A Large Dimensional Analysis of Least Squares Support Vector Machines

Zhenyu Liao, Romain Couillet

▶ **To cite this version:**

Zhenyu Liao, Romain Couillet. A Large Dimensional Analysis of Least Squares Support Vector Machines. IEEE Transactions on Signal Processing, 2019, 67 (4), pp.1065-1074. 10.1109/TSP.2018.2889954 . hal-02048984

HAL Id: hal-02048984

https://inria.hal.science/hal-02048984

Submitted on 19 May 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

A Large Dimensional Analysis of Least Squares Support Vector Machines

Zhenyu Liao, Romain Couillet

arXiv:1701.02967v2 [stat.ML] 22 Feb 2019

***Abstract*—In this article, a large dimensional performance analysis of kernel least squares support vector machines (LS-SVMs) is provided under the assumption of a two-class Gaussian mixture model for the input data. Building upon recent advances in random matrix theory, we show, when the dimension of data p and their number n are both large, that the LS-SVM decision function can be well approximated by a normally distributed random variable, the mean and variance of which depend explicitly on a local behavior of the kernel function. This theoretical result is then applied to the MNIST and Fashion-MNIST datasets which, despite their non-Gaussianity, exhibit a convincingly close behavior. Most importantly, our analysis provides a deeper understanding of the mechanism into play in SVM-type methods and in particular of the impact on the choice of the kernel function as well as some of their theoretical limits in separating high dimensional Gaussian vectors.**

***Index Terms*—High dimensional statistics, kernel methods, random matrix theory, support vector machines**

I. Introduction

In the past two decades, due to their surprising classification capability and simple implementation, kernel support vector machine (SVM) [1] and its variants [2–4] have been used in a wide variety of classification applications, such as face detection [5,6], handwritten digit recognition [7], and text categorization [8,9]. In all aforementioned applications, the dimension of data p and their number n are large: in the hundreds and even thousands. The significance of working in this large n, p regime is even more convincing in the Big Data paradigm today where handling data which are both numerous and large dimensional becomes increasingly common.

Firmly grounded in the framework of statistical learning theory [10], support vector machine has two main features: (i) in SVM, the training data $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p$ are mapped into some *feature space* through a non-linear function $\boldsymbol{\varphi}$, which, thanks to the so-called "kernel trick" [11], needs not be computed explicitly, so that some *kernel function* f is introduced in place of the inner product in the feature space: $f(\mathbf{x}, \mathbf{y}) = \boldsymbol{\varphi}(\mathbf{x})^\mathsf{T} \boldsymbol{\varphi}(\mathbf{y})$, and (ii) a standard (convex) optimization method is used to find the classifier that both minimizes the training error and yields a good generalization performance for unknown data.

This work is supported by the ANR Project RMT4GRAPH (ANR-14-CE28-0006).This paper was presented in part at the 42nd IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'17), New Orleans, USA, March 2017.

Z. Liao and R. Couillet are with the Laboratoire de Signaux et Systèmes, CNRS-CentraleSupélec-Université Paris-Sud, 3 rue Joliot-Curie, 91192 Gif-sur-Yvette, France (email: zhenyu.liao@l2s.centralesupelec.fr; romain.couillet@centralesupelec.fr).

As the training of SVMs involves a quadratic programming problem, the computation complexity of SVM training algorithms can be intensive when the number of training examples n becomes large (at least quadratic with respect to n). It is thus difficult to deal with large scale problems with traditional SVMs. To cope with this limitation, least squares SVM (LS-SVM, also later referred to as kernel regularized least-squares estimator or kernel ridge regression [12–14]) was proposed in [2], providing a more computationally efficient implementation of the traditional SVMs, by taking equality optimization constraints instead of inequalities, which results in an explicit solution (from a set of linear equations) rather than an implicit one in SVMs. This article is mostly concerned with this particular type of SVMs.

Trained SVMs are strongly data-dependent: the data with generally unknown statistics are passed through a nonlinear kernel function f and standard optimization methods are used to find the best classifier. All these features make the performance of SVM hardly traceable (at least within the classical finite n, p regime). To understand the mechanism of SVMs, the notion of VC dimension was introduced to provide bounds on the generalization performance of SVM [10], while a probabilistic interpretation of LS-SVM was discussed in [15] through a Bayesian inference approach. In other related works, connections between LS-SVMs and SVMs were revealed in [16], and more relationships were shown between SVM-type and other learning methods, e.g., LS-SVMs and extreme learning machines (ELMs) [17]; SVMs and regularization networks (RNs) [18], etc. Theoretical analyses on the generalization performance of LS-SVM have been developed, under the conventional asymptotic statistics framework (i.e., assuming $n \to \infty$), to obtain optimal convergence rates in [13,14]. Nonetheless, a proper adaptation to the large n, p setting to address LS-SVM performance for large dimensional datasets (of growing interest today) is still missing.

Similar to classical analysis of asymptotic statistics where $n \to \infty$ while p is fixed, where the diversity of the number of data provides convergence through laws of large numbers, working in the large n, p regime by letting in addition $p \to \infty$ helps exploit the diversity offered by the size of each data vector, providing us with another dimension to guarantee the convergence of some key objects in our analysis, and thus makes the asymptotic analysis of the elusive *kernel matrix* $\mathbf{K} = \{f(\mathbf{x}_i, \mathbf{x}_j)\}_{i,j=1}^n$ technically more accessible. Recent breakthroughs in random matrix theory have allowed one to overtake the theoretical difficulty posed by the nonlinearity of the aforementioned kernel function f [19,20] and thus make an in-depth analysis of LS-SVM possible in the large n, p regime.

These tools were notably used to assess the performance of the popular Ng-Weiss-Jordan kernel spectral clustering methods for large datasets [20], in the analysis of graphed-based semi-supervised learning [21] or for the development of novel kernel subspace clustering methods [22].

Similar to these works, in this article, we provide a performance analysis of LS-SVM, in the regime of $n, p \to \infty$ and $p/n \to \bar{c}_0 \in (0, \infty)$, under the assumption of a two-class Gaussian mixture model of means $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ and covariance matrices $\mathbf{C}_1, \mathbf{C}_2$ for the input data. The Gaussian assumption may seem artificial to the practitioners, but reveals first insights into how SVM-type methods deal with the information in means and covariances from a more quantitative point of view. Besides, the early investigations [20,21] have revealed that the behavior of some machine learning methods under Gaussian or deterministic practical input datasets are a close match, despite the obvious non-Gaussianity of the latter.

Our main finding is that, as in [20], in the large n, p regime and under suitable conditions on the input statistics, a non-trivial asymptotic classification error rate (i.e., neither 0 nor 1) can be obtained and the decision function of LS-SVM converges to a Gaussian random variable whose mean and variance depend on the statistics of the two different classes as well as on the behavior of the kernel function f evaluated at $2\operatorname{tr}(n_1\mathbf{C}_1 + n_2\mathbf{C}_2)/(np)$, with n_1 and n_2 the number of instances in each class. This brings novel insights into some key issues of SVM-type methods such as kernel function selection and parameter optimization (see for example [15,23–27] and the references therein), as far as large dimensional data are concerned. More importantly, we confirm through simulations that our theoretical findings closely match the performance obtained on the MNIST [28] and the Fashion-MNIST datasets [29], which conveys a strong applicative motivation for this work.

In the remainder of the article, we provide a rigorous statement of our main results. The problem of LS-SVM is discussed in Section II and our model and main results presented in Section III, while all proofs are deferred to the appendices in the Supplementary Material. In Section IV, attention will be paid on some special cases that are more analytically tractable. Section V concludes the paper by summarizing the main results and outlining future research directions.

Reproducibility: Python 3 codes to reproduce the results in this article are available at https://github.com/Zhenyu-LIAO/RMT4LSSVM.

Notations: Boldface lowercase (uppercase) characters stand for vectors (matrices), and scalars non-boldface respectively. $\mathbf{1}_n$ is the column vector of ones of size n, $\mathbf{0}_n$ the column vector of zeros, and $\mathbf{I}_n$ the $n \times n$ identity matrix. The notation $(\cdot)^\mathsf{T}$ denotes the transpose operator. The norm $\|\cdot\|$ is the Euclidean norm for vectors and the operator norm for matrices. The notation $\mathrm{P}(\cdot)$ denotes the probability measure of a random variable. The notation $\xrightarrow{d}$ denotes convergence in distribution and $\xrightarrow{\text{a.s.}}$ almost sure convergence, respectively. The operator $\mathcal{D}(\mathbf{v}) = \mathcal{D}\{v_a\}_{a=1}^k$ is the diagonal matrix having $v_a, \ldots, v_k$ as its ordered diagonal elements. We denote $\{v_a\}_{a=1}^k$ a column vector with a-th entry (or block entry) v_a (which may be a vector), while $\{V_{ab}\}_{a,b=1}^k$ denotes a square matrix with entry (or block-entry) (a, b) given by V_{ab} (which may be a matrix).

II. Problem statement

Least squares support vector machines (LS-SVMs) are a modification of the standard SVM introduced in [2] to overcome the drawbacks of SVM related to computational efficiency. The optimization problem has half the number of parameters and benefits from solving a linear system of equations instead of a quadratic programming problem as in standard SVM and is thus more practical for large dimensional learning tasks. In this article, we will focus on a binary classification problem using LS-SVM as described in the following paragraph.

Given a training set $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ of size n, where data $\mathbf{x}_i \in \mathbb{R}^p$ and labels $y_i \in \{-1, 1\}$, the objective of LS-SVM is to devise a decision function $g(\mathbf{x})$ that ideally maps all $\mathbf{x}_i$ in the training set to y_i and subsequently all unknown data $\mathbf{x}$ to their corresponding y value. Here we denote $\mathbf{x}_i \in \mathcal{C}_1$ if $y_i = -1$ and $\mathbf{x}_i \in \mathcal{C}_2$ if $y_i = 1$ and shall say that $\mathbf{x}_i$ belongs to class $\mathcal{C}_1$ or class $\mathcal{C}_2$, respectively. Due to the often nonlinear separability of these training data in the input space $\mathbb{R}^p$, in most cases, one associates the training data $\mathbf{x}_i$ to some feature space $\mathcal{H}$ through a nonlinear mapping $\boldsymbol{\varphi} : \mathbf{x}_i \mapsto \boldsymbol{\varphi}(\mathbf{x}_i) \in \mathcal{H}$. Constrained optimization methods are then used to define a separating hyperplane in $\mathcal{H}$ with direction vector $\mathbf{w}$ and correspondingly to find a function $g(\mathbf{x}) = \mathbf{w}^\mathsf{T}\boldsymbol{\varphi}(\mathbf{x}) + b$ that minimizes the training errors $e_i = y_i - \left(\mathbf{w}^\mathsf{T}\boldsymbol{\varphi}(\mathbf{x}_i) + b\right)$, and meanwhile yields good generalization performance by minimizing the norm of $\mathbf{w}$ [30]. More specifically, the LS-SVM approach consists in minimizing the squared errors e_i^2, thus resulting in[1]

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

where $\gamma > 0$ is a penalty factor that weights the structural risk $\|\mathbf{w}\|^2$ against the empirical one $\frac{1}{n}\sum_{i=1}^n e_i^2$.

The problem can be solved by introducing Lagrange multipliers $\alpha_i, i = 1, \ldots, n$ with solution $\mathbf{w} = \sum_{i=1}^n \alpha_i \boldsymbol{\varphi}(\mathbf{x}_i)$, where, letting $\mathbf{y} = [y_1, \ldots, y_n]^\mathsf{T}, \boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_n]^\mathsf{T}$, we obtain

$$\begin{cases} \boldsymbol{\alpha} &= \mathbf{S}^{-1}\left(\mathbf{I}_n - \frac{\mathbf{1}_n\mathbf{1}_n^\mathsf{T}\mathbf{S}^{-1}}{\mathbf{1}_n^\mathsf{T}\mathbf{S}^{-1}\mathbf{1}_n}\right)\mathbf{y} = \mathbf{S}^{-1}\left(\mathbf{y} - b\mathbf{1}_n\right) \\ b &= \frac{\mathbf{1}_n^\mathsf{T}\mathbf{S}^{-1}\mathbf{y}}{\mathbf{1}_n^\mathsf{T}\mathbf{S}^{-1}\mathbf{1}_n} \end{cases} \tag{2}$$

with $\mathbf{S} = \mathbf{K} + \frac{n}{\gamma}\mathbf{I}_n$ and $\mathbf{K} \triangleq \left\{\boldsymbol{\varphi}(\mathbf{x}_i)^\mathsf{T}\boldsymbol{\varphi}(\mathbf{x}_j)\right\}_{i,j=1}^n$ referred to as the kernel matrix [2].

Given $\boldsymbol{\alpha}$ and b, a new datum $\mathbf{x}$ is then classified into class $\mathcal{C}_1$ or $\mathcal{C}_2$ depending on the value of the following decision function

$$g(\mathbf{x}) = \boldsymbol{\alpha}^\mathsf{T}\mathbf{k}(\mathbf{x}) + b \tag{3}$$

[1] We include the bias term b as in the (classical) LS-SVM formulation [2], which may be different from kernel ridge regression in some literature [12,15] where no bias term is used.

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

With the "kernel trick" [11], as shown in (2) and (3) that, both in the "training" and "testing" steps, one only needs to evaluate the inner product $\boldsymbol{\varphi}(\mathbf{x}_i)^\mathsf{T}\boldsymbol{\varphi}(\mathbf{x}_j)$ or $\boldsymbol{\varphi}(\mathbf{x})^\mathsf{T}\boldsymbol{\varphi}(\mathbf{x}_j)$, and never needs to know explicitly the mapping $\boldsymbol{\varphi}(\cdot)$. In the rest of this article, we assume that the kernel is *translation invariant* and focus on kernel functions $f : \mathbb{R}^+ \to \mathbb{R}^+$ that satisfy $\boldsymbol{\varphi}(\mathbf{x}_i)^\mathsf{T}\boldsymbol{\varphi}(\mathbf{x}_j) = f(\|\mathbf{x}_i - \mathbf{x}_j\|^2/p)$ and shall redefine $\mathbf{K}$ and $\mathbf{k}(\mathbf{x})$ for data point $\mathbf{x}$ as[3]

$$\begin{aligned}\mathbf{K} &= \left\{f\left(\|\mathbf{x}_i - \mathbf{x}_j\|^2/p\right)\right\}_{i,j=1}^n \\ \mathbf{k}(\mathbf{x}) &= \left\{f\left(\|\mathbf{x} - \mathbf{x}_j\|^2/p\right)\right\}_{j=1}^n.\end{aligned} \tag{4}$$

Some commonly used kernel functions are the Gaussian radial basis (RGB) kernel $f(x) = \exp\left(-\frac{x}{2\sigma^2}\right)$ with $\sigma > 0$ and the polynomial kernel $f(x) = \sum_{i=0}^d a_i x^i$ with $d \geq 1$.

In the rest of this article, we will focus on the performance of LS-SVM, in the large n, p regime, by studying the asymptotic behavior of the decision function $g(\mathbf{x})$ defined in (3), in a binary classification problem with some statistical properties of the data, the model of which will be specified in the next section.

III. Main results

A. Model and assumptions

Evaluating the performance of LS-SVM is made difficult by the heavily data-driven aspect of the method. In this article, we assume that all $\mathbf{x}_i$'s are extracted from a Gaussian mixture, thereby allowing for a thorough theoretical analysis.

