Statistical Analysis and Improvement of Large Dimensional SVM

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

Abstract

In this work, the asymptotic classification performance of support vector machines, for simultaneously large and numerous (i.e., "large p, large n" with $p/n \to c \in (0, \infty)$) data arising from a Gaussian mixture model, is provided. The analysis provides as an aftermath a new parameter estimation method which we theoretically prove superior and much faster than traditional cross-validation schemes. This fact is furthermore verified on practical datasets. The article introduces those results and provides a step-by-step proof, based on two key ingredients: an original approach via the dual optimization problem and a quite versatile leave-one-out technique.

1. Introduction

With the advent of the big data era, a need for novel classification methods adapted to *large and numerous* datasets emerges. By means of a large dimensional statistical analysis, several recent works have notably assessed the asymptotic performance (sometimes even demonstrating as a by-product the asymptotic inconsistency) of many classical machine learning algorithms in this regime; this is particularly the case of several kernel-based methods for clustering (El Karoui, 2010; Couillet & Benaych-Georges, 2016; Cheng & Singer, 2013), classification (Elkhalil et al., 2017; Liao & Couillet, 2017; Mai & Couillet, 2017), and regression (El Karoui et al., 2013). The performance evaluation is the first step into the hyperparameter setting, the improvement, or the complete change of these approaches, as proposed in several occasions in these references. Yet, most of these works are so far anchored in the analysis of algorithms and methods assuming an *explicit formulation*; algorithms involving *implicit* solutions to convex (or non-convex) optimization schemes are much less tractable. A notable exception is the line of works initiated in (El Karoui et al., 2013) and rigorously concluded in (El Karoui, 2013; Donoho & Montanari, 2013) on the asymptotic performance of robust M-estimation and regression; in (El Karoui, 2013), the authors pursue a quite natural and convincing approach based on a "two-ways" leave-one-out method adapted to the "twice" large dimensional regime (in both the number and size of the data). Our technical derivations are inspired by this approach.

The present article concentrates on the asymptotic performance of the popular support vector machines (SVM) (Bishop, 2006) for simultaneously numerous and large dimensional data, statistically described through a Gaussian mixture model. Following up on (Liao & Couillet, 2017), where the asymptotic performance of the *explicit* least-square SVMs (LS-SVM) is studied and obtained in closed-form, we consider here the same growth rate assumptions and determine the asymptotic performance through an implicit (but numerically easy to evaluate) expression. This work notably generalizes (Huang, 2017), where the authors exploit a (non formally rigorous) statistical physics approach to evaluate asymptotic SVM performance in a simpler setting.[1] The present setting comprises more generic forms of Gaussian mixture models than in (Huang, 2017), provides a more intuitive proof approach as well as additional consequences. Notably, as opposed to (Huang, 2017), our proof technique relies on the dual formulation of the SVM convex optimization which naturally allows for a joint treatment of the asymptotic performance for both *hard-margin* and *soft-margin* SVM schemes (see (Bishop, 2006, Chapter 7) for instance).

More importantly, as an aftermath of the theoretical findings, for the soft-margin SVM method, we provide a consistent estimator for the asymptotically optimal margin parameter (hereby denoted τ) which is shown in numerical experiments – both for synthetic and realistic datasets – to outperform the standard cross-validation method with a comparatively negligible computational cost. This is explicitly detailed under the form of Algorithm 2.

One of our main objectives being to walk the reader through the quite versatile leave-one-out approach, applied here to the SVM dual optimization formulation (and which we believe could be applied to many other problems), the core of the article (Section 2) will consist in developing the technical details. The main theoretical and practical results

[1] It should be mentioned that, while the present work correctly recovers the findings of (Huang, 2017) in the hard-margin setting, our results slightly differ in the soft-margin case; the main theorem in (Huang, 2017) indeed mistakenly contains a factor 2 which should be replaced by α.

are subsequently found and discussed in Section 3, before concluding the article in Section 4.

2. Technical Aspects

2.1. System Model

As in classical supervised classification problems, we dispose of n observations (x_i, y_i), $x_i \in \mathbb{R}^p$ being data vectors and $y_i \in \{-1, 1\}$ indicating their classes. There are n_1 (resp., n_2) data in class $\mathcal{C}_1$ (resp., class $\mathcal{C}_2$), labeled by $y_i = -1$ (resp., $y_i = 1$). We consider datasets generated from Gaussian distributions with a so-called *spiked* covariance matrix model, where data variations are modeled by a few principal components plus i.i.d. background noise. Introduced by (Johnstone, 2001), spiked models are particularly suitable to analyze many high dimensional statistical inference problems (Hastie et al., 1995; Hoyle & Rattray, 2004; Telatar, 1999). The most general assumption is that of different means and different spiked covariances for the two classes. For technical convenience, we add a condition that all spikes live in the same eigenspace.[2]

In other words, for $k \in \{1, 2\}$,

$$\begin{aligned} &i \in \mathcal{C}_k \Leftrightarrow x_i \sim \mathcal{N}(\mu_k, C_k) \\ &C_k = \sigma^2 I_p + \sigma^2 \sum_{d=1}^{m} \left(l_{[k]d}\right)^2 v_d v_d^\mathsf{T} \end{aligned} \tag{1}$$

where $v_d^\mathsf{T} v_{d'} = \delta_{d,d'}$. With this model, we can have different "spikes" (i.e., eigenvectors) for two classes (e.g., $\lambda_{1d} = 0, \lambda_{2d} \neq 0$) as long as the spikes are orthogonal. In addition, and similar to (Liao & Couillet, 2017) for LS-SVM, we assume that, as $p \to \infty$, $\|\mu_1 - \mu_2\| = O(1)$, and that σ, m and the $l_{[k]d}$ remain fixed irrespective of p.

Note in passing that one can turn x_i into $x_i - (\mu_1 + \mu_2)/2$ without changing the classification problem. For this reason, we shall from now on take $\mu_1 = -\mu$, $\mu_2 = \mu$ without any loss of generality.

2.2. Technical Arguments

We aim to investigate the statistical behavior of the class separating hyperplane obtained at the output of the SVM optimization in the limit where $n, p \to \infty$ with $n/p \to \alpha > 0$. Our technical approach is inspired by the leave-one-out procedure (in both feature and sample dimensions) employed in (El Karoui et al., 2013) for a linear regression problem under convex constraints. The SVM problem is however different and raises many additional technical difficulties. Our method also fundamentally takes advantage of the dual formulation of the SVM optimization problem. The method is delineated step by step in Section 2, leading up to our main results in Section 3.

[2]This condition can be relaxed but leads to more sophisticated algebraic manipulations and shall only be treated in an extended version of the present article.

