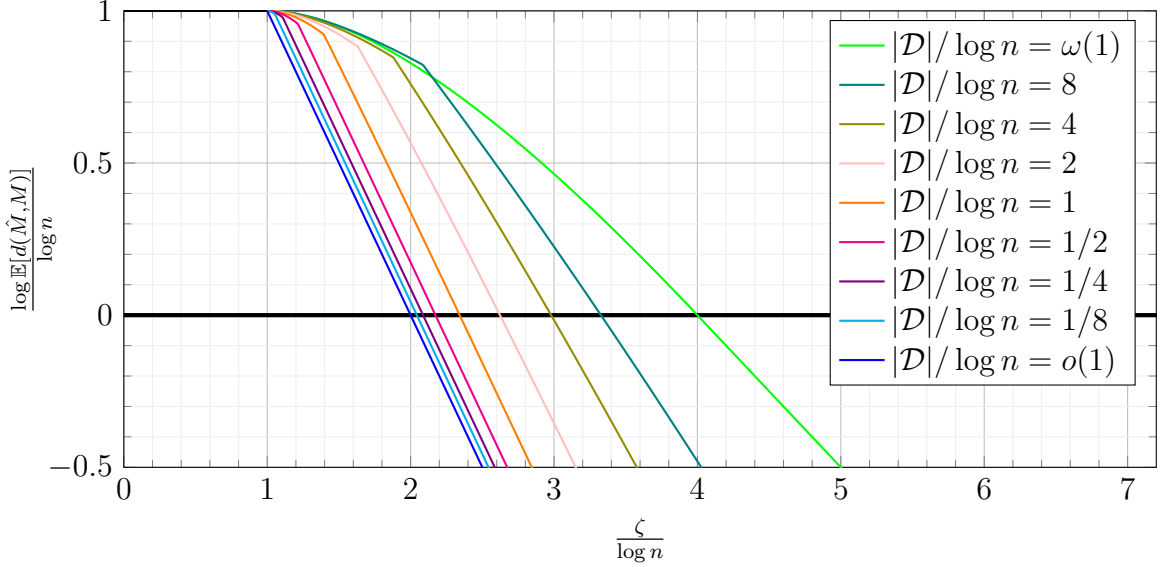


Figure 5.2: Comparison of boundaries for achievability regions for maximum row estimation under the fully covered case (i.e. $|\mathcal{U}| = |\mathcal{V}| = |M|$) for various values of the number of dimensions $|\mathcal{D}|$ as given by Theorem 5.2.1 for $|\mathcal{D}| \leq \mathcal{O}(\log n)$ and by Theorem 4.4.2 for $|\mathcal{D}| \geq \omega(\log n)$. The x-axis is to signal strength $\frac{\zeta}{\log n} = \frac{I_{XY}}{\log n}$ and y-axis is the order of magnitude of expected number of mismatched users. Achievability regions are areas above/right of curves. Note that the boundary from Theorem 4.4.2 intersects some of the boundaries from Theorem 5.2.1.

and third bounds respectively.

The achievability regions for various values of the number of dimensions $|\mathcal{D}|$ are illustrated in Fig. 5.2.

Note that, in the $\omega(\log n)$ -dimensional regime, in the balanced case where $|\mathcal{U}| = |\mathcal{V}| = |M|$, by Theorem 4.4.2, the achievability bound for maximum row estimation beats the bound for threshold testing at every value of I_{XY} . In Theorem 5.2.1, we simply use the threshold testing bounds from Theorem 5.1.1 for small values of I_{XY} . We conjecture that the bound for Theorem 5.1.1 is never tight for maximum row estimation, and that a more careful analysis (comparable to the one in Chapter 4 that gave the parabolic boundaries) would give us tighter bounds. The fact that the boundary from Theorem 4.4.2 for the $\omega(\log n)$ -dimensional case intersects some of the boundaries from Theorem 5.2.1 for the $\mathcal{O}(\log n)$ -dimensional case in Fig. 5.2 supports the claim that this achievability bound is

not tight for small values of I_{XY} .

Lemma 31. *Consider two pairs of correlated databases (\vec{A}_0, \vec{B}_0) and (\vec{A}_1, \vec{B}_1) between the same user sets \mathcal{U} and \mathcal{V} and with the same underlying mapping M . Suppose $|\mathcal{U}| = 1$ and $|\mathcal{V}| = 2$ and suppose the two pairs of databases have the same number of effective features, as defined in Definition 1. Let $R_0, R_1 : \mathbb{R}^{1 \times 2} \rightarrow \mathbb{R}$ denote the generating functions for the information density matrices corresponding to the two pairs of correlated databases as defined in Section 4.2.*

If R_0 corresponds to a pair of databases with mutual information distributed equally among all dimensions (Condition 2), then $R_0([1/2, 1/2]) \geq R_1([1/2, 1/2])$.

Proof. By Lemma 18,

$$\begin{aligned} R([1/2, 1/2]) &= \prod_{i \in \mathcal{D}} \left[\frac{1 - \rho_i^2}{1 - \rho_i^2/2} \right]^{\frac{1}{2}} \\ \implies \log R([1/2, 1/2]) &= -I_{XY} - \frac{1}{2} \sum_{i \in \mathcal{D}} \log \left(1 - \frac{\rho_i^2}{2} \right) \end{aligned} \quad (5.3)$$

The proof of the lemma statement can be shown by exactly following the proof for Lemma 26. We simply need to replace the expression $\log(1 - \rho_i^2 \theta^2)$ in (5.1) with $\log\left(1 - \frac{\rho_i^2}{2}\right)$. The proof for Lemma 26 holds for any value of $\theta \in [0, 1]$, which includes $\theta^2 = 1/2$.

(The $-I_{XY}\theta$ term in (5.1) and the $-I_{XY}$ term in (5.3) are not important and make no difference in the proofs since I_{XY} is fixed.) \square

Proof for Theorem 5.2.1. Assume that mutual information is distributed equally across each dimension as described in Condition 2, i.e. under the canonical form, the correlation vector $\vec{\rho} \in (-1, 1)^{\mathcal{D}}$ is such that

$$\rho_i = \pm \sqrt{1 - \exp\left(-\frac{2I_{XY}}{|\mathcal{D}|}\right)}$$

for each $i \in \mathcal{D}$. Then $\rho_i \neq 0$ as long as $I_{XY} > 0$. Henceforth, in this proof, we refer to this

value simply as ρ .

Consider a misalignment on a given row. Without loss of generality, we assume \mathcal{U} consists of a single user u (to which the row corresponds) and \mathcal{V} consists of two users, v which is the true pair of u , and v' , which is the user u is falsely mapped to. Let m be the correct mapping and m' the false mapping. By Corollary 11,

$$\Pr(\langle \mathbf{G}, \mathbf{m} \rangle \leq \langle \mathbf{G}, \mathbf{m}' \rangle) \leq R\left(\frac{\mathbf{m} + \mathbf{m}'}{2}\right) = R([1/2, 1/2]).$$

By Lemma 18,

$$\begin{aligned} \log R([1/2, 1/2]) &= \frac{1}{2} \sum_{i \in \mathcal{D}} \log(1 - \rho_i^2) - \log\left(1 - \frac{\rho_i^2}{2}\right) \\ &= \frac{|\mathcal{D}|}{2} \log(1 - \rho^2) - \frac{|\mathcal{D}|}{2} \log\left(1 - \frac{\rho^2}{2}\right) \\ &= -I_{XY} - \frac{|\mathcal{D}|}{2} \log\left(1 - \frac{\rho^2}{2}\right) \end{aligned}$$

By Lemma 31, this bound still holds if we remove the assumption that mutual information is distributed equally across each dimension as described in Condition 2.

There are $|\mathcal{V}|$ different misalignment events on a given row. Then, by the union bound, the log-expected number of errors on a given row is upper bounded by $\log |\mathcal{V}| - I_{XY} - \frac{|\mathcal{D}|}{2} \log\left(1 - \frac{\rho^2}{2}\right)$ where $\rho^2 = 1 - \exp\left(-\frac{2I_{XY}}{|\mathcal{D}|}\right)$. Summing up the expectation over all $|\mathcal{U}|$ rows, we get

$$\log \mathbb{E} \left[d\left(\hat{M}_{\text{MR}}, M\right) \right] \leq \log |\mathcal{U} \times \mathcal{V}| - I_{XY} - \frac{|\mathcal{D}|}{2} \log\left(\frac{1 + \exp\left(-\frac{2I_{XY}}{|\mathcal{D}|}\right)}{2}\right).$$

By Lemma 1, $d\left(\hat{M}_{\text{MR}}, M\right) \leq 2d\left(\hat{M}_{\text{TT}}, M\right)$. Then, by Theorem 5.1.1,

$$\log \mathbb{E} \left[d\left(\hat{M}_{\text{MR}}, M\right) \right] \leq \log n - \frac{|\mathcal{D}|}{2} \left[\sqrt{1 + \xi^2} - 1 + \log\left(\frac{\sqrt{1 + \xi^2} - 1}{\xi^2}\right) + \log 2 \right] + \log 2.$$

Evaluating either of these bounds at $|\mathcal{D}| \leq o(\log n)$ gives the expression for the special case.

□

5.3 Majorization of distributions

5.3.1 Majorization

Majorization is a partial ordering of multisets. Consider multisets of n real numbers $\mathcal{X} = \{x_1, x_2, \dots, x_n\}$ and $\mathcal{Y} = \{y_1, y_2, \dots, y_n\}$. Assume, without loss of generality, that $x_1 \geq x_2 \geq \dots \geq x_n$ and $y_1 \geq y_2 \geq \dots \geq y_n$.

We say \mathcal{X} **majorizes** \mathcal{Y} and write $\mathcal{X} \succ \mathcal{Y}$ if and only if

$$\sum_{k=1}^m x_k \geq \sum_{k=1}^m y_k, \quad \forall m \in \{1, 2, \dots, n-1\}$$

and

$$\sum_{k=1}^n x_k = \sum_{k=1}^n y_k.$$

Theorem 5.3.1 ([41]). $\mathcal{X} \succ \mathcal{Y}$ if and only if $\sum_{x \in \mathcal{X}} h(x) \geq \sum_{y \in \mathcal{Y}} h(y)$ for all convex continuous function h .

5.3.2 Lorenz ordering and majorization for distributions

Lorenz ordering is a scale-invariant generalization of majorization that applies to distributions. Given real-valued random variable X , let $F_X^{-1}(p) \triangleq \sup\{x \in \mathbb{R} : \Pr(X \leq x) \leq p\}$ denote its inverse cumulative distribution function.

Lorenz curve

Let X a random variable such that its expectation is finite and strictly positive. The Lorenz curve corresponding to X is defined as $L_X(p) = \frac{1}{\mathbb{E}[X]} \int_0^p F_X^{-1}(t) dt$ for any $p \in [0, 1]$.

We talk about a partial ordering of distributions if the Lorenz curve of one is greater than the Lorenz curve of the other at every point.

Non-standardized Lorenz curve

The Lorenz curve is scale-invariant, i.e. multiplying a random variable by a scalar does not change its Lorenz curve. Majorization, on the other hand, is scale dependent. Therefore, we define a non-standardized Lorenz curve to incorporate scale variance.

We define the non-standardized Lorenz curve as

$$\tilde{L}_X(p) = \int_0^p F_X^{-1}(t) dt. \quad (5.4)$$

We say X **majorizes** Y and write $X \succ Y$ if and only if

$$\begin{aligned} \tilde{L}_X(p) &\leq \tilde{L}_Y(p), \quad \forall p \in [0, 1] \\ \text{and } \tilde{L}_X(1) &= \tilde{L}_Y(1). \end{aligned}$$

The last line is equivalent to saying that $\mathbb{E}[X] = \mathbb{E}[Y]$.

Theorem 5.3.2 ([42]). *$X \succ Y$ if and only if $\mathbb{E}[h(X)] \geq \mathbb{E}[h(Y)]$ for all convex continuous function h .*

Lorenz ordering as a generalization of majorization

Let $\mathcal{X} = \{x_1, x_2, \dots, x_n\}$ be some multiset with $x_1 \leq x_2 \leq \dots \leq x_n$. Let X be a random variable sampled uniformly at random from \mathcal{X} . Then, given any $k \in \{1, 2, \dots, n-1\}$,

$$x < x_k \iff \Pr(X \leq x) < \frac{k}{n}.$$

So, it follows that,

$$\begin{aligned} \lim_{x \rightarrow x_k^-} \Pr(X \leq x) &< \frac{k}{n} \\ \text{and } \Pr(X \leq x_k) &\geq \frac{k}{n}. \end{aligned}$$

Then, for any $p \in [\frac{k-1}{n}, \frac{k}{n})$, $F_X^{-1}(p) = \sup\{x \in \mathbb{R} : \Pr[X \leq x] \leq p\} = x_k$. Therefore the

value of $F_X^{-1}(p)$ is given by

$$F_X^{-1}(p) = x_{\lceil np \rceil}.$$

It then follows that

$$\begin{aligned} \tilde{L}_X(p) &= \int_0^p F_X^{-1}(t) dt = \int_0^p x_{\lceil nt \rceil} dt \\ &= \frac{1}{n} \sum_{k=1}^{\lfloor np \rfloor} x_k + \left(p - \frac{\lfloor np \rfloor}{n} \right) x_{\lceil np \rceil} \end{aligned} \quad (5.5)$$

The value of this curve at k/n for some integer $k \in \{1, 2, \dots, n\}$ is equal to $1/n$ times the sum of the k smallest entries in \mathcal{X} . We obtain the rest of the curve by connecting the values at $0/n, 1/n, 2/n, \dots, n/n$ through straight lines.

Let \mathcal{X} and \mathcal{Y} multisets of same size and X and Y sampled uniformly at random from multisets \mathcal{X} and \mathcal{Y} respectively. We can check the Lorenz ordering between X and Y by only comparing the Lorenz curves at k/n for each integer $k \in \{1, 2, \dots, n\}$. This is equivalent to comparing the sums of the k smallest entries in \mathcal{X} and \mathcal{Y} for each k . Then $X \succ Y \iff \mathcal{X} \succ \mathcal{Y}$.

5.3.3 Lorenz ordering of terms that make up the generating function

By Lemma 16, given m_1 and m_2 two mappings of size n such that $(1 - \theta)\mathbf{m}_1 + \theta\mathbf{m}_2$ is a matrix block corresponding to a cycle of the type given in Fig. 2.3-I, the expression for the log of the generating function $R\left(\frac{\mathbf{m}_1 + \mathbf{m}_2}{2}\right)$ as defined in Definition 6 is given by

$$\log R\left(\frac{\mathbf{m}_1 + \mathbf{m}_2}{2}\right) = -\frac{I_{XY}}{2} - \frac{1}{2} \sum_{i \in \mathcal{D}} \sum_{k=0}^{n-1} \log \left(1 - \frac{\rho_i^2}{2} \left[1 + \cos\left(\frac{2k\pi}{n}\right) \right] \right).$$

This expression can be described as the expected value of some function of a random variable, as shown below:

For each $\ell \in \{2, 3, \dots\}$, define U_ℓ as a random variable distributed uniformly over the set $\{0, \frac{1}{\ell}, \frac{2}{\ell}, \dots, \frac{\ell-1}{\ell}\}$. Let U_∞ be the continuous random variable distributed uniformly over $[0, 1]$. U_ℓ converges to U_∞ in distribution as $\ell \rightarrow \infty$.

For each $\ell \in \{2, 3, \dots\}$, also define V_ℓ as the random variable a mixture distribution between that of U_2 and of U_∞ . Specifically,

$$\begin{aligned}\Pr(V_\ell = 0) &= \Pr(V_\ell = 1/2) = \frac{1}{\ell} \\ \Pr(a < V_\ell < b) &= \frac{\ell - 2}{\ell}(b - a) \\ \text{and } \Pr(a + 1/2 < V_\ell < b + 1/2) &= \frac{\ell - 2}{\ell}(b - a)\end{aligned}$$

for any $0 \leq a \leq b \leq 1/2$. U_2 and V_2 are identical in distribution. V_ℓ converges to U_∞ in distribution as $\ell \rightarrow \infty$.

Define $W_\ell \triangleq \cos(2\pi U_\ell)$. Then, $\log \log R\left(\frac{\mathbf{m}_1 + \mathbf{m}_2}{2}\right)$ can be expressed as the expected value of the following function of W_ℓ :

$$\log R\left(\frac{\mathbf{m}_1 + \mathbf{m}_2}{2}\right) = \mathbb{E}\left[-\frac{I_{XY}}{2} - \frac{1}{2n} \sum_{i \in \mathcal{D}} \log\left(1 - \frac{\rho_i^2}{2}(1 + W_\ell)\right)\right].$$

We are interested in the Lorenz ordering of $W_\ell \triangleq \cos(2\pi U_\ell)$ and $Z_\ell \triangleq \cos(2\pi V_\ell)$ for different values of ℓ . Z_ℓ is an auxiliary random variable that is to be useful in the analysis in Section 5.4. The distribution of each Z_ℓ corresponds to a mixture distribution between that of W_2 and that of W_∞ .

The plots for non-standardized Lorenz curves of W_2, W_3, W_4, W_5, W_6 and W_∞ are given in Figure 5.3. The plots for non-standardized Lorenz curves of Z_ℓ compared with that of W_2, W_ℓ and W_∞ are given in Figure 5.4. Note that $\tilde{L}_{W_2} = \tilde{L}_{Z_2}$ and $\tilde{L}_{W_\infty} = \tilde{L}_{Z_\infty}$ since (W_2, Z_2) and (W_∞, Z_∞) are identically distributed pairs.

Lemma 32. For any $\ell \in \{2, 3, \dots\} \cup \{\infty\}$, $\mathbb{E}[W_\ell] = \mathbb{E}[Z_\ell] = 0$.

Proof. Let $\text{Re} : \mathbb{C} \rightarrow \mathbb{R}$ be the function that returns the real component of a complex

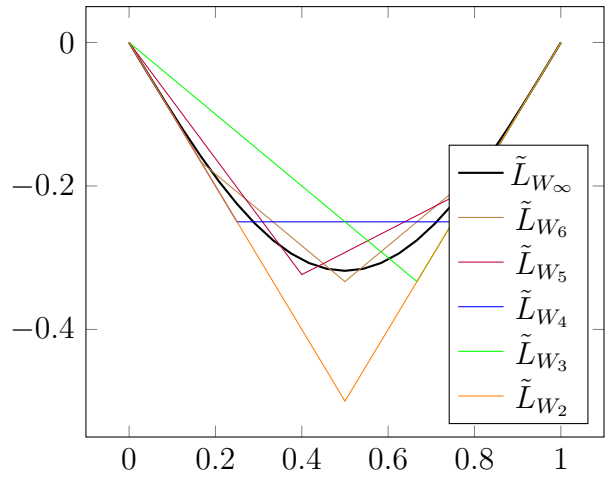


Figure 5.3: Comparison of non-standardized Lorenz curves for W_2, W_3, W_4, W_5, W_6 and W_∞

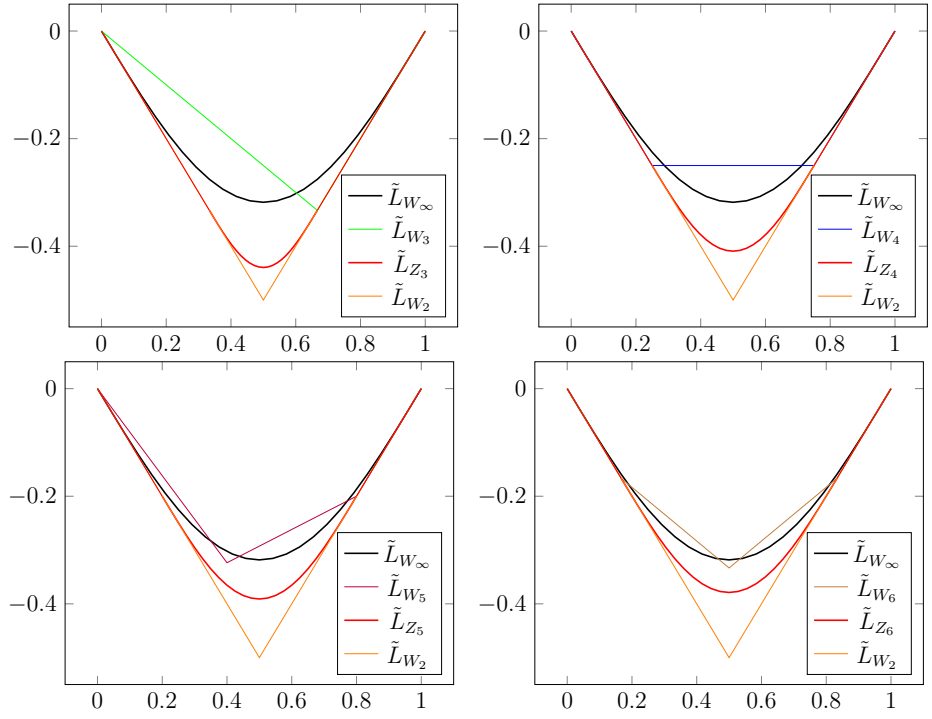


Figure 5.4: Comparison non-standardized Lorenz curves for W_2, Z_ℓ, W_ℓ and W_∞ for $\ell \in \{3, 4, 5, 6\}$.

number. For any $\ell \in \{2, 3, \dots\}$,

$$\begin{aligned}\mathbb{E}[W_\ell] &= \frac{1}{\ell} \sum_{k=0}^{\ell-1} \cos\left(2\pi \frac{k}{\ell}\right) = \frac{1}{\ell} \cdot \sum_{k=0}^{\ell-1} \operatorname{Re}\left(e^{\frac{2i\pi k}{\ell}}\right) \\ &= \frac{1}{\ell} \cdot \operatorname{Re}\left(\sum_{k=0}^{\ell-1} e^{\frac{2i\pi k}{\ell}}\right) = \frac{1}{\ell} \cdot \operatorname{Re}\left(\frac{1 - e^{\frac{2i\pi \ell}{\ell}}}{1 - e^{\frac{2i\pi 1}{\ell}}}\right) \\ &= \frac{1}{\ell} \operatorname{Re}\left(\frac{1 - 1}{1 - e^{\frac{2i\pi 1}{\ell}}}\right) = \frac{1}{\ell} \operatorname{Re}(0) = 0.\end{aligned}$$

For $\ell = \infty$,

$$\begin{aligned}\mathbb{E}[W_\infty] &= \int_0^1 \cos(2\pi x) dx \\ &= \left[\frac{\sin(2\pi x)}{2\pi}\right]_0^1 = 0.\end{aligned}$$

Finally, using the fact that the distribution of Z_ℓ is a mixture distribution between W_2 and W_∞ , we have $\mathbb{E}[Z_\ell] = \frac{2}{\ell}\mathbb{E}[W_2] + \frac{\ell-2}{\ell}\mathbb{E}[W_\infty] = 0$. \square

Lemma 33. *Given any $\ell \in \{2, 3, 4, 5, \dots\}$, W_2 majorizes W_ℓ , i.e.*

$$\begin{aligned}\tilde{L}_{W_2}(p) &\leq \tilde{L}_{W_\ell}(p), \quad \forall p \in [0, 1] \\ \text{and } \tilde{L}_{W_2}(1) &= \tilde{L}_{W_\ell}(1)\end{aligned}$$

Proof. The inverse cumulative distribution function F^{-1} for a random variable X is defined as $F_X^{-1}(p) = \sup\{x \in \mathbb{R} : \Pr[X \leq x] \leq p\}$.

Then, by (5.4), the slope of the non-standardized Lorenz curve of W_ℓ is given by

$$\left.\frac{d}{dp} \tilde{L}_{W_\ell}(p)\right|_{p=p_k} = F_{Z_\ell}^{-1}(p_k)$$

- The slope of $\tilde{L}_{W_2}(p)$ is less than the slope of $\tilde{L}_{W_\ell}(p)$ within the interval $p \in (0, 1/2)$:
 - $\Pr(W_2 < -1) = 0$ and $\Pr(W_2 \leq -1) = 1/2$. Then, for any $p \in (0, 1/2)$,

$F_{W_2}^{-1}(p) = -1$. Then the slope is -1 within the interval.

– $\Pr(W_\ell < -1) = 0$ and $\Pr(W_\ell \leq -1) \leq 1/2$. Then, for any $p \in (0, 1/2)$,

$F_{W_\ell}^{-1}(p) \geq -1$. Then the slope is at least -1 within the interval.

- The slope of $\tilde{L}_{W_2}(p)$ is greater than the slope of $\tilde{L}_{W_\ell}(p)$ within the interval $p \in (1/2, 1)$:

– $\Pr(W_2 < 1) = 1/2$ and $\Pr(W_2 \leq 1) = 1$. Then, for any $p \in (1/2, 1)$,

$F_{W_2}^{-1}(p) = 1$. Then the slope is 1 within the interval.

– $\Pr(W_\ell < 1) < 1$ and $\Pr(W_\ell \leq 1) = 1$. Then, for any $p \in (1/2, 1)$, $F_{W_\ell}^{-1}(p) \leq 1$. Then the slope is at most 1 within the interval.

Since $\tilde{L}_{W_2}(0) = \tilde{L}_{W_\ell}(0) = 0$, and since the slope of $\tilde{L}_{W_2}(p)$ is less than the slope of $\tilde{L}_{W_\ell}(p)$ within the interval $p \in (0, 1/2)$, it then follows that $\tilde{L}_{W_2}(p) \leq \tilde{L}_{W_\ell}(p)$ within that interval.

Similarly, since $\tilde{L}_{W_2}(1) = \tilde{L}_{W_\ell}(1) = 0$, and since the slope of $\tilde{L}_{W_2}(p)$ is greater than the slope of $\tilde{L}_{W_\ell}(p)$ within the interval $p \in (1/2, 1)$, it then follows that $\tilde{L}_{W_2}(p) \leq \tilde{L}_{W_\ell}(p)$ within that interval. \square

Lemma 34. *The non-standardized Lorenz curves of W_∞ and of Z_ℓ are given by the following expressions:*

- Given any $p \in [0, 1]$,

$$\tilde{L}_{W_\infty}(p) = -\frac{\sin(\pi p)}{\pi}.$$

- Given any $p \in [0, 1]$,

$$\tilde{L}_{Z_\ell} \left(\frac{1}{\ell} + \frac{\ell - 2}{\ell} \cdot p \right) = -\frac{1}{\ell} \left[1 + (\ell - 2) \cdot \frac{\sin(\pi p)}{\pi} \right]$$

and, given any $p \in [0, 1/\ell]$,

$$\begin{aligned}\tilde{L}_{Z_\ell}(p) &= -p \\ \tilde{L}_{Z_\ell}(1-p) &= -p.\end{aligned}$$

Proof. Given any $w \in [-1, 1]$ and $p \in [0, 1]$,

$$\begin{aligned}F_{W_\infty}(w) &= \Pr(W_\infty \leq w) \\ &= \Pr\left(\frac{\arccos(w)}{2\pi} \leq U_\infty \leq 1 - \frac{\arccos(w)}{2\pi}\right) \\ &= 1 - \frac{\arccos(w)}{\pi} \\ \implies F_{W_\infty}^{-1}(p) &= \cos(\pi(1-p)) = -\cos(\pi p) \\ \implies \tilde{L}_{W_\infty}(p) &= -\int_0^p \cos(\pi t) dt \\ &= -\frac{\sin(\pi p)}{\pi},\end{aligned}$$

which gives us the first part of the statement.

