Random Matrix-Inspired Improved Semi-Supervised Learning on Graphs

Xiaoyi Mai [1] **Romain Couillet** [1]

Abstract

The recent work (Mai & Couillet, 2017) proves that in the so-called "random matrix regime" (i.e., for large quantities of high dimensional data), classical Laplacian regularization for semi-supervised learning fails to effectively extract information from unlabelled dat. As an answer, this article proposes a kernel centering regularization, which eliminates the phenomenon known as classification scores 'flatness' (Nadler et al., 2009). Besides intuitive justification, this new scheme is both theoretically supported and shown empirically to outperform traditional Laplacian regularizations and their state-of-art variants on real datasets.

1. Introduction

Semi-supervised learning (SSL) is a machine learning method jointly exploiting labelled and unlabelled data. As data labelling is a time consuming manual process as opposed to the inexpensive collection of data, semi-supervised learning aims particularly to improve classification accuracy, using a large amount of unlabelled data in conjunction with few labelled data. Obviously, combining both labelled and unlabelled data should allow semi-supervised learning to consistently outperform both supervised or unsupervised learning taken alone. However, this outperformance is rarely observed in practice as discussed in (Chapelle et al., 2009), thereby hindering semi-supervised learning techniques from being more popular.

Among these techniques, the Laplacian regularization (Zhu et al., 2003b; Zhou et al., 2004) is a classical method of graph-based semi-supervised classification with underlying connections to label propagation, random walk and electrical networks. Although driven by a straightforward reasoning to learn from the data affinity graph (as presented in Subsection 2.2), the Laplacian regularization suffers the severe flaw of having *flat* (i.,e., asymptotically equal) classification "scores" assigned to unlabelled data in the limit of infinite unlabelled data (Nadler et al., 2009) or in the case of extremely large dimensional data (Mai & Couillet, 2017). Since firstly identified by (Nadler et al., 2009), this problem has been addressed in many works (El Alaoui et al., 2016; Zhou & Belkin, 2011; Bridle & Zhu, 2013), advocating for a higher order regularization (see Subsection 2.3 for more details).

Alternatively to higher order regularization, we propose here a kernel centering approach (introduced in Subsection 3.1), motivated by the recent contribution in (Mai & Couillet, 2017), where a comprehensive understanding of Laplacian regularization algorithms, obtained through a random matrix approach, allows to pinpoint the issue. Indeed, beyond the asymptotic 'flat' score limit evidenced in (Nadler et al., 2009), the more profound results of (Mai & Couillet, 2017) reveal that proper classification is still achievable, as in fact already confirmed by experimental observations in (Nadler et al., 2009). Nonetheless, despite the possibility of asymptotic non-trivial classification, the analysis in(Mai & Couillet, 2017) does identify the additional aforementioned issue of unlabelled data having negligible effect on improving the performance. In addition to solving the flatness of score issue, our present method, inspired by a random matrix analysis detailed in Section 5, seeks directly to boost the classification performance through an enhanced usage of unlabelled data.

The remainder of the article is thus organized as follows. Graph-based semi-supervised learning techniques are presented in Section 2. We introduce the proposed kernel centering regularization in the following section, along with the main arguments justifying its usage. Experimental results on real datasets are provided in Section 4, where the kernel centering algorithms are shown to outperform not only the Laplacian regularization but also its state-of-art variants. Finally, we theoretically support the superiority of the kernel centering regularization from a random matrix standpoint in Section 5, before concluding the article.

2. Background

2.1. Problem setup

Consider a set of n data vectors $X = [x_1, \ldots, x_n] \in \mathbb{R}^{p\times n}$ of dimension p, belonging to one of the similarity classes $\mathcal{C}_1$ or $\mathcal{C}_2$.[1] The dataset is divided in two subsets $X = [X_u \quad X_l]$ with 'u' and 'l' respectively standing for 'unlabelled' and

[1]For simplicity, we focus here on binary classification. Algorithms proposed in the article are adaptable to multiple classification in a tradional one-vesus-all manner.

'labelled', i.e., we dispose of a labeling y_l (-1 for $\mathcal{C}_1$, 1 for $\mathcal{C}_2$) for the data vectors composing X_l, while the class of the data in X_u remains to be determined. There are n_l elements in X_l and n_u in X_u.

Data points x_i are viewed as vertices on a graph, the similarity W_{ij} between two data points x_i and x_j is a non-negative value, often given by a kernel function $W_{ij} = \kappa(x_i, x_j)$ (e.g., $\kappa(x, y) = \exp(-\|x - y\|^2/\sigma^2)$).

The objective of graph-based semi-supervised leaning is to produce classification scores for unlabelled data X_u by making use of the pre-labelled data X_l and the data geometry induced by the graph structure (e.g., W_{ij}).

2.2. Laplacian regularization on graph

Laplacian regularization consists in finding a classification score function f in accordance with the known labels y_l of X_l and smooth over the graph, which is to say that f varies little between two close data points (i.e., when W_{ij} is large). The smoothness condition over the graph is incorporated in a penalty term $\sum_{ij=1}^n W_{ij}([f]_i - [f]_j)^2$ and the respect to labelled data is measured by $\|f_l - y_l\|^2$, where $f_l \in \mathbb{R}^{n_l}$ is the score vector of the labelled data, leading to the following optimization problem,

$$\min_{f \in \mathbb{R}^n} \frac{1}{2} \sum_{ij=1}^n W_{ij}([f]_i - [f]_j)^2 + \lambda \|f_l - y_l\|^2, \quad (1)$$

where the parameter λ determines the trade-off between forcing the known labels y_l on labelled nodes and ensuring the smoothness of f over the graph. As W_{ij} are non-negative, the optimization problem is convex with solution:

$$f = \left[S + \lambda^{-1} L\right]^{-1} \begin{bmatrix} 0 \\ y_l \end{bmatrix} \quad (2)$$

where $L = D - W$ with W being the affinity matrix composed of W_{ij} and D the diagonal 'degree' matrix with $D_{ii} = \sum_{j=1}^n W_{ij}$. The matrix L is referred to as Laplacian matrix, or more precisely the unnormalized Laplacian which is distinct from other versions of Laplacian matrices, such as the symmetric normalized Laplacian $L_s = I_n - D^{-1/2} W D^{-1/2}$, random walk normalized Laplacian $L_r = I_n - D^{-1} W$ and $L_g = I_n - D_g^{-1} W_g$ with $W_g = D^{-1} W D^{-1}$, D_g being the corresponding degree diagonal matrix $[D_g]_{ii} = \sum_{j=1}^n [W_g]_{ij}$, is referred to as the geometry Laplacian in (Zhou & Belkin, 2011). Alternative formulations of (1) through various normalizations lead to substituting L in (2) for one of these Laplacian versions.