2.2.1. Optimization Problem

Support vector machines build a hyperplane $\beta^\mathsf{T} x + \beta_0 = 0$ (with dummy variable $x \in \mathbb{R}^p$) which separates the set of training data in two subsets of $\mathbb{R}^p$ with maximal gap between the subsets. For linearly separable data, the so-called hard-margin problem is defined as

$$\begin{aligned} &\min_{\beta} \|\beta\|^2 \\ &\text{s.t.}\ \forall i \in \{1, \ldots, n\},\ y_i(\beta^\mathsf{T} x_i + \beta_0) \geq 1. \end{aligned} \tag{2}$$

When no such hyperplane exists, one usually resorts to a soft-margin alternative given by

$$\begin{aligned} &\min_{\beta} \frac{1}{2}\|\beta\|^2 + \frac{\tau}{n}\sum_{i=1}^{n} \xi_n \\ &\text{s.t.}\ \forall i \in \{1, \ldots, n\},\ y_i(\beta^\mathsf{T} x_i + \beta_0) \geq 1 - \xi_i,\ \xi_i \geq 0. \end{aligned} \tag{3}$$

For both cases, a Lagrange multipliers approach gives the dual problem

$$\begin{aligned} &\max_{c_1 \cdots c_n} \sum_{i=1}^{n} c_i - \frac{1}{2}\sum_{i,j=1}^{n} c_i c_j y_i y_j x_i^\mathsf{T} x_j \\ &\text{s.t.}\ \forall i \in \{1, \ldots, n\},\ 0 \leq c_i \leq U, \sum_{i=1}^{n} c_i y_i = 0 \end{aligned} \tag{4}$$

where $U \equiv U_h = +\infty$ for the hard-margin problem (2) or $U \equiv U_s = \tau/n$ for the soft-margin problem (3). With the dual solutions c_i, the hyperplane direction β is obtained as

$$\beta = \sum_{i=1}^{n} y_i c_i x_i. \tag{5}$$

Obviously, c_i are random variables intricately dependent on the data $x_1, \ldots, x_n$. Our approach precisely focuses on understanding this dependence to then retrieve β from (5).

Already, by the Karush-Kuhn-Tucker conditions, we have

1. for hard-margin SVM
   - $c_i = 0$ if $y_i(\beta^\mathsf{T} x_i + \beta_0) > 1$,
   - $c_i > 0$ if $y_i(\beta^\mathsf{T} x_i + \beta_0) = 1$;
2. for soft-margin SVM
   - $c_i = 0$ if $y_i(\beta^\mathsf{T} x_i + \beta_0) > 1$,
   - $\frac{\tau}{n} > c_i > 0$ if $y_i(\beta^\mathsf{T} x_i + \beta_0) = 1$,
   - $c_i = \frac{\tau}{n}$ if $y_i(\beta^\mathsf{T} x_i + \beta_0) < 1$;

thereby leading up to the following expression for c_i

$$c_i = \phi\left(\frac{1}{\|x_i\|^2}\left[1 - y_i(b_{(i)}^\mathsf{T} x_i + \beta_0)\right]\right) \tag{6}$$

where $b_{(i)} = \sum_{j\neq i} y_j c_j x_j$ and $\phi \equiv \phi_h(t) = \max\{0, t\}$ for hard-margin SVM, $\phi \equiv \phi_s(t) = \max\{0, \min\{t, \tau/n\}\}$ for soft-margin SVM.

As $\frac{1}{p}\|x_i\|^2 \to \sigma^2$ when $p \to \infty$ (since $\|\mu\| = O(1)$), the key to quantifying c_i lies in $b_{(i)}^\mathsf{T} x_i$. But $b_{(i)}$ depends on x_i in a complex manner and not much can be said at this stage.

2.2.2. Leave-one-out

To "induce" asymptotic independence between x_i and $b_{(i)}$, we resort to the 'leave-one-out' method. We first define $\beta_{(i)}$ as the SVM solution obtained with all data except x_i and $c_{(i)j}$ as the corresponding dual problem coefficients. Since $\beta_{(i)}$ is independent of x_i, $\beta_{(i)}^\mathsf{T} x_i$ simply follows a normal distribution. Our goal is to establish a relation between $b_{(i)}^\mathsf{T} x_i$ and $\beta_{(i)}^\mathsf{T} x_i$, the proof steps of which are sketched below.

As $n \to \infty$, it is intuitive that for $j \neq i$, $\beta^\mathsf{T} x_j \simeq \beta_{(i)}^\mathsf{T} x_j$,[3] as accepted in (El Karoui et al., 2013). Similarly, $c_j \simeq c_{(i)j}$. Combined to the above relations between c_i and $y_i(\beta^\mathsf{T} x_i + \beta_0)$, we get

- if $y_j(\beta^\mathsf{T} x_j + \beta_0) > 1$, then $y_j(\beta_{(i)}^\mathsf{T} x_j + \beta_0) > 1$, so that $c_j = c_{(i)j} = 0$;
- if $y_j(\beta^\mathsf{T} x_j + \beta_0) < 1$, then $y_j(\beta_{(i)}^\mathsf{T} x_j + \beta_0) < 1$, and then $c_j = c_{(i)j} = \tau/n$;
- if $y_j(\beta^\mathsf{T} x_j + \beta_0) = 1$, then $0 < c_{(i)j} \simeq c_j < U$, hence $y_j(\beta_{(i)}^\mathsf{T} x_j + \beta_0) = 1$.

In summary, if x_j is a support vector on the margin's boundary (i.e., $y_j(\beta_{(i)}^\mathsf{T} x_j + \beta_0) = 1$), then

$$c_j - c_{(i)j} = -\frac{\sum_{k\neq i,j} y_k y_j (c_k - c_{(i)k}) x_k^\mathsf{T} x_j + y_i y_j c_i x_i^\mathsf{T} x_j}{\|x_j\|^2} \tag{7}$$

as a result of expanding β, $\beta_{(i)}$ in the equality $y_j(\beta^\mathsf{T} x_j + \beta_0) = y_j(\beta_{(i)}^\mathsf{T} x_j + \beta_0) = 1$; otherwise $c_j - c_{(i)j} = 0$.

Denote the number of support vectors on the margin's boundary by n_b, the corresponding matrix $X_b \in \mathbb{R}^{p\times n_b}$ with columns $y_j x_j$, and $\Delta_{(i)}c \in \mathbb{R}^{n_b}$ the vector composed of the differences $c_j - c_{(i)j}$, for all x_j support vectors. Then (7) can be written in the following matrix form

$$\Delta_{(i)}c = -(\hat{X}_b^\mathsf{T}\hat{X}_b)^{-1}\hat{X}_b^\mathsf{T} x_i c_i. \tag{8}$$

[3] $a \simeq b$ informally stands here for $\|a - b\|/\|a\| = o(1)$, as $n, p \to \infty$.

