Journal of Machine Learning Research XX (2017) X-X Submitted XX/XX; Revised XX/XX; Published XX/XX

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: XX XX

Abstract

Semi-supervised Laplacian regularization, a standard graph-based approach for learning from both labelled and unlabelled data, is demonstrated by the recent work of (Mai and Couillet, 2017) to have an insignificant high dimensional learning efficiency with respect to unlabelled data, 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 is proposed as solution, which is shown both theoretically and empirically to have a superior performance over Laplacian regularization.

Keywords: semi-supervised learning, graph-based methods, high dimensional statistics, 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 datasets to achieve satisfying results.

The idea of semi-supervised learning (Chapelle et al., 2009) comes from the expectation of maximizing the learning performance by combining labelled and unlabelled data, which is of significant practical value when the cost of supervised learning is too high and the performances of unsupervised approaches is too weak. Somewhat surprisingly though, although quite natural, semi-supervised learning has not reached broad recognition. Due to the difficulty of properly uniting both labelled and unlabelled information, many standard semi-supervised learning techniques were found to exhibit worse performances than their one-sided counterparts (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 datasets. 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 the 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 the 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 present work tackles both aforementioned tractability and dimensionality difficulties at once, by precisely leveraging the large dimension of datasets to exploit recent advances in random matrix theory. In doing so, the asymptotic performance of a class of well-known semi-supervised approaches is precisely analyzed, the aforementioned weak performances understood, and a simple yet powerful correction for those algorithms proposed and corroborated by compelling simulation results.

The article concerns specifically *semi-supervised graph regularization* approaches (Zhu et al., 2003; Zhou et al., 2004), a major subset of semi-supervised learning methods (Chapelle et al., 2009), often referred to as Laplacian regularizations with their loss functions involving differently normalized Laplacian matrices (Avrachenkov et al., 2012). We refer the readers to Section 2.1 for an introduction of these techniques. Characterizing the distance concentration phenomenon under the conditions of a Gaussian mixture model, in a previous work by the authors (Mai and Couillet, 2017), it was made clear that among existing Laplacian regularization algorithms, only the one with the random walk normalized Laplacian matrix yields reasonable results, yet with asymptotically negligible contribution from unlabelled dataset (see Subsection 2.2 for more details). This last observation of the inefficiency of Laplacian regularization methods to learn from unlabelled data may cause it to be outperformed by a mere (unsupervised) spectral clustering approach (Von Luxburg, 2007) in the same high dimensional settings (Couillet and Benaych-Georges, 2016).

The main contribution of this paper is the proposition of a novel semi-supervised graph regularization algorithm to address the aforementioned inconsistency problem of the traditional Laplacian approach with respect to unlabelled data. The proposed improvement is simple to implement and very powerful. It is shown, through a rigorous theoretical analysis placed under the large dimensional random matrix setting of large and numerous data (similar to the previous work (Mai and Couillet, 2017) or to (Couillet and Benaych-Georges, 2016) in the context of spectral clustering), to induce a consistent learning from high dimensional data with labelled and unlabelled data learning efficiency lowered bounded respectively by Laplacian regularization and spectral clustering. As a matter of fact, the proposed method, featuring a tuning hyperparameter, consistently relates semi-supervised learning to both unsupervised and supervised learning in showing that, at the two extremes in the selection of the hyperparameter, the performance of unsupervised spectral clustering and that of Laplacian regularization, which is essentially a supervised learning method in high dimensions, are exactly recovered. With the hyperparameter optimally set somewhere between these two extremes, the algorithm fulfills precisely the semi-supervised learning goal of surpassing one-sided learning schemes by properly combining them, resulting in a significant advantage over the traditional Laplacian regularization. Apart from theoretical conclusions, the superiority of the new regularization method is also illustrated by simulations on various data sets.

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. High Dimensional Semi-Supervised Graph Regularization

2.1 Preliminaries

We begin this section by recalling the basics of graph learning methods, before delving into a high dimensional discussion of semi-supervised graph regularization. Consider a set $\{x_1, \ldots, x_n\} \in \mathbb{R}^p$ of p-dimensional input vectors belonging to either one of two affinity classes $\mathcal{C}_1$ or $\mathcal{C}_2$. In graph-based methods, data samples $x_1, \ldots, x_n$ are represented by vertices in a graph, upon which a weight matrix W is computed by

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

for some decreasing non-negative function h, so that nearby data vectors x_i, x_j are connected with a large weight w_{ij}, which can also be seen as a similarity measure between data samples. 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 samples $x_1, \ldots, x_n$, it varies little from x_i to x_j when w_{ij} has a large value. The graph smoothness assumption is usually characterized as minimizing a smoothness penalty term of the form

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

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

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

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 variations of the smoothness penalty term 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 $L_r = I_n - WD^{-1}$.

In the semi-supervised setting, we dispose of $n_{[l]}$ pairs of labelled points and 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

$$\begin{aligned} &\min_{f\in\mathbb{R}^n} f^{\mathsf{T}} L f \\ &s.t. \quad f_i = y_i, \quad 1 \le i \le n_{[l]}. \end{aligned} \tag{2}$$

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, which gives the following explicit solution

$$f_{[u]} = -L_{[uu]}^{-1} L_{[ul]} f_{[l]}. \tag{3}$$

Finally, the decision step consists in assigning unlabelled sample x_i to $\mathcal{C}_1$ (resp., $\mathcal{C}_2$) if $f_i < 0$ (resp., $f_i > 0$).

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 signals 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 $f^\mathsf{T} L_r f$ can lead to better classification results. To integrate all these different Laplacian regularization algorithms into a common framework, we define $L^{(a)} = I - D^{-1-a} W D^a$ as the a-normalized Laplacian matrix. Replacing L with $L^{(a)}$ in (3) to get

$$f_{[u]} = -\left(L^{(a)}_{[uu]}\right)^{-1} L^{(a)}_{[ul]} f_{[l]}, \tag{4}$$

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 $a = -1$.

Despite being a popular semi-supervised learning approach, Laplacian regularization algorithms are shown by (Mai and Couillet, 2017) to have a non-efficient learning capacity for high dimensional unlabelled data, as a direct consequence of the distance concentration phenomenon, hinted at in the introduction. A deeper examination of the results in (Mai and Couillet, 2017) allows us to discover that this problem of unlabelled data learning efficiency may in fact be settled through the usage of a centered similarity measure, as opposed to the current convention of non-negative similarities w_{ij}. In the following subsections, we will recall the findings of (Mai and Couillet, 2017), then move on to the proposition of the novel corrective algorithm, along with some general remarks explaining the effectiveness of the proposed algorithm, leaving the thorough performance analysis to the next section.

2.2 Behaviour of Laplacian Regularization

Conforming to the settings of (Mai and Couillet, 2017), 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 $\|C_k\| = O(1)$, $\|C_k^{-1}\| = O(1)$, $\|\mu_2 - \mu_1\| = O(1)$, $\operatorname{tr}C_1 - \operatorname{tr}C_2 = O(\sqrt{p})$ and $\operatorname{tr}(C_1 - C_2)^2 = O(\sqrt{p})$.

The ratios $c_0 = \frac{n}{p}$, $c_{[l]} = \frac{n_{[l]}}{p}$ and $c_{[u]} = \frac{n_{[u]}}{p}$ are uniformly bounded in $(0, +\infty)$ 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 $\|\mu_2 - \mu_1\| = O(1)$, $\operatorname{tr}C_1 - \operatorname{tr}C_2 = O(\sqrt{p})$, and $\operatorname{tr}(C_1 - C_2)^2 = O(\sqrt{p})$ 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 at the core of the reasons why Laplacian-based semi-supervised learning is bound to fail with large dimensional data.

Proposition 1 *Define $\tau = \operatorname{tr}(C_1 + C_2)/p$. 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 samples converge to the same value, thereby indicating that the presumed connection between *proximity and data affinity* is completely disrupted. In such situations, the performance of the Laplacian regularization approach (along with most distance-based classification methods), which normally works well in small dimensions, may be severely affected. Indeed, under some mild smooth conditions on the weight function h, the analysis of (Mai and Couillet, 2017) reveals several surprising and critical aspects of the high dimensional behavior of this approach. 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 all unlabelled data to be classified in the same class (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

$$f_{[l]} = \left(I_{n_{[l]}} - \frac{1}{n_{[l]}} 1_{n_{[l]}} 1_{n_{[l]}}^{\mathsf{T}} \right) y_{[l]} \tag{5}$$

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 (5), (Mai and Couillet, 2017) 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 $L^{(a)} = I - D^{-1-a} W D^{a}$) except for $a \simeq -1$. This indicates that among all existing Laplacian regularization algorithms proposed in the literature, only the random walk normalized Laplacian regularization yields non-trivial classification results for large dimensional data. We recall in the following theorem the exact statistical characterization of $f_{[u]}$ produced by the random walk normalized Laplacian regularization, which was firstly presented in (Mai and Couillet, 2017).

