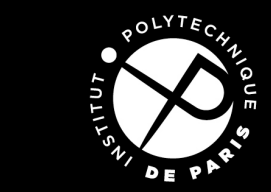
Contributions to structured high-dimensional inference

Suzanne Sigalla



INSTITUT POLYTECHNIQUE DE PARIS

NNT : 2022IPPAG013

Thèse de doctorat

ENSAE

IP PARIS

Contributions to structured high-dimensional inference

Thèse de doctorat de l'Institut Polytechnique de Paris
préparée à l'École Nationale de la Statistique et de l'Administration Économique

École doctorale n°574 École doctorale de mathématiques Hadamard (EDMH)
Spécialité de doctorat : Mathématiques appliquées

Thèse présentée et soutenue à Palaiseau, le 6 décembre 2022, par

SUZANNE SIGALLA

Composition du Jury :

Christophe Giraud Professeur, Université Paris-Saclay (Institut de Mathématiques d'Orsay)	Président
Marianna Pensky Professeur, University of Central Florida	Rapporteur
Mohamed Hebiri Maître de conférences, Université Paris-Est (Marne-la-Vallée)	Rapporteur
Katia Meziani Maître de conférences, Ceremade, Paris-Dauphine	Examinateur
Alexandre Tsybakov Directeur du département de statistiques du CREST, CREST-ENSAE	Directeur de thèse

Remerciements

Mes premiers remerciements vont à Sacha, mon directeur de thèse. Sacha, je ne saurais te remercier assez pour tout ce que tu m'as appris. Je te suis très reconnaissante de m'avoir donné ma chance pendant mon master et de m'avoir encouragée à persévérer en thèse. Je te remercie d'avoir toujours été si disponible et pédagogue tout au long de ma thèse. C'était un grand honneur de travailler avec toi : j'étais si intimidée par cette perspective au début et tu as su me donner une place en recherche. Je te remercie infiniment pour tout cela.

Un grand merci à Marianna Pensky et à Mohamed Hebiri, qui ont accepté d'être rapporteurs pour ma thèse. Je remercie également beaucoup Christophe Giraud et Katia Meziani d'avoir accepté de faire partie de mon jury de thèse. Je remercie l'ensemble du jury pour l'honneur qu'ils me font.

Merci à mes professeurs de Master 2, Christophe Giraud et Pierre Alquier, qui m'ont aidée dans mon parcours et m'ont encouragée à partir en thèse. Je remercie également chaleureusement tous ceux avec qui j'ai travaillé au cours de cette thèse. Simo Ndaoud, un grand merci, tu m'as accueillie au CREST et initiée à la recherche avec bienveillance. Merci à Olga Klopp pour ta patience et ta gentillesse. Merci à Maxim Panov pour tous nos échanges. Merci à Julien Chhor, toujours si enthousiaste.

Merci aux chercheurs et doctorants du CREST qui m'ont accompagnée dans cette expérience. Merci à Arnak pour ta bienveillance et ta sollicitude. Merci à Victor, "my statistical older brother", pour ta bonne humeur et tout le temps que tu consacres aux doctorants. Merci à Nicolas, Guillaume, Jaouad, Anna, Matthieu, Vianney, Cristina pour les nombreuses discussions et cafés du déjeuner.

Un merci tout particulier à Julien d'avoir été mon codoctorant pendant cette thèse, et de m'avoir épaulée de nombreuses fois. Nos moments de musique me manqueront énormément. Je te souhaite le meilleur aux États-Unis et dans la suite de ta carrière de chercheur. Merci aux anciens du CREST de m'y avoir si bien accueillie : Simo, Geoffrey, Badr, Avo, Amir, Lucie, Solenne, Lionel. Merci aux membres du bureau 3017, pour les nombreux cafés partagés ensemble : Flore, Yannis. Merci aux compagnons de thèse que je n'ai pas encore cités : Étienne, Meyer, Nicolas, Arshak, Arya, François-Pierre, Dang, Corentin, Gabriel D., Gabriel R., Gauthier, Jules, Martin. Merci à Djamila G. pour son aide. Merci à mon ancien professeur, Marie-Christine Pauzin, de m'avoir encouragée dans mes études.

Merci aux élèves à qui j'ai eu le plaisir d'enseigner pendant ma thèse.

Merci à mes amis de la Team Plante et LGs : Shu, Juliette, Valentin, Cédric, vous êtes les meilleurs amis dont je puisse rêver. Merci à Édith d'être une amie

si généreuse et intéressante. Merci à Hugo, Idriss, Aurélien, d'être présents depuis l'ENSAE. Merci à Charlotte P. de m'avoir soutenue dans des décisions difficiles. Merci à Juliette et Jean-Guillaume d'être des amis si fidèles. Merci à Nona pour son amitié constante. Merci à J., à qui je garde mon affection. Merci à mes amis de T. et de Stan.

Merci infiniment à mes parents de m'avoir toujours soutenue. Vous êtes pour moi des modèles de courage, d'intelligence et de finesse. Merci à mes soeurs, Laure, Marie, Claire, chacune si intéressante, amusante et présente à sa manière. Merci à Chantal, que j'embrasse affectueusement. Merci au reste de ma famille de m'avoir encouragée quand j'en avais besoin, en particulier à Manu, pour son expérience en recherche. Merci à Christian, si pressé de m'appeler "Docteur". Merci à Adeline de m'avoir accueillie et soutenue à chaque instant, sans qui je n'aurais pas pu y arriver.

"Dieu n'a créé que des énigmes." Dostoïevski, *Les Frères Karamazov*.

"I shall be telling this with a sigh
Somewhere ages and ages hence:
Two roads diverged in a wood, and I-
I took the one less traveled by,
And that has made all the difference."

Robert Frost, *The Road Not Taken*.

Contents

1 Introduction **1**

1 Clustering in Bipartite Stochastic Block Model 1
- 1.1 Erdös-Renyi model 1
- 1.2 Stochastic Block Model 2
- 1.3 Bipartite Stochastic Block Model 4

2 Estimation of topic-document matrix in topic model 8
- 2.1 Probabilistic Latent Semantic Indexing 8
- 2.2 Constraints 10
- 2.3 Overview of previous results 10
- 2.4 SPOC algorithm 12
- 2.5 Summary of the results 12

3 Benign overfitting in nonparametric regression 13
- 3.1 State of the art 13
- 3.2 Local polynomial estimator 15
- 3.3 Summary of the results 16

2 Introduction en français **18**

1 BSBM 18
- 1.1 Modèle d'Erdös-Rényi 18
- 1.2 Le Stochastic Block Model 19
- 1.3 Le Bipartite Stochastic Block Model 21

2 Topic Model 25
- 2.1 Probabilistic Latent Semantic Indexing 26
- 2.2 Contraintes 27
- 2.3 Vue d'ensemble des résultats précédents 28
- 2.4 Algorithme SPOC 29
- 2.5 Résultats 30

3 Benign Overfitting 30
- 3.1 État de l'art 31
- 3.2 Estimateur par polyômes locaux 33
- 3.3 Résultats 34

3 Improved clustering algorithms for the Bipartite Stochastic Block Model **35**
1 Introduction . . . 36
1.1 Definition of Bipartite Stochastic Block Model . . . 37
1.2 Recovery of communities . . . 37
2 Reduction to a spiked model . . . 39
3 Related work . . . 41
4 Main contributions . . . 44
5 Properties of the spectral method . . . 47
6 Exact recovery by the hollowed Lloyd's algorithm . . . 50
7 Impossibility result for a supervised oracle . . . 50
8 Gram matrix study . . . 51
9 Lower bound on the oracle . . . 56
10 Main proofs . . . 59
10.1 Proof of Theorem 1 . . . 59
10.2 Proof of Theorem 2 . . . 63
11 Numerical experiments . . . 68

4 Assigning topics to documents by successive projections **71**
1 Introduction . . . 72
2 Successive Projection Overlapping Clustering . . . 77
3 Main results . . . 79
3.1 Deterministic bounds . . . 79
3.2 Bounds with high probability . . . 80
3.3 Adaptive procedure when K is unknown. . . . 82
3.4 Minimax lower bound . . . 83
4 Related Work . . . 84
5 Numerical experiments . . . 85
5.1 Synthetic Data . . . 85
5.2 Corpus of NIPS abstracts . . . 89
6 Conclusion . . . 90
7 Tools . . . 91
7.1 Matrix Perturbation Bounds . . . 91
7.2 Noisy Separable Matrix Factorization . . . 93
7.3 Concentration Bounds for Multinomial Matrices . . . 94
8 Proofs of the Main Results . . . 97
8.1 Proof of Lemma 1 . . . 97
8.2 Proof of Lemma 2 . . . 98
8.3 Proof of Theorem 1 . . . 99
8.4 Proof of Theorem 2 and Corollary 3 . . . 100
8.5 Proof of Theorem 3 . . . 100
9 Auxiliary lemmas . . . 107
9.1 The anchor document assumption under the Dirichlet prior . . . 109
10 Additional Experiments: Estimation of topic-word matrix . . . 110

11 Additional Experiments: Empirical study of singular values of word-document and topic-document matrices . . . 111
12 Additional Experiments: Estimation for the $p = 2000$ . . . 116

5 Benign overfitting and adaptive nonparametric regression 118
1 Introduction . . . 118
2 Preliminaries . . . 120
2.1 Notation . . . 120
2.2 Model . . . 121
2.3 Hölder classes of functions . . . 121
3 Local polynomial estimators and interpolation . . . 123
4 Minimax optimal interpolating estimator . . . 125
5 Adaptive interpolating estimator . . . 127
6 Numerical experiment . . . 128
7 Proofs . . . 133
8 Conclusion . . . 147

6 Conclusion 148

Chapter 1

Introduction

In this thesis, we have studied different statistical problems: clustering in a bipartite graph, estimation in topic models and benign overfitting in nonparametric framework. We present these three problems in more detail in the following sections. Notation may change from a section to another. More detail about each problem are provided in the following chapters.

1 Clustering in Bipartite Stochastic Block Model

The first problem we considered is the clustering problem in a graph. It is considered in Chapter 3. For clustering problems, one can generally consider two approaches. Either the observations are composed of individuals/objects without interaction and we choose to model them by a mixture model. Either the observations are composed of individuals/objects with interactions and we choose to model them by a graph model. Graph models have applications in many disciplines and allow, for example, the study of social, biological and computer interactions. An essential model in statistics is the Stochastic Block Model (SBM). It is a suitable model in community detection.

1.1 Erdös-Renyi model

Before presenting the SBM, we first introduce the fundamental Erdös-Rényi model (ER) [Erdős and Rényi, 1959, Erdős et al., 1960]. Let n be an integer and $p \in (0, 1)$. A graph $G(n, p)$ generated according to the ER model is a non-oriented graph with n vertices that are randomly connected. The probability that two vertices are connected is p, independently of the other vertices. Although this model is very simple and often not realistic in applications, many statistical problems have been developed from this model. It has also allowed progress in the study of more complex graph models. In particular, the ER model is not a suitable model for community detection, since the probability of interaction of two vertices is homogeneous in the whole graph.

This model allows one to introduce the concept of phase transition. A phase transition phenomenon occurs when a threshold phenomenon is observed. The two parameters of the ER model are n and p. According to their relative values, the graph looks different. Thus, [Erdős et al., 1960] proved that for any $\varepsilon \in (0,1)$,

- if $p < \frac{(1-\varepsilon)\log n}{n}$, then $G(n,p)$ almost surely contains isolated vertices (i.e., is a disconnected graph),
- if $p > \frac{(1+\varepsilon)\log n}{n}$, then $G(n,p)$ is almost surely connected (i.e., does not contain any isolated vertex).

Thus, for fixed n, a small variation of p radically modifies the appearance of the graph and $\frac{\log n}{n}$ can be considered as a connectivity threshold of the graph.

1.2 Stochastic Block Model

The Stochastic Block Model [Holland et al., 1983] can be considered as an extension of the ER model. The main assumption of the SBM is that the vertices are not connected randomly but according to their respective community.

Let us consider the framework of the SBM with two communities. Let n be an integer and $(p,q) \in (0,1)^2$. A graph $G(n,p,q)$ generated according to a two-communities SBM model is a non-oriented graph with n vertices, such that the probability that two vertices belonging to the same community are connected is p, and the probability that two vertices belonging to different communities are connected is q.

We now give a formal definition of the SBM. Let n_+, n_- be two positive integers such that $n = n_+ + n_-$. Let V be the set of n vertices such that V contains n_+ vertices with label $+1$ and n_- vertices with label -1. For each vertex u of V, we denote by $\sigma(u) \in \{-1,1\}$ its label. We denote by A the adjacency matrix of the graph, i.e. the (n,n)-matrix such that its entries A_{ij} equal 1 if the corresponding vertices $i,j \in V$ are connected, and 0 otherwise.
We say that A is generated according to a $SBM(n_+,n_-,p,q)$ model if the entries A_{ij} are independent and if

- $A_{ij} \sim Ber(p)$ if $\sigma(i) = \sigma(j)$, i.e., two vertices with the same labels are connected with probability p,
- $A_{ij} \sim Ber(q)$ if $\sigma(i) \neq \sigma(j)$, i.e., two vertices with different labels are connected with probability q.

There, $Ber(p)$ denotes the Bernoulli distribution with parameter p.

We note that if $p = q$, we obtain the Erdös-Rényi model. If $p > q$, the SBM is called assortative and interactions are more frequent within a community than between communities. This is the most common situation in applications. Conversely, if $p < q$, the SBM is called disassortative and interactions are less frequent within

a community than between communities. Examples of disassortative graphs can be found in biology or in the link architecture of web pages.

We denote by $\eta \in \{\pm 1\}^n$ the vector of the labels of the vertices of V. Note that if the task is to classify the vertices in two communities, it can be achieved either by estimating η or by estimating $-\eta$.

Any measurable function $\hat{\eta}$ from A to $\{\pm 1\}^n$ is an estimator of η. In order to measure the loss of such an estimtor, we introduce the Hamming distance, which equals twice the number of coordinates where η and $\hat{\eta}$ differ:

$$\|\eta - \hat{\eta}\|_1 := \sum_{i=1}^{n} |\eta_i - \hat{\eta}_i| = 2 \sum_{i=1}^{n} \mathbf{1}(\eta_i \neq \hat{\eta}_i),$$

where η_i (resp., $\hat{\eta}_i$) designates the i^{th} coordinate of η (respectively, $\hat{\eta}$).

Since as mentioned above, it is equivalent for community detection to estimate η and $-\eta$, we consider the following loss

$$r(\eta, \hat{\eta}) = \min_{\nu \in \{-1,1\}} \|\hat{\eta} - \nu\eta\|_1. \tag{1.1}$$

There are various properties of interest in the study of SBM.

Definition 1 (*weak recovery* in SBM)**.** *The estimator $\hat{\eta}$ achieves weak recovery of η if there exists $\alpha \in (0,1)$ such that*

$$\lim_{n\to\infty} \sup_{SBM} \mathbf{P}\left(\frac{r(\eta, \hat{\eta})}{n} \geq \alpha\right) = 0, \tag{1.2}$$

where $\sup_{SBM}$ denotes the maximum over all distributions of A drawn from $SBM(n_+, n_-, p, q)$.

Weak recovery is also called *detection* in the literature.

Definition 2 (*almost full recovery* in SBM)**.** *The estimator $\hat{\eta}$ achieves almost full recovery of η if (2.2) holds for all $\alpha \in (0,1)$.*

Almost full recovery, also called *almost exact recovery* in the literature, means that $\hat{\eta}$ correctly classifies almost every vertex with high probability.

Definition 3 (*exact recovery* in SBM)**.** *The estimator $\hat{\eta}$ achieves exact recovery of η if*

$$\lim_{n\to\infty} \inf_{SBM} \mathbf{P}\Big(r(\eta, \hat{\eta}) = 0\Big) = 1.$$

Exact recovery means that $\hat{\eta}$ correctly classifies all the vertices with high probability.

A great deal of work was devoted to determine the phase transitions on n, p, q for these problems. Most results were obtained under the specific assumptions $a = pn$ and $b = qn$, $\alpha = pn/\log(n)$ and $\beta = qn/\log(n)$, that characterize the most interesting case. In particular, for the problem of *weak recovery*, [Massoulié, 2014] and [Mossel et al., 2018] proved that *weak recovery* is possible if and only if $(a-b)^2 > 2(a+b)$. For the problem of *exact recovery* when $p > q$, [Abbe et al., 2015] have proved the following phase transition phenomenon: *exact recovery* is possible if $\frac{\alpha+\beta}{2} - \sqrt{\alpha\beta} > 1$ and is impossible if $\frac{\alpha+\beta}{2} - \sqrt{\alpha\beta} < 1$.

1.3 Bipartite Stochastic Block Model

Many generalizations of the Stochastic Block Model have been considered. In this thesis, we are particularly interested in a non-symmetric generalization of the SBM, which is the Bipartite Stochastic Block Model (BSBM).

Introduced by [Feldman et al., 2015], the BSBM is adapted to the study of interactions between two sets, where each set is divided into several communities, and these two sets are most often composed of objects of different natures. Within a set, there may be inter- and intra-communities interactions, but most often these are inaccessible or uninformative. The BSBM is relevant for example for the study of object/user interactions within the framework of recommendation systems. The users constitute a first set, divided into several communities, and the objects constitute a second set, also divided into several communities. This model has several applications such as document/word interactions [Dhillon and Modha, 2001, Lancichinetti et al., 2014], gene/genetic sequences interactions [Eren et al., 2013, Larremore et al., 2013], and object/user interactions in the context of recommendation systems [Jang et al., 2007], and other.

Initially, the BSBM was introduced by [Feldman et al., 2015] in the context of Constraint Satisfaction Problems (CSP). CSPs are mathematical problems that consist in the study of states or objects satisfying certain criteria or constraints. Formally, a CSP is a triplet $(\mathscr{X}, \mathscr{D}, \mathscr{C})$ where $\mathscr{X}$ is a set of n variables, $\mathscr{D}$ is a set of n domains of values for each variable of $\mathscr{X}$, and $\mathscr{C}$ is a set of constraints. CSPs arise in many fields, for example in computer science and machine learning. In particular, the theory of CSPs is closely related to complexity theory in theoretical computer science. [Feldman et al., 2015] introduced the BSBM to unify several problems including the classical SBM problem and the random CSP k-SAT problem. The SAT problem, or Boolean satisfiability problem, is the CSP problem which, given a propositional logic formula, determines whether there is an assignment of propositional variables that makes the formula true. If such an assignment exists, the formula is said to be satisfiable. Recall that in Boolean logic, a literal is a Boolean variable or its negation, and a clause is a disjunction of literals. Clauses relate to constraints. A random k-SAT formula is a conjunction of m clauses of k Boolean variables randomly chosen from n Boolean variables. We are interested in the probability that a random k-SAT formula is satisfiable. [Feldman et al., 2015] have studied the problem of *planted* satisfiability. In this framework, a so-called

planted assignment is fixed in advance and the (random) clauses are drawn according to a distribution defined by this assignment. [Feldman et al., 2015] proved that the random planted k-SAT problem reduces to the BSBM problem.

Definition of BSBM

Let $n_{1+}, n_{1-}, n_{2+}, n_{2-}$ be four non-zero positive integers such that $n_1 := n_{1+} + n_{1-} \leq n_{2+} + n_{2-} := n_2$ and let $p \in (0, 1/2)$, $\delta \in (0, 2)$. Let V_1 and V_2 be two sets of vertices such that V_1 (respectively V_2) is composed of n_{1+} (resp. n_{2+}) vertices with label $+1$ and of n_{1-} (resp. n_{2-}) vertices with label -1. For each vertex u of V_1 or V_2, we denote by $\sigma(u) \in \{-1, 1\}$ its label.

We denote by A the biadjacency matrix, i.e., the (n_1, n_2)-matrix such that its entries A_{ij} equal 1 if the corresponding vertices $i \in V_1, j \in V_2$ are connected, and 0 otherwise.

Then, we say that A is generated according to a $BSBM(\delta, n_{1+}, n_{1-}, n_{2+}, n_{2-}, p)$ model if the entries A_{ij} are independent and if

- $A_{ij} \sim Ber(\delta p)$ if $\sigma(i) = \sigma(j)$ i.e. two vertices $i \in V_1$ et $j \in V_2$ with the same label are connected with probability δp,
- $A_{ij} \sim Ber((2 - \delta)p)$ if $\sigma(i) \neq \sigma(j)$ i.e. two vertices $i \in V_1$ et $j \in V_2$ with different labels are connected with probability $(2 - \delta)p$.

Note that the BSBM model is a generalization of the SBM model since if $V_1 = V_2$, we obtain the SBM.

Community detection

Suppose that we observe a matrix A generated according to a model $BSBM(\delta, n_{1+}, n_{1-}, n_{2+}, n_{2-}, p)$. We are interested in the problem of estimating the partition associated to V_1 from the observation of the biadjacency matrix A. Let $\eta_1 \in \{\pm 1\}^{n_1}$ be the vector of labels of the vertices of V_1. As in the SBM setting, it is equivalent to estimate η_1 and $-\eta_1$. Any measurable function $\hat{\eta}$ from A to $\{\pm 1\}^{n_1}$ is an estimator of η_1.

As for the SBM problem, we consider the problems of *weak recovery*, *almost full recovery* and *exact recovery*. We may easily adapt Definitions 4,5,6 to the BSBM setting. We use the Hamming loss r defined in (2.1) to characterise the loss of an estimator $\hat{\eta}$ of η_1.

Overview of previous results

If the SBM has been extensively studied, the BSBM remains less known. [Feldman et al., 2015, Florescu and Perkins, 2016, Cai et al., 2019] studied the phase transition phenomena for p. [Florescu and Perkins, 2016] proved that the phase transition

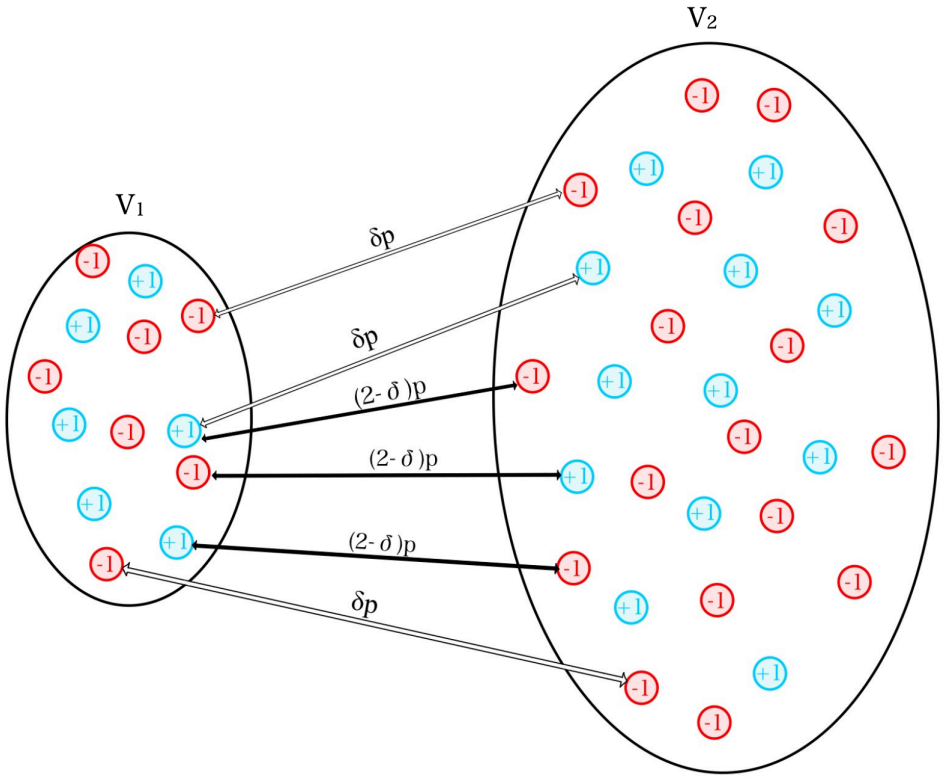


Figure 1.1: Illustration of BSBM.

phenomenon for *weak recovery* occurs for the critical probability $p_c = \frac{1}{(1-\delta)^2\sqrt{n_1 n_2}}$. In order to prove the sufficient condition, they used a reduction to the SBM and then an optimal "black-box" algorithm for the *weak recovery* in the SBM, as in [Bordenave et al., 2015, Massoulié, 2014, Mossel et al., 2018].

[Florescu and Perkins, 2016] also provided a sufficient condition for *almost full recovery* in the high dimensional framework i.e. for $n_2 \gg n_1$. [Florescu and Perkins, 2016] consider spectral methods, which are classical methods in the community estimation framework. In particular, [Florescu and Perkins, 2016] proved that modifying the classical SVD method allows a significant improvement in the high dimensional framework. Instead of considering the singular vectors of the biadjacency matrix A, [Florescu and Perkins, 2016] consider the eigenvectors of the associated Gram matrix, whose diagonal elements are all set to 0. The associated algorithm is called "Diagonal Deletion SVD". For $n_2 \geq n_1(\log n_1)^4$, [Florescu and Perkins, 2016] have proved with the "Diagonal Deletion SVD" that $p = \Omega\left(\frac{\log n_1}{\sqrt{n_1 n_2}}\right)$ is a sufficient condition to obtain *almost full recovery* in the BSBM model. Here and in what follows, we write $a_n = O(b_n)$ if there is a constant $c > 0$ such that $a_n \leq cb_n$, and we write $a_n = \Omega(b_n)$ if there is a constant $c > 0$ such that $a_n \geq cb_n$. We also write $a_n \asymp b_n$ if $a_n = O(b_n)$ and $a_n = \Omega(b_n)$.

[Feldman et al., 2015] have proved the best known conditions for *exact recovery* until our results. They proved in the high-dimensional setting $n_2 \geq n_1$ that the condition $p = \Omega\left(\frac{(1-\delta)^2 \log n_1}{\sqrt{n_1 n_2}}\right)$ is sufficient to achieve *exact recovery* using their algorithm named "Subsampled Power Iteration". [Feldman et al., 2015] conjecture that the condition $p = \Omega(\frac{\log n_1}{\sqrt{n_1 n_2}})$ is necessary for the *exact recovery*. We have disproved this conjecture as detailed below.

We notice that under very similar conditions, [Feldman et al., 2015] provide better results than [Florescu and Perkins, 2016], since the "Subsampled Power Iteration" algorithm allows *exact recovery* instead of *almost full recovery*.

Note that [Cai et al., 2019] have proved that the "Diagonal Deletion SVD" algorithm also allows *exact recovery* under conditions very close to [Feldman et al., 2015] for $n_2 \gtrsim n_1$, but these conditions degrade in the $n_2 \gg n_1$ frame, for example for $n_2 \asymp e^{n_1}$. To be more specific, [Cai et al., 2019] prove that it is possible to achieve *exact recovery* via Diagonal Deletion SVD if

$$p \geq C(1-\delta)^2 \left(\frac{\log(n_1+n_2)}{\sqrt{n_1 n_2}} \vee \frac{\log(n_1+n_2)}{n_1+n_2} \right), \tag{1.3}$$

where $C > 0$ is a constant independent of p, δ, n_1, n_2. [Zhou and Amini, 2019, Zhou and Amini, 2020, Neumann, 2018] have also provided results on the BSBM but in a more general framework. These results do not allow one to estimate the communities in the high dimensional framework when $n_2 \gg n_1$. [Zhou and Amini, 2019] use spectral methods on a well chosen matrix. Theorem 1 of [Zhou and Amini, 2020] adapted to our framework requires $n_2 = O(n_1^2)$, which is quite restrictive. [Neumann, 2018] focuses on the estimation of so-called tiny clusters, i.e., clusters of size n^ε for any $\varepsilon > 0$, where n is the number of vertices of the graph. However, the work of [Neumann, 2018] applied to our setting is suboptimal.

[Feldman et al., 2015, Florescu and Perkins, 2016] observed that the main difficulties of estimating communities arise in the high-dimensional regime $n_2 \gg n_1$. This regime is most adapted to real applications of the BSBM and it is the regime that presents theoretical difficulties. Indeed, the $n_2 \leq n_1$ regime is more standard, as we can apply classical SBM clustering methods to it. We obtain optimal results for $n_2 \leq n_1$ by applying SVD algorithms directly to the biadjacency matrix [Zhou and Amini, 2019, Zhou and Amini, 2020]. This is because we know how to control the spectral norm of the noise of the biadjacency matrix, but less so of the noise of the Gram matrix with a null diagonal, which we have to control for $n_2 \gg n_1$.

To obtain our results, we made an analogy with the Gaussian Mixture Model (abbreviated GMM). [Lu and Zhou, 2016, Ndaoud, 2018, Giraud and Verzelen, 2019, Loffler et al., 2019] have developed optimal clustering algorithms for the GMM. [Lu and Zhou, 2016] have proved that an iterative procedure similar to Lloyd's algorithm [Lloyd, 1982] allows clustering with optimal properties for GMM. [Ndaoud, 2018] provided a modified version of such an iterative clustering algorithm which has been shown to achieve the phase transition for *exact recovery* in this model. Thus, by analogy between GMM and BSBM, [Ndaoud, 2018] conjectured that similar algo-

rithms would allow *almost full recovery* and *exact recovery* for BSBM, and that the condition $p = \Omega\left((1-\delta)^2\sqrt{\frac{\log n_1}{n_1 n_2}}\right)$ is sufficient to obtain the *exact recovery* for the BSBM if $n_2 \geq n_1 \log n_1$. This goes against the heuristics made by analogy with the SBM that $p = \Omega\left(\frac{\log n_1}{\sqrt{n_1 n_2}}\right)$ is necessary for *exact recovery*, cf. [Feldman et al., 2015].

Summary of the results

In Chapter 3, we introduce an algorithm called the *hollowed Lloyd's algorithm*, which allows *exact recovery* under better conditions than those known in the previous work [Ndaoud et al., 2021]. We prove that, for all regimes of (n_1, n_2),

$$p \geq C(1-\delta)^2\left(\sqrt{\frac{\log n_1}{n_1 n_2}} \vee \frac{\log n_1}{n_2}\right) \tag{1.4}$$

is a sufficient condition for the *exact recovery* for the BSBM, with $C > 0$ a positive constant. This condition is better than (2.3). In particular, it does not degrade for $n_2 \asymp e^{n_1}$.

The condition (2.4) reveals an elbow at $n_2 \asymp n_1 \log n_1$ between the high and low dimension regimes. In the low-dimensional regime $n_2 \leq n_1 \log n_1$, we find the sufficient condition of [Zhou and Amini, 2019, Cai et al., 2019], $p = \Omega\left(\frac{\log n_1}{n_2}\right)$. In the high dimensional regime $n_2 \geq n_1 \log n_1$, we obtain the sufficient condition $p = \Omega\left(\sqrt{\frac{\log n_1}{n_1 n_2}}\right)$. We conjectured in [Ndaoud et al., 2021] that this condition is necessary in the regime $n_2 \geq n_1 \log n_1$ by exhibiting an oracle estimator that does not allow *exact recovery* if $p < c\sqrt{\frac{\log n_1}{n_1 n_2}}$ for $c > 0$ small enough.

Our results can also be read in[Ndaoud et al., 2021].

2 Estimation of topic-document matrix in topic model

The second problem we studied is the topic model problem. It is considered in Chapter 4.

In Statistics and Natural Language Processing (NLP), a topic model is a useful model for classifying documents by topics. For instance, it is necessary to know how to classify web pages in order to recommend them to users. It is also useful to be able to automatically classify scientific articles online. Topic models can be used to address these issues.

2.1 Probabilistic Latent Semantic Indexing

In this thesis, we consider the probabilistic Latent Semantix Indexing (pLSI) model. Introduced by [Hofmann, 1999], this model links three types of variables: documents, topics and words. We assume that we have a dictionary of p words and a corpus of

n documents, where p, n are non-zero natural numbers. A document is a sequence of words from the dictionary. We assume that the documents can be classified according to their topic. Let $K \in \mathbf{N}^*$ be the number of topics. We suppose $2 \leq K \leq \min(n, p)$. We typically have K very small compared to n and p.

The fundamental assumption of the pLSI model is that the probability of a word j appearing in a document on topic k is independent of the document. In other words, by the law of total probability,

$$\mathbf{P}(\text{word } j|\text{document } i) = \sum_{k=1}^{K} \mathbf{P}(\text{topic } k|\text{document } i)\mathbf{P}(\text{word } j|\text{topic } k).$$

We introduce the following notation

$$\begin{aligned} \Pi_{ij} &= \mathbf{P}(\text{word } j|\text{document } i), \\ W_{ik} &= \mathbf{P}(\text{topic } k|\text{document } i), \\ A_{kj} &= \mathbf{P}(\text{word } j|\text{topic } k). \end{aligned}$$

Then, we can write $\Pi_{ij} = W_i^{\mathrm{T}} A_j$, where $W_i = (W_{i1}, \ldots, W_{iK})^{\mathrm{T}} \in [0,1]^K$ is the topic probability vector of document i and $A_j = (A_{1j}, \ldots, A_{Kj})^{\mathrm{T}} \in [0,1]^K$ is the topic probabilityvector of word j, for each topic $k = 1, \ldots, K$. Thus, we can write in matrix form

$$\boldsymbol{\Pi} = \boldsymbol{W}\boldsymbol{A},$$

where $\boldsymbol{\Pi}$ is the (n, p)-document-word matrix with entries Π_{ij}, $\boldsymbol{W} := (W_1, \ldots, W_n)^{\mathrm{T}}$ is the (n, K)-document-topic matrix, $\boldsymbol{A} := (A_1, \ldots, A_p)$ is the (K, p)-topic-word matrix. The rows of these matrices being probability vectors, we have

$$\sum_{m=1}^{K} W_{im} = 1, \ \sum_{j=1}^{p} A_{kj} = 1, \ \sum_{j=1}^{p} \Pi_{ij} = 1 \text{ for all } i = 1, \ldots, n, \ k = 1, \ldots K.$$

Here, Π_{ij} is the probability of occurrence of word j in document i. In practice, we do not know Π_{ij} but we know the corresponding empirical frequency X_{ij}.

Thus, we have a (n, p)-document-word matrix $\boldsymbol{X} = (X_{ij})$ such that for each document i in $1, \ldots, n$, and each word j in $1, \ldots, p$, X_{ij} is the observed frequency of the word j in the document i. Let N_i denote the (deterministic) number of words sampled in document i. We assume that, for each document-word vector $X_i = (X_{i1}, \ldots, X_{ip})^{\top}$, the corresponding vector $N_i X_i$ of the number of occurrences of each word in the document i follows a multinomial distribution of dimension p, of parameters (N_i, Π_i), where $\Pi_i = \mathbf{E}(X_i) = (\Pi_{i1}, \ldots, \Pi_{ip})^{\top}$. We also assume that $X_1, \ldots, X_n$ are independent random vectors. We can write the model as a "signal + noise"-model:

$$\boldsymbol{X} = \boldsymbol{\Pi} + \boldsymbol{Z} = \boldsymbol{W}\boldsymbol{A} + \boldsymbol{Z}, \tag{1.5}$$

where $\boldsymbol{Z} := \boldsymbol{X} - \boldsymbol{\Pi}$ is a zero mean matrix.

The objective of topic modelling is to estimate the $\boldsymbol{A}$ and $\boldsymbol{W}$ matrices from the observation of the $\boldsymbol{X}$ matrix and the knowledge of $N_1, \ldots, N_n$. The estimation of $\boldsymbol{A}$ and the estimation of $\boldsymbol{W}$ answer different objectives. An estimator of the $\boldsymbol{A}$ matrix identifies the distribution of the topics on the dictionary. An estimator of the $\boldsymbol{W}$ matrix identifies the topics associated with each document. Our study is mainly focused on the estimation of the $\boldsymbol{W}$ matrix.

2.2 Constraints

It is standard in the context of topic models to introduce anchoring hypotheses. One widely used hypothesis in the literature is the *anchor word assumption*, cf., for example [Arora et al., 2013, Bing et al., 2020a, Ke and Wang, 2017]. This *anchor word assumption* consists in assuming that for each topic, there is at least one word associated only to this topic. This assumption is best suited to estimation of the $\boldsymbol{A}$ matrix, which has been extensively studied in the literature. Another assumption that we adopt below is the *anchor document assumption*, which postulates that for each topic, there is at least one document dealing exclusively with this topic.

A fundamental idea of high-dimensional statistics is that a large collection of data can be of very low rank. Thus, by reducing the dimension, it is possible to better exploit the data. This is the case for topic models, where there is an underlying topic structure of dimension K that is typically very small in comparison to n and p. There are several dimensionality reduction methods. We are particularly interested in matrix factorisation methods.

2.3 Overview of previous results

Non-negative matrix factorization (NMF) with no noise

We first place ourselves in a noiseless setting, where equation (2.5) becomes

$$\boldsymbol{X} = \boldsymbol{W}\boldsymbol{A}.$$

Non-negative Matrix Factorisation (NMF) methods aim to recover $\boldsymbol{W}$ and $\boldsymbol{A}$ from observing $\boldsymbol{X}$. A non-negative matrix is a matrix whose entries are non-negative. The NMF method was introduced by [Paatero and Tapper, 1994, Paatero, 1997, Lee and Seung, 1999a, Lee and Seung, 2000] and consists in factoring under suitable assumptions a non-negative matrix $\boldsymbol{X}$ into a product of two non-negative matrices $\boldsymbol{W}$, $\boldsymbol{A}$. In the case where $\boldsymbol{X}$ is a (n, p) matrix, $\boldsymbol{W}$ a (n, K) matrix and $\boldsymbol{A}$ a (K, p) matrix, this factorisation allows a dimension reduction. This method is useful in certain settings where the matrices studied are intrinsically non-negative, such as in NLP or image analysis. Classical methods like Principal Component Analysis (PCA) cannot be applied to such matrices. Indeed, the orthogonality constraint required in PCA cannot be respected. This is precisely the framework of the pLSI

model, where all matrices under consideration are non-negative. In general, the NMF problem is an NP-hard problem, but separability constraints such as the *anchor document assumption* allow for its resolution. These methods consist most often in the minimisation of a regularised cost function, cf. for example [Cichocki et al., 2009, Donoho and Stodden, 2004, Lee and Seung, 1999a, Recht et al., 2012]. These papers deal with non-noisy case and their result cannot be used under the noisy model in (2.5).

Bayesian perspective: Latent Dirichlet Allocation

We now turn to NMF in the noisy case specific to topic models. There is an abundant literature on estimation of matrices $\boldsymbol{A}, \boldsymbol{W}$ in this context. Probably the most famous and popular model in text classification is the Latent Dirichlet Allocation or LDA model. Introduced by [Blei et al., 2003], the LDA model imposes a Dirichlet prior on the $\boldsymbol{A}$ matrix and then estimates the $\boldsymbol{W}$ matrix by an EM algorithm. This line of work is mainly interested in the construction of algorithms and does not provide statistical guarantees for the obtained estimators. Moreover, the LDA algorithm is computationally slow and assumes that the subjects are uncorrelated, which is not realistic [Blei and Lafferty, 2007, Li and McCallum, 2006]. To face this last issue, [Lafferty and Blei, 2006] introduced another model, the Correlated Topic Model. Other related papers considered Gibbs-sampling [Porteous et al., 2008, Ramage et al., 2009] or variational Bayes techniques [Chien and Chueh, 2010, Zhai et al., 2012] to estimate $\boldsymbol{W}$, rather than using the EM algorithm.

Prior work on statistical guarantees for topic models

To estimate matrix $\boldsymbol{A}$, several papers have proposed algorithms with statistical guarantees under the *anchor word assumption*, cf. for example [Arora et al., 2012, Arora et al., 2013, Ding et al., 2013, Anandkumar et al., 2014, Bansal et al., 2014, Bing et al., 2020a, Bing et al., 2020c, Bing et al., 2021a, Ke and Wang, 2017]. These papers use different techniques including co-occurrence matrix analysis, tensors, or simplex methods via SVD. The work of [Bing et al., 2020a, Bing et al., 2020c, Bing et al., 2021a, Ke and Wang, 2017] develops methods attaining up to log-factors minimax optimal rates of convergence for estimator of $\boldsymbol{A}$ in ℓ_1-norm. These papers use the *anchor word assumption* but propose different estimators than ours, as they focus on estimation of matrix $\boldsymbol{A}$. The estimation method of matrix $\boldsymbol{A}$ in [Ke and Wang, 2017] is a simplex method in dimension p, with a computational cost p^K. Their procedure also allows for estimation of matrix $\boldsymbol{W}$ using their estimator $\hat{\boldsymbol{A}}$ of $\boldsymbol{A}$, via a least-squares method.

[Bing et al., 2020a] propose a fast SVD-based method for the estimation of $\boldsymbol{A}$ under the *anchor word assumption*. In follow-up works, [Bing et al., 2020c, Bing et al., 2021a] consider problem (2.5) in a sparse setting and under the *anchor word assumption*. [Bing et al., 2021a] study the estimation of the rows of matrix $\boldsymbol{W}$,

provided an estimator $\hat{\boldsymbol{A}}$ of $\boldsymbol{A}$. Denote by $(w_i)_{i=1}^n$ the rows of matrix $\boldsymbol{W}$ and by $(\hat{w}_i^{MLE})_{i=1}^n$ the corresponding maximum-likelihood estimators of [Bing et al., 2021a]. Assume that $N_i = N$ for all i. Let $1 \leq i \leq n$ be fixed. [Bing et al., 2021a] prove, for the ℓ_1-norm of ith row $\|\hat{w}_i^{MLE} - w_i\|_1$, a bound in probability of the order $\left(\sqrt{\frac{p}{nN}} \vee \frac{1}{\sqrt{N}}\right)$ (to within a weak factor, that is, a factor depending only on K or sparsity, and a logarithm of the main parameters p, n, N). Consequently, the rate of estimating $\boldsymbol{W}$ in the ℓ_1-norm obtained in [Bing et al., 2021a] scales as $\left(\sqrt{\frac{np}{N}} \vee \frac{n}{\sqrt{N}}\right)$, up to a weak factor. An analogous result is obtained in the 2022 version of the paper [Ke and Wang, 2017] for a different estimator, again under the *anchor word assumption*.

2.4 SPOC algorithm

In order to estimate $\boldsymbol{W}$, we have introduced an algorithm called Successive Projection Overlapping Clustering (SPOC), inspired by the Successive Projection Algorithm (SPA). The SPA algorithm was introduced by [Araújo et al., 2001] to solve the NMF problem and was widely used thereafter, due to its simplicity and speed. The SPOC algorithm applies the SPA algorithm not to the initial matrix $\boldsymbol{X}$, but rather to the matrix of its left singular vectors and uses the result to recover $\boldsymbol{W}$.

The SPA and SPOC algorithms are iterative projection algorithms. Starting from the SVD decomposition of the $\boldsymbol{X}$ matrix, we apply the following iterative procedure. At each step of the algorithm, we select the row with maximum norm of the matrix of left singular vectors, and we project our matrix onto the orthogonal complement of the space generated by this row. Geometrically, this procedure is explained by the fact that the rows of the matrix of left singular vectors of $\boldsymbol{\Pi}$ belong to a simplex whose vertices are *anchor documents*. The SPOC algorithm iteratively finds estimators of these vertices, which then allows the estimation of $\boldsymbol{W}$.

2.5 Summary of the results

In Chapter 4, we prove bounds on the accuracy of the SPOC algorithm in the Frobenius norm and the ℓ_1-norm [Klopp et al., 2021]. We obtain upper and lower bounds that match within a logarithmic factor, implying the quasi-optimality of our procedure. Specifically, assuming that $N_i = N$ for all i, we prove that the SPOC estimator of $\boldsymbol{W}$ converges in the Frobenius norm and in the ℓ_1-norm with the rates $\sqrt{n/N}$ and $n/\sqrt{N}$ respectively, up to a weak factor. We also prove minimax lower bounds of the order $\sqrt{n/N}$ and $n/\sqrt{N}$ respectively. We can observe that the rate in ℓ_1 is faster than $\left(\sqrt{\frac{np}{N}} \vee \frac{n}{\sqrt{N}}\right)$ following from [Bing et al., 2021a]. However, the assumptions in [Bing et al., 2021a] are different. They impose the *anchor word assumption* while we are working under the *anchor document assumption*. The potentially big term $\sqrt{\frac{np}{N}}$ in the rate appears in [Bing et al., 2021a] because of the use of preliminary

estimation of $\boldsymbol{A}$. It accounts for the rate of estimation of $\boldsymbol{A}$, which is an artefact of the method and is not reflected in the lower bound.

We also notice from our theoretical results and our simulations that the error of the SPOC algorithm does not increase significantly with p, unlike the error of the LDA algorithm. Moreover, our procedure is computationally fast and simple to implement. Compared to [Ke and Wang, 2017], whose algorithm has complexity p^K, our algorithm has complexity $\max(p,n)K + nK^2$. The first term $\max(p,n)K$ corresponds to the cost of truncated SVD, and the second term nK^2 correponds to the cost of SPA. Finally, our procedure is adaptive as it does not need to know the number of topics K in advance.

Our results can also be read in [Klopp et al., 2021].

3 Benign overfitting in nonparametric regression

The third problem we studied is the problem of benign overfitting in the nonparametric setting. It is considered in Chapter 5.

3.1 State of the art

Modern machine learning methods have shown some unexpected and seemingly contradictory properties to classical statistics. One of the most popular learning methods today is the neural network method. Initially inspired by the functioning of the human brain, neural networks mimic the transmission of a signal from one neuron to another. The theoretical functioning of neural networks is still not well understood. Empirically, the excellent performance of deep neural networks is no longer to prove (cf. for example [Gupta et al., 2015, Sagun et al., 2017, Huang et al., 2017]). However, deep neural networks have the property of being interpolative, i.e., with zero bias on the training data set, but still has a small prediction error on unknown datasets [Belkin et al., 2019a, Zhang et al., 2021]. This phenomenon goes against the classical bias-variance trade-off intuition, which has thus been questioned for some years (cf. for example [Ma et al., 2018]).

Recent work in benign overfitting in linear regression

In order to better understand this phenomenon, called benign overfitting, several papers have studied this problem in the context of linear regression, cf. for example [Bartlett et al., 2020, Tsigler and Bartlett, 2020, Chinot and Lerasle, 2020, Muthukumar et al., 2020, Bartlett and Long, 2021, Lecué and Shang, 2022]. The main conclusion of these works is that benign overfitting can only occur in a linear model if the model is over-parametrized and if the design matrix has an unbalanced spectrum - which is close to the nonparametric setting. The above papers show that the estimators are interpolative but do not attain the optimal rates. The very recent paper [Wang et al., 2022] obtains optimal rates for benign overvitting in the linear regression setting, with sparsity set to 1.

Benign overfitting in nonparametric regression model

Consider now the nonparametric regression framework. Assume that we have n independent pairs of random variables $(X_1, Y_1), \ldots, (X_n, Y_n)$ in $\mathbf{R}^d \times \mathbf{R}$ such that, for all $1 \leq i \leq n$,

$$Y_i = f(X_i) + \varepsilon_i, \text{ with } \mathbf{E}(\varepsilon_i) = 0, \tag{1.6}$$

where ε_i are random noises. Function $f : \mathbf{R}^d \to \mathbf{R}$ is called the regression function and is unknown. The problem of nonparametric regression is to estimate f given a priori that it belongs to a nonparametric class of functions $\mathscr{F}$.

On particular instance is the nonparametric ridge regression framework. Then, it is assumed that f belongs to $\mathscr{H}$, where $\mathscr{H}$ is a reproducing kernel Hilbert space (RKHS). In this setting, it is common to consider estimators of f based on regularized least squares [Alvarez et al., 2012, Golub et al., 1979, Smola and Schölkopf, 1998], that is, solutions of the problem

$$\min_{f \in \mathscr{H}} \frac{1}{n} \sum_{i=1}^{n} (f(X_i) - Y_i)^2 + \lambda \|f\|_{\mathscr{H}}^2,$$

where $\lambda > 0$ and $\|\cdot\|_{\mathscr{H}}$ denotes the norm in $\mathscr{H}$.

The role of the regularization term is to avoid overfitting. In contrast to that, [Zhang et al., 2021, Belkin et al., 2019b], [Liang and Rakhlin, 2020] proposed the "ridgeless" kernel regression estimator, showing that interpolative solutions with minimal norms provide an implicit regularization under some conditions. They proposed the following estimator

$$\hat{f} = \arg\min_{f \in \mathscr{H}} \|f\|_{\mathscr{H}}, \text{ s.t. } f(X_i) = Y_i, \ \forall i.$$

[Liang and Rakhlin, 2020] assume that the sample size n is of the same order as the dimension d of the data, i.e., $d \asymp n$. Subsequently, [Liang et al., 2020] extended this framework to $d \asymp n^\alpha$ for $\alpha \in (0, 1)$. These papers provide upper bounds on the risk that depend on the sample and can be small depending on the spectral properties of the data and the RHKS kernel. In the case where d is constant independent of n, [Rakhlin and Zhai, 2019] have shown that the minimum norm interpolative estimator depending on the Laplace kernel does not converge.

We now place ourselves in the general nonparametric estimation regression framework (2.6). [Belkin et al., 2019b] have provided statistically optimal interpolative kernel estimators, using the Nadaraya-Watson estimator with singular kernel. The Nadaraya-Watson estimator is defined as

$$\hat{f}_n^{NW}(x) = \frac{\sum_{i=1}^n Y_i K\left(\frac{X_i - x}{h}\right)}{K\left(\frac{X_i - x}{h}\right)},$$

where $h > 0$ is the bandwith and $K : \mathbf{R}^d \to \mathbf{R}$ is a kernel.

It is known since [Shepard, 1968] that using a singular kernel with the Nadaraya-Watson estimator allows one to obtain an interpolative estimator. Singular kernel was initially chosen as $K(u) = \|u\|^{-a}$ where $a > 0$ and $\|\cdot\|$ denotes the Euclidean norm and $u \in \mathbf{R}^d$. [Shepard, 1968] introduced his interpolating estimator for $d = a = 2$. Not aware of the work of [Shepard, 1968] and its extensive use in image processing, [Devroye et al., 1998] generalized Shepard's estimator for any d and proved that this estimator converges in probability but not almost surely. However, this kernel is not integrable and has the particular property that the smoothing parameter h vanishes in the definition of the estimator. So, the bias-variance trade-off cannot be achieved by the choice of smoothing parameter h. Thus, [Belkin et al., 2019b] suggest to define the kernel as follows:

$$K(u) = \|u\|^{-a}\mathbf{1}(\|u\| \leq 1), \ a \in (0, d/2).$$

The Nadaraya-Watson estimator with this modified kernel remains interpolative. [Belkin et al., 2019b] proved that this estimator reaches the minimax convergence rates for the Hölder class of functions with regularity parameter $\beta \in (0, 2]$. However, these results do not extend to other values of β and the obtained estimators are not adaptive to β.

3.2 Local polynomial estimator

We consider the local polynomial estimator (LPE), which is a generalization of the Nadaraya-Watson estimator. For simplicity, we give here the definition of LPE in dimension $d = 1$. The definition in any dimension d is detailed in Chapter 5.

If we consider a nonnegative kernel, the Nadaraya-Watson estimator satisfies

$$\hat{f}_n^{NW}(x) = \arg\min_{\theta \in \mathbf{R}} \sum_{i=1}^{n} (Y_i - \theta)^2 K\left(\frac{X_i - x}{h}\right). \tag{1.7}$$

Thus, $\hat{f}_n^{NW}$ can be seen as a local approximation of the outputs Y_i by a constant via least squares. The local character is determined by the kernel, which gives more weight to the X_i closer to x, while θ is a local constant to be adjusted. We can generalize this definition by replacing the constant θ in (2.7) by a polynomial of degree ℓ. We introduce

$$U(u) = \left(1, u, u^2/2!, \ldots, u^\ell/\ell!\right)^\top$$
$$\theta(x) = \left(f(x), f'(x)h, f''(x)h^2, \ldots, f^{(\ell)}h^\ell\right)^\top.$$

Then, vector $\hat{\theta}_n(x) \in \mathbf{R}^{\ell+1}$ defined by

$$\hat{\theta}_n(x) = \arg\min_{\theta \in \mathbf{R}^{\ell+1}} \sum_{i=1}^{n} \left(Y_i - \theta^\top U\left(\frac{X_i - x}{h}\right)\right)^2 K\left(\frac{X_i - x}{h}\right) \tag{1.8}$$

is called a local polynomial estimator (LPE) of order ℓ of $\theta(x)$. The statistic

$$\hat{f}_n(x) = U^\top(0)\hat{\theta}_n(x)$$

is called a local polynomial estimator of order ℓ (or LP(ℓ)) of $f(x)$.

We will use LPE with singular kernels K, which requires a slightly different definition of $\hat{f}_n$ due to the fact that the minimization problem (2.8) is not well-defined for $x = X_i$ (see Chapter 5 for more detail). Local polynomial estimators with singular kernels have been quite popular since [Lancaster and Salkauskas, 1981] as interpolation tool in numerical analysis. They have also been considered in the context of nonparametric estimation by [Katkovnik, 1985]. However, these works did not provide statistical guarantees and studied only the regularity properties of such estimators.

3.3 Summary of the results

In Chapter 5, we prove that using local polynomial estimators with a singular kernel, one can construct minimax optimal interpolative estimators on Hölder classes of regularity parameter β, for any $\beta > 0$. For $\beta > 0$, $L > 0$, we denote by $\Sigma(\beta, L)$ a suitably defined Hölder class of functions (see Chapter 5 for the precise definition of this class). We denote by $\bar{f}_n$ our estimator. We suppose that K is a non-negative integrable kernel with singularity at 0, which is compactly supported and continuous on $\mathbf{R}^d \backslash \{0\}$. Let $C > 0$ be a constant which does not depend on n. We prove that our estimator satisfies, for all x in the support of X_1,

$$\sup_{f \in \Sigma(\beta, L)} \mathbf{E}\left([\bar{f}_n(x) - f(x)]^2 \right) \leq C n^{-\frac{2\beta}{2\beta+d}},$$
$$\sup_{f \in \Sigma(\beta, L)} \mathbf{E}\left(\|\bar{f}_n - f\|_{L_2}^2 \right) \leq C n^{-\frac{2\beta}{2\beta+d}},$$

and that there exists a constant $c > 0$ such that with probability greater than $1 - ce^{-A_n/c}$, with $A_n = n^{\frac{2\beta}{2\beta+d}}$, our estimator is interpolative, i.e. satisfies $\bar{f}_n(X_i) = Y_i$ for every i, and $\bar{f}_n$ is continuous on the support of X_1. Here, $\|\cdot\|_{L_2}$ denotes the $L_2(P_X)$-norm, where P_X is the marginal distribution of X_1.

We have also shown that these estimators can be constructed adaptively to β if $\beta \in (0, \beta_{\max}]$, for all $\beta_{\max} > 0$, and adaptively to the boundary parameter $L > 0$ of the Hölder class. These adaptive procedures are constructed using aggregation methods.

As a by-product, we obtain non-asymptotic bounds on the quadratic risk of LPE estimators in the classical framework, with non-singular kernels. To the best of our knowledge, such bounds were not available in the literature, as the prior work on LPE was focused on asymptotic properties of LPE, such as convergence in probability or asymptotic normality [Stone, 1980, Stone, 1982, Tsybakov, 1986, Fan and Gijbels, 1996].

Our results can also be read in [Chhor et al., 2022].

Chapter 2

Introduction en français

Au cours de cette thèse, nous avons étudié différents problèmes statistiques : le clustering dans le Bipartite Stochastic Block Model, l'estimation dans les Topic Models et le problème de Benign Overfitting dans le cadre non-paramétrique. Nous présentons ces trois problèmes plus en détails dans les chapitres suivants. À noter que les notations peuvent varier d'un chapitre à l'autre.

1 Le clustering dans le Bipartite Stochastic Block Model

Le premier problème étudié est le problème de clustering dans un graphe. Ce problème est développé dans le Chapitre 3. Pour les problèmes de clustering, on peut généralement considérer deux approches. Soit les observations sont composées d'individus/objets sans interaction et on choisit de les modéliser par un modèle de mélange. Soit les observations sont composées d'individus/objets avec interactions et on choisit de les modéliser par un modèle de graphe. Les modèles de graphe ont des applications dans de nombreuses disciplines et permettent, par exemple, l'étude d'interactions sociales, biologiques et informatiques. Un modèle essentiel en statistiques est le Stochastic Block Model (SBM). Il s'agit d'un modèle adapté à la détection des communautés.

1.1 Modèle d'Erdös-Rényi

Avant de présenter le SBM, nous commençons par introduire le modèle fondamental d'Erdös-Rényi (ER) [Erdős and Rényi, 1959, Erdős et al., 1960]. Soit n un entier et $p \in (0, 1)$. Un graphe $G(n, p)$ généré selon le modèle ER est un graphe non orienté de n sommets connectés aléatoirement. La probabilité que deux sommets soient connectés est de p, indépendamment des autres sommets. Bien que ce modèle soit très simple et souvent peu réaliste en pratique, de nombreux problèmes statistiques ont été étudiés à partir de ce modèle. Il a également permis de progresser dans l'étude de modèles de graphes plus complexes. En particulier, le modèle ER n'est pas un

modèle adapté à la détection de communautés, puisque la probabilité d'interaction de deux sommets est homogène dans tout le graphe.

Ce modèle permet d'introduire le concept de transition de phase. Un phénomène de transition de phase se produit lorsqu'on observe un phénomène de seuil. Les deux paramètres du modèle ER sont n et p. En fonction de leurs valeurs relatives, le graphique a un aspect différent. Ainsi, [Erdős et al., 1960] ont prouvé que pour tout $\varepsilon \in (0,1)$,

- si $p < \frac{(1-\varepsilon)\log n}{n}$, alors $G(n,p)$ est presque sûrement non-connexe (i.e., contient des sommets isolés),
- si $p > \frac{(1-\varepsilon)\log n}{n}$, alors $G(n,p)$ est presque sûrement connexe (i.e., ne contient aucun sommet isolé).

Ainsi, pour n fixé, une petite variation de p modifie radicalement l'aspect du graphe et $\frac{\log n}{n}$ peut être considéré comme un seuil de connectivité du graphe.

1.2 Le Stochastic Block Model

Le Stochastic Block Model (SBM) [Holland et al., 1983] peut être considéré comme une extension du modèle ER. L'hypothèse principale du SBM est que les sommets ne sont pas connectés au hasard mais selon leur communauté respective.

Considérons le cadre du SBM à deux communautés. Soit n un entier et $(p,q) \in (0,1)^2$. Un graphe $G(n,p,q)$ généré selon un modèle SBM à deux communautés est un graphe non orienté avec n sommets, tel que la probabilité que deux sommets appartenant à la même communauté soient connectés est p, et la probabilité que deux sommets appartenant à des communautés différentes soient connectés est q.

Nous donnons maintenant une définition formelle du SBM. Soit n_+, n_- deux entiers positifs tels que $n = n_+ + n_-$. Soit V l'ensemble de n sommets tel que V contienne n_+ sommets d'étiquette $+1$ et n_- sommets d'étiquette -1. Pour chaque sommet u de V, on note $\sigma(u) \in \{-1,1\}$ son étiquette. On note A la matrice d'adjacence du graphe, i.e. la matrice (n,n) dont les coefficients A_{ij} valent 1 si les sommets correspondants $i,j \in V$ sont connectés, et 0 sinon.
On dit que A est générée selon un modèle $SBM(n_+,n_-,p,q)$ si les coefficients A_{ij} sont indépendants et si

- $A_{ij} \sim Ber(p)$ si $\sigma(i) = \sigma(j)$, i.e. si deux sommets de même étiquette sont connectés avec une probabilité p,
- $A_{ij} \sim Ber(q)$ si $\sigma(i) \neq \sigma(j)$, i.e. si deux sommets d'étiquettes différentes sont connectés avec une probabilité q.

Ici, $Ber(p)$ désigne la distribution de Bernoulli de paramètre p.

On note que si $p = q$, on obtient le modèle d'Erdös-Rényi. Si $p > q$, le SBM est dit assortatif et les interactions sont plus fréquentes au sein d'une communauté qu'entre communautés. C'est la situation la plus courante dans les applications. Inversement, si $p < q$, le SBM est dit disassortatif et les interactions sont moins fréquentes au sein d'une communauté qu'entre les communautés. On trouve des exemples de graphes disassortatifs en biologie ou dans l'architecture des liens des pages Web.

Nous désignons par $\eta \in \{\pm 1\}^n$ le vecteur de la partition de V. Notez que si l'objectif est de classer les sommets entre deux communautés, cet objectif peut être atteint aussi bien en estimant η que $-\eta$.

Toute fonction mesurable $\hat{\eta}$ de A vers $\{\pm 1\}^n$ est un estimateur de η. Afin de mesurer la perte d'un tel estimateur, nous introduisons la distance de Hamming, qui est égale au double du nombre de coordonnées où η et $\hat{\eta}$ diffèrent :

$$|\eta - \hat{\eta}| := \sum_{i=1}^{n} |\eta_i - \hat{\eta}_i| = 2 \sum_{i=1}^{n} \mathbf{1}(\eta_i \neq \hat{\eta}_i),$$

où η_i (resp., $\hat{\eta}_i$) désigne la i^{e} coordonnée de η (resp., $\hat{\eta}$).

Puisque, comme mentionné ci-dessus, il est équivalent pour la détection de communautés d'estimer η et $-\eta$, nous considérons la perte suivante :

$$r(\eta, \hat{\eta}) = \min_{\nu \in \{-1,1\}} |\hat{\eta} - \nu\eta|. \tag{2.1}$$

Il existe plusieurs propriétés pertinentes pour l'étude du SBM.

Definition 4 (*weak recovery* dans le SBM)**.** *L'estimateur $\hat{\eta}$ accomplit la weak recovery de η s'il existe $\alpha \in (0, 1)$ tel que*

$$\lim_{n \to \infty} \sup_{SBM} \mathbf{P}\left(\frac{r(\eta, \hat{\eta})}{n} \geq \alpha \right) = 0, \tag{2.2}$$

où $\sup_{SBM}$ désigne le maximum sur tous les tirages de A selon $SBM(n_+, n_-, p, q)$.

La *weak recovery* est également appelé *detection* dans la littérature.

Definition 5 (*almost full recovery* dans le SBM)**.** *L'estimateur $\hat{\eta}$ accomplit l'almost full recovery de η si (2.2) est vraie pour tout $\alpha \in (0, 1)$.*

L'*almost full recovery*, également appelée *almost exact recovery* dans la littérature, signifie que $\hat{\eta}$ classe correctement presque tous les sommets avec une probabilité élevée.

Definition 6 (*exact recovery* dans le SBM)**.** *L'estimateur $\hat{\eta}$ accomplit l'exact recovery de η si*

$$\lim_{n \to \infty} \inf_{SBM} \mathbf{P}\Big(r(\eta, \hat{\eta}) = 0 \Big) = 1.$$

L'*exact recovery* signifie que $\hat{\eta}$ classe correctement tous les sommets avec une forte probabilité.

Un grand nombre de travaux ont été consacrés à la détermination des transitions de phase sur n, p, q pour ces problèmes. La plupart des résultats ont été obtenus sous les hypothèses suivantes, $a = pn$ et $b = qn$, $\alpha = pn/\log(n)$ et $\beta = qn/\log(n)$, qui caractérisent le cas le plus intéressant. En particulier, pour le problème de *weak recovery*, [Massoulié, 2014] et [Mossel et al., 2018] ont prouvé que la *weak recovery* est possible si et seulement si $(a-b)^2 > 2(a+b)$. Pour le problème d'*exact recovery* lorsque $p > q$, [Abbe et al., 2015] ont prouvé le phénomène de transition de phase suivant : l'*exact recovery* est possible si $\frac{\alpha+\beta}{2} - \sqrt{\alpha\beta} > 1$ et est impossible si $\frac{\alpha+\beta}{2} - \sqrt{\alpha\beta} < 1$.

1.3 Le Bipartite Stochastic Block Model

De nombreuses généralisations du Stochastic Block Model ont été étudiées. Dans cette thèse, nous nous intéressons plus particulièrement à une généralisation non-symétrique du SBM : le Bipartite Stochastic Block Model (BSBM).

Introduit par [Feldman et al., 2015], le BSBM est adapté à l'étude des interactions entre deux ensembles : chaque ensemble est divisé en plusieurs communautés, et ces deux ensembles sont le plus souvent composés d'objets de natures différentes. Au sein d'un ensemble, il peut exister des interactions inter et intra-communautés, mais celles-ci sont le plus souvent inaccessibles ou peu informatives. Le BSBM est par exemple pertinent pour l'étude des interactions objet/utilisateur dans le cadre des systèmes de recommandation. Les utilisateurs constituent un premier ensemble, divisé en plusieurs communautés, et les objets constituent un second ensemble, également divisé en plusieurs communautés. Ce modèle a plusieurs applications telles que les interactions document/mot [Dhillon and Modha, 2001, Lancichinetti et al., 2014], les interactions gène/séquences génétiques [Eren et al., 2013, Larremore et al., 2013], et les interactions objet/utilisateur dans le cadre des systèmes de recommandation [Jang et al., 2007].

Initialement, le BSBM a été introduit par [Feldman et al., 2015] dans le contexte des problèmes de satisfaction de contraintes (CSP). Les CSP sont des problèmes mathématiques qui consistent en l'étude d'états ou d'objets satisfaisant certains critères ou contraintes. Formellement, un CSP est un triplet $(\mathscr{X}, \mathscr{D}, \mathscr{C})$ où $\mathscr{X}$ est un ensemble de n variables, $\mathscr{D}$ est un ensemble de n domaines de définition pour chaque variable de $\mathscr{X}$, et $\mathscr{C}$ est un ensemble de contraintes. Les CSP apparaissent dans de nombreux domaines, par exemple en informatique et en apprentissage automatique. En particulier, la théorie des CSP est étroitement liée à la théorie de la complexité en informatique théorique. [Feldman et al., 2015] ont introduit le BSBM pour unifier plusieurs problèmes, notamment le problème classique du SBM et le problème CSP k-SAT. Le problème SAT, ou problème de satisfiabilité booléenne, est le problème CSP qui, étant donné une formule de logique propositionnelle, détermine s'il existe une affectation de variables propositionnelles qui rende la formule vraie. Si une telle affectation existe, on dit que la formule est satisfaisable. Rappelons qu'en logique

booléenne, un littéral est une variable booléenne ou sa négation, et une clause est une disjonction de littéraux. Les clauses s'identifient aux contraintes. Une formule k-SAT aléatoire est une conjonction de m clauses de k variables booléennes choisies aléatoirement parmi n variables booléennes. Nous nous intéressons à la probabilité qu'une formule k-SAT aléatoire soit satisfaisable. [Feldman et al., 2015] ont étudié le problème de la satisfiabilité *plantée*. Dans ce cadre, une affectation dite *plantée* est fixée à l'avance et les clauses (aléatoires) sont tirées selon une distribution définie par cette affectation. [Feldman et al., 2015] ont prouvé que le problème k-SAT planté aléatoire se réduit au BSBM.