Looking closely now at the difference between $b_{(i)}^\mathsf{T} x_i$ and $\beta_{(i)}^\mathsf{T} x_i$, we find that, with $\mathcal{B}$ the set of support vectors,

$$\begin{aligned} b_{(i)}^\mathsf{T} x_i - \beta_{(i)}^\mathsf{T} x_i &= \sum_{j\neq i} (c_j - c_{(i)j}) y_j y_i x_j^\mathsf{T} x_i \\ &= \sum_{j\in\mathcal{B}\setminus\{i\}} (c_j - c_{(i)j}) y_j y_i x_j^\mathsf{T} x_i \\ &= \Delta_{(i)}c^\mathsf{T} \hat{X}_s^\mathsf{T} x_i c_i y_i \\ &= -c_i x_i^\mathsf{T} \hat{X}_b (\hat{X}_b^\mathsf{T}\hat{X}_b)^{-1}\hat{X}_b^\mathsf{T} x_i. \end{aligned} \tag{9}$$

Since $\hat{X}_b(\hat{X}_b^\mathsf{T}\hat{X}_b)^{-1}\hat{X}_b^\mathsf{T}$ is a projection matrix of rank n_b, and x_i is independent of x_j, $j \neq i$, by (Bai & Silverstein, 2010, Lemma B.26),

$$b_{(i)}^\mathsf{T} x_i - \beta_{(i)}^\mathsf{T} x_i \simeq -\sigma^2 n_b c_i. \tag{10}$$

An important remark at this point is that c_i is not defined if $p = n_b$. So from now on, when discussing c_i, it is automatically assumed that $p \neq n_b$. However this condition may not hold for hard-margin SVM and we refer the reader to Section 3.3.3 for a detailed discussion. Note also that $n_b \leq p$ almost surely since the x_i's are almost surely linearly independent and thus no more than p data points can fill the hyperplane (of dimension $p-1$). Combining (6) and (10), the dual variables c_i can then be expressed as

$$c_i \simeq \phi\left(\frac{\eta_i - (\kappa_i + y_i\beta_0 - 1)}{(p - n_b)\sigma^2}\right) \tag{11}$$

with $\kappa_i = \beta_{(i)}^\mathsf{T}\mu$ and $\eta_i = -\beta_{(i)}^\mathsf{T}(y_i x_i - \mu) \sim \mathcal{N}(0, e_i^2)$ for $e_i = \beta_{(i)}^\mathsf{T} C_k \beta_{(i)}$ such that $i \in \mathcal{C}_k$.

Equation (11) is crucial for our analysis, as it gives access to the statistical behavior of the dual coefficients c_i.

2.2.3. Statistics of the SVM Hyperplane

With c_i characterized by (11), the remaining task is to deal with the dependence between c_i and x_i so as to provide the statistical characterization of the SVM hyperplane. Let us begin by introducing some notations:

- $r = \|\mu\|$, $\rho = r/\sigma$,[4] $s_d = v_d^\mathsf{T}\mu/r$ for $d \in \{1, \ldots, m\}$, $s_{m+1} = \sqrt{1 - \sum_{d=1}^m s_d^2}$ and $s_d = 0$ for $d \in \{m+2, \ldots, p\}$;
- for $k \in \{1, 2\}$, $d \in \{1, \ldots, p\}$, $\lambda_{[k]d}^2$ is the d-th eigenvalue of C_k, i.e., $\lambda_{[k]d}^2 = \sigma^2(1 + l_{[k]d}^2)$ for $d \in \{1, \ldots, m\}$ and $\lambda_{[k]d}^2 = \sigma^2$ for $d \in \{m+1, \cdots, p\}$;
- $\kappa = \beta^\mathsf{T}\mu$, $e_{[k]}^2 = \beta^\mathsf{T} C_k \beta$ for $k \in \{1, 2\}$.

[4] for any symbol given in square, e.g., a^2, a is by default its positive root, i.e., $a = \sqrt{a^2}$, unless specified otherwise.

Before discussing the dependence between c_i and x_i, it is convenient to project the x_i's in a basis where they become statistically independent. To this end, consider an orthonormal basis $V = [v_1, \cdots, v_{m+1}, v_{m+2}, \cdots, v_p]$ where $v_1, \cdots, v_m$ are defined in (1); as for v_{m+1}, it is given by $v_{m+1} = \left(\mu - \sum_{d=1}^m (\mu^\mathsf{T} v_d) v_d\right) / \|(\mu - \sum_{d=1}^m (\mu^\mathsf{T} v_d) v_d\|$ if $\mu - \sum_{d=1}^m (\mu^\mathsf{T} v_d) v_d \neq 0$ and arbitrary otherwise. In particular, $\mu = \sum_{d=1}^p s_d v_d$ and $\beta = \sum_{d=1}^p t_d v_d$ (and the same is understood with $t_{(i)d}$ for $\beta_{(i)}$). By Gaussian invariance,

$$y_i x_i = r \sum_{d=1}^p s_d v_d + \sum_{d=1}^p \lambda_{[k]d} [z_i]_d v_d \tag{12}$$

for $i \in \mathcal{C}_k$, with $z_i = \left[[z_i]_1, \cdots, [z_i]_p\right] \sim \mathcal{N}(0, I_p)$. Thus, Equation (5) gives directly that

$$t_d = r s_d \sum_{i=1}^n c_i + \sum_{k=\{1,2\}} \lambda_{[k]d} \sum_{i \in \mathcal{C}_1} c_i [z_i]_d. \tag{13}$$

Also, from the definitions following (11), $g_i = \eta_i / e_i = -\sum_{d=1}^p \lambda_{[k]d} t_{(i)d} [z_i]_d / e_i \sim \mathcal{N}(0,1)$.

Still following a leave-one-out intuition, when $n \to \infty$, removing or adding one datum should have negligible impact on the hyperplane, i.e., $\kappa_i \simeq \kappa$, and $e_i \simeq e_{[k]}$ for $i \in \mathcal{C}_k$. It follows that for $i \in \mathcal{C}_k$,

$$c_i \simeq \frac{e_{[k]}}{(p - n_b)\sigma^2} \hat{\phi}_{[k]}(g_i - \gamma_k) \tag{14}$$

where $\gamma_1 = \frac{\kappa - 1 - \beta_0}{e_{[1]}}$, $\gamma_2 = \frac{\kappa - 1 + \beta_0}{e_{[2]}}$ and $\hat{\phi}_{[k]}(t) = \hat{\phi}_{h[k]}(t) \equiv \max\{0, t\}$ for hard-margin, $\hat{\phi}_{[k]}(t) = \hat{\phi}_{s[k]}(t) \equiv \max\left\{0, \min\left\{t, (p - n_b)\sigma^2 \tau / n e_{[k]}\right\}\right\}$ for soft-margin.

Let $n_b = n_{b[1]} + n_{b[2]}$ with $n_{b[k]}$ the number of support vectors for class k. As discussed in Section 2.2.1, $n_{b[k]} = \sum_{i \in \mathcal{C}_k} \mathbf{1}_{(0,+\infty)}(c_i)$ for hard-margin and $n_{b[k]} = \sum_{i \in \mathcal{C}_k} \mathbf{1}_{(0,\tau/n)}(c_i)$ for soft-margin. Then, by (14),

$$n_{b[k]} \simeq n_k \mathbb{E}[\delta_{[k]}(g_i - \gamma_k)] \tag{15}$$

with $\delta_{[k]}(t) = \delta_{h[k]}(t) \equiv \mathbf{1}_{(0,+\infty)}(t)$ for hard-margin, $\delta_{[k]}(t) = \delta_{s[k]}(t) \equiv \mathbf{1}_{(0,(p-n_b)\sigma^2\tau/ne_{[k]})}(t)$ for soft-margin.