Theorem 2 *Let Assumption 1 hold, the function h be three-times continuously differentiable in a neighborhood of τ, and the solution $f_{[u]}$ be given by (4) for $a = -1$. Then, for $n_{[l]} + 1 \leq i \leq n$ (i.e., x_i unlabelled) and $x_i \in \mathcal{C}_k$,*

$$p f_i = g_i + o_P(1)$$

where $g_i \sim \mathcal{N}(m_k, \sigma_k^2)$ with

$$m_k = (-1)^k (1 - \rho_k) \left[-\frac{2h'(\tau)}{h(\tau)} \|\mu_1 - \mu_2\|^2 + \left(\frac{h''(\tau)}{h(\tau)} - \frac{h'(\tau)^2}{h(\tau^2)} \right) (\operatorname{tr} C_1 - \operatorname{tr} C_2)^2 \right] \tag{6}$$

$$\begin{aligned} \sigma_k^2 = {} & \frac{4h'(\tau)^2}{h(\tau)^2} \left[(\mu_1 - \mu_2)^{\mathsf{T}} C_k (\mu_1 - \mu_2) + \left(\sum_{a=1}^{2} \frac{\operatorname{tr} C_a C_k}{\rho_a} \right) \frac{1}{c_{[l]}} \right] \\ & + \frac{2 \operatorname{tr} C_k^2}{p} \left(\frac{h''(\tau)}{h(\tau)} - \frac{h'(\tau)^2}{h(\tau^2)} \right)^2 (\operatorname{tr} C_1 - \operatorname{tr} C_2)^2 . \end{aligned} \tag{7}$$

Theorem 2 states that the classification scores 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

$$\mathcal{P}(x_i \to C_k | x_i \in C_k, i > n_{[l]}) = \Phi\left(\sqrt{m_k^2/\sigma_k^2}\right) + o_p(1) \tag{8}$$

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* $c_{[u]}$, which tells us 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 gain. 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 subsection.

2.3 Regularization with Centered Similarities

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 (4), 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 allows to capitalize on the global information contained in unlabelled data. However, as a consequence of the distance concentration "curse" stated in Proposition 1, the similarities (weights) w_{ij} between high dimensional data vectors are dominated by the constant value $h(\tau)$ plus some small fluctuations, which results in the collapse of the smoothness filter:

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

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 Subsection 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 if all elements of the signal s have the same value. Notice also that in perfect situations where the data points in different class subgraphs are connected with zero weights w_{ij}, class indicators (i.e., signals with constant values within class subgraphs which are different for each 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 for high dimensional learning is that when 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 global information in high dimensional unlabelled data through Laplacian regularizations.

In view of the above discussion, we shall try to eliminate the dominant advantages of constant signals, in an attempt to render detectable the discrimination between class-structured signals and other noised directions. As constant signals always have a smoothness penalty of zero, a very easy way to break their optimal smoothness is to introduce negative weights in the graph so that the values of the smoothness regularizer can go below zero. More specifically, in the cases where the intra-class similarities are averagely positive and the inter-class similarities are averagely negative, class-structured signals are bound to have a lower smoothness penalty than constant signals. However, the implementation of such idea using both positive and negative similarities 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 $\mathcal{P}_l(f_{[l]}) = D^{-1-a}_{[u]} W_{[ul]} D^{a}_{[l]} f_{[l]}$, 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 if the signs of the degree of that point and those of labelled data change, leading thus to extremely unstable classification results.

To cope with this problem, we propose here the usage of centered similarities $\hat{w}_{ij}$, for which the positive and negative weights are balanced out at any data point, i.e., for all

$i \in \{1, \ldots, n\}$, $d_i = \sum_{j=1}^{n} w_{ij} = 0$. 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:

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

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

$$\hat{Q}(s) = \sum_{i,j=1}^{n} \hat{w}_{ij}(s_i - s_j)^2 = -s^{\mathsf{T}} \hat{W} s, \tag{9}$$

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

Another merit of using centered similarities is that the distance between the intra-class similarities and inter-class similarities in the previous graph is preserved, in the sense that the average of inter-class similarities minus the average of inter-class similarities stays unchanged after centering. Since the total sum of centered similarities $\hat{w}_{ij}$ amounts to zero, the average of intra-class similarities is always positive while that of inter-class similarities negative as long as the former are greater on average than the latter, which remains a necessary condition for a functional semi-supervised graph regularization. Furthermore, in the common situations where the similarity matrices W are constructed through a kernel function, e.g., through the popular radial basis function (RBF) kernel $w_{ij} = e^{-\|x_i - x_j\|^2/t}$, there exists (by definition of kernel functions) a mapping $x \mapsto \phi(x)$ such that

$$w_{ij} = \phi(x_i)^{\mathsf{T}} \phi(x_j).$$

Since

$$\hat{w}_{ij} = \left(\phi(x_i) - \frac{1}{n}\sum_{t=1}^{n} \phi(x_t)\right)^{\mathsf{T}} \left(\phi(x_j) - \frac{1}{n}\sum_{t=1}^{n} \phi(x_t)\right),$$

the centering operation is equivalent to translating the feature vectors $\phi(x_i)$ by moving their center to the origin, meaning that the relative positions between feature vectors remain intact after the centering step.

This being said, a 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 (5), the new optimization problem may now be posed as follows:

$$\begin{aligned} &\min_{f_{[u]} \in \mathbb{R}^{n_{[u]}}} -f^{\mathsf{T}} \hat{W} f \\ &s.t. \|f_{[u]}\|^2 = n_{[u]} e^2. \end{aligned} \tag{10}$$

Naturally, the optimization can be solved by introducing a Laplacian multiplier α to the norm constraint $\|f_{[u]}\|^2 = n_{[u]} e^2$ and the solution is given by

$$f_{[u]} = \left(\alpha I_{n_{[u]}} - \hat{W}_{[uu]}\right)^{-1} \hat{W}_{[ul]} f_{[l]} \tag{11}$$

where α is determined by $\alpha > \|\hat{W}_{[uu]}\|$ and $\|f_{[u]}\|^2 = n_{[u]}e^2$. In practice, α can be used directly as a parameter for a more convenient implementation. We summarize the method in Algorithm 1.

Algorithm 1 Semi-Supervised Graph Regularization with Centered Similarities

```
Input: $n_{[l]}$ pairs of labelled points and 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 $\{x_{n_{[l]}+1}, \ldots, x_n\}$.
Output: Classification of unlabelled data $\{x_{n_{[l]}+1}, \ldots, x_n\}$.
Compute the similarity matrix $W$.
Compute the centered similarity matrix $\hat{W} = PWP$ with $P = I_n - \frac{1}{n}1_n 1_n^\mathsf{T}$, and define $\hat{W} = \begin{bmatrix} \hat{W}_{[ll]} & \hat{W}_{[lu]} \\ \hat{W}_{[ul]} & \hat{W}_{[uu]} \end{bmatrix}$.
Set $f_{[l]} = \left( I_{n_{[l]}} - \frac{1}{n_{[l]}} 1_{n_{[l]}} 1_{n_{[l]}}^\mathsf{T} \right) y_{[l]}$ with $y_{[l]}$ the vector containing labelled $y_i$.
Compute the class scores of unlabelled data $f_{[u]} = \left( \alpha I_{n_{[u]}} - \hat{W}_{[uu]} \right)^{-1} \hat{W}_{[ul]} f_{[l]}$ for some $\alpha > \|\hat{W}_{[uu]}\|$.
Classify unlabelled data $\{x_{n_{[l]}+1}, \ldots, x_n\}$ by the signs of $f_{[u]}$.
```

The proposed algorithm induces almost no extra cost to the classical Laplacian approach, except the addition of the parameter α controlling the norm of $f_{[u]}$. However, as will be demonstrated in the next section on performance analysis, the existence of this parameter, aside from making the regularization with centered similarities a well-posed problem, actually allows one to adjust the combination of labelled and unlabelled information in search for an optimal semi-supervised learning performance.

3. Performance Analysis

With the proposition of the centered similarities regularization intuitively justified in Subsection 2.3, the main purpose of this section is to provide mathematical support for its effective high dimensional learning capabilities 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 learning performance of the proposed method has an unlabelled data learning efficiency that is at least as good as spectral clustering, as opposed to Laplacian regularization.

We first enunciate the central theorem providing the statistical characterization of unlabelled data scores $f_{[u]}$ obtained by the proposed updated algorithm. As the new algorithm will be shown to draw both on labelled and unlabelled data information, the complex interactions between these two types of data generate more intricate outcomes than in (Mai and Couillet, 2017). To facilitate the interpretation of the theoretical results without cumbersome notations, we restrict the theorem to the homoscedastic case as considered in linear discrimination analysis (i.e., the class covariances are taken equal, $C_1 = C_2 = C$), without affecting the generality of the conclusions given subsequently. We refer the interested reader to the appendix for an extended version of the theorem along with its proof.