Définition du BSBM

Soient $n_{1+}, n_{1-}, n_{2+}, n_{2-}$ quatre entiers positifs non nuls tels que $n_1 := n_{1+} + n_{1-} \leq n_{2+} + n_{2-} := n_2$ et soient $p \in (0, 1/2)$, $\delta \in (0, 2)$. Soient V_1 et V_2 deux ensembles de sommets tels que V_1 (respectivement V_2) soit composé de n_{1+} (resp. n_{2+}) sommets d'étiquette $+1$ et de n_{1-} (resp. n_{2-}) sommets d'étiquette -1. Pour chaque sommet u de V_1 ou V_2, on note $\sigma(u) \in \{-1, 1\}$ son étiquette.

On note A la matrice de biadjacence, i.e. la matrice (n_1, n_2) telle que ses coefficients A_{ij} sont égaux à 1 si les sommets correspondants $i \in V_1, j \in V_2$ sont connectés, et 0 sinon.

On dit alors que A est générée selon un modèle $BSBM(\delta, n_{1+}, n_{1-}, n_{2+}, n_{2-}, p)$ si les coefficients A_{ij} sont indépendants et si

- $A_{ij} \sim Ber(\delta p)$ si $\sigma(i) = \sigma(j)$ i.e. si deux sommets $i \in V_1$ et $j \in V_2$ de même étiquette sont connectés avec une probabilité δp,
- $A_{ij} \sim Ber((2 - \delta)p)$ si $\sigma(i) \neq \sigma(j)$ i.e. si deux sommets $i \in V_1$ et $j \in V_2$ d'étiquettes différentes sont connectés avec une probabilité $(2 - \delta)p$.

Notons que le modèle BSBM est bien une généralisation du modèle SBM puisque si $V_1 = V_2$, on obtient le SBM.

Détection de communautés

Supposons que nous observons une matrice A générée selon un modèle $BSBM(\delta, n_{1+}, n_{1-}, n_{2+}, n_{2-}, p)$. On s'intéresse au problème de l'estimation de la partition associée à V_1 à partir de l'observation de la matrice de biadjacence A. Soit $\eta_1 \in \{\pm 1\}^{n_1}$ le vecteur des étiquettes des sommets de V_1. Comme dans le cadre du SBM, il est équivalent d'estimer η_1 et $-\eta_1$. Toute fonction mesurable $\hat{\eta}$ de A vers $\{\pm 1\}^{n_1}$ est un estimateur de η_1.

Comme pour le problème SBM, nous considérons les problèmes de *weak recovery*, *almost full recovery* et *exact recovery*. On adapte facilement les définitions 4,5,6 au cadre du BSBM. Nous utilisons la perte de Hamming r définie dans (2.1) pour caractériser la perte d'un estimateur $\hat{\eta}$ de η_1.

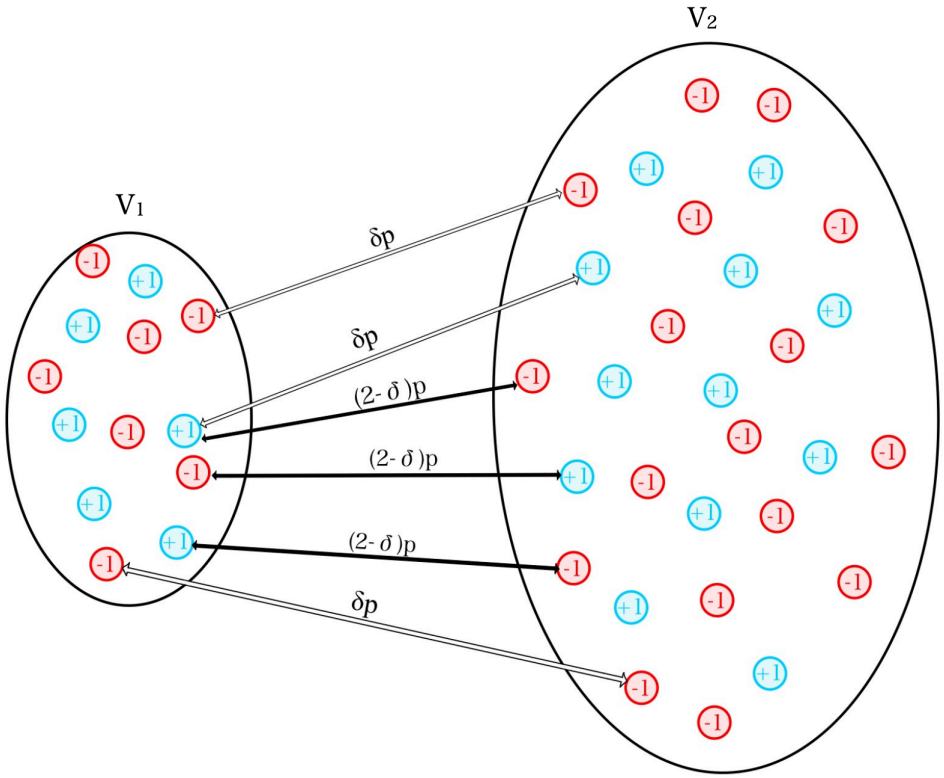


Figure 2.1: Illustration du BSBM.

Synthèse des résultats précédents

Si le SBM a été abondamment étudié, le BSBM reste moins connu. [Feldman et al., 2015, Florescu and Perkins, 2016, Cai et al., 2019] ont étudié les phénomènes de transition de phase pour p. [Florescu and Perkins, 2016] ont prouvé que le phénomène de transition de phase pour la *weak recovery* se produit pour la probabilité seuil $p_c = \frac{1}{(\delta-1)^2\sqrt{n_1 n_2}}$. Afin de prouver la condition suffisante, ils ont utilisé une réduction au SBM et ensuite un algorithme "black-box" optimal pour la *weak recovery* dans le SBM, comme dans [Bordenave et al., 2015, Massoulié, 2014, Mossel et al., 2018].

[Florescu and Perkins, 2016] ont également fourni une condition suffisante pour l'*almost full recovery* dans le cadre de grande dimension, i.e. pour $n_2 \gg n_1$. [Florescu and Perkins, 2016] utilisent des méthodes spectrales, qui sont des méthodes classiques dans le cadre de l'estimation de communautés. En particulier, [Florescu and Perkins, 2016] ont prouvé que modifier la méthode SVD classique permet une amélioration significative dans le cadre de grande dimension. En effet, au lieu de considérer les vecteurs singuliers de la matrice de biadjacence A, [Florescu and Perkins, 2016] considèrent les vecteurs propres de la matrice de Gram associée, dont les éléments diagonaux sont tous fixés à 0. L'algorithme associé est appelé "Diagonal Deletion SVD". Pour $n_2 \geq n_1(\log n_1)^4$, [Florescu and Perkins, 2016] ont prouvé avec le "Diagonal Deletion SVD" que $p = \Omega\left(\frac{\log n_1}{\sqrt{n_1 n_2}}\right)$ est une condition suffisante pour ac-

complir l'*almost full recovery* dans le modèle BSBM. Ici et dans ce qui suit, on note $a_n = O(b_n)$ s'il existe une constante $c > 0$ telle que $a_n \leq cb_n$, et on note $a_n = \Omega(b_n)$ s'il existe une constante $c > 0$ telle que $a_n \geq cb_n$. On note également $a_n \asymp b_n$ si $a_n = O(b_n)$ et $a_n = \Omega(b_n)$.

[Feldman et al., 2015] ont prouvé les meilleures conditions connues pour l'*exact recovery* jusqu'à nos résultats. Ils ont prouvé dans le cadre de grande dimension $n_2 \geq n_1$ que la condition $p = \Omega\left(\frac{(\delta-1)^2 \log n_1}{\sqrt{n_1 n_2}}\right)$ est suffisante pour accomplir l'*exact recovery* en utilisant leur algorithme nommé "Subsampled Power Iteration". [Feldman et al., 2015] conjecturent que la condition $p = \Omega(\frac{\log n_1}{\sqrt{n_1 n_2}})$ est nécessaire pour l'*exact recovery*. Nous avons réfuté cette conjecture (cf. ci-dessous).

Nous remarquons que dans des conditions très similaires, [Feldman et al., 2015] fournissent de meilleurs résultats que [Florescu and Perkins, 2016], puisque l'algorithme "Subsampled Power Iteration" permet l'*exact recovery* au lieu de l'*almost full recovery*.

[Cai et al., 2019] ont prouvé que l'algorithme "Diagonal Deletion SVD" permet également l'*exact recovery* dans des conditions très proches de [Feldman et al., 2015] pour $n_2 \gtrsim n_1$, mais ces conditions se dégradent dans le cadre $n_2 \gg n_1$, par exemple pour $n_2 \asymp e^{n_1}$. Pour être plus précis, [Cai et al., 2019] ont prouvé qu'il est possible d'accomplir l'*exact recovery* via le "Diagonal Deletion SVD" si

$$p \geq C(\delta - 1)^2 \left(\frac{\log(n_1 + n_2)}{\sqrt{n_1 n_2}} \vee \frac{\log(n_1 + n_2)}{n_1 + n_2} \right), \tag{2.3}$$

où $C > 0$ est une constante indépendante de p, δ, n_1, n_2. [Zhou and Amini, 2019, Zhou and Amini, 2020, Neumann, 2018] ont également fourni des résultats sur le BSBM mais dans un cadre plus général. Ces résultats ne permettent pas d'estimer les communautés dans le cadre de grande dimension lorsque $n_2 \gg n_1$. [Zhou and Amini, 2019] utilisent des méthodes spectrales sur une matrice bien choisie. Le théorème 1 de [Zhou and Amini, 2020] adapté à notre cadre nécessite $n_2 = O(n_1^2)$, ce qui est assez restrictif. [Neumann, 2018] se concentre sur l'estimation des clusters dits minuscules, i.e. les clusters de taille n^ε pour tout $\varepsilon > 0$, où n est le nombre de sommets du graphe. Cependant, les travaux de [Neumann, 2018] appliqués à notre cadre sont sous-optimaux.

[Feldman et al., 2015, Florescu and Perkins, 2016] ont observé que les principales difficultés d'estimation des communautés apparaissent dans le régime de grande dimension $n_2 \gg n_1$. Ce régime est le plus adapté aux applications réelles du BSBM et c'est celui qui présente des difficultés théoriques. En effet, le régime $n_2 \leq n_1$ est plus standard, car nous pouvons lui appliquer les méthodes classiques de clustering pour le SBM. Nous obtenons des résultats optimaux pour $n_2 \leq n_1$ en appliquant des algorithmes SVD directement à la matrice de biadjacence [Zhou and Amini, 2019, Zhou and Amini, 2020]. En effet, on sait contrôler la norme spectrale du bruit de la matrice de biadjacence, mais moins celle du bruit de la matrice de Gram à diagonale nulle, que l'on doit contrôler pour $n_2 \gg n_1$.

Pour obtenir nos résultats, nous avons fait une analogie avec le Modèle de Mélange Gaussien (GMM). [Lu and Zhou, 2016, Ndaoud, 2018, Giraud and Verzelen, 2019, Loffler et al., 2019] ont développé des algorithmes de clustering optimaux pour le GMM. [Lu and Zhou, 2016] ont prouvé qu'une procédure itérative similaire à l'algorithme de Lloyd [Lloyd, 1982] permet le clustering du GMM avec des propriétés optimales. [Ndaoud, 2018] a fourni une version modifiée de cet algorithme de clustering itératif, dont il a été démontré qu'il permet de réaliser la transition de phase pour l'*exact recovery* dans ce modèle GMM. Ainsi, par analogie entre le GMM et le BSBM, [Ndaoud, 2018] a conjecturé que des algorithmes similaires permettraient l'*almost full recovery* et l'*exact recovery* pour le BSBM, et que la condition $p = \Omega\left((\delta-1)^2\sqrt{\frac{\log n_1}{n_1 n_2}}\right)$ était suffisante pour obtenir l'*exact recovery* pour le BSBM si $n_2 \geq n_1 \log n_1$. Cela va à l'encontre de l'heuristique faite par analogie avec le SBM selon laquelle $p = \Omega\left(\frac{\log n_1}{\sqrt{n_1 n_2}}\right)$ est nécessaire pour obtenir l'*exact recovery*, cf. [Feldman et al., 2015].

Résultats

Dans le Chapitre 3, nous introduisons un algorithme appelé le *hollowed Lloyd's algorithm*, qui accomplit l'*exact recovery* dans de meilleures conditions que celles connues dans les travaux précédents. Nous prouvons que, pour tous les régimes de (n_1, n_2),

$$p \geq C(1-\delta)^2\left(\sqrt{\frac{\log n_1}{n_1 n_2}} \vee \frac{\log n_1}{n_2}\right) \tag{2.4}$$

est une condition suffisante pour l'*exact recovery* pour le BSBM, où $C > 0$ est une constante positive. Cette condition est meilleure que (2.3). En particulier, elle ne se dégrade pas pour $n_2 \asymp e^{n_1}$.

La condition (2.4) révèle un effet de coude à $n_2 \asymp n_1 \log n_1$ entre les régimes de grande et faible dimension. Dans le régime de faible dimension $n_2 \leq n_1 \log n_1$, nous retrouvons la condition suffisante de [Zhou and Amini, 2019, Cai et al., 2019], $p = \Omega\left(\frac{\log n_1}{n_2}\right)$. Dans le régime de grande dimension $n_2 \geq n_1 \log n_1$, nous obtenons la condition suffisante $p = \Omega\left(\sqrt{\frac{\log n_1}{n_1 n_2}}\right)$. Nous avons conjecturé dans [Ndaoud et al., 2021] que cette condition est nécessaire dans le régime $n_2 \geq n_1 \log n_1$ en exhibant un estimateur oracle qui n'accomplit pas l'*exact recovery* si $p < c\sqrt{\frac{\log n_1}{n_1 n_2}}$ pour $c > 0$ assez petit.

Nos résultats sont issus de [Ndaoud et al., 2021].

2 Estimation de la matrice topic-document dans le cadre de topic model

Le deuxième problème que nous avons étudié est celui de l'estimation dans le cadre de topic models. Ce problème est développé dans le Chapitre 4.

En statistiques et en traitement du langage naturel (NLP), un topic model (en français, "modèle thématique") est un modèle utile pour classer des documents par thèmes. Par exemple, il est nécessaire de savoir comment classer les pages Web afin de les recommander aux utilisateurs. Il est également utile de pouvoir classer automatiquement des articles scientifiques en ligne. Les topic models peuvent être utilisés pour résoudre ces problèmes.

2.1 Probabilistic Latent Semantic Indexing

Dans cette thèse, nous considérons le modèle probabilistic Latent Semantic Indexing (pLSI). Introduit par [Hofmann, 1999], ce modèle relie trois types de variables : les documents, les topics (ou sujets) et les mots. Nous supposons que nous disposons d'un dictionnaire de p mots et d'un corpus de n documents, où p, n sont des entiers strictement positifs. Un document est une suite de mots du dictionnaire. Nous supposons que les documents portent sur plusieurs sujets ou topics. Soit $K \in \mathbf{N}^*$ le nombre de sujets. Nous supposons que $2 \leq K \leq \min(n, p)$. Typiquement, K est très petit devant n et p.

L'hypothèse fondamentale du modèle pLSI est que la probabilité qu'un mot j apparaisse dans un document sur le sujet k est indépendante du document. En d'autres termes, selon la loi des probabilités totales,

$$\mathbf{P}(\text{mot } j|\text{document } i) = \sum_{k=1}^{K} \mathbf{P}(\text{topic } k|\text{document } i)\mathbf{P}(\text{mot } j|\text{topic } k).$$

Nous introduisons les notations suivantes :

$$\begin{aligned} \Pi_{ij} &= \mathbf{P}(\text{mot } j|\text{document } i), \\ W_{ik} &= \mathbf{P}(\text{topic } k|\text{document } i), \\ A_{kj} &= \mathbf{P}(\text{mot } j|\text{topic } k). \end{aligned}$$

Ensuite, nous pouvons écrire $\Pi_{ij} = W_i^{\mathrm{T}} A_j$, où $W_i = (W_{i1}, \ldots, W_{iK})^{\mathrm{T}} \in [0,1]^K$ est le vecteur de probabilité de chaque sujet dans le document i et $A_j = (A_{1j}, \ldots, A_{Kj})^{\mathrm{T}} \in [0,1]^K$ est le vecteur de probabilité de chaque sujet pour le mot j, pour chaque sujet $k = 1, \ldots, K$. Ainsi, nous pouvons écrire sous forme matricielle

$$\mathbf{\Pi} = \boldsymbol{W}\boldsymbol{A},$$

où $\mathbf{\Pi}$ est la matrice (n, p) documents-mots de coefficients Π_{ij}, $\boldsymbol{W} := (W_1, \ldots, W_n)^{\mathrm{T}}$ est la matrice (n, K) documents-sujets, $\boldsymbol{A} := (A_1, \ldots, A_p)$ est la matrice (K, p) sujets-mots. Les lignes de ces matrices étant des vecteurs de probabilité, nous avons

$$\sum_{m=1}^{K} W_{im} = 1, \ \sum_{j=1}^{p} A_{kj} = 1, \ \sum_{j=1}^{p} \Pi_{ij} = 1 \text{ pour tout } i = 1, \ldots, n, \ k = 1, \ldots K.$$

Ici, Π_{ij} est la probabilité d'occurrence du mot j dans le document i. En pratique, nous n'observons pas Π_{ij} mais la fréquence empirique correspondante X_{ij}.

Nous avons donc une matrice (n, p) documents-mots $\boldsymbol{X} = (X_{ij})$ telle que pour chaque document i de $1, \dots, n$, et chaque mot j de $1, \dots, p$, X_{ij} est la fréquence observée du mot j dans le document i. Soit N_i le nombre (déterministe) de mots échantillonnés dans le document i. Nous supposons que, pour chaque vecteur $X_i = (X_{i1}, \dots, X_{ip})^\top$ de mots du document i, le vecteur correspondant $N_i X_i$ du nombre d'occurrences de chaque mot dans le document i suit une distribution multinomiale de dimension p, de paramètres (N_i, Π_i), où $\Pi_i = \mathbf{E}(X_i) = (\Pi_{i1}, \dots, \Pi_{ip})^\top$. Nous supposons également que $X_1, \dots, X_n$ sont des vecteurs aléatoires indépendants. Nous pouvons écrire le modèle comme un modèle "signal + bruit" :

$$\boldsymbol{X} = \boldsymbol{\Pi} + \boldsymbol{Z} = \boldsymbol{W}\boldsymbol{A} + \boldsymbol{Z}, \tag{2.5}$$

où $\boldsymbol{Z} := \boldsymbol{X} - \boldsymbol{\Pi}$ est une matrice de moyenne nulle.

L'objectif des topic models est d'estimer les matrices $\boldsymbol{A}$ et $\boldsymbol{W}$ à partir de l'observation de la matrice $\boldsymbol{X}$ et de la connaissance de $N_1, \dots, N_n$. L'estimation de $\boldsymbol{A}$ et l'estimation de $\boldsymbol{W}$ répondent à des objectifs différents. Un estimateur de la matrice $\boldsymbol{A}$ identifie la distribution des sujets sur le dictionnaire. Un estimateur de la matrice $\boldsymbol{W}$ identifie les sujets associés à chaque document. Notre étude se concentre principalement sur l'estimation de la matrice $\boldsymbol{W}$.

2.2 Contraintes

Il est courant, dans le contexte des topic models, d'introduire des hypothèses d'ancrages. Une hypothèse largement utilisée dans la littérature est l'*anchor word assumption*, cf. par exemple [Arora et al., 2013, Bing et al., 2020a, Ke and Wang, 2017]. Cette *anchor word assumption* consiste à supposer que pour chaque sujet, il existe au moins un mot associé uniquement à ce sujet. Cette hypothèse est la mieux adaptée à l'estimation de la matrice $\boldsymbol{A}$, qui a été largement étudiée dans la littérature. Une autre hypothèse que nous adoptons ci-dessous est l'hypothèse *anchor document assumption*, qui postule que pour chaque sujet, il existe au moins un document portant exclusivement sur ce sujet.

Une idée fondamentale de la statistique en grande dimension est qu'un large ensemble de données peut être de très faible rang. Ainsi, en réduisant la dimension, il est possible de mieux exploiter les données. C'est le cas pour les topic models, où il existe une structure thématique sous-jacente de dimension K qui est généralement très petite par rapport à n et p. Il existe plusieurs méthodes de réduction de dimension. Nous sommes particulièrement intéressés aux méthodes de factorisation matricielle.

2.3 Vue d'ensemble des résultats précédents

Nonnegative Matrix Factorization (NMF) sans bruit

Nous considérons d'abord un modèle sans bruit, où l'équation (2.5) devient

$$\boldsymbol{X} = \boldsymbol{W}\boldsymbol{A}.$$

Les méthodes de Nonnegative Matrix Factorization (NMF) visent à estimer $\boldsymbol{W}$ et $\boldsymbol{A}$ à partir de l'observation de $\boldsymbol{X}$. Une matrice positive ("nonnegative" en anglais) est une matrice dont les entrées sont positives. La méthode NMF a été introduite par [Paatero and Tapper, 1994, Paatero, 1997, Lee and Seung, 1999a, Lee and Seung, 2000] et consiste à factoriser sous des hypothèses appropriées une matrice positive $\boldsymbol{X}$ en un produit de deux matrices positives $\boldsymbol{W}$, $\boldsymbol{A}$. Dans le cas où $\boldsymbol{X}$ est une matrice (n, p), $\boldsymbol{W}$ une matrice (n, K) et $\boldsymbol{A}$ une matrice (K, p), cette factorisation permet une réduction de dimension. Cette méthode est utile dans certains contextes où les matrices étudiées sont intrinsèquement positives, comme en NLP ou en analyse d'images. Les méthodes de réduction classiques comme l'analyse en composantes principales (ACP) ne peuvent pas être appliquées à de telles matrices. En effet, la contrainte d'orthogonalité requise dans l'ACP ne peut être respectée. C'est précisément le cadre du modèle pLSI, où toutes les matrices considérées sont positives. En général, le problème NMF est un problème NP-hard, mais des contraintes de séparabilité telles que l'hypothèse *anchor document assumption* permettent de le résoudre. Ces méthodes consistent le plus souvent en la minimisation d'une fonction de coût régularisée, cf. par exemple [Cichocki et al., 2009, Donoho and Stodden, 2004, Lee and Seung, 1999a, Recht et al., 2012]. Cependant, ces articles traitent du cas sans bruit et leurs résultats ne peuvent être utilisés dans le modèle avec bruit (2.5).

Perspective bayésienne : Latent Dirichlet Allocation

Nous nous intéressons à présent à la NMF dans le cas bruité spécifique aux topic models. Il existe une littérature abondante sur l'estimation des matrices $\boldsymbol{A}, \boldsymbol{W}$ dans ce contexte. Le modèle le plus célèbre et le plus populaire dans la classification de documents est probablement le modèle Latent Dirichlet Allocation ou LDA. Introduit par [Blei et al., 2003], le modèle LDA impose une prior de Dirichlet sur la matrice $\boldsymbol{A}$ et estime ensuite la matrice $\boldsymbol{W}$ par un algorithme EM. Cette perspective vise principalement à la construction d'algorithmes et ne fournit pas de garanties statistiques sur les estimateurs obtenus. De plus, l'algorithme LDA est computationnellement lent et suppose que les sujets sont non corrélés, ce qui n'est pas réaliste [Blei and Lafferty, 2007, Li and McCallum, 2006]. Pour résoudre ce dernier problème, [Lafferty and Blei, 2006] ont introduit un autre modèle, le Correlated Topic Model. D'autres articles connexes ont utilisé l'échantillonnage de Gibbs ([Porteous et al., 2008, Ramage et al., 2009]) ou les techniques de Bayes variationnelles ([Chien and Chueh, 2010, Zhai et al., 2012]) pour estimer $\boldsymbol{W}$, plutôt que d'utiliser l'algorithme EM.

Travaux antérieurs sur les garanties statistiques pour les topic models

Pour estimer la matrice $\boldsymbol{A}$, plusieurs articles ont proposé des algorithmes avec des garanties statistiques sous l'hypothèse *anchor word assumption*, cf. par exemple [Arora et al., 2012, Arora et al., 2013, Ding et al., 2013, Anandkumar et al., 2014, Bansal et al., 2014, Bing et al., 2020a, Bing et al., 2020c, Bing et al., 2021a, Ke and Wang, 2017]. Ces articles utilisent différentes techniques, notamment l'analyse de la matrice de co-occurrence, des tenseurs ou des méthodes de simplexes via SVD. Les travaux de [Bing et al., 2020a, Bing et al., 2020c, Bing et al., 2021a, Ke and Wang, 2017] développent des méthodes atteignant des taux de convergence minimax optimaux à des facteurs logarithmiques près pour l'estimateur de $\boldsymbol{A}$ pour la norme ℓ_1. Ces articles utilisent l'hypothèse *anchor word assumption* mais proposent des estimateurs différents des nôtres, car ils se concentrent sur l'estimation de la matrice $\boldsymbol{A}$. La méthode d'estimation de la matrice $\boldsymbol{A}$ dans [Ke and Wang, 2017] est une méthode de simplexe en dimension p, avec un coût de calcul p^K. Leur procédure permet également d'estimer la matrice $\boldsymbol{W}$ en utilisant leur estimateur $\hat{\boldsymbol{A}}$ de $\boldsymbol{A}$, via une méthode des moindres carrés.

[Bing et al., 2020a] proposent une méthode rapide fondée sur des méthodes SVD pour l'estimation de $\boldsymbol{A}$ sous l'hypothèse *anchor word assumption*. Dans les travaux suivants, [Bing et al., 2020c, Bing et al., 2021a] considèrent le problème (2.5) dans un cadre sparse et sous l'hypothèse *anchor word assumption*. [Bing et al., 2021a] étudient l'estimation des lignes de la matrice $\boldsymbol{W}$ à partir d'un estimateur $\hat{\boldsymbol{A}}$ de $\boldsymbol{A}$. Notons $(w_i)_{i=1}^n$ les lignes de la matrice $\boldsymbol{W}$ et $(\hat{w}_i^{MLE})_{i=1}^n$ les estimateurs par maximum de vraisemblance correspondants de [Bing et al., 2021a]. Supposons que $N_i = N$ pour tout i. On fixe $1 \leq i \leq n$. [Bing et al., 2021a] prouvent, pour la norme ℓ_1 du i^e rang $\|\hat{w}_i^{MLE} - w_i\|_1$, une borne en probabilité de l'ordre de $\left(\sqrt{\frac{p}{nN}} \vee \frac{1}{\sqrt{N}}\right)$ (à un faible facteur près, i.e. un facteur ne dépendant que de K ou de la sparsité, et un logarithme des paramètres principaux p, n, N). Par conséquent, le taux d'estimation de $\boldsymbol{W}$ en la norme ℓ_1 obtenue par [Bing et al., 2021a] est de l'ordre $\left(\sqrt{\frac{np}{N}} \vee \frac{n}{\sqrt{N}}\right)$, à un faible facteur près. Un résultat analogue est obtenu dans la version 2022 de l'article [Ke and Wang, 2017] pour un estimateur différent, toujours sous l'hypothèse *anchor word assumption*.

2.4 Algorithme SPOC

Afin d'estimer $\boldsymbol{W}$, nous avons introduit un algorithme appelé Successive Projection Overlapping Clustering (SPOC), inspiré par le Successive Projection Algorithm (SPA). L'algorithme SPA a été introduit par [Araújo et al., 2001] pour résoudre le problème NMF et a été largement utilisé par la suite, en raison de sa simplicité et de sa rapidité. L'algorithme SPOC applique l'algorithme SPA non pas à la matrice initiale $\boldsymbol{X}$, mais plutôt à la matrice de ses vecteurs singuliers gauches, et utilise le résultat pour estimer $\boldsymbol{W}$.

Les algorithmes SPA et SPOC sont des algorithmes itératifs de projection. En partant de la décomposition SVD de la matrice $\boldsymbol{X}$, nous appliquons la procédure

itérative suivante : à chaque étape de l'algorithme, nous sélectionnons la ligne de norme maximale de la matrice des vecteurs singuliers gauches, et nous projetons notre matrice sur le complément orthogonal de l'espace généré par cette ligne. Géométriquement, cette procédure s'explique par le fait que les lignes de la matrice des vecteurs singuliers de $\mathbf{\Pi}$ appartiennent à un simplexe dont les sommets sont les *anchor documents*. L'algorithme SPOC trouve itérativement des estimateurs de ces sommets, ce qui permet ensuite l'estimation de $\boldsymbol{W}$.

2.5 Résultats

Dans le Chapitre 4, nous fournissons des bornes sur l'algorithme SPOC en la norme de Frobenius et la norme ℓ_1. Nous obtenons des bornes supérieures et inférieures qui correspondent à un facteur logarithmique près, ce qui implique la quasi-optimalité de notre procédure. Plus précisément, en supposant que $N_i = N$ pour tout i, nous prouvons que l'estimateur SPOC de $\boldsymbol{W}$ converge en la norme de Frobenius et en la norme ℓ_1 aux taux $\sqrt{n/N}$ et $n/\sqrt{N}$ respectivement, à un faible facteur près. Nous prouvons également des bornes inférieures minimax d'ordres $\sqrt{n/N}$ et $n/\sqrt{N}$ respectivement. Nous pouvons observer que le taux ℓ_1 est plus rapide que le résultat $\left(\sqrt{\frac{np}{N}} \vee \frac{n}{\sqrt{N}}\right)$ de [Bing et al., 2021a]. Cependant, les hypothèses de [Bing et al., 2021a] sont différentes. Ils imposent l'hypothèse *anchor word assumption* alors que nous travaillons sous l'hypothèse *anchor document assumption*. Le terme potentiellement grand $\sqrt{\frac{np}{N}}$ dans le taux apparaît dans [Bing et al., 2021a] du fait de l'estimation préliminaire de $\boldsymbol{A}$. Elle tient compte de l'estimation de $\boldsymbol{A}$, qui est un artefact de la méthode et n'apparaît pas dans la borne inférieure.

Nous remarquons également, d'après nos résultats théoriques et nos simulations, que l'erreur de l'algorithme SPOC n'augmente pas significativement avec p, contrairement à l'erreur de l'algorithme LDA. De plus, notre procédure est computationnellement rapide et simple à mettre en oeuvre. Comparé à [Ke and Wang, 2017], dont l'algorithme a une complexité p^K, notre algorithme a une complexité $\max(p,n)K + nK^2$. Le premier terme $\max(p,n)K$ correspond au coût du SVD tronqué, et le second terme nK^2 correspond au coût du SPA. Enfin, notre procédure est adaptative : elle ne requiert pas la connaissance du nombre de sujets K.

Nos résultats sont issus de [Klopp et al., 2021].

3 Benign Overfitting en régression non paramétrique

Le troisième problème que nous avons étudié est le problème de Benign Overfitting dans le cadre de la régression non paramétrique. Ce problème est développé dans le Chapitre 5.

3.1 État de l'art

Les méthodes modernes d'apprentissage automatique ont manifesté des propriétés inattendues et en apparence contradictoires avec les statistiques classiques. L'une des méthodes d'apprentissage les plus populaires aujourd'hui est celle des réseaux de neurones. Initialement inspirés par le fonctionnement du cerveau humain, les réseaux de neurones imitent la transmission d'un signal d'un neurone à un autre. Le fonctionnement théorique des réseaux de neurones est encore mal compris. Empiriquement, les excellentes performances des réseaux de neurones profonds ne sont plus à démontrer (cf. par exemple [Gupta et al., 2015, Sagun et al., 2017, Huang et al., 2017]). Cependant, les réseaux de neurones profonds ont la particularité d'être interpolants, i.e. de biais nul sur le jeu de données d'apprentissage, mais tout en présentant une erreur de prédiction faible sur des jeux de données inconnus [Belkin et al., 2019a, Zhang et al., 2021]. Ce phénomène va à l'encontre de l'intuition classique du compromis biais-variance, qui est donc remis en question depuis quelques années (cf. par exemple [Ma et al., 2018]).

Travaux récents sur le Benign Overfitting en régression linéaire

Afin de mieux comprendre ce phénomène, appelé Benign Overfitting, plusieurs travaux ont étudié ce problème dans le contexte de la régression linéaire, cf. par exemple [Bartlett et al., 2020, Tsigler and Bartlett, 2020, Chinot and Lerasle, 2020, Muthukumar et al., 2020, Bartlett and Long, 2021, Lecué and Shang, 2022]. La principale conclusion de ces travaux est que le Benign Overfitting ne peut se produire dans un modèle linéaire que si le modèle est surparamétré et si la matrice de design a un spectre déséquilibré – ce qui est proche du cadre non paramétrique. Les articles ci-dessus montrent que les estimateurs sont interpolants mais n'atteignent pas les taux de convergence optimaux. L'article très récent [Wang et al., 2022] obtient des taux optimaux pour le Benign Overfitting dans le cadre de la régression linéaire, avec une sparsité fixée à 1.

Benign Overfitting dans le modèle de régression non paramétrique

Considérons maintenant le cadre de la régression non paramétrique. Supposons que nous disposons de n paires de variables aléatoires indépendantes $(X_1, Y_1), \dots, (X_n, Y_n)$ dans $\mathbf{R}^d \times \mathbf{R}$ tel que, pour tout $1 \leq i \leq n$,

$$Y_i = f(X_i) + \varepsilon_i, \text{ avec } \mathbf{E}(\varepsilon_i) = 0, \tag{2.6}$$

où ε_i sont des bruits aléatoires. La fonction $f : \mathbf{R}^d \to \mathbf{R}$ est appelée la fonction de régression et est inconnue. Le problème de régression non paramétrique est d'estimer f, étant donné qu'elle appartient à une classe non paramétrique de fonctions $\mathscr{F}$.

Le cadre non paramétrique de la régression ridge en est un exemple particulier. On suppose alors que f appartient à $\mathscr{H}$, où $\mathscr{H}$ est un espace de Hilbert à noyau reproducteur (RKHS). Dans ce cadre, il est courant de considérer des estimateurs de

f de moindres carrés régularisés [Alvarez et al., 2012, Golub et al., 1979, Smola and Schölkopf, 1998], i.e. des solutions du problème

$$\min_{f\in\mathscr{H}} \frac{1}{n}\sum_{i=1}^{n}(f(X_i)-Y_i)^2+\lambda\|f\|^2_{\mathscr{H}},$$

où $\lambda>0$ et $\|\cdot\|_{\mathscr{H}}$ désigne la norme dans $\mathscr{H}$.

Le rôle du terme de régularisation est d'éviter le surajustement. À l'opposé, [Zhang et al., 2021, Belkin et al., 2019b], [Liang and Rakhlin, 2020] ont proposé l'estimateur "ridgeless" de régression à noyau, montrant que les solutions interpolantes de normes minimales fournissent une régularisation implicite sous certaines conditions. Ils ont proposé l'estimateur suivant

$$\hat{f}=\arg\min_{f\in\mathscr{H}}\|f\|_{\mathscr{H}},\ \text{tel que}\ f(X_i)=Y_i,\ \forall i.$$

[Liang and Rakhlin, 2020] supposent que la taille de l'échantillon n est du même ordre que la dimension d des données, i.e. $d\asymp n$. Par la suite, [Liang et al., 2020] ont étendu ce cadre à $d\asymp n^\alpha$ pour $\alpha\in(0,1)$. Ces articles fournissent des bornes supérieures sur le risque qui dépendent de l'échantillon et peuvent être petites en fonction des propriétés spectrales des données et du noyau RHKS. Dans le cas où d est constant et indépendant de n, [Rakhlin and Zhai, 2019] ont montré que l'estimateur interpolant de norme minimale dépendant du noyau de Laplace ne converge pas.

Nous nous plaçons maintenant dans le cadre général de la régression non paramétrique (2.6). [Belkin et al., 2019b] ont fourni des estimateurs à noyau interpolants statistiquement optimaux, en utilisant l'estimateur de Nadaraya-Watson avec un noyau singulier. L'estimateur de Nadaraya-Watson est défini comme suit :

$$\hat{f}_n^{NW}(x)=\frac{\sum_{i=1}^n Y_i K\left(\frac{X_i-x}{h}\right)}{K\left(\frac{X_i-x}{h}\right)},$$

où $h>0$ est un paramètre de lissage et $K:\mathbf{R}^d\to\mathbf{R}$ est un noyau.

Il est connu depuis [Shepard, 1968] que l'utilisation d'un noyau singulier avec l'estimateur de Nadaraya-Watson permet d'obtenir un estimateur interpolant. Le noyau singulier initialement choisi était $K(u)=\|u\|^{-a}$ où $a>0$ et $\|\cdot\|$ désigne la norme euclidienne et $u\in\mathbf{R}^d$. [Shepard, 1968] a introduit son estimateur interpolant pour $d=a=2$. Ignorant le travail de [Shepard, 1968] et sa large utilisation en traitement d'images, [Devroye et al., 1998] ont développé l'estimateur de Shepard pour tout d et ont prouvé que cet estimateur converge en probabilité mais pas presque sûrement. Cependant, le noyau qu'ils considèrent n'est pas intégrable et a la propriété particulière que le paramètre de lissage h se simplifie dans la définition de l'estimateur. Ainsi, le compromis biais-variance ne peut pas être atteint par le choix du paramètre de lissage h. Par conséquent, [Belkin et al., 2019b] suggèrent de

définir le noyau comme suit :

$$K(u) = \|u|^{-a}\mathbf{1}(\|u\| \leq 1),\ a \in (0, d/2).$$

L'estimateur de Nadaraya-Watson avec ce noyau modifié reste interpolant. [Belkin et al., 2019b] ont prouvé que cet estimateur atteint les taux de convergence minimax pour la classe des fonctions de Hölder avec un paramètre de régularité $\beta \in (0, 2]$. Cependant, ces résultats ne s'étendent pas à d'autres valeurs de β et les estimateurs obtenus ne sont pas adaptatifs en β.

3.2 Estimateur par polyômes locaux

Nous considérons l'estimateur par polynômes locaux (LPE), qui est une généralisation de l'estimateur de Nadaraya-Watson. Pour des raisons de simplicité, nous donnons ici la définition du LPE en dimension $d = 1$. La définition en dimension d quelconque est détaillée dans le Chapitre 5.

Si l'on considère un noyau positif, l'estimateur de Nadaraya-Watson satisfait la condition suivante

$$\hat{f}_n^{NW}(x) = \arg\min_{\theta \in \mathbf{R}} \sum_{i=1}^{n} (Y_i - \theta)^2 K\left(\frac{X_i - x}{h}\right). \tag{2.7}$$

Ainsi, $\hat{f}_n^{NW}$ peut être vu comme une approximation locale des sorties Y_i par une constante via les moindres carrés. Le caractère local est déterminé par le noyau, qui donne plus de poids aux X_i plus proches de x, tandis que θ est une constante locale à ajuster. Nous pouvons généraliser cette définition en remplaçant la constante θ dans (2.7) par un polynôme de degré ℓ. Nous introduisons

$$U(u) = \left(1, u, u^2/2!, \ldots, u^\ell/\ell!\right)^\top$$
$$\theta(x) = \left(f(x), f'(x)h, f''(x)h^2, \ldots, f^{(\ell)}h^\ell\right)^\top.$$

Ensuite, le vecteur $\hat{\theta}_n(x)\ dans \mathbf{R}^{\ell+1}$ défini par

$$\hat{\theta}_n(x) = \arg\min_{\theta \in \mathbf{R}^{\ell+1}} \sum_{i=1}^{n} \left(Y_i - \theta^\top U\left(\frac{X_i - x}{h}\right)\right)^2 K\left(\frac{X_i - x}{h}\right) \tag{2.8}$$

est appelé un estimateur par polynômes locaux (LPE) d'ordre ℓ de $\theta(x)$. La statistique

$$\hat{f}_n(x) = U^\top(0)\hat{\theta}_n(x)$$

est appelé un estimateur par polynômes locaux d'ordre ℓ (ou LP(ℓ)) de $f(x)$.

Nous utiliserons les LPE avec des noyaux singuliers K, ce qui nécessite une définition légèrement différente de $\hat{f}_n$, du fait que le problème de minimisation (2.8) n'est pas bien défini pour $x = X_i$ (voir Chapitre 5 pour plus de détails). Les LPE à noyaux singuliers sont assez populaires depuis [Lancaster and Salkauskas, 1981] comme outils d'interpolation en analyse numérique. Ils ont également été considérés dans le contexte de l'estimation non paramétrique par [Katkovnik, 1985]. Cependant, ces travaux n'ont pas fourni de garanties statistiques et ont étudié uniquement les propriétés de régularité de tels estimateurs.

3.3 Résultats

Dans le Chapitre 5, nous prouvons qu'en utilisant des LPE avec un noyau singulier, on peut construire des estimateurs interpolants minimax-optimaux sur les classes de Hölder de paramètre de régularité β, pour tout $\beta > 0$. Pour $\beta > 0$, $L > 0$, nous désignons par $\Sigma(\beta, L)$ une classe de fonctions de Hölder convenablement définie (voir Chapitre 5 pour la définition précise de cette classe). Nous notons $\bar{f}_n$ notre estimateur. Nous supposons que K est un noyau intégrable positif avec une singularité en 0, qui est défini de manière compacte et continue sur $\mathbf{R}^d \backslash \{0\}$. Soit $C > 0$ une constante qui ne dépend pas de n. Nous prouvons que notre estimateur satisfait, pour tout x dans le support de X_1,

$$\sup_{f \in \Sigma(\beta, L)} \mathbf{E}\left([\bar{f}_n(x) - f(x)]^2 \right) \leq C n^{-\frac{2\beta}{2\beta+d}},$$
$$\sup_{f \in \Sigma(\beta, L)} \mathbf{E}\left(\|\bar{f}_n - f\|_{L_2}^2 \right) \leq C n^{-\frac{2\beta}{2\beta+d}},$$

et qu'il existe une constante $c > 0$ telle qu'avec une probabilité supérieure à $1 - ce^{-A_n/c}$, avec $A_n = n^{\frac{2\beta}{2\beta+d}}$, notre estimateur est interpolant, i.e. qu'il satisfait à $\bar{f}_n(X_i) = Y_i$ pour tout i, et $\bar{f}_n$ est continue sur le support de X_1. Ici, $\|\cdot\|_{L_2}$ désigne la norme $L_2(P_X)$, où P_X est la distribution marginale de X_1.

Nous avons également montré que ces estimateurs peuvent être construits de manière adaptative en β si $\beta \in (0, \beta_{\max}]$, pour tout $\beta_{\max} > 0$, et de manière adaptative au paramètre de borne $L > 0$ de la classe de Hölder. Ces procédures adaptatives sont construites en utilisant des méthodes d'agrégation.

En corollaire, nous obtenons des bornes non-asymptotiques sur le risque quadratique des estimateurs LPE dans le cadre classique, avec des noyaux non-singuliers. À notre connaissance, de telles bornes n'étaient pas disponibles dans la littérature, car les travaux antérieurs sur les LPE étaient axés sur les propriétés asymptotiques des LPE, comme la convergence en probabilité ou la normalité asymptotique [Stone, 1980, Stone, 1982, Tsybakov, 1986, Fan and Gijbels, 1996].

Nos résultats sont issus de [Chhor et al., 2022].

Chapter 3

Improved clustering algorithms for the Bipartite Stochastic Block Model

We establish sufficient conditions of exact and almost full recovery of the node partition in Bipartite Stochastic Block Model (BSBM) using polynomial time algorithms. First, we improve upon the known conditions of almost full recovery by spectral clustering algorithms in BSBM. Next, we propose a new computationally simple and fast procedure achieving exact recovery under milder conditions than the state of the art. Namely, if the vertex sets V_1 and V_2 in BSBM have sizes n_1 and n_2, we show that the condition $p = \Omega\left(\max\left(\sqrt{\frac{\log n_1}{n_1 n_2}}, \frac{\log n_1}{n_2}\right)\right)$ on the edge intensity p is sufficient for exact recovery witin V_1. This condition exhibits an elbow at $n_2 \asymp n_1 \log n_1$ between the low-dimensional and high-dimensional regimes. The suggested procedure is a variant of Lloyd's iterations initialized with a well-chosen spectral estimator leading to what we expect to be the optimal condition for exact recovery in BSBM. The optimality conjecture is supported by showing that, for a supervised oracle procedure, such a condition is necessary to achieve exact recovery. The key elements of the proof techniques are different from classical community detection tools on random graphs. Numerical studies confirm our theory, and show that the suggested algorithm is both very fast and achieves almost the same performance as the supervised oracle. Finally, using the connection between planted satisfiability problems and the BSBM, we improve upon the sufficient number of clauses to completely recover the planted assignment.

This chapter is based on [Ndaoud et al., 2021]: M. Ndaoud, S. Sigalla, and A. B. Tsybakov, *Improved clustering algorithms for the bipartite stochastic block model.* IEEE Transactions on Information Theory, 2021, vol. 68, no 3, p. 1960-1975.

---

1 Introduction 36

1.1 Definition of Bipartite Stochastic Block Model 37

1.2 Recovery of communities 37
2 Reduction to a spiked model 39
3 Related work 41
4 Main contributions 44
5 Properties of the spectral method 47
6 Exact recovery by the hollowed Lloyd's algorithm 50
7 Impossibility result for a supervised oracle 50
8 Gram matrix study 51
9 Lower bound on the oracle 56
10 Main proofs 59
10.1 Proof of Theorem 1 59
10.2 Proof of Theorem 2 63
11 Numerical experiments 68

1 Introduction

Unsupervised learning or clustering is a recurrent problem in statistics and machine learning. Depending on the objects we wish to classify, we can generally consider two approaches: either the observed objects are individuals without any interaction, which is often described by a mixture model, or the observed objects are individuals with interactions, which is described by a graph model. In the latter case, the individuals correspond to vertices of the graph and two vertices are connected if the two corresponding individuals interact. The clustering problem becomes then a node clustering problem, which means grouping the individuals by communities. The most known and studied framework for node clustering is the Stochastic Block Model (SBM), cf. [Holland et al., 1983]. In this paper, we focus on the Bipartite Stochastic Block Model (BSBM), cf. [Feldman et al., 2015], which is a non-symmetric generalization of the SBM. This model arises in several fields of applications. For example, it can be used to describe different types of interactions; documents/words [Dhillon, 2001, Lancichinetti et al., 2014], genes/genetic sequences [Eren et al., 2013, Larremore et al., 2013] and objects/users in recommendation systems [Jang et al., 2007]. Some other examples are related to random computational problems with planted solutions such as planted satisfiability problems, cf. [Feldman et al., 2018] for a general definition. As shown in [Feldman et al., 2015], three planted satisfiability problems reduce to solving the BSBM. Namely, this concerns planted hypergraph partitioning, planted random $k-$SAT, and Goldreich's planted CSP. Planted satisfiability can be viewed as a $k-$uniform hypegraph stochastic block model. The corresponding reduction to BSBM is characterized by a high imbalance between its two dimensions. For instance, one dimension is n while the other is n^{r-1}, where n is the number of

boolean literals and r (that can be large) is the distribution complexity of the model that we define later.

1.1 Definition of Bipartite Stochastic Block Model

Let n_{1+}, n_{1-}, n_{2+} and n_{2-} be four integers such that $n_1 := n_{1+} + n_{1-} \leq n_{2+} + n_{2-} := n_2$, where $n_1 \geq 2$, $n_2 \geq 2$, and let $\delta \in (0, 2)$, $p \in (0, 1/2)$. Consider two sets of vertices V_1 and V_2 such that:

- V_1 is composed of n_{1+} vertices with label $+1$ and of n_{1-} vertices with label -1;
- V_2 is composed of n_{2+} vertices with label $+1$ and of n_{2-} vertices with label -1.

We denote by $\sigma(u) \in \{-1, 1\}$ the label corresponding to vertex u. We call $|n_{1+} - n_{1-}|/n_1$ (respectively, $|n_{2+} - n_{2-}|/n_2$) the imbalance of the set V_1 (respectively, V_2). In what follows, it is assumed that there exist $\gamma_i \in (0, 1), i = 1, 2$, such that

$$|n_{1+} - n_{1-}|/n_1 \leq \gamma_1, \qquad |n_{2+} - n_{2-}|/n_2 \leq \gamma_2. \tag{3.1}$$

Let A denote the biadjacency matrix, i.e., a rectangular matrix of size $n_1 \times n_2$ whose entries A_{ij} take value 1 if the two corresponding vertices $i \in V_1$ and $j \in V_2$ are connected and take value $A_{ij} = 0$ otherwise.

We say that matrix A is drawn according to the $BSBM(\delta, n_{1+}, n_{1-}, n_{2+}, n_{2-}, p)$ model if the entries A_{ij} are independent and

- $A_{ij} \sim Ber(\delta p)$ if $\sigma(i) = \sigma(j)$, i.e., two vertices $i \in V_1$, $j \in V_2$ with the same label are connected with probability δp;
- $A_{ij} \sim Ber((2-\delta)p)$ if $\sigma(i) \neq \sigma(j)$, i.e., two vertices $i \in V_1$, $j \in V_2$ with different labels are connected with a probability $(2 - \delta)p$.

Here, $Ber(q)$ denotes the Bernoulli distribution with parameter $q \in (0, 1)$.

In this definition, p is proportional to the overall edge density. The Bipartite SBM is a generalization of the SBM in the sense that we obtain the SBM if $V_1 = V_2$. Another possible definition of BSBM is obtained by fixing only n_1 and n_2 and letting $n_{1+}, n_{1-}, n_{2+}, n_{2-}$ be random variables such that the expectations of n_{i+} and n_{i-} are both equal to $n_i/2$ for $i = 1, 2$ (then the partitions are called balanced). This is the case when the labels are independent Rademacher random variables as assumed, for example, in the previous work [Feldman et al., 2015, Florescu and Perkins, 2016].

1.2 Recovery of communities

Assume that we observe a biadjacency matrix A drawn according to a $BSBM(\delta, n_{1+}, n_{1-}, n_{2+}, n_{2-}, p)$ model. We consider the problem of recovering the node partition associated with V_1 from the observation of A. Denote by $\eta_1 \in \{\pm 1\}^{n_1}$ the vector of vertex labels in V_1. Recovering the node partition of V_1 is equivalent to retrieving either η_1 or $-\eta_1$.

As estimators of η_1 we consider any measurable functions $\hat{\eta}$ of A taking values in $\{\pm 1\}^{n_1}$. We characterize the loss of any such estimator $\hat{\eta}$ by the ℓ_1-distance between $\hat{\eta}$ and η_1, that is, by twice the number of positions at which $\hat{\eta}$ and η_1 differ:

$$\|\hat{\eta} - \eta_1\|_1 := \sum_{i=1}^{n_1} |\hat{\eta}_i - \eta_{1i}| = 2 \sum_{i=1}^{n_1} \mathbb{1}\left(\hat{\eta}_i \neq \eta_{1i}\right),$$

where $\hat{\eta}_i$ and η_{1i} denote the ith components of $\hat{\eta}$ and η_1, respectively. Since for community detection it is enough to determine either η_1 or $-\eta_1$ we consider the loss

$$r(\eta_1, \hat{\eta}) = \min_{\nu \in \{-1,1\}} \|\hat{\eta} - \nu \eta_1\|_1.$$

The performance of an estimator $\hat{\eta}$ is characterized by one of the three properties defined below. The limits in the following definitions and everywhere in the sequel are considered over sequences $n_1 \to \infty$ such that the first imbalance condition in (3.1) is satisfied for every n_1. The size of the second set of vertices n_2 need not tend to infinity and should only satisfy the second condition in (3.1). Since we consider the asymptotics as $n_1 \to \infty$ the values γ_i, p and δ are allowed to depend on n_1.

Definition 7 (weak recovery)**.** *The estimator $\hat{\eta}$ achieves weak recovery of η_1 if there exists $\alpha \in (0,1)$ such that*

$$\lim_{n_1 \to \infty} \sup_{BSBM} \mathbb{P}\left(\frac{r(\eta_1, \hat{\eta})}{n_1} \geq \alpha\right) = 0, \tag{3.2}$$

where $\sup_{BSBM}$ denotes the maximum over all distributions of A drawn from $BSBM(\delta, n_{1+}, n_{1-}, n_{2+}, n_{2-}, p)$.

If the communities in V_1 are balanced weak recovery with small α can be interpreted as the fact that $\hat{\eta}$ recovers the vertices better than chance. However, if there is a strong imbalance, Defintion 7 does not necessarily characterize good estimators as one can achieve weak recovery with small α using a trivial estimator that assigns all vertices to one community.

In this paper, the property stated in Definition 7 is not of interest on its own but rather as an auxiliary fact that we need to prove exact recovery. Namely, the initialization of the algorithm proposed below should satisfy the property stated in Definition 7.

Definition 8 (almost full recovery)**.** *The estimator $\hat{\eta}$ achieves almost full recovery of η_1 if for all $\alpha \in (0,1)$ we have*

$$\lim_{n_1 \to \infty} \sup_{BSBM} \mathbb{P}\left(\frac{r(\eta_1, \hat{\eta})}{n_1} \geq \alpha\right) = 0.$$

Almost full recovery means that $\hat{\eta}$ correctly classifies almost every vertex with high probability.

Definition 9 (exact recovery)**.** *The estimator $\hat{\eta}$ achieves exact recovery of η_1 if*

$$\lim_{n_1 \to \infty} \inf_{BSBM} \mathbb{P}\Big(r(\eta_1, \hat{\eta}) = 0\Big) = 1.$$

Exact recovery means that $\hat{\eta}$ correctly classifies all the vertices with high probability.

Notation

We will use the following notation. For given sequences a_n and b_n, we write that $a_n = O(b_n)$ (respectively, $a_n = \Omega(b_n)$) if there is an absolute constant c such that $a_n \leq cb_n$ (respectively, $a_n \geq cb_n$). We write $a_n \asymp b_n$ if $a_n = O(b_n)$ and $a_n = \Omega(b_n)$. For $a, b \in \mathbb{R}$, we denote by $a \vee b$ (respectively, $a \wedge b$) the maximum (respectively, minimum) of a and b. For $x, y \in \mathbb{R}^m$ for any $m \in \mathbb{N}$, we denote by $x^\top y$ the Euclidean scalar product, by $\|x\|_2$ the corresponding norm of x and by $\text{sign}(x)$ the vector of signs of the components of x. For any matrix $M \in \mathbb{R}^{m \times m}$, we denote by $\|M\|_\infty$ its spectral norm. Further, $\mathbf{I}_m$ denotes the $m \times m$ identity matrix and $\mathbb{1}(\cdot)$ denotes the indicator function. We denote by c positive constants that may vary from line to line.

2 Reduction to a spiked model

The biadjacency matrix A can be written as

$$A = \mathbb{E}(A) + W$$

where A is observed, $\mathbb{E}(A)$ is interpreted as the signal, and $W := A - \mathbb{E}(A)$ as the noise. It is easy to check that

$$\mathbb{E}(A) = p\mathbb{1}_{n_1}\mathbb{1}_{n_2}^\top + (\delta - 1)p\eta_1\eta_2^\top, \tag{3.3}$$

where $\mathbb{1}_{n_1}$ (respectively, $\mathbb{1}_{n_2}$) is the vector of ones with dimension n_1 (respectively, n_2) and η_1, η_2 are the vectors of labels corresponding to the sets of vertices V_1 and V_2, respectively. The second component on the right hand side of (3.3) contains information about the vector η_1 that we are interested in, while the first component $p\mathbb{1}_{n_1}\mathbb{1}_{n_2}^\top$ is non-informative about the labels. Assuming parameter p to be known we can simply subtract this component from A. From an adaptive perspective, one way to eliminate the non-informative component is by getting an estimator $\hat{p}$ of p, then considering $A - \hat{p}\mathbb{1}_{n_1}\mathbb{1}_{n_2}^\top$ as the new data matrix. Another way to disregard this component is to assume, as in [Feldman et al., 2015, Florescu and Perkins, 2016], that the partitions are balanced, which implies the orthogonality of $\mathbb{1}_{n_i}$ and η_i for $i = 1, 2$. This assures that η_1 and η_2 are the singular vectors of $\mathbb{E}(A)$ corresponding to the second largest singular value, which makes it possible to recover them with suitable accuracy from the observation of A.

In this paper, we follow the first approach where we estimate p by

$$\hat{p} = \frac{1}{n_1 n_2} \mathbb{1}_{n_1}^\top A \mathbb{1}_{n_2}. \tag{3.4}$$

Then we consider the corrected adjacency matrix

$$\hat{A} := A - \hat{p}\mathbb{1}_{n_1}\mathbb{1}_{n_2}^\top = (\delta - 1)p\eta_1\eta_2^\top + \underbrace{W + (p - \hat{p})\mathbb{1}_{n_1}\mathbb{1}_{n_2}^\top}_{noise}. \tag{3.5}$$

This is a special case of spiked matrix model where the underlying signal and the noise have a particular structure. In the rest of this paper, we assume that the observed matrix $\hat{A}$ is of the form (3.5).

A well-known approach to community detection is the spectral approach, i.e., clustering according to the signs of the entries of eigenvectors or singular vectors of the adjacency matrix or its modified version. In our case, η_1 is the left singular vector associated with the largest singular value of the signal matrix $(\delta - 1)p\eta_1\eta_2^\top$. Since $\mathbb{E}(\hat{A})$ is unknown – only $\hat{A}$ is observed – a natural algorithm for recovering η_1 would, at first sight, consist in computing the left singular vector of $\hat{A}$ corresponding to the top singular value and then taking the signs of the entries of this vector as estimators of the entries of η_1. However, such a method provides a good estimator of η_1 only if the top singular value $(\delta - 1)p$ of the signal matrix is much larger than the spectral norm of the noise term in (3.5) that is dominated by the spectral norm of W under mild assumptions on the imbalance $\gamma_1\gamma_2$. As noticed in [Florescu and Perkins, 2016], this approach suffers from a strict deterioration of sufficient conditions of recovery when n_2 grows larger than n_1. The problem can be avoided by applying the spectral approach to *hollowed matrix* $H(\hat{A}\hat{A}^\top)$ rather than to $\hat{A}$, where $H : \mathbb{R}^{n_1 \times n_1} \to \mathbb{R}^{n_1 \times n_1}$ is the linear operator defined by the relation

$$H(M) = M - \mathrm{diag}(M), \ \forall M \in \mathbb{R}^{n_1 \times n_1}.$$

Here, $\mathrm{diag}(M)$ is a diagonal matrix with the same diagonal as M. The corresponding spectral estimator of η_1 is

$$\eta_1^0 = \mathrm{sign}(\hat{v}), \tag{3.6}$$

where $\hat{v}$ is the eigenvector corresponding to the top eigenvalue of $H(\hat{A}\hat{A}^\top)$. We will further refer to η_1^0 as *spectral procedure on hollowed matrix*. The properties of η_1^0 are studied in Section 5. In particular, we show that η_1^0 achieves almost full recovery under milder conditions than previously established in [Florescu and Perkins, 2016] for a different method called the diagonal deletion SVD. However, it is not known whether η_1^0 can achieve exact recovery.

In order to grant exact recovery, we propose a new estimator. Namely, we run the sequence of iterations $(\hat{\eta}^k)_{k \geq 1}$ defined by the recursion

$$\hat{\eta}^{k+1} = \operatorname{sign}\left(H(\hat{A}\hat{A}^\top)\hat{\eta}^k\right), \qquad k = 0, 1, \ldots, \tag{3.7}$$

with the spectral estimator as initializer: $\hat{\eta}^0 = \eta_1^0$. Our final estimator is $\hat{\eta}^m$ with $m > \frac{\log n_1}{2\log 2} - \frac{3}{2}$. We call this procedure the *hollowed Lloyd's algorithm*. It is inspired by Lloyd's iterations, whose statistical guarantees were studied in the context of SBM and Gaussian Mixture Models by [Lu and Zhou, 2016]. More recently, this approach was used in [Ndaoud, 2018] to derive sharp optimal conditions for exact recovery in the Gaussian Mixture Model. It follows from those papers that the issue of proper initialization of Lloyd's algorithm is essential. The question of proving optimality of recovery by Lloyd's algorithm under random initialization is still open, both in the Gaussian Mixture Model and in the BSBM model.

3 Related work

While the literature about the classical SBM abounds (we refer to the paper [Abbe et al., 2015] and references therein), fewer results are known about the Bipartite SBM. Papers [Zhou and Amini, 2019, Zhou and Amini, 2020, Neumann, 2018] consider more general BSBM settings than ours. Being specified to our setting, their results guarantee consistency for clustering under conditions not covering the high-dimensional regime $n_2 \gg n_1$. In particular, paper [Zhou and Amini, 2019] shows that consistency can be achieved by spectral clustering on an appropriately regularized adjacency matrix when $n_2 \asymp n_1$. As an example of limitations used in [Zhou and Amini, 2020], we refer to the main theorem in [Zhou and Amini, 2020] (Theorem 1) that requires $p^2 = O(n_1/n_2^2)$ in our setting (cf. assumption (A3) in [Zhou and Amini, 2020]). This assumption combined with the necessary condition for weak recovery $p^2 = \Omega((n_1 n_2)^{-1})$ only allows for values of n_1, n_2 such that $n_2 = O(n_1^2)$. In [Neumann, 2018], the focus is on handling mutiple and possibly overlapping clusters. The recovery conditions from [Neumann, 2018] being specified to our setting (two non-overlapping clusters) are far from optimal.

On the other hand, papers [Feldman et al., 2015, Florescu and Perkins, 2016, Cai et al., 2019] study the problem on finding proper thresholds for p under conditions covering the high-dimensional regime $n_2 \gg n_1$. In particular, [Florescu and Perkins, 2016] proves that the sharp phase transition for the weak recovery problem occurs around the critical probability $p_c = \frac{(\delta-1)^{-2}}{\sqrt{n_1 n_2}}$. The sufficient condition in this case is based on a reduction to SBM then using any optimal "black-box" algorithm for detection in the SBM as in [Bordenave et al., 2015, Massoulié, 2014, Mossel et al., 2018].

For the problem of exact recovery, [Feldman et al., 2015] obtained what we will further call the state of the art sufficient conditions. Namely, using the Subsampled Power Iteration algorithm, [Feldman et al., 2015] shows that the condition $p = \Omega\left(\frac{(\delta-1)^{-2}\log n_1}{\sqrt{n_1 n_2}}\right)$ is sufficient to achieve exact recovery. Although no necessary

condition for this property is known, it is conjectured in [Feldman et al., 2015] that $p = \Omega\left(\frac{\log n_1}{\sqrt{n_1 n_2}}\right)$ is necessary for exact recovery. Our results below disprove this conjecture.

Spectral algorithms for BSBM were investigated in [Florescu and Perkins, 2016, Cai et al., 2019]. In particular, [Florescu and Perkins, 2016] compared sufficient conditions of almost full recovery for the classical SVD algorithm and for the diagonal deletion SVD. It was shown in [Florescu and Perkins, 2016] that, in the high-dimensional regime $n_2 \gg n_1$, the diagonal deletion SVD provides a strict improvement over the classical SVD. One way to explain this improvement is by observing that, in this regime, the spectral norm of the expectation of the noise term $WW^\top$ is much larger than its deviation. It was proved in [Florescu and Perkins, 2016] that $p = \Omega\left(\frac{\log n_1}{\sqrt{n_1 n_2}}\right)$ is sufficient to achieve almost full recovery through the diagonal deletion SVD algorithm. Note that [Feldman et al., 2015] proved that, under similar conditions, the Subsampled Power Iteration algorithm achieves a better result, i.e., it provides exact recovery rather than almost full recovery. The most recent paper [Cai et al., 2019] parallel to our work shows that the diagonal deletion SVD also upgrades from almost full to exact recovery under the conditions that are analogous to [Florescu and Perkins, 2016] for moderate $n_2 \geq n_1$ but deteriorate for very large n_2 (for example, if $n_2 \asymp e^{n_1}$). The results of [Feldman et al., 2015, Florescu and Perkins, 2016, Cai et al., 2019] are summarized in Table 1.

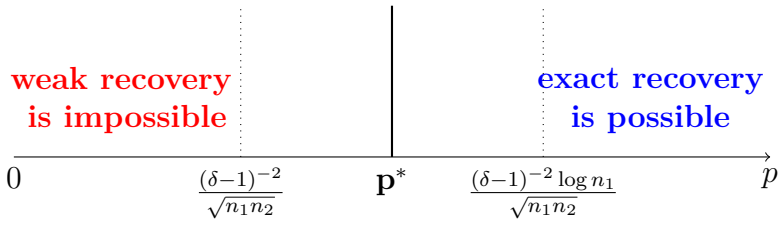
Ref.	Results	Conditions	Algorithm
[Feldman et al., 2015]	Exact recovery	$\begin{cases} \text{known } p,\ n_2 \geq n_1, \\ p \geq C(\delta-1)^{-2}\frac{\log n_1}{\sqrt{n_1 n_2}} \end{cases}$	Subsampled iterations
[Cai et al., 2019]	Exact recovery	$\begin{cases} \text{known } p, \\ p \geq C(\delta-1)^{-2}\left(\frac{\log(n_1+n_2)}{\sqrt{n_1 n_2}} \vee \frac{\log(n_1+n_2)}{n_2}\right) \end{cases}$	Diagonal deletion SVD
[Florescu and Perkins, 2016]	Almost full recovery	$\begin{cases} \text{unknown } p,\ n_2 \geq n_1(\log n_1)^4,\ \gamma_1=\gamma_2=0, \\ p \geq C_\delta \frac{\log n_1}{\sqrt{n_1 n_2}} \end{cases}$	Diagonal deletion SVD
[Florescu and Perkins, 2016]	Weak recovery	$\begin{cases} \text{unknown } p,\ n_2 \geq n_1,\ \gamma_1=\gamma_2=0, \\ p > \frac{(\delta-1)^{-2}}{\sqrt{n_1 n_2}} \end{cases}$	SBM reduction

Table 3.1: Summary of the results of [Feldman et al., 2015, Florescu and Perkins, 2016, Cai et al., 2019]. Here, $C_\delta > 0$ is a constant depending on δ. In this table, condition $\gamma_1 = \gamma_2 = 0$ means that the labels are independent Rademacher random variables.