Now, in order to tackle the dependence between c_i and $[z_i]_d$, we build the following variables, jointly Gaussian with g_i:

$$[h_i]_d = -\frac{\lambda_{[k]d} t_{(i)d}}{e_i \sqrt{e_i^2 - \lambda_{[k]d}^2 t_{(i)d}^2}} \sum_{q \neq d} \lambda_{[k]q} t_{(i)q} [z_i]_q + \frac{\sqrt{e_i^2 - \lambda_{[k]d}^2 t_{(i)d}^2}}{e_i} [z_i]_d \sim \mathcal{N}(0,1). \tag{16}$$

Since uncorrelated jointly Gaussian variables are independent and $\mathbb{E}[g_i [h_i]_d] = 0$, $[h_i]_d$ are indeed independent of g_i, hence asymptotically independent of c_i.

Writing $[z_i]_d = \frac{1}{e_i}\sqrt{e_i^2 - \lambda_{[k]d}^2 t_{(i)d}^2}[h_i]_d - \frac{1}{e_i}\lambda_{[k]d} t_{(i)d} g_i$, (13) becomes

$$t_d = r s_d \sum_{i=1}^n c_i - \sum_{k=\{1,2\}} \left[\lambda_{[k]d}^2 \sum_{i \in \mathcal{C}_k} \frac{t_{(i)d} c_i g_i}{e_i} + \lambda_{[k]d} \sum_{i \in \mathcal{C}_k} \frac{\sqrt{e_i^2 - \lambda_{[k]d}^2 t_{(i)d}^2} c_i [h_i]_d}{e_i} \right]. \tag{17}$$

By the same asymptotic arguments, $t_{(i)d} \simeq t_d$, from which it follows that for d such that $s_d \neq 0$, t_d approaches the expectation

$$t_d \simeq \frac{r s_d \sum_{k \in \{1,2\}} e_{[k]} n_k \mathbb{E}[\hat{\phi}_{[k]}(g_i - \gamma_k)]}{(p - n_b)\sigma^2 + \sum_{k \in \{1,2\}} \lambda_{[k]d}^2 n_{b[k]}}, \tag{18}$$

which follows from (17) and $\mathbb{E}[\hat{\phi}_{[k]}(g_i - \gamma_k) g_i] = \mathbb{E}[\delta_{[k]}(g_i - \gamma_k)]$ (obtained by integration by parts). Also, for d such that $s_d = 0$, the expectation of t_d is zero and

$$\sum_{d=m+2}^p t_d^2 \simeq \frac{\sum_{k \in \{1,2\}} e_{[k]}^2 n_{[k]} \mathbb{E}\{\hat{\phi}_{[k]}^2 (g_i - \gamma_k)\}}{p\sigma^2}. \tag{19}$$

Since $\kappa \simeq r \sum_{d=1}^{m+1} t_d s_d$ and $e_{[k]} \simeq \sum_{d=1}^{m+1} t_d^2 \lambda_{[k]d}^2 + \sigma^2 \sum_{d=m+2}^p t_d^2$, adding that

$$\frac{n_1}{n} e_{[1]} \mathbb{E}[\hat{\phi}(g_i - \gamma_1)] \simeq \frac{n_2}{n} e_{[2]} \mathbb{E}[\hat{\phi}(g_i - \gamma_2)] \tag{20}$$

as a result of $\sum_{i=1}^n y_i c_i = 0$, which is a constraint of the dual optimization given in (4), we can determine the asymptotic values of κ, $e_{[k]}$, β_0, and thus the asymptotic performance, as formally stated in Theorem 1.

3. Main Results

3.1. Asymptotic performance

Using the asymptotic statistical behavior of the hyperplane determined in Section 2.2.3, we are able to retrieve the asymptotic SVM performance.

Theorem 1. *Let x_i be defined as in (1) with n_k elements in class $\mathcal{C}_k$. Let $D(x) = \beta^\mathsf{T} x + \beta_0$ for β, β_0 solution to (2) (hard-margin) or (3) (soft-margin). Then, as $n, p \to \infty$ with $n/p \to \alpha > 0$ and $n_k/p \to \alpha_k > 0$, for $\tilde{x}_k \sim \mathcal{N}(\mu_k, C_k)$,*

$$\mathbb{P}\left[D(\tilde{x}_1) < 0\right] \to 1 - Q(\zeta_1) \tag{21}$$

$$\mathbb{P}\left[D(\tilde{x}_2) > 0\right] \to 1 - Q(\zeta_2) \tag{22}$$

where $Q(t) = \frac{1}{\sqrt{2\pi}} \int_t^\infty e^{-\frac{1}{2}u^2} du$ and, for $k \in \{1,2\}$,

$$\zeta_k = \gamma_k + \frac{1}{e_{[k]}} \tag{23}$$

with γ_k, $e_{[k]}$ defined, for hard-margin SVM as the solutions to the equations

$$M_h \sum_{d=1}^{m+1} \frac{s_d^2}{1+l_{[1]d}^2 R_{h[1]} + l_{[2]d}^2 R_{h[2]}} = \frac{e_{[1]}\gamma_1 + e_{[2]}\gamma_2 + 2}{2} \tag{24}$$

$$M_h^2 \sum_{d=1}^{m+1} \frac{s_d^2(1+l_{[1]d}^2)}{\left(1+l_{[1]d}^2 R_{h[1]} + l_{[2]d}^2 R_{h[2]}\right)^2} + L_h = e_{[1]}^2 \tag{25}$$

$$M_h^2 \sum_{d=1}^{m+1} \frac{s_d^2(1+l_{[2]d}^2)}{\left(1+l_{[1]d}^2 R_{h[1]} + l_{[2]d}^2 R_{h[2]}\right)^2} + L_h = e_{[2]}^2 \tag{26}$$

$$\alpha_1 e_{[1]} F_h(\gamma_1) = \alpha_2 e_{[2]} F_h(\gamma_2) \tag{27}$$

or, for soft-margin SVM as the solutions to the equations

$$M_s \sum_{d=1}^{m+1} \frac{s_d^2}{1+l_{[1]d}^2 R_{s[1]} + l_{[2]d}^2 R_{s[2]}} = \frac{e_{[1]}\gamma_1 + e_{[2]}\gamma_2 + 2}{2} \tag{28}$$

$$M_s^2 \sum_{d=1}^{m+1} \frac{s_d^2(1+l_{[1]d}^2)}{\left(1+l_{[1]d}^2 R_{s[1]} + l_{[2]d}^2 R_{s[2]}\right)^2} + L_s = e_{[1]}^2 \tag{29}$$

$$M_s^2 \sum_{d=1}^{m+1} \frac{s_d^2(1+l_{[2]d}^2)}{\left(1+l_{[1]d}^2 R_{s[1]} + l_{[2]d}^2 R_{s[2]}\right)^2} + L_s = e_{[2]}^2 \tag{30}$$

$$\alpha_1 e_{[1]} F_s(\gamma_1, b_u/e_{[1]}) = \alpha_2 e_{[2]} F_s(\gamma_2, b_u/e_{[2]}) \tag{31}$$

