Finite-Dimensional Projection for Classification and Statistical Learning
Gilles Blanchard, Laurent Zwald

To cite this version:

HAL Id: hal-00661865
https://hal.archives-ouvertes.fr/hal-00661865
Submitted on 20 Jan 2012

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.
Finite Dimensional Projection for Classification and Statistical Learning

Gilles Blanchard and Laurent Zwald

Abstract

A new method for the binary classification problem is studied. It relies on empirical minimization of the hinge risk over an increasing sequence of finite-dimensional spaces. A suitable dimension is picked by minimizing the regularized risk, where the regularization term is proportional to the dimension. An oracle-type inequality is established for the excess generalization risk (i.e. regret to Bayes) of the procedure, which ensures adequate convergence properties of the method.

We suggest to select the considered sequence of subspaces by applying kernel principal components analysis. In this case the asymptotical convergence rate of the method can be better than what is known for the Support Vector Machine. Exemplary experiments are presented on benchmark datasets where the practical results of the method are comparable to the SVM.

Index Terms

statistical learning, classification, support vector machine (SVM), kernel principal component analysis (KPCA), regularization, dimension reduction

Support from the IAP research network nr. P6/03 of the Federal Science Policy, Belgium, is acknowledged. This work was supported in part by the IST Programme of the European Community, under the PASCAL Network of Excellence, IST-2002-506778.
I. INTRODUCTION

A. The classification framework.

In this paper, we consider the framework of supervised binary classification. Let \((X, Y)\) denote a random variable with values in \(X \times \{-1, +1\}\) and probability distribution \(P\). The marginal distribution of \(X\) is denoted by \(Q\). \(Y\) is the label associated to the input variable \(X\). We observe a set of \(n\) independent and identically distributed (i.i.d.) pairs \((X_i, Y_i)\) sampled according to \(P\). These observations form the training set. (We will suppose \(n \geq 3\) to avoid inconsistencies in the sequel.)

A classifier is a mapping \(f\) from \(X\) to \(\{-1, +1\}\) assigning to every point \(x \in X\) a prediction of its label. The quality of such a classifier is naturally measured by its generalization error \(P[f(X) \neq Y]\). Consequently, the aim is to estimate (using only the information of the training set) the classifier having minimal generalization error, called Bayes classifier. We will denote this optimal classifier \(f^*\); it is well-known that \(f^*(x) = 21_{\{\eta(x) > 1/2\}} - 1\) a.s. on the set \(\{\eta(x) \neq 1/2\}\), where \(\eta(x) = P[Y = 1|X = x]\).

Given a certain set of classifiers \(C\) fixed in advance, the Empirical Risk Minimization procedure (see, e.g., [1]) consists in finding a classifier \(f \in C\) minimizing the empirical classification error \(\frac{1}{n} \sum_{i=1}^{n} 1_{(f(X_i) \neq Y_i)}\). Here, the function \(\ell(f, (x, y)) = 1_{(f(x) \neq y)}\), also called 0-1 loss, is the natural loss function for the classification framework. Unfortunately, minimizing exactly the empirical classification error is, in most cases, practically intractable, mainly because it is not a convex optimization procedure.

This is the reason why numerous classification algorithms, such as the support vector machine (SVM for short in the sequel), or boosting, do not consider the 0-1 loss, but minimize instead a convex surrogate loss function \(\gamma\) over some real-valued (instead of \(\{-1; 1\}\)-valued) function space \(F\), opening way to the use of efficient convex programming methods. A real output function \(f\) over \(X\) can then be transformed into a binary classifier by considering \(\text{sign}(f)\). We will in this work concentrate on the surrogate loss function used by the SVM, called the hinge loss:

\[\gamma_h(y, (x, y)) = (1 - yy (x))_+\]

where \((a)_+ = a 1_{\{a \geq 0\}}\) denotes the positive part.

B. Regularization in classification and in regression

For a given loss function we call risk the averaged loss over some data distribution. The averaged loss over the training set is called empirical risk and the averaged loss over the draw of a new independent random observation \((X, Y)\) is called generalization risk or sometimes just risk. Overfitting is the phenomenon of excessive discrepancy between empirical risk (observed on the training sample) and generalization risk on independent test data, and leads to performance degradation and inconsistency. To avoid overfitting, a common remedy is to consider regularization: it consists in adding to the empirical risk an additional balancing term \(\Omega(f)\), which, roughly speaking, should be representative of how irregular the considered function is. The function \(f \in F\) minimizing the sum of these two terms is then picked. For SVMs, the regularization \(\Omega(f) = C \|f\|^2\) is taken proportional to the squared norm of \(f\) in a certain functional Hilbert space \(H\).

The latter form of regularizer is also known as Arsenin-Tikhonov’s regularization (Tikhonov’s regularization for short in the sequel), and has already been widely used and studied in statistics in the framework of least squares regression. In this case, and if the proportionality constant \(C\) is chosen in a suitable way as a function of the training sample size, it can be shown that the resulting estimator enjoys minimaxity properties over the Hilbert balls \(B(R) = \{f \in H, \|f\| \leq R\}\) (see, e.g., [2] for a survey). Note that \(B(R)\) can equally be seen as an (Hilbert-Schmidt) ellipsoid of \(L_2(Q)\).

However, in the case of least square regression, an also widely used alternative strategy to Tikhonov’s regularization is to consider least squares fitting over an increasing sequence of linear, finite dimensional subspaces \(S_1 \subset S_2 \subset \ldots\), and to use a regularization term proportional to the subspace dimension (the “number of parameters”). The selected dimension then minimizes the sum of the residual least squares and of the regularization term (see, e.g., [3] for a extended study and [4] for regression on a random design).

The interesting thing here is that these two regularization approaches for regression can be compared. More precisely, the following has been shown in the particular case of Gaussian white noise regression (see [5], section 4.3): if the subspace \(S_D\) is taken to be the span of the \(D\) first principal axes of \(B(R)\) (as an ellipsoid in \(L_2(Q)\)), then the finite-projection estimator is also minimax over \(B(R)\). More than that, it is even minimax for any other Hilbert-Schmidt ellipsoid of \(L_2(Q)\) having the same principal axes. This, on the other hand, is not generally the case for the estimator obtained by Tikhonov’s regularization. In conclusion, the finite-dimensional approach can actually be more adaptive than Tikhonov’s regularization.
II. FINITE DIMENSIONAL PROJECTION FOR CLASSIFICATION.

A. The finite dimensional projection estimator.

Remember our main loss function is the hinge loss \( \gamma_h(f,(x,y)) = (1-yf(x))_+ \). The hinge loss is consistent in the sense that the Bayes classifier satisfies (see [9])

\[
f^* = \arg\min_g \mathbb{E} [\gamma_h(g)].
\]

It is straightforward that the hinge loss upper-bounds the classification error (0-1 loss):

\[
\gamma_h(g,(x,y)) \geq 1_{\{g(x) \neq y\}}.
\]

Moreover, the excess or regret hinge risk with respect to the Bayes classifier upper-bounds the excess classification error [10]:

\[
\mathbb{E} [\gamma_h(g)] - \mathbb{E} [\gamma_h(f^*)] \geq P(Yg(X) \leq 0) - P(Yf^*(X) \leq 0).
\]

This means that the rate of convergence to the Bayes classifier for the hinge risk is an upper bound on the classification error convergence rate.

We will now consider the following setting. Let \((\Psi_1, \Psi_2, \ldots)\) be an arbitrary family of functions fixed beforehand. We denote \(S_D = \text{span} \{\Psi_1, \ldots, \Psi_D\}\) the functional subspace spanned by the first \(D\) functions in the family. The following classifier is associated to each dimension \(D \in \mathbb{N}^*\) by minimization of the empirical hinge risk over \(S_D\):

\[
\hat{f}_D = \arg\min_{f \in S_D} \frac{1}{n} \sum_{i=1}^n (1 - Y_i f(X_i))_+.
\]

At an intuitive level, we can think about the function family as a smoothness basis (think, e.g., a Fourier basis): subspaces \(S_D\) with lower dimension \(D\) contain smoother functions. Note that there is no regularization term of any kind in the definition of \(\hat{f}_D\). Instead, it the dimension parameter \(D\) will play the role of a complexity penalty. We now devise a method for the selection of the dimension.

Some technical problems arising when analyzing the statistical properties of \(\hat{f}_D\) are caused by the unboundedness of the loss function \(\gamma_h\). In order to alleviate these, we introduce here the clip function (which was already considered in [11] in a regression framework, in [12], and in [13] in relation to the SVM):

\[
\text{clip}(g(x)) = \begin{cases} 
1 & \text{if } g(x) \geq 1 \\
g(x) & \text{if } -1 < g(x) < 1 \\
-1 & \text{if } g(x) \leq -1.
\end{cases}
\]

Note that the hinge loss of a clipped function corresponds to ‘trimming’ the hinge loss of the original function:

\[
\gamma_h(\text{clip}(f),(x,y)) = \min(\gamma_h(f,(x,y)), 2).
\]

We now apply the following dimension selection strategy: we first “clip” the estimated function \(\hat{f}_D\) for each \(D\), defining \(\bar{f}_D = \text{clip}(\hat{f}_D)\). We define our final estimator by performing the model selection step over values of \(D\) using the clipped estimators \(\bar{f}_D\). This is obtained by penalized minimization of the empirical risk: the final classifier is \(\text{sign} (\bar{f}_D)\) where

\[
\hat{D} = \arg\min_{D \geq 1} \left( \frac{1}{n} \sum_{i=1}^n (1 - Y_i \bar{f}_D(X_i))_+ + \lambda D \right),
\]

and the constant \(\lambda\) has to be suitably chosen. The linear penalty \(\lambda D\) will be justified from a statistical point of view in the next section.

