RANDOM MATRIX ANALYSIS TO BALANCE BETWEEN SUPERVISED AND UNSUPERVISED LEARNING UNDER THE LOW DENSITY SEPARATION ASSUMPTION

Anonymous authors
Paper under double-blind review

ABSTRACT

We propose a theoretical framework to analyze semi-supervised classification under the low density separation assumption in a high-dimensional regime. In particular, we introduce QLDS, a linear classification model, where the low density separation assumption is implemented via quadratic margin maximization. The algorithm has an explicit solution with rich theoretical properties, and we show that particular cases of our algorithm are the least-square support vector machine in the supervised case, the spectral clustering in the fully unsupervised regime, and a class of semi-supervised graph-based approaches. As such, QLDS establishes a smooth bridge between these supervised and unsupervised learning methods. Using recent advances in the random matrix theory, we formally derive a theoretical evaluation of the classification error in the asymptotic regime. As an application, we derive a hyperparameter selection policy that finds the best balance between the supervised and the unsupervised terms of our learning criterion. Finally, we provide extensive illustrations of our framework, as well as an experimental study on several benchmarks to demonstrate that QLDS, while being computationally more efficient, improves over cross-validation for hyperparameter selection, indicating a high promise of the usage of random matrix theory for semi-supervised model selection.

1 INTRODUCTION

Semi-supervised learning (SSL, Chapelle et al., 2010; van Engelen and Hoos, 2019) aims to learn using both labeled and unlabeled data at once. This machine learning approach received a lot of attention over the past decade due to its relevance to many real-world applications, where the annotation of data is costly and performed manually (Imran et al., 2020), while the data acquisition is cheap and may result in an abundance of unlabeled data (Fergus et al., 2009). As such, semi-supervised learning could be seen as a learning framework that lies in between the supervised and the unsupervised settings, where the former occurs when all the data is labeled, and the latter is restored when only unlabeled data is available. Generally, a semi-supervised algorithm is expected to outperform its supervised counterpart trained only on labeled data by efficiently extracting the information valuable to the prediction task from unlabeled examples.

In practice, integration of unlabeled observations to the learning process does not always affect the performance (Singh et al., 2008), since the marginal data distribution $p(\mathbf{x})$ must contain information on the prediction task $p(y|\mathbf{x})$. Consequently, most semi-supervised approaches rely on specific assumptions about how $p(\mathbf{x})$ and $p(y|\mathbf{x})$ are linked with each other. It is principally assumed that examples *similar* to each other tend to share the same class labels (van Engelen and Hoos, 2019), and implementation of this assumption results in different families of semi-supervised learning models. The first approaches aim to capture the intrinsic geometry of the data using a graph Laplacian (Chong et al., 2020; Song et al., 2022) and suppose that high-dimensional data points with the same label lie on the same low-dimensional *manifold* (Belkin and Niyogi, 2004). Another family of semi-supervised algorithms suggests that examples from a dense region belong to the same class. While some methods explicitly look for such regions by relying on a clustering algorithm (Rigollet, 2007; Peikari et al., 2018), another idea is to directly restrict the classification model to have a decision boundary that only passes through low density regions. This latter approach is said to rely on the *Low Density*

Separation (LDS) assumption (Chapelle and Zien, 2005; van Engelen and Hoos, 2019), and it has been widely used in practice in recent decades, combined with the support vector machine (Bennett and Demiriz, 1998; Joachims, 1999), ensemble methods (d'Alché-Buc et al., 2001; Feofanov et al., 2019) and deep learning methods (Sajjadi et al., 2016; Berthelot et al., 2019).

Despite its popularity, the study of the low density separation assumption still has many open questions. First, there is a deficiency of works devoted to theoretical analysis of the algorithm's performance under this assumption, and most approaches focus on the methodological part (van Engelen and Hoos, 2019). Second, in real applications, it always remains unclear how a semi-supervised algorithm should balance the importance of the labeled and the unlabeled examples in order to not degrade the performance with respect to supervised and unsupervised baselines. This implies in particular that the hyperparameter selection for a semi-supervised classification model is crucial, and it is known that using the cross-validation for model selection may be suboptimal in the semi-supervised case due to the lack of labeled examples (Madani et al., 2005).

Motivated by the aforementioned reasons, this paper proposes a framework to analyze semi-supervised classification under the low density separation assumption using the power of the random matrix theory (Paul and Aue, 2014; Marchenko and Pastur, 1967). We consider a simple yet insightful quadratic margin maximization problem, QLDS, that seeks for an optimal balance between the labeled part represented by the Least Square Support Vector Machine (LS-SVM, Suykens and Vandewalle, 1999) and the unlabeled part represented by the spectral clustering (Ng et al., 2001). In addition, the considered algorithm recovers the graph-based approach proposed by Mai and Couillet (2021) as a particular case.

The main contributions of this paper may be summarized as follows:

- We propose a large dimensional analysis of QLDS and derive a theoretical evaluation of the classification error in the asymptotic regime under the data concentration assumption (Louart and Couillet, 2018). The results allow a strong understanding of the interplay between the data statistics and the hyperparameters of the model.
- Based on the proposed theoretical result, we propose a hyperparameter selection approach to optimally balance the supervised and unsupervised term of QLDS. We empirically validate this approach on synthetic and real-world data showing that it outperforms a hyperparameter selection by the cross-validation both in terms of performance and running time.

The remainder of the article is structured as follows. In Section 2, we review the related work. Section 3 introduces the semi-supervised framework as well as the optimization problem of QLDS. Under mild conditions on the data distribution, Section 4 provides the large dimensional analysis of the proposed algorithm along with several insights and discusses its application for hyperparameter selection. Section 5 provides various numerical experiments to corroborate the pertinence of the theoretical analysis and to hyperparameter selection policy. Section 6 concludes the article.

2 Related Work

LDS in Semi-supervised Learning. Formally introduced by Chapelle and Zien (2005), the LDS assumption imposes the optimal class boundary to lie in a low density region. This assumption is usually implemented by margin maximization, which underlies either explicitly or implicitly many semi-supervised algorithms such as the Transductive SVM (TSVM) (Joachims, 1999; Ding et al., 2017), self-training (Tür et al., 2005; Feofanov et al., 2019) or entropy minimization approaches (Grandvalet and Bengio, 2004; Sajjadi et al., 2016). As the margin's signs for unlabeled data are unknown, various unsigned alternatives have been proposed (d'Alché-Buc et al., 2001; Grandvalet and Bengio, 2004), where the classical approach is to consider the margin's absolute value (Joachims, 1999; Amini et al., 2008). In practice, the latter is usually replaced by an exponential surrogate function for gradient-based optimization of TSVM (Chapelle and Zien, 2005; Gieseke et al., 2014). In this paper, we will consider TSVM with the quadratic margin that is both differentiable and convex, which allows us to perform theoretical analysis and obtain a graph-based semi-supervised learning as a particular case. A similar framework of the quadratic margins was considered by Belkin et al. (2006) whose work considered a more general case with a kernel-based SVM and the Laplacian matrix integrated to the objective, for which they proved a Representer theorem. While our work focuses on explicitly deriving a theoretical expression of the classification error, their paper may complement

us from the algorithmic point of view showing a direct extension of QLDS to a non-linear case. It is important to mention other theoretical studies of approaches based on the low density separation, including upper-bounds of the classification error of TSVM (Derbeko et al., 2004; Wang et al., 2007) and analysis of the self-training algorithm (Feofanov et al., 2021; Zhang et al., 2022).