$$\frac{\sigma^2\tau}{\alpha}\left(1 - R_{s[1]} - R_{s[2]}\right) = b_u \tag{32}$$

where we introduced the notations $\mu = (\mu_2 - \mu_1)/2$, ρ, s_d as in Section 2.2.3, $l_{[1]m+1} = l_{[2]m+1} = 0$,

$$F_h(t) = \mathbb{E}_{z\sim\mathcal{N}(0,1)}[\psi_h(z-t)] \tag{33}$$

$$G_h(t) = \mathbb{E}_{z\sim\mathcal{N}(0,1)}[\psi_h^2(z-t)] \tag{34}$$

$$F_s(t_1,t_2) = \mathbb{E}_{z\sim\mathcal{N}(0,1)}[\psi_s(z-t_1,t_2)] \tag{35}$$

$$G_s(t_1,t_2) = \mathbb{E}_{z\sim\mathcal{N}(0,1)}[\psi_s^2(z-t_1,t_2)] \tag{36}$$

$$\psi_h(\nu) = \max\{0,\nu\} \tag{37}$$

$$\psi_s(\nu_1,\nu_2) = \max\{0,\min\{\nu_1,\nu_2\}\} \tag{38}$$

$$M_h = \sum_{q\in\{1,2\}} \rho\alpha_q F_h(\gamma_q) e_{[q]} \tag{39}$$

$$L_h = \sum_{q\in\{1,2\}} e_{[q]}^2 \alpha_{[q]} G_h(\gamma_q) \tag{40}$$

$$R_{h[q]} = \alpha_q Q(\gamma_q) \tag{41}$$

$$M_s = \sum_{q=\{1,2\}} \rho\alpha_q F_s(\gamma_q, b_u/e_{[q]}) e_{[q]} \tag{42}$$

$$L_s = \sum_{q\in\{1,2\}} e_{[q]}^2 \alpha_{[q]} G_s(\gamma_q, b_u/e_{[q]}) \tag{43}$$

$$R_{s[q]} = \alpha_q \left[Q(\gamma_q) - Q(\gamma_q b_u/e_{[q]})\right]. \tag{44}$$

Besides, $\sqrt{\beta^\mathsf{T} C_k \beta} \to e_{[k]}$, $(\beta^\mathsf{T}\mu - 1 - \beta_0)/\sqrt{\beta^\mathsf{T} C_1 \beta} \to \gamma_1$ and $(\beta^\mathsf{T}\mu - 1 + \beta_0)/\sqrt{\beta^\mathsf{T} C_2 \beta} \to \gamma_2$.

We confirm our asymptotic results with simulations on finite dimensional Gaussian datasets, as displayed in Figure 1.

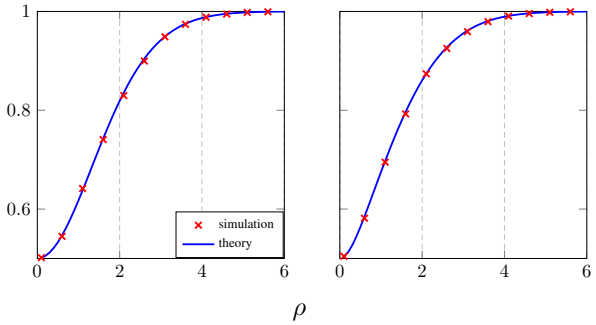

Figure 1. Theoretical and empirical average precision of (**left:** hard-margin, **right:** soft-margin with $\tau = 0.5$) SVM as a function of ρ for two-class Gaussian data with $n = p = 256$, $n_1 = n_2$, $m = 3$, $s_{1-3} = [1/\sqrt{2}, 1/\sqrt{8}, 1/\sqrt{8}]$, $l_{[1]1-3} = [\sqrt{4}, \sqrt{3}, 0]$, $l_{[2]1-3} = [\sqrt{4}, 0, \sqrt{3}]$. Averaged over 100 random iterations.

Note that if $C_1 = C_2$, we get directly $e_{[1]} = e_{[2]}$ for $e_{[k]}$ is the limit of $\sqrt{\beta^\mathsf{T} C_k \beta}$. Moreover, $\gamma_1 = \gamma_2$ is an immediate consequence of (27) or (31) when $n_1 = n_2$. In this case, the results are much simpler.

Corollary 1. *Let $C_1 = C_2$ and $n_1 = n_2$. Then, under the conditions of Theorem 1 and with the same notations, $\gamma_1 = \gamma_2$, $e_{[1]} = e_{[2]}$, and $\zeta_1 = \zeta_2$. Besides, for hard-margin SVM, γ_1 is determined by*

$$2\rho^2\alpha_1^2 F_h^2(\gamma_1) \sum_{d=1}^{m+1} \frac{s_d^2(1+l_{[1]d}^2)}{\left(1+2l_{[1]d}^2\alpha_1 Q(\gamma_1)\right)^2} + \alpha_1 G_h(\gamma_1) = 1, \tag{45}$$

from which $e_{[1]}$ is given explicitly by

$$e_{[1]} = \left(2\rho\alpha_1 F_h(\gamma_1) \sum_{d=1}^{m+1} \frac{s_d^2}{1+2l_{[1]d}^2\alpha_1 Q(\gamma_1)} - \gamma_1\right)^{-1} \tag{46}$$

and, for soft-margin SVM, γ_1 is determined by two non-linear equations

$$2\rho^2\alpha_1^2 F_s^2(\gamma_1, b_u') \sum_{d=1}^{m+1} \frac{s_d^2(1+l_{[1]d}^2)}{\left(1+l_{[1]d}^2 R_s\right)^2} + \alpha_1 G_s(\gamma_1, b_u') = 1 \tag{47}$$

$$\frac{\sigma^2\tau}{\alpha}(1-R_s)\left(2\rho\alpha_1 F_s(\gamma_1, b_u') \sum_{d=1}^{m+1} \frac{s_d^2}{1+l_{[1]d}^2 R_s} - \gamma_1\right) = b_u' \tag{48}$$

where $R_s = \alpha(Q(\gamma_1) - Q(\gamma_1 + b'_u))$, *and* $e_{[1]}$ *is given by*

$$e_{[1]} = \left(2\rho\alpha_1 F_s(\gamma_1, b'_u) \sum_{d=1}^{m+1} \frac{s_d^2}{1 + l_{[1]d}^2 R_s} - \gamma_1\right)^{-1}. \tag{49}$$

Figure 2 displays the soft-margin precision as a function of τ using Corollary 1. We find that when the 'spike' vector v_1 is aligned with μ (left of Figure 2), as assumed in (Huang, 2017), the average precision is a monotone decreasing function of τ, in which case small τ values are optimal; on the other hand, when v_1 is different in direction from μ (right of Figure 2), there exists a τ maximizing the average precision. This phenomenon is discussed in Subsection 3.3.1.