For any $z \in (-1, 1)$ and $p \in [0, 1]$,

$$\begin{aligned}F_{Z_\ell}(z) &= \Pr(Z_\ell \leq z) \\ &= \Pr\left(\frac{\arccos(z)}{2\pi} \leq V_\infty \leq 1 - \frac{\arccos(z)}{2\pi}\right) \\ &= \frac{1}{\ell} + \frac{\ell-2}{\ell} \left[1 - \frac{\arccos(w)}{\pi}\right] \\ \implies F_{Z_\ell}^{-1}\left(\frac{1}{\ell} + \frac{\ell-2}{\ell} \cdot p\right) &= \cos(\pi(1-p)) = -\cos(\pi p).\end{aligned}$$

and, since $\Pr(Z_\ell = -1) = \Pr(Z_\ell = 1) = 1/\ell$ and $\Pr(Z_\ell \in [-1, 1]) = 1$,

$$\begin{aligned}F_{Z_\ell}^{-1}(p) &= -1 \text{ for any } p \in [0, 1/\ell] \\ F_{Z_\ell}^{-1}(p) &= 1 \text{ for any } p \in [1 - 1/\ell, 1]\end{aligned}$$

It then follows that, for any $p \in [0, 1]$

$$\begin{aligned}
\tilde{L}_{Z_\ell} \left(\frac{1}{\ell} + \frac{\ell-2}{\ell} \cdot p \right) &= -\frac{1}{\ell} + \int_{\frac{1}{\ell}}^{\frac{1}{\ell} + \frac{\ell-2}{\ell} \cdot p} F_{Z_\ell}^{-1}(t) dt \\
&= -\frac{1}{\ell} + \frac{\ell-2}{\ell} \int_0^p F_{Z_\ell}^{-1} \left(\frac{1}{\ell} + \frac{\ell-2}{\ell} \cdot t \right) dt \\
&= \frac{1}{\ell} \left[-1 + (\ell-2) \int_0^p \cos(\pi t) dt \right] \\
&= \frac{1}{\ell} \left[-1 - (\ell-2) \cdot \frac{\sin(\pi p)}{\pi} \right]
\end{aligned}$$

and, for any $p \in [0, 1/\ell]$

$$\begin{aligned}
\tilde{L}_{Z_\ell}(p) &= -p \\
\tilde{L}_{Z_\ell}(1-p) &= -p.
\end{aligned}$$

□

Lemma 35. Given any $\ell \in \{2, 3, 4, 5, \dots\}$, Z_ℓ majorizes W_ℓ , i.e.

$$\begin{aligned}
\tilde{L}_{Z_\ell}(p) &\leq \tilde{L}_{W_\ell}(p), \quad \forall p \in [0, 1] \\
\text{and } \tilde{L}_{Z_\ell}(1) &= \tilde{L}_{W_\ell}(1)
\end{aligned}$$

Proof. By Lemma 32, $\mathbb{E}[W_\ell] = \mathbb{E}[Z_\ell] = 0$, which is equivalent to $\tilde{L}_{Z_\ell}(1) = \tilde{L}_{W_\ell}(1) = 0$.

a) Consider $\ell = 2$. Then W_ℓ and Z_ℓ are identically distributed therefore $\tilde{L}_{W_\ell} = \tilde{L}_{Z_\ell}$ at any $p \in [0, 1]$

b) Consider $\ell = 3$.

Then the curve \tilde{L}_{W_ℓ} consists of two line segments: From $(x, \tilde{L}) = (0, 0)$ to $(2/3, -1/3)$ and from $(2/3, -1/3)$ to $(1, 0)$.

$\tilde{L}_{Z_\ell}(0) = \tilde{L}_{Z_\ell}(1) = 0 = \tilde{L}_{W_\ell}(0)$ and $\tilde{L}_{Z_\ell}(2/3) = -1/3 = \tilde{L}_{W_\ell}(2/3)$. Since \tilde{L}_{Z_ℓ} is a convex function, it follows that $\tilde{L}_{W_\ell} \leq \tilde{L}_{Z_\ell}$ at any $p \in [0, 1]$.

c) Consider $\ell = 4$.

Then the curve \tilde{L}_{W_ℓ} consists of three line segments: From $(x, \tilde{L}) = (0, 0)$ to $(1/4, -1/4)$, from $(1/4, -1/4)$ to $(3/4, -1/4)$ and from $(3/4, -1/4)$ to $(1, 0)$.

$\tilde{L}_{Z_\ell}(0) = \tilde{L}_{Z_\ell}(1) = 0 = \tilde{L}_{W_\ell}(0)$ and $\tilde{L}_{Z_\ell}(1/4) = \tilde{L}_{Z_\ell}(3/4) = -1/4 = \tilde{L}_{W_\ell}(1/4) = \tilde{L}_{W_\ell}(3/4)$. Since \tilde{L}_{Z_ℓ} is a convex function, it follows that $\tilde{L}_{W_\ell} \leq \tilde{L}_{Z_\ell}$ at any $p \in [0, 1]$.

d) Suppose ℓ is even and greater than 5.

Then the non-standardized Lorenz curve $\tilde{L}_{W_\ell}(p)$ is symmetric around the axis $p = 1/2$. We only consider the left half of the curve (i.e. the segment corresponding to $p \in [0, 1/2]$) and show that, the slope of every line segment that forms \tilde{L}_{W_ℓ} is greater than the average slope of \tilde{L}_{Z_ℓ} within the interval corresponding to that line segment. Since both Lorenz curves start at $(p, \tilde{L}(p)) = (0, 0)$, it then follows that $\tilde{L}_{W_\ell}(p) \geq \tilde{L}_{Z_\ell}(p)$ for any value of $p \in [0, 1/2]$. Since both \tilde{L}_{W_ℓ} and \tilde{L}_{Z_ℓ} are symmetric, we then have $\tilde{L}_{W_\ell}(p) \geq \tilde{L}_{Z_\ell}(p)$ for all $p \in [0, 1]$, i.e. Z_ℓ majorizes W_ℓ .

W_ℓ can be described as a discrete random variable distributed uniformly over the multi-set (i.e. set with duplicates) that contains one copy of -1 , two copies of $-\cos\left(\frac{2\pi k}{\ell}\right)$ for each $k \in \{1, 2, \dots, \frac{\ell}{2} - 1\}$, and one copy of 1 . Then, given any $k \in \{1, 2, \dots, \frac{\ell}{2} - 1\}$ and $p_k \in \left(\frac{2(k-1)+1}{\ell}, \frac{2k+1}{\ell}\right)$,

$$F_{W_\ell}^{-1}(p_k) = \sup\{w \in \mathbb{R} : \Pr[W_\ell \leq w] \leq p_k\} = -\cos\left(\frac{2\pi k}{\ell}\right).$$

Then, by (5.4), the slope of the non-standardized Lorenz curve of W_ℓ is given by

$$\left. \frac{d}{dp} \tilde{L}_{W_\ell}(p) \right|_{p=p_k} = F_{W_\ell}^{-1}(p_k) = -\cos\left(\frac{2\pi k}{\ell}\right)$$

It is easy to show that, given any $p_0 \in (0, 1/\ell)$,

$$\left. \frac{d}{dp} \tilde{L}_{W_\ell}(p) \right|_{p=p_0} = F_{W_\ell}^{-1}(p_0) = -1$$

This exactly matches the slope of \tilde{L}_{Z_ℓ} at $p = p_0$ for each $p_0 \in (0, 1/\ell)$. Therefore \tilde{L}_{W_ℓ}

and \tilde{L}_{Z_ℓ} have the same slope within the intervals $p \in [0, 1/\ell]$.

Next we find the average slope of \tilde{L}_{Z_ℓ} for the interval $\left(\frac{2(k-1)+1}{\ell}, \frac{2k+1}{\ell}\right)$ for each $k \in \{1, 2, \dots, \frac{\ell}{2} - 1\}$. By Lemma 34:

$$\begin{aligned}
& \frac{\ell}{2} \left[\tilde{L}_{Z_\ell} \left(\frac{2k+1}{\ell} \right) - \tilde{L}_{Z_\ell} \left(\frac{2(k-1)+1}{\ell} \right) \right] \\
&= -\frac{\ell-2}{2\pi} \left[\sin \left(\frac{2\pi k}{\ell-2} \right) - \sin \left(\frac{2\pi(k-1)}{\ell-2} \right) \right] \\
&= -\frac{\ell-2}{\pi} \cdot \sin \left(\frac{\pi}{\ell-2} \right) \cdot \cos \left(\frac{2\pi(k-1/2)}{\ell-2} \right) \\
&= -\cos \left(\frac{2\pi(k-1/2)}{\ell-2} \right) + \left(1 - \frac{\ell-2}{\pi} \cdot \sin \left(\frac{\pi}{\ell-2} \right) \right) \cdot \cos \left(\frac{2\pi(k-1/2)}{\ell-2} \right) \\
&\leq -\cos \left(\frac{2\pi(k-1/2)}{\ell-2} \right) + \left(1 - \frac{\ell-2}{\pi} \cdot \sin \left(\frac{\pi}{\ell-2} \right) \right) \\
&\leq -\cos \left(\frac{2\pi(k-1/2)}{\ell-2} \right) + \frac{\pi^2}{6(\ell-2)^2}
\end{aligned}$$

Finally, we show that the difference between $\cos \left(\frac{2\pi(k-1/2)}{\ell-2} \right)$ and $\cos \left(\frac{2\pi k}{\ell} \right)$ is at least $\frac{\pi^2}{6(\ell-2)^2}$ within the range $k \in [1, \frac{\ell-1}{4}]$ for any $\ell \geq 5$.

Applying Lemma 54 with $x = \frac{k}{\ell} - \frac{1}{4}$ and $\alpha = \frac{\ell}{\ell-2}$ we get

$$\begin{aligned}
f \left(\frac{k}{\ell} - \frac{1}{4}, \frac{\ell}{\ell-2} \right) &= -\cos \left(\frac{2\pi k}{\ell} \right) - \sin \left(\frac{2\pi k}{\ell-2} - \frac{2\pi \ell}{4(\ell-2)} \right) \\
&= -\cos \left(\frac{2\pi k}{\ell} \right) + \cos \left(\frac{2\pi k}{\ell-2} - \frac{2\pi \ell}{4(\ell-2)} + \frac{\pi}{2} \right) \\
&= -\cos \left(\frac{2\pi k}{\ell} \right) + \cos \left(\frac{2\pi(k-1/2)}{\ell-2} \right)
\end{aligned}$$

which is exactly the difference we are interested in. Once again making the substitutions

$x = \frac{k}{\ell} - \frac{1}{4}$ and $\alpha = \frac{\ell}{\ell-2}$, we get $1 < \alpha \leq 5/3 \iff 5 \leq \ell$,

$$\begin{aligned}
& -\frac{1}{4\alpha} \leq x \leq -\frac{\alpha-1}{8\alpha} \\
\iff & -\frac{\ell-2}{4\ell} \leq \frac{k}{\ell} - \frac{1}{4} \leq -\frac{1}{4\ell} \\
\iff & \frac{1}{2} \leq k \leq \frac{\ell-1}{4}
\end{aligned}$$

and

$$f(x, \alpha) \geq \frac{\pi^2(\alpha - 1)^2}{24} = \frac{\pi^2}{6(\ell - 2)^2}.$$

Then, by Lemma 54, $\cos\left(\frac{2\pi(k-1/2)}{\ell-2}\right) - \cos\left(\frac{2\pi k}{\ell}\right)$ is at least $\frac{\pi^2}{6(\ell-2)^2}$ within the range $k \in [1, \frac{\ell-1}{4}]$ for any $\ell \geq 5$.

Consequently, the difference between the slope of $\tilde{L}_{W_\ell}(p)$ and the average slope of $\tilde{L}_{Z_\ell}(p)$ is positive within the interval $p \in \left(\frac{2(k-1)+1}{\ell}, \frac{2k+1}{\ell}\right)$ for each $k \in \{1, 2, \dots, \lfloor \frac{\ell-1}{4} \rfloor\}$:

$$\begin{aligned} & \left[-\cos\left(\frac{2\pi k}{\ell}\right) \right] - \left[-\cos\left(\frac{2\pi(k-1/2)}{\ell-2}\right) + \frac{\pi^2}{6(\ell-2)^2} \right] \\ &= \left[\cos\left(\frac{2\pi(k-1/2)}{\ell-2}\right) - \cos\left(\frac{2\pi k}{\ell}\right) \right] - \frac{\pi^2}{6(\ell-2)^2} \\ &\geq \frac{\pi^2}{6(\ell-2)^2} - \frac{\pi^2}{6(\ell-2)^2} = 0, \end{aligned}$$

i.e. the slope of $\tilde{L}_{W_\ell}(p)$ is greater than the average slope of $\tilde{L}_{Z_\ell}(p)$.

All Lorenz curves are convex. Since $\tilde{L}_{Z_\ell}(1/\ell) = \tilde{L}_{W_\ell}(1/\ell)$ and since \tilde{L}_{Z_ℓ} is a convex function, the inequality between the slope of \tilde{L}_{W_ℓ} and the average slope of \tilde{L}_{Z_ℓ} within each interval implies that $\tilde{L}_{W_\ell}(p) \geq \tilde{L}_{Z_\ell}(p)$ within each of the intervals. Combining each of these (and adding the interval $p \in [0, 1/\ell]$ which was shown at the beginning),

we have $\tilde{L}_{W_\ell}(p) \geq \tilde{L}_{Z_\ell}(p)$ within the interval $p \in \left[0, \frac{2\lfloor(\ell-1)/4\rfloor+1}{\ell}\right]$. Then, by symmetry,

we also have $\tilde{L}_{W_\ell}(p) \geq \tilde{L}_{Z_\ell}(p)$ within the interval $p \in \left[1 - \frac{2\lfloor(\ell-1)/4\rfloor+1}{\ell}, 1\right]$.

It just remains to show the inequality for the interval $p \in \left[\frac{2\lfloor(\ell-1)/4\rfloor+1}{\ell}, 1 - \frac{2\lfloor(\ell-1)/4\rfloor+1}{\ell}\right]$.

This interval is non-empty only if ℓ is a multiple of 4. Then the interval corresponds to $\left[\frac{1}{2} - \frac{1}{\ell}, \frac{1}{2} + \frac{1}{\ell}\right]$. The slope of \tilde{L}_{W_ℓ} is zero and the average slope of \tilde{L}_{Z_ℓ} is zero within this interval. Then, once again, by the convexity of \tilde{L}_{Z_ℓ} , $\tilde{L}_{W_\ell}(p) \geq \tilde{L}_{Z_\ell}(p)$ within this last interval.

e) Suppose ℓ is odd and at least 5.

We can describe Z_ℓ as a discrete random variable distributed uniformly over the multiset

(i.e. set with duplicates) two copies of $-\cos\left(\frac{2\pi(k-1/2)}{\ell}\right)$ for each $k \in \{1, 2, \dots, \frac{\ell-1}{2}\}$, and one copy of 1. Then, given any $k \in \{1, 2, \dots, \frac{\ell-1}{2}\}$ and $p_k \in \left(\frac{2(k-1)}{\ell}, \frac{2k}{\ell}\right)$

$$\begin{aligned} F_{Z_\ell}^{-1}(p_k) &= \sup\{z \in \mathbb{R} : \Pr[Z_\ell \leq z] \leq p_k\} \\ &= -\cos\left(\frac{2\pi(k-1/2)}{\ell}\right). \end{aligned}$$

Then, by (5.4), the slope of the non-standardized Lorenz curve of W_ℓ is given by

$$\begin{aligned} \left. \frac{d}{dp} \tilde{L}_{W_\ell}(p) \right|_{p=p_k} &= F_{Z_\ell}^{-1}(p_k) \\ &= -\cos\left(\frac{2\pi(k-1/2)}{\ell}\right) \end{aligned}$$

Next we find the average slope of \tilde{L}_{Z_ℓ} for the interval $\left(\frac{2(k-1)}{\ell}, \frac{2k}{\ell}\right)$ for each $k \in \{1, 2, \dots, \frac{\ell}{2} - 1\}$.

By Lemma 34:

$$\begin{aligned} &\frac{\ell}{2} \left[\tilde{L}_{Z_\ell}\left(\frac{2k}{\ell}\right) - \tilde{L}_{Z_\ell}\left(\frac{2(k-1)}{\ell}\right) \right] \\ &= -\frac{\ell-2}{2\pi} \left[\sin\left(\frac{\pi(2k-1)}{\ell-2}\right) - \sin\left(\frac{\pi(2k-3)}{\ell-2}\right) \right] \\ &= -\frac{\ell-2}{\pi} \cdot \sin\left(\frac{\pi}{\ell-2}\right) \cdot \cos\left(\frac{2\pi(k-1)}{\ell-2}\right) \\ &= -\cos\left(\frac{2\pi(k-1)}{\ell-2}\right) + \left(1 - \frac{\ell-2}{\pi} \cdot \sin\left(\frac{\pi}{\ell-2}\right)\right) \cdot \cos\left(\frac{2\pi(k-1)}{\ell-2}\right) \\ &\leq -\cos\left(\frac{2\pi(k-1)}{\ell-2}\right) + \left(1 - \frac{\ell-2}{\pi} \cdot \sin\left(\frac{\pi}{\ell-2}\right)\right) \\ &\leq -\cos\left(\frac{2\pi(k-1)}{\ell-2}\right) + \frac{\pi^2}{6(\ell-2)^2} \end{aligned}$$

This value is between $-\cos\left(\frac{2\pi(k-1)}{\ell-2}\right)$ and $-\cos\left(\frac{2\pi(k-1)}{\ell-2}\right) + \frac{\pi^2}{6(\ell-2)^2}$.

We compare the slope of \tilde{L}_{W_ℓ} , which is $-\cos\left(\frac{2\pi(k-1/2)}{\ell}\right)$, to the average slope of \tilde{L}_{Z_ℓ} , which is between $-\cos\left(\frac{2\pi(k-1)}{\ell-2}\right)$ and $-\cos\left(\frac{2\pi(k-1)}{\ell-2}\right) + \frac{\pi^2}{6(\ell-2)^2}$. Applying Lemma 54

with $x = \frac{k-1/2}{\ell} - \frac{1}{4}$ and $\alpha = \frac{\ell}{\ell-2}$ we get

$$\begin{aligned} f\left(\frac{k-1/2}{\ell} - \frac{1}{4}, \frac{\ell}{\ell-2}\right) &= -\cos\left(\frac{2\pi(k-1/2)}{\ell}\right) - \sin\left(\frac{2\pi(k-1/2)}{\ell-2} - \frac{2\pi\ell}{4(\ell-2)}\right) \\ &= -\cos\left(\frac{2\pi(k-1/2)}{\ell}\right) + \cos\left(\frac{2\pi(k-1/2)}{\ell-2} - \frac{2\pi\ell}{4(\ell-2)} + \frac{\pi}{2}\right) \\ &= -\cos\left(\frac{2\pi(k-1/2)}{\ell}\right) + \cos\left(\frac{2\pi(k-1)}{\ell-2}\right) \end{aligned}$$

which is exactly the difference we are interested in.

We study the left-hand and right-hand part of \tilde{L}_{W_ℓ} separately.

i) Left-hand part, specifically $p \in \left[0, \frac{2\lfloor(\ell+1)/4\rfloor+1}{\ell}\right]$:

Making the substitutions $x = \frac{k-1/2}{\ell} - \frac{1}{4}$ and $\alpha = \frac{\ell}{\ell-2}$, we get $1 < \alpha \leq 5/3 \iff 5 \leq \ell$,

$$\begin{aligned} -\frac{1}{4\alpha} \leq x \leq -\frac{\alpha-1}{8\alpha} \\ \iff 1 \leq k \leq \frac{\ell+1}{4} \end{aligned}$$

and

$$f(x, \alpha) \geq \frac{\pi^2(\alpha-1)^2}{24} = \frac{\pi^2}{6(\ell-2)^2}.$$

Then, by Lemma 54, $\cos\left(\frac{2\pi(k-1/2)}{\ell-2}\right) - \cos\left(\frac{2\pi k}{\ell}\right)$ is at least $\frac{\pi^2}{6(\ell-2)^2}$ within the range $k \in \left[1, \frac{\ell+1}{4}\right]$ for any $\ell \geq 5$.

Consequently, the difference between the slope of $\tilde{L}_{W_\ell}(p)$ and the average slope of $\tilde{L}_{Z_\ell}(p)$ is positive within the interval $p \in \left(\frac{2(k-1)+1}{\ell}, \frac{2k+1}{\ell}\right)$ for each $k \in \{1, 2, \dots, \lfloor \frac{\ell+1}{4} \rfloor\}$.

Since $\tilde{L}_{Z_\ell}(0) = \tilde{L}_{W_\ell}(0) = 0$ \tilde{L}_{W_ℓ} is a convex function, the inequality between the slope of \tilde{L}_{W_ℓ} and the average slope of \tilde{L}_{Z_ℓ} within each interval implies that $\tilde{L}_{W_\ell}(p) \geq \tilde{L}_{Z_\ell}(p)$ within each of the intervals. Combining each of these, we have $\tilde{L}_{W_\ell}(p) \geq \tilde{L}_{Z_\ell}(p)$ within the interval $p \in \left[0, \frac{2\lfloor(\ell+1)/4\rfloor+1}{\ell}\right]$.

ii) Right-hand part, specifically $p \in \left[\frac{2\lceil(\ell+2)/4\rceil-2}{\ell}, 1 \right]$:

Making the substitutions $x = \frac{k-1/2}{\ell} - \frac{1}{4}$ and $\alpha = \frac{\ell}{\ell-2}$, we get

$$\begin{aligned} 0 &\leq x \leq \frac{1}{2(\alpha+1)} \\ \iff 0 &\leq \frac{k-1/2}{\ell} - \frac{1}{4} \leq \frac{\ell-2}{\ell-1} = 1 - \frac{1}{\ell-1} \\ \iff \frac{\ell}{4} + \frac{1}{2} &\leq \ell \left(\frac{5}{4} - \frac{1}{\ell-1} \right) + \frac{1}{2} \end{aligned}$$

If $\ell \geq 5$, this implies $\frac{\ell+2}{4} \leq k \leq \ell$. Then, by Lemma 54, $\cos\left(\frac{2\pi(k-1/2)}{\ell-2}\right) - \cos\left(\frac{2\pi k}{\ell}\right)$ is at most 0 within the range $k \in \left[\frac{\ell+2}{4}, \ell\right]$ for any $\ell \geq 5$.

Consequently, the difference between the slope of $\tilde{L}_{W_\ell}(p)$ and the average slope of $\tilde{L}_{Z_\ell}(p)$ is negative within the interval $p \in \left(\frac{2(k-1)+1}{\ell}, \frac{2k+1}{\ell}\right)$ for each $k \in \left\{\lceil\frac{\ell+2}{4}\rceil, \dots, \frac{\ell-1}{2}\right\}$.

Since $\tilde{L}_{W_\ell}(1) = \tilde{L}_{Z_\ell}(1) = 0$ and since \tilde{L}_{Z_ℓ} is a convex function, the inequality between the slope of \tilde{L}_{W_ℓ} and the average slope of \tilde{L}_{Z_ℓ} implies that $\tilde{L}_{W_\ell}(p) \geq \tilde{L}_{Z_\ell}(p)$ within each of the intervals. Combining each of these, we have $\tilde{L}_{W_\ell}(p) \geq \tilde{L}_{Z_\ell}(p)$ within the interval $p \in \left[\frac{2\lceil(\ell+2)/4\rceil-2}{\ell}, 1\right]$.

For any odd ℓ , either

$$\left\lfloor \frac{\ell+1}{4} \right\rfloor = \left\lceil \frac{\ell+2}{4} \right\rceil - 1.$$

Then

$$\frac{2\lfloor(\ell+1)/4\rfloor}{\ell} = \frac{2\lceil(\ell+2)/4\rceil-2}{\ell}.$$

So our proofs for the left-hand part corresponding to the interval $p \in \left[0, \frac{2\lfloor(\ell+1)/4\rfloor+1}{\ell}\right]$ and the right-hand part corresponding to the interval $p \in \left[\frac{2\lceil(\ell+2)/4\rceil-2}{\ell}, 1\right]$ spans the entirety of the interval $[0, 1]$. So $\tilde{L}_{W_\ell}(p) \geq \tilde{L}_{Z_\ell}(p)$ for any $p \in [0, 1]$.

□

5.4 Maximum likelihood estimation

We present an achievability result that extends the linear boundary for maximum likelihood estimation from Theorem 4.4.2 beyond $I_{XY} < 2 \log n$ in the case where the number of dimensions is $\mathcal{O}(\log n)$.

Theorem 5.4.1. *Let $n \triangleq |M|$ and suppose $|\mathcal{U}| = |\mathcal{V}| = n$. Let m some bijective mapping between \mathcal{U} and \mathcal{V} . Consider a pair of correlated Gaussian databases with mutual information equal to $x \log n$ and effective number of dimensions (as defined in Definition 1) equal to $z \log n$.*

If $x \geq x_z$ where $x_z \triangleq z \log(2 \exp(1/z) - 1)$, then

$$\mathbb{E} \left[d \left(\hat{M}_{ML}, M \right) \right] \leq \frac{n^{2-x}}{1 - n^{1 - \frac{x+x_z}{2}}}.$$

In particular, if $n \geq \omega(1)$ and if $x \geq 1 + \omega(1/\log n)$, then

$$\mathbb{E} \left[d \left(\hat{M}_{ML}, M \right) \right] \leq n^{2-x} (1 + o(1)).$$

The achievability regions for various values of the number of dimensions $|\mathcal{D}|$ are illustrated in Fig. 5.5.

A sketch of the proof of this result is given in Subsection 5.4.1. A description of the contents and purpose of Subsections 5.4.3 and 5.4.4 can be found in this subsection. In Subsection 5.4.2, we present a lemma that shows that any bound applicable for the case where mutual information is distributed equally among all dimensions is applicable in general with no such constraint, which allows us to extend our results beyond the assumption that mutual information is distributed equally among all dimensions. Finally, in Subsection 5.4.5, the proof for Theorem 5.1.1 is given based on the results from Subsections 5.4.3 and 5.4.4.

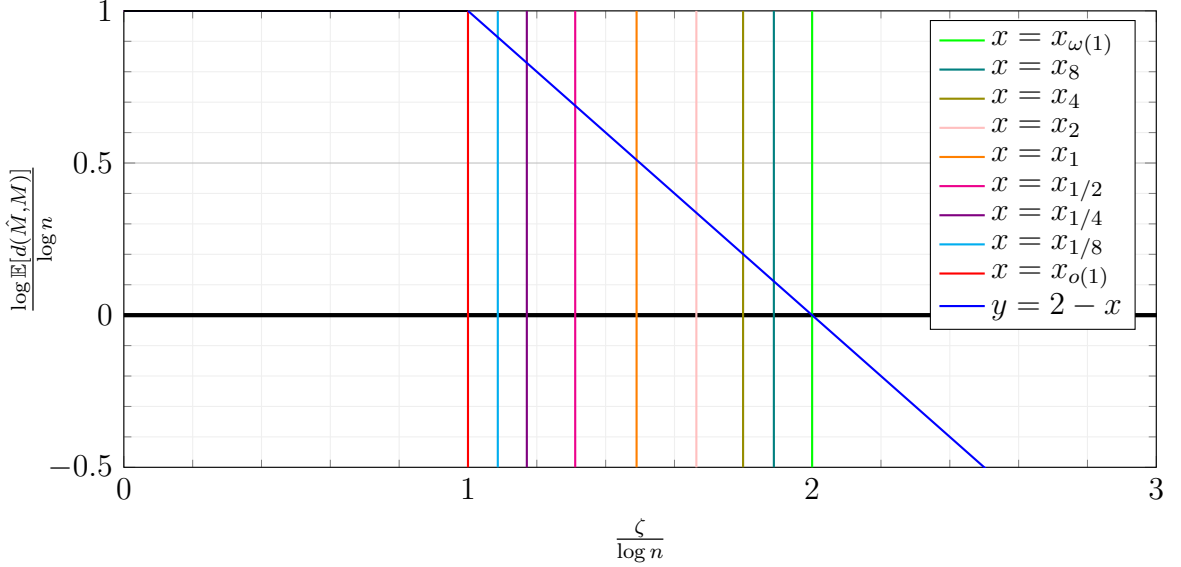


Figure 5.5: Comparison of boundaries for achievability regions for maximum likelihood estimation for various values of the number of dimensions under the fully covered case (i.e. $|\mathcal{U}| = |\mathcal{V}| = |M|$). The x-axis is to signal strength $\frac{\zeta}{\log n} = \frac{I_{XY}}{\log n}$ and y-axis is the order of magnitude of expected number of mismatched users. By Theorem 5.4.1, if the number of dimensions is $z \log n$, then the achievability region is the area to the right of $x = x_z \triangleq z \log(2 \exp(1/z) - 1)$ and to the right of $y = x - 2$.

5.4.1 Sketch of proof

In Subsection 5.4.3, we define some sequence of functions h_ℓ such that the log expected number of errors that result from elementary misalignments of size ℓ is bounded by $\ell \cdot (1 - x + h_\ell) \log n$ where $x = \frac{I_{XY}}{\log n}$. The plots for $\ell \cdot (1 - x + h_\ell)$ for $\ell \in \{2, 3, 4, 5, 6, 7, 8, 9, 10\}$ in the $\log n$, $0.15 \log n$ and $1000 \log n$ dimensional cases are given in Figures 5.6, 5.8 and 5.7 respectively.