2.3. Higher order regularization

There are mainly two forms of higher order regularizations. One is called l_p-based Laplacian regularization, which forces a stronger constrain on the smoothness by minimizing $\sum_{i,j=1}^n (W_{ij})^m |[f]_i - [f]_j|^m$ (Zhou & Schölkopf, 2005; El Alaoui et al., 2016). The other proposes to use iterated Laplacian regularizer $f^\mathsf{T} L^m f$ (Zhou & Belkin, 2011), allowing to penalize up to m derivatives of f instead of the first derivative. As opposed to standard Laplacian regularization, both regularizers are shown to yield non-flat solutions of f_u with $m > d + 1$ for the l_p-based Laplacian (El Alaoui et al., 2016) and $m > d/2$ for the iterated Laplacian (Zhou & Belkin, 2011). In comparison, the iterated Laplacian is more computationally efficient than l_p-based Laplacian regularization since it has an explicit solution. Also, the iterated Laplacian is found to outperform p-voltages Laplacian regularization (Bridle & Zhu, 2013), a dual version of l_p-based Laplacian in the context of electrical networks. In addition to avoiding the score 'flatness' issue, these higher order regularizers or their variations $f^\mathsf{T} g(L) f$ (Smola & Kondor, 2003) obviously benefit from more degrees of freedom to improve the classification performance.

2.4. Manifold based method

Another closely related method is proposed by (Belkin & Niyogi, 2003). Rather than regularizing f over the graph, this approach computes first the eigenmap of L, then uses a certain number of eigenvectors $E = [e_1, e_2, \cdots, e_s]$ associated with smallest eigenvalues except 0 to construct a linear subspace and search within this space for an f which minimizes $\|f_l - y_l\|$. By the method of least squares, $f = Ea$ with $a = (E_l^\mathsf{T} E_l)^{-1} E_l^\mathsf{T} y_l$.

3. Algorithms and motivation

3.1. Centered kernel regularization

The proposed method consists in a slight, yet fundamental, modification of Laplacian regularization introduced in Subsection 2.2. As subsequently discussed in Subsection 3.2 , this update is mainly inspired by the recent random matrix theoretical findings of (Mai & Couillet, 2017) and essentially allows to better exploit the full set of labelled and unlabelled data as well as to avoid the aforementioned asymptotic score flatness issue.

Our method can be interpreted as following the same reasoning as Laplacian regularization, but with a centered affinity (kernel) matrix $K \in \mathbb{R}^{n \times n}$ of the form

$$K = P \left\{\kappa(x_i, x_j)\right\}_{i,j=1}^n P \quad (3)$$

for some function $\kappa : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$ and $P \equiv I_n - \frac{1}{n} 1_n 1_n^\mathsf{T}$. The function κ is here *any* standard kernel function, such as the heat kernel ($\kappa(x, y) = \exp(-\|x - y\|^2/\sigma^2)$), or more simply a correlation function (e.g., $\kappa(x, y) = x^\mathsf{T} y$).

Recall that in standard Laplacian regularization, the affinity matrix W is defined by $W_{ij} = \kappa(x_i, x_j)$, so that K is

simply obtained by applying the projection matrix P at both sides of W. According to the definition of kernel functions, there exists a mapping $x \to \phi(x)\ \forall x \in \mathbb{R}^p$ such that $\kappa(x_i, x_j) = \phi(x_i)^\mathsf{T}\phi(x_j)$. In this case, $K_{ij} = \left(\phi(x_i) - \frac{1}{n}\sum_{k=1}^n \phi(x_k)\right)^T \left(\phi(x_j) - \frac{1}{n}\sum_{k=1}^n \phi(x_k)\right)$, and the difference between W and K reduces to a translation in the feature space, hence involving no loss of information relevant to the learning task. Notice that

$$\hat{\kappa}(x_i, x_j) = \left(\phi(x_i) - \frac{1}{n}\sum_{k=1}^n \phi(x_k)\right)^\mathsf{T} \left(\phi(x_j) - \frac{1}{n}\sum_{k=1}^n \phi(x_k)\right)$$

is still a kernel function (Scholkopf & Smola, 2001) with mapping $x \to \phi(x) - \frac{1}{n}\sum_{k=1}^n \phi(x_k)$; we call it a centered kernel and K the corresponding centered kernel matrix.

While the relative information between data points remains intact, using K_{ij} as similarity measure does generate a problem: since K has positive and negative elements, the optimization (1) is not necessarily convex and might admit no finite solution. This issue is settled by fixing the norm of f. The optimization problem is thus formulated as

$$\min_{f \in \mathbb{R}^n} \frac{1}{2}\sum_{ij=1}^n K_{ij}([f]_i - [f]_j)^2 + \lambda\|f_l - y_l\|^2 \quad s.t. \quad \|f\| = t. \tag{4}$$

Denote by S the $n \times n$ diagonal matrix with $S_{ii} = 1$ if $x_i \in X_l$, otherwise $S_{ii} = 0$. By the Lagrange multipliers method (α being the Lagrange multiplier), the solution of (4) satisfies

$$f = \left[S + \lambda^{-1}(\alpha I - K)\right]^{-1} \begin{bmatrix} 0 \\ y_l \end{bmatrix} \tag{5}$$

and $\|f\| = t$. In our algorithms, we consider directly α as a parameter (more convenient for implementation and analysis). Splitting K into labelled and unlabelled parts as

$$K = \begin{bmatrix} K_{uu} & K_{ul} \\ K_{lu} & K_{ll} \end{bmatrix}$$

and letting f_u the unlabelled data scores, (5) gives

$$f_u = \lambda(\alpha I_{n_u} - K_{uu} - K_{ul}R_{ll}K_{lu})^{-1}K_{ul}R_{ll}y_l \tag{6}$$

where $R_{ll} = [(\alpha + \lambda)I_{n_l} - K_{ll}]^{-1}$. If instead of penalizing the difference between f_l and y_l, we impose directly $f_l = y_l$ (i.e., $\lambda = +\infty$) as proposed in (Zhu et al., 2003b), f_u takes the shorter form

$$f_u = (\alpha I_{n_u} - K_{uu})^{-1}K_{ul}y_l. \tag{7}$$

The proposed centered kernel approach is summarized in Algorithm 1. It should be pointed out that, as discovered by theoretical analysis in Section 5, the decision threshold zero used in Algorithm 1 is a consistent but not necessarily optimal choice. Designing a better threshold can further improve the classification; this is discussed in Section 5.

Algorithm 1 Centered kernel graph-based SSL

1: **Input:** Pre-labelled dataset X_l with labeling y_l and unlabelled dataset X_u. Kernel function κ. Parameters $\lambda, \alpha > 0$.
2: **Output:** Classification of unlabelled data X_u.
3: Compute centered kernel matrix K by (3).
4: Let $y_l \leftarrow y_l - \frac{1}{n_l}1_{n_l}1_{n_l}^\mathsf{T}y_l$.
5: Compute f_u by (6) if allowing f_l to differ from y_l, or by (7) if not.
6: The unlabelled data classification is performed by affecting all x_i from X_u such that $[f]_i < 0$ to class $\mathcal{C}_1$ and to class $\mathcal{C}_2$ otherwise.