We emphasize that [Feldman et al., 2015, Florescu and Perkins, 2016] focus on the regime $n_2 \geq n_1$. It will be also the main challenge in the present paper even though our results are valid for all n_1, n_2 with no restriction. There are two reasons for that:

- The high-dimensional regime $n_2 \gg n_1$ is of dominant importance in the applications of bipartite SBM.
- The high-dimensional regime is challenging from the theoretical point of view. The case $n_2 \leq n_1$ is more direct to handle as it can be solved similarly to standard SBM. Indeed, optimal results for $n_2 \leq n_1$ are achieved by SVD type algorithms applied to the adjacency matrix A (cf. [Zhou and Amini, 2019, Zhou and Amini, 2020]). They are based on a control of the spectral norm of W. While the behavior of the spectral norm of W is well understood (cf. [Bandeira and Van Handel, 2016]), existing results for the spectral norms of $WW^\top - \mathbb{E}(WW^\top)$ or of $H(WW^\top)$ that one needs to control when $n_2 \gg n_1$ turn out to be suboptimal. It makes the regime $n_2 \gg n_1$ quite challenging.

Under the condition $n_2 \geq n_1 \log n_1$, the state of the art results can be summarized by the following diagram leaving open the optimal value $p = \mathbf{p}^*$ at which exact recovery can be achieved.



A related recent line of work developed optimal clustering algorithms for Gaussian Mixture Models (GMM) [Lu and Zhou, 2016, Giraud and Verzelen, 2019, Ndaoud, 2018, Loffler et al., 2019]. It was shown in [Lu and Zhou, 2016] that clustering with optimality properties in GMM can be achieved by an iterative algorithm analogous to Lloyd's procedure. Moreover, [Ndaoud, 2018] proved that a version of such iterative clustering algorithm attains the sharp phase transition for exact recovery in those models. Based on an analogy between the GMM and the BSBM, it is conjectured in [Ndaoud, 2018] that similar algorithms can achieve almost full recovery and exact recovery in bipartite graph models. Namely, comparing the first two moments of the matrices arising in the two models one may expect $p = \Omega\left((\delta-1)^{-2}\sqrt{\frac{\log n_1}{n_1 n_2}}\right)$ to be sufficient to achieve exact recovery in the BSBM, provided that $n_2 \geq n_1 \log n_1$. This heuristics suggests a logarithmic improvement over the state of the art condition presented in the diagram above. More interestingly, it goes against another, seemingly more natural, heuristics based on an analogy with standard SBM and conjecturing the right recovery condition in the form $p = \Omega\left(\frac{\log n_1}{\sqrt{n_1 n_2}}\right)$ (cf. [Feldman et al., 2015]). We show below that, surprisingly, the analogy with GMM and not with SBM (however, the "closest parent" of BSBM) appears to be correct.

Finally, some consequences were obtained for planted satisfiability problems. Reduction of those problems to BSBM allows one to get sufficient conditions of complete recovery of the planted assignment. We refer to [Feldman et al., 2015] for the details of this reduction. Namely, it is shown in [Feldman et al., 2015] that considering a planted satisfiability problem is equivalent to considering a BSBM where $n_1 = n$ and $n_2 = n^{r-1}$, with n and $r \geq 2$ defined below. For any satisfiability problem, we are interested in m, which is the sufficient number of k-clauses from C_k in order to recover completely the planted assignment σ. Here, C_k is the set of all ordered k-tuples of n literals $x_1, \ldots, x_n$ and their negations with no repetition of variables. For a k-tuple of literals C and an assignment $\sigma \in \{-1, +1\}^n$, $\sigma(C)$ denotes the vector of values that σ assigns to the literals in C. Given a planting distribution $Q : \{-1, +1\}^k \to [0, 1]$, and an assignment σ, we define the random constraint satisfaction problem $F_{Q,\sigma}(n, m)$ by drawing m k-clauses from C_k independently according to the distribution

$$Q_\sigma(C) = \frac{Q(\sigma(C))}{\sum_{C' \in C_k} Q(\sigma(C'))}.$$

A related class of problems is one in which for some fixed predicate $P : \{-1, 1\}^k \to \{-1, 1\}$, an instance is generated by choosing a planted assignment σ uniformly at random and generating a set of m random and uniform P-constraints. That is, each constraint is of the form $P(x_{i_1}, \ldots, x_{i_k}) = \mathrm{P}(\sigma_{i_1}, \ldots, \sigma_{i_k})$, where $(x_{i_1}, \ldots, x_{i_k})$ is a randomly and uniformly chosen k-tuple of variables (without repetitions).

In simpler words m plays the role of pn_1n_2 in the BSBM, and any sufficient condition on p leads to a sufficient condition for m. It was shown in [Feldman et al., 2015] that the following conditions are sufficient to achieve exact recovery in some of the satisfiability problems.

- For any planting distribution $Q : \{-1, 1\}^k \to [0, 1]$, there exists an algorithm that for any assignment $\sigma \in \{-1, 1\}^n$, given an instance of $F_{Q,\sigma}(n, m)$, completely recovers the planted assignment σ for $m = O(n^{r/2} \log n)$. Here, $r \geq 2$ is the smallest integer such that there is some $S \subseteq \{1, \ldots, k\}$ with $|S| = r$, for which the discrete Fourier coefficient $\hat{Q}(S)$ is non-zero.

- For any predicate $P : \{-1, 1\}^k \to \{-1, 1\}$, there exists an algorithm that for any assignment σ, given m random P-constraints, completely recovers the planted assignment σ for $m = O(n^{r/2} \log n)$ where $r \geq 2$ is the degree of the lowest-degree non-zero Fourier coefficient of P.

4 Main contributions

Our findings can be summarized as follows.

- *Exact recovery.* We present a novel method - the *hollowed Lloyd's algorithm* - that achieves exact recovery under strictly milder conditions than the state of

the art. Namely, we show that

$$p = \Omega\left(\sqrt{\frac{\log n_1}{n_1 n_2}} \vee \frac{\log n_1}{n_2}\right) \tag{3.8}$$

is sufficient to achieve exact recovery in the BSBM. Condition (3.8) exhibits an elbow at $n_2 \asymp n_1 \log n_1$ between the low-dimensional and high-dimensional regimes. In the low-dimensional regime $n_2 \leq n_1 \log n_1$, it takes the form $p = \Omega\left(\frac{\log n_1}{n_2}\right)$, the same as the sufficient condition in [Cai et al., 2019], that can be shown minimax optimal using similar lower bound techniques as in [Bandeira, 2015] for SBM (clustering oracle with side information). Such a lower bound was formalized, for the Bipartite SBM, in Theorem 2 of [Zhou and Amini, 2020] under the conditions $n_1 \leq n_2$ and $n_2 = O(n_1 \log n_1)$ that correspond to assumptions $(A1) - (A4)$ from [Zhou and Amini, 2020]. On the other hand, in the high-dimensional regime $n_2 \geq n_1 \log n_1$ the sufficient condition of exact recovery (3.8) reads as $p = \Omega\left(\sqrt{\frac{\log n_1}{n_1 n_2}}\right)$. We argue that this condition is tight by showing that even a supervised oracle procedure fails to achieve exact recovery in the regime $n_2 \geq n_1 \log n_1$ if $p < c\sqrt{\frac{\log n_1}{n_1 n_2}}$ for a constant $c > 0$ small enough. Importantly, our findings imply that the condition $p = \Omega\left(\frac{\log n_1}{\sqrt{n_1 n_2}}\right)$ common for all the related work and based on the analogy with usual SBM is not necessary for exact recovery in the BSBM when $n_2 \geq n_1 \log n_1$.

- *Almost full recovery.* We provide a new sufficient condition for almost full recovery by spectral techniques using the diagonal deletion device. Our spectral estimator and its analysis are different from [Florescu and Perkins, 2016], where another diagonal deletion method was suggested. The analysis uses an adapted version of matrix Bernstein inequality applied to a sum of hollowed rank one random matrices where bounding the corresponding moments, in operator norm, involve combinatorics arguments. This leads to an improvement upon the sufficient condition of [Florescu and Perkins, 2016]. We show that, unlike in the Gaussian case, hollowing the Gram matrix yields, both theoretically and empirically, a strict improvement over debiasing, i.e., subtracting the expecation of the Gram matrix.

- The hollowed Lloyd's algorithm that we propose is computationally faster than the previously known methods. Its analysis that we develop is novel and makes it possible to transform any estimator achieving weak recovery into another one achieving exact recovery. We expect this analysis to be useful to solve more general exact recovery problems for random graphs.

- In contrast to the related work, where simplifying assumptions of either zero imbalance ($\gamma_1 = \gamma_2 = 0$) as in [Florescu and Perkins, 2016] or known p as in [Feldman et al., 2015, Cai et al., 2019] were imposed, our approach is more general. In particular, our exact recovery result holds adaptively to p under a mild

assumption on $\gamma_1\gamma_2$. Notice that as $\gamma_1\gamma_2$ get closer to 1, then estimation of p gets harder. Our theoretical findings are supported by numerical experiments, where we show that our iterative procedure (with or without spectral initialization) outperforms spectral methods and achieves almost the same performance as the supervised oracle.

- Our results regarding almost full recovery based on the spectral estimator, exact recovery via the hollowed Lloyd's algorithm, and the impossibility of exact recovery via the supervised oracle are summarized in the table below.

Results	Conditions	Procedure
Almost full recovery *by spectral methods*	$\begin{cases} \text{unknown } p, \quad \gamma_1\gamma_2 \le 1/C'_{n_1} \\ p \ge C_{n_1}(\delta-1)^{-2}\left(\sqrt{\frac{\log n_1}{n_1 n_2}} \vee \frac{\log n_1}{n_2}\right) \end{cases}$	Spectral on hollowed matrix
Exact recovery	$\begin{cases} \text{unknown } p, \quad \gamma_1\gamma_2 < 1/480, \\ p \ge C(\delta-1)^{-2}\left(\sqrt{\frac{\log n_1}{n_1 n_2}} \vee \frac{\log n_1}{n_2}\right) \end{cases}$	Hollowed Lloyd's
Impossibility of exact recovery	$\begin{cases} n_2 \ge n_1 \log n_1, \ \gamma_1 = \gamma_2 = 0, \\ p < C_\delta \sqrt{\frac{\log n_1}{n_1 n_2}} \end{cases}$	Oracle

Table 3.2: Summary of our main contributions. Here, $C_\delta > 0$ is a positive constant depending on δ, $C > 0$ is an absolute constant and C_{n_1}, C'_{n_1} are any sequences such that $C_{n_1}, C'_{n_1} \to \infty$ as $n_1 \to \infty$.

- As a byproduct, we also improve upon sufficient conditions of [Feldman et al., 2015] for exact recovery in some of the satisfiability problems. Namely, our results imply the following.

  1. For any planting distribution $Q : \{-1,1\}^k \to [0,1]$, there exists an algorithm that for any assignment σ, given an instance of $F_{Q,\sigma}(n,m)$, completely recovers the planted assignment σ for $m = O(n^{r/2}\sqrt{\log n})$ where $r \ge 3$ is the smallest integer such that there is some $S \subseteq \{1,\dots,k\}$ with $|S| = r$, for which the discrete Fourier coefficient $\hat{Q}(S)$ is non-zero.
  2. For any predicate $P : \{-1,1\}^k \to \{-1,1\}$, there exists an algorithm that for any assignment σ, given m random P-constraints, completely recovers the planted assignment σ for $m = O(n^{r/2}\sqrt{\log n})$ where $r \ge 3$ is the degree of the lowest-degree non-zero Fourier coefficient of P.

5 Properties of the spectral method

In this section, we analyze the risk of the spectral initializer η_1^0. As in the case of SDP relaxations of the problem, the matrix of interest is the Gram matrix $\hat{A}\hat{A}^\top$. It is well known that it suffers from a bias that grows with n_2. In [Royer, 2017], a debiasing procedure is proposed using an estimator of the covariance of the noise. In this section, we consider a different approach that consists in removing the diagonal entries of the Gram matrix.

We give some intuition about this procedure when p is known. In this case the adjacency matrix can be replaced by

$$\tilde{A} = A - p\mathbb{1}_{n_1}\mathbb{1}_{n_2}^\top.$$

The general case follows similarly since one can show that $|\hat{p}-p|$ does not exceed the noise level arising when p is known (see the details below). The spectral norm of the expected noise matrix $\mathbb{E}(WW^\top)$ is of the order of $n_2 p$. If $n_2 \gg n_1$, which is the most interesting case in the applications, this is too large compared to the deviation, in the spectral norm, of the noise matrix from its expectation, cf. [Florescu and Perkins, 2016]. Since the expectation of the noise $WW^\top$ is a diagonal matrix, removing diagonal terms is expected to reduce the spectral norm of the noise and hence to make the recovery problem easier. Specifically, observe that the matrix $H(\tilde{A}\tilde{A}^\top)$ can be decomposed as follows:

$$H(\tilde{A}\tilde{A}^\top) = \underbrace{(\delta-1)^2 p^2 n_2 H(\eta_1\eta_1^\top)}_{signal} + \underbrace{H(WW^\top) + p(\delta-1)H(W\eta_2\eta_1^\top + \eta_1\eta_2^\top W^\top)}_{noise}. \tag{3.9}$$

It turns out that the main driver of the noise is $H(WW^\top)$. On the other hand, it is easy to see (cf., e.g., Lemma 17 in [Ndaoud, 2018]) that

$$\|H(WW^\top)\|_\infty \le 2\|WW^\top - \mathbb{E}(WW^\top)\|_\infty \tag{3.10}$$

for any random matrix W with independent columns. This shows that removing the diagonal terms is a good candidate to remove the bias induced by the noise. Thus, diagonal deletion can be viewed as an alternative to debiasing of the Gram matrix. Nevertheless, the operator $H(\cdot)$ may affect dramatically the signal. Fortunately, it does not happen in our case; the signal term is almost insensitive to this operation since it is a rank one matrix. In particular, we have:

$$\|H(\eta_1\eta_1^\top)\|_\infty = \left(1 - \frac{1}{n_1}\right)\|\eta_1\eta_1^\top\|_\infty.$$

Thus, as n_1 grows, the signal does not get affected by removing its diagonal terms while we get rid of the bias in the noise term. This motivates the spectral estimator

η_1^0 defined by (3.6), where $\hat{v}$ is the eigenvector corresponding to the top eigenvalue of $H(\hat{A}\hat{A}^\top)$. The next theorem gives sufficient conditions for the estimator η_1^0 to achieve weak and almost full recovery.

Theorem 1. *Let η_1^0 be the estimator given by (3.6) with $\hat{p}$ defined in (3.4) and let $\alpha \in (0,1)$. Let $(C_{n_1}), (C'_{n_1})$ be sequences of positive numbers that tend to infinity as $n_1 \to \infty$.*

(i) Let the following conditions hold:

$$\begin{cases} \gamma_1\gamma_2 \le \sqrt{\alpha}/96, \\ p \ge C(\delta-1)^{-2}\left(\sqrt{\frac{\log n_1}{n_1 n_2}} \vee \frac{\log n_1}{n_2}\right), \end{cases}$$

where $C > C_0/\sqrt{\alpha}$ for an absolute constant $C_0 > 0$ large enough. Then the estimator η_1^0 satisfies (3.2).

(ii) Let the following conditions hold:

$$\begin{cases} \gamma_1\gamma_2 \le 1/C'_{n_1}, \\ p \ge C_{n_1}(\delta-1)^{-2}\left(\sqrt{\frac{\log n_1}{n_1 n_2}} \vee \frac{\log n_1}{n_2}\right). \end{cases}$$

Then the estimator η_1^0 achieves almost full recovery of η_1.

Part (i) of Theorem 1 establishes the property of spectral initializer η_1^0 that we need to prove the exact recovery in Theorem 2 below. Part (ii) of Theorem 1 improves upon the existing sufficient conditions of almost full recovery *by spectral methods* [Florescu and Perkins, 2016], cf. Table 1 above. Theorem 1 covers any n_1, n_2 with no restriction, and scales as $\sqrt{\frac{\log n_1}{n_1 n_2}}$ rather than $\frac{\log n_1}{\sqrt{n_1 n_2}}$ in the regime $n_2 \ge n_1 \log n_1$.

The proof of Theorem 1 is given in Section 10. It is based on a variant of matrix Bernstein inequality applied to a sum of independent hollowed rank one random matrices (Theorem 4). As a consequence of this new matrix concentration result, we have the following improved bound for the spectral norm of the noise term.

Proposition 1. *Assume that $p \ge C\left(\sqrt{\frac{\log n_1}{n_1 n_2}} \vee \frac{\log n_1}{n_2}\right)$ for some constant $C > 0$. Then, there exists a constant $c_* > 0$ such that*

$$\mathbb{E}\left(\left\|H\left(WW^\top\right)\right\|_\infty^2\right) \le c_*\left(1 + \frac{n_1 \log n_1}{n_2}\right) n_1 n_2 p^2 \log n_1.$$

On the other hand, for the non-hollowed matrix $WW^\top$, using matrix Bernstein inequality, we can only obtain that

$$\mathbb{E}\left(\left\|WW^\top\right\|_\infty^2\right) = O\Big((n_1+n_2)^2 p^2 (\log n_1)^2\Big).$$

Comparing the above two bounds explains why our hollowed spectral method is superior to the standard SVD procedure even in the low-dimensional regime $n_2 = O(n_1 \log n_1)$.

Our next point is to explain why applying the hollowing operator $H(\cdot)$ is better than debiasing by subtraction of $\mathbb{E}(WW^\top)$. Inequality (3.10) is useful to bound the spectral norm of the hollowed Gram matrix, under the Gaussian Mixture Model (cf. [Ndaoud, 2018]). What is more, one can show that (3.10) is tight when the noise is isotropic and normal, suggesting that hollowing and debiasing are almost equivalent in the Gaussian Mixture Model. Suprisingly, the same inequality turns out to be loose in the the BSBM model. It turns out that hollowing the Gram matrix can be strictly better. Indeed, the next proposition shows that debiasing the Gram matrix through covariance subtraction can be suboptimal.

Proposition 2. *Let* $n_2 \geq n_1 \log n_1$ *and*

$$18\sqrt{\frac{\log n_1}{n_1 n_2}} \leq p \leq \frac{1}{206\, n_1 \log n_1}. \tag{3.11}$$

Then

$$\mathbb{E}\left(\|WW^\top - \mathbb{E}(WW^\top)\|_\infty^2\right) \geq \frac{n_2 p}{40}.$$

Proposition 2 deals with the high-dimensional regime $n_2 \geq n_1 \log n_1$ under the additional restriction $n_1(\log n_1)^3 = O(n_2)$ that follows from condition (3.11). Notice that for smaller p satisfying (3.11), we have $n_1 n_2 p^2 \log n_1 = o(n_2 p)$, so that inequality (3.10) is loose. This explains the suboptimality of debiased spectral estimator. We further check this fact through simulations in Section 11. Proposition 2 also explains why the result in [Cai et al., 2019] is suboptimal. Indeed, the assumptions in [Cai et al., 2019] are such that the spectral norm of matrix $\mathrm{diag}\Big(WW^\top - \mathbb{E}(WW^\top)\Big)$ (that scales as $\sqrt{n_2 p}$) is not bigger in order of magnitude than the spectral norm of the corresponding off-diagonal matrix. While this fact is true in several other settings, it is not in the high-dimensional regime of bipartite clustering, cf. Proposition 2.

The question of whether the spectral estimator η_1^0 can achieve exact recovery under the conditions of Theorem 1 remains open. Pursuing similar arguments as developed in [Abbe et al., 2017] for the case of SBM would lead to a logarithmic dependence of order $\log n_1$ or bigger in the sufficient condition (as it is the case in [Cai et al., 2019]), and not to the desired $\sqrt{\log n_1}$. By analogy to the Gaussian Mixture Model, for which it was shown recently in [Abbe et al., 2020] that the spectral estimator is optimal for exact recovery, we conjecture that the condition $p > C(\delta - 1)^{-2}\sqrt{\frac{\log n_1}{n_1 n_2}}$ is sufficient for η_1^0 to achieve exact recovery whenever $n_2 \geq n_1 \log n_1$. Proving such a result would most likely require developing novel concentration bounds for Bernoulli covariance matrices.

6 Exact recovery by the hollowed Lloyd's algorithm

In this section, we present sufficient conditions, under which the hollowed Lloyd's algorithm $(\hat{\eta}^k)_{k\geq 0}$ defined in (3.7) with spectral initialization achieves exact recovery for all k large enough.

Theorem 2. *Let $(\hat{\eta}^k)_{k\geq 0}$ be the recursion (3.7) initialized with the spectral estimator (3.6) for $\hat{p}$ given by (3.4). There exists an absolute constant $C > 0$ such that if the following conditions hold:*

$$\begin{cases} \gamma_1\gamma_2 \leq 1/480, \\ p \geq C(\delta-1)^{-2}\left(\sqrt{\frac{\log n_1}{n_1 n_2}} \vee \frac{\log n_1}{n_2}\right), \end{cases}$$

then the estimator $\hat{\eta}^m$ with $m = m(n_1) > \frac{\log n_1}{2\log 2} - \frac{3}{2}$ achieves exact recovery of η_1.

Some comments are in order here.

1. The approach that we developed to construct $\hat{\eta}^m$ is general. In fact, it is a tool that transforms any estimator achieving weak recovery into a new estimator achieving exact recovery under mild assumptions. This can be readily seen from the proof of Theorem 2.

2. Numerically, the procedure $(\hat{\eta}^k)_{k\geq 0}$ considered in Theorem 2 has the same complexity as the spectral initializer η_1^0. It remains an open question whether the result of Theorem 2 holds with random initialization, which would further bring down the complexity.

3. We conjecture that the conditions $p \geq C(\delta-1)^{-2}\sqrt{\frac{\log n_1}{n_1 n_2}}$ and $n_2 > n_1 \log n_1$ of Theorem 2 cannot be improved. In the next section, we provide a result supporting this fact. The imbalance condition $\gamma_1\gamma_2 = O(1)$ is only required to handle the estimation of p. If p is known the results of this paper remain valid with no assumption on γ_1 and γ_2.

7 Impossibility result for a supervised oracle

Motivated by the spiked reduction of the BSBM model when p is known, we define the supervised oracle as follows

$$\tilde{\eta}_1 = \operatorname{sign}(H(\tilde{A}\tilde{A}^\top)\eta_1). \tag{3.12}$$

Note that this oracle is extremely powerful. We set the definition of the oracle in a compact form using the hollowed matrix $H(\cdot)$ for the purpose of shorter writing. However, if one unfolds this definition, it turns out that the oracle makes a decision about one vertex by using the majority vote of all the other vertices. In other words,

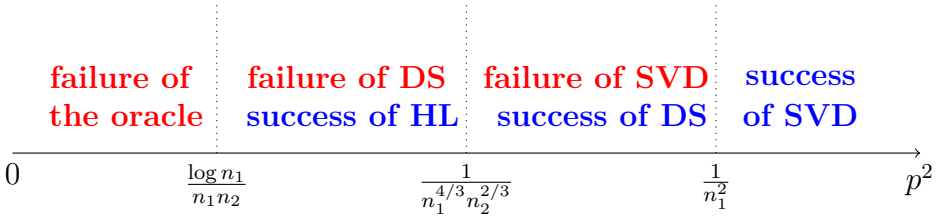
this oracle has access to *all except one* labels and uses these known labels to predict the remaining one. Specifically, for each label η_{1i} to estimate, the supervised oracle has access to the remaining labels $(\eta_{1j}, j \neq i)$ and to p. We refer the reader to [Ndaoud, 2018] for more discussion about such an oracle structure.

We state below an impossibility result corresponding to the supervised oracle.

Proposition 3. *Assume that* $n_2 \geq n_1 \log n_1$ *and* $\gamma_1 = \gamma_2 = 0$. *There exists* $c_\delta > 0$ *depending only on* δ *such that if* $p = \sqrt{c_\delta \frac{\log n_1}{n_1 n_2}}$ *then for the oracle* $\tilde{\eta}_1$ *we have*

$$\lim_{n_1 \to \infty} \sum_{i=1}^{n_1} \mathbb{P}(\tilde{\eta}_{1i} \neq \eta_{1i}) = \infty.$$

Proposition 3 shows that condition $p = \Omega\left(\sqrt{\frac{\log n_1}{n_1 n_2}}\right)$ is necessary for the supervised oracle to achieve exact recovery when $n_2 \geq n_1 \log n_1$. Combining this result with the sufficient conditions for exact recovery from Theorem 2, we can now complete the diagram in [Florescu and Perkins, 2016], which compares the exact recovery conditions for SVD and for the debiased spectral method when $n_2 \geq n_1 (\log n_1)^4$. We recall here that the SVD estimator is the one returning signs of the second eigenvector of $AA^\top$. In [Florescu and Perkins, 2016], a debiased spectral method is also considered, which uses as an estimator the signs of the second eigenvector of $AA^\top - \mathbb{E}(WW^\top)$. Under perfect balance (that is, $\gamma_1 = \gamma_2 = 0$), $\mathbb{E}(WW^\top)$ is proportional to $\mathbf{I}_{n_1}$ and hence SVD and debiased spectral method coincide in that case, while in general the debiased spectral method outperforms the SVD estimator. Comparison of the oracle and of the three methods: SVD, debiased spectral (DS) and hollowed Lloyd's (HL), in the general case of imbalance, and under the condition $n_2 \geq n_1 (\log n_1)^4$, can be summarized as follows :



This hierarchy of procedures becomes apparent in the simulations given in the next section.

8 Control of the spectral norm of the hollowed Gram matrix

This section is devoted to the control of the spectral norm of the hollowed matrix $H(WW^\top) = \sum_{j=1}^{n_2} H(W_j W_j^\top)$, where we denote by W_j the columns of W. The following theorem will be used in the proofs.

Theorem 3 (Matrix Bernstein inequality – adapted from [Tropp, 2012], Theorem 6.2)**.** *Let* $(Y_j)_{j=1}^n$ *be a sequence of independent symmetric random matrices of size*

$d \times d$, and $a, R > 0$. Assume that for all j in $\{1, \dots, n\}$ we have

$$\mathbb{E}(Y_j) = 0 \text{ and } \|\mathbb{E}(Y_j^q)\|_\infty \le \frac{q!}{2} R^{q-2} a^2 \text{ for } q = 2, 3, \dots .$$

Then, for all $t \ge 0$,

$$\mathbb{P}\left(\Big\| \sum_{j=1}^{n} Y_j \Big\|_\infty \ge t \right) \le d \exp\left(-\frac{t^2}{2\sigma^2 + 2Rt} \right) \text{ with } \sigma^2 = na^2.$$

We will show that in our case this theorem can be applied with $Y_j = H(W_j W_j^\top)$, $d = n_1$, $n = n_2$, $R = 3(1 + 2n_1 p)$ and $a^2 = 4p^2 n_1$. One can check that it gives a strict improvement over the matrix Hoeffding type inequality that uses only the fact that $\|H(W_j W_j^\top)\|_\infty \le n_1$ almost surely. Namely, we have the following theorem.

Theorem 4. *For all $t \ge 0$,*

$$\mathbb{P}\left(\Big\| \sum_{j=1}^{n_2} H(W_j W_j^\top) \Big\|_\infty \ge t \right) \le n_1 \exp\left(-\frac{t^2}{8n_1 n_2 p^2 + 6(1 + 2n_1 p)t} \right).$$

Proof. Fix j in $\{1, \dots, n_2\}$. In view of Theorem 3, it is enough to show that for all integers $q \ge 2$ we have

$$\Big\| \mathbb{E}(H(W_j W_j^\top)^q) \Big\|_\infty \le 2q!(3(1 + 2n_1 p))^{q-2} p^2 n_1. \tag{3.13}$$

We now prove (3.13). To alleviate the notation, we set $w = W_j$ and we denote by w_k the entries of w. Note that w_k are independent random variables taking value $1 - \mathsf{p}$ w.p. p and $-\mathsf{p}$ w.p. $1 - \mathsf{p}$, where p is either δp or $(2 - \delta)p$. We have $\mathbb{E}(w_k) = 0$ for all k. Furthermore, for any integer $m \ge 2$,

$$\|\mathbb{E}(w_k^m)\|_1 \le 2p. \tag{3.14}$$

Indeed,

$$\|\mathbb{E}(w_k^m)\|_1 \le \mathsf{p}(1 - \mathsf{p}) \max_{0 \le \mathsf{p} \le 1} ((1 - \mathsf{p})^{m-1} + \mathsf{p}^{m-1}) = \mathsf{p}(1 - \mathsf{p}).$$

Denote by $h_{ik}(q)$ the (i,k)th entry of matrix $H(ww^\top)^q$. Note that the (i,k)th entry of matrix $H(ww^\top)$ is $H(ww^\top)_{ik} = w_i w_k \mathbb{1}(i \ne k)$. It comes out that

$$h_{ik}(q) = \sum_{(i_2, i_3, \dots, i_q) \in J} w_i w_k \prod_{\ell=2}^{q} w_{i_\ell}^2,$$

where $J = \{(i_2, i_3, \dots, i_q) : i_2 \ne i_3, \dots, i_{q-1} \ne i_q; i_2 \ne i, i_q \ne k\}$ and indices i_ℓ take values in $\{1, \dots, n_1\}$. Thus,

$$\|\mathbb{E}(h_{ik}(q))\|_1 \le \sum_{(i_2, i_3, \dots, i_q) \in J} \left\| \mathbb{E}\Big(w_i w_k \prod_{\ell=2}^{q} w_{i_\ell}^2 \Big) \right\|_1. \tag{3.15}$$

First note that for $q = 2$ the terms in this sum are non-zero only if $i = k$ and in this case the sum is bounded by $4p^2 n_1$. Thus, (3.13) holds for $q = 2$. In order to prove (3.13) for $q \geq 3$, it suffices to show that for all i, k we have

$$\|\mathbb{E}(h_{ik}(q))\|_1 \leq 2q!(1 + 2pn_1)^{q-2}p^2, \quad i \neq k, \tag{3.16}$$

and

$$\|\mathbb{E}(h_{ii}(q))\|_1 \leq 4q!(1 + 2pn_1)^{q-2}p^2 n_1. \tag{3.17}$$

We start by showing (3.16) for all $q \geq 3$. Let $i \neq k$. We first bound the number of non-zero terms in the sum in (3.15). Since $w_1, \ldots, w_{n_1}$ are independent zero-mean random variables, the term in this sum corresponding to some fixed $(i_2, i_3, \ldots, i_q)$ can be non-zero only if both i and k belong to the set $\{i_2, i_3, \ldots, i_q\}$. In order to take into account equalities between different indices i_ℓ, consider all partitions π of the set $\{i_2, i_3, \ldots, i_q\}$ into s subsets, with equal indices in each subset, where s runs from 2 to $q - 1$ (the case $s = 1$, that is $i_2 = i_3 = \cdots = i_q$, is excluded since the corresponding expectation vanishes).

Assume a partition π in s subsets fixed. Then, for the expectation

$$\mathbb{E}\left(w_i w_k \prod_{\ell=2}^{q} w_{i_\ell}^2\right)$$

to be non-zero, two out of s subsets must contain variables with indices i and k, and in this case due to independence of w_m and (3.14) we have

$$\left\|\mathbb{E}\left(w_i w_k \prod_{\ell=2}^{q} w_{i_\ell}^2\right)\right\|_1 \leq (2p)^s. \tag{3.18}$$

Denote by $\mathscr{P}_{s,2}$ the set of all partitions π of $\{i_2, i_3, \ldots, i_q\}$ into s subsets such that for two of these subsets the indices i_ℓ are equal to i and k. To get an upper bound on the cardinality of $\mathscr{P}_{s,2}$, notice that any such partition can be obtained by choosing $s - 2$ distinct indices among the $q - 3$ possible values (other than i and k) and then allocating the remaining $q - s$ indices to s buckets. This leads to the bound $\mathrm{Card}(\mathscr{P}_{s,2}) \leq \binom{q-3}{s-2} s^{q-s}$. Denote by $i_1(\pi) \neq \cdots \neq i_{s-2}(\pi)$ the $s - 2$ distinct indices (other than i and k) corresponding to the partition $\pi \in \mathscr{P}_{s,2}$. Using (3.18) and the fact that the indices $i_\ell(\pi)$ can take values from 1 to n_1 we obtain

$$\begin{aligned}
\|\mathbb{E}(h_{ik}(q))\|_1 &\leq \sum_{s=2}^{q-1} \sum_{\pi \in \mathscr{P}_{s,2}} \sum_{i_1(\pi) \neq \cdots \neq i_{s-2}(\pi)} (2p)^s \\
&\leq \sum_{s=2}^{q-1} \binom{q-3}{s-2} s^{q-s} n_1^{s-2} (2p)^s \\
&\leq 2p^2 q! \sum_{s=2}^{q-1} \binom{q-2}{s-2} (2pn_1)^{s-2}
\end{aligned}$$

$$\leq 2q!(1+2pn_1)^{q-2}p^2,$$

where we have used the inequalities $s^{q-s} \leq (q-1)!/(s-1)! \leq q!/2$. Thus, the bound (3.16) is proved for all $q \geq 3$.

It remains to show that (3.17) holds for $q \geq 3$. Denote by $\mathscr{P}_{s,1}$ the set of all partitions π of $\{i_2, i_3, \ldots, i_q\}$ into s subsets such that for one of these subsets the index i_ℓ is equal to i. Similarly to the argument for $\mathscr{P}_{s,2}$, we obtain that $\mathrm{Card}(\mathscr{P}_{s,1}) \leq \binom{q-2}{s-1} s^{q-s}$ and

$$\begin{aligned}
\|\mathbb{E}(h_{ik}(q))\|_1 &\leq \sum_{s=2}^{q-1} \sum_{\pi \in \mathscr{P}_{s,1}} \sum_{i_1(\pi) \neq \cdots \neq i_{s-1}(\pi)} (2p)^s \\
&\leq \sum_{s=2}^{q-1} \binom{q-2}{s-1} s^{q-s} n_1^{s-1} (2p)^s \\
&\leq \sum_{s=2}^{q-1} \binom{q-2}{s-2} \frac{q-s}{s-1} \frac{(q-1)!}{(s-1)!} n_1^{s-1} (2p)^s \\
&\leq 4p^2 n_1 q! \sum_{s=2}^{q-1} \binom{q-2}{s-2} (2pn_1)^{s-2} \\
&\leq 4q!(1+2pn_1)^{q-2} p^2 n_1.
\end{aligned}$$

□

Proposition 1. *Assume that* $p \geq C\left(\sqrt{\frac{\log n_1}{n_1 n_2}} \vee \frac{\log n_1}{n_2}\right)$ *for some constant* $C > 0$. *Then, there exists a constant* $c_* > 0$ *such that*

$$\mathbb{E}\left(\left\|H\left(WW^\top\right)\right\|_\infty^2\right) \leq c_* \left(1 + \frac{n_1 \log n_1}{n_2}\right) n_1 n_2 p^2 \log n_1.$$

Proof. Introduce the notation $\mathsf{H} = \|H(WW^\top)\|_\infty^2$, $t_1 = 6\sqrt{n_1 n_2 p^2 \log n_1}$, $t_2 = 48(1+2n_1 p)\log n_1$ and $t_3 = t_1 \vee t_2$. Using Theorem 4 and the facts that $\exp(-a/(b+c)) \leq \exp(-a/(2b)) + \exp(-a/(2c))$ for all $a, b, c > 0$, we get

$$\begin{aligned}
\mathbb{E}\left(\mathsf{H}^2\right) = 2\int_0^\infty \mathbb{P}(\mathsf{H} > t)\, t dt &\leq t_3^2 + 2n_1 \int_{t_3}^\infty \exp\left(-\frac{t^2}{16 n_1 n_2 p^2}\right) t dt \\
&\quad + 2n_1 \int_{t_3}^\infty \exp\left(-\frac{t}{12(1+2n_1 p)}\right) t dt.
\end{aligned}$$

Since $n_1 n_2 p^2 \geq C \log n_1$ it comes out that, for some constants c_1, c_2, c_3 we have

$$\begin{aligned}
\mathbb{E}\left(\mathsf{H}^2\right) &\leq c_1 t_3^2 \\
&\leq c_2(n_1 n_2 p^2 \log n_1 + n_1^2 p^2 \log^2 n_1) \\
&\leq c_3 n_1 n_2 p^2 \log n_1 (1 + n_1 \log n_1 / n_2).
\end{aligned}$$

□

Proposition 2. *Let* $n_2 \geq n_1 \log n_1$ *and*

$$18\sqrt{\frac{\log n_1}{n_1 n_2}} \leq p \leq \frac{1}{206\, n_1 \log n_1}. \tag{3.11}$$

Then

$$\mathbb{E}\left(\|WW^\top - \mathbb{E}(WW^\top)\|_\infty^2\right) \geq \frac{n_2 p}{40}.$$

Proof. Set $\mathsf{p} = \max(\delta p, (2-\delta)p) \geq p$. Since $\gamma_1 < 1$ and $\gamma_2 < 1$ then at least one row of W has not less than $n_2/2$ entries that are centered Bernoulli variables with parameter p. Without loss of generality, let it be the first row of W. We denote this first row by X_1. We have

$$\begin{aligned}\|WW^\top - \mathbb{E}(WW^\top)\|_\infty &\geq \|WW^\top - \mathbb{E}(WW^\top)\|_\infty - \|H(WW^\top)\|_\infty \\ &\geq |\|X_1\|^2 - \mathbb{E}(\|X_1\|^2)| - \|H(WW^\top)\|_\infty,\end{aligned}$$

so that

$$\mathbb{E}\left(\|WW^\top - \mathbb{E}(WW^\top)\|_\infty^2\right) \geq \frac{1}{2}\mathbb{E}\left((\|X_1\|^2 - \mathbb{E}(\|X_1\|^2))^2\right) - \mathbb{E}\left(\|H(WW^\top)\|_\infty^2\right).$$

Denoting by η the centered Bernoulli variable with parameter p (η takes value $1-\mathsf{p}$ with probability p and value $-\mathsf{p}$ with probability $1-\mathsf{p}$) we get

$$\begin{aligned}\mathbb{E}\left((\|X_1\|^2 - \mathbb{E}(\|X_1\|^2))^2\right) &\geq \frac{n_2}{2}\mathrm{Var}(\eta^2) \\ &= \frac{n_2}{2}\mathsf{p}(1-\mathsf{p})(1-2\mathsf{p})^2 \geq \frac{9n_2 p}{20},\end{aligned}$$

where we have used the inequalities $p \leq \mathsf{p} \leq 2p \leq 1/70$.

Next, note that $2n_1 p \leq 1$ and introduce again the notation $\mathsf{H} = \|H(WW^\top)\|_\infty$, $t_1 = 4\sqrt{n_1 n_2 p^2 \log n_1}$, $t_2 = 4n_1 n_2 p^2/(3(1+2n_1 p)) > t_1$. From Theorem 4 and the facts that $t_2 \geq (2/3)n_1 n_2 p^2 \geq (2c/3)\log n_1$ with $c = 18^2$, and $n_1 \geq 2$ we get

$$\begin{aligned}\mathbb{E}\left(\mathsf{H}^2\right) &= 2\int_0^\infty \mathbb{P}(\mathsf{H} > t)\, t dt \leq t_1^2 + 2n_1 \int_{t_1}^{t_2} \exp\left(-\frac{t^2}{16 n_1 n_2 p^2}\right) t dt \\ &\quad + 2n_1 \int_{t_2}^\infty \exp\left(-\frac{t}{12(1+2n_1 p)}\right) t dt \\ &\leq 16 n_1 n_2 p^2 \log n_1 + 16 n_1 n_2 p^2 + 2n_1 \int_{t_2}^\infty \exp\Big(-\frac{t}{24}\Big) t dt \\ &\leq 16 n_1 n_2 p^2 (\log n_1 + 1) + 2n_1 (32 n_1 n_2 p^2 + (24)^2)\exp(-c(\log n_1)/36) \\ &\leq 16 n_1 n_2 p^2 (\log n_1 + 1) + 2^{-3} n_1 n_2 p^2 + (3/2)^2 \\ &\leq n_1 n_2 p^2 \log n_1 \Big(16 + \frac{17}{\log n_1}\Big)\end{aligned}$$

$$\leq \frac{n_2 p}{5},$$

where we have used the condition on p. Combining the above displays we get the proposition. □

9 Lower bound on the oracle

We prove below Proposition 3, which, we recall, states the impossibility of our oracle estimator (3.12) to achieve exact recovery if $p = \sqrt{c_\delta \frac{\log n_1}{n_1 n_2}}$. We re-write the oracle as follows

$$\tilde{\eta}_1 = \operatorname{sign}\left(H((A - p\mathbb{1}_{n_1}\mathbb{1}_{n_2}^\top)(A - p\mathbb{1}_{n_1}\mathbb{1}_{n_2}^\top)^\top)\eta_1\right).$$

Proof. Since $n_{1+} = n_{1-}$ and $n_{2+} = n_{2-}$ we obtain that the ith entry of vector $H((A - p\mathbb{1}_{n_1}\mathbb{1}_{n_2}^\top)(A - p\mathbb{1}_{n_1}\mathbb{1}_{n_2}^\top)^\top)\eta_1$ is equal to

$$h_i = \sum_{j=1}^{n_2}(A_{ij} - p)\sum_{k=1}^{n_1} A_{kj}\eta_{1k} - \eta_{1i}\sum_{j=1}^{n_2}(A_{ij} - p)^2.$$

For all i in $\{1, \dots, n_1\}$, since $\tilde{\eta}_{1i} \neq \eta_{1i}$ is equivalent to $h_i\eta_{1i} < 0$ we have

$$\mathbb{P}(\tilde{\eta}_{1i} \neq \eta_{1i}) = \mathbb{P}\left(\sum_{k\neq i}\sum_{j=1}^{n_2}\eta_{1i}\eta_{1k}(A_{ij} - p)A_{kj} < p\sum_{j=1}^{n_2}(p - A_{ij})\right).$$

Observe that

$$\begin{aligned}\sum_{k\neq i}\sum_{j=1}^{n_2}\eta_{1i}\eta_{1k}(A_{ij} - p)A_{kj} &= (1-p)\sum_{k\neq i,\ j:A_{ij}=1}\eta_{1i}\eta_{1k}A_{kj} - p\sum_{k\neq i,\ j:A_{ij}=0}\eta_{1i}\eta_{1k}A_{kj}\\ &= -(1-p)\sum_{k\neq i:\eta_{1k}\neq\eta_{1i}}\ \sum_{j:A_{ij}=1}A_{kj}\\ &\quad + (1-p)\sum_{k\neq i:\eta_{1k}=\eta_{1i}}\ \sum_{j:A_{ij}=1}A_{kj} - p\sum_{k\neq i,\ j:A_{ij}=0}\eta_{1i}\eta_{1k}A_{kj}.\end{aligned}$$

Hence

$$\begin{aligned}\mathbb{P}(\tilde{\eta}_{1i} \neq \eta_{1i}) &= \mathbb{P}\left(\sum_{k\neq i:\eta_{1k}\neq\eta_{1i}}\ \sum_{j:A_{ij}=1}A_{kj} > \beta\right) \geq \mathbb{P}\left(\alpha > \beta\right)\\ &\geq \mathbb{E}\Big[\mathbb{P}(\alpha > \beta\Big|A_i)\mathbb{1}_F\Big],\end{aligned}$$

where $A_i = (A_{ij})_{j=1}^{n_2}$,

$$\alpha = \sum_{k\neq i:\eta_{1k}\neq\eta_{1i}}\ \sum_{j:A_{ij}=1,\eta_{2j}=\eta_{1i}}A_{kj},$$

$$\beta = \sum_{k\neq i:\eta_{1k}=\eta_{1i}} \sum_{j:A_{ij}=1} A_{kj} + \frac{p}{1-p}\Big(\sum_{j=1}^{n_2} A_{ij} - \sum_{k\neq i,\ j:A_{ij}=0} \eta_{1i}\eta_{1k}A_{kj}\Big),$$

and

$$F = \Big\{\sum_{j=1}^{n_2} A_{ij} \leq 4n_2p\Big\} \cap \Big\{\sum_{j:\eta_{2j}=\eta_{1i}} A_{ij} \geq \delta p n_2/4\Big\}.$$

Note that F is an event of large enough probability for n_1 large enough. Indeed, as $\mathbb{E}(A_{ij}) \leq 2p$ and $\mathrm{Var}(A_{ij}) \leq 2p$ we get from Chebyshev inequality that

$$\begin{aligned}\mathbb{P}\Big(\sum_{j=1}^{n_2} A_{ij} > 4n_2p\Big) &\leq \mathbb{P}\Big(\sum_{j=1}^{n_2}(A_{ij} - \mathbb{E}(A_{ij})) > 2n_2p\Big) \\ &\leq \frac{1}{2pn_2} \leq \frac{1}{2\sqrt{c_\delta}\log n_1},\end{aligned} \tag{3.19}$$

where we have used the fact that $n_2 \geq n_1 \log n_1$. Similarly, using Chebyshev inequality and the facts that for any i we have $\mathrm{Card}\{j : \eta_{2j} = \eta_{1i}\} = n_2/2$ and that $\mathbb{E}(A_{ij}) = \delta p$ for $\eta_{2j} = \eta_{1i}$ we find

$$\begin{aligned}\mathbb{P}\Big(\sum_{j:\eta_{2j}=\eta_{1i}} A_{ij} < \delta p n_2/4\Big) &\leq \mathbb{P}\Big(\sum_{j:\eta_{2j}=\eta_{1i}} (\delta p - A_{ij}) > \delta p n_2/4\Big) \\ &\leq \frac{8}{\delta p n_2} \leq \frac{8}{\delta\sqrt{c_\delta}\log n_1}.\end{aligned} \tag{3.20}$$

It follows from (3.19) and (3.20) that

$$\mathbb{P}(F) \geq 1 - \frac{1}{\sqrt{c_\delta}\log n_1}\Big(\frac{1}{2} + \frac{8}{\delta}\Big). \tag{3.21}$$

Next, from Chebyshev inequality and the facts that $\mathbb{E}(A_{kj}) \leq 2p$, $\mathrm{Var}(A_{kj}) \leq 2p$, we obtain, conditionally on A_i,

$$\mathbb{P}\Big(\Big|\sum_{k\neq i,\ j:A_{ij}=0} \eta_{1i}\eta_{1k}A_{kj}\Big| \geq 4n_2n_1p\Big|\, A_i\Big) \leq \frac{1}{2n_2n_1p}\,.$$

Quite similarly, as $\mathrm{Card}\{k : \eta_{1k} = \eta_{1i}\} = n_1/2$ and for $A_i \in F$ we have $\mathrm{Card}\{j : A_{ij} = 1\} = \sum_{j=1}^{n_2} A_{ij} \leq 4pn_2$, the following inequality holds

$$\forall\, A_i \in F: \quad \mathbb{P}\left(\sum_{k\neq i:\eta_{1k}=\eta_{1i}} \sum_{j:A_{ij}=1} A_{kj} \geq 8n_2n_1p^2\Big|\, A_i\right) \leq \frac{1}{4n_2n_1p^2}\,.$$

Thus, for all n_1 large enough to have $p \leq 1/2$ we obtain

$$\forall\, A_i \in F: \quad \mathbb{P}(\beta \leq 24c_\delta \log n_1|\, A_i) = \mathbb{P}(\beta \leq 24n_1n_2p^2|\, A_i)$$

$$\geq 1 - \frac{3}{4n_2n_1p^2} = 1 - \frac{3}{4c_\delta \log n_1}.$$

Observe that random variables α and β are independent conditionally on A_i since the sums over (k, j) in their definitions are taken over disjoint sets of indices. Using this we get

$$\mathbb{P}(\tilde{\eta}_{1i} \neq \eta_{1i}) \geq \left(1 - \frac{3}{4c_\delta \log n_1}\right) \mathbb{E}\left[\mathbb{P}\left(\alpha \geq 24c_\delta \log n_1 \middle| A_i\right) \mathbb{1}_F\right]. \tag{3.22}$$

Note that, conditionally on A_i, the random variable α has a Binomial distribution with probability parameter $(2-\delta)p$. Moreover, if $A_i \in F$ then the number of terms in α denoted by n is such that $n \leq 4pn_1n_2$ and $n \geq (n_1/2-1)(\delta pn_2/4) \geq \delta pn_1n_2/12$ for $n_1 \geq 6$. It follows that, for any fixed $A_i \in F$, the assumptions of Lemma 1 are satisfied with $\mathsf{p} = (2-\delta)p$, $t = 24c_\delta \log n_1$ provided that $\sqrt{n_1/\log n_1} > 288\sqrt{c_\delta}/\delta$. Therefore, for n_1 large enough to satisfy this condition and $c_\delta \log n_1 \geq 1$, $n_1 \geq 6$, Lemma 1 implies that, for any $A_i \in F$,

$$\mathbb{P}(\alpha \geq 24c_\delta \log n_1 | A_i) \geq \frac{e^{-1/6}}{\sqrt{50\pi c_\delta \log n_1}} \exp\left(-25c_\delta \log n_1 \log\left(\frac{300}{\delta(2-\delta)}\right)\right).$$

With the choice $c_\delta = \left(50 \log\left(\frac{300}{\delta(2-\delta)}\right)\right)^{-1}$ this yields

$$\mathbb{P}(\alpha \geq 24c_\delta \log n_1 | A_i) \geq \frac{e^{-1/6}}{\sqrt{50\pi c_\delta n_1 \log n_1}}.$$

Combining this inequality with (3.21) and (3.22) we get the proposition. □

The following lemma is used to control the lower tail of binomial variables.

Lemma 1. *Let $\xi_1, \ldots, \xi_n$ be i.i.d. Bernoulli random variables with parameter p and $\alpha = \sum_{i=1}^n \xi_i$. Then for all $n\mathsf{p} < t < n$ we have*

$$\mathbb{P}(\alpha \geq t) \geq \frac{e^{-1/6}}{\sqrt{2\pi(t+1)}} \exp\left(-(t+1)\log\left(\frac{t+1}{n\mathsf{p}}\right)\right).$$

Proof. Set $k = \lceil t \rceil$. Since

$$\mathbb{P}\left(\sum_{i=1}^n \xi_i \geq t\right) \geq \mathbb{P}\left(\sum_{i=1}^n \xi_i = k\right)$$

for $k = n$ the result is trivial. Assume that $k \leq n-1$ and set $a = k/n$. Then $\mathsf{p} < a < 1$. By Stirling's approximation,

$$\sqrt{2\pi n}\,(n/e)^n \leq n! \leq \sqrt{2\pi n}\,(n/e)^n e^{1/12}.$$

Therefore,

$$\begin{aligned}\mathbb{P}\left(\sum_{i=1}^{n}\xi_i \geq t\right) &\geq \mathbb{P}\left(\sum_{i=1}^{n}\xi_i = k\right) = \frac{n!\mathsf{p}^k(1-\mathsf{p})^{n-k}}{k!(n-k)!} \\ &\geq \frac{\sqrt{2\pi n}\, n^n \mathsf{p}^k(1-\mathsf{p})^{n-k}}{e^{1/6}\sqrt{2\pi k}\, k^k \sqrt{2\pi(n-k)}\,(n-k)^{n-k}} \\ &\geq \frac{\mathsf{p}^k(1-\mathsf{p})^{n-k}}{e^{1/6}\sqrt{2\pi an}\; a^k\,(1-a)^{n-k}} \geq \frac{\mathsf{p}^k}{e^{1/6}\sqrt{2\pi an}\; a^k}.\end{aligned}$$

□

10 Main proofs

10.1 Proof of Theorem 1

Recall that

$$\eta_1^0 = \operatorname{sign}(\hat{v}),$$

where $\hat{v}$ is the eigenvector corresponding to the top eigenvalue of the matrix

$$H((A - \hat{p}\mathbb{1}_{n_1}\mathbb{1}_{n_2}^\top)(A - \hat{p}\mathbb{1}_{n_1}\mathbb{1}_{n_2}^\top)^\top)$$

with $\hat{p} = \frac{1}{n_1 n_2}\mathbb{1}_{n_1}^\top A \mathbb{1}_{n_2}$. Recall the notation $\tilde{A} = A - p\mathbb{1}_{n_1}\mathbb{1}_{n_2}^\top$. We have

$$H((A - \hat{p}\mathbb{1}_{n_1}\mathbb{1}_{n_2}^\top)(A - \hat{p}\mathbb{1}_{n_1}\mathbb{1}_{n_2}^\top)^\top) = H(\tilde{A}\tilde{A}^\top) + Z_4,$$

where (cf. (3.9))

$$H(\tilde{A}\tilde{A}^\top) = (\delta-1)^2 p^2 n_2 H(\eta_1\eta_1^\top) + H(WW^\top) + p(\delta-1)H(W\eta_2\eta_1^\top + \eta_1\eta_2^\top W^\top)$$

and

$$Z_4 := H((A - \hat{p}\mathbb{1}_{n_1}\mathbb{1}_{n_2}^\top)(A - \hat{p}\mathbb{1}_{n_1}\mathbb{1}_{n_2}^\top)^\top) - H(\tilde{A}\tilde{A}^\top).$$

Therefore,

$$H((A - \hat{p}\mathbb{1}_{n_1}\mathbb{1}_{n_2}^\top)(A - \hat{p}\mathbb{1}_{n_1}\mathbb{1}_{n_2}^\top)^\top) = (\delta-1)^2 p^2 n_2 \eta_1\eta_1^\top + Z,$$

where

$$Z = \underbrace{H(WW^\top)}_{Z_1} + \underbrace{p(\delta-1)H(W\eta_2\eta_1^\top + \eta_1\eta_2^\top W^\top)}_{Z_2} - \underbrace{(\delta-1)^2 p^2 n_2 I_{n_1}}_{Z_3} + Z_4.$$

Notice that since Z_3 is a multiple of the identity matrix, $\hat{v}$ is the eigenvector corresponding to the top eigenvalue of $H' = H((A - \hat{p}\mathbb{1}_{n_1}\mathbb{1}_{n_2}^\top)(A - \hat{p}\mathbb{1}_{n_1}\mathbb{1}_{n_2}^\top)^\top) + Z_3$. Thus,

$\hat{v}$ and $\frac{1}{\sqrt{n_1}}\eta_1$ are the eigenvectors of $\frac{1}{n_1}H'$ and $(\delta-1)^2p^2n_2\frac{\eta_1\eta_1^\top}{n_1}$ associated to their top eigenvalues, respectively. Since $\eta_1\eta_1^\top$ is rank one matrix, we get from Davis-Kahan Theorem (Theorem 4.5.5. in [Vershynin, 2018]) that

$$\min_{\nu\in\{-1,1\}}\left\|\frac{1}{\sqrt{n_1}}\eta_1-\nu\hat{v}\right\|_2^2\leq\frac{8\left\|Z_1+Z_2+Z_4\right\|_\infty^2}{(\delta-1)^4p^4n_1^2n_2^2}.$$

This implies (see Lemma 2 below) that

$$\frac{1}{n_1}r(\eta_1,\eta_1^0)\leq\frac{16}{(\delta-1)^4p^4n_1^2n_2^2}\|Z_1+Z_2+Z_4\|_\infty^2.$$

Thus, in order to bound $r(\eta_1,\eta_1^0)$, it remains to control the spectral norm of $Z_1+Z_2+Z_4$. Namely, we will prove that

$$\lim_{n_1\to\infty}\mathbb{P}\left(\|Z_i\|_\infty\geq\frac{\sqrt{\alpha}}{12}(\delta-1)^2p^2n_1n_2\right)=0,\quad i=1,2,4,$$

which implies the theorem.

- **Control of $\|Z_1\|_\infty$.**

  Recall that W is a random matrix with entries that are independent and distributed as $\zeta-\mathbb{E}(\zeta)$ where ζ is a Bernoulli random variable with parameter δp or $(2-\delta)p$. Therefore, both the expectation and the variance of each entry are bounded by $2p$. We now apply Theorem 4 with $t=\frac{\sqrt{\alpha}}{12}(\delta-1)^2p^2n_1n_2$. This yields

$$\begin{aligned}\mathbb{P}\left(\|Z_1\|_\infty\geq\frac{\sqrt{\alpha}}{12}(\delta-1)^2p^2n_1n_2\right)&\leq n_1\exp\left[-\frac{t^2}{(8n_1n_2p^2+6t)+2n_1pt}\right]\\&\leq n_1\exp\left[-12^{-2}\alpha(\delta-1)^4n_1n_2p^2/17\right]\\&\quad+n_1\exp\left[-\sqrt{\alpha}(\delta-1)^2pn_2/288\right],\end{aligned}$$

  where the last inequality uses the facts that $\exp(-a/(b+c))\leq\exp(-a/(2b))+\exp(-a/(2c))$ for all $a,b,c>0$, and $\alpha\in(0,1)$, $|\delta-1|<1$. Recall that $p\geq C(\delta-1)^{-2}\sqrt{\frac{\log n_1}{n_1n_2}}$ and $p\geq C(\delta-1)^{-2}\frac{\log n_1}{n_2}$ by the assumption of the theorem. Using these conditions and choosing $C\geq 289/\sqrt{\alpha}$ we obtain

$$\mathbb{P}\left(\|Z_1\|_\infty\geq\frac{\sqrt{\alpha}}{12}(\delta-1)^2p^2n_1n_2\right)\leq 2n_1^{-\frac{1}{288}}.$$

- **Control of $\|Z_2\|_\infty$.**

  In order to control Z_2, we first observe, using the inequality $\|H(M)\|_\infty\leq 2\|M\|_\infty$ valid for any matrix $M\in\mathbb{R}^{n_1\times n_1}$ (cf., e.g., Lemma 17 in [Ndaoud,

2018]), that

$$\begin{aligned}\left\|H\left(\eta_1\eta_2^\top W^\top + W\eta_2\eta_1^\top\right)\right\|_\infty &\le 2\left\|\eta_1\eta_2^\top W^\top + W\eta_2\eta_1^\top\right\|_\infty \\ &\le 2\left\|\eta_1\eta_2^\top W^\top\right\|_\infty + 2\|W\eta_2\eta_1^\top\|_\infty \\ &\le 4\sqrt{n_1}\|W\eta_2\|_2.\end{aligned}$$

Hence

$$\mathbb{E}(\|Z_2\|_\infty^2) \le 16(\delta-1)^2p^2n_1\mathbb{E}(\|W\eta_2\|_2^2). \tag{3.23}$$

Denote by $X_1,\dots,X_{n_1}$ the column vectors equal to the transposed rows of matrix W. Since $\mathbb{E}(X_iX_i^\top)$ is a diagonal matrix with positive entries bounded from above by $2p$ for all $i=1,\dots,n_1$, we obtain

$$\mathbb{E}(\|W\eta_2\|_2^2) = \eta_2^\top\mathbb{E}(W^\top W)\eta_2 = \sum_{i=1}^{n_1}\eta_2^\top\mathbb{E}(X_iX_i^\top)\eta_2 \le 2pn_1n_2. \tag{3.24}$$

Chebyshev's inequality combined with (3.23) and (3.24) yields the bound

$$\begin{aligned}\mathbb{P}\left(\|Z_2\|_\infty \ge \frac{\sqrt{\alpha}}{12}(\delta-1)^2p^2n_1n_2\right) &\le \frac{9\cdot 2^9}{\alpha(\delta-1)^2pn_2} \\ &\le \frac{9\cdot 2^9}{C\alpha\log n_1},\end{aligned}$$

where we have used the fact that $p \ge C(\delta-1)^{-2}\frac{\log n_1}{n_2}$ by the assumptions of the theorem.

- **Control of $\|Z_4\|_\infty$.**

  We have

$$\begin{aligned}Z_4 &= H((A-\hat p\mathbb{1}_{n_1}\mathbb{1}_{n_2}^\top)(A-\hat p\mathbb{1}_{n_1}\mathbb{1}_{n_2}^\top)^\top - (A-p\mathbb{1}_{n_1}\mathbb{1}_{n_2}^\top)(A-p\mathbb{1}_{n_1}\mathbb{1}_{n_2}^\top)^\top) \\ &= H((p-\hat p)(A\mathbb{1}_{n_2}\mathbb{1}_{n_1}^\top + \mathbb{1}_{n_1}\mathbb{1}_{n_2}^\top A^\top) + ((\hat p-p)^2 - 2p(p-\hat p))n_2\mathbb{1}_{n_1}\mathbb{1}_{n_1}^\top) \\ &= (p-\hat p)H((W\mathbb{1}_{n_2}\mathbb{1}_{n_1}^\top + \mathbb{1}_{n_1}\mathbb{1}_{n_2}^\top W^\top) \\ &+ (\delta-1)p(n_{2+}-n_{2-})(\eta_1\mathbb{1}_{n_1}^\top + \mathbb{1}_{n_1}\eta_1^\top) + (p-\hat p)n_2\mathbb{1}_{n_1}\mathbb{1}_{n_1}^\top).\end{aligned}$$

  Since

$$\hat p - p = \frac{(\delta-1)p(n_{1+}-n_{1-})(n_{2+}-n_{2-})}{n_1n_2} + \frac{1}{n_1n_2}\sum_{i,j}W_{ij}$$

  then, recalling that $|n_{i+}-n_{i-}|/n_i \le \gamma_i$ for $i=1,2$, and setting $y := \frac{1}{n_1n_2}\sum_{i,j}W_{ij}$, we have

$$|\hat p - p| \le |\delta-1|p\gamma_1\gamma_2 + |y|.$$

Thus, using again the inequality $\|H(M)\|_\infty \le 2\|M\|_\infty$ and introducing the notation $L = \|W\mathbb{1}_{n_2}\mathbb{1}_{n_1}^\top + \mathbb{1}_{n_1}\mathbb{1}_{n_2}^\top W^\top\|_\infty$ we obtain

$$\begin{aligned}\|Z_4\|_\infty &\le 2|p-\hat p|L + 4|\delta-1||p-\hat p|pn_1n_2 + |p-\hat p|^2 n_2\|H(\mathbb{1}_{n_1}\mathbb{1}_{n_1}^\top)\|_\infty \\ &\le 2|\hat p - p|L + 4(\delta-1)^2p^2n_1n_2\gamma_1\gamma_2 + 4|y|pn_1n_2 + |p-\hat p|^2n_1n_2 \\ &\le V + 6(\delta-1)^2p^2n_1n_2\gamma_1\gamma_2,\end{aligned}$$

where

$$V = 2|\hat p - p|L + 4|y|pn_1n_2 + 2y^2n_1n_2.$$

Now, note that since W_{ij} are zero mean random variables

$$\mathbb{E}(y^2) \le \frac{2p}{n_1n_2}, \quad \mathbb{E}(|\hat p - p|^2) \le p^2 + \frac{2p}{n_1n_2}.$$

Moreover, by the same argument as in the control of $\|Z_2\|_\infty$,

$$\mathbb{E}(L^2) = \mathbb{E}(\|W\mathbb{1}_{n_2}\mathbb{1}_{n_1}^\top + \mathbb{1}_{n_1}\mathbb{1}_{n_2}^\top W^\top\|_\infty^2) \le 32n_2n_1^2p^3 \le 8n_2n_1^2p.$$

Using these inequalities and the facts that $p \le 1/2$, $n_2 \ge 2$ and $\sqrt{p^2n_1n_2} \ge C(\delta-1)^{-2}\sqrt{\log n_1} \ge 289\sqrt{\log 2}$ we obtain

$$\begin{aligned}\mathbb{E}(V) &\le 2\sqrt{\mathbb{E}(|\hat p - p|^2)}\sqrt{\mathbb{E}(L^2)} + 4\sqrt{\mathbb{E}(y^2)}pn_1n_2 + 2\mathbb{E}(y^2)n_1n_2 \\ &\le 2n_1\sqrt{8n_2p}\sqrt{p^2 + \frac{2p}{n_1n_2}} + 4\sqrt{2n_1n_2p^3} + 4p \\ &\le 12\sqrt{p^2n_1n_2}(1+\sqrt{pn_1}).\end{aligned}$$

Putting the above arguments together and applying Markov inequality we get that, for $\gamma_1\gamma_2 \le \sqrt{\alpha}/96$,

$$\begin{aligned}\mathbb{P}\left(\|Z_4\|_\infty \ge \frac{\sqrt{\alpha}}{12}(\delta-1)^2p^2n_1n_2\right) &\le \mathbb{P}\left(V \ge \frac{\sqrt{\alpha}}{48}(\delta-1)^2p^2n_1n_2\right) \\ &\le \frac{576(1+\sqrt{pn_1})}{(\delta-1)^2\sqrt{\alpha}\sqrt{p^2n_1n_2}} \\ &\le \frac{576}{(\delta-1)^2\sqrt{\alpha}}((p^2n_1n_2)^{-1/2} + (pn_2)^{-1/2}).\end{aligned}$$

Recall that, by the assumptions of the theorem, we have $p \ge C(\delta-1)^{-2}\sqrt{\frac{\log n_1}{n_1n_2}}$, $p \ge C(\delta-1)^{-2}\frac{\log n_1}{n_2}$, and that we have chosen $C \ge 289/\sqrt{\alpha}$. Using these inequalities and the facts that $|\delta-1| < 1$, $\alpha \in (0,1)$ in the last display we find

$$\mathbb{P}\left(\|Z_4\|_\infty \ge \frac{\sqrt{\alpha}}{12}(\delta-1)^2p^2n_1n_2\right) \le \frac{36}{\alpha^{1/4}|\delta-1|\sqrt{\log n_1}}.$$

In conclusion, we have proved that, for any $\alpha \in (0,1)$, $C \geq 289/\sqrt{\alpha}$ and $\gamma_1\gamma_2 \leq \sqrt{\alpha}/96$ we have

$$\mathbb{P}\left(\frac{16}{(\delta-1)^4 p^4 n_1^2 n_2^2}\|Z_1+Z_2+Z_4\|_\infty^2 \geq \alpha\right) \leq 2n_1^{-\frac{1}{288}} + \frac{9\cdot 2^9}{C\alpha \log n_1} + \frac{36}{\alpha^{1/4}|\delta-1|\sqrt{\log n_1}}. \tag{3.25}$$

Hence, if $\alpha \in (0,1)$, $\gamma_1\gamma_2 \leq \sqrt{\alpha}/96$, there exists an absolute constant $C_0 > 0$ such that for $C > C_0/\sqrt{\alpha}$ we have

$$\lim_{n_1\to\infty} \mathbb{P}\left(\frac{1}{n_1} r(\eta_1, \eta_1^0) \geq \alpha\right) = 0. \tag{3.26}$$

This proves part (i) of the theorem. Next, if we assume that $\gamma_1\gamma_2 \leq 1/C'_{n_1}$ and set $C = C_{n_1}$ where C_{n_1}, C'_{n_1} are any positive sequences that tend to infinity then (3.26) holds simultaneously for all $\alpha \in (0,1)$, which proves almost full recovery.