The slope of the curve $\ell \cdot (1 - x + h_\ell)$ gets steeper as ℓ increases. In particular, it is exactly -1 for $\ell = 2$ and it is less than $-\ell/2$ for any ℓ . Eventually, as $\ell \rightarrow \infty$, the curve converges to a vertical line. We identify the point at which this line intersects the x-axis in Subsection 5.4.4. We refer to this as the critical point. The position of the critical point depends on the number of dimensions. Therefore we denote the value of this point by x_z where z refers to the ratio between the number of dimensions and $\log n$.

The achievability result relies on the fact that, as shown in in Subsection 5.4.4, the curve

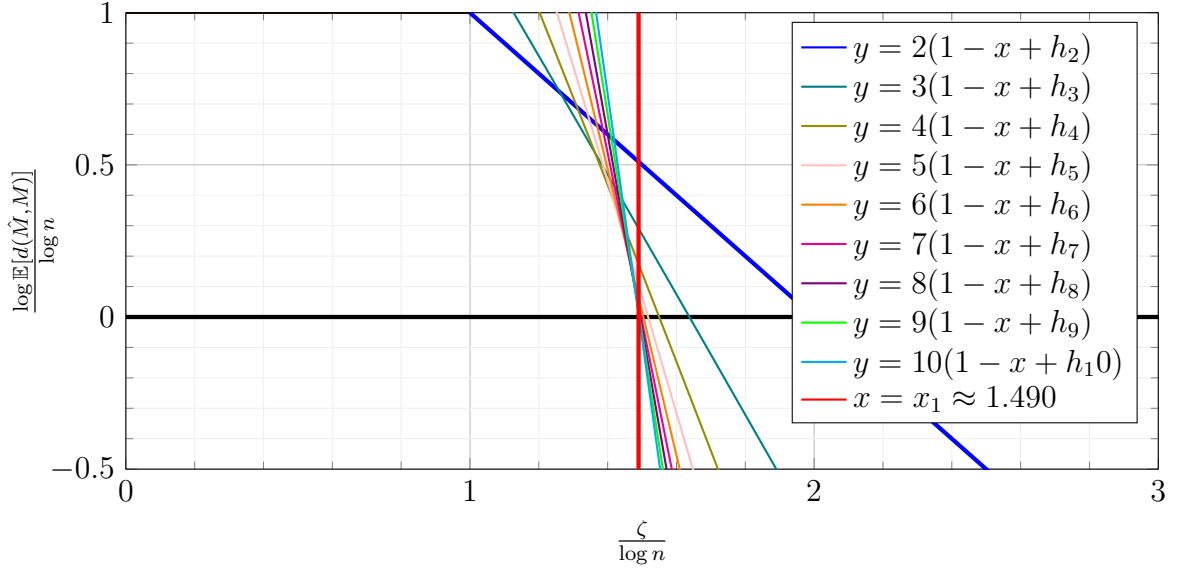


Figure 5.6: Comparison of error bound exponents for elementary misalignments of different sizes ℓ in the $\log n$ dimensional case. The curve for $\ell = 2$ dominates the others if $x > x_1$. The curve for $\ell(1 - x + h_\ell)$ converges to the line $x = x_1$ as $\ell \rightarrow \infty$.

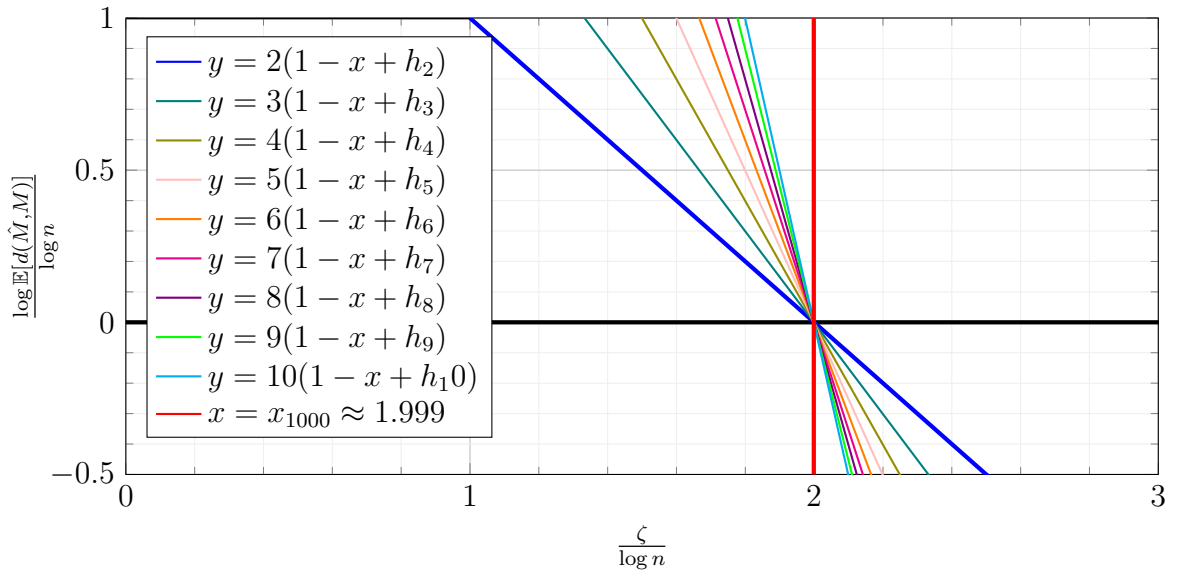


Figure 5.7: Comparison of error bound exponents for elementary misalignments of different sizes ℓ in the $1000 \log n$ dimensional case. The curve for $\ell = 2$ dominates the others if $x > x_{1000}$. The curve for $\ell(1 - x + h_\ell)$ converges to the line $x = x_{1000}$ as $\ell \rightarrow \infty$.

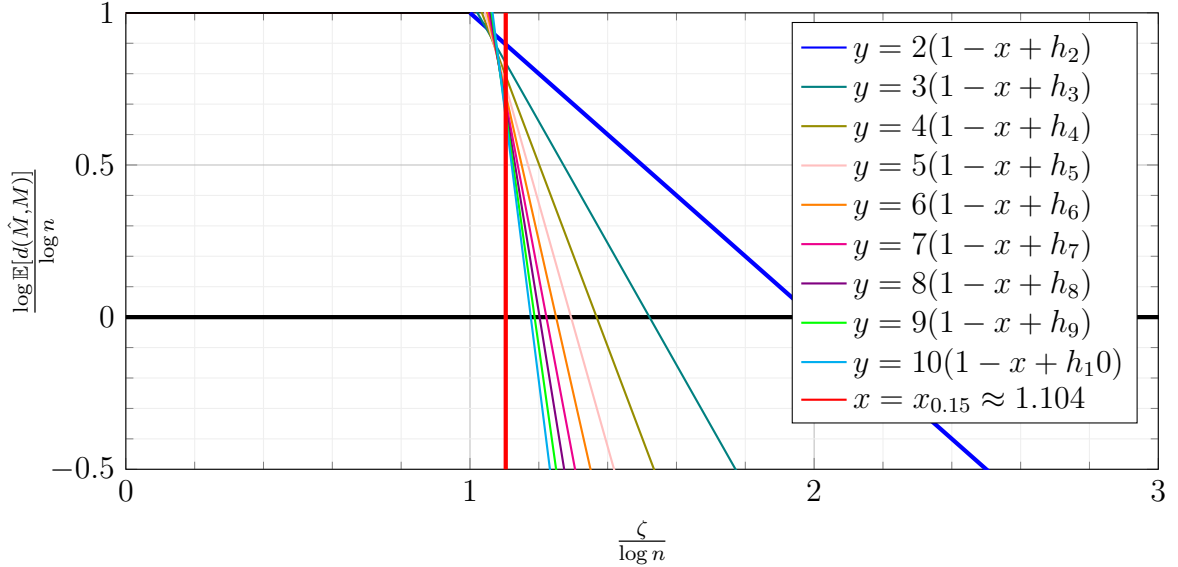


Figure 5.8: Comparison of error bound exponents for elementary misalignments of different sizes ℓ in the $0.15 \log n$ dimensional case. The curve for $\ell = 2$ dominates the others if $x > x_{0.15}$. The curve for $\ell(1 - x + h_\ell)$ converges to the line $x = x_{0.15}$ as $\ell \rightarrow \infty$.

for $\ell = 2$ dominates all the others if x is to the right of the critical point. At any point to the left of the critical point, given any ℓ (including $\ell = 2$), we can find some other ℓ' such that the very steep curve for ℓ' dominates the curve for ℓ . Therefore, within the region to the left of the critical point, we cannot bound the curves by some curve with finite slope. Therefore this analysis does not give any achievability result for $\frac{I_{XY}}{\log n} < x_z$. For that region, we need an analysis that better deals with correlations between error events, such as the one in Chapters 3 and 4 that gave us the elliptic and parabolic achievability boundaries for maximum likelihood estimation.

It is worth noting that, in the $\omega(\log n)$ dimensional case, the infinite slope boundary at the critical point ($x_{\omega(1)} = 2$) matches the achievability results in Theorems 3.3.2 and 4.4.2 and the converse result in Theorem 3.4.1, which include a vertical boundary at $x = 2$. This suggests that the infinite slope boundary might be meaningful, although we wouldn't expect this boundary to extend all the way to $y = 1$.

5.4.2 Convexity of generating function

Lemma 36. Consider two pairs of correlated databases (\vec{A}_0, \vec{B}_0) and (\vec{A}_1, \vec{B}_1) between the same user sets \mathcal{U} and \mathcal{V} and with the same underlying mapping M . Suppose $|\mathcal{U}| = |\mathcal{V}| = \ell$ and suppose the two pairs of databases have the same number of effective features, as defined in Definition 1. Let $R_0, R_1 : \mathbb{R}^{\ell \times \ell} \rightarrow \mathbb{R}$ denote the generating functions for the information density matrices corresponding to the two pairs of correlated databases as defined in Section 4.2.

Let m_1 and m_2 be two mappings of size n such that $\Theta = \frac{m_1 + m_2}{2}$ is a matrix block corresponding to a cycle of the type given in Fig. 2.3-I.

If R_0 corresponds to a pair of databases with mutual information distributed equally among all dimensions (Condition 2), then $R_0(\Theta) \geq R_1(\Theta)$.

Proof. By Lemma 16,

$$\begin{aligned} R(\Theta) &= -\ell \cdot I_{XY} - \frac{1}{2} \sum_{i \in \mathcal{D}} \sum_{k=0}^{\ell-1} \log \left(1 - \frac{\rho_i^2}{2} \left[1 + \cos \left(\frac{2k\pi}{\ell} \right) \right] \right) \\ &= \sum_{k=0}^{\ell-1} \left[-I_{XY} - \frac{1}{2} \sum_{i \in \mathcal{D}} \log \left(1 - \frac{\rho_i^2}{2} \left[1 + \cos \left(\frac{2k\pi}{\ell} \right) \right] \right) \right] \end{aligned} \quad (5.6)$$

The proof of the lemma statement can be shown by exactly following the proof for Lemma 26 for each value of k . We simply need to replace the expression $\log(1 - \rho_i^2 \theta^2)$ in (5.1) with $\log \left(1 - \rho_i^2 \cdot \frac{1 + \cos(\frac{2k\pi}{\ell})}{2} \right)$. The proof for Lemma 26 holds for any value of $\theta \in [0, 1]$, which includes $\theta^2 = \frac{1 + \cos(\frac{2k\pi}{\ell})}{2}$.

(The $-I_{XY}\theta$ term in (5.1) and the $-I_{XY}$ term in (5.6) are not important and make no difference in the proofs since I_{XY} is fixed.) \square

5.4.3 Normalized error exponent

Definition 7. Given any $\ell \in \mathbb{N}$, define $h_\ell : (0, \infty)^2 \rightarrow \mathbb{R}$ such that

$$h_\ell(x, z) \triangleq -\frac{z}{2\ell} \sum_{k=0}^{\ell-1} \log \left(1 - \frac{\rho^2}{2} \left(1 + \cos \left(\frac{2\pi k}{\ell} \right) \right) \right)$$

where $\rho^2 = \rho^2(x, z) \triangleq 1 - \exp \left(-\frac{2x}{z} \right)$.

In particular, for $\ell = 2$, $h_2(x, z) = \frac{x}{2}$ for any $z \in (0, \infty)$.

Lemma 37. Let $n \triangleq |M|$ and suppose $|\mathcal{U}| = |\mathcal{V}| = n$. Let m some bijective mapping between \mathcal{U} and \mathcal{V} . Given any $\ell \in \{2, 3, \dots, n\}$, let $\mathcal{M}_\ell^\circ(m)$ be the set of all bijective mappings between \mathcal{U} and \mathcal{V} such that $m' \in \mathcal{M}_\ell(m)$ if and only if the disagreement between m and m' corresponds to a type-I misalignment of size ℓ as illustrated in Fig. 2.2-I.

Consider a pair of correlated Gaussian databases with mutual information equal to $x \log n$ and effective number of dimensions (as defined in Definition 1) equal to $z \log n$.

Then, given any $\ell \in \{2, 3, \dots, n\}$,

$$\sum_{m' \in \mathcal{M}_\ell^\circ(m)} \ell \cdot \Pr (\langle \mathbf{G}, \mathbf{m} \rangle \leq \langle \mathbf{G}, \mathbf{m}' \rangle | M = m) \leq n^{\ell \cdot (1 - x + h_\ell(x, z))}.$$

Proof. The claimed result can be derived by the following sequence of inequalities:

$$\begin{aligned} & \sum_{m' \in \mathcal{M}_\ell^\circ(m)} \ell \cdot \Pr (\langle \mathbf{G}, \mathbf{m} \rangle \leq \langle \mathbf{G}, \mathbf{m}' \rangle | M = m) \\ & \stackrel{(i)}{\leq} \sum_{m' \in \mathcal{M}_\ell^\circ(m)} \ell \cdot R \left(\frac{\mathbf{m} + \mathbf{m}'}{2} \right) \\ & \stackrel{(ii)}{=} \sum_{m' \in \mathcal{M}_\ell^\circ(m)} \ell \cdot \exp (-\ell \cdot I_{XY} + \ell \cdot h_\ell (I_{XY}, |\mathcal{D}|)) \\ & \stackrel{(iii)}{\leq} n^\ell \cdot \exp (-\ell \cdot I_{XY} + \ell \cdot h_\ell (I_{XY}, |\mathcal{D}|)) \\ & \stackrel{(iv)}{=} n^{\ell \left(1 - \frac{I_{XY}}{\log n} + h_\ell \left(\frac{I_{XY}}{\log n}, \frac{|\mathcal{D}|}{\log n} \right) \right)}. \end{aligned}$$

The argument behind each inequality is given below:

i) By Corollary 11, $\Pr(\langle \mathbf{G}, \mathbf{m} \rangle < \langle \mathbf{G}, \mathbf{m}' \rangle | M = m) \leq R\left(\frac{\mathbf{m} + \mathbf{m}'}{2}\right)$.

ii) Let $\vec{\rho} \in (-1, 1)^{\mathcal{D}}$ denote the correlation vector under canonical form. By Lemma 16,

$$\log R\left(\frac{\mathbf{m} + \mathbf{m}'}{2}\right) \leq -\ell \cdot I_{XY} - \frac{1}{2} \sum_{i \in \mathcal{D}} \sum_{k=0}^{\ell-1} \log\left(1 - \frac{\rho_i^2}{2} \left(1 + \cos\left(\frac{2\pi k}{\ell}\right)\right)\right).$$

By Lemma 36, the generating function only gets larger if mutual information is distributed equally across each dimension as described in Condition 2, so that is a valid upper bound. In that case, $\rho_i^2 = \rho^2 \triangleq 1 - \exp\left(-\frac{2I_{XY}}{|\mathcal{D}|}\right)$ for any $i \in \mathcal{D}$ and the summation over \mathcal{D} can be written as a product $|\mathcal{D}| \sum_k \log\left(1 - \frac{\rho^2}{2} \left(1 + \cos\left(\frac{2\pi k}{\ell}\right)\right)\right)$. Then $\log R\left(\frac{\mathbf{m} + \mathbf{m}'}{2}\right) = -\ell \cdot I_{XY} + \ell \cdot h_\ell(I_{XY}, |\mathcal{D}|)$.

iii) By Lemma 2, $\mathcal{M}_\ell^{\circ}(m) \leq \frac{n^\ell}{\ell}$.

iv) $\frac{h_\ell(a, b)}{c} = h_\ell\left(\frac{a}{c}, \frac{b}{c}\right)$.

□

Lemma 38. Given any $\ell \in \{2, 3, 4, \dots\}$,

$$\frac{\partial}{\partial x} h_\ell(x, z) \leq \frac{\partial}{\partial x} h_2(x, z) = \frac{1}{2}$$

Proof. First we show that $\frac{\partial}{\partial x} h_\ell(x, z) \leq \frac{\partial}{\partial x} h_2(x, z)$.

Define $f(w, a) \triangleq \frac{1+w}{1-a(1+w)}$. We show that, given any $a \in (0, \infty)$, $f(w, a)$ is a convex function of w within the interval $w \in (-\infty, \frac{1-a}{a})$:

$$\frac{\partial^2}{\partial w^2} f(w, a) = \frac{2a}{(1-a(1+w))^3}.$$

The second derivative of $f(w, a)$ is non-negative for any $w \in (-\infty, \frac{1-a}{a})$ and therefore $f(w, a)$ is a convex function of w .

Consider W_ℓ , the sequence of random variables as defined in Subsection 5.3.3. Specifically, define U_ℓ as a random variable distributed uniformly over the set $\{0, \frac{1}{\ell}, \frac{2}{\ell}, \dots, \frac{\ell-1}{\ell}\}$

and $W_\ell \triangleq \cos(2\pi U_\ell)$. Then,

$$\begin{aligned}\frac{\partial}{\partial \rho^2} h_2(x, z) &= -\frac{z}{2} \mathbb{E} \left[\frac{1 + W_\ell}{2 - \rho^2(1 + W_\ell)} \right] \\ &= -\frac{z}{2} \mathbb{E} \left[f \left(W_\ell, \frac{\rho^2}{2} \right) \right]\end{aligned}$$

$f \left(x, \frac{\rho^2}{2} \right)$ is a convex function for any $x \in \left(-\infty, \frac{1 - \rho^2/2}{\rho^2/2} \right)$. For any $x > 0$ and $z > 0$, $\rho^2 = \rho^2(x, z) \triangleq 1 - \exp\left(-\frac{2x}{z}\right) \in (0, 1)$. Then $\frac{1 - \rho^2/2}{\rho^2/2} \geq 1$ and the interval $\left(-\infty, \frac{1 - \rho^2/2}{\rho^2/2} \right)$ contains $[-1, 1]$. So $f \left(x, \frac{\rho^2}{2} \right)$ is a convex function of x over the convex hull of the support of W_ℓ .

By Lemma 33, W_2 majorizes W_ℓ for any $\ell \in \{2, 3, \dots\}$. Then, by Theorem 5.3.2,

$$\begin{aligned}\mathbb{E} \left[f \left(W_2, \frac{\rho^2}{2} \right) \right] &\geq \mathbb{E} \left[f \left(W_\ell, \frac{\rho^2}{2} \right) \right] \\ \iff \frac{\partial}{\partial \rho^2} h_2(x, z) &\geq \frac{\partial}{\partial \rho^2} h_\ell(x, z)\end{aligned}$$

So $\frac{\partial}{\partial x} h_\ell(x, z) \leq \frac{\partial}{\partial x} h_2(x, z)$. We get $\frac{\partial}{\partial x} h_\ell(x, z) \leq \frac{\partial}{\partial x} h_2(x, z)$ by multiplying both sides by $\frac{\partial \rho^2}{\partial x}$.

Finally, since $h_2(x, z) = \frac{x}{2}$, we get $\frac{\partial}{\partial x} h_2(x, z) = \frac{1}{2}$. □

5.4.4 Critical point

We identify the point where $\lim 1 - x + h_\ell(x, z) = 0$ as $\ell \rightarrow \infty$, which we refer to as the critical point. We then show that $2[1 - x + h_2(x, z)] \geq \ell[1 - x + h_\ell(x, z)]$ at this point.

Lemma 39.

$$\lim_{\ell \rightarrow \infty} h_\ell(x, z) = -z \log \left(\frac{1 + \exp(-x/z)}{2} \right)$$

Proof.

$$\begin{aligned}
\lim_{\ell \rightarrow \infty} h_\ell(x, z) &= -\frac{z}{2} \left[\lim_{\ell \rightarrow \infty} \frac{1}{\ell} \sum_{k=0}^{\ell-1} \log \left(1 - \frac{\rho^2}{2} \left(1 + \cos \left(\frac{2\pi k}{\ell} \right) \right) \right) \right] \\
&\stackrel{(*)}{=} -\frac{z}{2} \int_0^1 \log \left(1 - \frac{\rho^2}{2} - \frac{\rho^2}{2} \cos(2\pi u) \right) du \\
&= -\frac{z}{2} \log \left(\left(1 - \frac{\rho^2}{2} \right) + \sqrt{\left(1 - \frac{\rho^2}{2} \right)^2 - \left(-\frac{\rho^2}{2} \right)^2} \right) + \frac{z}{2} \log 2 \\
&= -\frac{z}{2} \log \left(\left(1 - \frac{\rho^2}{2} \right) + \sqrt{1 - \rho^2} \right) + \frac{z}{2} \log 2 \\
&= -\frac{z}{2} \log \left(\frac{\left(1 + \sqrt{1 - \rho^2} \right)^2}{2} \right) + \frac{z}{2} \log(2) \\
&= -z \log \left(\frac{1 + \sqrt{1 - \rho^2}}{2} \right) \\
&= -z \log \left(\frac{1 + \exp(-x/z)}{2} \right)
\end{aligned}$$

where the inequality (*) follows from Lemma 55. □

Given any $z \in (0, \infty)$, $1 - x + \lim_{\ell \rightarrow \infty} h_\ell(x, z) = 0$ at point $x = z \log(2 \exp(1/z) - 1)$.

Henceforth we refer to this point as the critical point and denote it by $x_z \triangleq z \log(2 \exp(1/z) - 1)$.

Lemma 40. *Given any $\ell \in \{2, 3, \dots\}$ and $z \in (0, \infty)$,*

$$2 \cdot [1 - x_z + h_2(x_z, z)] \geq \ell \cdot [1 - x_z + h_\ell(x_z, z)]$$

where $x_z \triangleq z \log(2 \exp(1/z) - 1) \geq 1$

Proof. Let us denote $\lim_{\ell \rightarrow \infty} h_\ell(x, z)$ by $h_\infty(x, z)$. By the definition of x_z , $h_\infty(x_z, z) =$

$-1 + x_z$.

$$\begin{aligned}
& 2 \cdot [1 - x_z + h_2(x_z, z)] - \ell \cdot [1 - x_z + h_\ell(x_z, z)] \\
&= 2 \cdot [-h_\infty(x_z, z) + h_2(x_z, z)] - \ell \cdot [-h_\infty(x_z, z) + h_\ell(x_z, z)] \\
&= \ell \left[\frac{\ell - 2}{\ell} \cdot h_\infty(x_z, z) + \frac{2}{\ell} \cdot h_2(x_z, z) - h_\ell(x_z, z) \right].
\end{aligned}$$

Consider W_ℓ and Z_ℓ , the sequences of random variables as defined in Subsection 5.3.3.

Specifically, define U_ℓ as a random variable distributed uniformly over the set $\{0, \frac{1}{\ell}, \frac{2}{\ell}, \dots, \frac{\ell-1}{\ell}\}$

and $W_\ell \triangleq \cos(2\pi U_\ell)$. Define Z_ℓ as a mixture between W_2 and W_∞ : Given auxiliary

Bernoulli random variable Y_ℓ with mean $\frac{2}{\ell}$, given any set $\mathcal{S} \subseteq \mathbb{R}$, $\Pr(Z_\ell \in \mathcal{S} | Y_\ell = 1) =$

$\Pr(W_2 \in \mathcal{S})$ and $\Pr(Z_\ell \in \mathcal{S} | Y_\ell = 0) = \Pr(W_\infty \in \mathcal{S})$. Then, we can write

$$\begin{aligned}
& 2 \cdot [1 - x_z + h_2(x_z, z)] - \ell \cdot [1 - x_z + h_\ell(x_z, z)] \\
&= \ell \left[\frac{\ell - 2}{\ell} \cdot h_\infty(x_z, z) + \frac{2}{\ell} \cdot h_2(x_z, z) - h_\ell(x_z, z) \right] \\
&= \ell \cdot \frac{z}{2} \mathbb{E} \left[-\log \left(1 - \frac{\rho^2}{2} (1 + Z_\ell) \right) \right] - \ell \cdot \frac{z}{2} \mathbb{E} \left[-\log \left(1 - \frac{\rho^2}{2} (1 + W_\ell) \right) \right]
\end{aligned}$$

where $\rho^2 = \rho^2(x_z, z) \triangleq 1 - \exp\left(-\frac{2x_z}{z}\right) \in (0, 1)$.

The second derivative of $-\log\left(1 - \frac{\rho^2}{2}(1+w)\right)$ with respect to w is $\frac{\rho^4/4}{\left(1 - \frac{\rho^2}{2}(1+w)\right)^2}$, which is positive for any $w \leq 1$. Then $-\log\left(1 - \frac{\rho^2}{2}(1+w)\right)$ is a convex function of w .

By Lemma 35, Z_ℓ majorizes W_ℓ for any $\ell \in \{2, 3, \dots\}$. Then, by Theorem 5.3.2,

$$\mathbb{E} \left[-\log \left(1 - \frac{\rho^2}{2} (1 + Z_\ell) \right) \right] \geq \mathbb{E} \left[-\log \left(1 - \frac{\rho^2}{2} (1 + W_\ell) \right) \right].$$

It then follows that

$$\begin{aligned}
& 2 \cdot [1 - x_z + h_2(x_z, z)] - \ell \cdot [1 - x_z + h_\ell(x_z, z)] \\
&= \ell \cdot \frac{z}{2} \mathbb{E} \left[-\log \left(1 - \frac{\rho^2}{2} (1 + Z_\ell) \right) \right] - \ell \cdot \frac{z}{2} \mathbb{E} \left[-\log \left(1 - \frac{\rho^2}{2} (1 + W_\ell) \right) \right] \geq 0
\end{aligned}$$

□

5.4.5 Proof of achievability result

Proof. The claimed result can be derived by the following sequence of inequalities. We drop the second argument for h_ℓ and write $h_\ell(x)$ instead of $h_\ell(x, z)$ for brevity.

$$\begin{aligned}
\mathbb{E} \left[d \left(\hat{M}_{\text{ML}}, M \right) \right] &\stackrel{(i)}{\leq} \sum_{\ell=2}^n \sum_{m' \in \mathcal{M}_\ell^c(m)} \ell \cdot \Pr (\langle \mathbf{G}, \mathbf{m} \rangle \leq \langle \mathbf{G}, \mathbf{m}' \rangle \mid M = m) \\
&\stackrel{(ii)}{\leq} \sum_{\ell=2}^n n^{\ell \cdot (1-x+h_\ell(x))} \\
&= n^{2(1-x+h_2(x))} \sum_{\ell=2}^n n^{(\ell-2)(1-x)-(2 \cdot h_2(x)-\ell \cdot h_\ell(x))} \\
&= n^{2(1-x+h_2(x))} \sum_{\ell=2}^n n^{(\ell-2)(1-x)-2(h_2(x)-h_2(x_z))+\ell(h_\ell(x)-h_\ell(x_z))-(2 \cdot h_2(x_z)-\ell \cdot h_\ell(x_z))} \\
&\stackrel{(iii)}{\leq} n^{2(1-x+h_2(x))} \sum_{\ell=2}^n n^{(\ell-2)(1-x)-2(h_2(x)-h_2(x_z))+\ell(h_\ell(x)-h_\ell(x_z))} \\
&\stackrel{(iv)}{\leq} n^{2(1-x+h_2(x))} \sum_{\ell=2}^n n^{(\ell-2)(1-x)-2\left(\frac{x-x_z}{2}\right)+\ell\left(\frac{x-x_z}{2}\right)} \\
&= n^{2(1-x+h_2(x))} \sum_{\ell=2}^n \left[n^{1-x+\frac{x-x_z}{2}} \right]^{\ell-2} \\
&= n^{2(1-x+h_2(x))} \sum_{k=0}^n \left[n^{1-\frac{x+x_z}{2}} \right]^k \\
&\leq \frac{n^{2(1-x+h_2(x))}}{1-n^{1-\frac{x+x_z}{2}}} = \frac{n^{2-x}}{1-n^{1-\frac{x+x_z}{2}}}
\end{aligned}$$

The argument behind each inequality is given below:

1. If an elementary misalignment (as defined in Definition 3) of size ℓ is part of the difference between the estimated mapping and the true mapping, it contributes ℓ to the total number of errors. Suppose m' induces an elementary misalignment with m . That elementary misalignment cannot be part of the final misalignment unless $\langle \mathbf{G}, \mathbf{m} \rangle \leq \langle \mathbf{G}, \mathbf{m}' \rangle$.