Some preliminary comments are in order. First, note that the clipping operation does not alter the associated classifier function since clipping leaves the sign unchanged. As a second point, note that the clipping is only performed for the dimension
then the excess risk satisfies

\[ D \]

taken into consideration satisfies \( L \). Let \( C \) where \( L \)

Then, there exist universal constants \( \lambda \) and \( \gamma \) such that, for any \( K > 1 \), the following holds: provided that

\[ \forall D \geq 1, \quad \text{pen} \_n(D) \geq \frac{C_1 K D \log n}{h_0} n, \]  

then the excess risk satisfies

\[ \mathbb{E} \left[ L_h(\hat{f}_D, \gamma) \right] \leq \frac{K}{K - 1} \left( \inf_{D \geq 1} \left( \inf_{f \in S_D} L_h(f, \gamma) + 2\mathbb{E}[\text{pen} \_n(D)] \right) \right) + \frac{C_2 K}{h_0 n}, \]  

where \( L_h(g, \gamma) = \mathbb{E} [\gamma_h(g)] - \mathbb{E} [\gamma_h(\gamma)] \) is the excess hinge risk.

This result is proved in appendix in a more generic framework. We now give some comments.

- The above theorem takes the form of a so-called oracle inequality (8), where the excess risk of the penalized estimator can be compared to the excess risk of the best possible choice of function over each model. There is a tradeoff appearing between approximation error, decreasing with \( D \), and estimation error (represented by the penalization term on the right-hand side) which increases with \( D \). If we make additional assumptions about the approximation properties of \( S_D \) with respect to the target \( f^\star \), this can be used to derive rates of convergence (this will be elaborated in the next section). A crucial point, however, is that the oracle inequality itself, and the penalty, are independent of any such assumptions on \( f^\star \). It means that the resulting estimator enjoys adaptivity properties.

- This result means that the penalty \( \lambda D \) is linear with the dimension used in the criterion (3) is statistically justified. However, admittedly, the multiplicative constant \( C_1 \) obtained by this theoretical study is too large to be directly used in practice. This theorem therefore should be seen essentially as a theoretical guarantee that using this form of penalty is well-founded from a statistical point of view, and will have suitable convergence properties as the sample size \( n \) grows large.

- Using inequality (1), inequality (8) also provides an upper bound of the excess classification error (i.e. excess risk for the 0-1 loss) with respect to the Bayes classifier.

- One important drawback of this result is the dependence of the penalty on the unknown margin parameter \( h_0 \). Results of [14] show that this parameter plays a crucial role for minimax rates of convergence in classification. A related and more general noise condition is Tsybakov’s noise exponent (see [15], [16]) which similarly has a prominent influence on minimax rates. Procedures that can be shown to be adaptive to the noise margin parameter or to Tsybakov’s noise exponent have only been studied very recently [17]–[20].
In this section, we try to compare the rate of convergence that can be obtained for the finite-dimensional approach via Theorem 1 to the Tikhonov regularization approach. The natural candidate to compare against is therefore the Support Vector Machine (SVM), which uses this form of regularization with the same loss function as above.

If we draw a parallel to what can be shown in the least squares regression case [5], we must however consider several caveats. This will result in the picture of the present situation being unfortunately noticeably less complete.

- There is, up to our knowledge, no definite universal reference bound agreed upon that would accurately depict the behavior of the SVM, but there is a diversity of performance bounds to choose from in the recent literature (e.g., [13], [21], [22]). Here, we have chosen to compare the bound of Theorem 1 to the performance bound shown in [23]. The main reason for this choice is that the bounds in [23] have a very similar ‘oracle-type’ form involving approximation properties of models (in the case of SVMs, these models are ellipsoids) for the excess hinge risk. This makes a comparison easier.
- For least squares regression, minimax rates have been extensively studied, and provide a definite yardstick for establishing when a convergence bound is optimal and cannot be improved. For classification with the hinge loss, up to our knowledge, no minimax bounds have been established (and we do not know of lower bound results on the rate of convergence of the SVM). Here, the comparison will be therefore limited to an upper bound comparison.
- In regression again, the least squares loss is associated to function approximation in $L_2$ distance, which is particularly well suited to compare approximation properties of ellipsoids and finite dimensional subspaces. For the hinge loss, comparing such approximation properties is far from obvious. We will therefore resort to bounding the excess hinge risk by the $L_2$ distance and compare the obtained bounds.

A. Background material on the SVM.

In this subsection, we recall briefly some points crucial to the Support vector machine. Although it was not originally derived this way, the (soft-margin) SVM algorithm can be formulated as the minimization of the regularized empirical hinge risk [24], [25]:

$$
\hat{g} = \arg\min_{g \in \mathcal{H}_b} \frac{1}{n} \sum_{i=1}^{n} (1 - Y_i g(X_i))_+ + \lambda_n \|g\|_{\mathcal{H}}^2.
$$

(9)

Here $\mathcal{H}_b = \{g(x) + b, \ g \in \mathcal{H}, b \in \mathbb{R}\}$ and $\mathcal{H}$ is a reproducing kernel Hilbert space (RKHS) of functions on $X$, with Hilbert norm $\|\cdot\|_{\mathcal{H}}$. We will actually consider a restricted case of the above where the minimization is over $\mathcal{H}$ instead of $\mathcal{H}_b$, i.e., the arbitrary constant $b$ is set to zero. We will denote the resulting function $\hat{g}_0$. This simplified setting is frequently (although not always) used in theoretical studies of the SVM, among which [23], which exposes the bound we wish to compare ourselves to. It is expected that this modification does not change fundamentally the asymptotical properties of the SVM.

Note that the optimization problem (9) (with the above simplification to $b = 0$) can be rewritten in the following way: $\hat{g}_0 = \hat{g}_{\hat{R}}$ where

$$
\hat{g}_R = \arg\min_{g \in \mathcal{E}(R)} \frac{1}{n} \sum_{i=1}^{n} (1 - Y_i g(X_i))_+,
$$

and

$$
\hat{R} = \arg\min_{R \geq 0} \left( \frac{1}{n} \sum_{i=1}^{n} (1 - Y_i \hat{g}_R(X_i))_+ + C R^2 \right),
$$

(10)

where $\mathcal{E}(R)$ are balls of radius $R$ in $\mathcal{H}$. Thus, we can equivalently interpret the regularization (9) as model selection, where the models are balls $\mathcal{E}(R)$, $\hat{g}_R$ is the minimum empirical risk estimator on each model, and criterion (10) is used to select the radius $\hat{R}$.

At this point, it is interesting to recall that the balls in the RKHS space $\mathcal{H}$ can be viewed as ellipsoids in $L_2(Q)$. Let us recall some facts about the structure of a RKHS $\mathcal{H}$: such a space is uniquely characterized by a symmetric, positive definite kernel function $k : X \times X \rightarrow \mathbb{R}$. We will assume that $X$ is a metric compact space and $k$ a continuous kernel, which allows us to apply Mercer’s theorem. Let $T_k$ be the following integral kernel operator of $L_2(Q)$:

$$
T_k : g \rightarrow \int_X k(x, \cdot) g(x) dQ(x).
$$

(11)

The operator $T_k$ is compact and self-adjoint, and can therefore be diagonalized: let $(\Psi_i)_{i \geq 1}$ be an $L_2(Q)$-orthonormal basis of eigenfunctions corresponding to the non-increasing sequence of eigenvalues $(\lambda_i)_{i \geq 1}$. Mercer’s theorem allows to get a representation of $\mathcal{H}$ in terms of spectral quantities associated to $T_k$. Precisely, $\mathcal{H}$ can be characterized as

$$
\mathcal{H} = \left\{ g \in L_2(Q) : g = \sum_{i \geq 1} a_i \Psi_i \text{ such that } \|g\|_{\mathcal{H}}^2 = \sum_{i \geq 1} \frac{a_i^2}{\lambda_i} < \infty \right\}.
$$

(12)
We then have
\[ E(R) = \{ g \in \mathcal{H}, \|g\|_\mathcal{H} \leq R \} = \left\{ g \in L_2(Q), g = \sum_{i \geq 1} a_i \Psi_i ; \sum_{i \geq 1} a_i^2 \leq R^2 \right\}. \]
Consequently, a ball of the RKHS is an ellipsoid of \( L_2(Q) \) whose principal axes are precisely the eigenfunctions of \( T_k \).

