Journal of Machine Learning Research 22 (2021) 1-48 Submitted 1/19; Revised 2/21; Published 3/21

Consistent Semi-Supervised Graph Regularization for High Dimensional Data

Xiaoyi Mai[1] XIAOYI.MAI@L2S.CENTRALESUPELEC.FR
Romain Couillet[1,2] ROMAIN.COUILLET@CENTRALESUPELEC.FR
[1] *CentraleSupélec, Laboratoire des Signaux et Systèmes*
Université Paris-Saclay
3 rue Joliot Curie, 91192 Gif-Sur-Yvette

[2] *GIPSA-lab, GSTATS DataScience Chair*
Université Grenoble–Alpes
11 rue des Mathématiques, 38400 St Martin d'Hères.

Editor: Rina Foygel Barber

Abstract

Semi-supervised Laplacian regularization, a standard graph-based approach for learning from both labelled and unlabelled data, was recently demonstrated to have an insignificant high dimensional learning efficiency with respect to unlabelled data (Mai and Couillet, 2018), causing it to be outperformed by its unsupervised counterpart, spectral clustering, given sufficient unlabelled data. Following a detailed discussion on the origin of this inconsistency problem, a novel regularization approach involving centering operation is proposed as solution, supported by both theoretical analysis and empirical results.

Keywords: semi-supervised learning, graph-based methods, centered similarities, distance concentration, random matrix theory

1. Introduction

Machine learning methods aim to form a mapping from an input data space to an output characterization space (classification labels, regression vectors) by optimally exploiting the information contained in the collected data. Depending on whether the data fed into the learning model are *labelled* or *unlabelled*, the machine learning algorithms are respectively broadly categorized as *supervised* or *unsupervised*. Although the supervised approach has by now occupied a dominant place in real world applications thanks to its high-level accuracy, the cost of labelling process, overly high in comparison to the collection of data, continually compels researchers to develop techniques using unlabelled data with growing interest, as many popular learning tasks of these days, such as image classification, speech recognition and language translation, require enormous training data sets to achieve satisfying results.

The idea of semi-supervised learning (SSL) comes from the expectation of maximizing the learning performance by combining labelled and unlabelled data (Chapelle et al., 2010). It is of significant practical value when the cost of supervised learning is too high and the performances of unsupervised approaches is too weak. Despite its natural idea, semi-supervised learning has not reached broad recognition. As a matter of fact, many

standard semi-supervised learning techniques were found to be unable to learn effectively from unlabelled data (Shahshahani and Landgrebe, 1994; Cozman et al., 2002; Ben-David et al., 2008), thereby hindering the interest for these methods.

A first key reason for the underperformance of semi-supervised learning methods lies in the lack of understanding of such approaches, caused by the technical difficulty of a theoretical analysis. Indeed, even the simplest problem formulations, the solutions of which assume an explicit form, involve complicated-to-analyze mathematical objects (such as the resolvent of kernel matrices).

A second important aspect has to do with dimensionality. As most semi-supervised learning techniques are built upon low-dimensional reasonings, they suffer the transition to large dimensional data sets. Indeed, it has been long noticed that learning from data of intrinsically high dimensionality presents some unique problems, for which the term *curse of dimensionality* was coined. One important phenomenon of the curse of dimensionality is known as *distance concentration*, which is the tendency for distances between high dimensional data vectors to become indistinguishable. This problem has been studied in many works (Beyer et al., 1999; Aggarwal et al., 2001; Hinneburg et al., 2000; Francois et al., 2007; Angiulli, 2018), providing mathematical characterization of distance concentration under the conditions of intrinsically high dimensional data.

Since the strong agreement between geometric proximity and data affinity in low dimensional spaces is the foundation of similarity-based learning techniques, it is then questionable whether these traditional techniques will perform effectively on high dimensional data sets, and many counterintuitive phenomena may occur.

The aforementioned tractability and dimensionality difficulties can be tackled at once by exploiting recent advances in random matrix theory to analyze the performance of semi-supervised algorithms. With their weakness understood, it is then possible to propose fundamental corrections for these algorithms. The present article focuses on *semi-supervised graph regularization* approaches (Belkin and Niyogi, 2003; Zhu et al., 2003; Zhou et al., 2004), a major subset of semi-supervised learning methods (Chapelle et al., 2010), often referred to as Laplacian regularizations with their loss functions involving differently normalized Laplacian matrices (Avrachenkov et al., 2012). These semi-supervised learning algorithms of Laplacian regularization are presented in Section 2.1. It was made clear in a recent work of Mai and Couillet (2018) that among existing Laplacian regularization algorithms, only one (related to the PageRank algorithm) yields reasonable classification results, yet with asymptotically negligible contribution from the unlabelled data set. This last observation of the inefficiency of Laplacian regularization methods to learn from unlabelled data may cause them to be outperformed by a mere (unsupervised) spectral clustering approach (Von Luxburg, 2007) in the same high dimensional settings (Couillet and Benaych-Georges, 2016). We refer to Section 2.2 for a summary of the key mathematical results in the previous analysis of Mai and Couillet (2018), which motivate the present work.

The contributions of the present work start from Section 3: with the cause for the unlabelled data learning inefficiency of Laplacian regularization identified in Section 3.1, a new regularization approach with centered similarities is proposed in Section 3.2 as a cure, followed by arguments of high-level justification. This new regularization method

is simple to implement and its effectiveness supported by a rigorous analysis, in addition to heuristic arguments and empirical results which justify its usage in more general data settings. Specifically, the statistical analysis of Section 4, placed under a high dimensional Gaussian mixture model (as employed in the previous analysis of Mai and Couillet 2018, as well as that of Couillet and Benaych-Georges 2016 in the context of spectral clustering), proves the consistency of our proposed high dimensional semi-supervised learning method, with guaranteed performance gains over Laplacian regularization. The theoretical results of Section 4 are validated by simulations in Section 5.1. Broadening the perspective, the discussion in Section 3.1 suggests that the unlabelled data learning inefficiency of Laplacian regularization is due to the universal distance concentration phenomenon of high dimensional data. The advantage of our centered regularization, proposed as a countermeasure to the problem of distance concentration, should extend beyond the analyzed Gaussian mixture model. This claim is verified in Section 5.2 through experimentation on real-world data sets, where we observe that the proposed method tends to produce larger performance gains over the Laplacian approach under higher levels of distance concentration. The discussion is extended in Section 6 to related graph-based SSL methods. Although not suffering from the unlabelled data learning inefficiency problem like Laplacian regularization, these methods may still have a suboptimal semi-supervised learning performance on high dimensional data as they do not possess the same performance guarantees as our proposed method. This claim is verified in Section 6.2 thanks to a recent work of Lelarge and Miolane (2019) characterizing the optimal performance on isotropic Gaussian data. A higher-order version of centered regularization is proposed in Section 6.3.1, with a remarkable competitiveness demonstrated in Section 6.3.2 through experiments on several benchmark data sets. The subject of computational efficiency on sparse graphs is approached in Section 6.4.

Notations: 1_n is the column vector of ones of size n, I_n the $n \times n$ identity matrix. The norm $\|\cdot\|$ is the Euclidean norm for vectors and the operator norm for matrices. We follow the convention to use $o_P(1)$ for a sequence of random variables that convergences to zero in probability. For a random variable $x \equiv x_n$ and $u_n \geq 0$, we write $x = O(u_n)$ if for any $\eta > 0$ and $D > 0$, we have $n^D \mathrm{P}(x \geq n^\eta u_n) \to 0$.

2. Background

We will begin this section by recalling the basics of graph learning methods, before briefly reviewing the main results of Mai and Couillet (2018), which motivate the proposition of our centered regularization method in Section 3.

2.1 Laplacian Regularization Method

Consider a set $\{x_1, \ldots, x_n\} \in \mathbb{R}^p$ of p-dimensional data vectors belonging to either one of two affinity classes $\mathcal{C}_1$, $\mathcal{C}_2$. In graph-based methods, data points $x_1, \ldots, x_n$ are represented by vertices in a graph, upon which a weight matrix W is computed by

$$W = \{w_{ij}\}_{i,j=1}^n = \left\{ h\left(\frac{1}{p} \|x_i - x_j\|^2 \right) \right\}_{i,j=1}^n \tag{1}$$

for some decreasing non-negative function h, so that nearby data vectors x_i, x_j are connected with a large weight w_{ij}, also seen as a similarity measure between data vectors. A typical

kernel function for defining w_{ij} is the radial basis function kernel $w_{ij} = e^{-\|x_i - x_j\|^2/t}$. The connectivity of data point x_i is measured by its degree $d_i = \sum_{j=1}^n w_{ij}$, the diagonal matrix $D \in \mathbb{R}^{n\times n}$ having d_i as its diagonal elements is called the degree matrix.

Graph learning approach assumes that data points belonging to the same affinity group are "close" in a graph-proximity sense. In other words, if $f \in \mathbb{R}^n$ is a class signal of data points $x_1, \ldots, x_n$, it varies little from x_i to x_j when w_{ij} has a large value. The graph smoothness assumption translates into the minimization of a smoothness penalty term

$$\frac{1}{2}\sum_{i,j=1}^n w_{ij}(f_i - f_j)^2 = f^\mathsf{T} L f$$

where $L = D - W$ is referred to as the Laplacian matrix. Notice that the above smoothness penalty is minimized to zero for $f = 1_n$, a trivial solution containing no information about the data class. According to this remark, the popular unsupervised graph learning method of spectral clustering simply consists in finding a unit vector orthogonal to 1_n that minimizes the smoothness penalty term, as formalized below

$$\begin{aligned} &\min_{f\in\mathbb{R}^n} f^\mathsf{T} L f \\ s.t. \quad &\|f\| = 1 \quad f^\mathsf{T} 1_n = 0. \end{aligned}$$

It is easily shown by the spectral properties of Hermitian matrices that the solution to the above optimization is the eigenvector of L associated to the second smallest eigenvalue. There exist also other formulations of smoothness penalty involving differently normalized Laplacian matrices, such as the symmetric normalized Laplacian matrix $L_s = I_n - D^{-\frac{1}{2}} W D^{-\frac{1}{2}}$, and the random walk normalized Laplacian matrix $L_r = I_n - WD^{-1}$, which is related to the PageRank algorithm.

In the semi-supervised setting, we dispose of $n_{[l]}$ pairs of data vectors and their labels $\{(x_1, y_1), \ldots, (x_{n_{[l]}}, y_{n_{[l]}})\}$ with $y_i \in \{-1, 1\}$ the class label of x_i, and $n_{[u]}$ unlabelled data $\{x_{n_{[l]}+1}, \ldots, x_n\}$. To incorporate the prior knowledge on the class of labelled data into the class signal f, the semi-supervised graph regularization approach imposes deterministic scores at the labelled points of f, e.g., by letting $f_i = y_i$ for all x_i labelled. The mathematical formulation of the problem then becomes

$$\begin{aligned} &\min_{f\in\mathbb{R}^n} f^\mathsf{T} L f \\ s.t. \quad &f_i = y_i, \quad 1 \le i \le n_{[l]}. \end{aligned}$$

Denoting

$$f = \begin{bmatrix} f_{[l]} \\ f_{[u]} \end{bmatrix}, \; L = \begin{bmatrix} L_{[ll]} & L_{[lu]} \\ L_{[ul]} & L[uu] \end{bmatrix},$$

the above convex optimization problem with equality constrains on $f_{[l]}$ is realized by letting the derivative of the loss function with respect to $f_{[u]}$ equal zero, leading to the following explicit solution

$$f_{[u]} = -L_{[uu]}^{-1} L_{[ul]} f_{[l]}. \tag{2}$$

The typical decision step consists in classifying unlabelled sample x_i by the sign of classification score f_i.

The aforementioned method is frequently referred to as Laplacian regularization, for it finds the class scores of unlabelled data $f_{[u]}$ by regularizing them over the Laplacian matrix along with predefined class scores of labelled data $f_{[l]}$. It is often observed in practice that using other normalized Laplacian regularizers such as $f^\mathsf{T} L_s f$ or $f^\mathsf{T} L_r f$ lead to better classification results. Similarly to the work of Avrachenkov et al. (2012), we define

$$L^{(a)} = I - D^{-1-a} W D^a$$

as the a-normalized Laplacian matrix in order to integrate different Laplacian regularization algorithms into a common framework. Replacing L with $L^{(a)}$ in (2) to get

$$f_{[u]} = -\left(L^{(a)}_{[uu]}\right)^{-1} L^{(a)}_{[ul]} f_{[l]}, \tag{3}$$

we retrieve the solutions of standard Laplacian L, symmetric Laplacian L_s and random walk Laplacian L_r respectively at $a = 0$, $a = -1/2$ and $a = -1$.

Note additionally that the matrix $L^{(a)}_{[uu]}$ is invertible under the trivial condition that the graph represented by W is fully connected (i.e., with no isolated subgraph). To show this, note first that, under this condition, we have

$$u^\mathsf{T}_{[u]} D^{1+2a}_{[u]} L^{(a)}_{[uu]} u_{[u]} = \sum_{i,j=n_{[l]}+1}^{n} w_{ij}(d_i^a u_i - d_j^a u_j)^2 + \sum_{i=n_{[l]}+1}^{n} d_i^{2a} u_i^2 \sum_{m=1}^{n_{[l]}} w_{im} > 0$$

for any $u_{[u]} \neq 0_{n_{[u]}} \in \mathbb{R}^{n_{[u]}}$, as the first term on the right-hand side is strictly positive unless all $d_i^a u_i$ have the same positive value, in which case the second term is strictly positive for there is at least one $w_{im} > 0$. The matrix $L^{(a)}_{[uu]}$ is therefore positive definite. As will be shown in the following though, the fully connected condition is not required for the new algorithm proposed in this article to be well defined and to perform as expected.

Despite being a popular semi-supervised learning approach, Laplacian regularization algorithms are shown by Mai and Couillet (2018) to learn inefficiently from high dimensional unlabelled data, as a direct consequence of the distance concentration phenomenon briefly discussed in the introduction. A deeper examination of the analysis by Mai and Couillet (2018) allows us to discover that the unlabelled data learning inefficiency problem can in fact be settled through the usage of centered similarities, a new approach defying the current convention of non-negative similarities w_{ij}. We will present now the main findings by Mai and Couillet (2018), before the proposition of the novel corrective algorithm in Section 3, along with some general remarks explaining the effectiveness of the proposed algorithm, leaving the thorough performance analysis to Section 4.

2.2 High Dimensional Behaviour of Laplacian Regularization

Conforming to the settings employed by Mai and Couillet (2018), we adopt the following high dimensional data model for the theoretical discussions in this paper.

Assumption 1 *Data samples $x_1, \ldots, x_n$ are i.i.d. observations from a generative model such that, for $k \in \{1, 2\}$, $\mathbb{P}(x_i \in \mathcal{C}_k) = \rho_k$, and*

$$x_i \in \mathcal{C}_k \Leftrightarrow x_i \sim \mathcal{N}(\mu_k, C_k)$$

with $\|C_k\| = O(1)$, $\|C_k^{-1}\| = O(1)$, $\|\mu_2 - \mu_1\| = O(1)$, $\mathrm{tr}C_1 - \mathrm{tr}C_2 = O(\sqrt{p})$ and $\mathrm{tr}(C_1 - C_2)^2 = O(\sqrt{p})$.

The ratios $c_0 = \frac{n}{p}$, $c_{[l]} = \frac{n_{[l]}}{p}$ and $c_{[u]} = \frac{n_{[u]}}{p}$ are bounded away from zero for arbitrarily large p.

Here are some remarks to interpret the conditions imposed on the data means μ_k and covariance matrices C_k in Assumption 1. Firstly, as the discussion is placed under a large dimensional context, we need to ensure that the data vectors do not lie in a low dimensional manifold; the fact that $\|C_k\| = O(1)$ along with $\|C_k^{-1}\| = O(1)$ guarantees non-negligible variations in p linearly independent directions. Other conditions controlling the differences between the class statistics $\|\mu_2 - \mu_1\| = O(1)$, $\mathrm{tr}C_1 - \mathrm{tr}C_2 = O(\sqrt{p})$, and $\mathrm{tr}(C_1 - C_2)^2 = O(\sqrt{p})$ are made for the consideration of establishing *non-trivial classification* scenarios where the classification of unlabelled data does not become impossible or overly easy at extremely large values of p.

The first result concerns the distance concentration of high dimension data. This result is fundamentally responsible for the failure of semi-supervised Laplacian regularization on large dimensional data.

Proposition 1 *Define $\tau = \mathrm{tr}(C_1 + C_2)/p$. Under Assumption 1, we have that, for all $i, j \in \{1, \ldots, n\}$,*

$$\frac{1}{p}\|x_i - x_j\|^2 = \tau + O(p^{-\frac{1}{2}}).$$

The above proposition indicates that in large dimensional spaces, all pairwise distances of data vectors converge to the same value, thereby implying that the presumed relation between *proximity and data affinity* is completely disrupted. In such situations, the performance of Laplacian regularization (and also most distance-based classification methods) may be severely affected. Indeed, under some mild smooth conditions on the weight function h, the analysis of Mai and Couillet (2018) reveals several surprising and critical aspects of the high dimensional behavior of Laplacian regularization. The first conclusion is that all unlabelled data scores f_i for $n_{[l]} + 1 \leq i \leq n$ tend to have the same signs in the case of unequal class priors (i.e., $\rho_1 \neq \rho_2$), causing a meaningless classification of unlabelled data by the sign of their score, unless one normalizes the deterministic scores at labelled points so that they are balanced for each class. In accordance with this message, we shall use in the remainder of the article a class-balanced $f_{[l]}$ defined as below

$$f_{[l]} = \left(I_{n_{[l]}} - \frac{1}{n_{[l]}} 1_{n_{[l]}} 1_{n_{[l]}}^{\mathsf{T}} \right) y_{[l]} \tag{4}$$

where $y_{[l]} \in \mathbb{R}^{n_{[l]}}$ is the label vector composed of y_i for $1 \leq i \leq n_{[l]}$.

Nevertheless, even with balanced $f_{[l]}$ as per (4), the work of Mai and Couillet (2018) shows that the aforementioned "all data affected to the same class" problem still persists

for all Laplacian regularization algorithms under the framework of a-normalized Laplacian (i.e., for $L^{(a)} = I - D^{-1-a}WD^a$) except for $a \simeq -1$. This indicates that among the existing Laplacian regularization algorithms in the literature, only the random walk normalized Laplacian regularization yields non-trivial classification results for large dimensional data. We recall now the exact statistical characterization of $f_{[u]}$ produced by the random walk normalized Laplacian regularization, firstly presented by Mai and Couillet (2018).

Theorem 2 *Let Assumption 1 hold, the function h of (1) be three-times continuously differentiable in a neighborhood of τ, and the solution $f_{[u]}$ be given by (3) for $a = -1$. Then, for $n_{[l]} + 1 \leq i \leq n$ (i.e., x_i unlabelled) and $x_i \in \mathcal{C}_k$,*

$$p(c_0/2\rho_1\rho_2 c_{[l]}) f_i = \tilde{f}_i + o_P(1), \text{ where } \tilde{f}_i \sim \mathcal{N}(m_k, \sigma_k^2)$$

with

$$m_k = (-1)^k(1-\rho_k)\left[-\frac{2h'(\tau)}{h(\tau)}\|\mu_1-\mu_2\|^2 + \left(\frac{h''(\tau)}{h(\tau)} - \frac{h'(\tau)^2}{h(\tau)^2}\right)\frac{(\mathrm{tr}C_1 - \mathrm{tr}C_2)^2}{p}\right] \tag{5}$$

$$\sigma_k^2 = \frac{4h'(\tau)^2}{h(\tau)^2}\left[(\mu_1-\mu_2)^\mathsf{T} C_k(\mu_1-\mu_2) + \frac{1}{c_{[l]}}\frac{\sum_{a=1}^2(\rho_a)^{-1}\mathrm{tr}C_aC_k}{p}\right] + \left(\frac{h''(\tau)}{h(\tau)} - \frac{h'(\tau)^2}{h(\tau^2)}\right)^2 \frac{2\mathrm{tr}C_k^2\,(\mathrm{tr}C_1 - \mathrm{tr}C_2)^2}{p^2}. \tag{6}$$

Theorem 2 states that the classification score f_i for an unlabelled x_i follows approximately a Gaussian distribution at large values of p, with the mean and variance being explicitly dependent of the data statistics μ_k, C_k, the class proportions ρ_k, and the ratio of labelled data over dimensionality $c_{[l]}$. The asymptotic probability of correct classification for unlabelled data is then a direct result of Theorem 2, and reads

$$\mathcal{P}(x_i \to C_k | x_i \in C_k, i > n_{[l]}) = \Phi\left(\sqrt{m_k^2/\sigma_k^2}\right) + o_p(1)$$

where $\Phi(u) = \frac{1}{2\pi}\int_{-\infty}^{u} e^{-\frac{t^2}{2}}dt$ is the cumulative distribution function of the standard Gaussian distribution.

Of utmost importance here is the observation that, while m_k^2/σ_k^2 is an increasing function of $c_{[l]}$, suggesting an effective learning from the labelled set, it is *independent of the unlabelled data ratio* $c_{[u]}$, meaning that in the case of high dimensional data, the addition of unlabelled data, even in significant numbers with respect to the dimensionality p, produces negligible performance gains. Motivated by this crucial remark, we propose in this paper a *simple and fundamental update* to the classical Laplacian regularization approach, for the purpose of boosting high dimensional learning performance through an enhanced utilization of unlabelled data. The proposed algorithm will be presented and intuitively justified in the next section.

3. Proposed Regularization with Centered Similarities

As will be put forward in Section 3.1, we find that the unlabelled data learning inefficiency problem of Laplacian regularization, revealed by Mai and Couillet (2018), is rooted in the

concentration of pairwise distances between data vectors of high dimensionality. To counter the disastrous effect of the distance concentration problem, a new regularization method with centered similarities is proposed in Section 3.2. Some high-level interpretations of the proposed method from the perspective of label propagation and spectral information are given in Section 3.3, justifying its usage in general scenarios beyond the analyzed high dimensional regime.

3.1 Problem Identification

To gain perspective on the cause of inefficient learning from unlabelled data, we will start with a discussion linking the issue to the data high dimensionality.