When $|\mathcal{U}| = |\mathcal{V}| = |M|$, by Lemma 2, there are no elementary misalignment of type-II or type III, and elementary misalignments of type-I may only have size $\ell \in \{2, 3, \dots, n\}$.

2. By Lemma 37, $\sum_{m' \in \mathcal{M}_\ell^c(m)} \ell \cdot \Pr(\langle \mathbf{G}, \mathbf{m} \rangle \leq \langle \mathbf{G}, \mathbf{m}' \rangle | M = m) \leq n^{\ell \cdot (1-x+h_\ell(x,z))}$.
3. By Lemma 40, $2 \cdot h_2(x_z) - \ell \cdot h_\ell(x_z) \geq 0$.
4. By Lemma 38, the slope of h_ℓ is less than that of h_2 , which is equal to $1/2$. Then $h_\ell(x) - h_\ell(x_z) \leq h_2(x) - h_2(x_z) = \frac{1}{2}(x - x_z)$ for any $x \geq x_z$.

□

CHAPTER 6
ANALYSIS OF A CANONICAL LABELING ALGORITHM FOR THE
ALIGNMENT OF CORRELATED ERDŐS-RÉNYI GRAPHS

In this chapter we study the performance of a polynomial time canonical labeling algorithm for graph alignment under the correlated Erdős-Rényi graph model.

The algorithm proceeds in two phases: during the first phase, for a fixed threshold parameter h , the h highest-degree vertices in both graphs are matched in the natural way (highest degree to highest, second-highest to second-highest, and so forth). For convenience, we call these ‘anchors’. In the second phase, each remaining vertex is labeled with a binary vector of length h that encodes its adjacency to the set of anchor vertices. The final alignment is then generated via a minimum-distance matching over the labels in both graphs. Note that the second phase is equivalent to the matching of two bipartite graphs given the matching of one of their partite sets.

We evaluate the performance of the algorithm on the correlated random graph model of asymptotic size and determine conditions for the reliable performance of the algorithm. This result relies on an achievability result on the alignment of bipartite graphs as an intermediary step, which is of independent interest.

In Section 6.1, we present the problem setting, introduce the algorithm we study and the graph model we work with as well as present a sketch of the analysis in this chapter. In Section 6.2, we present our main result which provides a condition on the correlation and density parameters of the correlated graph model that guarantees perfect recovery of the underlying alignment with high probability. In Sections 6.3 and 6.4 we study the two stages of the canonical labeling algorithm and combine the analyses in Section 6.5 to provide the proof of our main result.

The analysis and results in this chapter have been presented at Sigmetrics 2019 [21].

6.1 Model

6.1.1 Notation

For a graph G we denote its vertex set and edge set as $V(G)$ and $E(G)$, respectively. Alternatively we write $G = (V; E)$ where $V = V(G)$ and $E = E(G)$. For a bipartite graph H we denote $H = (A, B; E)$ where A and B are the partite sets and $E = E(H)$. For any vertex $v \in V(G)$ let $N_G(v)$ be the set of its neighbors in G , $d_G(v)$ its degree and $d_{\overline{G}}(v)$ its complementary degree. The maximum degree in graph G is denoted by $\Delta(G)$. When referring to graphs distinguished by their subscript (e.g. G_a, G_b), we use a shorthand notation to denote neighborhoods, degrees etc. as follows: $N_a(v) = N_{G_a}(v)$, $d_a(v) = d_{G_a}(v)$, $d_{\overline{a}}(v) = d_{\overline{G_a}}(v)$. For a set X , let X^k be set of vectors of length k with entries from X . We will use $[k]$ as the index set for these vectors. We denote vectors in lower case bold font, e.g. $\vec{v} = (v_1, v_2, \dots, v_k) \in V^k$.

For any $n \in \mathbb{N}$, $[n]$ denotes the set of all integers from 1 to n . We denote by $\text{Bin}(n; p)$ the binomial distribution with n trials and event probability p .

6.1.2 Problem Definition

Let $G_a = (V_a; E_a)$ and $G_b = (V_b; E_b)$ be graphs and let $M : V_a \rightarrow V_b$ be a bijection between their vertex sets. We say that these graphs are correlated if the edge set of one provides information about the edge set of the other. We are interested in the case of simple positive correlation: conditioning on the event $\{u, v\} \in E_a$ makes the event $\{M(u), M(v)\} \in E_b$ more likely. The details of our random graph model are given in Section 6.1.4.

Graph Alignment Problem: For a pair of correlated random $G_a = (V_a; E_a)$ and $G_b = (V_b; E_b)$, recover $M : V_a \rightarrow V_b$, the bijection between the vertex pairs in the two graphs based on the correlation of the edge sets.

6.1.3 Alignment by Canonical Labeling

The classical graph isomorphism recovery problem, that is, finding the bijection between vertex sets of a pair of identical graphs, is often solved by canonical labeling based approaches. For a graph $G = (V; E)$ this approach returns a function ℓ_G from a set of vertices V to a set of labels called the canonical labeling of vertices, with the property that, any for any permutation σ of the vertex set and the graph H induced by this permutation on G , $\ell_G(v) = \ell_H(\sigma(v))$ for all vertices $v \in V$. In other words, the canonical labeling only depends on the structure of the graph and is invariant to permutations of the vertex set. This allows us to identify an underlying bijection. If ℓ_G is injective, then the labeling allows for recovery of the automorphism.

If the canonical labeling scheme is robust in the sense that small differences in the structure of the graph induce small perturbations on the labels of vertices, then the canonical labeling can still be used to align a pair of graphs that are “adequately” correlated. In this setting, we seek to find a matching between the label sets of the two graphs that minimizes an appropriately defined labeling distance.

Labeling is done in two steps: In the first step vertices are labeled with their degrees and the small subset of the vertices with high-degrees are identified. These are referred to as ‘anchors’ and form a basis for the alignment of the rest of the graph. In the second step, the remaining vertices are labeled with signature vectors based on their adjacencies with the anchors identified in step one.

This second step ignores all edges between unidentified vertices, effectively treating the graph as a bipartite graph. Therefore, the second step may be considered separately as an algorithm to align two bipartite graphs with one unidentified partite set. In the remainder of this paper, we refer to the first step as the *anchor alignment algorithm* and the second step as the *bipartite alignment algorithm*.

ALGORITHM 1 AnchorSignAlign

Input: $G_a = (V_a; E_a)$, $G_b = (V_b; E_b)$, h **Output:** Estimated alignment $\widehat{M} : V_b \rightarrow V_a$

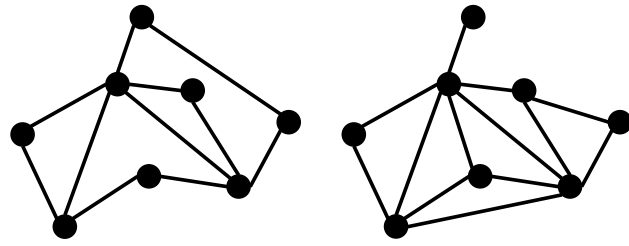
1: **Step 1: Anchor alignment**2: $\vec{w}_a = f_h(G_a)$ 3: $\vec{w}_b = f_h(G_b)$ 4: **for** $i \in [h]$ **do**5: $\widehat{M}(w_{b,i}) = w_{a,i}$ 6: **end for**7: **Step 2: Bipartite alignment**8: $H_a = \{w_{a,i} : i \in [h]\}$ 9: $H_b = \{w_{b,i} : i \in [h]\}$ 10: **for** vertex $v \in V_b \setminus H_b$ **do**11: $\widehat{M}(v) = \arg \min_{u \in V_a \setminus H_a} \|\text{sig}_a(u) - \text{sig}_b(v)\|$ 12: **end for**

The alignment algorithm uses the same canonical labeling scheme originally presented for the graph isomorphism problem by Babai, Erdős, and Selkow [43] and subsequently used for graph alignment in the adversarial setting [44]. (Note that the graph isomorphism algorithm runs in $\mathcal{O}(n^2)$ -time when graphs because the signature matching step can be accomplished by sorting the signatures. The variation for noisy signatures requires $\mathcal{O}(n^2h)$ -time.)

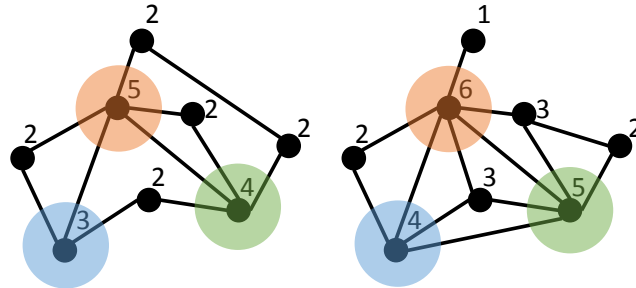
Definition 8. For a n -vertex graph G , let $\vec{\delta}_G = (\delta_{G,1}, \dots, \delta_{G,n})$ be the degree sequence of G in decreasing order.

Definition 9. The high-degree sorting function f_h takes as input a graph G on the vertex set V and lists the h highest-degree vertices, sorted by degree. More precisely, $f_h(G)$ is some vector $\vec{w} = (w_1, w_2, \dots, w_h) \in V^h$ of distinct vertices such that $d_G(w_i) = \delta_{G,i}$.

The degree sequence of G is always uniquely defined. $f_h(G)$ is uniquely defined only



(I) Anchor alignment with $h = 3$



(II) Bipartite alignment

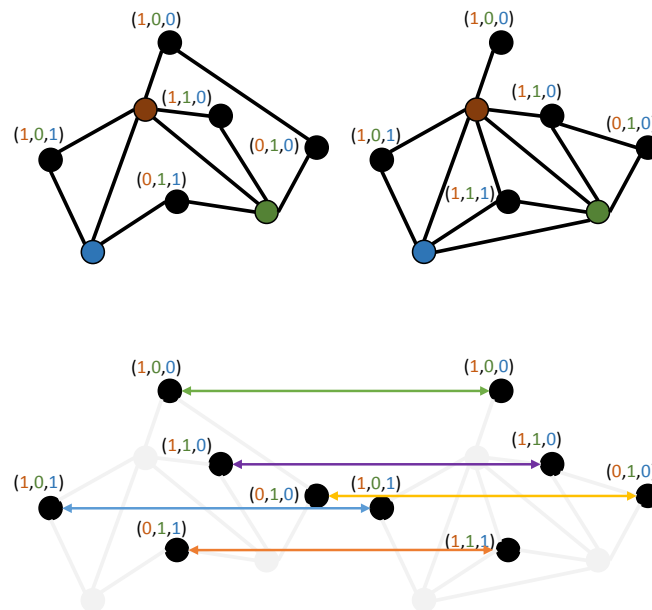


Figure 6.1: Illustration of AnchorSignAlign with $h = 3$. I) Label nodes by degrees and match the h nodes with the highest degrees. II) Label nodes by signatures that encode adjacencies to highest-degree vertices and match each node to the closest signature in other graph.

if the first h entries of δ_G are strictly decreasing. If multiple high-degree vertices have the same degree, $f_h(G)$ lists them in some arbitrary order.

Anchor alignment on graphs G_a and G_b corresponds to the index-by-index alignment of vertices of $f_h(G_a)$ and $f_h(G_b)$. We refer to the set of h vertices that appear in $f_h(G_a)$ as H_a , the set of h vertices that appear in $f_h(G_b)$ as H_b , and when they are the same we say $H_a = H_b = H$. The bipartite alignment algorithm labels each vertex in $V_a \setminus H_a$ by a binary vector encoding its adjacency with vertices in H_a . These labels, which we refer to as signatures, are defined as follows:

Definition 10. *Given graph G and anchor vector $f_h(G) = \vec{w} = (w_1, w_2, \dots, w_h)$, the signature function sig_G takes as input vertex $u \in V(G)$ and returns the signature label of the vertex such that,*

$$\text{sig}_G(u) \in \{0, 1\}^h \quad \text{and} \quad \text{sig}_G(u)_i = \mathbb{1}\{u, w_i\} \in E(G)$$

where $\mathbb{1}\cdot$ denotes the indicator function of an event. We use the shorthand notation $\text{sig}_a(u) = \text{sig}_{G_a}(u)$, $\text{sig}_b(u) = \text{sig}_{G_b}(u)$ when referring to graphs G_a and G_b .

The bipartite alignment algorithm aligns vertices in $V \setminus H$ such as to minimize the Hamming distance between pairs of signatures of aligned vertices. In our analysis we consider a naive approach, aligning each vertex in one graph to the vertex with the closest signature in the other graph. Notice that any graph alignment approach limited to signatures ignores all information pertaining to edges among the unidentified set of vertices.

The steps of the alignment algorithm are summarized in Algorithm 1. We refer to the estimated alignment as \widehat{M} . We say the algorithm is successful when $\widehat{M} = M$.

6.1.4 Correlated Erdős-Rényi Graphs

We perform our analysis on correlated Erdős-Rényi (ER) graphs [45]. Under the basic ER model of random graphs, $G \sim ER(n; p)$ is a random graph on n vertices where any two

vertices share an edge with probability p independent from the rest of the graph. Under the correlated graph model, $(G_a, G_b) \sim ER(n; (p_{11}, p_{10}, p_{01}, p_{00}))$ are a pair of graphs on the same set of n vertices where the occurrences of an edge $e = \{u, v\}$ between any pair of vertices u, v is independent and identically distributed with the following probabilities:

$$(\mathbb{1}e \in E(G_a), \mathbb{1}e \in E(G_b)) = \begin{cases} (1, 1) & \text{w.p. } p_{11} \\ (1, 0) & \text{w.p. } p_{10} \\ (0, 1) & \text{w.p. } p_{01} \\ (0, 0) & \text{w.p. } p_{00}. \end{cases} \quad (6.1)$$

The marginal probabilities are then defined as:

$$\begin{aligned} p_{1*} &= p_{11} + p_{10} & p_{*1} &= p_{11} + p_{01} \\ p_{0*} &= p_{01} + p_{00} & p_{*0} &= p_{10} + p_{00} \end{aligned}$$

We denote the vector of probabilities as $\vec{p} = (p_{11}, p_{10}, p_{01}, p_{00})$. Note that all probabilities are functions of n . We limit our interest to sparse graphs and only consider \vec{p} such that $\lim_{n \rightarrow \infty} p_{00} = 1$.

Two other variations of the correlated Erdős-Rényi model have appeared in the literature.

Subsampling model: This generates a pair of correlated graphs via subsampling of a parent graph $G_{\text{parent}} \sim ER(n; r)$. Each edge in G_{parent} is then included in G_a with probability s_a and in G_b with probability s_b . Each of these $2|E(G_{\text{parent}})|$ edge subsampling events

are independent. This results in $(G_a, G_b) \sim ER(n; (p_{11}, p_{10}, p_{01}, p_{00}))$ with

$$\begin{aligned} p_{11} &= r s_a s_b \\ p_{10} &= r s_a (1 - s_b) \\ p_{01} &= r (1 - s_a) s_b \\ p_{00} &= 1 - r (s_a + s_b - s_a s_b). \end{aligned}$$

This model appeared in Pedarsani and Grossglauser [46] in the symmetric case $s_a = s_b$. Observe that $\frac{p_{11}}{p_{1*} p_{*1}} = \frac{1}{r} \geq 1$, so this model can only represent non-negatively correlated graphs.

Perturbation model: This starts by generating a base graph $G_{\text{parent}} \sim ER(n; r)$. In the adversarial perturbation model considered by Mitzenmacher and Morgan [44], G_a and G_b are each created by making up to $d/2$ changes to the edge set of G_{base} . In the natural randomized version, G_a and G_b differ from G_{base} at each of the $\binom{n}{2}$ vertex pairs independently with probability $\delta = \frac{d}{n(n-1)}$. This results in $(G_a, G_b) \sim ER(n; (p_{11}, p_{10}, p_{01}, p_{00}))$ with

$$\begin{aligned} p_{11} &= r(1 - 2\delta) + \delta^2 \\ p_{10} &= \delta - \delta^2 \\ p_{01} &= \delta - \delta^2 \\ p_{00} &= (1 - r)(1 - 2\delta) + \delta^2. \end{aligned}$$

The models that we have just described generates a pair of graphs on the same vertex set V . To convert these graphs to a pair of correlated graphs on distinct vertex sets, the vertices of G_b can be relabeled using the bijection $M : V \rightarrow V_b$. This relabeling hides the association between the vertex sets and makes the alignment recovery problem nontrivial. For the analysis of Algorithm 1, it is more convenient to work with pairs of graphs on the same vertex sets rather than work with M explicitly, so we will do this for the remainder

of the paper.

In the case of bipartite graphs we use an analogous model. We denote the distribution as $ER(h, k; \vec{p})$ for pairs of correlated graphs with left vertex set of size h and right vertex set of size k . For random bipartite graphs $(B_a, B_b) \sim ER(h, k; \vec{p})$, a left vertex u , and a right vertex v , the pair of random variables $(\mathbb{1}(u, v) \in E(B_a), \mathbb{1}(u, v) \in E(B_b))$ have the same distribution as (6.1).

6.1.5 Outline and Intuition for the Analysis

The two steps of Algorithm 1 dictate opposing bounds on the value of the parameter h . The bipartite alignment phase requires distinct signatures, which is guaranteed only if the length of the signature vectors (h) is large enough. However, the performance of the anchor alignment phase degrades as h grows larger. Our analysis consists of determining upper and lower bounds on h and identifying the region for h which satisfies both bounds.

In Section 6.3 we present a sufficient condition to perfectly align the h highest-degree vertices in correlated ER graphs. This gives an upper bound on h . The result is derived by determining the conditions that guarantee, with high probability, that the h highest-degree vertices have large enough degree separation. It is then unlikely that any two high-degree vertices have their degree order reversed. Applying the Chernoff bound, we show that a degree separation of $\sigma\sqrt{\log h}$ is sufficient, where $\sigma^2 \approx n(p_{10} + p_{01})$ is the variance of a vertex degree in G_b given its degree in G_a . Trivially, independent of the variance, the degree separation must also be at least 1. Thus we get

$$\text{minimum degree separation} \geq \max \left\{ 1, \sigma\sqrt{\log h} \right\}. \quad (6.2)$$

The argument that gives this minimum degree separation bound is illustrated in Fig. 6.2.

Combining (6.2) with a known result on the degree separation of high-degree vertices

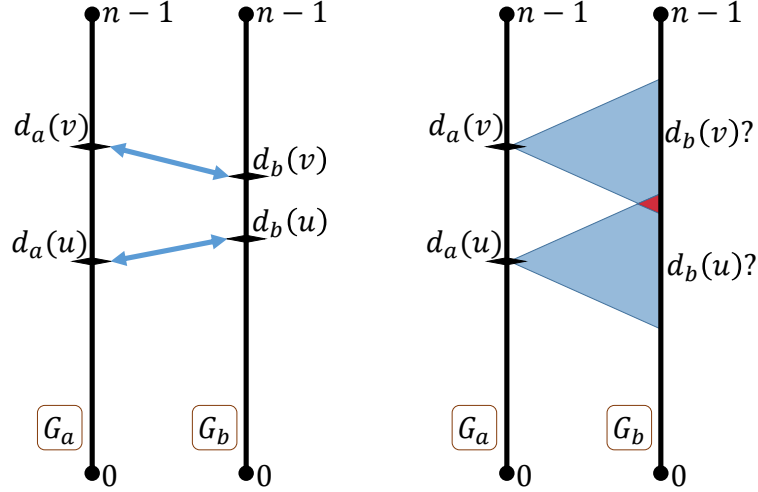


Figure 6.2: Illustration of the degree separation argument for the success of anchor alignment. (Left) Degrees of nodes u and v vary slightly between G_a and G_b . (Right) Without looking at G_b , we can identify the range of values with high likelihood for $d_b(u)$ and $d_b(v)$ based on the values for $d_a(v)$ and $d_a(u)$. When $d_a(v)$ and $d_a(u)$ are too close, these ranges overlap, introducing a significant risk that the order of degrees flips between the two graphs.

gives Theorem 6.3.2, which states a sufficient condition on h for high-degree matching.

Ignoring logarithmic terms, this condition can be simply written as

$$\frac{\sqrt{np_{11}}}{\max\{1, \sqrt{n(p_{10} + p_{01})}\}} \geq \omega(h^2).$$

The intuition behind this result is as follows: given that all vertex degrees are distributed within an interval of size roughly $\sqrt{np_{11}}$, the range of degrees can be partitioned into $\sqrt{np_{11}} / \max\{1, \sqrt{n(p_{10} + p_{01})}\}$ bins of size equal to the minimum degree separation. Two vertices in the same bin violate the degree separation requirement. If the degrees of the h high-degree vertices were to be distributed uniformly within this range, then by the birthday paradox, we would need the number of bins to be significantly larger than h^2 . Clearly high-degree vertices are not uniformly distributed. Nevertheless a rigorous analysis shows that this rough estimate is accurate in the leading term and differs from the actual necessary condition only in the logarithmic terms. This packing argument is illustrated in Fig. 6.3.

In order to understand the constraints on the bipartite matching phase, in Section 6.4 we

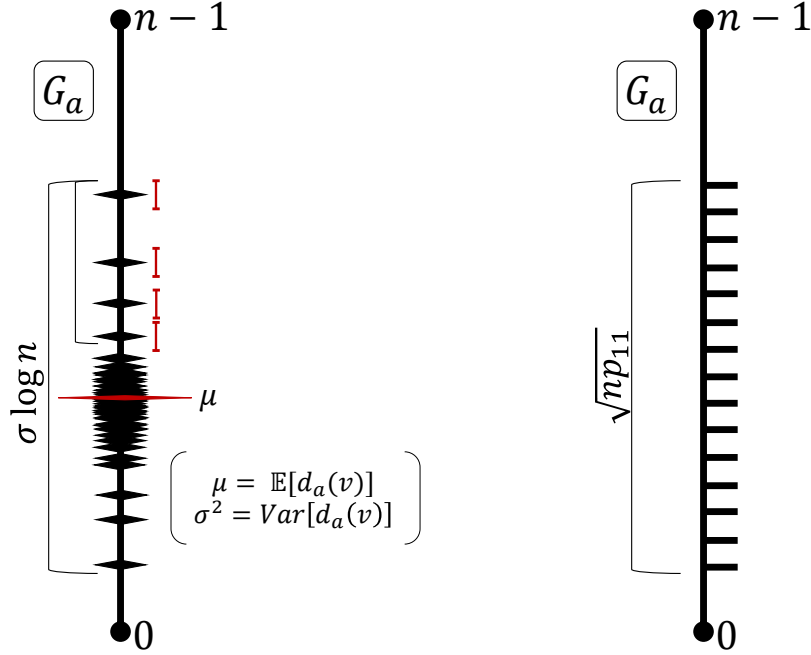


Figure 6.3: Illustration for the packing argument that gives intuition on the upper bound on h . (Left) Degrees are distributed around a region of width on the order of $\sigma \log n$, which is, ignoring logarithmic terms, equal to $\sqrt{np_{11}}$. (Right) We need to fit $\omega(h^2)$ bins within this region. Otherwise, we cannot hope that the randomly distributed h high degree vertices fall into different bins, and therefore they cannot satisfy the minimum degree separation requirement.

analyze the closely related problem of correlated random bipartite graphs. We try to match one of the partite sets based on the complete knowledge of the matching of the other partite set. The identified set is of size h . As in Algorithm 1, this matching is done through sparse binary signatures. The signatures of the copies of any vertex in the two graphs have around hp_{11} common ones. Thus $h \geq \Omega(1/p_{11})$ is a necessary condition for matching. Applying the Chernoff inequality and the union bound over all $\binom{n-h}{2} \approx n^2/2$ possible mismatches, we derive the result in Remark 2 which gives the sufficient condition as

$$h \geq \frac{2 \log n + \omega(1)}{p_{11}}.$$

This problem closely relates to the bipartite alignment phase of Algorithm 1; in both cases we assume to have complete knowledge of the alignment of one partite set (i.e. the set

of anchor vertices) and try to match the other side by only considering edges that connect these two sets. In the case of the correlated bipartite distribution, the analysis is straightforward since edge random variables are independent. But in the general case the edge random variables between the high-degree set and the remaining vertices are not independent. Fortunately the dependence is weak and it is possible to handle this issue by requiring the anchor set to be robust to the addition or removal of a pair of vertices. (This simply requires an additional degree separation of 2 between anchor vertices.)

6.2 Achievability result

Our main result is a condition under which Algorithm 1 successfully recovers the true graph alignment.

Theorem 6.2.1. *Let $G_a = (V; E_a)$ and $G_b = (V; E_b)$ such that $(G_a, G_b) \sim ER(n, \vec{p})$ where \vec{p} is a function of n with $p_{11} \leq o(1)$,*

$$p_{11} \geq \omega\left(\frac{\log^{7/5} n}{n^{1/5}}\right) \text{ and } p_{01} + p_{10} \leq o\left(\frac{p_{11}^5}{\log^6 n}\right),$$

Then Algorithm 1 with parameter h such that

$$\frac{\log n + \omega(1)}{p_{11}} \leq h \leq \mathcal{O}\left(\frac{\log n}{p_{11}}\right)$$

exactly recovers the alignment between the vertex sets of G_a and G_b with probability $1 - o(1)$.

Fig. 6.4 illustrates the asymptotic achievability region of Algorithm 1 as a function of graph density $\left(\frac{\log p_{11}}{\log n}\right)$ and noise $\left(\frac{\log p_{10}}{\log n}\right)$. We also include the achievability of the noiseless scenario [43], more challenging adversarial scenario [44], as well as the information theoretic achievability region [45]. We only consider the symmetric case, where $p_{10} = \Theta(p_{01})$. The x -axis shows $\frac{\log p_{01}}{\log n}$ and the y -axis shows $\frac{\log p_{11}}{\log n}$. Note that in the region

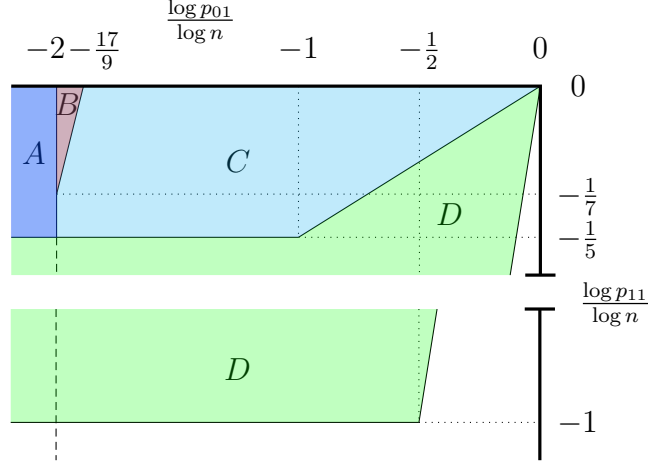


Figure 6.4: Comparison of regions of achievability for symmetric noise, i.e. $p_{10} = \Theta(p_{01})$: (A) region achievable by Algorithm 1 under no expected noise [47], (AUB) under the adversarial model [44], (AUBUC) under the random graph model, (AUBUCUD) theoretical achievability region for the random graph model [45].

$x < -2$, the number of edge differences between the pairs of graphs is zero under the adversarial model and is zero with high probability under the random graph model, so the alignment problem reduces to the graph isomorphism problem.

The adversarial model is defined as follows: Consider a random graph $G_a = ER(n; p)$ and its modified copy G_b obtained by the addition or deletion of at most d edges by an adversary where $d \geq 2$ is a deterministic function of n .