In [23], the following result was proved for the SVM: assuming \( k(x, x) \leq M \) for all \( x \in \mathcal{X} \), and under the same margin condition (5) as in Theorem 1, the function \( \hat{g}_0 \) defined in equation (9) satisfies with probability at least \( 1 - \delta \), over the i.i.d. draw of the training sample:
\[ L_h(\hat{g}_0, f^*) \leq 2 \left( \inf_{g \in \mathcal{H}} L_h(g, f^*) + C(M) \Lambda_n \|g\|_\mathcal{H}^2 \right) + C' \Lambda_n, \] (13)
provided that \( \Lambda_n \) is bounded from below by a certain function depending on the eigenvalue sequence \( (\lambda_i) \). If the eigenvalues are of the form \( \lambda_j = O(j^{-2\gamma}) \) for some \( \gamma \geq 1 \), then the corresponding condition reads \( \Lambda_n \geq C(M, h_0, \delta) n^{-\frac{2\gamma}{2\gamma + 1}} \). Finally, it is reported in [23] that a similar bound holds for the expected excess risk \( E L_h(\hat{g}_0, f^*) \) (averaged over the draw of the training sample), up to additional logarithmic factors in the penalty. We will use this averaged risk version for comparison with Theorem 1.

B. Bound comparison.

To set up the comparison with the finite-dimensional subspace method, we will consider the subspaces \( S_D \) spanned using the sequence of functions \( (\Psi_1, \Psi_2, \ldots) \) defined in the previous section as the eigenfunctions of operator \( T_k \). Note that in the present section, we assume that the marginal \( Q \) is known, and therefore that both \( (\Psi_i) \) and \( (\lambda_i) \) are known to be able to compare the performance bounds. (In the next section, we will discuss a practical procedure in order to approximate the sequence \( (\Psi_i) \) when \( Q \) is unknown.)

We will compare the obtained bounds in the following specific setting:
(a) the eigenvalues satisfy a polynomial decay \( \lambda_i = \Theta(i^{-2\gamma}) \). (Note that \( \gamma > \frac{1}{2} \) since the eigenvalues series must be summable, \( T_k \) being a trace-class operator.)
(b) the coefficients of the Bayes classifier \( f^* \) in the \( L_2(Q) \) orthogonal basis \( (\Psi_i) \) satisfy \( \langle f^*, \Psi_i \rangle = \Theta(i^{-\alpha}) \). (Note that \( \alpha > \frac{1}{2} \) since \( f^* \in L_2(Q) \).

As previously discussed in the caveats, in order to compare the approximation properties of the balls of \( \mathcal{H} \) and the subspaces \( S_D \), we will additionally upper-bound (for both compared bounds) the excess hinge risk in the following way:
\[ L_h(f, f^*) \leq \| f - f^* \|_{Q,1} \leq \| f - f^* \|_{Q,2}, \] (14)
where the first inequality holds because the hinge loss \( \gamma_h \) is Lipschitz.

We are now left with comparing the following bounds: for the SVM classifier \( \hat{g}_0 \), we have from (13), (14) and condition (a):
\[ E[\hat{L}(\hat{g}_0, f^*)] \leq \inf_{g \in \mathcal{H}} \left( \| g - f^* \|_{Q,2} + n^{-\frac{2\gamma}{2\gamma + 1}} \| g \|_{\mathcal{H}}^2 \right), \] (15)
while Theorem 1 together with (14) yields for the finite-dimensional estimator \( \bar{f} \):
\[ E[\bar{L}(\bar{f}, f^*)] \leq \inf_{D \geq 1} \left( \inf_{g \in S_D} \| g - f^* \|_{Q,2} + \frac{D}{n} \right), \] (16)
where \( \leftarrow \) means that the bound holds up to a fixed multiplicative constant and possibly an additional factor \( \log(n) \).

We sum up some necessary computations in the following lemma:

\textbf{Lemma 2:} Assume condition (b) above holds. Then the right-hand side of inequality (16) is bounded the following way:
\[ \inf_{D \geq 1} \left( \inf_{g \in S_D} \| g - f^* \|_{Q,2} + \frac{D}{n} \right) \leq n^{-\frac{2\gamma}{2\gamma + 1}}. \]

On the other hand, if conditions (a) and (b) above hold, and \( \alpha < 2\gamma + \frac{1}{2} \), the right-hand side of inequality (15) is at least
\[ \inf_{g \in \mathcal{H}} \left( \| g - f^* \|_{Q,2} + n^{-\frac{2\gamma}{2\gamma + 1}} \| g \|_{\mathcal{H}}^2 \right) \geq n^{-\frac{2(2\alpha-1)}{(2\alpha+1)(4\gamma-2\alpha+1)}}. \]
In can be checked easily that \( \frac{2\alpha-1}{2\alpha+1} > \frac{2(2\alpha-1)\gamma}{(2\alpha+1)(4\gamma-2\alpha+1)} \) is implied by the sufficient condition \( \alpha \leq \gamma + \frac{3}{2} \). Of course, since we are only comparing upper bounds, the above result does not imply that the finite dimensional projection classifier is necessarily better than the SVM: more to the point, whenever the above condition is satisfied, the known bound on the rate of convergence of the finite dimensional projection classifier outperforms the known bound on the rate of the SVM.

It is now legitimate to ask in what situation the condition \( \alpha \leq \gamma + \frac{3}{2} \) is satisfied. We argue that, if the eigenvalues and the expansion coefficients follow a polynomial decrease as assumed above, then the condition actually covers most of the possible
range of values for $\alpha$ and $\gamma$. A simple observation is namely that the Bayes classifier $f^*$ cannot belong to the RKHS $\mathcal{H}$, since it is discontinuous (except for the trivial case where it is constant), while we assumed that the kernel was continuous, implying that all functions in $\mathcal{H}$ are also continuous. This therefore means that the series $\sum_{i>0} f_i^2/\lambda_i$ must diverge, in turn necessarily that $\alpha < \gamma + \frac{1}{2}$. Therefore, the condition $\alpha \leq \gamma + \frac{1}{2}$ covers a very large part of the available range. It also trivially implies the condition needed for the second part of the lemma.

Let us finally briefly describe a very simple example illustrating the above situation. Assume that $\mathcal{X}$ is the real interval $[0, 1]$, that the marginal $Q$ is the Lebesgue measure; and consider the following kernel:

$$k(x, y) = 1 + 2 \sum_{j \geq 1} \lambda_j \left( \cos(2\pi j x) \cos(2\pi j y) + \sin(2\pi j x) \sin(2\pi j y) \right),$$

where, for $j \geq 2$, $\lambda_j = (2\pi j)^{-2\gamma}$ with $\gamma > \frac{1}{2}$. In this case, $\{\Psi_i\}_{i \geq 1}$ is the trigonometric basis of $L_2([0, 1])$. It is known (see e.g. [2], chapter 2) that, if $\gamma$ is an integer, the RKHS associated with this kernel is the Sobolev space of order $\gamma$ with periodic boundary conditions $H^{\gamma}_{\text{per}}$. Precisely, $H^{\gamma}_{\text{per}}$ is the set of functions of $L_2([0, 1])$ with $\gamma - 1$ continuous derivatives satisfying $f(0) = f(1), \ldots, f^{(\gamma-1)}(0) = f^{(\gamma-1)}(1)$, and with $f^{(\gamma)} \in L_2([0, 1])$. It is endowed with the Sobolev norm $\|f\|_{H^{\gamma}}^2 = \int_0^1 |f^{(\gamma)}(t)|^2 dt + \left( \int_0^1 f(t) dt \right)^2$. This norm coincides with the RKHS norm implicitly defined by $k$.

In this situation, the non-continuity of $f^*$ prevents normal convergence of its Fourier series, so that necessarily $\alpha \leq 1$. On the other hand, the parameter $\gamma$ represents the regularity of the kernel and it is not customary to choose very irregular kernels, so that we would expect in a reasonable practical situation that $\gamma > 1$; in this case, the sufficient condition giving the advantage to the finite dimensional bound is satisfied.

As a final note, it is reported in [23] that the quadratic penalty $\Lambda_n \|g\|_{\mathcal{H}}^2$ in the ‘standard’ SVM (i.e., Tikhonov’s regularization) could in principle be reduced to a lighter linear (instead of quadratic) penalty of the form $\|g\|_{\mathcal{H}}^2$. With this type of regularization, different from Tikhonov’s, it is still possible to ensure good statistical properties, i.e., an oracle inequality similar to (13) holds. It is possible that, for this modified SVM with linear penalty, the corresponding oracle bound would also give rise to a faster convergence rate. Here however, our point here was to compare the finite dimensional approach to the “standard” SVM only, so that we did not consider this alternative regularization.

The above bound comparison suggests that it should be a good idea to use the basis $\{\Psi_1, \Psi_2, \ldots\}$ of eigenfunctions of $T_k$ for the finite-dimensional projection approach. However, these functions are in general not available, since the marginal $Q$ is not known. In the next section, we will propose a practical procedure to approximate these functions using Kernel PCA.

IV. KERNEl PROJECTION MACHINE AND NUMERICAL RESULTS.

In this section, we will compare the finite dimensional projection approach to the classical SVM on real data by constructing an actual classification algorithm based on the finite dimensional projection principle. This algorithm is dubbed kernel projection machine (KPM, originally introduced in [8]).

A. Using kernel PCA to approximate eigenfunctions

The theoretical study in the previous section suggests to use a basis of functions which is, in some sense, adapted to the underlying distribution $Q$ of the input data, by considering

$$S_D = \text{span} \{\Psi_1, \ldots, \Psi_D\},$$

where $\{\Psi_1, \ldots, \Psi_D\}$ are the eigenfunctions (in order of decreasing eigenvalues) of the operator $T_k$ given by (11). In practice, $Q$ is not known exactly, so that the functions $\{\Psi_i\}$ are also unknown. A standard approach, the so-called Nyström approximation [26], consists in replacing $Q$ by the empirical distribution of the $X_i$ in the definition of the kernel operator.