It is easy to check that all exponents of a centered kernel matrix K are themselves centered kernel matrices. It is thus possible to apply polynomial functions to K in the search of a better data graph construction. Finding the right parametrization for the polynomial function can be challenging. Comparing (2) and (5), we notice that the matrix $Q = \alpha I - K$ plays the same role in centered kernel SSL as the Laplacian matrix L in Laplacian regularization. We then borrow the idea from iterated Laplacian so as to propose a higher order regularization for centered kernel SSL, which consists in simply replacing $Q = \alpha I - K$ in (5) with its exponents $Q^{(m)} = (\alpha I - K)^m$, leading to

$$f_u = \lambda\left(Q_{uu}^{(m)} - Q_{ul}^{(m)}R_{ll}Q_{lu}^{(m)}\right)^{-1}Q_{ul}^{(m)}R_{ll}y_l \tag{8}$$

where $R_{ll}^{(m)} = (\lambda I_{n_l} + Q_{ll}^{(m)})^{-1}$. If imposing $f_l = y_l$,

$$f_u = \left(Q_{uu}^{(m)}\right)^{-1}Q_{ul}^{(m)}y_l. \tag{9}$$

The matrices $Q_{uu}^{(m)}$,$Q_{ul}^{(m)}$,$Q_{ll}^{(m)}$ and $Q_{lu}^{(m)}$ in (8) and (9) are understood similarly to K_{uu},K_{ul},K_{ll} and K_{lu}. The iterated centered kernel approach is formalized in Algorithm 2.

3.2. Motivation

It was pointed out in (Mai & Couillet, 2017) that, for many standard kernels, as $n, p \to \infty$ at the same rate, the similarities W_{ij} tend to the same value irrespective of the class of x_i, x_j. This nonetheless does not prevent non-trivial classification performance for an appropriately chosen normalized Laplacian matrix. Yet, growing numbers of unlabelled data were shown to have insignificant contribution to the classification. This problem can in fact be easily solved using the aforementioned centered kernel matrix K as the affinity matrix. We present in this section an intuitive discussion, leaving more formal analysis to Section 5.

Algorithm 2 Iterated centered kernel graph-based SSL

1: **Input:** Pre-labelled dataset X_l with labeling y_l and unlabelled dataset X_u. Kernel function κ. Parameters $\lambda, \alpha > 0$, $m \in \mathbb{N}^*$.
2: **Output:** Classification of unlabelled data X_u.
3: Compute centered kernel matrix K by (3).
4: Let $y_l \leftarrow y_l - \frac{1}{n_l} 1_{n_l} 1_{n_l}^\mathsf{T} y_l$.
5: Set $Q^{(m)} = (\alpha I - K)^m$.
6: Compute f_u by (8) if allowing f_l to differ from y_l, or by (9) if not.
7: The unlabelled data classification is performed by affecting all x_i from X_u such that $[f]_i < 0$ to class $\mathcal{C}_1$ and to class $\mathcal{C}_2$ otherwise.

For the sake of argumentation, we adopt here a simplistic deterministic model for W. Conforming to the large dimensional assumption in (Mai & Couillet, 2017), we suppose that $W_{ij} = 1$ if x_i, x_j belong to different classes, otherwise $W_{ij} = 1 + \epsilon$ with $\epsilon \ll 1$. Notice that if we remove the penalty term $\lambda \|f_l - y_l\|^2$ from (2), the optimization objective is minimized by letting all $[f]_i$ have the same value, which is evidently not the intended solution. To push f as far away as possible from this undesirable outcome, we let $\lambda \to +\infty$, forcing consequently $f_l = y_l$. Then, the solution is $f_u = 0_{n_u}$ if all W_{ij} have exactly the same value ($\epsilon = 0$). But since $\epsilon \ll 1$, f_u is essentially 0_{n_u} with a small fluctuation. Even though a 'correct' classification is still possible thanks to the information contained in the small fluctuation (Mai & Couillet, 2017), this causes unlabelled data to be 'non expressive' on the graph as they have almost non discriminative nodes in comparison with labelled data points. More precisely, the optimization solution is found to coincide with the stationary point of a label propagation(Zhu & Ghahramani, 2002):

$$f_u \leftarrow W_{uu} f_u + W_{ul} y_l.$$

The vector f_u can therefore be seen as receiving scores of classes from y_l and itself through the graph. Since f_u is 0_{n_u} plus an arbitrary small fluctuation, its contribution is negligible when compared with y_l.

Interestingly, the same phenomenon is observed by (Nadler et al., 2009) on finite dimension datasets in the limit of infinite unlabelled data, where it is referred to as the 'flatness' of estimated labels. Again, we try to illustrate this point without going into the technical details of (Nadler et al., 2009). We consider this time $W_{ij} = a$ if x_i, x_j belong to the same class, otherwise $W_{ij} = b$, and $a \neq b$. Similarly as before, $[f]_i$ having constant value is the optimization solution if discarding constraints on labelled data, we then impose $f_l = y_l$. The optimization problem becomes

$$\min \sum_{i,j \in \mathcal{U}} W_{ij} ([f]_i - [f]_j)^2 + 2 \sum_{i \in \mathcal{U}, j \in \mathcal{L}} W_{ij} \|[f]_i - [f]_j\|^2$$

where $f_l = y_l$ and $\mathcal{U}$, $\mathcal{L}$ stand respectively for the index set of unlabelled and labelled data.

As the number of unlabelled data goes to infinity, we have

$$\sum_{i,j \in \mathcal{U}} W_{ij} ([f]_i - [f]_j)^2 \gg \sum_{i \in \mathcal{U}, j \in \mathcal{L}} W_{ij} ([f]_i - [f]_j)^2$$

unless f_u has asymptotically equal elements. Since $\sum_{i,j \in \mathcal{U}} W_{ij} ([f]_i - [f]_j)^2$ is minimized when the entries of f_u are equal, the classification scores of unlabelled data go to the same limit, hence the score 'flatness'.

From the discussion above, we find that in both regimes (large dimension or huge quantity of unlabelled data), the problem stems from the fact that the Laplacian regularizer $\sum_{i,j=1}^n W_{ij} ([f]_i - [f]_j)^2$ reaches its minimum at meaningless solutions ($f = c 1_n$). This is no longer the case with our centered kernel regularizer $\sum_{i,j=1}^n K_{ij} ([f]_i - [f]_j)^2$. In fact, since $\sum_{i,j=1}^n K_{ij} ([f]_i - [f]_j)^2 = -f^\mathsf{T} K f \leq 0$ as K is a definite positive (by the properties of kernel matrices), $[f]_i$ with equal entries is the lest favored scenario by the centered kernel regularizer as it corresponds to the maximum (zero) of the regularizer.

4. Experimental evidence

As discussed in Subsection 3.2 and theoretically demonstrated in Section 3.2, the advantage of the centered kernel regularization over the classical Laplacian regularization is that the classification scores of unlabelled data are not pulled towards the same value regardless of their actual classes, thus allowing for a considerably better utilization of unlabelled data. Beside the centered kernel approach, it is shown in (Zhou & Belkin, 2011) that the iterated Laplacian method is also effective in combating this problem; furthermore, iterated Laplacian yield quite competitive results in comparison with other SSL methods. We note that the manifold based method proposed by (Belkin & Niyogi, 2003) (Subsection 2.4) also offers non-flat solutions for f_u.