Graph-based Semi-Supervised Learning. The principle of a graph-based approach is to 1) build a suitable graph with all the labeled and the unlabeled examples as the nodes connected by the weighted edges measuring the pairwise similarities (graph construction step), 2) search for a function f over the graph that is close as possible to the given labels, and that is smooth on the entire constructed graph (label inference step). The graph structure can be naturally used as a reflection for the manifold assumption in SSL that suggests that samples located near to each other on a low-dimensional manifold should share similar labels. Among the graph construction methods, the K-nearest neighbor (KNN) graph (Ozaki et al., 2011; Vega-Oliveros et al., 2014) and b-Matching methods (Jebara et al., 2009; Dhillon et al., 2010), along with their extensions, are the most popular ones. Several extensions have considered labeled samples as prior knowledge to refine the generated graph (Rohban and Rabiee, 2012; Berton and Lopes, 2014). Depending on the particular choice of loss functions, the label inference methods can be divided in label propagation approaches (Xiaojin and Zoubin, 2002; Zhou et al., 2003), manifold regularization (Belkin et al., 2006; Xu et al., 2010), Poisson learning (Calder et al., 2020) and deformed Laplacian regularization (Gong et al., 2015). Recently, Mai and Couillet (2021) proposed a theoretical analysis of a unified framework for label inference in a graph that encompasses label propagation, manifold, and Laplacian regularization as special cases. In this paper, we recover (Mai and Couillet, 2021) as a special case of QLDS.

Large Dimensional Analysis for Machine Learning. Recently, Random Matrix Theory (RMT) has received particular attention in the machine learning community for studying the asymptotic performance in a regime when the dimension is of the same order of magnitude as the sample size. Recent advances include analysis of the linear discriminant (Niyazi et al., 2021), spectral clustering (Couillet and Benaych-Georges, 2016), least square SVM (Liao and Couillet, 2019), graph-based semi-supervised learning (Mai and Couillet, 2021). In this paper, we show that theoretical findings of the last three aforementioned works are recovered from our theoretical analysis of QLDS provided in Section 4. We derive our theoretical results under the assumption that observations follow a vector-concentration inequality (Louart and Couillet, 2018), which can be particularly interesting for deep learning representations that preserve concentration property (Seddik et al., 2020). It is interesting to mention that a number of machine learning algorithms have been theoretically analyzed using methods from theoretical physics, especially glassy physics (Agliari et al., 2020; Carleo et al., 2019; Loureiro et al., 2021; Cui et al., 2021; d'Ascoli et al., 2020). To continue with physical statistics-based methods, we highlight (Lelarge and Miolane, 2019) which derived Asymptotic Bayes risk using theoretical information and a replica method. Although statistical physics and RMT-based approaches share the same objectives, the techniques used and the interpretations make them two different but complementary methods. To the best of our knowledge, we are not aware of any analysis of the algorithm studied in this paper using a statistical physics approach, which we believe is however possible and could be an interesting future work. For completeness, let us also mention the works based on the Convex Gaussian MinMax Theorem (Thrampoulidis et al., 2015; 2016) that allows the analysis of many machine learning algorithms but is mathematically different from the approach used in this paper.

3 Framework

Notations Matrices will be represented by bold capital letters (*e.g.,* matrix $\mathbf{A}$). Vectors will be represented in bold minuscule letters (*e.g.,* vector $\mathbf{v}$) and scalars will be represented without bold letters (*e.g.,* variable a). The *canonical vector* of size n is denoted by $\mathbf{e}_m^{[n]} \in \mathbb{R}^n$, $1 \leq m \leq n$, where the i-th element is 1 if $i = m$, and 0 otherwise. The diagonal matrix with diagonal $\mathbf{x}$ and 0 elsewhere is denoted by $\mathcal{D}_{\mathbf{x}}$, while $A_{i:}$ denotes the i-th line of the matrix A.

Semi-supervised Setting We consider binary classification problems, where an observation $\mathbf{x} \in \mathbb{R}^d$ is described by d features and belongs either to the class $\mathcal{C}_1$ with a label $y = -1$ or to the class $\mathcal{C}_2$ with a label $y = +1$. We assume that training data consists of n_l labeled examples $(\mathbf{X}_\ell, \mathbf{y}_\ell) =$

$(\mathbf{x}_i, y_i)_{i=1}^{n_\ell} \in \mathbb{R}^{d\times n_\ell} \times \{-1,+1\}^{n_\ell}$ and n_u unlabeled examples $\mathbf{X}_u = (\mathbf{x}_i)_{i=n_\ell+1}^{n_\ell+n_u} \in \mathbb{R}^{d\times n_u}$ given without labels. Following the transductive setting (Vapnik, 1982), we formulate the goal of semi-supervised learning as to learn a classification model $\mathbb{R}^d \to \{-1,+1\}$ that yields the minimal error on the unlabeled data $\mathbf{X}_u$. For convenience, we denote the concatenation of labeled and unlabeled observations by $\mathbf{X} = [\mathbf{X}_\ell, \mathbf{X}_u]$. For each class $\mathcal{C}_j,\ j \in \{1,2\}$, we denote the observations from this class as $\mathbf{X}_\ell^{(j)} = [\mathbf{x}_1^{(j)}, \ldots, \mathbf{x}_{n_{\ell j}}^{(j)}]$, where $\mathbf{X}_\ell = [\mathbf{X}_\ell^{(1)}, \mathbf{X}_\ell^{(2)}]$ and $n_{\ell 1}+n_{\ell 2}=n_\ell$. The same convention is used for the unlabeled data $\mathbf{X}_u$. By $n_j = n_{\ell j}+n_{uj}$ we denote the total number of samples in class $\mathcal{C}_j,\ j \in \{1,2\}$.

QLDS Based on the training set $[\mathbf{X}_\ell, \mathbf{X}_u]$, we seek for a separating hyperplane (linear decision boundary) $\boldsymbol{\omega}^\star$ that is a solution of the following optimization problem:

$$\boldsymbol{\omega}^\star = \arg\min_{\boldsymbol{\omega}} \underbrace{\frac{\alpha_\ell}{2}\sum_{i=1}^{n_\ell}\left(y_i - \frac{\mathbf{x}_i^\top \boldsymbol{\omega}}{\sqrt{n}}\right)^2}_{\text{label fidelity term}} - \underbrace{\frac{\alpha_u}{2}\sum_{i=n_\ell+1}^{n_\ell+n_u}\left(\boldsymbol{\omega}^\top \frac{\mathbf{x}_i}{\sqrt{n}}\right)^2}_{\text{low density separation}} + \underbrace{\frac{\lambda}{2}\|\boldsymbol{\omega}\|^2}_{\text{regularization}} . \tag{1}$$

The first term is the label fidelity term that involves the labeled data only, and it represents the classical least-square loss used in the LS-SVM. The second term implements the low density separation regularization by maximizing the square of the margin of each unlabeled example, thereby pushing the decision boundary away from the unlabeled points. The third term is the classical Tikhonov regularization although we do fix λ to the maximum eigenvalue of $\mathbf{X} = [\mathbf{X}_\ell, \mathbf{X}_u]$ (see Appendix C.1). The first two terms are considered up to a $(1/\sqrt{n})$ factor in order to ease the notations of the theoretical derivations of Section 4.

