Finite dimensional projection for classification and statistical learning

Laurent Zwald, Gilles Blanchard

Finite Dimensional Projection for Classification and Statistical Learning

Laurent Zwald and Gilles Blanchard

September 26, 2006

Abstract

A new method for the binary classification problem is studied. It relies on empirical minimization of the hinge loss over an increasing sequence of finite-dimensional spaces. A suitable dimension is picked by minimizing the regularized loss, where the regularization term is proportional to the dimension. An oracle-type inequality is established, 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.

1 Introduction.

1.1 The classification framework.

In this paper, we consider the framework of supervised binary classification. Let (X, Y) denote a random variable with values in $\mathcal{X} \times \{-1, +1\}$ 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 $(X_i, Y_i)_{i=1}^n$ 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 $\mathcal{X}$ to $\{-1, +1\}$ assigning to every point $x \in \mathcal{X}$ a prediction of its label. The quality of such a classifier is naturally measured by its *generalization error* $\mathbb{P}[f(X) \neq Y]$. 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 $f^*(x) = 2\mathbf{1}_{\{\eta(x)>1/2\}} - 1$ a.s. on the set $\{\eta(x) \neq 1/2\}$, where $\eta(x) = \mathbb{P}[Y = 1|X = x]$.

Given a certain set of classifiers $\mathcal{C}$ fixed in advance, the *Empirical Risk Minimization* procedure (see, e.g., [1]) consists in finding a classifier $f \in \mathcal{C}$ minimizing the empirical classification error $\frac{1}{n}\sum_{i=1}^n \mathbf{1}_{\{f(X_i)\neq Y_i\}}$. Here, the function $\ell(f, (x, y)) = \mathbf{1}_{\{f(x)\neq y\}}$, 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 γ over some real-valued (instead of $\{-1; 1\}$-valued) function space $\mathcal{F}$, opening way to the use of efficient convex programming methods. A real output function f over $\mathcal{X}$ can then be transformed into a binary classifier by considering sign(f). We will in this work concentrate on the surrogate loss function used by the SVM, called the *hinge loss*:

$$\gamma_h(g, (x, y)) = (1 - yg(x))_+ ,$$

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

1.2 Regularization in classification and in regression

Overfitting is the phenomenon of excessive discrepancy between empirical loss (observed on the sample) and generalization loss on fresh samples, and leads to performance degradation and inconsistency. To avoid overfitting, a common remedy is to consider regularization: it consists in adding to the empirical loss an additional balancing term $\Omega(f)$, which, roughly speaking, should be representative of how irregular the considered function is. The function $f \in \mathcal{F}$ 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 $\mathcal{H}$.

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 $B(R) = \{f \in \mathcal{H}, \|f\| \leq R\}$ (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 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.

The reason for this long discussion of the regression setting is to motivate the goals of this paper. Namely, it does not appear that an approach similar to the finite-dimensional projection has been studied for classification. This is precisely the aim of the present work.

The paper is organized as follows. In section 2, we present the finite-dimensional projection method and give our main theoretical result, an oracle inequality about the performance of the method. For this, we use theoretical tools of M-estimation ([6], [7]). In section 3, we compare the bound obtained to those obtained for SVMs. In section 4, we propose to use Kernel principal component analysis to estimate the subspaces used for projection. The resulting algorithm is dubbed the kernel projection machine (KPM): the model selection used in this new algorithm is guided by the previous theoretical result. Testing it on some benchmark datasets, we find results comparable to the SVM. While we do not outperform the SVM, the main point we want to convey in the present work is to show that finite dimensional projection is a viable alternative to Tikhonov's regularization.

2 Finite dimensional projection for classification.

2.1 The finite dimensional projection estimator.

Remember our main loss function is the hinge loss $\gamma_h(g,(x,y)) = (1 - yg(x))_+$. The hinge loss is consistent in the sense that the Bayes classifier satisfies (see [8])

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

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

$$\gamma_h(g,(x,y)) \geq \mathbf{1}_{\{g(x)\neq y\}}\,.$$

Moreover, the *excess* hinge loss with respect to the Bayes classifier upper-bounds the excess classification error:

$$\mathbb{E}\left[\gamma_h(g)\right] - \mathbb{E}\left[\gamma_h(f^*)\right] \geq P(Yg(X)\leq 0) - P(Yf^*(X)\leq 0)\,. \tag{1}$$

This means that the rate of convergence to the Bayes classifier for the hinge loss implies the same rate for the classification error excess risk.