Developing (3), we get

$$f_{[u]} = L_{[uu]}^{(a)-1} D_{[u]}^{-1-a} W_{[ul]} D_{[l]}^{a} f_{[l]}$$

where

$$W = \begin{bmatrix} W_{[ll]} & W_{[lu]} \\ W_{[ul]} & W_{[uu]} \end{bmatrix} \text{ and } D = \begin{bmatrix} D_{[l]} & 0 \\ 0 & D_{[u]} \end{bmatrix}.$$

From a graph-signal processing perspective (Shuman et al., 2013), since $L_{[uu]}^{(a)}$ is the Laplacian matrix on the subgraph of unlabelled data, and a smooth signal $s_{[u]}$ on the unlabelled data subgraph typically induces large values for the inverse smoothness penalty $s_{[u]}^{\mathsf{T}} L_{[uu]}^{(a)-1} s_{[u]}$, we may consider the operator $\mathcal{P}_u(s_{[u]}) = L_{[uu]}^{(a)-1} s_{[u]}$ as a "smoothness filter" strengthening smooth signals on the unlabelled data subgraph. The unlabelled scores $f_{[u]}$ can be therefore seen as obtained by a two-step procedure:

1. propagating the predetermined labelled scores $f_{[l]}$ through the graph with the a-normalized weight matrix $D_{[u]}^{-1-a} W_{[ul]} D_{[l]}^{a}$ through the label propagation operator $\mathcal{P}_l(f_{[l]}) = D_{[u]}^{-1-a} W_{[ul]} D_{[l]}^{a} f_{[l]}$;
2. passing the received scores at unlabelled points through the smoothness filter $\mathcal{P}_u(s_{[u]}) = L_{[uu]}^{(a)-1} s_{[u]}$ to finally get $f_{[u]} = \mathcal{P}_u\left(\mathcal{P}_l(f_{[l]})\right)$.

It is easy to see that the first step is essentially a supervised learning process, whereas the second one capitalizes on the unlabelled data information. However, as a consequence of the distance concentration "curse" stated in Proposition 1, the similarities (weights) w_{ij} between high dimensional data vectors have essentially a constant value $h(\tau)$ plus some small fluctuations, which results in the collapse of the smoothness filter:

$$\mathcal{P}_u(s_{[u]}) = L_{[uu]}^{(a)-1} s_{[u]} \simeq \left(I_{n_{[u]}} - \frac{1}{n} 1_{n_{[u]}} 1_{n_{[u]}}^{\mathsf{T}} \right)^{-1} s_{[u]} = s_{[u]} + \frac{1}{n_{[l]}} (1_{n_{[u]}}^{\mathsf{T}} s_{[u]}) 1_{n_{[u]}},$$

meaning that, at large values of p, only the constant signal direction $1_{n_{[u]}}$ is amplified by the smoothness filter $\mathcal{P}_u$.

To understand such behavior of the smoothness filter $\mathcal{P}_u$, we recall that, as mentioned in Section 2.1, constant signals with the same value at all points are always considered to

be the most smooth on the graph. This comes from the fact that all weights w_{ij} have non-negative value, so the smoothness penalty term $\mathcal{Q}(s) = \sum_{i,j} w_{[ij]}(s_i - s_j)^2$ is minimized at the value of zero when all elements of the signal s have the same value. Notice also that in perfect situations where all data points in different classes are connected with zero weights $w_{ij} = 0$, class indicators (with non-zero equal values at all data points of a certain class) are just as smooth as constant signals for they also minimize the smoothness penalty term to zero. Even though such scenarios almost never happen in real life, it is hoped that the inter-class similarities are sufficiently weak so that the smoothness filter $\mathcal{P}_u$ is still effective. What is problematic in high dimensional learning is that as the similarities w_{ij} tend to be indistinguishable due to the distance concentration issue of high dimensional data vectors, constant signals have overwhelming advantages to the point that they become the only direction privileged by the smoothness filter $\mathcal{P}_u$, with almost no discrimination between all other directions. In consequence, there is nearly no utilization of the unlabelled data information through Laplacian regularization.

In view of the above discussion, we shall try to eliminate the dominant advantage of constant signals, in an attempt to render detectable the discrimination between class-structured signals and other noisy directions. As constant signals always have a smoothness penalty of zero, a straightforward way to break their optimal smoothness is to introduce negative weights in the graph so that the values of smoothness regularizers can go below zero. More specifically, in the cases where the intra-class similarities are averagely positive and the inter-class similarities averagely negative, class-structured signals are bound to have a lower smoothness penalty than constant signals. However, implementing such idea is hindered by the fact that the positivity of the data points degrees $d_i = \sum_{j=1}^{n} w_{ij}$ is no longer ensured, and having negative degrees can lead to severely unstable results. Take for instance the label propagation step $\mathcal{P}_l(f_{[l]}) = D_{[u]}^{-1-a} W_{[ul]} D_{[l]}^{a} f_{[l]}$, at an unlabelled point x_i, the sum of the received scores after that step equals to $d_i^{-1-a} \sum_{j=1}^{n_{[l]}} (w_{ij} d_j^a) f_j$, the sign of which obviously alters with the sign of the degree at that point, leading thus to extremely unstable classification results.

3.2 Approach of Centered Similarities

To cope with the problem identified above, we propose here to use centered similarities $\hat{w}_{ij}$, for which the positive and negative weights are balanced out at any data point, i.e., for all $i \in \{1, \ldots, n\}$, $d_i = \sum_{j=1}^{n} \hat{w}_{ij} = 0$. Given any similarity matrix W, its centered version $\hat{W}$ is easily obtained by applying a projection matrix $P = \left(I_n - \frac{1}{n} 1_n 1_n^{\mathsf{T}}\right)$ on both sides:

$$\hat{W} = PWP. \tag{7}$$

As a first advantage, the centering approach allows one to remove the degree matrix altogether (for the degrees are exactly zero now) from the updated smoothness penalty

$$\hat{Q}(s) = \sum_{i,j=1}^{n} \hat{w}_{ij}(s_i - s_j)^2 = -s^{\mathsf{T}} \hat{W} s,$$

securing thus a stable behavior of graph regularization with both positive and negative weights.

Another problematic consequence of regularization procedures employing positive and negative weights is that the optimization problem is no longer convex and may have an infinite solution. To deal with this issue, we add a constraint on the norm of the solution. Letting $f_{[l]}$ be given by (4), the new optimization problem is now posed as follows:

$$\min_{f_{[u]} \in \mathbb{R}^{n_{[u]}}} -f^\mathsf{T} \hat{W} f$$
$$s.t. \quad \|f_{[u]}\|^2 = n_{[u]} e^2 \tag{8}$$

for some $e > 0$.

The optimization can be solved by introducing a Lagrange multiplier $\lambda = \lambda(e)$ to the norm constraint $\|f_{[u]}\|^2 = n_{[u]} e^2$ and the solution reads

$$f_{[u]} = \left(\lambda I_{n_{[u]}} - \hat{W}_{[uu]} \right)^{-1} \hat{W}_{[ul]} f_{[l]} \tag{9}$$

with $\lambda > \|\hat{W}_{[uu]}\|$ uniquely given by

$$\left\| \left(\lambda I_{n_{[u]}} - \hat{W}_{[uu]} \right)^{-1} \hat{W}_{[ul]} f_{[l]} \right\|^2 = n_{[u]} e^2. \tag{10}$$

To see that (9) is the unique solution to the optimization problem (8), it is useful to remark that, by the properties of convex optimization, (9) is the unique solution to the unconstrained convex optimization problem $\min_{f_{[u]}} \lambda \|f_{[u]}\|^2 - f^\mathsf{T} \hat{W} f$ for some $\lambda > \|\hat{W}_{[uu]}\|$. When Equation (10) is satisfied, we get (through a proof by contradiction) that (9) is the only solution that minimizes $-f^\mathsf{T} \hat{W} f$ in the subspace defined by $\|f_{[l]}\|^2 = n_{[u]} e^2$.

In practice, λ can be used directly as a hyperparameter for a more convenient implementation. We summarize the method in Algorithm 1 where, for the sake of normalization, we consider a change of variable $\alpha = \lambda / \|\hat{W}_{[uu]}\| - 1$, which is allowed to take any positive value.

Algorithm 1 Graph-Based Centered Regularization

1: **Input:** $n_{[l]}$ pairs of labelled points and labels $\{(x_i, y_i)\}_{i=1}^{n_{[l]}}$, $n_{[u]}$ unlabelled data $\{x_i\}_{i=n_{[l]}+1}^{n}$, parameter $\alpha \in \mathbb{R}^+$.
2: **Output:** Classification score vector of unlabelled data $f_{[u]} \in \mathbb{R}^{n_{[u]}}$
3: Define the similarity matrix $W = \{w_{i,j}\}_{i,j=1}^{n}$ with w_{ij} reflecting the closeness between x_i and x_j.
4: Compute the centered similarity matrix $\hat{W}$ by (7) and the balanced labels $f_{[l]}$ by (4).
5: Set $\lambda = (\alpha + 1)\|W_{[uu]}\|$ and compute $f_{[u]}$ by (9).

The proposed algorithm induces almost no extra cost to the classical Laplacian approach, except the addition of the hyperparameter α controlling the norm of $f_{[u]}$. The performance analysis in Section 4 will help demonstrate that the existence of this hyperparameter, aside from making the regularization with centered similarities a well-posed problem, allows one to adjust the combination of labelled and unlabelled information in search for an optimal semi-supervised learning performance. Roughly speaking, with small α, the algorithm puts a greater weight on the unlabelled data information; conversely, large values of α correspond to a stressed impact of the labelled data.

3.3 Interpretation

Before the performance analysis of Section 4, which establishes the effectiveness of the proposed method for a consistent high dimensional semi-supervised learning, we provide here some high-level arguments that help understand the proposed method and its applicability in a more general context.

3.3.1 Viewpoint of Label Propagation

Similarly to Laplacian regularization, the centered regularization method can also be interpreted from the perspective of label propagation (Zhu and Ghahramani, 2002). Setting $f_{[u]}^{(0)} \leftarrow \lambda^{-1}\hat{W}_{[ul]}f_{[l]}$, we retrieve the solution (9) of centered regularization at the stationary point $f_{[u]}^{(\infty)}$ of the following iterative procedure:

$$f_{[u]}^{(t+1)} \leftarrow \lambda^{-1}\begin{bmatrix}0_{n_{[u]}\times n_{[l]}} & I_{n_{[u]}}\end{bmatrix} PWP \begin{bmatrix} f_{[l]} \\ f_{[u]}^{(t)}\end{bmatrix}.$$

Denoting $f^{(t)} = [f_{[l]}, f_{[u]}^{(t)}]^\mathsf{T}$, the above process can be seen as propagating the centered score vector $\hat{f}^{(t)} = Pf^{(t)}$ through the weight matrix W, then recentering the received scores $\eta^{(t)} = W\hat{f}^{(t)}$ before outputting $f_{[u]}^{(t+1)}$ as the subset of $\hat{\eta}^{(t)} = P\eta^{(t)}$ corresponding to the unlabelled points.

Recall from the discussion in Section 3.1 that the extremely amplified constant signal $1_{n_{[u]}}$ in the outcome $f_{[u]}$ of Laplacian regularization is closely related to the ineffective unlabelled data learning problem. In the proposed approach, the constant signal is cancelled thanks to the recentering operations before and after the label propagation over W. The existence of the multiplier λ^{-1} allows us to magnify the score vector, after its norm was significantly reduced due to the recentering operations.

3.3.2 Spectral Information of Regularizers

As explained in Section 3.1, the motivation behind centered regularization is to propose a smoothness regularizer that penalizes the constant score vector 1_n. With centered similarities, we exchange the Laplacian regularizer $f^\mathsf{T}Lf$ with $-f^\mathsf{T}\hat{W}f$, so that the smoothness penalty is no longer minimized at 1_n. It is easy to see that under the setting of constant degrees $d_i = \sum_{j=1}^n w_{ij} = d$, such as in the case of directed KNN graphs where d equals the number of neighbors of each node, the Laplacian matrix $L = D - W$ has the same ordering of eigenvalue-eigenvector pairs as $-\hat{W}$ except for the constant direction 1_n.

One may argue that the spectral information of L is kept intact in centered regularization, while the potentially counterproductive tendency to favor constant vectors is removed. However, in practice, we often deal with heterogeneous degrees. The observed performance gains by using differently normalized Laplacian matrices suggests the importance of taking into account the effect of heterogeneous degrees on the spectral information. It is thus hard to predict how the centered regularization method may behave in comparison, without a deep understanding on the impact of heterogeneous degrees. As the analysis in Section 4 establishes the superiority of centered regularization in the limit of very high dimensional data learning, the benefit of our proposed method is expected to manifest itself when the

advantage of effective high dimensional data learning outweighs the impact of heterogeneous degrees (which may work against the algorithm of centered regularization). Future studies are envisioned to propose normalized versions of centered regularization better adapted to heterogeneous degrees.

4. Theoretical Guarantees

The main purpose of this section is to provide mathematical support for an effective high dimensional learning of the centered regularization algorithm from not only labelled data but also from unlabelled data, allowing for a theoretically guaranteed performance gain over the classical Laplacian approach (through an enhanced utilization of unlabelled data). The theoretical results also point out that the proposed method has an unlabelled data learning efficiency that is at least as good as spectral clustering, as opposed to Laplacian regularization.

4.1 Precise Performance Analysis

We provide here the statistical characterization of unlabelled data scores $f_{[u]}$ obtained by the proposed algorithm. As the new algorithm will be shown to draw on both labelled and unlabelled data, the complex interactions between these two types of data generate more intricate outcomes than in the analysis of Mai and Couillet (2018). To facilitate the interpretation of the theoretical results without cumbersome notations, we present the theorem here under the data homoscedasticity, i.e., $C_1 = C_2 = C$, without affecting the generality of the conclusions given subsequently. We refer interested readers to the appendix for the generalized theorem along with its proof.

We introduce first two positive functions $m(\xi)$ and $\sigma^2(\xi)$ which are crucial for describing the statistical distribution of unlabelled scores:

$$m(\xi) = \frac{2c_{[l]}\theta(\xi)}{c_{[u]}\big(1-\theta(\xi)\big)} \tag{11}$$

$$\sigma^2(\xi) = \frac{\rho_1\rho_2(2c_{[l]} + m(\xi)c_{[u]})^2 s(\xi) + \rho_1\rho_2(4c_l + m(\xi)^2 c_{[u]})\omega(\xi)}{c_{[u]}\left(c_{[u]} - \omega(\xi)\right)} \tag{12}$$

where

$$\begin{aligned}
\theta(\xi) &= \rho_1\rho_2\xi(\mu_1-\mu_2)^\mathsf{T}\left(I_p - \xi C\right)^{-1}(\mu_1-\mu_2)\\
\omega(\xi) &= \xi^2 p^{-1}\operatorname{tr}\left[\left(I_p-\xi C\right)^{-1}C\right]^2\\
s(\xi) &= \rho_1\rho_2\xi^2(\mu_1-\mu_2)^\mathsf{T}\left(I_p-\xi C\right)^{-1}C\left(I_p-\xi C\right)^{-1}(\mu_1-\mu_2).
\end{aligned}$$

Here the positive functions $m(\xi)$ and $\sigma^2(\xi)$ are defined respectively on the domains $(0,\xi_m)$ and $(0,\xi_{\sigma^2})$ with $\xi_m, \xi_{\sigma^2} > 0$ uniquely given by $\theta(\xi_m) = 1$ and $\omega(\xi_{\sigma^2}) = c_{[u]}$. Additionally, we define

$$\xi_{\text{sup}} = \min\{\xi_m, \xi_{\sigma^2}\}. \tag{13}$$

These definitions may at first glance seem complicated, but it suffices to keep in mind a few key messages to understand the theoretical results and their implications:

- $\theta(\xi)$, $\omega(\xi)$ and $s(\xi)$ are all positive and strictly increasing functions for $\xi \in (0, \xi_{\text{sup}})$; consequently so are $m(\xi)$ and $\sigma^2(\xi)$.
- ξ_m does not depend on $c_{[l]}$ or $c_{[u]}$; as for ξ_{σ^2}, it is constant with $c_{[l]}$ but increases as $c_{[u]}$ increases.
- $\rho_1\rho_2 m^2(\xi) + \sigma^2(\xi)$ monotonously increases from zero to infinity as ξ increases from zero to ξ_{sup}.

The above remarks are derived directly from the definitions of the involved mathematical objects.

Theorem 3 *Let Assumption 1 hold with $C_1 = C_2 = C$, the function h of (1) be three-times continuously differentiable in a neighborhood of τ, $f_{[u]}$ be the solution of (8) with fixed norm $n_{[u]}e^2$ and with the notations of $m(\xi)$, $\sigma^2(\xi)$, ξ_{sup} given in (11), (12), (13). Then, for $n_{[l]} + 1 \leq i \leq n$ (i.e., x_i unlabelled) and $x_i \in \mathcal{C}_k$,*

$$f_i = \tilde{f}_i + o_P(1), \text{ where } \tilde{f}_i \sim \mathcal{N}((-1)^k(1-\rho_k)\hat{m}, \hat{\sigma}^2)$$

with

$$\hat{m} = m(\xi_e), \quad \hat{\sigma}^2 = \sigma^2(\xi_e)$$

for $\xi_e \in (0, \xi_{\text{sup}})$ uniquely given by $\rho_1\rho_2 m(\xi_e)^2 + \sigma^2(\xi_e) = e^2$.

4.2 Consistent Learning from Labelled and Unlabelled Data

Theorem 3 implies that the performance of the proposed method is controlled by both $c_{[l]}$ and $c_{[u]}$ (the number of labelled and unlabelled samples per dimension), as $m(\xi)$, $\sigma^2(\xi)$ (given by (11), (12)) are dependent of $c_{[l]}$ and $c_{[u]}$. It is however hard to see directly a consistently increasing performance with both $c_{[l]}$ and $c_{[u]}$ from these results. As a first objective of this section, we translate the theorem into more interpretable results.

First, it should be pointed out that, with the approach of centered similarities, the norm of the unlabelled data score vector $f_{[u]}$ is adjustable via the hyperparameter e, as opposed to the Laplacian regularization methods. As will be demonstrated later in this section, the norm of $f_{[u]}$, or more precisely the norm of its deterministic part $\mathbb{E}\{f_{[u]}\}$, directly affects how much the learning process relies on the unlabelled (versus labelled) data. With $\mathbb{E}\{f_{[u]}\}$ given by Theorem 3 for high dimensional data, we indeed note that

$$\frac{\|\mathbb{E}\{f_{[u]}\}\|}{\|f_{[l]}\| + \|\mathbb{E}\{f_{[u]}\}\|} = \frac{c_{[u]}\hat{m}}{2c_{[l]} + c_{[u]}\hat{m}} + o_P(1) = \theta(\xi_e) + o_P(1)$$

as it can be obtained from (11) that

$$\theta(\xi) = \frac{c_{[u]}m(\xi)}{2c_{[l]} + c_{[u]}m(\xi)}.$$

In the following discussion, we shall use the variance over square mean ratio

$$r_{\text{ctr}} \equiv \hat{\sigma}^2/\hat{m}^2 \tag{14}$$

as the inverse performance measure for the method of centered regularization (i.e., smaller values of r_{ctr} imply better classification results for high dimensional data). A reorganization of the results in Theorem 3 leads to the corollary below.

Corollary 4 *Under the conditions and notations of Theorem 3, and with r_{ctr} defined in* (14)*, we have*

$$\frac{r_{\text{ctr}}}{\rho_1\rho_2} = \frac{s(\xi_e)}{\theta^2(\xi_e)} + \frac{\omega(\xi_e)}{\theta^2(\xi_e)}\left[\frac{\theta^2(\xi_e)}{c_{[u]}}\left(1+\frac{r_{\text{ctr}}}{\rho_1\rho_2}\right) + \frac{(1-\theta(\xi_e))^2}{c_{[l]}}\right] \tag{15}$$

where we recall $\theta(\xi) = \frac{c_{[u]}m(\xi)}{2c_{[l]}+c_{[u]}m(\xi)} \in (0,1)$.

Equation (15) suggests a growing performance with more labelled or unlabelled data, as the last two terms on the right-hand side have respectively $c_{[u]}$ and $c_{[l]}$ in their denominators. These two terms are actually quite similar, except for the pair of $\theta^2(\xi_e)$ and $[1-\theta(\xi_e)]^2$ each associated to one of them, and a factor of $1 + r_{\text{ctr}}/\rho_1\rho_2 \geq 1$ in the term with $c_{[u]}$. As said earlier, the quantity $\theta(\xi_e) = c_{[u]}\hat{m}/(2c_{[l]} + c_{[u]}\hat{m}) \in (0,1)$ reflects how much the learning relies on unlabelled data. Indeed, it can be observed from (15) that r_{ctr} tends to be only dependent of $c_{[l]}$ (resp., $c_{[u]}$) in the limit $\theta(\xi_e) \to 0$ (resp., $\theta(\xi_e) \to 1$). The factor $1 + r_{\text{ctr}}/\rho_1\rho_2 \geq 1$ translates into the fact that unlabelled data are less informative than the labelled ones. According to the definition of r_{ctr}, this factor goes to 1 when the scores of unlabelled data tend to deterministic values, indicating an equivalence between labelled and unlabelled data in this extreme scenario. In a way, the factor of $1 + r_{\text{ctr}}/\rho_1\rho_2$ quantifies how much labelled samples are more helpful than unlabelled data to the learning process.

To demonstrate an effective learning from labelled and unlabelled data, we now show that, for a well-chosen e, r_{ctr} decreases with $c_{[u]}$ and $c_{[l]}$. Recall that the expressions of $\theta(\xi)$, $\omega(\xi)$ and $s(\xi)$ do not involve $c_{[u]}$ or $c_{[l]}$. It is then easy to see that, at some fixed ξ_e, $r_{\text{ctr}} > 0$ is a strictly decreasing function of both $c_{[u]}$ and $c_{[l]}$. Adding to this argument the fact that the attainable range $(0, \xi_{\text{sup}})$ of ξ_e over $e > 0$ is independent of $c_{[l]}$ and only enlarges with greater $c_{[u]}$ (as can be derived from the definition (13) of ξ_{sup}), we conclude that the performance of the proposed method consistently benefits from the addition of input data, *whether labelled or unlabelled*, as formally stated in Proposition 5. These remarks are illustrated in Figure 1, where we plot the probability of correct classification as $\theta(\xi_e)$ varies from 0 to 1.