Note that the label fidelity term can be alternatively represented by the hinge loss or the log-loss, which slightly alters the overall behavior of the algorithm. Our choice of the least square loss is primarily motivated by the possibility of obtaining more explicit, tractable and insightful results, let alone numerically cheaper implementation. As a matter of fact, by a now well-established universality argument in the large dimensional statistics, it has been shown by closely related works (Mai and Liao, 2019; Tiomoko et al., 2020) that quadratic cost functions are asymptotically optimal (as the data dimension and sample size increase) and uniformly outperform alternative costs such as SVM or logistic approaches, even in a classification setting (the proof is obtained in the precise large dimensional setting which we consider in this paper). Table 4 in Appendix summarizes the classification error by using three losses for labelled parts (hinge, logistic, and quadratic) and two losses for unlabelled parts (quadratic and absolute value). This table shows that the selection of losses presented in the article has a competitive performance.

The optimization problem in Equation (1) is convex and admits a unique solution (all details are given in the supplementary material, Section A) given by

$$\boldsymbol{\omega}^\star = \frac{1}{\sqrt{n}}\left(\lambda \mathbf{I}_d - \alpha_u \frac{\mathbf{X}_u\mathbf{X}_u^\top}{n} + \alpha_\ell \frac{\mathbf{X}_\ell\mathbf{X}_\ell^\top}{n}\right)^{-1} \mathbf{X}_\ell \mathbf{y}_\ell . \tag{2}$$

It is worth remarking that for the fully-supervised case $(\alpha_\ell, \alpha_u) = (1,0)$, we recover the Least Square SVM (Suykens and Vandewalle, 1999). Another extreme case is to take $(\alpha_\ell, \alpha_u) = (0,1)$ that leads to the optimal decision boundary of the graph-based approach proposed by Mai and Couillet (2021) (further denoted by *GB-SSL*). Moreover, if additionally to $(\alpha_\ell, \alpha_u) = (0,1)$ take λ as the maximum eigenvalue of the unlabeled data $(1/n)\,\mathbf{X}_u\mathbf{X}_u^\top$, we recover spectral clustering (See Section B of the supplementary material for a complete derivation).

Given the optimal decision boundary as per Equation (2), the decision score function for any example $\mathbf{x} \in \mathbb{R}^d$ is given as

$$f(\mathbf{x}) = \frac{1}{\sqrt{n}}\boldsymbol{\omega}^{\star\top}\mathbf{x} = \frac{1}{n}\mathbf{y}_\ell^\top \mathbf{X}_\ell^\top \left(\lambda \mathbf{I}_d - \alpha_u \frac{\mathbf{X}_u\mathbf{X}_u^\top}{n} + \alpha_\ell \frac{\mathbf{X}_\ell\mathbf{X}_\ell^\top}{n}\right)^{-1} \mathbf{x} . \tag{3}$$

4 Theoretical Analysis and Its Application

In this section, we theoretically analyze the statistical behavior of QLDS and its decision function $f(\mathbf{x})$. First, we state the assumptions used for theoretical analysis. Then, we present the main results and describe an application for hyperparameter selection.

4.1 ASSUMPTIONS

In the following, we assume the following classical concentration property.

Assumption 1 (Concentration of $\mathcal{D}(\mathbf{X})$) *For two classes $\mathcal{C}_j,\ j \in \{1,2\}$, we assume that all vectors $\mathbf{x}_1^{(j)}, \ldots, \mathbf{x}_{n_j}^{(j)} \in \mathcal{C}_j$ are i.i.d. and in particular $\mathrm{Cov}(\mathbf{x}_i^{(j)}) = \mathbf{I}_d$. Moreover we assume that there exist two constants $C, c > 0$ (independent of n, d) such that, for any 1-Lipschitz function $f : \mathbb{R}^d \to \mathbb{R}$,*

$$\forall t > 0, \qquad \mathbb{P}_{\mathbf{x} \sim \mathcal{D}(\mathbf{X})} \left(|f(\mathbf{x}) - m_{f(\mathbf{x})}| \geq t \right) \leq C e^{-(t/c)^2}$$

where m_Z is a median of the random variable Z.

Assumption 1 notably encompasses the following scenarios: the columns of $\mathbf{X}$ are (a) independent Gaussian random vectors with identity covariance, (b) independent random vectors uniformly distributed on the $\mathbb{R}^d$ sphere of radius $\sqrt{d}$, and, most importantly, (c) any Lipschitz continuous transformation thereof, such as GAN as it has been recently theoretically shown in (Seddik et al., 2020). In the appendix (Section D), we have further explained the concentrated vector assumption and complemented its relevance and generality for the study of machine learning algorithms. In Assumption 1, we only consider identity covariance matrix to keep this presentation simple. The more general case of arbitrary covariance matrix $\mathbf{\Sigma}_j$ is fully derived in the supplementary material, Section C. We should mention that it is convenient to "center" the data $\mathbf{X}$ for the sake of simplicity. This centering operation is performed on the whole data set $\mathbf{X}$ by substracting the global mean from the training points *i.e.,* $\mathbf{X} \leftarrow \mathbf{X} - \mathbb{E}[\mathbf{X}]$. Furthermore, we place ourselves into the following large dimensional regime:

Assumption 2 (High-dimensional asymptotics)) *As $n \to \infty$, we consider the regime where $d = \mathcal{O}(n)$ and assume $d/n \to c_0 > 0$. Furthermore, for $j = 1, 2$, $n_{\ell j}/n \to c_{\ell j}$ and $n_{uj}/n \to c_{uj}$. We denote by $\mathbf{c}_\ell = [c_{\ell 1}, c_{\ell 2}]$ and $\mathbf{c}_u = [c_{u1}, c_{u2}]$.*

This assumption of the commensurable relationship between the number of samples and their dimension corresponds to a realistic regime and differs from classical asymptotic where the number of samples is often assumed to be exponentially larger than the feature size. Note that this chosen asymptotic regime classical in Random Matrix Theory fits most real-life applications and has been successfully applied in telecommunications (Couillet and Debbah, 2011a), finance (Potters et al., 2005) and more recently in machine learning (Liao, 2019; Mai and Couillet, 2021; Tiomoko et al., 2020).

