Random matrices meet machine learning: A large dimensional analysis of LS-SVM

Zhenyu Liao, Romain Couillet

RANDOM MATRICES MEET MACHINE LEARNING: A LARGE DIMENSIONAL ANALYSIS OF LS-SVM

Zhenyu Liao, Romain Couillet

CentraleSupélec, Université Paris-Saclay, Gif-sur-Yvette, France

ABSTRACT

This article proposes a performance analysis of kernel least squares support vector machines (LS-SVMs) based on a random matrix approach, in the regime where both the dimension of data p and their number n grow large at the same rate. Under a two-class Gaussian mixture model for the input data, we prove that the LS-SVM decision function is asymptotically normal with means and covariances shown to depend explicitly on the derivatives of the kernel function. This provides improved understanding along with new insights into the internal workings of SVM-type methods for large datasets.

Index Terms— kernel methods, machine learning, random matrices, support vector machines

1. INTRODUCTION

One of the salient features of the Big Data paradigm lies in handling data which are both numerous and large dimensional – in applications such as computer vision and natural language processing, the number of data n and their dimension p are classically more than hundreds or even thousands. The objective of this article is to investigate the performance of classical non-linear classification methods based on kernel approaches in the large n, p regime. Kernel methods consist in modifying the data vectors $\mathbf{x} \in \mathbb{R}^p$ for some smartly chosen $\boldsymbol{\varphi}$ as $\boldsymbol{\varphi}(\mathbf{x}) \in \mathcal{H}$ with $\mathcal{H}$ some (possibly infinite dimensional) Hilbert space. We shall assume here that the kernel is *radial* in the sense that there exists f such that $\boldsymbol{\varphi}(\mathbf{x})^T \boldsymbol{\varphi}(\mathbf{y}) = f\left(\|\mathbf{x}-\mathbf{y}\|^2/p\right)$.[1] Our focus is on the popular classification method known as kernel support vector machines (SVMs) [1] and more precisely on least-squares kernel SVMs (LS-SVMs) [2] which have the interesting feature of offering an explicit decision function for classification.[2] The performance of SVMs has been widely studied for small dimensional data [3] or under the assumption of linearly independent datasets, in the regime where $p \to \infty$ and n fixed [4]. The regime $n, p \to \infty$ of central interest here however has thus far remained open.

Recent breakthroughs in random matrix theory have allowed one to overtake the theoretical difficulty to evaluate kernel methods posed by the non-linearity of the aforementioned kernel function f [5, 6]. These tools have notably been used to assess the performance of the popular Ng-Weiss-Jordan kernel spectral clustering methods for large datasets [6]. In this article, following up on [6], we provide a performance analysis of LS-SVMs, under a two-class Gaussian mixture model of means $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ and covariances $\mathbf{C}_1, \mathbf{C}_2$ in the regime where $n, p \to \infty$ and $n/p \to c_0 \in (0, \infty)$. Similar to [6], we find that there exists a critical growth rate regime (with n and p) of the aforementioned means and covariances for which a non-trivial asymptotic classification error rate is obtained. We also notice that, just as in [6] for kernel spectral clustering, only a very local aspect of the kernel function drives the classification performance in the large n, p regime. Precisely, we find that the decision function of LS-SVM converges to a Gaussian random variable with means and covariances depending explicitly on the derivatives of the kernel function evaluated at $2\left(n_1 \cdot \operatorname{tr} \mathbf{C}_1 + n_2 \cdot \operatorname{tr} \mathbf{C}_2\right)/(np)$, with n_1 and n_2 the number of instances in each class. This brings new insights into questions such as kernel function selection and parameter optimization for LS-SVMs with large dimensional data. Because of space limitation, only proof sketches are provided for our main results, the complete derivations being available in an extended version of the article.

Notations: Boldface lowercase (uppercase) characters stand for vectors (matrices), and scalars non-boldface respectively. $\mathbf{1}_n$ is the column vector of ones, and $\mathbf{I}_n$ the $n \times n$ identity matrix. The notation $(\cdot)^T$ denotes the transpose operator. The norm $\|\cdot\|$ is the Euclidean norm for vectors and the operator norm for matrices.