Theorem 3 *Let Assumption 1 hold with $C_1 = C_2 = C$, h be three-times continuously differentiable in a neighborhood of τ, and $f_{[u]}$ be the solution of (10) with fixed norm $n_{[u]}e^2$. Then, for $n_{[l]} + 1 \leq i \leq n$ (i.e., x_i unlabelled) and $x_i \in \mathcal{C}_k$,*

$$f_i = g_i + o_P(1)$$

where

$$g_i \sim \mathcal{N}\left((-1)^k(1-\rho_k)m, \sigma^2\right)$$

for some $m, \sigma^2 > 0$. More precisely, defining

$$\theta = \frac{c_{[u]}m}{2c_{[l]}} \tag{12}$$

and letting $s : (0, \|C + \rho_1\rho_2(\mu_1 - \mu_2)(\mu_1 - \mu_2)^\mathsf{T}\|) \to (0, +\infty)$ be the injective function given by

$$s(\xi) = \xi(\mu_1 - \mu_2)^\mathsf{T}\left\{I_p - \xi\left[C + \rho_1\rho_2(\mu_1 - \mu_2)(\mu_1 - \mu_2)^\mathsf{T}\right]\right\}^{-1}(\mu_1 - \mu_2). \tag{13}$$

the values of m and σ^2 are determined by $\rho_1\rho_2 m^2 + \sigma^2 = e^2$ and

$$\frac{\sigma^2}{m^2} = \left[1 - \left(\frac{\theta}{1+\theta}\right)^2 \frac{q(\theta)}{(\rho_1\rho_2)^2 c_{[u]}}\right]^{-1}\left[\omega(\theta) + \left(\frac{\theta}{1+\theta}\right)^2 \frac{q(\theta)}{\rho_1\rho_2 c_{[u]}} + \left(\frac{1}{1+\theta}\right)^2 \frac{q(\theta)}{\rho_1\rho_2 c_{[l]}}\right] \tag{14}$$

where

$$q(\theta) = \frac{\operatorname{tr}\left[(I_p - s^{-1}(\theta)C)^{-1}C\right]^2}{p\left[(\mu_1 - \mu_2)^\mathsf{T}(I_p - s^{-1}(\theta)C)^{-1}(\mu_1 - \mu_2)\right]^2}$$
$$\omega(\theta) = \frac{(\mu_1 - \mu_2)^\mathsf{T}(I_p - s^{-1}(\theta)C)^{-1}C(I_p - s^{-1}(\theta)C)^{-1}(\mu_1 - \mu_2)}{\left[(\mu_1 - \mu_2)^\mathsf{T}(I_p - s^{-1}(\theta)C)^{-1}(\mu_1 - \mu_2)\right]^2}.$$

In the special cases where $C_1 = C_2 = \lambda^2 I_p$, the above theorem has much simpler form.

Corollary 4 *Under the conditions and notations of Theorem 3, let $C_1 = C_2 = \lambda^2 I_p$, then the values of m, σ^2 are given by $\rho_1\rho_2 m^2 + \sigma^2 = e^2$ and*

$$\frac{\sigma^2}{m^2} = \left[1 - \left(\frac{\theta}{1+\theta}\right)^2 \frac{\lambda^4}{\|\mu_1 - \mu_2\|^4(\rho_1\rho_2)^2 c_{[u]}}\right]^{-1}\left[\frac{\lambda^2}{\|\mu_1 - \mu_2\|^2} + \left(\frac{\theta}{1+\theta}\right)^2 \frac{\lambda^4}{\|\mu_1 - \mu_2\|^4\rho_1\rho_2 c_{[u]}} + \left(\frac{1}{1+\theta}\right)^2 \frac{\lambda^4}{\|\mu_1 - \mu_2\|^4\rho_1\rho_2 c_{[l]}}\right]. \tag{15}$$

Like the centered similarities regularization, the random walk normalized Laplacian algorithm, which is the only one ensuring non-trivial classification results among existing Laplacian algorithms for high dimensional data, gives also $g_i \sim \mathcal{N}\left((-1)^k(1-\rho_k)m', \sigma'^2\right)$ for some other $m', \sigma'^2 > 0$ under the homoscedasticity assumption. We shall use the variance over square mean ratio $r = \sigma^2/m^2$ as the inverse performance measure (i.e., lower r indicates better classification results for high dimensional data) in the following discussion. Denote by r_{lap} the ratio of the random walk normalized Laplacian algorithm, which is obtained from Theorem 2 as

$$r_{\text{lap}} = \frac{(\mu_1 - \mu_2)^{\mathsf{T}} C(\mu_1 - \mu_2)}{\|\mu_1 - \mu_2\|^4} + \frac{\operatorname{tr} C^2}{p\|\mu_1 - \mu_2\|^4 \rho_1 \rho_2 c_{[l]}} \tag{16}$$

and by r_{ctr} the ratio for the centered similarities method, the expression of which has a rather complicated form given by (14).

Note importantly that the quantity $\theta(e)$ in fact reflects *the ratio between the labelled data scores* $f_{[l]}$ *and the unlabelled data scores* $f_{[u]}$ as

$$\theta = \frac{c_{[u]} m}{2c_{[l]}} \simeq \sqrt{\frac{\|\mathbb{E}\{f_{[u]}\}\|^2}{\|f_{[l]}\|^2}}.$$

We observe notably that, when $\|f_{[l]}\|^2 \gg \|\mathbb{E}\{f_{[u]}\}\|^2$, θ goes to zero, at which value the unlabelled data over dimension ratio $c_{[u]} = n_{[u]}/p$ disappears from the expression of r_{ctr}, suggesting the performance relies solely on the labelled data. Inversely, if θ goes to infinity, then it is the labelled data ratio $c_{[l]} = n_{[l]}/p$ that will be left out from (14), and the learning is only guided by the unlabelled data. In other words, the quantity θ can be seen as *a variable tuning the impacts of the two types of data* on the learning process, which is modified by changing the parameter e in the equality constraint $\|f_{[u]}\| = n_{[u]}e^2$ of the optimization problem (10).

As stated in Subsection 2.3, the proposed method can be more conveniently implemented by Algorithm 1, with $f_{[u]}$ computed by $f_{[u]} = \left(\alpha I_{n_{[u]}} - \hat{W}_{[uu]}\right)^{-1} \hat{W}_{[ul]} f_{[l]}$ for some $\alpha > \|\hat{W}_{[uu]}\|$. Obviously, the norm of $f_{[u]}$ is controlled by the hyperparameter α with large α implying small $\|f_{[u]}\|^2$, and consequently small θ, indicating that the labelled data information is emphasized at great values of α. By the same reasoning, the unlabelled data information becomes more influential as α gets close to its minimal limit $\alpha_{\inf} = \|\hat{W}_{[uu]}\|$. Actually, taking $\alpha \in \left(\|\hat{W}_{[uu]}\|, +\infty\right)$ infinitely near the two extremes of its admissible range allows to retrieve respectively the performances of Laplacian regularization and spectral clustering, as will be demonstrated in the following.

Firstly, following the argument in Subsection 2.3 that using centered similarities should cause no loss of information as the difference between the intra-class and inter-class similarities is preserved, we find indeed, by comparing (14) and (16), that

$$\lim_{\theta \to 0} r_{\text{ctr}} = r_{\text{lap}},$$

meaning that the performance of the classical Laplacian regularization can be perfectly retrieved with the centered similarities approach by letting its learning process be completely guided with labelled data. In practice, this is achieved by letting $\alpha \to +\infty$, at which $\|\mathbb{E}\{f_{[u]}\}\|^2 < \|f_{[u]}\|^2 \to 0$, leading to $\theta \to 0$. We remark thus that, with an appropriately set α, the performance of the proposed method is *lowered bounded by that of Laplacian regularization.*

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

Recall that the regular graph smoothness penalty term $Q(s)$ of a signal s can be written as $Q(s) = s^{\mathsf{T}}Ls$. In an unsupervised spectral learning manner, we therefore 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 by the eigenvector associated with the second smallest eigenvalue. Instead, by (9), the updated smoothness penalty term with centered similarities, that is $\hat{Q}(s) = s^{\mathsf{T}}\hat{W}s$, 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 $L_s = I_n - D^{-\frac{1}{2}}WD^{-\frac{1}{2}}$, so fair comparison should be made versus L_s rather than L.

Let us define $d_{\text{inter}}(v)$ as the inter-cluster distance operator that takes as input a real-value vector v of dimension n, then returns the distance between the centroids of the clusters formed by the set of points $\{v_i | 1 \le i \le n, x_i \in \mathcal{C}_k\}$, for $k \in \{1, 2\}$; and $d_{\text{intra}}(v)$ be the intra-cluster distance operator that returns the standard deviation within clusters. As the purpose of clustering analysis is to produce clusters conforming to the intrinsic classes of data points, with low variance within a cluster and large distance between clusters, the following proposition (see the proof in the appendix) shows that the performance of the classical normalized spectral clustering is practically the same as the one with centered similarities for high dimensional data. In other terms, the high dimensional performance of Laplacian spectral clustering on data samples of size $n_{[u]}$ is retrieved from the limiting results in Theorem 3 at $\theta \to +\infty$ (when spectral clustering leads to non-trivial partitioning). This remark is subsequently validated on simulations in Figure 2, where the empirical performance of Laplacian spectral clustering is found to closely match the theoretical performance of the centered similarity approach when letting the learning process guided completely by unlabelled data.

Proposition 5 *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,*

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

for non-trivial clustering results in the sense that $d_{\text{inter}}(v_{\text{lap}}), d_{\text{inter}}(v_{\text{ctr}}) \neq 0$.