4.2 MAIN RESULTS

We introduce the mean matrix $\mathbf{M} = [\boldsymbol{\mu}_1, \boldsymbol{\mu}_2] \in \mathbb{R}^{d \times 2}$, where $\boldsymbol{\mu}_j = \mathbb{E}_{\mathbf{x} \in \mathcal{C}_j}[\mathbf{x}] \in \mathbb{R}^d$ is the theoretical mean of the class $\mathcal{C}_j, j \in \{1, 2\}$. Further, we define matrices $\mathcal{M}$ and $\mathcal{G}$ that will play an important role at the core formulation of the statistics of $f(\mathbf{x})$.

Definition 1 (Data statistics matrices $\mathcal{M}$ and $\mathcal{G}$) *We define data matrices $\mathcal{M}$ and $\mathcal{G}$ as*

$$\mathcal{M} = \left(\mathcal{D}_{\boldsymbol{\kappa}}^{-1} + \delta \mathbf{M}^\top \mathbf{M} \right)^{-1}, \quad \mathcal{G} = -\left(\frac{n_u}{n(1 - \alpha_u \delta)} + \mathbf{a}^\top \mathbf{d} \right) \delta \mathbf{M}^\top \mathbf{M},$$

where the vectors $\mathbf{a}$, $\mathbf{d}$ and $\boldsymbol{\kappa}$ are the unique positive solution of the following fixed point equations

$$a_j = \frac{c_{\ell j} \alpha_\ell^2}{(1 + \alpha_\ell \delta)^2} + \frac{c_{uj} \alpha_u^2}{(1 - \alpha_u \delta)^2}, \quad d_j = -\frac{\delta^2}{(1 - \alpha_u \delta)^2} \frac{c_0 n_u}{n(1 - c_0 \delta^2 a_j)},$$
$$\delta = \frac{1}{\lambda + \kappa_1 + \kappa_2}, \quad \kappa_j = \frac{c_{\ell j} \alpha_\ell}{1 + \alpha_\ell \delta} - \frac{c_{uj} \alpha_u}{1 - \alpha_u \delta}.$$

The existence of $\mathbf{a}, \mathbf{d}, \boldsymbol{\kappa}, \delta$ are a direct application of (Louart and Couillet, 2018, Proposition 3.8)). These quantities are common in Random Matrix Theory in order to correct large biases in high dimensions (for more details, we refer to the supplementary material, Section D). We are now in position to introduce the asymptotic theoretical analysis of the score $f(\mathbf{x})$ of any unlabeled sample $\mathbf{x}$.

Theorem 1 *Let $\mathbf{X} \in \mathbb{R}^{d \times n}$ be a data set that follows Assumptions 1 and 2. For any $\mathbf{x} \in \mathbf{X}_u$ with $\mathbf{x} \in \mathcal{C}_j$ and $f(\mathbf{x}) = \frac{1}{\sqrt{n}} \boldsymbol{\omega}^{\star\top} \mathbf{x}$ defined by Equation (3), we have almost surely for both classes j*

$$f(\mathbf{x}|\mathbf{x} \in \mathcal{C}_j) - \mathfrak{f}_j \xrightarrow{\text{a.s.}} 0, \quad \textit{where} \quad \mathfrak{f}_j \sim \mathcal{N}\left(m_j, \sigma^2\right).$$

The mean m_j and the variance σ^2 are defined as

$$m_j = \frac{(-1)^j \left(c_{\ell j} - (\mathbf{e}_1^{[2]} - \mathbf{e}_2^{[2]})^\top \mathcal{D}_{\mathbf{c}_\ell} \mathcal{D}_{\boldsymbol{\kappa}}^{-1} \mathcal{M} \mathbf{e}_j^{[2]} \right)}{\kappa_j (1 - \alpha_u \delta)(1 + \alpha_\ell \delta)},$$

$$\sigma^2 = \left(\mathbf{e}_1^{[2]} - \mathbf{e}_2^{[2]} \right)^\top \left(\mathcal{D}_{\mathbf{s}} \mathcal{M} \mathcal{G} \mathcal{M} \mathcal{D}_{\mathbf{s}} + \mathcal{D}_{\mathbf{d}} \mathcal{D}_{\mathbf{c}_\ell} \right) \left(\mathbf{e}_1^{[2]} - \mathbf{e}_2^{[2]} \right),$$

with $\mathbf{s} = [c_{\ell 1}/(\kappa_1(1 + \alpha_\ell \delta)), \;\; c_{\ell 2}/(\kappa_2(1 + \alpha_\ell \delta)]$.

Finally, the theoretical classification error is asymptotically given by

$$\varepsilon_\star = \frac{1}{2} \left(1 - \operatorname{erf} \left(\frac{m_1 - m_2}{2\sqrt{2}\sigma} \right) \right), \tag{4}$$

where $\operatorname{erf}(z) = 2/\sqrt{\pi} \int_0^z e^{-t^2} dt$ *is the Gauss error function.*

A fundamental aspect of Theorem 4 is that the performance of the *large dimensional* (large n, large d) classification problem under consideration merely concentrates into two-dimensional *sufficient statistics*, as all objects defined in the theorem are at most of size 2. All quantities defined in Theorem 4 are a priori known, apart from the proportion of classes in $\mathbf{X}_u$ and the matrix $\mathbf{M}^\top \mathbf{M} \in \mathbb{R}^{2 \times 2}$, whose (i, j)-entries are the inner products $\boldsymbol{\mu}_i^\top \boldsymbol{\mu}_j$ that have to be estimated from data. From a practical perspective, these inner products are easily amenable to fast and efficient estimation as per Proposition 2, requiring a few training data samples.

Proposition 2 (On the estimation of m_j and σ) *The following estimates holds:*

$$\left[\mathbf{M}^\top \mathbf{M}\right]_{ij} = \begin{cases} \left(4/n_{\ell i}^2\right) \mathbf{1}_{n_{\ell i}}^\top \mathbf{X}_{\ell;1}^{(i)\top} \mathbf{X}_{\ell;2}^{(i)} \mathbf{1}_{n_{\ell i}} + \mathcal{O}\left(1/\sqrt{d\, n_{\ell i}}\right) & \textit{if } i = j\,, \\ \left(1/\left(n_{\ell i}\, n_{\ell j}\right)\right) \mathbf{1}_{n_{\ell i}}^\top \mathbf{X}_{\ell}^{(i)\top} \mathbf{X}_{\ell}^{(j)} \mathbf{1}_{n_{\ell j}} + \mathcal{O}\left(\left(d \min\{n_{\ell i}, n_{\ell j}\}\right)^{-\frac{1}{2}} \right) & \textit{otherwise.} \end{cases}$$

with $\mathbf{X}_\ell^{(j)} = [\mathbf{X}_{\ell;1}^{(j)}, \mathbf{X}_{\ell;2}^{(j)}]$ *an even-sized division of* $\mathbf{X}_\ell^{(j)}$.