Lemma 2. *For $\eta_1^0 = \operatorname{sign}(\hat{v})$ we have*

$$\frac{1}{n_1} r(\eta_1, \eta_1^0) \leq 2 \min_{\nu\in\{-1,1\}} \left\| \frac{\nu}{\sqrt{n_1}}\eta_1 - \hat{v}\right\|_2^2.$$

Proof. By definition, $r(\eta_1, \eta_1^0) = 2\min_{\nu\in\{-1,1\}} \sum_{i=1}^{n_1} \mathbb{1}\left(\nu\eta_{1i} \neq \eta_{1i}^0\right)$. Set $\hat{b} = \hat{v}\sqrt{n_1}$. Then $\eta_1^0 = \operatorname{sign}(\hat{b})$ and, for any $\nu \in \{-1,1\}$,

$$\left\| \frac{\nu}{\sqrt{n_1}}\eta_1 - \hat{v}\right\|_2^2 = \frac{1}{n_1}\|\nu\eta_1 - \hat{b}\|_2^2 \geq \frac{1}{n_1}\sum_{i=1}^{n_1} \mathbb{1}\left(\nu\eta_{1i} \neq \eta_{1i}^0\right),$$

where the last inequality is due to the fact that $(x-y)^2 \geq \mathbb{1}\left(x \neq \operatorname{sign}(y)\right)$ for any $x \in \{-1,1\}$ and $y \in \mathbb{R}$. □

10.2 Proof of Theorem 2

Note that the assumptions of Theorem 1(i) are satisfied with $\alpha = 1/25$. Note also that $\|\eta_1 - \eta_1^0\|_1 = n_1 - \eta_1^\top \eta_1^0$. It follows from Theorem 1 and the definition of $r(\hat{\eta}_1, \eta_1)$ that with probability that tends to 1 as $n_1 \to \infty$ we have either $\frac{1}{n_1}\eta_1^\top\eta_1^0 \geq 3/4$ or $\frac{1}{n_1}\eta_1^\top\eta_1^0 \leq -3/4$. Next, recall that

$$\Gamma := H((A - \hat{p}\mathbb{1}_{n_1}\mathbb{1}_{n_2}^\top)(A - \hat{p}\mathbb{1}_{n_1}\mathbb{1}_{n_2}^\top)^\top) = (\delta-1)^2 p^2 n_2 \eta_1\eta_1^\top + Z.$$

From (3.25) we have

$$\lim_{n_1\to\infty} \mathbb{P}\left(\|Z_1+Z_2+Z_4\|_\infty \geq \frac{1}{20}(\delta-1)^2 p^2 n_1 n_2\right) = 0,$$

using the same notation as in the proof of Theorem 1. Observing that $\|Z_3\|_\infty = (\delta-1)^2 p^2 n_2$ we get moreover that

$$\lim_{n_1\to\infty} \mathbb{P}\left(\|Z\|_\infty \geq \frac{1}{16}(\delta-1)^2 p^2 n_1 n_2\right) = 0. \tag{3.27}$$

Define the following random events:

$$O_i = \left\{\left(\frac{1}{n_1}\Gamma_i \eta_1\right)\eta_{1i} \geq \frac{(\delta-1)^2}{2} p^2 n_2\right\}, \quad i = 1,\dots,n_1,$$
$$B = \left\{\frac{1}{n_1}\|Z\|_\infty \leq \frac{1}{16}(\delta-1)^2 p^2 n_2\right\},$$

where Γ_i denotes the ith row of matrix Γ. From (3.27) we have that the probability of B tends to 1 as $n_1 \to \infty$. We call O_i the oracle events since they are similar to the events arising in the analysis of the supervised oracle procedure that, given the labels $(\eta_{1j}, j \neq i)$, estimates the label η_{1i}. The proof is decomposed in three steps that we detail in what follows.

- **Proving the contraction.**
  We place ourselves on the random event $B \cap O_1 \cap \cdots \cap O_{n_1}$. Our first goal is to prove that if $\frac{1}{n_1}\eta_1^\top \hat{\eta}^k \geq 3/4$, then $\|\hat{\eta}^{k+1} - \eta_1\|_1 \leq \frac{1}{4}\|\hat{\eta}^k - \eta_1\|_1$ and $\frac{1}{n_1}\eta_1^\top \hat{\eta}^{k+1} \geq 3/4$. We have

$$\begin{aligned}\frac{1}{n_1}\Gamma_i \hat{\eta}^k =& \frac{1}{n_1} z_i^\top(\hat{\eta}^k - \eta_1) + \frac{1}{n_1}\Gamma_i \eta_1 \\ &- (\delta-1)^2 p^2 n_2 \eta_{1i}\left(1 - \frac{1}{n_1}\eta_1^\top \hat{\eta}^k\right),\end{aligned}$$

  where z_i is a column vector equal to the transposed ith row of matrix Z. Hence, if $\eta_{1i} = -1$ then

$$\frac{1}{n_1}\Gamma_i \hat{\eta}^k \leq \frac{1}{n_1} z_i^\top(\hat{\eta}^k - \eta_1) - \frac{(\delta-1)^2}{4} p^2 n_2.$$

  It follows that

$$\mathbb{1}_{\left\{\frac{1}{n_1}\Gamma_i\hat{\eta}^k \geq 0\right\}} \leq \mathbb{1}_{\left\{\frac{1}{n_1} z_i^\top(\hat{\eta}^k - \eta_1) \geq \frac{(\delta-1)^2}{4} p^2 n_2\right\}} \leq \left(\frac{4 z_i^\top(\hat{\eta}^k - \eta_1)}{n_1(\delta-1)^2 p^2 n_2}\right)^2.$$

  Similarly, if $\eta_{1i} = 1$ then

$$\mathbb{1}_{\left\{\frac{1}{n_1}\Gamma_i\hat{\eta}^k \leq 0\right\}} \leq \left(\frac{4 z_i^\top(\hat{\eta}^k - \eta_1)}{n_1(\delta-1)^2 p^2 n_2}\right)^2.$$

Now,

$$\frac{1}{2}\|\hat{\eta}^{k+1}-\eta_1\|_1=\sum_{i=1}^{n_1}\mathbb{1}_{\left\{\frac{1}{n_1}\Gamma_i\hat{\eta}^k\geq 0\right\}}\mathbb{1}_{\eta_{1i}=-1}+\sum_{i=1}^{n_1}\mathbb{1}_{\left\{\frac{1}{n_1}\Gamma_i\hat{\eta}^k\leq 0\right\}}\mathbb{1}_{\eta_{1i}=1}.$$

Hence, we get

$$\frac{1}{2n_1}\|\hat{\eta}^{k+1}-\eta_1\|_1\leq\left(\frac{4\|Z\|_\infty}{n_1(\delta-1)^2p^2n_2}\right)^2\ \frac{\|\hat{\eta}^k-\eta_1\|_2^2}{n_1}\leq\frac{1}{8n_1}\|\hat{\eta}^k-\eta\|_1. \quad (3.28)$$

The fact that $\frac{1}{n_1}\eta_1^\top\hat{\eta}^{k+1}\geq 3/4$ follows immediately from the inequality $\|\hat{\eta}^{k+1}-\eta_1\|_1\leq\frac{1}{4}\|\hat{\eta}^k-\eta_1\|_1$ and the relation $\|\hat{\eta}^k-\eta_1\|_1=n_1-\eta_1^\top\hat{\eta}^k$.
Quite analogously, we find that that if $\frac{1}{n_1}\eta_1^\top\hat{\eta}^k\leq -3/4$, then $\|\hat{\eta}^{k+1}+\eta_1\|_1\leq\frac{1}{4}\|\hat{\eta}^k+\eta_1\|_1$.

- **Reduction to the oracle events.**
Assume that the event $B\cap O_1\cap\cdots\cap O_{n_1}$ holds. Let first $\frac{1}{n_1}\eta_1^\top\eta_1^0\geq 3/4$. Since $\|\eta_1^0-\eta_1\|_1=n_1-\eta_1^\top\eta_1^0$ we get

$$\frac{1}{n_1}\|\hat{\eta}^k-\eta_1\|_1\leq\frac{1}{n_1}\|\eta_1^0-\eta_1\|_1\left(\frac{1}{4}\right)^k\leq\left(\frac{1}{4}\right)^{k+1}.$$

For $k>\frac{\log n_1}{2\log 2}-\frac{3}{2}$ we have

$$\left(\frac{1}{4}\right)^{k+1}<\frac{2}{n_1},$$

so that

$$\|\hat{\eta}^k-\eta_1\|_1=0.$$

Quite similarly we prove that if $\frac{1}{n_1}\eta_1^\top\hat{\eta}^k\leq -3/4$ then, for $k>\frac{\log n_1}{2\log 2}-\frac{3}{2}$,

$$\|\hat{\eta}^k+\eta_1\|_1=0.$$

Recalling the definition of $r(\hat{\eta}^k,\eta_1)$ we conclude that

$$\mathbb{P}\left(r(\hat{\eta}^k,\eta_1)\neq 0\right)\leq\mathbb{P}(B^c)+\sum_{i=1}^{n_1}\mathbb{P}(O_i^c).$$

It follows from (3.27) that $\lim_{n_1\to\infty}\mathbb{P}(B^c)=0$. Thus, the proof of the theorem will be complete if we show that

$$\lim_{n_1\to\infty}\sum_{i=1}^{n_1}\mathbb{P}(O_i^c)=0. \quad (3.29)$$

- **Control of the oracle events.**
We proceed now to the proof of (3.29). Let $G_1, \dots, G_{n_1}$ be the column vectors equal to the transposed rows of matrix $G := A - \hat{p}\mathbb{1}_{n_1}\mathbb{1}_{n_2}^\top = (p-\hat{p})\mathbb{1}_{n_1}\mathbb{1}_{n_2}^\top + (\delta-1)p\eta_1\eta_2^\top + W$. For all $i = 1, \dots, n_1$, we have

$$\mathbb{P}(O_i^c) = \mathbb{P}\left(\eta_{1i}G_i^\top\left(\sum_{k\neq i}\eta_{1k}G_k\right) < \frac{(\delta-1)^2}{2}p^2n_2n_1\right).$$

Denoting by $X_1, \dots, X_{n_1}$ the column vectors equal to the transposed rows of matrix W we may write $\eta_{1i}G_i = v_i + \eta_{1i}X_i$, where $v_i = \eta_{1i}(p-\hat{p})\mathbb{1}_{n_2} + (\delta-1)p\eta_2$. Therefore,

$$\begin{aligned}\eta_{1i}G_i^\top\left(\sum_{k\neq i}\eta_{1k}G_k\right) &= (v_i^\top + \eta_{1i}X_i^\top)\left(\sum_{k\neq i}v_k + \sum_{k\neq i}\eta_{1k}X_k\right)\\ &= (\delta-1)^2p^2n_2(n_1-1) + T_1 + T_2 + T_3 + T_4,\end{aligned}$$

where

$$T_1 = \eta_{1i}\sum_{k\neq i}X_i^\top v_k, \quad T_2 = \sum_{k\neq i}\eta_{1k}v_i^\top X_k,$$

$$T_3 = \eta_{1i}\sum_{k\neq i}\eta_{1k}X_i^\top X_k, \quad T_4 = \sum_{k\neq i}v_i^\top v_k - (\delta-1)^2p^2n_2(n_1-1)$$

and we obtain

$$\mathbb{P}(O_i^c) = \mathbb{P}\left(-T_1 - T_2 - T_3 - T_4 > (\delta-1)^2p^2n_2(n_1/2-1)\right).$$

We now bound from above the four corresponding probabilities. First, recall that

$$|\hat{p}-p| \le |\delta-1|p\gamma_1\gamma_2 + \left|\frac{1}{n_1n_2}\sum_{i,j}W_{ij}\right| \le \frac{|\delta-1|p}{480} + \left|\frac{1}{n_1n_2}\sum_{i,j}W_{ij}\right|.$$

The entries W_{ij} of matrix W are independent zero-mean random variables distributed as $\zeta - \mathbb{E}(\zeta)$ where ζ is a Bernoulli random variable with parameter δp or $(2-\delta)p$. As W_{ij} are bounded in absolute value by 1 and have variances bounded by $2p$ we get from Bernstein's inequality that

$$\mathbb{P}(|\hat{p}-p| \ge |\delta-1|p/64) \le 2e^{-c(\delta-1)^2n_1n_2p}. \tag{3.30}$$

Here and below we denote by c absolute positive constants that may vary from line to line. Next, on the event $|\hat{p}-p| \le |\delta-1|p/64$ we have

$$\begin{aligned}|T_1| &\le |\mathbb{1}_{n_2}^\top X_i|\left|\sum_{k\neq i}\eta_{1k}(p-\hat{p})\right| + |\delta-1|(n_1-1)p|\eta_2^\top X_i|\\ &\le |\delta-1|pn_1(|\mathbb{1}_{n_2}^\top X_i| + |\eta_2^\top X_i|).\end{aligned}$$

Here, $\mathbb{1}_{n_2}^\top X_i$ and $\eta_2^\top X_i$ are two sums of n_2 independent zero-mean random variables bounded in absolute value by 1 and with variances bounded by $2p$. Using these remarks, Bernstein's inequality and (3.30) we obtain that, for $n_1 \geq 4$,

$$\begin{aligned}
&\mathbb{P}\left(|T_1| \geq \frac{1}{4}(\delta-1)^2 p^2 n_2 (n_1/2-1)\right) \\
&\leq \mathbb{P}\left(|\mathbb{1}_{n_2}^\top X_i| + |\eta_2^\top X_i| \geq \frac{1}{16}|\delta-1| p n_2\right) + \mathbb{P}(|\hat{p}-p| \geq |\delta-1|p/64) \\
&\leq 4\exp\left(-c(\delta-1)^2 p n_2\right) + \mathbb{P}(|\hat{p}-p| \geq |\delta-1|p/64) \\
&\leq 6\exp\left(-c\,C\log n_1\right) \leq \frac{1}{n_1^2}
\end{aligned}$$

where we have used the assumption that $p \geq C(\delta-1)^{-2}\frac{\log n_1}{n_2}$ for some $C>0$ large enough. Quite analogous application of Bernstein's inequality, this time to two sums of $n_2(n_1-1)$ random variables, yields the bound

$$\begin{aligned}
&\mathbb{P}\left(-T_2 \geq \frac{1}{4}(\delta-1)^2 p^2 n_2 (n_1/2-1)\right) \\
&\qquad \leq \mathbb{P}\left(-\sum_{k\neq i} \eta_{1k} v_i^\top X_k \geq \frac{1}{16}(\delta-1)^2 p^2 n_1 n_2\right) \\
&\qquad \leq 6\exp\left(-c(\delta-1)^2 p n_1 n_2\right) \\
&\qquad \leq 6\exp\left(-c\,C n_1 \log n_1\right) \leq \frac{1}{n_1^2}.
\end{aligned}$$

Next, we consider the term $T_3 = \eta_{1i}\sum_{k\neq i}\eta_{1k} X_i^\top X_k$. We have

$$\begin{aligned}
&\mathbb{P}\left(-T_3 \geq \frac{1}{4}(\delta-1)^2 p^2 n_2 n_1\right) \\
&\qquad \leq \mathbb{E}\left[\mathbb{P}\left(-T_3 \geq \frac{1}{4}(\delta-1)^2 p^2 n_2 n_1 \middle| X_i\right)\mathbb{1}_{F_i}\right] + \mathbb{P}(F_i^c),
\end{aligned}$$

where $F_i = \{\|X_i\|_2^2 \leq 6 n_2 p\}$. Recall that $\|X_i\|_2^2 = \sum_{j=1}^{n_2} W_{ij}^2$ where W_{ij} are the elements of matrix W. We now apply Bernstein's inequality conditionally on X_i to the random variable T_3, which is (conditionally on X_i) a sum of $n_2(n_1-1)$ independent zero-mean random variables bounded in absolute value by 1 and with the sum of variances bounded by $2p(n_1-1)\|X_i\|_2^2$. It follows from Bernstein's inequality that for any fixed $X_i \in F_i$ we have

$$\begin{aligned}
&\mathbb{P}\left(-\eta_{1i}\sum_{k\neq i}\eta_{1k} X_i^\top X_k \geq \frac{1}{4}(\delta-1)^2 p^2 n_2 n_1 \middle| X_i\right) \\
&\qquad \leq \exp\left(-\frac{c(\delta-1)^4 p^4 n_2^2 n_1^2}{p n_1 \|X_i\|_2^2 + (\delta-1)^2 p^2 n_2 n_1}\right)
\end{aligned}$$

$$\leq \exp\left(-c(\delta-1)^4 p^2 n_2 n_1\right) \leq \frac{1}{n_1^2},$$

where the last inequality is valid if $C > 0$ is large enough. Applying once more Bernstein's inequality we obtain the bound

$$\mathbb{P}(F_i^c) \leq \mathbb{P}\left(\sum_{j=1}^{n_2}(W_{ij}^2 - \mathbb{E}(W_{ij}^2)) \geq 4n_2 p\right) \leq \exp\left(-cn_2 p\right) \leq \frac{1}{n_1^2}$$

if $C > 0$ is large enough. Finally, we consider the term $T_4 = \sum_{k\neq i} v_i^\top v_k - (\delta - 1)^2 p^2 n_2(n_1 - 1)$. We have

$$\begin{aligned}
|T_4| &\leq \left|\eta_{1i}(p-\hat{p})^2 n_2 \sum_{k\neq i}\eta_{1k}\right| + \left|(\delta-1)p(p-\hat{p})(\eta_2^\top \mathbb{1}_{n_2})\sum_{k\neq i}\eta_{1k}\right| \\
&\quad + \left|\eta_{1i}(\delta-1)p(p-\hat{p})(n_1-1)(\eta_2^\top \mathbb{1}_{n_2})\right| \\
&\leq n_1 n_2 (p-\hat{p})^2 + 2|\delta-1|pn_1n_2|\hat{p}-p|.
\end{aligned}$$

Therefore, on the event $|\hat{p}-p| \leq p|\delta-1|/64$ we have $|T_4| < \frac{1}{16}(\delta-1)^2p^2n_1n_2$, which implies that for $n_1 \geq 4$ and $C > 0$ large enough,

$$\begin{aligned}
&\mathbb{P}\left(-T_4 \geq \frac{1}{4}(\delta-1)^2p^2n_2(n_1/2-1)\right) \\
&\qquad \leq \mathbb{P}(|\hat{p}-p| \geq p|\delta-1|/64) \leq 2\exp\left(-c\,Cn_1\log n_1\right) \leq \frac{1}{n_1^2},
\end{aligned}$$

where we have used (3.30) and the assumption that $p \geq C(\delta-1)^{-2}\frac{\log n_1}{n_2}$ for some $C > 0$ large enough. Combining the above inequalities we find that, for $C > 0$ large enough,

$$\sum_{i=1}^{n_1}\mathbb{P}(O_i^c) \leq \frac{4}{n_1} \underset{n_1\to\infty}{\to} 0.$$

This proves (3.29) and hence the theorem.

11 Numerical experiments

The goal of this section is to provide numerical evidence to our theory. We compare the performance of methods defined previously, namely:

- SVD estimator (SVD),
- debiased spectral estimator (DS),
- diagonal deletion SVD estimator (DD),
- hollowed Lloyd's algorithm with spectral initialization (HL),

- the oracle procedure (O).

In what follows, we fix the number of labels $n = 300$, the imbalance $\gamma_1 = 0$, $\gamma_2 = 0.5$ and $\delta = 0.5$. For the sake of readability of plots, we define the parameters a and b such that

$$p = \sqrt{a}/n_1 \quad \text{and} \quad b = n_1(\log n_1)/n_2.$$

According to our improved sufficient conditions and using the above parameterization we expect the phase transition for exact recovery to happen at

$$a \geq C_\delta(b \vee b^2)$$

for some $C_\delta > 0$. We set up the simulations as follows. We consider $b \in \{0.1, 0.5, 5\}$ and we take a on a uniform grid of 20 points in a region where the phase transition occurs. For each such (a, b), we repeat the simulation 1000 times. Figure 1 presents the empirical probabilities of exactly recovering the vector of true labels η_1.

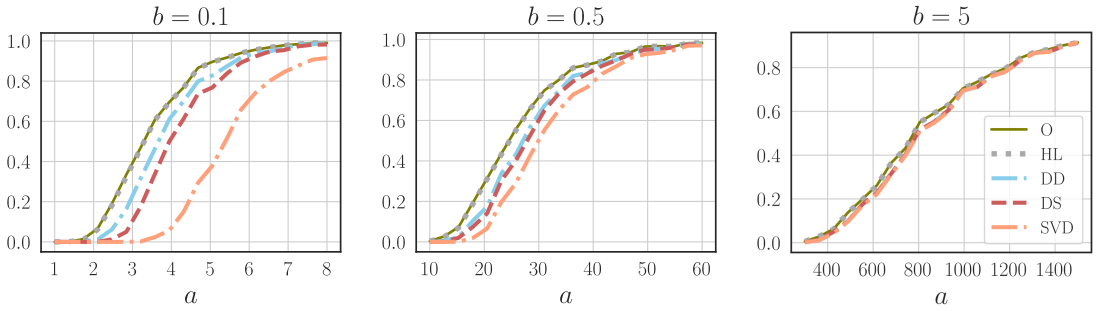


Figure 3.1: Empirical probability of success over 1000 runs of the experiment for: $b = 0.1$ (left), $b = 0.5$ (center) and $b = 5$ (right).

Overall, numerical experiments match our theoretical findings and provide some interesting insights:

1. Hollowed Lloyd's algorithm with spectral initialization achieves a performance remarkably close to the oracle without any prior knowledge about the true labels. Notice that this holds also when only a fraction of labels can be recovered, i.e. when the probability of wrong recovery is not exactly zero. This, in particular, suggests that the theoretical comparison we established between the above algorithms can be extended beyond the problem of exact recovery. Further simulations show that randomly initialized hollowed Lloyd's algorithm achieves the same performace as well (we omit these simulations since such an algorithm is not covered by our theory).

2. In the case $b = 0.1$ (high dimension), we recover empirically the diagram of Section 7. Observe that as b gets larger (moderate and small dimension) all the estimators converge to almost indistinguishable performance. In other words,

the ranking of estimators given in Section 7 only accentuates for the high-dimensional regime. This agrees with the fact that the conclusions of Section 7 are restricted to the zone $n_2 \geq n_1(\log n_1)^4$.

3. In high dimensions, the DD method outperforms the DS, which supports the argument that, under the BSBM model, hollowing is more beneficial than debiasing (cf. Proposition 2 and the corresponding discussion).

Chapter 4

Assigning topics to documents by successive projections

Topic models provide a useful tool to organize and understand the structure of large corpora of text documents, in particular, to discover hidden thematic structure. Clustering documents from big unstructured corpora into topics is an important task in various fields, such as image analysis, e-commerce, social networks, population genetics. Since the number of topics is typically substantially smaller than the size of the corpus and of the dictionary, the methods of topic modeling can lead to a dramatic dimension reduction. We study the problem of estimating the topic-document matrix, which gives the topics distribution for each document in a given corpus, that is we focus on the clustering aspect of the problem. We introduce an algorithm that we call Successive Projection Overlapping Clustering (SPOC) inspired by the Successive Projection Algorithm for separable matrix factorization. This algorithm is simple to implement and computationally fast. We establish upper bounds on the performance of SPOC algorithm for estimation of topic-document matrix, as well as near matching minimax lower bounds. We also propose a method that achieves analogous results when the number of topics is unknown and provides an estimate of the number of topics. Our theoretical results are complemented with a numerical study on synthetic and semi-synthetic data.

This chapter is based on [Klopp et al., 2021]: O. Klopp, M. Panov, S. Sigalla, and al. *Assigning Topics to Documents by Successive Projections.* ArXiv preprint arXiv:2107.03684, 2021.

---

1 Introduction **72**
2 Successive Projection Overlapping Clustering **77**
3 Main results **79**
3.1 Deterministic bounds 79
3.2 Bounds with high probability 80

3.3 Adaptive procedure when K is unknown. 82
3.4 Minimax lower bound 83
4 Related Work 84
5 Numerical experiments 85
5.1 Synthetic Data 85
5.2 Corpus of NIPS abstracts 89
6 Conclusion 90
7 Tools 91
7.1 Matrix Perturbation Bounds 91
7.2 Noisy Separable Matrix Factorization 93
7.3 Concentration Bounds for Multinomial Matrices 94
8 Proofs of the Main Results 97
8.1 Proof of Lemma 1 97
8.2 Proof of Lemma 2 98
8.3 Proof of Theorem 1 99
8.4 Proof of Theorem 2 and Corollary 3 100
8.5 Proof of Theorem 3 100
9 Auxiliary lemmas 107
9.1 The anchor document assumption under the Dirichlet prior 109
10 Additional Experiments: Estimation of topic-word matrix 110
11 Additional Experiments: Empirical study of singular values of word-document and topic-document matrices . 111
12 Additional Experiments: Estimation for the $p = 2000$ . . 116

1 Introduction

Assigning topics to documents is an important task in several applications. For example, press agencies need to identify articles of interest to readers based on the topics of articles that they have read in the past. Analogous goals are pursued by many other text-mining applications such as, for example, recommending blogs from among the millions of blogs available. A popular approach to the problem of estimating hidden thematic structures in a corpus of documents is based on topic modeling. Topic models have attracted a great deal of attention in the past two decades. Beyond text mining, they were used in areas, such as population genetics [Bicego et al., 2012, Pritchard et al., 2000], social networks [McCallum et al., 2005, Curiskis et al., 2020], image analysis [Li et al., 2010, Zhu et al., 2017], e-commerce [Palese and Usai, 2018, Yuan et al., 2018].

We adopt the *probabilistic Latent Semantic Indexing* (pLSI) model [Hofmann, 1999]. The pLSI model deals with three types of variables, namely, documents, topics and words. Topics are latent variables, while the observed variables are words and documents. Assume that we have a dictionary of p words and a collection of n documents. Documents are sequences of words from the dictionary. The number of topics is denoted by K. Usually, $K \ll \min(p, n)$ and we will assume that $2 \leq K \leq \min(p, n)$. The pLSI model assumes that the probability of occurrence of word j in a document discussing topic k is independent of the document. Therefore, by the total probability formula,

$$\mathbb{P}(\text{word } j \mid \text{document } i) = \sum_{k=1}^{K} \mathbb{P}(\text{topic } k \mid \text{document } i)\, \mathbb{P}(\text{word } j \mid \text{topic } k).$$

Let $\Pi_{ij} := \mathbb{P}(\text{word } j \mid \text{document } i)$, $W_{ik} := \mathbb{P}(\text{topic } k \mid \text{document } i)$ and $A_{kj} := \mathbb{P}(\text{word } j \mid \text{topic } k)$. We can write $\Pi_{ij} = W_i^{\mathrm{T}} A_j$, where $W_i = (W_{i1}, \ldots, W_{iK})^{\mathrm{T}} \in [0,1]^K$ is the topic probability vector for document i and $A_j = (A_{1j}, \ldots, A_{Kj})^{\mathrm{T}} \in [0,1]^K$ is the vector of word j probabilities under topics $k = 1, \ldots, K$. Then,

$$\boldsymbol{\Pi} = \boldsymbol{W}\boldsymbol{A}, \tag{4.1}$$

where $\boldsymbol{\Pi}$ is $n \times p$ document-word matrix with entries Π_{ij}, $\boldsymbol{W} := (W_1, \ldots, W_n)^{\mathrm{T}}$ is $n \times K$ document-topic matrix and $\boldsymbol{A} := (A_1, \ldots, A_p)$ is $K \times p$ topic-word matrix. The rows of $\boldsymbol{\Pi}, \boldsymbol{W}$ and $\boldsymbol{A}$ are probability vectors:

$$\sum_{m=1}^{K} W_{im} = 1,\ \sum_{j=1}^{p} A_{kj} = 1,\ \sum_{j=1}^{p} \Pi_{ij} = 1 \text{ for any } i = 1, \ldots, n,\ k = 1, \ldots K. \tag{4.2}$$

Unless otherwise stated, we will assume throughout the paper that $\boldsymbol{\Pi}, \boldsymbol{W}, \boldsymbol{A}$ are matrices with non-negative entries satisfying (4.2). The value Π_{ij} is the probability of occurrence of word j in document i. It is not available but we have access to the corresponding empirical frequency X_{ij}. Thus, we have a document-word matrix $\boldsymbol{X} = (X_{ij})$ of size $n \times p$ such that for each document i in $1, \ldots, n$, and each word j in $1, \ldots, p$, the entry X_{ij} is the observed frequency of word j in document i. Let N_i denote the (non-random) number of sampled words in document i. We assume that, for each document-word vector $X_i = (X_{i1}, \ldots, X_{ip})^{\mathrm{T}}$, the corresponding vector of cumulative counts $N_i X_i$ follows a Multinomial$_p(N_i, \Pi_i)$ distribution, where $\Pi_i := \mathbb{E}(X_i) = (\Pi_{i1}, \ldots, \Pi_{ip})^{\mathrm{T}}$. We also assume that $X_1, \ldots, X_n$ are independent. We will denote by $\mathbb{P}_{\boldsymbol{\Pi}}$ the probability measure corresponding to the distribution of $\boldsymbol{X}$. We can write the observation model in a "signal + noise" form:

$$\boldsymbol{X} = \boldsymbol{\Pi} + \boldsymbol{Z} = \boldsymbol{W}\boldsymbol{A} + \boldsymbol{Z}, \tag{4.3}$$

where $\boldsymbol{Z} := \boldsymbol{X} - \boldsymbol{\Pi}$ is a zero mean noise. In topic modeling, the objective is to estimate the matrices $\boldsymbol{A}$ and $\boldsymbol{W}$ based on the observed frequency matrix $\boldsymbol{X}$ and on

the known $N_1, \dots, N_n$. The recovery of $\boldsymbol{A}$ and the recovery of $\boldsymbol{W}$ address different purposes. An estimator of $\boldsymbol{A}$ identifies the topic distribution on the dictionary. An estimator of $\boldsymbol{W}$ indicates the topics associated to each document.

The estimation of $\boldsymbol{W}$ has multiple applications and has been discussed mainly in the Bayesian perspective. The focus was on Latent Dirichlet Allocation (LDA) and related techniques (see Section 4 for more details and references). These methods are computationally slow and, to the best of our knowledge, no theoretical guarantees on their performance are available. On the other hand, the estimation of matrix $\boldsymbol{A}$ is well-studied in the theory. Several papers provide bounds on the performance of different estimators of $\boldsymbol{A}$. We give a more detailed account of this work in Section 4. Most of the results [Arora et al., 2013, Bing et al., 2020b, Bing et al., 2020d, Ke and Wang, 2017] use the *anchor word assumption* postulating that, for every topic, there is at least one word which occurs only in this topic.

At first sight, it seems that results on estimation of $\boldsymbol{A}$ can be applied to estimation of $\boldsymbol{W}$ by simply taking the transpose of (4.2) and inverting the roles of these two matrices. However, such an argument is not valid since the resulting models are different. Indeed, the rows of the matrix $\boldsymbol{X}^{\mathrm{T}}$ are not independent and the rows of the matrices $\boldsymbol{\Pi}^{\mathrm{T}}, \boldsymbol{A}^{\mathrm{T}}, \boldsymbol{W}^{\mathrm{T}}$ do not sum up to 1, which leads to a different statistical analysis.

Note that in some works on topic modeling, authors chose to estimate $\boldsymbol{A}$ first and treat the estimation of $\boldsymbol{W}$ as an easy problem, for example, by using least squares, given an estimator $\widehat{\boldsymbol{A}}$. The argument used to justify this approach is that the $K \times p$ matrix $\boldsymbol{A}$ can be learned more accurately as we have more documents (see, for example, [Bing et al., 2020b, Ke and Wang, 2017]) but the number of parameters in $n \times K$ matrix $\boldsymbol{W}$ increases as we increase n. However, this approach is questionable for several reasons. First, it is not always possible to get an accurate estimate of matrix $\boldsymbol{A}$, as it will be the case when we deal with a relatively large sized dictionary (cf. Remark 2 below). In such a situation, the existing algorithms for estimation of matrix $\boldsymbol{A}$ may be quite slow, and the error of estimating $\boldsymbol{A}$ passes on the estimation of $\boldsymbol{W}$ making it suboptimal (see the comments at the end of Section 4). Secondly, many of the existing methods of estimating $\boldsymbol{A}$ are based on the anchor word assumption that has been pointed out as a major limitation of spectral topic models. This assumption is not needed if we estimate matrix $\boldsymbol{W}$ directly.

In the present paper, we change the framework by focusing on estimation of $\boldsymbol{W}$ rather than $\boldsymbol{A}$. We introduce the following assumption:

Assumption 1 (Anchor document assumption)**.** *For each topic $k = 1, \dots, K$, there exists at least one document i (called an anchor document) such that $W_{ik} = 1$ and $W_{il} = 0$ for all $l \neq k$.*

Since each document is identified with a mixture of K topics, the anchor document assumption means that, for each topic, there is a document devoted solely to this topic. To illustrate the anchor document assumption, consider the Associated Press data set [Harman, 1993], which is a collection of 2246 articles published by this press agency mostly around 1988. An application of the pLSI model fitted via the

Document	Finance	Politics
1	0.248	0.752
2	0.362	0.638
3	0.527	0.473
4	0.357	0.643
5	0.181	0.819
6	0.001	0.999
7	0.773	0.227
8	0.004	0.996
9	0.967	0.033
10	0.147	0.953

Table 4.1: The first ten rows of the estimated matrix $\boldsymbol{W}$ for the Associated Press data set.

LDA method with $K = 2$ leads to two well-shaped topics "finance" and "politics". For the details of the analysis see [Silge and Robinson, 2020]. The first 9 rows of the estimator of $\boldsymbol{W}$ are presented in Table 4.1. Notice that documents 6 and 8 in Table 4.1 can be considered as anchor documents. For example, document 6 has the weight of the second topic estimated as 0.999. A closer look at the most frequent words in this document (Noriega, Panama, Jackson, Powell, administration, economic, general) tells us that, indeed, this article corresponds solely to the topic "politics" – it is about the relationship between the American government and the Panamanian leader Manuel Noriega.

Our approach to estimation of $\boldsymbol{W}$, that we call Successive Projection Overlapping Clustering (SPOC), is inspired by the Successive Projection Algorithm (SPA) initially proposed for non-negative matrix factorization [Araujo et al., 2001] and further used in the context of mixed membership stochastic block models [Gillis and Vavasis, 2014, Panov et al., 2017, Mao et al., 2020]. The idea of our method is to start with the singular value decomposition (SVD) of $\boldsymbol{X}$ and launch the SPA on the matrix of singular vectors. This gives an iterative procedure that, at each step, chooses the maximum norm row of the matrix of singular vectors and then projects on the linear subspace orthogonal to the selected row. From a geometric perspective, the rows of the matrix of singular vectors of $\boldsymbol{\Pi}$ belong to a simplex in $\mathbb{R}^K$. The documents can be identified with some points in this simplex and the anchor documents with its vertices. Our algorithm iteratively finds estimators of the vertices, based on which we estimate the topic-document matrix $\boldsymbol{W}$.

The idea of exploiting simplex structures was previously applied for estimation of matrix $\boldsymbol{A}$, see [Arora et al., 2013, Ding et al., 2013, Ke and Wang, 2017], among others. For example, the method to estimate $\boldsymbol{A}$ suggested in [Ke and Wang, 2017] is based on an exhaustive search over all size K subsets of $\{1, \dots, p\}$. Its goal is to select K vertices of a p-dimensional simplex and its computational cost is at least of the order p^K. Our algorithm recovers the vertices of a K-dimensional simplex

(recall that $K \ll p$), and its computational cost is of the order $\max(p, n)K + nK^2$. Here, $\max(p, n)K$ and nK^2 are the costs of performing a truncated SVD and SPA, respectively.

Our theoretical results deal with the problem of estimating the document-topic matrix $\boldsymbol{W}$. In practice, our method can be used for estimation of $\boldsymbol{A}$ as well. Based on the SPOC estimator of $\boldsymbol{W}$, we can obtain an estimator of matrix $\boldsymbol{A}$ by a computationally fast procedure (see Section 2). Our simulation studies (see Section 10) indicate that this estimator exhibits a behavior similar to the LDA on average while being more stable.

Assuming that $N_i = N$ for $i = 1, \dots, n$, we prove that the SPOC estimator of $\boldsymbol{W}$ converges in the Frobenius norm and in the ℓ_1-norm with the rates $\sqrt{n/N}$ and $n/\sqrt{N}$, up to a weak factor, respectively. (We mean by *weak factor* a small power of K multiplied by a term logarithmic in the parameters of the problem. We will ignore weak factors when discussing the convergence rates.) We also prove lower bounds of the order $\sqrt{n/N}$ and $n/\sqrt{N}$, respectively, implying near optimality of the proposed method. One of the conclusions, both from the theory and the numerical experiments, is that the error of the SPOC algorithm does not grow significantly with the size of the dictionary p, in contrast to what one observes for Latent Dirichlet Allocation. We also introduce an estimator for the number K of topics, which is usually unknown in practice. We show that SPOC algorithm using the estimator of K preserves its optimal properties in this more challenging setting.

We stress that the minimax convergence rates for estimation of matrix $\boldsymbol{W}$ established in this paper cannot be improved by *any* estimation method. In particular, our results imply that an accurate estimation of matrix $\boldsymbol{W}$ requires the number of words per document N to be large. In practice, when the topic-word matrix $\boldsymbol{A}$ can be accurately estimated from the data one can estimate document-topic matrix $\boldsymbol{W}$ via, for example, least squares. However, accurate estimators of $\boldsymbol{A}$ are only available when p (the size of the dictionary) is relatively small while n (the total number of documents) and N are large. On the other hand, when p is large compared to everything else, the error of estimating $\boldsymbol{A}$ could be too high to allow for a good estimator of $\boldsymbol{W}$ in the above scheme. But our algorithm that estimates $\boldsymbol{W}$ directly still works.

The rest of the chapter is organized as follows. In Section 2, we introduce the SPOC algorithm. Section 3 contains the main results on the convergence rate of the algorithm and the minimax lower bound for estimation of $\boldsymbol{W}$. In Section 5, we present numerical experiments for synthetic and real-world data in order to illustrate our theoretical findings. Finally, in Section 6 we summarize the outcomes of the study. Last sections contains proofs, additional simulations and a detailed discussion of related work.

Notation. For any matrix $\boldsymbol{M} = (M_{ij}) \in \mathbb{R}^{n \times k}$, $\|\boldsymbol{M}\|$ denotes its spectral norm, i.e., its maximal singular value, $\|\boldsymbol{M}\|_F$ its Frobenius norm, and $\|\boldsymbol{M}\|_1 = \sum_{i=n}^{k} \sum_{j=1}^{k} |M_{ij}|$ its ℓ_1-norm. We also consider the maximum ℓ_1-norm of its rows $\|\boldsymbol{M}\|_{1,\infty} = \max\limits_{1 \le i \le n} \sum_{j=1}^{k} |M_{ij}|$. We denote by $\lambda_j(\boldsymbol{M})$ the jth singular value and by

$\lambda_{\min}(\boldsymbol{M})$ the smallest singular value of $\boldsymbol{M}$. Assuming that matrix $\boldsymbol{M}$ has rank K we consider its singular values $\lambda_1(\boldsymbol{M}) \geq \lambda_2(\boldsymbol{M}) \geq \cdots \geq \lambda_K(\boldsymbol{M}) > 0$ and its condition number $\kappa(\boldsymbol{M}) = \lambda_1(\boldsymbol{M})/\lambda_K(\boldsymbol{M})$. If J is a non-empty subset of rows of matrix $\boldsymbol{M}$ the notation $\boldsymbol{M}_J$ is used for a matrix in $\mathbb{R}^{|J|\times k}$ obtained from $\boldsymbol{M}$ by keeping only the rows in J. We denote by $\boldsymbol{I}_K$ the $K \times K$ identity matrix, and by $(\mathbf{e}_1, \ldots, \mathbf{e}_n)$ the canonical basis of $\mathbb{R}^n$. For any vector $u \in \mathbb{R}^d$, we denote by $\|u\|_2$ its Euclidean norm. Throughout the paper, we use the notation $\mathbf{O}$ for orthogonal matrices and $\mathbf{P}$ for permutation matrices. $\mathscr{P}$ denotes the set of all permutation matrices in $\mathbb{R}^{K\times K}$ and c, C positive constants that may vary from line to line.

2 Successive Projection Overlapping Clustering

We start by introducing the *Successive Projection Overlapping Clustering (SPOC)* algorithm. It is an analog, in the context of topic models, of the algorithm proposed in [Panov et al., 2017] for the problem of parameter estimation in Mixed Membership Stochastic Block Model. To explain the main idea of the algorithm, we consider the singular value decomposition (SVD) of matrix $\boldsymbol{\Pi}$. Using assumption $K \leq \min(p, n)$ and that $\operatorname{rank} \boldsymbol{\Pi} \leq K$ we get:

$$\boldsymbol{\Pi} = \mathbf{U}\mathbf{L}\mathbf{V}^{\mathrm{T}}, \tag{4.4}$$

where $\mathbf{U} = [U_1, \ldots, U_K] \in \mathbb{R}^{n\times K}$ and $\mathbf{V} = [V_1, \ldots, V_K] \in \mathbb{R}^{p\times K}$ are matrices of left and right singular vectors and $\mathbf{L} \in \mathbb{R}^{K\times K}$ is the diagonal matrix of the corresponding singular values. Under Assumption 1, if $\lambda_K(\boldsymbol{\Pi}) > 0$, the key observation is that $\mathbf{U}$ can be represented as

$$\mathbf{U} = \boldsymbol{W}\mathbf{H}, \tag{4.5}$$

where $\mathbf{H} \in \mathbb{R}^{K\times K}$ is a full rank matrix (cf. Lemma 6). Thus, the rows of matrix $\mathbf{U}$ belong to a simplex in $\mathbb{R}^K$ with vertices given by the rows of the matrix $\mathbf{H}$. The empirical counterparts of $\mathbf{U}, \mathbf{L}, \mathbf{V}$ are obtained from the SVD of $\boldsymbol{X}$:

$$\boldsymbol{X} = \hat{\mathbf{U}}\hat{\mathbf{L}}\widehat{\mathbf{V}}^{\mathrm{T}} + \hat{\mathbf{U}}_{\mathbf{1}}\hat{\mathbf{L}}_{\mathbf{1}}\widehat{\mathbf{V}}_{\mathbf{1}}^{\mathrm{T}}, \tag{4.6}$$

where $\hat{\mathbf{U}} = [\widehat{U}_1, \ldots, \widehat{U}_K]$ and $\widehat{\mathbf{V}} = [\widehat{V}_1, \ldots, \widehat{V}_K]$ are, respectively, the matrices of left and right singular vectors of $\boldsymbol{X}$ corresponding to its K leading singular values $\widehat{\lambda}_1 \geq \cdots \geq \widehat{\lambda}_K$; $\widehat{\mathbf{L}} = \operatorname{diag}\{\widehat{\lambda}_1, \ldots, \widehat{\lambda}_K\}$, and $\hat{\mathbf{U}}_1\hat{\mathbf{L}}_1\widehat{\mathbf{V}}_1^{\mathrm{T}}$ is the singular value decomposition of $\boldsymbol{X} - \hat{\mathbf{U}}\hat{\mathbf{L}}\widehat{\mathbf{V}}^{\mathrm{T}}$. It follows from the matrix perturbation theory (see 7.1) that there exists an orthogonal matrix $\mathbf{O}$ such that $\hat{\mathbf{U}}$ is a good approximation for $\mathbf{U}\mathbf{O}$ and we can write

$$\hat{\mathbf{U}} = \mathbf{U}\mathbf{O} + \mathbf{N} = \boldsymbol{W}\mathbf{H}\mathbf{O} + \mathbf{N}, \tag{4.7}$$

where $\mathbf{N}$ is a "small enough" noise matrix. Having obtained $\hat{\mathbf{U}}$ from the SVD of $\boldsymbol{X}$, we then apply the *Successive Projection Algorithm* (SPA) [Araujo et al., 2001, Gillis

and Vavasis, 2014] to estimate matrix $\mathbf{HO}$ in (4.7). Applied to matrix $\boldsymbol{M} = \hat{\mathbf{U}}$ and $r = K$ this algorithm finds the rows of matrix $\hat{\mathbf{U}}$ with the maximum Euclidean norm and then projects on the subspace orthogonal to these rows and repeats the procedure until K rows are selected. The main idea underlying the SPA is that the maximum of the Euclidean norm of a vector on a simplex is attained at one of its vertices.

Algorithm 1 SPA

Input: Matrix $\boldsymbol{M} \in \mathbb{R}^{n \times K}$ and integer $r \leq n$.
Output: Set of indices $J \subseteq \{1, \dots, n\}$.
1: Initialize: $\boldsymbol{S}_0 = \boldsymbol{M}^{\mathrm{T}}$, $J_0 = \emptyset$.
2: For $t = 1, \dots, r$ do:
– Find $i(t) = \arg\max_{i=1,\dots,n} \|\mathbf{s}_i\|_2$, where $\mathbf{s}_i$'s are the column vectors of $\boldsymbol{S}_{t-1}$.
– Set $\boldsymbol{S}_t = \left(\boldsymbol{I}_K - \frac{\mathbf{s}_{i(t)}\mathbf{s}_{i(t)}^{\mathrm{T}}}{\|\mathbf{s}_{i(t)}\|_2^2}\right)\boldsymbol{S}_{t-1}, \quad J_t = J_{t-1} \cup \{i(t)\}$.
3: Set $J = J_r$.

If Assumption 1 holds in the noiseless case (i. e. when $\mathbf{N} = 0$), it can be shown that $\hat{\mathbf{U}}_J = \mathbf{HO}$, where J is the set of K rows of $\hat{\mathbf{U}}$ selected after K steps of SPA. In the noisy case we need additional assumptions on the noise level to ensure that SPA extracts documents close to anchor ones, which leads to an accurate enough estimator $\hat{\mathbf{H}}$ of $\mathbf{HO}$ (see 7.2 for the precise statement). Once we have such estimator, the final step is to define our estimator of matrix $\boldsymbol{W}$ as $\hat{\boldsymbol{W}} = \hat{\mathbf{U}}\hat{\mathbf{H}}^{-1}$. This definition is valid only if matrix $\hat{\mathbf{H}}$ is non-degenerate, which is true with high probability under suitable assumptions (cf. Section 3). An additional potentially useful step is to apply preconditioning to matrix $\hat{\mathbf{U}}$, which leads to improved bounds on the performance of the algorithm in the presence of noise, see [Gillis and Vavasis, 2015, Mizutani, 2016]. Preconditioned SPA is defined as follows. Let $r = K$ and let $\mathbf{a}_1, \dots, \mathbf{a}_n$ be the column vectors of matrix $\boldsymbol{M}^{\mathrm{T}}$. Let $\boldsymbol{L}^* \in \mathbb{R}^{K \times K}$ be the solution of the following minimization problem

$$\min_{\boldsymbol{L} \succ 0:\ \max_i \mathbf{a}_i^{\mathrm{T}} \boldsymbol{L} \mathbf{a}_i \leq 1} -\log\det \boldsymbol{L}. \tag{4.8}$$

Matrix $\boldsymbol{L}^*$ defines the minimum volume ellipsoid centered at the origin that contains $\mathbf{a}_1, \dots, \mathbf{a}_n$. The preconditioned SPA is defined by Algorithm 1 initialized with $\boldsymbol{S}_0 = (\boldsymbol{L}^*)^{1/2}\boldsymbol{M}^{\mathrm{T}}$ rather than with $\boldsymbol{S}_0 = \boldsymbol{M}^{\mathrm{T}}$. The SPOC algorithm for topic modeling is summarized in Algorithm 2.

Based on the SPOC estimator $\hat{\boldsymbol{W}}$ of matrix $\boldsymbol{W}$, it is possible to construct an estimator for matrix $\boldsymbol{A}$ in a straightforward way. Indeed, given the decompositions (4.4) and (4.5), we can use the definition $\boldsymbol{\Pi} = \boldsymbol{W}\boldsymbol{A}$ and deduce that $\boldsymbol{A} = \mathbf{H}\mathbf{L}\mathbf{V}^{\mathrm{T}}$. A direct sample-based estimator of $\boldsymbol{A}$ is then given by

$$\hat{\boldsymbol{A}} = \hat{\mathbf{H}}\hat{\mathbf{L}}\widehat{\mathbf{V}}^{\mathrm{T}}. \tag{4.9}$$

Algorithm 2 SPOC (respectively, preconditioned SPOC)

Input: Observed matrix $\boldsymbol{X}$ and number of topics K.
Output: Estimated document-topic matrix $\hat{\boldsymbol{W}}$.
1: Get the rank K SVD of $\boldsymbol{X}$: $\hat{\mathbf{U}}\hat{\mathbf{L}}\hat{\mathbf{V}}^{\mathrm{T}}$.
2: Run SPA (respectively, preconditioned SPA) with input $(\hat{\mathbf{U}}, K)$, which outputs a set of indices J with cardinality K.
3: Set $\hat{\mathbf{H}} := \hat{\mathbf{U}}_J$.
4: Set $\hat{\boldsymbol{W}} := \hat{\mathbf{U}}\hat{\mathbf{H}}^{-1}$.

In order to illustrate the performances of this new estimator, we have performed experiments to compare it with LDA, see Section 10.

3 Main results

In this section, we provide bounds on the performance of the SPOC algorithm. We first prove deterministic bounds assuming that $\boldsymbol{X}$ is some fixed matrix close enough to $\boldsymbol{\Pi}$ in the spectral norm. Next, we combine these results with a concentration inequality for $\|\boldsymbol{X} - \boldsymbol{\Pi}\|$ when $\boldsymbol{X}$ is distributed according to $\mathbb{P}_{\boldsymbol{\Pi}}$ in order to obtain a bound on the estimation error with high probability under our statistical model.

3.1 Deterministic bounds

A key step in analyzing the performance of the SPOC algorithm is to show that $\hat{\mathbf{U}}$ is close to an orthogonal transformation $\mathbf{UO}$ of the population matrix $\mathbf{U}$. The next lemma gives a bound on the maximal ℓ_2-distance between the rows of $\hat{\mathbf{U}}$ and $\mathbf{UO}$ for some orthogonal matrix $\mathbf{O}$. This lemma will allow us to deduce an upper bound on the error of SPA (see 7.2 for the details). Recall that $\lambda_1(\boldsymbol{W})$ is the maximum singular value of matrix $\boldsymbol{W}$, $\lambda_K(\boldsymbol{\Pi})$ is the Kth singular value of matrix $\boldsymbol{\Pi}$, and $\kappa(\boldsymbol{W})$, $\kappa(\boldsymbol{\Pi})$ are the condition numbers of matrices $\boldsymbol{W}$ and $\boldsymbol{\Pi}$, respectively. Assuming that $\lambda_K(\boldsymbol{\Pi}) > 0$ (that is, $\boldsymbol{\Pi}$ is a rank K matrix) we define

$$\beta_i(\boldsymbol{X}, \boldsymbol{\Pi}) = K^{1/2}\kappa^2(\boldsymbol{\Pi}) \, \frac{\|\mathbf{e}_i^{\mathrm{T}}\boldsymbol{X}\|_2 \|\boldsymbol{X} - \boldsymbol{\Pi}\|}{\lambda_K^2(\boldsymbol{\Pi})} + \frac{\|\mathbf{e}_i^{\mathrm{T}}(\boldsymbol{X} - \boldsymbol{\Pi})\|_2}{\lambda_K(\boldsymbol{\Pi})}, \quad i = 1, \ldots, n,$$

where $(\mathbf{e}_1, \ldots, \mathbf{e}_n)$ is the canonical basis of $\mathbb{R}^n$.

Lemma 1. *Assume that $\boldsymbol{\Pi} \in \mathbb{R}^{n \times p}$ is a rank K matrix, and $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ is any matrix such that $\|\boldsymbol{X} - \boldsymbol{\Pi}\| \leq \lambda_K(\boldsymbol{\Pi})/2$. Let $\hat{\mathbf{U}}, \mathbf{U}$ be the $n \times K$ matrices of left singular vectors corresponding to the top K singular values of $\boldsymbol{X}$ and $\boldsymbol{\Pi}$, respectively. Then, there exist an orthogonal matrix $\mathbf{O}$ and a constant $C > 0$ such that, for any $i = 1, \ldots, n$,*

$$\|\mathbf{e}_i^{\mathrm{T}}(\hat{\mathbf{U}} - \mathbf{UO})\|_2 \leq C\beta_i(\boldsymbol{X}, \boldsymbol{\Pi}).$$

Define now $\beta(\boldsymbol{X}, \boldsymbol{\Pi}) = \max_{i=1,\dots,n} \beta_i(\boldsymbol{X}, \boldsymbol{\Pi})$. We will need the following condition:

Assumption 2. *For a constant $\bar{C} > 0$ we have*

$$\beta(\boldsymbol{X}, \boldsymbol{\Pi}) \leq \frac{\bar{C}}{\lambda_1(\boldsymbol{W})\kappa(\boldsymbol{W})K\sqrt{K}}.$$

Assumption 2 is satisfied with high probability for $\boldsymbol{X} \sim \mathbb{P}_{\boldsymbol{\Pi}}$ for N, the sample size, large enough (see Section 8.3). Under Assumption 2, we can derive from Lemma 1 the following deterministic bound on the error of estimating the document-topic matrix by the SPOC algorithm:

Lemma 2. *Let Assumptions 1 and 2 be satisfied with constant $\bar{C}$ small enough. Assume that $\boldsymbol{\Pi} \in \mathbb{R}^{n\times p}$ is a rank K matrix, and $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ is any matrix such that $\|\boldsymbol{X} - \boldsymbol{\Pi}\| \leq \lambda_K(\boldsymbol{\Pi})/2$. Then, matrix $\hat{\mathbf{H}}$ is non-degenerate and the preconditioned SPOC algorithm outputs matrix $\hat{\boldsymbol{W}}$ such that*

$$\min_{\mathbf{P}\in\mathscr{P}} \|\hat{\boldsymbol{W}} - \boldsymbol{W}\mathbf{P}\|_F \leq CK^{1/2}\left\{\lambda^2_{\max}(\boldsymbol{W})\kappa(\boldsymbol{W})\beta(\boldsymbol{X}, \boldsymbol{\Pi}) + \frac{\kappa(\boldsymbol{\Pi})\lambda_1(\boldsymbol{W})\|\boldsymbol{X} - \boldsymbol{\Pi}\|}{\lambda_K(\boldsymbol{\Pi})}\right\},$$

where $\mathscr{P}$ denotes the set of all permutation matrices.

Inspection of the proof shows that, for this lemma to hold, it is enough to choose the constant $\bar{C} \leq \min(C_*, C_0^{-1})$ where C_*, C_0 are the constants from Theorem 4 and Corollary 5.

3.2 Bounds with high probability

Lemma 2 combined with a concentration inequality for $\|\boldsymbol{X} - \boldsymbol{\Pi}\|$ (cf. Lemma 4 in the Section) allows us to derive a bound for the estimation error that holds with high probability when $\boldsymbol{X}$ is sampled from distribution $\mathbb{P}_{\boldsymbol{\Pi}}$. Introduce the value

$$\Delta(\boldsymbol{W}, \boldsymbol{\Pi}) = \left(\frac{\lambda_1(\boldsymbol{W})}{\lambda_K(\boldsymbol{\Pi})}\right)^2 \kappa(\boldsymbol{W})\kappa^2(\boldsymbol{\Pi}).$$

The main result is summarized in the next theorem.

Theorem 1. *Let Assumption 1 hold, and $N_i = N$ for $i = 1, \dots, n$. Assume that $N \geq \log(n + p)$ and*

$$\lambda_K(\boldsymbol{\Pi}) \geq \sqrt{\frac{10}{\bar{C}}}\, K \left(\frac{n\log(n+p)}{N}\right)^{1/4} \kappa(\boldsymbol{\Pi})\sqrt{\lambda_1(\boldsymbol{W})\kappa(\boldsymbol{W})}. \tag{4.10}$$

Then, with probability at least $1 - 2(n+p)^{-1}$*, matrix* $\hat{\mathbf{H}}$ *is non-degenerate and the output* $\hat{\boldsymbol{W}}$ *of preconditioned SPOC algorithm satisfies, for some constant* $\mathrm{C}_1 > 0$*,*

$$\min_{\mathbf{P}\in\mathscr{P}} \left\| \hat{\boldsymbol{W}} - \boldsymbol{W}\mathbf{P} \right\|_F \leq \mathrm{C}_1 K \sqrt{\frac{n\log(n+p)}{N}} \Delta(\boldsymbol{W}, \boldsymbol{\Pi}),$$

where $\mathscr{P}$ *denotes the set of all permutation matrices.*

Condition (4.10) in Theorem 1 guarantees that Assumption 2 is satisfied. This condition holds if N is large enough and it quantifies the separation of the spectrum of the matrix $\boldsymbol{\Pi}$ from zero. The bound of Theorem 1 depends on the singular values of matrices $\boldsymbol{W}$ and $\boldsymbol{\Pi}$. We now further detail this bound for the balanced case where matrices $\boldsymbol{W}$ and $\boldsymbol{\Pi}$ are well conditioned and the smallest non-zero singular value of $\boldsymbol{\Pi}$ is of the same order as the largest singular value of $\boldsymbol{W}$. It follows from Lemma 7 that in this case both $\lambda_K(\boldsymbol{\Pi})$ and $\lambda_1(\boldsymbol{W})$ are of the order of $\sqrt{n/K}$. This is coherent with the behavior of the singular values of $\boldsymbol{\Pi}$ and $\boldsymbol{W}$ that we observed in the simulation study (see Section 11). The balanced case is formally described by the following assumption.

Assumption 3. *There exist two constants* $C > 1$ *and* $c > 0$ *such that*

$$\lambda_K(\boldsymbol{\Pi}) \geq C\lambda_1(\boldsymbol{W}) \quad \textit{and} \quad \max\left\{\kappa(\boldsymbol{\Pi}), \kappa(\boldsymbol{W})\right\} \leq c.$$

The second condition in Assumption 3 is quite standard and just states that matrices $\boldsymbol{\Pi}$ and $\boldsymbol{W}$ are well-conditioned. The first condition is more restrictive. It holds, in particular, if matrix $\boldsymbol{A}$ is well-conditioned with large enough $\lambda_K(\boldsymbol{A})$. For example, it will be the case if $\boldsymbol{A}$ satisfies the *anchor word assumption* (see Section 1) with the probabilities of anchor words uniformly above the probabilities of other words. This is detailed in Lemma 9. Noteworthy, the lower bound of Theorem 3 below is attained with such choice of matrix $\boldsymbol{A}$, see the proof of Theorem 3 in Section 8.5. We can interpret it as the fact that, in a minimax sense, such matrices $\boldsymbol{A}$ are associated with the least favorable models. The following corollary quantifies the behavior of SPOC estimator in the balanced case:

Corollary 1 (Upper bound in the balanced case)**.** *Let Assumptions 1 and 3 hold, and* $N_i = N$ *for* $i = 1, \ldots, n$*. Let also*

$$N \geq CK^5 \log(n+p) \tag{4.11}$$

for some $C > 0$ *large enough. Then, with probability at least* $1 - 2(n+p)^{-1}$*, matrix* $\hat{\mathbf{H}}$ *is non-degenerate and the output* $\hat{\boldsymbol{W}}$ *of preconditioned SPOC algorithm satisfies, for some constant* $\mathrm{C}_2 > 0$*,*

$$\min_{\mathbf{P}\in\mathscr{P}} \left\| \hat{\boldsymbol{W}} - \boldsymbol{W}\mathbf{P} \right\|_F \leq \mathrm{C}_2 K \sqrt{\frac{n\log(n+p)}{N}},$$

where $\mathscr{P}$ *denotes the set of all permutation matrices.*

To prove Corollary 1, it is enough to notice that under Assumption 3 we have $\Delta(\boldsymbol{W}, \boldsymbol{\Pi}) \leq C'$ for some constant $C' > 0$, and condition (4.10) follows from (4.11), Assumption 3 and the inequality $\lambda_1(\boldsymbol{W}) \geq \sqrt{n/K}$ (see Lemma 7). Note that from Theorem 1 and Corollary 1 we can derive bounds in other norms. Thus, using the inequalities $\left\|\hat{\boldsymbol{W}} - \boldsymbol{W}\mathbf{P}\right\|_1 \leq \sqrt{Kn}\left\|\hat{\boldsymbol{W}} - \boldsymbol{W}\mathbf{P}\right\|_F$ and $\left\|\hat{\boldsymbol{W}} - \boldsymbol{W}\mathbf{P}\right\|_{1,\infty} \leq \sqrt{K}\left\|\hat{\boldsymbol{W}} - \boldsymbol{W}\mathbf{P}\right\|_F$ we obtain the following corollary:

Corollary 2. *If the assumptions of Theorem 1 are satisfied then, with probability at least* $1 - 2(n+p)^{-1}$*, matrix* $\hat{\mathbf{H}}$ *is non-degenerate and the output* $\hat{\boldsymbol{W}}$ *of preconditioned SPOC algorithm satisfies*

$$\min_{\mathbf{P}\in\mathscr{P}}\left\|\hat{\boldsymbol{W}} - \boldsymbol{W}\mathbf{P}\right\|_{1,\infty} \leq \mathrm{C}_1 K^{3/2}\sqrt{\frac{n\log(n+p)}{N}}\Delta(\boldsymbol{W}, \boldsymbol{\Pi}) \quad \textit{and}$$
$$\min_{\mathbf{P}\in\mathscr{P}}\left\|\hat{\boldsymbol{W}} - \boldsymbol{W}\mathbf{P}\right\|_{1} \leq \mathrm{C}_1 K^{3/2} n\sqrt{\frac{\log(n+p)}{N}}\Delta(\boldsymbol{W}, \boldsymbol{\Pi}).$$

If the assumptions of Corollary 1 are satisfied then with probability at least $1 - 2(n+p)^{-1}$ *matrix* $\hat{\mathbf{H}}$ *is non-degenerate and the output* $\hat{\boldsymbol{W}}$ *of preconditioned SPOC algorithm satisfies*

$$\min_{\mathbf{P}\in\mathscr{P}}\left\|\hat{\boldsymbol{W}} - \boldsymbol{W}\mathbf{P}\right\|_{1,\infty} \leq \mathrm{C}_2 K^{3/2}\sqrt{\frac{n\log(n+p)}{N}} \quad \textit{and}$$
$$\min_{\mathbf{P}\in\mathscr{P}}\left\|\hat{\boldsymbol{W}} - \boldsymbol{W}\mathbf{P}\right\|_{1} \leq \mathrm{C}_2 K^{3/2} n\sqrt{\frac{\log(n+p)}{N}}.$$

It follows from Corollaries 1 and 2 that the rate of estimating $\boldsymbol{W}$ (to within a weak factor) is determined by two parameters, which are the number of documents n and the sample size N. The dependence on the size of the dictionary p is weak. This is confirmed by the numerical experiments, see Section 5.

3.3 Adaptive procedure when K is unknown.

We now propose an adaptive variant of the SPOC algorithm when the number of topics K is unknown. It is obtained by replacing K in Algorithm 2 by the estimator

$$\hat{K} = \max\left\{j : \ \lambda_j(\boldsymbol{X}) > 4\sqrt{\frac{n\log(n+p)}{N}}\right\}.$$

In the sequel, the resulting procedure will be called the adaptive (preconditioned) SPOC algorithm. The following analogs of Theorem 1 and Corollary 1 hold.

Theorem 2. *Let the assumptions of Theorem 1 be satisfied and*

$$\lambda_1(\boldsymbol{W}) > \frac{32\bar{C}}{5K^2}\sqrt{\frac{n\log(n+p)}{N}}. \tag{4.12}$$

Then, with probability at least $1 - 2(n+p)^{-1}$, *matrix* $\hat{\mathbf{H}}$ *is non-degenerate,* $\hat{K} = K$, *and the output* $\hat{\boldsymbol{W}}$ *of the adaptive preconditioned SPOC algorithm satisfies*

$$\min_{\mathbf{P}\in\mathscr{P}} \left\| \hat{\boldsymbol{W}} - \boldsymbol{W}\mathbf{P} \right\|_F \leq \mathrm{C}_1 K \sqrt{\frac{n\log(n+p)}{N}} \Delta(\boldsymbol{W}, \boldsymbol{\Pi}).$$

Corollary 3. *Let the assumptions of Corollary 1 and (4.12) be satisfied. Then, with probability at least* $1 - 2(n+p)^{-1}$, *matrix* $\hat{\mathbf{H}}$ *is non-degenerate,* $\hat{K} = K$, *and the output* $\hat{\boldsymbol{W}}$ *of the adaptive preconditioned SPOC algorithm satisfies*

$$\min_{\mathbf{P}\in\mathscr{P}} \left\| \hat{\boldsymbol{W}} - \boldsymbol{W}\mathbf{P} \right\|_F \leq \mathrm{C}_2 K \sqrt{\frac{n\log(n+p)}{N}}.$$

Note that condition (4.12) introduced in Theorem 2 and Corollary 3 additionally to the conditions of Theorem 1 and Corollary 1 is rather mild. Indeed, due to inequality (4.42) we have $\lambda_1(\boldsymbol{W}) \geq \sqrt{n/K}$. Therefore, it is sufficient that $N > C\log(n+p)/K^3$ to grant (4.12).