2. MODEL AND ASSUMPTIONS

Let $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p$ be independent vectors belonging to two distribution classes $\mathcal{C}_1, \mathcal{C}_2$, with $\mathbf{x}_1, \ldots, \mathbf{x}_{n_1} \in \mathcal{C}_1$ and $\mathbf{x}_{n_1+1}, \ldots, \mathbf{x}_n \in \mathcal{C}_2$ (so that class $\mathcal{C}_1$ has cardinality n_1 and class $\mathcal{C}_2$ has cardinality $n_2 = n - n_1$). We assume that $\mathbf{x}_i \in \mathcal{C}_a$

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

[1] As shall be seen later, the division by p here is the necessary normalization in the large n, p regime.

[2] Unlike the standard SVM which is based on the implicit solution to a quadratic programming problem.

for $a \in \{1, 2\}$ if

$$\mathbf{x}_i \sim \mathcal{N}(\boldsymbol{\mu}_a, \mathbf{C}_a)$$

for some $\boldsymbol{\mu}_a \in \mathbb{R}^p$ and nonnegative definite $\mathbf{C}_a \in \mathbb{R}^{p \times p}$.

To achieve an asymptotically non-trivial misclassification rates (i.e., neither 0 nor 1), we show (similar to [6]) that one needs to work under the following growth rate assumptions:

Assumption 1 (Growth rate)**.** *As* $n \to \infty$, $\frac{p}{n} \to c_0 > 0$, $\frac{n_i}{n} \to c_i > 0$ *(we shall write* $\mathbf{c} = [c_1, c_2]^T$*). Besides,*

1. *For* $\boldsymbol{\mu}^\circ \triangleq \frac{n_1}{n}\boldsymbol{\mu}_1 + \frac{n_2}{n}\boldsymbol{\mu}_2$ *and* $\boldsymbol{\mu}_a^\circ \triangleq \boldsymbol{\mu}_a - \boldsymbol{\mu}$, $\|\boldsymbol{\mu}_a^\circ\| = O(1)$.
2. *For* $\mathbf{C}^\circ \triangleq \frac{n_1}{n}\mathbf{C}_1 + \frac{n_2}{n}\mathbf{C}_2$ *and* $\mathbf{C}_a^\circ \triangleq \mathbf{C}_a - \mathbf{C}^\circ$, $\|\mathbf{C}_a\| = O(1)$ *and* $\mathrm{tr}\mathbf{C}_a^\circ = O(\sqrt{n})$.
3. *As* $n \to \infty$, $\frac{2}{p} \operatorname{tr} \mathbf{C}^\circ \to \tau > 0$.

Let $\boldsymbol{\varphi} : \mathbb{R}^p \to \mathcal{H}$ be a function such that there exists $f : \mathbb{R}_+ \to \mathbb{R}_+$ for which $\boldsymbol{\varphi}(\mathbf{x}_i)^T\boldsymbol{\varphi}(\mathbf{x}_j) = f\left(\|\mathbf{x}_i - \mathbf{x}_j\|^2/p\right)$, with f satisfying the following assumptions:

Assumption 2 (Kernel function)**.** *The function* f *is a three-times differentiable function in a neighborhood of* τ.

Under this assumption, we define $\mathbf{K} \in \mathbb{R}^{n \times n}$ as the kernel matrix

$$\mathbf{K} \triangleq \left\{ f\left(\frac{1}{p}\|\mathbf{x}_i - \mathbf{x}_j\|^2\right)\right\}_{i,j=1}^{n}. \tag{1}$$

The value τ introduced in Assumption 1 is important since in the regime where both $n, p \to \infty$, from Assumption 1, for all pairs $i \neq j$, $\|\mathbf{x}_i - \mathbf{x}_j\|^2/p \to \tau$, almost surely, which makes it possible to perform a Taylor expansion around $f(\tau)$ of $f\left(\|\mathbf{x}_i - \mathbf{x}_j\|^2/p\right)$. We shall see that only the local behavior of the matrix $\mathbf{K}$ around the matrix $f(\tau)\mathbf{1}_n\mathbf{1}_n^T$ plays a significant role in the classification of LS-SVMs. As such, the intractable non-linear kernel matrix $\mathbf{K}$ can be (asymptotically) linearized and, in turn, the decision function of LS-SVMs (which, as shown next, is an explicit function of $\mathbf{K}$) becomes tractable as $n, p \to \infty$.