Note that the parameters in this problem formulation relate to the correlated random graph problem through:

$$p = p_{1*} = p_{11} + p_{10} = p_{11} (1 + o(1)) \quad \text{and} \quad d = (p_{10} + p_{01}) \binom{n}{2}. \quad (6.3)$$

By Theorem 5.3 in [44], there exists an appropriate choice of parameter h for which Algorithm 1 perfectly aligns the vertex sets of the two graphs with probability at least $1 - o(1)$ as long as $p = \omega \left(d \log n \left(\frac{d^2}{n} \right)^{1/7} \right)$. By the equalities in (6.3), this condition is satisfied when

$$\Omega(n^{-2}) \leq (p_{01} + p_{10})^{\frac{9}{7}} \leq o \left(\frac{p_{11}}{\log n} n^{-17/7} \right).$$

Recall that the x -axis shows $\frac{\log p_{01}}{\log n}$ and the y -axis shows $\frac{\log p_{11}}{\log n}$. Taking the logarithm of both sides and dividing by $\log n$, in the symmetric case, results in the triangular region defined by the inequality

$$-2 \leq \frac{9}{7}x \leq y - \frac{17}{7} - o(1).$$

Note that $d = o(1)$ for $p_{01} + p_{10} \leq o(n^{-2})$, so the adversarial scenario with a fixed number of edge changes reduces to the graph isomorphism problem and under the random graph model the graphs are isomorphic with high likelihood. The condition to guarantee successful alignment for that problem, given in Theorem 3.17 in [47], is $p = \omega(n^{-1/5} \log n)$, which corresponds to the region where

$$y \geq -\frac{1}{5} + o(1) \quad \text{and} \quad x \leq -2.$$

The achievability region is derived similarly. Theorem 2 in [45] gives the following achievability condition as

$$p_{11} \geq 2 \frac{\log n + \omega(1)}{n} \quad \text{and} \quad p_{01}p_{10} = o(p_{11}p_{00}),$$

which for $p_{00} = 1 - o(1)$ and $p_{10} = \Theta(p_{01})$ corresponds to

$$p_{11} = \Omega\left(\frac{\log n}{n}\right) \quad \text{and} \quad p_{01}^2 = o(p_{11}).$$

This gives the the region defined by

$$y \geq -1 + o(1) \quad \text{and} \quad 2x \leq y.$$

In Section 6.3 we analyze the performance of the anchor alignment stage of the algorithm. In Section 6.4 we present the result on the performance of bipartite graph alignment stage of the algorithm. Finally, in Section 6.5 the results from these two analyses are com-

bined to provide a proof on performance of the alignment algorithm.

6.3 Anchor alignment

The expected performance on the alignment of the anchors (i.e. high-degree vertices) is a function of the sparsity of the graph, its size, and the number of anchors to be matched. We first present a result on the required minimum degree separation between a pair of vertices in one graph to guarantee a given degree separation on the other graph with high probability. We remind the reader of our shorthand notation where for any vertex $v \in V$, $d_a(v)$ and $d_{\bar{a}}(v) = |V| - d_a(v) - 1$ denote v 's degree and inverse degree in G_a , respectively. Similarly $d_b(v)$, $d_{\bar{b}}(v)$ denote the degree and inverse degree in G_b .

Lemma 41. *Let $(G_a, G_b) \sim ER(n; \vec{p})$. Given $u, v \in V(G)$ such that $d_a(u) > d_a(v)$, define $\varphi \triangleq d_a(u) \frac{p_{10}}{p_{1*}} + d_{\bar{a}}(v) \frac{p_{01}}{p_{0*}}$ and $\varepsilon = \frac{p_{01}}{p_{0*}} + \frac{p_{10}}{p_{1*}}$. Let $\eta \in (0, \infty)$ be a function of n . If*

$$d_a(u) - d_a(v) \geq (1 - \varepsilon)^{-1} \left(k + 4 \max(\eta, \sqrt{\varphi \cdot \eta}) \right),$$

Then $P[d_b(u) - d_b(v) \leq k] \leq e^{-\eta}$.

Proof. Let us denote the degree separations in the two graphs by $\alpha \triangleq d_a(u) - d_a(v)$ and $\beta \triangleq d_b(u) - d_b(v)$. Observe that the presence of the edge $\{u, v\}$ in G_a does not affect α .

Thus we define

$$\begin{aligned} d_a^u &\triangleq |N_a(u) \setminus \{v\}| & d_{\bar{a}}^u &\triangleq n - 2 - d_a^u \\ d_a^v &\triangleq |N_a(v) \setminus \{u\}| & d_{\bar{a}}^v &\triangleq n - 2 - d_a^v. \end{aligned}$$

The error event in the degree sequence, i.e. $d_b(u) - d_b(v) \leq k$, corresponds to $\beta \leq k$. By the Chernoff bound:

$$P[\beta \leq k | d_a^u, d_a^v] \leq z^{-k} \mathbb{E}[z^\beta | d_a^u, d_a^v] \quad \forall 0 < z \leq 1.$$

By Lemma 42, the expression for the probability generating function $F_\beta(z) \triangleq \mathbb{E} [z^\beta | d_a^u, d_a^v]$ is given by

$$F_\beta(z) = z^\alpha \left(1 + \frac{p_{10}}{p_{1*}}(z-1)\right)^{d_a^v} \left(1 + \frac{p_{01}}{p_{0*}}(z-1)\right)^{d_a^u} \\ \times \left(1 + \frac{p_{10}}{p_{1*}}(z^{-1}-1)\right)^{d_a^u} \left(1 + \frac{p_{01}}{p_{0*}}(z^{-1}-1)\right)^{d_a^v}.$$

By applying $1+x \leq e^x$ we get

$$F_\beta(z) \leq \exp \left\{ \alpha \log z + \left(\frac{p_{10}}{p_{1*}} d_a^v + \frac{p_{01}}{p_{0*}} d_a^u \right) (z-1) \right\} \\ \times \exp \left\{ \left(\frac{p_{10}}{p_{1*}} d_a^u + \frac{p_{01}}{p_{0*}} d_a^v \right) (z^{-1}-1) \right\}$$

Furthermore applying $\log x \leq x-1$ we have

$$z^{-k} F_\beta(z) \leq \exp \left\{ \left(\alpha - k + \frac{p_{10}}{p_{1*}} d_a^v + \frac{p_{01}}{p_{0*}} d_a^u \right) (z-1) \right\} \\ \times \exp \left\{ \left(\frac{p_{10}}{p_{1*}} d_a^u + \frac{p_{01}}{p_{0*}} d_a^v \right) (z^{-1}-1) \right\} \quad (6.4)$$

Denote the coefficients by

$$r' \triangleq \alpha - k + \frac{p_{10}}{p_{1*}} d_a^v + \frac{p_{01}}{p_{0*}} d_a^u \quad \text{and} \quad r \triangleq \frac{p_{10}}{p_{1*}} d_a^u + \frac{p_{01}}{p_{0*}} d_a^v.$$

Denote their difference as

$$\Delta r \triangleq r' - r = \alpha - k + \frac{p_{10}}{p_{1*}} d_a^v - \frac{p_{10}}{p_{1*}} \left((n-2) - d_a^v \right) \\ - \frac{p_{10}}{p_{1*}} d_a^u + \frac{p_{01}}{p_{0*}} \left((n-2) - d_a^u \right) \\ = \alpha - k - \left(\frac{p_{01}}{p_{0*}} + \frac{p_{10}}{p_{1*}} \right) (d_a^u - d_a^v) \\ = \alpha \left(1 - \frac{p_{01}}{p_{0*}} - \frac{p_{10}}{p_{1*}} \right) - k$$

The right hand side of the inequality in (6.4) is minimized at $z^* \triangleq \sqrt{r/r'}$. Taking the logarithm of both sides in (6.4) and evaluating it at $z = z^*$ we get

$$\log F_\beta(z^*) - k \log z^* \leq - \left(\sqrt{r'} - \sqrt{r} \right)^2 = -\Delta r \left(\sqrt{1 + r/\Delta r} - \sqrt{r/\Delta r} \right)^2.$$

The inequality $\sqrt{1 + x^2} - x \geq \frac{1}{1+2x}$ holds for any $x \geq 0$. Specifically the choice of $x = \sqrt{r/\Delta r}$ results in

$$- [\log F_\beta(z^*) - k \log z^*] \geq \frac{\Delta r}{(1 + 2\sqrt{r/\Delta r})}$$

Note that:

$$\begin{aligned} \Delta r \geq 4 \max \{ \eta, \sqrt{r\eta} \} &\implies \left(1 + 2\sqrt{\frac{r}{\Delta r}} \right)^2 \leq 4 \min \left\{ 1, \frac{r}{\eta}, \sqrt{\frac{r}{\eta}} \right\} \\ &\implies \frac{\Delta r}{(1 + 2\sqrt{\frac{r}{\Delta r}})^2} \geq \frac{1}{4} \max \left\{ \Delta r', \Delta r \frac{\eta}{r}, \Delta r \sqrt{\frac{\eta}{r}} \right\} \\ &\geq \frac{\Delta r}{4} \geq \eta \end{aligned}$$

which implies $(z^*)^{-k} F_\beta(z^*) \leq e^{-\eta}$. Finally observe that $\varphi \triangleq d_a(u) \frac{p_{10}}{p_{1*}} + d_a(v) \frac{p_{01}}{p_{0*}}$ is at least r . Therefore the condition in the statement of the lemma implies $\Delta r \geq 4 \max \{ \eta, \sqrt{r\eta} \}$. \square

Lemma 42. $\beta \triangleq d_b(u) - d_b(v)$, $d_a^u \triangleq |N_a(u) \setminus \{v\}|$ and $d_a^v \triangleq |N_a(v) \setminus \{u\}|$, the expression for the conditional generating function probability generating function $F_\beta(z) \triangleq \mathbb{E} [z^\beta | d_a^u, d_a^v]$ is given by

$$\begin{aligned} F_\beta(z) &= z^\alpha \left(1 + \frac{p_{10}}{p_{1*}}(z - 1) \right)^{d_a^v} \left(1 + \frac{p_{01}}{p_{0*}}(z - 1) \right)^{d_a^u} \\ &\quad \times \left(1 + \frac{p_{10}}{p_{1*}}(z^{-1} - 1) \right)^{d_a^u} \left(1 + \frac{p_{01}}{p_{0*}}(z^{-1} - 1) \right)^{d_a^v} \end{aligned}$$

Proof. Given the random graphs $(G_a, G_b) \sim ER(n; \vec{p})$ and for a given pair of vertices

u, v , define $d_a^u = |N_a(u) \setminus \{v\}|$, $d_a^v = |N_a(v) \setminus \{u\}|$ and d_b^u, d_b^v analogous for G_b . We seek to find $F_\beta(z) = \mathbb{E} [z^\beta | d_a^u, d_a^v]$ where $\beta = d_b(u) - d_b(v)$.

Let us denote the degree separation in G_a as $\alpha = d_a(u) - d_a(v)$. Note that $d_a^u - d_a^v = d_a(u) - d_a(v) = \alpha$ and $d_b^u - d_b^v = d_b(u) - d_b(v) = \beta$. Let us denote the number edges of x in $G_a \setminus \{u\}$ that are non-edges in G_b (i.e. number of edges *exclusive* to G_a) as $e_a^u = \|N_a(u) \setminus N_b(u) \setminus \{v\}\|$ and vice versa as $e_b^u = \|N_b(u) \setminus N_a(u) \setminus \{v\}\|$. It can be shown that $d_b^u = d_a^u - e_a^u + e_b^u$. Similarly define e_a^v and e_b^v by ignoring the edge $\{u, v\}$. We then have

$$d_b^u - d_b^v = d_a^u - d_a^v - e_a^u + e_a^v + e_b^u - e_b^v$$

or simply $\beta = \alpha - e_a^u + e_a^v + e_b^u - e_b^v$. Notice that given d_a^u and d_a^v , α is deterministic. Also notice that the remaining terms $e_a^u, e_a^v, e_b^u, e_b^v$ are mutually independent binomially distributed random variables with distribution:

$$\begin{aligned} e_a^u &\sim B\left(d_a^u, \frac{p_{10}}{p_{1*}}\right), & e_b^u &\sim B\left(d_a^u, \frac{p_{01}}{p_{0*}}\right), \\ e_a^v &\sim B\left(d_a^v, \frac{p_{10}}{p_{1*}}\right), & e_b^v &\sim B\left(d_a^v, \frac{p_{01}}{p_{0*}}\right) \end{aligned}$$

where $d_a^u = n - 2 - d_a^v$ and $d_a^v = n - 2 - d_a^u$. The probability generating function of a binomially distributed random variable $X \sim \text{Bin}(n, p)$ is given by $[1 + p(z - 1)]^n$. Thus we get the probability generating function of β as

$$\begin{aligned} F_\beta(z) &= z^\alpha \left(1 + \frac{p_{10}}{p_{1*}}(z^{-1} - 1)\right)^{d_a^u} \left(1 + \frac{p_{10}}{p_{1*}}(z - 1)\right)^{d_a^v} \\ &\quad \times \left(1 + \frac{p_{01}}{p_{0*}}(z - 1)\right)^{d_a^u} \left(1 + \frac{p_{01}}{p_{0*}}(z^{-1} - 1)\right)^{d_a^v} \end{aligned}$$

□

Lemma 41 involves two lower bounds on the gap between degrees: one depending on

η and the other on $\sqrt{\varphi \cdot \eta}$. The quantity φ is the expected number of edges ‘lost’ by u and ‘gained’ by v when moving from G_a to G_b . A larger φ implies higher likelihood for the degree gap to be ‘bridged’ moving from G_a to G_b . At the dense high-noise performance limit, the $\sqrt{\varphi \cdot \eta}$ lower bound is dominant. The η lower bound arises from the discreteness of the degrees. This bound is dominant at the sparse low-noise limit.

Lemma 41 only concerns pairs of vertices. Next we present a condition on the graph sequence of G_a that guarantees with high probability the desired degree separation among high-degree vertices in G_b . Recall that, by Definition 8, δ_a and δ_b denote the degree sequences in G_a and G_b respectively.

Corollary 43. *Let $(G_a, G_b) \sim ER(n; \vec{p})$ where $G_a = (V; E_a)$ and $G_b = (V; E_b)$. Define $\varphi \triangleq \Delta(G_a) \frac{p_{10}}{p_{1*}} + n \frac{p_{01}}{p_{0*}}$ and $\varepsilon \triangleq \frac{p_{01}}{p_{0*}} + \frac{p_{10}}{p_{1*}}$. Let $h \in [n]$ and η be functions of n . Let s be an integer such that $s \geq h + \frac{1}{\eta} \log\left(\frac{n}{h}\right) + 1$. If*

$$\forall i \in [s], \quad \delta_{a,i} - \delta_{a,i+1} \geq (1 - \varepsilon)^{-1} \left(k + 4 \max\{\eta, \sqrt{\varphi \cdot \eta}\} \right) \quad (6.5)$$

then, with probability at least $\frac{1 - (2h+1)e^{-\eta}}{1 - e^{-\eta}}$, $f_h(G_a) = f_h(G_b)$ and $\delta_{b,i} - \delta_{b,i+1} > k$ for any $i \in [h]$.

Proof. Let H_a and S_a be the set of h and s highest-degree vertices in G_a respectively and define H_b analogously for G_b . The following two events collectively imply $f_h(G_a) = f_h(G_b)$ and $\delta_{b,i} - \delta_{b,i+1} > k$ for any $i \in [h]$.

- Let $\mathcal{E}^{\text{high}}$ be the event that vertices in H_a have the same degree ordering in G_a and in G_b as well as a minimum degree separation larger than k in G_b . Note that this does not guarantee $H_a = H_b$.
- Let \mathcal{E}^{low} be the event that all vertices in $V \setminus H_a$ have degree less than $\delta_{b,h} - k$ in G_b , i.e. no vertex from $V \setminus H_a$ is in H_b and all have a sufficiently large degree separation with the h -th highest-degree vertex.

First we consider $\mathcal{E}^{\text{high}}$, i.e. the event where $\delta_{b,i} - \delta_{b,j} > k$ for any $i < j$ with $i, j \in [h]$. Notice that it is sufficient to check this condition for consecutive pairs of vertices in the degree sequence. Given the condition in (6.5), Lemma 41 states that for any pair of vertices $v_i, v_{i+1} \in H_a$, v_i and v_{i+1} in G_b have the same degree ordering as well as a degree separation larger than k with probability at least $e^{-\eta}$. Thus, by the union bound, we get $P[\overline{\mathcal{E}^{\text{high}}}] \leq 1 - he^{-\eta}$.

Second we consider \mathcal{E}^{low} , i.e. the event where $\delta_{b,h} - \delta_{b,i} > k$ for any $i \in [n] \setminus [h]$. By the condition in (6.5) we have, $\forall i \in [s] \setminus [h]$,

$$\begin{aligned} \delta_{a,h} - \delta_{a,i} &\geq (i - h)(k + 4 \max\{\eta, \sqrt{\varphi \cdot \eta}\})(1 - \varepsilon)^{-1} \\ &\geq \left(k + 4 \max\left\{(i - h)\eta, \sqrt{(i - h)\varphi\eta}\right\}\right) (1 - \varepsilon)^{-1} \end{aligned}$$

and $\forall i \in [n] \setminus [s]$,

$$\begin{aligned} \delta_{a,h} - \delta_{a,i} &\geq (s + 1 - h)(k + 4t \max\{\eta, \sqrt{\varphi \cdot \eta}\})(1 - \varepsilon)^{-1} \\ &\geq \left(k + 4 \max\left\{(s + 1 - h)\eta, \sqrt{(s + 1 - h)\varphi\eta}\right\}\right) (1 - \varepsilon)^{-1}. \end{aligned}$$

By Lemma 41 we then have

$$P[\delta_{a,h} - \delta_{a,i} \leq k] \leq \exp(-\eta \min\{i - h, s + 1 - h\}).$$

Then, by the union bound,

$$\begin{aligned} P[\overline{\mathcal{E}^{\text{low}}}] &\leq \sum_{i=h+1}^s e^{-\eta(i-h)} + \sum_{i=s+1}^n e^{-\eta(s+1-h)} \\ &\leq \frac{e^{-\eta}}{1 - e^{-\eta}} + (n - s) \frac{h}{n} e^{-\eta} \end{aligned}$$

Applying the union bound again we obtain

$$P \left[\overline{\mathcal{E}^{\text{high}}} \vee \overline{\mathcal{E}^{\text{low}}} \right] \leq (2h + 1)e^{-\eta}/(1 - e^{-\eta}).$$

□

The φ term in Corollary 43 corresponds to an upper bound for the same term in Lemma 41 that we obtain by replacing the vertex degree with the max degree in the graph, and the inverse degree with n . We then need the following upper bound on the maximum degree of a random graph.

Lemma 44. *Let $G \sim ER(n; p)$ with $p \geq \omega\left(\frac{\log n}{n}\right)$. For any constant $\epsilon > 0$, we have $P[\Delta(G) \geq pn(1 + \epsilon)] \leq o(1)$.*

Proof of Lemma 44. For any vertex $u \in V(G)$, $d_G(u) \sim \text{Bin}(n - 1; p)$. By the Chernoff bound, for any $D \in \mathbb{N}$ and $z \in [1, \infty]$

$$P[d_G(u) \geq D] \leq z^{-D} \mathbb{E} \left[z^{-d_G(u)} \right] \leq z^{-D} [1 + p(z - 1)]^{n-1}.$$

Applying $1 + x \leq e^x$ to both terms this becomes

$$\log P[d_G(u) \geq D] \leq D(z^{-1} - 1) + p(n - 1)(z - 1).$$

The right hand side is minimized for $z^* = \sqrt{\frac{D}{p(n-1)}}$ which gives us

$$\log P[d_G(u) \geq D] \leq - \left(\sqrt{D} - \sqrt{p(n-1)} \right)^2.$$

Let $D = (1 + \epsilon)p(n - 1)$. By the union bound, the probability that the maximum degree

is at least D is at most

$$\begin{aligned} nP[d_G(u) \geq D] &\leq n \exp\left(-p(n-1)\left(\sqrt{1+\epsilon}-1\right)^2\right) \\ &\leq n \exp(-\omega(\log n)) \leq o(1). \end{aligned}$$

□

Corollary 43 relies on G_a having a degree sequence whose largest terms are sufficiently separated. We now present a condition that guarantees a given degree separation for almost all random graphs.

Theorem 6.3.1. ([47] Theorem 3.15) *Let $h \in \mathbb{N}$ and $c \in \mathbb{R}^+$ functions of n such that $h = o(n)$ and $c = o(1)$. Then, with probability $1 - o(1)$, in $G \sim ER(n, p)$*

$$\delta_i - \delta_{i+1} \geq \frac{c}{h^2} \left(\frac{np(1-p)}{\log n} \right)^{1/2} \quad \text{for each } i \in [h].$$

We are now in a position to present a result on the performance of the high-degree matching step of our algorithm. First we define the three events that are needed to be able to successfully align the high-degree vertices: the set of high-degree vertices must be the same in the two graphs and in each graph the high-degree vertices must have sufficiently separated degrees. Distinct degrees are clearly required, but we require the stronger condition that degrees have difference of at least 3. This allows us to establish the independence of this stage of the algorithm with the bipartite matching stage later in Subsection 6.5.

Definition 11. *Let \mathcal{E}^H be the event that the lists of the h highest-degree vertices in G_a and G_b are the same, i.e. $f_h(G_a) = f_h(G_b)$. This is the “high-degree match” event. Let \mathcal{E}_a^S be the event that $\delta_{a,i} > \delta_{a,i+1} + 2$ for all $i \in [h]$. Define \mathcal{E}_b^S analogously for δ_b . These are the “degree separation” events.*

Theorem 6.3.2. Let $(G_a, G_b) \sim ER(n, \vec{p})$ where \vec{p} is a function of n such that $p_{00} = 1 - o(1)$. Moreover let $h \in [n]$ such that $\omega(\log n) \leq h \leq o(n)$. If

$$\max \{(\log h)^2, n(p_{01} + p_{10}) \log h\} \leq o\left(\frac{np_{11}}{h^4 \log n} \cdot \frac{p_{11}}{p_{1*}}\right), \quad (6.6)$$

then $P[\mathcal{E}^H \wedge \mathcal{E}_a^S \wedge \mathcal{E}_b^S] \geq 1 - o(1)$.

Proof. To apply Corollary 43, η and s must satisfy $s \geq h + \frac{1}{\eta} \log\left(\frac{n}{h}\right)$. We pick η such that $s = \left\lceil h + \frac{1}{\eta} \log\left(\frac{n}{h}\right) \right\rceil$ and $\log h + \omega(1) \leq \eta \leq \mathcal{O}(\log h)$. The condition $h \geq \omega(\log n)$ guarantees that $s \leq h(1 + o(1))$.

Applying Lemma 44, we have

$$\varphi = \Delta(G_a) \frac{p_{10}}{p_{1*}} + n \frac{p_{01}}{p_{0*}} \leq (1 + \epsilon)np_{10} + n \frac{p_{01}}{p_{0*}} \leq (1 + \epsilon + o(1))n(p_{10} + p_{01}). \quad (6.7)$$

Define $c \triangleq \left(\frac{s^4 \log n}{np_{1*}}\right)^{1/2} \left(1 - \frac{p_{01}}{p_{0*}} - \frac{p_{10}}{p_{1*}}\right)^{-1} (2 + 4 \max\{\eta, (r\eta)^{1/2}\})$. By $p_{00} = 1 - o(1)$ we have $\left(1 - \frac{p_{10}}{p_{1*}} - \frac{p_{01}}{p_{0*}}\right)^{-1} = \left(\frac{p_{11}}{p_{1*}} - o(1)\right)^{-1} = \Theta\left(\frac{p_{1*}}{p_{11}}\right)$. Together with the upper bounds on η , h , and s , we get

$$c \leq \mathcal{O}(1) \left(\frac{h^4 \log n}{np_{11}} \cdot \frac{p_{1*}}{p_{11}}\right)^{1/2} \max \{ \log h, (n(p_{01} + p_{10}) \log h)^{1/2} \}.$$

From (6.6), we have $c \leq o(1)$.

By Theorem 6.3.1, with probability $1 - o(1)$, we have a minimum separation of $2 + 4 \max\{\eta, (\varphi \cdot \eta)^{1/2}\}$ among the top s degrees in $G_a \sim ER(n; p_{1*})$. Then Corollary 43 implies that the probability that $f_h(G_a) \neq f_h(G_b)$ is at most $se^{-\eta} \leq (1 + o(1))he^{-\log h - \omega(1)} \leq o(1)$. \square

6.4 Bipartite graph alignment

We will need the following method of specifying an induced bipartite subgraph. Let G be a graph on the vertex set V and let $U \subseteq V$. Let \vec{w} be a vector of h distinct vertices in $V \setminus U$.

Define $G[U, \vec{w}]$ to be the bipartite graph with left vertex set U , right vertex set $[h]$, and edge set

$$E(G[U, \vec{w}]) = \{(u, j) \in U \times [h] : (u, w_j) \in E(G)\}.$$

Recall that in Algorithm 1, we have $\vec{w}_a = f_h(G_a)$ and $\vec{w}_b = f_h(G_b)$. By Definition 10, the signature of any $u \in U$ is the edge indicator function for $G_a[\{u\}, \vec{w}_a]$:

$$\text{sig}_a(u) \in \{0, 1\}^h \quad \text{and} \quad \text{sig}_a(u)_i = \mathbb{1}(u, i) \in E(G_a[\{u\}, \vec{w}_a]).$$

We define an analogous signature scheme for bipartite graphs to be used for the bipartite alignment step.

Definition 12. Given the bipartite graph $B = (V, [h]; E)$, the bipartite signature function sig'_B takes as input vertex $u \in V$ and returns the signature label of the vertex such that

$$\text{sig}'_B(u) \in \{0, 1\}^h \quad \text{and} \quad \text{sig}'_B(u)_i = \mathbb{1}(u, i) \in E$$

When referring to signatures on bipartite graphs that are distinguished only by their subscripts (e.g. B_a and B_b) we only denote the signatures in shorthand notation, e.g. $\text{sig}'_a(u) = \text{sig}'_{B_a}(u)$, $\text{sig}'_b(u) = \text{sig}'_{B_b}(u)$.

We restate the second half of Algorithm 1 as the bipartite graph alignment algorithm in Algorithm 2

ALGORITHM 2 Bipartite Graph alignment

Input: $B_a = (V_a, [h]; E_a)$, $B_b = (V_b, [h]; E_b)$

Output: Estimated alignment $\widehat{M} : V_b \rightarrow V_a$

- 1: **for** vertex $v \in V_b$ **do**
 - 2: $\widehat{M}(v) = \arg \min_{u \in V_a} \|\text{sig}'_a(u) - \text{sig}'_b(v)\|$
 - 3: **end for**
-

Suppose that we have bipartite graphs $B_a = (V_a, [h]; E_a)$ and $B_b = (V_b, [h]; E_b)$ such

that $|V_a| = |V_b|$. Assume there is an exact correspondence between the vertex sets, expressed by the alignment $M : V_b \rightarrow V_a$. Algorithm 2 is guaranteed to map vertex $u \in V_b$ to $M(u) \in V_a$ if

$$\|\text{sig}'_a(M(v)) - \text{sig}'_b(u)\| > \|\text{sig}'_a(M(u)) - \text{sig}'_b(u)\| \quad (6.8)$$

for any $v \in V_b \setminus \{u\}$. Hence verifying the equality above for any ordered pair of vertices $(u, v) \in V_b^2$ guarantees that the algorithm perfectly aligns all vertices.

In the remainder of the section, in order to avoid cumbersome notation, we assume that, without loss of generality, $V_a = V_b = V$ and the true alignment is the trivial alignment $M(v) = v$ for any $v \in V$.

To analyze Algorithm 2 for random bipartite graphs, we need the following lemma which bounds the probability that a pair of vertices are misaligned. This corresponds to the failure of (6.8) for either one of the vertices.