As explained before, the solution $f_{[u]}$ of the centered similarities regularization can be expressed as $f_{[u]} = \left(\alpha I_{n_{[u]}} - \hat{W}_{[uu]}\right)^{-1}\hat{W}_{[ul]}f_{[l]}$ for some $\alpha > \|\hat{W}_{[uu]}\|$. Clearly, as $\alpha \downarrow \|\hat{W}_{[uu]}\|$, $f_{[u]}$ tends to align to the eigenvector of $\hat{W}_{[uu]}$ associated with the largest eigenvalue, and we thus retrieve *the performance of spectral clustering on the unlabelled data subgraph.*

It is worth pointing out that, according to the results of (Couillet and Benaych-Georges, 2016), it may occur that the solution $v_{[u]}$ obtained by spectral clustering be pure noise, i.e., $\mathbb{E}\{v_{[u]}\} \simeq 0_{n_{[u]}}$ for all large n, p. For example, with $C_1 = C_2 = I$, we have $\mathbb{E}\{v_{[u]}\} \simeq 0_{n_{[u]}}$ unless

$$c_{[u]} > \frac{1}{(\rho_1\rho_2)^2\|\mu_1 - \mu_2\|^4},$$

suggesting that there exists a threshold for $c_{[u]}$ under which spectral clustering performs equally as random guess. This behavior of spectral clustering relates to an important phase transition phenomenon on spiked random matrix models discussed in (Couillet and Benaych-Georges, 2016) (see, e.g., (Baik and Silverstein, 2006; Benaych-Georges and Nadakuditi, 2012)). Such a phase transition phenomenon is actually associated with the fact that the proposed semi-supervised learning scheme cannot produce reasonable classification results (i.e., bounded values of r_{ctr}) by solely relying on unlabelled data information (i.e., $\theta \to +\infty$) below the phase transition threshold. Indeed, we observe from (14) in the appendix that r_{ctr} has a well-defined positive value whenever the following condition of θ is satisfied:

$$1 - \left(\frac{\theta}{1+\theta}\right)^2 \frac{q(\theta)}{(\rho_1\rho_2)^2 c_{[u]}} > 0. \tag{17}$$

Letting $\theta \to +\infty$ in the case of $C_1 = C_1 = C$, for which $q(\theta) = 1/\|\mu_1 - \mu_2\|^4$ according to (15), we find the inequality condition (17) of $c_{[u]}$ to coincide with the phase transition threshold in (3), as expected. Generally speaking, a certain value θ' of θ is attainable through the adjustment of α if the inequality (17) is satisfied at $\theta = \theta'$. As such, we note importantly that *the attainable range of* θ *can only enlarge with greater* $c_{[u]}$.

It is obvious by looking at (14) that, at the same value of θ, r_{ctr} is *a strictly decreasing function of both* $c_{[l]}$ *and* $c_{[u]}$. Combining this observation with the remark that the attainable range of θ can only broaden with larger $c_{[u]}$ and is not affected by the value of $c_{[l]}$, we deduce straightforwardly that, with an appropriately chosen α, the performance of the proposed method consistently benefits from the addition of input data, *whether labelled or unlabelled*, as illustrated in Figure 1.

The following conclusion summarizes the main remarks obtained above.

Conclusion 1 *The proposed centered similarities regularization, implemented by Algorithm 1 with the hyperparameter* α*, allows one to*

1. *recover the high dimensional performance of Laplacian regularization at* $\alpha \to +\infty$*;*
2. *recover the high dimensional performance of spectral clustering at* $\alpha \downarrow \|\hat{W}_{[uu]}\|$*;*
3. *accomplish a consistent high dimensional semi-supervised learning for* α *set between the two extremes, thus leading to an increasing performance gain over Laplacian regularization with greater amounts of unlabelled data.*

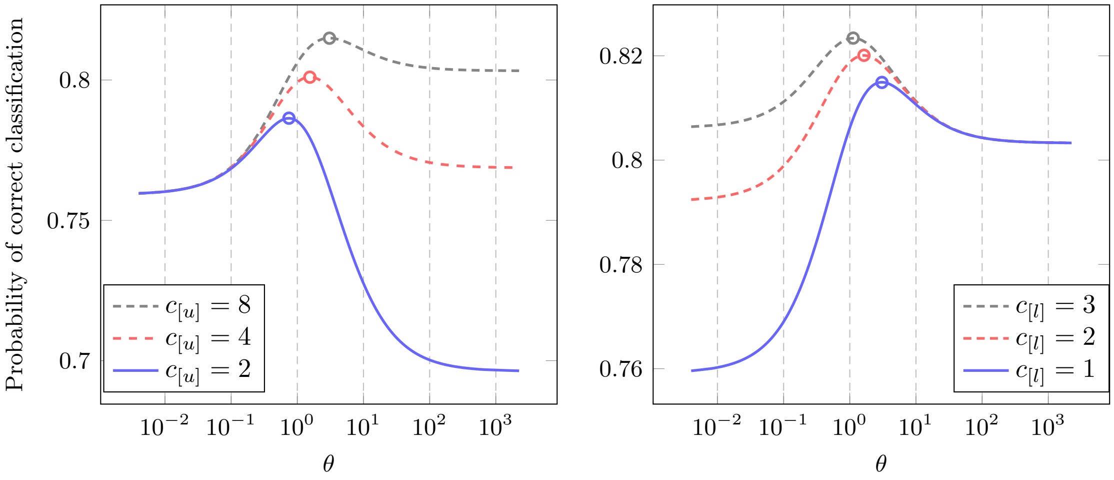

Figure 1: Asymptotic probability of correct classification as a function of θ with $\rho_1 = \rho_2$, $p = 100$, $\mu_1 = -\mu_2 = [-1, 0, \ldots, 0]^\mathsf{T}$, $\{C\}_{i,j} = .1^{|i-j|}$. Left: various $c_{[u]}$ with $c_{[l]} = 1$. Right: various $c_{[l]}$ with $c_{[u]} = 8$. Optimal values marked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

4. Experimentation

The objective of this section is to provide empirical evidence to support the proposition of centered similarities regularization, by comparing it with Laplacian regularization through simulations under and beyond the settings of the theoretical analysis.

4.1 Validation on Finite-Size Systems

We first validate the asymptotic results of the above section on finite data sets of relatively small sizes ($n, p \sim 100$). Recall from Section 3 that the asymptotic performance of Laplacian regularization and spectral clustering are recovered by centered similarities regularization at extreme values of the hyperparameter θ. In other words, the high dimensional accuracies of Laplacian regularization and spectral clustering are given by Equation (14) of Theorem 3, respectively in the limit $\theta = 0$ and $\theta = +\infty$ (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 $L^{(a)} = I - D^{-1-a} W D^{a}$ for Laplacian regularization and spectral clustering, and for the optimal (oracle) choice of the hyperparameter α for centered similarities 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 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 similarities regularization, on the other hand, is consistently above that of spectral cluster-

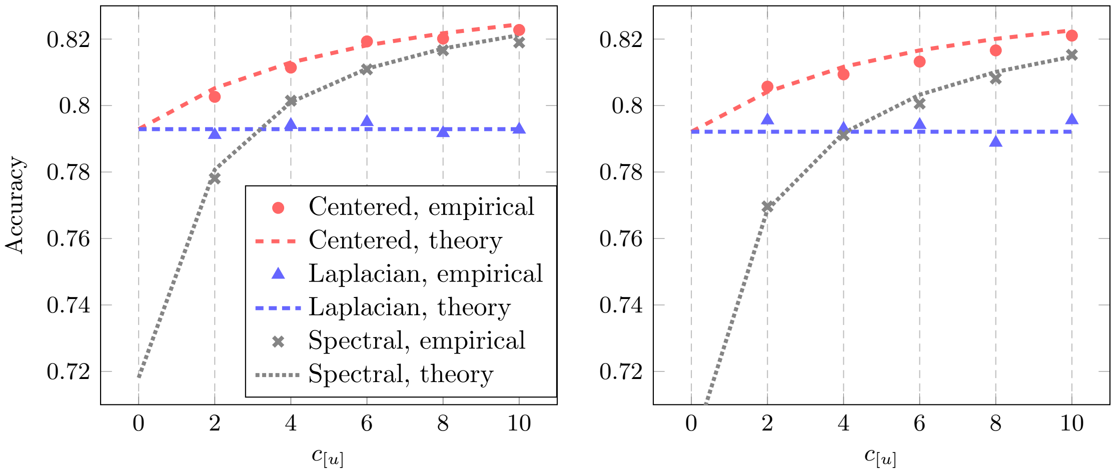

Figure 2: Empirical and theoretical accuracy as a function of $c_{[u]}$ with $c_{[l]} = 2$, $\rho_1 = \rho_2$, $p = 100$, $-\mu_1 = \mu_2 = [-1, 0, \ldots, 0]^\mathsf{T}$, $C = I_p$ (left) or $\{C\}_{i,j} = .1^{|i-j|}$ (right). Graph constructed with $w_{ij} = e^{-\|x_i - x_j\|^2/p}$. Averaged over $50000/n_{[u]}$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

ing, 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 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 of the above section.

4.2 Beyond the Model Assumptions

After verifying the advantage of the proposed centered similarities regularization in a finite (not too large) sample setting, we are now interested in examining the extent of its superiority beyond the analysis framework.

As thoroughly discussed in Subsection 2.3, the key element causing the unlabelled data learning inefficiency of Laplacian regularization is the negligible distinction between inter-class and intra-class similarities, induced by the *distance concentration* of high dimensional data. It is important to understand that this concentration phenomenon is essentially irrespective of the Gaussianity of the data. 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 $x_i = \mu_k + C_k z_i$, 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.