Proposition 5 *Under the conditions and notations of Corollary 4, we have that, for any $e > 0$, there exists an $e' > 0$ such that $r_{\text{ctr}}(c_{[l]}, c_{[u]}, e) > r'_{\text{ctr}}(c'_{[l]}, c'_{[u]}, e')$ if $c'_{[l]} \geq c_{[l]}$, $c'_{[u]} \geq c_{[u]}$ and $c'_{[l]} + c'_{[u]} > c_{[l]} + c_{[u]}$.*

Not only is the proposed method of centered regularization able to achieve an effective semi-supervised learning on high dimensional data, it does so with *a labelled data learning*

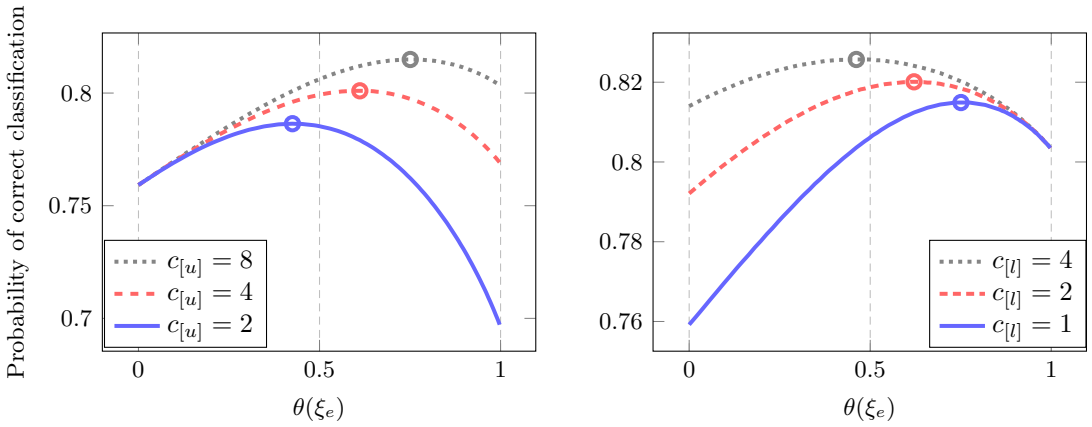

Figure 1: Asymptotic probability of correct classification as $\theta(\xi_e)$ varies, for $\rho_1 = \rho_2$, $p = 100$, $\mu_1 = -\mu_2 = [-1, 0, \ldots, 0]^\mathsf{T}$, $\{C\}_{i,j} = .1^{|i-j|}$. Left: various $c_{[u]}$ with $c_{[l]} = 1$. Right: various $c_{[l]}$ with $c_{[u]} = 8$. Optimal values marked in circle.

efficiency lower bounded by that of Laplacian regularization (which is reduced to supervised learning in high dimensions), and *an unlabelled data learning efficiency lower bounded by that of spectral clustering*, a standard unsupervised learning algorithm on graphs. The focus of the following discussion is to establish this second remark, which implies the superiority of centered regularization over the methods of Laplacian regularization and spectral clustering.

Observe from Theorem 2 that under the homoscedasticity assumption, the random walk normalized Laplacian algorithm (the only one ensuring non-trivial high dimensional classification among existing Laplacian algorithms) gives (similarly to the centered regularization method) $\tilde{f}_i \sim \mathcal{N}\left((-1)^k(1-\rho_k)m', \sigma'^2\right)$ for $m' = (2\rho_1\rho_2 c_{[l]}/pc_0)(m_2 - m_1), \sigma' = (2\rho_1\rho_2 c_{[l]}/pc_0)\sigma_1 = (2\rho_1\rho_2 c_{[l]}/pc_0)\sigma_2$ with $m_k, \sigma_k, k \in \{1, 2\}$ given in Theorem 2. Similarly to the definition of r_{ctr}, we denote

$$r_{\text{Lap}} \equiv \sigma'^2/m'^2. \tag{16}$$

Since $\theta(\xi_e) \to 0$ as $\xi_e \to 0$ and $\xi_e \to 0$ as $e \to 0$, we obtain the following proposition from the results of Theorem 2 and Corollary 4.

Proposition 6 *Under the conditions and notations of Theorem 2 and Corollary 4, letting* r_{Lap} *be defined by* (16)*, we have that*

$$\lim_{e\to 0} r_{\text{ctr}} = r_{\text{Lap}} = \frac{(\mu_1-\mu_2)^\mathsf{T} C(\mu_1-\mu_2)}{\|\mu_1-\mu_2\|^4} + \frac{\text{tr}C^2}{p\|\mu_1-\mu_2\|^4\rho_1\rho_2 c_{[l]}}.$$

We thus remark that *the performance of Laplacian regularization is retrieved by the proposed method in the limit* $e \to 0$.

After ensuring the superiority of the new regularization method over the original Laplacian approach, we now proceed to provide further guarantee on its unlabelled data learning efficiency by comparing it to the unsupervised method of spectral clustering.

Recall that the regular graph smoothness penalty term $Q(s)$ of a signal s can be written as $Q(s) = s^\mathsf{T} L s$. In an unsupervised learning setting, we shall seek the unit-norm vector that minimizes the smoothness penalty, which is the eigenvector of L associated with the smallest eigenvalue. However, as $Q(s)$ reaches its minimum at the clearly non-informative flat vector $s = 1_n$, the sought-for solution is provided instead by the eigenvector associated with the second smallest eigenvalue. In contrast, the updated smoothness penalty term $\hat{Q}(s) = -s^\mathsf{T} \hat{W} s$ with centered similarities does not achieves its minimum for "flat" signals, and thus the eigenvector associated with the smallest eigenvalue is here a valid solution. Another important aspect is that spectral clustering based on the unnormalized Laplacian matrix $L = D - W$ has long been known to behave unstably (Von Luxburg et al., 2008), as opposed to the symmetric normalized Laplacian $L_s = I_n - D^{-\frac{1}{2}} W D^{-\frac{1}{2}}$, so here we compare against L_s rather than L.

Let us define $d_{\text{inter}}(v)$ as the inter-cluster distance operator that takes as input a real-valued vector v of dimension n, then returns the distance between the centroids of the clusters formed by the set of points in the same class $\{v_i | 1 \leq i \leq n, x_i \in \mathcal{C}_k\}$, for $k \in \{1, 2\}$; and $d_{\text{intra}}(v)$ be the intra-cluster distance operator that returns the standard deviation within clusters. Namely,

$$d_{\text{inter}}(v) = |j_1^\mathsf{T} v / n_1 - j_2^\mathsf{T} v / n_2|$$
$$d_{\text{intra}}(v) = \|v - (j_1^\mathsf{T} v / n_1) j_1 - (j_2^\mathsf{T} v / n_2) j_2\| / \sqrt{n}$$

where $j_k \in \mathbb{R}^n$ with $k \in \{1, 2\}$ is the indicator vector of class k with $[j_k]_i = 1$ if $x_i \in \mathcal{C}_k$, otherwise $[j_k]_i = 0$; and n_k the number of ones in the vector j_k. As the purpose of clustering analysis is to produce clusters conforming to the intrinsic classes of data points, with low variance within each cluster and large distance between clusters, the following proposition (see the proof in Appendix B) shows that the spectral clustering algorithm based on the normalized Laplacian matrix L_s, which has been studied by Couillet and Benaych-Georges (2016) under the high dimensional setting, has practically the same performance as the one with the centered similarity matrix $\hat{W}$.

Proposition 7 *Under the conditions of Theorem 3, let v_{Lap} be the eigenvector of L_s associated with the second smallest eigenvalue, and v_{ctr} the eigenvector of $\hat{W}$ associated with the largest eigenvalue. Then,*

$$\frac{d_{\text{inter}}(v_{\text{Lap}})}{d_{\text{intra}}(v_{\text{Lap}})} = \frac{d_{\text{inter}}(v_{\text{ctr}})}{d_{\text{intra}}(v_{\text{ctr}})} + o_P(1)$$

for non-trivial clustering with $d_{\text{inter}}(v_{\text{Lap}})/d_{\text{intra}}(v_{\text{Lap}}), d_{\text{inter}}(v_{\text{ctr}})/d_{\text{intra}}(v_{\text{ctr}}) = O(1)$.

As explained before, the solution $f_{[u]}$ of the centered similarities regularization can be expressed as $f_{[u]} = \left(\lambda I_{n_{[u]}} - \hat{W}_{[uu]}\right)^{-1} \hat{W}_{[ul]} f_{[l]}$ for some $\lambda > \|\hat{W}_{[uu]}\|$ (dependent of e as indicated in (10)). Clearly, as $\lambda \downarrow \|\hat{W}_{[uu]}\|$, $f_{[u]}$ tends to align with the eigenvector of $\hat{W}_{[uu]}$ associated with the largest eigenvalue. Therefore, *the performance of spectral clustering on the unlabelled data subgraph is retrieved at $e \to +\infty$.*

In view of the above discussion, we conclude that the proposed regularization method with centered similarities

- recovers the high dimensional performance of Laplacian regularization at $e \to 0$;
- recovers the high dimensional performance of spectral clustering at $e \to +\infty$;
- accomplishes a consistent high dimensional semi-supervised learning for e appropriately set between the two extremes, thus leading to an increasing performance gain over Laplacian regularization with greater amounts of unlabelled data.

5. Numerical Evidence

The main objective of this section is to provide empirical evidence for the superiority of the proposed regularization method on high dimensional data, in addition to the theoretical guarantees provided in Section 4. Firstly, we check the validity of our asymptotic results on moderately large data sets. The figures of Section 5.1 show that our performance prediction matches the empirical value with great precision on data sets of $n, p \sim 100$. Beyond our theoretical model, we provide an empirical study in Section 5.2 of how Laplacian and centered regularizations behave under different levels of distance concentration through simulations on real-world data. The numerical results confirm a positive correlation between the severity of distance concentration and the advantage of centered regularization over Laplacian regularization, as we observe a consistently effective unlabelled data learning of the former while the latter fails under strong influence of distance concentration.

5.1 Validation of Precise Performance Analysis on Finite-Size Systems

We first validate the asymptotic results of Section 4 on finite data sets of only moderately large sizes ($n, p \sim 100$). Recall from Section 4 that the asymptotic performance of Laplacian regularization and spectral clustering are recovered by centered regularization at extreme values of the parameter θ, respectively in the limit $\theta = 0$ and $\theta = 1$ (when spectral clustering yields non-trivial solutions); this is how the theoretical values of both methods are computed in Figure 2. The finite-sample results are given for the best (oracle) choice of the hyperparameter a in the generalized Laplacian matrix $L^{(a)} = I - D^{-1-a}WD^a$ for Laplacian regularization and spectral clustering, and for the optimal (oracle) choice of the hyperparameter α for centered regularization.

Under a non-trivial Gaussian mixture model setting (see caption) with $p = 100$, Figure 2 demonstrates a sharp prediction of the average empirical performance by the asymptotic analysis. As revealed by the theoretical results, the Laplacian regularization method fails to learn effectively from unlabelled data, causing it to be outperformed by the purely unsupervised spectral clustering approach (for which the labelled data are treated as unlabelled ones) for sufficiently numerous unlabelled data. The performance curve of the proposed centered regularization algorithm, on the other hand, is consistently above that of spectral clustering, with a growing advantage over Laplacian regularization as the number of unlabelled data increases.

Figure 2 also interestingly shows that the unsupervised performance of spectral clustering is noticeably reduced when the covariance matrix of the data distribution changes from the identity matrix to a slightly disrupted model (here for $\{C\}_{i,j} = .1^{|i-j|}$). On the contrary, the Laplacian regularization, the high dimensional performance of which relies

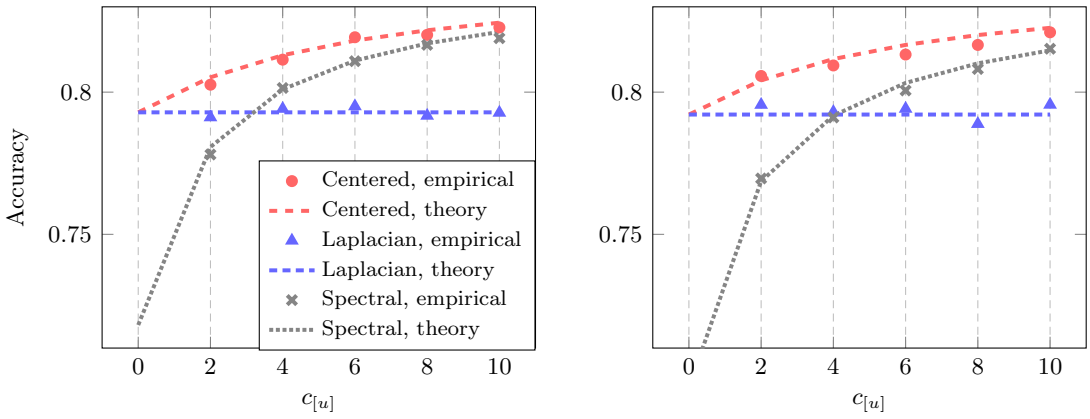

Figure 2: Empirical and theoretical accuracy as a function of $c_{[u]}$ with $c_{[l]} = 2$, $\rho_1 = \rho_2$, $p = 100$, $-\mu_1 = \mu_2 = [-1, 0, \ldots, 0]^\mathsf{T}$, $C = I_p$ (left) or $\{C\}_{i,j} = .1^{|i-j|}$ (right). Graph constructed with $w_{ij} = e^{-\|x_i - x_j\|^2/p}$. Averaged over $50000/n_{[u]}$ iterations.

essentially on labelled data, is barely affected. This is explained by the different impacts labelled and unlabelled data have on the learning process, which can be understood from the theoretical results in Section 4.

5.2 Effect of Distance Concentration Beyond Model Assumptions

While our technical derivation relies on the Gaussianity of data vectors, we expect the advantage of an effective semi-supervised learning by the proposed method to hold in a broader context of high dimensional learning, as the distance concentration phenomenon (which, as we recall from Section 3.1, is responsible for the unlabelled data learning inefficiency of Laplacian regularization) is essentially *irrespective of the data Gaussianity*. Proposition 1 can indeed be generalized to a wider statistical model by a mere law of large numbers; this is the case for instance of all high dimensional data vectors x_i of the form $x_i = \mu_k + C_k^{\frac{1}{2}} z_i$, for $k \in \{1, 2\}$, where $\mu_k \in \mathbb{R}^p$, $C_k \in \mathbb{R}^{p \times p}$ are means and covariance matrices as specified in Assumption 1 and $z_i \in \mathbb{R}^p$ any random vector of independent elements with zero mean, unit variance and bounded fourth order moment. Beyond this model of z_i with independent entries, the recent work by Louart and Couillet (2018) strongly suggests that Proposition 1 remains valid for the wider class of *concentrated vectors* x_i (Ledoux, 2005), including in particular generative models of the type $x_i = F(z_i)$ for $z_i \sim \mathcal{N}(0, I_p)$ and $F : \mathbb{R}^p \to \mathbb{R}^p$ any 1-Lipschitz mapping (for instance, artificial images produced by generative adversarial networks, Goodfellow et al., 2014).

The main objective of this section is to provide an actual sense of how the Laplacian regularization approach and the proposed method behave under *different levels of distance concentration*. We focus here, as a real-life example, on the MNIST data of handwritten digits (LeCun, 1998).

For a fair comparison of Laplacian and centered regularizations, the results displayed here are obtained on their respective best performing graphs, selected among the $k-$nearest

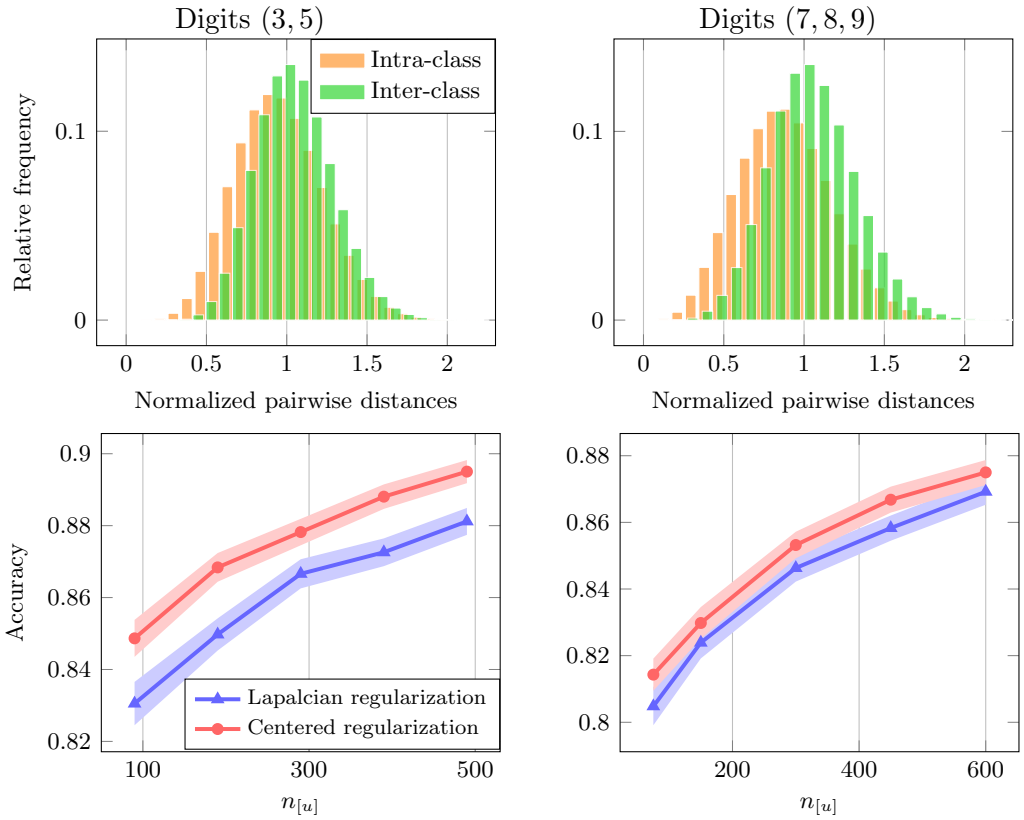

Figure 3: Top: distribution of normalized pairwise distances $\|x_i - x_j\|^2/\bar{\delta}$ $(i \neq j)$ with $\bar{\delta}$ the average of $\|x_i - x_j\|^2$ for MNIST data. Bottom: average accuracy as a function of $n_{[u]}$ with $n_{[l]} = 15$ (left) or $n_{[l]} = 10$ (right), computed over 1000 random realizations with 99% confidence intervals represented by shaded regions.

neighbors graphs (which were observed to yield very competitive performance on MNIST data) with various numbers of neighbors $k = \{2^1, \ldots, 2^q\}$, for q the largest integer such that $2^q < n$. The hyperparameters of Laplacian and centered regularizations are set optimally within the admissible range.[1] It worth pointing out that the popular KNN graphs, constructed by letting $w_{ij} = 1$ if data points x_i or x_j is among the k nearest (k being the parameter to be set beforehand) to the other data point, and $w_{ij} = 0$ if not, are not covered by the present analytic framework. Our study only deals with graphs where w_{ij} is exclusively determined by the distance between x_i and x_j, while in the KNN graphs, w_{ij} is dependent of all pairwise distances in the whole data set. Nonetheless, KNN graphs evidently suffer the same problem of distance concentration, for they are still based on the distances between data points. It is thus natural to expect the proposed centering procedure to be also advantageous on KNN graphs.

1. Specifically, the hyperparameter a of Laplacian regularization is searched from -2 to 0 with a step of 0.02, and the hyperparameter α of centered regularization within the grid $\alpha = 10^{\{-3,-2.9,\ldots,2.9,3\}}$. The results outside these ranges are observed to be non-competitive.

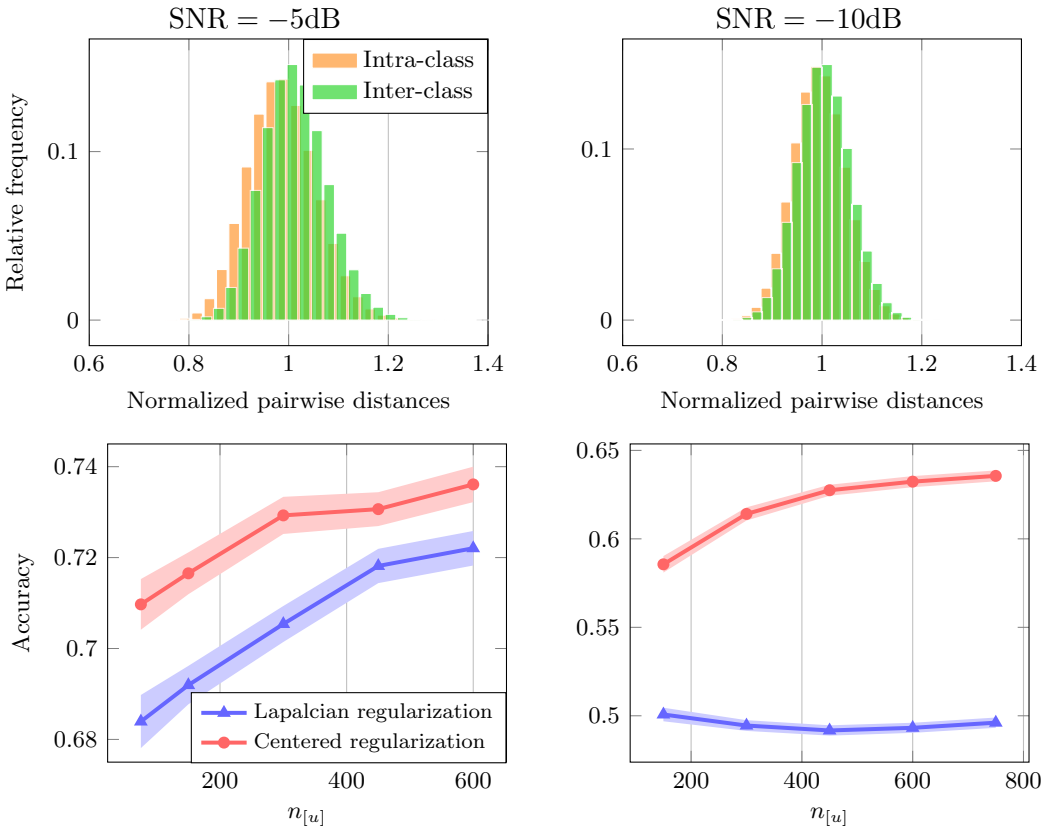

Figure 4: Top: distribution of normalized pairwise distances $\|x_i - x_j\|^2/\bar{\delta}$ $(i \neq j)$ with $\bar{\delta}$ the average of $\|x_i - x_j\|^2$ for noisy MNIST data (7,8,9). Bottom: average accuracy as a function of $n_{[u]}$ with $n_{[l]} = 15$, computed over 1000 random realizations with 99% confidence intervals represented by shaded regions.