In this section, we compare the centered kernel methods (Algorithms 1-2) with the classical Laplacian regularization (Subection 2.2), as well as the iterated Laplacian (Subection 2.3) and the manifold based method (Subsection 2.4), on MNIST hand-written digits (LeCun, 1998) and German Traffic Sign databases (Stallkamp et al., 2012). Since we focus here on binary classification, pairs of categories are selected to build binary datasets.[2]

To emphasize the influence of using centered kernel regularization, we use simple settings: $\kappa(x, y) = \exp(-\|x - y\|^2/\sigma^2)$ with σ being the average Euclidean distances be-

[2] -10dB background Gaussian noises are added to MNIST datasets so as to render the classification more difficult. Similar traffic signs are chosen to form the binary learning tasks.

tween data vectors, $\lambda = +\infty$ (i.e., $f_l = y_l$, known labels are forced on labelled data). As for other parameters, parameters $m \in \mathbb{N}^*$ for iterated Laplacian and iterated centered kernel and $s \in \{1, \cdots, n_l\}$ for manifold based method are selected among all possible values,[3] parameter α in centered kernel algorithms is searched within the range $[l_{\max}/3, 10l_{\max}]$ with $l_{\max}$ the largest eigenvalue of K_{uu}. Furthermore, the Laplacian and iterated Laplacian regularization are tested on all four Laplacian matrices presented in Subsection 2.2; we also use the class mass normalization technique in (Zhu et al., 2003b) to further improve the performance of Laplacian and iterated Laplacian regularization. The results are reported in Table 1 and Table 2, where we observe that the Laplacian regularization is consistently outperformed by the centered kernel regularization (Algorithms 1) and so is the iterated Laplacian regularization by the iterated centered kernel algorithm (Algorithms 1), notably on MNIST datasets, confirming the advantage of using centered kernel matrices for regularization.

Digits	(0,8)	(2,7)	(6,9)
	$n_u = 100$		
Centered kernel	**89.5±3.6**	**89.5±3.4**	**85.3±5.9**
Iterated centered kernel	**89.5±3.6**	**89.5±3.4**	**85.3±5.9**
Laplacian	75.5±5.6	74.2±5.8	70.0±5.5
Iterated Laplacian	87.2±4.7	86.0±5.2	81.4±6.8
Manifold	88.0±4.7	88.4±3.9	82.8±6.5
	$n_u = 500$		
Centered kernel	**91.7±1.3**	**92.2±1.3**	91.6±2.2
Iterated centered kernel	**91.8±1.4**	**92.2±1.3**	**92.0±2.1**
Laplacian	75.6±4.1	74.4±4.0	69.5±3.7
Iterated Laplacian	**91.6±1.5**	91.9±1.4	90.6±2.7
Manifold	90.7±2.1	91.2±1.9	90.1±3.7

Table 1. Comparison of classification accuracy (%) on MNIST datasets with $n_l = 10$. Computed over 1000 random iterations for $n_u = 100$ and 500 for $n_u = 500$.

5. Theoretical support

5.1. Main results

In order to demonstrate that our proposed algorithm outperforms standard kernel semi-supervised learning methods, such as the Page-Rank algorithm and derived versions (Avrachenkov et al., 2012) or the related harmonic function approach (Zhu et al., 2003a), we perform here a statistical analysis of the asymptotic algorithm performance under a simple Gaussian mixture data model and inner-product kernel in sample space in the simultaneously large p, n regime

[3]Theoretically, m can be indefinitely large. However, beyond a certain point, it produces instable solutions due to the ill-condition of regularization matrices.

Class ID	(2,7)	(9,10)	(11,18)
	$n_u = 100$		
Centered kernel	79.0±10.4	77.5±9.2	78.5±7.1
Iterated centered kernel	**85.3±5.9**	**89.2±5.6**	**90.1±6.7**
Laplacian	73.8±9.8	77.3±9.5	78.6±7.2
Iterated Laplacian	83.7±7.2	88.0±6.8	87.1±8.8
Manifold	77.6±8.9	81.4±10.4	82.3±10.8
	$n_u = 500$		
Centered kernel	82.5±4.0	82.6±6.4	79.2±18.0
Iterated centered kernel	**84.4±4.2**	**88.9±5.7**	**95.8±3.2**
Laplacian	72.7±8.9	77.6±8.3	79.1±6.3
Iterated Laplacian	82.7±5.7	88.1±7.4	92.4±6.7
Manifold	77.4±5.9	83.5±10.4	89.3±9.2

Table 2. Comparison of classification accuracy (%) on German Traffic Sign datasets with $n_l = 10$. Computed over 1000 random iterations for $n_u = 100$ and 500 for $n_u = 500$.

(sometimes referred to as the random matrix regime).

We begin with some notations. X is conveniently written

$$X = \begin{bmatrix} X_{u1} & X_{u2} & X_{l1} & X_{l2} \end{bmatrix}$$

with $X_{ui} \in \mathbb{R}^{p \times n_{ui}}$, $X_{li} \in \mathbb{R}^{p \times n_{li}}$ with 'u' and 'l' respectively standing for 'unlabelled' and 'labelled' as before, and '1', '2' denoting the data class index. As such, the total number n_u of unlabelled data is $n_u = n_{u1} + n_{u2}$ and the total number n_l of labelled data is $n_l = n_{l1} + n_{l2}$. The total number of data n_i in $\mathcal{C}_i$ is $n_i = n_{ui} + n_{li}$.

For readability we shall extensively use the shortcut notations $c_{ui} = \frac{n_{ui}}{n}$, $c_{li} = \frac{n_{li}}{n}$, $c_u = \frac{n_u}{n}$, $c_l = \frac{n_l}{n}$, $c_i = \frac{n_i}{n}$, $c_0 = \frac{p}{n}$, along with the vector notations

$$j_1 = \begin{bmatrix} 1_{n_{u1}} \\ 0_{n_{u2}} \\ 1_{n_{l1}} \\ 0_{n_{l2}} \end{bmatrix}, \; j_1^{(u)} = \begin{bmatrix} 1_{n_{u1}} \\ 0_{n_{u2}} \end{bmatrix}, \; j_1^{(l)} = \begin{bmatrix} 1_{n_{l1}} \\ 0_{n_{l2}} \end{bmatrix}$$

and symmetrically for class $\mathcal{C}_2$. We finally define $D_i^{(u)} = \mathrm{diag}(j_i^{(u)}) \in \mathbb{R}^{n_u \times n_u}$.

Write the labeling for the labelled data X_l as $y_l = s_1 j_1^{(l)} + s_2 j_2^{(l)} \in \mathbb{R}^{n_l}$. To balance input scores, we also demand that $s_1 n_1 + s_2 n_2 = 0$ (i.e., $\sum_{i=1}^{n_l} [f_l]_i = 0$), as in Algorithm 1-2. In order to facilitate the analysis, we consider directly that $f_l = y_l$, in which case f_u is given by (7):

$$f_u = (\alpha I_{n_u} - K_{uu})^{-1} K_{ul} y_l.$$

Assuming $s_1 < 0$, the unlabelled data classification is then performed by affecting all x_i such that $[f_u]_i < 0$ to class $\mathcal{C}_1$ and to class $\mathcal{C}_2$ otherwise.

Note that our objective here is merely to get insights into the behavior and demonstrate the gains of the proposed approach, so that more elaborate kernel choices or data settings

are irrelevant at this point. Deeper future investigations, following the lines of recent works on kernel asymptotics (El Karoui, 2010; Couillet & Benaych-Georges, 2016; Liao & Couillet, 2017) will provide a more accurate picture of the method performance on more realistic settings; these are nonetheless technically more demanding and out of scope here.