As a side comment, it worth pointing out that the $k-$nearest neighbors (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 of the whole data sets. 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 that the proposed centering procedure may also be advantageous to KNN graphs.

Upon the above remarks, we expect the advantage of the proposed method to manifest itself on practical datasets, whenever a weak difference between inter-class and intra-class similarities is observed (and whenever the data themselves or the relevant features to classify are obviously not too far from a mixture model). The exact convergence of all distances to a common limit is of course an extreme mathematically ideal scenario; to gain an actual sense of how the Laplacian regularization and the proposed centered similarities approaches behave under different levels of distance concentration, we provide first, as a real-life example, simulations on datasets from the standard MNIST database of handwritten digits (LeCun, 1998). These are depicted in Figures 3–4.

As the performance of the methods tends to depend on the similarity graph, for a fair and extensive comparison of Laplacian and centered similarities regularizations, the results displayed here are obtained on their respective best performing graphs, selected among commonly used graphs including KNN graphs with various numbers of neighbors $k = \{2^1, \ldots, 2^q\}$, for q the largest integer such that $2^q < n$, and graphs constructed by Gaussian (also called RBF) kernels, i.e., $w_{ij} = e^{-\|x_i - x_j\|^2/\sigma^2}$, with bandwidth σ set to the average data vector distance. The hyperparameters of the Laplacian and centered similarities regularization approaches are set optimally within the admissible range.[1]

Figure 3 shows that high classification accuracy is easily obtained on MNIST data, even with the classical Laplacian approach. However, it exhibits an unsatisfactory learning efficiency when compared to the proposed method. We also find that the benefit of the proposed algorithm is more perceptible on the classification task displayed in the right of Figure 3 (digits 3 versus 5) than on the left task (digits 8 versus 9), for which the difference between inter-class and intra-class distances is more apparent (and thus, in our setting, too "trivial"). To further evidence the impact of non-trivial classification, 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, and so is the performance gain generated by the centered similarities approach; this is indeed corroborated by Figure 4, demonstrating extremely large performance gains produced by the proposed method. In the right of Figure 4 where the similarity information is seriously disrupted by the noise, we observe the anticipated saturation effect when increasing $n_{[u]}$ for the Laplacian regularization, in contrast to the growing performance of the proposed approach. This suggests, in conclusion, that the

[1] Specifically, the hyperparameter a of Laplacian regularization is searched among the values from -2 to 0 with a step of 0.02, and the hyperparameter α of centered similarities regularization within the grid $\alpha = (1 + 10^t)\|\hat{W}_{[uu]}\|$ where t varies from -3 to 3 with a step of 0.1. The results outside these ranges are observed to be non-competitive.

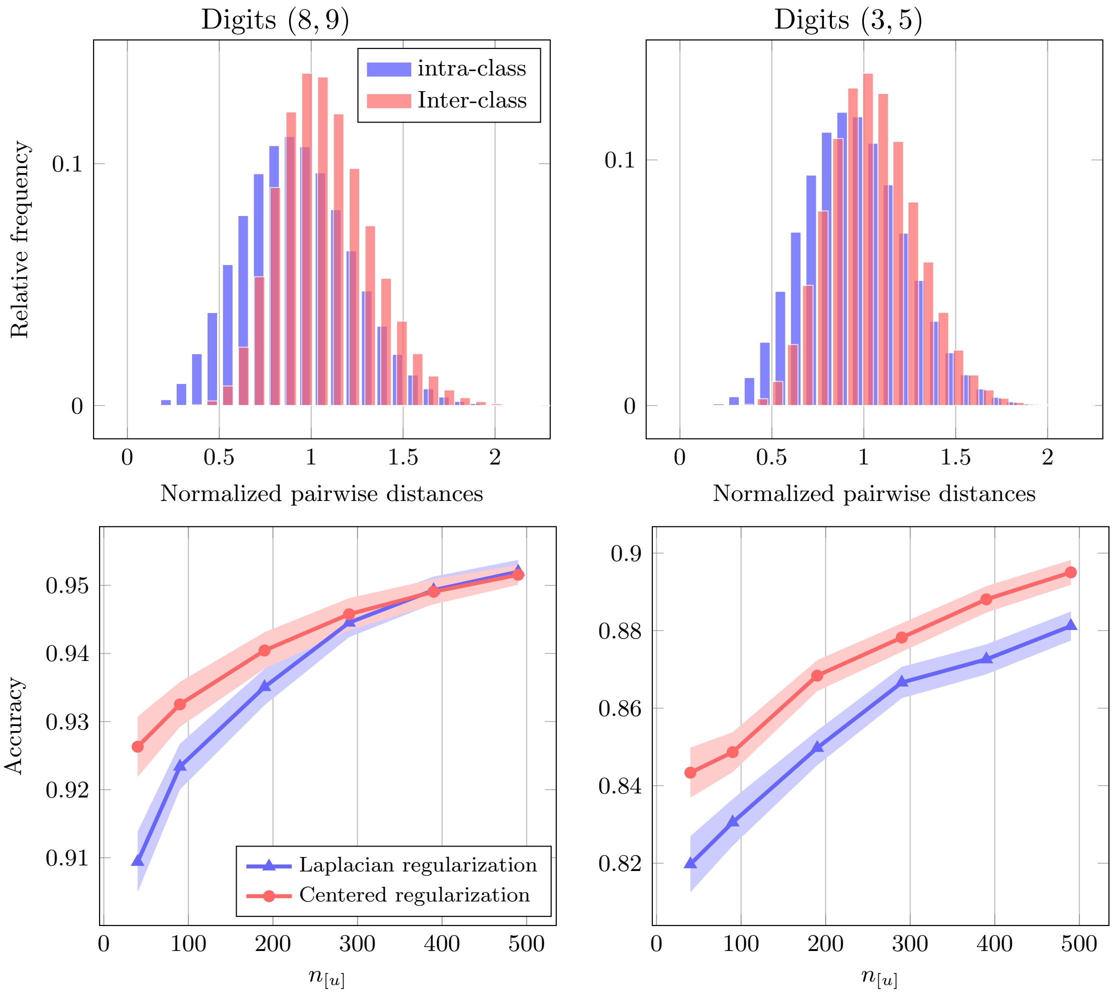

Figure 3: Top: distribution of normalized pairwise distances $\|x_i - x_j\|^2/\bar{\delta}$ $(i \neq j)$ with $\bar{\delta}$ the average of $\|x_i - x_j\|^2$ for 2-class MNIST data. Bottom: average accuracy as a function of $n_{[u]}$ with $n_{[l]} = 10$, computed over 1000 random realizations with 99%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

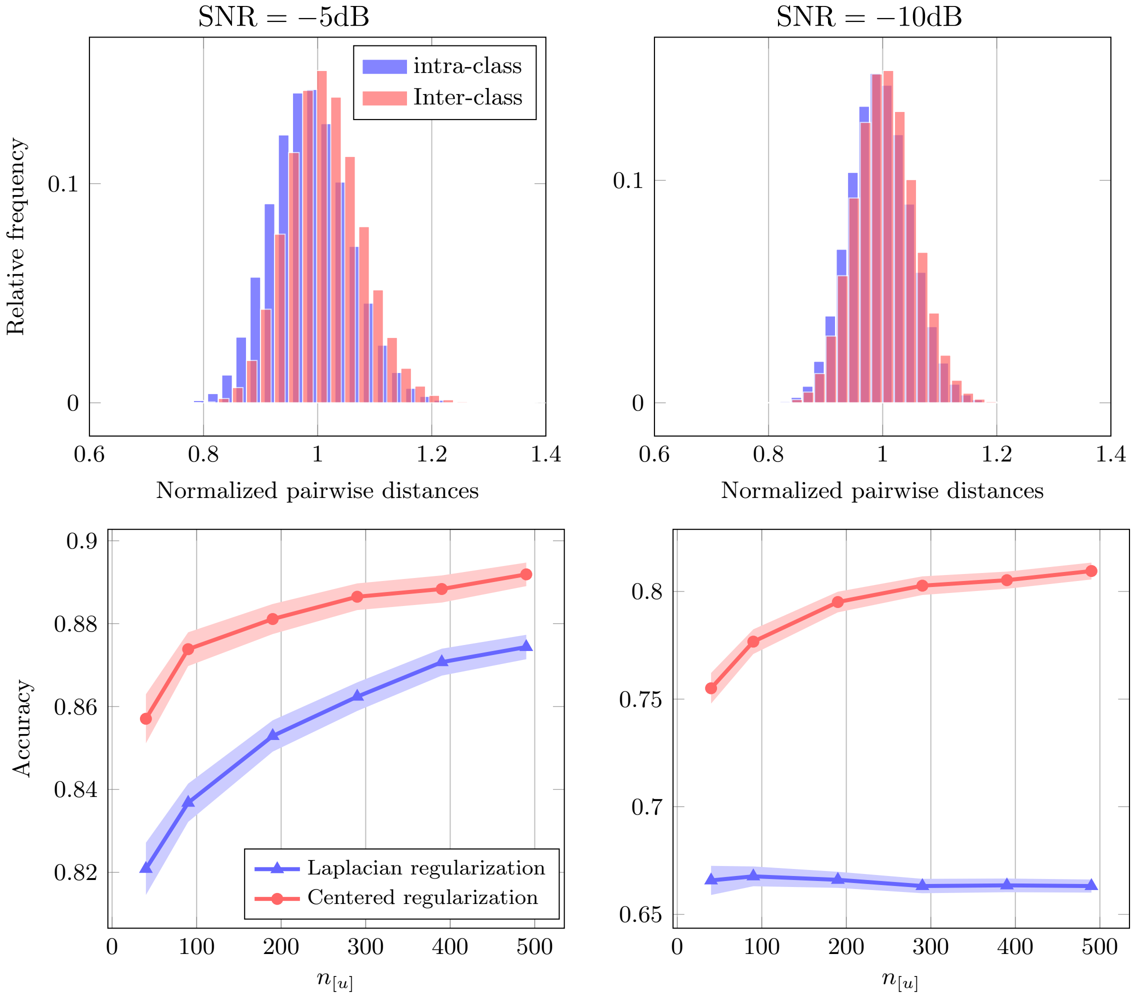

Figure 4: Top: distribution of normalized pairwise distances $\|x_i - x_j\|^2/\bar{\delta}$ $(i \neq j)$ with $\bar{\delta}$ the average of $\|x_i - x_j\|^2$ for noised MNIST data (8,9). Bottom: average accuracy as a function of $n_{[u]}$ with $n_{[l]} = 10$, computed over 1000 random realizations with 99% confidence intervals represented by shaded regions.