Note that a single sample (two when $i = j$) per class is sufficient to obtain a consistent estimate for all quantities as long as d is large. In the semi-supervised setting, when only few labeled examples are available, it is thus still possible to estimate $\mathbf{M}^\top \mathbf{M}$. It is important to remark that the convergence rate of the estimation is a quadratic improvement over the convergence rate of the usual central-limit theorem. Finally, to estimate the proportion of classes in the unlabeled set, not known *a priori*, we assume that the distribution of classes to be the same for the labeled and unlabeled data, so that we have $c_{uj} = c_{\ell j} \frac{n_u}{n_\ell}$ for $j \in \{1, 2\}$. We show in the supplementary material (Section E) that this assumption has little impact on the theoretical insights as well as in the experiments.

4.3 Application to hyperparameter selection

Following the discussion in Section 3, we obtain that the theorem allows us to recover the asymptotic performance of the spectral clustering, the graph-based approach GB-SSL of Mai and Couillet (2021) and the LS-SVM (Suykens and Vandewalle, 1999). This generality of the theorem represents an important asset in the unification of some SSL learning schemes. In particular, as the theoretical error can be regarded as a function of α_ℓ and α_u, below we propose to use Equation (4) as a criterion to automatically select α_ℓ and α_u through the grid search over different values. This leads to Algorithm 1. Note that the classification error is invariant to a scaling of λ (see Equation (3)). Thus, we fix the value of λ in our experiments to be $\lambda_{\max}$ with $\lambda_{\max}$ the maximum eigenvalue of $\mathbf{X} = [\mathbf{X}_\ell, \mathbf{X}_u]$, and optimize only α_ℓ and α_u. The fixed value corresponds to the one also proposed in (Mai and Couillet, 2021). We give more details about this choice for λ and its numerical stability in Appendix C.2.

Our proposition to select α_ℓ and α_u by the theorem is motivated by several practical reasons. Firstly, the importance of labeled and unlabeled examples varies, making the graph-based learning more

Table 1: Running time comparison between theory-based hyperparameter selector and cross-validation based with 10 folds, with $n_{\ell j} = n_{uj} = d$ for $j \in \{1, 2\}$.

dimension	5	8	16	32	64	128	256	512	1024	2048
`QLDS(th)`	.01 s	.01 s	.01 s	.01 s	.017 s	.02 s	.04 s	.17 s	.95 s	8.58 s
`QLDS(cv)`	.32 s	.38 s	.43 s	.64 s	1.02 s	2.71 s	14 s	95 s	587 s	4970 s

effective in some cases, and the LS-SVM more effective in the others. By properly choosing α_ℓ and α_u, one can find the best balance between the GB-SSL and LS-SVM. Secondly, the classical approach of selecting hyperparameters by cross-validation suffers from high computational time and prone to bias in the semi-supervised setting due to the scarcity of the labeled set (Madani et al., 2005).

Algorithm 1 QLDS algorithm with optimal selection of α_ℓ and α_u

Input: labeled data $\mathbf{X}_\ell$ and unlabeled data $\mathbf{X}_u$
grid of hyperparameter values $\{(\alpha_\ell^{(t)}, \alpha_u^{(t)})\}_{t=1}^T$
Preprocessing: center data $\mathbf{X} \leftarrow \mathbf{X} - \mathbb{E}[\mathbf{X}]$, where $\mathbf{X} = [\mathbf{X}_\ell, \mathbf{X}_u]$
Output: estimated label $\hat{y} \in \{-1, 1\}$ for each unlabeled example $\mathbf{x} \in \mathbf{X}_u$

estimate inner product $\mathbf{M}^\top \mathbf{M}$ using Proposition 2
choose λ as the maximum eigenvalue of $\frac{1}{n}\mathbf{X}_u \mathbf{X}_u^\top$
for $t = 1, \ldots, T$ grid-search steps **do**
  take $\alpha_\ell^{(t)}$ and $\alpha_u^{(t)}$
  estimate classification error $\varepsilon_\star^{(t)}$ by Theorem 4 with $\alpha_u = \alpha_u^{(t)}$ and $\alpha_\ell = \alpha_\ell^{(t)}$
end for
select $\alpha_u^\star$ and $\alpha_\ell^\star$ by finding t that yields minimal classification error $\varepsilon_\star^{(t)}$
compute the decision score $f(\mathbf{x})$ using Equation (3) with $\alpha_u = \alpha_u^\star$ and $\alpha_\ell = \alpha_\ell^\star$
return label $\hat{y} = \begin{cases} -1 & \text{if } f(\mathbf{x}) < 0 \\ 1 & \text{otherwise} \end{cases}$

Complexity analysis. Algorithm 1 or `QLDS(th)` may be sequentially described as in 1) training of `QLDS`, 2) estimation of $\mathbf{M}^\top \mathbf{M}$, 3) selection of α_ℓ and α_u over the grid. As `QLDS` has an explicit solution, its complexity is equivalent to the computation of the decision scores $f(\mathbf{x})$ which requires solving a system of n linear equations, yielding complexity $\mathcal{O}(n^3)$. The computation of $\mathbf{M}^\top \mathbf{M}$ is of complexity $\mathcal{O}(dn + d)$ (estimation + product). Hyperparameter selection consists of iterating T times the error estimation from Theorem 4 and its complexity is $\mathcal{O}(T)$. Finally, the global complexity of `QLDS(th)` is $\mathcal{O}(n^3)$ in the regime of Assumption 2.

It is important to mention that an alternative way to optimize α_ℓ and α_u is cross-validation (`QLDS(cv)`, which requires optimizing `QLDS` for each candidate $(\alpha_\ell^t, \alpha_u^t)$ for K folds, leading to a complexity of $\mathcal{O}(TKn^3)$. This indicates a clear advantage of using Theorem 4 in terms of time complexity as highlighted in Table 1.

5 Experimental Results

In this section, we illustrate the robustness of the different algorithms and the optimization of the hyperparameters α_ℓ and α_u proposed in the previous section. More specifically, Section 5.1 confirms empirically the robustness of the concentrated random vector assumption on real data by comparing the empirical distribution of the decision score with the theoretical prediction of Theorem 4. Section 5.2 analyzes the performance of QLDS when increasing the number of labeled examples, and Section 5.3 is a benchmark with several baselines for a wide range of real data sets. We perform comparison between the following methods:

- `QLDS(0, 1)` with $\alpha_\ell = 0, \alpha_u = 1$ which stands for the graph-based approach proposed in (Mai and Couillet, 2021);
- `QLDS(1, 0)` with $\alpha_\ell = 1, \alpha_u = 0$ which stands for LS-SVM;

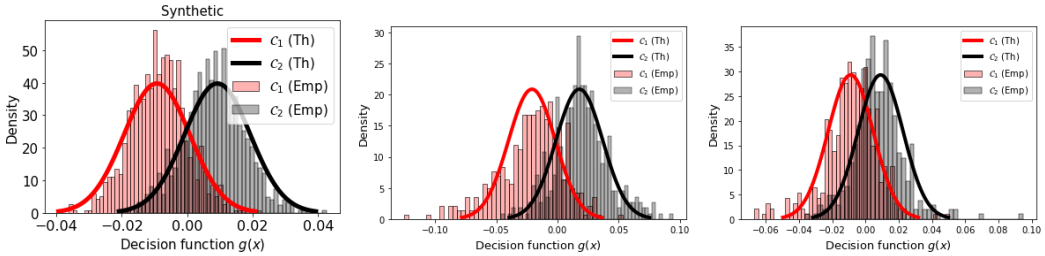

Figure 1: Empirical versus theoretical density of decision score $f(\mathbf{x})$ for (**Left**) Synthetic data set with $d = 100$, $n_{\ell 1} = n_{\ell 2} = 100$, $n_{u1} = n_{u2} = 1\ 000$ (**Center**) Review-kitchen classification $d = 400$, $n_{\ell 1} = n_{\ell 2} = 100$ (**Right**) Review-books classification $d = 400$, $n_{\ell 1} = n_{\ell 2} = 100$. For both review data sets, the empirical histogram is computed using 400 unlabeled samples.

