Journal of Machine Learning Research 19 (2018) 1-27 Submitted 7/17; Revised 11/18; Published 11/18

A Random Matrix Analysis and Improvement of Semi-Supervised Learning for Large Dimensional Data

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

Editor: Nicolas Vayatis

Abstract

This article provides an original understanding of the behavior of a class of graph-oriented semi-supervised learning algorithms in the limit of large and numerous data. It is demonstrated that the intuition at the root of these methods collapses in this limit and that, as a result, most of them become inconsistent. Corrective measures and a new data-driven parametrization scheme are proposed along with a theoretical analysis of the asymptotic performances of the resulting approach. A surprisingly close behavior between theoretical performances on Gaussian mixture models and on real data sets is also illustrated throughout the article, thereby suggesting the importance of the proposed analysis for dealing with practical data. As a result, significant performance gains are observed on practical data classification using the proposed parametrization.

Keywords: semi-supervised learning, kernel methods, random matrix theory, high dimensional statistics

1. Introduction

Semi-supervised learning consists in classification schemes combining few labelled and numerous unlabelled data. With the advent of the big-data paradigm, where supervised learning implies the impossible pre-labelling of sometimes millions of samples, these so-far marginal methods are attracting a renewed attention. Its appeal also draws on its providing an alternative to unsupervised learning which excludes the possibility to exploit known data. We refer to Chapelle et al. (2006) for an overview.

An important subset of semi-supervised learning methods concerns graph-based approaches. In these, one considers data instances $x_1, \cdots, x_n \in \mathbb{R}^p$ as vertices on a graph with edge weights W_{ij} encoding their similarity, which is usually defined through a kernel function f, as with radial kernels of the type $W_{ij} = f(\|x_i - x_j\|^2/p)$ which we shall focus on in this article. The motivation follows from one's expectation that two instances with a strong edge weight tend to belong to the same class and thus vertices of a common class tend to aggregate. Standard methods for recovering the classes of the unlabelled data then consist in various random walk (Jaakkola and Szummer, 2002) or label propagation (Zhu and Ghahramani, 2002) algorithms on the graph which softly allocate "scores" for each

node to belong to a particular class. These scores are then compared for each class in order to obtain a hard decision on the individual unlabelled node class. A popular, and widely recognized as highly performing, example is the PageRank approach (Avrachenkov et al., 2011).

Many of these algorithms also have the particularity of having a closed-form and quite interrelated expression for their stationary points. These stationary points are also often found to coincide with the solutions to optimization problems under constraints, independently established. This is notably the case of Zhu et al. (2003) under equality constraints for the labelled nodes or of Belkin et al. (2004) where a relaxation approach is used instead to allow for modifications of the value of labelled nodes – this ensuring that erroneously labelled data or poorly informative labelled data do not hinder the algorithm performance. As is often the case in graph-related optimization, a proper choice of the matrix representative of the inter-data affinity is at the core of scientific research and debates and mainly defines the differences between any two schemes. In particular, Joachims et al. (2003) suggests the use of a standard Laplacian representative, where Zhou et al. (2004) advises for a normalized Laplacian approach. These individual choices correspondingly lead to different versions of the label propagation methods on the graph, as discussed in Avrachenkov et al. (2011).

There also exists another branch of manifold based semi-supervised learning (Belkin and Niyogi, 2004; Goldberg et al., 2009; Moscovich et al., 2016). In contrast to the methods discussed in this paper, these approaches involve a step of manifold learning, which plays a decisive role in the success of the learning task. While there exist many articles providing theoretical analyses for such methods (Wasserman and Lafferty, 2008; Bickel et al., 2007; Moscovich et al., 2016; Globerson et al., 2017), a comprehensive comparison to the graph-based methods presently discussed is beyond current analytical reach. This being said, while the Gaussian-mixture data model under study in the present article violates the manifold assumption, given appropriate feature (kernel function) mapping, there exists a low dimensional manifold where data demonstrate a clustering behavior, as shown by Couillet and Benaych-Georges (2015); as such, when the classes are very well separated and sufficient data are available to estimate the manifold, manifold-based methods in this setting should lead to competitive performance. While clearly out of our present scope, future investigations might allow for a comparative study of manifold versus graph approaches. Another recent line of alternative works consider SSL from a graph signal processing perspective (Narang et al., 2013a,b; Gadde et al., 2014; Anis et al., 2015), where the classification scores are viewed as smooth signals on the similarity graph and the learning task then consists in recovering a bandlimited (understood in the graph Fourier transform domain) graph signal from its known sample values.

Returning to graph-based SSL, a likely key reason for the open-ended question of a most natural choice for the graph representative arises from these methods being essentially built upon intuitive reasoning arising from low dimensional data considerations rather than from mostly inaccessible theoretical results. Indeed, the non-linear expression of the affinity matrix W as well as the rather involved form assumed by the algorithm output (although explicit) hinder the possibility to statistically evaluate the algorithm performances for all finite n, p, even for simple data assumptions. The present article is placed instead under a large dimensional data assumption, thus appropriate to the present big-data paradigm,

and proposes instead to derive, for the first time to the best of the authors' knowledge, theoretical results on the performance of the aforementioned algorithms in the large n, p limit for a certain class of statistically distributed data $x_1, \ldots, x_n \in \mathbb{R}^p$. Precisely due to the large data assumption, as we shall observe, most of the intuition leading up to the aforementioned algorithms collapse as $n, p \to \infty$ at a similar rate, and we shall prove that few algorithms remain consistent in this regime.

Specifically, recall that the idea behind graph-based semi-supervised learning is to exploit the similarity between data points and thus expect a clustering behavior of close-by data nodes. In the large data assumption (i.e., $p \gg 1$), this similarity-based approach suffers a curse of dimensionality. As the span of $\mathbb{R}^p$ grows exponentially with the data dimension p, when p is large, the data points x_i (if not too structured) are in general so sparsely distributed that their pairwise distances tend to be similar regardless of their belonging to the same class or not. The Gaussian mixture model that we define in Subsection 3 and will work on is a telling example of this phenomenon; as we show, in a regime where the classes ought to be separable (even by unsupervised methods as shown by Couillet and Benaych-Georges 2015), the normalized distance $\|x_i - x_j\|/\sqrt{p}$ of two random different data instances x_i and x_j generated from this model converges to a constant *irrespective of the class of* x_i *and* x_j in the Gaussian mixture and, consequently, the similarity defined by $W_{ij} = f(\|x_i - x_j\|^2/p)$ is asymptotically the same for all pairs of data instances. This behavior should therefore invalidate the intuition behind semi-supervised classification, hence likely render graph-based methods ineffective. As a direct consequence, the scores are *flat* in the sense that they have the same asymptotic values, irrespective of the class. Nonetheless, we will show that sensible classification on data sets generated from this model can still be achieved provided that appropriate amendments to the classification algorithms are enforced, due to the small fluctuations around these flat asymptotic limit of scores. This flat limit is reminiscent of the work by Nadler et al. (2009) where the authors show that the scores indeed share the same limit, irrespective of the class, in the presence of infinitely many unlabelled samples but for $p \geq 2$ fixed. Yet, despite the scores flatness, the authors experimentally observed non-trivial classification in binary tasks thanks to the small difference between scores; they however did not provide any theoretical support for such behavior, for they analysis failed to recover the small fluctuations.

Inspired by Avrachenkov et al. (2011), we generalize here the algorithm proposed in Zhu et al. (2003) by introducing a normalization parameter α in the cost function in order to design a large class of regularized affinity-based methods, among which are found the traditional Laplacian- and normalized Laplacian-based algorithms. The generalized optimization framework is presented in Section 2.

The main contribution of the present work is to provide a quantitative performance study of the generalized graph-based semi-supervised algorithm for large dimensional Gaussian-mixture data and radial kernels, technically following the random matrix approach developed by Couillet and Benaych-Georges (2015). Our main findings are summarized as follows:

- Irrespective of the choice of the data affinity matrix, the classification outcome is strongly biased by *the number of labelled data from each class* and unlabelled data tend

to be classified into the class with most labelled nodes; we propose a normalization update of the standard algorithms to correct this limitation.

- Once the aforementioned bias corrected, the choice of the affinity matrix (and thus of the parameter α) strongly impacts the performances; most importantly, within our framework, both *standard Laplacian* ($\alpha = 0$ here) and *normalized Laplacian-based* ($\alpha = -\frac{1}{2}$) methods, although widely discussed in the literature, fail in the large dimensional data regime. Of the family of algorithms discussed above, only the *PageRank* approach ($\alpha = -1$) is shown to provide asymptotically acceptable results.

- The scores of belonging to each class attributed to individual nodes by the algorithms are shown to asymptotically follow a *Gaussian distribution* with mean and covariance depending on the statistical properties of classes, the ratio of labelled versus unlabelled data, and the value of the first derivatives of the kernel function at the limiting value τ of $\frac{1}{p}\|x_i - x_j\|^2$ (which we recall is irrespective of the genuine classes of x_i, x_j). This last finding notably allows one to *predict the asymptotic performances* of the semi-supervised learning algorithms.

- From the latter result, three main outcomes unfold:

  - when three classes or more are considered, there exist Gaussian mixture models for which classification is shown to be *impossible*;
  - despite PageRank's consistency, we further justify that the choice $\alpha = -1$ is not in general optimal. For the case of 2-class learning, we provide a method to approach the optimal value of α; this method is demonstrated on real data sets to convey sometimes *dramatic improvements* in correct classification rates.
  - for a 2-class learning task, necessary and sufficient conditions for asymptotic consistency are: $f'(\tau) < 0$, $f''(\tau) > 0$ and $f''(\tau)f(\tau) > f'(\tau)^2$; in particular, Gaussian kernels, failing to meet the last condition, cannot deal with the large dimensional version of the "concentric spheres" task.

Throughout the article, theoretical results and related discussions are confirmed and illustrated with simulations on Gaussian-mixture data as well as the popular MNIST data (LeCun et al., 1998), which serves as a comparison for our theoretical study on real world data sets. The consistent match of our theoretical findings on MNIST data, despite their departing from the very large dimensional and Gaussian-mixture assumption, suggests that our results have a certain robustness to these assumptions and can be applied to a larger range of data. We indeed believe that, while only the limiting behavior of Gaussian mixture inputs is characterized in this article (mostly for technical reasons), the analysis reveals certain properties inherent to graph-based SSL methods, which extend well beyond the Gaussian hypothesis.

Notations: $\boldsymbol{\delta}_a^b$ is a binary function taking the value of 1 if $a = b$ or that of 0 if not. 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. The operator $\mathrm{diag}(v) = \mathrm{diag}\{v_a\}_{a=1}^k$ is the diagonal matrix having $v_1, \ldots, v_k$ as its ordered diagonal elements. $O(\cdot)$ is the same as specified in the work of Couillet and Benaych-Georges (2015): 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 $n^D \mathrm{P}(x \geq n^\eta u_n) \to 0$. When multidimensional objects are concerned, for a vector (or a diagonal matrix)v, $v = O(u_n)$ means the maximum entry in absolute value is $O(u_n)$ and for a square matrix M, $M = O(u_n)$ means that the operator norm of M is $O(u_n)$.

2. Optimization Framework

Let $x_1, \ldots, x_n \in \mathbb{R}^p$ be n data vectors belonging to K classes $\mathcal{C}_1, \ldots, \mathcal{C}_K$. The class association of the $n_{[l]}$ vectors $x_1, \ldots, x_{n_{[l]}}$ is known (these vectors will be referred to as *labelled*), while the class of the remaining $n_{[u]}$ vectors $x_{n_{[l]}+1}, \ldots, x_n$ ($n_{[l]} + n_{[u]} = n$) is unknown (these are referred to as *unlabelled* vectors). Within both labelled and unlabelled subsets, the data are organized in such a way that the $n_{[l]1}$ first vectors $x_1, \ldots, x_{n_{[l]1}}$ belong to class $\mathcal{C}_1$, $n_{[l]2}$ subsequent vectors to $\mathcal{C}_2$, and so on, and similarly for the $n_{[u]1}, n_{[u]2}, \ldots$ first vectors of the set $x_{n_{[l]}+1}, \ldots, x_n$. Note already that this ordering is for notational convenience and shall not impact the generality of our results.