3. MAIN RESULTS

For $\mathbf{x}_1, \ldots, \mathbf{x}_n$ defined previously, let $y_i = -1$ if $\mathbf{x}_i \in \mathcal{C}_1$ and $y_i = 1$ if $\mathbf{x}_i \in \mathcal{C}_2$. The objective of LS-SVM is to separate the classes $\mathcal{C}_1$ and $\mathcal{C}_2$ in the kernel space $\mathcal{H}$, via a "hyperplane" of the form $\mathbf{w}^T\boldsymbol{\varphi}(\mathbf{x}) + b = 0$, where $\mathbf{w}$ and b are the solutions of the following optimization problem[2]:

$$\begin{aligned} \underset{\mathbf{w}}{\arg\min} \quad & J(\mathbf{w}, e) = \|\mathbf{w}\|^2 + \frac{\gamma}{n}\sum_{i=1}^{n} e_i^2 \\ \text{such that} \quad & y_i = \mathbf{w}^T\boldsymbol{\varphi}(\mathbf{x}_i) + b + e_i,\ i = 1, \ldots, n \end{aligned} \tag{2}$$

where $\gamma > 0$ is a penalty factor on the square deviations e_i^2 from the hyperplane. The solution of (2) is $\mathbf{w} = \sum_{i=1}^n \alpha_i\boldsymbol{\varphi}(\mathbf{x}_i)$, where

$$\begin{cases} \boldsymbol{\alpha} &= \mathbf{S}^{-1}\left(\mathbf{I}_n - \frac{\mathbf{1}_n\mathbf{1}_n^T\mathbf{S}^{-1}}{\mathbf{1}_n^T\mathbf{S}^{-1}\mathbf{1}_n}\right)\mathbf{y} = \mathbf{S}^{-1}\left(\mathbf{y} - b\mathbf{1}_n\right) \\ b &= \frac{\mathbf{1}_n^T\mathbf{S}^{-1}\mathbf{y}}{\mathbf{1}_n^T\mathbf{S}^{-1}\mathbf{1}_n} \end{cases} \tag{3}$$

with $\mathbf{S}^{-1} = \left(\mathbf{K} + \frac{n}{\gamma}\mathbf{I}_n\right)^{-1}$, $\mathbf{K}$ given by Equation (1), $\mathbf{y} = [y_1, \ldots, y_n]^T$ and $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_n]^T$.

Given $\boldsymbol{\alpha}$ and b, each new datum $\mathbf{x}$ is classified using LS-SVM, based on the following decision function[2]

$$g(\mathbf{x}) = \boldsymbol{\alpha}^T\mathbf{k}(\mathbf{x}) + b \tag{4}$$

where $\mathbf{k}(\mathbf{x}) = \left[f\left(\|\mathbf{x}_j - \mathbf{x}\|^2/p\right)\right]_{j=1}^n \in \mathbb{R}^n$. More precisely, $\mathbf{x}$ is associated to class $\mathcal{C}_1$ if $g(\mathbf{x})$ takes a small value (below a certain threshold) and to class $\mathcal{C}_2$ otherwise.

We are concerned here with assessing the performance of LS-SVM, under the setting of Assumption 1, as $n, p \to \infty$. The idea is to study the random variable $g(\mathbf{x})$ in Equation (4) for $\mathbf{x} \in \mathcal{C}_1$ or $\mathbf{x} \in \mathcal{C}_2$, which then makes it possible to evaluate the error rate of classification. Since $g(\mathbf{x})$ is explicitly defined as a function of $\mathbf{K}$ (through $\boldsymbol{\alpha}$ and b), with $\mathbf{K}$ linearizable in the large n, p regime, one can work out an asymptotic linearization of $g(\mathbf{x})$ as a function of f and the statistics of the known vectors $\mathbf{x}_i$'s. We provide next a sketch of this derivation.

Let us start by Taylor-expanding $\mathbf{S}^{-1}$. Under the settings of Section 2, we notice that the leading term $f(\tau)\mathbf{1}_n\mathbf{1}_n^T$ in the Taylor expansion of $\mathbf{K}$ (with respect to the operator norm) as well as $\frac{n}{\gamma}\mathbf{I}_n$ are of norm $O(n)$. As such, and after a basic algebraic manipulation,

$$\mathbf{S}^{-1} = \frac{\gamma}{n}\left(\mathbf{I}_n - \frac{\gamma f(\tau)}{1 + \gamma f(\tau)}\frac{\mathbf{1}_n\mathbf{1}_n^T}{n}\right) + O(n^{-\frac{3}{2}}) \tag{5}$$