3.4 Minimax lower bound

The following lower bound shows that the rate obtained in Corollary 1 is near minimax optimal. Denote by $\mathscr{M}$ the class of all matrices $\boldsymbol{\Pi}$ satisfying the assumptions stated in the Section 1 and Assumption 3.

Theorem 3 (Lower bound)**.** *Assume that* $N_i = N$ *for* $i = 1, \ldots, n$ *and* $2 \leq K \leq \min(p/4, N/2, n/2)$. *Then, there exist two constants* $C > 0$ *and* $c \in (0,1)$ *such that, for any estimator and* $\overline{\boldsymbol{W}}$ *of* $\boldsymbol{W}$ *we have*

$$\sup_{\boldsymbol{\Pi}\in\mathscr{M}} \mathbb{P}_{\boldsymbol{\Pi}} \left\{ \min_{\mathbf{P}\in\mathscr{P}} \|\overline{\boldsymbol{W}} - \boldsymbol{W}\mathbf{P}\|_F \geq C\sqrt{\frac{n}{N}} \right\} \geq c, \tag{4.13}$$

and

$$\sup_{\boldsymbol{\Pi}\in\mathscr{M}} \mathbb{P}_{\boldsymbol{\Pi}} \left\{ \min_{\mathbf{P}\in\mathscr{P}} \|\overline{\boldsymbol{W}} - \boldsymbol{W}\mathbf{P}\|_1 \geq Cn\sqrt{\frac{K}{N}} \right\} \geq c, \tag{4.14}$$

where $\mathscr{P}$ *denotes the set of all permutation matrices.*

Combining Corollary 1 and (4.13) we find that the minimax optimal rate of estimation of $\boldsymbol{W}$ on the class $\mathscr{M}$ in the Frobenius norm scales as $\sqrt{n/N}$. On the other hand, (4.14) and Corollary 2 imply that the minimax optimal rate of estimation of $\boldsymbol{W}$ in the l_1-norm on $\mathscr{M}$ scales as $n/\sqrt{N}$.

Remark 1. *Inspection of the proof of Theorem 3 shows that the lower bound is in fact established for a subset of* $\mathscr{M}$ *composed of matrices satisfying both anchor word and anchor document assumptions.*

Remark 2. *Under the same observation model and the anchor word assumption, the minimax optimal rate for estimation of matrix* $\boldsymbol{A}$ *in the* ℓ_1*-norm scales as* $\sqrt{p/nN}$, *see [Ke and Wang, 2017, Bing et al., 2020b]. Note that this rate is determined by all*

the three main parameters of the problem - the size of the dictionary p, the number of documents n and the sample size N. This is quite different from the minimax ℓ_1-rate $n/\sqrt{N}$ of estimation $\boldsymbol{W}$, which remains valid under anchor word assumption, cf. Remark 1 and the remark after Assumption 3. It shows that there is a significant difference between the problems of estimating matrices $\boldsymbol{A}$ and $\boldsymbol{W}$ in topic models.

4 Related Work

There exists an extensive literature on topic modeling and several algorithms have been proposed to estimate the matrices $\boldsymbol{A}$ and $\boldsymbol{W}$. As the problem of recovering these two matrices when there is no noise is an instance of non-negative matrix factorization, several papers propose algorithms based on minimization of a regularized cost function, see, e.g., [Lee and Seung, 1999b, Donoho and Stodden, 2004, Cichocki et al., 2009, Recht et al., 2012]. Such methods result in non-convex optimization and often fail when many words do not appear in a single document, that is, when $N \ll p$. Also, spectral analysis methods have long been used in related problems, see, for example, [Azar et al., 2001].

Another approach is to use Bayesian methods such as the popular *Latent Dirichlet Allocation* (LDA) introduced in [Blei et al., 2003]. LDA proceeds by imposing a Dirichlet prior on $\boldsymbol{A}$ and then computing an estimator of $\boldsymbol{W}$ by a variational EM-algorithm. The original paper [Blei et al., 2003] and the subsequent line of work do not provide statistical guarantee on the recovery of $\boldsymbol{W}$. LDA avoids two issues of the pLSI that are the risk of overfitting and the difficulty of classifying a new document outside the corpus, see [Blei et al., 2003] for more details. Yet, LDA is computationally slow and makes the assumption that topics are uncorrelated, which may be unrealistic [Blei and Lafferty, 2007, Li and McCallum, 2006]. This last issue has been addressed in [Lafferty and Blei, 2006] by introducing *Correlated topic models*. LDA has been extended to relax some assumptions such as the *bag-of-words* hypothesis ("order of words does not matter") [Wallach, 2006], the exchangeability of documents ("topics do not vary in time") [Blei and Lafferty, 2006], the assumption that the number of topics is known [Teh et al., 2005]. Also, to recover $\boldsymbol{W}$ in the LDA setting, some papers used Gibbs-sampling [Ramage et al., 2009, Porteous et al., 2008] or variational Bayes techniques [Zhai et al., 2012, Chien and Chueh, 2010] rather than the EM-algorithm. However, these works do not provide statistical guarantees on the estimation of $\boldsymbol{W}$ and the associated algorithms are computationally slow.

Paper [Li et al., 2015] adopts a more general approach than LDA model by considering a statistical mixture model, which includes topic models with LDA. Using spectral and transportation techniques, they provide an estimation method for $\boldsymbol{A}$, which does not require any assumption of structure such as the *anchor word assumption* but has weak statistical guarantees.

In [Arora et al., 2016], the authors focus on labeling a single document when $\boldsymbol{A}$ is known, which means finding the proportion of each topic in this document, with

a fixed set of topics. They solve this problem assuming the true topic proportion vector is sparse and the number of topics is large (typically $K = 100$).

For the problem of estimation of matrix $\boldsymbol{A}$, papers [Arora et al., 2012, Anandkumar et al., 2012, Arora et al., 2013, Ding et al., 2013, Anandkumar et al., 2014, Bansal et al., 2014, Ke and Wang, 2017, Bing et al., 2020b], to mention but a few, provide provable statistical guarantees under the *anchor word assumption*. They propose various techniques based, for example, on analyzing co-occurrence matrices, tensors, or on recovering vertices of a simplex using SVD. Most of these papers, except for [Ke and Wang, 2017, Bing et al., 2020b], do not work under the same statistical model as ours (cf. Section 1). Thus, [Arora et al., 2012, Arora et al., 2013, Ding et al., 2013] assume that topic-document matrix $\boldsymbol{W}$ is randomly generated from some prior distribution. For a setting with no randomness, [Mizutani, 2014] proposes ellipsoidal rounding algorithm with application to topic models. Paper [Mao et al., 2018] develops a generalized method to bind overlapping clustering models, including topic models. Moreover, for some classes of matrices, minimax optimal algorithms for estimating $\boldsymbol{A}$ have been introduced in [Ke and Wang, 2017, Bing et al., 2020b]. Both [Ke and Wang, 2017, Bing et al., 2020b] impose the *anchor word assumption* but their estimators are different. Thus, [Ke and Wang, 2017] performs SVD on properly normalized matrix $\boldsymbol{X}$ followed by an exhaustive search over a p-dimensional simplex, while [Bing et al., 2020b] proceeds by first recovering the anchor words and then deriving estimators of $\boldsymbol{A}$ from a scaled version of matrix $\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}$. A work parallel to ours [Bing et al., 2021b] considers the problem estimating a single row of matrix $\boldsymbol{W}$ under sparsity assumptions based on a preliminary estimator of matrix $\boldsymbol{A}$. Specifically, [Bing et al., 2021b] establishes the rate $\max\{\sqrt{p/nN}, 1/\sqrt{N}\}$ (to within logarithmic factors) for estimating a single row of $\boldsymbol{W}$ in the ℓ_1-norm, which implies the ℓ_1-error of the order $\max\{\sqrt{np/N}, n/\sqrt{N}\}$ for estimation of the whole matrix $\boldsymbol{W}$. On the other hand, we obtain a better rate $n/\sqrt{N}$. The additional term $\sqrt{np/N}$ corresponds to the accuracy of estimating $\boldsymbol{A}$, and seems to be inevitable for methods of recovering $\boldsymbol{W}$ based on a preliminary estimator of $\boldsymbol{A}$.

Additionally, we mention the series of works [Lee et al., 2015, Lee et al., 2019, Lee et al., 2020], which look on the estimation in topic models from a different perspective and provide algorithms having strong practical performance. In particular, [Lee et al., 2020] considers influence of the prior on the estimation and [Lee et al., 2015] explores the use of word-word co-occurrence matrix for extraction of topics. The latter idea, potentially combined with our approach, could lead to new algorithms for estimation in topic models.

5 Numerical experiments

5.1 Synthetic Data

We first present the results of experiments on synthetic data. We have performed simulations with different values of the parameters n, p, N and the number of topics

K. Our aim was to observe the effect of each of these parameters on the Frobenius error between $\boldsymbol{W}$ and its estimator $\hat{\boldsymbol{W}}$ obtained by the SPOC algorithm. We report the results for the SPOC algorithm without the preconditioning step as it had a negligible impact on the performance of the method while being computationally demanding. As a benchmark, we use the LDA algorithm [Blei et al., 2003]. For the experiments we use the Python implementation of SPOC[1] and an implementation of the LDA algorithm available in Sklearn [Pedregosa et al., 2011].

Figures 4.3-4.4 present an example of results that we have typically obtained in simulations. We take $K = 3$. We set K rows of $\boldsymbol{W}$ as the canonical basis vectors and each of the remaining $N - K$ rows is generated independently using the Dirichlet distribution with parameter $\alpha = (0.1, 0.15, 0.2)$. In Figure 4.4, where K must vary, we define $\boldsymbol{W}$ in a different way. Namely, for the $N - K$ rows that are not canonical basis vectors, each element W_{kj} is generated from the uniform distribution on $[0, 1]$ and then each row of the matrix is normalized so as to have $\sum_{k=1}^{K} W_{ik} = 1$. For the matrix $\boldsymbol{A}$, we take K columns proportional to canonical basis vectors with coefficients equal to random variables $U_k, k = 1, \dots, K$ uniformly distributed on $[0, 1]$. The elements A_{kj} of matrix $\boldsymbol{A}$ in the remaining $p - K$ columns are obtained by generating numbers from the uniform distribution on $[0, 1]$ and then normalizing each row of the matrix to have $\sum_{j=K+1}^{p} A_{kj} = 1 - U_k, k = 1, \dots, K$. The resulting matrix $\boldsymbol{A}$ has normalized rows, i.e. $\sum_{j=1}^{p} A_{kj} = 1$ (we also performed experiments in the case when Assumption 3 is violated, see below). For given $\boldsymbol{W}$ and $\boldsymbol{A}$, the data matrix $\boldsymbol{X}$ is generated according to the pLSI model defined in Section 1. For each value on the x-axes of the figures, we present the averaged result over 10 simulations.

We clearly retrieve the patterns indicated in Theorem 1, Corollary 1 and (4.13). Thus, the plots have a near $\sqrt{n}$ and a near $1/\sqrt{N}$ behaviour in Figures 4.1 and 4.3. Figures 4.2 and 4.4 show weak dependence of the error of the SPOC algorithm on the size of the dictionary p, which agrees with the bound obtained in Corollary 1. Note that choosing $p = 5000$ in three plots of Figures 4.1 and 4.2 corresponds to a rare case where LDA slightly outperforms SPOC (seen the left plot of Figure 4.2). Additional experiments for $p = 2000$ are presented in Section 12 and show a clear advantage of SPOC over LDA.

We also studied the influence of Assumption 3 on the performance of the algorithms. In Figures 4.3 and 4.4 we generate data in the same way as above except for the definition of matrix $\boldsymbol{A}$. Namely, we take K columns of $\boldsymbol{A}$ proportional to the canonical basis vectors, while the elements in the remaining $p - K$ columns are obtained by generating numbers from the uniform distribution on $[0, 1]$. Then we normalize each row of the matrix to have $\sum_{j=1}^{p} A_{kj} = 1$. With this definition, Assumption 3 is violated, which presents an unfavorable case according to our theory. Nevertheless, we still observe the dependence on the key parameters outlined in Theorem 1, Corollary 1 and (4.13). Moreover, we observe that SPOC still performs better than LDA. Generally, in all the experiments we observe that the SPOC algorithm is very competitive with LDA while being much more stable.

[1]The code of SPOC algorithm is available at https://github.com/stat-ml/SPOC

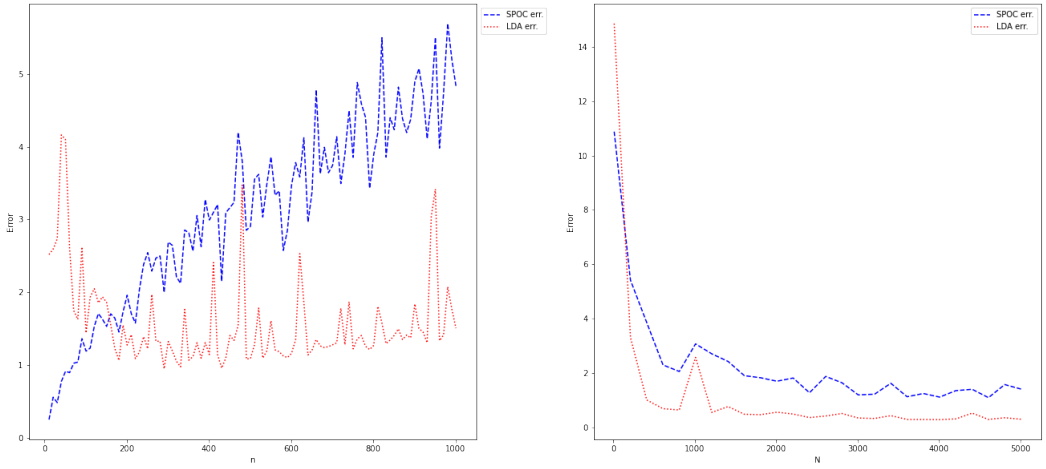


Figure 4.1: On the left (respectively, on the right), the n-dependence (respectively, the N-dependence) of $\min_{\mathbf{P}\in\mathscr{P}} \|\boldsymbol{W} - \hat{\boldsymbol{W}}\mathbf{P}\|_F$ using SPOC and LDA algorithms when the total number of words is $p = 5000$. The number of sampled words is $N = 200$ on the left, the number of documents on the right is $n = 1000$. Matrix $\boldsymbol{A}$ is generated in a way that Assumption 3 is satisfied.

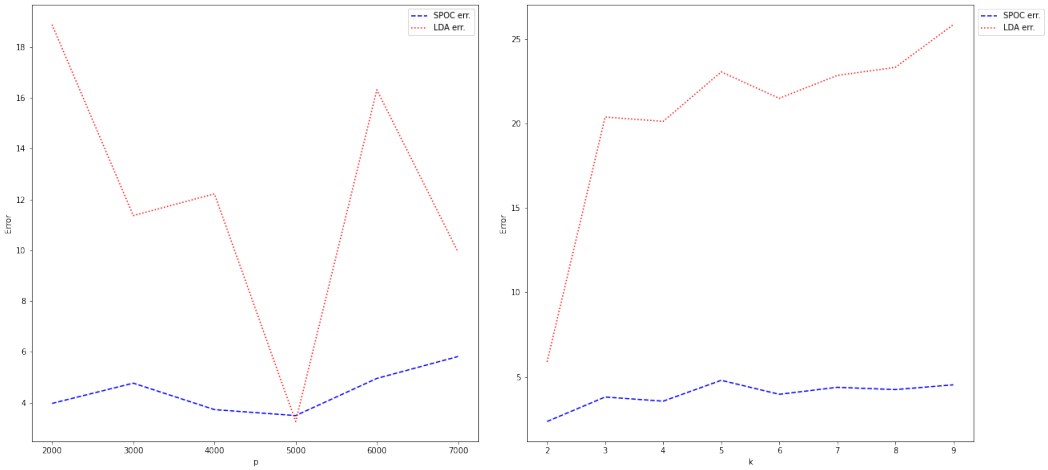


Figure 4.2: On the left (respectively, on the right), the p-dependence (respectively, the k-dependence) of $\min_{\mathbf{P}\in\mathscr{P}} \|\boldsymbol{W} - \hat{\boldsymbol{W}}\mathbf{P}\|_F$ using SPOC and LDA algorithms. Number of documents $n = 1000$ on the left, number of sampled words $N = 200$ on the left and $N = 5000$ on the right, total number of words $p = 5000$ on the right. Matrix $\boldsymbol{A}$ is generated in a way that Assumption 3 is satisfied.

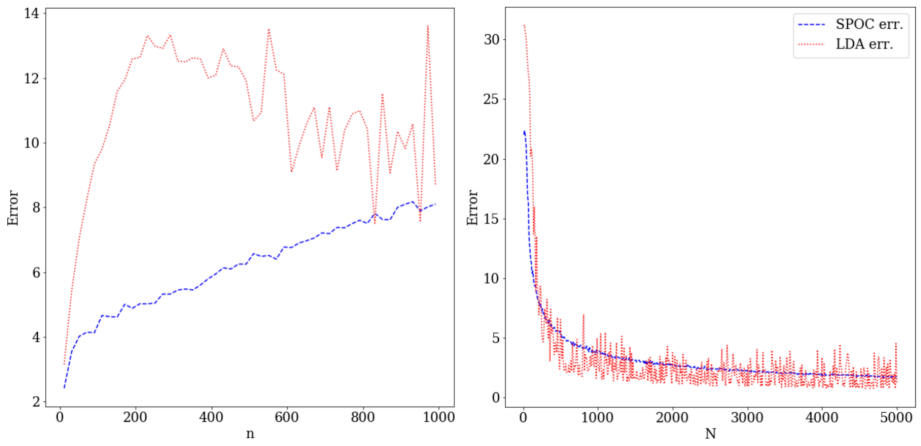


Figure 4.3: On the left (respectively, on the right), the n-dependence (respectively, the N-dependence) of $\min_{\mathbf{P}\in\mathscr{P}} \|\boldsymbol{W} - \hat{\boldsymbol{W}}\mathbf{P}\|_F$ using SPOC and LDA algorithms when the total number of words is $p = 5000$. The number of sampled words is $N = 200$ on the left, the number of documents on the right is $n = 1000$. Matrix $\boldsymbol{A}$ is generated in a way that Assumption 3 is not satisfied.

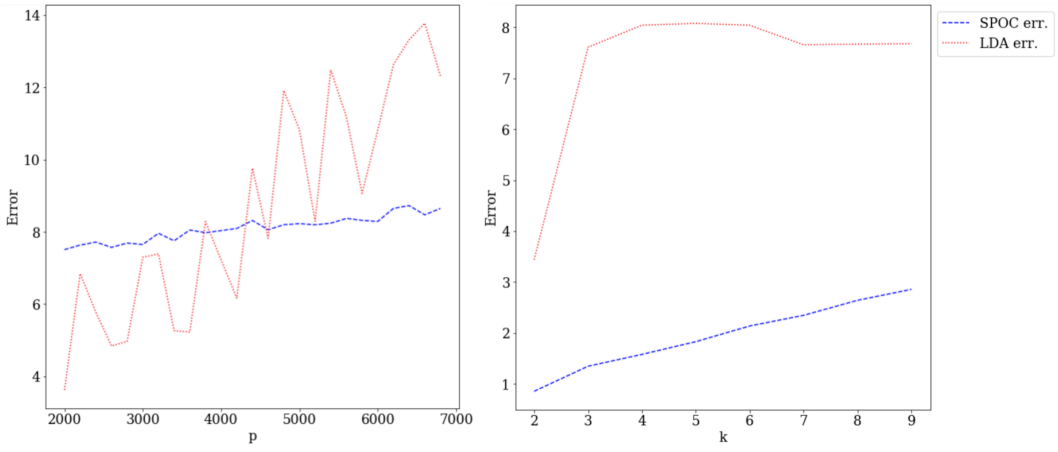


Figure 4.4: On the left (respectively, on the right), the p-dependence (respectively, the k-dependence) of $\min_{\mathbf{P}\in\mathscr{P}} \|\boldsymbol{W} - \hat{\boldsymbol{W}}\mathbf{P}\|_F$ using SPOC and LDA algorithms. Number of documents $n = 1000$ on the left, number of sampled words $N = 200$ on the left and $N = 5000$ on the right, total number of words $p = 5000$ on the right. Matrix $\boldsymbol{A}$ is generated in a way that Assumption 3 is not satisfied.

A numerical study of the SPOC estimator $\hat{\boldsymbol{A}}$ of matrix $\boldsymbol{A}$ is deferred to Section 10. It shows that $\hat{\boldsymbol{A}}$ behaves similarly to the corresponding LDA estimator while being more stable.

5.2 Corpus of NIPS abstracts

We now illustrate the performance of our algorithm applying it to the data set of full texts of NIPS papers[2] [Perrone et al., 2017]. This data set contains the distribution of words in the full text of the NIPS conference papers published from 1987 to 2015. It has the form of a 11463×5811 matrix of word counts containing 11463 words and 5811 NIPS conference papers. Each column contains the number of times each word appears in the corresponding document.

We start by pre-processing the data. We first remove all the documents with less than 150 words. Then we remove from the resulting dictionary the stop words and the words that appear in less that 150 documents. This results in a database of 5801 documents with a dictionary of 6380 words. In order to compare our method to LDA, we proceed as follows. For each value of $K = 3, \dots, 10$, we first compute the LDA estimator $\tilde{\boldsymbol{W}}$ of the document-topic matrix and the LDA estimator $\tilde{\boldsymbol{A}}$ of the topic-word matrix. Next, with the underlying matrix $\tilde{\boldsymbol{\Pi}} = \tilde{\boldsymbol{W}}\tilde{\boldsymbol{A}}$, for each value of K we simulate 10 matrices $\tilde{\boldsymbol{X}}$ with $N = 200$ sampled words according to pLSI model. For each matrix $\tilde{\boldsymbol{X}}$, we estimate $\tilde{\boldsymbol{W}}$ using both LDA and SPOC algorithms. Finally, for each K we compute the mean error over 10 simulations. The resulting comparison as function of K is presented in Figure 4.5. We can observe that SPOC systematically outperforms the LDA algorithm, except for $K = 2$.

Next, we apply the SPOC estimator of matrix $\boldsymbol{W}$ to the problem classifying the documents in the NIPS corpus. We apply the simplest possible classifier: we assign to the article i the topic that has the maximum value of $\hat{W}_{ik}$ for $k = 1, \dots, K$. We consider the number of topics $K = 3$. While some words (such as "learning" or "model") are very frequent in the whole corpus, other words are more frequent for particular topics. Therefore, for each topic, we choose the words that have the highest difference between their frequency for this topic and their maximum frequency for other topics. We clearly see that the obtained topics are semantically well separated, see Table 4.2.

Additionally, we can note that the anchor documents extracted by the SPOC algorithm are quite adequate. For the topic "Algorithms and Theory", the selected paper is [Beygelzimer et al., 2015]. This is clearly a theoretical paper, it does not mention neural networks at all and also does not speak much about statistical learning or modeling. The paper [Park et al., 2013] was selected as an anchor document for the "Statistical Learning" topic. It is a purely modeling paper with again no mentioning of neural networks and no theoretical results. Finally, the paper [Liu et al., 1994] was chosen as an anchor document for the "Neural Networks" topic.

[2]The link to the dataset: https://archive.ics.uci.edu/ml/datasets/NIPS+Conference+Papers+1987-2015

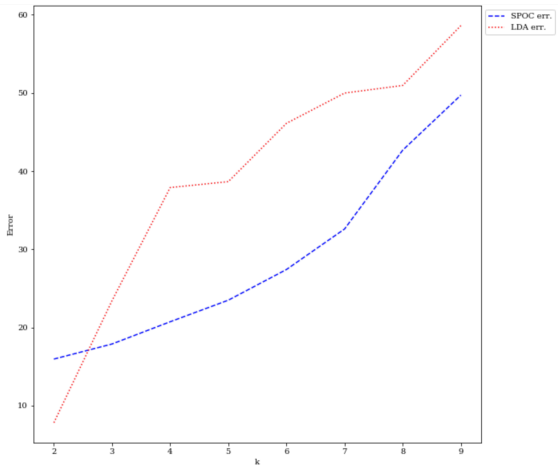


Figure 4.5: The K-dependence of $\min_{\mathbf{P}\in\mathscr{P}} \|\boldsymbol{W} - \hat{\boldsymbol{W}}\mathbf{P}\|_F$ using SPOC and LDA algorithms on semi-synthetic data. Matrix $\tilde{\boldsymbol{W}}$ is the LDA estimator on the NIPS data set ($n = 5081$ documents, $p = 6380$ words), and $\hat{\boldsymbol{W}}$ is the LDA or SPOC estimator on data simulated from $\tilde{\boldsymbol{\Pi}}$.

Table 4.2: Top 10 words, which have the highest difference in frequency for each topic compared to other topics. The three topics were identified by SPOC method.

	"Neural Networks"	"Statistical Learning"	"Algorithms and Theory"
1	network	model	algorithm
2	input	data	learning
3	neural	image	function
4	neurons	distribution	problem
5	units	inference	set
6	output	likelihood	theorem
7	layer	latent	bound
8	neuron	prior	matrix
9	system	Gaussian	loss
10	synaptic	parameters	error

This paper deals with neural networks and control. In the context of the extracted three topics (see Table 4.2), it is heavily a neural networks paper as it does not develop any theory and also does not talk about statistical modeling.

6 Conclusion

In the present paper we proposed a new algorithm for estimating the document-topic matrix in the topic model, the SPOC algorithm. Our algorithm is computationally efficient even in the case of an unknown number of topics. It is based on the Successive Projection Algorithm used to recover the vertices of a K-dimensional simplex in

the context of separable matrix factorization. We developed the statistical analysis of the SPOC algorithm under the *anchor document assumption* requiring that, for each topic, there is a document devoted solely to this topic. We proved that the proposed method is near minimax optimal under the Frobenius norm and the ℓ_1-norm. As an element of our analysis, we derived a bound on the concentration of matrices with independent multinomial columns that may be of independent interest. The theoretical results are supported by empirical evidence demonstrating a good performance of the SPOC algorithm and its advantages compared to the LDA.

7 Tools

7.1 Matrix Perturbation Bounds

In this section, we provide some facts about matrix perturbation that will be used in the proofs. We start with the following lemma, which is a variant of Davis-Kahan theorem.

Proposition 1 (Lemma 5.1 [Lei and Rinaldo, 2015]). *Let $\mathbf{M} \in \mathbb{R}^{n\times n}$ be a rank K symmetric matrix with smallest nonzero eigenvalue $\lambda_K(\mathbf{M})$, and let $\hat{\mathbf{M}} \in \mathbb{R}^{n\times n}$ be any symmetric matrix. Let $\hat{\mathbf{U}}(\hat{\mathbf{M}}) \in \mathbb{R}^{n\times K}$ and $\mathbf{U}(\mathbf{M}) \in \mathbb{R}^{n\times K}$ be the matrices of K leading eigenvectors of $\hat{\mathbf{M}}$ and $\mathbf{M}$, respectively. Then there exists a $K\times K$ orthogonal matrix $\mathbf{O}$ such that*

$$\|\hat{\mathbf{U}}(\hat{\mathbf{M}}) - \mathbf{U}(\mathbf{M})\mathbf{O}\|_F \le \frac{2\sqrt{2K}\|\hat{\mathbf{M}} - \mathbf{M}\|}{\lambda_K(\mathbf{M})}.$$

Corollary 4. *Let $\mathbf{\Pi}$ and $\boldsymbol{X}$ be matrices with singular value decompositions given by (4.4) and (4.6). Then there exist $K\times K$ orthogonal matrices $\mathbf{O}$ and $\widetilde{\mathbf{O}}$ such that*

$$\|\hat{\mathbf{U}} - \mathbf{U}\mathbf{O}\|_F \le \frac{2\sqrt{2K}(\|\boldsymbol{X}\| + \|\mathbf{\Pi}\|)\|\boldsymbol{X} - \mathbf{\Pi}\|}{\lambda_K^2(\mathbf{\Pi})} \tag{4.15}$$

and

$$\|\hat{\mathbf{V}} - \mathbf{V}\widetilde{\mathbf{O}}\|_F \le \frac{2\sqrt{2K}(\|\boldsymbol{X}\| + \|\mathbf{\Pi}\|)\|\boldsymbol{X} - \mathbf{\Pi}\|}{\lambda_K^2(\mathbf{\Pi})}. \tag{4.16}$$

Furthermore, if $\|\boldsymbol{X} - \mathbf{\Pi}\| \le \frac{1}{2}\lambda_K(\mathbf{\Pi})$ then

$$\max\left(\|\hat{\mathbf{U}} - \mathbf{U}\mathbf{O}\|_F, \|\hat{\mathbf{V}} - \mathbf{V}\widetilde{\mathbf{O}}\|_F\right) \le \frac{5\sqrt{2K}\kappa(\mathbf{\Pi})\|\boldsymbol{X} - \mathbf{\Pi}\|}{\lambda_K(\mathbf{\Pi})}. \tag{4.17}$$

Proof. Applying Proposition 1 to matrices $\mathbf{\Pi}\mathbf{\Pi}^{\mathrm{T}}$ and $\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}$ we get

$$\|\hat{\mathbf{U}} - \mathbf{U}\mathbf{O}\|_F \le \frac{2\sqrt{2K}\|\mathbf{\Pi}\mathbf{\Pi}^{\mathrm{T}} - \boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\|}{\lambda_K(\mathbf{\Pi}\mathbf{\Pi}^{\mathrm{T}})} \le \frac{2\sqrt{2K}(\|\boldsymbol{X}\| + \|\mathbf{\Pi}\|)\|\boldsymbol{X} - \mathbf{\Pi}\|}{\lambda_K^2(\mathbf{\Pi})}.$$

Similarly, inequality (4.16) is obtained by applying Proposition 1 to matrices $\mathbf{\Pi}^{\mathrm{T}}\mathbf{\Pi}$ and $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$. Next, if $\|\boldsymbol{X}-\mathbf{\Pi}\| \le \frac{1}{2}\lambda_K(\mathbf{\Pi})$ then due to the triangle inequality we have $\|\boldsymbol{X}\| \le \|\mathbf{\Pi}\| + \frac{1}{2}\lambda_K(\mathbf{\Pi}) \le \frac{3}{2}\|\mathbf{\Pi}\|$. Combining this fact with (4.15) and (4.16) we obtain (4.17). □

We will also need the following bounds for matrices of singular values $\hat{\mathbf{L}}$ and $\mathbf{L}$:

Lemma 3. *Let the assumptions of Corollary 4 hold. Let $\hat{\mathbf{L}}$ and $\mathbf{L}$ be diagonal $K\times K$-matrices of K largest singular values of $\boldsymbol{X}$ and $\mathbf{\Pi}$, respectively, cf. (4.4) and (4.6). If $\|\boldsymbol{X}-\mathbf{\Pi}\| \le \frac{1}{2}\lambda_K(\mathbf{\Pi})$ then*

$$\|\hat{\mathbf{L}} - \mathbf{O}^{\mathrm{T}}\mathbf{L}\widetilde{\mathbf{O}}\| \le C\kappa^2(\mathbf{\Pi})\sqrt{K}\ \|\boldsymbol{X}-\mathbf{\Pi}\|$$

and

$$\|\hat{\mathbf{L}}^{-1} - \widetilde{\mathbf{O}}^{\mathrm{T}}\mathbf{L}^{-1}\mathbf{O}\| \le C\kappa^2(\mathbf{\Pi})\sqrt{K}\frac{\|\boldsymbol{X}-\mathbf{\Pi}\|}{\lambda_K^2(\mathbf{\Pi})},$$

where the orthogonal matrices $\mathbf{O}$, $\widetilde{\mathbf{O}}$ are the same as in Corollary 4.

Proof. Applying Weyl's inequality [Giraud, 2015, Theorem C.6] we get

$$\|\hat{\mathbf{U}}\hat{\mathbf{L}}\hat{\mathbf{V}}^{\mathrm{T}} - \mathbf{U}\mathbf{L}\mathbf{V}^{\mathrm{T}}\| \le 2\|\mathbf{\Pi}-\boldsymbol{X}\|$$

and further

$$\begin{aligned}\|\hat{\mathbf{U}}\hat{\mathbf{L}}\hat{\mathbf{V}}^{\mathrm{T}} - \mathbf{U}\mathbf{L}\mathbf{V}^{\mathrm{T}}\| \ge{}& \|\mathbf{U}\mathbf{O}(\hat{\mathbf{L}} - \mathbf{O}^{\mathrm{T}}\mathbf{L}\widetilde{\mathbf{O}})\hat{\mathbf{V}}^{\mathrm{T}}\| \\ &- \|(\hat{\mathbf{U}}-\mathbf{U}\mathbf{O})\hat{\mathbf{L}}\hat{\mathbf{V}}^{\mathrm{T}}\| - \|\mathbf{U}\mathbf{L}\widetilde{\mathbf{O}}(\hat{\mathbf{V}}-\mathbf{V}\widetilde{\mathbf{O}})^{\mathrm{T}}\|.\end{aligned}$$

Therefore

$$\begin{aligned}\|\hat{\mathbf{L}} - \mathbf{O}^{\mathrm{T}}\mathbf{L}\widetilde{\mathbf{O}}\| &\le \|\boldsymbol{X}-\mathbf{\Pi}\| + \|(\hat{\mathbf{U}}-\mathbf{U}\mathbf{O})\hat{\mathbf{L}}\hat{\mathbf{V}}^{\mathrm{T}}\| + \|\mathbf{U}\mathbf{L}\widetilde{\mathbf{O}}(\hat{\mathbf{V}}-\mathbf{V}\widetilde{\mathbf{O}})^{\mathrm{T}}\| \\ &\le \|\boldsymbol{X}-\mathbf{\Pi}\| + \|\boldsymbol{X}\|\|\hat{\mathbf{U}}-\mathbf{U}\mathbf{O}\| + \|\mathbf{\Pi}\|\|\hat{\mathbf{V}}-\mathbf{V}\widetilde{\mathbf{O}}\| \\ &\le \left(\frac{2\sqrt{2K}(\|\boldsymbol{X}\|+\|\mathbf{\Pi}\|)^2}{\lambda_K^2(\mathbf{\Pi})} + 1\right)\|\boldsymbol{X}-\mathbf{\Pi}\|,\end{aligned}$$

where the last inequality is due to Corollary 4. Next,

$$\begin{aligned}\|\hat{\mathbf{L}}^{-1} - \widetilde{\mathbf{O}}^{\mathrm{T}}\mathbf{L}^{-1}\mathbf{O}\| &= \|\hat{\mathbf{L}}^{-1}(\mathbf{O}^{\mathrm{T}}\mathbf{L}\widetilde{\mathbf{O}} - \hat{\mathbf{L}})\widetilde{\mathbf{O}}^{\mathrm{T}}\mathbf{L}^{-1}\mathbf{O}\| \\ &\le \|\hat{\mathbf{L}}^{-1}\|\ \|\hat{\mathbf{L}} - \mathbf{O}^{\mathrm{T}}\mathbf{L}\widetilde{\mathbf{O}}\|\ \|\mathbf{L}^{-1}\| \\ &\le \left(\frac{2\sqrt{2K}(\|\boldsymbol{X}\|+\|\mathbf{\Pi}\|)^2}{\lambda_K^2(\mathbf{\Pi})} + 1\right)\frac{\|\boldsymbol{X}-\mathbf{\Pi}\|}{\lambda_K(\boldsymbol{X})\lambda_K(\mathbf{\Pi})},\end{aligned}$$

where $\lambda_K(\boldsymbol{X})$ is the K-th largest singular value of matrix $\boldsymbol{X}$. Due to Weyl's inequality and the fact that $\|\boldsymbol{X}-\mathbf{\Pi}\| \le \lambda_K(\mathbf{\Pi})/2$ we have $\lambda_K(\boldsymbol{X}) \ge \lambda_K(\mathbf{\Pi})/2$ and

$\|\boldsymbol{X}\| \leq \|\boldsymbol{\Pi}\| + (\lambda_K(\boldsymbol{\Pi})/2) \leq 3\|\boldsymbol{\Pi}\|/2$. Plugging these inequalities in the last two displays we obtain the lemma. □

7.2 Noisy Separable Matrix Factorization

In this section, we give a bound on the error of preconditioned SPA in Noisy Separable Matrix Factorization model. Assume that we observe

$$\tilde{\mathbf{G}} = \mathbf{G} + \mathbf{N} = \boldsymbol{W}\mathbf{Q} + \mathbf{N},$$

where $\mathbf{N} \in \mathbb{R}^{n\times K}$ is a perturbation (noise) matrix, and

$$\mathbf{G} = \boldsymbol{W}\mathbf{Q},$$

where $\boldsymbol{W} \in \mathbb{R}_+^{n\times K}$ and $\mathbf{Q} \in \mathbb{R}^{K\times K}$. If we assume that $\boldsymbol{W}$ satisfies Assumption 1 then we obtain the setting usually referred to as Noisy Separable Matrix Factorization (NSMF). The following theorem holds for preconditioned SPA in the NSMF model, see [Gillis and Vavasis, 2015, Mizutani, 2016]:

Theorem 4. *Let $K \geq 2$ and let Assumption 1 hold. Assume that matrix $\mathbf{Q}$ is non-degenerate and the entries W_{im} of matrix $\boldsymbol{W}$ satisfy the condition $\sum_{m=1}^{K} W_{im} \leq 1$ for $i = 1, \ldots, n$. Moreover, assume that for any $i = 1, \ldots, n$, the norms of the rows of matrix $\mathbf{N}$ satisfy $\|\mathbf{e}_i^{\mathrm{T}}\mathbf{N}\|_2 \leq \epsilon$ with*

$$\epsilon \leq C_* \frac{\lambda_{\min}(\mathbf{Q})}{K\sqrt{K}}$$

for some constant $C_* > 0$ small enough. Let J be the set of indices returned by the preconditioned SPA with input $(\tilde{\mathbf{G}}, K)$. Then, there exist a constant $C_0 > 0$ and a permutation π such that, for all $j \in J$,

$$\|\tilde{\mathbf{g}}_j - \mathbf{q}_{\pi(j)}\|_2 \leq C_0 \kappa(\mathbf{Q})\epsilon,$$

where $\tilde{\mathbf{g}}_k$ and $\mathbf{q}_k$ are the k-th rows of matrices $\tilde{\mathbf{G}}$ and $\mathbf{Q}$, respectively.

Note that this error bound depends on the upper bound on the individual errors $\|\mathbf{e}_i^{\mathrm{T}}\mathbf{N}\|_2$. From the statistical point of view, one might expect that there should be an algorithm, which improves upon this error bound if there are many nearly "pure" rows in matrix $\mathbf{G}$, so that the value of the error is diminished by averaging. However, to the best of our knowledge, no such algorithm complemented with a performance analysis can be found in the literature.

We now consider a specific instance of NSMF model given by (4.7). In this case, $\tilde{\mathbf{G}} = \hat{\mathbf{U}}$ and $\mathbf{Q} = \mathbf{HO}$ for an orthogonal matrix $\mathbf{O}$. Specifically, $\mathbf{O}$ is the orthogonal matrix, for which (4.15) holds (it is the same matrix $\mathbf{O}$, for which the bound of Lemma 1 is valid). Combining Theorem 4 with Lemma 1 we get the following corollary.

Corollary 5. *Let Assumptions 1 and 2 be satisfied with constant $\bar{C} \le C_*$. Consider the matrices $\mathbf{\Pi}$, $\boldsymbol{X}$, $\mathbf{H}$, $\hat{\mathbf{U}}$ as in (4.4) – (4.6) such that $\lambda_K(\mathbf{\Pi}) > 0$ and $\|\boldsymbol{X} - \mathbf{\Pi}\| \le \lambda_K(\mathbf{\Pi})/2$. Let $\mathbf{O}$ be the orthogonal matrix, for which (4.15) holds. Let J be the set of indices returned by the preconditioned SPA with input $(\hat{\mathbf{U}}, K)$, and let $\hat{\mathbf{H}} = \hat{\mathbf{U}}_J$. Then, there exist a constant $C_0 > 0$ and a permutation π such that, for all $j = 1, \dots, K$,*

$$\left\|\hat{\mathbf{h}}_j - \mathbf{q}_{\pi(j)}\right\|_2 \le C_0 \kappa(\mathbf{H}) \beta(\boldsymbol{X}, \mathbf{\Pi}), \tag{4.18}$$

where $\hat{\mathbf{h}}_k$ and $\mathbf{q}_k$ are the k-th rows of matrices $\hat{\mathbf{H}}$ and $\mathbf{HO}$, respectively. Furthermore,

$$\left\|\hat{\mathbf{H}} - \tilde{\mathbf{P}}\mathbf{HO}\right\|_F \le C_0 K^{1/2} \kappa(\boldsymbol{W}) \beta(\boldsymbol{X}, \mathbf{\Pi}), \tag{4.19}$$

where $\tilde{\mathbf{P}}$ is a permutation matrix corresponding to the permutation π.

Proof. Taking into account equations (4.4) – (4.7) we apply Theorem 4 with $\mathbf{Q} = \mathbf{HO}$, $\mathbf{N} = \hat{\mathbf{U}} - \mathbf{UO}$. By Lemma 1,

$$\|\mathbf{e}_i^{\mathrm{T}} \mathbf{N}\|_2 = \|\mathbf{e}_i^{\mathrm{T}} (\hat{\mathbf{U}} - \mathbf{UO})\|_2 \le \epsilon, \quad i = 1, \dots, n,$$

where $\epsilon = \beta(\boldsymbol{X}, \mathbf{\Pi})$. Therefore, using Assumption 3, (4.39), and the fact that $\bar{C} \le C^*$ we have

$$\epsilon \le \frac{\bar{C}}{\lambda_1(\boldsymbol{W}) K \sqrt{K}} = \frac{\bar{C} \lambda_{\min}(\mathbf{H})}{K \sqrt{K}} \le \frac{C_* \lambda_{\min}(\mathbf{HO})}{K \sqrt{K}}.$$

Thus, the assumptions of Theorem 4 are satisfied and we deduce from Theorem 4 that

$$\left\|\hat{\mathbf{h}}_j - \mathbf{q}_{\pi(j)}\right\|_2 \le C_0 \kappa(\mathbf{HO}) \epsilon = C_0 \kappa(\mathbf{H}) \beta(\boldsymbol{X}, \mathbf{\Pi}),$$

where $\hat{\mathbf{h}}_k$ and $\mathbf{q}_k$ are the k-th rows of matrices $\hat{\mathbf{H}}$ and $\mathbf{HO}$, respectively. Thus, (4.18) follows. Inequality (4.19) is an immediate consequence of (4.18) and of the equality $\kappa(\mathbf{H}) = \kappa(\boldsymbol{W})$ (cf. (4.40)). □

7.3 Concentration Bounds for Multinomial Matrices

In this section, we provide a bound with high probability on the spectral norm of matrix $\boldsymbol{X} - \mathbf{\Pi}$. Recall that, by definition, $\boldsymbol{X}^{\mathrm{T}} = [X_1, \dots, X_n]$ is such that $NX_i \in \mathbb{R}^p$ are independent random vectors distributed according to p-dimensional multinomial distribution with parameters (N, Π_i). We will use matrix Bernstein inequality (cf. Theorem 6.1.1 [Tropp, 2015]):

Proposition 2 (Matrix Bernstein inequality)**.** *Let $\boldsymbol{Z}_1, \dots, \boldsymbol{Z}_N$ be independent zero-mean $n \times p$ random matrices such that $\|\boldsymbol{Z}_m\| \le L$ for $m = 1, \dots, N$. Then, for all $t > 0$ we have*

$$\mathbb{P}\left(\left\|\frac{1}{N}\sum_{m=1}^{N} \boldsymbol{Z}_m\right\| \ge t\right) \le (n+p) \exp\left(-\frac{t^2 N^2}{2(\sigma^2 + LtN/3)}\right),$$

where

$$\sigma^2 = \max\left\{\left\|\sum_{m=1}^{N}\mathbb{E}\left(\boldsymbol{Z}_m\boldsymbol{Z}_m^{\mathrm{T}}\right)\right\|, \left\|\sum_{m=1}^{N}\mathbb{E}\left(\boldsymbol{Z}_m^{\mathrm{T}}\boldsymbol{Z}_m\right)\right\|\right\}.$$

Applying Proposition 2 to our setting we obtain the following result.

Proposition 3. *Let $\boldsymbol{X}^{\mathrm{T}} = [X_1, \dots, X_n]$ be such that $NX_i \in \mathbb{R}^p$ are independent random vectors distributed according to p-dimensional multinomial distribution with parameters (N, Π_i). Then, for all $t > 0$ we have*

$$\mathbb{P}\left(\|\boldsymbol{X} - \boldsymbol{\Pi}\| \geq t\right) \leq (n+p)\exp\left(-\frac{t^2 N}{2\sqrt{2}(n + t\sqrt{n}/3)}\right). \tag{4.20}$$

Proof. We prove (4.20) for $\boldsymbol{X}^{\mathrm{T}} - \boldsymbol{\Pi}^{\mathrm{T}}$ rather than $\boldsymbol{X} - \boldsymbol{\Pi}$, which is equivalent. Matrix $\boldsymbol{Z}^{\mathrm{T}} = \boldsymbol{X}^{\mathrm{T}} - \boldsymbol{\Pi}^{\mathrm{T}}$ has the form $\boldsymbol{Z}^{\mathrm{T}} = [Z_1, \dots, Z_n]$ with independent column vectors $Z_i = \frac{1}{N}\sum_{m=1}^{N}(T_{im} - \mathbb{E}(T_{im}))$, where vectors T_{im} are distributed according to p-dimensional multinomial distribution with parameters $(1, \Pi_i)$ and independent over m for any fixed i. Here, we have used the fact that $\mathrm{Multinomial}_p(N, \Pi_i)$ is a sum of N independent $\mathrm{Multinomial}_p(1, \Pi_i)$ vectors. We also have $\boldsymbol{\Pi}^{\mathrm{T}} = [\Pi_1, \dots, \Pi_n]$. Thus, we can write

$$\boldsymbol{Z}^{\mathrm{T}} = \frac{1}{N}\sum_{m=1}^{N}(\boldsymbol{T}_m - \mathbb{E}(\boldsymbol{T}_m)) = \frac{1}{N}\sum_{m=1}^{N}\boldsymbol{Z}_m, \tag{4.21}$$

where $\boldsymbol{T}_m = [T_{1m}, \dots, T_{nm}]$ and $\boldsymbol{Z}_m = \boldsymbol{T}_m - \mathbb{E}(\boldsymbol{T}_m)$ are independent zero-mean random matrices.

We apply Proposition 2 to the sum (4.21). The first step is to evaluate $\left\|\sum_{m=1}^{N}\mathbb{E}\left(\boldsymbol{Z}_m\boldsymbol{Z}_m^{\mathrm{T}}\right)\right\|$. Let $T_{im}(k)$ denote the k-th component of T_{im}, $k = 1, \dots, p$. We have $\mathbb{E}(T_{im}(k)) = \Pi_{ik}$, $\mathrm{Var}(T_{im}(k)) = \Pi_{ik}(1 - \Pi_{ik})$, $\mathrm{Cov}(T_{im}(k), T_{im}(j)) = -\Pi_{ik}\Pi_{ij}$ for $i \neq j$. Therefore,

$$\begin{aligned}\mathbb{E}\left(\boldsymbol{Z}_m\boldsymbol{Z}_m^{\mathrm{T}}\right) &= \mathbb{E}\left(\boldsymbol{T}_m\boldsymbol{T}_m^{\mathrm{T}}\right) - \mathbb{E}\left(\boldsymbol{T}_m\right)\mathbb{E}\left(\boldsymbol{T}_m^{\mathrm{T}}\right)\\ &= \mathbb{E}\sum_{i=1}^{n}T_{im}T_{im}^{\mathrm{T}} - \boldsymbol{\Pi}\boldsymbol{\Pi}^{\mathrm{T}} = \sum_{i=1}^{n}\boldsymbol{Y}_i,\end{aligned}$$

where

$$\boldsymbol{Y}_i = \mathrm{diag}(\Pi_{i1}, \dots, \Pi_{ip}) - \Pi_i\Pi_i^{\mathrm{T}}.$$

The spectral norm of $\boldsymbol{Y}_i$ satisfies

$$\|\boldsymbol{Y}_i\|^2 \leq \|\boldsymbol{Y}_i\|_F^2 = \sum_{k=1}^{p}\Pi_{ik}^2 + \left(\sum_{k=1}^{p}\Pi_{ik}^2\right)^2 - 2\sum_{k=1}^{p}\Pi_{ik}^3 \leq 2,$$

where we have used the fact that $\sum_{k=1}^{p}\Pi_{ik}^2 \leq \sum_{k=1}^{p}\Pi_{ik} = 1$. Thus, $\left\|\mathbb{E}\left(\boldsymbol{Z}_m\boldsymbol{Z}_m^{\mathrm{T}}\right)\right\| \leq \sqrt{2}n$ and

$$\left\|\sum_{m=1}^{N}\mathbb{E}\left(\boldsymbol{Z}_m\boldsymbol{Z}_m^{\mathrm{T}}\right)\right\| \leq \sqrt{2}Nn. \tag{4.22}$$

Next, we derive an upper bound on $\left\|\sum_{m=1}^{N}\mathbb{E}\left(\boldsymbol{Z}_m^{\mathrm{T}}\boldsymbol{Z}_m\right)\right\|$. Note that $\mathbb{E}(\boldsymbol{T}_m^{\mathrm{T}}\boldsymbol{T}_m)$ is a matrix with diagonal entries $\mathbb{E}(T_{im}^{\mathrm{T}}T_{im}) = \sum_{k=1}^{p}\Pi_{ik} = 1$ while its off-diagonal entries are $\mathbb{E}(T_{im}^{\mathrm{T}}T_{jm}) = [\mathbb{E}(T_{im})]^{\mathrm{T}}\,\mathbb{E}(T_{jm}) = \Pi_i^{\mathrm{T}}\Pi_j$ due to independence between T_{im} and T_{jm} for $i \neq j$. Also, $\mathbb{E}(\boldsymbol{T}_m^{\mathrm{T}})\,\mathbb{E}(\boldsymbol{T}_m) = \boldsymbol{\Pi}^{\mathrm{T}}\boldsymbol{\Pi}$ is a matrix with entries $\Pi_i^{\mathrm{T}}\Pi_j$. Hence,

$$\mathbb{E}\left(\boldsymbol{Z}_m^{\mathrm{T}}\boldsymbol{Z}_m\right) = \mathbb{E}(\boldsymbol{T}_m^{\mathrm{T}}\boldsymbol{T}_m) - \mathbb{E}(\boldsymbol{T}_m^{\mathrm{T}})\,\mathbb{E}(\boldsymbol{T}_m) = \operatorname{diag}\Big(1 - \|\Pi_1\|_2^2, \ldots, 1 - \|\Pi_n\|_2^2\Big). \tag{4.23}$$

It follows that $\left\|\mathbb{E}\left(\boldsymbol{Z}_m^{\mathrm{T}}\boldsymbol{Z}_m\right)\right\| \leq 1$, and thus $\left\|\sum_{m=1}^{N}\mathbb{E}\left(\boldsymbol{Z}_m^{\mathrm{T}}\boldsymbol{Z}_m\right)\right\| \leq N$. Combining this inequality with (4.22) we obtain that σ^2 defined in Proposition 2 satisfies $\sigma^2 \leq \sqrt{2}Nn$.

Finally, we specify the constant L that gives an upper bound on $\|\boldsymbol{Z}_m\|$. Let $u \in S^{p-1}$ be an element of the unit sphere in $\mathbb{R}^p$. Since for any i vector T_{im} has only one component equal to 1 and all other components 0 we have $\|T_{im} - \mathbb{E}(T_{im})\|_2^2 = \|T_{im} - \Pi_i\|_2^2 \leq 2$ and thus

$$\left|u^{\mathrm{T}}(T_{im} - \mathbb{E}(T_{im}))\right| \leq \sqrt{2}.$$

It follows that

$$\begin{aligned}\|\boldsymbol{T}_m - \mathbb{E}(\boldsymbol{T}_m)\|^2 &= \sup_{u\in S^{p-1}}\left\|u^{\mathrm{T}}(\boldsymbol{T}_m - \mathbb{E}(\boldsymbol{T}_m))\right\|_2^2 \\ &= \sup_{u\in S^{p-1}}\sum_{i=1}^{n}\left|u^{\mathrm{T}}(T_{im} - \mathbb{E}(T_{im}))\right|^2 \leq 2n\end{aligned}$$

and we get $\|\boldsymbol{Z}_m\| \leq \sqrt{2n} =: L$ for any $m = 1, \ldots, n$. The desired result now follows by applying Proposition 2 with $\sigma^2 \leq \sqrt{2}Nn$ and $L = \sqrt{2n}$. □

The next lemma is a corollary of Proposition 3:

Lemma 4. *Let the assumptions of Proposition 3 be satisfied. Assume that* $N \geq \log(n+p)$ *and* $\min(n,p) \geq 2$. *Then*

$$\mathbb{P}\left(\|\boldsymbol{X} - \boldsymbol{\Pi}\| \geq 4\sqrt{\frac{n\log(n+p)}{N}}\right) \leq (n+p)^{-1}. \tag{4.24}$$

Furthermore,

$$\mathbb{P}\left(\max_{1\leq i\leq n}\|\mathbf{e}_i^{\mathrm{T}}(\boldsymbol{X} - \boldsymbol{\Pi})\|_2 \geq 5\sqrt{\frac{\log(n+p)}{N}}\right) \leq (n+p)^{-1}. \tag{4.25}$$

Proof. Inequality (4.24) follows easily from Proposition 3 by setting $t = 4\sqrt{\frac{n\log(n+p)}{N}}$ and using the assumptions $N \geq \log(n+p)$. In order to prove (4.25), we bound each probability $\mathbb{P}\left(\|\mathbf{e}_i^{\mathrm{T}}(\boldsymbol{X} - \boldsymbol{\Pi})\|_2 \geq 5\sqrt{\log(n+p)/N}\right)$ via Proposition 3 with $n = 1$

(that is, we apply Proposition 3 to $1 \times p$ matrices $\mathbf{e}_i^{\mathrm{T}}\boldsymbol{X}$, $\mathbf{e}_i^{\mathrm{T}}\boldsymbol{\Pi}$) and then use the union bound. This yields

$$\mathbb{P}\left(\max_{1\le i\le n}\|\mathbf{e}_i^{\mathrm{T}}(\boldsymbol{X}-\boldsymbol{\Pi})\|_2 \ge 5\sqrt{\frac{\log(n+p)}{N}}\right) \le n(p+1)\exp\left(-\frac{75\log(n+p)}{16\sqrt{2}}\right).$$

The right hand side of this inequality does not exceed $(n+p)^{-1}$. □

8 Proofs of the Main Results

8.1 Proof of Lemma 1

Using the fact that $\widehat{\mathbf{V}}_1^{\mathrm{T}}\widehat{\mathbf{V}}^{\mathrm{T}} = 0$ we obtain

$$\begin{aligned}
\|\mathbf{e}_i^{\mathrm{T}}(\hat{\mathbf{U}}-\mathbf{U}\mathbf{O})\|_2 =& \|\mathbf{e}_i^{\mathrm{T}}(\boldsymbol{X}\hat{\mathbf{V}}\hat{\mathbf{L}}^{-1} - \boldsymbol{\Pi}\mathbf{V}\mathbf{L}^{-1}\mathbf{O})\|_2 \\
=& \|\mathbf{e}_i^{\mathrm{T}}\boldsymbol{X}\hat{\mathbf{V}}(\hat{\mathbf{L}}^{-1} - \widetilde{\mathbf{O}}^{\mathrm{T}}\mathbf{L}^{-1}\mathbf{O}) \\
&+ \mathbf{e}_i^{\mathrm{T}}\boldsymbol{X}(\hat{\mathbf{V}} - \mathbf{V}\widetilde{\mathbf{O}})\widetilde{\mathbf{O}}^{\mathrm{T}}\mathbf{L}^{-1}\mathbf{O} + \mathbf{e}_i^{\mathrm{T}}(\boldsymbol{X}-\boldsymbol{\Pi})\mathbf{V}\mathbf{L}^{-1}\mathbf{O}\|_2 \\
\le& \|\mathbf{e}_i^{\mathrm{T}}\boldsymbol{X}\hat{\mathbf{V}}(\hat{\mathbf{L}}^{-1} - \widetilde{\mathbf{O}}^{\mathrm{T}}\mathbf{L}^{-1}\mathbf{O})\|_2 \\
&+ \|\mathbf{e}_i^{\mathrm{T}}\boldsymbol{X}(\hat{\mathbf{V}} - \mathbf{V}\widetilde{\mathbf{O}})\widetilde{\mathbf{O}}^{\mathrm{T}}\mathbf{L}^{-1}\mathbf{O}\|_2 \\
&+ \|\mathbf{e}_i^{\mathrm{T}}(\boldsymbol{X}-\boldsymbol{\Pi})\mathbf{V}\mathbf{L}^{-1}\mathbf{O}\|_2 \\
&= G_1 + G_2 + G_3.
\end{aligned}$$

We now bound the values G_1, G_2 and G_3 separately. We have

$$\begin{aligned}
G_1 &= \|\mathbf{e}_i^{\mathrm{T}}\boldsymbol{X}\hat{\mathbf{V}}(\hat{\mathbf{L}}^{-1} - \widetilde{\mathbf{O}}^{\mathrm{T}}\mathbf{L}^{-1}\mathbf{O})\|_2 \le \|\mathbf{e}_i^{\mathrm{T}}\boldsymbol{X}\|_2\,\|\hat{\mathbf{V}}\|\,\|\hat{\mathbf{L}}^{-1} - \widetilde{\mathbf{O}}^{\mathrm{T}}\mathbf{L}^{-1}\mathbf{O}\| \\
&\le CK^{1/2}\kappa^2(\boldsymbol{\Pi})\,\frac{\|\mathbf{e}_i^{\mathrm{T}}\boldsymbol{X}\|_2\,\|\boldsymbol{X}-\boldsymbol{\Pi}\|}{\lambda_K^2(\boldsymbol{\Pi})},
\end{aligned}$$

where the last inequality is due to Lemma 3. The values G_2 and G_3 can be controlled using the bounds for the norm of matrix product and Corollary 4:

$$\begin{aligned}
G_2 &= \|\mathbf{e}_i^{\mathrm{T}}\boldsymbol{X}(\hat{\mathbf{V}} - \mathbf{V}\widetilde{\mathbf{O}})\widetilde{\mathbf{O}}^{\mathrm{T}}\mathbf{L}^{-1}\mathbf{O}\|_2 \le \|\mathbf{e}_i^{\mathrm{T}}\boldsymbol{X}\|_2\,\|\hat{\mathbf{V}} - \mathbf{V}\widetilde{\mathbf{O}}\|\,\|\mathbf{L}^{-1}\| \\
&= \frac{\|\mathbf{e}_i^{\mathrm{T}}\boldsymbol{X}\|_2\,\|\hat{\mathbf{V}} - \mathbf{V}\widetilde{\mathbf{O}}\|}{\lambda_K(\boldsymbol{\Pi})} \le 5\sqrt{2}\kappa(\boldsymbol{\Pi})\frac{\|\mathbf{e}_i^{\mathrm{T}}\boldsymbol{X}\|_2\,\|\boldsymbol{X}-\boldsymbol{\Pi}\|}{\lambda_K^2(\boldsymbol{\Pi})}
\end{aligned}$$

$$G_3 = \|\mathbf{e}_i^{\mathrm{T}}(\boldsymbol{X}-\boldsymbol{\Pi})\mathbf{V}\mathbf{L}^{-1}\mathbf{O}\|_2 \le \|\mathbf{e}_i^{\mathrm{T}}(\boldsymbol{X}-\boldsymbol{\Pi})\|_2\,\|\mathbf{V}\|\,\|\mathbf{L}^{-1}\| = \frac{\|\mathbf{e}_i^{\mathrm{T}}(\boldsymbol{X}-\boldsymbol{\Pi})\|_2}{\lambda_K(\boldsymbol{\Pi})}.$$

Combining these bounds proves the lemma.

8.2 Proof of Lemma 2

We first prove that matrix $\hat{\mathbf{H}}$ is non-degenerate. In this proof, we denote by $\mathbf{O}$ the orthogonal matrix, for which (4.15) holds, and by $\tilde{\mathbf{P}}$ the permutation matrix, for which the bound of Corollary 5 holds. Using Weyl's inequality [Giraud, 2015, Theorem C.6] and Corollary 5 we obtain

$$\begin{aligned}\lambda_{\min}(\hat{\mathbf{H}}) &\geq \lambda_{\min}(\tilde{\mathbf{P}}\mathbf{H}\mathbf{O}) - \|\hat{\mathbf{H}} - \tilde{\mathbf{P}}\mathbf{H}\mathbf{O}\| \\ &\geq \lambda_{\min}(\mathbf{H}) - C_0 K^{1/2}\kappa(\boldsymbol{W})\beta(\boldsymbol{X}, \boldsymbol{\Pi}).\end{aligned}$$

Using this inequality, Assumption 2 with $\bar{C} \leq C_0^{-1}$, and equations (4.39), (4.40) we find

$$\lambda_{\min}(\hat{\mathbf{H}}) \geq \lambda_{\min}(\mathbf{H}) - \frac{1}{2\lambda_1(\boldsymbol{W})} = \frac{1}{2\lambda_1(\boldsymbol{W})}, \tag{4.26}$$

which proves that $\hat{\mathbf{H}}$ is invertible. Then, for the estimator $\hat{\boldsymbol{W}} = \hat{\mathbf{U}}\hat{\mathbf{H}}^{-1}$ and the permutation matrix $\mathbf{P} = \tilde{\mathbf{P}}^{-1}$ we have

$$\begin{aligned}\|\hat{\boldsymbol{W}} - \boldsymbol{W}\mathbf{P}\|_F &= \|\hat{\mathbf{U}}\hat{\mathbf{H}}^{-1} - \mathbf{U}\mathbf{H}^{-1}\mathbf{P}\|_F \\ &\leq \|\hat{\mathbf{U}}(\hat{\mathbf{H}}^{-1} - \mathbf{O}^{\mathrm{T}}\mathbf{H}^{-1}\mathbf{P})\|_F + \|(\hat{\mathbf{U}} - \mathbf{U}\mathbf{O})[\mathbf{P}^{-1}\mathbf{H}\mathbf{O}]^{-1}\|_F \\ &= I_1 + I_2.\end{aligned}$$

We now bound separately I_1 and I_2. Due to (4.26) and (4.39) we have

$$\|\hat{\mathbf{H}}^{-1}\| = \frac{1}{\lambda_{\min}(\hat{\mathbf{H}})} \leq 2\lambda_1(\boldsymbol{W}), \quad \|\mathbf{H}^{-1}\| = \lambda_1(\boldsymbol{W}). \tag{4.27}$$

Using the fact that $\|\mathbf{A}^{-1} - \mathbf{B}^{-1}\|_F \leq \|\mathbf{A}^{-1}\|\,\|\mathbf{B}^{-1}\|\,\|\mathbf{A} - \mathbf{B}\|_F$ with $\mathbf{A} = \hat{\mathbf{H}}$, $\mathbf{B} = \mathbf{P}^{-1}\mathbf{H}\mathbf{O} = \tilde{\mathbf{P}}\mathbf{H}\mathbf{O}$, inequality (4.27) and Corollary 5 we find

$$\begin{aligned}I_1 = \|\hat{\mathbf{U}}(\hat{\mathbf{H}}^{-1} - \mathbf{O}^{\mathrm{T}}\mathbf{H}^{-1}\mathbf{P})\|_F &\leq \|\hat{\mathbf{H}}^{-1} - \mathbf{O}^{\mathrm{T}}\mathbf{H}^{-1}\mathbf{P}\|_F \\ &\leq \|\hat{\mathbf{H}}^{-1}\|\,\|\mathbf{O}^{\mathrm{T}}\mathbf{H}^{-1}\mathbf{P}\|\,\|\hat{\mathbf{H}} - \tilde{\mathbf{P}}\mathbf{H}\mathbf{O}\|_F \\ &\leq 2C_0 K^{1/2}\lambda_1^2(\boldsymbol{W})\kappa(\boldsymbol{W})\beta(\boldsymbol{X}, \boldsymbol{\Pi}).\end{aligned}$$

On the other hand,

$$\begin{aligned}I_2 = \|(\hat{\mathbf{U}} - \mathbf{U}\mathbf{O})[\mathbf{P}^{-1}\mathbf{H}\mathbf{O}]^{-1}\|_F &\leq \|\hat{\mathbf{U}} - \mathbf{U}\mathbf{O}\|_F\,\|\mathbf{H}^{-1}\| \\ \leq \frac{5\sqrt{2K}\kappa(\boldsymbol{\Pi})\|\boldsymbol{X} - \boldsymbol{\Pi}\|}{\lambda_K(\boldsymbol{\Pi})}&\|\mathbf{H}^{-1}\|,\end{aligned}$$

where the last inequality follows from Corollary 4. Combining the above bounds we get

$$\|\hat{\boldsymbol{W}} - \boldsymbol{W}\mathbf{P}\|_F \leq CK^{1/2}\lambda_1(\boldsymbol{W})\left\{\lambda_1(\boldsymbol{W})\kappa(\boldsymbol{W})\beta(\boldsymbol{X},\boldsymbol{\Pi}) + \frac{\kappa(\boldsymbol{\Pi})\|\boldsymbol{X}-\boldsymbol{\Pi}\|}{\lambda_K(\boldsymbol{\Pi})}\right\}.$$

8.3 Proof of Theorem 1

We apply Lemma 2 combined with the concentration inequalities of Lemma 4. First, we check that Assumption 2 holds with probability at least $1 - 2(n+p)^{-1}$. For $N \geq \log(n+p)$ we get from Lemma 4 that

$$\|\boldsymbol{X} - \boldsymbol{\Pi}\| \leq 4\sqrt{\frac{n\log(n+p)}{N}}$$

with probability at least $1 - 1/(n+p)$, and

$$\max_{1\leq i\leq n} \|\mathbf{e}_i^{\mathrm{T}}(\boldsymbol{X} - \boldsymbol{\Pi})\|_2 \leq 5\sqrt{\frac{\log(n+p)}{N}}$$

with probability at least $1-1/(n+p)$. Notice also that $\max_i \|\mathbf{e}_i^{\mathrm{T}}\boldsymbol{X}\|_2 = \max_i \sqrt{\sum_{j=1}^p X_{ij}^2} \leq \max_i \sqrt{\sum_{j=1}^p X_{ij}} = 1$. Putting together the above remarks we deduce that, with probability at least $1 - 2(n+p)^{-1}$,

$$\begin{aligned}\beta(\boldsymbol{X},\boldsymbol{\Pi}) &\leq 5\left\{\kappa^2(\boldsymbol{\Pi})\sqrt{Kn} + \lambda_K(\boldsymbol{\Pi})\right\}\frac{\sqrt{\log(n+p)}}{\lambda_K^2(\boldsymbol{\Pi})\sqrt{N}} \\ &\leq 10\kappa^2(\boldsymbol{\Pi})\sqrt{Kn}\frac{\sqrt{\log(n+p)}}{\lambda_K^2(\boldsymbol{\Pi})\sqrt{N}}\end{aligned}$$

where we have used the inequality $\lambda_K(\boldsymbol{\Pi}) \leq \sqrt{n/K}$ proved in Lemma 7. Since $\lambda_K(\boldsymbol{\Pi})$ is chosen to satisfy (4.10) we get that, with probability at least $1 - 2(n+p)^{-1}$,

$$\beta(\boldsymbol{X},\boldsymbol{\Pi}) \leq \frac{\bar{C}}{\lambda_1(\boldsymbol{W})\kappa(\boldsymbol{W})K\sqrt{K}}.$$

Thus, on an event that has probability at least $1 - 2(n+p)^{-1}$, Assumption 2 is satisfied and we can apply Lemma 2. This yields that, with probability at least $1 - 2(n+p)^{-1}$,

$$\min_{\mathbf{P}\in\mathscr{P}}\left\|\hat{\boldsymbol{W}} - \boldsymbol{W}\mathbf{P}\right\|_F \leq CK^{1/2}\lambda_1(\boldsymbol{W})\left\{\lambda_1(\boldsymbol{W})\kappa(\boldsymbol{W})\beta(\boldsymbol{X},\boldsymbol{\Pi}) + \frac{\kappa(\boldsymbol{\Pi})\|\boldsymbol{X}-\boldsymbol{\Pi}\|}{\lambda_K(\boldsymbol{\Pi})}\right\}$$

$$\leq C\frac{\lambda_1(\boldsymbol{W})\sqrt{nK\log(n+p)}}{\sqrt{N}\lambda_K(\boldsymbol{\Pi})}\left\{\lambda_1(\boldsymbol{W})\kappa(\boldsymbol{W})\frac{K^{1/2}\kappa^2(\boldsymbol{\Pi})}{\lambda_K(\boldsymbol{\Pi})}+\kappa(\boldsymbol{\Pi})\right\}$$
$$\leq CK\sqrt{\frac{n\log(n+p)}{N}}\left(\frac{\lambda_1(\boldsymbol{W})}{\lambda_K(\boldsymbol{\Pi})}\right)^2\kappa(\boldsymbol{W})\kappa^2(\boldsymbol{\Pi}),$$

where we have used the inequalities $\lambda_K(\boldsymbol{\Pi}) \leq \sqrt{n/K}$ and $\lambda_1(\boldsymbol{\Pi}) \leq \sqrt{K}\lambda_1(\boldsymbol{W})$ (see Lemma 7).