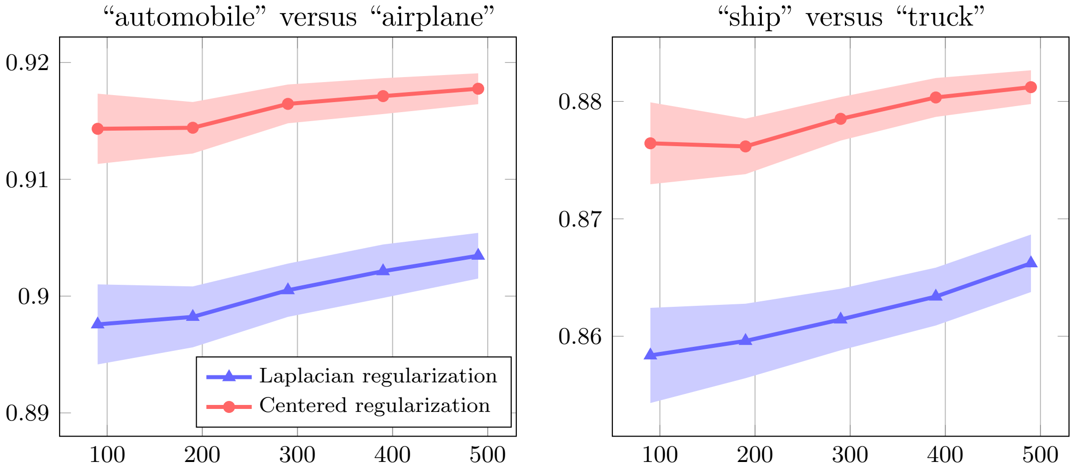

Figure 5: Average accuracy on two-class Cifar10 data as a function of $n_{[u]}$ with $n_{[l]} = 10$, computed over 1000 random realizations with 99% confidence intervals represented by shaded regions.

centered similarities approach is a privileged solution in all situations, but is especially meaningful when the distinction between intra-class and inter-class similarities is quite subtle.

In order to further stress the advantage of the proposed method on more challenging datasets, we subsequently compare the Laplacian and centered similarities regularization methods on the popular Cifar10 database (Krizhevsky et al., 2014). To obtain meaningful results, the data went through a feature extraction step using the standard pre-trained ResNet-50 network (He et al., 2016). Other experimental settings are the same as for the above MNIST data. The simulations are reported in Figure 5, where the findings confirm again the superiority of the proposed centered similarities approach.

5. Concluding Remarks

The key to the proposed semi-supervised learning method lies in the replacement of conventional Laplacian regularizations by a centering operation on similarities. The motivation behind this operation is rooted in the large dimensional concentration of pairwise-data distances and thus likely extends beyond the present graph-based semi-supervised learning schemes. It would in particular be interesting to know whether other advanced learning models involving Laplacian regularizations benefit from the same update. A specific example is Laplacian support vector machines (Laplacian SVMs) (Belkin et al., 2006), which is another widespread semi-supervised learning algorithm. Answering this question about Laplacian SVMs is however not a straightforward extension of the present analysis. Unlike the outcomes of Laplacian regularization, Laplacian SVMs are learned through an optimization problem without an explicit solution; additional technical tools, such as those recently devised in the work of El Karoui et al. (2013), to deal with implicit objects are required for analyzing their performance.

As already anticipated by the theoretical results, it is not surprising that the proposed centered similarities regularization empirically produces large performance gains over the standard Laplacian regularization when the aforementioned distance concentration problem is severe on the experimented data. However, it is quite illuminating to observe that even on datasets with weak distance concentration, for which the standard Laplacian approach exhibits a clear performance growth with respect to unlabelled data, the advantage of the proposed algorithm is still preserved. This attests to the general potential of such high dimensional studies for improving machine learning algorithms by identifying and settling some underlying issues compromising the learning performance, which would be difficult to spot if not through high dimensional analyses.

6. Acknowledgements

This work is supported by the ANR Project RMT4GRAPH (ANR-14-CE28-0006) and by the IDEX GSTATS Chair at University Grenoble Alpes.

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.

Theorem 6 *Let Assumption 1 hold, h be three-times continuously differentiable in a neighborhood of τ, and $f_{[u]}$ be the solution of* (10) *with fixed norm $n_{[u]}e^2$. Then, for $n_{[l]}+1 \leq i \leq n$ (i.e., x_i unlabelled) and $x_i \in \mathcal{C}_k$,*

$$f_i = g_i + o_P(1)$$

where

$$g_i \sim \mathcal{N}\left((-1)^k(1-\rho_k)m, \sigma_k\right)$$

for some $m, \sigma_k^2 > 0$. More precisely, defining

$$\theta = \frac{c_{[u]}m}{2c_{[l]}},$$

letting

$$\begin{aligned} \nu_k &= \begin{bmatrix} -2h'(\tau)\mu_k^{\mathsf{T}} & h''(\tau)\operatorname{tr} C_k \end{bmatrix}^{\mathsf{T}} \\ \Sigma_k &= \begin{bmatrix} 4h'(\tau)^2 C_k & 0_{p\times 1} \\ 0_{1\times p} & h''(\tau)^2 \operatorname{tr} {C_k}^2 \end{bmatrix} \end{aligned}$$

and $s : (0, \|(\rho_1\sigma_1 + \rho_2\Sigma_2) + \rho_1\rho_2(\nu_1-\nu_2)(\nu_1-\nu_2)^{\mathsf{T}}\|) \to (0,+\infty)$ be the injective function given by

$$s(\xi) = \xi(\nu_1-\nu_2)^{\mathsf{T}}\left\{I_{p+1} - \xi\left[(\rho_1\sigma_1 + \rho_2\Sigma_2) + \rho_1\rho_2(\nu_1-\nu_2)(\nu_1-\nu_2)^{\mathsf{T}}\right]\right\}^{-1}(\nu_1-\nu_2),$$

the values of m *and* σ_k^2 *are determined by* $\rho_1\rho_2 m^2 + \rho_1\sigma_1^2 + \rho_2\sigma_2^2 = e^2$ *and*

$$\frac{\sigma_k^2}{m^2} = \omega_k(\theta) + \left(\frac{\theta}{1+\theta}\right)^2 \frac{q_k(\theta)}{\rho_1\rho_2 c_{[u]}}\left(1 + \frac{\rho_1\sigma_1^2+\rho_2\sigma_2^2}{\rho_1\rho_2 m^2}\right) + \left(\frac{1}{1+\theta}\right)^2 \frac{q_k(\theta)}{\rho_1\rho_2 c_{[l]}}$$

where

$$q_k(\theta) = \frac{\operatorname{tr}\left(Q(\theta)^{-1}\Sigma_k\right)^2}{p\left[(\nu_1-\nu_2)^{\mathsf{T}}Q(\theta)^{-1}(\nu_1-\nu_2)\right]^2}$$

$$\omega_k(\theta) = \frac{(\nu_1-\nu_2)^{\mathsf{T}}Q(\theta)^{-1}\Sigma_k Q(\theta)^{-1}(\nu_1-\nu_2)}{\left[(\nu_1-\nu_2)^{\mathsf{T}}Q(\theta)^{-1}(\nu_1-\nu_2)\right]^2}$$

with $Q(\theta) = I_{p+1} - s^{-1}(\theta)\left[(\rho_1\sigma_1 + \rho_2\Sigma_2) + \rho_1\rho_2(\nu_1-\nu_2)(\nu_1-\nu_2)^{\mathsf{T}}\right]$.