and thus the terms making the classification possible are hidden in the $O(n^{-\frac{3}{2}})$ term, which needs to be thoroughly developed next. From the expansion (5) of $\mathbf{S}^{-1}$, we further have:

$$\begin{cases} \boldsymbol{\alpha} &= \frac{\gamma}{n}\left(\mathbf{y} - (c_2 - c_1)\mathbf{1}_n\right) + O(n^{-\frac{3}{2}}) \\ b &= c_2 - c_1 + O(n^{-\frac{1}{2}}) \end{cases} \tag{6}$$

and

$$\mathbf{k}(\mathbf{x}) = f(\tau)\mathbf{1}_n + O(n^{-\frac{1}{2}})$$

so that, similar to $\mathbf{S}^{-1}$, the structural (class) information of the new data $\mathbf{x}$ is carried by the term $O(n^{-\frac{1}{2}})$ which also needs to be carefully developed. This is performed following the technique elaborated in [5, 6].

Putting all the approximations together brings the main result of the article as follows:

Theorem 1 (Gaussian approximation of $g(\mathbf{x})$)**.** *Let Assumptions 1 and 2 hold, and* $g(\mathbf{x})$ *be defined by* (4)*. Then for* $\mathbf{x} \in \mathcal{C}_a$*,* $a \in \{1, 2\}$*,* $n(g(\mathbf{x}) - G_a) \to 0$*, where*

$$G_a \sim \mathcal{N}(\mathrm{E}_a, \mathrm{Var}_a)$$

with

$$\mathrm{E}_a = \begin{cases} c_2 - c_1 - 2c_2 \cdot c_1 c_2 \gamma \mathfrak{D} \,, & a = 1 \\ c_2 - c_1 + 2c_1 \cdot c_1 c_2 \gamma \mathfrak{D} \,, & a = 2 \end{cases}$$

$$\mathrm{Var}_a = 8\gamma^2 c_1^2 c_2^2 \left(\mathcal{V}_1^a + \mathcal{V}_2^a + \mathcal{V}_3^a \right)$$

and

$$\mathfrak{D} = -\frac{2f'(\tau)}{p} \|\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1\|^2 + \frac{f''(\tau)}{p^2} \left(\mathrm{tr}\left(\mathbf{C}_2 - \mathbf{C}_1 \right) \right)^2 + \frac{2f''(\tau)}{p^2} \mathrm{tr}\left((\mathbf{C}_2 - \mathbf{C}_1)^2 \right) \tag{7}$$

$$\mathcal{V}_1^a = \frac{\left(f''(\tau) \right)^2}{p^4} \left(\mathrm{tr}\left(\mathbf{C}_2 - \mathbf{C}_1 \right) \right)^2 \mathrm{tr} \mathbf{C}_a^2$$

$$\mathcal{V}_2^a = \frac{2\left(f'(\tau) \right)^2}{p^2} \left(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1 \right)^T \mathbf{C}_a \left(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1 \right)$$

$$\mathcal{V}_3^a = \frac{2\left(f'(\tau) \right)^2}{np^2} \left(\frac{\mathrm{tr}\, \mathbf{C}_1 \mathbf{C}_a}{c_1} + \frac{\mathrm{tr}\, \mathbf{C}_2 \mathbf{C}_a}{c_2} \right)$$

Theorem 1 states that the problem of classification using LS-SVM is asymptotically equivalent to a simple thresholding to separate two monovariate Gaussian random variables, the means and covariances of which depend on the (normalized) inter-class mean-deviation, traces of class covariances and the derivatives of the kernel function f at τ, when both $n, p \to \infty$. Letting $Q(x) = \frac{1}{2\pi} \int_x^\infty \exp(-t^2/2) dt$, we have in particular the following immediate corollary of Theorem 1:

Corollary 1 (Asymptotic error rates)**.** *Under the setting of Theorem 1, for ξ_n possibly depending on n,*

$$\mathrm{P}(g(\mathbf{x}) > \xi_n \mid \mathbf{x} \in \mathcal{C}_1) - Q\left(\frac{\xi_n - \mathrm{E}_1}{\sqrt{\mathrm{Var}_1}} \right) \to 0 \tag{8a}$$

$$\mathrm{P}(g(\mathbf{x}) < \xi_n \mid \mathbf{x} \in \mathcal{C}_2) - Q\left(\frac{\mathrm{E}_2 - \xi_n}{\sqrt{\mathrm{Var}_2}} \right) \to 0. \tag{8b}$$

Note here the importance of a proper setting of the function f. For instance, if $f'(\tau) = 0$, the term $\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1$ vanishes from the mean and variance of G_a, meaning that the classification of LS-SVM will not rely (at least asymptotically and under the key Assumption 1) on differences in means. Similarly, if $f''(\tau) = 0$, all terms of covariance matrices vanish in E_a but remain in Var_a.