We will now consider the following setting. Let $(\Psi_1, \Psi_2, \ldots)$ be an arbitrary family of functions fixed beforehand. We denote $S_D = \operatorname{span}\{\Psi_1, \cdots, \Psi_D\}$ 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 loss over S_D:

$$\widehat{f}_D = \arg\min_{f\in S_D} \frac{1}{n}\sum_{i=1}^{n}(1 - Y_i f(X_i))_+\,. \tag{2}$$

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 $\widehat{f}_D$. 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 $\widehat{f}_D$ are caused by the unboundedness of the loss function γ_h. In order to alleviate these, we introduce here the *clip* function (which was already considered in [9] in a regression framework, in [10], and in [11] in relation to the SVM) :

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

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

$$\gamma_h(\operatorname{clip}(f),(x,y)) = \min(\gamma_h(f,(x,y)),2)\,.$$

We now apply the following dimension selection strategy: we first "clip" the estimated function $\widehat{f}_D$ for each D, defining $\widetilde{f}_D = \operatorname{clip}(\widehat{f}_D)$. We define our final estimator by performing the model selection step over values of D using the clipped estimators $\widetilde{f}_D$. This is obtained by penalized minimization of the empirical loss: the final classifier is $\operatorname{sign}(f_{\widehat{D}})$ where

$$\widehat{D} = \arg\min_{D\geq 1}\left(\frac{1}{n}\sum_{i=1}^{n}(1 - Y_i\widetilde{f}_D(X_i))_+ + \lambda D\right)\,, \tag{3}$$

and the constant λ has to be suitably chosen. The linear penalty λD 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 selection step. We could, of course, consider the option of minimizing the 'trimmed' loss $\min(\gamma_h, 2)$ over each model S_D to get completely rid of boundedness issues. But since the trimmed loss is not convex, it would be difficult to devise a practical procedure to minimize this function over a model S_D. This is why we first minimize the *true* hinge loss γ_h on every model S_D (which is a convex optimization problem, hence amenable to convex programming techniques), and then choose the dimension D by minimizing the penalized trimmed loss of the functions $\widehat{f}_D$. This, again, is a practically feasible step, since we only have to compare a finite – and relatively small – number of functions (we typically expect that the maximum dimension D_{max} taken into consideration satisfies $D_{max} \leq n$). To sum up, the above procedure is practically feasible.

2.2 Main result.

We are now ready to formulate our main theoretical result: it aims at giving a precise statistical justification to the model selection procedure (3) involved in the finite dimensional regularization, and will be used for theoretical comparison with the SVM in the next section.

Theorem 1. *Let $(S_D)_{D\geq 1}$ be a family of linear subspaces of $L_2(P)$ where S_D is of dimension at most D and $\widehat{f}_D$ denotes the minimizer of the empirical loss:*

$$\widehat{f}_D = \arg\min_{f\in S_D} \sum_{i=1}^{n} (1 - Y_i f(X_i))_+ \,. \tag{4}$$

Let η be defined as $\eta(x) = P[Y = 1|X = x]$. We suppose that the following "noise margin" condition holds:

$$\exists\, h_0 > 0,\, \forall x \in \mathcal{X},\, \left|\eta(x) - \frac{1}{2}\right| \geq h_0 \,. \tag{5}$$

Let $\widetilde{f}_D = \operatorname{clip}(\widehat{f}_D)$. The dimension of the final estimator is selected by

$$\widehat{D} = \arg\min_{D\geq 1} \left(\frac{1}{n} \sum_{i=1}^{n} (1 - Y_i \widetilde{f}_D(X_i))_+ + \operatorname{pen}_n(D) \right) \,. \tag{6}$$