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 $n - n_1 = n_2$). We assume that $\mathbf{x}_i \in \mathcal{C}_a$ for $a \in \{1, 2\}$ if

$$\mathbf{x}_i = \boldsymbol{\mu}_a + \sqrt{p}\boldsymbol{\omega}_i$$

for some $\boldsymbol{\mu}_a \in \mathbb{R}^p$ and $\boldsymbol{\omega}_i \sim \mathcal{N}(0, \mathbf{C}_a/p)$, with $\mathbf{C}_a \in \mathbb{R}^{p\times p}$ some positive definite matrix.

As the $\boldsymbol{\mu}_a$'s and $\mathbf{C}_a$'s scale with p, to avoid asymptotic trivial misclassification rates (i.e., neither 0 or 1 in the limit of $n, p \to \infty$), we shall (as in [20,32]) technically place ourselves under the following controlled growth rate assumption:

Assumption 1 (Growth Rate)**.** *As $n \to \infty$, for $a \in \{1, 2\}$, the following conditions hold.*

- ***Data scaling***: $\frac{p}{n} \triangleq c_0 \to \bar{c}_0 > 0$.
- ***Class scaling***: $\frac{n_a}{n} \triangleq c_a \to \bar{c}_a > 0$.
- ***Mean scaling***: $\|\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1\| = O(1)$.
- ***Covariance scaling***: $\|\mathbf{C}_a\| = O(1)$ *and* $\operatorname{tr}(\mathbf{C}_2 - \mathbf{C}_1) = O(\sqrt{p})$.
- *for* $\mathbf{C}^\circ \triangleq \frac{n_1}{n}\mathbf{C}_1 + \frac{n_2}{n}\mathbf{C}_2$, $\frac{2}{p}\operatorname{tr}\mathbf{C}^\circ \to \tau > 0$ *as* $n, p \to \infty$.

From a practical aspect, where p and n are fixed quantities, the dual condition $n \to \infty$ and $\frac{p}{n} \to \bar{c}_0 > 0$ must be understood as requesting that both p and n be large and such that the ratio $\frac{p}{n}$ is sufficiently distinct from 0 and ∞.[4]

Aside from the last assumption, stated here mostly for technical convenience, it can be shown that the growth rate demanded in Assumption 1 is rate-optimal in the sense that an oracle Neyman–Pearson hypothesis testing procedure (with known $\boldsymbol{\mu}_a$ and $\mathbf{C}_a$) is (in general) ineffective at any smaller distance rates (so that the misclassification rate will constantly be 1), as discussed in the following remark.

Remark 1 (Optimal Growth Rate)**.** Assume that both $\|\mathbf{C}_a\|$ and $\|\mathbf{C}_a^{-1}\|$ are of order $O(1)$ and let $\mathbf{x}$ be a vector belonging to class $\mathcal{C}_1$, i.e., $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}_1, \mathbf{C}_1)$. Then, for perfectly known means $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ and covariances $\mathbf{C}_1, \mathbf{C}_2$, the Neyman–Pearson test for $\mathbf{x}$ to belong to $\mathcal{C}_1$ consists in the following comparison,

$$(\mathbf{x} - \boldsymbol{\mu}_2)^\mathsf{T}\mathbf{C}_2^{-1}(\mathbf{x} - \boldsymbol{\mu}_2) - (\mathbf{x} - \boldsymbol{\mu}_1)^\mathsf{T}\mathbf{C}_1^{-1}(\mathbf{x} - \boldsymbol{\mu}_1) \lessgtr \log\frac{\det\mathbf{C}_1}{\det\mathbf{C}_2}$$

which is further equivalent to

$$\begin{aligned}t(\mathbf{x}) \triangleq\ & \boldsymbol{\omega}^\mathsf{T}\left(\mathbf{C}_2^{-1} - \mathbf{C}_1^{-1}\right)\boldsymbol{\omega} + \frac{2}{\sqrt{p}}\Delta\boldsymbol{\mu}^\mathsf{T}\mathbf{C}_2^{-1}\boldsymbol{\omega} + \frac{1}{p}\Delta\boldsymbol{\mu}^\mathsf{T}\mathbf{C}_2^{-1}\Delta\boldsymbol{\mu} \\ & - \frac{1}{p}\log\frac{\det\mathbf{C}_1}{\det\mathbf{C}_2} \lessgtr 0\end{aligned}$$

where we denote $\Delta\boldsymbol{\mu} \triangleq \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$, $\boldsymbol{\omega} \triangleq \frac{1}{\sqrt{p}}(\mathbf{x} - \boldsymbol{\mu}_1)$ and thus $\boldsymbol{\omega} \sim \mathcal{N}(0, \mathbf{C}_1/p)$. To explore the difference in means $\Delta\boldsymbol{\mu}$ we take $\mathbf{C}_1 = \mathbf{C}_2 = \mathbf{C}$ and by Lyapunov's CLT [33, Theorem 27.3] we have, as $p \to \infty$,

$$t(\mathbf{x}) - \hat{t} \xrightarrow{d} 0.$$

where $\hat{t} \sim \mathcal{N}\left(\frac{1}{p}\Delta\boldsymbol{\mu}^\mathsf{T}\mathbf{C}^{-1}\Delta\boldsymbol{\mu}, \frac{2}{p}\Delta\boldsymbol{\mu}^\mathsf{T}\mathbf{C}^{-1}\Delta\boldsymbol{\mu}\right)$.

For a non-trivial classification rate, the mean of $\hat{t}$ must scale with p at least at the same rate as its standard deviation and thus, since $\|\mathbf{C}_a^{-1}\| = O(1)$, this implies that $\|\boldsymbol{\Delta\mu}\|$ be at least of order $O(1)$. Similar analysis can be performed to obtain the rate $\|\mathbf{C}_1 - \mathbf{C}_2\| = O(1/\sqrt{p})$ and consequently $\operatorname{tr}(\mathbf{C}_2 - \mathbf{C}_1) = O(\sqrt{p})$. We refer the readers to [32] for more discussions in this respect.

A key observation, also made in [20], is that, as a consequence of Assumption 1, for all pairs $i \neq j$,

$$\|\mathbf{x}_i - \mathbf{x}_j\|^2/p \xrightarrow{\text{a.s.}} \tau \tag{5}$$

and the convergence is even uniform across all $i \neq j$. This remark is the crux of all subsequent results (note that, surprisingly at first, it states that all data are essentially at the same distance from one another, irrespective of classes, and that the matrix $\mathbf{K}$ defined in (4) has all its entries essentially equal "in the limit" due to the the high dimensional nature of the data; this can be seen as a manifestation of the "curse

[2] Since data from $\mathcal{C}_1$ are labeled -1 while data from $\mathcal{C}_2$ are labeled 1.

[3] As shall be seen later, the division by p here is a convenient normalization in the large n, p regime. For example, we have the (normalized) norm $\|\mathbf{x}_i\|/\sqrt{p}$ is of order $O(1)$ with high probability for large n, p. The motivation of studying "translation invariant" kernel is that, being one of the most popular types of kernel used in practice, it offers (additionally) technical tractability as a result of the "concentration" phenomenon of large dimensional Gaussian vector, as we shall see later for example in (5). Similar results can be obtained for "inner-product" kernel of the type $f(\mathbf{x}_i^\mathsf{T}\mathbf{x}_j/p)$ as presented in [19,31].

[4] As a matter of fact, as our results will demonstrate, the case where $\frac{p}{n} \to \bar{c}_0 = 0$ is also valid as an extension by continuity through $\bar{c}_0 \to 0$.

of dimensionality" with respect to the Euclidean distance in high-dimensional space).

The function f defining the kernel matrix $\mathbf{K}$ in (4) shall be requested to satisfy the following assumption:

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

The objective of this article is to assess the performance of LS-SVM, under the setting of Assumptions 1 and 2, by studying the asymptotic behavior of the decision function $g(\mathbf{x})$ defined in (3). Following the work of [19] and [20], under our basic settings, the convergence in (5) makes it possible to linearize the kernel matrix $\mathbf{K}$ around the matrix $f(\tau)\mathbf{1}_n\mathbf{1}_n^\mathsf{T}$, and thus the intractable nonlinear kernel matrix $\mathbf{K}$ can be asymptotically linearized in the large n, p regime. As such, since the decision function $g(\mathbf{x})$ is explicitly defined as a function of $\mathbf{K}$ (through $\boldsymbol{\alpha}$ and b as defined in (2)), one can work out an asymptotic linearization of $g(\mathbf{x})$ as a function of the kernel function f and the statistics of the data. This analysis, presented in detail in Appendix A of the Supplementary Material, allows one to reveal the relationship between the performance of LS-SVM and the kernel function f as well as the given learning task, for Gaussian input data as $n, p \to \infty$, as presented in the following subsection.

B. Asymptotic behavior of the decision function $g(\mathbf{x})$

Before going into our main results, a few notations need to be introduced. In the remainder of the article, we shall use the following deterministic and random elements notations:

$$\mathbf{P} \triangleq \mathbf{I}_n - \mathbf{1}_n\mathbf{1}_n^\mathsf{T}/n \in \mathbb{R}^{n\times n}, \quad \boldsymbol{\Omega} \triangleq [\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_n] \in \mathbb{R}^{p\times n}$$
$$\boldsymbol{\psi} \triangleq \left\{\|\boldsymbol{\omega}_i\|^2 - \mathbb{E}\left[\|\boldsymbol{\omega}_i\|^2\right]\right\}_{i=1}^n \in \mathbb{R}^n.$$

Under Assumptions 1 and 2, following up [20], one can approximate the kernel matrix $\mathbf{K}$ by $\hat{\mathbf{K}}$ in such a way that

$$\|\mathbf{K} - \hat{\mathbf{K}}\| \xrightarrow{\text{a.s.}} 0$$

with $\hat{\mathbf{K}} = -2f'(\tau)(\mathbf{M} + \mathbf{V}\mathbf{V}^\mathsf{T}) + (f(0) - f(\tau) + \tau f'(\tau))\,\mathbf{I}_n$ for some matrices $\mathbf{M}$ and $\mathbf{V}$, where $\mathbf{M}$ is a standard random matrix model (of operator norm $O(1)$) and $\mathbf{V}\mathbf{V}^\mathsf{T}$ a small rank matrix (of operator norm $O(n)$), which depends both on $\mathbf{P}, \boldsymbol{\Omega}, \boldsymbol{\psi}$ and on the class statistics $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ and $\mathbf{C}_1, \mathbf{C}_2$. The same analysis is applied to the vector $\mathbf{k}(\mathbf{x})$ by similarly defining the following random variables for a new datum $\mathbf{x} \in \mathcal{C}_a$, $a \in \{1, 2\}$:

$$\boldsymbol{\omega}_\mathbf{x} \triangleq (\mathbf{x} - \boldsymbol{\mu}_a)/\sqrt{p} \in \mathbb{R}^p, \quad \psi_\mathbf{x} \triangleq \|\boldsymbol{\omega}_\mathbf{x}\|^2 - \mathbb{E}\left[\|\boldsymbol{\omega}_\mathbf{x}\|^2\right] \in \mathbb{R}.$$

Based on the (operator norm) approximation $\mathbf{K} \approx \hat{\mathbf{K}}$, a Taylor expansion is then performed on $\mathbf{S}^{-1} = (\mathbf{K} + n\mathbf{I}_n/\gamma)^{-1}$ to obtain an (asymptotic) approximation of $\mathbf{S}^{-1}$, and subsequently on $\boldsymbol{\alpha}$ and b which depend explicitly on $\mathbf{S}^{-1}$. At last, plugging these results into (3), one finds the main technical result of this article as follows.

Theorem 1 (Asymptotic Approximation)**.** *Let Assumptions 1 and 2 hold, and $g(\mathbf{x})$ be defined by* (3)*. Then, as $n, p \to \infty$, $n(g(\mathbf{x}) - \hat{g}(\mathbf{x})) \xrightarrow{\text{a.s.}} 0$, where*

$$\hat{g}(\mathbf{x}) = \begin{cases} c_2 - c_1 + \gamma\left(\mathfrak{P} - 2c_1c_2^2\mathfrak{D}\right), & \text{if } \mathbf{x} \in \mathcal{C}_1 \\ c_2 - c_1 + \gamma\left(\mathfrak{P} + 2c_1^2c_2\mathfrak{D}\right), & \text{if } \mathbf{x} \in \mathcal{C}_2 \end{cases} \tag{6}$$

with

$$\mathfrak{P} = -\frac{2f'(\tau)}{n}\mathbf{y}^\mathsf{T}\mathbf{P}\boldsymbol{\Omega}^\mathsf{T}\boldsymbol{\omega}_\mathbf{x} - \frac{4c_1c_2f'(\tau)}{\sqrt{p}}\left(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1\right)^\mathsf{T}\boldsymbol{\omega}_\mathbf{x} + 2c_1c_2f''(\tau)\psi_\mathbf{x}\frac{\operatorname{tr}\left(\mathbf{C}_2 - \mathbf{C}_1\right)}{p} \tag{7}$$

$$\mathfrak{D} = -\frac{2f'(\tau)}{p}\|\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1\|^2 + \frac{f''(\tau)}{p^2}\left(\operatorname{tr}\left(\mathbf{C}_2 - \mathbf{C}_1\right)\right)^2 + \frac{2f''(\tau)}{p^2}\operatorname{tr}\left(\left(\mathbf{C}_2 - \mathbf{C}_1\right)^2\right). \tag{8}$$

Leaving the proof to Appendix A in the Supplementary Material, Theorem 1 tells us that the decision function $g(\mathbf{x})$ has an asymptotic equivalent $\hat{g}(\mathbf{x})$ that consists of three parts:

1) the deterministic term $c_2 - c_1$ of order $O(1)$ that depends on the number of instances in each class of the training set, which essentially comes from the term $\mathbf{1}_n^\mathsf{T}\mathbf{y}/n$ in b;
2) the "noisy" term $\mathfrak{P}$ of order $O(n^{-1})$ which is a function of the zero mean random variables $\boldsymbol{\omega}_\mathbf{x}$ and $\psi_\mathbf{x}$, thus in particular $\mathbb{E}[\mathfrak{P}] = 0$;
3) the "informative" term containing $\mathfrak{D}$, also of order $O(n^{-1})$, which features the deterministic differences between the two classes.

From Theorem 1, under the basic settings of Assumption 1, for Gaussian data $\mathbf{x} \in \mathcal{C}_a$, $a \in \{1, 2\}$, we can show that $\hat{g}(\mathbf{x})$ (and therefore $g(\mathbf{x})$) converges to a random Gaussian variable the mean and variance of which are given in the following theorem. The proof is deferred to Appendix B.

Theorem 2 (Gaussian Approximation)**.** *Under the setting of Theorem 1, $n(g(\mathbf{x}) - G_a) \xrightarrow{d} 0$, where*

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

with

$$\mathrm{E}_a = \begin{cases} c_2 - c_1 - 2c_2 \cdot c_1c_2\gamma\mathfrak{D}, & a = 1 \\ c_2 - c_1 + 2c_1 \cdot c_1c_2\gamma\mathfrak{D}, & a = 2 \end{cases}$$
$$\mathrm{Var}_a = 8\gamma^2c_1^2c_2^2\left(\mathcal{V}_1^a + \mathcal{V}_2^a + \mathcal{V}_3^a\right)$$

and

$$\mathcal{V}_1^a = \frac{\left(f''(\tau)\right)^2}{p^4}\left(\operatorname{tr}\left(\mathbf{C}_2 - \mathbf{C}_1\right)\right)^2\operatorname{tr}\mathbf{C}_a^2$$
$$\mathcal{V}_2^a = \frac{2\left(f'(\tau)\right)^2}{p^2}\left(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1\right)^\mathsf{T}\mathbf{C}_a\left(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1\right)$$
$$\mathcal{V}_3^a = \frac{2\left(f'(\tau)\right)^2}{np^2}\left(\frac{\operatorname{tr}\mathbf{C}_1\mathbf{C}_a}{c_1} + \frac{\operatorname{tr}\mathbf{C}_2\mathbf{C}_a}{c_2}\right).$$

Theorem 2 is our main practical result as it allows one to evaluate the large n, p performance of LS-SVM for Gaussian data. While dwelling on the implications of Theorem 1 and 2, several remarks and discussions are in order.