Figure 3 shows that high classification accuracy is easily obtained on MNIST data, even with the classical Laplacian approach. However, it exhibits an lower learning efficiency compared to the proposed method. We also find that the benefit of the proposed algorithm is more perceptible on the binary classification task displayed on the left side of Figure 3 than the multiclassification task on the right side, for which the difference between inter-class and intra-class distances is more apparent. This suggests that the advantage of the proposed method is more related to a subtle distinction between inter-class and intra-class distances than to the number of classes.

Figure 4 presents situations where the learning problem becomes more challenging in the presence of additive noise. Understandably, the distance concentration phenomenon is more acute in this noise-corrupted setting, causing more subtle distinction between inter-class and intra-class distances. As a result, the performance gain generated by the proposed method should be more significant for being robust to the negative influence of distance concentration. This is corroborated by Figure 4, where larger performance gains are observed on noised data from the same task on the right side of Figure 3. Moreover, on the

right display of Figure 4, where the similarity information is seriously disrupted by the additive noise, we observe the anticipated saturation effect when increasing $n_{[u]}$ for Laplacian regularization, in contrast to the growing performance of the proposed approach. This suggests, in conclusion, that regularization with centered similarities has a competitive, if not superior, performance in various situations, and yields particularly significant performance gains when the distinction between intra-class and inter-class similarities is quite subtle.

6. Further Discussion and Support

We start this section by presenting other graph-based semi-supervised learning methods and discussing them in relation to the regularization approaches investigated in this article. To evaluate the ability of these SSL methods to optimally exploit the information in partially labelled data sets, we use the recent results of Lelarge and Miolane (2019) as a reference point, where the best achievable semi-supervised learning performance on high dimensional Gaussian mixture data with identity covariance matrices was characterized. For a broader discussion on the applicability of the centering approach, we propose in Section 6.3.1 a higher-order version of centered regularization adapted from the iterated Laplacian method (Zhou and Belkin, 2011), which serves as an example for applying centered similarities beyond the standard Laplacian regularization. Our experiments in Section 6.3.2 demonstrate the practical interest of centered similarities on various benchmark data sets, and show in particular that the algorithm combining centered simliarities with the iterated technique of higher-order regularization helps substantially increase the competitiveness of graph regularization as an SSL approach on challenging data sets. We conclude by remarking in Section 6.4 that the sparsity of the weight matrix can also be exploited for improving the computational efficiency of centered regularization with the help of Woodbury's inversion formula.

6.1 Related Methods

We review first some competitive variants of Laplacian regularization before presenting another major approach of graph-based semi-supervised learning which relies explicitly on the spectral decomposition of Laplacian matrices.

6.1.1 Variants of Laplacian Regularization

The method of Laplacian regularization has been found by Nadler et al. (2009) to suffer from the "score flatness" problem where unlabelled data scores f_i concentrate around the same value (i.e., $f_i = c + o(1)$ for some constant c) when the number of unlabelled samples is exceedingly large compared to that of labelled ones (i.e., $n_{[u]}/n_{[l]} \to \infty$). Following this discovery, several regularization techniques have been proposed to adapt the original method to ensure well-behaved scores, which will be presented in this section. On a related note, the analysis of Mai and Couillet (2018) (which motivated the present study) pointed out that, in high dimensions, the phenomenon of flat unlabelled data scores occurs even when the number of unlabelled samples is comparable to that of labelled ones. As can be easily deduced from our study, the problem of flat unlabelled scores is addressed by the centered regularization method in the more challenging setting of high dimensional learning.

Higher Order Regularization. A well-studied approach to address the problem of flat unlabelled scores revealed by Nadler et al. (2009) is higher-order regularization. There exist mainly two types of higher-order regularization: iterated Laplacian and ℓ_p-based Laplacian. The method of iterated Laplacian regularization consists in using the powers of Laplacian matrices for constructing high-order regularizers $f^\mathsf{T} L^q f$ of graph smoothness. It was shown by Zhou and Belkin (2011) to yield non-flat unlabelled scores. The ℓ_p-based Laplacian regularization achieves this by forcing a stronger constrain $\sum_{i,j=1}^n w_{ij}|f_i - f_j|^q$ on the smoothness (Zhou and Schölkopf, 2005; El Alaoui et al., 2016). Another method in the same vein is *game-theoretic p*-Laplacian (Rios et al., 2019), which does not arise through an optimization problem and tends to be numerically better conditioned. In comparison, the iterated approach is more computationally efficient as it assumes an explicit solution, whereas the method of ℓ_p-based Laplacian regularization calls for practical algorithms to solve more efficiently the implicit optimization (Rios et al., 2019). Also, the iterated Laplacian is found to outperform p-voltages Laplacian regularization (Bridle and Zhu, 2013), a dual version of ℓ_p-based Laplacian in the context of electrical networks. In addition to avoiding the score 'flatness' issue, high-order regularizers $f^\mathsf{T} L^q f$ and their extensions $f^\mathsf{T} g(L) f$ (Smola and Kondor, 2003) obviously benefit from more degrees of freedom to improve the classification performance.

Weighted Nonlocal Laplacian. Other than the methods of high-order regularization, the approach of weighted nonlocal Laplacian (Shi et al., 2017) has also been observed to be effective in combating the score flatness problem. By changing the optimization to

$$\min_{f_{[u]} \in \mathbb{R}^{n_{[u]}}} \sum_{i=n_{[l]}+1}^{n} \left[\sum_{j=n_{[l]}+1}^{n} w_{ij}(f_i - f_j)^2 + \frac{n}{n_{[l]}} \sum_{j=1}^{n_{[l]}} w_{ij}(f_i - f_j)^2 \right],$$

it places a higher weight of $n/n_{[l]}$ on the smoothness penalty between unlabelled data scores and fixed non-flat ones on labelled points.

6.1.2 Eigenvector-Based Methods

Aside from graph regularization methods, another popular graph-based semi-supervised approach exists which takes advantage of the spectral information of Laplacian matrices (Belkin and Niyogi, 2003). Rather than regularizing f over the graph, this method computes first the eigenmap of Laplacian matrices, then uses a certain number s of eigenvectors $E = [e_1, \ldots, e_s]$ associated with the smallest eigenvalues to build a linear subspace and search within this space for an f which minimizes $\|f_{[l]} - y_{[l]}\|$. By the method of least squares, $f = Ea$ with $a = (E_{[l]}^\mathsf{T} E_{[l]})^{-1} E_{[l]}^\mathsf{T} y_{[l]}$.

As an advantage of using the spectral information, this eigenvector-based method is guaranteed to achieve at least the performance of spectral clustering, as opposed to the Laplacian regularization approach. On the other hand, the regularization approach does not have a performance which depends crucially on how well the class signal is captured by a small number of eigenvectors, as it uses the graph matrix as a whole. Another benefit of the graph regularization approach is that it can be easily incorporated into other algorithms as an additional term in the loss function (e.g., Laplacian SVMs). With our proposed algorithm of centered regularization, a consistent learning of unlabelled data, related to

the performance of spectral clustering, can also be achieved by the graph regularization approach. Moreover, the proposed method has a theoretically-proven efficient usage of labelled data which is absent in the eigenvector-based method.

6.2 Optimal Performance on Isotropic Gaussian Data of High Dimensionality

A very recent work of Lelarge and Miolane (2019) has established the optimal performance of semi-supervised learning on a high dimensional Gaussian mixture data model $\mathcal{N}(\pm\mu, I_p)$, with identity covariance matrices.[2] In this work, a method of Bayesian estimation is identified as the one achieving the optimal performance. However, as pointed out by the authors, this method is computationally expensive except on fully labelled data sets and approximations are needed for practical usage.

By comparing the results of Lelarge and Miolane (2019) with our performance analysis in Section 4, we find that the method of centered regularization achieves an optimal performance on fully labelled data sets and a nearly optimal one on partially labelled sets.[3] Numerical results are given in Figure 5, where the classification accuracy of the centered regularization method, computed from Theorem 3 and maximized over the hyperparameter e, is observed to be extremely close to the optimal performance provided by Lelarge and Miolane (2019). Hence, the centered regularization method can be used as a computationally efficient alternative to the Bayesian approach which yields the best achievable performance. In contrast, other graph-based semi-supervised learning algorithms are much less effective in reaching the optimal performance, as can be observed from Figure 6.

We remark also that the iterated Laplacian regularization method appears to be less efficient in exploiting unlabelled data and so is the eigenvector-based method in learning from labelled data. As can be observed in Figure 6, the iterated Laplacian regularization method falls notably short of approaching the optimal performance when the value of m yielding the highest accuracy is further away from 1 (scenarios depicted by the blue curves in the figure). Since we retrieve the standard Laplacian regularization at $m = 1$, which gives the optimal performance in the absence of unlabelled data, the performance gain yielded by the iterated Laplacian regularization technique over the Laplacian method is mainly brought by the utilization of unlabelled data at higher m. However, as demonstrated in Figure 6, the utilization of unlabelled data at higher m is unsatisfactory in allowing the method to reach the optimal semi-supervised learning performance. Since the eigenvector-based approach is reduced to the purely unsupervised method of spectral clustering at $s = 1$, the same remark can be made with respect to its labelled data learning efficiency.

6.3 Applicability of Centered Similarities and High-Order Regularization

The focus of this article is to promote the usage of centered similarities in graph regularization for semi-supervised learning. This fundamental idea can also be applied together with other regularization techniques. In this section, we start by presenting a higher-order version of centered regularization that borrows from iterated Laplacian. To investigate the practical potential of centered similarities, we test the proposed method of centered regu-

2. To the authors' knowledge, more general results (e.g., with arbitrary covariance matrices) are currently out-of-reach.
3. We refer to Appendix D for some theoretical details.

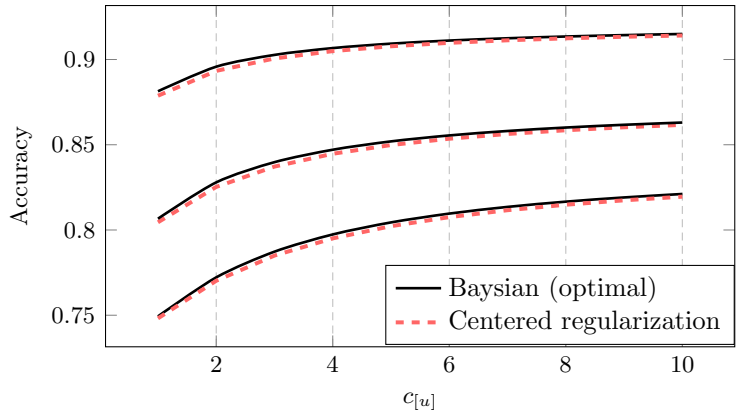

Figure 5: Asymptotic accuracy on isotropic Gaussian mixture data. Performance curves as a function of $c_{[u]}$ with $c_{[l]} = 1/2$, for (from top to bottom) $\|\mu\|^2 = 2, 4/3,$ or 1.

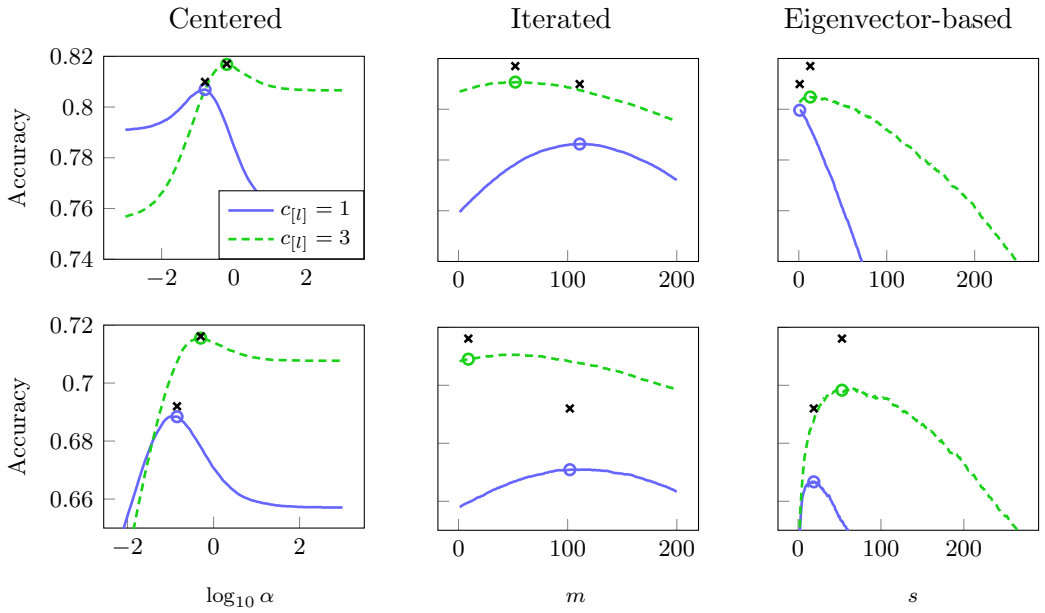

Figure 6: Empirical accuracy of graph-based SSL algorithms at different values of hyperparameters for isotropic Gaussian mixture data with $p = 60$, $n = 360$ and $\|\mu\|^2 = 1$ (bottom) or $\|\mu\|^2 = 2$ (top). Averaged over 1000 realizations. Best empirical value marked in circle and the asymptotic optimum in cross.

larization and its higher-order variant on several benchmark data sets. Our experiments support the benefit of centered similarities, especially on challenging data sets. We find importantly that the proposed higher-order centered regularization, which combines the ideas of centered similarities and iterated regularization, yields a very competitive performance when compared to not only related graph-based methods but also a wide range of popular SSL approaches.

6.3.1 Higher-Order Centered Regularization

It is easy to check that all exponents of a centered simliarity matrix $\hat{W}$ enjoy the same property of being orthogonal to the constant vector 1_n. It is thus possible to apply polynomial functions of $\hat{W}$ in the search of a better centered regularizer. Finding the right parametrization for the polynomial function can be challenging. Comparing (2) and (9), we notice that the matrix $\hat{L} = \lambda I_n - \hat{W}$ plays the same role in centered regularization as the Laplacian matrix L in Laplacian regularization. We then borrow the idea from iterated Laplacian to propose a higher-order centered regularization method, which consists in simply replacing $\hat{L} = \lambda I_n - \hat{W}$ in (9) with its exponents $\hat{L}^{(q)} = (\lambda I_n - \hat{W})^q$, leading to

$$f_{[u]} = -\hat{L}^{(q)-1}_{[uu]} \hat{L}^{(q)}_{[ul]} f_{[l]} \tag{17}$$

The method of iterated centered regularization is formalized in Algorithm 2.

Algorithm 2 Graph-Based Iterated Centered Regularization

1: **Input:** $n_{[l]}$ pairs of labelled points and labels $\{(x_i, y_i)\}_{i=1}^{n_{[l]}}$, $n_{[u]}$ unlabelled data $\{x_i\}_{i=n_{[l]}+1}^{n}$, parameters $\alpha \in \mathbb{R}^+$, $q \in \mathbb{N}^+$.
2: **Output:** Classification score vector of unlabelled data $f_{[u]} \in \mathbb{R}^{n_{[u]}}$
3: Define the similarity matrix $W = \{w_{i,j}\}_{i,j=1}^{n}$ with w_{ij} reflecting the closeness between x_i and x_j.
4: Compute the centered similarity matrix $\hat{W}$ by (7) and the balanced labels $f_{[l]}$ by (4).
5: Set $\lambda = (\alpha + 1)\|W\|$, $\hat{L}^{(q)} = (\lambda I_n - \hat{W})^q$, and compute $f_{[u]}$ by (17).

6.3.2 Experimental Results and Comparison to Other Methods

To investigate the practical potential of centered similarities, we test in this section the proposed methods of standard and higher-order centered regularization on several benchmark data. We first focus on the comparison with the related graph-based SSL methods presented in Section 6.1, before moving on to a wider range of competitors. Our results attest to the general interest of centered similarities for improving over the classical Laplacian approach. Remarkably, the algorithm of iterated centered regularization, which inherits the strengths of both iterated and centered approaches, is identified as a powerful competitor not only against other graph-based algorithms but also against a wide range of SSL methods surveyed by Chapelle et al. (2010), with an advantage which tends to be more observable on challenging data sets.

Firstly, to compare graph-based SSL methods, we conduct experiments on the MINST data of handwritten digits and the RCV1 data of categorized newswire stories. For MNIST

n	$N/6$	$N/2$	N
		MNIST	
Laplacian	53.9 ± 11.2	44.7 ± 10.6	33.5 ± 14.0
Centered	81.4 ± 2.6	85.3 ± 2.5	86.3 ± 3.9
Iterated Laplacian	81.7 ± 3.5	83.7 ± 3.6	86.4 ± 5.5
Iterated Centered	$\mathbf{83.2 \pm 2.7}$	$\mathbf{86.8 \pm 3.7}$	$\mathbf{88.4 \pm 3.2}$
ℓ_p-based Laplacian	72.7 ± 3.5	75.9 ± 4.0	76.0 ± 2.7
WN Laplacian	78.0 ± 3.2	79.0 ± 4.3	79.1 ± 4.1
Eigenvector-based	79.9 ± 3.7	85.4 ± 4.2	86.6 ± 3.8
		RCV1	
Laplacian	36.8 ± 12.9	34.2 ± 9.7	33.4 ± 9.6
Centered	78.7 ± 2.7	79.0 ± 3.2	79.1 ± 2.3
Iterated Laplacian	81.0 ± 2.1	81.4 ± 3.2	80.8 ± 5.6
Iterated Centered	$\mathbf{83.8 \pm 2.4}$	$\mathbf{84.5 \pm 2.1}$	$\mathbf{84.8 \pm 2.7}$
WN Laplacian	70.1 ± 3.4	71.7 ± 4.9	71.8 ± 4.8
Eigenvector-based	81.8 ± 5.4	82.1 ± 4.5	82.8 ± 3.0

Table 1: Classification accuracy (%) of graph-based SSL algorithms averaged over 10 random splits, for $n_{[l]} = 5K$ with K the number of classes ($K = 10$ for MNIST, $K = 4$ for RCV1) and $n = \{N/6, N/2, N\}$ with N the total sample number ($N = 60000$ for MNIST, $N = 19000$ for RCV1).

data, we use, as in the experiments of Section 5.2, the raw version[4] of vectorized pixels (LeCun, 1998), and for RCV1 data, we retrieve from the paper of Cai and He (2011) a preprocessed four-class version[5] of tf-idf feature vectors. As suggested in the previous studies (Belkin and Niyogi, 2003; Zhou and Belkin, 2011; Johnson and Zhang, 2007), satisfying performance can be observed on MNIST data for KNN graphs with $k \sim 10$ and on RCV1 data for $k \sim 100$; we will test with k over $10 \times 2^{\{0,1,\ldots,6\}}$. To judge the potential of graph-based methods, we report the best model performance for each method. We test the three common versions of Laplacian matrices L, L_s, L_r presented in Section 2.1 for Laplacian methods. The hyperparameter q of higher-order regularization algorithms is tried on $2^{\{1,2,3\}}$, the hyperparameter s of the eigenvector-based method is searched among all possible values (i.e., all integers from 1 to $n_{[l]}$), and the hyperparameter α of centered regularization takes its values in $10^{\{-3,-2,-1,0\}}$. The results are displayed in Table 1[6], where the advantage of centered similarities is supported by performance gains over the Laplacian methods. We also find that the iterated technique is powerful for improving the performance of centering regularization, which alone can be insufficient for producing superior results.

To evaluate the competitiveness of centered regularization beyond the family of graph-based methods, we report its performance on SSL benchmark data sets[7] established by

4. Available for download at `http://yann.lecun.com/exdb/mnist/` .
5. Available for download at `http://www.cad.zju.edu.cn/home/dengcai/Data/TextData.html` .
6. The performance of the ℓ_p-based Laplacian is not reported on RCV1 data due to the unsatisfying performance at small k and the very slow computation at large k.
7. Available for download at `http://olivier.chapelle.cc/ssl-book/benchmarks.html` .

	g241c	g241d	Digit1	USPS	BCI	Text
Laplacian+CMN (Chapelle et al., 2010)	77.95	71.80	96.85	93.64	53.78	74.29
Best over 13 Algo (Chapelle et al., 2010)	86.51	**95.05**	**97.56**	95.32	68.64	**76.91**
Iterated Laplacian (Zhou and Belkin, 2011)	85.18	89.45	**97.78**	**96.04**	56.22	74.23
Iterated Centered	**87.18**	86.31	**97.50**	94.40	**70.78**	**76.96**

Table 2: Classification accuracy (%) on SSL benchmarks with $n_{[l]} = 100$ and $n = 1500$, averaged over 12 random splits.

Chapelle et al. (2010) and tested for an extensive variety of SSL methods including Laplacian regularization (with a class mass normalization technique (CMN) for performance improvement). The same trials were also carried out by Zhou and Belkin (2011) for the method of iterated Laplacian regularization.

To produce our results, we simply use a KNN graph with $k = 10$ on the "Digit1" and "USPS"[8] tasks of image data; for the rest, we switch to fully connected graphs $w_{ij} = \exp(-\|x_i - x_j\|^2/2\sigma^2)$ as in the experiments of Chapelle et al. (2010); Zhou and Belkin (2011). Also following the settings for the Laplacian and iterated Laplacian methods, the hyperparameters are determined by a (10-fold) cross-validation on the first split of each data set, searched over the grid $\sigma = \{d/3, 3d\}$ for d the average pairwise distance, $q = \{2^1, 2^4, 2^7, 2^{10}\}$, and with α set to 10^{-3}. Our results are reported in Table 2, along with the performances of the iterated Laplacian and Laplacian regularization methods obtained by Chapelle et al. (2010); Zhou and Belkin (2011), as well as the best performance over 13 algorithms tested by Chapelle et al. (2010).