The affinity relation between the vectors $x_1, \ldots, x_n$ is measured from the weight matrix W defined by

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

for some function f. The matrix W may be seen as the adjacency matrix of the n-node graph indexed by the vectors $x_1, \ldots, x_n$. We further denote by D the diagonal matrix with $D_{ii} \equiv d_i = \sum_{j=1}^n W_{ij}$ the degree of the node associated to x_i.

We next define a score matrix $F \in \mathbb{R}^{n \times K}$ with F_{ik} representing the evaluated score for x_i to belong to $\mathcal{C}_k$. In particular, following the conventions typically used in graph-based semi-supervised learning (Chapelle et al., 2006), we shall affect a unit score $F_{ik} = 1$ if x_i is a labelled data of class $\mathcal{C}_k$ and a null score for all $F_{ik'}$ with $k' \neq k$. In order to attribute classes to the unlabelled data, scores are first affected by means of the resolution of an optimization framework. We propose here

$$\begin{aligned} F &= \mathrm{argmin}_{F \in \mathbb{R}^{n \times k}} \sum_{k=1}^K \sum_{i,j=1}^n W_{ij} \left\| d_i^\alpha F_{ik} - d_j^\alpha F_{jk} \right\|^2 \\ \text{s.t. } F_{ik} &= \begin{cases} 1, & \text{if } x_i \in \mathcal{C}_k \\ 0, & \text{otherwise} \end{cases}, \; 1 \leq i \leq n_{[l]}, \; 1 \leq k \leq K \end{aligned} \tag{1}$$

where $\alpha \in \mathbb{R}$ is a given parameter. The interest of this generic formulation is that it coincides with the standard Laplacian-based approach for $\alpha = 0$ and with the normalized Laplacian-based approach for $\alpha = -\frac{1}{2}$, both discussed in Section 1. Note importantly that Equation (1) is naturally motivated by the observation that large values of W_{ij} enforce close values for F_{ik} and F_{jk} while small values for W_{ij} allow for more freedom in the choice of F_{ik} and F_{jk}.

By denoting

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

with $F_{[l]} \in \mathbb{R}^{n_{[l]}}$, $W_{[ll]} \in \mathbb{R}^{n_{[l]} \times n_{[l]}}$, $D_{[l]} \in \mathbb{R}^{n_{[l]} \times n_{[l]}}$, one easily finds (since the problem is a convex quadratic optimization with linear equality constraints) the solution to (1) is explicitly given by

$$F_{[u]} = \left(I_{n_u} - D_{[u]}^{-1-\alpha} W_{[uu]} D_{[u]}^{\alpha} \right)^{-1} D_{[u]}^{-1-\alpha} W_{[ul]} D_{[l]}^{\alpha} F_{[l]}. \tag{2}$$

Once these scores are affected, a mere comparison between all scores $F_{i1}, \ldots, F_{iK}$ for unlabelled data x_i (i.e., for $i > n_{[l]}$) is performed to decide on its class, i.e., the allocated class index $\hat{\mathcal{C}}_{x_i}$ for vector x_i is given by

$$\hat{\mathcal{C}}_{x_i} = \mathcal{C}_{\hat{k}} \text{ for } \hat{k} = \operatorname{argmax}_{1 \leq k \leq K} F_{ik}.$$

Note in passing that the formulation (2) implies in particular that

$$F_{[u]} = D_{[u]}^{-1-\alpha} W_{[uu]} D_{[u]}^{\alpha} F_{[u]} + D_{[u]}^{-1-\alpha} W_{[ul]} D_{[l]}^{\alpha} F_{[l]} \tag{3}$$

$$F_{[l]} = \{ \boldsymbol{\delta}_{x_i \in \mathcal{C}_k} \}_{\substack{1 \leq i \leq n_{[l]} \\ 1 \leq k \leq K}} \tag{4}$$

and thus the matrix F is a stationary point for the algorithm constituted of the updating rules (3) and (4) (when replacing the equal signs by affectations). In particular, for $\alpha = -1$, the algorithm corresponds to the standard label propagation method found in the PageRank algorithm for semi-supervised learning as discussed in Avrachenkov et al. (2011), with the major difference that $F_{[l]}$ is systematically reset to its known value while in the study of Avrachenkov et al. (2011), $F_{[l]}$ is allowed to evolve (for reasons related to robustness to pre-labeling errors).

The technical objective of the article is to analyze the behavior of $F_{[u]}$ in the large n, p regime for a Gaussian mixture model for the data $x_1, \ldots, x_n$. To this end, we shall first need to design appropriate growth rate conditions for the Gaussian mixture statistics as $p \to \infty$ (in order to avoid trivializing the classification problem as p grows large) before proceeding to the evaluation of the behavior of W, D, and thus F.

3. Model and Theoretical Results

3.1. Model and Assumptions

In the remainder of the article, we shall assume that the data $x_1, \ldots, x_n$ are extracted from a Gaussian mixture model composed of K classes. Specifically, for $k \in \{1, \ldots, K\}$,

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

Consistently with the previous section, for each k, there are n_k instances of vectors of class $\mathcal{C}_k$, among which $n_{[l]k}$ are labelled and $n_{[u]k}$ are unlabelled.

As pointed out above, in the regime where $n, p \to \infty$, special care must be taken to ensure that the classes $\mathcal{C}_1, \ldots, \mathcal{C}_K$, the statistics of which evolve with p, remain at a "somewhat constant" distance from each other. This is to ensure that the classification problem does not become asymptotically infeasible nor trivially simple as $p \to \infty$. Based on the earlier work (Couillet and Benaych-Georges, 2015) where similar considerations were made, the behavior of the class means, covariances, and cardinalities will follow the prescription below:

Assumption 1 (Growth Rate) *As* $n \to \infty$, $\frac{p}{n} \to c_0 > 0$ *and* $\frac{n_{[l]}}{n} \to c_{[l]} > 0$, $\frac{n_{[u]}}{n} \to c_{[u]} > 0$. *For each* k, $\frac{n_k}{n} \to c_k > 0$, $\frac{n_{[l]k}}{n} \to c_{[l]k} > 0$, $\frac{n_{[u]k}}{n} \to c_{[u]k} > 0$. *Besides,*

1. For $\mu^\circ \triangleq \sum_{k=1}^K \frac{n_k}{n} \mu_k$ and $\mu_k^\circ \triangleq \mu_k - \mu^\circ$, $\|\mu_k^\circ\| = O(1)$.
2. For $C^\circ \triangleq \sum_{k=1}^K \frac{n_k}{n} C_k$ and $C_k^\circ \triangleq C_k - C^\circ$, $\|C_k\| = O(1)$ and $\mathrm{tr} C_k^\circ = O(\sqrt{p})$.
3. As $n \to \infty$, $\frac{2}{p} \mathrm{tr} C^\circ \to \tau \neq 0$.
4. As $n \to \infty$, $\alpha = O(1)$.

It will also be convenient in the following to define

$$t_k \equiv \frac{1}{\sqrt{p}} \operatorname{tr} C_k^\circ$$
$$T_{kk'} \equiv \frac{1}{p} \operatorname{tr} C_k C_{k'}$$

as well as the labelled-data centered notations

$$\tilde{\mu}_k \equiv \mu_k - \sum_{k'=1}^K \frac{n_{[l]k'}}{n_{[l]}} \mu_{k'}$$
$$\tilde{C}_k \equiv C_k - \sum_{k'=1}^K \frac{n_{[l]k'}}{n_{[l]}} C_{k'}$$
$$\tilde{t}_k \equiv \frac{1}{\sqrt{p}} \operatorname{tr} \tilde{C}_k$$
$$\tilde{T}_{kk'} \equiv \frac{1}{p} \operatorname{tr} \tilde{C}_k \tilde{C}_{k'}.$$

A few comments on Assumption 1 are in order. First note that, unlike in the previous works (Nadler et al., 2009; Globerson et al., 2017) where the number of labelled data $n_{[l]}$ and data dimension p are considered fixed and the number of unlabelled data $n_{[u]}$ is supposed to be infinite, we assume a regime where $n_{[l]}, n_{[u]}$ and p are simultaneously large. Letting p large allows us to investigate SSL in the context of large dimensional data. Further imposing that $n_{[l]}, n_{[u]}$ grow at a controlled rate with respect to p (here at the same rate) allows for an *exact characterization* of the limiting SSL performances, as a function of the hyperparameters α, f and data statistics μ_k, C_k, in non-trivial classification scenarios (i.e., when classification is neither asymptotically perfect nor impossible), instead of solely retrieving consistency bounds as a function of growth rates in $p, n_{[l]}, n_{[u]}$. This in turn allows for possible means of precise parameter setting to reach optimal performances (which is not possible with results based on bounds). While it may be claimed that SSL in practice often handles scenarios where $n_{[u]} \gg n_{[l]}$, assuming that $n_{[u]}, n_{[l]}$ are of the same order but that $n_{[u]}$ is multiple times $n_{[l]}$ actually maintains the validity of our results so long that $n_{[l]}$ is not too small. To be more exact, our results are still valid in the limit where $c_{[l]} \to 0$, but then become trivial, as numerically confirmed by Figure 5. To consider the setting where

$n_{[l]}$ is fixed while $p, n_{[u]}$ grow large would demand a change in the statistical assumptions of the input data sets, which goes beyond the scope of the present investigation.

Item 3. of Assumption 1 is mostly a technical convenience that shall simplify our analysis, but our results naturally extend as long as both liminf and limsup of $\frac{2}{p}\mathrm{tr}C^\circ$ are away from zero or infinity. The necessity of Item 1. only appears through a detailed analysis of spectral properties of the weight matrix W for large n, p, carried out later in the article. As for Item 2., note that if $\mathrm{tr}\, C_k^\circ = O(\sqrt{p})$ were to be relaxed, it is easily seen that a mere (unsupervised) comparison of the values of $\|x_i\|^2$ would asymptotically provide an almost surely perfect classification.

As a by-product of imposing the growth constraints on the data to ensure non-trivial classification, Assumption 1 induces the following seemingly unsettling implication, easily justified by a simple concentration of measure argument

$$\max_{1\leq i,j\leq n} \left| \frac{1}{p}\|x_i - x_j\|^2 - \tau \right| \xrightarrow{\text{a.s.}} 0 \tag{5}$$

as $p \to \infty$. Equation (5) is the cornerstone of our analysis and states that all vector pairs x_i, x_j are essentially at the same distance from one another as p gets large, *irrespective of their classes*. This striking result evidently is in sharp opposition to the very motivation for the optimization formulation (1) as discussed in the introduction. It thus immediately entails that the solution (2) to (1) is bound to produce asymptotically inconsistent results. We shall see that this is indeed the case for all but a short range of values of α.

This being said, Equation (5) has an advantageous side as it allows for a Taylor expansion of $W_{ij} = f(\frac{1}{p}\|x_i - x_j\|^2)$ around $f(\tau)$, provided f is sufficiently smooth around τ, which is ensured by our subsequent assumption.

Assumption 2 (Kernel function) *The function $f : \mathbb{R}^+ \to \mathbb{R}$ is three-times continuously differentiable in a neighborhood of τ.*

Note that Assumption 2 does not constrain f aside from its local behavior around τ. In particular, we shall not restrict ourselves to matrices W arising from nonnegative definite kernels as standard machine learning theory would advise (Schölkopf and Smola, 2002).

The core technical part of the article now consists in expanding W, and subsequently all terms intervening in (2), in a Taylor expansion of successive matrices of *non-vanishing operator norm*. Note indeed that the magnitude of the individual entries in the Taylor expansion of W needs not follow the magnitude of the operator norm of the resulting matrices;[1] rather, great care must be taken to only retain those matrices of non-vanishing operator norm. These technical details call for advanced random matrix considerations and are discussed in the appendix and in Couillet and Benaych-Georges (2015).

We are now in position to introduce our main technical results.

3.2. Main Theoretical Results

In the course of this section, we provide in parallel a series of technical results under the proposed setting (notably under Assumption 1) along with simulation results both on a

1. For instance, $\|I_n\| = 1$ while $\|1_n 1_n^\mathsf{T}\| = n$ despite both matrices having entries of similar magnitude.