Remark 2 (Dominant Bias)**.** From Theorem 1, under the key Assumption 1, both the random noise $\mathfrak{P}$ and the deterministic "informative" term $\mathfrak{D}$ are of order $O(n^{-1})$, which means that the decision function $g(\mathbf{x}) = c_2 - c_1 + O(n^{-1})$. This result somehow contradicts the classical decision criterion proposed in [2], based on the sign of $g(\mathbf{x})$, i.e., $\mathbf{x}$ is associated to class $\mathcal{C}_1$ if $g(\mathbf{x}) < 0$ and to class $\mathcal{C}_2$ otherwise. When $c_1 \neq c_2$,

this would lead to an asymptotic classification of all new data $\mathbf{x}$'s in the same class as $n \to \infty$. Practically speaking, this means for n, p large that the decision function $g(\mathbf{x})$ of a new datum $\mathbf{x}$ lies (sufficiently) away from 0 (0 being the classically considered threshold), so that the sign of $g(\mathbf{x})$ is constantly positive (in the case of $\bar{c}_2 > \bar{c}_1$) or negative (in the case of $\bar{c}_2 < \bar{c}_1$). As such, all new data will be trivially classified into the same class. Instead, a first result of Theorem 1 is that the decision threshold ξ should be taken as $\xi = \xi_n = c_2 - c_1 + O(n^{-1})$ for imbalanced classification problem.

The conclusion of Remark 2 was in fact already known since the work of [15] who reached the same conclusion through a Bayesian inference analysis, *for all finite* n, p. From their Bayesian perspective, the term $c_2 - c_1$ appears in the "bias term" b under the form of prior class probabilities $\mathrm{P}(y = -1)$, $\mathrm{P}(y = 1)$ and allows for adjusting classification problems with different prior class probabilities in the training and test sets. This idea of a (static) bias term correction has also been applied in [34] in order to improve the validation set performance. Here we confirm the problem of imbalanced datasets in Remark 2 by Figure 1 with $c_1 = 1/4$ and $c_2 = 3/4$, where the histograms of $g(\mathbf{x})$ for $\mathbf{x} \in \mathcal{C}_1$ and $\mathcal{C}_2$ center somewhere close to $c_2 - c_1 = 0.5$, thus resulting in a trivial classification by assigning all new data to $\mathcal{C}_2$ if one takes $\xi = 0$ because $\mathrm{P}(g(\mathbf{x}) < \xi \mid \mathbf{x} \in \mathcal{C}_1) \to 0$ and $\mathrm{P}(g(\mathbf{x}) > \xi \mid \mathbf{x} \in \mathcal{C}_2) \to 1$ as $n, p \to \infty$ (the convergence being in fact an equality for finite n, p in this particular figure).

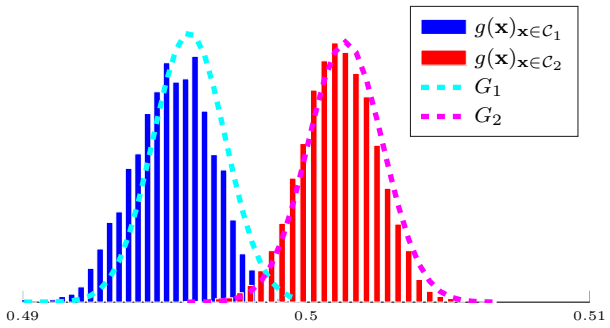


Fig. 1. Gaussian approximation of $g(\mathbf{x})$, $n = 256$, $p = 512$, $c_1 = 1/4, c_2 = 3/4, \gamma = 1$, Gaussian kernel with $\sigma^2 = 1$, $\mathbf{x} \in \mathcal{N}(\boldsymbol{\mu}_a, \mathbf{C}_a)$ with $\boldsymbol{\mu}_a = [\mathbf{0}_{a-1}; 3; \mathbf{0}_{p-a}]$, $\mathbf{C}_1 = \mathbf{I}_p$ and $\{\mathbf{C}_2\}_{i,j} = .4^{|i-j|}(1 + 5/\sqrt{p})$.

An alternative to alleviate this imbalance issue is to normalize the label vector $\mathbf{y}$. From the proof of Theorem 1 in Appendix A we see the term $c_2 - c_1$ is due to the fact that in b one has $\mathbf{1}_n^\mathsf{T}\mathbf{y}/n = c_2 - c_1 \neq 0$. Thus, one may normalize the labels y_i as $y_i^* = -1/c_1$ if $\mathbf{x}_i \in \mathcal{C}_1$ and $y_i^* = 1/c_2$ if $\mathbf{x}_i \in \mathcal{C}_2$, so that the relation $\mathbf{1}_n^\mathsf{T}\mathbf{y}^* = 0$ is satisfied. This formulation is also referred to as the *Fishers targets*: $\{-n/n_1, n/n_2\}$ in the context of kernel fisher discriminant analysis [35,36]. With the aforementioned normalized labels $\mathbf{y}^*$, we have the following lemma that reveals the connection between the corresponding decision function $g^*(\mathbf{x})$ and $g(\mathbf{x})$.

Lemma 1. *Let* $g(\mathbf{x})$ *be defined by* (3) *and* $g^*(\mathbf{x})$ *be defined as* $g^*(\mathbf{x}) = (\boldsymbol{\alpha}^*)^\mathsf{T}\mathbf{k}(\mathbf{x}) + b^*$, *with* $(\boldsymbol{\alpha}^*, b^*)$ *given by* (2) *for* $\mathbf{y}^*$ *in the place of* $\mathbf{y}$, *where* $y_i^* = -1/c_1$ *if* $\mathbf{x}_i \in \mathcal{C}_1$ *and* $y_i^* = 1/c_2$ *if* $\mathbf{x}_i \in \mathcal{C}_2$. *Then,*

$$g(\mathbf{x}) - (c_2 - c_1) = 2c_1c_2 g^*(\mathbf{x}).$$

Proof: From (2) and (3) we get

$$g(x) = \mathbf{y}^\mathsf{T}\left(\mathbf{S}^{-1} - \frac{\mathbf{S}^{-1}\mathbf{1}_n\mathbf{1}_n^\mathsf{T}\mathbf{S}^{-1}}{\mathbf{1}_n^\mathsf{T}\mathbf{S}^{-1}\mathbf{1}_n}\right)\mathbf{k}(\mathbf{x}) + \frac{\mathbf{y}^\mathsf{T}\mathbf{S}^{-1}\mathbf{1}_n}{\mathbf{1}_n^\mathsf{T}\mathbf{S}^{-1}\mathbf{1}_n} = \mathbf{y}^\mathsf{T}\boldsymbol{\varpi}$$

with $\boldsymbol{\varpi} = \left(\mathbf{S}^{-1} - \frac{\mathbf{S}^{-1}\mathbf{1}_n\mathbf{1}_n^\mathsf{T}\mathbf{S}^{-1}}{\mathbf{1}_n^\mathsf{T}\mathbf{S}^{-1}\mathbf{1}_n}\right)\mathbf{k}(\mathbf{x}) + \frac{\mathbf{S}^{-1}\mathbf{1}_n}{\mathbf{1}_n^\mathsf{T}\mathbf{S}^{-1}\mathbf{1}_n}$. Besides, note that $\mathbf{1}_n^\mathsf{T}\boldsymbol{\varpi} = 1$. We thus have

$$\begin{aligned} g(\mathbf{x}) - (c_2 - c_1) &= \mathbf{y}^\mathsf{T}\boldsymbol{\varpi} - (c_2 - c_1)\mathbf{1}_n^\mathsf{T}\boldsymbol{\varpi} \\ &= 2c_1c_2\left(\frac{\mathbf{y} - (c_2 - c_1)\mathbf{1}_n}{2c_1c_2}\right)^\mathsf{T}\boldsymbol{\varpi} \\ &= 2c_1c_2(\mathbf{y}^*)^\mathsf{T}\boldsymbol{\varpi} = 2c_1c_2 g^*(\mathbf{x}) \end{aligned}$$

which concludes the proof. ∎

As a consequence of Lemma 1, instead of Theorem 2 for standard labels $\mathbf{y}$, one would have the following corollary for the corresponding Gaussian approximation of $g^*(\mathbf{x})$ when normalized labels $\mathbf{y}^*$ are used.

Corollary 1 (Gaussian Approximation of $g^*(\mathbf{x})$)**.** *Under the setting of Theorem 1, and with* $g^*(\mathbf{x})$ *defined in Lemma 1,* $n(g^*(\mathbf{x}) - G_a^*) \xrightarrow{d} 0$, *where*

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

with

$$\mathrm{E}_a^* = \begin{cases} -c_2\gamma\mathfrak{D}, & a = 1 \\ +c_1\gamma\mathfrak{D}, & a = 2 \end{cases}$$
$$\mathrm{Var}_a^* = 2\gamma^2\left(\mathcal{V}_1^a + \mathcal{V}_2^a + \mathcal{V}_3^a\right)$$

and $\mathfrak{D}$ *is defined by* (8), $\mathcal{V}_1^a, \mathcal{V}_2^a$ *and* $\mathcal{V}_3^a$ *as in Theorem 2.*

Figure 2 illustrates this result in the same settings as Figure 1. Compared to Figure 1, one can observe that in Figure 2 both histograms are now centered close to 0 (at distance $O(n^{-1})$ from zero) instead of $c_2 - c_1 = 1/2$. Still, even in the case where normalized labels $\mathbf{y}^*$ are used as observed in Figure 2 (where the histograms cross at about $-0.004 \approx 1/n$), taking $\xi = 0$ as a decision threshold may not be an appropriate choice, as $\mathrm{E}_1^* \neq -\mathrm{E}_2^*$.

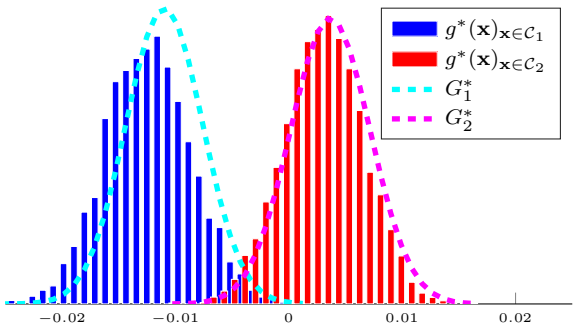


Fig. 2. Gaussian approximation of $g^*(\mathbf{x})$, $n = 256$, $p = 512$, $c_1 = 1/4, c_2 = 3/4, \gamma = 1$, Gaussian kernel with $\sigma^2 = 1$, $\mathbf{x} \in \mathcal{N}(\boldsymbol{\mu}_a, \mathbf{C}_a)$, with $\boldsymbol{\mu}_a = [\mathbf{0}_{a-1}; 3; \mathbf{0}_{p-a}]$, $\mathbf{C}_1 = \mathbf{I}_p$ and $\{\mathbf{C}_2\}_{i,j} = .4^{|i-j|}(1 + 5/\sqrt{p})$.

Remark 3 (Insignificance of γ)**.** As a direct result of Theorem 1 and Remark 2, note in (6) that $\hat{g}(\mathbf{x}) - (c_2 - c_1)$ is proportional to the hyperparameter γ, which indicates that, rather surprisingly, the tuning of γ is (asymptotically) of no importance when $n, p \to \infty$ since it does not alter the classification statistics when one uses the sign of $g(\mathbf{x}) - (c_2 - c_1)$ for the decision.[5]

Letting $Q(x) = \frac{1}{2\pi}\int_x^\infty \exp\left(-t^2/2\right) dt$, from Theorem 2 and Corollary 1, we now have the following immediate corollary for the (asymptotic) classification error rate.

Corollary 2 (Asymptotic Error Rate)**.** *Under the setting of Theorem 1, for a threshold ξ_n possibly depending on n, as $n \to \infty$,*

$$\mathrm{P}(g(\mathbf{x}) > \xi_n \mid \mathbf{x} \in \mathcal{C}_1) - Q\left(\frac{\xi_n - \mathrm{E}_1}{\sqrt{\mathrm{Var}_1}}\right) \to 0 \tag{9}$$

$$\mathrm{P}(g(\mathbf{x}) < \xi_n \mid \mathbf{x} \in \mathcal{C}_2) - Q\left(\frac{\mathrm{E}_2 - \xi_n}{\sqrt{\mathrm{Var}_2}}\right) \to 0 \tag{10}$$

with E_a and Var_a given in Theorem 2.

Obviously, Corollary 2 is only meaningful when $\xi_n = c_2 - c_1 + O(n^{-1})$ as recalled earlier. Besides, it is clear from Lemma 1 and Corollary 1 that $\mathrm{P}(g(\mathbf{x}) > \xi_n \mid \mathbf{x} \in \mathcal{C}_a) = \mathrm{P}(g^*(\mathbf{x}) > \xi_n - (c_2 - c_1) \mid \mathbf{x} \in \mathcal{C}_a)$, so that Corollary 2 extends naturally to $g^*(\mathbf{x})$ when normalized labels $\mathbf{y}^*$ are applied.

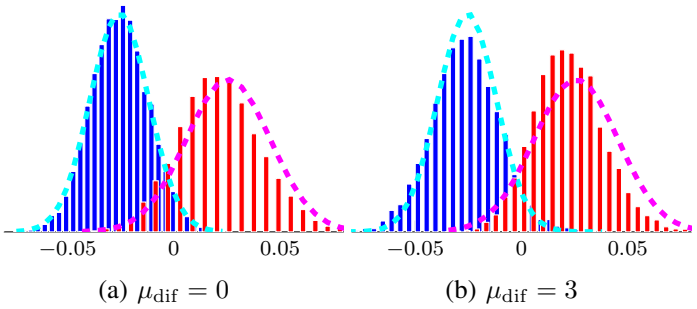
Corollary 2 allows one to compute the asymptotic misclassification rate as a function of $\mathrm{E_a}, \mathrm{Var_a}$ and the threshold ξ_n. Combined with Theorem 2, one may note the significance of a proper choice of the kernel function f. For instance, if $f'(\tau) = 0$, the term $\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1$ vanishes from the mean and variance of G_a, meaning that the classification of LS-SVM will not rely (at least asymptotically and under Assumption 1) on the differences in means of the two classes. Figure 3 corroborates this finding with the same theoretical Gaussian approximations G_1 and G_2 in subfigures (a) and (b). When $\|\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1\|^2$ varies from 0 in (a) to 18 in (b), the distribution of $g(\mathbf{x})$, and in particular, the overlap between two classes, remain almost the same in (a) and (b).

More traceable special cases and discussions on the choice of kernel function f will be given in the next section.