As can be observed in Table 2, the combined approach of iterated and centering techniques has a remarkable competitiveness overall. We note in particular its superiority on the three most difficult tasks "g241c", "BCI" and "Text" with lowest best accuracy, where the iterated technique alone fails to approach the best performance over 13 algorithms. The task on which the proposed method yields comparably worst results is "g241d", observed to be quite unstable with huge gaps between the best performing algorithm and the rest (Chapelle et al., 2010).

6.4 Computational Cost on Sparse Graphs

Sparse graphs such as KNN graphs are commonly used in graph-based learning. As our proposed algorithm involves a centering operation on the weight matrix W, it disrupts the sparsity of W and may cause increased computational cost in comparison to the original Laplacian approach. We would like to point out that, even though the centered weight matrix $\hat{W}$ is not sparse, it can be written as a sum of W and a matrix of rank two:

8. The KNN graph is kept directed for USPS data to cope with the effect of imbalanced classes. And for the same reason we conduct a k-means clustering of the classification scores in order to decide the affinity group of unlabelled data.

$$\hat{W} = W + \begin{bmatrix} 1_n & v \end{bmatrix} A \begin{bmatrix} 1_n^\mathsf{T} \\ v^\mathsf{T} \end{bmatrix}$$

where $v = W1_n$ and $A = \begin{bmatrix} (1_n^\mathsf{T} W 1_n)/n^2 & -1/n \\ -1/n & 0 \end{bmatrix}$. Using Woodbury's inversion formula, we can then decompose the inverse of $\lambda I_{n_{[u]}} - \hat{W}_{[uu]}$ as the inverse of $\lambda I_{n_{[u]}} - W_{[uu]}$ plus a matrix of rank two as:

$$\left(\lambda I_{n_{[u]}} - \hat{W}_{[uu]}\right)^{-1} = Q - Q \begin{bmatrix} 1_{n_{[u]}} & v_{[u]} \end{bmatrix} \left(A^{-1} + \begin{bmatrix} 1_{n_{[u]}}^\mathsf{T} \\ v_{[u]}^\mathsf{T} \end{bmatrix} Q \begin{bmatrix} 1_{n_{[u]}} & v_{[u]} \end{bmatrix} \right)^{-1} \begin{bmatrix} 1_{n_{[u]}}^\mathsf{T} \\ v_{[u]}^\mathsf{T} \end{bmatrix} Q$$

where $Q = (\lambda I_{n_{[u]}} - W_{[uu]})^{-1}$. Therefore, the complexity of computing the solution of centered regularization can be reduced to that of computing $QW_{[ul]}f_{[l]}$, which benefits from the sparsity of W. Similar reasoning can also be made for the iterated regularization algorithm presented in section 6.3.1.

7. Concluding Remarks

The key to the proposed semi-supervised learning method lies in the replacement of conventional Laplacian regularizers by a new graph smoothness regularizer with centered similarities. The motivation is rooted in the failure of the Laplacian regularization approach to learn effectively from unlabelled data due to the concentration of pairwise distances between high dimensional data vectors. As anticipated by our theoretical results, the proposed centered regularization method produced large performance gains over the standard Laplacian regularization method when the aforementioned distance concentration problem is severe, thanks to an effective learning from both labelled and unlabelled data learning. Moreover, it was observed that the proposed algorithm can further improve the performance even on data sets with weak distance concentration, for which the standard Laplacian approach exhibits a clear performance growth with respect to unlabelled data. Other than improving upon the Laplacian regularization method, the proposed algorithm is also theoretically proven to enjoy a near-optimal performance of semi-supervised learning on isotropic Gaussian mixture data. From a general perspective, the extended advantage of centered regularization beyond the initial motivation suggests that placing oneself under the challenging setting of high dimensional learning can help identify and address some underlying issues compromising the performance of popular learning heuristics.

As the usage of centered similarities constitutes a fundamental update to the classical Laplacian regularization approach, it would be interesting to investigate whether other algorithms involving Laplacian regularizers benefit from the same update. In this article we proposed additionally a higher-order version of centered regularization based on the iterated Laplacian method (Zhou and Belkin, 2011). The empirical analysis on various real-world data sets showed that adopting centered similarities can improve substantially the performance on difficult tasks where the Laplacian methods are observed to underperform. Future studies can be devoted to analysing and adapting more involved methods like Laplacian support vector machines (Belkin et al., 2006), which combines the optimization

of SVMs with the Laplacian regularization approach. The absence of explicit solutions in the method of Laplacian SVMs calls for additional technical tools to conduct similar precise performance analyses to the present one, such as the leave-one-out procedure devised in the work of El Karoui et al. (2013) to deal with implicit optimization solutions.

8. Acknowledgements

Couillet's work is partially supported by the ANR-MIAI Large-DATA chair at University Grenoble-Alpes, and the HUAWEI LarDist project.

Appendix A. Generalization of Theorem 3 and Proof

A.1 Generalized Theorem

We first present an extended version of Theorem 3 for the general setting where C_1 may differ from C_2. The functions $m(\xi)$, $\sigma^2(\xi)$ defined in (11) and (12) for describing the statistical distribution of unlabelled scores in the case of $C_1 = C_2$ need be adapted as follows:

$$m(\xi) = \frac{2c_{[l]}\theta(\xi)}{c_{[u]}\big(1-\theta(\xi)\big)} \tag{18}$$

$$\sigma^2(\xi) = \frac{\rho_1\rho_2(2c_{[l]} + m(\xi)c_{[u]})^2 s(\xi) + \rho_1\rho_2(4c_l + m(\xi)^2 c_{[u]})\omega(\xi)}{c_{[u]}\left(c_{[u]} - \omega(\xi)\right)}, \quad k \in \{1,2\} \tag{19}$$

where

$$\begin{aligned}
\theta(\xi) &= \rho_1\rho_2\xi(\nu_1-\nu_2)^\mathsf{T}\left(I_p - \xi\bar{\Sigma}\right)^{-1}(\nu_1-\nu_2)\\
\omega(\xi) &= \xi^2 p^{-1}\operatorname{tr}\left[\left(I_p - \xi\bar{\Sigma}\right)^{-1}\bar{\Sigma}\right]^2\\
s(\xi) &= \rho_1\rho_2\xi^2(\nu_1-\nu_2)^\mathsf{T}\left(I_p - \xi\bar{\Sigma}\right)^{-1}\bar{\Sigma}\left(I_p - \xi\bar{\Sigma}\right)^{-1}(\nu_1-\nu_2),
\end{aligned} \tag{20}$$

with

$$\begin{aligned}
\nu_k &= \left[\sqrt{-2h'(\tau)}\mu_k^\mathsf{T} \quad \sqrt{h''(\tau)}\operatorname{tr} C_k/\sqrt{p}\right]^\mathsf{T}\\
\Sigma_k &= \begin{bmatrix} -2h'(\tau)C_k & 0_{p\times 1} \\ 0_{1\times p} & 2h''(\tau)\operatorname{tr} {C_k}^2/p \end{bmatrix}
\end{aligned}$$

and $\bar{\Sigma} = \rho_1\Sigma_1 + \rho_2\Sigma_2$.

Notice that the adaptation is made here through the redefinitions of $\theta(\xi)$, $\omega(\xi$ and $s(\xi)$; the expressions of $m(\xi)$ and $\sigma^2(\xi)$ are kept identical. As in the case of $C_1 = C_2$, the positive functions $m(\xi)$ and $\sigma^2(\xi)$ are defined respectively on the domains $(0, \xi_m)$ and $(0, \xi_{\sigma^2})$ with $\xi_m, \xi_{\sigma^2} > 0$ uniquely given by $\theta(\xi_m) = 1$ and $\omega(\xi_{\sigma^2}) = c_{[u]}$. We define ξ_{sup} as

$$\xi_{\text{sup}} = \min\{\xi_m, \xi_{\sigma^2}\} \tag{21}$$

With these adapted notations, we present the generalized results in the theorem below.

Theorem 8 *Let Assumption 1 hold, the function h of (1) be three-times continuously differentiable in a neighborhood of τ, $f_{[u]}$ be the solution of (8) with fixed norm $n_{[u]}e^2$, and with the notations of $m(\xi)$, $\sigma^2(\xi)$, $\xi_{\sup}$ given in (18), (19), (21). Then, for $n_{[l]}+1 \leq i \leq n$ (i.e., x_i unlabelled) and $x_i \in \mathcal{C}_k$,*

$$f_i \xrightarrow{\mathcal{L}} \mathcal{N}\left((-1)^k(1-\rho_k)\hat{m}, \hat{\sigma}_k^2\right)$$

where

$$\begin{aligned}\hat{m} &= m(\xi_e)\\ \hat{\sigma}_k^2 &= c_{[u]}^{-2}\rho_1\rho_2\left[(2c_{[l]}+m(\xi_e)c_{[u]})^2 s_k(\xi_e) + (4c_l + m(\xi_e)^2 c_{[u]} + \sigma^2(\xi_e)c_{[u]})\omega(\xi_e)\right]\end{aligned}$$

with

$$s_a(\xi_e) = \rho_1\rho_2\xi_e^2(\nu_1-\nu_2)^\mathsf{T}\left(I_p - \xi_e\bar{\Sigma}\right)^{-1}\Sigma_k\left(I_p-\xi_e\bar{\Sigma}\right)^{-1}(\nu_1-\nu_2), \quad a\in\{1,2\},$$

and $\xi_e \in (0, \xi_{\sup})$ uniquely given by

$$\rho_1\rho_2 m(\xi_e)^2 + \sigma^2(\xi_e) = e^2.$$

A.2 Proof of Generalized Theorem

The proof of Theorem 8 relies on a leave-one-out approach, in the spirit of El Karoui et al. (2013), along with arguments from previous related analyses (Couillet and Benaych-Georges, 2016; Mai and Couillet, 2018) based on random matrix theory .

A.2.1 Main Idea

The main idea of the proof is to first demonstrate that for unlabelled data scores f_i (i.e., with $i > n_{[l]}$),

$$f_i = \gamma\beta^{(i)\mathsf{T}}\phi_c(x_i) + o_P(1) \tag{22}$$

where γ is a finite constant, ϕ_c a certain mapping from the data space that we shall define, and $\beta^{(i)}$ a random vector independent of $\phi_c(x_i)$. Additionally, we shall show that

$$\beta^{(i)} = \frac{1}{p}\sum_{j=1}^{n} f_j\phi_c(x_j) + \epsilon \tag{23}$$

with $\|\epsilon\|/\|\beta^{(i)}\| = o_P(1)$.

As a consequence of (22), the statistical behavior of the unlabelled data scores can be understood through that of $\beta^{(i)}$, which itself depends on the unlabelled data scores as described by (23). By combining (22) and (23), we thus establish the equations ruling the asymptotic statistical behavior (i.e., mean and variance) of the unlabelled data scores f_i.

A.2.2 Detailed Arguments

In addition to the notations given in the end of the introduction (Section 1), we specify that when multidimensional objects are concerned, $O(u_n)$ is understood entry-wise. The notation $O_{\|\cdot\|}$ is understood as follows: for a vector v, $v = O_{\|\cdot\|}(u_n)$ means its Euclidean norm is $O(u_n)$ and for a square matrix M, $M = O_{\|\cdot\|}(u_n)$ means that the operator norm of M is $O(u_n)$.

First note that, as $w_{ij} = h(\|x_i - x_j\|^2/p) = h(\tau) + O(p^{-\frac{1}{2}})$, Taylor-expanding w_{ij} around $h(\tau)$ gives (see Appendix C for a detailed proof) $\hat{W} = O_{\|\cdot\|}(1)$ and

$$\hat{W} = \frac{1}{p}\hat{\Phi}^{\mathsf{T}}\hat{\Phi} + [h(0) - h(\tau) + \tau h'(\tau)]P_n + O_{\|\cdot\|}(p^{-\frac{1}{2}}) \tag{24}$$

where $P_n = I_n - \frac{1}{n}1_n 1_n^{\mathsf{T}}$, and $\hat{\Phi} = [\hat{\phi}(x_1), \ldots, \hat{\phi}(x_n)] = [\phi(x_1), \ldots, \phi(x_n)]P_n$ with

$$\phi(x_i) = \left[\sqrt{-2h'(\tau)}x_i^{\mathsf{T}} \quad \sqrt{h''(\tau)}\|x_i\|^2/\sqrt{p}\right]^{\mathsf{T}}.$$

Define $\nu_k = \mathbb{E}\{\phi(x_i)\}$, $\Sigma_k = \mathrm{cov}\{\phi(x_i)\}$ for $x_i \in \mathcal{C}_k$, $k \in \{1, 2\}$, and let $Z = [z_1, \ldots, z_n]$ with $z_i = \phi(x_i) - \nu_k$ (i.e., $\mathbb{E}\{z_i\} = 0$). We also write the labelled versus unlabelled divisions $\Phi = \begin{bmatrix}\Phi_{[l]} & \Phi_{[u]}\end{bmatrix}$, $Z = \begin{bmatrix}Z_{[l]} & Z_{[u]}\end{bmatrix}$ and $\hat{\Phi} = \begin{bmatrix}\hat{\Phi}_{[l]} & \hat{\Phi}_{[u]}\end{bmatrix}$.

Recall that $f_{[u]} = \left(\lambda I_{n_{[u]}} - \hat{W}_{[uu]}\right)^{-1}\hat{W}_{[ul]}f_{[l]}$. To proceed, we need to show that $\frac{1}{n}1_{n_{[u]}}^{\mathsf{T}}f_{[u]} = O(p^{-\frac{1}{2}})$. Applying (24), we can express $f_{[u]}$ as

$$f_{[u]} = \left(\tilde{\lambda}I_{n_{[u]}} - \frac{1}{p}\hat{\Phi}_{[u]}^{\mathsf{T}}\hat{\Phi}_{[u]} + \frac{r}{n}1_{n_{[u]}}1_{n_{[u]}}^{\mathsf{T}}\right)^{-1}\left(\frac{1}{p}\hat{\Phi}_{[u]}^{\mathsf{T}}\hat{\Phi}_{[l]} - \frac{r}{n}1_{n_{[u]}}1_{n_{[l]}}^{\mathsf{T}}\right)f_{[l]} + O(p^{-\frac{1}{2}})$$

where $\tilde{\lambda} = \lambda - h(0) + h(\tau) - \tau h'(\tau)$, $r = h(0) - h(\tau) + \tau h'(\tau)$. Since $1_{[l]}^{\mathsf{T}}f_{[l]} = 0$ from its definition given in (4),

$$f_{[u]} = \left(\tilde{\lambda}I_{n_{[u]}} - \frac{1}{p}\hat{\Phi}_{[u]}^{\mathsf{T}}\hat{\Phi}_{[u]} + \frac{r}{n}1_{n_{[u]}}1_{n_{[u]}}^{\mathsf{T}}\right)^{-1}\frac{1}{p}\hat{\Phi}_{[u]}^{\mathsf{T}}\Phi_{[l]}f_{[l]} + O(p^{-\frac{1}{2}}). \tag{25}$$

Write $\hat{\Phi}_{[u]} = \mathbb{E}\{\hat{\Phi}_{[u]}\} + Z_{[u]} - (Z1_n/n)1_{n_{[u]}}^{\mathsf{T}}$. Evidently, $\mathbb{E}\{\hat{\Phi}_{[u]}\} = (\nu_1 - \nu_2)s^{\mathsf{T}}$ where $s \in \mathbb{R}^{n_{[u]}}$ with $s_i = (-1)^k(n - n_k)/n$ for $x_i \in \mathcal{C}_k$, $k \in \{1, 2\}$. By the large number law, $s = \zeta + O(p^{-\frac{1}{2}})$ where $\zeta \in \mathbb{R}^{n_{[u]}}$ with $\zeta_i = (-1)^k(1 - \rho_k)$ for $x_i \in \mathcal{C}_k$, therefore

$$\begin{aligned}\frac{1}{p}\hat{\Phi}_{[u]}^{\mathsf{T}}\hat{\Phi}_{[u]} =& \frac{1}{p}\Big\{\|\nu_1 - \nu_2\|^2\zeta\zeta^{\mathsf{T}} + Z_{[u]}^{\mathsf{T}}Z_{[u]} + (1_n^{\mathsf{T}}Z^{\mathsf{T}}Z1_n/n^2)1_{n_{[u]}}1_{n_{[u]}}^{\mathsf{T}} + [Z_{[u]}^{\mathsf{T}}(\nu_1 - \nu_2)]\zeta^{\mathsf{T}} \\ &+ \zeta[Z_{[u]}^{\mathsf{T}}(\nu_1 - \nu_2)]^{\mathsf{T}} - (Z_{[u]}^{\mathsf{T}}Z1_n/n)1_{n_{[u]}}^{\mathsf{T}} - 1_{n_{[u]}}(Z_{[u]}^{\mathsf{T}}Z1_n/n)^{\mathsf{T}}\Big\} + O_{\|\cdot\|}(p^{-\frac{1}{2}}).\end{aligned}$$

Invoking Woodbury's identity (Woodbury, 1950) expressed as

$$\left(R - UNU^{\mathsf{T}}\right)^{-1} = R + RU(N^{-1} - U^{\mathsf{T}}RU)^{-1}U^{\mathsf{T}}R,$$

we get

$$\left(\tilde{\lambda} I_{n_{[u]}} - \frac{1}{p}\hat{\Phi}_{[u]}^{\mathsf{T}}\hat{\Phi}_{[u]} + \frac{r}{n}1_{n_{[u]}}1_{n_{[u]}}^{\mathsf{T}}\right)^{-1} = \left(\tilde{\lambda} I_{n_{[u]}} - \frac{1}{p}Z_{[u]}^{\mathsf{T}}Z_{[u]} - UNU^{\mathsf{T}}\right)^{-1} + O_{\|\cdot\|}(p^{-\frac{1}{2}})$$
$$= R + RU(N^{-1} - U^{\mathsf{T}}RU)^{-1}U^{\mathsf{T}}R + O_{\|\cdot\|}(p^{-\frac{1}{2}}) \quad (26)$$

by letting $R = \left(\tilde{\lambda} I_{n_{[u]}} - \frac{1}{p}Z_{[u]}^{\mathsf{T}}Z_{[u]}\right)^{-1}$ and

$$U = \frac{1}{\sqrt{p}}\begin{bmatrix}\zeta & Z_{[u]}^{\mathsf{T}}(\nu_1 - \nu_2) & 1_{n_{[u]}} & Z_{[u]}^{\mathsf{T}}Z1_n/n\end{bmatrix}$$
$$N = \begin{bmatrix}\|\nu_1 - \nu_2\|^2 & 1 & 0 & 0\\ 1 & 0 & 0 & 0\\ 0 & 0 & (1_n^{\mathsf{T}}Z^{\mathsf{T}}Z1_n/n^2) - \frac{r}{c_0} & -1\\ 0 & 0 & -1 & 0\end{bmatrix}. \quad (27)$$

Note also that

$$\frac{1}{p}\hat{\Phi}_{[u]}^{\mathsf{T}}\Phi_{[l]}f_{[l]} = \sqrt{p}U\begin{bmatrix}(\nu_2 - \nu_1)^{\mathsf{T}}\frac{1}{p}\Phi_{[l]}f_{[l]}\\ 2c_{[l]}\rho_1\rho_2\\ 0\\ 0\end{bmatrix} + \frac{1}{p}Z_{[u]}^{\mathsf{T}}Z_{[l]}f_{[l]} + O(p^{-\frac{1}{2}}). \quad (28)$$

Now we want to prove that $U^{\mathsf{T}}RU$ is of the form

$$U^{\mathsf{T}}RU = \begin{bmatrix}A & 0_{2\times 2}\\ 0_{2\times 2} & B\end{bmatrix} + O(p^{-\frac{1}{2}}), \quad (29)$$

for some matrices $A, B \in \mathbb{R}^{2\times 2}$ with elements of $O(1)$. First it should be pointed out that z_i is a Gaussian vector if the last element is ignored. Since ignoring the last element of z_i will not change the concentration results given subsequently to prove the form of $U^{\mathsf{T}}RU$, we shall treat z_i as Gaussian vectors for simplicity. As there exists a deterministic matrix $\bar{R}$ of the form $cI_{n_{[u]}}$ such that

$$a^{\mathsf{T}}Rb - a^{\mathsf{T}}\bar{R}b = O(p^{-\frac{1}{2}})$$

for any $a, b = O_{\|\cdot\|}(1)$ independent of R (Benaych-Georges and Couillet, 2016, Proposition 5), we get immediately that

$$U_{\cdot 1}^{\mathsf{T}}RU_{\cdot 3} = \frac{1}{p}\zeta^{\mathsf{T}}R1_{n_{[u]}} = \frac{1}{p}\zeta^{\mathsf{T}}\bar{R}1_{n_{[u]}} + O(p^{-\frac{1}{2}}) = O(p^{-\frac{1}{2}}).$$

In order to prove the rest, we begin by showing that

$$\frac{1}{\sqrt{p}}a^{\mathsf{T}}Z_{[u]}Rb = O(p^{-\frac{1}{2}}) \quad (30)$$

for any $a, b = O_{\|\cdot\|}(1)$ independent of $Z_{[u]}$. First let us set $a' = \mathrm{Cov}\{z_i\}^{\frac{1}{2}} a$ and denote by $P_{a'}$ the projection matrix orthogonal to a'. We write then

$$\begin{aligned} z_i &= \mathrm{Cov}\{z_i\}^{\frac{1}{2}} P_{a'} \mathrm{Cov}\{z_i\}^{-\frac{1}{2}} z_i + \mathrm{Cov}\{z_i\}^{\frac{1}{2}} \frac{a' a'^{\mathsf{T}}}{\|a'\|^2} \mathrm{Cov}\{z_i\}^{-\frac{1}{2}} z_i \\ &= \tilde{z}_i + \frac{a^{\mathsf{T}} z_i}{\|a'\|^2} \mathrm{Cov}\{z_i\} a \end{aligned}$$

where $\tilde{z}_i = \mathrm{Cov}\{z_i\}^{\frac{1}{2}} P_{a'} \mathrm{Cov}\{z_i\}^{-\frac{1}{2}} z_i$. Note that in this decomposition of z_i, the two terms are independent. Indeed, since

$$\mathrm{Cov}\{\tilde{z}_i, a^{\mathsf{T}} z_i\} = \mathbb{E}\{\mathrm{Cov}\{z_i\}^{\frac{1}{2}} P_{a'} \mathrm{Cov}\{z_i\}^{-\frac{1}{2}} z_i z_i^{\mathsf{T}} a\} = \mathrm{Cov}\{z_i\}^{\frac{1}{2}} P_{a'} \mathrm{Cov}\{z_i\}^{\frac{1}{2}} a = 0_p,$$