Precisely, we consider the Gram matrix kernel model

$$K = \frac{1}{n} P X^\mathsf{T} X P$$

and we assume that vector x_i belongs to class $\mathcal{C}_j$, $j \in \{1, 2\}$, if $x_i \sim \mathcal{N}(\mu_j, I_p)$ for some mean vector μ_j. As such, we can write

$$\begin{aligned} X &= \tilde{M} + W \\ \tilde{M} &= \mu_1 j_1^\mathsf{T} + \mu_2 j_2^\mathsf{T} + \begin{bmatrix} W_u & W_l \end{bmatrix} \end{aligned}$$

with W having i.i.d. $\mathcal{N}(0,1)$ entries, and $W_u \in \mathbb{R}^{p \times n_u}$, $W_l \in \mathbb{R}^{p \times n_l}$.

Remark 1 (Centering of μ_1 and μ_2)**.** *Since $P^2 = P$, $\tilde{M}P = (\tilde{M}P)P$ and it is thus equivalent to assume that*

$$\begin{aligned} X &= M + W \\ M &= c_2 \mu j_1^\mathsf{T} - c_1 \mu j_2^\mathsf{T} \end{aligned}$$

where $\mu \equiv \mu_1 - \mu_2$.

To obtain simpler expressions, we also request that $c_{ui} = c_u c_i + o(1)$ and $c_{li} = c_l c_i + o(1)$, i.e., labelled and unlabelled data from each class are sampled in proportion of the class sizes; note that this occurs with high probability if sampling is performed uniformly at random.

Under these basic assumptions, exploiting advanced random matrix tools summarized in the supplementary material, the following result is obtained.

Theorem 1 (Asymptotic mean and variance)**.** *Define $\alpha_- \equiv (\sqrt{c_u} - \sqrt{c_0})^2$ and $\alpha_+ \equiv (\sqrt{c_u} + \sqrt{c_0})^2$, and let $\alpha \in \mathbb{R} \setminus \{(\alpha_-, \alpha_+) \cup \{0\}\}$. Then, as $n \to \infty$ in such a way that the liminf and limsup of c_0, c_u and c_1 differ from zero and infinity,*

$$\frac{j_i^{(u)\mathsf{T}} f_u}{n_{ui}} - m_i \xrightarrow{\text{a.s.}} 0$$

$$\frac{(f_u - m_i 1_{n_u})^\mathsf{T} D_i^{(u)} (f_u - m_i 1_{n_u})}{n_{ui}} - \sigma_i^2 \xrightarrow{\text{a.s.}} 0$$

where, for $i = 1, 2$,

$$\begin{aligned} m_i &\equiv -\frac{c_l}{c_u} s_i \left(1 - \frac{1}{1 + \frac{c_u c_1 c_2 \|\mu\|^2}{c_0} \frac{\delta}{1+\delta}} \right) \\ \sigma_i^2 &\equiv \frac{s_i^2 c_l^2 c_i^2 \|\mu\|^2 \delta^2}{c_0^2 (1+\delta)^2 - c_u c_0 \delta^2} \frac{1 + \frac{c_u c_1 c_2 \|\mu\|^2}{c_0} \frac{\delta^2}{(1+\delta)^2}}{\left(1 + \frac{c_u c_1 c_2 \|\mu\|^2}{c_0} \frac{\delta}{1+\delta} \right)^2} \\ &\quad + \frac{s_i^2 c_l c_i}{1 - c_i} \frac{\delta^2}{c_0 (1+\delta)^2 - c_u \delta^2} \end{aligned}$$

with δ defined as

$$\delta \equiv -\frac{1}{2} + \frac{c_u - c_0 + \operatorname{sign}(\alpha) \sqrt{(\alpha - \alpha_-)(\alpha - \alpha_+)}}{2\alpha}.$$

Theorem 1 provides asymptotic, almost sure, limits for the *empirical averages and variances* of the labels genuinely belonging to $\mathcal{C}_i$. It is confirmed by Theorem 1 that the problem of score 'flatness' is avoided with the centered kernel regularization.

Conclusion 1 (Distinct classification scores)**.** *As given by Theorem 1, the average score of unlabelled data in the same class is proportional to the value of initial label (s_i for $\mathcal{C}_i$) given to the labelled data in that class. The centered kernel regularization thus yields non-flat solutions of f_u.*

To understand precisely the displays in Theorem 1, a few remarks are in order.

Remark 2 (The parameter δ)**.** *The expert reader shall recognize $\delta = \delta(\alpha)$ as the Stieltjes transform $\int \frac{\mu(dt)}{t - \alpha}$ of (a scaled version of) the celebrated Marčenko–Pastur distribution μ (Marčenko & Pastur, 1967), corresponding here to the limiting eigenvalue distribution of the random matrix $\frac{1}{n} W_u W_u^\mathsf{T}$ at the core of our analysis. In particular, $\delta(\alpha)$ is only defined for α outside the support $[\alpha_-, \alpha_+] \cup \{0\}$ if μ (0 being in the support if $\sqrt{c_u} < \sqrt{c_0}$), is increasing on its definition domain, and is positive on $(-\infty, \alpha_-)$ and negative on (α_+, ∞). Figure 1 depicts the eigenvalue distribution of K_{uu} versus the Marčenko–Pastur distribution.*

Remark 3 (Singularities)**.** *Note that $|m_i|, \sigma_i^2 \to \infty$ if $\delta(\alpha)$ approaches $\delta_0 \equiv -(1 + c_u c_1 c_2 c_0^{-1} \|\mu\|^2)^{-1}$. This singularity is reminiscent of an important phase transition phenomenon related to spiked random matrix models (see e.g., (Baik & Silverstein, 2006; Benaych-Georges & Nadakuditi, 2012)). To characterize those singular values, note that, since $\delta(\alpha)$ is negative and increasing on (α_+, ∞) (and positive on (∞, α_-)), δ_0 is only reached if $\delta_0 > \lim_{\alpha \downarrow \alpha_+} \delta(\alpha) = -(1 + \sqrt{\frac{c_u}{c_0}})^{-1}$, so only if*

$$\|\mu\|^2 > \frac{1}{c_1 c_2} \sqrt{\frac{c_0}{c_u}}.$$

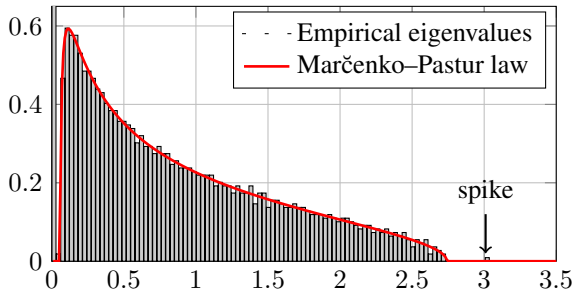

Figure 1. Eigenvalue distribution of K_{uu} versus the (scaled) Marčenko–Pastur law with Stieltjes transform δ, for $c_u = \frac{9}{10}$, $c_0 = \frac{1}{2}$. The value $\|\mu\| = 2.5$ ensures the presence of a leading isolated eigenvalue (spike).