IV. Special Cases and Further Discussions

A. More discussions on the kernel function f

Following the discussion at the end of Section III, if $f'(\tau) = 0$, the information about the statistical means of the two different classes is lost and will not help perform the classification. Nonetheless, we find that, rather surprisingly, if one further assumes $\mathrm{tr}\,\mathbf{C}_1 = \mathrm{tr}\,\mathbf{C}_2 + o(\sqrt{p})$ (which is beyond the minimum "distance" rate in Assumption 1), using a kernel f that satisfies $f'(\tau) = 0$ results in $\mathrm{Var}_a = 0$ while E_a may remain non-zero, thereby ensuring a vanishing misclassification rate (as long as $f''(\tau) \neq 0$). Intuitively speaking, the kernels with $f'(\tau) = 0$ play an important role in extracting the information of "shape" of both classes, making the classification extremely accurate even in cases that are deemed impossible to classify according to Remark 1. This phenomenon was also remarked in [20] and deeply investigated in [22]. Figure 4 substantiates this finding for $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$, $\mathbf{C}_1 = \mathbf{I}_p$ and $\{\mathbf{C}_2\}_{i,j} = .4^{|i-j|}$, for which $\mathrm{tr}\,\mathbf{C}_1 = \mathrm{tr}\,\mathbf{C}_2 = p$. We observe a rapid drop of the classification error as $f'(\tau)$ gets close to 0.



(a) $\mu_{\mathrm{dif}} = 0$ (b) $\mu_{\mathrm{dif}} = 3$

Fig. 3. Gaussian approximation of $g(\mathbf{x})$, $n = 256$, $p = 512$, $c_1 = c_2 = 1/2, \gamma = 1$, polynomial kernel with $f(\tau) = 4$, $f'(\tau) = 0$, and $f''(\tau) = 2$. $\mathbf{x} \in \mathcal{N}(\boldsymbol{\mu}_a, \mathbf{C}_a)$, with $\boldsymbol{\mu}_a = [\mathbf{0}_{a-1}; \mu_{\mathrm{dif}}; \mathbf{0}_{p-a}]$, $\mathbf{C}_1 = \mathbf{I}_p$ and $\{\mathbf{C}_2\}_{i,j} = .4^{|i-j|}(1 + 5/\sqrt{p})$.

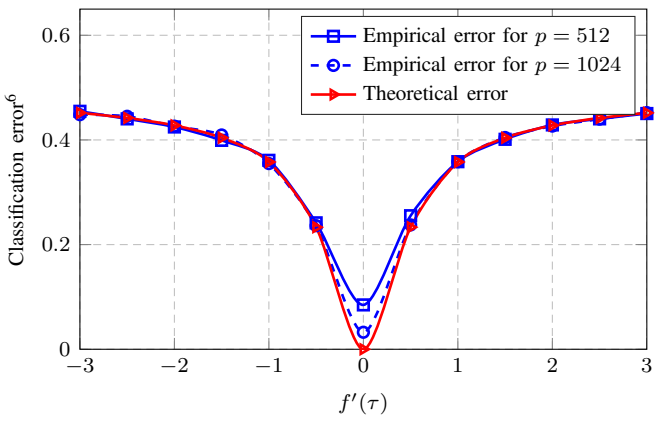


Fig. 4. Performance of LS-SVM, $c_0 = 1/4$, $c_1 = c_2 = 1/2, \gamma = 1$, polynomial kernel with $f(\tau) = 4$, $f''(\tau) = 2$. $\mathbf{x} \in \mathcal{N}(\boldsymbol{\mu}_a, \mathbf{C}_a)$, with $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \mathbf{0}_p$, $\mathbf{C}_1 = \mathbf{I}_p$ and $\{\mathbf{C}_2\}_{i,j}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

Remark 4 (Condition on Kernel Function f)**.** From Theorem 2 and Corollary 1, one observes that $|\mathrm{E}_1 - \mathrm{E}_2|$ is always proportional to the "informative" term $\mathfrak{D}$ and should, for fixed Var_a, be made as large as possible to avoid the overlap of $g(\mathbf{x})$ for $\mathbf{x}$ from different classes. Since Var_a does not depend on the signs of $f'(\tau)$ and $f''(\tau)$, it is easily deduced that, to achieve optimal classification performance, one needs to

[5]This remark is only valid only under Assumption 1 and $\gamma = O(1)$, i.e., γ is considered to remain a constant as $n, p \to \infty$. Recall that this is in sharp contrast with [13] where $\gamma = O(\sqrt{n})$ (or $O(n)$, depending on the problem) is claimed optimal in the large n only regime. From Remark 1 on the growth rate optimality reached by LS-SVM, we see here that $\gamma = O(1)$ is rate-optimal under the present large n, p setting; yet we believe that more elaborate kernels (such as those explored in [37]) may allow for improved performances (not in the rate but in the constants), possibly for different scales of γ. This intuition will be explored in future investigations.

[6]Unless particularly stated, the classification error will be understood as $c_1\mathrm{P}(\mathrm{g}(\mathbf{x}) > \xi_\mathrm{n} | \mathbf{x} \in \mathcal{C}_1) + c_2\mathrm{P}(\mathrm{g}(\mathbf{x}) < \xi_\mathrm{n} | \mathbf{x} \in \mathcal{C}_2)$.

choose the kernel function f such that $f(\tau) > 0, f'(\tau) < 0$ and $f''(\tau) > 0$.

Incidentally, the condition in Remark 4 is naturally satisfied for Gaussian kernel $f(x) = \exp\left(-x/(2\sigma^2)\right)$ for any σ, meaning that, even without specific tuning of the kernel parameter σ through cross validation or other techniques, LS-SVM is expected to perform rather well with a Gaussian kernel (as shown in Figure 5), which is not always the case for polynomial kernels. This especially entails, for a second-order polynomial kernel given by $f(x) = a_2x^2 + a_1x + a_0$, that attention should be paid to meeting the aforementioned condition when tuning the kernel parameters a_2, a_1 and a_0. Figure 6 attests of this remark with Gaussian input data. A rapid increase in classification error rate can be observed both in theory and in practice as soon as the condition $f'(\tau) < 0, f''(\tau) > 0$ is no longer satisfied.

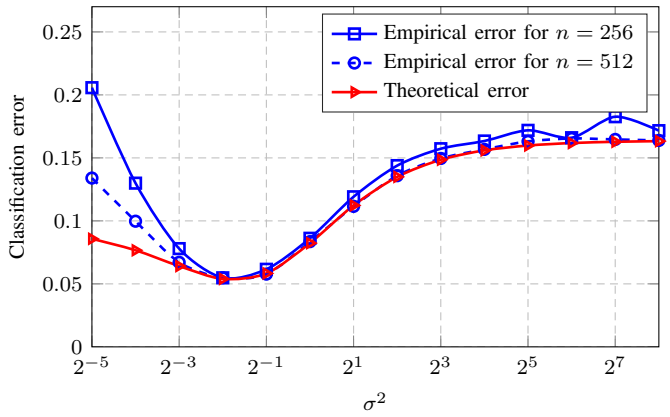
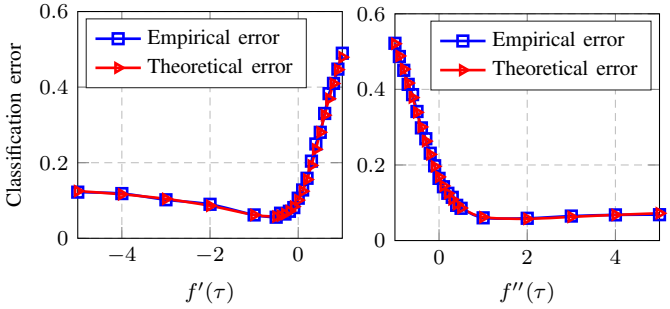


Fig. 5. Performance of LS-SVM, $c_0 = 2, c_1 = c_2 = 1/2, \gamma = 1$, Gaussian kernel. $\mathbf{x} \in \mathcal{N}(\boldsymbol{\mu}_a, \mathbf{C}_a)$, with $\boldsymbol{\mu}_a = [\mathbf{0}_{a-1}; 2; \mathbf{0}_{p-a}]$, $\mathbf{C}_1 = \mathbf{I}_p$ and $\{\mathbf{C}_2\}_{i,j} = .4^{|i-j|}(1 + 4/\sqrt{p})$.

Note also from both Figure 4 and Figure 5 that, when n, p are doubled (from $2048, 512$ to $4\,096, 1\,024$ in Figure 4 and from $256, 512$ to $512, 1\,024$ in Figure 5), the empirical error becomes closer to the theoretical one, which confirms the asymptotic result as $n, p \to \infty$.



(a) $f(\tau) = 4,\ f''(\tau) = 1$ (b) $f(\tau) = 4,\ f'(\tau) = -1$

Fig. 6. Performance of LS-SVM, $n = 256,\ p = 512,\ c_1 = c_2 = 1/2, \gamma = 1$, polynomial kernel. $\mathbf{x} \in \mathcal{N}(\boldsymbol{\mu}_a, \mathbf{C}_a)$, with $\boldsymbol{\mu}_a = [\mathbf{0}_{a-1}; 2; \mathbf{0}_{p-a}]$, $\mathbf{C}_1 = \mathbf{I}_p$ and $\{\mathbf{C}_2\}_{i,j} = .4^{|i-j|}(1 + 4/\sqrt{p})$.

Clearly, for practical use, one needs to know in advance the value of τ before training so that the kernel f can be properly chosen during the training step. The estimation of τ is possible, in the large n, p regime, with the following lemma.

Lemma 2. *Under Assumptions 1 and 2, as $n \to \infty$,*

$$\frac{2}{n}\sum_{i=1}^{n}\frac{\|\mathbf{x}_i - \bar{\mathbf{x}}\|^2}{p} \xrightarrow{\text{a.s.}} \tau \tag{11}$$

with $\bar{\mathbf{x}} \triangleq \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i$.

Proof: Since

$$\frac{2}{n}\sum_{i=1}^{n}\frac{\|\mathbf{x}_i - \bar{\mathbf{x}}\|^2}{p} = \frac{2c_1c_2\|\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1\|^2}{p} + \frac{2}{n}\sum_{i=1}^{n}\|\boldsymbol{\omega}_i - \bar{\boldsymbol{\omega}}\|^2 + \kappa$$

with $\kappa = \frac{4}{n\sqrt{p}}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^{\mathsf{T}}\left(-c_2\sum_{\mathbf{x}_i \in \mathcal{C}_1}\boldsymbol{\omega}_i + c_1\sum_{\mathbf{x}_j \in \mathcal{C}_2}\boldsymbol{\omega}_j\right)$ and $\bar{\boldsymbol{\omega}} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\omega}_i$.

According to Assumption 1 we have $\frac{2c_1c_2}{p}\|\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1\|^2 = O(n^{-1})$. The term κ is a linear combination of independent zero-mean Gaussian variables and thus $\kappa \sim \mathcal{N}(0, \mathrm{Var}[\kappa])$ with $\mathrm{Var}[\kappa] = \frac{16c_1c_2}{np^2}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^{\mathsf{T}}(c_2\mathbf{C}_1 + c_1\mathbf{C}_2)(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) = O(n^{-3})$. We thus deduce from Chebyshev's inequality and Borel-Cantelli lemma that $\kappa \xrightarrow{\text{a.s.}} 0$.

We then work on the last term $\frac{2}{n}\sum_{i=1}^{n}\|\boldsymbol{\omega}_i - \bar{\boldsymbol{\omega}}\|^2$ as

$$\frac{2}{n}\sum_{i=1}^{n}\|\boldsymbol{\omega}_i - \bar{\boldsymbol{\omega}}\|^2 = \frac{2}{n}\sum_{i=1}^{n}\|\boldsymbol{\omega}_i\|^2 - 2\|\bar{\boldsymbol{\omega}}\|^2.$$

Since $\bar{\boldsymbol{\omega}} \sim \mathcal{N}(\mathbf{0}, \mathbf{C}^\circ/np)$, we deduce that $\|\bar{\boldsymbol{\omega}}\|^2 \xrightarrow{\text{a.s.}} 0$. Ultimately by the strong law of large numbers, the term $\frac{2}{n}\sum_{i=1}^{n}\|\boldsymbol{\omega}_i\|^2 \xrightarrow{\text{a.s.}} \tau$, which concludes the proof. ∎

B. Some limiting cases

1) Dominant information in means: When $\|\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1\|^2$ is largely dominant over $(\mathrm{tr}(\mathbf{C}_2 - \mathbf{C}_1))^2/p$ and $\mathrm{tr}((\mathbf{C}_2 - \mathbf{C}_1)^2)/p$, from Theorem 2, both $\mathrm{E_a} - (\mathrm{c_2} - \mathrm{c_1})$ and $\sqrt{\mathrm{Var_a}}$ are (approximately) proportional to $f'(\tau)$, which eventually makes the choice of the kernel irrelevant (as long as $f'(\tau) \neq 0$). This result also holds true for $\mathrm{E_a^*}$ and $\sqrt{\mathrm{Var_a^*}}$ when normalized labels $\mathbf{y}^*$ are applied, as a result of Lemma 1.

2) c_0 large or small: Note that, different from both $\mathcal{V}_1$ and $\mathcal{V}_2$, $\mathcal{V}_3$ is a function of c_0 as it can be rewritten as

$$\mathcal{V}_3^a = \frac{2c_0\left(f'(\tau)\right)^2}{p^3}\left(\frac{\mathrm{tr}\,\mathbf{C}_1\mathbf{C}_a}{c_1} + \frac{\mathrm{tr}\,\mathbf{C}_2\mathbf{C}_a}{c_2}\right)$$

which indicates that the variance of $g(\mathbf{x})$ grows as c_0 becomes large. This result is easily understood since, with p fixed, a small c_0 means a larger n, and with more training samples, one may "gain" more information of the two different classes, which reduces the "uncertainty" of the classifier. When $n \to \infty$ with a fixed p, we have $c_0 \to 0$ and the LS-SVM is considered "well-trained" and its performance can be described with Theorem 2 by taking $\mathcal{V}_3 = 0$. However, it is worthy noting that the misclassification rate may not be 0 even in this case, since $\mathcal{V}_1$ and $\mathcal{V}_2$ may differ from 0, which indicated the theoretical limitation of LS-SVM in separating high dimensional Gaussian vectors. On the contrary,

when $c_0 \to \infty$, with few training data, LS-SVM does not sample sufficiently the high dimensional space of the $\mathbf{x}$'s, thus resulting in a classifier with arbitrarily large variance (for fixed means). Moreover, since the term $\mathcal{V}_3^a$ is proportional to n^{-1}, we see that for $f'(\tau)$ away from zero and fixed large p, as n grows large, the two Gaussians G_1 and G_2 in Theorem 2 separate from each other at a rate of $n^{-\frac{1}{2}}$, the overlapping section of the Gaussian tails then provides the misclassification rate via Corollary 2. Figure 7 confirms this result with p fixed to 256 while n varies from 8 to 8 192.

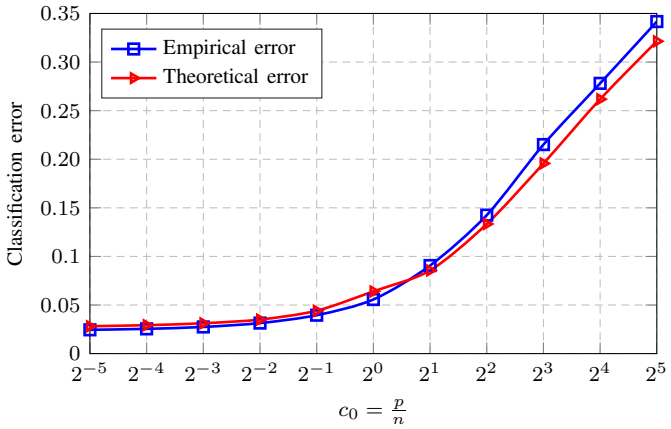


Fig. 7. Performance of LS-SVM, $p = 256$, $c_1 = c_2 = 1/2, \gamma = 1$, Gaussian kernel with $\sigma^2 = 1$. $\mathbf{x} \in \mathcal{N}(\boldsymbol{\mu}_a, \mathbf{C}_a)$, with $\boldsymbol{\mu}_a = [\mathbf{0}_{a-1}; 2; \mathbf{0}_{p-a}]$, $\mathbf{C}_1 = \mathbf{I}_p$ and $\{\mathbf{C}_2\}_{i,j} = .4^{|i-j|}(1 + 4/\sqrt{p})$.