- Self-training with the Least-Square SVM as the base classifier, where the confidence threshold is optimized as proposed by Feofanov et al. (2019) denoted `ST (LS-SVM)`;
- `QLDS (cv)` with model selection of α_ℓ and α_u by the 10-fold cross-validation;
- `QLDS (th)` with model selection of α_ℓ and α_u performed theoretically using Theorem 4;
- `QLDS (or)` Oracle to measure the efficiency of the proposed algorithm: QLDS, where model selection of α_ℓ and α_u is performed on the ground truth (as if the labels for the unlabeled examples would be known). It represents the error-classification lower-bound for the previous approaches.

Through the experimental part, we will use several data sets described as follows (see more details in Section E of the supplementary material):

- `Synthetic`: Gaussian mixture model with $\mathbf{x}_i^{(j)} \sim \mathcal{N}(\boldsymbol{\mu}_j, \mathbf{I}_d)$ with $\boldsymbol{\mu}_1 = -\boldsymbol{\mu}_2$;
- `Amazon Review` data set (McAuley et al., 2015; He and McAuley, 2016) from textual user reviews, positive or negative, on books (`books`), DVDs (`dvd`), electronics (`electronics`), and kitchen (`kitchen`) items respectively. The data is encoded as $d = 400$-dimensional tf-idf feature vectors of bag-of-words unigrams and bigrams;
- `Adult` data set (Kohavi et al., 1996) which consists in predicting whether income exceeds $50\,000$ per year based on census data;
- `Mushrooms` data set from UCI Machine Learning repository (Dua and Graff, 2017) which classifies between poisonous and edible mushrooms based on their physical characteristics;
- `Splice` data set from UCI Machine Learning repository (Dua and Graff, 2017) which aims to recognize two types of splice junctions in DNA sequences.

5.1 Robustness of theoretical analysis to real data

This section illustrates the close fit of the theoretical performance (*i.e.,* Theorem 4) on the synthetic and two real-life data sets. To do so, we compare the empirical decision function represented by the histograms on Figure 1 versus the Gaussian statistics m_j and σ^2 from Theorem 4.

5.2 Analysis of sample size

Figure 2 represents the classification error as a function of the number of labeled examples. The picture shows that the theoretical model selection outperforms the cross-validation scheme and is close to the oracle selection (which uses the ground truth labels). In general, we observe that `QLDS (th)` is very stable in comparison with the cross-validation selector `QLDS (cv)`.

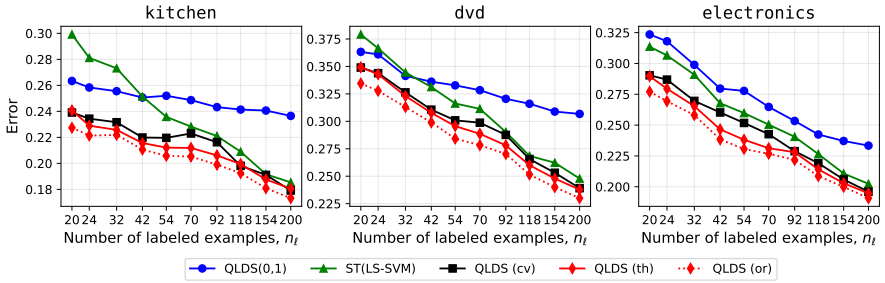

Figure 2: Classification error depending on the number of labeled examples on different data sets. Positive vs. negative review for different products (**Left**) `kitchen` (**Center**) `dvd` and (**Right**) `electronics` with $n_{u1} = n_{u2} = 200$, $d = 400$.

5.3 Comparative performance on several data sets

This section compares through Table 2 the performance obtained by fixing the number of labeled and unlabeled data on several data sets to analyze the performance of the hyperparameter selection and also to validate the theoretical intuitions formulated in this article.

Table 2: The classification error of different methods under consideration on the real benchmark data sets. $^{\downarrow}$ indicates statistically significantly worse performance than the best result (shown in bold), according to the Mann-Whitney U test ($p < 0.01$) (Mann and Whitney, 1947).

Data set	Baselines				Model Selection		Oracle
	QLDS (1,0) (LS-SVM)	QLDS (0,1) (GB-SSL)	QLDS (1,1)	ST (LS-SVM)	QLDS (cv)	QLDS (th)	QLDS (or)
books	$37.47^{\downarrow} \pm 2.25$	26.47 ± 0.72	$49.13^{\downarrow} \pm 0.65$	$35.83^{\downarrow} \pm 2.48$	27.91 ± 3.32	$\mathbf{26.03} \pm 0.79$	25.7 ± 0.93
dvd	$38.33^{\downarrow} \pm 1.72$	29.12 ± 1.35	$49.25^{\downarrow} \pm 0.68$	$36.46^{\downarrow} \pm 1.94$	29.53 ± 3.48	$\mathbf{28.53} \pm 1.33$	26.94 ± 1.47
electronics	$34.15^{\downarrow} \pm 3.25$	19.4 ± 0.29	$48.67^{\downarrow} \pm 1.05$	$31.69^{\downarrow} \pm 3.56$	$20.1^{\downarrow} \pm 1.03$	$\mathbf{19.41} \pm 0.46$	19.11 ± 0.58
kitchen	$32.39^{\downarrow} \pm 3.02$	19.31 ± 0.16	$49.07^{\downarrow} \pm 0.64$	$29.62^{\downarrow} \pm 3.03$	$19.98^{\downarrow} \pm 2.28$	$\mathbf{19.11} \pm 0.32$	18.67 ± 0.43
splice	$39.81^{\downarrow} \pm 2.93$	35.48 ± 0.86	$44.36^{\downarrow} \pm 2.3$	$39.36^{\downarrow} \pm 3.12$	37.02 ± 3.04	$\mathbf{35.35} \pm 1.26$	33.63 ± 1.75
adult	33.35 ± 0.68	$36.28^{\downarrow} \pm 0.06$	32.55 ± 1.47	$35.45^{\downarrow} \pm 0.75$	$\mathbf{32.25} \pm 1.92$	32.88 ± 2.46	31.9 ± 1.74
mushrooms	$6.55^{\downarrow} \pm 2.07$	$11.33^{\downarrow} \pm 0.04$	$33.94^{\downarrow} \pm 10.67$	$6.62^{\downarrow} \pm 2.39$	$\mathbf{2.57} \pm 1.86$	$8.49^{\downarrow} \pm 3.63$	1.75 ± 1.31

The experimental results show that

- QLDS benefits from both labelled and unlabelled data and significantly outperforms `LS-SVM` and `GB-SSL` on datasets 5 and 2 respectively.
- Fine tunning of α_ℓ and α_u provides better results than setting them to the default values.
- Hyperparameter selection using Theorem 4 outperforms or is comparable to cross-validation, at the same time being more robust according to the error's standard deviation.
- There is still room for improvement when we compare `QLDS (or)` with `QLDS (th)`.