This can be interpreted as a phase transition phenomenon around which m_i suddenly changes sign (moving from, say, $-\infty$ to $+\infty$). In practice, for finite n, p, this transition occurs as α approaches the isolated largest eigenvalue of K_{uu}, call it α^∞ (emphasized in Figure 1), making $K_{uu} - \alpha I_{n_u}$ no longer invertible. Figure 3 provides a picture of this phenomenon for two values of $\|\mu\|$ right below and right above the phase transition threshold.

Also, σ_i^2 displays the denominator $c_0(1+\delta)^2 - c_u\delta^2$. With the same arguments, for $\alpha > \alpha_+$, this term is shown to be positive (tending to zero as $\alpha \downarrow \alpha_+$), and similarly for $\alpha < \alpha_-$. Therefore, this denominator never cancels.

Beyond the empirical average and variance of the class-wise block components of f_u, our main interest here is on the asymptotic classification performance of the proposed algorithm. This is the focus of the subsequent section.

5.2. Asymptotic performance

Theorem 1 does not strictly provide guarantees of classification performances as the entries $[f_u]_j$ for x_j in a given class need not be independent and identically distributed, thereby not allowing for a central limit theorem on the number of correctly assigned labels to strictly apply. Nonetheless, our random matrix experience on similar problems strongly suggests that the $[f_u]_j$ are asymptotically independent and that, as a result, the number of correctly assigned labels likely does satisfy a central limit $\mathcal{N}(m_i, \sigma_i^2)$. As such, we have the following claimed result, illustrated in Figures 3–4 and sustained by simulations in Figure 5.

Claim 1 (Asymptotic algorithm performance)**.** *Under the setting of Theorem 1 and with $s_1 < 0$,*

$$\frac{1}{n_{u1}} \left| \{ [f_u]_j < 0,\ j \in \{1, \ldots, n_{u1}\} \} \right| - \Phi\left(\frac{m_1}{\sigma_1} \right) \xrightarrow{\text{a.s.}} 0$$

with $\Phi(x) \equiv \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{1}{2}t^2} dt$ the Gaussian distribution function, and symmetrically for class $\mathcal{C}_2$.

A few interesting aspects of the asymptotic correct classification rate are worth commenting.

First, recall from Remark 3 that, for $\|\mu\|^2 > \sqrt{c_0/c_u}/(c_1 c_2)$, as $\alpha \to \alpha^\infty$ (the singularity point defined in Remark 3), both m_i and σ_i^2 diverge to infinity. Interestingly, as $\alpha \to \alpha^\infty$, one easily obtains

$$\left(\frac{m_i}{\sigma_i} \right)^2 \to \frac{c_u c_1^2 c_2^2 \|\mu\|^4 - c_0}{c_u c_1^2 c_2^2 \|\mu\|^2 + c_0 c_1 c_2}$$

which is an increasing function of c_u (and thus a decreasing function of c_l). As a consequence, and perhaps surprisingly at first sight, for $\alpha = \alpha^\infty$, semi-supervised learning reduces to unsupervised learning. This in particular explains why in the neighborhood of α^∞ a sudden change in the sign of f_u is observed; this is reminiscent of spectral clustering where classification vectors are defined up to their sign.

On the contrary, for all $\alpha \neq \alpha^\infty$, m_i/σ_i vanishes as $c_l \to 0$, so that $\Phi(m_i/\sigma_i) \to 0.5$, i.e., leads to random guess. More precisely, from the expression of m_i/σ_i, α must be taken close to α^∞ to avoid m_i/σ_i to vanish.

Letting now $c_u \to 0$, $\delta(\alpha) \to -c_0 \alpha^{-1}$ so that, for $\alpha > c_0$,

$$\left(\frac{m_i}{\sigma_i} \right)^2 \to \frac{c_1^2 c_2^2 (1 - c_i) \|\mu\|^4}{c_i (c_0 + c_1 c_2 \|\mu\|^2)}$$

which is independent of α. This is to be expected as K_{uu} is then much less informative than K_{ul}, thereby letting the choice of α in the resolvent $(K_{uu} - \alpha I_{n_u})^{-1}$ asymptotically irrelevant.

More interestingly, if now we let $n_u \to \infty$ while maintaining all other parameters fixed (this thus implies both $c_0 \to 0$ and $c_u \to 1$ in a jointly controlled manner), α^∞ is always eventually defined and it can be shown that, by letting $\alpha \to \alpha^\infty$, up to a large constant, f_u is essentially equal to the dominant eigenvector of K_{uu}, thereby again retrieving spectral clustering performance. Similarly, letting $n_l \to \infty$ with all other parameters fixed $\alpha \to \infty$ leads the algorithm to behave similar to supervised learning.

An important consequence is thus that an increase in either labelled or unlabelled datasets leads to the growth of classification accuracy (since both supervised and unsupervised learning do improve with more data), as displayed in Figure 2. Interestingly, in the same figure, we further observe that when dealing with a difficult learning task (i.e., small $\|\mu\|$), including labelled data (even of small quantity) in the learning process has important value; this further reinforces the need to ensure a good balance between labelled and unlabelled data by setting a proper α.

The above discussion leads to the following two key conclusions supporting the motivation in Section 3.

Conclusion 2 (Learning with labelled and unlabelled data)**.** *The performance of the centered kernel approach is consistently improved by the augmentation of both labelled and unlabelled data, serving thereby the purpose of semi-supervised learning.*

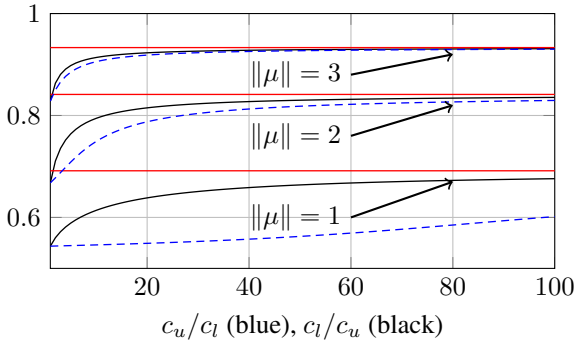

Figure 2. Correct classification rate, at optimal α, as a function of (i) n_u for fixed $p/n_l = 5$ (blue) and (ii) n_l for fixed $p/n_u = 5$ (black); $c_1 = c_2 = \frac{1}{2}$; different values for $\|\mu\|$. Comparison to optimal Neyman–Pearson performance for known μ (in red).

Conclusion 3 (Consistency with respect to spectral clustering)**.** *The centered kernel method consistently outperforms spectral clustering (unsupervised graph-based learning technique). In fact, it retrieves spectral clustering performance with $\alpha \to \alpha^{\infty}$.*

We conclude our theoretical analysis with a last important remark of notable importance for future improvements of the proposed method.