2-class Gaussian mixture data model with $\mu_1 = [4; 0_{p-1}]$, $\mu_2 = [0; 4; 0_{p-2}]$, $C_1 = I_p$ and $\{C_2\}_{i,j} = .4^{|i-j|}(1 + \frac{3}{\sqrt{p}})$, as well as on real data sets, here images of eights and nines from the MNIST database (LeCun et al., 1998), for $f(t) = \exp(-\frac{1}{2}t)$, i.e., the classical Gaussian (or heat) kernel. For reasons that shall become clear in the following discussion, these figures will depict the (size n) vectors

$$\left[F^\circ_{[u]}\right]_{\cdot k} \equiv \left[F_{[u]}\right]_{\cdot k} - \frac{1}{K}\sum_{k'=1}^{K} \left[F_{[u]}\right]_{\cdot k'}$$

for $k \in \{1, 2\}$. Obviously, the decision rule on $F^\circ_{[u]}$ is the same as that on $F_{[u]}$.

Our first hinging result concerns the behavior of the score matrix F in the large n, p regime, as per Assumption 1, and reads as follows.

Proposition 1 *Let Assumptions 1–2 hold. Then, for $i > n_{[l]}$ (i.e., for x_i an unlabelled vector),*

$$F_{ik} = \frac{n_{[l]k}}{n}\left[1 + \underbrace{(1+\alpha)\frac{f'(\tau)}{f(\tau)}\frac{t_k}{\sqrt{p}} + z_i}_{O(n^{-\frac{1}{2}})} + O(n^{-1})\right] \tag{6}$$

where $z_i = O(n^{-\frac{1}{2}})$ is a random variable, function of x_i, but independent of k.

The proof of Proposition 1 is given as an intermediary result of the proof of Theorem 5 in the appendix.

Proposition 1 provides a clear overview of the outcome of the semi-supervised learning algorithm. First note that $F_{ik} = c_{[l]k} + O(n^{-\frac{1}{2}})$. Therefore, irrespective of x_i, F_{ik} is strongly biased towards $c_{[l]k}$. If the values $n_{[l]1}, \ldots, n_{[l]k}$ differ by $O(n)$, this induces a systematic asymptotic allocation of every x_i to the class having largest $c_{[l]k}$ value. Figure 1 illustrates this phenomenon, observed both on synthetic and real data sets, here for $n_{[l]1} = 3n_{[l]2}$.

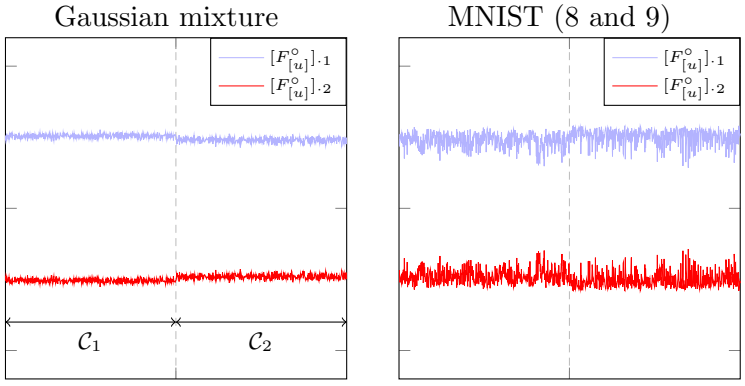

Figure 1: $[F^\circ_{[u]}]_{\cdot 1}$ and $[F^\circ_{[u]}]_{\cdot 2}$ for 2-class data, $n = 1024$, $p = 784$, $n_l/n = 1/16$, $n_{[u]1} = n_{[u]2}$, $n_{[l]1} = 3n_{[l]2}$, $\alpha = -1$, Gaussian kernel.

Pursuing the analysis of Proposition 1 by now assuming that $n_{[l]1} = \ldots = n_{[l]K}$, the comparison between $F_{i1}, \ldots, F_{iK}$ next revolves around the term of order $O(n^{-\frac{1}{2}})$. Since z_i only depends on x_i and not on k, it induces a constant offset to the vector $F_{i\cdot}$, thereby not intervening in the class allocation. On the opposite, the term t_k is independent of x_i but may vary with k, thereby possibly intervening in the class allocation, again an undesired effect. Figure 2 depicts the effect of various choices of α for equal values of $n_{[l]k}$. This deleterious outcome can be avoided either by letting $f'(\tau) = O(n^{-\frac{1}{2}})$ or $\alpha = -1 + O(n^{-\frac{1}{2}})$. But, as discussed in the study of Couillet and Benaych-Georges (2015) and later in the article, the choice of f such that $f'(\tau) \simeq 0$, if sometimes of interest, is generally inappropriate.

The discussion above thus induces two important consequences to adapt the semi-supervised learning algorithm to large data.

1. The final comparison step *must* be made upon the normalized scores

$$\hat{F}_{ik} \equiv \frac{n}{n_{[l]k}} F_{ik} \tag{7}$$

   rather than upon the scores F_{ik} directly.

2. The parameter α *must* be chosen in such a way that

$$\alpha = -1 + O(n^{-\frac{1}{2}}).$$

Under these two amendments of the algorithm, according to Proposition 1, the performance of the semi-supervised learning algorithm now relies upon terms of magnitude $O(n^{-1})$, which are so far left undefined. A thorough analysis of these terms allows for a complete understanding of the asymptotic behavior of the normalized scores $\hat{F}_{i\cdot} = (\hat{F}_{i1}, \ldots, \hat{F}_{iK})$, as presented in our next result.

Theorem 2 *Let Assumptions 1–2 hold. For $i > n_{[l]}$ (i.e., x_i unlabelled) with $x_i \in \mathcal{C}_b$, let $\hat{F}_{ia}$ be given by (7) with F defined in (2) and $\alpha = -1 + \frac{\beta}{\sqrt{p}}$ for $\beta = O(1)$. Then,*

$$p\hat{F}_{i\cdot} = p(1 + z_i)1_K + G_i + o_P(1) \tag{8}$$

where $z_i = O(\sqrt{p})$ is as in Proposition 1 and $G_i \sim \mathcal{N}(m_b, \Sigma_b)$, $i > n_{[l]}$, are independent with

$$[m_b]_a = -\frac{2f'(\tau)}{f(\tau)} \tilde{\mu}_a^\mathsf{T} \tilde{\mu}_b + \left(\frac{f''(\tau)}{f(\tau)} - \frac{f'(\tau)^2}{f(\tau)^2} \right) \tilde{t}_a \tilde{t}_b + \frac{2f''(\tau)}{f(\tau)} \tilde{T}_{ab} + \frac{\beta}{c_{[l]}} \frac{f'(\tau)}{f(\tau)} t_a \tag{9}$$

$$[\Sigma_b]_{a_1 a_2} = 2 \left(\frac{f''(\tau)}{f(\tau)} - \frac{f'(\tau)^2}{f(\tau)^2} \right)^2 T_{bb} t_{a_1} t_{a_2} + 4 \frac{f'(\tau)^2}{f(\tau)^2} \left[\mu_{a_1}^{\circ\mathsf{T}} C_b \mu_{a_2}^\circ + \boldsymbol{\delta}_{a_1 a_2} \frac{c_0 T_{b,a_1}}{c_{[l]} c_{[l]a_1}} \right]. \tag{10}$$

Besides, there exists $\mathcal{A} \subset \sigma(\{\{x_1, \ldots, x_{n_{[l]}}\}, p = 1, 2, \ldots\})$ (the σ-field induced by the labelled variables) with $P(\mathcal{A}) = 1$ over which (8) also holds conditionally to $\{\{x_1, \ldots, x_{n_{[l]}}\}, p = 1, 2, \ldots\}$.

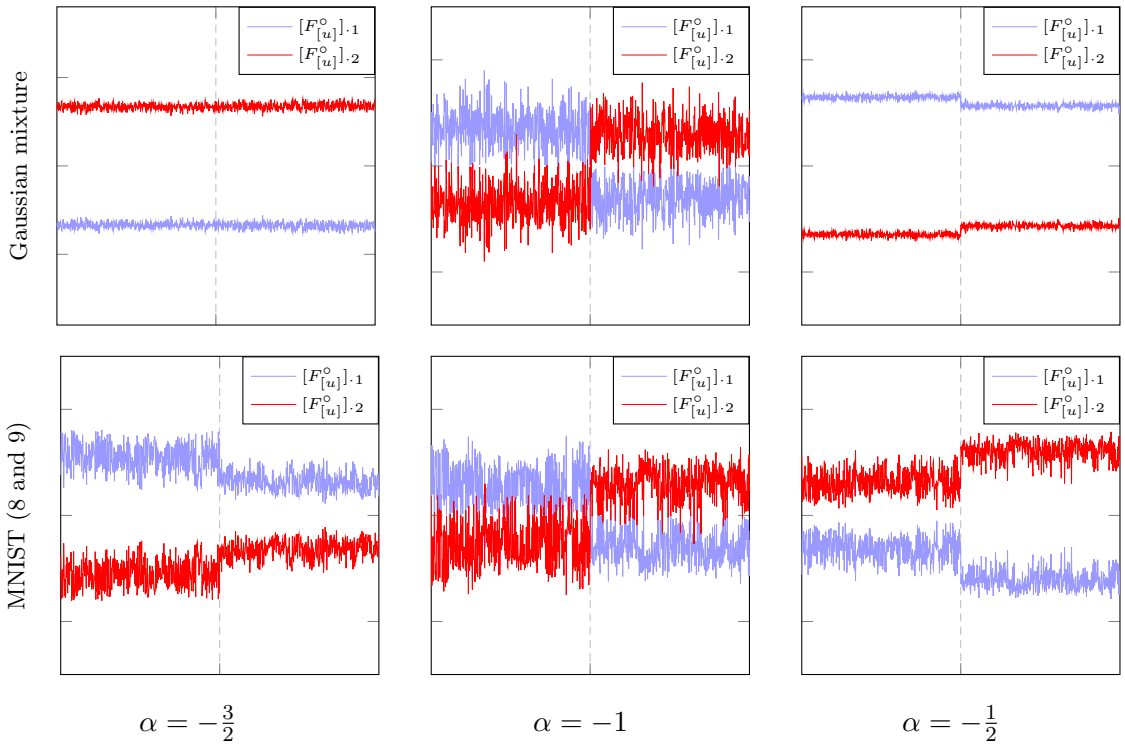

Figure 2: $[F^\circ_{[u]}]_{\cdot 1}$, $[F^\circ_{[u]}]_{\cdot 2}$ for 2-class data, $n = 1024$, $p = 784$, $n_l/n = 1/16$, $n_{[u]1} = n_{[u]2}$, $n_{[l]1} = n_{[l]2}$, Gaussian kernel.

Note that the statistics of G_i are independent of the realization of $x_1, \ldots, x_{[l]}$ when $\alpha = -1 + O(\frac{1}{\sqrt{p}})$. This in fact no longer holds when α is outside this regime, as pointed out by Theorem 5 in the appendix which provides the asymptotic behavior of $\hat{F}_{i\cdot}$ for all values of α (and thus generalizes Theorem 2).

Since the ordering of the entries of $\hat{F}_{i\cdot}$ is the same as that of $\hat{F}_{i\cdot} - (1 + z_i)$, Theorem 2 amounts to saying that the probability of correctly classifying unlabeled vectors x_i genuinely belonging to class $\mathcal{C}_b$ is asymptotically given by the probability of $[G_i]_b$ being the maximal element of G_i, which, as mentioned above, is the same whether conditioned or not on $x_1, \ldots, x_{[l]}$ for $\alpha = -1 + O(\frac{1}{\sqrt{p}})$. This is formulated in the following corollary.