8.4 Proof of Theorem 2 and Corollary 3

We will use the following lemma.

Lemma 5. *Let $\boldsymbol{\Pi} \in \mathbb{R}^{n\times p}$ be a rank K matrix with smallest non-zero singular value $\lambda_K(\boldsymbol{\Pi})$, and $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ be a matrix such that $\|\boldsymbol{X}-\boldsymbol{\Pi}\| \leq \tau$ for some $\tau > 0$. Let $\hat{K} = \max\{j : \lambda_j(\boldsymbol{X}) > \tau\}$. If $\lambda_K(\boldsymbol{\Pi}) > 2\tau$ then $\hat{K} = K$.*

Proof. By Weyl's inequality, we have $|\lambda_j(\boldsymbol{X}) - \lambda_j(\boldsymbol{\Pi})| \leq \tau$ for all j. Since $\lambda_j(\boldsymbol{\Pi}) = 0$ for $j \geq K+1$ we deduce that $\hat{K} \leq K$. On the other hand, $\hat{K} \geq K$. Indeed, condition $\lambda_K(\boldsymbol{\Pi}) > 2\tau$ implies that $\lambda_K(\boldsymbol{X}) \geq \lambda_K(\boldsymbol{\Pi}) - |\lambda_j(\boldsymbol{X}) - \lambda_j(\boldsymbol{\Pi})| > \tau$. □

Theorem 2 is obtained by combining Theorem 1 with Lemma 5. Indeed, notice that the bound of Theorem 1 is proved on the event $\mathscr{A} := \left\{\|\boldsymbol{X}-\boldsymbol{\Pi}\| \leq 4\sqrt{n\log(n+p)/N}\right\}$. Set $\tau = 4\sqrt{n\log(n+p)/N}$. It follows from Lemma 5 that if

$$\lambda_K(\boldsymbol{\Pi}) > 8\sqrt{\frac{n\log(n+p)}{N}} \tag{4.28}$$

then on the event $\mathscr{A}$ we have $\hat{K} = K$. But condition (4.28) is implied by (4.10) and (4.12). Therefore, the proof of Theorem 1 goes through verbatim if we replace K by $\hat{K}$. This yields Theorem 2. Corollary 3 is deduced from Theorem 2 in the same way as Corollary 1 was deduced from Theorem 1.

8.5 Proof of Theorem 3

We use the techniques of proving minimax lower bounds based on a reduction to the problem of testing multiple hypotheses [Tsybakov, 2008, Chapter 2]. The hypotheses correspond to probability measures $\mathbb{P}_{\boldsymbol{\Pi}^{(j)}}$, where $\boldsymbol{\Pi}^{(j)} = \boldsymbol{W}^{(j)}\boldsymbol{A}$ with carefully chosen matrix $\boldsymbol{A}$ and matrices $\boldsymbol{W}^{(j)}$, $j = 0, 1, \ldots, T$. The construction of these matrices borrows some elements from the proofs of the lower bounds in papers [Ke and Wang, 2017, Bing et al., 2020b]. An additional subtlety is related to the fact that we need to grant Assumption 3 on the singular values. Without loss of generality we assume that n is a multiple of K and that K is even.

1. *Construction of the set of matrices* $\boldsymbol{W}^{(j)}$.
We first introduce the basic matrix $\boldsymbol{W}^{(0)}$ and then define matrices $\boldsymbol{W}^{(j)}$, $j = 1, \ldots, T$ as slightly perturbed versions of $\boldsymbol{W}^{(0)}$.

Let $\boldsymbol{D}_1$ be a $n \times K$ matrix composed of n/K blocks, each of which is the identity matrix $\boldsymbol{I}_K$ of size K:

$$\boldsymbol{D}_1^{\mathrm{T}} = \left[\begin{array}{c|c|c|c} \boldsymbol{I}_K & \boldsymbol{I}_K & \text{\_\_\_\_} & \boldsymbol{I}_K \end{array} \right].$$

We have that $\boldsymbol{D}_1^{\mathrm{T}} \boldsymbol{D}_1 = (n/K)\boldsymbol{I}_K$ and $\sigma(\boldsymbol{D}_1) = \{\sqrt{n/K}, 0\}$, where $\sigma(\boldsymbol{D}_1)$ denotes the set of singular values of $\boldsymbol{D}_1$. Set

$$\gamma_1 = \frac{1}{4\,K}$$

and define the $n \times K$ matrix $\boldsymbol{D}_2$ by the relation

$$\boldsymbol{D}_2^{\mathrm{T}} = \gamma_1 \left[\begin{array}{c|c} \mathbf{0}_{K,K} & \mathbf{1}_{K,(n-K)} \end{array} \right]$$

where we denote by $\mathbf{1}_{n,p}$ (respectively, $\mathbf{0}_{n,p}$) the $n \times p$ matrix with all entries 1 (respectively, 0). Then, $\boldsymbol{D}_2^{\mathrm{T}} \boldsymbol{D}_2 = (n-K)\gamma_1^2 \mathbf{1}_{K,K}$ and $\sigma(\boldsymbol{D}_2) = \{\gamma_1\sqrt{K(n-K)}, 0\}$. We will further consider the matrix $\boldsymbol{D}_3 = \boldsymbol{D}_1 + \boldsymbol{D}_2$ given by the relation

$$\begin{aligned} \boldsymbol{D}_3^{\mathrm{T}} &= \left[\begin{array}{c|cccc|c|cccc} 1 \cdots 0 & (1+\gamma_1) & \gamma_1 & \ldots & \gamma_1 & \ldots\ldots & (1+\gamma_1) & \gamma_1 & \ldots & \gamma_1 \\ \vdots \quad \vdots & \gamma_1 & (1+\gamma_1) & \ldots & \gamma_1 & \ldots\ldots & \gamma_1 & (1+\gamma_1) & \ldots & \gamma_1 \\ \vdots \quad \vdots & \vdots & & & \vdots & \ldots\ldots & \vdots & & & \vdots \\ 0 \cdots 1 & \gamma_1 & \gamma_1 & \ldots & (1+\gamma_1) & \ldots\ldots & \gamma_1 & \gamma_1 & \ldots & (1+\gamma_1) \end{array} \right] \\ &= \left[\begin{array}{c|c|c|c} \boldsymbol{I}_K & \boldsymbol{I}_K + \gamma_1 \mathbf{1}_{K,K} & \ldots & \boldsymbol{I}_K + \gamma_1 \mathbf{1}_{K,K}. \end{array} \right]. \end{aligned}$$

Applying Weyl's inequality [Giraud, 2015, Theorem C.6], we get

$$\left\{ \begin{array}{l} \lambda_1(\boldsymbol{D}_3) \leq \sqrt{n/K} + \frac{1}{4}\sqrt{(n-K)/K} \leq \frac{5}{4}\sqrt{n/K}, \\ \lambda_K(\boldsymbol{D}_3) \geq \sqrt{n/K} - \frac{1}{4}\sqrt{(n-K)/K} \geq \frac{3}{4}\sqrt{n/K}. \end{array} \right.$$

Finally, the basic matrix $\boldsymbol{W}^{(0)}$ is defined by the relation

$$\begin{aligned} (\boldsymbol{W}^{(0)})^{\mathrm{T}} &= \boldsymbol{D}_3^{\mathrm{T}} - \left[\begin{array}{c|c|c|c} \mathbf{0}_{K,K} & K\gamma_1 \boldsymbol{I}_K & \ldots & K\gamma_1 \boldsymbol{I}_K \end{array} \right] \\ &= \left[\begin{array}{c|c|c|c} \boldsymbol{I}_K & (1-K\gamma_1)\boldsymbol{I}_K + \gamma_1 \mathbf{1}_{K,K} & \ldots & (1-K\gamma_1)\boldsymbol{I}_K + \gamma_1 \mathbf{1}_{K,K} \end{array} \right]. \end{aligned}$$

Clearly, $\boldsymbol{W}^{(0)}$ satisfies Assumption 1, all entries of $\boldsymbol{W}^{(0)}$ are non-negative and its rows sum up to 1. Applying Weyl's inequality to matrix $\boldsymbol{W}^{(0)}$ yields

$$\left\{ \begin{array}{l} \lambda_1(\boldsymbol{W}^{(0)}) \leq \frac{5}{4}\sqrt{n/K} + K\gamma_1(\sqrt{n/K - 1}) \leq \frac{3}{2}\sqrt{n/K}, \\ \lambda_K(\boldsymbol{W}^{(0)}) \geq \frac{3}{4}\sqrt{n/K} - K\gamma_1(\sqrt{n/K - 1}) \geq \frac{1}{2}\sqrt{n/K}, \end{array} \right. \tag{4.29}$$

implying that $\kappa(\boldsymbol{W}^{(0)}) \leq 3$.

Our next step is to define the matrices $\boldsymbol{W}^{(j)}$, $j = 1, \ldots, T$. Consider the set of binary sequences

$$M = \{0,1\}^{K(n-K)/2}.$$

Applying the Varshamov-Gilbert bound [Tsybakov, 2008, Lemma 2.9] we get that there exist $w^{(j)} \in M$, $j = 1, \ldots, T$, such that:

$$\|w^{(i)} - w^{(j)}\|_1 = \|w^{(i)} - w^{(j)}\|_2^2 \geq \frac{K(n-K)}{16}, \text{ for any } 0 \leq i \neq j \leq T, \tag{4.30}$$

with $w^{(0)} = 0$ and

$$\log T \geq \frac{\log 2}{16} K(n-K). \tag{4.31}$$

We divide each $w^{(j)}$ into $(n-K)$ chunks as $w^{(j)} = (w_1^{(j)}, w_2^{(j)}, \ldots, w_{n-K}^{(j)})$ with $w_i^{(j)} \in \{0,1\}^{K/2}$. Next, for each $w_i^{(j)}$, we introduce its augmented counterpart defined as $\tilde{w}_i^{(j)} = (w_i^{(j)}, -w_i^{(j)}) \in \{-1, 0, 1\}^K$. In what follows, we set

$$\gamma = c_* \sqrt{\frac{N}{K(N-K)^2}},$$

where $c_* > 0$ is a small enough absolute constant. For $1 \leq j \leq T$, define the $(n-K) \times K$ matrix $\boldsymbol{\Omega}^{(\mathbf{j})}$ and the $n \times K$ matrix $\boldsymbol{\Delta}^{(j)}$ as follows:

$$\boldsymbol{\Omega}^{(\mathbf{j})} = \gamma \begin{bmatrix} \tilde{w}_1^{(j)} \\ \tilde{w}_2^{(j)} \\ \vdots \\ \tilde{w}_{n-K}^{(j)} \end{bmatrix} \text{ and } \boldsymbol{\Delta}^{(j)} = \begin{bmatrix} \mathbf{0}_{K,K} \\ \hline \boldsymbol{\Omega}^{(\mathbf{j})} \end{bmatrix}.$$

Note that all the entries of $(\boldsymbol{\Delta}^{(j)})^{\mathrm{T}} \boldsymbol{\Delta}^{(j)}$ are bounded in absolute value by $\gamma^2(n-K)$, which yields

$$\|\boldsymbol{\Delta}^{(j)}\| = \sqrt{\|(\boldsymbol{\Delta}^{(j)})^{\mathrm{T}} \boldsymbol{\Delta}^{(j)}\|} \leq \sqrt{\|(\boldsymbol{\Delta}^{(j)})^{\mathrm{T}} \boldsymbol{\Delta}^{(j)}\|_F} \leq \gamma \sqrt{(n-K)K}.$$

Thus, choosing c^* small enough and using the assumption that $N \geq 2K$ we obtain

$$\|\boldsymbol{\Delta}^{(j)}\| \leq \frac{1}{4}\sqrt{n/K}.$$

Now, for $1 \leq j \leq T$, we define $\boldsymbol{W}^{(j)}$ as

$$\boldsymbol{W}^{(j)} = \boldsymbol{W}^{(0)} + \boldsymbol{\Delta}^{(j)}. \tag{4.32}$$

It is easy to check that, for each $1 \leq j \leq T$, the rows of $\boldsymbol{W}^{(j)}$ are probability vectors if c^* is chosen small enough, and $\boldsymbol{W}^{(j)}$ satisfies Assumption 1. Moreover, using (4.29) and applying Weyl's inequality once again, we obtain

$$\begin{cases} \lambda_1(\boldsymbol{W}^{(j)}) \leq \frac{7}{4}\sqrt{n/K}, \\ \lambda_K(\boldsymbol{W}^{(j)}) \geq \frac{1}{4}\sqrt{n/K}, \end{cases} \tag{4.33}$$

so that $\kappa(\boldsymbol{W}^{(j)}) \leq 7$.

2. *Constructing matrix $\boldsymbol{A}$ and checking the fact that $\boldsymbol{\Pi}^{(j)} \in \mathscr{M}$, $j = 0, 1, \ldots, T$.* Assume that p is a multiple of K (if it is not the case the definition of $\boldsymbol{A}$ should be modified by adding a block of zeros of the size of the residual). Define the following block matrix:

$$\boldsymbol{A}^0 = \{\underbrace{\mathbf{e}_1, 0_K, \ldots, 0_K}_{p/K}, \underbrace{\mathbf{e}_2, 0_K, \ldots, 0_K}_{p/K}, \ldots, \underbrace{\mathbf{e}_K, 0_K, \ldots, 0_K}_{p/K}\} \in \mathbb{R}^{K \times p},$$

where $(\mathbf{e}_1, \ldots, \mathbf{e}_K)$ is the canonical basis of $\mathbb{R}^K$ and $0_K \in \mathbb{R}^K$ is the vector with all entries 0. Define

$$\boldsymbol{A} := \frac{N-K}{N}\boldsymbol{A}^0 + \frac{K}{pN}\mathbf{1}_{K,p}.$$

All entries of $\boldsymbol{A}$ are non-negative and the rows of $\boldsymbol{A}$ sum up to 1. We have that $\sigma\left(\frac{N-K}{N}\boldsymbol{A}^0\right) = \left\{\frac{N-K}{N}, 0\right\}$ and $\sigma\left(\frac{K}{pN}\mathbf{1}_{K,p}\right) = \left\{\frac{K^{3/2}}{\sqrt{p}N}, 0\right\}$. Using the assumption that $K \leq p/4$ and Weyl's inequality we get

$$\begin{cases} \lambda_1(\boldsymbol{A}) \leq \frac{N-K}{N} + \frac{K^{3/2}}{\sqrt{p}N} \leq 1, \\ \lambda_K(\boldsymbol{A}) \geq \frac{N-K}{N} - \frac{K^{3/2}}{\sqrt{p}N} \geq 1/4, \end{cases} \tag{4.34}$$

which implies that $\kappa(\boldsymbol{A}) \leq 4$.

For $0 \leq j \leq T$, define $\boldsymbol{\Pi}^{(j)} = \boldsymbol{W}^{(j)}\boldsymbol{A}$. Using Lemma 8, (4.29), (4.33) and (4.34) we obtain

$$\lambda_K(\boldsymbol{\Pi}^{(j)}) = \lambda_K(\boldsymbol{W}^{(j)}\boldsymbol{A}) \geq \lambda_K(\boldsymbol{W}^{(j)})\lambda_K(\boldsymbol{A}) \geq \frac{1}{16}\sqrt{n/K}. \tag{4.35}$$

It follows from (4.29), (4.33) and (4.35) that the first inequality in Assumption 3 is satisfied for $\boldsymbol{W} = \boldsymbol{W}^{(j)}$ and $\boldsymbol{\Pi} = \boldsymbol{\Pi}^{(j)} = \boldsymbol{W}^{(j)}\boldsymbol{A}$, $j = 0, 1, \ldots, T$. Next, using the first inequality in (4.35) and the fact that $\lambda_1(\boldsymbol{W}^{(j)}\boldsymbol{A}) \leq \lambda_1(\boldsymbol{W}^{(j)})\lambda_1(\boldsymbol{A})$ yields

$$\kappa(\boldsymbol{W}^{(j)}\boldsymbol{A}) \leq \kappa(\boldsymbol{W}^{(j)})\kappa(\boldsymbol{A}) \leq C. \tag{4.36}$$

Thus, Assumption 3 is satisfied for $\boldsymbol{W} = \boldsymbol{W}^{(j)}$ and $\boldsymbol{\Pi} = \boldsymbol{\Pi}^{(j)} = \boldsymbol{W}^{(j)}\boldsymbol{A}$, $j = 0, 1, \ldots, T$. In conclusion, we have proved that $\boldsymbol{\Pi}^{(j)} \in \mathscr{M}$, $j = 0, 1, \ldots, T$.

To prove Theorem 3, we now use Theorem 2.5 [Tsybakov, 2008], according to which the lower bounds (4.13) and (4.14) hold if the following conditions are satisfied:

(a) $\mathrm{KL}(\mathbb{P}_{\mathbf{\Pi}^{(j)}}, \mathbb{P}_{\mathbf{\Pi}^{(0)}}) \leq \frac{\log T}{16}$, for each $j = 1, \ldots, T$, where $\mathrm{KL}(\mathbb{P}, \mathbb{Q})$ denotes the Kullback-Leibler divergence between the probability measures $\mathbb{P}$ and $\mathbb{Q}$.

(b) For $0 \leq j < \ell \leq T$ we have $\min_{\mathbf{P} \in \mathscr{P}} \|\boldsymbol{W}^{(\ell)} - \boldsymbol{W}^{(j)}\mathbf{P}\|_F \geq c\sqrt{\frac{n}{N}}$ and $\min_{\mathbf{P} \in \mathscr{P}} \|\boldsymbol{W}^{(j)} - \boldsymbol{W}^{(\ell)}\mathbf{P}\|_1 \geq c\, n\sqrt{\frac{K}{N}}$. where $\mathscr{P}$ is the set of all permutation matrices and c is a positive constant.

(c) The maps $(\boldsymbol{M}_1, \boldsymbol{M}_2) \mapsto \min_{\mathbf{P} \in \mathscr{P}} \|\boldsymbol{M}_1 - \boldsymbol{M}_2\mathbf{P}\|_F$ and $(\boldsymbol{M}_1, \boldsymbol{M}_2) \mapsto \min_{\mathbf{P} \in \mathscr{P}} \|\boldsymbol{M}_1 - \boldsymbol{M}_2\mathbf{P}\|_1$ are semi-distances.

The rest of the proof is devoted to checking that these conditions (a) – (c) are indeed satisfied.

3. *Proof of (a).*
Our aim now is to derive an upper bound on the Kullback-Leibler divergence between $\mathbb{P}_{\mathbf{\Pi}^{(j)}}$ and $\mathbb{P}_{\mathbf{\Pi}^{(0)}}$, where

$$\mathbf{\Pi}^{(j)} = \boldsymbol{W}^{(j)}\boldsymbol{A} \quad \text{and} \quad \mathbf{\Pi}^{(0)} = \boldsymbol{W}^{(0)}\boldsymbol{A}.$$

To shorten the notation, we set

$$\alpha := \frac{N-K}{N} + \frac{K}{pN}, \qquad \beta := \frac{K}{pN}.$$

For any $1 \leq i \leq n$, $1 \leq \ell \leq p$, we have $\Pi^{(0)}_{i\ell} = \sum_{k=1}^{K} W^{(0)}_{ik} A_{k\ell}$. If $i \geq K+1$, for the entries in the ith row of matrix $\mathbf{\Pi}^{(0)}$ the following holds.

- For the columns ℓ such that $(\ell - 1)$ is a multiple of p/K:
  - $\Pi^{(0)}_{i\ell}$ takes once the value $\alpha + (K-1)\gamma_1(\beta - \alpha)$,
  - $\Pi^{(0)}_{i\ell}$ takes $K-1$ times the value $\beta + \gamma_1(\alpha - \beta)$.
- For all other columns: $\Pi^{(0)}_{i\ell} \in \{\alpha, \beta\}$.

On the other hand, for any $1 \leq j \leq T$, by the definition of $\boldsymbol{W}^{(j)}$ in (4.32) we have

$$\begin{aligned} \mathbf{\Pi}^{(j)} &= \mathbf{\Pi}^{(0)} + \mathbf{\Delta}^{(j)} \\ &= \mathbf{\Pi}^{(0)} + \begin{bmatrix} \mathbf{0}_{K,p} \\ \boldsymbol{\Omega}^{(j)}\boldsymbol{A} \end{bmatrix}. \end{aligned}$$

Therefore, for any $1 \leq \ell \leq p$, if $i \leq K$, $\Pi_{i\ell}^{(j)} = \Pi_{i\ell}^{(0)}$, and if $i \geq K+1$, $\Pi_{i\ell}^{(j)} = \Pi_{i\ell}^{(0)} + \Delta_{i\ell}^{(j)}$, where

$$\Delta_{i\ell}^{(j)} = \gamma \left[\sum_{k=1}^{K/2} w_{i-K}^{(j)}(k) A_{k\ell} - \sum_{k=K/2+1}^{K} w_{i-K}^{(j)}(k - K/2) A_{k\ell} \right] .$$

If $i \geq K+1$, for the entries in the ith row of matrix $\mathbf{\Delta}^{(j)}$ the following holds.

- For the columns ℓ such that $(\ell - 1)$ is a multiple of p/K:
  - $\Delta_{i\ell}^{(j)}$ is $K/2$ times equal to $\gamma(\alpha - \beta)$,
  - $\Delta_{i\ell}^{(j)}$ is $K/2$ times equal to $-\gamma(\alpha - \beta)$.
- For all other ℓ: $\Delta_{i\ell}^{(j)} = 0$.

We are now ready to bound the Kullback-Leibler divergence between $\mathbb{P}_{\mathbf{\Pi}^{(j)}}$ and $\mathbb{P}_{\mathbf{\Pi}^{(0)}}$. Denote by $M_p(N, q)$ the multinomial distribution with parameters (N, q) where q is a probability vector in $\mathbb{R}^p$. We recall that the Kullback-Leibler divergence between two multinomial distributions $M_p(N, q_1)$ and $M_p(N, q_2)$ is equal to $N \sum_{\ell=1}^{p} q_{1\ell} \log(q_{1\ell}/q_{2\ell})$. Hence, we have

$$\begin{aligned}
\mathrm{KL}\left(\mathbb{P}_{\mathbf{\Pi}^{(j)}}, \mathbb{P}_{\mathbf{\Pi}^{(0)}}\right) &= N \sum_{i=1}^{n} \sum_{\ell=1}^{p} \Pi_{i\ell}^{(j)} \log\left(\frac{\Pi_{i\ell}^{(j)}}{\Pi_{i\ell}^{(0)}}\right) \\
&= N \sum_{i=K+1}^{n} \sum_{\ell=1}^{p} \Pi_{i\ell}^{(j)} \log\left(\frac{\Pi_{i\ell}^{(j)}}{\Pi_{i\ell}^{(0)}}\right) \\
&= N \sum_{i=K+1}^{n} \sum_{\ell=1}^{p} \left(\Pi_{i\ell}^{(0)} + \Delta_{i\ell}^{(j)}\right) \log\left(1 + \frac{\Delta_{i\ell}^{(j)}}{\Pi_{i\ell}^{(0)}}\right) \\
&\leq N \sum_{i=K+1}^{n} \sum_{\ell=1}^{p} \left(\Delta_{i\ell}^{(j)} + \frac{(\Delta_{i\ell}^{(j)})^2}{\Pi_{i\ell}^{(0)}}\right) .
\end{aligned}$$

Note that, by construction, $\sum_{\ell=1}^{p} \Delta_{i\ell}^{(j)} = 0$. Therefore,

$$\begin{aligned}
\mathrm{KL}\left(\mathbb{P}_{\mathbf{\Pi}^{(j)}}, \mathbb{P}_{\mathbf{\Pi}^{(0)}}\right) &\leq N \sum_{i=K+1}^{n} \sum_{\ell=1}^{p} \frac{(\Delta_{i\ell}^{(j)})^2}{\Pi_{i\ell}^{(0)}} \\
&= N \sum_{i=K+1}^{n} \sum_{(\ell-1) \text{ multiple of } p/K} \frac{(\Delta_{i\ell}^{(j)})^2}{\Pi_{i\ell}^{(0)}} \\
&\leq N \sum_{i=K+1}^{n} \sum_{(\ell-1) \text{ multiple of } p/K} \frac{\gamma^2(\alpha - \beta)^2}{\Pi_{i\ell}^{(0)}} \\
&\leq \frac{4c^*}{K} \sum_{i=K+1}^{n} \sum_{(\ell-1) \text{ multiple of } p/K} \frac{(N-K)^2}{N^2 \, \Pi_{i\ell}^{(0)}}
\end{aligned}$$

$$\begin{aligned}
&\leq \frac{4c^*}{K} \sum_{i=K+1}^{n} \left[\frac{3N}{N-K} + \frac{2(K-1)KN}{N-K} \right] \\
&\leq \frac{c}{K}(n-K)\frac{K^2 N}{N-K} \\
&\leq cK(n-K) \\
&\leq \frac{\log T}{16},
\end{aligned}$$

where we have used (4.31) and we have chosen c^* small enough, such that the constant c in the penultimate line does not exceed $(\log 2)/256$.

4. *Proof of (b).*

Note that for any $0 \leq j < \ell \leq T$ we have $\min_{\mathbf{P}\in\mathscr{P}} \|\boldsymbol{W}^{(\ell)} - \boldsymbol{W}^{(j)}\mathbf{P}\|_F = \|\boldsymbol{W}^{(\ell)} - \boldsymbol{W}^{(j)}\|_F$ since the first K rows are the same for matrices $\mathbf{W}^{(\ell)}$ and $\mathbf{W}^{(j)}$. Then,

$$\begin{aligned}
\|\boldsymbol{W}^{(j)} - \boldsymbol{W}^{(\ell)}\|_F^2 = \|\boldsymbol{\Omega}^{(\mathbf{j})} - \boldsymbol{\Omega}^{(\ell)}\|_F^2 &= \sum_{i=1}^{n-K} \|\Omega_{i\cdot}^{(j)} - \Omega_{i\cdot}^{(\ell)}\|_2^2 \\
&= 2\gamma^2 \sum_{i=1}^{n-K} \|w_i^{(j)} - w_i^{(\ell)}\|_2^2 = 2\gamma^2 \|w^{(j)} - w^{(\ell)}\|_2^2 \\
&\geq \frac{\gamma^2}{8} K(n-K) \quad \text{(using (4.30))} \\
&= \frac{c_*^2}{8} \frac{N(n-K)}{(N-K)^2} \geq c\frac{n}{N} \quad \text{(since } K \leq n/2\text{)}, \qquad (4.37)
\end{aligned}$$

which proves (b) for the Frobenius norm. Quite analogously, for the ℓ_1-norm we get

$$\begin{aligned}
\|\boldsymbol{W}^{(j)} - \boldsymbol{W}^{(\ell)}\|_1 = \|\boldsymbol{\Omega}^{(\mathbf{j})} - \boldsymbol{\Omega}^{(\ell)}\|_1 &= \sum_{i=1}^{n-K} \|\Omega_{i\cdot}^{(j)} - \Omega_{i\cdot}^{(\ell)}\|_1 \\
&\geq 2\gamma \|w^{(j)} - w^{(\ell)}\|_1 \\
&\geq c\, n\sqrt{\frac{K}{N}}.
\end{aligned}$$

5. *Proof of (c).*

We now prove that the map $(\boldsymbol{M}_1, \boldsymbol{M}_2) \mapsto \min_{\mathbf{P}\in\mathscr{P}} \|\boldsymbol{M}_1 - \boldsymbol{M}_2\mathbf{P}\|_F$ satisfies the triangle inequality. For any matrices $\boldsymbol{M}_1, \boldsymbol{M}_2, \boldsymbol{M}_3$, we have

$$\begin{aligned}
\min_{\mathbf{P}\in\mathscr{P}} \|\boldsymbol{M}_1 - \boldsymbol{M}_2\mathbf{P}\|_F &= \min_{\mathbf{P},\mathbf{P}'\in\mathscr{P}} \|\boldsymbol{M}_1\mathbf{P}' - \boldsymbol{M}_2\mathbf{P}\|_F \\
&\leq \min_{\mathbf{P},\mathbf{P}'\in\mathscr{P}} (\|\boldsymbol{M}_1\mathbf{P}' - \boldsymbol{M}_3\|_F + \|\boldsymbol{M}_3 - \boldsymbol{M}_2\mathbf{P}\|_F) \\
&= \min_{\mathbf{P}'\in\mathscr{P}} \|\boldsymbol{M}_1\mathbf{P}' - \boldsymbol{M}_3\|_F + \min_{\mathbf{P}\in\mathscr{P}} \|\boldsymbol{M}_3 - \boldsymbol{M}_2\mathbf{P}\|_F \\
&= \min_{\mathbf{P}'\in\mathscr{P}} \|\boldsymbol{M}_1 - \boldsymbol{M}_3\mathbf{P}'\|_F + \min_{\mathbf{P}\in\mathscr{P}} \|\boldsymbol{M}_3 - \boldsymbol{M}_2\mathbf{P}\|_F.
\end{aligned}$$

The same calculation holds with the ℓ_1-norm in place of the Frobenius norm. This completes the proof of Theorem 3.

9 Auxiliary lemmas

Lemma 6. *Let Assumption 1 be satisfied. Then,*

$$\lambda_K(\boldsymbol{W}) \geq 1. \tag{4.38}$$

If, in addition, $\lambda_K(\boldsymbol{\Pi}) > 0$ *then the matrix* $\mathbf{U}$ *of left singular vectors of* $\boldsymbol{\Pi}$ *can be represented in the form* (4.5)*, where* $\mathbf{H}$ *is a rank* K *matrix with singular values*

$$\lambda_1(\mathbf{H}) = \frac{1}{\lambda_K(\boldsymbol{W})}, \qquad \lambda_{\min}(\mathbf{H}) = \lambda_K(\mathbf{H}) = \frac{1}{\lambda_1(\boldsymbol{W})}, \tag{4.39}$$

and the condition number satisfying

$$\kappa(\mathbf{H}) = \kappa(\boldsymbol{W}). \tag{4.40}$$

Proof. Let $J^* \subseteq \{1, \ldots, n\}$ be the set of K row indices of $\boldsymbol{W}$ corresponding to anchor documents. By Assumption 1 we have $\boldsymbol{W}_{J^*} = \boldsymbol{I}_K$. Hence,

$$\lambda_K(\boldsymbol{W}) = \min_{\|a\|_2=1} \|\boldsymbol{W}a\|_2 \geq \min_{\|a\|_2=1} \|\boldsymbol{W}_{J^*}a\|_2 = 1,$$

which proves (4.38). Next, if $\lambda_K(\boldsymbol{\Pi}) > 0$ then matrix $\mathbf{L}$ is positive definite and we define $\mathbf{H} := \boldsymbol{A}\mathbf{V}\mathbf{L}^{-1}$. In view of (4.4) we have $\boldsymbol{W}\mathbf{H} = \boldsymbol{\Pi}\mathbf{V}\mathbf{L}^{-1} = \mathbf{U}$, which yields (4.5). We now prove that $\mathbf{H}$ is non-degenerate. Indeed, (4.38) implies that matrix $\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W} \in \mathbb{R}^{K\times K}$ is positive definite, so that $\mathbf{H} = (\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W})^{-1}\boldsymbol{W}^{\mathrm{T}}\mathbf{U}$. Then for the minimal singular value $\lambda_{\min}(\mathbf{H})$ of matrix $\mathbf{H}$ we have

$$\begin{aligned}\lambda_{\min}(\mathbf{H}) &= \min_{\|a\|_2=1} \|(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W})^{-1}\boldsymbol{W}^{\mathrm{T}}\mathbf{U}a\|_2 \\ &\geq \min_{x\in\mathbb{R}^n:\|x\|_2=1} \|(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W})^{-1}\boldsymbol{W}^{\mathrm{T}}x\|_2 = \frac{1}{\lambda_1(\boldsymbol{W})} > 0.\end{aligned}$$

Thus, $\mathbf{H}$ is non-degenerate and we can write $\boldsymbol{W} = \mathbf{U}\mathbf{H}^{-1}$ implying (4.39). Equality (4.40) is an immediate consequence of (4.39). □

Lemma 7. *Let* $\boldsymbol{W}, \boldsymbol{A}$ *and* $\boldsymbol{\Pi} = \boldsymbol{W}\boldsymbol{A}$ *be matrices with non-negative entries satisfying* (4.2)*. Then the singular values of matrices* $\boldsymbol{W}$ *and* $\boldsymbol{\Pi}$ *satisfy the inequalities*

$$\lambda_K(\boldsymbol{\Pi}) \leq \sqrt{n/K}, \tag{4.41}$$

$$\sqrt{n/K} \leq \lambda_1(\boldsymbol{W}) \leq \sqrt{n}, \tag{4.42}$$

$$\lambda_1(\boldsymbol{\Pi}) \leq \sqrt{K}\lambda_1(\boldsymbol{W}). \tag{4.43}$$

Proof. Inequality (4.41) follows from the fact that

$$K\lambda_K^2(\mathbf{\Pi}) \le \lambda_1^2(\mathbf{\Pi}) + \cdots + \lambda_K^2(\mathbf{\Pi}) = \|\mathbf{\Pi}\|_F^2 \le n.$$

Next, using (4.2) we obtain

$$\lambda_1(\boldsymbol{W}) \le \|\boldsymbol{W}\|_F = \sqrt{\sum_{i=1}^{n}\sum_{k=1}^{K} W_{ik}^2} \le \sqrt{\sum_{i=1}^{n}\sum_{k=1}^{K} W_{ik}} = \sqrt{n}. \tag{4.44}$$

On the other hand, for $a = (1/\sqrt{K}, \ldots, 1/\sqrt{K})^{\mathrm{T}} \in \mathbb{R}^K$ we have

$$\lambda_1(\boldsymbol{W}) \ge \|\boldsymbol{W}a\|_2 = \sqrt{\sum_{i=1}^{n} \frac{1}{K}\left(\sum_{k=1}^{K} W_{ik}\right)^2} = \sqrt{\frac{n}{K}}.$$

Quite similarly to (4.44), using (4.2) we get $\|\boldsymbol{A}\| = \lambda_1(\boldsymbol{A}) \le \|\boldsymbol{A}\|_F \le \sqrt{K}$, which implies (4.43):

$$\|\mathbf{\Pi}\| = \|\boldsymbol{W}\boldsymbol{A}\| \le \|\boldsymbol{W}\|\|\boldsymbol{A}\| \le \sqrt{K}\lambda_1(\boldsymbol{W}).$$

□

Lemma 8. *Let $K \le \min(n, p)$. For any two matrices $\boldsymbol{W} \in \mathbb{R}^{n\times K}$ and $\boldsymbol{A} \in \mathbb{R}^{K\times p}$ we have*

$$\lambda_K(\boldsymbol{W}\boldsymbol{A}) \ge \lambda_K(\boldsymbol{W})\lambda_K(\boldsymbol{A}). \tag{4.45}$$

Proof. We consider only the case $\lambda_K(\boldsymbol{A}) > 0$ since otherwise (4.45) is trivial. By Courant-Fischer min-max formula (see [Giraud, 2015, Theorem C.3] e.g.) we have

$$\lambda_K(\boldsymbol{W}\boldsymbol{A}) = \max_{S:\dim(S)=K} \min_{y\in S\setminus\{0\}} \frac{\|\boldsymbol{W}\boldsymbol{A}y\|_2}{\|y\|_2},$$

where the maximum is taken over all linear spans S of K vectors in $\mathbb{R}^p$. Since $\lambda_K(\boldsymbol{A}) > 0$ and $\boldsymbol{A}y \in \mathbb{R}^K$ we can write

$$\begin{aligned}
\lambda_K(\boldsymbol{W}\boldsymbol{A}) &= \max_{S:\dim(S)=K} \min_{y\in S\setminus\{0\}} \frac{\|\boldsymbol{W}\boldsymbol{A}y\|_2}{\|\boldsymbol{A}y\|_2}\frac{\|\boldsymbol{A}y\|_2}{\|y\|_2} \\
&\ge \min_{x\in\mathbb{R}^K\setminus\{0\}} \frac{\|\boldsymbol{W}x\|_2}{\|x\|_2} \max_{S:\dim(S)=K} \min_{y\in S\setminus\{0\}} \frac{\|\boldsymbol{A}y\|_2}{\|y\|_2} \\
&= \lambda_K(\boldsymbol{W})\lambda_K(\boldsymbol{A}).
\end{aligned}$$

□

Lemma 9. *Let $\boldsymbol{A}^0$ be a matrix with the following block structure:*

$$\boldsymbol{A}^0 = \left[\alpha_1 \mathbf{e}_1, \ldots, \alpha_K \mathbf{e}_K, \underbrace{0_K, \ldots, 0_K}_{p-K}\right] \in \mathbb{R}^{K \times p},$$

where $(\mathbf{e}_1, \ldots, \mathbf{e}_K)$ is the canonical basis of $\mathbb{R}^K$, $\alpha_i \in (0,1)$ and $0_K \in \mathbb{R}^K$ is the vector with all entries 0. Let

$$\boldsymbol{A} = \mathbf{P}_1(\boldsymbol{A}^0 + \boldsymbol{A}^1)\mathbf{P}_2 \quad \textit{and} \quad \boldsymbol{\Pi} = \boldsymbol{W}\boldsymbol{A}$$

where $\mathbf{P}_1, \mathbf{P}_2$ are permutation matrices, $\|\boldsymbol{A}^1\| \leq \beta$, and $\boldsymbol{W} \in \mathbb{R}^{n \times K}$. If $\min_{1 \leq i \leq K} \alpha_i - \beta \geq C$ then $\lambda_K(\boldsymbol{\Pi}) \geq C\lambda_K(\boldsymbol{W})$.

Proof. Matrix $\boldsymbol{A}^0$ has K top non-zero singular values $\alpha_1, \ldots, \alpha_K$. Using Weyl's inequality (see [Giraud, 2015, Theorem C.6] e.g.) we get

$$\lambda_K(\boldsymbol{A}) = \lambda_K(\boldsymbol{A}^0 + \boldsymbol{A}^1) \geq \min_{1 \leq i \leq K} \alpha_i - \beta \geq C.$$

Combining this inequality with (4.45) yields the result. □

9.1 The anchor document assumption under the Dirichlet prior

In this section, we provide a simple result that shows that the anchor document Assumption 1 is approximately satisfied with high probability under the Dirichlet prior on the document-topic matrix $\boldsymbol{W}$ if the number of documents devoted to a particular topic is large enough and the Dirichlet prior is putting a weight of at least $1/2$ on one of the topics.

Lemma 10. *Assume that we have m documents such that the corresponding document-topic vectors $w_1, \ldots, w_m \in \mathbb{R}^K$ are i.i.d. following the Dirichlet distribution parametrized by $(\alpha_1, \ldots, \alpha_K)$ with $\sum_{i=1}^K \alpha_i = 1$. Let $\alpha_K = 1 - \alpha$, where $\alpha \in (0, 1/2)$, and let $\varepsilon \in (0,1)$ and $\beta \in (0,1)$. If*

$$m \geq -\frac{2\log(1-\beta)}{\varepsilon^\alpha},$$

then

$$\mathbb{P}\left(\exists i : w_{iK} > 1 - \varepsilon\right) \geq \beta,$$

where w_{iK} denotes the Kth component of w_i.

Proof. We have

$$\mathbb{P}\left(\exists i : w_{iK} > 1 - \varepsilon\right) = 1 - \left[\mathbb{P}\left(w_{iK} < 1 - \varepsilon\right)\right]^m.$$

Here,

$$\begin{aligned}\mathbb{P}\left(w_{iK} < 1-\varepsilon\right) &= \frac{\sin(\pi\alpha_K)}{\pi}\int_0^{1-\varepsilon} x^{\alpha_K-1}(1-x)^{-\alpha_K}\mathrm{d}x \\ &= 1 - \frac{\sin(\pi\alpha_K)}{\pi}\int_{1-\varepsilon}^{1} x^{-\alpha}(1-x)^{\alpha-1}\mathrm{d}x.\end{aligned}$$

Since $\alpha < 1/2$ we have

$$\sin(\pi\alpha_K) = \sin(\pi(1-\alpha_K)) \geq \frac{\pi\alpha}{2},$$

so that

$$\frac{\sin(\pi\alpha_K)}{\pi}\int_{1-\varepsilon}^{1} x^{-\alpha}(1-x)^{\alpha-1}\mathrm{d}x \geq \frac{\alpha}{2}\int_{1-\varepsilon}^{1}(1-x)^{\alpha-1}\mathrm{d}x = \frac{1}{2}\varepsilon^{\alpha}.$$

This yields

$$\mathbb{P}\left(w_{iK} < 1-\varepsilon\right) \leq 1 - \frac{\varepsilon^{\alpha}}{2}.$$

Using the inequality $\log(1-\varepsilon^{\alpha}/2) \leq -\varepsilon^{\alpha}/2$ we finally get

$$\mathbb{P}\left(\exists i : w_{iK} > 1-\varepsilon\right) \geq 1 - \left(1 - \frac{\varepsilon^{\alpha}}{2}\right)^{m} \geq \beta.$$

□

10 Additional Experiments: Estimation of topic-word matrix

In this section, we investigate the SPOC estimator of topic-word matrix $\boldsymbol{A}$ using the sequence of experiments on synthetic data similar to those of Section 5. Figures 4.6-4.9 below present the results of simulations with different values of parameters n, p, N and the number of topics K. The generation of matrices $\boldsymbol{W}$ and $\boldsymbol{A}$ was performed in the same way as in Section 5. For each value on the x-axes of the figures, we present the averaged result over 3 simulations. Our objective is to assess the effect of each of parameters n, p, N, K on the Frobenius error between $\boldsymbol{A}$ and the estimator $\hat{\boldsymbol{A}}$ derived from SPOC algorithm via (4.9). For comparison, we provide the same simulation study for the LDA estimator of $\boldsymbol{A}$ and also Joint Stochastic Matrix Factorization algorithm (JSMF; [Lee et al., 2015]). The experiments show that the SPOC estimator is very competitive with JSMF and they show quite similar results in many regimes. The LDA estimator outperforms both SPOC and JSMF in some experiments, but it has serious issues with stability when the number of words increases; see Figure 4.8. Finally, somewhat surprisingly, SPOC is very robust when the number of topics

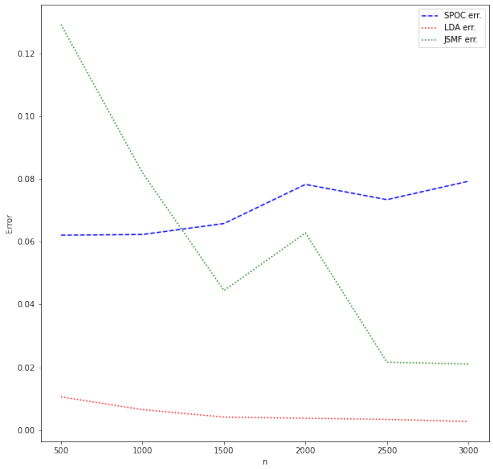


Figure 4.6: The n-dependence of the Frobenius error of $\hat{\boldsymbol{A}}$ using SPOC, LDA and JSMF algorithms. Total number of words $p = 5000$, number of sampled words in each document $N = 1000$, number of latent topics $K = 3$.

increases (see Figure 4.9), while both LDA and JSMF have clear error-increasing trend.

Additionally, we compared the runtime for the considered algorithms. For the moderate size task with $n = 1000$ documents, $p = 5000$ words in the vocabulary and $K = 3$ latent topics to be extracted, SPOC required less than 0.1 second to calculate, while for LDA and JSMF the execution time was 4 and 20 seconds respectively. Such a difference is not surprising as SPOC is non-iterative method unlike LDA and JSMF. However, of course, it has limitations and might become very computationally and memory demanding for large vocabularies and document corpora.

11 Additional Experiments: Empirical study of singular values of word-document and topic-document matrices

Most conditions and assumptions used throughout the paper are satisfied for fairly general choices of parameters. However, Assumption 3 enforces certain bounds on matrices $\boldsymbol{W}$ and $\boldsymbol{\Pi}$ which might seem restrictive. The goal of this section is to experimentally show that singular values and quotients appearing in Assumption 3 admit reasonably small upper bounds.

We consider matrices $\boldsymbol{W}, \boldsymbol{\Pi}$ and $\boldsymbol{A}$ generated in the following way. In most experiments we take $K = 3$ and the matrix $\boldsymbol{W}$ has the following structure: K rows of $\boldsymbol{W}$ are canonical basis vectors, each of the remaining $N - K$ rows is generated independently using the Dirichlet distribution with parameter $\alpha = (0.1, 0.15, 0.2)$. In the experiments where K must vary, we define $\boldsymbol{W}$ in a different way. Namely, for the $N - K$ rows that are not canonical basis vectors, each element W_{kj} is generated from

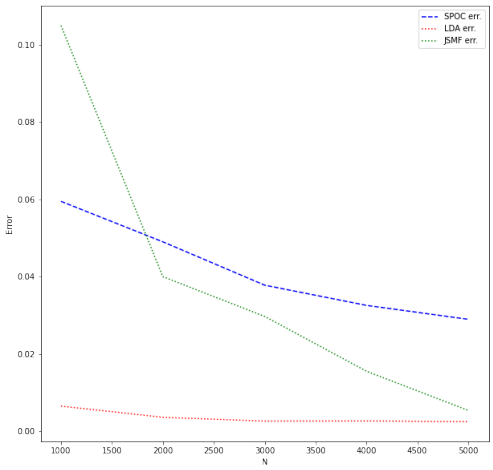


Figure 4.7: The N-dependence of the Frobenius error of $\hat{\boldsymbol{A}}$ using SPOC and LDA algorithms. Total number of words $p = 5000$, number of documents $n = 1000$, number of latent topics $K = 3$.

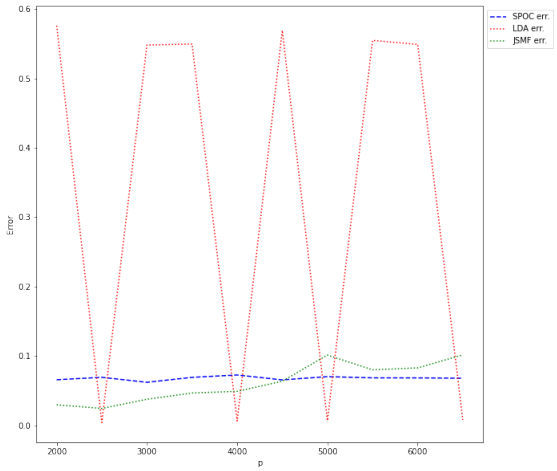


Figure 4.8: The p-dependence of the Frobenius error of $\hat{\boldsymbol{A}}$ using SPOC and LDA algorithms. Total number of words $n = 1000$, number of sampled words in each document $N = 1000$, number of latent topics $K = 3$.

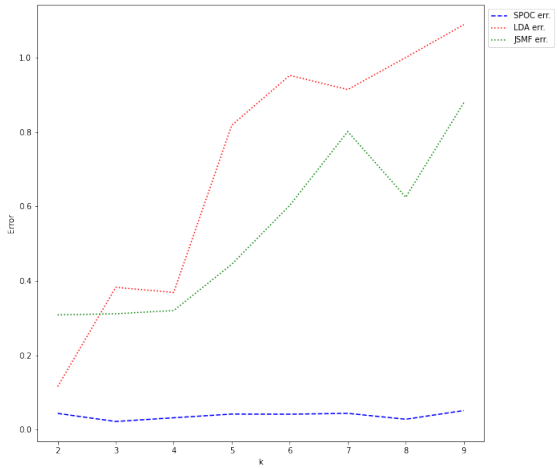


Figure 4.9: The K-dependence of the Frobenius error of $\hat{\boldsymbol{A}}$ using SPOC and LDA algorithms. Total number of words $p = 5000$, number of sampled words in each document $N = 1000$, number of documents $n = 1000$.

the uniform distribution on $[0, 1]$ and then each row of the matrix is normalized to have $\sum_{k=1}^{K} W_{ik} = 1$. For the matrix $\boldsymbol{A}$, we take K columns proportional to canonical basis vectors with coefficients equal to random variables $U_k, k = 1, \ldots, K$ uniformly distributed on $[0, 1]$. The elements A_{kj} of matrix $\boldsymbol{A}$ in the remaining $p - K$ columns are obtained by generating numbers from the uniform distribution on $[0, 1]$ and then normalizing each row of the matrix to have $\sum_{j=K+1}^{p} A_{kj} = 1 - U_k, k = 1, \ldots, K$. The resulting matrix $\boldsymbol{A}$ has normalized rows, i.e. $\sum_{j=1}^{p} A_{kj} = 1$. We essentially use the same parameters as in the experiments reported in Section 5. The dependencies of the condition numbers $\kappa(\boldsymbol{\Pi})$ and $\kappa(\boldsymbol{W})$ on parameters n, p and K are presented on Figures 4.10 and 4.11. All the condition numbers have small to moderate values for a quite wide range of parameters n and p, while the dependence on K is stronger. Additionally, we study the ratio $\lambda_1(\boldsymbol{W})/\lambda_K(\boldsymbol{\Pi})$ also appearing in Assumption 3. As presented on Figure 4.12 it shows the tendencies similar to the condition numbers.

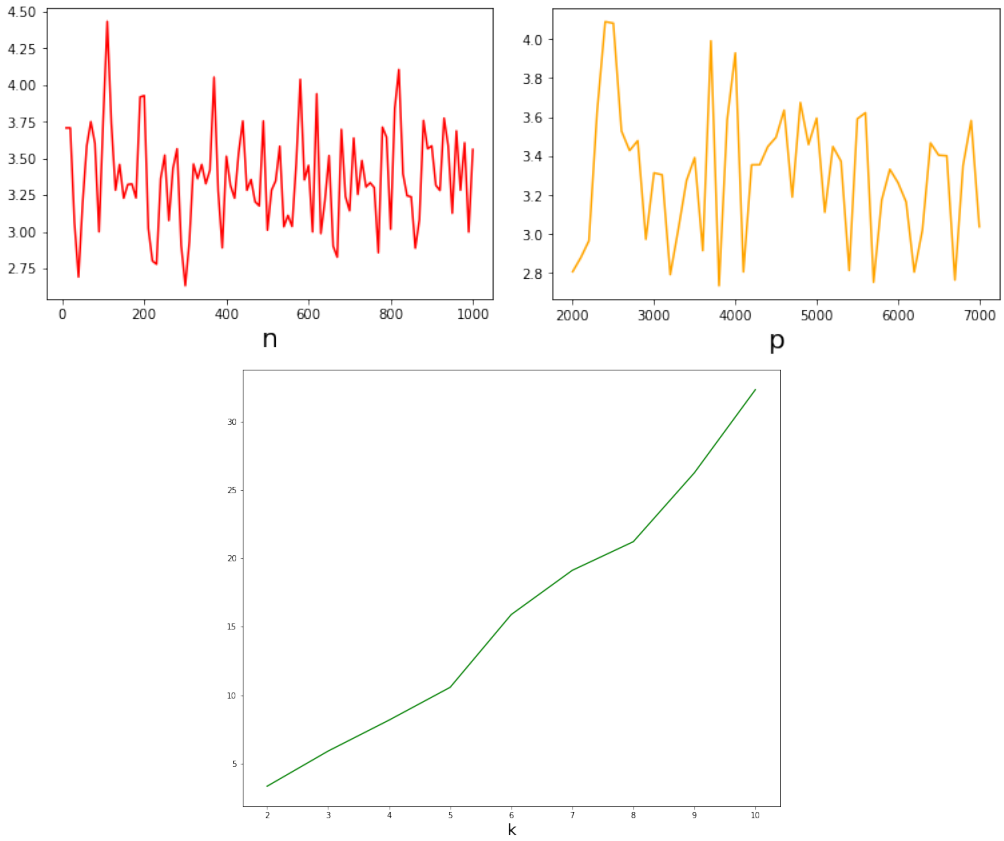


Figure 4.10: The dependence of $\kappa(\mathbf{\Pi})$ on parameters n, p and K.

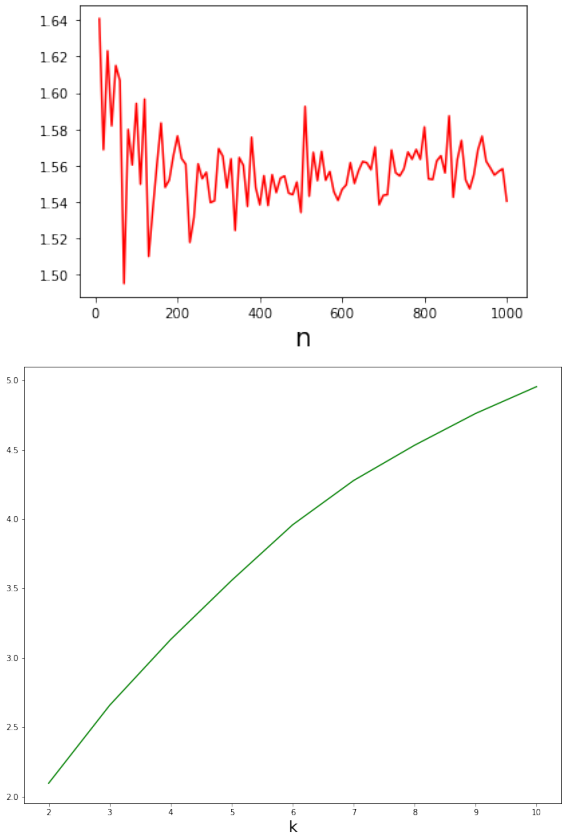


Figure 4.11: The dependence of $\kappa(\boldsymbol{W})$ on parameters n and K.

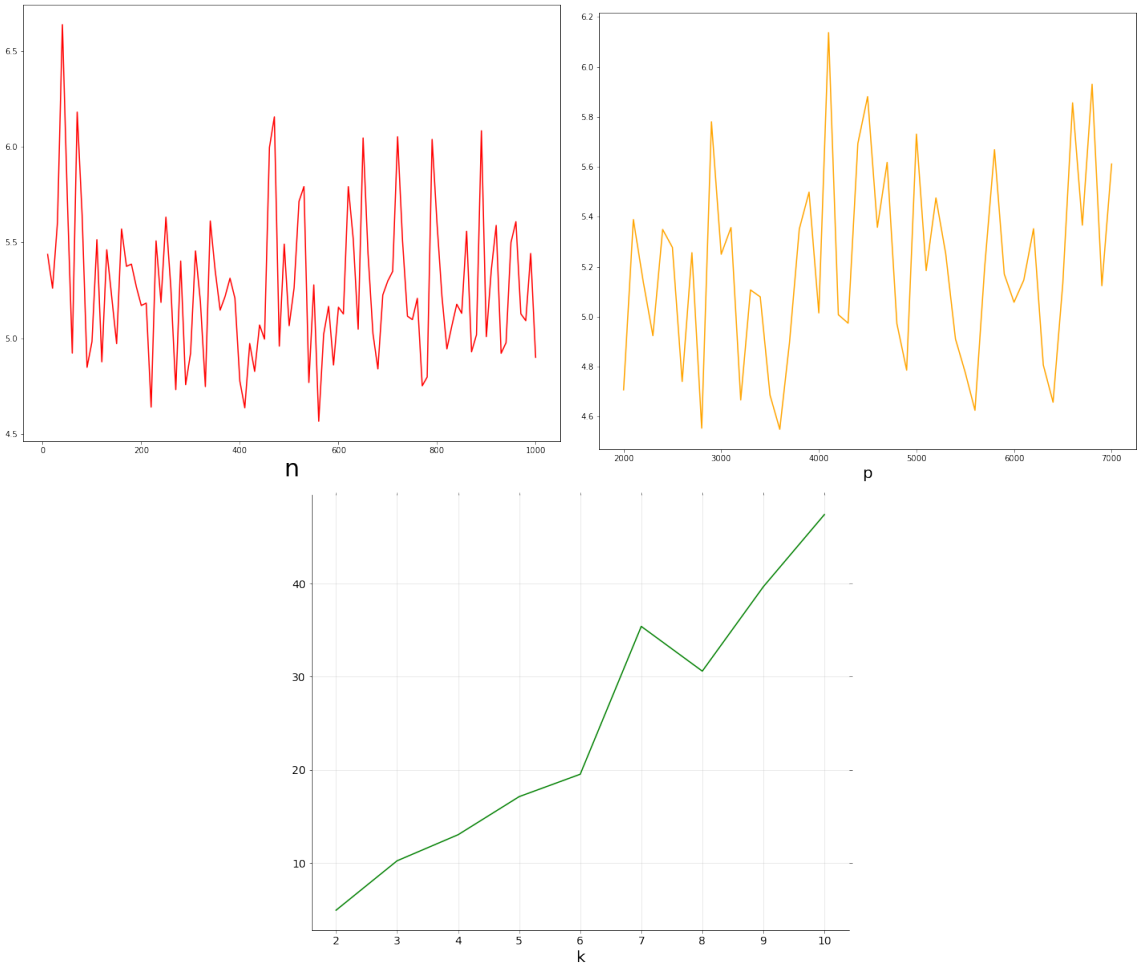


Figure 4.12: The dependence of $\lambda_1(\boldsymbol{W})/\lambda_K(\boldsymbol{\Pi})$ on parameters n, p and K.

12 Additional Experiments: Estimation for the $p = 2000$

Additionally, we decided to look on the behaviour of estimators in the scenario with a smaller size of the dictionary. We took $p = 2000$ and performed the same experiments as in Section 5. We clearly observe on Figures 4.13 and 4.14 that in a situation is more favorable for the SPOC algorithm and it significantly outperforms LDA.

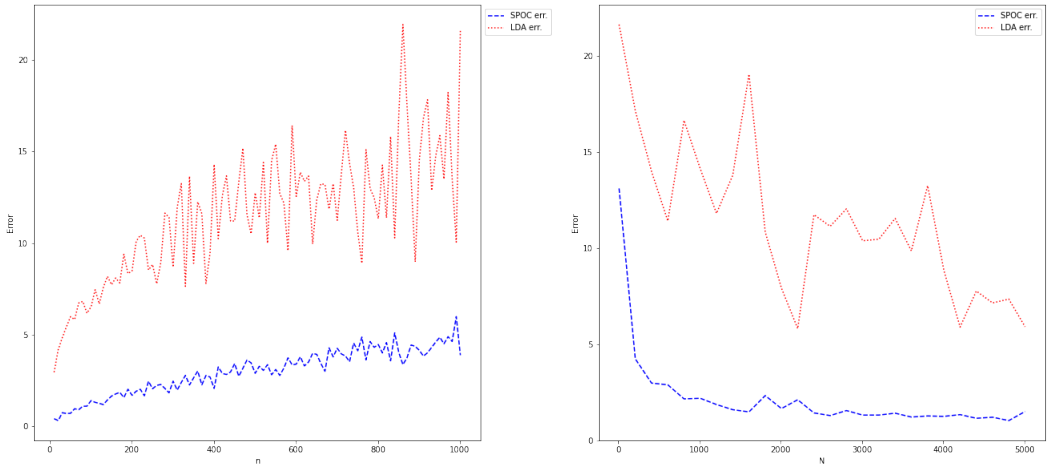


Figure 4.13: On the left (respectively, on the right), the n-dependence (respectively, the N-dependence) of $\min_{\mathbf{P}\in\mathscr{P}} \|\boldsymbol{W} - \hat{\boldsymbol{W}}\mathbf{P}\|_F$ using SPOC and LDA algorithms. Total number of words $p = 2000$ on right and left, number of sampled words $N = 200$ on the left, number of documents on the right $n = 1000$. Matrix $\boldsymbol{A}$ is generated in a way that Assumption 3 is satisfied.

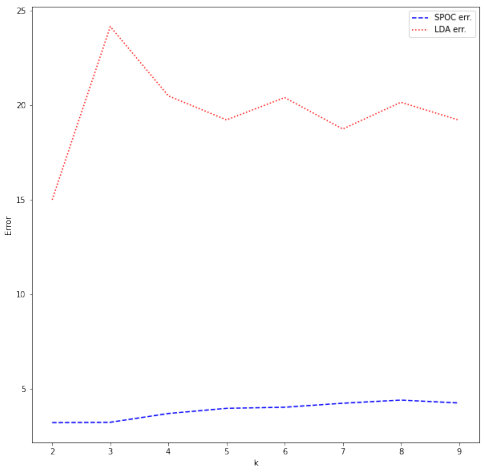


Figure 4.14: The k-dependence) of $\min_{\mathbf{P}\in\mathscr{P}} \|\boldsymbol{W} - \hat{\boldsymbol{W}}\mathbf{P}\|_F$ using SPOC and LDA algorithms. Number of documents $n = 1000$, number of sampled words $N = 5000$, and total number of words $p = 2000$. Matrix $\boldsymbol{A}$ is generated in a way that Assumption 3 is satisfied.

Chapter 5

Benign overfitting and adaptive nonparametric regression

In nonparametric regression setting, we construct an estimator, which is a continuous function interpolating the data points with high probability while attaining minimax optimal rates under mean squared risk on the scale of Hölder classes adaptively to the unknown smoothness.

This chapter is based on: J. Chhor, S. Sigalla, and A. B. Tsybakov, *Benign overfitting and adaptive nonparametric regression.* ArXiv preprint arXiv:2206.13347, 2022.

1 Introduction 118
2 Preliminaries 120
2.1 Notation 120
2.2 Model 121
2.3 Hölder classes of functions 121
3 Local polynomial estimators and interpolation 123
4 Minimax optimal interpolating estimator 125
5 Adaptive interpolating estimator 127
6 Numerical experiment 128
7 Proofs 133
8 Conclusion 147

1 Introduction

Benign overfitting has attracted a great deal of attention in the recent years. It was initially motivated by the fact that deep neural networks have good predictive

properties even when perfectly interpolating the training data [Belkin et al., 2019a], [Belkin et al., 2018b], [Zhang et al., 2021], [Belkin, 2021]. Such a behavior stands in strong contrast with the classical point of view that perfectly fitting the data points is not compatible with predicting well. With the aim of understanding this new phenomenon, a series of recent papers studied benign overfitting in linear regression setting, see [Bartlett et al., 2020], [Tsigler and Bartlett, 2020], [Chinot and Lerasle, 2020], [Muthukumar et al., 2020], [Bartlett and Long, 2021], [Lecué and Shang, 2022] and the references therein. The main conclusion for the linear model is that an unbalanced spectrum of the design matrix and over-parametrization, which in a sense approaches the model to non-parametric setting, are essential for benign overfitting to occur in linear regression. Extensions to kernel ridgeless regression were considered in [Liang and Rakhlin, 2020] when the sample size n and the dimension d were assumed to satisfy $n \asymp d$, and in [Liang et al., 2020] for a more general case $d \asymp n^\alpha$ for $\alpha \in (0, 1)$. These papers give data-dependent upper bounds on the risk that can be small assuming favorable spectral properties of the data and the kernel matrix. On the other hand, if d is constant (independent of n) then the least-norm interpolating estimator with respect to the Laplace kernel is inconsistent [Rakhlin and Zhai, 2019].