Formally, we will therefore consider the eigenfunctions $\hat{\Psi}_i$ of the operator

$$T_{k,n} f(x) = \frac{1}{n} \sum_{i=1}^n f(x_i) k(x, x_i)$$

(17)

to define the models $S_D$. Finding the eigenfunctions of $T_{k,n}$ is equivalent to performing the well-known *kernel principal components analysis* (KPCA) algorithm [27]. A very convenient fact used to perform this step efficiency is to note that it suffices to diagonalize the $n \times n$ kernel Gram matrix $K_{1,n} = \frac{1}{n} \langle k(X_i, X_j) \rangle_{1 \leq i, j \leq n}$ to obtain the eigenfunctions of $T_{k,n}$. Indeed, for $j \geq 1$ such that $\hat{\lambda}_j > 0$,

$$\hat{\Psi}_j(x) = \frac{1}{\sqrt{\hat{\lambda}_j n}} \sum_{i=1}^n V_j^{(i)} k(x_i, x),$$

(18)

where $\{V_1\}_{1 \leq j \leq n}$ is an orthonormal basis of eigenvectors of $K_{1,n}$ associated to the eigenvalues $(\hat{\lambda}_j)_{1 \leq j \leq n}$, sorted in decreasing order. The above normalization ensures $\|\hat{\Psi}_j\|_{\mathcal{H}} = 1$. 

### References

[2] [Name], [Title], [Year].

[23] [Name], [Title], [Year].

[26] [Name], [Title], [Year].

[27] [Name], [Title], [Year].
This choice of functions and models $S_D$ is now data-dependent; for this reason, strictly speaking, this does not enter in the framework of Theorem 1 where models are assumed to be fixed. An exact study of the whole data-dependent procedure, also taking precisely into account the variability in the data-dependent models, is however out of the scope of the present paper. Let us only mention here that it is well-known that $Ψ$ converges to $Ψ_i$ as $n→∞$ and details of this convergence have been studied (see, e.g., [28]–[30]), so that using this approximation is a well-founded heuristic.

Furthermore, there is a particular data-dependent setting where Theorem 1 strictly applies, namely when the eigenfunctions are estimated on a distinct data sample. Let us assume that we have an independent (unlabeled) sample $(X'_i)$ drawn according to the same input distribution $Q$ . This situation is not uncommon in a lot of practical applications where only a part of the available data has been labeled. The eigenfunctions $(Ψ_i)$ are estimated using this unlabeled sample. These are then used to perform the finite-dimensional projection estimation on the original sample $(X_i,Y_i)$. In this case, conditionally to $(X'_i)$ Theorem 1 applies. We therefore have the following:

**Corollary 1:** Assume that the noise margin condition (5) holds. Let $D = (X_i,Y_i)_{i=1,...,n}$ be an i.i.d. sample drawn from $P$ and $D' = (X'_i)_{j=1,...,m}$ be an independent, unlabeled i.i.d. sample.

Assume the unlabeled sample $D'$ is used to construct a family of functions $(Ψ_i)$ and put $S_D = \text{span} \{Ψ_1,...,Ψ_D\}$.

Follow the same estimation procedure as in Theorem 1 with the data-dependent subspace family $S_D$ used in place of $S_D$, i.e. definitions (4), (6), giving rise to estimator $f_D'$. Then, there exist universal constants $C_1,C_2$ such that for any $K>1$, if condition (7) on the penalty function holds, then

$$\mathbb{E}_{D,D'} \left[ L_h(f_D',f^*) \right] \leq \frac{K}{K-1} \mathbb{E}_{D'} \left[ \inf \left\{ \inf_{D\geq 1} L_h(f,f^*) + 2\mathbb{E}_{D} [\text{pen}_n(D)] \right\} \right] + \frac{C_2 K}{K_0 n}.$$  \hfill (19)

**Proof:** Conditionally to the unlabeled sample $D'$ the functions $(Ψ_i)$, and the models $S_D$ are fixed; therefore Theorem 1 applies. We then take the expectation of (8) with respect to $D'$. \hfill \blacksquare

In the next section, we will however only consider the situation where we use only one sample for simplicity and give a detailed account of the obtained algorithm.

**B. The kernel projection machine (KPM) algorithm.**

Using the approximate eigenfunctions, the first step consists in computing the empirical minimizers over $(1,Ψ_1,···,Ψ_D)$:

$$f_D = \arg \min_{f\in(1,Ψ_1,···,Ψ_D)} \sum_{i=1}^{n} (1-Y_i f(X_i))_+.$$ \hfill (20)

Note that the constant function 1 equal to one is systematically included in the models in order to take into account translation on the data: this function corresponds to the threshold $b$ in the original SVM algorithm. This optimization problem can be put under the form of a linear programming (LP) problem (see (21) below).

If we adopt the parametrization by $(\hat{γ},\hat{b})$ of $\hat{f}$, of the form $\hat{f}_D = \sum_{j=1}^{D} \frac{\hat{γ}_j}{\sqrt{n}\lambda_j} \hat{ϕ}_j + \hat{b}$, then equation (18) leads to

$$\hat{ϕ}_j(X_i) = \left( \sqrt{\frac{1}{n}\frac{1}{K_{1,n}}} V_j \right)^{(i)} = \sqrt{n}\lambda_j V_j^{(i)}(\hat{γ},\hat{b}) \text{ are given as the solutions of}

\frac{\hat{γ}_j}{\sqrt{n}\lambda_j} V_j^{(i)} + \hat{b} \right) \right).$$  \hfill (21)

To conclude, the KPM algorithm can be summarized as follows:

1) given data $X_1,\ldots,X_n \in \chi$ and a positive kernel $k$ defined on $\chi \times \chi$, compute the kernel matrix $K_{1,n}$ and its eigenvectors $V_1,\ldots,V_n$ together with its eigenvalues in decreasing order $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$.

2) for each dimension $D$ such that $\lambda_D > 0$ solve the linear optimization problem

$$\left( \hat{γ},\hat{b} \right) = \arg \min_{\gamma,\delta,\xi} \sum_{i=1}^{n} \xi_i \text{ under the constraints: } \forall i=1,\ldots,n : \xi_i \geq 0, \text{ and } Y_i \left( \sum_{j=1}^{D} \gamma_j V_j^{(i)} + \hat{b} \right) \geq 1 - \xi_i.$$ \hfill (21)

Next, put $\hat{α}_j = \sum_{j=1}^{D} \frac{\hat{γ}_j}{n\lambda_j} V_j^{(i)}$, and finally $\hat{f}_D = \sum_{i=1}^{n} \hat{α}_i k(x_i,\cdot) + \hat{b}$.
3) The last step is the model selection problem consisting in choosing the dimension $D$; for this step, we use the penalized trimmed hinge loss as studied earlier:

$$
\hat{D} = \arg\min_{D \geq 1} \left( \frac{1}{n} \sum_{i=1}^{n} (1 - Y_{i} \tilde{f}_{D}(X_{i}))_{+} + \lambda D \right),
$$

where $\tilde{f}_{D} = \text{clip}(\hat{f}_{D})$.

C. Some numerical results.

We first illustrate the Nyström approximation on a controlled example where the theoretical eigenfunctions are known. Then, in our main experiment, the performances of KPM are compared with the SVM on several benchmark classification datasets.

First, we consider the idealized case of a perfect model selection step. This amounts to choosing the regularization parameter for both methods (denoted $C$ in equation (9) for the SVM and $\lambda$ for the KPM) on the test set. This allows to compare directly the best estimators within the families considered by the SVM and the KPM, respectively. Remember that the “KPM classifier family” is formed of empirical risk minimization (ERM) estimators on linear subspaces of increasing dimension, while the “SVM classifier family” can be understood as ERM estimators on RKHS balls of increasing radii. Although selecting the model on the test set does not correspond to a realistic situation, this comparison is useful to decouple and understand separately the quality of the classifier families considered, independently of the additional error introduced by model selection.

In our final experiments, we consider in turn a comparison for a realistic practical scenario: in this case, 5-fold cross-validation is used to select the regularization parameter for each method.

1) Numerical experiments for the Nyström approximation: The KPM algorithm relies on the Nyström approximation formulated in equations (17), (18). In order to illustrate it, a Gaussian kernel is considered along with a Gaussian input distribution $Q$. In this case, Theorem 9, recalled for completeness in the appendix, gives explicitly the eigenfunctions; they are of multiplicity 1. The empirical eigenfunctions are computed by using formula (18) and a random draw of 500 points. Figure 1 is obtained with $a = \frac{1}{2}$ and $b = \frac{1}{18}$ where $b$ determines the width of the gaussian kernel $k(x,y) = e^{-b(x-y)^2}$ and $a$ the gaussian law $dP(x) = \frac{1}{\sqrt{2\pi}} e^{-2ax^2} dx$ of $X$.

The straight lines (resp. the circles) represent the theoretical (resp. empirical) eigenfunctions. These graphics highlight a consequence of results of [28]: the accuracy of Nyström approximation decreases with the eigenvalues, suggesting that the approximation of the theoretical eigenfunctions by the empirical ones is more suitable for large eigenvalues, i.e., for the first eigenfunctions.

2) Numerical results for the KPM algorithm: The KPM was implemented in Matlab using the free library GLPK for solving the linear optimization problem. Since the algorithm involves the eigendecomposition of the kernel matrix, only relatively small datasets have been considered at this point.