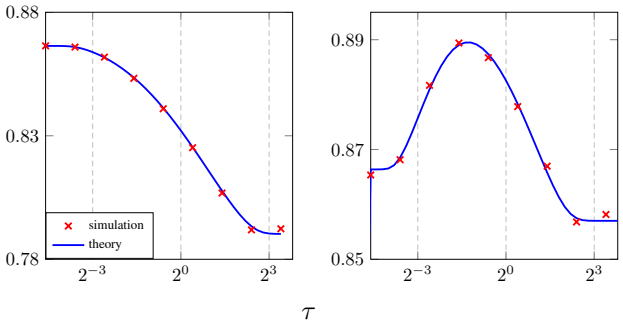

Figure 2. Theoretical and empirical average precision as a function of τ for two-class Gaussian data with $n = p = 256$, $n_1 = n_2$, $m = 1$, $s_1 = 1$, $l_{[1]1} = l_{[2]1} = \sqrt{2}$ **(left)** or $n = p = 256$, $n_1 = n_2$, $m = 1$, $s_1 = 1/\sqrt{2}$, $l_{[1]1} = l_{[2]1} = \sqrt{4}$ **(right)**. Averaged over 100 random iterations.

3.2. Practical SVM improvement

Theorem 1 allows us to obtain the value of τ yielding the maximum precision. In practice though, this is only achieved if ρ, s_d, $l_{[q]d}$ can be estimated from the data. Letting $\hat{\rho} = \hat{r}/\hat{\sigma}$ with

$$\hat{\sigma}^2 = \frac{1}{np}\sum_{i=1}^{n} x_i^\mathsf{T} x_i = \sigma^2 + O(n^{-1}) \tag{50}$$

$$\hat{r}^2 = \frac{1}{4}\left(\left\|\frac{1}{n_2}\sum_{i\in\mathcal{C}_2} x_i - \frac{1}{n_1}\sum_{i\in\mathcal{C}_1} x_i\right\|^2 - \left(\frac{p}{n_1} + \frac{p}{n_2}\right)\hat{\sigma}^2\right) \tag{51}$$

$$= r^2 + O(n^{-1}) \tag{52}$$

we have that $\hat{\rho} = \rho + O(n^{-1})$. The estimation of s_d and $l_{[k]d}$ are only possible if $l_{[k]d}^2 > \sqrt{1/\alpha}$, where $\alpha = \lim n/p$. Based on (Baik & Silverstein, 2006; Paul, 2007), s_d, $l_{[k]d}$ are consistently estimated by the output of Algorithm 1.

Algorithm 1 Estimation of data parameters

Input: Data set $(x_1, y_1), \cdots, (x_n, y_n)$. Parameter ϵ.
Output: Estimations of data parameters $\hat{r}$, s_d, $l_{[k]d}$.
Let $\hat{\sigma}$, $\hat{r}$ be defined as in (50)–(51) and write $\hat{\mu} = \frac{1}{2}\left(\frac{1}{n_2}\sum_{i\in\mathcal{C}_2} x_i - \frac{1}{n_1}\sum_{i\in\mathcal{C}_1} x_i\right)$.
For $k \in \{1, 2\}$, compute the eigenvalues and eigenvectors of the sample covariance matrices $\hat{C}_k = \frac{1}{n_k}\sum_{i\in\mathcal{C}_k}\left(x_i - \frac{1}{n_k}\sum_{j\in\mathcal{C}_k} x_j\right)\left(x_i - \frac{1}{n_k}\sum_{j\in\mathcal{C}_k} x_j\right)^\mathsf{T}$ and store all eigenvalues greater than $(1 + \sqrt{1/\alpha})^2$ as $\tilde{\lambda}_{[k]} = [\tilde{\lambda}_{[k]1}, \cdots, \tilde{\lambda}_{[k]m_k}]$ and their corresponding eigenvectors $\tilde{V}_{[k]} = [\tilde{v}_{[k]1}, \cdots, \tilde{v}_{[k]m_k}]$. Set $\tilde{l}_{[k]} = [\tilde{l}_{[k]1} \cdots, \tilde{l}_{[k]m_k}]$ with $\tilde{l}_{[k]d} = \sqrt{\tilde{\lambda}_{[k]d}/\hat{\sigma}^2 - 1}$.
Define
$$H(t_1, t_2) = \frac{1}{2}\left(t_1 - \frac{1}{t_2} + \sqrt{\left(t_1 - \frac{1}{t_2}\right)^2 - \frac{4}{t_2}}\right)$$
$$P(t_1, t_2) = \sqrt{\left(1 - \frac{1}{t_2 t_1^2}\right)\Big/\left(1 + \frac{1}{t_2 t_1}\right)}$$
for $d \in \{1, \cdots, m_1\}$ **do**
  $\hat{l}_{[1]d}^2 = H(\tilde{l}_{[1]d}^2, \alpha_1)$ and $\hat{s}_d = \frac{\hat{\mu}^\mathsf{T}\tilde{v}_{[1]d}}{\hat{r}P(\tilde{l}_{[1]d}^2, \alpha_1)}$
  do $d' = 1$ **while** $d' \in \{1, \cdots, m_2\}$
    $\hat{l}^2 = H(\tilde{l}_{[2]d'}^2, \alpha_2)$ and $q = \frac{|\tilde{v}_{[1]d}^\mathsf{T}\tilde{v}_{[2]d'}|}{P(\tilde{l}_{[1]d}^2, \alpha_1)P(\tilde{l}, \alpha_2)}$
    if $q > 1 - \epsilon$
      $\hat{l}_{[2]d}^2 = \hat{l}^2$, remove $\tilde{\lambda}_{[2]d'}$, $\tilde{v}_{[2]d'}$ from $\tilde{\lambda}_{[2]}$, $\tilde{V}_{[2]}$
      $m_2 \leftarrow m_2 - 1$
    else $\hat{l}_{[2]d}^2 = 0$
for $d \in \{1, \cdots, m_2\}$ **do**
  $\hat{l}_{[1]d+m_1}^2 = 0$, $\hat{l}_{[2]d+m_1}^2 = H(\tilde{l}_{[2]d}^2, \alpha_2)$
  $\hat{s}_{d+m_1} = \frac{\hat{\mu}^\mathsf{T}\tilde{v}_{[2]d}}{\hat{r}P(\tilde{l}_{[2]d+m_1}^2, \alpha_2)}$
Return $\hat{r}$, $\hat{s}$, $\hat{l}_{[1]}$ and $\hat{l}_{[2]}$.

Algorithm 2 Optimization of τ

Input: Data set $(x_1, y_1), \cdots, (x_n, y_n)$. Performance measure (e.g. average precision).
Output: the value of τ maximizing the performance measure.
Obtain data parameters estimations $\hat{r}$, $\hat{s}$, $\hat{l}_{[1]}$ and $\hat{l}_{[2]}$ by Algorithm 1.
Compute the performance measure $P(\tau)$ as a function of τ using Theorem 1.
Find $\hat{\tau} = \arg\max_\tau P(\tau)$.
Return $\hat{\tau}$.

Optimally tuning the hyperparameter τ to improve the generalization performance of SVMs requires in general to perform cross-validation on the x_i's by partitioning them into multiple subsets, leaving one subset out of the training process as the test set, and iterating over all such subsets.