Lemma 45. *Let bipartite graphs $B_a = (\{u, v\}, [h]; E_a)$ and $B_b = (\{u, v\}, [h]; E_b)$ be distributed according to $(B_a, B_b) \sim ER(2, h; \vec{p})$.*

Define $\mathcal{E}^M(B_a, B_b)$ to be the “misalignment event” i.e. the event where either of the following inequalities hold:

$$\begin{aligned} & \|\text{sig}'_a(v) - \text{sig}'_b(u)\| \leq \|\text{sig}'_a(u) - \text{sig}'_b(u)\| \\ \text{or} \quad & \|\text{sig}'_a(u) - \text{sig}'_b(v)\| \leq \|\text{sig}'_a(v) - \text{sig}'_b(v)\|. \end{aligned}$$

Then $P[\mathcal{E}^M(B_a, B_b)] \leq 2 \exp(-h\rho^2)$ where

$$\rho \triangleq \sqrt{p_{00}p_{1*} + p_{11}p_{0*}} - \sqrt{p_{10}p_{0*} + p_{01}p_{1*}}.$$

Proof of Lemma 45. Define the random variable

$$\gamma \triangleq \|\text{sig}'_a(v) - \text{sig}'_b(u)\| - \|\text{sig}'_a(u) - \text{sig}'_b(u)\|.$$

We bound the probability of $\gamma \leq 0$ using the Chernoff bound: $P[\gamma \leq 0] \leq \mathbb{E}[z^\gamma]$ for all $0 < z \leq 1$. By Lemma 46, the generating function $F_\gamma(z) \triangleq \mathbb{E}[z^\gamma]$ is given as

$$F_\gamma(z) = [1 + q_0(z - 1) + q_1(z^{-1} - 1)]^h$$

where $q_0 = p_{00}p_{1*} + p_{11}p_{0*}$ and $q_1 = p_{10}p_{0*} + p_{01}p_{1*}$.

Applying $1 + x \leq e^x$ and evaluating the function at $z^* = \sqrt{\frac{q_1}{q_0}}$, we get $\log F_\gamma(z^*) \leq -h(\sqrt{q_0} - \sqrt{q_1})^2$. Hence for $\rho = \sqrt{q_0} - \sqrt{q_1}$ we have $P[\gamma \leq 0] \leq \exp(-h\rho^2)$.

Notice that for the analogous

$$\gamma' = \|\text{sig}'_a(u) - \text{sig}'_b(v)\| - \|\text{sig}'_a(v) - \text{sig}'_b(v)\|$$

the same bound holds. The event $\mathcal{E}^M(B_a, B_b)$ is equivalent to

$\{\gamma \leq 0 \vee \gamma' \leq 0\}$. Thus by the union bound $P[\mathcal{E}^M(B_a, B_b)] \leq 2 \exp(-h\rho^2)$. \square

Lemma 46. Given $\gamma \triangleq \|\text{sig}'_a(v) - \text{sig}'_b(u)\| - \|\text{sig}'_a(u) - \text{sig}'_b(u)\|$, the expression for the probability generating function $F_\gamma(z) = \mathbb{E}[z^\gamma]$ is given by

$$F_\gamma(z) = [1 + q_0(z - 1) + q_1(z^{-1} - 1)]^h$$

where $q_0 = p_{00}p_{1*} + p_{11}p_{0*}$ and $q_1 = p_{10}p_{0*} + p_{01}p_{1*}$

Proof. Consider the random bipartite graphs $B_a = (V, H; E_a)$, $B_b = (V, H; E_b)$ distributed according to $(B_a, B_b) \sim ER(h, n; \vec{p})$. For a given pair of vertices $u, v \in V$ let us define the relative signature distance of u to v observed from G_b as $\gamma(u, v) = \|\text{sig}'_a(v) - \text{sig}'_b(u)\| - \|\text{sig}'_a(u) - \text{sig}'_b(u)\|$. $\gamma(u, v)$ can be expressed as the sum of the contributions of each

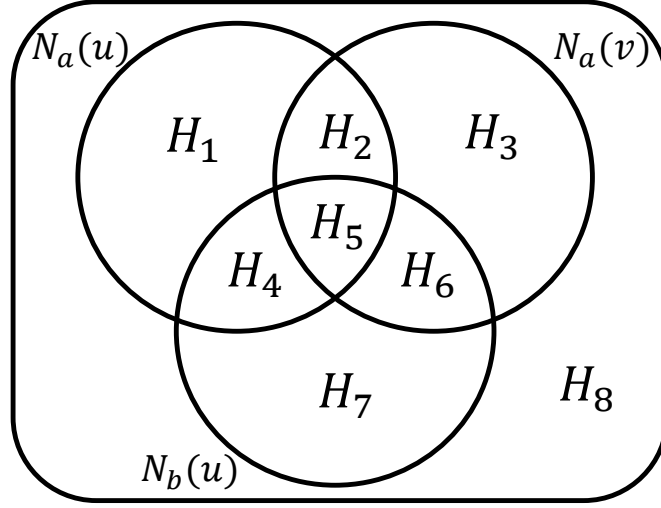


Figure 6.5: Venn diagram representation of $N_a(u)$, $N_a(v)$ and $N_b(u)$

high-degree vertex $w \in H$. The neighborhoods $N_a(v)$, $N_a(u)$ and $N_b(u)$ partition the set of high-degree vertices in 8 disjoint sets as given in Fig. 6.5. We then have $\gamma(u, v) = \sum_{w \in H} \mathbb{1}_{w \in H_3 \cup H_4} - \mathbb{1}_{w \in H_1 \cup H_6}$.

Notice that for any $w \in H$, $P[w \in H_3 \cup H_4] = p_{00}p_{1*} + p_{11}p_{0*}$ and $P[w \in H_1 \cup H_6] = p_{10}p_{0*} + p_{01}p_{1*}$. In fact the random variables $\left\{ \mathbb{1}_{w \in H_3 \cup H_4} - \mathbb{1}_{w \in H_1 \cup H_6} \right\}_{w \in H}$ are mutually independent and identically distributed. Let us define $q_0 = p_{00}p_{1*} + p_{11}p_{0*}$ and $q_1 = p_{10}p_{0*} + p_{01}p_{1*}$. This gives us the following generating function

$$F_\gamma(z) = \mathbb{E} [z^{\gamma(u,v)}] = [1 + q_0(z - 1) + q_1(z^{-1} - 1)]^h. \quad (6.9)$$

□

The quantity ρ is a measure of the correlation between the pair of graphs. The likelihood of misalignment between a pair of vertices can be upper bounded in terms of h , the size of the readily identified set, and ρ , the strength of the correlation between the new graphs. Applying this result over the entire graph gives us the following result.

Remark 2. Let $(B_a, B_b) \sim ER(n, h; \vec{p})$. Then for each $u, v \in [n]$, the subgraphs induced by $\{u, v\}$ and $[h]$ have joint distribution $ER(2, h; \vec{p})$. By Lemma 45, the probability that

Algorithm 2 misaligns u with v or v with u is at most $2 \exp(-h\rho^2)$. Then, by the union bound over all $\binom{n}{2}$ pairs of vertices, Algorithm 2 correctly recovers the alignment between B_a and B_b with probability at least $1 - n(n-1) \exp(-h\rho^2)$ and the algorithm is correct with probability $1 - o(1)$ when

$$h \geq \frac{2 \log n + \omega(1)}{\rho^2}. \quad (6.10)$$

In our analysis of Algorithm 1, the situation is similar yet not quite as simple as the one described in Remark 2. After we find the lists of anchors in G_a and G_b , we obtain a pair of induced bipartite subgraphs: $G_a[V_a \setminus H_a, \vec{w}_a]$ and $G_b[V_b \setminus H_b, \vec{w}_b]$. When the anchor lists are the same, i.e. $\vec{w}_a = \vec{w}_b$, Algorithm 2 can be applied, but bipartite graphs do not have the joint distribution $ER(n-h, h, \vec{p})$, required for Remark 2. This is due to the fact that we used edge information to partition the original vertex set, so the edges are not independent of this partition. However, this dependence is weak. In Section 6.5 we will apply Lemma 45 after careful conditioning.

6.5 General alignment algorithm

In this section we first show that the anchor alignment stage is independent from the alignment of any pair of non-anchor vertices in the bipartite alignment step. We do this by considering the subgraph obtained by removing any pair of vertices and show that the anchor set is sufficiently stable due to the degree separation of at least 3 as guaranteed by Theorem 6.3.2. This then allows us to combine results on both stages to get the condition for successful alignment of pairs of random graphs.

Recall that $\vec{w}_a = f_h(G_a)$ and $\vec{w}_b = f_h(G_b)$. For $U = \{u_1, u_2\} \subseteq V$, the induced bipartite subgraphs $(G_a[U, \vec{w}_a], G_b[U, \vec{w}_b])$ determine whether Algorithm 1 misaligns u_1 with u_2 or u_2 with u_1 . However, these graphs do not have a correlated ER joint distribution, so we define a related pair of induced bipartite subgraphs.

Definition 13. Let G_a and G_b be graphs on vertex set V . For set $U = \{u_1, u_2\} \subseteq V$, and $h \in \mathbb{N}$, define

$$\vec{w}_a^U = f_h(G_a[V \setminus \{u_1, u_2\}]) \quad \text{and} \quad B_a^U = G_a[U, \vec{w}_a^U],$$

i.e. $(u, i) \in E(B_a^U) \iff \{u, w_i^U\} \in E(G_a)$ for any $u \in U$ and $i \in [h]$. Define \vec{w}_b^U and B_b^U analogously. Let $\mathcal{E}^H(U)$ be the event $\vec{w}_a^U = \vec{w}_b^U$.

We emphasize that in both B_a^U and B_b^U the left vertex set is $\{u_1, u_2\}$ and the right vertex set is $[h]$, so the vertex sets are not random variables.

We start by stating a result on conditional independence of the high-degree neighborhoods of a pair of vertices.

Lemma 47. Let $(G_a, G_b) \sim ER(n; \vec{p})$ be correlated graphs on the vertex set V and let $U = \{u_1, u_2\} \subseteq V$. Then

$$B_a^U \sim ER(2, h, p_{1*}), \quad B_b^U \sim ER(2, h, p_{*1})$$

$$\text{and} \quad (B_a^U, B_b^U) | \mathcal{E}^H(U) \sim ER(2, h, \vec{p}),$$

where B_a^U and B_b^U are as defined in Definition 13

Proof. Recall that, by definition, $B_a^U = G_a[U, \vec{w}_a^U]$ and $\mathbb{1}(u, j) \in E(B_a^U) = \mathbb{1}\{u, w_{a,j}^U\} \in E(G_a)$. We will show that despite being defined using \vec{w}_a^U , the random variable B_a^U is independent of the random variable \vec{w}_a^U . Observe that $B_a^U = G_a[U, \vec{w}_a^U]$ is independent of $G_a[V \setminus U]$ because they have no edge random variables in common. Because $\vec{w}_a^U = f_h(G_a[V \setminus U])$, B_a^U is independent of \vec{w}_a^U as well.

Similarly, B_b^U is independent of \vec{w}_b^U and $\mathbb{1}(u, j) \in E(B_b^U) = \mathbb{1}\{u, w_{b,j}\} \in E(G_b)$. As long as $\vec{w}_a^U = \vec{w}_b^U$ holds, $\mathbb{1}(u, j) \in E(B_a^U)$ and $\mathbb{1}(u, j) \in E(B_b^U)$ have the joint distribution of a pair of corresponding edges in the correlated Erdős-Rényi model. \square

This result may be counterintuitive because we are selecting the right vertex set of B_a^U using high degree vertices, but there the edge density of B_a^U is the same as G_a . For a fixed $(u, j) \in U \times [h]$, the random variable $\mathbb{1}(u, j) \in E(B_a^U)$ is not determined by any single edge random variable from G_a , but is a mixture of $\mathbb{1}\{u, v\} \in E(G_a)$ over all $v \in V \setminus U$ because \vec{w}_b^U is random. It is helpful to compare with $G_a[U^{\vec{w}_a}, \vec{w}_a]$, where $U^{\vec{w}_a} = \{u_1, u_2\}$ is a uniformly random subset of $V \setminus H_a$. This bipartite graph is not distributed as $ER(n, p_{1*})$ because edges of G_a are slightly more likely to be sampled than non-edges.

Recall from Definition 11 that \mathcal{E}_a^S is defined as the event that $\delta_{a,i} > \delta_{a,i+1} + 2$ for all $i \in [h]$ and \mathcal{E}_b^S is the corresponding event for \vec{w}_b and G_b .

Lemma 48. *The event \mathcal{E}_a^S implies $\vec{w}_a = \vec{w}_a^U$ for all $U \subseteq V$ pair of vertices that do not include any from \vec{w}_a . Similarly \mathcal{E}_b^S implies $\vec{w}_b = \vec{w}_b^U$.*

Proof. For any $v \in V$, the degree of v in G_a differs by at most 2 from the degree of v in $G_a[V \setminus U]$. The same holds for G_b . □

Finally we prove our main theorem:

Proof of Theorem 6.2.1. Theorem 6.3.2 provides the condition on the correlation of graphs required to successfully align a given number h of high-degree vertices. From the inequalities $h \leq \mathcal{O}\left(\frac{\log n}{p_{11}}\right)$, $\log h \leq \log n$, and the conditions in the theorem statement, $p_{11} \geq \omega\left(n^{-1/5} \log^{7/5} n\right)$ and $p_{01} + p_{10} \leq o\left(\frac{p_{11}^5}{\log^6 n}\right)$, we have

$$\max\{(\log h)^2, n(p_{10} + p_{01}) \log h\} \leq o\left(\frac{n p_{11}}{h^4 \log n} \cdot \frac{p_{11}}{p_{1*}}\right).$$

Thus $P[\mathcal{E}^H \wedge \mathcal{E}_a^S \wedge \mathcal{E}_b^S] \geq 1 - o(1)$, where \mathcal{E}^H , \mathcal{E}_a^S and \mathcal{E}_b^S are events as defined in Definition 11. These events imply $H_a = H_b = H$.

Recall the definition of $\mathcal{E}^M(B_a, B_b)$ from Lemma 45 and $\mathcal{E}^H(U)$ from Definition 13. Applying the union bound to error events in the bipartite alignment stage of the algorithm

results in the following:

$$\begin{aligned}
& P[\widehat{M} \neq M | \mathcal{E}^H \wedge \mathcal{E}_a^S \wedge \mathcal{E}_b^S] \\
& \leq \sum_{\{u_1, u_2\} \subseteq V \setminus H} P[\mathcal{E}^M(G_a[U, \vec{w}_a], G_b[U, \vec{w}_b]) \wedge \mathcal{E}^H \wedge \mathcal{E}_a^S \wedge \mathcal{E}_b^S] \\
& \stackrel{(a)}{\leq} \sum_{\{u_1, u_2\} \subseteq V \setminus H} P[\mathcal{E}^M(B_a^U, B_b^U) \wedge \mathcal{E}^H(U)] \\
& \stackrel{(b)}{\leq} \sum_{\{u_1, u_2\} \subseteq V} P[\mathcal{E}^M(B_a^U, B_b^U) | \mathcal{E}^H(U)] \\
& \stackrel{(c)}{\leq} \sum_{\{u_1, u_2\} \subseteq V} \exp(-h\rho^2).
\end{aligned}$$

The inequality (a) is derived by applying Lemma 48 twice, which gives $G_a[U, \vec{w}_a] = B_a^U$, $G_b[U, \vec{w}_b] = B_b^U$, and $\vec{w}_a^U = \vec{w}_b^U$. (Recall that the event $\{\vec{w}_a^U = \vec{w}_b^U\}$ is denoted by $\mathcal{E}^H(U)$.) In (b), we use $P[\mathcal{E}^U] \leq 1$ and also extend the sum to include pairs $\{u_1, u_2\}$ that include members of H . Because u_1 and u_2 are now arbitrary vertices with no conditioning, from Lemma 47 we have that $(B_a^U, B_b^U) \sim ER(2, h, \vec{p})$. Observe that for any $U = \{u_1, u_2\} \subseteq V \setminus H$, the signatures in Lemma 45 are the same as the signatures in Algorithm 1: $\text{sig}_{G_a[U, \vec{w}_a]}(u_i) = \text{sig}'_a(u_i)$. Finally, (c) follows from Lemma 45. Note that the final bound is the same as the one stated earlier in Remark 2.

We have

$$\begin{aligned}
\rho &= \sqrt{p_{00}p_{1*} + p_{11}p_{0*}} - \sqrt{p_{10}p_{0*} + p_{01}p_{1*}} \\
&= \sqrt{p_{11}p_{00}} \left(\sqrt{2 + \frac{p_{10}}{p_{11}} + \frac{p_{01}}{p_{00}}} - \sqrt{2 \frac{p_{01}p_{10}}{p_{11}p_{00}} + \frac{p_{10}}{p_{11}} + \frac{p_{01}}{p_{00}}} \right) \\
&\geq \sqrt{2p_{11}} \left(1 - \mathcal{O}\left(\frac{1}{\log n}\right) \right)
\end{aligned}$$

because $\frac{p_{10}}{p_{11}} \leq o\left(\frac{1}{\log^6 n}\right)$ and $\frac{p_{01}}{p_{00}} \leq o\left(\frac{1}{\log^6 n}\right)$. The logarithm of the probability of an

incorrect alignment in $V \setminus H$ is at most

$$\begin{aligned} & \log(n(n-1)\exp(-h\rho^2)) \\ & \leq 2\log n - \frac{\log n + \omega(1)}{p_{11}} \cdot 2p_{11} \left(1 - \mathcal{O}\left(\frac{1}{\log n}\right)\right) \\ & \leq 2\log n - 2\log n + \mathcal{O}(1) - \omega(1) \left(1 - \mathcal{O}\left(\frac{1}{\log n}\right)\right) = -\omega(1). \end{aligned}$$

□

Appendices

APPENDIX A

STATISTICS OF THE INFORMATION DENSITY MATRIX FOR GAUSSIAN

DATABASE ALIGNMENT

Recall the definition of \mathbf{G} : Let f_{XY} , f_X and f_Y denote the joint and marginal distributions for correlated features in \vec{A} and \vec{B} .

$$G_{u,v} = \log \frac{f_{XY}(\vec{A}(v), \vec{B}(v))}{f_X(\vec{A}(u))f_Y(\vec{B}(v))}$$

Lemma 49. *The expressions for the first and second moments of \mathbf{G} in terms of the correlation vector $\vec{\rho}$ as defined in Appendix 2.4 are given below.*

(a) *Mean and variance of information density of a true pair:*

$$\mathbb{E} \left[G_{u,v} | u \stackrel{M}{\sim} v \right] = \sum_{i \in \mathcal{D}} -\frac{1}{2} \log (1 - \rho_i^2)$$

$$\text{Var} \left(G_{u,v} | u \stackrel{M}{\sim} v \right) = \sum_{i \in \mathcal{D}} \rho_i^2$$

(b) *Mean and variance of information density of a false pair:*

$$\mathbb{E} \left[G_{u,v'} | u \not\stackrel{M}{\sim} v' \right] = \sum_{i \in \mathcal{D}} -\frac{1}{2} \log (1 - \rho_i^2) - \frac{\rho_i^2}{1 - \rho_i^2}$$

$$\text{Var} \left(G_{u,v'} | u \not\stackrel{M}{\sim} v' \right) = \sum_{i \in \mathcal{D}} \frac{\rho_i^2(1 + \rho_i^2)}{(1 - \rho_i^2)^2}$$

(c) *Covariance between information density of a true match and a false match with a user in common:*

$$\text{Cov} \left(G_{u,v}, G_{u,v'} | u \stackrel{M}{\sim} v \right) = \sum_{i \in \mathcal{D}} -\frac{\rho_i^4}{2(1 - \rho_i^2)}$$

(d) Covariance between information density of two false matches with a user in common:

$$\text{Cov} \left(G_{u,v'}, G_{u,v''} | u \stackrel{M}{\sim} v \right) = \sum_{i \in \mathcal{D}} \frac{\rho_i^4}{4(1 - \rho_i^2)^2}$$

(e) Covariance between information density of two false matches that break apart two true matches:

$$\text{Cov} \left(G_{u_2,v_1}, G_{u_1,v_2} | u_1 \stackrel{M}{\sim} v_1, u_2 \stackrel{M}{\sim} v_2 \right) = \sum_{i \in \mathcal{D}} \frac{\rho_i^4}{1 - \rho_i^2}$$

(f) Covariance between information density of two false matches that break apart one true match:

$$\text{Cov} \left(G_{u_2,v_1}, G_{u_1,v_2} | u_1 \stackrel{M}{\sim} v_1, u_2 \not\sim v_2 \right) = \sum_{i \in \mathcal{D}} \frac{\rho_i^4(1 + \rho_i^2)}{2(1 - \rho_i^2)^2}$$

(g) Covariance between information density of two matches that do not break apart any true match:

$$\text{Cov} \left(G_{u_2,v_1}, G_{u_1,v_2} | u_1 \not\sim v_1, u_2 \not\sim v_2 \right) = 0$$

Proof. Assume \vec{A} and \vec{B} are given in canonical form with correlation vector $\vec{\rho} \in [-1, 1]^{\mathcal{D}}$.

Then

$$\begin{aligned}
\log \frac{f_{XY}(\vec{x}, \vec{y})}{f_X(\vec{x})f_Y(\vec{y})} &= \log \frac{\prod_i f_{X_i Y_i}(x_i, y_i)}{\prod_i f_{X_i}(x_i) \prod_i f_{Y_i}(y_i)} \\
&= \sum_{i \in \mathcal{D}} \log \frac{f_{X_i Y_i}(x_i, y_i)}{f_{X_i}(x_i) f_{Y_i}(y_i)} \\
&= \sum_{i \in \mathcal{D}} \log \frac{\frac{1}{2\pi\sqrt{1-\rho_i^2}} \exp\left(-\frac{1}{2} \begin{bmatrix} x_i \\ y_i \end{bmatrix}^\top \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}^{-1} \begin{bmatrix} x_i \\ y_i \end{bmatrix}\right)}{\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x_i^2}{2}\right) \cdot \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y_i^2}{2}\right)} \\
&= \sum_{i \in \mathcal{D}} -\frac{1}{2} \log(1-\rho_i^2) - \frac{\rho_i^2(x_i^2 + y_i^2) - 2\rho_i x_i y_i}{2(1-\rho_i^2)} \\
&= \sum_{i \in \mathcal{D}} -\frac{1}{2} \log(1-\rho_i^2) - \frac{\rho_i^2(x_i - y_i)^2}{2(1-\rho_i^2)} + \frac{\rho_i x_i y_i}{1+\rho_i} \tag{A.1}
\end{aligned}$$

Then $G_{u,v}$ can be written as the sum of $|\mathcal{D}|$ independent random variables, each a function of only $A_i(u)$, $B_i(v)$ and ρ_i . For the rest of this section we assume $|\mathcal{D}| = 1$ and drop all subscripts i for simplicity of notation. The means, variances or covariances for the case $|\mathcal{D}| > 1$ can be found by summing over the corresponding expression for different values of ρ .

For the derivations (a) to (d), assume $u \stackrel{M}{\sim} v$ and let $X \triangleq A(u)$, $Y \triangleq B(v)$. Then $(X, Y) \sim \mathcal{N}\left(\vec{0}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right)$. Define $Z \triangleq \frac{Y-\rho X}{\sqrt{1-\rho^2}}$. Then $Y = \rho X + \sqrt{1-\rho^2}Z$. Also define $W = W_0 \triangleq B(v')$ and $W_1 \triangleq B(v'')$. Then W_0, W_1, X and Z are i.i.d. standard normal random variables.

(a) $\mathbb{E}\left[G_{u,v}|u \stackrel{M}{\sim} v\right]$ and $\text{Var}\left(G_{u,v}|u \stackrel{M}{\sim} v\right)$:

We want to find the mean and variance of $-\frac{1}{2} \log(1-\rho^2) - \frac{\rho^2(X-Y)^2}{2(1-\rho^2)} + \frac{\rho XY}{1+\rho}$. The first term is constant. We find the mean, variance and covariance of the two other terms.

$X - Y = (1-\rho)X - \sqrt{1-\rho^2}Z \sim \mathcal{N}(0, 2(1-\rho))$, which implies that $\frac{(X-Y)^2}{2(1-\rho)} \sim \chi^2(1)$ and therefore $\frac{\rho^2(X-Y)^2}{2(1-\rho^2)}$ has mean $\frac{\rho^2}{1+\rho}$ and variance $\frac{2\rho^4}{(1+\rho)^2}$.

$XY = \rho X^2 + \sqrt{1 - \rho^2} XZ$. Then XY has mean ρ .

$$\begin{aligned}
 (XY - \rho)^2 &= \rho^2 X^4 + 2\rho\sqrt{1 - \rho^2} X^3 Z + (1 - \rho^2) X^2 Z^2 \\
 &\quad - 2\rho^2 X^2 - 2\rho\sqrt{1 - \rho^2} XZ + \rho^2 \\
 \implies \text{Var}(XY) &= \mathbb{E}[(XY - \rho)^2] \\
 &= 3\rho^2 + (1 - \rho^2) - 2\rho^2 + \rho^2 \\
 &= 1 + \rho^2
 \end{aligned}$$

So $\frac{\rho XY}{1 + \rho}$ has mean $\frac{\rho^2}{1 + \rho}$ and variance $\frac{\rho^2(1 + \rho^2)}{(1 + \rho)^2}$.

$$\begin{aligned}
 \text{Cov}((X - Y)^2, XY) &= \mathbb{E} [((X - Y)^2 - 2(1 - \rho)) (XY - \rho)] \\
 ((X - Y)^2 - 2(1 - \rho)) (XY - \rho) &= -2(XY - \rho)^2 - 2(XY - \rho) \\
 &\quad + (X^2 + Y^2)(XY - \rho) \\
 &= -2(XY - \rho)^2 - 2(XY - \rho) \\
 &\quad + X^3 Y + XY^3 - \rho X^2 - \rho Y^2 \\
 \implies \text{Cov}((X - Y)^2, XY) &= -2 \text{Var}(XY) + 2\mathbb{E}[X^3 Y] - 2\rho\mathbb{E}[X^2] \\
 &= -2(1 + \rho^2) + 2\mathbb{E} [\rho X^4 + \sqrt{1 - \rho^2} X^3 Z] - 2\rho \\
 &= -2 - 2\rho^2 + 6\rho - 2\rho \\
 &= -2(1 - \rho)^2
 \end{aligned}$$

The covariance between $\frac{\rho^2(X - Y)^2}{2(1 - \rho^2)}$ and $\frac{\rho XY}{1 + \rho}$ equals $\frac{-\rho^3(1 - \rho)}{(1 + \rho)^2}$.

Then

$$\begin{aligned}
\mathbb{E} \left[G_{u,v} | u \stackrel{M}{\sim} v \right] &= \mathbb{E} \left[-\frac{1}{2} \log(1 - \rho^2) - \frac{\rho^2(X - Y)^2}{2(1 - \rho^2)} + \frac{\rho XY}{1 + \rho} \right] \\
&= -\frac{1}{2} \log(1 - \rho^2) - \frac{\rho^2}{1 + \rho} + \frac{\rho^2}{1 + \rho} \\
&= -\frac{1}{2} \log(1 - \rho^2) \\
\text{Var} \left(G_{u,v} | u \stackrel{M}{\sim} v \right) &= \text{Var} \left(\frac{\rho^2(X - Y)^2}{2(1 - \rho^2)} \right) + \text{Var} \left(\frac{\rho XY}{1 + \rho} \right) \\
&\quad - 2 \text{Cov} \left(\frac{\rho^2(X - Y)^2}{2(1 - \rho^2)}, \frac{\rho XY}{1 + \rho} \right) \\
&= \frac{2\rho^4}{(1 + \rho)^2} + \frac{\rho^2(1 + \rho^2)}{(1 + \rho)^2} + \frac{2\rho^3(1 - \rho)}{(1 + \rho)^2} \\
&= \frac{\rho^2 + 2\rho^3 + \rho^4}{(1 + \rho)^2} = \rho^2
\end{aligned}$$

(b) $\mathbb{E} \left[G_{u,v'} | u \stackrel{M}{\not\sim} v' \right]$ and $\text{Var} \left(G_{u,v'} | u \stackrel{M}{\not\sim} v' \right)$:

Once again, we find the mean, variance and covariance of the random terms in $-\frac{1}{2} \log(1 - \rho^2) - \frac{\rho^2(X - W)^2}{2(1 - \rho^2)} + \frac{\rho XW}{1 + \rho}$.

$X - W \sim \mathcal{N}(0, 2)$, which implies that $\frac{(X - W)^2}{2} \sim \chi^2(1)$ and therefore $\frac{\rho^2(X - W)^2}{2(1 - \rho^2)}$ has mean $\frac{\rho^2}{1 - \rho^2}$ and variance $\frac{2\rho^4}{(1 - \rho^2)^2}$.

$\mathbb{E}[XW] = 0$ and $\text{Var}(XW) = \mathbb{E}[X^2W^2] = \mathbb{E}[X^2]\mathbb{E}[W^2] = 1$. Then $\frac{\rho XW}{1 + \rho}$ has mean 0 and variance $\frac{\rho^2}{(1 + \rho)^2}$.

$$\begin{aligned}
\text{Cov}((X - W)^2, XW) &= \mathbb{E} \left[((X - W)^2 - 2) XW \right] \\
((X - W)^2 - 2) XW &= X^3W + XW^3 - 2X^2W^2 - 2XW \\
\implies \text{Cov}((X - W)^2, XW) &= -2
\end{aligned}$$

The covariance between $\frac{\rho^2(X - W)^2}{2(1 - \rho^2)}$ and $\frac{\rho XW}{1 + \rho}$ equals $\frac{-\rho^3}{(1 + \rho)(1 - \rho^2)}$.