3) $c_1 \to 0$ *:* As revealed in Remark 2, the ratio c_1/c_2 plays a significant role in the performance of classification. A natural question arises: what happens when one class is strongly dominant over the other? Take the case of $c_1 \to 0, c_2 \to 1$. From Corollary 1, one has $\mathrm{E}_1^* \to -\gamma\mathfrak{D}$, $\mathrm{E}_2^* \to 0$ and $\mathcal{V}_3^a \to \infty$ because of $c_1 \to 0$ in the denominator, which then makes the ratio $\frac{\mathrm{E}_a^*}{\sqrt{\mathrm{Var}_a^*}}$ (and thus $\frac{\mathrm{E}_a - (c_2 - c_1)}{\sqrt{\mathrm{Var}_a}}$) go to zero, resulting in a poorly-performing LS-SVM. The same occurs when $c_1 \to 1$ and $c_2 \to 0$. Figure 8 collaborates this remark with $c_1 = 1/32$ in subfigure (a) and $1/2$ in (b). Note that in subfigure (a), even with a smartly chosen threshold ξ, LS-SVM is impossible to perform as well as in the case $c_1 = c_2$, as a result of the significant overlap between the two histograms.

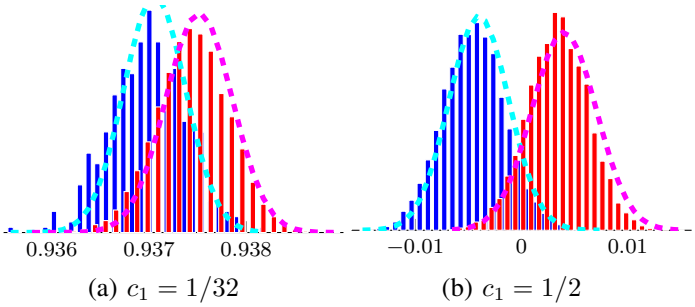


Fig. 8. Gaussian approximation of $g(\mathbf{x})$, $n = 256$, $p = 512$, $c_2 = 1 - c_1, \gamma = 1$, Gaussian kernel with $\sigma^2 = 1$. $\mathbf{x} \in \mathcal{N}(\boldsymbol{\mu}_a, \mathbf{C}_a)$, with $\boldsymbol{\mu}_a = [\mathbf{0}_{a-1}; 2; \mathbf{0}_{p-a}]$, $\mathbf{C}_1 = \mathbf{I}_p$ and $\{\mathbf{C}_2\}_{i,j} = .4^{|i-j|}(1 + 4/\sqrt{p})$.

C. Applying to real-world datasets

When the classification performance of real-world datasets is concerned, our theory may be limited by: i) the fact that it is an asymptotic result and allows for an estimation error of order $O(n^{-\frac{1}{2}})$ between theory and practice and ii) the strong Gaussian assumption for the input data.

However, when applied to real-world datasets, here to the popular MNIST [28] and Fashion-MNIST [29] datasets, our asymptotic results, which are theoretically only applicable for Gaussian data, show an unexpectedly similar behavior. Here we consider a two-class classification problem with a training set of $n = 256$ vectorized images of size $p = 784$ randomly selected from the MNIST and Fashion-MNIST datasets (numbers 8 and 9 in both cases as an example). Then a test set of $n_{\text{test}} = 256$ is used to evaluate the classification performance. Means and covariances are empirically obtained from the full set of 11 800 MNIST images (5 851 images of number 8 and 5 949 of number 9) and of 11 800 Fashion-MNIST images (5 851 images of number 8 and 5 949 of number 9), respectively. Despite the obvious non-Gaussianity as well as the clearly different nature of the input data (from the two datasets), the distribution of $g(\mathbf{x})$ is still surprisingly close to its Gaussian approximation computed from Theorem 2, as shown in Figure 9 and 10 for MNIST and Fashion-MNIST, respectively. In both cases we plot the results from (a) raw images as well as (b) when Gaussian white noise is artificially added to the image vectors.

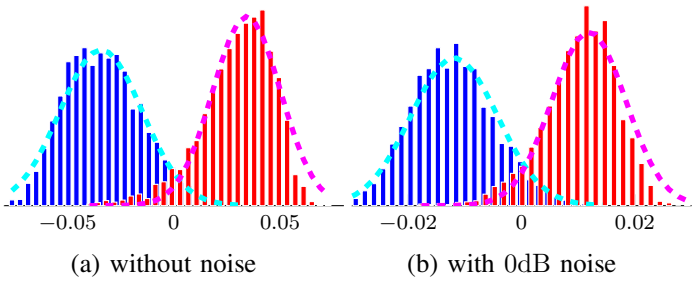


Fig. 9. Gaussian approximation of $g(\mathbf{x})$, $n = 256$, $p = 784$, $c_1 = c_2 = \frac{1}{2}, \gamma = 1$, Gaussian kernel with $\sigma = 1$, MNIST data (numbers 8 and 9) without and with 0dB noise.

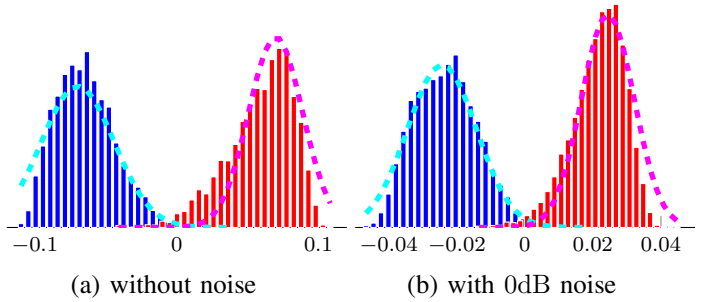


Fig. 10. Gaussian approximation of $g(\mathbf{x})$, $n = 256$, $p = 784$, $c_1 = c_2 = \frac{1}{2}, \gamma = 1$, Gaussian kernel with $\sigma = 1$, Fashion-MNIST data (numbers 8 and 9) without and with 0dB noise.

In Figure 11 we plot the misclassification rate as a function of the decision threshold ξ for MNIST and Fashion-MNIST data (number 8 and 9). We observe that although derived from a Gaussian mixture model, the conclusion from Remark 2, Lemma 1 and Corollary 1 that the decision threshold should approximately be $c_2 - c_1$ rather than 0 approximately holds true in both cases.

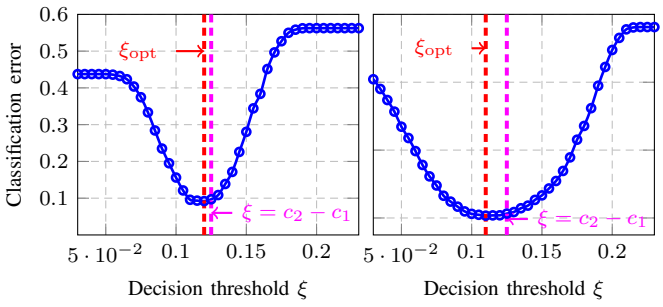


Fig. 11. $n = 512$, $p = 784$, $c_2 - c_1 = 0.125, \gamma = 1$, Gaussian kernel with $\sigma = 1$ for MNIST (left) and Fashion-MNIST data (right). With optimal decision threshold $\xi_{\text{opt}} = 0.12$ (left) and 0.11 (right) in red.

In Figure 12 and 13 we evaluated the performance of LS-SVM on the MNIST and Fashion-MNIST datasets (with and without noise) as a function of the kernel parameter σ of Gaussian kernel $f(x) = \exp(-x/2\sigma^2)$. Surprisingly, compared to Figure 5, we face the situation where there is little difference in the performance of LS-SVM as soon as σ^2 is away from 0, 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 influence of covariances as mentioned in the first paragraph of Section IV-B. This argument is numerically sustained by Table I. The gap between theory and practice observed as $\sigma^2 \to 0$ is likely a result of the finite n, p (as in Figure 5) rather than of the Gaussian assumption of the input data, since we observe a similar behavior even when Gaussian white noise is added.

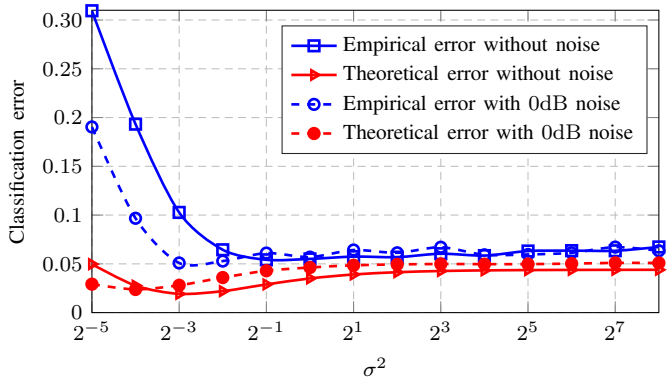


Fig. 12. $n = 256$, $p = 784$, $c_1 = c_2 = 1/2, \gamma = 1$, Gaussian kernel, MNIST data (numbers 8 and 9) with and without noise.

V. Concluding Remarks

In this work, through a performance analysis of LS-SVM for large dimensional data, we reveal the significance of balanced

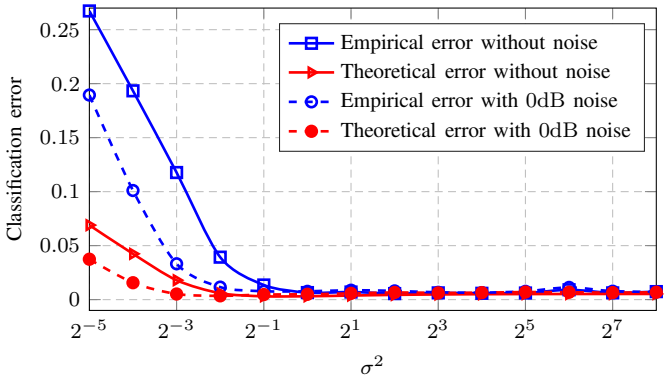


Fig. 13. $n = 256$, $p = 784$, $c_1 = c_2 = 1/2, \gamma = 1$, Gaussian kernel, Fashion-MNIST data (numbers 8 and 9) with and without noise.

TABLE I
Empirical estimation of differences in means and covariances of MNIST/Fashion-MNIST data (numbers 8 and 9)

	MNIST/Fashion-MNIST without noise	MNIST/Fashion-MNIST with 0dB noise
$\|\boldsymbol{\mu}_2 - \boldsymbol{\mu}_2\|^2$	251/483	96/197
$\frac{1}{p}\left(\text{tr}\left(\mathbf{C}_2 - \mathbf{C}_1\right)\right)^2$	19/89	3/13
$\frac{1}{p}\text{tr}\left((\mathbf{C}_2 - \mathbf{C}_1)^2\right)$	30/86	5/13

dataset with $c_1 = c_2$, as well as the interplay between the pivotal kernel function f and the statistical structure of the data. The normalized labels $y_i^* \in \{-1/c_1, 1/c_2\}$ are proposed to mitigate the damage of $c_2 - c_1$ in the decision function. We prove the irrelevance of γ when it is considered to remain constant in the large n, p regime; however, this argument is not guaranteed to hold true when γ scales with n, p. Our theoretical results, even though built upon the assumption of Gaussian data, provide similar results when tested on real-world large dimensional datasets, which offers a possible application despite the strong Gaussian assumption in the general context of large scale supervised learning.

The major difference of the present work compared to other theoretical analyses (for example [13]) is that, by studying the rather simple problem of a two-class Gaussian mixture separation with comparably large instance number and data dimension, together with sufficiently smooth kernel function f and regularization parameter γ of order $O(1)$, we deduce *explicit* results for the output of LS-SVM which surprisingly coincide with observations on some large dimensional real-world datasets (including MNIST and beyond) and therefore allowing for novel insights into the behavior of LS-SVM for large dimensional datasets. Of interest to future work is the remark that, unlike in the work of [13] where, in the large n alone asymptotics, γ is best scaled large with n, in the present large p, large n setting, where we demonstrate rate-optimality of LS-SVM for $\gamma = O(1)$. This apparent paradox could be deciphered through the analysis of more advanced (normalized inner product) kernels of the type $f(\mathbf{x}_i^\mathsf{T}\mathbf{x}_j/\sqrt{p})$, studied notably in [37], for which we believe that other scalings for γ would be optimal; it is also importantly believed that such kernels could lead to improved performances (not in rate, as those are already optimal in the present setting,

but possibly in absolute performance). These technically more involved considerations are left for future investigations.

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 $\mathbf{K}$. Additional technical tools are thus required to cope with this dependence structure.

The link between LS-SVM and extreme learning machine (ELM) was brought to light in [17] and the performance analysis of ELM in large dimension has been investigated in the recent article [38]. Together with these works, we have the possibility to identify the tight but subtle relation between the kernel function and the activation function in the context of some simple structured neural networks. This is notably of interest when the datasets are so large that computing $\mathbf{K}$ and the decision function $g(\mathbf{x})$ becomes prohibitive, a problem largely alleviated by neural networks with controllable number of neurons. This link also generally opens up a possible direction of research into the complex neural networks realm.

REFERENCES

[1] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

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

[3] Y.-J. Lee and O. L. Mangasarian, "Rsvm: Reduced support vector machines," in *Proceedings of the 2001 SIAM International Conference on Data Mining*. SIAM, 2001, pp. 1–17.

[4] G. M. Fung and O. L. Mangasarian, "Multicategory proximal support vector machine classifiers," *Machine learning*, vol. 59, no. 1-2, pp. 77–97, 2005.

[5] E. Osuna, R. Freund, and F. Girosit, "Training support vector machines: an application to face detection," in *Computer vision and pattern recognition, 1997. Proceedings., 1997 IEEE computer society conference on*. IEEE, 1997, pp. 130–136.

[6] C. Papageorgiou and T. Poggio, "A trainable system for object detection," *International Journal of Computer Vision*, vol. 38, no. 1, pp. 15–33, 2000.

[7] Y. LeCun, L. Jackel, L. Bottou, C. Cortes, J. S. Denker, H. Drucker, I. Guyon, U. Muller, E. Sackinger, P. Simard *et al.*, "Learning algorithms for classification: A comparison on handwritten digit recognition," *Neural networks: the statistical mechanics perspective*, vol. 261, p. 276, 1995.

[8] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features," in *European conference on machine learning*. Springer, 1998, pp. 137–142.

[9] D. Sculley and G. M. Wachman, "Relaxed online svms for spam filtering," in *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2007, pp. 415–422.

[10] V. Vapnik, *The nature of statistical learning theory*. Springer science & business media, 2013.

[11] B. Schölkopf and A. J. Smola, *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002.

[12] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*. MIT Press, 2012.

[13] A. Caponnetto and E. De Vito, "Optimal rates for the regularized least-squares algorithm," *Foundations of Computational Mathematics*, vol. 7, no. 3, pp. 331–368, 2007.

[14] I. Steinwart, D. R. Hush, C. Scovel *et al.*, "Optimal rates for regularized least squares regression." in *COLT*, 2009.