From our analysis, since the optimal value for τ depends on the ratio n/p, cross-validation affects this ratio and is thus suboptimal. In contrast, our proposed estimation approach for τ, consisting in estimating $\rho, s_d, l_{[k]d}$ through Algorithm 1 and then evaluating the performance for all τ from Theorem 1, is *asymptotically* optimal. This procedure is formalized in Algorithm 2. To validate this remark, we apply Algorithm 2 and cross-validation to search for the optimal τ in terms of average precision, and compare their effectiveness measured by the difference between the average precision obtained using these methods and an oracle value. Table 1 provides the comparison on Gaussian data and a noisy version of the MNIST dataset.[5] It is observed that our method works significantly better and offers more stability on these datasets. Indeed, our method allows to close the gap with oracle performance by more than 50% in comparison with the 6-fold cross validation.

Gaussian data	Algorithm 2	3-fold	6-fold
p=128	**0.33±0.38**	0.71±0.86	0.64±0.67
p=256	**0.038±0.039**	0.25±0.27	0.15±0.18
p=512	**0.036±0.032**	0.16±0.12	0.14±0.16
MNIST data	Algorithm 2	3-fold	6-fold
(1,7)	**0.099± 0.063**	0.35± 0.35	0.23±0.19
(6,9)	**0.076±0.061**	0.21±0.21	0.14±0.18
(0,8)	**0.087± 0.092**	0.28±0.23	0.29±0.29

Table 1. Difference (in %) between the oracle average precision and those obtained using various methods for Gaussian data with $n = 3p$, $n_1 = n_2$, $m = 3$, $s_{1-3} = [1/\sqrt{2}, 1/\sqrt{8}, 1/\sqrt{8}]$, $l_{[1]1-3} = [\sqrt{4}, \sqrt{3}, 0]$, $l_{[2]1-3} = [\sqrt{4}, 0, \sqrt{3}]$ (**top**) and MINST data, $p = 784$, $n_1 = n_2 = 500$ (**bottom**). Averaged over 20 random iterations.

3.3. Insights and Remarks

Other than allowing to obtain the asymptotic SVM performance, many findings in Subsection 2.2 are of independent interest. In the following, we discuss their conceptual significance and practical implications.

3.3.1. Soft-margin Parameter

A key question in understanding and fine-tuning SVMs lies in the impact of the soft-margin parameter τ. Our derivation provides insights on how τ affects the classifier. Let us discuss this question in terms of bias and variance, as commonly done when analyzing the quality of classifiers. To facilitate the discussion, we consider that $C_1 = C_2 = C$.

For two-class Gaussian data $\mathcal{N}(\pm\mu, C)$, the Bayes optimal hyperplane is $\mu^\mathsf{T} C^{-1} x = 0$, so $\beta = C^{-1}\mu$ is optimal. Understandably, the bias is measured by the angle between $C^{-1}\mu$ and $\mathbb{E}[\beta]$, the variance by $\mathbb{E}[\|\beta - \mathbb{E}\{\beta\}\|^2]/\mathbb{E}[\|\beta\|^2]$, both of which can be computed by using the results of Section 2.2.

[5] −10dB background Gaussian noises are added to the data.

Proposition 1. *Under the conditions of Theorem 1 and with the same notations, let $C_1 = C_2 \equiv C$, define $B_p = \mu^\mathsf{T} C^{-1}\mathbb{E}\{\beta\}/\|C^{-1}\mu\|\|\mathbb{E}\{\beta\}\|$ and $V_p = \mathbb{E}\left\{\|\beta - \mathbb{E}\{\beta\}\|^2\right\}/\mathbb{E}\{\|\beta\|^2\}$. Then,*

$$B_p \to B = 1 - \frac{\sum_{d=1}^{m+1} \frac{s_d^2}{\left(1+l_{[1]d}^2\right)\left(1+l_{[1]d}^2 R_s\right)}}{\sqrt{\sum_{d=1}^{m+1} \frac{s_d^2}{\left(1+l_{[1]d}^2\right)^2}} \sqrt{\sum_{d=1}^{m+1} \frac{s_d^2}{\left(1+l_{[1]d}^2 R_s\right)^2}}} \tag{53}$$

$$V_p \to V = \frac{1}{1 + \frac{M_s^2}{L_s} \sum_{d=1}^{m+1} \frac{s_d^2}{\left(1+l_{[1]d}^2 R_s\right)^2}} \tag{54}$$

where R_s, M_s and L_s are determined by Theorem 1.

Figure 3 depicts the bias and variance for the Gaussian mixture model of Figure 2. For the case $s_1 = 1$ (i.e., $v_1 = \mu/\|\mu\|$) the bias is null for all τ, while it decreases as τ increases when $s_1 \notin \{0, 1\}$; in both cases, the variance is minimum as $\tau \to 0$. In fact, the variance always tends to its lower bound as $\tau \to 0$, regardless of data parameters. Indeed, the variance is minimized if β is simply the sum of all data vectors x_i and, as $\tau \to 0$, the margin widens, so more data vectors fall inside it with dual coefficients equal to τ/n, namely, they tend to the same value when $\tau \to 0$. Also, it is easily deduced from (53) that if spike vectors $v_1, \ldots, v_m$ are either aligned with or orthogonal to μ, the bias is always zero, otherwise it decreases with R_s. While it is not directly clear how R_s varies with τ, intuitively, β should be more aligned to $C^{-1}\mu$ when τ increases, as it tends to minimize the hinge loss of training data, minimized when $\beta = C^{-1}\mu$. In summary, when spike vectors $v_1, \ldots, v_m$ are either aligned with or orthogonal to μ (as in the data model used in (Huang, 2017)), the performance of SVMs increases as $\tau \to 0$. Otherwise, τ performs a bias-variance trade-off.

3.3.2. Class proportions

Besides bias and variance of β, the SVM performance also depends on the offset β_0. Consider again the two-class mixture $\mathcal{N}(\pm\mu, C)$ for which it seems best to separate data with a hyperplane going through the origin, i.e., $\beta_0 = 0$. Equation (20) states that this is only the case when $n_1 = n_2$; otherwise the hyperplane is pulled away from the class with more training data.

Proposition 2. *Under the conditions of Theorem 1 and with the same notations, let $C_1 = C_2 \equiv C$. Then, for α fixed, the maximum average precision is achieved when*

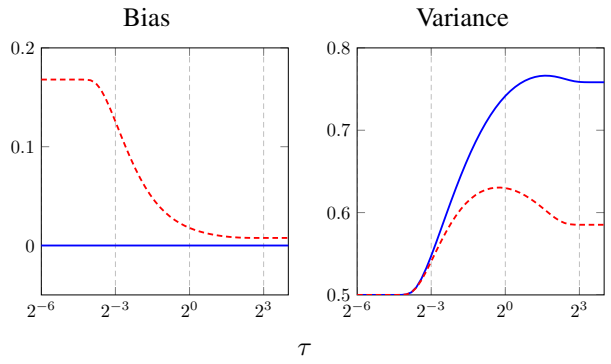

Figure 3. Bias and variance as functions of τ for two-class Gaussian data with $n = p = 256$, $n_1 = n_2$, $m = 1$, $s_1 = 1$, $l_{[1]1} = l_{[2]1} = \sqrt{2}$ (**solid blue**) or $n = p = 256$, $n_1 = n_2$, $m = 1$, $s_1 = 1/\sqrt{2}$, $l_{[1]1} = l_{[2]1} = \sqrt{4}$ (**dashed red**).