Then, there exist universal constants C_1 and C_2 such that, for any $K > 1$, the following holds: provided that

$$\forall\, D \geq 1,\, \operatorname{pen}_n(D) \geq \frac{C_1 K}{h_0} \frac{D \log n}{n} \,, \tag{7}$$

then

$$\mathbb{E}\left[L_h(\widetilde{f}_{\widehat{D}}, f^*) \right] \leq \frac{K}{K-1} \left(\inf_{D\geq 1} \left(\inf_{f\in S_D} L_h(f, f^*) + 2\mathbb{E}[\operatorname{pen}_n(D)] \right) \right) + \frac{C_2 K}{h_0 n} \,, \tag{8}$$

where $L_h(g, f^*) = \mathbb{E}\left[\gamma_h(g)\right] - \mathbb{E}\left[\gamma_h(f^*)\right]$ is the excess hinge loss.

This result is proved in appendix A 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 risk of the penalized estimator can be compared to the 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^*, 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^*. It means that the resulting estimator enjoys *adaptivity* properties.

- This result means that the penalty λD 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 (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 [12] (see also [13, 14]) show that this parameter plays a crucial role for rates of convergence in classification. Procedures that can be shown to be adaptive to this parameter (or a related condition) have only been studied very recently [15, 16].

3 Comparison with the risk bound for the SVM.

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 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., [11,17,18]). Here, we have chosen to compare the bound of Theorem 1 to the performance bound shown in [19]. The main reason for this choice is that the bounds in [19] 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 loss. 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 loss by the L_2 distance and compare the obtained bounds.

3.1 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 loss [20], [21]:

$$\widehat{g} = \operatorname*{arg\,min}_{g \in \mathcal{H}^b} \frac{1}{n} \sum_{i=1}^{n} (1 - Y_i g(X_i))_+ + \Lambda_n \|g\|_{\mathcal{H}}^2 . \tag{9}$$

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 $\mathcal{X}$, 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 $\widehat{g}_0$. This simplified setting is frequently (although not always) used in theoretical studies of the SVM, among which [19], which exposes the bound we wish to compare ourselves to. It is expected that this modification does no 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: $\widehat{g}_0 = \widehat{g}_{\widehat{R}}$ where

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

and

$$\widehat{R} = \operatorname*{arg\,min}_{R \geq 0} \left(\frac{1}{n} \sum_{i=1}^{n} (1 - Y_i \widehat{g}_R(X_i))_+ + CR^2 \right) , \tag{10}$$

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 $\mathcal{E}(R)$, $\widehat{g}_R$ is the minimum empirical risk estimator on each model, and criterion (10) is used to select the radius 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 $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. We will assume that $\mathcal{X}$ 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 $L_2(Q)$:

$$T_k : g \to \int_{\mathcal{X}} k(x,.)g(x)dQ(x)\,. \tag{11}$$

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

$$\mathcal{H} = \left\{ g \in L_2(Q) \,:\, g = \sum_{i\geq 1} a_i \Psi_i \,\text{such that}\, \|g\|^2_{\mathcal{H}} = \sum_{i\geq 1} \frac{a_i^2}{\lambda_i} < \infty \right\}. \tag{12}$$

We then have

$$\mathcal{E}(R) = \{g \in \mathcal{H}, \|g\|_{\mathcal{H}} \leq R\} = \left\{ g \in L_2(Q), g = \sum_{i\geq 1} a_i \Psi_i\,;\, \sum_{i\geq 1} \frac{a_i^2}{\lambda_i} \leq R^2 \right\}.$$

Consequently, a ball of the RKHS is an ellipsoid of $L_2(P)$ whose principal axes are precisely the eigenfunctions of T_k.

In [19], 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 $\widehat{g}_0$ defined in equation (9) satisfies with probability at least $1-\delta$, over the i.i.d. draw of the training sample:

$$L_h(\widehat{g}_0, f^*) \leq 2 \left(\inf_{g\in\mathcal{H}} L_h(g, f^*) + C(M)\Lambda_n \|g\|^2_{\mathcal{H}} \right) + C'\Lambda_n\,, \tag{13}$$

provided that Λ_n is bounded from below by a certain function depending on the eigenvalue sequence (λ_i) . If the eigenvalues are of the form $\lambda_j = \mathcal{O}(j^{-2\gamma})$ 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 [19] that a similar bound holds for the expected excess loss $\mathbb{E}L_h(\widehat{g}_0, f^*)$ (averaged over the draw of the training sample), up to additional logarithmic factors in the penalty. We will use this averaged loss version for comparison with Theorem 1.