A.2 Sketch of Proof

From the form of the solution (3), Theorem 6 can be proved in all accuracy exploiting advanced tools from random matrix theory, by merging for instance the arguments from the work of Couillet and Benaych-Georges (2016) and and that of Mai and Couillet (2017), however at the cost of very cumbersome and not fully insightful mathematical details. In the present section, we instead propose a more "intuitive" approach based on a perturbation argument of the "leave-one-out" type. Additionally to the notations given in the end of Introduction, 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 (through standard high dimensional statistical manipulations not detailed here) that $\hat{W} = O_{\|\cdot\|}(1)$ and

$$\hat{W} = \frac{1}{p}\hat{\Phi}^{\mathsf{T}}\hat{\Phi} + O_{\|\cdot\|}(p^{-\frac{1}{2}})$$

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

$$\phi(x_i) = \begin{bmatrix} -2h'(\tau)x_i^{\mathsf{T}} & h''(\tau)\|x_i\|^2 \end{bmatrix}^{\mathsf{T}}.$$

Define $\nu_k = \mathbb{E}\{\phi(x_i)\}$ for $x_i \in \mathcal{C}_k$, $k \in \{1, 2\}$, and let $w_i = \phi(x_i) - \nu_k$ (i.e., $\mathbb{E}\{w_i\} = 0$) with $\Sigma_k = \mathbb{E}\{w_i w_i^{\mathsf{T}}\}$.

Recall that $\alpha f_{[u]} = \hat{W}_{[uu]}f_{[u]} + \hat{W}_{[ul]}f_{[l]}$ according to (11). Then

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

where $\hat{\Phi} = \begin{bmatrix}\hat{\Phi}_{[l]} & \hat{\Phi}_{[u]}\end{bmatrix}$. Decomposing the above equation for any $i > n_{[l]}$ (i.e., x_i unlabelled) leads to

$$\begin{aligned}\alpha f_i &= \frac{1}{p}\hat{\phi}(x_i)^{\mathsf{T}}\hat{\phi}(x_i)f_i + \frac{1}{p}\hat{\phi}(x_i)^{\mathsf{T}}\hat{\Phi}_{[u]}^{\{i\}}f_{[u]}^{\{i\}} + \frac{1}{p}\hat{\phi}(x_i)^{\mathsf{T}}\hat{\Phi}_{[l]}f_{[l]} + O(p^{-\frac{1}{2}})\\ \alpha f_{[u]}^{\{i\}} &= \frac{1}{p}\hat{\Phi}_{[u]}^{\{i\}\mathsf{T}}\hat{\phi}(x_i)f_i + \frac{1}{p}\hat{\Phi}_{[u]}^{\{i\}\mathsf{T}}\hat{\Phi}_{[u]}^{\{i\}}f_{[u]}^{\{i\}} + \frac{1}{p}\hat{\Phi}_{[u]}^{\{i\}\mathsf{T}}\hat{\Phi}_{[l]}f_{[l]} + O(p^{-\frac{1}{2}}) \qquad (18)\end{aligned}$$

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

Define the "leave-one-out" dataset $X_{(i)} = \{x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n\} \in \mathbb{R}^{(n-1)\times p}$ 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 $\hat{W}$,

$$\hat{W}^{(i)} = \frac{1}{p}\hat{\Phi}^{(i)\mathsf{T}}\hat{\Phi}^{(i)} + O_{\|\cdot\|}(p^{-\frac{1}{2}})$$

where $\hat{\Phi}^{(i)} = [\hat{\phi}^{(i)}(x_1), \ldots, \hat{\phi}^{(i)}(x_{i-1}), \hat{\phi}^{(i)}(x_{i+1}), \ldots, \hat{\phi}^{(i)}(x_n)] = [\phi(x_1), \ldots, \phi(x_{i-1}), \phi(x_{i+1}), \ldots, \phi(x_n)]P_{n-1}$. The solution $f_{[u]}^{(i)}$ of the centered similarities regularization on the "leave-one-out" dataset $X_{(i)}$ is given by

$$\alpha f_{[u]}^{(i)} = \frac{1}{p}\hat{\Phi}_{[u]}^{(i)\mathsf{T}}\hat{\Phi}_{[u]}^{(i)} f_{[u]}^{(i)} + \frac{1}{p}\hat{\Phi}_{[u]}^{(i)\mathsf{T}}\hat{\Phi}_{[l]} f_{[l]} + O(p^{-\frac{1}{2}}) \tag{19}$$

where $\hat{\Phi}^{(i)} = \begin{bmatrix}\hat{\Phi}_{[l]}^{(i)} & \hat{\Phi}_{[u]}^{(i)}\end{bmatrix}$.

As

$$\frac{1}{p}\hat{\Phi}_{[u]}^{(i)\mathsf{T}}\hat{\Phi}_{[u]}^{(i)} - \frac{1}{p}\hat{\Phi}_{[u]}^{\{i\}\mathsf{T}}\hat{\Phi}_{[u]}^{\{i\}} = O(n^{-1}), \tag{20}$$

subtracting (19) from (18) gives

$$\left(\alpha I_{n-1} - \frac{1}{p}\hat{\Phi}_{[u]}^{(i)\mathsf{T}}\hat{\Phi}_{[u]}^{(i)}\right)\left(f_{[u]}^{\{i\}} - f_{[u]}^{(i)}\right) = \frac{1}{p}\hat{\Phi}_{[u]}^{(i)\mathsf{T}}\hat{\phi}(x_i)f_i + O(p^{-\frac{1}{2}}). \tag{21}$$

Set $\beta = \frac{1}{p}\hat{\Phi}f = O_{\|\cdot\|}(1)$, and its "leave-one-out" version $\beta^{(i)} = \frac{1}{p}\hat{\Phi}^{(i)}f^{(i)}$. Using (20) and (21) leads to

$$\alpha f_i = \beta^{(i)\mathsf{T}}\hat{\phi}(x_i) + \frac{1}{p^2}\hat{\phi}(x_i)^{\mathsf{T}}\hat{\Phi}_{[u]}^{(i)}\left(\alpha I_{n-1} - \frac{1}{p}\hat{\Phi}_{[u]}^{(i)\mathsf{T}}\hat{\Phi}_{[u]}^{(i)}\right)^{-1}\hat{\Phi}_{[u]}^{(i)\mathsf{T}}\hat{\phi}(x_i)f_i + \frac{1}{p}\|\hat{\phi}(x_i)\|^2 f_i + O(p^{-\frac{1}{2}}). \tag{22}$$

By standard random matrix arguments, we have

$$\begin{aligned}
&\frac{1}{p^2}\hat{\phi}(x_i)^{\mathsf{T}}\hat{\Phi}_{[u]}^{(i)}\left(\alpha I_{n-1} - \frac{1}{p}\hat{\Phi}_{[u]}^{(i)\mathsf{T}}\hat{\Phi}_{[u]}^{(i)}\right)^{-1}\hat{\Phi}_{[u]}^{(i)\mathsf{T}}\hat{\phi}(x_i)f_i + \frac{1}{p}\|\hat{\phi}(x_i)\|^2 f_i \\
&= \frac{\alpha}{p}\hat{\phi}(x_i)^{\mathsf{T}}\left(\alpha I_p - \frac{1}{p}\hat{\Phi}_{[u]}^{(i)}\hat{\Phi}_{[u]}^{(i)\mathsf{T}}\right)^{-1}\hat{\phi}(x_i)f_i \\
&= \kappa f_i + O(p^{-\frac{1}{2}})
\end{aligned} \tag{23}$$

where κ is a deterministic value dependent of $C_1, C_2, c_{[u]}, h, \alpha$. Specifically, for $C_1 = C_2 = I_p$,

$$\kappa = \frac{\alpha c_{[u]}}{2h'(\tau)} - \frac{c_{[u]} - 1}{2} - \frac{\sqrt{(\alpha h'(\tau)^{-1}c_{[u]} - (1 + c_{[u]})^2)(\alpha h'(\tau)^{-1}c_{[u]} - (1 - c_{[u]})^2)}}{2c_{[u]}}.$$

Also, it is easily shown that

$$\begin{aligned}\beta^{(i)\mathsf{T}}\hat{\phi}(x_i) &= \beta^{(i)\mathsf{T}}\left(\mathbb{E}\{\hat{\phi}(x_i)\} + \frac{n-1}{n}w_i - \frac{1}{n}\sum_{m\neq i} w_m\right)\\ &= \beta^{(i)\mathsf{T}}\left(\mathbb{E}\{\hat{\phi}(x_i)\} + \frac{n-1}{n}w_i\right) + \frac{1}{np}f^{(i)\mathsf{T}}\hat{\Phi}^{(i)\mathsf{T}}W^{(i)}1_{n-1}\\ &= \beta^{(i)\mathsf{T}}(\mathbb{E}\{\hat{\phi}(x_i)\} + w_i) + \frac{1}{np}f^{(i)\mathsf{T}}\mathbb{E}\big\{\hat{\Phi}^{(i)\mathsf{T}}W^{(i)}\big\}1_{n-1} + O(p^{-\frac{1}{2}})\\ &= \beta^{(i)\mathsf{T}}(\mathbb{E}\{\hat{\phi}(x_i)\} + w_i) + O(p^{-\frac{1}{2}})\end{aligned}$$

where $W^{(i)} = [w_1 \ldots, w_{i-1}, w_{i+1}, \ldots, w_n]$. Let us define

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

Notice that $\mathbb{E}\{\hat{\phi}(x_i)\} = \mathbb{E}\{\phi_c(x_i)\} + O(n^{-1})$, but unlike $\hat{\phi}(x_i)$, $\phi_c(x_i)$ is independent of all x_j with $j \neq i$.