$a^{\mathsf{T}} z_i$ and $\tilde{z}_i$ are uncorrelated, and thus independent by the property that uncorrelated jointly Gaussian variables are independent. Applying this decomposition of z_i, we have, by letting $\tilde{Z} = [\tilde{z}_1, \ldots, \tilde{z}_n]$ and $q = [a^{\mathsf{T}} z_1 \|\mathrm{Cov}\{z_1\} a\| / \|a'\|^2, \ldots, a^{\mathsf{T}} z_n \|\mathrm{Cov}\{z_n\} a\| / \|a'\|^2]$, that

$$Z_{[u]}^{\mathsf{T}} Z_{[u]} = \tilde{Z}_{[u]}^{\mathsf{T}} \tilde{Z}_{[u]} + q q^{\mathsf{T}}.$$

Then with the help of Sherman-Morrison's formula (Sherman and Morrison, 1950), we get

$$R = \tilde{R} - \frac{\tilde{R} q q^{\mathsf{T}} \tilde{R} / p}{1 + q^{\mathsf{T}} \tilde{R} q / p}.$$

Similarly to R, we have also for $\tilde{R}$ a deterministic equivalent $\bar{\tilde{R}} = \tilde{c} I_{n_{[u]}}$ with some constant $\tilde{c}$ such that

$$u^{\mathsf{T}} \tilde{R} v - u^{\mathsf{T}} \bar{\tilde{R}} v = O(p^{-\frac{1}{2}})$$

for any $u, v = O_{\|\cdot\|}(1)$ independent of $\tilde{R}$ (Benaych-Georges and Couillet, 2016, Proposition 5). Since $Z_{[u]}^{\mathsf{T}} a$ and q are independent of $\tilde{R}$, we prove $\frac{1}{\sqrt{p}} a^{\mathsf{T}} Z_{[u]} R b = O(p^{-\frac{1}{2}})$ with

$$\begin{aligned} \frac{1}{\sqrt{p}} a^{\mathsf{T}} Z_{[u]} R b &= \frac{1}{\sqrt{p}} a^{\mathsf{T}} Z_{[u]} \tilde{R} b - \frac{\frac{1}{\sqrt{p}} a^{\mathsf{T}} Z_{[u]} \tilde{R} q q^{\mathsf{T}} \tilde{R} b}{1 + q^{\mathsf{T}} \tilde{R} q} \\ &= \frac{1}{\sqrt{p}} \tilde{c} a^{\mathsf{T}} Z_{[u]} b - \frac{\frac{1}{\sqrt{p}} \tilde{c}^2 a^{\mathsf{T}} Z_{[u]} q q^{\mathsf{T}} b}{1 + \tilde{c} \|q\|^2} + O(p^{-\frac{1}{2}}) \\ &= O(p^{-\frac{1}{2}}). \end{aligned}$$

This leads directly to

$$U_{\cdot 2}^{\mathsf{T}} R U_{\cdot 3} = \frac{1}{\sqrt{p}} (\nu_1 - \nu_2) Z_{[u]} R 1_{n_{[u]}} / \sqrt{p} = O(p^{-\frac{1}{2}}).$$

With the same argument, we have also

$$\begin{aligned} U_{\cdot 1}^{\mathsf{T}} R U_{\cdot 4} &= \frac{1}{p} \zeta^{\mathsf{T}} R \left(Z_{[u]}^{\mathsf{T}} Z_{[u]} 1_{n_{[u]}} / n + Z_{[u]}^{\mathsf{T}} Z_{[l]} 1_{n_{[l]}} / n \right) \\ &= \zeta^{\mathsf{T}} (\tilde{\lambda} R - I_{n_{[u]}}) 1_{n_{[u]}} / n + \frac{1}{p} \zeta^{\mathsf{T}} R Z_{[u]}^{\mathsf{T}} \left(Z_{[l]} 1_{n_{[l]}} / n \right) = O(p^{-\frac{1}{2}}); \end{aligned}$$

and

$$\begin{aligned} U_{\cdot 2}^{\mathsf{T}} R U_{\cdot 4} =& \frac{1}{p}(\nu_1-\nu_2)^{\mathsf{T}} Z_{[u]} R \left(Z_{[u]}^{\mathsf{T}} Z_{[u]} 1_{n_{[u]}}/n + Z_{[u]}^{\mathsf{T}} Z_{[l]} 1_{n_{[l]}}/n \right) \\ =& \tilde{\lambda}(\nu_1-\nu_2)^{\mathsf{T}} Z_{[u]} R 1_{n_{[u]}}/n - (\nu_1-\nu_2)^{\mathsf{T}} Z_{[u]} 1_{n_{[u]}}/n \\ &+ (\nu_1-\nu_2)^{\mathsf{T}} \left(\frac{1}{p} Z_{[u]} R Z_{[u]}^{\mathsf{T}} \right) \left(Z_{[l]} 1_{n_{[l]}}/n \right) = O(p^{-\frac{1}{2}}). \end{aligned}$$

We conclude thus that $U^{\mathsf{T}} R U$ is of the form (29).

Substituting (26) and (28) into (25) and using the fact that $p^{-\frac{3}{2}} \| U^{\mathsf{T}} R Z_{[u]}^{\mathsf{T}} Z_{[l]} f_{[l]} \| = O(p^{-\frac{1}{2}})$ derived by similar reasoning to the above, we obtain

$$\frac{1}{n} 1_{n_{[u]}}^{\mathsf{T}} f_{[u]} = c_0^{-1} \begin{bmatrix} 0 & 0 & 1 & 0 \end{bmatrix} K \begin{bmatrix} (\nu_2-\nu_1)^{\mathsf{T}} \frac{1}{p} \Phi_{[l]} f_{[l]} \\ 2 c_{[l]} \rho_1 \rho_2 \\ 0 \\ 0 \end{bmatrix} + O(p^{-\frac{1}{2}}) \tag{31}$$

with

$$K = U^{\mathsf{T}} R U + U^{\mathsf{T}} R U (N^{-1} - U^{\mathsf{T}} R U)^{-1} U^{\mathsf{T}} R U.$$

Since $U^{\mathsf{T}} R U$ is of the form (29), we find from classical algebraic arguments that K is also of the same diagonal block matrix form. We thus finally get from (31) that

$$\frac{1}{n} 1_{n_{[u]}}^{\mathsf{T}} f_{[u]} = O(p^{-\frac{1}{2}}).$$

Now that we have shown that $\frac{1}{n} 1_{n_{[u]}}^{\mathsf{T}} f_{[u]} = O(p^{-\frac{1}{2}})$, multiplying both sides of (25) with $\tilde{\lambda} I_{n_{[u]}} - \frac{1}{p} \hat{\Phi}_{[u]}^{\mathsf{T}} \hat{\Phi}_{[u]} + \frac{r}{n} 1_{n_{[u]}} 1_{n_{[u]}}^{\mathsf{T}}$ from the left gives

$$\tilde{\lambda} f_{[u]} = \frac{1}{p} \hat{\Phi}_{[u]}^{\mathsf{T}} \hat{\Phi}_{[u]} f_{[u]} + \frac{1}{p} \hat{\Phi}_{[u]}^{\mathsf{T}} \hat{\Phi}_{[l]} f_{[l]} + O(p^{-\frac{1}{2}}).$$

Decomposing this equation for any $i > n_{[l]}$ (i.e., x_i unlabelled) leads to

$$\tilde{\lambda} f_i = \frac{1}{p} \hat{\phi}(x_i)^{\mathsf{T}} \hat{\Phi} f + O(p^{-\frac{1}{2}}) \tag{32}$$

$$\tilde{\lambda} f_{[u]}^{\{i\}} = \frac{1}{p} \hat{\Phi}_{[u]}^{\{i\}\mathsf{T}} \hat{\phi}(x_i) f_i + \frac{1}{p} \hat{\Phi}_{[u]}^{\{i\}\mathsf{T}} \hat{\Phi}_{[u]}^{\{i\}} f_{[u]}^{\{i\}} + \frac{1}{p} \hat{\Phi}_{[u]}^{\{i\}\mathsf{T}} \hat{\Phi}_{[l]} f_{[l]} + O(p^{-\frac{1}{2}}) \tag{33}$$

with $f_{[u]}^{\{i\}}$ standing for the vector obtained by removing f_i from $f_{[u]}$, $\hat{\Phi}_{[u]}^{\{i\}}$ for the matrix obtained by removing $\hat{\phi}(x_i)$ from $\hat{\Phi}_{[u]}$.

Our objective is to compare the behavior of the vector $f_{[u]}$ decomposed as $\{f_i, f_{[u]}^{\{i\}}\}$ to the "leave-x_i-out" version $f_{[u]}^{(i)}$ to be introduced next. To this end, define the leave-one-out data set $X^{(i)} = \{x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n\} \in \mathbb{R}^{(n-1)\times p}$ for any $i > n_{[l]}$ (i.e., x_i unlabelled),

and $\hat{W}^{(i)} \in \mathbb{R}^{(n-1)\times(n-1)}$ the corresponding centered similarity matrix, for which we have, similarly to $\hat{W}$,

$$\hat{W}^{(i)} = \frac{1}{p}\hat{\Phi}^{(i)\mathsf{T}}\hat{\Phi}^{(i)} + [h(0) - h(\tau) + \tau h'(\tau)]P_{n-1} + O_{\|\cdot\|}(p^{-\frac{1}{2}}) \tag{34}$$

where $\hat{\Phi}^{(i)} = [\hat{\phi}^{(i)}(x_1), \ldots, \hat{\phi}^{(i)}(x_{i-1}), \hat{\phi}^{(i)}(x_{i+1}), \ldots, \hat{\phi}^{(i)}(x_n)] = [\phi(x_1), \ldots, \phi(x_{i-1}), \phi(x_{i+1}), \ldots, \phi(x_n)]P_{n-1}$. Denote by $f^{(i)}_{[u]}$ the solution of the centered similarities regularization on the "leave-one-out" data set $X_{(i)}$, i.e.,

$$f^{(i)}_{[u]} = \left(\lambda I_{n_{[u]}-1} - \hat{W}^{(i)}_{[uu]}\right)^{-1} \hat{W}^{(i)}_{[ul]} f_{[l]}. \tag{35}$$

Substituting (34) into (35) leads to

$$\tilde{\lambda} f^{(i)}_{[u]} = \frac{1}{p}\hat{\Phi}^{(i)\mathsf{T}}_{[u]}\hat{\Phi}^{(i)}_{[u]} f^{(i)}_{[u]} + \frac{1}{p}\hat{\Phi}^{(i)\mathsf{T}}_{[u]}\hat{\Phi}_{[l]} f_{[l]} + O(p^{-\frac{1}{2}}) \tag{36}$$

where $\hat{\Phi}^{(i)} = \begin{bmatrix}\hat{\Phi}^{(i)}_{[l]} & \hat{\Phi}^{(i)}_{[u]}\end{bmatrix}$. From the definitions of $\hat{\Phi}^{(i)}_{[u]}$ and $\hat{\Phi}^{\{i\}}_{[u]}$, which essentially differ by the addition of the $O(1/\sqrt{p})$-norm term $\phi(x_i)/n$ to every column, we easily have

$$\frac{1}{\sqrt{p}}\hat{\Phi}^{(i)}_{[u]} - \frac{1}{\sqrt{p}}\hat{\Phi}^{\{i\}}_{[u]} = O_{\|\cdot\|}(p^{-1}),$$

which entails

$$\frac{1}{p}\hat{\Phi}^{(i)\mathsf{T}}_{[u]}\hat{\Phi}^{(i)}_{[u]} - \frac{1}{p}\hat{\Phi}^{\{i\}\mathsf{T}}_{[u]}\hat{\Phi}^{\{i\}}_{[u]} = O_{\|\cdot\|}(p^{-1}), \tag{37}$$

Thus, subtracting (36) from (33) gives

$$M^{(i)}\left(f^{\{i\}}_{[u]} - f^{(i)}_{[u]}\right) = \frac{1}{p}\hat{\Phi}^{(i)\mathsf{T}}_{[u]}\hat{\phi}(x_i) f_i + O(p^{-\frac{1}{2}}) \tag{38}$$

with

$$M^{(i)} = \tilde{\lambda} I_{(n_{[u]}-1)} - \frac{1}{p}\hat{\Phi}^{(i)\mathsf{T}}_{[u]}\hat{\Phi}^{(i)}_{[u]}.$$

Set $\beta = \frac{1}{p}\hat{\Phi} f = O_{\|\cdot\|}(1)$, the unlabelled data "regression vector" which gives unlabelled data scores by $f_i = \tilde{\lambda}^{-1}\beta^{\mathsf{T}}\hat{\phi}(x_i)$, and its "leave-one-out" version $\beta^{(i)} = \frac{1}{p}\hat{\Phi}^{(i)} f^{(i)}$ with $f^{(i)} = \begin{bmatrix} f_{[l]} & f^{(i)}_{[u]}\end{bmatrix}$. Applying (37) and (38), we get that

$$\beta - \beta^{(i)} = \left(I_p + \frac{1}{p}\hat{\Phi}^{(i)}_{[u]}\left(M^{(i)}\right)^{-1}\hat{\Phi}^{(i)\mathsf{T}}_{[u]}\right)\frac{1}{p} f_i \hat{\phi}(x_i) + O_{\|\cdot\|}(p^{-1}) = O_{\|\cdot\|}(p^{-\frac{1}{2}}). \tag{39}$$

By the above result, Equation (32) can be expanded as

$$\tilde{\lambda} f_i = \beta^{(i)\mathsf{T}}\hat{\phi}(x_i) + \frac{1}{p}\hat{\phi}(x_i)^{\mathsf{T}}\left(I_p + \frac{1}{p}\hat{\Phi}^{(i)}_{[u]}\left(M^{(i)}\right)^{-1}\hat{\Phi}^{(i)\mathsf{T}}_{[u]}\right)\hat{\phi}(x_i) f_i + O(p^{-\frac{1}{2}}). \tag{40}$$

To go further in the development of (40), we first need to evaluate the quadratic form

$$\kappa_i \equiv \frac{1}{p}\hat{\phi}(x_i)^\mathsf{T} T^{(i)} \hat{\phi}(x_i)$$

where

$$T^{(i)} = I_p + \frac{1}{p}\hat{\Phi}_{[u]}^{(i)} \left(M^{(i)}\right)^{-1} \hat{\Phi}_{[u]}^{(i)\mathsf{T}}.$$

Since $\frac{1}{p}\hat{\Phi}_{[u]}^{(i)\mathsf{T}}\hat{\Phi}_{[u]}^{(i)} = O_{\|\cdot\|}(1)$, it is easy to see that $T^{(i)} = O_{\|\cdot\|}(1)$. As $\hat{\phi}(x_i)$ is independent of $T^{(i)}$, it unfolds from the "trace lemma" (Couillet and Debbah, 2011, Theorem 3.4) that

$$\kappa_i - \frac{1}{p}\operatorname{tr}\Sigma_k T^{(i)} \xrightarrow{\text{a.s.}} 0.$$

Notice that

$$\begin{aligned} T^{(i)} &= \tilde{\lambda}\left(\tilde{\lambda} I_p - \frac{1}{p}\hat{\Phi}_{[u]}^{(i)}\hat{\Phi}_{[u]}^{(i)\mathsf{T}}\right)^{-1} = \tilde{\lambda}\left(\tilde{\lambda} I_p - \frac{1}{p}\hat{\Phi}_{[u]}^{\{i\}}\hat{\Phi}_{[u]}^{\{i\}\mathsf{T}}\right)^{-1} + O_{\|\cdot\|}(p^{-1}) \\ &= T - \frac{\frac{\tilde{\lambda}}{p}T^{(i)}\hat{\phi}(x_i)\hat{\phi}(x_i)^\mathsf{T}T^{(i)}}{1 - \frac{1}{p}\frac{\kappa_i}{\tilde{\lambda}}} + O_{\|\cdot\|}(p^{-1}) \end{aligned}$$

where

$$T = \tilde{\lambda}\left(\tilde{\lambda} I_p - \frac{1}{p}\hat{\Phi}_{[u]}\hat{\Phi}_{[u]}^\mathsf{T}\right)^{-1} = T^{(i)} + \frac{\frac{\tilde{\lambda}}{p}T^{(i)}\hat{\phi}(x_i)\hat{\phi}(x_i)^\mathsf{T}T^{(i)}}{1 - \frac{1}{p}\frac{\kappa_i}{\tilde{\lambda}}}$$

by Sherman-Morrison's formula (Sherman and Morrison, 1950). We get consequently

$$\frac{1}{p}\operatorname{tr}\Sigma_k T^{(i)} = \frac{1}{p}\operatorname{tr}\Sigma_k T + O(p^{-1}),$$

κ_i converges thus to a deterministic limit κ independent of i at large n, p.

Equation (40) then becomes

$$f_i = \gamma\beta^{(i)\mathsf{T}}\hat{\phi}(x_i) + O(p^{-\frac{1}{2}}). \tag{41}$$

where $\gamma = (\tilde{\lambda} - \kappa)^{-1}$.

We focus now on the term $\beta^{(i)\mathsf{T}}\hat{\phi}(x_i)$ in (41). To discard the "weak" dependence between $\beta^{(i)\mathsf{T}}$ and $\hat{\phi}(x_i)$, let us define

$$\phi_c(x_i) = (-1)^k(1-\rho_k)(\nu_2 - \nu_1) + z_i.$$

As $n_k/n = \rho_k + O(n^{-\frac{1}{2}})$, by the law of large numbers, $\mathbb{E}\{\hat{\phi}(x_i)\} = (-1)^k[(n-n_k)/n](\nu_2 - \nu_1) = \mathbb{E}\{\phi_c(x_i)\} + O_{\|\cdot\|}(n^{-\frac{1}{2}})$. Remark that, unlike $\hat{\phi}(x_i)$, $\phi_c(x_i)$ is independent of all x_j with $j \neq i$, and therefore independent of $\beta^{(i)}$. We thus now have

$$\beta^{(i)\mathsf{T}}\hat{\phi}(x_i) = \beta^{(i)\mathsf{T}}\left(\mathbb{E}\{\hat{\phi}(x_i)\} + z_i - \frac{1}{n}\sum_{m=1}^{n} z_m\right) = \beta^{(i)\mathsf{T}}\phi_c(x_i) + \frac{1}{n}\beta^\mathsf{T} Z 1_n + O(p^{-\frac{1}{2}}).$$

We get from (39) that $\frac{1}{n}\beta^{(i)\mathsf{T}} Z 1_n = \frac{1}{n}\beta^{\mathsf{T}} Z 1_n + O(p^{-\frac{1}{2}})$, leading to

$$f_i = \gamma\beta^{(i)\mathsf{T}}\phi_c(x_i) + \frac{1}{n}\beta^{\mathsf{T}} Z 1_n + O(p^{-\frac{1}{2}}). \tag{42}$$

Since $\phi_c(x_i)$ is independent of $\beta^{(i)}$, according to the central limit theorem, $\beta^{(i)\mathsf{T}}\phi_c(x_i)$ asymptotically follows a Gaussian distribution.

To demonstrate that $\frac{1}{n}\beta^{\mathsf{T}} Z 1_n$ is negligibly small, notice fist that, by summing (42) for all $i > n_{[u]}$, we have

$$\frac{1}{n} 1_{n_{[u]}}^{\mathsf{T}} f_{[u]} = \frac{1}{n}\sum_{i=n_{[l]}+1}^{n} \beta^{(i)\mathsf{T}}\phi_c(x_i) + c_{[u]}(\beta^{(i)\mathsf{T}} Z 1_n/n) + O(p^{-\frac{1}{2}}).$$

Since $\frac{1}{n} 1_{n_{[u]}}^{\mathsf{T}} f_{[u]} = O(p^{-\frac{1}{2}})$, it suffices to prove $\frac{1}{n}\sum_{i=n_{[l]}+1}^{n} \beta^{(i)\mathsf{T}}\phi_c(x_i) = O(p^{-\frac{1}{2}})$ to consequently show that $\frac{1}{n}\beta^{\mathsf{T}} Z 1_n = O(p^{-\frac{1}{2}})$ from the above equation. To this end, we shall examine the correlation between $\beta^{(i)\mathsf{T}}\phi_c(x_i)$ and $\beta^{(j)\mathsf{T}}\phi_c(x_j)$ for $i \neq j > n_{[l]}$. Consider $\beta^{(ij)}, \hat{\Phi}_{[u]}^{(ij)}, M^{(ij)}$ obtained in the same way as $\beta^{(i)}, \hat{\Phi}_{[u]}^{(i)}, M^{(i)}$, but this time by leaving out the two unlabelled samples x_i, x_j. Similarly to (39), we have

$$\beta^{(i)} - \beta^{(ij)} = \left(I_p + \frac{1}{p}\hat{\Phi}_{[u]}^{(ij)}\left(M^{(ij)}\right)^{-1}\hat{\Phi}_{[u]}^{(ij)\mathsf{T}}\right)\frac{1}{p} f_j \hat{\phi}(x_j) + O_{\|\cdot\|}(p^{-1}) = O_{\|\cdot\|}(p^{-\frac{1}{2}}). \tag{43}$$

It follows from the above equation that, for $i \neq j > n_{[l]}$,

$$\begin{aligned}
&\mathrm{Cov}\{\beta^{(i)\mathsf{T}}\phi_c(x_i), \beta^{(i)\mathsf{T}}\phi_c(x_j)\} \\
&= \mathbb{E}\{\beta^{(i)\mathsf{T}}\phi_c(x_i)\beta^{(i)\mathsf{T}}\phi_c(x_j)\} - \mathbb{E}\{\beta^{(i)\mathsf{T}}\phi_c(x_i)\}\mathbb{E}\{\beta^{(j)\mathsf{T}}\phi_c(x_j)\} \\
&= \mathbb{E}\{\beta^{(ij)\mathsf{T}}\phi_c(x_i)\beta^{(ij)\mathsf{T}}\phi_c(x_j)\} - \mathbb{E}\{\beta^{(i)\mathsf{T}}\phi_c(x_i)\}\mathbb{E}\{\beta^{(j)\mathsf{T}}\phi_c(x_j)\} + O(p^{-1}) \\
&= \mathbb{E}\{\beta^{(ij)\mathsf{T}}\phi_c(x_i)\}\mathbb{E}\{\beta^{(ij)\mathsf{T}}\phi_c(x_j)\} - \mathbb{E}\{\beta^{(i)\mathsf{T}}\phi_c(x_i)\}\mathbb{E}\{\beta^{(j)\mathsf{T}}\phi_c(x_j)\} + O(p^{-1}) \\
&= O(p^{-1}),
\end{aligned} \tag{44}$$

leading to the conclusion that $\frac{1}{n_{[u]}}\sum_{i=n_{[l]}+1}^{n} \beta^{(i)\mathsf{T}}\phi_c(x_i) = \frac{1}{n_{[u]}}\sum_{i=n_{[l]}+1}^{n} \mathbb{E}\{\beta^{(i)\mathsf{T}}\phi_c(x_i)\} + O(p^{-\frac{1}{2}}) = O(p^{-\frac{1}{2}})$. Hence, $\frac{1}{n}\beta^{\mathsf{T}} Z 1_n = O(p^{-\frac{1}{2}})$. Finally, we have that, for $i > n_{[l]}$,

$$f_i = \gamma\beta^{(i)\mathsf{T}}\phi_c(x_i) + O(p^{-\frac{1}{2}}), \tag{45}$$

indicating that, up to the constant γ, f_i asymptotically follows the same Gaussian distribution as $\beta^{(i)\mathsf{T}}\phi_c(x_i)$.