It has been tested on benchmark datasets taken from [31]: they consist in some data originally coming from the UCI repository, to which some standardization transforms have been applied. All datasets consist of 100 samples, each sample being split into a training sample and a test sample; in all datasets except “banana” the 100 samples actually contain the same data points and only differ in the training/test split. The results obtained by applying several state-of-art classification algorithms have been reported in [31], including the SVM with Gaussian kernel $k(x,y) = e^{-\frac{\|x-y\|^2}{2\sigma^2}}$. In this case, suitable values (chosen by cross-validation) for parameters $\sigma_G$ (kernel width) and $C_G$ (SVM regularization parameter) are also reported. These values are specific to each benchmark dataset $G$.

In all experiments presented here, we used the Gaussian kernel with parameter $\sigma_G$ fixed to the value reported in [31]. Without selecting the regularization parameter, the family of classifiers obtained by the KPM algorithm is $\{f_{D}\}_{D \geq 1}$ and the family obtained by the SVM algorithm is $\{\hat{f}_C\}_{C \in C}$. The set $C$ of possible regularization constants that we consider for the SVM are tailored to each benchmark $G$ by forming a geometric sequence of 101 points running over $C_G/100$ to $100C_G$ and containing the “optimal” $C_G$.

Comparison of the family of SVM classifiers and KPM classifiers. Remember that here we aim at comparing the two families by shunting off the model selection procedure. In table I, for each sample, the smallest test error of the KPM (w.r.t. the parameter $D$) is compared with the smallest test error of the SVM (w.r.t. the parameter $C$). Each time, the winner is given one point.

Parameter selection by cross-validation. Table II presents results of SVM (resp. KPM) where the regularization parameters $C$ (resp. $\lambda$) is chosen by 5 fold-cross-validation separately on each of the samples. The results are presented in the form $\{\text{mean of the 100 test errors} \} \pm \{\text{variance of the 100 test errors} \}$.

These two tables highlight that the performances of KPM are comparable with the SVM. Considering table I, the SVM appears to have a slight advantage over the KPM which also appears in table II when the parameters are selected by cross-validation. However, note that average differences are quite small, in particular relative to the variance. Moreover, it is worth noticing that the same fixed parameter $\sigma_G$ is used for the KPM and the SVM, whereas it was originally tailored in [31] for good performance of the SVM only.
Fig. 1. From the left to the right: approximation of the first, the second, the third eigenfunctions of the kernel integral operator for a Gaussian kernel and Gaussian input distribution.

![Graphs showing quality of approximation of the first three eigenfunctions with n=500.]

**TABLE I**

<table>
<thead>
<tr>
<th>KPM</th>
<th>SVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Banana</td>
<td>31</td>
</tr>
<tr>
<td>Breast Cancer</td>
<td>44</td>
</tr>
<tr>
<td>Diabetis</td>
<td>55</td>
</tr>
<tr>
<td>Flare Solar</td>
<td>19</td>
</tr>
<tr>
<td>German</td>
<td>43</td>
</tr>
<tr>
<td>Heart</td>
<td>27</td>
</tr>
</tbody>
</table>

**TABLE II**

<table>
<thead>
<tr>
<th>Test Errors</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
</tr>
<tr>
<td>Banana ($\sigma = 0.7071$)</td>
</tr>
<tr>
<td>Breast Cancer ($\sigma = 5$)</td>
</tr>
<tr>
<td>Diabetis ($\sigma = 3.1623$)</td>
</tr>
<tr>
<td>Flare Solar ($\sigma = 3.8730$)</td>
</tr>
<tr>
<td>German ($\sigma = 5.2440$)</td>
</tr>
<tr>
<td>Heart ($\sigma = 7.7460$)</td>
</tr>
</tbody>
</table>

Finally, from an algorithmic optimization point of view, a nice property of the KPM is that the classifier $\hat{f}_D$ can be used as a “hot start” point in the optimization search for $\hat{f}_{D+1}$ since $\hat{f}_D \in S_{D+1}$. This will generally make the procedure faster than restarting separately the optimization for each value of $D$. 
A. Highlight of the present work.

We described the finite dimensional approach in the classification framework and deduced an effective algorithm: the KPM. The model selection aspect is tackled using a penalized criterion: we gave theoretical results justifying the use of a penalty which is a linear function of the dimension. We presented a theoretical study comparing known bounds on convergence rates of the finite-dimensional projection approach as compared to the SVM. We also compared performances of the KPM against the SVM in a realistic scenario in which the KPM appeared to be almost as efficient as the SVM although some parameters shared by both methods were chosen to optimize the SVM performance.

The main point of all the presented results is to highlight that

regularization can be performed thanks to a dimensionality reduction method such as Kernel-PCA.

Consequently, the finite dimensional projection is a credible alternative to the Tikhonov’s regularization used, for example, in the SVM algorithm.

An interesting view of the KPM is is that the training labels are used to select the optimal dimension \( D \) in a dimension-reduction method – optimal means that the resulting \( D \)-dimensional representation of the data contains the right amount of information needed to classify the inputs. To sum up, the KPM can be seen as a dimensionality-reduction-based classification method that takes into account the labels for choosing the right dimension in the dimensionality reduction step.

B. Comparison with other work.

We provided a detailed comparison of the theoretical bounds obtained for the KPM and of the bounds obtained in [23] for the SVM. It is more difficult to draw a meaningful comparison with other known bounds on the SVM; for example, in [21], a Gaussian kernel with width depending on the sample size \( n \) as well as on some “geometric” assumptions on \( P(Y|X) \) is considered. Here our focus was on a fixed kernel.

In [18], estimators with a finite expansion on a fixed function basis are considered, which is related to the present setting. A \( \ell_1 \)-penalty in the coefficients is considered, and the procedure is shown to be adaptive to the “Tsybakov noise exponent” parameter (a more general version of (5)). Here, we study a \( \ell_0 \)-penalty and our results are not adaptive to the noise margin parameter. However, [18] needs additional hypotheses on the \( L_2(Q) \) structure of the function family considered as well as on their supremum norm. In contrast, our focus here for the theoretical part was to obtain results on arbitrary function subspaces \( S_D \) without additional hypotheses.

C. Discussion: inverse problems and the spectral point of view.

It is interesting to note that, in the case of least squares regression, both Tikhonov’s regularization and the finite-dimensional projection (when the projection dimension is fixed) can be seen as special cases of a large class of linear estimators that have a diagonal form when expanded on the eigenfunction basis of a certain self-adjoint operator \( A \). In the present case, \( A \) is the kernel integral operator, but in a broader point of view, \( A \) could be more general. This is precisely the setting which is traditionally the basis of the inverse problems litterature. A recent and very general account of this point of view for linear least squares estimation can be found in [32], where a much broader situation is studied (i.e. the operator \( A \) is not assumed to be compact, so that the spectrum may not be discrete, in which case the most elegant way to describe linear estimators is really directly by their action on the spectrum).

Interpreting classification problems as inverse problems is not new (seminal ideas are found in [33]), but has received renewed attention recently due in particular to the strong link of support vector machines with inverse problems, as studied notably in [22]. In the present paper, what we called finite dimensional projection is generally referred to as spectral cut-off method in the inverse problems litterature.

A major part of the difficulties arising in the study of such methods for classification problems resides in the non-least squares cost function, so that the penalized empirical risk minimizer does not give rise to a linear estimator. In the present work, we have studied convergence rates for an equivalent of spectral cut-off in this setting (more precisely, combined with adequate model selection). Clearly, it would be of great interest to develop this point of view in more generality for classification problems and possibly try to recover the extent of results available for inverse problems in regression.

D. Ending remarks and future work.

The main drawback of Theorem 1 is the lack of adaptivity to the noise margin: the penalty function involves the unknown noise margin parameter. This leads to difficulties to calibrate the penalization constant in practice. We plan to investigate other calibration techniques for \( \lambda \) than cross-validation. One interesting direction is the so-called “slope heuristic” where the behavior of the empirical error of \( \hat{f}_D \) as a function of \( D \) is used to select a suitable parameter \( \lambda \).

An interesting potential advantage of the KPM with respect to the SVM is that it can easily be extended to use different kernels simultaneously by considering finite dimensional spaces spanned by eigenfunctions of kernel operators associated to...
A. Clipped empirical risk minimization on one model.

In this section, we obtain an excess risk bound for a clipped empirical error minimizer $\tilde{s}_D$ over a fixed vector space $S_D$ of dimension at most $D$ with a generic loss function $\gamma$:

$$\tilde{s}_D = \arg \min_{t \in S_D} \frac{1}{n} \sum_{i=1}^{n} \gamma(t, (X_i, Y_i)),$$

and

$$\tilde{s}_D = \text{clip}(\tilde{s}_D).$$

The goal is to estimate the target function $s$ minimizing the risk over a “large” class $S \supset S_D$:

$$s^* = \arg \min_{s \in S} \mathbb{E}[\gamma(s, (X, Y))].$$

The excess risk with respect to $\gamma$ is:

$$L(g, s^*) = \mathbb{E}[\gamma(g, (X, Y))] - \mathbb{E}[\gamma(s^*, (X, Y))].$$

In the sequel, $s^*$ is supposed to take values in $[-1, 1]$.