Remark 1 (Dominant bias)**.** *From Theorem 1, we have $\mathfrak{D} = O(n^{-1})$, which means that $\mathrm{E}_a = c_2 - c_1 + O(n^{-1})$. As such, from Corollary 1, it appears natural to set $\xi_n = c_2 - c_1$, rather than $\xi_n = 0$ as one would naturally do (because if $c_2 - c_1 > 0$ then $\mathrm{P}(g(\mathbf{x}) > 0 \mid \mathbf{x} \in \mathcal{C}_1) \to 1$ and $\mathrm{P}(g(\mathbf{x}) < 0 \mid \mathbf{x} \in \mathcal{C}_2) \to 0$). It has been shown in[7] through a Bayesian approach that this phenomenon is due to b, which is also referred to as the "bias term", that depends on the prior class probabilities.*

One may notice in Theorem 1 that in the case when $\mathrm{Var}_1 = \mathrm{Var}_2$, the performance of classification depends on deviation in means, which is proportional to $\mathfrak{D}$ as defined in (7). To achieve optimal classification performance, we thus need $f(\tau), f''(\tau) > 0$ and $f'(\tau) < 0$. Coincidentally, it is naturally the case with the popular Gaussian kernel $f(x) = \exp\left(-x/2\sigma^2\right)$, but not necessarily for other kernel functions. For the second-order polynomial kernel given by $f(x) = ax^2 + bx + c$, we shall have the following constraints:

Corollary 2 (Polynomial kernels)**.** *Under the setting of Theorem 1, with $f(x) = ax^2 + bx + c$, the following conditions are necessary to maximize $\mid \mathrm{E}_1 - \mathrm{E}_2 \mid$ while ensuring $f > 0$,*

1. $f(\tau) > 0, f'(\tau) < 0, f''(\tau) > 0$
2. $\left(f'(\tau) \right)^2 < 2f(\tau) f''(\tau)$.

The corollary above may give us some inspiration in the choice of kernel functions.

A particularly surprising outcome of Theorem 1 is that, when $\mathrm{tr}\, \mathbf{C}_1 = \mathrm{tr}\, \mathbf{C}_2$ and one chooses f in such a way that $f'(\tau) = 0$, then $\mathrm{Var}_a = 0$ while E_a may remain non-zero, thereby ensuring a vanishing error rate, i.e., the second left-hand side terms of (8a) and (8b) equal zero. Figure 1 corroborates this finding for $\mathbf{C}_1 = \mathbf{I}_p$ and $\{\mathbf{C}_2\}_{i,j} = .4^{|i-j|}$.

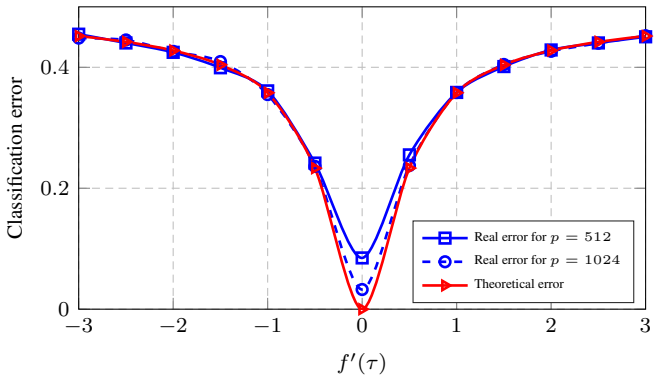


Fig. 1. Performance of LS-SVM, $c_0 = 1/4$, $c_1 = c_2 = 1/2, \gamma = 1$, polynomial kernel with $f(\tau) = 4$, $f''(\tau) = 2$. $\mathbf{x} \in \mathcal{N}(\boldsymbol{\mu}_a, \mathbf{C}_a)$, with $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \mathbf{0}_p$, $\mathbf{C}_1 = \mathbf{I}_p$ and $\{\mathbf{C}_2\}_{i,j} = .4^{|i-j|}$.