Corollary 3 *Let Assumptions 1–2 hold. Let $i > n_{[l]}$ and $\alpha = -1 + \frac{\beta}{\sqrt{p}}$. Then, under the notations of Theorem 2,*

$$\mathbb{P}\left(x_i \to \mathcal{C}_b | x_i \in \mathcal{C}_b, x_1, \cdots, x_{n_{[l]}}\right) - \mathbb{P}\left(x_i \to \mathcal{C}_b | x_i \in \mathcal{C}_b\right) \to 0$$

$$\mathbb{P}\left(x_i \to \mathcal{C}_b | x_i \in \mathcal{C}_b\right) - \mathbb{P}\left([G_i]_b > \max_{a \neq b}\{[G_i]_a\} | x_i \in \mathcal{C}_b\right) \to 0.$$

In particular, for $K = 2$, and $a \neq b \in \{1, 2\}$,

$$\mathbb{P}\left([G_i]_b > \max_{a \neq b} \{[G_i]_a\} | x_i \in \mathcal{C}_b \right) = \Phi(\theta_b^a), \quad \textit{with } \theta_b^a \equiv \frac{[m_b]_b - [m_b]_a}{\sqrt{[\Sigma_b]_{bb} + [\Sigma_b]_{aa} - 2[\Sigma_b]_{ab}}}$$

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

With G_i being independent, Corollary 3 allows us to approach the empirical classification accuracy as it is consistently estimated by the probability of correct classification given in the corollary. As with Theorem 2 which can be appended to Theorem 5 for a large set of values of α, Corollary 3 is similarly generalized by Corollary 6 in the appendix. Using both corollaries, Figure 3 displays a comparison between simulated accuracies from various pairs of digits from the MNIST data against our theoretical results; to apply our results, a 2-class Gaussian mixture model is assumed with means and covariances equal to the empirical means and covariances of the individual digits, evaluated from the full 60 000-image MNIST database. It is quite interesting to observe that, despite the obvious inadequacy of a Gaussian mixture model for this image database, the theoretical predictions are in strong agreement with the practical performances. Also surprising is the strong adequacy of the theoretical prediction of Corollary 3 beyond the range of values of α in the neighborhood of -1.

4. Consequences

4.1. Semi-Supervised Learning beyond Two Classes

An immediate consequence of Corollary 3 is that, for $K > 2$, there exists a Gaussian mixture model for which the semi-supervised learning algorithms under study necessarily fail to classify at least one class. To see this, we consider $K = 3$ and let $\mu_3 = 3\mu_2 = 6\mu_1$, $C_1 = C_2 = C_3$, $n_1 = n_2 = n_3$, $n_{[l]1} = n_{[l]2} = n_{[l]3}$. First, it follows from Corollary 3 that,

$$\begin{aligned}
\mathbb{P}\left(x_i \to \mathcal{C}_2 | x_i \in \mathcal{C}_2\right) &\leq \mathbb{P}\left([G_i]_2 > [G_i]_1 | x_i \in \mathcal{C}_2\right) + o(1) = \Phi(\theta_2^1) + o(1) \\
\mathbb{P}\left(x_i \to \mathcal{C}_3 | x_i \in \mathcal{C}_3\right) &\leq \mathbb{P}\left([G_i]_3 > [G_i]_1 | x_i \in \mathcal{C}_3\right) + o(1) = \Phi(\theta_3^1) + o(1)
\end{aligned}$$

Then, under Assumptions 1–2 and the notations of Corollary 3,

$$\begin{aligned}
\theta_2^1 &= \operatorname{sgn}(f'(\tau)) \frac{\mu_1^2}{\sqrt{(\Sigma_2)_{22} + (\Sigma_2)_{11} - 2(\Sigma_2)_{12}}} \\
\theta_3^1 &= -\operatorname{sgn}(f'(\tau)) \frac{15\mu_1^2}{\sqrt{(\Sigma_3)_{33} + (\Sigma_3)_{11} - 2(\Sigma_3)_{13}}}
\end{aligned}$$

so that $f'(\tau) < 0 \Rightarrow \theta_2^1 < 0$, $f'(\tau) > 0 \Rightarrow \theta_3^1 < 0$, while $f'(\tau) = 0 \Rightarrow \theta_2^1 = \theta_3^1 = 0$. As such, the correct classification rate of elements of $\mathcal{C}_2$ and $\mathcal{C}_3$ cannot be simultaneously greater than $\frac{1}{2}$, leading to necessarily inconsistent classifications.

It is nonetheless easy to check that this kind of inconsistency cannot occur if μ_1, μ_2 and μ_3 are mutually orthogonal (which is often bound to occur with large dimensional data). Indeed, note that all first three terms at the right-hand side of (9) can be viewed as products of

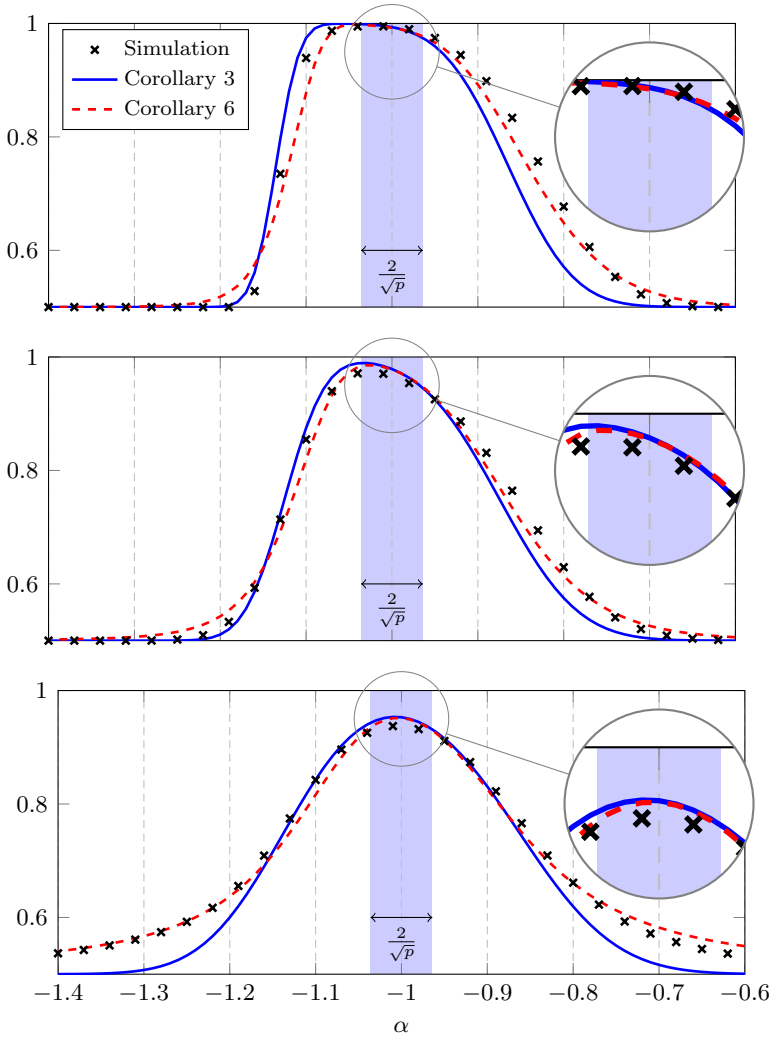

Figure 3: Theoretical and empirical accuracy as a function of α for 2-class MNIST data (top: digits (0,1), middle: digits (1,7), bottom: digits (8,9)), $n = 1024$, $p = 784$, $n_{[l]}/n = 1/16$, $n_{[u]1} = n_{[u]2}$, Gaussian kernel. Averaged over 50 iterations.

some centered vectors $\tilde{v}_k = v_k - \sum_{k'=1}^K \gamma_{k'} v_{k'}$ where $\sum_{k'=1}^K \gamma_{k'} = 1$.[2] Inconsistency occurs to class k if there exist $a, b \neq k$ such that $\tilde{v}_k^T \tilde{v}_b > \tilde{v}_k^T \tilde{v}_k > \tilde{v}_k^T \tilde{v}_a$. To better understand the cause of this inconsistency, let us consider two extreme scenarios: (i) the v_k differ by 'intensity', i.e., $v_k = r_k v$ for $k \in \{1, \cdots, K\}$, or (ii) the v_k differ by 'direction', i.e, $v_k = v + u_k$ with orthogonal u_k's. In scenario (i), let $s_{\min} = \operatorname{argmin}_{k \in \{1,\cdots,K\}} r_k$ and $s_{\max} = \operatorname{argmax}_{k \in \{1,\cdots,K\}} r_k$; then, for $k \neq \{s_{\min}, s_{\max}\}$, $\min\{\tilde{v}_k^T \tilde{v}_{s_{\min}}, \tilde{v}_k^T \tilde{v}_{s_{\max}}\} < \tilde{v}_k^T \tilde{v}_k < \max\{\tilde{v}_k^T \tilde{v}_{s_{\min}}, \tilde{v}_k^T \tilde{v}_{s_{\max}}\}$ and inconsistency is thus observed for classes $k \neq \{s_{\min}, s_{\max}\}$. Contrarily, in scenario (ii), for all

2. The third term of (9) can be seen in this way since for any two symmetric matrices $A = \{a_{ij}\}_{i,j=1}^m$ and $B = \{b_{ij}\}_{i,j=1}^m$ of same dimensions, $\operatorname{tr} AB = \sum_{i,j}^m a_{ij} b_{ij} = a_v^T b_v$ with $a_v = [a_{11}, \cdots, a_{1m}, \cdots, a_{m1}, \cdots, a_{mm}]$, $b_v = [b_{11}, \cdots, b_{1m}, \cdots, b_{m1}, \cdots, b_{mm}]$.

$k \neq k' \in \{1, \cdots, K\}$, $\tilde{v}_k^T \tilde{v}_k \geq \tilde{v}_k^T \tilde{v}_{k'}$ since $\tilde{v}_k^T \tilde{v}_k \geq 0$ and $\tilde{v}_k^T \tilde{v}_{k'} \leq 0$. As such, inconsistency is less likely to occur if the v_k's have very different directions.

4.2. Choice of f and Suboptimality of the Heat Kernel

As a consequence of the previous section, we shall from here on concentrate on the semi-supervised classification of $K = 2$ classes. In this case, it is easily seen that,

$$(K = 2) \quad \forall a \neq b \in \{1, 2\}, \ \|\tilde{\mu}_b\|^2 \geq \tilde{\mu}_b^{\mathsf{T}} \tilde{\mu}_a, \quad \tilde{t}_b^2 \geq \tilde{t}_a \tilde{t}_b, \quad \tilde{T}_{bb} \geq \tilde{T}_{ab}$$

with equalities respectively for $\mu_a = \mu_b$, $t_a = t_b$, and $\operatorname{tr} C_a C_b = \operatorname{tr} C_b^2$. This result, along with Corollary 3, implies the necessity of the conditions

$$f'(\tau) < 0, \quad f''(\tau) f(\tau) > f'(\tau)^2, \quad f''(\tau) > 0$$

to fully discriminate Gaussian mixtures. As such, from Corollary 3, by letting $\alpha = -1$, semi-supervised classification of $K = 2$ classes is always consistent under these conditions.

Since only the first three derivatives of f are involved, one may design a simple kernel for any desired values of $f'(\tau)$, $f''(\tau)f(\tau) - f'(\tau)^2$ and $f''(\tau)$ with a second degree polynomial $f(t) = ax^2 + bx + c$ in such a way that $a\tau + b = f'(\tau)$, $a(a\tau^2 + b\tau + c) - (a\tau + b)^2 = f''(\tau)f(\tau) - f'(\tau)^2$ and $a = f''(\tau)$, i.e.,

$$a = f''(\tau) \quad b = f'(\tau) - f''(\tau)\tau \quad c = (f''(\tau)f(\tau) - f'(\tau)\tau^2)/f''(\tau).$$

Since τ can be consistently estimated in practice by (11) (see the discussion in Subsection 5.1), so can a, b, and c.

A quite surprising outcome of the necessary conditions on the derivatives of f is that the widely used Gaussian (or heat) kernel $f(t) = \exp(-\frac{t}{2\sigma^2})$, while fulfilling the condition $f'(t) < 0$ and $f''(t) > 0$ for all t (and thus $f'(\tau) < 0$ and $f''(\tau) > 0$), only satisfies $f''(t)f(t) = f'(t)^2$. This indicates that discrimination over $t_1, \ldots, t_K$, under the conditions of Assumption 1, is asymptotically *not* possible with a Gaussian kernel. This remark is illustrated in Figure 4 for a discriminative task between two centered isotropic Gaussian classes only differing by the trace of their covariance matrices. There, irrespective of the choice of the bandwidth σ, the Gaussian kernel leads to a constant $1/2$ accuracy, where a mere second order polynomial kernel selected upon its derivatives at τ demonstrates good performances. Since p-dimensional isotropic Gaussian vectors tend to concentrate "close to" the surface of a sphere, this thus suggests that Gaussian kernels are not inappropriate to solve the large dimensional generalization of the "concentric spheres" task (for which they are very efficient in small dimensions). In passing, the right-hand side of Figure 4 confirms the need for $f''(\tau)f(\tau) - f'(\tau)^2$ to be positive (there $|f'(\tau)| < 1$) as an accuracy lower than $1/2$ is obtained for $f''(\tau)f(\tau) - f'(\tau)^2 < 0$.