All the results will be stated under the following assumption on the loss which we will refer to as “assumption (A)” in the sequel:

(A1) $\forall y \in \mathcal{Y}, \gamma(y, \cdot)$ is Lipschitz.

(A2) $\forall s \in S, \forall y \in \mathcal{Y}, \forall x \in \mathcal{X}, \gamma(y, s(x)) \geq \gamma(y, \text{clip}(s(x))).$

We will prove two results; the first one (Theorem 5) provides an excess risk bound for a fixed model $S_D$; then, a similar technique will be used to prove a model selection result (Theorem 6). The proofs of these two theorems rely on a very fundamental result coming from [7], providing a control of the empirical processes over a function class using localized Rademacher averages. We recall it now.

We need to introduce some notation first. A function $\Psi : [0, \infty) \rightarrow [0, \infty]$ is called sub-root if it is non-negative, non-decreasing and if $r \rightarrow \sqrt[\Psi(r)]{r}$ is non-increasing for $r > 0$. It can be shown that the fixed point equation $\Psi(r) = r$ has a unique positive solution (except for the trivial case $\Psi \equiv 0$). Moreover, a sub-root function is a continuous function on $[0, +\infty[$ satisfying $\Psi(r) > r$ for $r \in [0, r^*[$ and $\Psi(r) < r$ for $r \in [r^*, +\infty[$.

Let $F$ be a class of functions. The following notation for the Rademacher average of $F$ will be useful:

$$R_nF = \sup_{f \in F} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i f(X_i),$$

where $(\varepsilon_i)_{i=1,\ldots,n}$ are independent and identically distributed Rademacher variables ($\mathbb{P}[\varepsilon_1 = 1] = \mathbb{P}[\varepsilon_1 = -1] = 1/2$). The notation $\mathbb{E}_\varepsilon$ means that the expectation is considered only with respect to $\varepsilon$: the variables $X_1, \ldots, X_n$ are “fixed”.

We now recall the following result:

Theorem 3 ([7, Theorem 4.1]): Let $F$ be a class of functions with ranges in $[-1, 1]$ and assume that there is some constant $\kappa$ such that for every $f \in F$, $Pf^2 \leq \kappa Pf$. Denote $\text{star}(F) = \{\lambda f, 0 \leq \lambda \leq 1, f \in F\}$. Let $\tilde{\Psi}_n$ be a (possibly data-dependent) sub-root function and let $\tilde{\lambda}^*$ be the fixed point of $\tilde{\Psi}_n$. Fix $\xi > 0$ and assume that $\tilde{\Psi}_n$ satisfies, for any $r \geq \tilde{\lambda}^*$,

$$\tilde{\Psi}_n(r) \geq 2(10 \vee \kappa)\mathbb{E}_\varepsilon R_n\{f \in \text{star}(F); Pf^2 \leq 2r\} + (2(10 \vee \kappa) + 11)\frac{\xi}{n}. \quad (23)$$
Then, for any $K > 1$ with probability at least $1 - 3e^{-\xi},$
\[
\forall f \in \mathcal{F}, Pf \leq \frac{K}{K-1} P_{n} f + \frac{6K}{\kappa} r^* + \frac{\xi(11 + 5K)}{n}.
\]

The crux of the results to come will rely on the application of Theorem 3 to the excess loss function class on a model $S_D,$ defined as
\[
\mathcal{F}_D = \{(x, y) \rightarrow \gamma(s(x), y) - \gamma(s^*(x), y), s \in \text{clip}(S_D)\}.
\]

An important technical result is therefore to have an estimate of the fixed point $r^*$ appearing in Theorem 3, for the above class. This is the goal of the following result, whose proof is postponed to section D.

**Theorem 4:** Let $\mathcal{F}_D$ be defined as in (24); assume $|s^*(x)| \leq 1,$ and that (A1) is satisfied. Let $r^*_{D,n}$ denote the fixed point of the sub-root function $\tilde{\psi}(r) = \mathbb{E}_x \mathcal{R}_n \{ f \in \text{star}(\mathcal{F}_D), P_{n} f^2 \leq 2r \}.$ Then, the following holds:
\[
\tilde{r}^*_{D,n} \leq A_1 \frac{D + 1}{n} \left( \left( \log \frac{n}{D} \right) + 1 \right),
\]
where $A_1$ is a constant ($A_1 = 1200$ is suitable).

With these prerequisites in hand, we are now in a position to state and prove our main results. The first one concerns estimation on a fixed model $S_D.$

**Theorem 5:** Let $S_D$ be a vector space of dimension at most $D.$ Assume the following conditions are met:

(i) The target function is bounded by $1$: $|s(x)| \leq 1.$
(ii) $\gamma$ satisfies assumption (A).
(iii) $P_{n} f^2 \leq \kappa P_{n} f$ for all $f \in \mathcal{F}_D = \{ \gamma(t) - \gamma(s^*), t \in \text{clip}(S_D) \}.$

Then, for all $K > 1,$ the following inequality holds:
\[
\mathbb{E}[L(\tilde{s}_D, s^*)] \leq \frac{K}{K-1} \left( \inf_{t \in S_D} L(t, s^*) + C_3 K \left( \frac{10}{\kappa} \right)^2 \frac{D}{n} \log n \right) + C_4 \frac{K(\kappa + 10)}{n},
\]
where $C_3$ and $C_4$ are numerical constants.

**Proof of Theorem 5.** Let $s_D$ be an arbitrary element of $S_D.$ First, note that
\[
P_{n}(\gamma(\tilde{s}_D) - \gamma(s)) \leq P_{n}(\gamma(\tilde{s}_D) - \gamma(s)) \leq P_{n}(\gamma(s_D) - \gamma(s)),
\]
where the first inequality follows from assumption (A2) and the last from the definition of the empirical minimizer. Thus,
\[
L(\tilde{s}_D, s^*) = P(\gamma(\tilde{s}_D) - \gamma(s^*)) \leq (P - P_n)(\gamma(\tilde{s}_D) - \gamma(s^*)) + P_n(\gamma(s_D) - \gamma(s^*)).
\]

In order to control $(P - P_n)(\gamma(\tilde{s}_D) - \gamma(s^*)),$ we apply as announced Theorem 3 to the class of functions $\mathcal{F}_D = \{ \gamma(t) - \gamma(s^*), t \in \text{clip}(S_D) \}.$ Note that by assumptions (i) and (A1), all functions in $\mathcal{F}_D$ have range in $[-1, 1]$ which is the first requirement to apply Theorem 3. Then, assumption (iii) is the second condition required to apply Theorem 3. We now need to find the fixed point of the function appearing in (23). Let $\xi > 0$ be fixed. Theorem 4 gives us a bound on the fixed point $r^*_{D,n}$ of the sub-root function $\hat{\psi}(r) = \mathbb{E}_x \mathcal{R}_n \{ f \in \text{star}(\mathcal{F}_D), P_{n} f^2 \leq 2r \}.$ In sight of (23), we need a bound on the fixed point $\tilde{r}^*_{D,n}$ of an affine transform $a\hat{\psi}(x) + b$ of $\hat{\psi}$ (with $a = 2(10 \vee \kappa), b = (2(10 \vee \kappa) + 11) \xi^2$). Elementary arguments not reproduced here (see [34], Lemma 4.10) then show that
\[
\tilde{r}^*_{D,n} \leq 4a^2 r^*_{D,n} + 2b \leq 16(10 \vee \kappa)^2 A_1 \frac{D + 1}{n} \left( \left( \log \frac{n}{D} \right) + 1 \right) + 2\xi(2(10 \vee \kappa) + 11)
\leq C \frac{(10 \vee \kappa)^2 D \log n}{n} + C' \frac{\xi(10 \vee \kappa)}{n}.
\]
Consequently, Theorem 3 implies that $\forall K > 1,$ with probability at least $1 - 3e^{-\xi}, \forall f \in \mathcal{F}_D,$
\[
(P - P_n)f \leq \frac{1}{K - 1} P_{n} f + \frac{6K}{\kappa} \tilde{r}^*_{D,n} + \frac{\xi(11 + 5K)}{n}.
\]

Since the bound is available simultaneously for all functions in $\mathcal{F}_D,$ we can apply it to the random function $f = \gamma(\tilde{s}_D) - \gamma(s^*) \in \mathcal{F}_D.$ This yields that, with probability at least $1 - 3e^{-\xi},$
\[
(P - P_n)(\gamma(\tilde{s}_D) - \gamma(s^*)) \leq \frac{1}{K - 1} P_{n} (\gamma(\tilde{s}_D) - \gamma(s^*)) + \frac{6K}{\kappa} \tilde{r}^*_{D,n} + \frac{\xi(11 + 5K)}{n}
\leq \frac{1}{K - 1} P_{n} (\gamma(s_D) - \gamma(s^*)) + \frac{6K}{\kappa} \tilde{r}^*_{D,n} + \frac{\xi(11 + 5K)}{n},
\]
where we have used again inequality (26) for the last step. We now plug this inequality into (27) to obtain
\[
L(\tilde{s}_D, s^*) \leq \frac{K}{K-1} P_{n} (\gamma(s_D) - \gamma(s^*)) + \frac{6K}{\kappa} \tilde{r}^*_{D,n} + \frac{\xi(11 + 5K)}{n}.
\]
This concludes the proof of Theorem 5 by integrating with respect to the sample, bounding $\tilde{r}^*_{D,n}$ using (28), then taking the infimum over $s_D \in S_D.$ □
B. Model selection by penalization.