Remark 2 (Dominant deviation in means)**.** *A direct result from Theorem 1 is that, in the case when $\|\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1\|^2$ is largely dominant over $\mathrm{tr}\left(\mathbf{C}_2 - \mathbf{C}_1 \right)/\sqrt{p}$ and $\mathrm{tr}\left((\mathbf{C}_2 - \mathbf{C}_1)^2 \right)/p$, e.g., $\mathrm{tr}\left(\mathbf{C}_2 - \mathbf{C}_1 \right) = o(\sqrt{p})$ and $\mathrm{tr}\left((\mathbf{C}_2 - \mathbf{C}_1)^2 \right) = o(p)$ while $\|\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1\|^2 = O(1)$, then $\mathrm{E}_a - (c_2 - c_1)$ and $\sqrt{\mathrm{Var}_a}$ are (asymptotically) proportional to $f'(\tau)$, which makes the choice of kernel function of little importance (if $f'(\tau) \neq 0$).*

Remark 3 (Insignificance of γ)**.** *The parameter γ appears as a scale factor of both $\mathrm{E}_a - (c_2 - c_1)$ and $\sqrt{\mathrm{Var}_a}$ which, along with Corollary 1 and Remark 1, indicates the (asymptotic) independence of γ in the error rates $\mathrm{P}(g(\mathbf{x}) > c_2 - c_1 \mid \mathbf{x} \in \mathcal{C}_1)$ and $\mathrm{P}(g(\mathbf{x}) < c_2 - c_1 \mid \mathbf{x} \in \mathcal{C}_2)$.*

4. SIMULATIONS

We complete this article by demonstrating that our results, which apply theoretically only to Gaussian $\mathbf{x}_i$'s, show a behavior unexpectedly close to theory when applied to real-world datasets. Here, we consider the classification problem with a training set of $n_1 = n_2 = 128$ vectorized images of size $p = 784$ from the popular MNIST dataset[8] (numbers 1 and 7, as shown in Figure 2). Then a test set of size $n_{\text{test}} = 128 \times 2$ is used to evaluate the performance of LS-SVM. Means and covariances are empirically obtained from the full set of 13 007 MNIST images (6 742 images of number 1 and 6 265 of number 7).

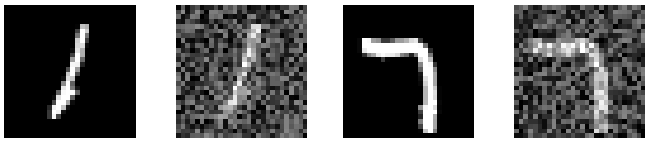
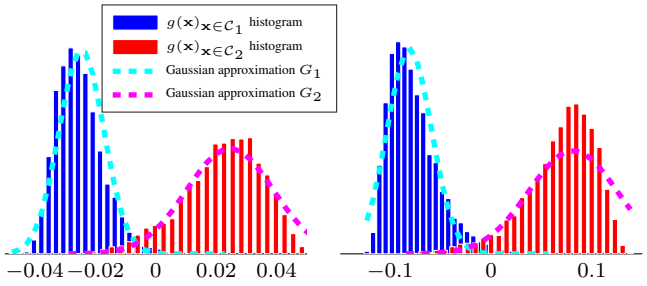


Fig. 2. Samples from the MNIST database, without and with 0dB noise.

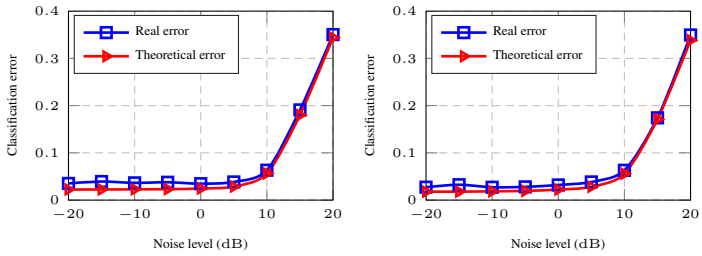
Figure 3 shows that, despite the obvious non-Gaussianity of the MNIST data, the distribution of the decision function $g(\mathbf{x})$ is surprisingly close to its Gaussian approximation G_a. When added a 0dB Gaussian white noise, the two agree with each other even better, in an almost perfect match, especially in the overlapping areas from which the classification error rates are computed.