Another interesting fact lies in the choice $f'(\tau) = 0$ (while $f''(\tau) \neq 0$). As already identified by Couillet and Benaych-Georges (2015) and further thoroughly investigated by Couillet and Kammoun (2016), if $t_1 = t_2$ (which can be enforced by normalizing the data set) and $\tilde{T}_{bb} > \tilde{T}_{ba}$ for all $a \neq b \in \{1, 2\}$, then $\Sigma_b = 0$ while $[m_b]_b > [m_b]_a$ for all $b \neq a \in \{1, 2\}$ and thus leading asymptotically to a perfect classification. As such, while Assumption 1 was

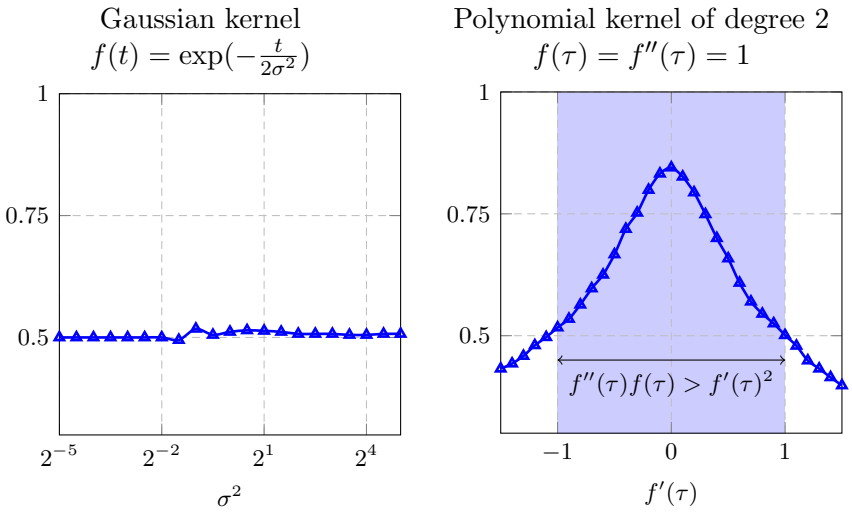

Figure 4: Empirical accuracy for 2-class Gaussian data with $\mu_1 = \mu_2$, $C_1 = I_p$ and $C_2 = (1 + \frac{3}{\sqrt{p}})I_p$, $n = 1024$, $p = 784$, $n_l/n = 1/16$, $n_{[u]1} = n_{[u]2}$, $n_{[l]1} = n_{[l]2}$, $\alpha = -1$.

claimed to ensure a "non-trivial" growth rate regime, asymptotically perfect classification may be achieved by choosing f such that $f'(\tau) = 0$, under the aforementioned statistical conditions. One must nonetheless be careful that this *asymptotic result* does not necessarily entail outstanding performances in practical finite dimensional scenarios. Indeed, note that taking $f'(\tau) = 0$ discards the visibility of differing means $\mu_1 \neq \mu_2$ (from the expression of $[m_b]_a$ in Theorem 2); for finite n, p, cancelling the differences in means (often larger than differences in covariances) may not be compensated for by the reduction in variance. Trials on MNIST particularly emphasize this remark.

4.3. Impact of Class Sizes

A final remark concerns the impact of $c_{[l]}$ and c_0 on the asymptotic performances. Note that $c_{[l]}$ and c_0 only act upon the covariance Σ_b and precisely on its diagonal elements. Both a reduction in c_0 (by increasing n) and an increase in $c_{[l]}$ reduce the diagonal terms in the variance, thereby mechanically increasing the classification performances (if in addition $[m_b]_b > [m_b]_a$ for $a \neq b$). In the opposite case of few labeled data, i.e., $c_{[l]} \to 0$, the variance diverges and the performance tends to that of random classification, as shown in Figure 5.

5. Parameter Optimization in Practice

5.1. Estimation of τ

In previous sections, we have emphasized the importance of selecting the kernel function f so as to meet specific conditions on its derivatives at the quantity τ. In practice however, τ is an unknown quantity. A mere concentration of measure argument nonetheless shows

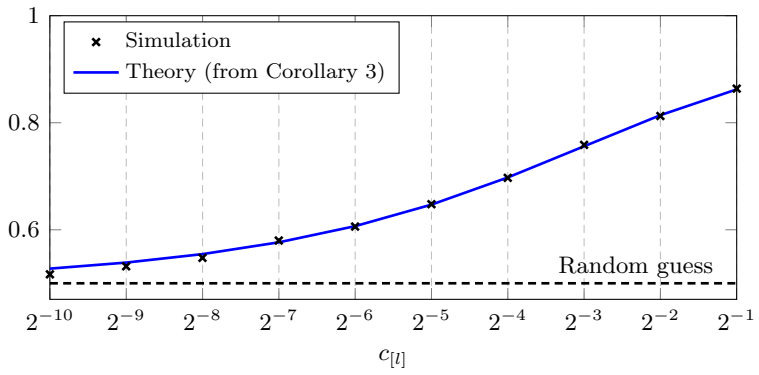

Figure 5: Theoretical and empirical accuracy as a function of $c_{[l]}$ for 2-class Gaussian data with $\mu_2 = 2\mu_1 = [6, 0, \cdots, 0]$, $C_1 = I_p$ and $\{C_2\}_{i,j} = .4^{|i-j|}$, $p = 2048$, $c_0 = 1$, $n_{[u]1} = n_{[u]2}$, $n_{[l]1} = n_{[l]2}$, $\alpha = -1$, Gaussian kernel. Averaged over 50 iterations.

that

$$\hat{\tau} \equiv \frac{1}{n(n-1)} \sum_{i,j=1}^{n} \frac{1}{p} \|x_i - x_j\|^2 \xrightarrow{\text{a.s.}} \tau. \tag{11}$$

As a consequence, the results of Theorem 2 and subsequently of Corollary 3 hold verbatim with $\hat{\tau}$ in place of τ. One thus only needs to design f in such as way that its derivatives at $\hat{\tau}$ meet the appropriate conditions.

5.2. Optimization of α

In Section 4.2, we have shown that the choice $\alpha = -1$, along with an appropriate choice of f, ensures the asymptotic consistency of semi-supervised learning for $K = 2$ classes, in the sense that non-trivial asymptotic accuracy (> 0.5) can be achieved. This choice of α may however not be optimal in general. This subsection is devoted to the optimization of α so as to maximize the average precision, a criterion often used in absence of prior information to favor one class over the other. While not fully able to estimate the optimal $\alpha^\star$ of α, we shall discuss here a heuristic means to select a close-to-optimal α, subsequently denoted α_0.

As per Theorem 2, α must be chosen as $\alpha = -1 + \frac{\beta}{\sqrt{p}}$ for some $\beta = O(1)$. In order to set $\beta = \beta^\star$ in such a way that the classification accuracy is maximized, Corollary 3 further suggests the need to estimate the θ_b^a terms which in turn requires the evaluation of a certain number of quantities appearing in the expressions of m_b and Σ_b. Most of these are however not directly accessible from simple statistics of the data. Instead, we shall propose here a heuristic and simple method to retrieve a reasonable choice β_0 for β, which we claim is often close to optimal and sufficient for most needs.

To this end, first observe from (9) that the mappings $\beta \mapsto [m_b]_a$ satisfy

$$\frac{d}{d\beta}([m_b]_b - [m_b]_a) = \frac{f'(\tau)}{f(\tau)c_{[l]}}(t_b - t_a) = -\frac{d}{d\beta}([m_a]_a - [m_a]_b).$$

Hence, changes in β induce a simultaneous reduction and increase of either one of $[m_b]_b - [m_b]_a$ and $[m_a]_a - [m_a]_b$. Placing ourselves again in the case $K = 2$, we define β_0 to be the value for which both differences (with $a \neq b \in \{1, 2\}$) are the same, leading to the following Proposition–Definition.

Proposition 4 *Let $K = 2$ and $[m_b]_a$ be given by (9). Then*

$$\beta_0 \equiv \frac{f(\tau)}{f''(\tau)} \frac{c_{[l]1} - c_{[l]2}}{t_1 - t_2} \Delta m \tag{12}$$

where

$$\Delta m = -\frac{2f'(\tau)}{f(\tau)} \|\mu_1 - \mu_2\|^2 + \left(\frac{f''(\tau)}{f(\tau)} - \frac{f'(\tau)^2}{f(\tau)^2} \right) (t_1 - t_2)^2 + \frac{2f''(\tau)}{f(\tau)} (T_{11} + T_{22} - 2T_{12})$$

is such that, for $\alpha = -1 + \frac{\beta_0}{\sqrt{p}}$, $[m_1]_1 - [m_1]_2 = [m_2]_2 - [m_2]_1$.

By choosing $\alpha = \alpha_0 \equiv -1 + \frac{\beta_0}{\sqrt{p}}$, one ensures that $\mathbb{E}_{x_i \in \mathcal{C}_1}[\hat{F}_{i1} - \hat{F}_{i2}] = -\mathbb{E}_{x_i \in \mathcal{C}_2}[\hat{F}_{i1} - \hat{F}_{i2}] + o(1)$ $(i > n_{[l]})$, thereby evenly balancing the *average* "resolution" of each class. An even balance typically produces the desirable output of the central displays of Figure 2 (as opposed to the largely undesirable bottom of top displays, there for very offset values of α). Obviously though, since the variances of $\hat{F}_{i1} - \hat{F}_{i2}$ for $x_i \in \mathcal{C}_1$ or $x_i \in \mathcal{C}_2$ are in general not the same, this choice of α may not be optimal. Nonetheless, in most experimental scenarios of practical interest, the score variances tend to be sufficiently similar for the choice of α_0 to be quite appealing.

This heuristic motivation made, note that β_0 is proportional to $c_{[l]b} - c_{[l]a}$. This indicates that the more unbalanced is the labelled data set, the more deviated from zero is β_0. In particular, for $n_{[l]1} = n_{[l]2}$, $\alpha_0 = -1$. As we shall subsequently observe in simulations, this remark is of dramatic importance in practice where taking $\alpha = -1$ (the PageRank method) in place of $\alpha = \alpha_0$ leads to significant performance losses.

Of utmost importance here is the fact that, unlike θ_b^a which are difficult to assess empirically, a consistent estimate of β_0 can be obtained through a rather simple method, which we presently elaborate on.

While an estimate for t_a and T_{ab} can be obtained empirically from the labelled data themselves, $\|\mu_1 - \mu_2\|^2$ is not directly accessible (note indeed that $\frac{1}{n_{[l]a}} \sum_{\mathcal{C}_a} x_i = \mu_a + \frac{1}{n_{[l]a}} \sum_{\mathcal{C}_a} w_i$, for some $w_i \sim \mathcal{N}(0, C_a)$, and the central limit theorem guarantees that $\| \frac{1}{n_{[l]a}} \sum_{\mathcal{C}_a} w_i \| = O(1)$, the same order of magnitude as $\|\mu_a - \mu_b\|$). However, one may access an estimate for Δm by running two instances of the PageRank algorithm ($\alpha = -1$), resulting in the method described in Algorithm 1. It is easily shown that, under Assumptions 1–2,

$$\hat{\beta}_0 - \beta_0 \xrightarrow{\text{a.s.}} 0.$$

Figure 6 provides a performance comparison, in terms of average precision, between the PageRank ($\alpha = -1$) method and the proposed heuristic improvement for $\alpha = \alpha_0$, versus the oracle estimator for which $\alpha = \alpha^\star$, the precision-maximizing value. The curves are here

Algorithm 1 Estimate $\hat{\beta}_0$ of β_0.