We now present a relatively general result about penalized minimization of the trimmed empirical risk over finite dimensional vector spaces. Theorem 1 will then be derived as a corollary.

Theorem 6: Let \( \{S_D\}_{D \geq 1} \) be a collection of vector spaces such that \( \text{dim}(S_D) \leq D \). Assume the following:

(i) The target function is bounded by 1: \( |s^*(x)| \leq 1 \).
(ii) \( \gamma \) satisfies assumption (A).
(iii) \( P f^2 \leq \kappa Pf \) for all \( f \in F_D = \{ \gamma(t) - \gamma(s^*), t \in \text{clip}(S_D) \} \).

Let \( K > 1 \). Choosing the dimension with the following penalized criterion

\[
D = \arg\min_{D \geq 1} \left( \frac{1}{n} \sum_{i=1}^{n} \gamma(\tilde{s}_D, (X_i, Y_i)) + \text{pen}_n(D) \right),
\]

with a possibly data dependent penalty function \( \text{pen}_n \) such that

\[
\forall D \geq 1, \text{pen}_n(D) \geq C_5 K \left( \kappa \cup 10 \right) \frac{D}{n} \log n,
\]

the following inequality holds

\[
E[L(\tilde{s}_D, s^*)] \leq \frac{K}{K-1} \left( \inf_{D \geq 1} \left( \inf_{s \in S_D} L(s, s^*) + E[\text{pen}_n(D)] \right) \right) + C_6 K \left( \kappa \cup 10 \right) \frac{n}{n},
\]

where \( C_5 \) and \( C_6 \) are numerical constants.

Proof: Let \( s_D \) be a fixed element of \( S_D \). The definition of \( \tilde{s}_D \) leads to the following chain of inequalities:

\[
P_n(\gamma(\tilde{s}_D) - \gamma(s^*)) + P_n(\gamma(\tilde{s}_D) - \gamma(s^*)) \leq (P_n(\gamma(\tilde{s}_D) - \gamma(s^*)) + P_n(\gamma(s) - \gamma(s^*)) + P_n(\gamma(s_D) - \gamma(s^*)) + \text{pen}_n(D) - \text{pen}_n(\tilde{D}).
\]

Let \( D' \geq 1 \). Let \( x_D \) be such that \( \sum_{D \geq 1} e^{-x_D} \leq 1 \). We now follow the same reasoning leading to inequality (29) in the proof of Theorem 5, and also use (28), obtaining that \( \forall K > 1 \), with probability at least \( 1 - 3e^{-\xi - x_D} \),

\[
(P - P_n)(\gamma(\tilde{s}_D') - \gamma(s^*)) \leq \frac{1}{K-1} P_n(\gamma(\tilde{s}_D') - \gamma(s^*)) + C \frac{K(10 \cup \kappa)^2 D'}{n} \log(n) + C' K(\kappa + x_D')(10 \cup \kappa).
\]

We now use a union bound to obtain the previous inequality simultaneously for all \( D' \geq 1 \) and apply it to \( D' = \tilde{D} \). With probability at least \( 1 - 3e^{-\xi} \),

\[
(P - P_n)(\gamma(\tilde{s}_D) - \gamma(s^*)) \leq \frac{1}{K-1} P_n(\gamma(\tilde{s}_D) - \gamma(s^*)) + C \frac{K(10 \cup \kappa)^2 \tilde{D}}{n} \log(n) + C' K(\kappa + x_D)(10 \cup \kappa).
\]

Plugging this inequality into (32) and using again (31) leads to:

\[
L(\tilde{s}_D, s^*) \leq C \frac{K(10 \cup \kappa)^2 \tilde{D}}{n} \log(n) + C' \frac{K(\kappa + x_D)(10 \cup \kappa)}{n} + \frac{K}{K-1} \left( P_n(\gamma(s_D) - \gamma(s^*)) + \text{pen}_n(D) - \text{pen}_n(\tilde{D}) \right).
\]

Choosing \( x_D = 2 \log(D + 1) \), condition (30) entails

\[
L(\tilde{s}_D, s^*) \leq C' \frac{K(10 \cup \kappa)}{n} + \frac{K}{K-1} \left( P_n(\gamma(s_D) - \gamma(s^*)) + \text{pen}_n(D) \right).
\]

Taking the infimum over \( D \geq 1 \) and integrating with respect to the sample:

\[
L(\tilde{s}_D, s^*) \leq \frac{K}{K-1} \mathbb{E} \left[ \inf_{D \geq 1} \left( P_n(\gamma(s_D) - \gamma(s^*)) + \text{pen}_n(D) \right) \right] + C' \frac{3K(10 \cup \kappa)}{n},
\]

and finally:

\[
L(\tilde{s}_D, s^*) \leq \frac{K}{K-1} \mathbb{E} \left[ \inf_{D \geq 1} \left( P_n(\gamma(s_D) - \gamma(s^*)) + \text{pen}_n(D) \right) \right] + C' \frac{3K(10 \cup \kappa)}{n}.
\]

This concludes the proof of Theorem 6.
C. Application to classification: proof of Theorem 1.

Theorem 1 is now a simple consequence of Theorem 6 whose conditions are met for the hinge loss with $S = L_2(Q)$ and $s^* = f^*$ (the Bayes classifier). Checking for conditions (i)-(ii) of Theorem 6 is straightforward, and the lemma below ensures that assumption (iii) is met with $\kappa = \frac{1}{h_0}$. (A related result appears in [23], but we provide a proof here for completeness.) This concludes the proof of Theorem 1.

Lemma 7: Let $f : \mathcal{X} \rightarrow [-1, 1]$. We suppose that $|\eta(x) - \frac{1}{2}| \geq h_0$ where $\eta(x) = P[Y = 1|X = x]$. Then

$$
\|\gamma_h(f) - \gamma_h(f^*)\|^2 \leq \frac{1}{p_0} L_h(f, f^*).
$$

Proof of Lemma 7. Since $f, f^*$ take their values in the interval $[-1, 1]$, we have $(1 \pm f(x))_+ = 1 \pm f(x)$ and $(1 \pm f^*(x)) = 1 \pm f^*(x)$, so that we can write explicitly

\[
L_h(f, f^*) = \int \eta(x)[(1-f(x))_+ - (1-f^*(x))_+] + (1-\eta(x))[1+f(x)_- - (1+f^*(x))_+] \, dP(x)
= \int \eta(x)[f^*(x) - f(x)] + (1-\eta(x))(f(x) - f^*(x)) \, dP(x)
= \int (2\eta(x) - 1)||f^*(x) - f(x)|| \, dP(x),
\]

(\text{where the introduction of absolute values in the last line can be checked to hold separately for the two cases (p(x) > 1/2, f^*(x) = 1) and (p(x) < 1/2, f^*(x) = -1)).

$$
\|\gamma_h(f) - \gamma_h(f^*)\|^2 = \int \eta(x)[(1-f(x))_+ - (1-f^*(x))_+]^2 + (1-\eta(x))[1+f(x)_- - (1+f^*(x))_+]^2 \, dP(x).
= \int (f(x) - f^*(x))^2 \, dP(x).
$$

Hence the ratio of the two integrands is just

$$
\frac{|f(x) - f^*(x)|}{2|\eta(x) - 1|} \leq \frac{1}{h_0},
$$

and the statement follows by integrating with respect to $x$. \hfill \Box

D. Proof of Theorem 4.

Below we will use the standard covering number notation: let $\mathcal{N}(\epsilon, \mathcal{G}, d)$ (resp. $\mathcal{M}(\epsilon, \mathcal{G}, d)$) denote the covering number (resp. packing number) of set $\mathcal{G}$ for distance $d$. The proof of Theorem 4 is inspired by the work of L. Györfi and al. [11]. In particular, it relies on Theorem 9.4 of this reference for the control of packing numbers on VC-subgraph sets of functions. We reproduce it here:

Theorem 8 (Györfi et al.): Let $\mathcal{G}$ be a class of functions $g : \mathcal{X} \longrightarrow [0, A]$ with $V(\mathcal{G}) \geq 2$. Let $p \geq 1$ and $\nu$ be a probability measure on $\mathcal{X}$. Let $\epsilon$ such that $0 < \epsilon < \frac{A}{2}$. Then the following holds :

$$
\mathcal{M}(\epsilon, \mathcal{G}, ||\cdot||_{L_p(\nu)}) \leq 3\left(\frac{2\epsilon A^p}{\nu} \log \frac{3\epsilon A^p}{\nu}\right)^{V(\mathcal{G})},
$$

where $V(\mathcal{G})$ is the VC-dimension of all subgraphs of $\mathcal{G}$, i.e., the set $\{(z, t) \in \mathcal{X} \times \mathbb{R}: t \leq g(z); g \in \mathcal{G}\}$.

To begin with, we control the $L_2(P_n)$-covering numbers of $\text{star}(\mathcal{F}_D)$. Note that since this set contains the null function and has diameter bounded by 1, any covering number for $\epsilon > 1$ is equal to 1. In what follows we therefore assume $\epsilon \leq 1$.