[15] T. V. Gestel, J. A. Suykens, G. Lanckriet, A. Lambrechts, B. D. Moor, and J. Vandewalle, "Bayesian framework for least-squares support vector machine classifiers, gaussian processes, and kernel fisher discriminant analysis," *Neural computation*, vol. 14, no. 5, pp. 1115–1147, 2002.

[16] J. Ye and T. Xiong, "Svm versus least squares svm," in *Artificial Intelligence and Statistics*, 2007, pp. 644–651.

[17] G.-B. Huang, H. Zhou, X. Ding, and R. Zhang, "Extreme learning machine for regression and multiclass classification," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 42, no. 2, pp. 513–529, 2012.

[18] T. Evgeniou, M. Pontil, and T. Poggio, "Regularization networks and support vector machines," *Advances in computational mathematics*, vol. 13, no. 1, p. 1, 2000.

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

[20] R. Couillet, F. Benaych-Georges *et al.*, "Kernel spectral clustering of large dimensional data," *Electronic Journal of Statistics*, vol. 10, no. 1, pp. 1393–1454, 2016.

[21] X. Mai and R. Couillet, "A random matrix analysis and improvement of semi-supervised learning for large dimensional data," *arXiv preprint arXiv:1711.03404*, 2017.

[22] R. Couillet and A. Kammoun, "Random matrix improved subspace clustering," in *Signals, Systems and Computers, 2016 50th Asilomar Conference on*. IEEE, 2016, pp. 90–94.

[23] V. Cherkassky and Y. Ma, "Practical selection of svm parameters and noise estimation for svm regression," *Neural networks*, vol. 17, no. 1, pp. 113–126, 2004.

[24] O. Chapelle, V. Vapnik, O. Bousquet, and S. Mukherjee, "Choosing multiple parameters for support vector machines," *Machine learning*, vol. 46, no. 1-3, pp. 131–159, 2002.

[25] N.-E. Ayat, M. Cheriet, and C. Y. Suen, "Automatic model selection for the optimization of svm kernels," *Pattern Recognition*, vol. 38, no. 10, pp. 1733–1745, 2005.

[26] J. Weston, S. Mukherjee, O. Chapelle, M. Pontil, T. Poggio, and V. Vapnik, "Feature selection for svms," in *Advances in neural information processing systems*, 2001, pp. 668–674.

[27] C.-L. Huang and C.-J. Wang, "A ga-based feature selection and parameters optimizationfor support vector machines," *Expert Systems with applications*, vol. 31, no. 2, pp. 231–240, 2006.

[28] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[29] H. Xiao, K. Rasul, and R. Vollgraf, "Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms," *arXiv preprint arXiv:1708.07747*, 2017.

[30] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statistics and computing*, vol. 14, no. 3, pp. 199–222, 2004.

[31] H. T. Ali, A. Kammoun, and R. Couillet, "Random matrix asymptotics of inner product kernel spectral clustering," in *2018 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'18), Calgary (AB)*, 2018.

[32] R. Couillet, Z. Liao, and X. Mai, "Classification Asymptotics in the Random Matrix Regime," in *26th European Signal Processing Conference (EUSIPCO'2018)*. IEEE, 2018.

[33] P. Billingsley, *Probability and measure*. John Wiley & Sons, 2008.

[34] T. Evgeniou, M. Pontil, C. Papageorgiou, and T. Poggio, "Image representations for object detection using kernel classifiers," in *Asian Conference on Computer Vision*. Citeseer, 2000, pp. 687–692.

[35] G. Baudat and F. Anouar, "Generalized discriminant analysis using a kernel approach," *Neural computation*, vol. 12, no. 10, pp. 2385–2404, 2000.

[36] S. Mika, G. Ratsch, J. Weston, B. Scholkopf, and K.-R. Mullers, "Fisher discriminant analysis with kernels," in *Neural networks for signal processing IX, 1999. Proceedings of the 1999 IEEE signal processing society workshop.* Ieee, 1999, pp. 41–48.

[37] X. Cheng and A. Singer, "The spectrum of random inner-product kernel matrices," *Random Matrices: Theory and Applications*, vol. 2, no. 04, p. 1350010, 2013.

[38] C. Louart, Z. Liao, R. Couillet *et al.*, "A random matrix approach to neural networks," *The Annals of Applied Probability*, vol. 28, no. 2, pp. 1190–1248, 2018.

[39] Z. Bai and J. W. Silverstein, "Spectral analysis of large dimensional random matrices," 2010.

Supplementary Material

A Large Dimensional Analysis of Least Squares Support Vector Machines

APPENDIX A
PROOF OF THEOREM 1

Our key interest here is on the decision function of LS-SVM: $g(\mathbf{x}) = \boldsymbol{\alpha}^\mathsf{T}\mathbf{k}(\mathbf{x}) + b$ with $(\boldsymbol{\alpha}, b)$ given by

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

and $\mathbf{S}^{-1} = \left(\mathbf{K} + \frac{n}{\gamma}\mathbf{I}_n\right)^{-1}$.

Before going into the detailed proof, as we will frequently deal with random variables evolving as n, p grow large, we shall use the extension of the $O(\cdot)$ notation introduced in [20]: for a random variable $x \equiv x_n$ and $u_n \geq 0$, we write $x = O(u_n)$ if for any $\eta > 0$ and $D > 0$, we have $n^D\mathrm{P}(x \geq n^\eta u_n) \to 0$. Note that under Assumption 1 it is equivalent to use either $O(u_n)$ or $O(u_p)$ since n, p scales linearly. In the following we shall use constantly $O(u_n)$ for simplicity.

When multidimensional objects are concerned, $\mathbf{v} = O(u_n)$ means the maximum entry of a vector (or a diagonal matrix) $\mathbf{v}$ in absolute value is of order $O(u_n)$ and $\mathbf{M} = O(u_n)$ means that the operator norm of $\mathbf{M}$ is of order $O(u_n)$. We refer the reader to [20] for more discussions on these practical definitions.

Under the growth rate settings of Assumption 1, from [20], the approximation of the kernel matrix $\mathbf{K}$ is given by

$$\mathbf{K} = -2f'(\tau)\left(\mathbf{P}\boldsymbol{\Omega}^\mathsf{T}\boldsymbol{\Omega}\mathbf{P} + \mathbf{A}\right) + \beta\mathbf{I}_n + O(n^{-\frac{1}{2}}) \tag{12}$$

with $\beta = f(0) - f(\tau) + \tau f'(\tau)$ and $\mathbf{A} = \mathbf{A}_n + \mathbf{A}_{\sqrt{n}} + \mathbf{A}_1$, $\mathbf{A}_n = -\frac{f(\tau)}{2f'(\tau)}\mathbf{1}_n\mathbf{1}_n^\mathsf{T}$ and $\mathbf{A}_{\sqrt{n}}$, $\mathbf{A}_1$ given by (18) and (19) at the top of next page, where we denote

$$t_a \triangleq \frac{\mathrm{tr}(\mathbf{C}_a - \mathbf{C}^\circ)}{\sqrt{p}} = O(1)$$
$$(\boldsymbol{\psi})^2 \triangleq [(\boldsymbol{\psi}_1)^2, \ldots, (\boldsymbol{\psi}_n)^2]^\mathsf{T}.$$

We start with the term $\mathbf{S}^{-1}$. The terms of leading order in $\mathbf{K}$, i.e.,$-2f'(\tau)\mathbf{A}_n$ and $\frac{n}{\gamma}\mathbf{I}_n$ are both of operator norm $O(n)$. Therefore a Taylor expansion can be performed as

$$\begin{aligned}\mathbf{S}^{-1} &= \left(\mathbf{K} + \frac{n}{\gamma}\mathbf{I}_n\right)^{-1} = \frac{1}{n}\Bigg[\mathbf{L}^{-1} - \frac{2f'(\tau)}{n} \\ &\left(\mathbf{A}_{\sqrt{n}} + \mathbf{A}_1 + \mathbf{P}\boldsymbol{\Omega}^\mathsf{T}\boldsymbol{\Omega}\mathbf{P}\right) + \frac{\beta\mathbf{I}_n}{n} + O(n^{-\frac{3}{2}})\Bigg]^{-1} \\ &= \frac{\mathbf{L}}{n} + \frac{2f'(\tau)}{n^2}\mathbf{L}\mathbf{A}_{\sqrt{n}}\mathbf{L} + \mathbf{L}\left(\mathbf{Q} - \frac{\beta}{n^2}\mathbf{I}_n\right)\mathbf{L} + O(n^{-\frac{5}{2}})\end{aligned}$$

with $\mathbf{L} = \left(f(\tau)\frac{\mathbf{1}_n\mathbf{1}_n^\mathsf{T}}{n} + \frac{\mathbf{I}_n}{\gamma}\right)^{-1}$ of order $O(1)$ and $\mathbf{Q} = \frac{2f'(\tau)}{n^2}\left(\mathbf{A}_1 + \mathbf{P}\boldsymbol{\Omega}^\mathsf{T}\boldsymbol{\Omega}\mathbf{P} + \frac{2f'(\tau)}{n}\mathbf{A}_{\sqrt{n}}\mathbf{L}\mathbf{A}_{\sqrt{n}}\right)$.

With the Sherman-Morrison formula we are able to compute explicitly $\mathbf{L}$ as

$$\begin{aligned}\mathbf{L} &= \left(f(\tau)\frac{\mathbf{1}_n\mathbf{1}_n^\mathsf{T}}{n} + \frac{\mathbf{I}_n}{\gamma}\right)^{-1} = \gamma\left(\mathbf{I}_n - \frac{\gamma f(\tau)}{1+\gamma f(\tau)}\frac{\mathbf{1}_n\mathbf{1}_n^\mathsf{T}}{n}\right) \\ &= \frac{\gamma}{1+\gamma f(\tau)}\mathbf{I}_n + \frac{\gamma^2 f(\tau)}{1+\gamma f(\tau)}\mathbf{P} = O(1).\end{aligned} \tag{13}$$

Writing $\mathbf{L}$ as a linear combination of $\mathbf{I}_n$ and $\mathbf{P}$ is useful when computing $\mathbf{L}\mathbf{1}_n$ or $\mathbf{1}_n^\mathsf{T}\mathbf{L}$, because by the definition of $\mathbf{P} = \mathbf{I}_n - \frac{\mathbf{1}_n\mathbf{1}_n^\mathsf{T}}{n}$, we have $\mathbf{1}_n^\mathsf{T}\mathbf{P} = \mathbf{P}\mathbf{1}_n = \mathbf{0}$.

We shall start with the term $\mathbf{1}_n^\mathsf{T}\mathbf{S}^{-1}$, since it is the basis of several other terms appearing in $\boldsymbol{\alpha}$ and b,

$$\begin{aligned}\mathbf{1}_n^\mathsf{T}\mathbf{S}^{-1} = \frac{\gamma\mathbf{1}_n^\mathsf{T}}{1+\gamma f(\tau)}\left[\frac{\mathbf{I}_n}{n} + \frac{2f'(\tau)}{n^2}\mathbf{A}_{\sqrt{n}}\mathbf{L} + \left(\mathbf{Q} - \frac{\beta}{n^2}\mathbf{I}_n\right)\mathbf{L}\right] \\ + O(n^{-\frac{3}{2}})\end{aligned}$$

since $\mathbf{1}_n^\mathsf{T}\mathbf{L} = \frac{\gamma}{1+\gamma f(\tau)}\mathbf{1}_n^\mathsf{T}$.

With $\mathbf{1}_n^\mathsf{T}\mathbf{S}^{-1}$ at hand, we next obtain,

$$\begin{aligned}\mathbf{1}_n\mathbf{1}_n^\mathsf{T}\mathbf{S}^{-1} = \frac{\gamma}{1+\gamma f(\tau)}\Bigg[\underbrace{\frac{\mathbf{1}_n\mathbf{1}_n^\mathsf{T}}{n}}_{O(1)} + \underbrace{\frac{2f'(\tau)}{n^2}\mathbf{1}_n\mathbf{1}_n^\mathsf{T}\mathbf{A}_{\sqrt{n}}\mathbf{L}}_{O(n^{-1/2})} \\ + \underbrace{\mathbf{1}_n\mathbf{1}_n^\mathsf{T}\left(\mathbf{Q} - \frac{\beta}{n^2}\mathbf{I}_n\right)\mathbf{L}}_{O(n^{-1})}\Bigg] + O(n^{-\frac{3}{2}})\end{aligned} \tag{14}$$

$$\begin{aligned}\mathbf{1}_n^\mathsf{T}\mathbf{S}^{-1}\mathbf{y} = \frac{\gamma}{1+\gamma f(\tau)}\Bigg[\underbrace{c_2 - c_1}_{O(1)} + \underbrace{\frac{2f'(\tau)}{n^2}\mathbf{1}_n^\mathsf{T}\mathbf{A}_{\sqrt{n}}\mathbf{L}\mathbf{y}}_{O(n^{-1/2})} \\ + \underbrace{\mathbf{1}_n^\mathsf{T}\left(\mathbf{Q} - \frac{\beta}{n^2}\mathbf{I}_n\right)\mathbf{L}\mathbf{y}}_{O(n^{-1})}\Bigg] + O(n^{-\frac{3}{2}})\end{aligned}$$

$$\begin{aligned}\mathbf{1}_n^\mathsf{T}\mathbf{S}^{-1}\mathbf{1}_n = \frac{\gamma}{1+\gamma f(\tau)}\Bigg[\underbrace{\frac{1}{O(1)}}_{O(1)} + \underbrace{\frac{2f'(\tau)}{n^2}\frac{\gamma\mathbf{1}_n^\mathsf{T}\mathbf{A}_{\sqrt{n}}\mathbf{1}_n}{1+\gamma f(\tau)}}_{O(n^{-1/2})} \\ + \underbrace{\frac{\gamma}{1+\gamma f(\tau)}\mathbf{1}_n^\mathsf{T}\left(\mathbf{Q} - \frac{\beta}{n^2}\mathbf{I}_n\right)\mathbf{1}_n}_{O(n^{-1})}\Bigg] + O(n^{-\frac{3}{2}}).\end{aligned}$$