Moreover, taking the expectation and the variance of the both sides of (45) for $x_i \in \mathcal{C}_k$ yields

$$\begin{aligned}
\mathbb{E}\{f_i | i > n_{[l]}, x \in \mathcal{C}_k\} &= \gamma\mathbb{E}\{\beta^{(i)\mathsf{T}}\}(-1)^k(1-\rho_k)(\nu_2-\nu_1) + O(p^{-\frac{1}{2}}) \\
\mathrm{var}\{f_i | i > n_{[l]}, x \in \mathcal{C}_k\} &= \gamma^2 \mathrm{tr}\big[\mathrm{cov}\{\beta^{(i)}\}\Sigma_k\big] + \gamma^2\mathbb{E}\{\beta^{(i)}\}^{\mathsf{T}}\Sigma_k\mathbb{E}\{\beta^{(i)}\} + O(p^{-\frac{1}{2}}).
\end{aligned}$$

Since $\beta - \beta^{(i)} = O_{\|\cdot\|}(p^{-\frac{1}{2}})$ as per (39), we obtain

$$\mathbb{E}\{f_i | i > n_{[l]}, x \in \mathcal{C}_k\} = \gamma \mathbb{E}\{\beta^{\mathsf{T}}\}(-1)^k(1-\rho_k)(\nu_2-\nu_1) + O(p^{-\frac{1}{2}}) \tag{46}$$

$$\mathrm{var}\{f_i | i > n_{[l]}, x \in \mathcal{C}_k\} = \gamma^2 \mathrm{tr}\big[\mathrm{cov}\{\beta\}\Sigma_k\big] + \gamma^2 \mathbb{E}\{\beta\}^{\mathsf{T}}\Sigma_k \mathbb{E}\{\beta\} + O(p^{-\frac{1}{2}}). \tag{47}$$

After linking the distribution parameters of unlabelled scores to those of β with Equation (46) and Equation (47), we now turn our attention to the statistical behaviour of β. Substituting (45) into $\beta = \frac{1}{p}\hat{\Phi} f$ yields

$$\begin{aligned}\beta &= \frac{1}{p}\sum_{i=1}^{n_{[l]}} f_i \hat{\phi}(x_i) + \frac{1}{p}\sum_{i=n_{[l]}+1}^{n} \gamma \beta^{(i)\mathsf{T}} \phi_c(x_i)\hat{\phi}(x_i) + O_{\|\cdot\|}(p^{-\frac{1}{2}}) \\ &= \frac{1}{p}\sum_{i=1}^{n_{[l]}} f_i \phi_c(x_i) + \frac{1}{p}\sum_{i=n_{[l]}+1}^{n} \gamma \beta^{(i)\mathsf{T}} \phi_c(x_i)\phi_c(x_i) + O_{\|\cdot\|}(p^{-\frac{1}{2}}).\end{aligned} \tag{48}$$

For $i > n_{[l]}$ and $x_i \in \mathcal{C}_k$, we decompose $\phi_c(x_i)$ as

$$\phi_c(x_i) = \mathbb{E}\{\phi_c(x_i)\} + \frac{\Sigma_k \beta^{(i)}}{\beta^{(i)\mathsf{T}} z_i} + \tilde{z}_i \tag{49}$$

where

$$\tilde{z}_i = z_i - \frac{\Sigma_k \beta^{(i)}}{\beta^{(i)\mathsf{T}} z_i}.$$

By substituting the expression (49) of $\phi_c(x_i)$ into (48) and using the fact that $\beta - \beta^{(i)} = O_{\|\cdot\|}(p^{-\frac{1}{2}})$, we obtain

$$\begin{aligned}\left(I_p - \gamma c_{[u]} \sum_{a=1}^{2} \rho_a \Sigma_a\right)\beta &= \frac{1}{p}\sum_{i=1}^{n_{[l]}} f_i \mathbb{E}\{\phi_c(x_i)\} + \frac{1}{p}\sum_{i=n_{[l]}+1}^{n} \gamma \beta^{(i)\mathsf{T}} \phi_c(x_i)\mathbb{E}\{\phi_c(x_i)\} \\ &\quad + \frac{1}{p}\sum_{i=1}^{n_{[l]}} f_i z_i + \frac{1}{p}\sum_{i=n_{[l]}+1}^{n} \gamma \beta^{(i)\mathsf{T}} \phi_c(x_i)\tilde{z}_i + O_{\|\cdot\|}(p^{-\frac{1}{2}}).\end{aligned} \tag{50}$$

Recall that $f_{[l]}$ is a deterministic vector (given in (4)) and note that

$$\mathbb{E}\{\beta^{(i)\mathsf{T}}\phi_c(x_i)\tilde{z}_i\} = \mathbb{E}\{\beta^{(i)\mathsf{T}} z_i [z_i - \Sigma_k \beta^{(i)}/(\beta^{(i)\mathsf{T}} z_i)]\} = \mathbb{E}\{\beta^{(i)\mathsf{T}} z_i z_i\} - \Sigma_k \mathbb{E}\{\beta^{(i)}\} = 0.$$

Taking the expectation of both sides of (50) thus gives

$$\begin{aligned}&\left(I_p - \gamma c_{[u]} \sum_{a=1}^{2} \rho_a \Sigma_a\right)\mathbb{E}\{\beta\} \\ &= \frac{1}{p}\sum_{i=1}^{n_{[l]}} f_i \mathbb{E}\{\phi_c(x_i)\} + \frac{1}{p}\sum_{i=n_{[l]}+1}^{n} \gamma \mathbb{E}\{\beta^{(i)}\}^{\mathsf{T}}\mathbb{E}\{\phi_c(x_i)\}\mathbb{E}\{\phi_c(x_i)\} + O_{\|\cdot\|}(p^{-\frac{1}{2}}) \\ &= \frac{1}{p}\sum_{i=1}^{n_{[l]}} f_i \mathbb{E}\{\phi_c(x_i)\} + \frac{1}{p}\sum_{i=n_{[l]}+1}^{n} \gamma \mathbb{E}\{\beta\}^{\mathsf{T}}\mathbb{E}\{\phi_c(x_i)\}\mathbb{E}\{\phi_c(x_i)\} + O_{\|\cdot\|}(p^{-\frac{1}{2}}).\end{aligned} \tag{51}$$

Let $Q = I_p - \gamma c_{[u]}\bar{\Sigma}$ with $\bar{\Sigma} = \rho_1\Sigma_1 + \rho_2\Sigma_2$ and denote $\hat{m} \equiv \gamma(\nu_2 - \nu_1)^\mathsf{T}\mathbb{E}\{\beta\}$. With these notations, we get directly from the above equation that

$$\hat{m} = \gamma\rho_1\rho_2(2c_{[l]} + mc_{[u]})(\nu_2 - \nu_1)^\mathsf{T}Q^{-1}(\nu_2 - \nu_1) + o_P(1). \tag{52}$$

With the notation m, (46) notably becomes

$$\mathbb{E}\{f_i | i > n_{[l]}, x \in \mathcal{C}_k\} = (-1)^k(1 - \rho_k)\hat{m} + O(p^{-\frac{1}{2}}).$$

In addition, we get from (51) that

$$\gamma^2\mathbb{E}\{\beta\}^\mathsf{T}\Sigma_k\mathbb{E}\{\beta\} = \left[\gamma\rho_1\rho_2(2c_{[l]} + \hat{m}c_{[u]})\right]^2(\nu_2 - \nu_1)^\mathsf{T}Q^{-1}\Sigma_kQ^{-1}(\nu_2 - \nu_1). \tag{53}$$

Furthermore, we have from (50) and (51)

$$\begin{aligned}\mathrm{tr}[\mathrm{cov}\{\beta\}\Sigma_k] &= \mathbb{E}\left\{(\beta - \mathbb{E}\{\beta\})^\mathsf{T}\Sigma_k(\beta - \mathbb{E}\{\beta\})\right\} \\ &= \frac{1}{p^2}\sum_{i=1}^{n_{[l]}} f_i^2\mathbb{E}\{z_i^\mathsf{T}Q^{-1}\Sigma_kQ^{-1}z_i\} + \frac{1}{p^2}\sum_{i=n_{[l]}+1}^{n}\gamma^2\mathbb{E}\{(\beta^{(i)\mathsf{T}}\phi_c(x_i))^2\tilde{z}_i^\mathsf{T}Q^{-1}\Sigma_kQ^{-1}\tilde{z}_i\} \\ &\quad + O(p^{-\frac{1}{2}}).\end{aligned}$$

Since $\frac{1}{p}z_i^\mathsf{T}Q^{-1}\Sigma_kQ^{-1}z_i = \frac{1}{p}\mathrm{tr}(Q^{-1}\bar{\Sigma})^2 + O(p^{-\frac{1}{2}})$ and $\frac{1}{p}\tilde{z}_i^\mathsf{T}Q^{-1}\Sigma_kQ^{-1}\tilde{z}_i = \frac{1}{p}\mathrm{tr}(Q^{-1}\bar{\Sigma})^2 + O(p^{-\frac{1}{2}})$, by the trace lemma (Couillet and Debbah, 2011, Theorem 3.4) and Assumption 1,

$$\begin{aligned}\gamma^2\mathrm{tr}[\mathrm{cov}\{\beta\}\Sigma_k] =& \gamma^2\left[\rho_1\rho_2(4c_{[l]} + \hat{m}^2c_{[u]}) + c_{[u]}\sum_{a=1}^{2}\rho_a\mathrm{var}\{f_i | i > n_{[l]}, x \in \mathcal{C}_a\}\right]\frac{1}{p}\mathrm{tr}(Q^{-1}\bar{\Sigma})^2 \\ &+ O(p^{-\frac{1}{2}}).\end{aligned} \tag{54}$$

Using the shortcut notation $\hat{\sigma}_k^2 \equiv \mathrm{var}\{f_i | i > n_{[l]}, x \in \mathcal{C}_k\}$ for $k \in \{1, 2\}$, we get by substituting (53) and (54) into (47) that

$$\begin{aligned}\hat{\sigma}_k^2 =& \left[\gamma\rho_1\rho_2(2c_{[l]} + \hat{m}c_{[u]})\right]^2(\nu_2 - \nu_1)^\mathsf{T}Q^{-1}\Sigma_kQ^{-1}(\nu_2 - \nu_1) \\ &+ \gamma^2\left[\rho_1\rho_2(4c_{[l]} + \hat{m}^2c_{[u]}) + c_{[u]}\sum_{a=1}^{2}\rho_a\hat{\sigma}_a^2\right]\frac{1}{p}\mathrm{tr}(Q^{-1}\bar{\Sigma})^2 + o_P(1).\end{aligned} \tag{55}$$

Letting $\xi \equiv c_{[u]}\gamma$, we get by multiplying the both sides of (52) with $c_{[u]}$ that

$$c_{[u]}\hat{m} = \xi\rho_1\rho_2(2c_{[l]} + \hat{m}c_{[u]})(\nu_2 - \nu_1)^\mathsf{T}\left(I_p - \gamma c_{[u]}\bar{\Sigma}\right)^{-1}(\nu_2 - \nu_1) + o_P(1).$$

And multiplying the both sides of (55) with $c_{[u]}^2$ leads to

$$\begin{aligned}c_{[u]}^2\hat{\sigma}_k^2 =& \left[\rho_1\rho_2(2c_{[l]} + \hat{m}c_{[u]})\right]^2\xi^2(\nu_2 - \nu_1)^\mathsf{T}Q^{-1}\Sigma_kQ^{-1}(\nu_2 - \nu_1) \\ &+ \left[\rho_1\rho_2(4c_{[l]} + \hat{m}^2c_{[u]}) + c_{[u]}\sum_{a=1}^{2}\rho_a\hat{\sigma}_a^2\right]\xi^2p^{-1}\mathrm{tr}(Q^{-1}\bar{\Sigma})^2 + o_P(1).\end{aligned} \tag{56}$$

Set $\hat{\sigma}^2 = \sum_{a=1}^2 \rho_a \hat{\sigma}_a^2$, we obtain

$$c_{[u]}^2 \hat{\sigma}^2 = \left[\rho_1\rho_2(2c_{[l]} + \hat{m}c_{[u]})\right]^2 \xi^2 (\nu_2-\nu_1)^\mathsf{T} Q^{-1}\bar{\Sigma}Q^{-1}(\nu_2-\nu_1) + \left[\rho_1\rho_2(4c_{[l]} + \hat{m}^2 c_{[u]}) + c_{[u]}\hat{\sigma}^2\right]\xi^2 p^{-1}\mathrm{tr}(Q^{-1}\bar{\Sigma})^2 + o_P(1).$$

It is derived from the above equations that there exists a $\xi \in \mathbb{R}$ such that $(\hat{m}, \hat{\sigma}^2) = (m(\xi), \sigma^2(\xi))$ with $m(\xi), \sigma^2(\xi)$ as given in (18) and (19). Let us denote by ξ_e the value of ξ that allows us to access $\hat{m}, \hat{\sigma}^2$ at some given value of the hyperparameter $e > 0$ (which, as we recall, was introduced in (8)). Notice that, as a direct consequence of (44) and (45), we have

$$\mathrm{Cov}\{f_i, f_j\} = O(p^{-1})$$

for $i, j > n_{[l]}$. With the same arguments, we get easily

$$\mathrm{Cov}\{f_i^2, f_j^2\} = O(p^{-1}),$$

which entails

$$\frac{1}{n_{[u]}}\|f_{[u]}\|^2 = \frac{1}{n_{[u]}}\sum_{i=n_{[l]}+1}^{n} f_i^2 = \frac{1}{n_{[u]}}\sum_{i=n_{[l]}+1}^{n} \mathbb{E}\{f_i^2\} + O(p^{-\frac{1}{2}}) = \rho_1\rho_2 m^2 + \sigma^2 + O(p^{-\frac{1}{2}}).$$

Therefore, the value ξ_e should satisfy, up to some asymptotically negligible terms, the equation

$$\rho_1\rho_2 m(\xi_e)^2 + \sigma^2(\xi_e) = e^2.$$

Note that the above equation does not give an unique ξ_e if ξ_e is allowed to take any value in $\mathbb{R}$. We need thus to further specify the admissible range of ξ_e as e goes from zero to infinity. We start by showing that m has always a positive value. With small adjustment to (31), we have

$$\frac{1}{n}\zeta^\mathsf{T} f_{[u]} = c_0^{-1}\begin{bmatrix}1 & 0 & 0 & 0\end{bmatrix} K \begin{bmatrix}(\nu_2-\nu_1)^\mathsf{T}\frac{1}{p}\Phi_{[l]}f_{[l]} \\ 2c_{[l]}\rho_1\rho_2 \\ 0 \\ 0\end{bmatrix} + O(p^{-\frac{1}{2}})$$

with

$$K = U^\mathsf{T}RU + U^\mathsf{T}RU(N^{-1} - U^\mathsf{T}RU)^{-1}U^\mathsf{T}RU.$$

We recall $U^\mathsf{T}RU$ is of the form (29), and further remark that the matrix A in (29) is of the form $A = \begin{bmatrix}a_{11} & 0 \\ 0 & a_{22}\end{bmatrix}$ as we have $U_{\cdot 1}^\mathsf{T} R U_{\cdot 2}$ by applying (30). As indicated in Section 3.2, for any $e > 0$, λ has a value greater than which is determined by (10). The matrix $\left(\lambda I_{n_{[u]}} - \hat{W}_{[uu]}\right)^{-1}$ is thus definite positive. Since

$$\begin{aligned}\left(\lambda I_{n_{[u]}} - \hat{W}_{[uu]}\right)^{-1} &= \left(\tilde{\lambda} I_{n_{[u]}} - \frac{1}{p}\hat{\Phi}_{[u]}^\mathsf{T}\hat{\Phi}_{[u]} + \frac{r}{n}1_{n_{[u]}}1_{n_{[u]}}^\mathsf{T}\right)^{-1} \\ &= R + RU(N^{-1} - U^\mathsf{T}RU)^{-1}U^\mathsf{T}R + O_{\|\cdot\|}(p^{-\frac{1}{2}}),\end{aligned}$$

K is definite positive with high probability. Notice also that

$$K = U^\mathsf{T} RU + U^\mathsf{T} RU(N^{-1} - U^\mathsf{T} RU)^{-1} U^\mathsf{T} RU = \left[\left(U^\mathsf{T} RU\right)^{-1} - N\right]^{-1},$$

meaning that

$$K_{12} = \frac{N_{12}}{\det\left\{(U^\mathsf{T} RU)^{-1} - N\right\}} = \frac{1}{\det\left\{(U^\mathsf{T} RU)^{-1} - N\right\}}.$$

We get thus $K_{12} > 0$ since $\det\left\{\left(U^\mathsf{T} RU\right)^{-1} - N\right\} = \det\left\{K^{-1}\right\} > 0$ due to the definite positiveness of K, which implies that all the eigenvalues of K are positive. The fact that K is definite positive implies also $K_{11} > 0$, otherwise we would have $\begin{bmatrix}1 & 0\end{bmatrix} K \begin{bmatrix}1\\0\end{bmatrix} = K_{11} \leq 0$. Since

$$\begin{aligned}\frac{1}{n}\zeta^\mathsf{T} f_{[u]} &= c_0^{-1}\left(K_{11}(\nu_2 - \nu_1)^\mathsf{T}\frac{1}{p}\Phi_{[l]} f_{[l]} + K_{12} 2c_{[l]}\rho_1\rho_2\right) + O(p^{-\frac{1}{2}}) \\ &= 2\rho_1\rho_2(K_{11}\|\nu_2 - \nu_1\|^2 + K_{12} l n_{[l]}/n) + O(p^{-\frac{1}{2}}),\end{aligned}$$

we get $\frac{1}{n}\zeta^\mathsf{T} f_{[u]} > 0$ at large p. As

$$\frac{1}{n}\zeta^\mathsf{T} f_{[u]} = \frac{1}{n}\zeta^\mathsf{T}\mathbb{E}\{f_{[u]}\} + O(p^{-\frac{1}{2}}) = \rho_1\rho_2\hat{m} + O(p^{-\frac{1}{2}})$$

as a result of $\mathrm{Cov}\{f_i, f_j\} = O(p^{-1})$. We remark thus that $\hat{m} > 0$ holds asymptotically for any $e > 0$. Since $\sigma^2 > 0$ by definition, we have necessarily $\xi_e \in (0, \xi_{\sup})$ for any e, as at least one of $m(\xi_e), \sigma^2(\xi_e)$ is negative (or not well defined) outside this range. It can also be observed from the expressions (18)–(19) of $m(\xi_e)$ and $\sigma^2(\xi_e)$ that $\rho_1\rho_2 m^2(\xi) + \sigma^2(\xi)$ monotonously increases from zero to infinity as ξ increases from zero to $\xi_{\sup}$. Therefore, $\xi_e \in (0, \xi_{\sup})$ is uniquely given by

$$\rho_1\rho_2 m(\xi_e)^2 + \sigma^2(\xi_e) = e^2.$$

In summary, for any $e \in (0, +\infty)$, we have that $\hat{m} = m(\xi_e)$, $\hat{\sigma}^2 = \sigma^2(\xi_e)$ with functions $m(\xi), \sigma^2(\xi)$ as defined in (18)–(19) and $\xi_e \in (0, \xi_{\sup})$ the unique solution of $\rho_1\rho_2 m(\xi_e)^2 + \sigma^2(\xi_e) = e^2$; we get also from (56) the value of $\hat{\sigma}_k^2$ as

$$\begin{aligned}\hat{\sigma}_k^2 =& c_{[u]}^{-2}\left[\rho_1\rho_2(2c_{[l]} + m(\xi_e)c_{[u]})\right]^2 \xi_e^2 (\nu_2 - \nu_1)^\mathsf{T} Q^{-1}\Sigma_k Q^{-1}(\nu_2 - \nu_1) \\ &+ c_{[u]}^{-2}\left[\rho_1\rho_2(4c_{[l]} + m(\xi_e)^2 c_{[u]}) + c_{[u]}\sum_{a=1}^{2}\rho_a\sigma^2(\xi_e)_a\right]\xi_e^2 p^{-1}\mathrm{tr}(Q^{-1}\bar{\Sigma})^2\end{aligned}$$

The proof of theorem 8 is thus concluded.

Appendix B. Proof of Proposition 7

As the eigenvector of L_s associated with the smallest eigenvalue is $D^{\frac{1}{2}}1_n$, we consider

$$L'_s = nD^{-\frac{1}{2}}WD^{-\frac{1}{2}} - n\frac{D^{\frac{1}{2}}1_n 1_n^\mathsf{T} D^{\frac{1}{2}}}{1_n^\mathsf{T} D 1_n}.$$

Note that $\|L'_s\| = O(1)$ according to (Couillet and Benaych-Georges, 2016, Theorem 1), and if v is an eigenvector of L_s associated with the eigenvalue u, then it is also an eigenvector of L'_s associated with the eigenvalue $-u+1$, except for the eigenvalue-eigenvector pair $(n, D^{\frac{1}{2}}1_n)$ of L_s turned into $(0, D^{\frac{1}{2}}1_n)$ for L'_s. The second smallest eigenvector v_{Lap} of L_s is the same as the largest eigenvector of L'_s.