Remark 4 (Bias in f_u)**.** *In the previous analysis, the coefficient $(1 - c_i)$ in the limiting expression of m_i/σ_i suggests that a bias is induced in the output vector entries $[f_u]_j$ which may call for the choice of a better threshold than zero in the decision rule $[f_u]_j \lessgtr 0$. This threshold in general depends in a much elaborate fashion on the parameters c_u, c_i, etc. By the previous analysis, if c_0 is small, an appropriate threshold choice is $(c_2 - c_1)/2$ (if $\mathrm{sign}(s_1) < 0$).*

6. Conclusion

The article has provided a new method for semi-supervised learning, based on an appropriate centralization and scaling of historical kernel-based approaches, motivated by recent random matrix theory advances in machine learning. Beyond avoiding the "flatness of scores" issue popular in the literature, the method also better accounts for the unlabelled data which improve the classification performance as their number grows (unlike most previous schemes).

Future works will further develop the theoretical analysis, notably by extending the classification model to a more generic Gaussian mixture model and to a larger class of kernels. Particular attention will be steered towards an online estimation of the optimal regularization α parameter, as well as towards the behavior (and possible resulting extensions) of the semi-supervised learning scheme in the presence of multiple outlying spikes which naturally occur in more advanced statistical models.

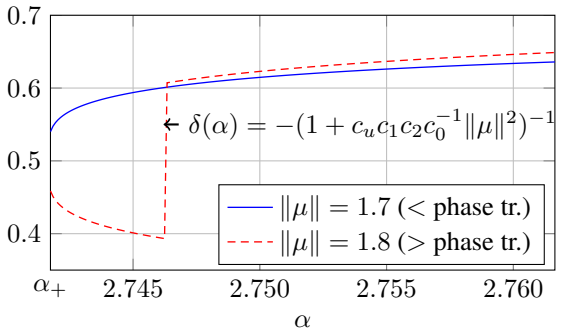

Figure 3. Asymptotic correct classification probability $\Phi\left(\frac{m_1}{\sigma_1}\right)$ as a function of α for $c_u = \frac{9}{10}$, $c_0 = \frac{1}{2}$, $c_1 = \frac{1}{2}$, two different values of $\|\mu\|$, below and above phase transition.

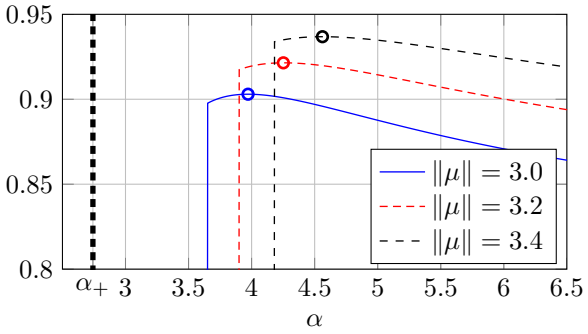

Figure 4. Asymptotic correct classification probability $\Phi\left(\frac{m_1}{\sigma_1}\right)$ as a function of α for $c_u = \frac{9}{10}$, $c_0 = \frac{1}{2}$, $c_1 = \frac{1}{2}$, three different values of $\|\mu\|$. Optimal values marked in circles.

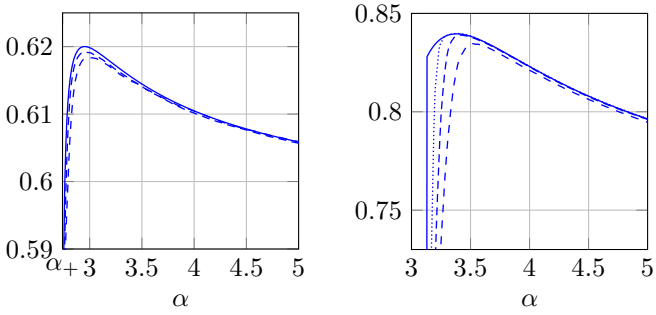

Figure 5. Theory (solid) versus practice (dashed; from right to left: $n = 400, 1000, 4000$): correct classification probability as a function of α for $c_u = \frac{9}{10}$, $c_0 = \frac{1}{2}$, $c_1 = \frac{1}{2}$, and **left**: $\|\mu\| = 1.5$ (below phase transition); **right**: $\|\mu\| = 2.5$ (above phase transition). Different values of n.

References

Avrachenkov, Konstantin, Gonçalves, Paulo, Mishenin, Alexey, and Sokol, Marina. Generalized optimization framework for graph-based semi-supervised learning. In Proceedings of SIAM Conference on Data Mining (SDM 2012), volume 9. SIAM, 2012.

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

Belkin, Mikhail and Niyogi, Partha. Using manifold stucture for partially labeled classification. In Advances in neural information processing systems, pp. 953–960, 2003.

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

Bridle, Nick and Zhu, Xiaojin. p-voltages: Laplacian regularization for semi-supervised learning on high-dimensional data. In Eleventh Workshop on Mining and Learning with Graphs (MLG2013), 2013.

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

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

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

El Karoui, N. The spectrum of kernel random matrices. The Annals of Statistics, 38(1):1–50, 2010.

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

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

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

Marčenko, V. A. and Pastur, L. A. Distribution of eigenvalues for some sets of random matrices. Math USSR-Sbornik, 1(4):457–483, 1967.

Nadler, Boaz, Srebro, Nathan, and Zhou, Xueyuan. Semi-supervised learning with the graph laplacian: The limit of infinite unlabelled data. In Proceedings of the 22nd International Conference on Neural Information Processing Systems, pp. 1330–1338. Curran Associates Inc., 2009.

Scholkopf, Bernhard and Smola, Alexander J. Learning with kernels: support vector machines, regularization, optimization, and beyond. MIT press, 2001.

Smola, Alexander J and Kondor, Risi. Kernels and regularization on graphs. In Learning theory and kernel machines, pp. 144–158. Springer, 2003.

Stallkamp, Johannes, Schlipsing, Marc, Salmen, Jan, and Igel, Christian. Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition. Neural networks, 32:323–332, 2012.

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

Zhou, Denny, Bousquet, Olivier, Lal, Thomas N, Weston, Jason, and Schölkopf, Bernhard. Learning with local and global consistency. In Advances in neural information processing systems, pp. 321–328, 2004.

Zhou, Xueyuan and Belkin, Mikhail. Semi-supervised learning by higher order regularization. In Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, pp. 892–900, 2011.

Zhu, X., Ghahramani, Z., and Lafferty, J. Semi-supervised learning using gaussian fields and harmonic functions. In ICML, volume 3, pp. 912–919, 2003a.

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

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

Supplementary material to "Random Matrix-Inspired Improved Semi-Supervised Learning on Graphs"

February 9, 2018

Sketch of proof for Theorem 1

Without generality restriction, we focus here on the empirical means and covariances for data in class $\mathcal{C}_1$.