1: Let $\hat{\tau}$ be given by (11).
2: Let

$$\widehat{\Delta t} = \frac{1}{2\sqrt{p}} \left(\frac{\sum_{i,j=1}^{n_{[l]1}} \|x_i - x_j\|^2}{n_{[l]1}(n_{[l]1} - 1)} - \frac{\sum_{i,j=n_{[l]1}+1}^{n_{[l]}} \|x_i - x_j\|^2}{n_{[l]2}(n_{[l]2} - 1)} \right)$$

3: Set $\alpha = -1$ and define $J \equiv p \sum_{i=n_{[l]}+1}^{n} \hat{F}_{i1} - \hat{F}_{i2}$.
4: Still for $\alpha = -1$, reduce the set of labelled data to $n'_{[l]1} = n'_{[l]2} = \min\{n_{[l]1}, n_{[l]2}\}$ and, with obvious notations, let $J' \equiv p \sum_{i=n'_{[l]}+1}^{n} \hat{F}'_{i1} - \hat{F}'_{i2}$.
5: Return $\hat{\beta}_0 \equiv \frac{c_{[l]} f(\hat{\tau})}{f'(\hat{\tau}) \widehat{\Delta t}} \frac{J' - J}{n_{[u]}}$.

snapshots of typical classification precision obtained from examples of $n = 4096$ images with $c_{[l]} = 1/16$. As expected, the gain in performance is largest as $|c_{[l]1} - c_{[l]2}|$ is large. More surprisingly, the performances obtained are impressively close to optimal. It should be noted though that simulations revealed more unstable estimates of $\hat{\beta}_0$ for smaller values of n.

Note that the method for estimating β_0 provided in Algorithm 1 implicitly exploits the resolution of two equations (through the observation of J, J' obtained for different values of $n_{[l]1}, n_{[l]2}$) to retrieve the value of Δm defined in Proposition 4. Having access to Δm further allows access to $\|\mu_1 - \mu_2\|^2$, for instance by setting f so that $f''(\tau) = 0$ and $f''(\tau) f(\tau) = f'(\tau)^2$. This in turn allows access to all terms intervening in $[m_b]_a$ (as per (9)), making it possible to choose f so to maximize the distances $|[m_1]_1 - [m_1]_2|$ and $|[m_2]_2 - [m_2]_1|$. However, in addition to the cumbersome aspect of the induced procedure (and the instability implied by multiple evaluations of the scores F under several settings for f and $c_{[l]a}$), such operations also alter the values of the variances in (10) for which not all terms are easily estimated. It thus seems more delicate to derive a simple method to optimize f in addition to α.

6. Concluding Remarks

This article is part of a series of works consisting in evaluating the performance of kernel-based machine learning methods in the large dimensional data regime (Couillet and Benaych-Georges, 2015; Liao and Couillet, 2017; Couillet and Kammoun, 2016). Relying on the derivations of Couillet and Benaych-Georges (2015) that provide a Taylor expansion of radial kernel matrices around the limiting common value τ of $\frac{1}{p}\|x_i - x_j\|^2$ for $i \neq j$ and $p \to \infty$, we observed that the choice of the kernel function f merely affects the classification performances through the successive derivatives of f at τ. In particular, similar to the earlier analyses (Couillet and Benaych-Georges, 2015; Liao and Couillet, 2017; Couillet and Kammoun, 2016), we found that the case $f'(\tau) = 0$ induces a sharp phase transition on normalized data by which the asymptotic classification error rate vanishes. However, unlike the works (Couillet and Benaych-Georges, 2015; Liao and Couillet, 2017), the exact

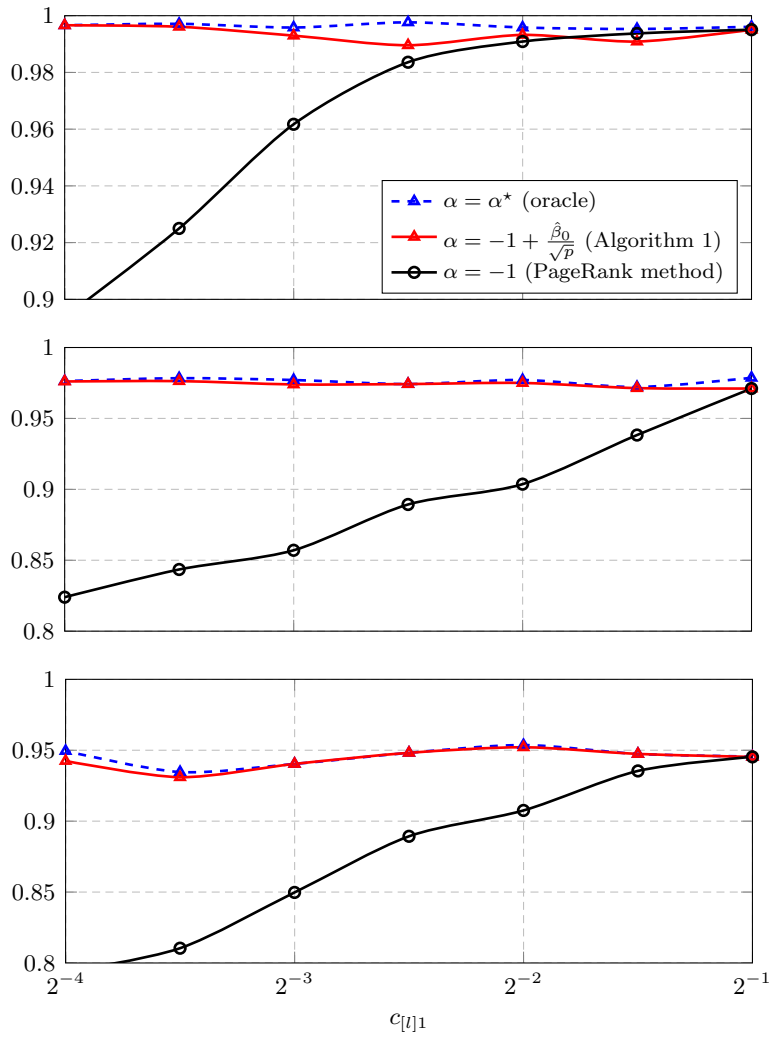

Figure 6: Average precision varying with $c_{[l]1}$ for 2-class MNIST data (**top:** digits (0,1), **middle:** digits (1,7), **bottom:** digits (8,9)), $n = 4096$, $p = 784$, $n_{[l]}/n = 1/16$, $n_{[u]1} = n_{[u]2}$, Gaussian kernel.

expression at the core of the limiting performance assumes a different form. Of importance is the finding that, under a heat kernel assumption $f(t) = \exp(-\frac{t}{2\sigma^2})$, the studied semi-supervised learning method fails to classify Gaussian mixtures of the type $\mathcal{N}(0, C_k)$ with $\operatorname{tr} C_k^\circ = O(\sqrt{p})$ and $\operatorname{tr} C_k C_{k'} - \operatorname{tr} C_k^2 = o(p)$, which unsupervised learning or LS-SVM are able to do (Couillet and Benaych-Georges, 2015; Liao and Couillet, 2017). This paradox may deserve a more structural way of considering together methods on the spectrum from unsupervised to supervised learning.

The very fact that the kernel matrix W is essentially equivalent to the matrix $f(\tau)1_n 1_n^\mathsf{T}$ (the $n \times n$ matrix filled with $f(\tau)$ values), thereby strongly disrupting with the expected

natural behavior of kernels, essentially follows from the Gaussian mixture model we assumed as well as from the decision to compare vectors by means of a mere Euclidean distance. We believe that this simplistic (although widely used) method explains the strong coincidence between performances on the Gaussian mixture model and on real data sets. Indeed, as radial functions are not specially adapted to image vectors (as would be wavelet or convolutional filters), the kernel likely operates on first order statistics of the input vectors, hence similar to its action on a Gaussian-mixture data. It would be interesting to generalize our result, and for that matter the set of works (Couillet and Benaych-Georges, 2015; Liao and Couillet, 2017; Couillet and Kammoun, 2016), to more involved data-oriented kernels, so long that the data contain enough exploitable degrees of freedom.

It is also quite instructive to note that, from the proof of our main results, the terms remaining after the expansion of $D_{[u]}^{-1-\alpha} W_{[uu]} D_{[u]}^{\alpha}$ appearing in (2) almost all vanish, strongly suggesting that similar results would be obtained if the inverse matrix in (2) were discarded altogether. This implies that the intra-unlabelled data kernel $W_{[uu]}$ is of virtually no asymptotic use. Also, the remark of Section 4.3 according to which $c_{[l]} \to 0$ implies a vanishing classification rate suggests that even the (unsupervised) clustering performance obtained by Couillet and Benaych-Georges (2015) is not achieved, despite the presence of possibly numerous unlabelled data. This, we believe, is due to a mismatched scaling in the SSL problem definition. A promising avenue of investigation would consist in introducing appropriate scaling parameters in the label propagation method or the optimization (1) to ensure that $W_{[uu]}$ is effectively used in the algorithm. Early simulations do suggest that elementary amendments to (2) indeed result in possibly striking performance improvements.

These considerations are left to future works.

Acknowledgments

This work is supported by the ANR Project RMT4GRAPH (ANR-14-CE28-0006).

Appendix A. Preliminaries

We begin with some additional notations that will be useful in the proofs.

- For $x_i \in \mathcal{C}_k$, $\omega_i \equiv (x_i - \mu_k)/\sqrt{p}$, and $\Omega \equiv [\omega_1, \cdots, \omega_n]^\mathsf{T}$
- $j_k \in \mathbb{R}^n$ is the canonical vector of $\mathcal{C}_k$, in the sense that its i-th element is 1 if $x_i \in \mathcal{C}_k$ or 0 otherwise. $j_{[l]k}$ and $j_{[u]k}$ are respectively the canonical vectors for labelled and unlabelled data of $\mathcal{C}_k$.
- $\psi_i \equiv \|\omega_i\|^2 - \mathrm{E}[\|\omega_i\|^2]$, $\psi \equiv [\psi_1, \cdots, \psi_n]^\mathsf{T}$ and $(\psi)^2 \equiv [(\psi_1)^2, \cdots, (\psi_n)^2]^\mathsf{T}$.

With these notations at hand, we introduce next the generalized version of Theorem 2 for all $\alpha = O(1)$ (rather than $\alpha = -1 + O(1/\sqrt{n})$).

Theorem 5 *For $x_i \in \mathcal{C}_b$ an unlabelled vector (i.e., $i > n_{[l]}$), let $\hat{F}_{ia}$ be given by (7) with F defined in (2) for $\alpha = O(1)$. Then, under Assumptions 1–2,*

$$p\hat{F}_{i\cdot} = p(1+z_i)1_K + G_i + o_P(1)$$
$$G_i \sim \mathcal{N}(m_b, \Sigma_b)$$

where z_i is as in Theorem 2 and

(i) for $F_{i\cdot}$ considered on the σ-field induced by the random variables $x_{[l]+1}, \ldots, x_n$, $p = 1, 2, \ldots$,

$$[m_b]_a = H_{ab} + \frac{1}{n_{[l]}} \sum_{d=1}^{K} (\alpha n_d + n_{[u]d}) H_{ad} + (1+\alpha)\frac{n}{n_{[l]}} \left[\Delta_a + \frac{p}{n_{[l]a}} \frac{f'(\tau)}{f(\tau)} \psi_{[l]}^T j_{[l]a} - \alpha \frac{f'(\tau)^2}{f(\tau)^2} t_a t_b \right] \tag{13}$$

$$[\Sigma_b]_{a_1 a_2} = \left(\frac{(-\alpha^2 - \alpha)n - n_{[l]}}{n_{[l]}} \frac{f'(\tau)^2}{f(\tau)^2} + \frac{f''(\tau)}{f(\tau)} \right)^2 T_{bb} t_{a_1} t_{a_2} + \boldsymbol{\delta}_{a_1}^{a_2} \frac{f'(\tau)^2}{f(\tau)^2} \frac{4c_0 T_{ba_1}}{c_{[l]a_1}} + \frac{4f'(\tau)^2}{f(\tau)^2} \mu_{a_1}^\circ C_b \mu_{a_2}^\circ \tag{14}$$

where

$$H_{ab} = \frac{f'(\tau)}{f(\tau)} \|\mu_b^\circ - \mu_a^\circ\|^2 + \left(\frac{f''(\tau)}{f(\tau)} - \frac{f'(\tau)^2}{f(\tau)^2} \right) t_a t_b + \frac{2f''(\tau)}{f(\tau)} T_{ab} \tag{15}$$

$$\Delta_a = \frac{\sqrt{p} f'(\tau)}{f(\tau)} t_a + \frac{\alpha f'(\tau)^2 + f(\tau) f''(\tau)}{2f(\tau)^2} \left(2T_{aa} + t_a^2 \right) + \frac{1}{n_{[l]}} \left(\frac{f'(\tau)}{f(\tau)} \right)^2 \left(\sum_{d=1}^{K} n_{[u]d} t_d \right) t_a. \tag{16}$$

(ii) for $F_{i\cdot}$ considered on the σ-field induced by the random variables $x_1, \ldots, x_n$,

$$[m_b]_a = H_{ab} + \frac{1}{n_{[l]}} \sum_{d=1}^{K} (\alpha n_d + n_{[u]d}) H_{ad} + (1+\alpha)\frac{n}{n_{[l]}} \left[\Delta_a - \alpha \frac{f'(\tau)^2}{f(\tau)^2} t_a t_b \right]$$

$$[\Sigma_b]_{a_1 a_2} = \left(\frac{(-\alpha^2 - \alpha)n - n_{[l]}}{n_{[l]}} \frac{f'(\tau)^2}{f(\tau)^2} + \frac{f''(\tau)}{f(\tau)} \right)^2 T_{bb} t_{a_1} t_{a_2} + \boldsymbol{\delta}_{a_1}^{a_2} \frac{f'(\tau)^2}{f(\tau)^2} \left(\frac{(1+\alpha)^2}{c_{[l]}^2} \frac{2c_0 T_{aa}}{c_{[l]a_1}} + \frac{4c_0 T_{ba_1}}{c_{[l]a_1}} \right) + \frac{4f'(\tau)^2}{f(\tau)^2} \mu_{a_1}^\circ C_b \mu_{a_2}^\circ$$

with H_{ab} given in (15) and Δ_a in (16).