Finally, we have that for $i > n_{[l]}$

$$f_i = \gamma\beta^{(i)\mathsf{T}}\phi_c(x_i) + O(p^{-\frac{1}{2}}) \tag{24}$$

with $\gamma = (\alpha - \kappa)^{-1}$. The interest of the above equation lies in that, since $\phi_c(x_i)$ is independent of $\beta_{(i)}$, the unlabelled scores f_i follow asymptotically a Gaussian distribution (indeed $g_i = \gamma\beta^{(i)\mathsf{T}}\phi_c(x_i)$ converges to a Gaussian variable according to the central limit theorem).

Moreover, taking the expectation and the variance of both sides of (24) for $x_i = \mathcal{C}_k$ yields

$$\begin{aligned}&\mathbb{E}\{f_i\} = \gamma\mathbb{E}\{\beta^{(i)}\}\nu_k + O(p^{-\frac{1}{2}})\\ &\operatorname{var}(f_i) = \mathbb{E}\{f_i^2\} - m_k^2 = \gamma^2\operatorname{tr}\big[\operatorname{cov}\{\beta^{(i)}\}(\nu_k\nu_k^{\mathsf{T}} + \Sigma_k)\big] + \gamma^2\operatorname{tr}\big[\mathbb{E}\{\beta^{(i)}\}\mathbb{E}\{\beta^{(i)}\}^{\mathsf{T}}\Sigma_k\big] + O(p^{-\frac{1}{2}}).\end{aligned}$$

Let m_k, σ_k^2 be the asymptotic limits of $\mathbb{E}\{f_i\}$, $\operatorname{var}(f_i)$ for $x_i \in \mathcal{C}_k$ unlabelled. As $\beta - \beta^{(i)} = O(p^{-\frac{1}{2}})$, we obtain

$$m_k \simeq \gamma\eta^{\mathsf{T}}\nu_k \tag{25}$$

$$\sigma_k^2 \simeq \gamma^2\operatorname{tr}\big[B(\nu_k\nu_k^{\mathsf{T}} + \Sigma_k)\big] + \gamma^2\operatorname{tr}\big[\eta\eta^{\mathsf{T}}\Sigma_k\big] \tag{26}$$

for η, B some asymptotic equivalents of $\mathbb{E}\{\beta\}$, $\operatorname{cov}\{\beta\}$ such that $\eta - \mathbb{E}\{\beta\} = O(p^{-\frac{1}{2}})$, $\operatorname{tr}B - \operatorname{tr}\mathbb{E}\{\beta\} = O(p^{-\frac{1}{2}})$.

Additionally, substituting (24) into $\beta = \frac{1}{p}\hat{\Phi}f$ gives

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

Write $w_i = \tilde{w}_i + \Sigma_k \beta^{(i)}/(\beta^{(i)\mathsf{T}} w_i)$ such that $\mathbb{E}\{\beta^{(i)\mathsf{T}}\phi_c(x_i)w_i\} = \mathbb{E}\{\beta^{(i)\mathsf{T}} w_i w_i\} = \mathbb{E}\{^{(i)\mathsf{T}} w_i e_i\} = \Sigma_k \mathbb{E}\{\beta^{(i)}\}$ for $x_i \in \mathcal{C}_k$. In other words, $\mathbb{E}\{\beta_{(i)}^{\mathsf{T}}\phi_c(x_i)\tilde{w}_i\} = 0$. Then,

$$\beta = \frac{1}{p}\sum_{i=1}^{n_{[l]}} f_i \phi_c(x_i) + \frac{1}{p}\sum_{i=n_{[l]}+1}^{n} \gamma(\beta^{(i)\mathsf{T}}\nu_{x_i}\nu_{x_i} + \beta^{(i)\mathsf{T}}\nu_{x_i} w_i + \beta^{(i)\mathsf{T}} w_i \tilde{w}_i + \Sigma_{x_i}\beta^{(i)}) + O_{\|\cdot\|}(p^{-\frac{1}{2}}) \tag{27}$$

where ν_{x_i}, Σ_{x_i} stand respectively for the mean and the covariance matrix of $\phi_c(x_i)$ (i.e., $\nu_{x_i} = \nu_k$, $\Sigma_{x_i} = \Sigma_k$ for $x_i \in \mathcal{C}_k$). As $\beta = \beta^{(i)} + O(p^{-\frac{1}{2}})$, Equation (27) becomes

$$\left(I_p - \gamma c_{[u]}\sum_{a=1}^{2}\rho_a\Sigma_a\right)\beta = \frac{1}{p}\sum_{i=1}^{n_{[l]}} f_i\phi_c(x_i) + \frac{1}{p}\sum_{i=n_{[l]}+1}^{n}\gamma(\beta^{(i)\mathsf{T}}\nu_{x_i}\nu_{x_i} + \beta^{(i)\mathsf{T}}\phi_c(x_i)\tilde{w}_i) + O_{\|\cdot\|}(p^{-\frac{1}{2}}).$$

Computing the expectation and the covariance matrix of both sides of the above equation, we get

$$\begin{aligned}
\eta &\simeq Q^{-1}\rho_1\rho_2[2c_{[l]} + (m_2 - m_1)c_{[u]}](\nu_2 - \nu_1) \\
B &\simeq Q^{-1}\left[\rho_1\rho_2[4c_{[l]} + (m_2 - m_1)^2 c_{[u]}]\sum_{a=1}^{2}(1-\rho_a)\Sigma_a + c_{[u}\sum_{a=1}^{2}(1-\rho_a)\sigma_a^2\Sigma_a\right]Q^{-1}
\end{aligned}$$

where $Q = I_p - \gamma c_{[u]}\sum_{a=1}^{2}\rho_a\Sigma_a$.

Substituting the above expressions of η, B into (25) and (26) yields four asymptotic equations of the five unknown variables m_k, σ_k^2 and γ. Adding $\sum_{a=1}^{2}\rho_a(m_a^2 + \sigma_a^2) = e^2$ to these equation allows to determine the asymptotic values of m_k, σ_k^2 and γ. Letting $\xi = c_{[u]}\gamma$, we retrieve the equations of Theorem 6.

Appendix B. Sketch of Proof for Proposition 5

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

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

Note that $\|L'\| = O(1)$, and if v is an eigenvector of L_s associated with the eigenvalue u, then it is also an eigenvector of L' 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'. The second smallest eigenvector v_{lap} of L_s is the same as the largest eigenvector of L'.

From the random matrix equivalent of L' given in (Couillet and Benaych-Georges, 2016), we have

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

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

For $k \in \{1,2\}$, define $j_k \in \mathbb{R}^n$ the indicator vector of class k with $[j_k]_i = 1$ if $x_i \in \mathcal{C}_k$, otherwise $[j_k]_i = 0$. Then, we have

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

for some $v \in \mathbb{R}^n$. According to the spectral analysis given in (Couillet and Benaych-Georges, 2016), there can be only one eigenvector of L' ($\hat{W}$, resp.) whose limiting scalar product with j_k for $k \in \{1,2\}$ is bounded away from zero: this vector is v_{lap} (v_{ctr}, resp.). Denote by λ_{lap} the eigenvalue of $h(\tau)L'$ associated with v_{lap}, and λ_{ctr} the eigenvalue of $\hat{W}$ associated with v_{ctr}. The theoretical results in (Couillet and Benaych-Georges, 2016) point out that the eigenvalue λ_{lap} of $h(\tau)L'$ is at macroscopic distance from other eigenvalues of $h(\tau)L'$. The same can be said about $\hat{W}$ and its eigenvalue λ_{ctr}.

Let γ be a positively oriented closed path circling only around λ_{lap} and λ_{ctr}. Then by Cauchy's formula,

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

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

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

As $\frac{1}{\sqrt{n_k}} j_k^\mathsf{T}(h(\tau)L' - zI_n)^{-1}\psi = o_P(1)$ by the arguments in (Couillet and Benaych-Georges, 2016), we get

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

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 5.

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. URL http://jmlr.org/papers/v18/17-151.html.

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

J. Baik and J. W. Silverstein. Eigenvalues of large sample covariance matrices of spiked population models. *Journal of Multivariate Analysis*, 97(6):1382–1408, 2006.

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 *COLT*, pages 33–44, 2008.

F. Benaych-Georges and R. R. Nadakuditi. The singular values and vectors of low rank perturbations of large rectangular random matrices. *Journal of Multivariate Analysis*, 111:120–135, 2012.

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.

Olivier Chapelle, Bernhard Scholkopf, and Alexander Zien. Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews]. *IEEE Transactions on Neural Networks*, 20 (3):542–542, 2009.

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

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

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.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Alexander Hinneburg, Charu C Aggarwal, and Daniel A Keim. What is the nearest neighbor in high dimensional spaces? In *26th Internat. Conference on Very Large Databases*, pages 506–515, 2000.

Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. The cifar-10 dataset. *online: http://www. cs. toronto. edu/kriz/cifar. html*, 2014.

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

Xiaoyi Mai and Romain Couillet. A random matrix analysis and improvement of semi-supervised learning for large dimensional data. *arXiv preprint arXiv:1711.03404*, 2017.

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.

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.

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.

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.

Xiaojin Zhu, Zoubin Ghahramani, and John D Lafferty. Semi-supervised learning using gaussian fields and harmonic functions. In *Proceedings of the 20th International conference on Machine learning (ICML-03)*, pages 912–919, 2003.