Then

$$\begin{aligned}
\mathbb{E} \left[G_{u,v'} | u \stackrel{M}{\sim} v' \right] &= \mathbb{E} \left[-\frac{1}{2} \log(1 - \rho^2) - \frac{\rho^2(X - W)^2}{2(1 - \rho^2)} + \frac{\rho XW}{1 + \rho} \right] \\
&= -\frac{1}{2} \log(1 - \rho^2) - \frac{\rho^2}{1 - \rho^2} \\
\text{Var} \left(G_{u,v'} | u \stackrel{M}{\sim} v' \right) &= \text{Var} \left(\frac{\rho^2(X - W)^2}{2(1 - \rho^2)} \right) + \text{Var} \left(\frac{\rho XW}{1 + \rho} \right) \\
&\quad - 2 \text{Cov} \left(\frac{\rho^2(X - W)^2}{2(1 - \rho^2)}, \frac{\rho XW}{1 + \rho} \right) \\
&= \frac{2\rho^4}{(1 - \rho^2)^2} + \frac{\rho^2}{(1 + \rho)^2} + \frac{2\rho^3}{(1 + \rho)(1 - \rho^2)} \\
&= \frac{2\rho^4 + \rho^2(1 - \rho)^2 + 2\rho^3(1 - \rho)}{(1 - \rho^2)^2} \\
&= \frac{\rho^2(1 + \rho^2)}{(1 - \rho^2)^2}
\end{aligned}$$

(c) $\text{Cov} \left(G_{u,v}, G_{u,v'} | u \stackrel{M}{\sim} v \right)$:

We want to find the covariance between $-\frac{1}{2} \log(1 - \rho^2) - \frac{\rho^2(X - Y)^2}{2(1 - \rho^2)} + \frac{\rho XY}{1 + \rho}$ and $-\frac{1}{2} \log(1 - \rho^2) - \frac{\rho^2(X - W)^2}{2(1 - \rho^2)} + \frac{\rho XW}{1 + \rho}$.

We've already shown that $\mathbb{E}[(X - Y)^2] = 2(1 - \rho)$, $\mathbb{E}[XY] = \rho$, $\mathbb{E}[(X - W)^2] = 2$

and $\mathbb{E}[XW] = 0$.

- $\text{Cov} \left(\frac{\rho^2(X-Y)^2}{2(1-\rho^2)}, \frac{\rho^2(X-W)^2}{2(1-\rho^2)} \right) = \frac{\rho^4}{2(1+\rho)^2}$:

$$\begin{aligned}
& (X-Y)^2(X-W)^2 \\
&= \left(X(1-\rho) - Z\sqrt{1-\rho^2} \right)^2 (X-W)^2 \\
&= (1-\rho)^2 X^4 - 2 \left(W(1-\rho) + Z\sqrt{1-\rho^2} \right) (1-\rho) X^3 \\
&\quad + (W^2(1-\rho)^2 + Z^2(1-\rho^2)) X^2 - 4X^2 W Z (1-\rho) \sqrt{1-\rho^2} \\
&\quad - 2 \left(W^2 Z (1-\rho) \sqrt{1-\rho^2} + W Z^2 (1-\rho^2) \right) X + W^2 Z^2 (1-\rho^2) \\
&\implies \mathbb{E}[(X-Y)^2(X-W)^2] \\
&= 3(1-\rho)^2 + (1-\rho)^2 + (1-\rho^2) + (1-\rho^2) \\
&= 6 - 8\rho + 2\rho^2 \\
&\implies \text{Cov} \left((X-Y)^2, (X-W)^2 \right) \\
&= \mathbb{E}[(X-Y)^2(X-W)^2] - \mathbb{E}[(X-Y)^2] \mathbb{E}[(X-W)^2] \\
&= 6 - 8\rho + 2\rho^2 - 4 + 4\rho \\
&= 2(1-\rho)^2
\end{aligned}$$

- $\text{Cov} \left(\frac{\rho XY}{1+\rho}, \frac{\rho^2(X-W)^2}{2(1-\rho^2)} \right) = \frac{\rho^4}{(1+\rho)^2(1-\rho)}$:

$$\begin{aligned}
XY(X-W)^2 &= X \left(\rho X + \sqrt{1-\rho^2} Z \right) (X^2 - 2XW + W^2) \\
&= \rho X^4 + \left(\sqrt{1-\rho^2} Z - 2\rho W \right) X^3 \\
&\quad + \left(\rho W^2 - 2\sqrt{1-\rho^2} W Z \right) X^2 + \sqrt{1-\rho^2} X Z W^2 \\
&\implies \mathbb{E}[XY(X-W)^2] = 3\rho + \rho = 4\rho \\
&\implies \text{Cov} \left(XY, (X-W)^2 \right) = \mathbb{E}[XY(X-W)^2] - \mathbb{E}[XY] \mathbb{E}[(X-W)^2] \\
&= 4\rho - 2\rho = 2\rho
\end{aligned}$$

- $\text{Cov}\left(\frac{\rho^2(X-Y)^2}{2(1-\rho^2)}, \frac{\rho XW}{1+\rho}\right) = 0:$

$$\mathbb{E}[(X - Y)^2 XW] = \mathbb{E}[(X - Y)^2 X] \mathbb{E}[W] = 0$$

$$\implies \text{Cov}((X - Y)^2, XW) = \mathbb{E}[(X - Y)^2 XW] - \mathbb{E}[(X - Y)^2] \mathbb{E}[XW] = 0$$

- $\text{Cov}\left(\frac{\rho XY}{1+\rho}, \frac{\rho XW}{1+\rho}\right) = 0:$

$$\mathbb{E}[XYXW] = \mathbb{E}[X^2 Y] \mathbb{E}[W] = 0$$

$$\implies \text{Cov}(XY, XW) = \mathbb{E}[XYXW] - \mathbb{E}[XY] \mathbb{E}[XW] = 0$$

Then

$$\begin{aligned} \text{Cov}\left(G_{u,v}, G_{u,v'} | u \stackrel{M}{\sim} v\right) &= \text{Cov}\left(\frac{\rho^2(X - Y)^2}{2(1 - \rho^2)}, \frac{\rho^2(X - W)^2}{2(1 - \rho^2)}\right) - \text{Cov}\left(\frac{\rho XY}{1 + \rho}, \frac{\rho^2(X - W)^2}{2(1 - \rho^2)}\right) \\ &= \frac{\rho^4}{2(1 + \rho)^2} - \frac{\rho^4}{(1 + \rho)^2(1 - \rho)} \\ &= -\frac{\rho^4}{2(1 - \rho^2)} \end{aligned}$$

(d) $\text{Cov}\left(G_{u,v'}, G_{u,v''} | u \stackrel{M}{\sim} v\right):$

We want to find the covariance between $-\frac{1}{2} \log(1 - \rho^2) - \frac{\rho^2(X - W_0)^2}{2(1 - \rho^2)} + \frac{\rho XW_0}{1 + \rho}$ and $-\frac{1}{2} \log(1 - \rho^2) - \frac{\rho^2(X - W_1)^2}{2(1 - \rho^2)} + \frac{\rho XW_1}{1 + \rho}$.

We've already shown that $\mathbb{E}[(X - W_0)^2] = \mathbb{E}[(X - W_1)^2] = 2$ and $\mathbb{E}[XW_0] = \mathbb{E}[XW_1] = 0$. It can also be shown that $\text{Cov}((X - W_0)^2, XW_1) = \text{Cov}(XW_0, (X -$

$W_1)^2) = \text{Cov}(XW_0, XW_1) = 0$. Then we only need to compute $\text{Cov}\left(\frac{\rho^2(X-W_0)^2}{2(1-\rho^2)}, \frac{\rho^2(X-W_1)^2}{2(1-\rho^2)}\right)$.

$$\begin{aligned} (X - W_0)^2, (X - W_1)^2 &= X^4 - 2X^3(W_0 + W_1) + X^2(W_0^2 + W_1^2) \\ &\quad - 4X^2W_0W_1 - 2X(W_0^2W_1 + W_0W_1^2) + W_0^2W_1^2 \\ \implies \mathbb{E}[(X - W_0)^2, (X - W_1)^2] &= 3 + 1 + 1 = 5 \\ \implies \text{Cov}((X - W_0)^2, (X - W_1)^2) & \\ &= \mathbb{E}[(X - W_0)^2, (X - W_1)^2] - \mathbb{E}[(X - W_0)^2]\mathbb{E}[(X - W_1)^2] = 1 \end{aligned}$$

Then

$$\begin{aligned} \text{Cov}\left(G_{u,v'}, G_{u,v''} | u \stackrel{M}{\sim} v\right) &= \text{Cov}\left(\frac{\rho^2(X - W_0)^2}{2(1 - \rho^2)}, \frac{\rho^2(X - W_1)^2}{2(1 - \rho^2)}\right) \\ &= \frac{\rho^4}{4(1 - \rho^2)^2} \end{aligned}$$

(e) $\text{Cov}\left(G_{u_2,v_1}, G_{u_1,v_2} | u_1 \stackrel{M}{\sim} v_1, u_2 \stackrel{M}{\sim} v_2\right)$:

Let $X_k \triangleq A(u_k)$, $Y_k \triangleq B(v_k)$ and $Z_k \triangleq \frac{Y_k - \rho X_k}{\sqrt{1 - \rho^2}}$ for $k = \{1, 2\}$. Then X_1, X_2, Z_1, Z_2 are i.i.d. standard normal random variables and $\{X_2, Y_1\}$ and $\{X_1, Y_2\}$ are two pairs of independent standard normal variables.

We want to find the covariance between $-\frac{1}{2} \log(1 - \rho^2) - \frac{\rho^2(X_2 - Y_1)^2}{2(1 - \rho^2)} + \frac{\rho X_2 Y_1}{1 + \rho}$ and $-\frac{1}{2} \log(1 - \rho^2) - \frac{\rho^2(X_1 - Y_2)^2}{2(1 - \rho^2)} + \frac{\rho X_1 Y_2}{1 + \rho}$. We know that $\mathbb{E}[(X_2 - Y_1)^2] = \mathbb{E}[(X_2 - Y_1)^2] = 2$ and $\mathbb{E}[X_2 Y_1] = \mathbb{E}[X_1 Y_2] = 0$.

- $\text{Cov}\left(\frac{\rho^2(X_2 - Y_1)^2}{2(1 - \rho^2)}, \frac{\rho^2(X_1 - Y_2)^2}{2(1 - \rho^2)}\right) = 0$:

$$\begin{aligned} (X_2 - Y_1)^2(X_1 - Y_2)^2 &= \left(X_2 - \rho X_1 - \sqrt{1 - \rho^2} Z_1\right)^2 \\ &\quad \times \left(X_1 - \rho X_2 - \sqrt{1 - \rho^2} Z_2\right)^2 \end{aligned}$$

Ignoring the terms with an X_1, X_2, Z_1 or Z_2 factor of power 1, we get the following expansion:

$$\begin{aligned}
(X_2 - Y_1)^2(X_1 - Y_2)^2 &= \rho^2(X_1^4 + X_2^4) + (1 + \rho^4 - 4\rho^2)X_1^2X_2^2 \\
&\quad + (1 - \rho^2)(X_1^2Z_1^2 + X_2^2Z_2^2) \\
&\quad + \rho^2(1 - \rho^2)(X_1^2Z_2^2 + X_2^2Z_1^2) \\
&\quad + (1 - \rho^2)^2 Z_1^2Z_2^2 + (\dots) \\
\implies \mathbb{E}[(X_2 - Y_1)^2(X_1 - Y_2)^2] &= 6\rho^2 + (1 + \rho^4 - 4\rho^2) + 2(1 - \rho^2) \\
&\quad + 2\rho^2(1 - \rho^2) + (1 - \rho^2)^2 \\
&= 4 \\
\implies \text{Cov}((X_2 - Y_1)^2, (X_1 - Y_2)^2) &= \mathbb{E}[(X_2 - Y_1)^2(X_1 - Y_2)^2] \\
&\quad - \mathbb{E}[(X_2 - Y_1)^2]\mathbb{E}[(X_1 - Y_2)^2] \\
&= 0
\end{aligned}$$

$$\bullet \text{Cov}\left(\frac{\rho^2(X_2 - Y_1)^2}{2(1 - \rho^2)}, \frac{\rho X_1 Y_2}{1 + \rho}\right) = \text{Cov}\left(\frac{\rho X_2 Y_1}{1 + \rho}, \frac{\rho^2(X_1 - Y_2)^2}{2(1 - \rho^2)}\right) = -\frac{\rho^5}{(1 + \rho)^2(1 - \rho)}$$

$$(X_2 - Y_1)^2 X_1 Y_2 = \left(X_2 - \rho X_1 - \sqrt{1 - \rho^2} Z_1\right)^2 X_1 \left(\rho X_2 + \sqrt{1 - \rho^2} Z_2\right)$$

Ignoring the terms with an X_1, X_2, Z_1 or Z_2 factor of power 1, we get the following expansion:

$$\begin{aligned}
(X_2 - Y_1)^2 X_1 Y_2 &= -2\rho^2 X_1^2 X_2 + (\dots) \\
\implies \mathbb{E}[(X_2 - Y_1)^2 X_1 Y_2] &= -2\rho^2 \\
\implies \text{Cov}((X_2 - Y_1)^2, X_1 Y_2) &= \mathbb{E}[(X_2 - Y_1)^2 X_1 Y_2] - \mathbb{E}[(X_2 - Y_1)^2]\mathbb{E}[X_1 Y_2] \\
&= -2\rho^2
\end{aligned}$$

$$\bullet \text{Cov} \left(\frac{\rho X_2 Y_1}{1+\rho}, \frac{\rho X_1 Y_2}{1+\rho} \right) = \frac{\rho^4}{(1+\rho)^2}$$

$$\begin{aligned} \mathbb{E}[X_2 Y_1 X_1 Y_2] &= \mathbb{E}[X_1 Y_1] \mathbb{E}[X_2 Y_2] = \rho^2 \\ \implies \text{Cov}(X_2 Y_1, X_1 Y_2) &= \mathbb{E}[X_2 Y_1 X_1 Y_2] - \mathbb{E}[X_2 Y_1] \mathbb{E}[X_1 Y_2] \\ &= \rho^2 \end{aligned}$$

Then

$$\begin{aligned} \text{Cov} \left(G_{u_2, v_1}, G_{u_1, v_2} \mid u_1 \stackrel{M}{\sim} v_1, u_2 \stackrel{M}{\sim} v_2 \right) &= \text{Cov} \left(\frac{\rho^2 (X_2 - Y_1)^2}{2(1-\rho^2)}, \frac{\rho^2 (X_1 - Y_2)^2}{2(1-\rho^2)} \right) \\ &\quad - 2 \text{Cov} \left(\frac{\rho^2 (X_2 - Y_1)^2}{2(1-\rho^2)}, \frac{\rho X_1 Y_2}{1+\rho} \right) \\ &\quad + \text{Cov} \left(\frac{\rho X_2 Y_1}{1+\rho}, \frac{\rho X_1 Y_2}{1+\rho} \right) \\ &= 0 + \frac{2\rho^5}{(1+\rho)^2(1-\rho)} + \frac{\rho^4}{(1+\rho)^2} \\ &= \frac{\rho^4}{1-\rho^2} \end{aligned}$$

(f) $\text{Cov} \left(G_{u_2, v_1}, G_{u_1, v_2} \mid u_1 \stackrel{M}{\sim} v_1, u_2 \stackrel{M}{\sim} v_2 \right)$:

Let $X \triangleq A(u_1)$, $Y \triangleq B(v_1)$ and $Z \triangleq \frac{Y - \rho X}{\sqrt{1-\rho^2}}$. Furthermore let $W_x \triangleq A(u_2)$ and $W_y \triangleq B(v_2)$. Then W_x, W_y, X and Z are i.i.d. standard normal random variables.

We want to find the covariance between $-\frac{1}{2} \log(1-\rho^2) - \frac{\rho^2(W_x - Y)^2}{2(1-\rho^2)} + \frac{\rho W_x Y}{1+\rho}$ and $-\frac{1}{2} \log(1-\rho^2) - \frac{\rho^2(X - W_y)^2}{2(1-\rho^2)} + \frac{\rho X W_y}{1+\rho}$. We know that $\mathbb{E}[(W_x - Y)^2] = \mathbb{E}[(X - W_y)^2] = 2$ and $\mathbb{E}[W_x Y] = \mathbb{E}[X W_y] = 0$. It can also be shown that $\text{Cov}((W_x - Y)^2, X W_y) = \text{Cov}(W_x Y, (X - W_y)^2) = \text{Cov}(W_x Y, X W_y) = 0$. Then we only need to compute $\text{Cov} \left(\frac{\rho^2(W_x - Y)^2}{2(1-\rho^2)}, \frac{\rho^2(X - W_y)^2}{2(1-\rho^2)} \right)$.

$$(W_x - Y)^2 (X - W_y)^2 = \left(W_x - \rho X - \sqrt{1-\rho^2} Z \right)^2 (X - W_y)^2$$

Ignoring the terms with an W_x, W_y, X or Z factor of power 1, we get the following expansion:

$$\begin{aligned}
(W_x - Y)^2(X - W_y)^2 &= \rho^2 X^4 + X^2 W_x^2 + \rho^2 X^2 W_y^2 \\
&\quad + (1 - \rho^2) X^2 Z^2 + W_x^2 W_y^2 + (1 - \rho^2) Z^2 W_y^2 + (\dots) \\
\implies \mathbb{E}[(W_x - Y)^2(X - W_y)^2] &= 3\rho^2 + 1 + \rho^2 + (1 - \rho^2) + 1 + (1 - \rho^2) \\
&= 2(1 + \rho^2)
\end{aligned}$$

Then

$$\begin{aligned}
\text{Cov} \left(G_{u_2, v_1}, G_{u_1, v_2} \mid u_1 \stackrel{M}{\sim} v_1, u_2 \not\sim v_2 \right) &= \text{Cov} \left(\frac{\rho^2 (W_x - Y)^2}{2(1 - \rho^2)}, \frac{\rho^2 (X - W_y)^2}{2(1 - \rho^2)} \right) \\
&= \frac{\rho^4 (1 + \rho^2)}{2(1 - \rho^2)^2}
\end{aligned}$$

$$(g) \text{ Cov} \left(G_{u_2, v_1}, G_{u_1, v_2} \mid u_1 \not\sim v_1, u_2 \not\sim v_2 \right)$$

Conditioned on $u_1 \not\sim v_1, u_2 \not\sim v_2$, $(A(u_2), B(u_1))$ is independent from $(A(u_1), B(u_2))$.

Since G_{u_2, v_1} is a function of the former and G_{u_1, v_2} a function of the latter with no additional randomness, it follows that G_{u_2, v_1} and G_{u_1, v_2} are independent and therefore have no correlation.

□

Lemma 50. *Under Condition 4,*

(a) *Mean and variance of information density of a true pair:*

$$\begin{aligned}
\mathbb{E} \left[G_{u, v} \mid u \stackrel{M}{\sim} v \right] &= I_{XY} \\
2I_{XY} &\geq \text{Var} \left(G_{u, v} \mid u \stackrel{M}{\sim} v \right) \geq 2I_{XY}(1 - o(1))
\end{aligned}$$

(b) *Mean and variance of information density of a false pair:*

$$\begin{aligned} -I_{XY} &\geq \mathbb{E} \left[G_{u,v'} | u \not\stackrel{M}{\sim} v' \right] \geq -I_{XY}(1 + o(1)) \\ 2I_{XY} &\leq \text{Var} \left(G_{u,v'} | u \not\stackrel{M}{\sim} v' \right) \leq 2I_{XY}(1 + o(1)) \end{aligned}$$

(c) *Covariance between information density of a true match and a false match with a user in common:*

$$0 > \text{Cov} \left(G_{u,v}, G_{u,v'} | u \stackrel{M}{\sim} v \right) \geq -\mathcal{O} \left(\rho_{\max}^2 I_{XY} \right)$$

(d) *Covariance between information density of two false matches with a user in common:*

$$0 < \text{Cov} \left(G_{u,v'}, G_{u,v''} | u \stackrel{M}{\sim} v \right) \leq \mathcal{O} \left(\rho_{\max}^2 I_{XY} \right)$$

(e) *Covariance between information density of two false matches that break apart two true matches:*

$$0 < \text{Cov} \left(G_{u_2,v_1}, G_{u_1,v_2} | u_1 \stackrel{M}{\sim} v_1, u_2 \stackrel{M}{\sim} v_2 \right) \leq \mathcal{O} \left(\rho_{\max}^2 I_{XY} \right)$$

(f) *Covariance between information density of two false matches that break apart one true match:*

$$0 < \text{Cov} \left(G_{u_2,v_1}, G_{u_1,v_2} | u_1 \stackrel{M}{\sim} v_1, u_2 \not\stackrel{M}{\sim} v_2 \right) \leq \mathcal{O} \left(\rho_{\max}^2 I_{XY} \right)$$

(g) *Covariance between information density of two matches that do not break apart any true match:*

$$\text{Cov} \left(G_{u_2,v_1}, G_{u_1,v_2} | u_1 \not\stackrel{M}{\sim} v_1, u_2 \not\stackrel{M}{\sim} v_2 \right) = 0$$

Proof. Exact expressions for the statistics are given in Lemma 49.

Next we show the following statements which help us derive the asymptotic forms of the expressions for the statistics. As $x \rightarrow 0$,

- (i) $-\log(1-x) \geq x \geq -(1-o(1))\log(1-x)$,
- (ii) $-\log(1-x) \leq \frac{x}{1-x} \leq -(1+o(1))\log(1-x)$,
- (iii) $-\log(1-x) \leq \frac{x(1+x)}{(1-x)^2} \leq -(1+o(1))\log(1-x)$.

(i) By the Taylor series expansion of \log , given $x \in (0, 1)$,

$$\begin{aligned} -\frac{1}{x} \log(1-x) &= \frac{1}{x} \sum_{k=1}^{\infty} \frac{x^k}{k} \\ &> \frac{1}{x} \sum_{k=1}^1 x^k = 1. \end{aligned}$$

So $-\log(1-x) > x$. Furthermore, the limit of $-\frac{1}{x} \log(1-x)$ and the limit of $\frac{1}{1-x}$ are both 1 as $x \rightarrow 0$. Then, given $x \leq o(1)$, $-\log(1-x) \geq x \geq -(1-o(1))\log(1-x)$.

(ii) Similarly, by the Taylor series expansion of \log , given $x \in (0, 1)$,

$$\begin{aligned} -\frac{1}{x} \log(1-x) &= \frac{1}{x} \sum_{k=1}^{\infty} \frac{x^k}{k} \\ &< \frac{1}{x} \sum_{k=1}^{\infty} x^k = \frac{1}{1-x}. \end{aligned}$$

So $-\log(1-x) < \frac{x}{1-x}$. Furthermore, the limit of $-\frac{1}{x} \log(1-x)$ and the limit of $\frac{1}{1-x}$ are both 1 as $x \rightarrow 0$. Then, given $x \leq o(1)$, $-\log(1-x) \leq \frac{x}{1-x} \leq -(1+o(1))\log(1-x)$.

- (iii) Finally, $-\log(1-x) < \frac{x}{1-x} < \frac{x(1+x)}{(1-x)^2}$ and the limits of $-\frac{1}{x} \log(1-x)$ and $\frac{x(1+x)}{(1-x)^2}$ are both 1 as $x \rightarrow 0$. So $-\log(1-x) \leq \frac{x(1+x)}{(1-x)^2} \leq -(1+o(1))\log(1-x)$.

Next we derive the asymptotic forms of the expressions for the statistics. Define $\rho_{\max}^2 \triangleq \max_i \rho_i^2$.

- (a) The expression for $\mathbb{E} \left[G_{u,v} | u \stackrel{M}{\sim} v \right]$ is exactly equal to the expression for mutual in-

- formation $I_{XY} = -\frac{1}{2} \sum \log(1 - \rho_i^2)$. By (i), $-I_{XY} \geq \mathbb{E} \left[G_{u,v'} | u \stackrel{M}{\not\sim} v' \right] \geq I_{XY} - 2I_{XY}(1 + o(1)) = -I_{XY}(1 + o(1))$.
- (b) By (ii), $-I_{XY} \geq \mathbb{E} \left[G_{u,v'} | u \stackrel{M}{\not\sim} v' \right] \geq I_{XY} - 2I_{XY}(1 + o(1)) = -I_{XY}(1 + o(1))$. By (iii), $2I_{XY} \leq \sum \frac{\rho_i^2(1+\rho_i^2)}{(1-\rho_i^2)^2} \leq 2I_{XY}(1 + o(1))$.
- (c) By (ii), $0 < \frac{x^2}{1-x} \leq -(1 + o(1))x \log(1 - x)$ which is $\mathcal{O}(-x \log(1 - x))$. Then $0 > \text{Cov} \left(G_{u,v}, G_{u,v'} | u \stackrel{M}{\sim} v \right) \geq -\mathcal{O}(\rho_{\max}^2 I_{XY})$.
- (d) By (ii), $\frac{x^2}{(1-x)^2} = \left(\frac{x}{1-x} \right)^2 \leq (1 + o(1))(-\log(1 - x))^2$. By (i), $(-\log(1 - x))^2 \leq -(1+o(1))x \log(1-x)$, which is $\mathcal{O}(-x \log(1-x))$. Then $0 < \text{Cov} \left(G_{u,v'}, G_{u,v''} | u \stackrel{M}{\sim} v \right) \leq \mathcal{O}(\rho_{\max}^2 I_{XY})$.
- (e) By the same argument as for (c), $\frac{x^2}{1-x} \leq \mathcal{O}(-x \log(1 - x))$ and therefore $0 < \text{Cov} \left(G_{u_2,v_1}, G_{u_1,v_2} | u_1 \stackrel{M}{\sim} v_1, u_2 \stackrel{M}{\sim} v_2 \right) \leq \mathcal{O}(\rho_{\max}^2 I_{XY})$.
- (f) By the same argument as for (d), $\frac{x^2}{(1-x)^2} \leq \mathcal{O}(-x \log(1 - x))$. Then $\frac{x^2}{(1-x)^2}(1 + x) \leq \frac{x^2}{(1-x)^2}(1+o(1)) \leq \mathcal{O}(-x \log(1-x))$. So $0 < \text{Cov} \left(G_{u_2,v_1}, G_{u_1,v_2} | u_1 \stackrel{M}{\sim} v_1, u_2 \stackrel{M}{\not\sim} v_2 \right) \leq \mathcal{O}(\rho_{\max}^2 I_{XY})$.
- (g) $\text{Cov} \left(G_{u_2,v_1}, G_{u_1,v_2} | u_1 \stackrel{M}{\not\sim} v_1, u_2 \stackrel{M}{\not\sim} v_2 \right) = 0$

□

APPENDIX B
OTHER LEMMAS

Here we present some auxiliary lemmas that are instrumental for the proofs of the main results in this thesis, but have not been included in the corresponding chapters in order to have a cleaner presentation.

Lemmas 51, 52 and 53 are used in Section 4.2 to simplify the expressions for the generating functions under the low per-feature correlation regime. Lemma 54 studies a function that happens to be instrumental in proving (in Subsection 5.3.3) the majorization ordering between the distributions that appear in Section 5.4. Lemma 55 gives the expression for the limit of the normalized error exponent in Subsection 5.4.3.

Lemma 51 (Bernoulli's inequality). *Let $a \in (-1, \infty)$. Then*

- $(1 + ax) > (1 + a)^x$ if $x \in (0, 1)$.
- $(1 + ax) < (1 + a)^x$ if $x \in \mathbb{R} \setminus [0, 1]$.
- If $|ax| \leq o(1)$,
then $\left| \frac{1+ax}{(1+a)^x} - 1 \right| \leq o(1)$.