$\alpha_1 = \alpha_2$. *Besides, the class with more training data has higher precision.*

As a consequence of Proposition 2, the performance may degrade when adding more training data, as depicted at the right side of Figure 4. There must then exist degrees of improvement of SVMs obtained when addressing this problem.

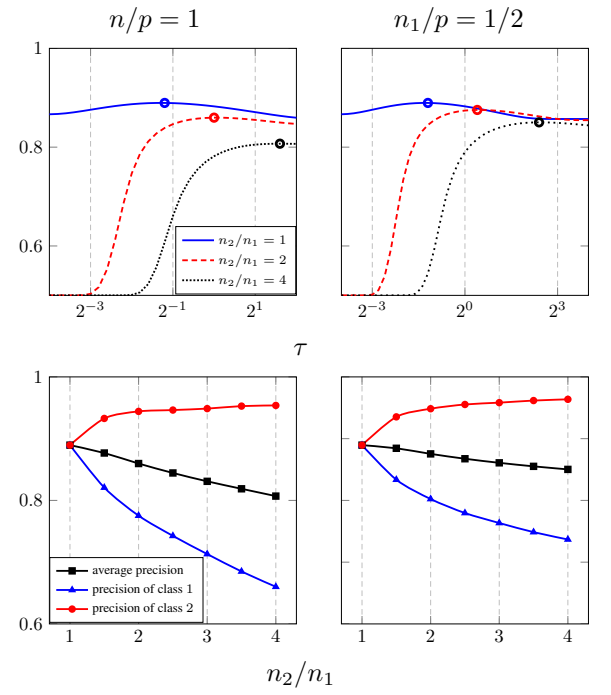

Figure 4. Precisions for two-class Gaussian data with $p = 256$, $m = 1$, $s_1 = 1/\sqrt{2}$, $l_{[1]1} = l_{[2]1} = \sqrt{4}$ at various n_2/n_1.

3.3.3. Limitations of Hard-Margin SVM

From Section 2.2.2, $c_i \to \infty$ as $n_b/p \to 1$. For soft-margin SVM, this never occurs. Indeed, if we let $n_b/p \simeq 1$, for any τ, $n_b \simeq n_1 \mathbb{E}\{\delta_{[1]}(g_i - \gamma_1)\} + n_2 \mathbb{E}\{\delta_{[2]}(g_i - \gamma_2)\} \simeq 0$ (since $\delta_{s[k]} = \mathbf{1}_{(0,(p-n_b)\sigma^2\tau/ne_{[k]} \simeq 0)}$). Hence, by contradiction, $n_b/p \simeq 1$ never occurs. This conclusion is compatible with intuitive thinking, as in the soft-margin case, a solution to the optimization problem always exists, so well defined c_i are always found.

On the opposite, there might be no viable solution for hard-margin SVM because it is possible that no hyperplane can separate perfectly the training data, especially when n is large. Consequently, the dual problem is ill-posed, as the optimization objective continually increases when $c_i \to \infty$. As the phenomenon inevitably happens when $n/p \to \infty$, given a data model, there must exist a threshold α_{th} for n/p beyond which hard-margin SVM cannot be solved with high probability. Then α_{th} corresponds to the value for which c_i grows unbounded, i.e., when $n_b/p \simeq 1$.

As a result, we have the following proposition.

Proposition 3. *Under the conditions of Theorem 1 and with the same notations, the threshold α_{th} for α beyond which there is almost surely no solution to hard-margin SVM is given by the solution of the following equation in α*

$$\alpha \left[\frac{\alpha_1}{\alpha} Q(\gamma_1) + \left(1 - \frac{\alpha_1}{\alpha}\right) Q(\gamma_2) \right] = 1 \qquad (55)$$

where γ_1, γ_2 are given by Theorem 1.

4. Conclusion

By means of a leave-one-out approach operated in the dual optimization problem, the article has provided a systematic method for the analysis of optimization schemes involving simultaneously large and numerous data, here in the case of support vector machines. Beside the asymptotic performance formulation, admittedly leaving room to little interpretation in practical settings, the main practical outcome of the study lies in the possibility to easily estimate the asymptotically optimal hyperparameters (the optimal margin), as confirmed by simulations both for synthetic and practical data.

This analysis opens the path to the understanding, hyperparameter estimation, and overall improvement of similarly elaborate machine learning methods based on implicit solutions to optimization problems, such as ℓ_1-norm optimization approaches in compressive sensing, non-linear regression (such as logistic regression-based classification), estimation in generalized linear mixed models, etc.

References

Bai, Zhidong and Silverstein, Jack W. Spectral analysis of large dimensional random matrices, volume 20. Springer, 2010.

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

Bishop, C. M. Pattern recognition and machine learning. springer, 2006.

Cheng, Xiuyuan and Singer, Amit. The spectrum of random inner-product kernel matrices. Random Matrices: Theory and Applications, 2(04):1350010, 2013.

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

Donoho, D. and Montanari, A. High dimensional robust m-estimation: Asymptotic variance via approximate message passing. Probability Theory and Related Fields, pp. 1–35, 2013.

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

El Karoui, N. Asymptotic behavior of unregularized and ridge-regularized high-dimensional robust regression estimators: rigorous results. arXiv preprint arXiv:1311.2445, 2013.

El Karoui, Noureddine, Bean, Derek, Bickel, Peter J, Lim, Chinghway, and Yu, Bin. On robust regression with high-dimensional predictors. Proceedings of the National Academy of Sciences, 110(36):14557–14562, 2013.

Elkhalil, Khalil, Kammoun, Abla, Couillet, Romain, Al-Naffouri, Tareq Y, and Alouini, Mohamed-Slim. A large dimensional analysis of regularized discriminant analysis classifiers. arXiv preprint arXiv:1711.00382, 2017.

Hastie, Trevor, Buja, Andreas, and Tibshirani, Robert. Penalized discriminant analysis. The Annals of Statistics, pp. 73–102, 1995.

Hoyle, David and Rattray, Magnus. Limiting form of the sample covariance eigenspectrum in pca and kernel pca. In Advances in Neural Information Processing Systems, pp. 1181–1188, 2004.

Huang, Hanwen. Asymptotic behavior of support vector machine for spiked population model. Journal of Machine Learning Research, 18(45):1–21, 2017.

Johnstone, Iain M. On the distribution of the largest eigenvalue in principal components analysis. Annals of statistics, pp. 295–327, 2001.

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

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

Paul, Debashis. Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. Statistica Sinica, pp. 1617–1642, 2007.

Telatar, Emre. Capacity of multi-antenna gaussian channels. Transactions on Emerging Telecommunications Technologies, 10(6):585–595, 1999.