The inverse of $\mathbf{1}_n^\mathsf{T}\mathbf{S}^{-1}\mathbf{1}_n$ can consequently be computed using a Taylor expansion around its leading order, allowing an error term of $O(n^{-\frac{3}{2}})$ as

$$\begin{aligned}\frac{1}{\mathbf{1}_n^\mathsf{T}\mathbf{S}^{-1}\mathbf{1}_n} = \frac{1+\gamma f(\tau)}{\gamma}\Bigg[\underbrace{\frac{1}{O(1)}}_{O(1)} - \underbrace{\frac{2f'(\tau)}{n^2}\frac{\gamma\mathbf{1}_n^\mathsf{T}\mathbf{A}_{\sqrt{n}}\mathbf{1}_n}{1+\gamma f(\tau)}}_{O(n^{-1/2})} \\ - \underbrace{\frac{\gamma}{1+\gamma f(\tau)}\mathbf{1}_n^\mathsf{T}\left(\mathbf{Q} - \frac{\beta}{n^2}\mathbf{I}_n\right)\mathbf{1}_n}_{O(n^{-1})}\Bigg] + O(n^{-\frac{3}{2}}).\end{aligned} \tag{15}$$

$$\mathbf{A}_{\sqrt{n}} = -\frac{1}{2}\left[\boldsymbol{\psi}\mathbf{1}_n^\mathsf{T} + \mathbf{1}_n\boldsymbol{\psi}^\mathsf{T} + \left\{t_a\frac{\mathbf{1}_{n_a}}{\sqrt{p}}\right\}_{a=1}^2\mathbf{1}_n^\mathsf{T} + \mathbf{1}_n\left\{t_b\frac{\mathbf{1}_{n_b}^\mathsf{T}}{\sqrt{p}}\right\}_{b=1}^2\right] \tag{18}$$

$$\begin{aligned}\mathbf{A}_1 = &-\frac{1}{2}\left[\left\{\|\boldsymbol{\mu}_a-\boldsymbol{\mu}_b\|^2\frac{\mathbf{1}_{n_a}\mathbf{1}_{n_b}^\mathsf{T}}{p}\right\}_{a,b=1}^2 + 2\left\{\frac{(\boldsymbol{\Omega}\mathbf{P})_a^\mathsf{T}(\boldsymbol{\mu}_b-\boldsymbol{\mu}_a)\mathbf{1}_{n_b}^\mathsf{T}}{\sqrt{p}}\right\}_{a,b=1}^2 - 2\left\{\frac{\mathbf{1}_{n_a}(\boldsymbol{\mu}_b-\boldsymbol{\mu}_a)^\mathsf{T}(\boldsymbol{\Omega}\mathbf{P})_b}{\sqrt{p}}\right\}_{a,b=1}^2\right]\\
&-\frac{f''(\tau)}{4f'(\tau)}\left[(\boldsymbol{\psi})^2\mathbf{1}_n^\mathsf{T} + \mathbf{1}_n[(\boldsymbol{\psi})^2]^\mathsf{T} + \left\{t_a^2\frac{\mathbf{1}_{n_a}}{p}\right\}_{a=1}^2\mathbf{1}_n^\mathsf{T} + \mathbf{1}_n\left\{t_b^2\frac{\mathbf{1}_{n_b}^\mathsf{T}}{p}\right\}_{b=1}^2 + 2\left\{t_at_b\frac{\mathbf{1}_{n_a}\mathbf{1}_{n_b}^\mathsf{T}}{p}\right\}_{a,b=1}^2 + 2\mathcal{D}\{t_a\mathbf{I}_{n_a}\}_{a=1}^2\boldsymbol{\psi}\frac{\mathbf{1}_n^\mathsf{T}}{\sqrt{p}}\right.\\
&\left.+\,2\boldsymbol{\psi}\left\{t_b\frac{\mathbf{1}_{n_b}^\mathsf{T}}{\sqrt{p}}\right\}_{b=1}^2 + 2\frac{\mathbf{1}_n}{\sqrt{p}}(\boldsymbol{\psi})^\mathsf{T}\mathcal{D}\{t_a\mathbf{1}_{n_a}\}_{a=1}^2 + 2\left\{t_a\frac{\mathbf{1}_{n_a}}{\sqrt{p}}\right\}_{a=1}^2(\boldsymbol{\psi})^\mathsf{T} + 4\left\{\mathrm{tr}(\mathbf{C}_a\mathbf{C}_b)\frac{\mathbf{1}_{n_a}\mathbf{1}_{n_b}^\mathsf{T}}{p^2}\right\}_{a,b=1}^2 + 2\boldsymbol{\psi}(\boldsymbol{\psi})^\mathsf{T}\right]\end{aligned} \tag{19}$$

$$\begin{aligned}\tilde{\mathbf{k}}(\mathbf{x}) = &f'(\tau)\left[\left\{\frac{\|\boldsymbol{\mu}_b-\boldsymbol{\mu}_a\|^2}{p}\mathbf{1}_{n_b}\right\}_{b=1}^2 - \frac{2}{\sqrt{p}}\left\{\mathbf{1}_{n_b}(\boldsymbol{\mu}_b-\boldsymbol{\mu}_a)^\mathsf{T}\right\}_{b=1}^2\boldsymbol{\omega}_\mathbf{x} + \frac{2}{\sqrt{p}}\mathcal{D}\left(\left\{\mathbf{1}_{n_b}(\boldsymbol{\mu}_b-\boldsymbol{\mu}_a)^\mathsf{T}\right\}_{b=1}^2\boldsymbol{\Omega}\right)\right]\\
&+\frac{f''(\tau)}{2}\left[\left\{\frac{(t_a+t_b)^2}{p}\mathbf{1}_{n_b}\right\}_{b=1}^2 + 2\mathcal{D}\left(\left\{\frac{t_a+t_b}{\sqrt{p}}\mathbf{1}_{n_b}\right\}_{b=1}^2\right)\boldsymbol{\psi} + 2\left\{\frac{t_a+t_b}{\sqrt{p}}\mathbf{1}_{n_b}\right\}_{b=1}^2\psi_\mathbf{x} + (\boldsymbol{\psi})^2 + 2\psi_\mathbf{x}\boldsymbol{\psi} + \psi_\mathbf{x}^2\mathbf{1}_n\right.\\
&\left.+\left\{\frac{4}{p^2}\,\mathrm{tr}(\mathbf{C}_a\mathbf{C}_b)\mathbf{1}_{n_b}\right\}_{b=1}^2\right]\end{aligned} \tag{20}$$

Combing (14) with (15) we deduce

$$\begin{aligned}\frac{\mathbf{1}_n\mathbf{1}_n^\mathsf{T}\mathbf{S}^{-1}}{\mathbf{1}_n^\mathsf{T}\mathbf{S}^{-1}\mathbf{1}_n} = &\underbrace{\frac{\mathbf{1}_n\mathbf{1}_n^\mathsf{T}}{n}}_{O(1)} + \underbrace{\frac{2f'(\tau)}{n^2}\mathbf{1}_n\mathbf{1}_n^\mathsf{T}\mathbf{A}_{\sqrt{n}}\left[\mathbf{L}-\frac{\gamma\frac{\mathbf{1}_n\mathbf{1}_n^\mathsf{T}}{n}}{1+\gamma f(\tau)}\right]}_{O(n^{-1/2})}\\
&+\underbrace{\mathbf{1}_n\mathbf{1}_n^\mathsf{T}\left(\mathbf{Q}-\frac{\beta}{n^2}\mathbf{I}_n\right)\left[\mathbf{L}-\frac{\gamma\frac{\mathbf{1}_n\mathbf{1}_n^\mathsf{T}}{n}}{1+\gamma f(\tau)}\right]}_{O(n^{-1})} + O(n^{-\frac{3}{2}})\end{aligned} \tag{16}$$

and similarly the following approximation of b as

$$\begin{aligned}b = &\underbrace{c_2-c_1}_{O(1)} - \underbrace{\frac{2\gamma}{\sqrt{p}}c_1c_2f'(\tau)(t_2-t_1)}_{O(n^{-1/2})} - \underbrace{\frac{\gamma f'(\tau)}{n}\mathbf{y}^\mathsf{T}\mathbf{P}\boldsymbol{\psi}}_{O(n^{-1})}\\
&\underbrace{-\frac{\gamma f''(\tau)}{2n}\mathbf{y}^\mathsf{T}\mathbf{P}(\boldsymbol{\psi})^2 + \frac{4\gamma c_1c_2}{p}[c_1T_1+(c_2-c_1)D-c_2T_2]}_{O(n^{-1})}\\
&+O(n^{-\frac{3}{2}})\end{aligned} \tag{17}$$

where

$$D = \frac{f'(\tau)}{2}\|\boldsymbol{\mu}_2-\boldsymbol{\mu}_1\|^2 + \frac{f''(\tau)}{4}(t_1+t_2)^2 + f''(\tau)\frac{\mathrm{tr}\,\mathbf{C}_1\mathbf{C}_2}{p}$$

$$T_a = f''(\tau)t_a^2 + f''(\tau)\frac{\mathrm{tr}\,\mathbf{C}_1\mathbf{C}_2}{p}$$

which gives the asymptotic approximation of b.

Moving to $\boldsymbol{\alpha}$, note from (13) that $\mathbf{L}-\frac{\gamma}{1+\gamma f(\tau)}\frac{\mathbf{1}_n\mathbf{1}_n^\mathsf{T}}{n} = \gamma\mathbf{P}$, and we can thus rewrite:

$$\begin{aligned}\frac{\mathbf{1}_n\mathbf{1}_n^\mathsf{T}\mathbf{S}^{-1}}{\mathbf{1}_n^\mathsf{T}\mathbf{S}^{-1}\mathbf{1}_n} = &\frac{\mathbf{1}_n\mathbf{1}_n^\mathsf{T}}{n} + \frac{2\gamma f'(\tau)}{n^2}\mathbf{1}_n\mathbf{1}_n^\mathsf{T}\mathbf{A}_{\sqrt{n}}\mathbf{P}\\
&+\gamma\mathbf{1}_n\mathbf{1}_n^\mathsf{T}\left(\mathbf{Q}-\frac{\beta}{n^2}\mathbf{I}_n\right)\mathbf{P} + O(n^{-\frac{3}{2}}).\end{aligned}$$

At this point, for $\boldsymbol{\alpha} = \mathbf{S}^{-1}\left(\mathbf{I}_n - \frac{\mathbf{1}_n\mathbf{1}_n^\mathsf{T}\mathbf{S}^{-1}}{\mathbf{1}_n^\mathsf{T}\mathbf{S}^{-1}\mathbf{1}_n}\right)\mathbf{y}$, we have

$$\begin{aligned}\boldsymbol{\alpha} = \mathbf{S}^{-1}\Big[&\mathbf{I}_n - \frac{2\gamma f'(\tau)}{n^2}\mathbf{1}_n\mathbf{1}_n^\mathsf{T}\mathbf{A}_{\sqrt{n}}\\
&-\gamma\mathbf{1}_n\mathbf{1}_n^\mathsf{T}\left(\mathbf{Q}-\frac{\beta}{n^2}\mathbf{I}_n\right)\Big]\mathbf{P}\mathbf{y} + O(n^{-\frac{5}{2}}).\end{aligned}$$

Here again, we use $\mathbf{1}_n^\mathsf{T}\mathbf{L} = \frac{\gamma}{1+\gamma f(\tau)}\mathbf{1}_n^\mathsf{T}$ and $\mathbf{L} - \frac{\gamma}{1+\gamma f(\tau)}\frac{\mathbf{1}_n\mathbf{1}_n^\mathsf{T}}{n} = \gamma\mathbf{P}$, to eventually get

$$\begin{aligned}\boldsymbol{\alpha} = &\underbrace{\frac{\gamma}{n}\mathbf{P}\mathbf{y}}_{O(n^{-1})} + \underbrace{\gamma^2\mathbf{P}\left(\mathbf{Q}-\frac{\beta}{n^2}\mathbf{I}_n\right)\mathbf{P}\mathbf{y}}_{O(n^{-2})}\\
&\underbrace{-\frac{\gamma^2}{1+\gamma f(\tau)}\left(\frac{2f'(\tau)}{n^2}\right)^2\mathbf{L}\mathbf{A}_{\sqrt{n}}\mathbf{1}_n\mathbf{1}_n^\mathsf{T}\mathbf{A}_{\sqrt{n}}\mathbf{P}\mathbf{y}}_{O(n^{-2})} + O(n^{-\frac{5}{2}}).\end{aligned} \tag{21}$$

Note here the absence of a term of order $O(n^{-3/2})$ in the expression of $\boldsymbol{\alpha}$ since $\mathbf{P}\mathbf{A}_{\sqrt{n}}\mathbf{P} = 0$ from (18).

We shall now work on the vector $\mathbf{k}(\mathbf{x})$ for a new datum $\mathbf{x}$, following the same analysis as in [20] for the kernel matrix $\mathbf{K}$, assuming that $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}_a, \mathbf{C}_a)$ and recalling the random variables definitions,

$$\begin{aligned}\boldsymbol{\omega}_\mathbf{x} &\triangleq (\mathbf{x}-\boldsymbol{\mu}_a)/\sqrt{p}\\ \psi_\mathbf{x} &\triangleq \|\boldsymbol{\omega}_\mathbf{x}\|^2 - \mathbb{E}\|\boldsymbol{\omega}_\mathbf{x}\|^2\end{aligned}$$

we show that the j-th entry of $\mathbf{k}(\mathbf{x})$ can be written as

$$[\mathbf{k}(\mathbf{x})]_j = \underbrace{f(\tau)}_{O(1)} + f'(\tau)\Bigg[\underbrace{\frac{t_a+t_b}{\sqrt{p}} + \psi_x + \psi_j - 2(\boldsymbol{\omega}_{\mathbf{x}})^{\mathsf{T}}\boldsymbol{\omega}_j}_{O(n^{-1/2})} + \underbrace{\frac{\|\boldsymbol{\mu}_b - \boldsymbol{\mu}_a\|^2}{p} + \frac{2}{\sqrt{p}}(\boldsymbol{\mu}_b - \boldsymbol{\mu}_a)^{\mathsf{T}}(\boldsymbol{\omega}_j - \boldsymbol{\omega}_{\mathbf{x}})}_{O(n^{-1})}\Bigg] + \frac{f''(\tau)}{2}$$
$$\underbrace{\left[\left(\frac{t_a+t_b}{\sqrt{p}} + \psi_j + \psi_{\mathbf{x}}\right)^2 + \frac{4}{p^2}\operatorname{tr}\mathbf{C}_a\mathbf{C}_b\right]}_{O(n^{-1})} + O(n^{-\frac{3}{2}}). \quad (22)$$

Combining (21) and (22), we deduce

$$\boldsymbol{\alpha}^{\mathsf{T}}\mathbf{k}(\mathbf{x}) = \underbrace{\frac{2\gamma}{\sqrt{p}}c_1c_2f'(\tau)(t_2-t_1)}_{O(n^{-1/2})} + \underbrace{\frac{\gamma}{n}\mathbf{y}^{\mathsf{T}}\mathbf{P}\tilde{\mathbf{k}}(\mathbf{x})}_{O(n^{-1})} + \underbrace{\frac{\gamma f'(\tau)}{n}\mathbf{y}^{\mathsf{T}}\mathbf{P}(\boldsymbol{\psi} - 2\mathbf{P}\boldsymbol{\Omega}^{\mathsf{T}}\boldsymbol{\omega}_{\mathbf{x}})}_{O(n^{-1})} + O(n^{-\frac{3}{2}}) \quad (23)$$

with $\tilde{\mathbf{k}}(\mathbf{x})$ given in (20).

At this point, note that the term of order $O(n^{-\frac{1}{2}})$ in the final object $g(\mathbf{x}) = \boldsymbol{\alpha}^{\mathsf{T}}\mathbf{k}(\mathbf{x}) + b$ disappears because in both (17) and (23) the term of order $O(n^{-1/2})$ is $\frac{2\gamma}{\sqrt{p}}c_1c_2f'(\tau)(t_2-t_1)$ but of opposite signs. Also, we see that the leading term $c_2 - c_1$ in b will remain in $g(\mathbf{x})$ as stated in Remark 2.