Proof. Let $f(x) \triangleq (1 + a)^x - (1 + ax)$. $f(0) = f(1) = 0$. The first derivative is given by $f'(x) = (1 + a)^x \log(1 + a) - a$. $f'(0)$ is strictly negative for $x = 0$ since $\log(1 + a) < a$ and $f'(1)$ is strictly positive for $x = 1$ since $\log(1 + a) \geq \frac{a}{1+a}$. Furthermore the second derivative $f''(x) = (1 + a)^x \log^2(1 + a)$ is strictly positive. Then this function has a global minimum at some $x^* \in (0, 1)$, is strictly decreasing for $x < x^*$ and strictly increasing for $x > x^*$. Then $f(x) < 0$ for $x \in (0, 1)$ and $f(x) > 0$ for $x \in (-\infty, 0) \cup (1, \infty)$.

The Taylor series expansion for $f(x)$ at $x = 0$ is given by

$$f(x) = -ax + \sum_{k=1}^{\infty} \frac{x^k \log^k(1+a)}{k!},$$

which is $-o(1) + o(1)$ if $|ax| \leq o(1)$. Then $\left| \frac{1+ax}{(1+a)^x} - 1 \right| \leq o(1)$.

□

Lemma 52. Let $\tau \in (0, \infty)$, $\sigma \in (-\tau, \tau)$, $x_1, \dots, x_n \in [-1, 1]$ and $s = \sum x_i$. Then

$$\prod_{i=1}^n (\tau + \sigma x_i) \geq (\tau - \sigma)^{\frac{n-s}{2}} (\tau + \sigma)^{\frac{n+s}{2}}$$

Proof. Define $\theta_i = \frac{x_i+1}{2} \in [0, 1]$. By the concavity of \log ,

$$\begin{aligned} \log(\tau + \sigma x_i) &= \log(\theta_i(\tau + \sigma) + (1 - \theta_i)(\tau - \sigma)) \\ &\geq \theta_i \log(\tau + \sigma) + (1 - \theta_i) \log(\tau - \sigma) \\ &= \frac{x_i}{2} \log\left(\frac{\tau + \sigma}{\tau - \sigma}\right) + \frac{1}{2} \log(\tau^2 - \sigma^2) \end{aligned}$$

Then

$$\begin{aligned} \prod_{i=1}^n (\tau + \sigma x_i) &\geq \exp\left(\sum_{i=1}^n \frac{x_i}{2} \log\left(\frac{\tau + \sigma}{\tau - \sigma}\right) + \frac{1}{2} \log(\tau^2 - \sigma^2)\right) \\ &= \exp\left(\frac{s}{2} \log\left(\frac{\tau + \sigma}{\tau - \sigma}\right) + \frac{n}{2} \log(\tau^2 - \sigma^2)\right) \\ &= (\tau + \sigma)^{\frac{n+s}{2}} (\tau - \sigma)^{\frac{n-s}{2}} \end{aligned}$$

□

Lemma 53. Let $\rho_{\max} \triangleq \max_i |\rho_i|$ and $I_{XY} \triangleq -\frac{1}{2} \sum_i \log(1 - \rho_i^2)$. Then

$$-\sum_i \log(1 - \rho_i^2) + \log(1 + \rho_i^2) \leq \sum_i \rho_i^2 \log(1 - \rho_i^2) \leq \rho_{\max}^2 I_{XY}$$

Proof. First we show that, for any $x \in [0, 1]$, $-x \log(1 - x) \geq -\log(1 - x) - \log(1 + x)$:

By the Taylor series expansion,

$$\begin{aligned} -x \log(1 - x) &= x \sum_{k=1}^{\infty} \frac{x^k}{k} = \sum_{k=2}^{\infty} \frac{x^k}{k-1} \\ &= \sum_{k=1}^{\infty} \frac{x^{2k}}{2k-1} + \frac{x^{2k+1}}{2k} \end{aligned}$$

For $x \in [0, 1]$, $\frac{x^{2k}}{2k-1} + \frac{x^{2k+1}}{2k} \geq x^{2k} \left(\frac{1}{2k-1} + \frac{1}{2k} \right) = \frac{x^{2k}}{k} \left(1 + \frac{1}{2(2k-1)} \right)$, which is strictly greater than $\frac{x^{2k}}{k}$ for any $k > 1/2$. Then

$$\begin{aligned} -x \log(1 - x) &= \sum_{k=1}^{\infty} \frac{x^{2k}}{2k-1} + \frac{x^{2k+1}}{2k} \\ &> \sum_{k=1}^{\infty} \frac{x^{2k}}{k} \\ &\stackrel{(*)}{=} -\log(1 - x^2) \\ &= -\log(1 - x) - \log(1 + x) \end{aligned}$$

where $(*)$ follows from the Taylor series expansion. Consequently, we have

$$-\rho^2 \log(1 - \rho^2) \geq -\log(1 - \rho^2) - \log(1 + \rho^2)$$

for any $\rho \in [-1, 1]$.

The result follows from the fact that $\sum \rho_i^2 \log(1 - \rho_i^2) \leq \max_i \rho_i^2 \log(1 - \rho_i^2)$. \square

Lemma 54. Define $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ such that

$$f(x, \alpha) \triangleq \sin(2\pi x) - \sin(2\pi \alpha x).$$

Then

- If $\alpha \geq 1$,

then $f(x, \alpha) \leq 0$ at any $x \in \left[0, \frac{1}{2(\alpha+1)}\right]$
and $f(x, \alpha) \geq 0$ at any $x \in \left[-\frac{1}{2(\alpha+1)}, 0\right]$.

- If $\alpha \in \left(1, \frac{5}{3}\right]$,
then $f(x, \alpha) \leq -\frac{\pi^2(\alpha-1)^2}{24}$ at any $x \in \left[\frac{\alpha-1}{8\alpha}, \frac{1}{4\alpha}\right]$
and $f(x, \alpha) \geq \frac{\pi^2(\alpha-1)^2}{24}$ at any $x \in \left[-\frac{1}{4\alpha}, -\frac{\alpha-1}{8\alpha}\right]$.

Proof. All lemma statements are for the case where $\alpha \geq 1$, so we always assume $\alpha \geq 1$.

$f(x, \alpha)$, as a function of x , is the difference of two sinusoidal functions, one with frequency $2\pi\alpha$ and the other with frequency 2π . Let us refer to these functions as

$$f_\alpha(x) \triangleq \sin(2\pi\alpha x)$$

$$f_1(x) \triangleq \sin(2\pi x).$$

Then $f(x, \alpha) = f_1(x) - f_\alpha(x)$. Since f_1 and f_α are both odd functions, it follows that $f(x, \alpha)$ is an odd function of x . Then we only study the function for $x \geq 0$ and derive results for $x < 0$ based on this property of f .

- At $x = 0$, $f_1(x) = f_\alpha(x) = 0$, so $f(0, \alpha) = 0$.
- At $x = \frac{1}{2(\alpha+1)}$, $f_1(x) = f_\alpha(x)$, so $f\left(\frac{1}{2(\alpha+1)}, \alpha\right) = 0$.
- Within the interval $x \in \left[0, \frac{1}{2(\alpha+1)}\right]$, $f_\alpha(x) \geq f_1(x)$:
 - Within the interval $x \in \left[0, \frac{1}{4\alpha}\right]$, $2\pi\alpha x$ and $2\pi x$ are both within the interval $\left[0, \frac{\pi}{2}\right]$. Then, since $2\pi\alpha x \geq 2\pi x$, it follows that $f_\alpha(x) \geq f_1(x)$.
 - Within the interval $\left[\frac{1}{4\alpha}, \frac{1}{2(\alpha+1)}\right]$, $2\pi\alpha x \in \left[\frac{\pi}{2}, \pi\right]$ and $2\pi x \in \left[0, \frac{\pi}{2}\right]$. Then, since $\pi - 2\pi\alpha x \geq 2\pi x$, it follows that $f_\alpha(x) \geq f_1(x)$.
- Within the interval $x \in \left[0, \frac{1}{2(\alpha+1)}\right]$, $2\pi x \in [0, \pi]$, so $f_1(x) \geq 0$. Then we can bound

the second derivative of f with respect to x as

$$\begin{aligned}\frac{\partial^2}{\partial x^2} f(x, \alpha) &= -4\pi^2 [f_1(x) - \alpha^2 f_\alpha(x)] \\ &\geq -4\pi^2 [\alpha^2 f_1(x) - \alpha^2 f_\alpha(x)] \\ &= -4\pi^2 \alpha^2 f(x, \alpha).\end{aligned}$$

It then follows that, within the interval $x \in \left[0, \frac{1}{2(\alpha+1)}\right]$, f is a convex function.

- Given any $\alpha \in \left(1, \frac{5}{3}\right]$, at $x = \frac{1}{4\alpha}$, $f(x, \alpha) \leq -\frac{\pi^2(\alpha-1)^2}{24}$:
 - Consider $\alpha = \frac{5}{3}$ and $x = \frac{1}{4\alpha} = \frac{3}{20}$. Then $f(x, \alpha) < -0.19 < -0.183 < -\frac{\pi^2(\alpha-1)^2}{24}$.
 - At $x = \frac{1}{4\alpha}$, the expression for $\frac{f\left(\frac{1}{4\alpha}, \alpha\right)}{(\alpha-1)^2}$ is given by

$$\begin{aligned}\frac{f\left(\frac{1}{4\alpha}, \alpha\right)}{(\alpha-1)^2} &= \frac{1}{(\alpha-1)^2} \left[\sin\left(\frac{\pi}{2\alpha}\right) - \sin\left(\frac{\pi}{2}\right) \right] \\ &= \frac{\sin\left(\frac{\pi}{2\alpha}\right) - 1}{(\alpha-1)^2}.\end{aligned}$$

The numerator $\sin\left(\frac{\pi}{2\alpha}\right) - 1$ is negative and decreasing as a function of α within the interval $\alpha \in (1, \infty)$. The denominator $(\alpha-1)^2$ is positive and increasing as a function of α within the same interval. Then, $\frac{f\left(\frac{1}{4\alpha}, \alpha\right)}{(\alpha-1)^2}$ is a negative and decreasing function of α within the interval $\alpha \in (1, \infty)$.

Since $\frac{f\left(\frac{1}{4\alpha}, \alpha\right)}{(\alpha-1)^2} < -\frac{\pi^2}{24}$ at $\alpha = \frac{5}{3}$, it then follows that it also satisfies $\frac{f\left(\frac{1}{4\alpha}, \alpha\right)}{(\alpha-1)^2} < -\frac{\pi^2}{24}$ at any $\alpha \in \left(1, \frac{5}{3}\right]$.

- Given any $\alpha \in \left(1, \frac{5}{3}\right]$, at $x = \frac{\alpha-1}{8\alpha}$, $f(x, \alpha) \leq -\frac{\pi^2(\alpha-1)^2}{24}$:
 - Consider $\alpha = \frac{5}{3}$ and $x = \frac{\alpha-1}{8\alpha} = \frac{1}{20}$. Then $f(x, \alpha) < -0.19 < -0.183 < -\frac{\pi^2(\alpha-1)^2}{24}$.
 - At $x = \frac{\alpha-1}{8\alpha}$, the expression for $\frac{f\left(\frac{\alpha-1}{8\alpha}, \alpha\right)}{(\alpha-1)^2}$ is given by

$$\frac{f\left(\frac{\alpha-1}{8\alpha}, \alpha\right)}{(\alpha-1)^2} = \frac{1}{(\alpha-1)^2} \left[\sin\left(\frac{\pi(\alpha-1)}{4\alpha}\right) - \sin\left(\frac{\pi(\alpha-1)}{4}\right) \right]$$

The numerator $\sin\left(\sin\left(\frac{\pi(\alpha-1)}{4\alpha}\right) - \sin\left(\frac{\pi(\alpha-1)}{4}\right)\right)$ is negative and decreasing as a function of α within the interval $\alpha \in (1, 2]$. The denominator $(\alpha - 1)^2$ is positive and increasing as a function of α within the same interval. Then, $\frac{f\left(\frac{1}{4\alpha}, \alpha\right)}{(\alpha-1)^2}$ is a negative and decreasing function of α within the interval $\alpha \in (1, 2]$. Since $\frac{f\left(\frac{\alpha-1}{8\alpha}, \alpha\right)}{(\alpha-1)^2} < -\frac{\pi^2}{24}$ at $\alpha = \frac{5}{3}$, it then follows that it also satisfies $\frac{f\left(\frac{\alpha-1}{8\alpha}, \alpha\right)}{(\alpha-1)^2} < -\frac{\pi^2}{24}$ at any $\alpha \in \left(1, \frac{5}{3}\right]$.

- Since $f(x, \alpha)$ is a convex function of x within the interval $\left[0, \frac{1}{2(\alpha+1)}\right]$, and since $f(0, \alpha) = f\left(\frac{1}{2(\alpha+1)}, \alpha\right) = 0$, it follows that $f(x, \alpha) \leq 0$ at any $x \in \left[0, \frac{1}{2(\alpha+1)}\right]$.
- $f(x, \alpha)$ is a convex function of x within the interval $\left[0, \frac{1}{2(\alpha+1)}\right]$ which contains the interval $\left[\frac{\alpha-1}{8\alpha}, \frac{1}{4\alpha}\right]$. Furthermore, if $\alpha \in \left(1, \frac{5}{3}\right]$, then $f\left(\frac{\alpha-1}{8\alpha}\right) \leq -\frac{\pi^2(\alpha-1)^2}{24}$ and $f\left(\frac{1}{4\alpha}\right) \leq -\frac{\pi^2(\alpha-1)^2}{24}$. It then follows that $f(x, \alpha) \leq -\frac{\pi^2(\alpha-1)^2}{24}$ at any $x \in \left[\frac{\alpha-1}{8\alpha}, \frac{1}{4\alpha}\right]$.
- The statements for $x < 0$ follow directly from the fact that f is an odd function.

□

Lemma 55. Given any $b \in (0, \infty)$ and $c \in (-b, b)$,

$$\int_0^1 \log(b + c \cos(2\pi x)) dx = \log\left(b + \sqrt{b^2 - c^2}\right) - \log 2$$

Proof. • $f(0) = 0$

- $\log(1 + a \cos(2x\pi))$ is continuous and bounded over x for any value of $a \in (-1, 1)$.

Then,

$$\begin{aligned} \frac{d}{da} f(a) &= \int_0^1 \frac{d}{da} \log(1 + a \cos(2x\pi)) dx \\ &= \int_0^1 \frac{\cos(2x\pi)}{1 + a \cos(2x\pi)} dx \\ &= \frac{1}{a} - \frac{1}{a} \int_0^1 \frac{1}{1 + a \cos(2x\pi)} dx \\ &= \frac{1}{a} - \frac{1}{a} \int_{-1/2}^{1/2} \frac{1}{1 + a \cos(2x\pi)} dx \end{aligned}$$

Change of variables:

$$u = \tan(x\pi),$$

$$x = \frac{1}{\pi} \arctan(u),$$

$$\frac{dx}{du} = \frac{1}{\pi(1+u^2)}$$

Then, $\cos(2x\pi) = \frac{1-u^2}{1+u^2}$. So

$$\begin{aligned} \frac{d}{da} f(a) &= \frac{1}{a} - \frac{1}{a} \int_{-1/2}^{1/2} \frac{1}{1+a \cos(2x\pi)} dx \\ &= \frac{1}{a} - \frac{1}{a} \int_{-\infty}^{\infty} \frac{1}{1+a \cdot \frac{1-u^2}{1+u^2}} \cdot \frac{1}{\pi(1+u^2)} du \\ &= \frac{1}{a} - \frac{1}{a\pi} \int_{-\infty}^{\infty} \frac{1}{1+u^2+a(1-u^2)} du \\ &= \frac{1}{a} - \frac{1}{a\pi} \int_{-\infty}^{\infty} \frac{1}{1+a+u^2(1-a)} du \\ &= \frac{1}{a} - \frac{1}{a\pi(1+a)} \int_{-\infty}^{\infty} \frac{1}{1+u^2 \cdot \frac{1-a}{1+a}} du \\ &= \frac{1}{a} - \frac{1}{a\pi\sqrt{1-a^2}} \int_{-\infty}^{\infty} \frac{1}{1+u^2 \cdot \frac{1-a}{1+a}} du \\ &= \frac{1}{a} - \frac{1}{a\pi\sqrt{1-a^2}} \int_{-\infty}^{\infty} \frac{1}{1+\left(u\sqrt{\frac{1-a}{1+a}}\right)^2} du \sqrt{\frac{1-a}{1+a}} \\ &= \frac{1}{a} - \frac{1}{a\pi\sqrt{1-a^2}} \int_{-\infty}^{\infty} \frac{1}{1+y^2} dy \\ &= \frac{1}{a} - \frac{1}{a\pi\sqrt{1-a^2}} [\arctan y]_{-\infty}^{\infty} \\ &= \frac{1}{a} - \frac{1}{a\pi\sqrt{1-a^2}} [\pi] \\ &= \frac{1}{a} - \frac{1}{a\sqrt{1-a^2}} \\ &= \frac{1}{a} \left(1 - \frac{1}{\sqrt{1-a^2}} \right) \end{aligned}$$

- Then, for $a \in (-1, 0)$

$$\begin{aligned}
f(a) &= f(0) - \int_a^0 \frac{d}{dx} f(x) dx \\
&= 0 - \int_a^0 \frac{1}{x} \left(1 - \frac{1}{\sqrt{1-x^2}} \right) dx \\
&= - \left[\frac{1}{2} \log \left(\frac{1 + \sqrt{1-x^2}}{1 - \sqrt{1-x^2}} \right) + \log(x) \right]_a^0 \\
&= - \left[\frac{1}{2} \log \left(\frac{(1 + \sqrt{1-x^2})^2}{1 - (1-x^2)} \right) + \log(x) \right]_a^0 \\
&= - \left[\log \left(1 + \sqrt{1-x^2} \right) \right]_a^0 \\
&= \log \left(1 + \sqrt{1-a^2} \right) - \log(2)
\end{aligned}$$

and for $a \in (0, 1)$

$$\begin{aligned}
f(a) &= f(0) + \int_0^a \frac{d}{dx} f(x) dx \\
&= 0 + \int_0^a \frac{1}{x} \left(1 - \frac{1}{\sqrt{1-x^2}} \right) dx \\
&= \left[\frac{1}{2} \log \left(\frac{1 + \sqrt{1-x^2}}{1 - \sqrt{1-x^2}} \right) + \log(x) \right]_0^a \\
&= \left[\frac{1}{2} \log \left(\frac{(1 + \sqrt{1-x^2})^2}{1 - (1-x^2)} \right) + \log(x) \right]_0^a \\
&= \left[\log \left(1 + \sqrt{1-x^2} \right) \right]_0^a \\
&= \log \left(1 + \sqrt{1-a^2} \right) - \log 2
\end{aligned}$$

So $f(a) \triangleq \int_0^1 \log(1 + a \cos(2x\pi)) dx = \log(1 + \sqrt{1-a^2}) - \log 2$.

It then follows that for any $b > 0$ and $c \in (-b, b)$,

$$\begin{aligned}\int_0^1 \log(b + c \cos(2\pi x)) dx &= \log b + \int_0^1 \log\left(1 + \frac{c}{b} \cos(2\pi x)\right) dx \\ &= \log b + \log\left(1 + \sqrt{1 - \frac{c^2}{b^2}}\right) - \log 2 \\ &= \log\left(b + \sqrt{b^2 - c^2}\right) - \log 2\end{aligned}$$

□

REFERENCES

- [1] R. Singh, J. Xu, and B. Berger, “Global alignment of multiple protein interaction networks with application to functional orthology detection,” *Proceedings of the National Academy of Sciences*, vol. 105, no. 35, pp. 12 763–12 768, 2008.
- [2] D. Conte, P. Foggia, C. Sansone, and M. Vento, “Thirty years of graph matching in pattern recognition,” *International journal of pattern recognition and artificial intelligence*, vol. 18, no. 03, pp. 265–298, 2004.
- [3] A. Narayanan and V. Shmatikov, “Robust de-anonymization of large sparse datasets,” in *2008 IEEE Symposium on Security and Privacy (sp 2008)*, IEEE, 2008, pp. 111–125.
- [4] C. Culnane, B. I. Rubinstein, and V. Teague, “Health data in an open world,” *arXiv preprint arXiv:1712.05627*, 2017.
- [5] Y. Tian and J. Patel, “TALE: A Tool for Approximate Large Graph Matching,” *Data Engineering, International Conference on*, vol. 0, pp. 963–972, 2008.
- [6] S. Zhang, J. Yang, and W. Jin, “SAPPER: Subgraph Indexing and Approximate Matching in Large Graphs,” *PVLDB*, vol. 3, no. 1, pp. 1185–1194, 2010.
- [7] F. Shirani, S. Garg, and E. Erkip, “An information theoretic framework for active de-anonymization in social networks based on group memberships,” in *2017 55th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, IEEE, 2017, pp. 470–477.
- [8] E. Onaran, S. Garg, and E. Erkip, “Optimal de-anonymization in random graphs with community structure,” in *2016 50th Asilomar Conference on Signals, Systems and Computers*, IEEE, 2016, pp. 709–713.
- [9] L. Fu, X. Fu, Z. Hu, Z. Xu, and X. Wang, “De-anonymization of social networks with communities: When quantifications meet algorithms,” *arXiv preprint arXiv:1703.09028*, 2017.
- [10] R. Singh, J. Xu, and B. Berger, “Global alignment of multiple protein interaction networks with application to functional orthology detection,” *Proceedings of the National Academy of Sciences*, vol. 105, no. 35, pp. 12 763–12 768, 2008.
- [11] O. Kuchaiev, T. Milenković, V. Memišević, W. Hayes, and N. Pržulj, “Topological network alignment uncovers biological function and phylogeny,” *Journal of the Royal Society Interface*, pp. 1341–1354, 2010.

- [12] N. Malod-Dognin and N. Pržulj, “L-graal: Lagrangian graphlet-based network aligner,” *Bioinformatics*, vol. 31, no. 13, pp. 2182–2189, 2015.
- [13] V. Saraph and T. Milenković, “Magna: Maximizing accuracy in global network alignment,” *Bioinformatics*, vol. 30, no. 20, pp. 2931–2940, 2014.
- [14] A. E. Aladag and C. Erten, “SPINAL: scalable protein interaction network alignment,” *Bioinformatics*, vol. 29, no. 7, pp. 917–924, 2013.
- [15] P. Pedarsani and M. Grossglauser, “On the privacy of anonymized networks,” in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2011, pp. 1235–1243.
- [16] V. Lyzinski, D. E. Fishkind, and C. E. Priebe, “Seeded graph matching for correlated erdős-rényi graphs,” *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 3513–3540, 2014.
- [17] L. Ganassali, L. Massoulié, and M. Lelarge, “Impossibility of partial recovery in the graph alignment problem,” in *Conference on Learning Theory*, PMLR, 2021, pp. 2080–2102.
- [18] G. Hall and L. Massoulié, “Partial recovery in the graph alignment problem,” *Operations Research*, 2022.
- [19] F. Shirani, S. Garg, and E. Erkip, “Seeded graph matching: Efficient algorithms and theoretical guarantees,” in *2017 51st Asilomar Conference on Signals, Systems, and Computers*, IEEE, 2017, pp. 253–257.
- [20] B. Barak, C.-N. Chou, Z. Lei, T. Schramm, and Y. Sheng, “(nearly) efficient algorithms for the graph matching problem on correlated random graphs,” *Advances in Neural Information Processing Systems*, vol. 32, 2019.
- [21] O. E. Dai, D. Cullina, N. Kiyavash, and M. Grossglauser, “Analysis of a canonical labeling algorithm for the alignment of correlated erdős-rényi graphs,” *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, vol. 3, no. 2, pp. 1–25, 2019.
- [22] J. Ding, Z. Ma, Y. Wu, and J. Xu, “Efficient random graph matching via degree profiles,” *Probability Theory and Related Fields*, vol. 179, pp. 29–115, 2021.
- [23] Z. Fan, C. Mao, Y. Wu, and J. Xu, “Spectral graph matching and regularized quadratic relaxations i algorithm and gaussian analysis,” *Foundations of Computational Mathematics*, pp. 1–55, 2022.

- [24] Z. Fan, C. Mao, Y. Wu, and J. Xu, “Spectral graph matching and regularized quadratic relaxations ii: Erdős-rényi graphs and universality,” *Foundations of Computational Mathematics*, pp. 1–51, 2022.
- [25] C. Mao, Y. Wu, J. Xu, and S. H. Yu, “Random graph matching at otter’s threshold via counting chandeliers,” *arXiv preprint arXiv:2209.12313*, 2022.
- [26] Y. Wu, J. Xu, and H. Y. Sophie, “Settling the sharp reconstruction thresholds of random graph matching,” *IEEE Transactions on Information Theory*, vol. 68, no. 8, pp. 5391–5417, 2022.
- [27] E. Mossel and J. Xu, “Seeded graph matching via large neighborhood statistics,” *Random Structures & Algorithms*, vol. 57, no. 3, pp. 570–611, 2020.
- [28] L. Ganassali, “Sharp threshold for alignment of graph databases with gaussian weights,” in *Mathematical and Scientific Machine Learning*, PMLR, 2022, pp. 314–335.
- [29] D. Cullina, P. Mittal, and N. Kiyavash, “Fundamental limits of database alignment,” in *2018 IEEE International Symposium on Information Theory (ISIT)*, IEEE, 2018, pp. 651–655.
- [30] F. Shirani, S. Garg, and E. Erkip, “A concentration of measure approach to database de-anonymization,” in *2019 IEEE International Symposium on Information Theory (ISIT)*, IEEE, 2019, pp. 2748–2752.
- [31] K. Zeynep and B. Nazer, “Detecting correlated gaussian databases,” in *2022 IEEE International Symposium on Information Theory (ISIT)*, IEEE, 2022, pp. 2064–2069.
- [32] M. Chertkov, L. Kroc, F. Krzakala, M. Vergassola, and L. Zdeborová, “Inference in particle tracking experiments by passing messages between images,” *Proceedings of the National Academy of Sciences*, vol. 107, no. 17, pp. 7663–7668, 2010.
- [33] G. Semerjian, G. Sicuro, and L. Zdeborová, “Recovery thresholds in the sparse planted matching problem,” *Physical Review E*, vol. 102, no. 2, p. 022 304, 2020.
- [34] M. Moharrami, C. Moore, and J. Xu, “The planted matching problem: Phase transitions and exact results,” *The Annals of Applied Probability*, vol. 31, no. 6, pp. 2663–2720, 2021.
- [35] J. Ding, Y. Wu, J. Xu, and D. Yang, “The planted matching problem: Sharp threshold and infinite-order phase transition,” *arXiv preprint arXiv:2103.09383*, 2021.
- [36] L. Ramshaw and R. E. Tarjan, “On minimum-cost assignments in unbalanced bipartite graphs,” *HP Labs, Palo Alto, CA, USA, Tech. Rep. HPL-2012-40R1*, 2012.

- [37] O. E. Dai, D. Cullina, and N. Kiyavash, “Gaussian database alignment and gaussian planted matching,” *arXiv preprint arXiv:2307.02459*, 2023.
- [38] O. E. Dai, D. Cullina, and N. Kiyavash, “Database alignment with gaussian features,” in *The 22nd International Conference on Artificial Intelligence and Statistics*, PMLR, 2019, pp. 3225–3233.
- [39] O. E. Dai, D. Cullina, and N. Kiyavash, “Achievability of nearly-exact alignment for correlated gaussian databases,” in *2020 IEEE International Symposium on Information Theory (ISIT)*, IEEE, 2020, pp. 1230–1235.
- [40] B. Hajek, *Random Processes for Engineers*. Cambridge University Press, 2015, ISBN: 9781316241240.
- [41] A. W. Marshall, I. Olkin, and B. C. Arnold, “Inequalities: Theory of majorization and its applications,” 1979.
- [42] B. C. Arnold, *Majorization and the Lorenz order: A brief introduction*. Springer Science & Business Media, 2012, vol. 43.
- [43] L. Babai, P. Erdős, and S. M. Selkow, “Random graph isomorphism,” *SIAM Journal on Computing*, vol. 9, no. 3, pp. 628–635, Aug. 1980.
- [44] M. Mitzenmacher and T. Morgan, *Reconciling graphs and sets of sets*, 2017. eprint: arXiv:1707.05867.
- [45] D. Cullina and N. Kiyavash, “Improved achievability and converse bounds for erdos-renyi graph matching,” in *Proceedings of the 2016 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Science - SIGMETRICS 2016*, ACM Press, 2016.
- [46] P. Pedarsani and M. Grossglauser, “On the privacy of anonymized networks,” in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, 2011, pp. 1235–1243.
- [47] B. Bollobas, *Random Graphs*. Cambridge University Press, 2001.