6 Concluding Remarks

In this paper, we proposed a theoretical analysis of a simple yet powerful linear semi-supervised classifier that relies on the low density separation assumption. Moreover, our approach builds a bridge between several existing approaches such as the least square support vector machine, the spectral clustering, and graph-based semi-supervised learning. The key approach to our analysis was to use modern large dimensional statistics to quantify the classification error through the data statistics of the decision function. Based on this result, we proposed a hyperparameter selection criterion that demonstrated promising experimental results compared to the time-consuming cross-validation. The proposed theoretical study opens broad perspectives for analysis of the LDS assumption in more challenging settings such as the multi-class classification, the non-linear case, or fully unsupervised domain adaptation.

REFERENCES

Agliari, E., Barra, A., Sollich, P., and Zdeborova, L. (2020). Machine learning and statistical physics: theory, inspiration, application. *Journal of Physics A: Special*, (2020).

Amini, M., Laviolette, F., and Usunier, N. (2008). A transductive bound for the voted classifier with an application to semi-supervised learning. In *Advances in Neural Information Processing Systems*, pages 65–72.

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

Belkin, M., Niyogi, P., and Sindhwani, V. (2006). Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of machine learning research*, 7(11).

Bennett, K. and Demiriz, A. (1998). Semi-supervised support vector machines. *Advances in Neural Information processing systems*, 11.

Berthelot, D., Carlini, N., Goodfellow, I., Papernot, N., Oliver, A., and Raffel, C. A. (2019). Mixmatch: A holistic approach to semi-supervised learning. *Advances in Neural Information Processing Systems*, 32.

Berton, L. and Lopes, A. D. A. (2014). Graph construction based on labeled instances for semi-supervised learning. In *2014 22nd International Conference on Pattern Recognition*, pages 2477–2482. IEEE.

Calder, J., Cook, B., Thorpe, M., and Slepcev, D. (2020). Poisson learning: Graph based semi-supervised learning at very low label rates. In *International Conference on Machine Learning*, pages 1306–1316. PMLR.

Carleo, G., Cirac, I., Cranmer, K., Daudet, L., Schuld, M., Tishby, N., Vogt-Maranto, L., and Zdeborová, L. (2019). Machine learning and the physical sciences. *Rev. Mod. Phys.*, 91:045002.

Chapelle, O., Schölkopf, B., and Zien, A. (2010). *Semi-Supervised Learning*. The MIT Press, 1st edition.

Chapelle, O. and Zien, A. (2005). Semi-supervised classification by low density separation. In *International workshop on artificial intelligence and statistics*, pages 57–64. PMLR.

Chong, Y., Ding, Y., Yan, Q., and Pan, S. (2020). Graph-based semi-supervised learning: A review. *Neurocomputing*, 408:216–230.

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

Couillet, R. and Debbah, M. (2011a). *Random matrix methods for wireless communications*. Cambridge University Press.

Couillet, R. and Debbah, M. (2011b). *Random matrix methods for wireless communications*. Cambridge University Press, New York, NY, USA, first edition.

Cui, H., Loureiro, B., Krzakala, F., and Zdeborová, L. (2021). Generalization error rates in kernel regression: The crossover from the noiseless to noisy regime. *Advances in Neural Information Processing Systems*, 34:10131–10143.

d'Alché-Buc, F., Grandvalet, Y., and Ambroise, C. (2001). Semi-supervised marginboost. *Advances in Neural Information Processing Systems*, 14:553–560.

Derbeko, P., El-Yaniv, R., and Meir, R. (2004). Explicit learning curves for transduction and application to clustering and compression algorithms. *Journal of Artificial Intelligence Research*, 22:117–142.

Dhillon, P. S., Talukdar, P., and Crammer, K. (2010). Learning better data representation using inference-driven metric learning. In *Proceedings of the acl 2010 conference short papers*, pages 377–381.

Ding, S., Zhu, Z., and Zhang, X. (2017). An overview on semi-supervised support vector machine. *Neural Computing and Applications*, 28(5):969–978.

Dua, D. and Graff, C. (2017). UCI machine learning repository.

d'Ascoli, S., Refinetti, M., Biroli, G., and Krzakala, F. (2020). Double trouble in double descent: Bias and variance (s) in the lazy regime. In *International Conference on Machine Learning*, pages 2280–2290. PMLR.

Feofanov, V., Devijver, E., and Amini, M.-R. (2019). Transductive bounds for the multi-class majority vote classifier. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):3566–3573.

Feofanov, V., Devijver, E., and Amini, M.-R. (2021). Multi-class probabilistic bounds for self-learning. *arXiv preprint arXiv:2109.14422*.

Fergus, R., Weiss, Y., and Torralba, A. (2009). Semi-supervised learning in gigantic image collections. *Advances in neural information processing systems*, 22.

Fleury, B., Guédon, O., and Paouris, G. (2007). A stability result for mean width of lp-centroid bodies. *Advances in Mathematics*, 214(2):865–877.

Gieseke, F., Airola, A., Pahikkala, T., and Kramer, O. (2014). Fast and simple gradient-based optimization for semi-supervised support vector machines. *Neurocomputing*, 123:23–32.

Gong, C., Liu, T., Tao, D., Fu, K., Tu, E., and Yang, J. (2015). Deformed graph laplacian for semisupervised learning. *IEEE transactions on neural networks and learning systems*, 26(10):2261–2274.

Grandvalet, Y. and Bengio, Y. (2004). Semi-supervised learning by entropy minimization. *Advances in Neural Information Processing System*, 17.

He, R. and McAuley, J. (2016). Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, page 507–517, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Imran, A.-A.-Z., Huang, C., Tang, H., Fan, W., Xiao, Y., Hao, D., Qian, Z., and Terzopoulos, D. (2020). Partly supervised multi-task learning. In *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 769–774.

Jebara, T., Wang, J., and Chang, S.-F. (2009). Graph construction and b-matching for semi-supervised learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 441–448.

Joachims, T. (1999). Transductive inference for text classification using support vector machines. In *Proceedings of the Sixteenth International Conference on Machine Learning*, ICML '99, pages 200–209, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015*.

Klartag, B. (2007). A central limit theorem for convex sets. *Inventiones mathematicae*, 168(1):91–131.