From the random matrix equivalent of L'_s given by Couillet and Benaych-Georges (2016, Theorem 1) and that of $\hat{W}$ expressed in (24), we have

$$\hat{W} = h(\tau)L'_s + \frac{5h'(\tau)^2}{4}\psi\psi^\mathsf{T} + O(p^{-\frac{1}{2}})$$

where $\psi = [\psi_1, \ldots, \psi_n]^\mathsf{T}$ with $\psi_i = \|x_i\|^2 - \mathbb{E}[\|x_i\|^2]$.

Recall that

$$\begin{aligned} d_{\text{inter}}(v) &= |j_1^\mathsf{T} v/n_1 - j_2^\mathsf{T} v/n_2| \\ d_{\text{intra}}(v) &= \|v - (j_1^\mathsf{T} v/n_1)j_1 - (j_2^\mathsf{T} v/n_2)j_2\|/\sqrt{n} \end{aligned}$$

for some $v \in \mathbb{R}^n$, and $j_k \in \mathbb{R}^n$ with $k \in \{1,2\}$ the indicator vector of class k with $[j_k]_i = 1$ if $x_i \in \mathcal{C}_k$, otherwise $[j_k]_i = 0$.

Denote by λ_{Lap} the eigenvalue of $h(\tau)L'_s$ associated with v_{Lap}, and λ_{ctr} the eigenvalue of $\hat{W}$ associated with v_{ctr}. Under the condition of non-trivial clustering upon v_{Lap} with $d_{\text{inter}}(v_{\text{Lap}})/d_{\text{intra}}(v_{\text{Lap}}) = O(1)$, we have $j_k^\mathsf{T} v_{\text{Lap}}/\sqrt{n_k} = O(1)$ from the above expressions of $d_{\text{inter}}(v)$ and $d_{\text{intra}}(v)$. The fact that $j_k^\mathsf{T} v_{\text{Lap}}/\sqrt{n_k} = O(1)$ implies that the eigenvalue λ_{Lap} of $h(\tau)L'_s$ remains at a non vanishing distance from other eigenvalues of $h(\tau)L'_s$ (Couillet and Benaych-Georges, 2016, Theorem 4). The same can be said about $\hat{W}$ and its eigenvalue λ_{ctr}.

Let γ be a positively oriented complex closed path circling only around λ_{Lap} and λ_{ctr}. Since there can be only one eigenvector of L'_s ($\hat{W}$, resp.) whose limiting scalar product with j_k for $k \in \{1,2\}$ is bounded away from zero (Couillet and Benaych-Georges, 2016, Theorem 4), which is v_{Lap} (resp., v_{ctr}), we have, by Cauchy's formula (Walter, 1987, Theorem 10.15),

$$\begin{aligned} \frac{1}{n_k}(j_k^\mathsf{T} v_{\text{Lap}})^2 &= -\frac{1}{2\pi i}\oint_\gamma \frac{1}{n_k} j_k^\mathsf{T}(h(\tau)L'_s - zI_n)^{-1} j_k dz + o_P(1) \\ \frac{1}{n_k}(j_k^\mathsf{T} v_{\text{ctr}})^2 &= -\frac{1}{2\pi i}\oint_\gamma \frac{1}{n_k} j_k^\mathsf{T}(\hat{W} - zI_n)^{-1} j_k dz + o_P(1) \end{aligned}$$

for $k \in \{1,2\}$. Since $\hat{W}$ is a low-rank perturbation of $\hat{L}$, invoking Sherman-Morrison's formula (Sherman and Morrison, 1950), we further have

$$j_k^\mathsf{T}(\hat{W} - zI_n)^{-1} j_k = j_k^\mathsf{T}(h(\tau)L'_s - zI_n)^{-1} j_k - \frac{(5h'(\tau)^2/4)\left(j_k^\mathsf{T}(h(\tau)L'_s - zI_n)^{-1}\psi\right)^2}{1 + (5h'(\tau)^2/4)\psi^\mathsf{T}(h(\tau)L'_s - zI_n)^{-1}\psi} + o_P(n_k).$$

As $\frac{1}{\sqrt{n_k}} j_k^\mathsf{T}(h(\tau)L_s' - zI_n)^{-1}\psi = o_P(1)$ (Couillet and Benaych-Georges, 2016, Equation 7.6), we get

$$\frac{1}{n_k} j_k^\mathsf{T}(\hat{W} - zI_n)^{-1} j_k = \frac{1}{n_k} j_k^\mathsf{T}(h(\tau)L_s' - zI_n)^{-1} j_k + o_P(1),$$

and thus

$$\frac{1}{n_k}(j_k^\mathsf{T} v_{\text{Lap}})^2 = \frac{1}{n_k}(j_k^\mathsf{T} v_{\text{ctr}})^2 + o_P(1),$$

which concludes the proof of Proposition 7.

Appendix C. Asymptotic Matrix Equivalent for $\hat{W}$

The objective of this section is to prove the asymptotic matrix equivalent for $\hat{W}$ expressed in (24). Some additional notations that will be useful in the proof:

- for $x_i \in \mathcal{C}_k$, $k \in \{1, 2\}$, $\theta_i \equiv x_i - \mu_k$, and $\theta \equiv [\theta_1, \cdots, \theta_n]^\mathsf{T}$;
- $\mu_k^\circ = \mu_k - \frac{1}{n}\sum_{k'=1}^2 n_{k'}\mu_{k'}$, $t_k = \left(\text{tr}C_k - \frac{1}{n}\sum_{k'=1}^2 n_{k'}\text{tr}C_{k'}\right)/\sqrt{p}$;
- $j_k \in \mathbb{R}^n$ is the canonical vector of $\mathcal{C}_k$, i.e., $[j_k]_i = 1$ if $x_i \in \mathcal{C}_k$ and $[j_k]_i = 0$ otherwise;
- $\psi_i \equiv \left(\|\theta_i\|^2 - \text{E}[\|\theta_i\|^2]\right)/\sqrt{p}$, $\psi \equiv [\psi_1, \cdots, \psi_n]^\mathsf{T}$ and $(\psi)^2 \equiv [(\psi_1)^2, \cdots, (\psi_n)^2]^\mathsf{T}$.

As $w_{ij} = h(\|x_i - x_j\|^2/p = h(\tau) + O(p^{-\frac{1}{2}})$ for all $i \neq j$, we can Taylor-expand $w_{ij} = h(\|x_i - x_j\|^2/p$ around $h(\tau)$ to obtain the following expansion for W, which can be found in the paper of Couillet and Benaych-Georges (2016):

$$\begin{aligned}
W = {} & h(\tau)1_n 1_n^\mathsf{T} + \frac{h'(\tau)}{\sqrt{p}}\left[\psi 1_n^\mathsf{T} + 1_n\psi^\mathsf{T} + \sum_{b=1}^2 t_b j_b 1_n^\mathsf{T} + 1_n\sum_{a=1}^2 t_a j_a^\mathsf{T}\right] \\
& + \frac{h'(\tau)}{p}\left[\sum_{a,b=1}^2 \|\mu_a^\circ - \mu_b^\circ\|^2 j_b j_a^\mathsf{T} - 2\theta\sum_{a=1}^2 \mu_a^\circ j_a^\mathsf{T} + 2\sum_{b=1}^2 \text{diag}(j_b)\theta\mu_b^\circ 1_n^\mathsf{T}\right. \\
& \left. - 2\sum_{b=1}^2 j_b\mu_b^{\circ\mathsf{T}}\theta^\mathsf{T} + 21_n\sum_{a=1}^2 \mu_a^{\circ\mathsf{T}}\theta^\mathsf{T}\text{diag}(j_a) - 2\theta\theta^\mathsf{T}\right] \\
& + \frac{h''(\tau)}{2p}\left[(\psi)^2 1_n^\mathsf{T} + 1_n[(\psi)^2]^\mathsf{T} + \sum_{b=1}^2 t_b^2 j_b 1_n^\mathsf{T} + 1_n\sum_{a=1}^2 t_a^2 j_a^\mathsf{T}\right. \\
& + 2\sum_{a,b=1}^2 t_a t_b j_b j_a^\mathsf{T} + 2\sum_{b=1}^2 \text{diag}(j_b)t_b\psi 1_n^\mathsf{T} + 2\sum_{b=1}^2 t_b j_b\psi^\mathsf{T} + 2\sum_{a=1}^2 1_n\psi^\mathsf{T}\text{diag}(j_a)t_a \\
& \left. + 2\psi\sum_{a=1}^2 t_a j_a^\mathsf{T} + 2\psi\psi^\mathsf{T}\right] + (h(0) - h(\tau) + \tau h'(\tau))I_n + O_{\|\cdot\|}(p^{-\frac{1}{2}}).
\end{aligned}$$

Applying $P_n = \left(I_n - \frac{1}{n}1_n 1_n^\mathsf{T}\right)$ on both sides of the above equation, we get

$$\begin{aligned}
\hat{W} &= P_n W P_n \\
&= \frac{-2h'(\tau)}{p}\left[\sum_{a,b=1}^{2}(\mu_a^\circ{}^\mathsf{T}\mu_b^\circ) j_b j_a^\mathsf{T} + P_n\theta\sum_{a=1}^{2}\mu_a^\circ j_a^\mathsf{T} + \sum_{b=1}^{2} j_b\mu_b^\circ{}^\mathsf{T}\theta^\mathsf{T}P_n + P_n\theta\theta^\mathsf{T}P_n\right] \\
&+ \frac{h''(\tau)}{p}\left[\sum_{a,b=1}^{2} t_a t_b j_b j_a^\mathsf{T} + \sum_{b=1}^{2} t_b j_b\psi^\mathsf{T}P_n + P_n\psi\sum_{a=1}^{2} t_a j_a^\mathsf{T} + P_n\psi\psi^\mathsf{T}P_n\right] \\
&+ (h(0) - h'(\tau) + \tau h''(\tau))P_n + O(p^{-\frac{1}{2}}) \\
&= \frac{1}{p}\hat{\Phi}^\mathsf{T}\hat{\Phi} + (h(0) - h(\tau) + \tau h'(\tau))P_n + O_{\|\cdot\|}(p^{-\frac{1}{2}})
\end{aligned}$$

where the last equality is justified by

$$\begin{aligned}
\frac{1}{p}\hat{\Phi}^\mathsf{T}\hat{\Phi} = &\frac{-2h'(\tau)}{p}\left[\sum_{a,b=1}^{2}(\mu_a^\circ{}^\mathsf{T}\mu_b^\circ) j_b j_a^\mathsf{T} + P_n\theta\sum_{a=1}^{2}\mu_a^\circ j_a^\mathsf{T} + \sum_{b=1}^{2} j_b\mu_b^\circ{}^\mathsf{T}\theta^\mathsf{T}P_n + P_n\theta\theta^\mathsf{T}P_n\right] \\
&+ \frac{h''(\tau)}{p}\left[\sum_{a,b=1}^{2} t_a t_b j_b j_a^\mathsf{T} + \sum_{b=1}^{2} t_b j_b\psi^\mathsf{T}P_n + P_n\psi\sum_{a=1}^{2} t_a j_a^\mathsf{T} + P_n\psi\psi^\mathsf{T}P_n\right].
\end{aligned}$$

Equation (24) is thus proved.

Appendix D. Guarantee for approaching the optimal performance on isotropic Gaussian data

The purpose of this section is to provide some general guarantee for the proposed centered regularization method to approach the best achievable performance on isotropic high dimensional Gaussian data, which was characterized in the recent work of Lelarge and Miolane (2019). In this work, the considered isotropic data model is a special case of our analytical framework, in which $-\mu_1 = \mu_2 = \mu$, $C_1 = C_2 = I_p$ and $\rho_1 = \rho_2$. Reorganizing the results of Lelarge and Miolane (2019), the optimally achievable classification accuracy in the limit of large p is equal to

$$1 - Q(\sqrt{q_*})$$

with $q_* > 0$ satisfying the fixed point equation

$$q_* = \|\mu\|^2 - \frac{p\|\mu\|^2}{p + \|\mu\|^2\left(n_{[l]} + \mathbb{E}_{z\sim\mathcal{N}(q_*,q_*)}\{\tanh(z)\}n_{[u]}\right)}. \tag{57}$$

It is easy to see that the optimal accuracy is higher with greater q_*. In parallel, reformulating the results of Corollary 4 for some value of the hyperparameter $e > 0$ such that

$$m(\xi_e) = \frac{m(\xi_e)^2}{m(\xi_e)^2 + \sigma^2(\xi_e)}, \tag{58}$$

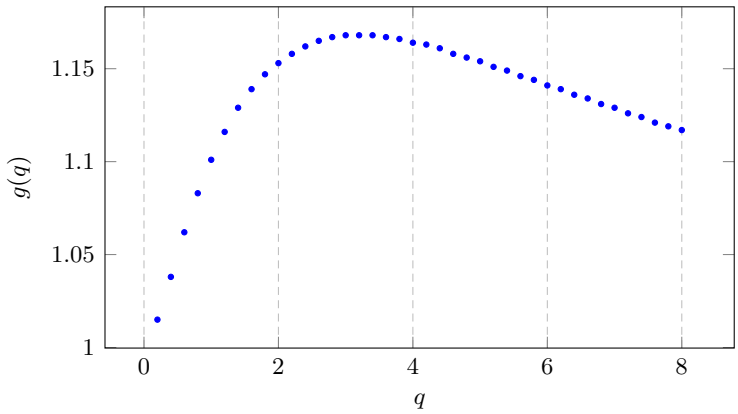

Figure 7: Values of $g(q)$ at various q.

the high dimensional classification accuracy achieved by the centered regularization method is asymptotically equal to

$$1 - Q(\sqrt{q_c})$$

with $q_c > 0$ satisfying the fixed point equation

$$q_c = \|\mu\|^2 - \frac{p\|\mu\|^2}{p + \|\mu\|^2 \left(n_{[l]} + \frac{q_c}{q_c+1} n_{[u]}\right)} \tag{59}$$

Obviously, the fixed-point equations (57)–(59) are identical at $n_{[u]} = 0$, meaning that the centered regularization method achieves the optimal performance on fully labelled sets. For partially labelled sets, the difference between (57) and (59) resides in the multiplying factors before $n_{[u]}$. This means that, for a best achievable accuracy of $1 - Q(\sqrt{q_*})$ at some $n_{[l]}$ and $n_{[u]}$, the centered regularization method achieves, with the hyperparameter e set to satisfy (58), the same level of accuracy with the same amount of labelled samples and $g(q_*)n_{[u]}$ unlabelled ones where $g(q_*) = \mathbb{E}_{z\sim\mathcal{N}(q_*,q_*)}\{\tanh(z)\}(q_*+1)/q_*$. The ratio function

$$g(q) = \frac{\mathbb{E}_{z\sim\mathcal{N}(q,q)}\{\tanh(z)\}(q+1)}{q}$$

is plotted in Figure 7. We remark also that $\lim_{q\to 0^+} g(q) = 1$ and $\lim_{q\to+\infty} g(q) = 0$. Although the value of $g(q)$ can get up to 1.168, the number of unlabelled samples required to reach the optimal performance can be reduced with an optimally chosen e (which generally does not satisfy (58)). In fact, even with the same numbers of labelled and unlabelled data, the performance of centered regularization method at an optimally set e is often very close to the best achievable one, as shown in Figures 5–6.

References

Charu C Aggarwal, Alexander Hinneburg, and Daniel A Keim. On the surprising behavior of distance metrics in high dimensional space. In *International Conference on Database Theory*, pages 420–434. Springer, 2001.

Fabrizio Angiulli. On the behavior of intrinsically high-dimensional spaces: distances, direct and reverse nearest neighbors, and hubness. *Journal of Machine Learning Research*, 18 (170):1–60, 2018.

Konstantin Avrachenkov, Alexey Mishenin, Paulo Gonçalves, and Marina Sokol. Generalized optimization framework for graph-based semi-supervised learning. In *International Conference on Data Mining*, pages 966–974. SIAM, 2012.

Mikhail Belkin and Partha Niyogi. Using manifold stucture for partially labeled classification. In *Advances in Neural Information Processing Systems*, pages 953–960, 2003.

Mikhail Belkin, Partha Niyogi, and Vikas Sindhwani. Manifold regularization: a geometric framework for learning from labeled and unlabeled examples. *Journal of Machine Learning Research*, 7(Nov):2399–2434, 2006.

Shai Ben-David, Tyler Lu, and Dávid Pál. Does unlabeled data provably help? worst-case analysis of the sample complexity of semi-supervised learning. In *Conference on Learning Theory*, pages 33–44, 2008.

Florent Benaych-Georges and Romain Couillet. Spectral analysis of the gram matrix of mixture models. *ESAIM: Probability and Statistics*, 20:217–237, 2016.

Kevin Beyer, Jonathan Goldstein, Raghu Ramakrishnan, and Uri Shaft. When is "nearest neighbor" meaningful? In *International Conference on Database Theory*, pages 217–235. Springer, 1999.

Nick Bridle and Xiaojin Zhu. p-voltages: Laplacian regularization for semi-supervised learning on high-dimensional data. In *Workshop on Mining and Learning with Graphs*, 2013.

Deng Cai and Xiaofei He. Manifold adaptive experimental design for text categorization. *IEEE Transactions on Knowledge and Data Engineering*, 24(4):707–719, 2011.

Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien. *Semi-supervised Learning*. MIT Press, 2010.

Romain Couillet and Florent Benaych-Georges. Kernel spectral clustering of large dimensional data. *Electronic Journal of Statistics*, 10(1):1393–1454, 2016.

Romain Couillet and Merouane Debbah. *Random Matrix Methods for Wireless Communications*. Cambridge University Press, 2011.

Fabio Gagliardi Cozman, Ira Cohen, and M Cirelo. Unlabeled data can degrade classification performance of generative classifiers. In *Flairs Conference*, pages 327–331, 2002.

Ahmed El Alaoui, Xiang Cheng, Aaditya Ramdas, Martin J Wainwright, and Michael I Jordan. Asymptotic behavior of ℓ_p-based laplacian regularization in semi-supervised learning. In *Conference on Learning Theory*, pages 879–906, 2016.

Noureddine El Karoui, Derek Bean, Peter J Bickel, Chinghway Lim, and Bin Yu. On robust regression with high-dimensional predictors. *Proceedings of the National Academy of Sciences*, page 201307842, 2013.

Damien Francois, Vincent Wertz, and Michel Verleysen. The concentration of fractional distances. *IEEE Transactions on Knowledge and Data Engineering*, 19(7):873–886, 2007.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680, 2014.

Alexander Hinneburg, Charu C Aggarwal, and Daniel A Keim. What is the nearest neighbor in high dimensional spaces? In *International Conference on Very Large Databases*, pages 506–515, 2000.

Rie Johnson and Tong Zhang. On the effectiveness of laplacian normalization for graph semi-supervised learning. *Journal of Machine Learning Research*, 8(7), 2007.

Yann LeCun. The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*, 1998.

Michel Ledoux. *The Concentration of Measure Phenomenon*. Number 89. American Mathematical Soc., 2005.

Marc Lelarge and Leo Miolane. Asymptotic Bayes risk for gaussian mixture in a semi-supervised setting. *arXiv preprint arXiv:1907.03792*, 2019.

Cosme Louart and Romain Couillet. Concentration of measure and large random matrices with an application to sample covariance matrices. *arXiv preprint arXiv:1805.08295*, 2018.

Xiaoyi Mai and Romain Couillet. A random matrix analysis and improvement of semi-supervised learning for large dimensional data. *Journal of Machine Learning Research*, 19(1):3074–3100, 2018.

Boaz Nadler, Nathan Srebro, and Xueyuan Zhou. Semi-supervised learning with the graph laplacian: the limit of infinite unlabelled data. In *Advances in Neural Information Processing Systems*, pages 1330–1338, 2009.

Mauricio Flores Rios, Jeff Calder, and Gilad Lerman. Algorithms for ℓ_p-based semi-supervised learning on graphs. *arXiv preprint arXiv:1901.05031*, 2019.

Behzad M Shahshahani and David A Landgrebe. The effect of unlabeled samples in reducing the small sample size problem and mitigating the hughes phenomenon. *IEEE Transactions on Geoscience and Remote Sensing*, 32(5):1087–1095, 1994.

Jack Sherman and Winifred J Morrison. Adjustment of an inverse matrix corresponding to a change in one element of a given matrix. *The Annals of Mathematical Statistics*, 21(1): 124–127, 1950.

Zuoqiang Shi, Stanley Osher, and Wei Zhu. Weighted nonlocal laplacian on interpolation from sparse data. *Journal of Scientific Computing*, 73(2):1164–1177, 2017.

David I Shuman, Sunil K Narang, Pascal Frossard, Antonio Ortega, and Pierre Vandergheynst. The emerging field of signal processing on graphs: extending high-dimensional data analysis to networks and other irregular domains. *IEEE Signal Processing Magazine*, 30(3):83–98, 2013.

Alexander J Smola and Risi Kondor. Kernels and regularization on graphs. In *Learning Theory and Kernel Machines*, pages 144–158. Springer, 2003.

Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4): 395–416, 2007.

Ulrike Von Luxburg, Mikhail Belkin, and Olivier Bousquet. Consistency of spectral clustering. *The Annals of Statistics*, pages 555–586, 2008.

Rudin Walter. *Real and Complex Analysis*. McGraw Hill Book Company, 1987.

Max A Woodbury. Inverting modified matrices. *Memorandum Report*, 42(106):336, 1950.

Dengyong Zhou and Bernhard Schölkopf. Regularization on discrete spaces. In *Joint Pattern Recognition Symposium*, pages 361–368. Springer, 2005.

Denny Zhou, Olivier Bousquet, Thomas N Lal, Jason Weston, and Bernhard Schölkopf. Learning with local and global consistency. In *Advances in Neural Information Processing Systems*, pages 321–328, 2004.

Xueyuan Zhou and Mikhail Belkin. Semi-supervised learning by higher order regularization. In *International Conference on Artificial Intelligence and Statistics*, pages 892–900, 2011.

Xiaojin Zhu and Zoubin Ghahramani. Learning from labeled and unlabeled data with label propagation. 2002.

Xiaojin Zhu, Zoubin Ghahramani, and John D Lafferty. Semi-supervised learning using gaussian fields and harmonic functions. In *International Conference on Machine Learning*, pages 912–919, 2003.