Following [7], since any $g \in \text{star}(\mathcal{F})$ is of the form $g = \lambda f$ with $\lambda \in [0, 1], f \in \mathcal{F}$, we can construct an $\epsilon$-cover of $\text{star}(\mathcal{F})$ by taking the direct product of an $\frac{\epsilon}{2}$-cover for $\mathcal{F}$ and an $\frac{\epsilon}{2}$-cover for the interval $[0, 1]$, which implies

$$
\mathcal{N}(\epsilon, \text{star}(\mathcal{F}), L_2(P_n)) \leq \left[\frac{2}{\epsilon}\right] + 1 \mathcal{N}\left(\epsilon, \frac{\mathcal{F}}{2}, L_2(P_n)\right).
$$

(33)

Moreover,

$$
\mathcal{N}\left(\frac{\epsilon}{2}, \mathcal{F}, L_2(P_n)\right) \leq \mathcal{M}\left(\frac{\epsilon}{2}, \mathcal{F}, L_2(P_n)\right)
\leq \mathcal{M}\left(\epsilon, \{t - s, t \in \text{clip}(S_D)\}, L_2(P_n)\right)
= \mathcal{M}\left(\frac{\epsilon}{2}, \{t + 1, t \in \text{clip}(S_D)\}, L_2(P_n)\right),
$$

where the second inequality holds because $\gamma$ is Lipschitz, and the last equality holds because covering and packing numbers are translation invariant.
We now apply Theorem 8 with $A = 2$, thus, for $0 < \varepsilon < 1$,
\[
\mathcal{M} \left( \frac{\varepsilon}{2}, \{t + 1, t \in \text{clip}(S_D)\}, L_2(P_n) \right) \leq 3 \left( \frac{32e}{\varepsilon^2} \log \left( \frac{48e}{\varepsilon^2} \right) \right)^{V(\text{clip}(S_D) + 1)} \leq \left( \frac{11}{\varepsilon} \right)^{4V(\text{clip}(S_D) + 1)} .
\]

The second inequality is obtained by using $\log(x) \leq \frac{x}{e}$.

Moreover, $V(\text{clip}(S_D) + 1) = V(\text{clip}(S_D)) \leq V(S_D) \leq D + 1$. The first equality holds because the subgraph VC-dimension is translation invariant. The last inequality is a well-known property of finite dimensional function spaces (see, e.g., Lemma 2.6.15 of [35]). Finally, the middle inequality comes from the observation that a discrete set shattered by the subgraph class of $\text{clip}(G)$ is also shattered by the subgraph class of $G$. Namely, let $(z_i, t_i)_{1 \leq i \leq K}$ be a finite set shattered by the subgraph class of $\text{clip}(G)$. We must have $-1 < t_i \leq 1$ for all $i$, since otherwise the corresponding point would belong to (resp. be outside of) the subgraph of all functions. Using this property, it is easy to check that, if a function family $(\text{clip}(g_j))_{1 \leq j \leq 2^K}$ is a witness of the shattering in $\text{clip}(G)$, then so is $(g_j)$ in $G$.

Now, gathering inequality (33) and the previous ones, we get for $0 < \varepsilon < 1$,
\[
\log N(\varepsilon, \text{star}(\mathcal{F}), L_2(P_n)) \leq \log \left( \frac{4}{\varepsilon} \right) + 4(D + 1) \log \left( \frac{20}{\varepsilon} \right) \leq 5(D + 1) \log \left( \frac{11}{\varepsilon} \right).
\]

We now use Dudley’s entropy integral (also known as chaining technique, see [36]) theorem, which states that
\[
E_C R_n(G) \leq 4\sqrt{\frac{\varepsilon}{n}} \int_{0}^{\infty} \sqrt{\log N(u, \mathcal{G}, L_2(P_n))} du.
\]

Therefore, using the above upper bound on the covering numbers and standard calculus,
\[
\hat{\psi}(r) = E_C R_n(f \in \text{star}(\mathcal{F}), P_n f^2 \leq 2r) \leq \frac{4\sqrt{\frac{\varepsilon}{n}}}{\sqrt{\varepsilon}} \int_{0}^{\sqrt{2r} \wedge 1} \sqrt{\log N(\varepsilon, \text{star}(\mathcal{F}), L_2(P_n))} du \\
\leq \frac{4\sqrt{10}}{\sqrt{\varepsilon}} \sqrt{D + 1} \int_{0}^{\sqrt{2r} \wedge 1} \sqrt{\log \left( \frac{20}{u} \right)} du \\
\leq 22 \left( r n^{-1} (D + 1) \log \left( \frac{11}{\sqrt{2r} \wedge 1} \right) \right)^{\frac{1}{2}} := \hat{\phi}(r).
\]

We easily check that $\hat{\phi}$ is a sub-root function. Hence, the fixed point $\hat{\psi}^*$ of $\hat{\psi}$ is an upper bound on the fixed point $\hat{r}_{D,n}^*$ of $\hat{\phi}$. Our goal is now to upper bound $\hat{r}^*$ by $\hat{r}_{D,n}^* = A_1 \frac{D + 1}{n} \left( \log \frac{n}{D} \right) + 1$; once this is established the proof is finished. By properties of sub-root functions, the inequality $\hat{r}^* \leq \hat{r}_{D,n}^*$ is equivalent to the following inequality being satisfied:

\[
\hat{\phi} \left( \hat{r}_{D,n}^* \right) \leq \hat{r}_{D,n}^* .
\]

If $A_1 \geq 11^2/(2e)$, then it can be checked that
\[
\log \left( \frac{11}{\sqrt{2r_{D,n}^*} \wedge 1} \right) = \log \left( \frac{11}{\sqrt{2r_{D,n}^*}} \right) \wedge \log(11) \leq \log(11) \left( \log \frac{n}{D} \right) + 1 ,
\]

and therefore
\[
\hat{\phi} \left( A_1 \frac{D + 1}{n} \left( \log \frac{n}{D} \right) + 1 \right) \leq C \frac{D + 1}{n} \left( \log \frac{n}{D} \right) + 1 ,
\]

where $C = 22 \sqrt{A_1 \log(11)}$. We obtain (34) by noting that $C \leq A_1$ if $A_1 \geq 1200$.

PROOF OF LEMMA 2.

Let $\beta^n_i = \langle f^*, \phi_i \rangle_{L_2(Q)}$. Obviously $\inf_{g \in S_D} \|g - f^*\|_{L_2}$ is attained for $g = \sum_{i=1}^{D} \beta^n_i \phi_i$, therefore
\[
\inf_{D \geq 1} \left( \inf_{g \in S_D} \|g - f^*\|_{L_2} + \frac{D}{n} \right) = \inf_{D \geq 1} \left( \sum_{i=1}^{D} \beta_i^2 \right)^{\frac{1}{2}} + \frac{D}{n} \lesssim \inf_{D \geq 1} \left( \frac{D^2}{D} + \frac{D}{n} \right) \lesssim n^{-\frac{2a+1}{2a+2}} ,
\]
Finally solving this inequality for \(2\alpha - 1 < 4\gamma\) entails
\[
\mu_{\beta_0} \gtrsim n^{-\frac{4(2\alpha - 1)\gamma}{4\gamma + 2(\alpha - 1)\gamma + 1}}.
\]
so that
\[
\inf_{g \in \mathcal{H}} (\|g - f^*\|_2 + \Lambda_n \|g\|_{\mathcal{H}}^2) \gtrsim n^{\frac{-4(2\alpha - 1)\gamma}{4\gamma + 2(\alpha - 1)\gamma + 1}}.
\]
This concludes the proof of the lemma. \(\square\)

**ADDITIONAL MATERIAL: EIGENFUNCTIONS IN THE GAUSSIAN CASE**

We use the following normalization for the Hermite polynomials: \(H_n\) is an orthogonal system of \(L_2(e^{-x^2})\) i.e. \(e^{2\lambda x - \lambda^2} = \sum_{n \geq 0} H_n(x) e^{-\frac{x^2}{2}}\). In this case, if \(f_n(x) = H_n(x) e^{-\frac{x^2}{2}}\) then \(\langle f_n, f_m \rangle_{L_2(\mathbb{R})} = \delta_{n,m} \sqrt{\pi} 2^n n!\).

**Theorem 9 (\cite{37} and \cite{38}):** Let \(d\mu(x) = \frac{1}{\sqrt{2\pi}} e^{-2ax^2} dx\) and \(T_k\) be the integral operator associated with the Gaussian kernel \(k(x, y) = e^{-b(x-y)^2}\).

\[
T_k : L_2(\mu) \to L_2(\mu), \quad f \to \int_{\mathbb{R}} f(x) e^{-b(x-y)^2} d\mu(x).
\]

An explicit orthonormal basis of \(L_2(\mu)\) of eigenvectors of \(T_k\) associated to \(\lambda_j = \sqrt{\frac{1}{2\lambda}} \left(\frac{b}{A}\right)^{j-1}\) is given by:

\[
\Psi_j(x) = \frac{(4c)^{\frac{1}{2}} e^{-b \lambda x^2} H_{j-1}(\sqrt{2bx}) (2)^{-1}(j-1)!)^{\frac{1}{2}}}{(2)\left(\frac{b}{A}\right)^{j-1}}
\]
where \(c = \sqrt{a^2 + 2ab}\) and \(A = a + b + c\) and \(j \geq 1\).

**ACKNOWLEDGEMENTS**

We are extremely grateful to P. Massart and R. Vert who contributed equally with us in the original introduction of the KPM algorithm in \cite{8} and kindly allowed us to reuse the code and algorithm presentation for the purposes of the present paper.