Kohavi, R. et al. (1996). Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid. In *Kdd*, volume 96, pages 202–207.

Lelarge, M. and Miolane, L. (2019). Asymptotic bayes risk for gaussian mixture in a semi-supervised setting. In *2019 IEEE 8th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, pages 639–643. IEEE.

Liao, Z. (2019). *A random matrix framework for large dimensional machine learning and neural networks*. PhD thesis, Université Paris-Saclay.

Liao, Z. and Couillet, R. (2019). A large dimensional analysis of least squares support vector machines. *IEEE Transactions on Signal Processing*, 67(4):1065–1074.

Louart, C. and Couillet, R. (2018). Concentration of measure and large random matrices with an application to sample covariance matrices. *arXiv preprint arXiv:1805.08295*.

Loureiro, B., Gerbelot, C., Cui, H., Goldt, S., Krzakala, F., Mezard, M., and Zdeborová, L. (2021). Learning curves of generic features maps for realistic datasets with a teacher-student model. *Advances in Neural Information Processing Systems*, 34:18137–18151.

Madani, O., Pennock, D. M., and Flake, G. W. (2005). Co-validation: Using model disagreement on unlabeled data to validate classification algorithms. In *Advances in neural information processing systems*, pages 873–880.

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

Mai, X. and Couillet, R. (2021). Consistent semi-supervised graph regularization for high dimensional data. *J. Mach. Learn. Res.*, 22:94:1–94:48.

Mai, X. and Liao, Z. (2019). High dimensional classification via regularized and unregularized empirical risk minimization: Precise error and optimal loss. *arXiv preprint arXiv:1905.13742*.

Mann, H. B. and Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, 18(1):50–60.

Marchenko, V. A. and Pastur, L. A. (1967). Distribution of eigenvalues for some sets of random matrices. *Matematicheskii Sbornik*, 114(4):507–536.

McAuley, J., Targett, C., Shi, Q., and van den Hengel, A. (2015). Image-based recommendations on styles and substitutes. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '15, page 43–52, New York, NY, USA. Association for Computing Machinery.

Ng, A., Jordan, M., and Weiss, Y. (2001). On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems*, 14.

Niyazi, L. B., Kammoun, A., Dahrouj, H., Alouini, M.-S., and Al-Naffouri, T. (2021). Weight vector tuning and asymptotic analysis of binary linear classifiers. *arXiv preprint arXiv:2110.00567*.

Ozaki, K., Shimbo, M., Komachi, M., and Matsumoto, Y. (2011). Using the mutual k-nearest neighbor graphs for semi-supervised classification on natural language data. In *Proceedings of the fifteenth conference on computational natural language learning*, pages 154–162.

Paul, D. and Aue, A. (2014). Random matrix theory in statistics: A review. *Journal of Statistical Planning and Inference*, 150:1–29.

Peikari, M., Salama, S., Nofech-Mozes, S., and Martel, A. L. (2018). A cluster-then-label semi-supervised learning approach for pathology image classification. *Scientific reports*, 8(1):1–13.

Potters, M., Bouchaud, J.-P., and Laloux, L. (2005). Financial applications of random matrix theory: Old laces and new pieces. *Acta Physica Polonica B*, 36(9):2767.

Rigollet, P. (2007). Generalization error bounds in semi-supervised classification under the cluster assumption. *Journal of Machine Learning Research*, 8(7).

Rohban, M. H. and Rabiee, H. R. (2012). Supervised neighborhood graph construction for semi-supervised classification. *Pattern Recognition*, 45(4):1363–1372.

Sajjadi, M., Javanmardi, M., and Tasdizen, T. (2016). Mutual exclusivity loss for semi-supervised deep learning. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 1908–1912. IEEE.

Seddik, M. E. A., Louart, C., Couillet, R., and Tamaazousti, M. (2021). The unexpected deterministic and universal behavior of large softmax classifiers. In *International Conference on Artificial Intelligence and Statistics*, pages 1045–1053. PMLR.

Seddik, M. E. A., Louart, C., Tamaazousti, M., and Couillet, R. (2020). Random matrix theory proves that deep learning representations of gan-data behave as gaussian mixtures. In *International Conference on Machine Learning*, pages 8573–8582. PMLR.

Singh, A., Nowak, R., and Zhu, J. (2008). Unlabeled data: Now it helps, now it doesn't. *Advances in neural information processing systems*, 21.

Song, Z., Yang, X., Xu, Z., and King, I. (2022). Graph-based semi-supervised learning: A comprehensive review. *IEEE transactions on neural networks and learning systems*, PP.

Suykens, J. A. and Vandewalle, J. (1999). Least squares support vector machine classifiers. *Neural processing letters*, 9(3):293–300.

Thrampoulidis, C., Abbasi, E., and Hassibi, B. (2016). Precise high-dimensional error analysis of regularized m-estimators. In *2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 410–417. IEEE.

Thrampoulidis, C., Oymak, S., and Hassibi, B. (2015). Regularized linear regression: A precise analysis of the estimation error. In *Conference on Learning Theory*, pages 1683–1709. PMLR.

Tiomoko, M., Ali, H. T., and Couillet, R. (2020). Deciphering and optimizing multi-task learning: a random matrix approach. In *International Conference on Learning Representations*.

Tür, G., Hakkani-Tür, D. Z., and Schapire, R. E. (2005). Combining active and semi-supervised learning for spoken language understanding. *Speech Communication*, 45:171–186.

van Engelen, J. E. and Hoos, H. H. (2019). A survey on semi-supervised learning. *Machine Learning*, 109:373–440.

Vapnik, V. (1982). *Estimation of Dependences Based on Empirical Data: Springer Series in Statistics (Springer Series in Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA.

Vega-Oliveros, D. A., Berton, L., Eberle, A. M., de Andrade Lopes, A., and Zhao, L. (2014). Regular graph construction for semi-supervised learning. In *Journal of physics: Conference series*, volume 490. IOP Publishing.

Wang, J., Shen, X., and Pan, W. (2007). On transductive support vector machines. *Contemporary Mathematics*, 443:7–20.

Xiaojin, Z. and Zoubin, G. (2002). Learning from labeled and unlabeled data with label propagation. *Tech. Rep., Technical Report CMU-CALD-02–107, Carnegie Mellon University*.

Xu, Z., King, I., Lyu, M. R.-T., and Jin, R. (2010). Discriminative semi-supervised feature selection via manifold regularization. *IEEE Transactions on Neural networks*, 21(7):1033–1047.

Zhang, S., Wang, M., Liu, S., Chen, P.-Y., and Xiong, J. (2022). How does unlabeled data improve generalization in self-training? a one-hidden-layer theoretical analysis. *arXiv preprint arXiv:2201.08514*.

Zhang, T. and Oles, F. J. (2001). Text categorization based on regularized linear classification methods. *Information retrieval*, 4(1):5–31.

Zhou, D., Bousquet, O., Lal, T., Weston, J., and Schölkopf, B. (2003). Learning with local and global consistency. *Advances in neural information processing systems*, 16.