In the line of work cited above, benign overfitting was understood as achieving simultaneously interpolation and prediction consistency, or possibly, consistency with some suboptimal rates. On the other hand, it was shown that, in non-parametric regression setting, interpolating estimators can attain minimax optimal rates [Belkin et al., 2019b]. Namely, it is proved in [Belkin et al., 2019b] that interpolation with minimax optimal rates can be achieved by Nadaraya-Watson estimator with a singular kernel.

The idea of using singular kernels can be traced back to [Shepard, 1968] giving start to popular techniques in image processing referred to as Shepard interpolation. In statistical language, Shepard interpolant is nothing else but the Nadaraya-Watson estimator with kernel $K(u) = 1/\|u\|^2$, where $\|\cdot\|$ denotes the Euclidean norm and $u \in \mathbf{R}^2$. Unaware of Shepard's work and its subsequent extensive use in image processing, [Devroye et al., 1998] considered the same estimator in general dimension d, that is, with the kernel $K(u) = \|u\|^{-d}$ for $u \in \mathbf{R}^d$, and proved that the Nadaraya-Watson estimator with such a kernel is consistent in probability but fails to be pointwise almost surely consistent. However, this kernel is not integrable and has a peculiar property that the bandwidth cancels out from the definition of the estimator. Thus, the bias cannot be controlled and the bias-variance trade-off argument based on bandwidth selection does not apply. It remains unclear whether some rates of convergence can be achieved by such an estimator. Therefore, it was suggested in [Belkin et al., 2019b] to localize and modify the kernel as $K(u) = \|u\|^{-a}\mathbf{1}(\|u\| \le 1)$ where $0 < a < d/2$ rather than $a = d$ and $\mathbf{1}(\cdot)$ denotes the indicator function. The estimator with such a weaker type of singularity is also interpolating, and it was shown in [Belkin et al., 2019b] that it achieves the minimax rates of convergence on the β-Hölder classes with $0 < \beta \le 2$. Also, [Belkin et al., 2018a] proved a similar claim for the k nearest neighbor analog of this estimator with $0 < \beta \le 1$. However,

those results were restricted to functions with low smoothness β and the suggested estimators were not adaptive to β.

In this paper, we show that:

(i) interpolating estimators attaining minimax optimal rates on β-Hölder classes can be obtained for any smoothness $\beta > 0$,

(ii) estimators with such properties can be constructed adaptively to the unknown smoothness $\beta \in (0, \beta_{\max}]$, for any $\beta_{\max} > 0$, and to the unknown parameter $L > 0$ of the Hölder class of regression functions.

The estimators that we consider to achieve (i) are local polynomial estimators (LPE) with singular kernels. In order to obtain adaptive estimators achieving (ii), we apply aggregation techniques to a family of LPE with singular kernels.

As a by-product, we obtain non-asymptotic bounds for the squared risk of LPE in classical setting with non-singular kernels. To the best of our knowledge, such bounds are missing in the existing literature on LPE that was mainly focused on asymptotic properties such as convergence in probability or pointwise asymptotic normality, cf. [Stone, 1980, Stone, 1982, Tsybakov, 1986, Fan and Gijbels, 1996].

Note that local polynomial method with singular kernels has been used as interpolation tool in numerical analysis, starting from [Lancaster and Salkauskas, 1981]. It was also invoked in the context of non-parametric regression in [Katkovnik, 1985]. However, [Lancaster and Salkauskas, 1981, Katkovnik, 1985] only discussed functional properties, such as the smoothness of interpolants, rather than their statistical behavior.

2 Preliminaries

2.1 Notation

For any vector $x = (x_1, \dots, x_d) \in \mathbf{R}^d$ and any multi-index $s = (s_1, \dots, s_d) \in \mathbf{N}^d$, we define

$$|s| = \sum_{i=1}^{d} s_i, \qquad s! = s_1! \dots s_d!$$

$$x^s = x_1^{s_1} \dots x_d^{s_d} \qquad D^s = \frac{\partial^{s_1+\dots+s_d}}{\partial x_1^{s_1} \dots \partial x_d^{s_d}}.$$

We denote by $\|\cdot\|$ the Euclidean norm, and by $\mathrm{Card}(J)$ the cardinality of set J. For any integer $k \in \mathbf{N}^*$, we set $[k] = \{1, \dots, k\}$. For any $x \in \mathbf{R}^d$, $r > 0$, we denote by $\mathscr{B}_d(x, r)$ the closed Euclidean ball centered at x with radius r. We set for brevity $\mathscr{B}_d = \mathscr{B}_d(0, 1)$. For any $\beta > 0$, we denote by $\lfloor \beta \rfloor$ the maximal integer less than β, and by $\lceil \beta \rceil$ the minimal integer greater than β. We use symbols C, C' to denote positive constants that can vary from line to line.

For any $k > 0$, we denote by I_k the identity matrix of size k. For any square matrix M, the writing $M \succ 0$ means that M is positive definite. For any matrix M, we denote by M^+ its Moore-Penrose inverse, and by $\|M\|_\infty$ its spectral norm.

2.2 Model

Let (X, Y) be a pair of random variables in $\mathbf{R}^d \times \mathbf{R}$ with distribution P_{XY} and assume that we are given n i.i.d. observations $\mathscr{D} := \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ with distribution P_{XY}. We denote by P_X the marginal distribution of X and assume that it admits a density p with respect to the Lebesgue measure on the compact set $\mathrm{Supp}(p)$. We assume that for all $x \in \mathrm{Supp}(p)$, the regression function $f(x) = \mathbf{E}(Y|X = x)$ exists and is finite. Set $\xi(X) = Y - \mathbf{E}(Y|X)$. Equivalently, the model can be written as $Y_i = f(X_i) + \xi(X_i)$, where $\mathbf{E}(\xi(X_i)|X_i) = 0$. We make the following assumptions.

Assumption (A1). $\mathbf{E}(|\xi(X)|^{2+\delta}|X = x) \leq C_\xi$ *for all* $x \in$ *Supp*(p)*, where* δ *and* C_ξ *are positive constants.*

Assumption (A2). *The random vector* X *is distributed with Lebesgue density* $p(\cdot)$ *such that* $p \in [p_{\min}, p_{\max}]$ *where* $p_{\max} \geq p_{\min} > 0$*. The support Supp*$(p)$ *of* p *is a convex compact set contained in* $\mathscr{B}_d$*.*

For any estimator f_n of f based on the sample $\mathscr{D}$, we consider the following L_2-loss :

$$\|f_n - f\|_{L_2}^2 = \mathbf{E}_X\left([f_n(X) - f(X)]^2\right) = \int [f_n(x) - f(x)]^2 p(x)dx,$$

where $\mathbf{E}_X$ denotes the expectation with respect to P_X. We define the expected risk as $\mathbf{E}\left[\|f_n - f\|_{L_2}^2\right]$, where $\mathbf{E}$ denotes the expectation with respect to the distribution of $\mathscr{D}$.

Definition 10 (Interpolating estimator)**.** *An estimator* f_n *of* f *based on a sample* $\mathscr{D} = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ *is called interpolating over* $\mathscr{D}$ *if* $f_n(X_i) = Y_i$ *for* $i = 1, \ldots, n$*.*

2.3 Hölder classes of functions

For any k-linear form $A : (\mathbf{R}^d)^k \longrightarrow \mathbf{R}$, we define its norm as follows

$$\|A\|_* := \sup\left\{\left|A[h_1, \ldots, h_k]\right| : \|h_j\| \leq 1, j \in [k]\right\}. \tag{5.1}$$

Given a k-times continuously differentiable function $f : \mathbf{R}^d \longrightarrow \mathbf{R}$ and $x \in \mathbf{R}^d$, we denote by $f^{(k)}(x) : (\mathbf{R}^d)^k \longrightarrow \mathbf{R}$ the following k-linear form

$$f^{(k)}(x)[h_1, \ldots, h_k] = \sum_{|m_j|=1, \forall j \in [k]} D^{m_1+\cdots+m_k} f(x) h_1^{m_1} \ldots h_k^{m_k}, \quad \forall h_1, \ldots, h_k \in \mathbf{R}^d,$$

where $m_1, \ldots, m_k \in \mathbf{N}^d$ are multi-indices. Throughout the paper, we will consider the following Hölder class of functions.

Definition 11. *Let* $\beta > 0$*,* $L > 0$*, and let* $f : \mathscr{B}_d \longrightarrow \mathbf{R}$ *be a* $\ell = \lfloor\beta\rfloor$ *times continuously differentiable function. We denote by* $\Sigma(\beta, L)$ *the set of all functions* f

defined on $\mathscr{B}_d$ such that

$$\max_{0\le k\le \ell}\sup_{x\in\mathscr{B}_d}\left\|f^{(k)}(x)\right\|_* + \sup_{x,x'\in\mathscr{B}_d}\frac{\left\|f^{(\ell)}(x)-f^{(\ell)}(x')\right\|_*}{\|x-x'\|^{\beta-\ell}} \le L.$$

These classes of functions have nice embedding properties that will be needed to prove our result on adaptive estimation. For $\beta' \le \beta \le 1$, we clearly have $\Sigma(\beta, L) \subseteq \Sigma(\beta', L)$. Analogous embedding is valid for $\beta > 1$ as stated in the next lemma proved in Section 7.

Lemma 3. *For any $0 < \beta' \le \beta$ and $L > 0$ we have $\Sigma(\beta, L) \subseteq \Sigma(\beta', 2L)$.*

The class $\Sigma(\beta, L)$ is closely related to several differently defined Hölder classes used in the literature. One of them is based on Taylor approximation, cf., for example, [Stone, 1980]. For any $x \in \mathbf{R}^d$ and any ℓ times continuously differentiable real-valued function f on $\mathbf{R}^d$, we denote by Tf_x its Taylor polynomial of degree ℓ at point x:

$$Tf_x(x') = \sum_{0\le|s|\le\ell}\frac{(x-x')^s}{s!}D^s f(x').$$

Lemma 4. *Let $\beta > 0$, $L > 0$ and $f \in \Sigma(\beta, L)$. Then for all $x, y \in \mathscr{B}_d$, and $\ell = \lfloor\beta\rfloor$ it holds that*

$$|f(x) - Tf_y(x)| \le \frac{L}{\ell!}\|x-y\|^\beta.$$

Thus, we have $\Sigma(\beta, L) \subseteq \Sigma'(\beta, L/\lfloor\beta\rfloor!)$, where $\Sigma'(\beta, L')$ stands for the class of all functions f satisfying the relation $|f(x) - Tf_y(x)| \le L'\|x-y\|^\beta$.

Next, considering one more definition of Hölder class:

$$\widetilde{\Sigma}(\beta, L) = \left\{f : \mathscr{B}_d \to \mathbf{R} : \sup_{x,x'}\frac{\left\|f^{(\ell)}(x)-f^{(\ell)}(x')\right\|_*}{\|x-x'\|^{\beta-\ell}} \le L\right\}$$

we also immediately have that $\Sigma(\beta, L) \subseteq \widetilde{\Sigma}(\beta, L)$. It follows from [Stone, 1982] that the minimax estimation rate on the class $\widetilde{\Sigma}(\beta, L)$ under the squared loss that we consider below is $n^{-\frac{2\beta}{2\beta+d}}$ up to constants depending only on β and d. Notice that the functions in $\widetilde{\Sigma}(\beta, L)$ used in the lower bound construction in [Stone, 1982] can be rescaled into functions in $\Sigma(\beta, L)$ by multiplying by a factor depending only on β and d. Hence, the lower bound construction in [Stone, 1982] remains valid for the class $\Sigma(\beta, L)$. It implies that the minimax rate of estimation on the class $\Sigma(\beta, L)$ is $n^{-\frac{2\beta}{2\beta+d}}$. In conclusion, though $\Sigma(\beta, L)$ is a subclass of suitable Hölder classes Σ' and $\widetilde{\Sigma}$ it is not substantially smaller, in the sense that estimation over these classes is essentially equally difficult.

3 Local polynomial estimators and interpolation

For $\ell \in \mathbf{N}$ let $C_{\ell,d} = \binom{\ell+d}{d}$ be the cardinality of the set of multi-indices $\{s = (s_1, \dots, s_d) \in \mathbf{N}^d, 0 \leq |s| \leq \ell\}$. We assume that the elements $s^{(1)}, \dots, s^{(C_{\ell,d})}$ of this set are ordered according to the increasing values of $|s|$, and in an arbitrary way for equal values of $|s|$. In particular, $s^{(1)} = (0, \dots, 0)$. For any $u \in \mathbf{R}^d$, define the vector $U(u) \in \mathbf{R}^{C_{\ell,d}}$ as follows:

$$U(u) := \left(\frac{u^s}{s!}\right)_{|s| \leq \ell},$$

where the components of $U(u)$ are ordered in the same way as $s^{(i)}$'s. In particular, the first component of $U(u)$ is 1 for any u.

The definition of local polynomial estimator usually given in the literature is as follows, cf., e.g., [Tsybakov, 2008]. Let $K : \mathbf{R}^d \to \mathbf{R}_+$ be a kernel, $h > 0$ be a bandwith and $\ell \geq 0$ be an integer. Consider a vector $\hat{\theta}_n(x) \in \mathbf{R}^{C_{\ell,d}}$ such that

$$\hat{\theta}_n(x) \in \operatorname*{argmin}_{\theta \in \mathbf{R}^{C_{\ell,d}}} \sum_{i=1}^{n} \left[Y_i - \theta^\top U\left(\frac{X_i - x}{h}\right)\right]^2 K\left(\frac{X_i - x}{h}\right) \tag{5.2}$$

Then

$$f_n(x) = U^\top(0)\hat{\theta}_n(x) \tag{5.3}$$

is called a local polynomial estimator of order ℓ of $f(x)$. Note that $f_n(x)$ is the first component of $\hat{\theta}_n(x)$.

However, this definition is not convenient for our purposes. First, $\hat{\theta}_n(x)$ is not uniquely defined for such $x \in \mathbf{R}^d$ that the matrix

$$B_{nx} := \frac{1}{nh^d} \sum_{i=1}^{n} U\left(\frac{X_i - x}{h}\right) U^\top\left(\frac{X_i - x}{h}\right) K\left(\frac{X_i - x}{h}\right) \in \mathbf{R}^{C_{\ell,d} \times C_{\ell,d}}$$

is degenerate. Furthermore, $\hat{\theta}_n(x)$ is not defined for $x = X_i$ if the kernel K has a singularity at 0, which will be the main case of interest in what follows. Therefore, we adopt the following slightly different definition.

Definition 12 (Local polynomial estimator)**.** *If the kernel K is bounded then the local polynomial estimator of order ℓ (or shortly, LP(ℓ) estimator) of $f(x)$ at point x is defined as*

$$f_n(x) = \sum_{i=1}^{n} Y_i W_{ni}(x), \tag{5.4}$$

where, for $i = 1, \dots, n$, the weights $W_{ni}(x)$ are given by

$$W_{ni}(x) = \frac{U^\top(0)}{nh^d} B_{nx}^+ U \left(\frac{X_i - x}{h}\right) K\left(\frac{X_i - x}{h}\right). \tag{5.5}$$

If the kernel K has a singularity at 0, that is, $\lim_{u\to 0} K(u) = +\infty$, then the LP($\ell$) estimator of $f(x)$ at point $x \notin \{X_1, \dots, X_n\}$ is still defined by (5.4) while we set, for $j = 1, \dots, n$,

$$f_n(X_j) = \limsup_{z \to X_j} f_n(z). \tag{5.6}$$

The purpose of (5.6) is to provide a valid definition for kernels with singularity at 0. We introduce lim sup in (5.6) for formal reasons. In the cases of our interest described in the next lemma there exists an exact limit in (5.6): $\lim_{x\to X_j} f_n(x) = Y_j$ for all $j \in [n]$, which means that the estimator f_n is interpolating.

Lemma 5. *[Interpolation property of LPE] Let f_n be an LP(ℓ) estimator with kernel $K : \mathbf{R}^d \to \mathbf{R}_+$ having a singularity at 0, that is, $\lim_{u\to 0} K(u) = +\infty$, and continuous on $\mathbf{R}^d \setminus \{0\}$. In particular, there exist $c_0 > 0$ and $\Delta > 0$ such that*

$$K(u) \geq c_0 \mathbf{1}(\|u\| \leq \Delta), \quad \forall u \in \mathbf{R}^d. \tag{5.7}$$

Assume that $X_1, \dots, X_n$ are distinct points in $\mathbf{R}^d$ and there exists a constant $\lambda_1 > 0$ such that

$$\sum_{j=1}^n U\left(\frac{X_j - x}{h}\right) U^\top \left(\frac{X_j - x}{h}\right) \mathbf{1}\left(\left\|\frac{X_j - x}{h}\right\| \leq \Delta\right) \succ \lambda_1 I_{C_{\ell,d}} \tag{5.8}$$

for all x in some neighborhood of X_i, where $I_{C_{\ell,d}}$ denotes the identity matrix. Then $f_n(X_i) = Y_i$.

For $\ell = 0$ (corresponding to the Nadaraya-Watson estimator) condition (5.8) is trivially satisfied since the expression on the left hand side is a positive scalar for any x in a neighborhood of X_i. For general ℓ, this condition is satisfied with high probability if X_j's are distributed with a density bounded away from zero on its support. Indeed, we have the following result. For $\Delta > 0$ consider the matrix

$$\overline{B}_{nx} := \frac{1}{nh^d} \sum_{i=1}^n U\left(\frac{X_i - x}{h}\right) U^\top \left(\frac{X_i - x}{h}\right) \mathbf{1}\left(\left\|\frac{X_i - x}{h}\right\| \leq \Delta\right) \in \mathbf{R}^{C_{\ell,d} \times C_{\ell,d}}.$$

Lemma 6. *Let $h \leq \alpha$, where $\alpha > 0$. Let Assumption (A2) be satisfied. Then, the following holds.*

(i) For any $\Delta > 0$ there exist constants $\lambda_0(\ell) > 0$, $c > 0$ independent of n and x and depending only on $\ell, \alpha, \Delta, d, p(\cdot)$ such that

$$\mathbf{P}\Big(\inf_{x\in \mathrm{Supp}(p)} \lambda_{\min}(\overline{B}_{nx}) \geq \lambda_0(\ell)\Big) \geq 1 - c(h^{-d^2-d}e^{-nh^d/c} + e^{-n^3h^{2d}/c}),$$

where $\lambda_{\min}(\overline{B}_{nx})$ is the minimal eigenvalue of $\overline{B}_{nx}$. Moreover, $\lambda_0(\ell) \geq \lambda_0(\ell')$ if $\ell \leq \ell'$.

(ii) If K is a kernel satisfying (5.7) then there exist constants $\lambda_0'(\ell) > 0$, $c' > 0$ independent of n and x and depending only on $\ell, \alpha, \Delta, d, p(\cdot)$ such that

$$\mathbf{P}\Big(\inf_{x\in \mathrm{Supp}(p)} \lambda_{\min}(B_{nx}) \geq \lambda_0'(\ell)\Big) \geq 1 - c'(h^{-d^2-d}e^{-nh^d/c'} + e^{-n^3h^{2d}/c'}).$$

Note that part (ii) of Lemma 6 is an immediate consequence of its part (i) and the fact that $B_{nx} \succ c_0\overline{B}_{nx}$ if (5.7) holds. Also, the next corollary follows immediately from Lemmas 5 and 6.

Corollary 1. *Let f_n be an LP(ℓ) with kernel $K : \mathbf{R}^d \to \mathbf{R}_+$ having a singularity at 0, that is, $\lim_{u\to 0} K(u) = +\infty$, and continuous on $\mathbf{R}^d \setminus \{0\}$. Let $h = \alpha n^{-\frac{1}{2\beta+d}}$, where $\alpha, \beta > 0$ and let Assumption (A2) be satisfied. Then, there exists a constant $c' > 0$ such that, with probability at least $1 - c'e^{-A_n/c'}$, where $A_n = n^{\frac{2\beta}{2\beta+d}}$, the LPE f_n is interpolating, that is, $f_n(X_i) = Y_i$ for $i = 1, \dots, n$, and $f_n(\cdot)$ is a continuous function on $\mathrm{Supp}(p)$. Furthermore, the LP(0) estimator is interpolating with probability 1.*

Note that the kernels $K(u) = \|u\|^{-a}\mathbb{1}(\|u\| \leq 1)$ with $a \in (0, d/2)$ considered in [Belkin et al., 2019b] are not continuous on $\mathbf{R}^d \setminus \{0\}$ and thus do not satisfy the conditions of Lemma 5 and Corollary 1. On the other hand, these conditions are met for the kernels $K(u) = \|u\|^{-a}\cos^2(\pi\|u\|/2)\mathbb{1}(\|u\| \leq 1)$ or $K(u) = \|u\|^{-a}(1 - \|u\|)_+$ with $a > 0$.

4 Minimax optimal interpolating estimator

In this section, we show that for any $\beta > 0$, one can construct an interpolating local polynomial estimator reaching the minimax rate $n^{-\frac{2\beta}{2\beta+d}}$ on the Hölder class $\Sigma(\beta, L)$.

In what follows, we assume that we know a constant L_0 such that $|f(x)| \leq L_0$ for all $x \in \mathrm{Supp}(p)$. We denote the class of all such functions f by $\mathscr{F}_0$. This assumption is not crucial and can be avoided at the expense of slightly more involved dependence of the result on the noise distribution (see Remark 1 below).

Let f_n be an LP(ℓ) estimator of order $\ell = \lfloor\beta\rfloor$. Set $\mu := L_0 \vee \max_{1\leq i\leq n} |Y_i|$ and consider the truncated estimator

$$\bar{f}_n(x) = \big[f_n(x)\big]_{-\mu}^{\mu}, \tag{5.9}$$

where for all $y \in \mathbf{R}$ and $a \leq b$ the truncation of y between a and b is defined as $[y]_a^b := (y \vee a) \wedge b$.

Theorem 5. *Let Assumptions (A1) and (A2) be satisfied. Let $f \in \Sigma(\beta, L)$ for $\beta > 0, L > 0$, and $|f(x)| \le L_0$ for all $x \in \mathrm{Supp}(p)$ and a constant $L_0 > 0$. Consider the estimator $\bar{f}_n$ defined in (5.9), where f_n is the LP(ℓ) estimator with $\ell = \lfloor \beta \rfloor$, $h = \alpha n^{-\frac{1}{2\beta+d}}$, for some $\alpha > 0$, and kernel K.*

(i) If K is a compactly supported kernel satisfying (5.7) and $\int K^2(u)du < \infty$ then

$$\mathbf{E}\left([\bar{f}_n(x) - f(x)]^2\right) \le C n^{-\frac{2\beta}{2\beta+d}}, \quad \forall x \in \mathrm{Supp}(p), \tag{5.10}$$

$$\mathbf{E}\left(\|\bar{f}_n - f\|_{L_2}^2\right) \le C n^{-\frac{2\beta}{2\beta+d}}, \tag{5.11}$$

where $C > 0$ is a constant depending only on $\beta, L, L_0, d, C_\xi, K, p_{\max}, p_{\min}$ and α.

(ii) If, in addition, $\lim_{u\to 0} K(u) = +\infty$ and K is continuous on $\mathbf{R}^d \setminus \{0\}$, then there exists a constant $c' > 0$ such that, with probability at least $1 - c'e^{-A_n/c'}$, where $A_n = n^{\frac{2\beta}{2\beta+d}}$, the estimator $\bar{f}_n$ is interpolating, that is, $\bar{f}_n(X_i) = Y_i$ for $i = 1, \dots, n$, and $\bar{f}_n(\cdot)$ is a continuous function on $\mathrm{Supp}(p)$.

Note that, for the examples of singular kernels given at the end of the previous section, we need $a \in (0, d/2)$ to grant the condition $\int K^2(u)du < \infty$ required in Theorem 5. Moreover, Shepard kernel $K(u) = \|u\|^{-d}$ does not satisfy the assumptions of Theorem 5.

Remark 1. The value $\max_{1\le i\le n} |Y_i|$ is introduced in the threshold μ only with the aim to preserve the interpolation property. Inspection of the proof shows that Theorem 5(i) remains valid when $\max_{1\le i\le n} |Y_i|$ is dropped from the definition of μ, so that $\mu = L_0$, but in this case data interpolation is not granted. On the other hand, by setting $\mu = 2\max_{1\le i\le n} |Y_i|$ it is possible to obtain both items (i) and (ii) of Theorem 5 for an estimator that does not require the knowledge of L_0. We do not state this result here since we are able to prove it with the constant C in (5.10) - (5.11) depending not only on C_ξ but also on a tail property of the distribution of $\xi(X)$ given X.

Remark 2. Theorem 5(i) completes the existing literature on LPE in the classical setting when the kernel is non-singular. To the best of our knowledge, non-asymptotic bounds on the mean squared error of LPE were not obtained. The previous work was mainly focused on asymptotic properties such as convergence in probability or pointwise asymptotic normality, cf. [Stone, 1980, Stone, 1982, Tsybakov, 1986, Fan and Gijbels, 1996]. For binary $Y \in \{0, 1\}$ specific to classification setting, non-asymptotic deviation bounds for LPE were obtained in [Audibert and Tsybakov, 2007]. However, the techniques of [Audibert and Tsybakov, 2007] cannot be extended beyond the case of bounded Y.

Remark 3. Inspection of the proof shows that Theorem 5 extends to kernels K that are not necessarily compactly supported. It suffices to assume that the integrals $\int(1 + \|u\|^\beta)K(u)du$ and $\int(1 + \|u\|^{2\beta})K^2(u)du$ are finite.

5 Adaptive interpolating estimator

In this section, we will use the following assumption on the noise $\xi(X)$.

Assumption (A3). *Conditionally on $X = x$, the random variable $\xi(X)$ is a zero-mean σ_ξ-subgaussian random variable for all $x \in Supp(p)$.*

We propose an adaptive estimator that does not need the knowledge of β, L, C_ξ, achieves the minimax L_2-rate of convergence on classes $\Sigma(\beta, L)$ for all $L > 0$ and $\beta \in (0, \beta_{\max}]$, where $\beta_{\max} > 0$ is an arbitrary given value, and is interpolating with high probability. Our adaptive estimator is based on least squares aggregation. We refer to [Wegkamp, 2003] for the study of such aggregation procedures.

Assume without loss of generality that n is even. We split the sample $\mathscr{D} = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$ into two independent subsamples $\mathscr{D}_1 = \left\{(X_1, Y_1), \dots, (X_{\frac{n}{2}}, Y_{\frac{n}{2}})\right\}$ and $\mathscr{D}_2 = \left\{(X_{\frac{n}{2}+1}, Y_{\frac{n}{2}+1}), \dots, (X_n, Y_n)\right\}$, and we proceed in two steps.

1. Choose a finite grid $(\beta_j)_{j\in J}$ on the values of β. Let $f_{n,j}$ denote a $\mathrm{LP}(\ell_j)$ estimator (with $\ell_j = \lfloor \beta_j \rfloor$) based on the subsample $\mathscr{D}_1$ with bandwidth $h = \alpha n^{-\frac{1}{2\beta_j + d}}$, $\alpha > 0$, and kernel K satisfying the assumptions of Theorem 5. Set $\mu := L_0 \vee \max_{1\le i\le n/2} |Y_i|$ and construct $|J|$ truncated local polynomial estimators:

$$\bar{f}_{n,j}(x) = \left[f_{n,j}(x)\right]_{-\mu}^{\mu}, \qquad j \in J. \tag{5.12}$$

   By Theorem 5, each estimator $\bar{f}_{n,j}$ is interpolating over $\mathscr{D}_1$ with high probability, and satisfies

$$\sup_{f\in\Sigma(\beta_j,L)\cap\mathscr{F}_0} \mathbf{E}_1\left[\|\bar{f}_{n,j} - f\|_{L_2}^2\right] \le C n^{-\frac{2\beta_j}{2\beta_j+d}}, \tag{5.13}$$

   where $\mathbf{E}_1$ denotes the expectation with respect to the distribution of $\mathscr{D}_1$.

2. From the collection $(f_{n,j})_{j\in J}$, we select an estimator $\widetilde{f}_n$ that minimizes the sum of squares over the second subsample $\mathscr{D}_2$, that is, we set $\widetilde{f}_n = \bar{f}_{n,\widetilde{j}}$ with

$$\widetilde{j} \in \operatorname*{argmin}_{j\in J} \sum_{k=\frac{n}{2}+1}^{n} \left(Y_k - \bar{f}_{n,j}(X_k)\right)^2.$$

As each of the estimators among $(\bar{f}_{n,j})_{j\in J}$ is interpolating over $\mathscr{D}_1$, the estimator $\widetilde{f}_n$ is also interpolating over $\mathscr{D}_1$, but not over $\mathscr{D}_2$. We therefore introduce the estimator $\widetilde{g}_n$ obtained in the same way as $\widetilde{f}_n$ by interchanging $\mathscr{D}_1$ and $\mathscr{D}_2$. Thus, $\widetilde{g}_n$ is interpolating over $\mathscr{D}_2$. Next, we define an estimator interpolating over $\mathscr{D}_1 \cup \mathscr{D}_2$ by combining $\widetilde{f}_n$ and $\widetilde{g}_n$ as follows.

For any $x \in \mathbf{R}^d$ and any set $A \subseteq \mathbf{R}^d$, denote by $d(x, A) = \inf_{y\in A} \|x - y\|$ the distance between x and A. Let $\lambda : \mathbf{R}^d \to [0, 1]$ be any continuous function such that $\lambda(x) \to 0$ as $d(x, \mathscr{D}_2) \to 0$ and $\lambda(x) \to 1$ as $d(x, \mathscr{D}_1) \to 0$. For example, take

$\lambda(x) = \frac{2}{\pi} \arctan\left(\frac{d(x,\mathscr{D}_2)}{d(x,\mathscr{D}_1)}\right)$ with $\frac{1}{0} = \infty$ and $\arctan(\infty) = 1$ by convention. We define our final estimator as

$$\hat{\mathsf{f}}_n(x) = \lambda(x)\widetilde{f}_n(x) + (1 - \lambda(x))\widetilde{g}_n(x). \tag{5.14}$$

Theorem 6. *Let $n \geq 3$, $\beta_{\max} > 1$. Consider the grid points β_j defined as follows:*

$$\beta_j = \left(1 + \frac{1}{\log n}\right)^j, \quad j = -M, \dots, M_{\max},$$

where $M = 2\lfloor \log(n) \log\log(n) \rfloor$ and $M_{\max} = M \wedge \lfloor \log(n) \log(\beta_{\max}) \rfloor$. Let Assumptions (A1) and (A3) be satisfied. If kernel K satisfies the assumptions of Theorem 5(i), then for any $\beta \in (0, \beta_{\max}]$ and $L > 0$ for the estimator $\hat{\mathsf{f}}_n$ defined by (5.14) we have

$$\sup_{f \in \Sigma(\beta,L) \cap \mathscr{F}_0} \mathbf{E}\left[\|\hat{\mathsf{f}}_n - f\|_{L_2}^2\right] \leq C n^{-\frac{2\beta}{2\beta+d}}, \tag{5.15}$$

where $C > 0$ is a positive constant depending only on $\beta, L, L_0, d, \beta_{\max}, \sigma_\xi, K, p_{\max}, p_{\min}$ and α.

If, in addition, kernel K satisfies the assumptions of Theorem 5(ii), then the estimator $\hat{\mathsf{f}}_n$ is an interpolating continuous function with probability at least $1 - c'' \exp(-n^{\frac{2}{2+d}}/c'')$, where c'' is a positive constant depending only on $L, L_0, d, \beta_{\max}, K$, $p_{\max}, p_{\min}$ and α.

6 Numerical experiment

In this section, we report some results of our numerical experiment with singular kernel local polynomial estimators. We ran simulations with various kernels and various regression functions in dimension $d = 1$. We present below some examples of obtained results for two regression functions:

$$f(x) = x^3 - x \quad \text{and} \quad g(x) = x + \cos(3x).$$

We generated $X_1, \dots, X_n$ according to a uniform law on $[-2, 2]$ with $n = 80$. We set, for all $i \in [n]$, $Y_i = f(X_i) + \varepsilon_i$ or $Y_i = g(X_i) + \varepsilon_i$, where ε_i's are independent normal random variables with mean 0 and variance 0.5. We considered three singular kernels and the rectangular kernel:

$$\begin{aligned} K_1(u) &= |u|^{-a} \mathbf{1}(|u| \leq 1), \\ K_2(u) &= |u|^{-a} (1 - |u|)_+^2, \\ K_3(u) &= |u|^{-a} \cos^2(\pi |u|/2) \mathbf{1}(|u| \leq 1), \end{aligned}$$

$$K_{\text{rect}}(u) = \mathbf{1}(|u| \leq 1)$$

for various choices of $a \in (0, 1/2)$. Below we only present the results for $a = 0.2$.

Both f and g belong to Hölder classes with any smoothness β. We take $\beta = 8$ and we compute LP(ℓ) estimators with $\ell = 7$ and with bandwidth h chosen, for each kernel, to minimize the mean squared error (MSE) over a dense enough grid. For each singular kernel estimator, we also compute its smoothed version (named Smooth LPE), which is a result of applying the running median with a short window to the initial LPE.

The results are presented below. For comparison, we reproduce in each figure the LPE with rectangular kernel on the right hand graph. Note that K_1 is not continuous on $\mathbf{R}^d \setminus \{0\}$ and therefore does not satisfy the assumptions of Lemma 5 ensuring the interpolation property. Nevertheless, our simulations show that the corresponding LPE does interpolate the data.

The tables present the averaged MSE values for 100 simulations. We note that they are bigger for singular kernel estimators than for rectangular kernel ones but not excessively big. It supports the fact that singular kernel LPE achieves the minimax optimal rate, with probably worse constant factor than for its non-singular kernel counterparts. Reasonable MSE values for singular kernel LPE's are obtained in spite of the fact that visually they are very spiky. The best results are observed for smoothed singular kernel method that cleans out the small spikes. Finally, note that the MSE values are better for function f, which itself is a polynomial, than for function g.

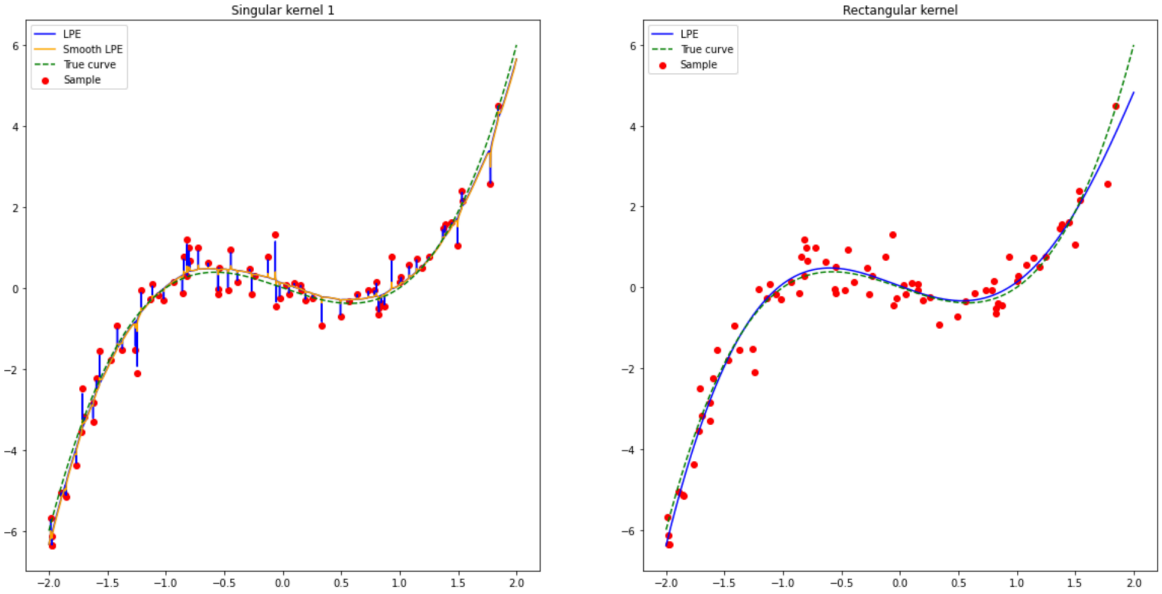


Figure 5.1: Local polynomial estimator of regression function f with singular kernel K_1 and rectangular kernel.

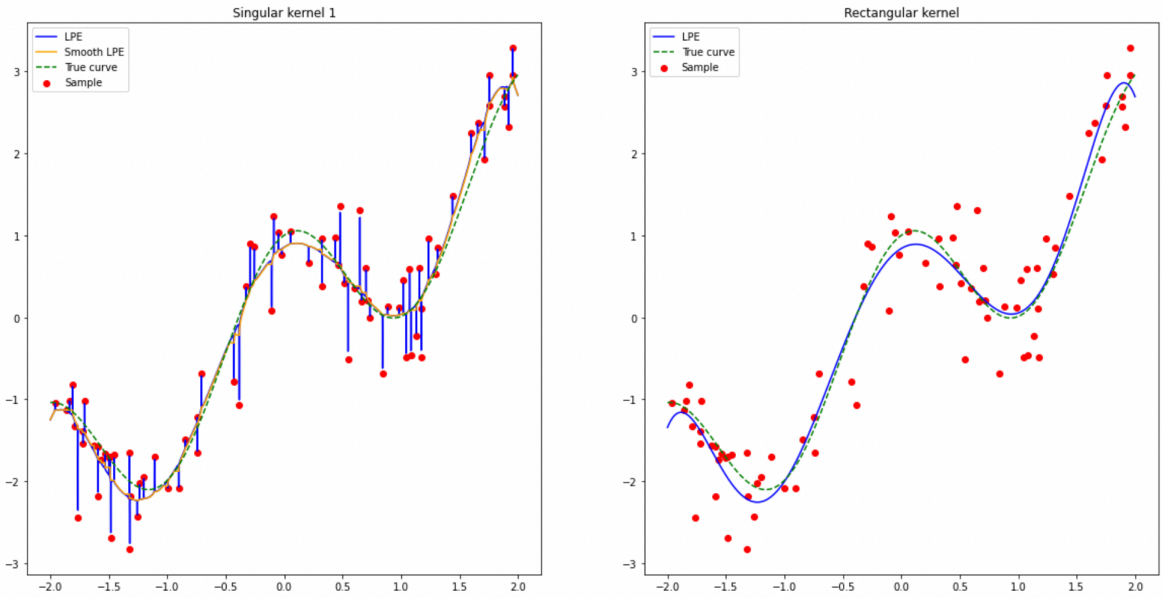


Figure 5.2: Local polynomial estimator of regression function g with singular kernel K_1 and rectangular kernel.

	Singular kernel K_1	Singular Kernel K_1 + Smooth	Rectangular kernel K_{rect}
Function f	0.0559	0.0252	0.0159
Function g	0.0584	0.0323	0.0265

Table 5.1: averaged MSE values for different kernels and functions, over 10 simulations.

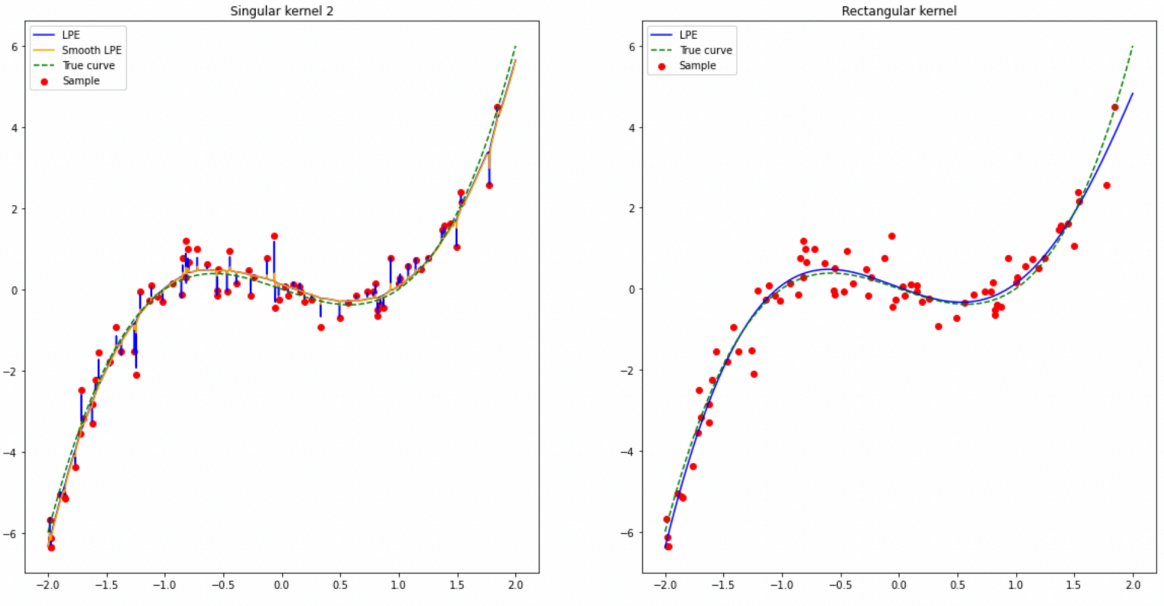


Figure 5.3: Local polynomial estimator of regression function f with singular kernel K_2 and rectangular kernel.

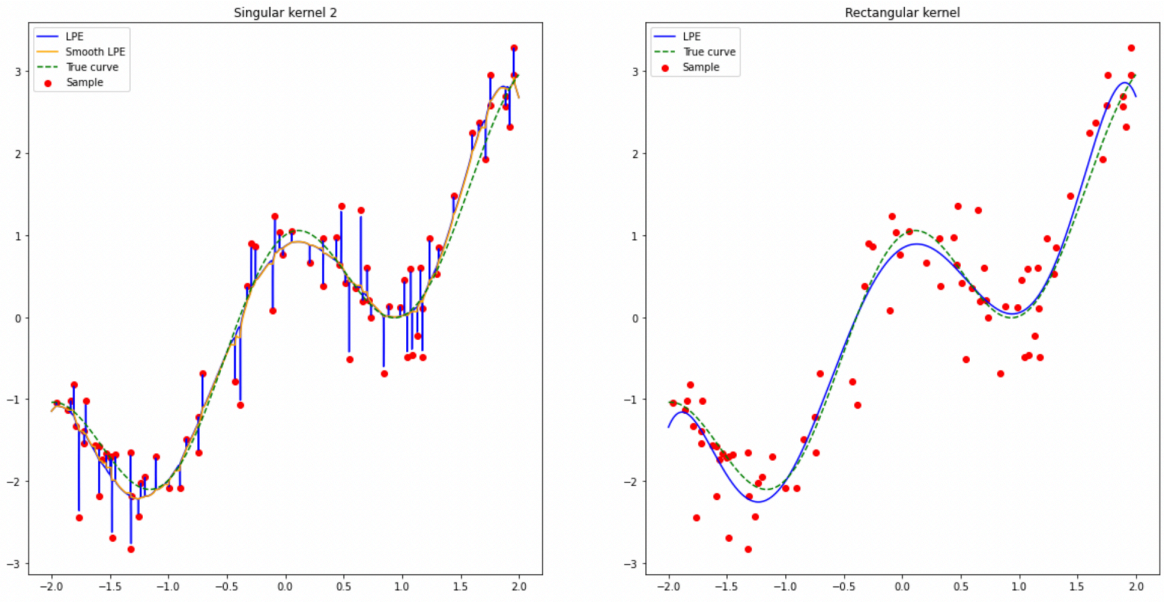


Figure 5.4: Local polynomial estimator of regression function g with singular kernel K_2 and rectangular kernel.

	Singular kernel K_2	Singular kernel K_2 + Smooth	Rectangular kernel K_{rect}
Function f	0.0675	0.0378	0.0259
Function g	0.0604	0.0328	0.0260

Table 5.2: averaged MSE values for different kernels and functions, over 10 simulations.

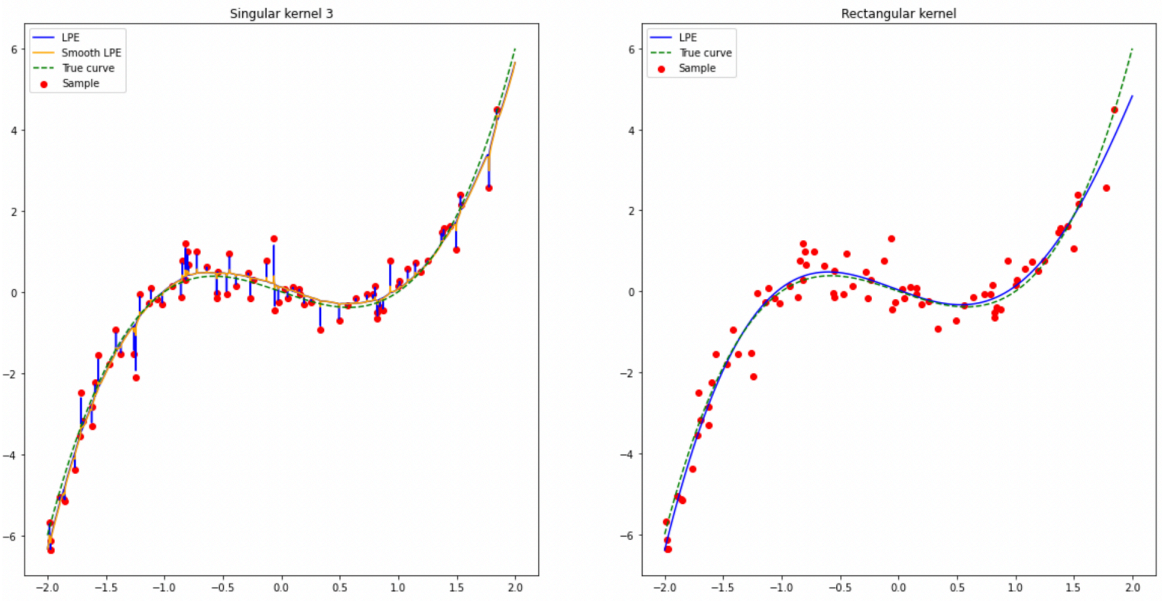


Figure 5.5: Local polynomial estimator of regression function f with singular kernel K_3 and rectangular kernel.

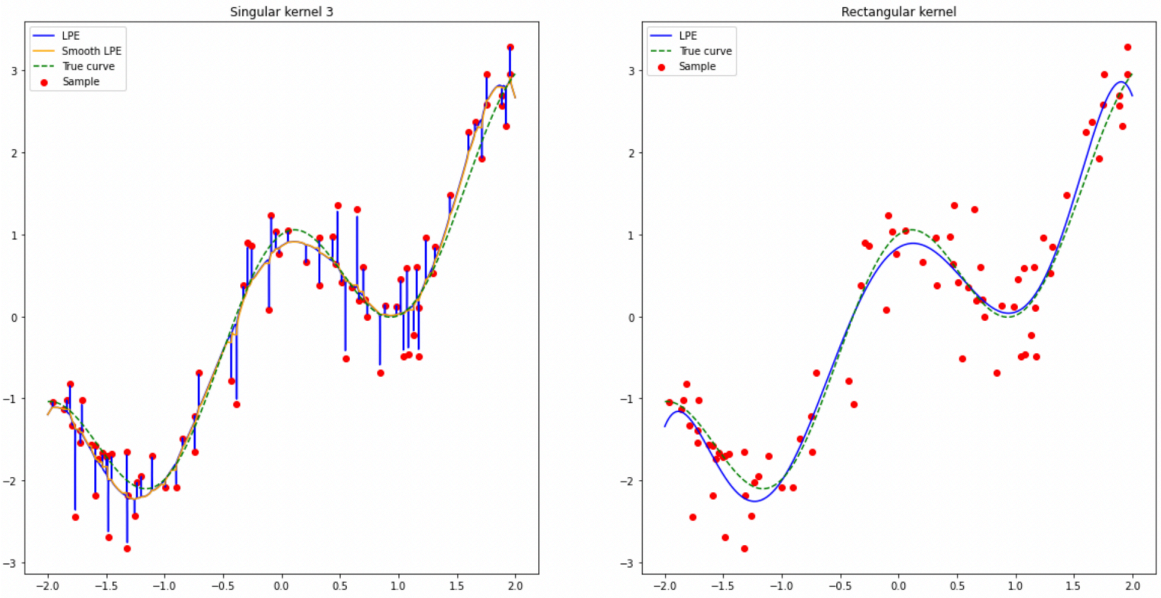


Figure 5.6: Local polynomial estimator of regression function g with singular kernel K_3 and rectangular kernel.

	Singular kernel K_3	Singular kernel K_3 + Smooth	Rectangular kernel K_{rect}
Function f	0.0640	0.0350	0.0249
Function g	0.0659	0.0385	0.0323

Table 5.3: averaged MSE values for different kernels and functions, over 10 simulations.

7 Proofs

Proof of Lemma 3. The result is straightforward if there exists an integer $\ell \geq 0$ such that $\ell < \beta' \leq \beta \leq \ell + 1$. Indeed, for any integer $\ell \geq 0$,

$$\ell < \beta' \leq \beta \leq \ell + 1 \quad \Longrightarrow \quad \Sigma(\beta, L) \subseteq \Sigma(\beta', L). \tag{5.16}$$

Thus, it remains to consider the case $\ell < \beta' \leq \ell + 1 < \beta$ for an integer ℓ. Handling this case will be based on the following embedding:

$$\Sigma(\beta, L) \subseteq \Sigma(\ell', 2L), \quad \forall \ell' \in \mathbf{N} \text{ such that } \ell' < \beta. \tag{5.17}$$

We now prove (5.17). Indeed, let $f \in \Sigma(\beta, L)$ and let ℓ' be an integer less than β. Then, in particular, $\max_{0 \leq s \leq \ell'} \sup_{x \in \mathscr{B}_d} \left\| f^{(s)}(x) \right\|_* \leq L$. Consider $x, y \in \mathscr{B}_d$ and $h = y - x$. Denote by h_i the ith component of h and by e_i the ith canonical basis vector in $\mathbf{R}^d$.

Set $k = \ell' - 1$. Then for any multi-indices $m_1, \dots, m_k \in \mathbf{N}^d$ we have

$$\begin{aligned} D^{m_1+\dots+m_k} f(y) - D^{m_1+\dots+m_k} f(x) &= \int_0^1 \Big\langle \nabla D^{m_1+\dots+m_k} f(x+th), h \Big\rangle \, dt \\ &= \int_0^1 \sum_{i=1}^d D^{m_1+\dots+m_k+e_i} f(x+th) h_i dt \\ &= \sum_{i=1}^d \int_0^1 D^{m_1+\dots+m_k+e_i} f(x+th) dt \, h^{e_i}. \end{aligned}$$

Writing for brevity $G_{m_1,\dots,m_k,e_i}(x,h) = \int_0^1 D^{m_1+\dots+m_k+e_i} f(x+th) dt$ we obtain

$$\begin{aligned} \Big\| f^{(k)}(y) - f^{(k)}(x) \Big\|_* &= \sup_{\substack{\|u_j\| \le 1, \\ j \in [k]}} \Big| \sum_{|m_j|=1, \forall j \in [k]} \sum_{i=1}^d G_{m_1,\dots,m_k,e_i}(x,h) \, h^{e_i} u_1^{m_1} \dots u_k^{m_k} \Big| \\ &= \|h\| \sup_{\substack{\|u_j\| \le 1, \\ j \in [k]}} \Big| \sum_{|m_j|=1, \forall j \in [k]} \sum_{i=1}^d G_{m_1,\dots,m_k,e_i}(x,h) \left(\frac{h}{\|h\|} \right)^{e_i} u_1^{m_1} \dots u_k^{m_k} \Big| \\ &\le \|h\| \sup_{\substack{\|u_j\| \le 1, \\ j \in [k+1]}} \Big| \sum_{|m_j|=1, \forall j \in [k+1]} \int_0^1 D^{m_1+\dots+m_{k+1}} f(x+th) dt \, u_1^{m_1} \dots u_{k+1}^{m_{k+1}} \Big| \\ &\le \|h\| \int_0^1 \sup_{\substack{\|u_j\| \le 1, \\ j \in [k+1]}} \Big| f^{(k+1)}(x+th)[u_1, \dots, u_{k+1}] \Big| dt \\ &\le \|h\| \sup_{z \in \mathscr{B}_d} \|f^{(k+1)}(z)\|_* \le L \|x-y\|, \end{aligned}$$

which, together with bound $\max_{0 \le s \le \ell'-1} \sup_{x \in \mathscr{B}_d} \Big\| f^{(s)}(x) \Big\|_* \le L$ implies that $f \in \Sigma(\ell', 2L)$. Thus, we have proved (5.17).

It follows from (5.17) that if $\ell < \beta' \le \ell + 1 < \beta$ for an integer ℓ then $\Sigma(\beta, L) \subseteq \Sigma(\ell+1, 2L)$, while taking $\beta = \ell + 1$ in (5.16) implies that $\Sigma(\ell+1, 2L) \subseteq \Sigma(\beta', 2L)$. This proves the lemma when $\ell < \beta' \le \ell + 1 < \beta$ for an integer ℓ. □

Proof of Lemma 4. The result is clear for $\beta \le 1$. Assume that $\beta > 1$ and fix some $x, y \in \mathscr{B}_d$. By Taylor expansion, there exists $c \in (0,1)$ such that

$$f(x) = \sum_{0 \le |k| \le \ell-1} \frac{1}{k!} D^k f(y)(x-y)^k + \sum_{|k|=\ell} \frac{1}{k!} D^k f(y + c(x-y))(x-y)^k,$$

and

$$\left| f(x) - \sum_{|k| \le \ell} \frac{1}{k!} D^k f(y)(x-y)^k \right| = \left| \sum_{|k|=\ell} \frac{1}{k!} \Big[D^k f\Big(y + c(x-y)\Big) - D^k f(y) \Big] (x-y)^k \right|.$$

By a standard combinatorial argument, it is not hard to check that, for any $h, z \in \mathbf{R}^d$,

$$f^{(k)}(z)[h]^k := \sum_{|m_1|=\ldots=|m_\ell|=1} D^{m_1+\ldots+m_\ell} f(z) h^{m_1+\ldots+m_\ell} = \sum_{|k|=\ell} \frac{\ell!}{k!} D^k f(z) h^k \ .$$

It follows that

$$\begin{aligned} &\left| \sum_{|k|=\ell} \frac{1}{k!} \Big[D^k f\Big(y+c(x-y)\Big) - D^k f(y) \Big] (x-y)^k \right| \qquad (5.18) \\ &\quad = \frac{1}{\ell!} \left| f^{(\ell)}\Big(y+c(x-y)\Big)[x-y]^\ell - f^{(\ell)}(y)[x-y]^\ell \right| \\ &\quad \le \frac{1}{\ell!} \left\| f^{(\ell)}\Big(y+c(x-y)\Big) - f^{(\ell)}(y) \right\|_* \|x-y\|^\ell \\ &\quad \le \frac{L}{\ell!} \|x-y\|^\ell \|c(x-y)\|^{\beta-l} \le \frac{L}{\ell!} \|x-y\|^\beta. \end{aligned}$$

□

Proof of Lemma 5. In this proof, we fix $i \in [n]$, and our aim is to prove that $\lim_{x \to X_i} f_n(x) = Y_i$. Let $\mathscr{V}$ be the neighborhood of X_i where (5.8) holds. Since $X_1, \ldots, X_n$ are distinct, we assume w.l.o.g. that $\mathscr{V}$ does not contain $(X_j)_{j \neq i}$. Due to conditions (5.7) and (5.8), we have that $B_{nx} \succ 0$ for all x in $\mathscr{V}_- := \mathscr{V} \setminus \{X_i\}$. Thus, for all $x \in \mathscr{V}_-$ the vector $\hat{\theta}_n(x)$ is the unique solution of (5.2), and $f_n(x)$ is given by (5.3):

$$\begin{aligned} \hat{\theta}_n(x) &= \operatorname*{argmin}_{\theta \in \mathbf{R}^{C_{\ell,d}}} \sum_{i=1}^n \left[Y_i - \theta^\top U\left(\frac{X_i - x}{h}\right) \right]^2 K\left(\frac{X_i - x}{h}\right), \\ f_n(x) &= U^\top(0) \hat{\theta}_n(x). \end{aligned}$$

Define $g_i(x) = \left(Y_i - \hat{\theta}_n(x)^\top U\left(\frac{X_i - x}{h}\right)\right)^2$. First, we prove by contradiction that $\lim\limits_{x \to X_i} g_i(x) = 0$ for any $i \in [n]$. Indeed, suppose that $\lim_{x \to X_i} g_i(x) \neq 0$. Then, there is a sequence $(x_k)_k$ in $\mathbf{R}^d$ converging to X_i as $k \to \infty$ such that $\lim_{k \to \infty} g_i(x_k) = +\infty$ or $\lim_{k \to \infty} g_i(x_k) = \text{const} > 0$. In both cases,

$$\lim_{k \to \infty} \sum_{j=1}^n g_j(x_k) K\left(\frac{X_j - x_k}{h}\right) = +\infty \qquad (5.19)$$

since the kernel K has a singularity at 0. On the other hand, the definition of $\hat{\theta}_n(x_k)$ implies that, for any k and any $\theta_* \in \mathbf{R}^{C_{\ell,d}}$,

$$\sum_{j=1}^n g_j(x_k) K\left(\frac{X_j - x_k}{h}\right) \le \sum_{j=1}^n \left(Y_j - \theta_*^\top U\left(\frac{X_j - x_k}{h}\right)\right)^2 K\left(\frac{X_j - x_k}{h}\right).$$

In particular, for $\theta_*^\top = (Y_i \quad 0 \dots 0)$ we have

$$\begin{aligned}\sum_{j=1}^n \left(Y_j - \theta_*^\top U\left(\frac{X_j - x_k}{h}\right)\right)^2 K\left(\frac{X_j - x_k}{h}\right) &= \sum_{j=1}^n (Y_j - Y_i)^2 K\left(\frac{X_j - x_k}{h}\right) \\ &= \sum_{j\neq i} (Y_j - Y_i)^2 K\left(\frac{X_j - x_k}{h}\right) \\ &\underset{k\to+\infty}{\to} \sum_{j\neq i} (Y_j - Y_i)^2 K\left(\frac{X_j - X_i}{h}\right) < +\infty,\end{aligned}$$

which is in contradiction with (5.19). Therefore, for any $i \in [n]$ we have $\lim\limits_{x\to X_i} g_i(x) = 0$.

A similar argument yields that $\limsup\limits_{x\to X_i} g_j(x) < +\infty$ for any $j \neq i$. Indeed, if for some $j \neq i$ this relation does not hold then there is a sequence $(x_k)_k$ in $\mathbf{R}^d$ converging to X_i as $k \to \infty$ such that $\lim_{k\to\infty} g_j(x_k) = +\infty$. It implies (5.19), which is not possible as shown above.

Next, we prove that $\|\hat\theta_n(x)\|$ is bounded for all x in a neighborhood of X_i. Since $\lim\limits_{x\to X_i} g_i(x) = 0$, and for any $j \neq i$ we have $\limsup\limits_{x\to X_i} g_j(x) < +\infty$ the values $g_j(x)$ are bounded for all $j \in [n]$ and all x in a neighborhood of X_i. We will further denote this neighborhood by $\mathscr{V}'$. It follows that $\varphi_j(x) = \hat\theta_n(x)^\top U\left(\frac{X_j - x}{h}\right)$, $j = 1, \dots, n$, are bounded for $x \in \mathscr{V}'$ and thus the sum $\sum_{j=1}^n \varphi_j^2(x)$ is bounded as well. On the other hand, by assumption (5.8), for all $x \in \mathscr{V}_-$,

$$\begin{aligned}\sum_{j=1}^n \varphi_j^2(x) &\geq \sum_{j=1}^n \hat\theta_n(x)^\top U\left(\frac{X_j - x}{h}\right) U^\top\left(\frac{X_j - x}{h}\right) \mathbf{1}\left(\left\|\frac{X_j - x}{h}\right\| \leq \Delta\right) \hat\theta_n(x) \\ &\geq \lambda_1 \|\hat\theta_n(x)\|^2,\end{aligned}$$

where $\lambda_1 > 0$. It follows that $\|\hat\theta_n(x)\|$ is bounded for all $x \in \mathscr{V}' \cap \mathscr{V}_-$.

Let $\hat\theta_{n,(1)}(x) = f_n(x)$ denote the first component of $\hat\theta_n(x)$ and $\hat\theta_{n,(2)}(x)$ the vector of its remaining $C_{\ell,d} - 1$ components, so that $\hat\theta_n(x)^\top = \left(\hat\theta_{n,(1)}(x), \hat\theta_{n,(2)}(x)^\top\right)$. Recall that the first component of $U(u)$ is equal to 1 for all $u \in \mathbf{R}^d$. Denote by $U_{(2)}(u)$ the vector of its remaining $C_{\ell,d} - 1$ components, so that $U(u)^\top = \left(1, U_{(2)}(u)^\top\right)$. With this notation, the relation $\lim\limits_{x\to X_i} g_i(x) = 0$ proved above can be written as:

$$g_i(x) = \left(Y_i - \hat\theta_{n,(1)}(x) - \hat\theta_{n,(2)}(x)^\top U_{(2)}\left(\frac{X_i - x}{h}\right)\right)^2 \underset{x\to X_i}{\to} 0.$$

Since $\|\hat\theta_n(x)\|$ is bounded for $x \in \mathscr{V}' \cap \mathscr{V}_-$ we get that $|\hat\theta_{n,(1)}(x)|$ and $\|\hat\theta_{n,(2)}(x)\|$ are also bounded for $x \in \mathscr{V}' \cap \mathscr{V}_-$. The definition of $U(u)$ implies the convergence

$\lim_{x\to X_i} \|U_{(2)}\left(\frac{X_i-x}{h}\right)\| = 0$. It follows that

$$\hat{\theta}_{n,(2)}(x)^\top U_{(2)}\left(\frac{X_i-x}{h}\right) \underset{x\to X_i}{\to} 0$$

and therefore

$$\hat{\theta}_{n,(1)}(x) \underset{x\to X_i}{\to} Y_i,$$

which concludes the proof since $\hat{\theta}_{n,(1)}(x) = f_n(x)$. □

Proof of Lemma 6. We prove only part (i) of the lemma since part (ii) is its immediate consequence. We have

$$\overline{B}_{nx} = \frac{1}{nh^d}\sum_{i=1}^n U\left(\frac{X_i-x}{h}\right)U^\top\left(\frac{X_i-x}{h}\right)\mathbf{1}\left(\frac{\|X_i-x\|}{\Delta}\le h\right)$$

and, for any $\lambda_0 > 0$,

$$\begin{aligned}\mathbf{P}\left(\inf_{x\in\mathrm{Supp}(p)}\lambda_{\min}(\overline{B}_{nx}) < \lambda_0\right) &= \mathbf{P}\left(\inf_{x\in\mathrm{Supp}(p)}\inf_{\|v\|=1} v^\top\overline{B}_{nx}v < \lambda_0\right)\\ &\le \mathbf{P}\left(\inf_{x\in\mathrm{Supp}(p)}\inf_{\|v\|=1} v^\top\overline{B}(x)v - \sup_{x\in\mathrm{Supp}(p)}\|\overline{B}_{nx}-\overline{B}(x)\|_\infty < \lambda_0\right) \qquad (5.20)\end{aligned}$$

where $\overline{B}(x) := \mathbf{E}(\overline{B}_{nx})$. Set $S(x,h,\Delta) = \{u \in \mathscr{B}_d(0,\Delta) : x+uh \in \mathrm{Supp}(p)\}$. Then we have

$$\begin{aligned}v^\top\overline{B}(x)v &= \frac{1}{h^d}\int\left[v^\top U\left(\frac{z-x}{h}\right)\right]^2\mathbf{1}\left(\left\|\frac{z-x}{h}\right\|\le\Delta\right)p(z)\mathrm{d}z\\ &\ge p_{\min}v^\top\left[\int_{S(x,h,\Delta)}U(u)U(u)^\top du\right]v\\ &\ge p_{\min}v^\top\left[\int_{S(x,\alpha,\Delta)}U(u)U(u)^\top du\right]v,\end{aligned}$$

where for the last inequality we used the fact that $S(x,\alpha,\Delta) \subset S(x,h,\Delta)$ since $h \le \alpha$ and $\mathrm{Supp}(p)$ is a convex set. Notice that $S(x,\alpha,\Delta)$ is also a convex set and it is not reduced to one point x as $\mathrm{Supp}(p)$ is a convex set with positive Lebesgue measure. Thus, $S(x,\alpha,\Delta)$ is of infinite cardinality for any $x \in \mathrm{Supp}(p)$.

Denote by $S_d(0,1)$ the unit sphere in $\mathbf{R}^d$ centered at 0. Note that, for fixed Δ and α, the function

$$\left\{\begin{array}{ccc}\mathrm{Supp}\ p\times S_d(0,1) & \longrightarrow & \mathbf{R}\\ (x,v) & \mapsto & v^\top\left[\int_{S(x,\alpha,\Delta)}U(u)U(u)^\top du\right]v\end{array}\right.$$

is continuous and defined on a compact set. Therefore, it attains its minimum at some (x_0, v_0), where $x_0 \in \mathrm{Supp}(p)$ and $\|v_0\| = 1$. We argue now that the value of this minimum is positive. Indeed, it is clearly non-negative, and if it were 0 we would have:

$$0 = v_0^\top U(u) = \sum_{|k|\le \ell} v_0(k) \frac{u^k}{k!}, \quad \forall u \in S(x_0, \alpha, \Delta). \tag{5.21}$$

As observed above, $S(x_0, \alpha, \Delta)$ is a set of infinite cardinality. On the other hand, the expression in (5.21) is a polynomial in u, so that for $v_0 \neq 0$ it can vanish only in a finite number of points. Thus, (5.21) is impossible. It follows that

$$\lambda_1(\ell) := \min_{v \in S_d(0,1), x \in \mathrm{Supp}(p)} v^\top \left[\int_{S(x,\alpha,\Delta)} U(u) U(u)^\top du \right] v > 0.$$

Next, note that the vector $U(u) = U_\ell(u)$ depends on ℓ, and that for $\ell \le \ell'$ and any fixed x, the matrix $\int_{S(x,\alpha,\Delta)} U_\ell(u) U_\ell(u)^\top du$ is an extraction of the matrix $\int_{S(x,\alpha,\Delta)} U_{\ell'}(u) U_{\ell'}(u)^\top du$. Hence, the smallest eigenvalue of the former matrix is necessarily not less than that of the latter. Thus, $\lambda_1(\ell) \ge \lambda_1(\ell')$ for $\ell \le \ell'$.