Let $P(x_i \to \mathcal{C}_b | x_i \in \mathcal{C}_b, x_1, \cdots, x_{n_{[l]}})$ denote the probability of correct classification of $x_i \in \mathcal{C}_b$ unlabelled, conditioned on $x_1, \ldots, x_{n_{[l]}}$, and $P(x_i \to \mathcal{C}_b | x_i \in \mathcal{C}_b)$ the unconditional

probability. Recall that the probability of correct classification of $x_i \in \mathcal{C}_b$ is the same as the probability of $\hat{F}_{ib} > \max_{a \neq b} \hat{F}_{ib}$, which, according to the above theorem, is asymptotically the probability that $[G_i]_b$ is the greatest element of G_i. Particularly for $K = 2$, we have the following corollary.

Corollary 6 *Under the conditions of Theorem 1, and with $K = 2$, we have, for $a \neq b \in \{1, 2\}$,*

(i) Conditionally on $x_1, \cdots, x_{n_{[l]}}$,

$$\mathbb{P}\left(x_i \to \mathcal{C}_b | x_i \in \mathcal{C}_b, x_1, \cdots, x_{n_{[l]}}\right) - \Phi(\theta_b^a) \to 0$$
$$\theta_b^a = \frac{[m_b]_b - [m_b]_a}{\sqrt{[\Sigma_b]_{bb} + [\Sigma_b]_{aa} - 2[\Sigma_b]_{ab}}}$$

where $\Phi(u) = \frac{1}{2\pi} \int_{-\infty}^{u} \exp(-t^2/2) dt$ and m_b, Σ_b are given in (i) of Theorem 5.

(ii) Unconditionally,

$$\mathbb{P}\left(x_i \to \mathcal{C}_b | x_i \in \mathcal{C}_b\right) - \Phi(\theta_b^a) \to 0$$
$$\theta_b^a = \frac{[m_b]_b - [m_b]_a}{\sqrt{[\Sigma_b]_{bb} + [\Sigma_b]_{aa} - 2[\Sigma_b]_{ab}}}$$

where here m_b, Σ_b are given in (ii) of Theorem 5.

The remainder of the appendix is dedicated to the proof of Theorem 5 and Corollary 6 from which the results of Section 3.2 directly unfold.

Appendix B. Proof of Theorems 5

The proof of Theorem 5 is divided into two steps: first, we Taylor-expand the normalized scores for unlabelled data $\hat{F}_{[u]}$ using the convergence $\frac{1}{p}\|x_i - x_j\|^2 \xrightarrow{\text{a.s.}} \tau$ for all $i \neq j$; this expansion yields a random equivalent $\hat{F}_{[u]}^{\text{eq}}$ in the sense that $p(\hat{F}_{[u]} - \hat{F}_{[u]}^{\text{eq}}) \xrightarrow{\text{a.s.}} 0$. Proposition 1 is directly obtained from $\hat{F}_{[u]}^{\text{eq}}$. We then complete the proof by demonstrating the convergence to Gaussian variables of $\hat{F}_{[u]}^{\text{eq}}$ by means of a central limit theorem argument.

B.1. Step 1: Taylor expansion

In the following, we provide a sketch of the development of $F_{[u]}$; most unshown intermediary steps can be retrieved from simple, yet painstaking algebraic calculus.

Recall from (2) the expression of the unnormalized scores for unlabelled data

$$F_{[u]} = (I_{n_u} - D_{[u]}^{-1-\alpha} W_{[uu]} D_{[u]}^{\alpha})^{-1} D_{[u]}^{-1-\alpha} W_{[ul]} D_{[l]}^{\alpha} F_{[l]}.$$

We first proceed to the development of the terms $W_{[ul]}$, $W_{[uu]}$, subsequently to $D_{[l]}, D_{[u]}$, to then reach an expression for $F_{[u]}$. To this end, owing to the convergence $\|x_i - x_j\|^2/p \xrightarrow{\text{a.s.}} \tau$

for all $i \neq j$, we first Taylor-expand $W_{ij} = f(\|x_i - x_j\|^2/p)$ around $f(\tau)$ to obtain the following expansion for W, already evaluated by Couillet and Benaych-Georges (2015),

$$W = W^{(n)} + W^{(\sqrt{n})} + W^{(1)} + O(n^{-\frac{1}{2}}) \tag{17}$$

where $\|W^{(n)}\| = O(n)$, $\|W^{(\sqrt{n})} = O(\sqrt{n})$ and $\|W^{(1)}\| = O(1)$, with the definitions

$$\begin{aligned}
W^{(n)} &= f(\tau) 1_n 1_n^\mathsf{T} \\
W^{(\sqrt{n})} &= f'(\tau) \left[\psi 1_n^\mathsf{T} + 1_n \psi^\mathsf{T} + \left(\sum_{b=1}^K \frac{t_b}{\sqrt{p}} j_b \right) 1_n^\mathsf{T} + 1_n \sum_{a=1}^K \frac{t_a}{\sqrt{p}} j_a^\mathsf{T} \right] \\
W^{(1)} &= f'(\tau) \left[\sum_{a,b=1}^K \frac{\|\mu_a^\circ - \mu_b^\circ\|^2}{p} j_b j_a^\mathsf{T} - \frac{2}{\sqrt{p}} \Omega \sum_{a=1}^K \mu_a^\circ j_a^\mathsf{T} + \frac{2}{\sqrt{p}} \sum_{b=1}^K \mathrm{diag}(j_b) \Omega \mu_b^\circ 1_n^\mathsf{T} \right. \\
&\quad \left. - \frac{2}{\sqrt{p}} \sum_{b=1}^K j_b \mu_b^{\circ\mathsf{T}} \Omega^\mathsf{T} + \frac{2}{\sqrt{p}} 1_n \sum_{a=1}^K \mu_a^{\circ\mathsf{T}} \Omega^\mathsf{T} \mathrm{diag}(j_a) - 2\Omega\Omega^\mathsf{T} \right] \\
&\quad + \frac{f''(\tau)}{2} \left[(\psi)^2 1_n^\mathsf{T} + 1_n [(\psi)^2]^\mathsf{T} + \sum_{b=1}^K \frac{t_b^2}{p} j_b 1_n^\mathsf{T} + 1_n \sum_{a=1}^K \frac{t_a^2}{p} j_a^\mathsf{T} \right. \\
&\quad + 2 \sum_{a,b=1}^K \frac{t_a t_b}{p} j_b j_a^\mathsf{T} + 2 \sum_{b=1}^K \mathrm{diag}(j_b) \frac{t_b}{\sqrt{p}} \psi 1_n^\mathsf{T} + 2 \sum_{b=1}^K \frac{t_b}{\sqrt{p}} j_b \psi^\mathsf{T} + 2 \sum_{a=1}^K 1_n \psi^\mathsf{T} \mathrm{diag}(j_a) \frac{t_a}{\sqrt{p}} \\
&\quad \left. + 2\psi \sum_{a=1}^K \frac{t_a}{\sqrt{p}} j_a^\mathsf{T} + 4 \sum_{a,b=1}^K \frac{T_{ab}}{p} j_b j_a^\mathsf{T} + 2\psi\psi^\mathsf{T} \right] + (f(0) - f(\tau) + \tau f'(\tau)) I_n.
\end{aligned}$$

As $W_{[ul]}$, $W_{[uu]}$ are sub-matrices of W, their approximated expressions are obtained directly by extracting the corresponding subsets of (17). Applying then (17) in $D = \mathrm{diag}(W 1_n)$, we next find

$$D = n f(\tau) \left[I_n + \frac{1}{n f(\tau)} \mathrm{diag}(W^{(\sqrt{n})} 1_n + W^{(1)} 1_n) \right] + O(n^{-\frac{1}{2}}).$$

Thus, for any $\sigma \in \mathbb{R}$, $(n^{-1}D)^\sigma$ can be Taylor-expanded around $f(\tau)^\sigma I_n$ as

$$\begin{aligned}
(n^{-1}D)^\sigma &= f(\tau)^\sigma \left[I_n + \frac{\sigma 1}{n f(\tau)} \mathrm{diag}((W^{(\sqrt{n})} + W^{(1)}) 1_n) + \frac{\sigma(\sigma-1)}{2n^2 f(\tau)^2} \mathrm{diag}^2(W^{(\sqrt{n})} 1_n) \right] \\
&\quad + O(n^{-\frac{3}{2}})
\end{aligned} \tag{18}$$

where $\mathrm{diag}^2(\cdot)$ stands for the squared diagonal matrix. The Taylor-expansions of $(n^{-1}D_{[u]})^\alpha$ and $(n^{-1}D_{[l]})^\alpha$ are then directly extracted from this expression for $\sigma = \alpha$, and similarly for $(n^{-1}D_{[u]})^{-1-\alpha}$ with $\sigma = -1-\alpha$. Since

$$D_{[u]}^{-1-\alpha} W_{[ul]} D_{[l]}^\alpha = \frac{1}{n} (n^{-1} D_{[u]})^{-1-\alpha} W_{[ul]} (n^{-1} D_{[l]})^\alpha$$

it then suffices to multiply the Taylor-expansions of $(n^{-1}D_{[u]})^\alpha$, $(n^{-1}D_{[l]})^\alpha$, and $W_{[ul]}$, given respectively in (18) and (17), normalize by n and then organize the result in terms of order $O(1)$, $O(1/\sqrt{n})$, and $O(1/n)$.

The term $D_{[u]}^{-1-\alpha} W_{[uu]} D_{[u]}^{\alpha}$ is dealt with in the same way. In particular,

$$D_{[u]}^{-1-\alpha} W_{[uu]} D_{[u]}^{\alpha} = \frac{1}{n} 1_{n_{[u]}} 1_{n_{[l]}} + O(n^{-\frac{1}{2}}).$$

Therefore, $(I_{n_{[u]}} - D_{[u]}^{-1-\alpha} W_{[uu]} D_{[u]}^{\alpha})^{-1}$ may be simply written as

$$\left(I_{n_{[u]}} - \frac{1}{n} 1_{n_{[u]}} 1_{n_{[u]}} + O(n^{-\frac{1}{2}}) \right)^{-1} = I_{n_{[u]}} + \frac{1}{n_{[l]}} 1_{n_{[u]}} 1_{n_{[u]}} + O(n^{-\frac{1}{2}}).$$

Combining all terms together completes the full linearization of $\hat{F}_{[u]}$.