1 Limiting means

We start by observing that

$$f_u = -\left([I_{n_u}\ 0]K\begin{bmatrix}I_{n_u}\\0\end{bmatrix} - \alpha I_{n_u}\right)^{-1}[I_{n_u}\ 0]K\begin{bmatrix}0\\I_{n_l}\end{bmatrix} f_l.$$

Recalling the definition of K, and with the property $(ZZ^T+I)^{-1}Z = Z(Z^TZ+I)^{-1}$, this is

$$f_u = -\begin{bmatrix}I & 0\end{bmatrix}\frac{1}{n}PX^T\left(\frac{1}{n}XP\begin{bmatrix}I & 0\\0 & 0\end{bmatrix}PX^T - \alpha I\right)^{-1} \times XP\begin{bmatrix}0\\I\end{bmatrix} f_l.$$

As a consequence, to study the empirical means $\frac{1}{n_{u1}} j_i^{(u)T} f_u$, we need to evaluate a bilinear form of the type

$$a^\intercal \frac{1}{n}X^T\left(\frac{1}{n}XP\begin{bmatrix}I & 0\\0 & 0\end{bmatrix}PX^T - \alpha I\right)^{-1} Xb$$

for $a \in \mathbb{R}^p$, $b \in \mathbb{R}^n$ deterministic vectors of bounded norm (precisely here, for $a = \sqrt{n}(-\frac{j_{u1}}{n_{u1}} + \frac{1_n}{n})$ and $b = \frac{1}{\sqrt{n}}(s_1 j_{l1} + s_2 j_{l2})$).

To this end, we decompose X as $X = M + W$ with $W = [W_u W_l]$ and we isolate the rank-1 contributions due to μ, 1_n (in P) and the various '$j_\cdot$' vectors.

After careful bookkeeping, we find that

$$\frac{1}{n}XP\begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix}PX^T = \frac{1}{n}W_uW_u^T + UNU^T$$

where

$$U = \begin{bmatrix} \mu & \frac{1}{n}Wj_{u1} & \frac{1}{n}Wj_{u2} & \frac{1}{n}Wj_l \end{bmatrix}$$
$$N = \begin{bmatrix} c_uc_1c_2 & c_2 & -c_1 & 0 \\ c_2 & -1-c_l & -1-c_l & -c_l \\ -c_1 & -1-c_l & -1-c_l & -c_l \\ 0 & -c_l & -c_l & c_u \end{bmatrix}$$

This expression isolates the Gram matrix $W_uW_u^T$, W_u with i.i.d. $\mathcal{N}(0,1)$ entries, from a rank-4 matrix involving the deterministic parameters of the model, thereby leading to a spiked random matrix model akin to [1, 3]. From there, invoking Woodbury's identity, we next have

$$\begin{aligned} &\left(\frac{1}{n}XP\begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix}PX^T - \alpha I\right)^{-1} \\ &= Q - QUN(I_4 + U^TQUN)^{-1}UQ \end{aligned} \tag{1}$$

with the notation $Q \equiv (\frac{1}{n}W_uW_u^T - \alpha I_p)^{-1}$, i.e., the much-studied resolvent of $\frac{1}{n}W_uW_u^T$ [4, 5].

In this expression, the size-4 inner matrix $N(I_4 + U^TQUN)^{-1}$ has a deterministic limit, which needs be evaluated. To this end, one needs to figure the limit of U^TQU. Standard random matrix results, e.g., [2], provide the following so-called deterministic equivalents: denoting $\tilde{Q} \equiv (\frac{1}{n}W_u^TW_u - \alpha I_{n_u})^{-1}$,

$$\begin{aligned} Q \leftrightarrow \bar{Q} &\equiv -\frac{1+\delta}{\alpha(1+\delta) - c_u}I_p \\ \tilde{Q} \leftrightarrow \bar{\tilde{Q}} &\equiv -\frac{1}{\alpha(1+\delta)}I_{n_u} \end{aligned}$$

with the notation '$X \leftrightarrow Y$' standing for the fact that, for all a, b, A bounded, $a^T(X-Y)b \xrightarrow{\text{a.s.}} 0$ and $\frac{1}{n}\operatorname{tr} A(X-Y) \xrightarrow{\text{a.s.}} 0$. Here δ is the unique real solution to $\delta = -\frac{c_0(1+\delta)}{\alpha(1+\delta)-c_u}$ defining an increasing function of α, which has the explicit form given in Theorem 1.

With these results, along with classical algebraic arguments, we find that, almost surely,

$$\begin{aligned} U^TQU &= \bar{D} + o(1) \\ &\equiv \begin{bmatrix} \mu^T\bar{Q}\mu & & & \\ & \frac{c_uc_1\delta}{1+\delta} & & \\ & & \frac{c_uc_2\delta}{1+\delta} & \\ & & & c_l\delta \end{bmatrix} + o(1) \end{aligned}$$

thereby allowing, after some cumbersome calculus, for the evaluation of the limit $N(I_4+\bar{D}N)^{-1}$ of $N(I_4+U^TQUN)^{-1}$. The same deterministic equivalent method as for the study of U^TQU next provides the limiting characterization of the bilinear forms $\frac{1}{n}a^TX^TQXb$ and $\frac{1}{\sqrt{n}}a^TX^TQU$ appearing in the expanded expression of $\frac{1}{n_{u1}}j_1^{(u)T}f_u$ based on (1). It then remains to take all required sums and products to finally reach the limiting expression for the empirical mean provided in Theorem 1.

2 Limiting variances

To evaluate the class-wise limiting variances (here for class $\mathcal{C}_1$), we need to assess the quantity

$$\begin{aligned}&\frac{1}{n_{u1}}f_u^TD_1^{(u)}f_u\\&=\frac{b^T}{n_{u1}n^2}X^T\left(\frac{1}{n}XP\begin{bmatrix}I & 0\\0 & 0\end{bmatrix}PX^T-\alpha I\right)^{-1}\\&\times XPD_{u1}PX^T\left(\frac{1}{n}XP\begin{bmatrix}I & 0\\0 & 0\end{bmatrix}PX^T-\alpha I\right)^{-1}Xb\end{aligned}$$

where $b=\frac{1}{\sqrt{n}}(s_1j_{l1}+s_2j_{l2})$ as above, and

$$D_{u1}=\begin{bmatrix}I_{n_{u1}} & 0\\0 & 0_{n_{u2}+n_l}\end{bmatrix},\ D_1^{(u)}=\begin{bmatrix}I_{n_{u1}} & 0\\0 & 0_{n_{u2}}\end{bmatrix}.$$

Performing the same expansion based on (1) as above, we need the following additional deterministic equivalents, found notably in [2]:

$$\begin{aligned}\frac{1}{n}QW_uD_1^{(u)}W_u^TQ&\leftrightarrow\frac{1}{c_0}\frac{c_uc_1\delta^2}{c_0(1+\delta^2)-c_u\delta^2}I_p\\\frac{1}{n^2}W_u^TQW_uD_1^{(u)}W_u^TQW_u&\leftrightarrow\frac{\delta^2}{(1+\delta)^2}D_1^{(u)}\\&\quad+\frac{c_uc_1\delta^2}{c_0(1+\delta)^2-c_u\delta^2}I_{n_u}.\end{aligned}$$

The result is finally obtained through a quite tedious, yet straightforward application of these results.

References

[1] 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.

[2] F. Benaych-Georges and R. Couillet. Spectral analysis of the gram matrix of mixture models. ESAIM: Probability and Statistics, 20:217–237, 2016.

[3] 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.

[4] V. A. Marčenko and L. A. Pastur. Distribution of eigenvalues for some sets of random matrices. Math USSR-Sbornik, 1(4):457–483, 1967.

[5] J. W. Silverstein and Z. D. Bai. On the empirical distribution of eigenvalues of a class of large dimensional random matrices. Journal of Multivariate Analysis, 54(2):175–192, 1995.