Setting $\lambda_0 = \lambda_0(\ell) := p_{\min} \lambda_1(\ell)/2$ and using (5.20) we find:

$$\mathbf{P}\left(\inf_{x \in \mathrm{Supp}(p)} \lambda_{\min}(\overline{B}_{nx}) < \lambda_0 \right) \le \mathbf{P}\left(\sup_{x \in \mathrm{Supp}(p)} \|\overline{B}_{nx} - \overline{B}(x)\|_\infty > \lambda_0 \right). \tag{5.22}$$

It remains now to bound the probability on the right hand side of (5.22).

By Assumption (A2), the convex compact set $\mathrm{Supp}(p)$ is included in $\mathscr{B}_d = \mathscr{B}_d(0,1)$. For $\varepsilon > 0$, let $\{x_1, \dots, x_N\} \subset \mathscr{B}_d^N$ be the minimal ε-net on $\mathscr{B}_d$ in the Euclidean metric. Then we have:

$$\begin{aligned} \sup_{x \in \mathrm{Supp}(p)} \|\overline{B}(x) - \overline{B}_{nx}\|_\infty \le & \sup_{x \in \mathscr{B}_d} \min_{1 \le k \le N} \|\overline{B}(x) - \overline{B}(x_k)\|_\infty \\ & + \max_{1 \le k \le N} \|\overline{B}(x_k) - \overline{B}_{nx_k}\|_\infty + \sup_{\substack{x, x' \in \mathscr{B}_d, \\ \|x - x'\| \le \varepsilon}} \|\overline{B}_{nx} - \overline{B}_{nx'}\|_\infty. \end{aligned}$$

Thus,

$$\mathbf{P}\left(\sup_{x \in \mathrm{Supp}(p)} \|\overline{B}_{nx} - \overline{B}(x)\|_\infty > \lambda_0 \right) \le P_1 + P_2 + P_3, \quad \text{where} \tag{5.23}$$

$$P_1 = \mathbf{P}\left(\sup_{x \in \mathscr{B}_d} \min_{1 \le k \le N} \|\overline{B}(x) - \overline{B}(x_k)\|_\infty > \frac{\lambda_0}{3} \right),$$

$$P_2 = \mathbf{P}\left(\max_{1 \le k \le N} \|\overline{B}(x_k) - \overline{B}_{nx_k}\|_\infty > \frac{\lambda_0}{3} \right),$$

$$P_3 = \mathbf{P}\left(\sup_{\substack{x,x' \in \mathscr{B}_d: \\ \|x-x'\| \le \varepsilon}} \|\overline{B}_{nx} - \overline{B}_{nx'}\|_\infty > \frac{\lambda_0}{3} \right).$$

In the rest of the proof, we control the terms P_1, P_2, P_3.

Control of P_2. Since all norms in the space of $C_{\ell,d} \times C_{\ell,d}$ matrices are equivalent there exists a constant $c_1 > 0$ depending only on ℓ, d such that, for all $k \in \{1, \dots, N\}$,

$$\|\overline{B}(x_k) - \overline{B}_{nx_k}\|_\infty \le c_1 \max_{1 \le i,j \le C_{\ell,d}} |b_{nx_k}(i,j) - b_{x_k}(i,j)|$$

where $b_{nx_k}(i,j)$ and $b_{x_k}(i,j)$ are the elements of $\overline{B}_{nx_k}$ and $\overline{B}(x_k)$, respectively. Then, for any $k \in \{1, \dots, N\}$,

$$\mathbf{P}\left(\|\overline{B}(x_k) - \overline{B}_{nx_k}\|_\infty > \frac{\lambda_0}{3} \right) \le C_{\ell,d}^2 \max_{1 \le i,j \le C_{\ell,d}} \mathbf{P}\left(|b_{nx_k}(i,j) - b_{x_k}(i,j)| > \frac{\lambda_0}{3c_1} \right).$$

We recall that $b_{x_k}(i,j) = \mathbf{E}\left[b_{nx_k}(i,j) \right]$. Setting $s = s^{(i)}$ and $r = s^{(j)}$ we have

$$b_{nx_k}(i,j) = \frac{1}{nh^d} \sum_{m=1}^{n} \frac{(X_m - x_k)^s}{h^s s!} \frac{(X_m - x_k)^r}{h^r r!} \mathbf{1}\left(\left\| \frac{X_m - x_k}{h} \right\| \le \Delta \right).$$

This is a sum of n i.i.d. random variables, each of which is bounded in absolute value by $\frac{C}{nh^d}$ and has variance not exceeding $\frac{C}{n^2 h^d}$, where $C > 0$ is a constant depending only on ℓ, d, Δ. By Bernstein's inequality,

$$\mathbf{P}\left(|b_{nx_k}(i,j) - b_{x_k}(i,j)| > \frac{\lambda_0}{3c_1} \right) \le 2\exp(-c_2 n h^d),$$

where $c_2 > 0$ only depends on ℓ, d, Δ and not on n, k, i, j. It follows from the above inequalities and the union bound that

$$P_2 \le 2N C_{\ell,d}^2 \exp(-c_2 n h^d). \tag{5.24}$$

Control of P_3. For any $x, x' \in \mathscr{B}_d$,

$$\begin{aligned} \overline{B}_{nx} - \overline{B}_{nx'} = \frac{1}{nh^d} \sum_{i=1}^{n} \Big[& U\left(\frac{X_i - x}{h} \right) U^\top \left(\frac{X_i - x}{h} \right) \mathbf{1}\left(\left\| \frac{X_i - x}{h} \right\| \le \Delta \right) - \\ & U\left(\frac{X_i - x'}{h} \right) U^\top \left(\frac{X_i - x'}{h} \right) \mathbf{1}\left(\left\| \frac{X_i - x'}{h} \right\| \le \Delta \right) \Big]. \end{aligned}$$

For any $u \in \mathbf{R}^d$ consider the matrix

$$V(u) = U(u) U^\top(u) \mathbf{1}\{\|u\| \le \Delta\}. \tag{5.25}$$

Notice that $U(u) \in \mathbf{R}^{C_{\ell,d}}$ is Lipschitz continuous in u on the ball $\mathscr{B}_d(0,\Delta)$ since the components of vector $U(u)$ are polynomials in u. Thus, there exists a constant $\widetilde{L} > 0$ depending only on ℓ and d such that for any $u, u' \in \mathbf{R}^d$, if either $\|u\| \le \Delta, \|u'\| \le \Delta$ or $\|u\| > \Delta, \|u'\| > \Delta$, then

$$\Big\|V(u) - V(u')\Big\|_\infty \le \widetilde{L}\|u - u'\|,$$

and if (u, u') belongs to the set

$$\widetilde{\Delta} := \{(u,u') : \|u\| \le \Delta, \|u'\| > \Delta\} \cup \{(u,u') : \|u\| > \Delta, \|u'\| \le \Delta\}$$

then

$$\Big\|V(u) - V(u')\Big\|_\infty \le \widetilde{L},$$

taking $\widetilde{L} \ge \max\limits_{\|u\|\le\Delta} \Big\|U(u)U(u)^\top\Big\|_\infty$. It follows that

$$\Big\|V(u) - V(u')\Big\|_\infty \le \widetilde{L}\Big\{\|u - u'\| + \mathbf{1}((u,u') \in \widetilde{\Delta})\Big\}, \tag{5.26}$$

which implies the bound

$$\|\overline{B}_{nx} - \overline{B}_{nx'}\|_\infty \le \frac{\widetilde{L}}{h^{d+1}}\|x - x'\| + \frac{\widetilde{L}}{nh^d}\mathrm{Card}\Big\{i \in [n] : X_i \in \widetilde{\Delta}(x, x', h\Delta)\Big\},$$

where we denote by $\widetilde{\Delta}(x, x', h\Delta)$ the symmetric difference $\mathscr{B}_d(x, h\Delta) \triangle \mathscr{B}_d(x', h\Delta)$. Thus,

$$\sup_{\substack{x,x' \in \mathscr{B}_d: \\ \|x-x'\| \le \varepsilon}} \|\overline{B}_{nx} - \overline{B}_{nx'}\|_\infty \le \frac{\widetilde{L}\varepsilon}{h^{d+1}} + \frac{\widetilde{L}}{nh^d} \sup_{\substack{x,x' \in \mathscr{B}_d: \\ \|x-x'\| \le \varepsilon}} \sum_{i=1}^n \mathbf{1}\Big(X_i \in \widetilde{\Delta}(x, x', h\Delta)\Big), \tag{5.27}$$

If $\|x - x'\| \le \varepsilon$ then

$$\widetilde{\Delta}(x, x', h\Delta) \subseteq \{z : h\Delta < \|z - x\| \le h\Delta + \varepsilon\} \cup \{z : h\Delta < \|z - x'\| \le h\Delta + \varepsilon\}.$$

Therefore, for $\|x - x'\| \le \varepsilon$ we have $\Big|\widetilde{\Delta}(x, x', h\Delta)\Big| \le C_* h^{d-1}\varepsilon$, where we denote by $|S|$ the Lebesgue measure of a measurable set $S \subset \mathbf{R}^d$, and $C_* > 0$ is a constant depending only on Δ and d. Set $\varepsilon = c_0 h^{d+1}$, where the constant c_0 satisfies $0 < c_0 \le \dfrac{\lambda_0}{6\widetilde{L}}$. Then for $\|x - x'\| \le \varepsilon$ we get $\mathbf{P}(X_1 \in \Delta(x, x', h\Delta)) \le p_{\max} C_* c_0 h^{2d}$. Choose c_0 small enough (and depending only on $\ell, d, p_{\min}p_{\max}, \Delta$) to satisfy $p_{\max} C_* c_0 \alpha^d \le \dfrac{\lambda_0}{12\widetilde{L}}$.

Consider the random event

$$\mathscr{A} = \Big\{ \sup_{\substack{x,x' \in \mathscr{B}_d: \\ \|x-x'\| \le \varepsilon}} \sum_{i=1}^{n} \mathbf{1}\Big(X_i \in \widetilde{\Delta}(x,x',h\Delta)\Big) \le A \Big\},$$

where $A = \dfrac{\lambda_0}{6\widetilde{L}} nh^d$. Due to the choice of c_0 and the fact that $h \le \alpha$ the bound $\mathbf{P}(X_1 \in \Delta(x,x',h\Delta)) \le A/2$ holds whenever $\|x-x'\| \le \varepsilon$. Hence,

$$\mathbf{P}\Big(\overline{\mathscr{A}}\Big) \le \mathbf{P}\Big\{ \sup_{\substack{x,x' \in \mathscr{B}_d: \\ \|x-x'\| \le \varepsilon}} \Big| \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\Big(X_i \in \widetilde{\Delta}(x,x',h\Delta)\Big) - \mathbf{P}\Big(X_1 \in \Delta(x,x',h\Delta)\Big) \Big| \ge A/2 \Big\}. \tag{5.28}$$

The class of all balls in $\mathbf{R}^d$ has a VC-dimension at most $d+2$, cf. Corollary 13.2 in [Devroye et al., 1996]. Consequently, the class of all intersections of two balls in $\mathbf{R}^d$ has a VC-dimension at most Cd where $C > 0$ is an absolute constant [van der Vaart and Wellner, 2009]. This allows us to apply the Vapnik-Chervonenkis inequality to bound the probability in (5.28). Indeed, we can use the decomposition

$$\begin{aligned} \mathbf{1}\Big(X_i \in \widetilde{\Delta}(x,x',h\Delta)\Big) = {} & \mathbf{1}\Big(X_i \in \mathscr{B}_d(x,h\Delta)\Big) + \mathbf{1}\Big(X_i \in \mathscr{B}_d(x',h\Delta)\Big) \\ & - 2 \cdot \mathbf{1}\Big(X_i \in \mathscr{B}_d(x,h\Delta) \cap \mathscr{B}_d(x',h\Delta)\Big) \end{aligned} \tag{5.29}$$

and bound from above the probability in (5.28) by the three probabilities corresponding to the three terms on the right hand side of (5.29). Applying the Vapnik-Chervonenkis inequality [Devroye et al., 1996, Theorem 12.5] to each of these probabilities we get

$$\mathbf{P}\Big(\overline{\mathscr{A}}\Big) \le c_3 n^{c_3} \exp(-nA^2/128) \le c_3 n^{c_3} \exp(-c_4 n^3 h^{2d}),$$

where $c_3 > 0, c_4 > 0$ are constants depending only on $d, \ell, p(\cdot), \Delta$. On the other hand, due to (5.27) and the definitions of ε and A, on the event $\mathscr{A}$ we have

$$\sup_{\substack{x,x' \in \mathscr{B}_d: \\ \|x-x'\| \le \varepsilon}} \|\overline{B}_{nx} - \overline{B}_{nx'}\|_\infty \le \frac{\lambda_0}{3}.$$

Thus, we have proved that

$$P_3 \le c_3 n^{c_3} \exp(-c_4 n^3 h^{2d}). \tag{5.30}$$

Control of P_1. Fix $x \in \mathscr{B}_d$ and let $k \in \{1,\dots,N\}$ be such that $\|x - x_k\| \le \varepsilon$. Using (5.26) we obtain

$$\|\overline{B}(x) - \overline{B}(x_k)\|_\infty \le \frac{1}{h^d} \int_{\mathbf{R}^d} \left\| V\left(\frac{z-x}{h}\right) - V\left(\frac{z-x_k}{h}\right) \right\|_\infty p(z)dz$$

$$
\begin{aligned}
&\leq \frac{\widetilde{L}}{h^d}\int_{\mathbf{R}^d}\left[\frac{\epsilon}{h}+\mathbf{1}(z\in\widetilde{\Delta}\big(x,x_k,h\Delta)\big)\right]p(z)dz\\
&\leq \widetilde{L}\varepsilon\left(\frac{1}{h^{d+1}}+\frac{C_*p_{\max}}{h}\right)\quad(\text{since }\left|\widetilde{\Delta}(x,x_k,h\Delta)\right|\leq C_*h^{d-1}\varepsilon)\\
&=\widetilde{L}c_0\left(1+C_*p_{\max}h^d\right)\leq\widetilde{L}c_0\left(1+C_*p_{\max}\alpha^d\right)<\frac{\lambda_0}{3}
\end{aligned}
$$

provided that c_0 is chosen small enough (depending only on $\ell, d, p(\cdot), \Delta, \alpha$). Thus, $P_1 = 0$ under this choice of c_0. Combining this remark with (5.22), (5.24) and (5.30) we conclude that

$$
\mathbf{P}\left(\inf_{x\in\mathrm{Supp}(p)}\lambda_{\min}(\overline{B}_{nx})<\lambda_0\right)\leq 2NC_{\ell,d}^2\exp(-c_2nh^d)+c_3n^{c_3}\exp(-c_4n^3h^{2d}).
$$

Recall that the cardinality N of the minimal ε-net on the ball $\mathscr{B}_d = \mathscr{B}_d(0,1)$ satisfies $N \leq \left(\frac{2}{\varepsilon}+1\right)^d$. The result of the lemma now follows by observing that under our choice of ε we have $N \leq Ch^{-d^2-d}$, where the constant $C > 0$ depends only on $\ell, d, p(\cdot), \Delta, \alpha$. □

In the proof of Theorem 5 below, we will use the fact that an LP(ℓ) estimator reproduces the polynomials of degree $\leq \ell$ for all $x \in \mathbf{R}^d$ such that $B_{nx} \succ 0$. We state this property in the next proposition. The proof is omitted. It follows the same lines as the proof of Proposition 1.12 in [Tsybakov, 2008] dealing with the case $d = 1$.

Proposition 4. *Let $x \in \mathbf{R}^d$ such that $B_{nx} \succ 0$ and let Q be a polynomial of degree $\leq \ell$. Then the LP(ℓ) weights W_{ni} are such that*

$$
\sum_{i=1}^n Q(X_i)W_{ni}(x) = Q(x).
$$

In particular,

$$
\sum_{i=1}^n W_{ni}(x) = 1 \text{ and } \sum_{i=1}^n (X_i - x)^k W_{ni}(x) = 0 \text{ for } |k| \leq \ell. \tag{5.31}
$$

Proof of Theorem 5. Part (ii) of the theorem follows from Corollary 1. Also, note that (5.11) is an immediate consequence of (5.10) and Assumption (A2). Therefore, we need only to prove (5.10).

Fix $x \in \mathrm{Supp}(p)$ and define the random events $\mathscr{E}_0 = \Big\{x \notin \{X_1, \dots, X_n\}\Big\}$, and

$$
\mathscr{E} = \{\lambda_{\min}(B_{nx}) \geq \lambda_0'\} \cap \mathscr{E}_0.
$$

where $\lambda_0' = \lambda_0'(\ell)$ is a constant from Lemma 6 that does not depend on n and x. From Assumption (A2) we get that $\mathbf{P}(\mathscr{E}_0) = 1$. This and Lemma 6 with our choice of h yield:

$$\mathbf{P}(\overline{\mathscr{E}}) \le c' e^{-A_n/c'}, \tag{5.32}$$

where $A_n = n^{\frac{2\beta}{2\beta+d}}$ and $c' > 0$ does not depend on x and n.

Since $|\bar{f}_n(x)| \le \mu = \max_{1\le i\le n} |Y_i| \vee L_0$ we obtain

$$\begin{aligned}\mathbf{E}\left(\left[\bar{f}_n(x) - f(x)\right]^2\right) &\le \mathbf{E}\Big(\left[\bar{f}_n(x) - f(x)\right]^2 \mathbf{1}(\mathscr{E})\Big) + \mathbf{E}\Big([L_0+\mu]^2 \mathbf{1}(\overline{\mathscr{E}})\Big) \\ &\le \mathbf{E}\Big(\left[f_n(x) - f(x)\right]^2 \mathbf{1}(\mathscr{E})\Big) + \mathbf{E}\Big([L_0+\mu]^{2+\delta}\Big)^{\frac{2}{2+\delta}} \mathbf{P}(\overline{\mathscr{E}})^{\frac{\delta}{2+\delta}},\end{aligned}$$

where we have used Hölder's inequality and the fact that $|\bar{f}_n(x) - f(x)| \le |f_n(x) - f(x)|$ for all $x \in \mathrm{Supp}(p)$. Next,

$$\mathbf{E}\Big([L_0+\mu]^{2+\delta}\Big) \le \mathbf{E}\Big([2L_0 + \max_{1\le i\le n} |\xi(X_i)|]^{2+\delta}\Big) \le C\Big[1 + n\mathbf{E}\Big(|\xi(X_1)|^{2+\delta}\Big)\Big].$$

Using this inequality and Assumption (A1) we get

$$\mathbf{E}\left(\left[\bar{f}_n(x) - f(x)\right]^2\right) \le \mathbf{E}\Big(\left[f_n(x) - f(x)\right]^2 \mathbf{1}(\mathscr{E})\Big) + Cn^{\frac{2}{2+\delta}} \mathbf{P}(\overline{\mathscr{E}})^{\frac{\delta}{2+\delta}}, \tag{5.33}$$

We now bound the main term $\mathbf{E}\Big(\left[f_n(x) - f(x)\right]^2 \mathbf{1}(\mathscr{E})\Big)$ on the right hand side of (5.33). Writing for brevity $\mathbf{E}[\cdot|X_1,\dots,X_n] = \widetilde{\mathbf{E}}[\cdot]$ we have

$$\begin{aligned}\mathbf{E}\left(\left[f_n(x) - f(x)\right]^2 \mathbf{1}(\mathscr{E})\right) &\le 2\mathbf{E}\left(\Big(f_n(x) - \widetilde{\mathbf{E}}[f_n(x)]\Big)^2 \mathbf{1}(\mathscr{E})\right) \\ &\quad + 2\mathbf{E}\left(\Big(\widetilde{\mathbf{E}}[f_n(x)] - f(x)\Big)^2 \mathbf{1}(\mathscr{E})\right).\end{aligned} \tag{5.34}$$

We analyze separately the two terms (bias and variance terms) on the right hand side of (5.34).

Bound on the variance term. On the event $\mathscr{E}$ we have

$$\widetilde{\mathbf{E}}[f_n(x)] = \sum_{i=1}^n f(X_i) W_{ni}(x),$$

where

$$W_{ni}(x) = \frac{1}{nh^d} U^\top(0) B_{nx}^{-1} U\left(\frac{X_i - x}{h}\right) K\left(\frac{X_i - x}{h}\right).$$

Thus, using Assumption (A1) the variance term can be bounded as follows:

$$\mathbf{E}\left(\Big(f_n(x) - \widetilde{\mathbf{E}}[f_n(x)]\Big)^2 \mathbf{1}(\mathscr{E})\right) = \mathbf{E}\left(\left(\sum_{i=1}^n \xi(X_i) W_{ni}(x)\right)^2 \mathbf{1}(\mathscr{E})\right)$$

$$= \mathbf{E}\left(\sum_{i=1}^{n} \mathbf{E}\left[\xi^2(X_i)\Big|X_i\right] W_{ni}^2(x)\mathbf{1}(\mathscr{E})\right) \le C\sigma^2(x),$$

where

$$\sigma^2(x) = \mathbf{E}\left(\sum_{i=1}^{n} W_{ni}^2(x)\mathbf{1}(\mathscr{E})\right).$$

In what follows, we assume w.l.o.g. that $\mathrm{Supp}(K) \subseteq \mathscr{B}_d$. On the event $\mathscr{E}$, we have $\|B_{nx}^{-1}v\| \le \|v\|/\lambda_0'$ for any $v \in \mathbf{R}^{C_{\ell,d}}$. This inequality and the fact that $\|U(0)\| = 1$ imply

$$\begin{aligned}
|W_{ni}(x)| &\le \frac{1}{nh^d}\left\|B_{nx}^{-1}U\left(\frac{X_i - x}{h}\right)K\left(\frac{X_i - x}{h}\right)\right\| \\
&\le \frac{1}{nh^d\lambda_0'}\left\|U\left(\frac{X_i - x}{h}\right)\right\| K\left(\frac{X_i - x}{h}\right) \\
&\le \frac{1}{nh^d\lambda_0'}K\left(\frac{X_i - x}{h}\right)\sqrt{\sum_{0\le|s|\le\ell}\frac{1}{(s!)^2}} \qquad \text{(since } \mathrm{Supp}(K) \subseteq \mathscr{B}_d\text{)} \\
&\le \frac{c_5}{nh^d}K\left(\frac{X_i - x}{h}\right) =: \zeta_i,
\end{aligned}$$

where $c_5 > 0$ is a constant that does not depend on n and x. Using Assumption (A2) and the compactness of the support of K we get

$$\mathbf{E}(\zeta_1^2) \le \frac{c_5^2 p_{\max}}{n^2h^d}\int K^2(u)\mathrm{d}u \le \frac{C}{n^2h^d}, \tag{5.35}$$

$$\mathbf{E}(\zeta_1) \le \frac{c_5 p_{\max}}{n}\int K(u)\mathrm{d}u \le \frac{C}{n}\left(\int K^2(u)\mathrm{d}u\right)^{1/2} \le \frac{C}{n}. \tag{5.36}$$

It follows that

$$\sigma^2(x) \le \mathbf{E}\left(\sum_{i=1}^{n}\zeta_i^2\right) \le \frac{C}{nh^d}$$

and

$$\mathbf{E}\left(\left(f_n(x) - \widetilde{\mathbf{E}}[f_n(x)]\right)^2\mathbf{1}(\mathscr{E})\right) \le \frac{C}{nh^d}. \tag{5.37}$$

Bound on the bias term. On the event $\mathscr{E}$ we have

$$\begin{aligned}
\widetilde{\mathbf{E}}[f_n(x)] - f(x) &= \sum_{i=1}^{n} f(X_i)W_{ni}(x) - f(x) \\
&= \sum_{i=1}^{n}[f(X_i) - f(x)]W_{ni}(x),
\end{aligned}$$

so that the bias term in (5.34) can be written as

$$\mathbf{E}\left(\left(\widetilde{\mathbf{E}}[f_n(x)]-f(x)\right)^2\mathbf{1}(\mathscr{E})\right)=\mathbf{E}\left(\left[\sum_{i=1}^n[f(X_i)-f(x)]W_{ni}(x)\right]^2\mathbf{1}(\mathscr{E})\right)=:b^2(x).$$

Using (5.31) and the Taylor expansion of f we get that for some $\tau_i\in[0,1]$,

$$\begin{aligned}\sum_{i=1}^n[f(X_i)-f(x)]W_{ni}(x)&=\sum_{i=1}^n\sum_{|k|=\ell}\frac{D^kf(x+\tau_i(X_i-x))}{k!}(X_i-x)^kW_{ni}(x)\\&=\sum_{i=1}^n\sum_{|k|=\ell}\frac{(D^kf(x+\tau_i(X_i-x))-D^kf(x)}{k!}(X_i-x)^kW_{ni}(x).\end{aligned}$$

Since f belongs to $\Sigma(\beta,L)$ we can apply (5.18), which yields

$$\begin{aligned}b^2(x)&\le\mathbf{E}\left[\left(\sum_{i=1}^n\frac{L}{\ell!}\|X_i-x\|^\beta|W_{ni}(x)|\right)^2\mathbf{1}(\mathscr{E})\right]\\&=\mathbf{E}\left[\left(\sum_{i=1}^n\frac{L}{\ell!}\|X_i-x\|^\beta\,|W_{ni}(x)|\,\mathbf{1}(\|X_i-x\|\le h)\right)^2\mathbf{1}(\mathscr{E})\right]\ \text{(as }\mathrm{supp}(K)\subset\mathscr{B}_d)\\&\le\mathbf{E}\left[\left(\sum_{i=1}^n\frac{L}{\ell!}h^\beta|W_{ni}(x)|\right)^2\mathbf{1}(\mathscr{E})\right].\end{aligned}$$

As $|W_{ni}(x)|\le\zeta_i$ we further get

$$\begin{aligned}b^2(x)&\le Ch^{2\beta}\mathbf{E}\left[\left(\sum_{i=1}^n\zeta_i\right)^2\right]=Ch^{2\beta}\left[\sum_{i=1}^n\mathbf{E}(\zeta_i^2)+\sum_{i\ne j}^n\mathbf{E}(\zeta_i)\mathbf{E}(\zeta_j)\right]\\&=Ch^{2\beta}\left[n\mathbf{E}(\zeta_1^2)+n(n-1)\mathbf{E}(\zeta_1)^2\right]\le Ch^{2\beta},\end{aligned}$$

where the last inequality follows from (5.35), (5.36) and the fact that $h=\alpha n^{-\frac{1}{2\beta+d}}$. Combining this bound on $b^2(x)$ with (5.32), (5.33), (5.34) and (5.37) we finally obtain

$$\mathbf{E}\left(\left[\bar f_n(x)-f(x)\right]^2\right)\le C\left(\frac{1}{nh^d}+h^{2\beta}+n^{\frac{2}{2+\delta}}e^{-n^a/C}\right),$$

where $a=\frac{2\beta}{2\beta+d}$. Since $h=\alpha n^{-\frac{1}{2\beta+d}}$ the desired bound (5.10) follows. □

Proof of Theorem 6. If K satisfies the assumptions of Theorem 5(ii) then each estimator $\bar f_{n,j}$ is interpolating on $\mathscr{D}_1$ with probability at least

$$1-C\exp(-n^{-2\beta_j/(2\beta_j+d)}/C)\ge1-C\exp(-n^{-\frac{2}{2+d}}/C)$$

if $\beta_j > 1$, and with probability 1 if $0 < \beta_j \le 1$. Hence all of them are simultaneously interpolating with probability at least

$$1 - CM_{\max}\exp(-n^{-\frac{2}{2+d}}/C) \ge 1 - C'\exp(-n^{-\frac{2}{2+d}}/C'),$$

and the same holds true for the estimator $\widetilde{f}_n$. Analogously, the estimator $\widetilde{g}_n$ is interpolating on $\mathscr{D}_2$ with the same probability. These remarks and the definition of $\hat{\mathsf{f}}_n$ in (5.14) ensure that $\hat{\mathsf{f}}_n$ is interpolating on the whole sample $\mathscr{D}$ with probability at least $1 - 2C'\exp(-n^{-\frac{2}{2+d}}/C')$.

We now prove the bound (5.15). First, we show that such a bound holds for the estimator $\widetilde{f}_n$. Set $B = L_0 + \mu$. Then $\|\bar{f}_{n,j} - f\|_\infty \le B$ for all $j = -M, \dots, M_{\max}$, where $\|\cdot\|_\infty$ denotes the L_∞-norm on $\mathrm{Supp}(p)$. Fix the subsample $\mathscr{D}_1$. Then $\bar{f}_{n,j}$'s become fixed functions, and applying Theorem 2.1 in [Wegkamp, 2003] with $a = 1$, $\lambda_j = 0$, $\forall j = -M, \dots, M_{\max}$, and $K = M + M_{\max} + 1 \le C\log^2(n)$, we get

$$\mathbf{E}_2\left[\|\widetilde{f}_n - f\|_{L_2}^2\right] \le 2\min_{-M\le j\le M_{\max}} \|\bar{f}_{n,j} - f\|_{L_2}^2 + \frac{C(B^2\log\log n + \log^2(n))}{n}, \quad (5.38)$$

where we denote by $\mathbf{E}_2$ the expectation over the distribution of the sample $\mathscr{D}_2$, and we have used the fact that $M_{\max} \le M$. Note that under Assumption (A3) we have $\mathbf{E}_1(B^2) \le C\log n$ (see, e.g., Lemma 1.6 in [Tsybakov, 2008]). Therefore, taking the expectations over $\mathscr{D}_1$ on both sides of (5.38) we get

$$\mathbf{E}_1\mathbf{E}_2\left[\|\widetilde{f}_n - f\|_{L_2}^2\right] \le 2\min_{-M\le j\le M_{\max}} \mathbf{E}_1\left[\|\bar{f}_{n,j} - f\|_{L_2}^2\right] + C\frac{\log^2(n)\log\log n}{n}. \quad (5.39)$$

Assume now that $\beta \in [\beta_j, \beta_{j+1}]$ for some $j \in \{-M, \dots, M_{\max} - 1\}$. Lemma 3 implies that $\Sigma(\beta, L) \subseteq \Sigma(\beta_j, 2L)$. Hence, using (5.13), we obtain:

$$\sup_{f\in\Sigma(\beta,L)\cap\mathscr{F}_0} \mathbf{E}_1\left[\|\bar{f}_{n,j} - f\|_{L_2}^2\right] \le \sup_{f\in\Sigma(\beta_j,2L)\cap\mathscr{F}_0} \mathbf{E}_1\left[\|\bar{f}_{n,j} - f\|_{L_2}^2\right] \le Cn^{-\frac{2\beta_j}{2\beta_j+d}}. \quad (5.40)$$

Combining (5.39) and (5.40) we get that, for $\beta \in [\beta_j, \beta_{j+1}]$,

$$\sup_{f\in\Sigma(\beta,L)\cap\mathscr{F}_0} \mathbf{E}_1\mathbf{E}_2\left[\|\widetilde{f}_n - f\|_{L_2}^2\right] \le Cn^{-\frac{2\beta_j}{2\beta_j+d}}. \quad (5.41)$$

Notice that if $\beta \in [\beta_j, \beta_{j+1}]$ for some $j \in \{-M, \dots, M_{\max} - 1\}$ then

$$n^{-\frac{2\beta_j}{2\beta_j+d}} \le en^{-\frac{2\beta}{2\beta+d}}.$$

Indeed,

$$\frac{\beta}{2\beta+d} - \frac{\beta_j}{2\beta_j+d} \le \frac{\beta_{j+1} - \beta_j}{(2\beta+d)(2\beta_j+d)} = \frac{\beta_j}{(2\beta_j+d)(2\beta+d)\log n}$$

$$\leq \frac{\beta}{(2\beta_j + d)(2\beta + d)\log n} \leq \frac{1}{2\log n}.$$

The case $\beta \in [\beta_{M_{\max}}, \beta_{\max}]$ is treated analogously. These remarks and (5.41) imply thta for each $\beta \in [\beta_{-M}, \beta_{\max}]$ there exists a constant $C > 0$ such that

$$\sup_{f \in \Sigma(\beta, L) \cap \mathscr{F}_0} \mathbf{E}\left[\|\widetilde{f}_n - f\|_{L_2}^2\right] \leq C n^{-\frac{2\beta}{2\beta+d}}. \tag{5.42}$$

Next, recalling the definition of M and β_{-M} as functions of n we note that for any fixed $\beta > 0$ it is possible to have $\beta < \beta_{-M}$ only for n not exceeding some finite number $n_0(\beta)$. For such values of n the estimation error of $\widetilde{f}_n$ is bounded by a constant depending only on β, d and L_0:

$$\mathbf{E}\Big[\|\widetilde{f}_n - f\|_{L_2}^2\Big] \leq 4\mathbf{E}_1\Big[\max_{i=1,\dots,n_0(\beta)/2} Y_i^2\Big] + 2L_0^2 \leq C(\log(n_0(\beta)) + L_0^2).$$

Consequently, (5.42) also holds for $0 < \beta < \beta_{-M}$ (and thus for all $\beta \in (0, \beta_{\max}]$) if we take the constant $C > 0$ in (5.42) large enough.

By the same argument, we deduce that the bound (5.42) holds for the estimator $\widetilde{g}_n$. Combining both bounds and using the fact that function $\lambda(\cdot)$ appearing in (5.14) takes values in $[0, 1]$ we get the desired bound (5.15) for the final estimator $\hat{\mathsf{f}}_n$. □

8 Conclusion

We have shown that local polynomial estimators with singular kernels can achieve minimax optimal rates of convergence (with respect to the mean squared risk) while perfectly interpolating the data, and moreover, can do it adaptively to the the smoothness of the regression function. This seemingly surprising conclusion is indeed not surprising at all because the mean squared risk is used as a criterion. Indeed, by adding "by hand" extremely small spikes to an accurate enough regression estimator we can always get a function interpolating the data and having a reasonably good mean squared risk. Of course, such a construction is very artificial. It makes no sense in practice and it is problematic to achieve adaptation in this way. The miracle of singular kernel LPE is to provide such an effect automatically, including adaptation, as we have outlined above. The resulting interpolating estimators have quite a reasonable behavior in terms of mean squared criterion but not in terms of visual criteria. Note that the interpolating procedures developed in different contexts in the recent literature, in particular, in deep learning are analyzed only in terms of mean squared error and expectedly share the same drawback. The difference from our setting is that, in those models, the resulting estimators are not easy to visualize, so that this sort of "spiky" behavior is not made explicit.

Chapter 6

Conclusion

To conclude this thesis, we have worked on three different problems. For the first problem, the clustering problem in the Bipartite Stochastic Block Model, we have introduced a new algorithm, the Hollowed Lloyd's algorithm, which allowed us to improve the estimation conditions of the BSBM. This algorithm, a modified version of the classical Lloyd's algorithm, is an iterative algorithm, fast and simple to implement. We have provided statistical guarantees on the result of this algorithm, improving the conditions on almost full recovery and exact recovery, in particular in the high dimensional framework. We have exhibited an oracle estimator to support the optimality of our conditions - optimality that was later proved. It remains to show whether our algorithm works with a random initialization, and whether the spectral estimator we introduced allows for exact recovery. For the second problem, the topic model problem, we have also introduced a new algorithm, the Successive Projection Overlapping Clustering algorithm. This algorithm is a modified version of the Successive Projection Algorithm, and is fast and simple to implement. To guarantee the results of this algorithm, we introduced a new anchor assumption, the document anchor assumption. We proved that the SPOC algorithm allows under this assumption the estimation of the document-topic matrix within the pLSI model. We have provided statistical guarantees on the estimation of the document-topic matrix for the Frobenius and ℓ_1 norms. Such guarantees were previously not available in the literature. Our procedure is adaptive in the number of topics. We have obtained upper and lower bounds matching up to weak factors. Finally, for the third problem, we have illustrated the Benign Overfitting phenomenon in nonparametric regression using local polynomial estimators with singular kernels. We proved that these estimators are minimax optimal on the Hölder classes of regression functions, interpolative and adaptive to the smoothness. This result questions the relevance of the classical mean square criterion since such estimators, despite their statistical optimality, are visually improper.

Bibliography

[Abbe et al., 2015] Abbe, E., Bandeira, A. S., and Hall, G. (2015). Exact recovery in the stochastic block model. *IEEE Transactions on Information Theory*, 62(1):471–487.

[Abbe et al., 2020] Abbe, E., Fan, J., and Wang, K. (2020). An ℓ_p theory of pca and spectral clustering.

[Abbe et al., 2017] Abbe, E., Fan, J., Wang, K., and Zhong, Y. (2017). Entrywise eigenvector analysis of random matrices with low expected rank. *arXiv preprint arXiv:1709.09565*.

[Alvarez et al., 2012] Alvarez, M. A., Rosasco, L., Lawrence, N. D., et al. (2012). Kernels for vector-valued functions: A review. *Foundations and Trends® in Machine Learning*, 4(3):195–266.

[Anandkumar et al., 2012] Anandkumar, A., Foster, D. P., Hsu, D. J., Kakade, S. M., and Liu, Y.-K. (2012). A spectral algorithm for latent Dirichlet allocation. In *Advances in Neural Information Processing Systems*, pages 917–925.

[Anandkumar et al., 2014] Anandkumar, A., Ge, R., Hsu, D., Kakade, S. M., and Telgarsky, M. (2014). Tensor decompositions for learning latent variable models. *Journal of Machine Learning Research*, 15:2773–2832.

[Araújo et al., 2001] Araújo, M. C. U., Saldanha, T. C. B., Galvao, R. K. H., Yoneyama, T., Chame, H. C., and Visani, V. (2001). The successive projections algorithm for variable selection in spectroscopic multicomponent analysis. *Chemometrics and Intelligent Laboratory Systems*, 57(2):65–73.

[Araujo et al., 2001] Araujo, M. C. U., Saldanha, T. C. B., Galvao, R. K. H., Yoneyama, T., Chame, H. C., and Visani, V. (2001). The successive projections algorithm for variable selection in spectroscopic multicomponent analysis. *Chemometrics and Intelligent Laboratory Systems*, 57(2):65 – 73.

[Arora et al., 2013] Arora, S., Ge, R., Halpern, Y., Mimno, D., Moitra, A., Sontag, D., Wu, Y., and Zhu, M. (2013). A practical algorithm for topic modeling with provable guarantees. In *International Conference on Machine Learning*, pages 280–288.

[Arora et al., 2016] Arora, S., Ge, R., Koehler, F., Ma, T., and Moitra, A. (2016). Provable algorithms for inference in topic models. In *International Conference on Machine Learning*, pages 2859–2867. PMLR.

[Arora et al., 2012] Arora, S., Ge, R., and Moitra, A. (2012). Learning topic models–going beyond svd. In *2012 IEEE 53rd Annual Symposium on Foundations of Computer Science*, pages 1–10. IEEE.

[Audibert and Tsybakov, 2007] Audibert, J.-Y. and Tsybakov, A. B. (2007). Fast learning rates for plug-in classifiers. *The Annals of Statistics*, 35(2):608–633.

[Azar et al., 2001] Azar, Y., Fiat, A., Karlin, A., McSherry, F., and Saia, J. (2001). Spectral analysis of data. In *Proceedings of the thirty-third annual ACM symposium on Theory of computing*, pages 619–626.

[Bandeira, 2015] Bandeira, A. S. (2015). Concentration inequalities, scalar and matrix versions. *Lectures notes*. http://math.mit.edu/~bandeira/2015_18.S096_4_Concentration_Inequalities.pdf.

[Bandeira and Van Handel, 2016] Bandeira, A. S. and Van Handel, R. (2016). Sharp nonasymptotic bounds on the norm of random matrices with independent entries. *The Annals of Probability*, 44(4):2479–2506.

[Bansal et al., 2014] Bansal, T., Bhattacharyya, C., and Kannan, R. (2014). A provable svd-based algorithm for learning topics in dominant admixture corpus. In *Advances in Neural Information Processing Systems*, pages 1997–2005.

[Bartlett and Long, 2021] Bartlett, P. L. and Long, P. M. (2021). Failures of model-dependent generalization bounds for least-norm interpolation. *Journal of Machine Learning Research*, 22(204):1–15.

[Bartlett et al., 2020] Bartlett, P. L., Long, P. M., Lugosi, G., and Tsigler, A. (2020). Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070.

[Belkin, 2021] Belkin, M. (2021). Fit without fear: Remarkable mathematical phenomena of deep learning through the prism of interpolation. *Acta Numerica*, 30:203–248.

[Belkin et al., 2019a] Belkin, M., Hsu, D., Ma, S., and Mandal, S. (2019a). Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854.

[Belkin et al., 2018a] Belkin, M., Hsu, D. J., and Mitra, P. (2018a). Overfitting or perfect fitting? Risk bounds for classification and regression rules that interpolate. *Advances in Neural Information Processing Systems*, 31.

[Belkin et al., 2018b] Belkin, M., Ma, S., and Mandal, S. (2018b). To understand deep learning we need to understand kernel learning. In *International Conference on Machine Learning*, pages 541–549. PMLR.

[Belkin et al., 2019b] Belkin, M., Rakhlin, A., and Tsybakov, A. B. (2019b). Does data interpolation contradict statistical optimality? In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1611–1619. PMLR.

[Beygelzimer et al., 2015] Beygelzimer, A., Hazan, E., Kale, S., and Luo, H. (2015). Online gradient boosting. *Advances in neural information processing systems*, 28.

[Bicego et al., 2012] Bicego, M., Lovato, P., Perina, A., Fasoli, M., Delledonne, M., Pezzotti, M., Polverari, A., and Murino, V. (2012). Investigating topic models' capabilities in expression microarray data classification. *IEEE/ACM transactions on computational biology and bioinformatics*, 9(6):1831–1836.

[Bing et al., 2021a] Bing, X., Bunea, F., Strimas-Mackey, S., and Wegkamp, M. (2021a). Likelihood estimation of sparse topic distributions in topic models and its applications to wasserstein document distance calculations. *arXiv preprint arXiv:2107.05766*.

[Bing et al., 2021b] Bing, X., Bunea, F., Strimas-Mackey, S., and Wegkamp, M. (2021b). Likelihood estimation of sparse topic distributions in topic models and its applications to Wasserstein document distance calculations. *arXiv preprint arXiv:2107.05766*.

[Bing et al., 2020a] Bing, X., Bunea, F., and Wegkamp, M. (2020a). A fast algorithm with minimax optimal guarantees for topic models with an unknown number of topics. *Bernoulli*, 26(3):1765–1796.

[Bing et al., 2020b] Bing, X., Bunea, F., and Wegkamp, M. (2020b). A fast algorithm with minimax optimal guarantees for topic models with an unknown number of topics. *Bernoulli 26(3): 1765-1796*.

[Bing et al., 2020c] Bing, X., Bunea, F., and Wegkamp, M. (2020c). Optimal estimation of sparse topic models. *Journal of machine learning research*, 21(177).

[Bing et al., 2020d] Bing, X., Bunea, F., and Wegkamp, M. (2020d). Optimal estimation of sparse topic models. *J. Mach. Learn. Res.*, 21(1).

[Blei and Lafferty, 2006] Blei, D. M. and Lafferty, J. D. (2006). Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, pages 113–120.

[Blei and Lafferty, 2007] Blei, D. M. and Lafferty, J. D. (2007). A correlated topic model of science. *The Annals of Applied Statistics*, 1(1):17–35.

[Blei et al., 2003] Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

[Bordenave et al., 2015] Bordenave, C., Lelarge, M., and Massoulié, L. (2015). Non-backtracking spectrum of random graphs: community detection and non-regular ramanujan graphs. In *2015 IEEE 56th Annual Symposium on Foundations of Computer Science*, pages 1347–1357. IEEE.

[Cai et al., 2019] Cai, C., Li, G., Chi, Y., Poor, H. V., and Chen, Y. (2019). Subspace estimation from unbalanced and incomplete data matrices: $\ell_{2,\infty}$ statistical guarantees. *arXiv preprint arXiv:1910.04267*.

[Chhor et al., 2022] Chhor, J., Sigalla, S., and Tsybakov, A. B. (2022). Benign overfitting and adaptive nonparametric regression. *arXiv preprint arXiv:2206.13347*.

[Chien and Chueh, 2010] Chien, J.-T. and Chueh, C.-H. (2010). Dirichlet class language models for speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(3):482–495.

[Chinot and Lerasle, 2020] Chinot, G. and Lerasle, M. (2020). On the robustness of the minimum ℓ_2 interpolator. *arXiv preprint arXiv:2003.05838*.

[Cichocki et al., 2009] Cichocki, A., Zdunek, R., Phan, A. H., and ichi Amari, S. (2009). *Nonnegative Matrix and Tensor Factorizations - Applications to Exploratory Multi-way Data Analysis and Blind Source Separation.* Wiley.

[Curiskis et al., 2020] Curiskis, S. A., Drake, B., Osborn, T. R., and Kennedy, P. J. (2020). An evaluation of document clustering and topic modelling in two online social networks: Twitter and reddit. *Information Processing & Management*, 57(2):102034.

[Devroye et al., 1998] Devroye, L., Györfi, L., and Krzyżak, A. (1998). The Hilbert kernel regression estimate. *Journal of Multivariate Analysis*, 65(2):209–227.

[Devroye et al., 1996] Devroye, L., Györfi, L., and Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition*. Springer, NY e.a.

[Dhillon, 2001] Dhillon, I. S. (2001). Co-clustering documents and words using bipartite spectral graph partitioning. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 269–274.

[Dhillon and Modha, 2001] Dhillon, I. S. and Modha, D. S. (2001). Concept decompositions for large sparse text data using clustering. *Machine learning*, 42(1-2):143–175.

[Ding et al., 2013] Ding, W., Rohban, M. H., Ishwar, P., and Saligrama, V. (2013). Topic discovery through data dependent and random projections. In Dasgupta, S. and McAllester, D., editors, *Proceedings of the 30th International Conference*

on Machine Learning, volume 28(3) of *Proceedings of Machine Learning Research*, pages 1202–1210, Atlanta, Georgia, USA. PMLR.

[Donoho and Stodden, 2004] Donoho, D. and Stodden, V. (2004). When does non-negative matrix factorization give a correct decomposition into parts? In Thrun, S., Saul, L., and Schölkopf, B., editors, *Advances in Neural Information Processing Systems*, volume 16. MIT Press.

[Erdős and Rényi, 1959] Erdős, P. and Rényi, A. (1959). On random graphs i. *Publicationes mathematicae*, 6(1):290–297.

[Erdős et al., 1960] Erdős, P., Rényi, A., et al. (1960). On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci*, 5(1):17–60.

[Eren et al., 2013] Eren, K., Deveci, M., Küçüktunç, O., and Çatalyürek, Ü. V. (2013). A comparative analysis of biclustering algorithms for gene expression data. *Briefings in bioinformatics*, 14(3):279–292.

[Fan and Gijbels, 1996] Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and its Applications*. Chapman and Hall, NY.

[Feldman et al., 2015] Feldman, V., Perkins, W., and Vempala, S. (2015). Subsampled power iteration: a unified algorithm for block models and planted csp's. In *Advances in Neural Information Processing Systems*, pages 2836–2844.

[Feldman et al., 2018] Feldman, V., Perkins, W., and Vempala, S. (2018). On the complexity of random satisfiability problems with planted solutions. *SIAM Journal on Computing*, 47(4):1294–1338.

[Florescu and Perkins, 2016] Florescu, L. and Perkins, W. (2016). Spectral thresholds in the bipartite stochastic block model. In *Conference on Learning Theory*, pages 943–959.

[Gillis and Vavasis, 2014] Gillis, N. and Vavasis, S. A. (2014). Fast and robust recursive algorithmsfor separable nonnegative matrix factorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(4):698–714.

[Gillis and Vavasis, 2015] Gillis, N. and Vavasis, S. A. (2015). Semidefinite programming based preconditioning for more robust near-separable nonnegative matrix factorization. *SIAM Journal on Optimization*, 25(1):677–698.

[Giraud, 2015] Giraud, C. (2015). *Introduction to high-dimensional statistics*. Chapman and Hall.

[Giraud and Verzelen, 2019] Giraud, C. and Verzelen, N. (2019). Partial recovery bounds for clustering with the relaxed k-means. *Mathematical Statistics and Learning*, 1(3):317–374.

[Golub et al., 1979] Golub, G. H., Heath, M., and Wahba, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21(2):215–223.

[Gupta et al., 2015] Gupta, S., Agrawal, A., Gopalakrishnan, K., and Narayanan, P. (2015). Deep learning with limited numerical precision. In *International conference on machine learning*, pages 1737–1746. PMLR.

[Harman, 1993] Harman, D. (1993). Overview of the first trec conference. In *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '93, page 3647, New York, NY, USA. Association for Computing Machinery.

[Hofmann, 1999] Hofmann, T. (1999). Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 50–57.

[Holland et al., 1983] Holland, P. W., Laskey, K. B., and Leinhardt, S. (1983). Stochastic blockmodels: First steps. *Social networks*, 5(2):109–137.

[Huang et al., 2017] Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708.

[Jang et al., 2007] Jang, S. W., Kim, S., and Ha, J. (2007). Graph-based recommendation systems: Comparison analysis between traditional clustering techniques and neural embedding.

[Katkovnik, 1985] Katkovnik, V. Y. (1985). *Nonparametric Identification and Data Smoothing*. Nauka, Moscow (in Russian).

[Ke and Wang, 2017] Ke, Z. T. and Wang, M. (2017). A new svd approach to optimal topic estimation. *arXiv preprint arXiv:1704.07016v1*.

[Klopp et al., 2021] Klopp, O., Panov, M., Sigalla, S., and Tsybakov, A. (2021). Assigning topics to documents by successive projections. *arXiv preprint arXiv:2107.03684*.

[Lafferty and Blei, 2006] Lafferty, J. D. and Blei, D. M. (2006). Correlated topic models. In *Advances in neural information processing systems*, pages 147–154.

[Lancaster and Salkauskas, 1981] Lancaster, P. and Salkauskas, K. (1981). Surfaces generated by moving least squares methods. *Mathematics of Computation*, 37(155):141–158.

[Lancichinetti et al., 2014] Lancichinetti, A., Sirer, M. I., Wang, J. X., Acuna, D., Körding, K., and Amaral, L. A. N. (2014). A high-reproducibility and high-accuracy method for automated topic classification. *arXiv preprint arXiv:1402.0422*.

[Larremore et al., 2013] Larremore, D. B., Clauset, A., and Buckee, C. O. (2013). A network approach to analyzing highly recombinant malaria parasite genes. *PLoS Comput Biol*, 9(10):e1003268.

[Lecué and Shang, 2022] Lecué, G. and Shang, Z. (2022). A geometrical viewpoint on the benign overfitting property of the minimum l_2-norm interpolant estimator. *arXiv preprint arXiv:2203.05873*.

[Lee and Seung, 2000] Lee, D. and Seung, H. S. (2000). Algorithms for non-negative matrix factorization. *Advances in neural information processing systems*, 13.

[Lee and Seung, 1999a] Lee, D. D. and Seung, H. S. (1999a). Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791.

[Lee and Seung, 1999b] Lee, D. D. and Seung, H. S. (1999b). Learning the parts of objects by nonnegative matrix factorization. *Nature*, 401:788–791.

[Lee et al., 2015] Lee, M., Bindel, D., and Mimno, D. (2015). Robust spectral inference for joint stochastic matrix factorization. *Advances in neural information processing systems*, 28.

[Lee et al., 2020] Lee, M., Bindel, D., and Mimno, D. (2020). Prior-aware composition inference for spectral topic models. In *International Conference on Artificial Intelligence and Statistics*, pages 4258–4268. PMLR.

[Lee et al., 2019] Lee, M., Cho, S., Bindel, D., and Mimno, D. (2019). Practical correlated topic modeling and analysis via the rectified anchor word algorithm. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4991–5001.

[Lei and Rinaldo, 2015] Lei, J. and Rinaldo, A. (2015). Consistency of spectral clustering in stochastic block models. *Ann. Statist.*, 43(1):215–237.

[Li et al., 2015] Li, J., Rabani, Y., Schulman, L. J., and Swamy, C. (2015). Learning arbitrary statistical mixtures of discrete distributions. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pages 743–752.

[Li et al., 2010] Li, L.-J., Wang, C., Lim, Y., Blei, D. M., and Fei-Fei, L. (2010). Building and using a semantivisual image hierarchy. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3336–3343. IEEE.

[Li and McCallum, 2006] Li, W. and McCallum, A. (2006). Pachinko allocation: Dag-structured mixture models of topic correlations. In *Proceedings of the 23rd international conference on Machine learning*, pages 577–584.

[Liang and Rakhlin, 2020] Liang, T. and Rakhlin, A. (2020). Just interpolate: Kernel ridgeless regression can generalize. *The Annals of Statistics*, 48(3):1329–1347.

[Liang et al., 2020] Liang, T., Rakhlin, A., and Zhai, X. (2020). On the multiple descent of minimum-norm interpolants and restricted lower isometry of kernels. In *Conference on Learning Theory*, pages 2683–2711. PMLR.

[Liu et al., 1994] Liu, K., Tokar, R., and McVey, B. (1994). An integrated architecture of adaptive neural network control for dynamic systems. *Advances in neural information processing systems*, 7.

[Lloyd, 1982] Lloyd, S. (1982). Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137.

[Loffler et al., 2019] Loffler, M., Zhang, A. Y., and Zhou, H. H. (2019). Optimality of spectral clustering for gaussian mixture model. *arXiv preprint arXiv:1911.00538*.

[Lu and Zhou, 2016] Lu, Y. and Zhou, H. H. (2016). Statistical and computational guarantees of lloyd's algorithm and its variants. *arXiv preprint arXiv:1612.02099*.

[Ma et al., 2018] Ma, S., Bassily, R., and Belkin, M. (2018). The power of interpolation: Understanding the effectiveness of sgd in modern over-parametrized learning. In *International Conference on Machine Learning*, pages 3325–3334. PMLR.

[Mao et al., 2018] Mao, X., Sarkar, P., and Chakrabarti, D. (2018). Overlapping clustering models, and one (class) svm to bind them all. In *Advances in Neural Information Processing Systems*, pages 2126–2136.

[Mao et al., 2020] Mao, X., Sarkar, P., and Chakrabarti, D. (2020). Estimating mixed memberships with sharp eigenvector deviations. *Journal of the American Statistical Association*.

[Massoulié, 2014] Massoulié, L. (2014). Community detection thresholds and the weak ramanujan property. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, pages 694–703.

[McCallum et al., 2005] McCallum, A., Corrada-Emmanuel, A., and Wang, X. (2005). The author-recipient-topic model for topic and role discovery in social networks: Experiments with enron and academic email. *Computer Science Department Faculty Publication Series*, page 44.

[Mizutani, 2014] Mizutani, T. (2014). Ellipsoidal rounding for nonnegative matrix factorization under noisy separability. *Journal of Machine Learning Research*, 15:1011–1039.

[Mizutani, 2016] Mizutani, T. (2016). Robustness analysis of preconditioned successive projection algorithm for general form of separable NMF problem. *Linear Algebra and its Applications*, 497:1 – 22.

[Mossel et al., 2018] Mossel, E., Neeman, J., and Sly, A. (2018). A proof of the block model threshold conjecture. *Combinatorica*, 38(3):665–708.

[Muthukumar et al., 2020] Muthukumar, V., Vodrahalli, K., Subramanian, V., and Sahai, A. (2020). Harmless interpolation of noisy data in regression. *IEEE Journal on Selected Areas in Information Theory*, 1(1):67–83.

[Ndaoud, 2018] Ndaoud, M. (2018). Sharp optimal recovery in the two component gaussian mixture model. *arXiv preprint arXiv:1812.08078*.

[Ndaoud et al., 2021] Ndaoud, M., Sigalla, S., and Tsybakov, A. B. (2021). Improved clustering algorithms for the bipartite stochastic block model. *IEEE Transactions on Information Theory*, 68(3):1960–1975.

[Neumann, 2018] Neumann, S. (2018). Bipartite stochastic block models with tiny clusters. In *Advances in Neural Information Processing Systems*, pages 3867–3877.

[Paatero, 1997] Paatero, P. (1997). Least squares formulation of robust non-negative factor analysis. *Chemometrics and intelligent laboratory systems*, 37(1):23–35.

[Paatero and Tapper, 1994] Paatero, P. and Tapper, U. (1994). Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics*, 5(2):111–126.

[Palese and Usai, 2018] Palese, B. and Usai, A. (2018). The relative importance of service quality dimensions in e-commerce experiences. *International Journal of Information Management*, 40:132–140.

[Panov et al., 2017] Panov, M., Slavnov, K., and Ushakov, R. (2017). Consistent estimation of mixed memberships with successive projections. In *International Conference on Complex Networks and their Applications*, pages 53–64. Springer.

[Park et al., 2013] Park, I. M., Archer, E. W., Latimer, K., and Pillow, J. W. (2013). Universal models for binary spike patterns using centered dirichlet processes. *Advances in neural information processing systems*, 26.

[Pedregosa et al., 2011] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

[Perrone et al., 2017] Perrone, V., Jenkins, P. A., Spano, D., and Teh, Y. W. (2017). Poisson random fields for dynamic feature models. *Journal of Machine Learning Research (JMLR)*.

[Porteous et al., 2008] Porteous, I., Newman, D., Ihler, A., Asuncion, A., Smyth, P., and Welling, M. (2008). Fast collapsed gibbs sampling for latent Dirichlet allocation. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 569–577.

[Pritchard et al., 2000] Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959.

[Rakhlin and Zhai, 2019] Rakhlin, A. and Zhai, X. (2019). Consistency of interpolation with Laplace kernels is a high-dimensional phenomenon. In *Conference on Learning Theory*, pages 2595–2623. PMLR.

[Ramage et al., 2009] Ramage, D., Hall, D., Nallapati, R., and Manning, C. D. (2009). Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 conference on empirical methods in natural language processing*, pages 248–256.

[Recht et al., 2012] Recht, B., Re, C., Tropp, J., and Bittorf, V. (2012). Factoring nonnegative matrices with linear programs. In Pereira, F., Burges, C. J. C., Bottou, L., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc.

[Royer, 2017] Royer, M. (2017). Adaptive clustering through semidefinite programming. In *Advances in Neural Information Processing Systems*, pages 1795–1803.

[Sagun et al., 2017] Sagun, L., Evci, U., Guney, V. U., Dauphin, Y., and Bottou, L. (2017). Empirical analysis of the hessian of over-parametrized neural networks. *arXiv preprint arXiv:1706.04454*.

[Shepard, 1968] Shepard, D. (1968). A two-dimensional interpolation function for irregularly-spaced data. In *Proceedings of the 1968 23rd ACM national conference*, pages 517–524. ACM.

[Silge and Robinson, 2020] Silge, J. and Robinson, D. (2020). *Text Mining with R: A Tidy Approach*.

[Smola and Schölkopf, 1998] Smola, A. J. and Schölkopf, B. (1998). *Learning with kernels*, volume 4. Citeseer.

[Stone, 1980] Stone, C. J. (1980). Optimal rates of convergence for nonparametric estimators. *The Annals of Statistics*, 8:1348–1360.

[Stone, 1982] Stone, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics*, 10:1040–1053.

[Teh et al., 2005] Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2005). Sharing clusters among related groups: Hierarchical Dirichlet processes. In *Advances in neural information processing systems*, pages 1385–1392.

[Tropp, 2012] Tropp, J. A. (2012). User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12(4):389–434.

[Tropp, 2015] Tropp, J. A. (2015). An introduction to matrix concentration inequalities. *Foundations and Trends in Machine Learning*, 8(1–2):1–230.

[Tsigler and Bartlett, 2020] Tsigler, A. and Bartlett, P. L. (2020). Benign overfitting in ridge regression. *arXiv preprint arXiv:2009.14286*.

[Tsybakov, 1986] Tsybakov, A. B. (1986). Robust reconstruction of functions by the local-approximation method. *Problems of Information Transmission*, 22:133–146.

[Tsybakov, 2008] Tsybakov, A. B. (2008). *Introduction to nonparametric estimation*. Springer.

[van der Vaart and Wellner, 2009] van der Vaart, A. and Wellner, J. A. (2009). A note on bounds for VC dimensions. In *High Dimensional Probability*, volume 5, pages 103–107. IMS Collections.

[Vershynin, 2018] Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge University Press.

[Wallach, 2006] Wallach, H. M. (2006). Topic modeling: beyond bag-of-words. In *Proceedings of the 23rd international conference on Machine learning*, pages 977–984.

[Wang et al., 2022] Wang, G., Donhauser, K., and Yang, F. (2022). Tight bounds for minimum ℓ_1-norm interpolation of noisy data. In *International Conference on Artificial Intelligence and Statistics*, pages 10572–10602. PMLR.

[Wegkamp, 2003] Wegkamp, M. (2003). Model selection in nonparametric regression. *The Annals of Statistics*, 31(1):252–273.

[Yuan et al., 2018] Yuan, H., Xu, W., Li, Q., and Lau, R. (2018). Topic sentiment mining for sales performance prediction in e-commerce. *Annals of Operations Research*, 270(1-2):553–576.

[Zhai et al., 2012] Zhai, K., Boyd-Graber, J., Asadi, N., and Alkhouja, M. L. (2012). Mr. lda: A flexible large scale topic modeling package using variational inference in mapreduce. In *Proceedings of the 21st international conference on World Wide Web*, pages 879–888.

[Zhang et al., 2021] Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. (2021). Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115.

[Zhou and Amini, 2019] Zhou, Z. and Amini, A. A. (2019). Analysis of spectral clustering algorithms for community detection: the general bipartite setting. *J. Mach. Learn. Res.*, 20:47–1.

[Zhou and Amini, 2020] Zhou, Z. and Amini, A. A. (2020). Optimal bipartite network clustering. *Journal of Machine Learning Research*, 21(40):1–68.

[Zhu et al., 2017] Zhu, Q., Zhong, Y., Zhang, L., and Li, D. (2017). Scene classification based on the fully sparse semantic topic model. *IEEE Transactions on Geoscience and Remote Sensing*, 55(10):5525–5538.

ECOLE
DOCTORALE
DE MATHEMATIQUES
HADAMARD

Titre : Contributions à l'inférence en grande dimension structurée

Mots clefs : Bipartite stochastic block model, topic model, benign overfitting, régression non-paramétrique, statistique en grande dimension, estimation adaptative

Résumé : Dans cette thèse, nous considérons les trois problèmes suivants : le problème de clustering dans le Bipartite Stochastic Block Model, le problème de classification de documents dans le cadre des topic models, et le problème de benign overfitting dans le cadre de régression non paramétrique. Tout d'abord, nous considérons le problème de clustering dans le Bipartite Stochastic Block Model (BSBM). Le BSBM est une généralisation non symétrique du Stochastic Block Model, avec deux ensembles de sommets. Nous introduisons un algorithme appelé le *Hollowed Lloyd's algorithm*, qui permet de classer les sommets du plus petit ensemble avec grande probabilité. Nous fournissons des garanties statistiques sur cet algorithme, qui est rapide et simple à implémenter. Nous établissons une condition suffisante pour le clustering dans le BSBM. Nos résultats améliorent les travaux précédents sur le BSBM, en particulier dans le cadre de grande dimension. Deuxièmement, nous étudions le problème de la classification de documents dans le cadre des topic models. Les topic models permettent d'exploiter des structures sous-jacentes dans un grand corpus de documents et ainsi de réduire la dimension du problème considéré. Chaque topic est vu comme une distribution de probabilité sur le dictionnaire de mots du corpus, et chaque document est vu comme un mélange de topics. Nous introduisons un algorithme appelé *Successive Projection Overlapping Clustering* (SPOC), inspiré du Successive Projection Algorithm pour le problème de Nonnegative Matrix Factorization. L'algorithme SPOC est rapide et simple à implémenter. Nous fournissons des garanties statistiques sur le résultat de l'algorithme SPOC. En particulier, nous fournissons des bornes minimax inférieures et supérieures sur son risque d'estimation pour les normes de Frobenius et ℓ_1, bornes correspondant à de faibles facteurs près. Notre procédure de clustering est adaptative en le nombre de topics. Enfin, le troisième problème étudié lors de cette thèse porte sur la régression non paramétrique. Nous considérons des estimateurs par polynômes locaux avec des noyaux singuliers. Nous prouvons que ces estimateurs sont minimax optimaux, adaptatifs en la régularité et interpolants avec une probabilité élevée. Cette propriété est appelée benign overfitting.

Title : Contributions to structured high-dimensional inference

Keywords : Bipartite stochastic block model, topic model, benign overfitting, nonparametric regression, high-dimensional statistics, adaptive estimation

Abstract : In this thesis, we consider the three following problems: clustering in Bipartite Stochastic Block Model, estimation of topic-document matrix in topic model, and benign overfitting in nonparametric regression. First, we consider the graph clustering problem in the Bipartite Stochastic Block Model (BSBM). The BSBM is a non-symmetric generalization of the Stochastic Block Model, with two sets of vertices. We provide an algorithm called the *Hollowed Lloyd's algorithm*, which allows one to classify vertices of the smallest set with high probability. We provide statistical guarantees on this algorithm, which is computationnally fast and simple to implement. We establish a sufficient condition for clustering in BSBM. Our results improve on previous works on BSBM, in particular in the high-dimensional regime. Second, we study the problem of assigning topics to documents using topic models. Topic models allow one to discover hidden structures in a large corpus of documents through dimension reduction. Each topic is considered as a probability distribution on the dictionary of words, and each document is considered as a mixture of topics. We introduce an algotihm called the *Successive Projection Overlapping Clustering* (SPOC) algorithm, inspired by the Successive Projection Algorithm for Non-negative Matrix Factorization. The SPOC algorithm is computationnally fast and simple to implement. We provide statistical guarantees on the outcome of the algorithm. In particular, we provide near matching minimax upper and lower bounds on its estimation risk under the Frobenius and the ℓ_1-norm. Our clustering procedure is adaptive in the number of topics. Finally, the third problem we study is a nonparametric regression problem. We consider local polynomial estimators with singular kernel, which we prove to be minimax optimal, adaptive to unknown smoothness, and interpolating with high probability. This property is called benign overfitting.

Institut Polytechnique de Paris
91120 Palaiseau, France