This last derivation, which we do not provide in full here, is simpler than it appears and is in fact quite instructive in the overall behavior of $F^{[u]}$. Indeed, only product terms in the development of $(I_{n_{[u]}} - D_{[u]}^{-1-\alpha} W_{[uu]} D_{[u]}^{\alpha})^{-1}$ and $D_{[u]}^{-1-\alpha} W_{[ul]} D_{[l]}^{\alpha} F^{[l]}$ of order at least $O(1)$ shall remain, which discards already a few terms. Now, in addition, note that for any vector v, $v 1_{n_{[l]}}^{\mathsf{T}} F^{[l]} = v 1_k^{\mathsf{T}}$ so that such matrices are non informative for classification (they have identical score columns); these terms are all placed in the intermediary variable z, the entries z_i of which are irrelevant and thus left as is (these are the z_i's of Proposition 1 and Theorem 2). It is in particular noteworthy to see that *all* terms of $W_{[uu]}^{(1)}$ that remain after taking the product with $D_{[u]}^{-1-\alpha} W_{[ul]} D_{[l]}^{\alpha} F^{[l]}$ are precisely those multiplied by $f(\tau) 1_{n_{[u]}} 1_{n_{[l]}}^{\mathsf{T}} F^{[l]}$ and thus become part of the vector z. Since most informative terms in the kernel matrix development are found in $W^{(1)}$, this means that the algorithm under study shall make little use of the *unsupervised* information about the data (those found in $W_{[uu]}^{(1)}$). This is an important remark which, as discussed in Section 6, opens up the path to further improvements of the semi-supervised learning algorithms which would use more efficiently the information in $W_{[uu]}^{(1)}$.

All calculus made, this development finally leads to $F_{[u]} = F_{[u]}^{\text{eq}}$ with, for $a, b \in \{1, \ldots, K\}$ and $x_i \in \mathcal{C}_b$, $i > n_{[l]}$,

$$\begin{aligned}
\hat{F}_{ia}^{\text{eq}} = 1 &+ \frac{1}{p} \left[H_{ab} + \frac{1}{n_{[l]}} \sum_{d=1}^{K} H_{ad} (\alpha n_d + n_{[u]d}) \right] + (1+\alpha) \frac{n}{p n_{[l]}} \left[\Delta_a - \alpha \frac{f'(\tau)^2}{f(\tau)^2} t_a t_b \right] \\
&+ \left(\frac{(-\alpha^2 - \alpha) n - n_{[l]}}{n_{[l]}} \frac{f'(\tau)^2}{f(\tau)^2} + \frac{f''(\tau)}{f(\tau)} \right) \frac{t_a}{\sqrt{p}} \psi_i + \frac{2 f'(\tau)}{f(\tau) \sqrt{p}} \mu_a^{\circ} \omega_i \\
&+ \frac{f'(\tau)}{f(\tau)} \left(\frac{(1+\alpha) n}{n_{[l]} n_{[l]a}} \psi_{[l]}^{\mathsf{T}} j_{[l]a} + \frac{4}{n_{[l]a}} j_{[l]a}^T \Omega_{[l]} \omega_i \right) + z_i
\end{aligned} \tag{19}$$

where H_{ab} is as specified in (15), Δ_a as in (16), and $z_i = O(\sqrt{p})$ is some residual random variable only dependent on x_i. Gathering the terms in successive orders of magnitude, Proposition 1 is then straightforwardly proven from (19).

B.2. Step 2: Central limit theorem

The focus of this step is to examine $\mathcal{G}_i = p(\hat{F}_i^{\text{eq}} - (1 + z_i) 1_K)$. Theorem 5 can be proven by showing that $\mathcal{G}_i = G_i + o_P(1)$.

First consider Item (i) of Theorem 5, which describes the behavior of $\hat{F}_{[u]}$ conditioned on $x_1, \ldots, x_{n_{[l]}}$. Recall that a necessary and sufficient condition for a vector v to be a Gaussian vector is that all linear combinations of the elements of v are Gaussian variables. Thus, for given $x_1, \ldots, x_{n_{[l]}}$ deterministic, according to (19), $\mathcal{G}_i$ is asymptotically Gaussian if, for all $g_1 \in \mathbb{R}$, $g_2 \in \mathbb{R}^p$, $g_1\psi_i + g_2^T\omega_i$ has a central limit.

Letting $\omega_i = \frac{C_b^{\frac{1}{2}}}{\sqrt{p}} r$, with $r \sim \mathcal{N}(0, I_p)$, $g_1\psi_i + g_2\omega_i$ can be rewritten as $r^\mathsf{T} A r + br + c$ with $A = g_1\frac{C_b}{p}$, $b = g_2\frac{C_b}{p}$, $c = -g_1\frac{\mathrm{tr} C_b}{p}$. Since A is symmetric, there exists an orthonormal matrix U and a diagonal Λ such that $A = U^\mathsf{T}\Lambda U$. We thus get

$$r^\mathsf{T} A r + br + c = r^T U^T \Lambda U r + bU^T U r + c = \tilde{r}^T \Lambda \tilde{r} + \tilde{b}\tilde{r} + c$$

with $\tilde{r} = Ur$ and $\tilde{b} = bU^\mathsf{T}$. By unitary invariance, we have $\tilde{r} \sim \mathcal{N}(0, I_p)$ so that $g_1\psi_i + g_2\omega_i$ is thus the sum of the independent but not identically distributed random variables $q_j = \lambda_j \tilde{r}_j^2 + \tilde{b}_j\tilde{r}_j$, $i = 1, \ldots, p$. From Lyapunov's central limit theorem (Billingsley, 1995, Theorem 27.3), it remains to find a $\delta > 0$ such that $\frac{\sum_j \mathbb{E}|q_j - \mathbb{E}[q_j]|^{2+\delta}}{\left(\sum_j \mathrm{Var}[q_j]\right)^{1+\delta/2}} \to 0$ to ensure the central limit theorem. For $\delta = 1$, we have $\mathbb{E}[q_j] = \lambda_j$, $\mathrm{Var}[q_j] = 2\lambda_j^2 + \tilde{b}_j^2$ and $\mathbb{E}\left[(q_j - \mathbb{E}[q_j])^3\right] = 8\lambda_j^3 + 6\lambda_j\tilde{b}_j^2$, so that $\frac{\sum_j \mathbb{E}\left[|q_j - \mathbb{E}[q_j]|^3\right]}{\left(\sum_j \mathrm{Var}[q_j]\right)^{3/2}} = O(n^{-\frac{1}{2}})$.

It thus remains to evaluate the expectation and covariance matrix of $\mathcal{G}_i$ conditioned on $x_1, \ldots, x_{n_{[l]}}$ to obtain (i) of Theorem 5. For $x_i \in \mathcal{C}_b$, we have

$$\begin{aligned}
\mathbb{E}\{[\mathcal{G}_i]_a\} &= H_{ab} + \frac{1}{n_{[l]}} \sum_{d=1}^{K} (\alpha n_d + n_{[u]d}) H_{ad} \\
&\quad + (1+\alpha)\frac{n}{n_{[l]}} \left[\Delta_a + \frac{p}{n_{[l]a}} \frac{f'(\tau)}{f(\tau)} \psi_{[l]}^T j_{[l]a} - \alpha \frac{f'(\tau)^2}{f(\tau)^2} t_a t_b \right] \\
\mathrm{Cov}\{[\mathcal{G}_i]_{a_1}[\mathcal{G}_i]_{a_2}\} &= \left(\frac{(-\alpha^2 - \alpha)n - n_{[l]}}{n_{[l]}} \frac{f'(\tau)^2}{f(\tau)^2} + \frac{f''(\tau)}{f(\tau)} \right)^2 T_{bb} t_{a_1} t_{a_2} \\
&\quad + \boldsymbol{\delta}_{a_1}^{a_2} \frac{f'(\tau)^2}{f(\tau)^2} \frac{4c_0 T_{ba_1}}{c_{[l]a_1}} + \frac{4f'(\tau)^2}{f(\tau)^2} \mu_{a_1}^\circ C_b \mu_{a_2}^\circ + o(1).
\end{aligned}$$

From the above equations, we retrieve the asymptotic expressions of $[m_b]_a$ and $[\Delta_b]_{a_1 a_2}$ given in (13) and (14). This completes the proof of Item (i) of Theorem 5. Item (ii) is easily proved by following the same reasoning.

References

Aamir Anis, Aly El Gamal, Salman Avestimehr, and Antonio Ortega. Asymptotic justification of bandlimited interpolation of graph signals for semi-supervised learning. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 5461–5465. IEEE, 2015.

Konstantin Avrachenkov, Paulo Gonçalves, Alexey Mishenin, and Marina Sokol. Generalized optimization framework for graph-based semi-supervised learning. *arXiv preprint arXiv:1110.4278*, 2011.

Mikhail Belkin and Partha Niyogi. Semi-supervised learning on riemannian manifolds. *Machine learning*, 56(1-3):209–239, 2004.

Mikhail Belkin, Irina Matveeva, and Partha Niyogi. Regularization and semi-supervised learning on large graphs. In *International Conference on Computational Learning Theory (COLT)*, pages 624–638. Springer, 2004.

Peter J Bickel, Bo Li, et al. Local polynomial regression on unknown manifolds. In *Complex Datasets and Inverse Problems*, pages 177–186. Institute of Mathematical Statistics, 2007.

P. Billingsley. *Probability and Measure*. John Wiley and Sons, Inc., Hoboken, NJ, third edition, 1995.

Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien. *Semi-Supervised Learning*. MIT press, 2006.

Romain Couillet and Florent Benaych-Georges. Kernel spectral clustering of large dimensional data. *arXiv preprint arXiv:1510.03547*, 2015.

Romain Couillet and Abla Kammoun. Random matrix improved subspace clustering. In *Asilomar Conference on Signals, Systems and Computers*, pages 90–94. IEEE, 2016.

Akshay Gadde, Aamir Anis, and Antonio Ortega. Active semi-supervised learning using sampling theory for graph signals. In *International Conference on Knowledge Discovery and Data Mining*, pages 492–501. ACM, 2014.

Amir Globerson, Roi Livni, and Shai Shalev-Shwartz. Effective semi-supervised learning on manifolds. In *International Conference on Learning Theory (COLT)*, pages 978–1003, 2017.

Andrew B Goldberg, Xiaojin Zhu, Aarti Singh, Zhiting Xu, and Robert Nowak. Multi-manifold semi-supervised learning. 2009.

Martin Szummer Tommi Jaakkola and Martin Szummer. Partially labeled classification with markov random walks. *International Conference in Neural Information Processing Systems*, 14:945–952, 2002.

Thorsten Joachims et al. Transductive learning via spectral graph partitioning. In *International Conference on Machine Learning*, volume 3, pages 290–297, 2003.

Yann LeCun, Corinna Cortes, and Christopher JC Burges. The mnist database of handwritten digits, 1998.

Zhenyu Liao and Romain Couillet. A large dimensional analysis of least squares support vector machines. *arXiv preprint arXiv:1701.02967*, 2017.

Amit Moscovich, Ariel Jaffe, and Boaz Nadler. Minimax-optimal semi-supervised regression on unknown manifolds. *arXiv preprint arXiv:1611.02221*, 2016.

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

Sunil K Narang, Akshay Gadde, and Antonio Ortega. Signal processing techniques for interpolation in graph structured data. In *IEEE International Conference Acoustics, Speech and Signal Processing*, pages 5445–5449. IEEE, 2013a.

Sunil K Narang, Akshay Gadde, Eduard Sanou, and Antonio Ortega. Localized iterative methods for interpolation in graph structured data. In *Global Conference on Signal and Information Processing*, pages 491–494. IEEE, 2013b.

Bernhard Schölkopf and Alexander J Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT press, 2002.

Larry Wasserman and John D Lafferty. Statistical analysis of semi-supervised regression. In *International Conference on Neural Information Processing Systems*, pages 801–808, 2008.

Dengyong Zhou, Olivier Bousquet, Thomas Navin Lal, Jason Weston, and Bernhard Schölkopf. Learning with local and global consistency. volume 16, pages 321–328, 2004.

Xiaojin Zhu and Zoubin Ghahramani. Learning from labeled and unlabeled data with label propagation. Technical report, Citeseer, 2002.

Xiaojin Zhu, Zoubin Ghahramani, John Lafferty, et al. Semi-supervised learning using gaussian fields and harmonic functions. In *International Conference on Machine Learning*, volume 3, pages 912–919, 2003.