The development of $\mathbf{y}^{\mathsf{T}}\mathbf{P}\tilde{\mathbf{k}}(\mathbf{x})$ induces many simplifications, since i) $\mathbf{P}\mathbf{1}_n = \mathbf{0}$ and ii) random variables as $\boldsymbol{\omega}_{\mathbf{x}}$ and $\boldsymbol{\psi}$ in $\tilde{\mathbf{k}}(\mathbf{x})$, once multiplied by $\mathbf{y}^{\mathsf{T}}\mathbf{P}$, thanks to probabilistic averaging of independent zero-mean terms, are of smaller order and thus become negligible. We thus get

$$\frac{\gamma}{n}\mathbf{y}^{\mathsf{T}}\mathbf{P}\tilde{\mathbf{k}}(\mathbf{x}) = 2\gamma c_1c_2f'(\tau)\Bigg[\frac{\|\boldsymbol{\mu}_2 - \boldsymbol{\mu}_a\|^2 - \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_a\|^2}{p} - 2(\boldsymbol{\omega}_{\mathbf{x}})^{\mathsf{T}}\frac{\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1}{\sqrt{p}}\Bigg] + \frac{\gamma f''(\tau)}{2n}\mathbf{y}^{\mathsf{T}}\mathbf{P}(\boldsymbol{\psi})^2 + \gamma c_1c_2f''(\tau)\Bigg[2\left(\frac{t_a}{\sqrt{p}} + \psi_{\mathbf{x}}\right)\frac{t_2-t_1}{\sqrt{p}} + \frac{t_2^2-t_1^2}{p} + \frac{4}{p^2}\operatorname{tr}(\mathbf{C}_a\mathbf{C}_2 - \mathbf{C}_a\mathbf{C}_1)\Bigg] + O(n^{-\frac{3}{2}}). \quad (24)$$

This result, together with (23), completes the analysis of the term $\boldsymbol{\alpha}^{\mathsf{T}}\mathbf{k}(\mathbf{x})$. Combining (23)-(24) with (17) we conclude the proof of Theorem 1.

Appendix B
Proof of Theorem 2

This section is dedicated to the proof of the central limit theorem for

$$\hat{g}(\mathbf{x}) = c_2 - c_1 + \gamma\left(\mathfrak{P} + c_{\mathbf{x}}\mathfrak{D}\right)$$

with the shortcut $c_{\mathbf{x}} = -2c_1c_2^2$ for $\mathbf{x} \in \mathcal{C}_1$ and $c_{\mathbf{x}} = 2c_1^2c_2$ for $\mathbf{x} \in \mathcal{C}_2$, and $\mathfrak{P}, \mathfrak{D}$ as defined in (7) and (8).

Our objective is to show that for $a \in \{1,2\}$, $n(\hat{g}(\mathbf{x}) - G_a) \xrightarrow{d} 0$ with

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

where E_a and Var_a are given in Theorem 2. We recall that $\mathbf{x} = \boldsymbol{\mu}_a + \sqrt{p}\boldsymbol{\omega}_{\mathbf{x}}$ with $\boldsymbol{\omega}_{\mathbf{x}} \sim \mathcal{N}(0, \mathbf{C}_a/p)$.

Letting $\mathbf{z}_{\mathbf{x}}$ such that $\boldsymbol{\omega}_{\mathbf{x}} = \mathbf{C}_a^{1/2}\mathbf{z}_{\mathbf{x}}/\sqrt{p}$, we have $\mathbf{z}_{\mathbf{x}} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$ and we can rewrite $\hat{g}(\mathbf{x})$ in the following quadratic form (of $\mathbf{z}_{\mathbf{x}}$) as

$$\hat{g}(\mathbf{x}) = \mathbf{z}_{\mathbf{x}}^{\mathsf{T}}\mathbf{A}\mathbf{z}_{\mathbf{x}} + \mathbf{z}_{\mathbf{x}}^{\mathsf{T}}\mathbf{b} + c$$

with

$$\mathbf{A} = 2\gamma c_1c_2f''(\tau)\frac{\operatorname{tr}(\mathbf{C}_2 - \mathbf{C}_1)}{p}\frac{\mathbf{C}_a}{p}$$
$$\mathbf{b} = -\frac{2\gamma f'(\tau)}{n}\frac{(\mathbf{C}_a)^{\frac{1}{2}}}{\sqrt{p}}\boldsymbol{\Omega}\mathbf{P}\mathbf{y} - \frac{4c_1c_2\gamma f'(\tau)}{\sqrt{p}}\frac{(\mathbf{C}_a)^{\frac{1}{2}}}{\sqrt{p}}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)$$
$$c = c_2 - c_1 + \gamma c_{\mathbf{x}}\mathfrak{D} - 2\gamma c_1c_2f''(\tau)\frac{\operatorname{tr}(\mathbf{C}_2 - \mathbf{C}_1)}{p}\frac{\operatorname{tr}\mathbf{C}_a}{p}.$$

Since $\mathbf{z}_{\mathbf{x}}$ is (standard) Gaussian and has the same distribution as $\mathbf{U}\mathbf{z}_{\mathbf{x}}$ for any orthogonal matrix $\mathbf{U}$ (i.e., such that $\mathbf{U}^{\mathsf{T}}\mathbf{U} = \mathbf{U}\mathbf{U}^{\mathsf{T}} = \mathbf{I}_n$), we choose $\mathbf{U}$ that diagonalize $\mathbf{A}$ such that $\mathbf{A} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^{\mathsf{T}}$, with $\boldsymbol{\Lambda}$ diagonal so that $\hat{g}(\mathbf{x})$ and $\tilde{g}(\mathbf{x})$ have the same distribution where

$$\tilde{g}(\mathbf{x}) = \mathbf{z}_{\mathbf{x}}^{\mathsf{T}}\boldsymbol{\Lambda}\mathbf{z}_{\mathbf{x}} + \mathbf{z}_{\mathbf{x}}^{\mathsf{T}}\tilde{\mathbf{b}} + c = \sum_{i=1}^{n}\left(z_i^2\lambda_i + z_i\tilde{b}_i + \frac{c}{n}\right)$$

and $\tilde{\mathbf{b}} = \mathbf{U}^{\mathsf{T}}\mathbf{b}$, λ_i the diagonal elements of $\boldsymbol{\Lambda}$ and z_i the elements of $\mathbf{z}_{\mathbf{x}}$.

Conditioning on $\boldsymbol{\Omega}$, we thus result in the sum of independent but not identically distributed random variables $r_i = z_i^2\lambda_i + z_i\tilde{b}_i + \frac{c}{n}$. We then resort to the Lyapunov CLT [33, Theorem 27.3].

We begin by estimating the expectation and the variance

$$\mathbb{E}[r_i|\boldsymbol{\Omega}] = \lambda_i + \frac{c}{n}$$
$$\mathrm{Var}[r_i|\boldsymbol{\Omega}] = \sigma_i^2 = 2\lambda_i^2 + \tilde{b}_i^2$$

of r_i, so that

$$\sum_{i=1}^{n}\mathbb{E}[r_i|\boldsymbol{\Omega}] = c_2 - c_1 + \gamma c_{\mathbf{x}}\mathfrak{D} = \mathrm{E}_a$$

$$s_n^2 = \sum_{i=1}^{n}\sigma_i^2 = 2\operatorname{tr}(\mathbf{A}^2) + \mathbf{b}^{\mathsf{T}}\mathbf{b}$$
$$= 8\gamma^2c_1^2c_2^2\left(f''(\tau)\right)^2\frac{\left(\operatorname{tr}\left(\mathbf{C}_2 - \mathbf{C}_1\right)\right)^2}{p^2}\frac{\operatorname{tr}\mathbf{C}_a^2}{p^2}$$
$$+ 4\gamma^2\left(\frac{f'(\tau)}{n}\right)^2\mathbf{y}^{\mathsf{T}}\mathbf{P}\boldsymbol{\Omega}^{\mathsf{T}}\frac{\mathbf{C}_a}{p}\boldsymbol{\Omega}\mathbf{P}\mathbf{y}$$
$$+ \frac{16\gamma^2c_1^2c_2^2(f'(\tau))^2}{p}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^{\mathsf{T}}\frac{\mathbf{C}_a}{p}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)$$
$$+ O(n^{-\frac{5}{2}}).$$

We shall rewrite $\boldsymbol{\Omega}$ into two blocks as:

$$\boldsymbol{\Omega} = \left[\frac{(\mathbf{C}_1)^{\frac{1}{2}}}{\sqrt{p}}\mathbf{Z}_1, \quad \frac{(\mathbf{C}_2)^{\frac{1}{2}}}{\sqrt{p}}\mathbf{Z}_2\right]$$

where $\mathbf{Z}_1 \in \mathbb{R}^{p\times n_1}$ and $\mathbf{Z}_2 \in \mathbb{R}^{p\times n_2}$ with i.i.d. Gaussian entries with zero mean and unit variance. Then

$$\boldsymbol{\Omega}^{\mathsf{T}}\frac{\mathbf{C}_a}{p}\boldsymbol{\Omega} = \frac{1}{p^2}\begin{bmatrix}\mathbf{Z}_1^{\mathsf{T}}(\mathbf{C}_1)^{\frac{1}{2}}\mathbf{C}_a(\mathbf{C}_1)^{\frac{1}{2}}\mathbf{Z}_1 & \mathbf{Z}_1^{\mathsf{T}}(\mathbf{C}_1)^{\frac{1}{2}}\mathbf{C}_a(\mathbf{C}_2)^{\frac{1}{2}}\mathbf{Z}_2 \\ \mathbf{Z}_2^{\mathsf{T}}(\mathbf{C}_2)^{\frac{1}{2}}\mathbf{C}_a(\mathbf{C}_1)^{\frac{1}{2}}\mathbf{Z}_1 & \mathbf{Z}_2^{\mathsf{T}}(\mathbf{C}_2)^{\frac{1}{2}}\mathbf{C}_a(\mathbf{C}_2)^{\frac{1}{2}}\mathbf{Z}_2\end{bmatrix}$$

and with $\mathbf{Py} = \mathbf{y} - (c_2 - c_1)\mathbf{1}_n$, we deduce

$$\begin{aligned}\mathbf{y}^\mathsf{T}\mathbf{P}\boldsymbol{\Omega}^\mathsf{T}\frac{\mathbf{C}_a}{p}\boldsymbol{\Omega}\mathbf{P}\mathbf{y} = \frac{4}{p^2}\Big(&c_2^2\mathbf{1}_{n_1}^\mathsf{T}\mathbf{Z}_1^\mathsf{T}(\mathbf{C}_1)^{\frac{1}{2}}\mathbf{C}_a(\mathbf{C}_1)^{\frac{1}{2}}bZ_1\mathbf{1}_{n_1}\\ &-2c_1c_2\mathbf{1}_{n_1}^\mathsf{T}\mathbf{Z}_1^\mathsf{T}(\mathbf{C}_1)^{\frac{1}{2}}\mathbf{C}_a(\mathbf{C}_2)^{\frac{1}{2}}\mathbf{Z}_2\mathbf{1}_{n_2}\\ &+c_2^2\mathbf{1}_{n_1}^\mathsf{T}\mathbf{Z}_2^\mathsf{T}(\mathbf{C}_2)^{\frac{1}{2}}\mathbf{C}_a(\mathbf{C}_2)^{\frac{1}{2}}\mathbf{Z}_2\mathbf{1}_{n_2}\Big).\end{aligned}$$

Since $\mathbf{Z}_i\mathbf{1}_{n_i} \sim \mathcal{N}(\mathbf{0}, n_i\mathbf{I}_{n_i})$, by applying the trace lemma [39, Lemma B.26] we get

$$\mathbf{y}^\mathsf{T}\mathbf{P}\boldsymbol{\Omega}^\mathsf{T}\frac{\mathbf{C}_a}{p}\boldsymbol{\Omega}\mathbf{P}\mathbf{y} - \frac{4nc_1^2c_2^2}{p^2}\left(\frac{\operatorname{tr}\mathbf{C}_1\mathbf{C}_a}{c_1} + \frac{\operatorname{tr}\mathbf{C}_2\mathbf{C}_a}{c_2}\right) \overset{\text{a.s.}}{\to} 0. \tag{25}$$

Consider now the events

$$E = \left\{\left|\mathbf{y}^\mathsf{T}\mathbf{P}\boldsymbol{\Omega}^\mathsf{T}\frac{\mathbf{C}_a}{p}\boldsymbol{\Omega}\mathbf{P}\mathbf{y} - \rho\right| < \epsilon\right\}$$
$$\bar{E} = \left\{\left|\mathbf{y}^\mathsf{T}\mathbf{P}\boldsymbol{\Omega}^\mathsf{T}\frac{\mathbf{C}_a}{p}\boldsymbol{\Omega}\mathbf{P}\mathbf{y} - \rho\right| > \epsilon\right\}$$

for any fixed ϵ with $\rho = \frac{4nc_1^2c_2^2}{p^2}\left(\frac{\operatorname{tr}\mathbf{C}_1\mathbf{C}_a}{c_1} + \frac{\operatorname{tr}\mathbf{C}_2\mathbf{C}_a}{c_2}\right)$ and write

$$\begin{aligned}\mathbb{E}\left[\exp\left(iun\frac{\tilde{g}(\mathbf{x}) - \mathrm{E}_a}{s_n}\right)\right] &= \mathbb{E}\left[\exp\left(iun\frac{\tilde{g}(\mathbf{x}) - \mathrm{E}_a}{s_n}\right)\middle|E\right]\\ \mathrm{P}(E) &+ \mathbb{E}\left[\exp\left(iun\frac{\tilde{g}(\mathbf{x}) - \mathrm{E}_a}{s_n}\right)\middle|\bar{E}\right]\mathrm{P}(\bar{E})\end{aligned} \tag{26}$$

We start with the variable $\tilde{g}(\mathbf{x})|E$ and check that Lyapunov's condition for $\bar{r}_i = r_i - \mathbb{E}[r_i]$, conditioning on E,

$$\lim_{n\to\infty}\frac{1}{s_n^4}\sum_{i=1}^n\mathbb{E}[|\bar{r}_i|^4] = 0$$

holds by rewriting

$$\lim_{n\to\infty}\frac{1}{s_n^4}\sum_{i=1}^n\mathbb{E}[|\bar{r}_i|^4] = \lim_{n\to\infty}\sum_{i=1}^n\frac{60\lambda_i^4 + 12\lambda_i^2\tilde{b_i}^2 + 3\tilde{b_i}^4}{s_n^4} = 0$$

since both λ_i and $\tilde{b}_i$ are of order $O(n^{-3/2})$.

As a consequence of the above, we have the CLT for the random variable $\tilde{g}(\mathbf{x})|E$, thus

$$\mathbb{E}\left[\exp\left(iun\frac{\tilde{g}(\mathbf{x}) - \mathrm{E}_a}{s_n}\right)\middle|E\right] \to \exp(-\frac{u^2}{2}).$$

Next, we see that the second term in (26) goes to zero because $\left|\mathbb{E}\left[\exp\left(iun\frac{\tilde{g}(\mathbf{x})-\mathrm{E}_a}{s_n}\right)|\bar{E}\right]\right| \leq 1$ and $\mathrm{P}(\bar{E}) \to 0$ from (25) and we eventually deduce

$$\mathbb{E}\left[\exp\left(iun\frac{\tilde{g}(\mathbf{x}) - \mathrm{E}_a}{s_n}\right)\right] \to \exp(-\frac{u^2}{2}).$$

With the help of Lévy's continuity theorem, we thus prove the CLT of the variable $n\frac{\tilde{g}(\mathbf{x})-\mathrm{E}_a}{s_n}$. Since $s_n^2 \to \mathrm{Var}_a$, with Slutsky's theorem, we have the CLT for $n\frac{\tilde{g}(\mathbf{x})-\mathrm{E}_a}{\sqrt{\mathrm{Var}_a}}$ (thus for $n\frac{\hat{g}(\mathbf{x})-\mathrm{E}_a}{\sqrt{\mathrm{Var}_a}}$), and eventually for $n\frac{g(\mathbf{x})-\mathrm{E}_a}{\sqrt{\mathrm{Var}_a}}$ by Theorem 1 which completes the proof.