(a) with 0dB noise (b) without noise

Fig. 3. Gaussian approximation of $g(\mathbf{x})$, $n = 256$, $p = 784$, $c_1 = c_2 = 1/2$, $\gamma = 1$, Gaussian kernel with $\sigma^2 = 1$, MNIST data (numbers 1 and 7) without and with 0dB noise.

In Figure 4, we evaluate the performance of LS-SVM as the function of the noise level (in dB), for Gaussian kernel $f(x) = \exp\left(-x/2\sigma^2\right)$ with $\sigma^2 = 1$ in subfigure (a), and $\sigma^2 = 10$ in (b). Surprisingly, we face the situation where there is little difference in the performance of LS-SVM with different values of σ, which likely comes from the fact that the difference in means $\|\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1\|$ is so large that it becomes predominant over the covariances as mentioned in Remark 2. This is numerically confirmed in Table 1.



(a) $\sigma^2 = 10$ (b) $\sigma^2 = 1$

Fig. 4. Performance of LS-SVM, $n = 256$, $p = 784$, $c_1 = c_2 = 1/2, \gamma = 1$,Gaussian kernel, MNIST data.

	Without noise	With 0dB noise
$\|\boldsymbol{\mu}_2 - \boldsymbol{\mu}_2\|^2$	429	178
$\left(\text{tr}\left(\mathbf{C}_2 - \mathbf{C}_1\right)\right)^2/p$	63	11
$\text{tr}\left(\left(\mathbf{C}_2 - \mathbf{C}_1\right)^2\right)/p$	35	6

Table 1. Empirical estimation of (normalized) differences in means and covariances of MNIST data.

5. CONCLUDING REMARKS

In this article, under a random matrix growth regime, we reexhibit the bias $c_2 - c_1$ in the decision function threshold of LS-SVM already identified in [7], and find out how information is retrieved from the means and covariances of data from two different classes, as well as the influence of the kernel function f. This notably allows us to have a deeper understanding of the mechanism into play and in particular the impact of the choice of the kernel function as well as some theoretical limits of the method.

The extension of the present work to the asymptotic performance analysis of the classical SVM requires more efforts since, there, the decision function $g(\mathbf{x})$ depends *implicitly* (through the solution to a quadratic programming problem) rather than explicitly on the underlying kernel matrix $\mathbf{K}$. A possible approach is to bound precisely the solution of the optimization problem with two random variables whose difference will asymptotically vanish as $n, p \to \infty$, similar to the method devised in [9] for the random matrix analysis of robust estimators of scatter. Also, while the theoretical formulas of Theorem 1 are simple, applications to practical datasets sometimes reveal larger discrepancies than observed in Figures 3 and 4. These are likely due to a too strong Gaussianity assumption on the input data, along with an important need for n and p to be significantly larger than in classical random matrix applications for the asymptotic results to be accurate. Nonetheless, in classical applications of signal processing such as radar and sonar, the system models are often based on (practically validated) Gaussian models, which provides the possibility of applying our theoretical results in real-world engineering problems.

6. REFERENCES

[1] Vladimir Vapnik, *The nature of statistical learning theory*, Springer Science & Business Media, 2013.

[2] Johan AK Suykens and Joos Vandewalle, "Least squares support vector machine classifiers," *Neural processing letters*, vol. 9, no. 3, pp. 293–300, 1999.

[3] S Sathiya Keerthi and Chih-Jen Lin, "Asymptotic behaviors of support vector machines with gaussian kernel," *Neural computation*, vol. 15, no. 7, pp. 1667–1689, 2003.

[4] Jieping Ye and Tao Xiong, "Svm versus least squares svm," in *International Conference on Artificial Intelligence and Statistics*, 2007, pp. 644–651.

[5] Noureddine El Karoui et al., "The spectrum of kernel random matrices," *The Annals of Statistics*, vol. 38, no. 1, pp. 1–50, 2010.

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

[7] Johan AK Suykens, Tony Van Gestel, Jos De Brabanter, Bart De Moor, Joos Vandewalle, JAK Suykens, and T Van Gestel, *Least squares support vector machines*, vol. 4, World Scientific, 2002.

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

[9] Romain Couillet, Frédéric Pascal, and Jack W Silverstein, "The random matrix regime of maronna's m-estimator with elliptically distributed samples," *Journal of Multivariate Analysis*, vol. 139, pp. 56–78, 2015.