3.2 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 (Ψ_i) and (λ_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 (Ψ_i) when Q is unknown.)

We will compare the obtained bounds in the following specific setting:

(a) the eigenvalues satisfy a polynomial decay $\lambda_j = \mathcal{O}(j^{-2\gamma})$. (Note that $\gamma > \frac{1}{2}$ since the eigenvalues series must be summable.)

(b) the coefficients of the Bayes classifier f^* in the $L_2(Q)$ orthogonal basis (Ψ_i) satisfy $\langle f^*, \Psi_j \rangle = \mathcal{O}(j^{-\alpha})$. (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 loss in the following way:

$$L_h(f,f^*) \leq \|f-f^*\|_{Q,1} \leq \|f-f^*\|_{Q,2} \ , \tag{14}$$

where the first inequality holds because the hinge loss γ_h is Lipschitz.

We are now left with comparing the following bounds: for the SVM classifier $\widehat{g}_0$, we have from (13), (14) and condition (a):

$$\mathbb{E}[L(\widehat{g}_0, f^*)] \lesssim \inf_{g\in\mathcal{H}} \left(\|g-f^*\|_{Q,2} + n^{-\frac{2\gamma}{2\gamma+1}} \|g\|^2_{\mathcal{H}} \right) \ , \tag{15}$$

while Theorem 1 together with (14) yields for the finite-dimensional estimator $\widetilde{f}$:

$$\mathbb{E}[L(\widetilde{f}, f^*)] \lesssim \inf_{D\geq 1} \left(\inf_{g\in S_D} \|g-f^*\|_{Q,2} + \frac{D}{n} \right) \ , \tag{16}$$

where $\lesssim$ 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:

Lemma 2. *Assume condition (b) above holds.Then the right-hand side of inequality* (16) *is bounded the following way:*

$$\inf_{D\geq 1} \left(\inf_{g\in S_D} \|g-f^*\|_{Q,2} + \frac{D}{n} \right) \leq \mathcal{O}\left(n^{-\frac{2\alpha-1}{2\alpha+1}} \right) \ .$$

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*

$$\inf_{g\in\mathcal{H}} \left(\|g-f^*\|_{Q,2} + n^{-\frac{2\gamma}{2\gamma+1}} \|g\|^2_{\mathcal{H}} \right) \geq \mathcal{O}\left(n^{-\frac{2(2\alpha-1)\gamma}{(2\gamma+1)(4\gamma-2\alpha+1)}} \right) \ .$$

In can be checked easily that $\frac{2\alpha-1}{2\alpha+1} > \frac{2(2\alpha-1)\gamma}{(2\gamma+1)(4\gamma-2\alpha+1)}$ is implied by the sufficient condition $\alpha \leq \gamma + \frac{3}{8}$. 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 $\alpha \leq \gamma + \frac{3}{8}$ 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 α and γ. 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, implying in turn necessarily that $\alpha < \gamma + \frac{1}{2}$. Therefore, the condition $\alpha \leq \gamma + \frac{3}{8}$ 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:

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

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 γ is an integer, the RKHS associated with this kernel is the Sobolev space of order γ with periodic boundary conditions $H^{(\gamma)}_{\mathrm{per}}$.

Precisely, $H^{(\gamma)}_{\text{per}}$ is the set of functions of $L_2([0,1])$ with $\gamma-1$ continuous derivatives satisfying $f(0)=f(1),\cdots,f^{(\gamma-1)}(0)=f^{(\gamma-1)}(0)$, and with $f^{(\gamma)}\in L_2([0,1])$. It is endowed with the Sobolev norm $\|f\|^2_{\text{Sob}}=\int_0^1|f^{(\gamma)}(t)|^2dt+\left(\int_0^1 f(t)dt\right)^2$. 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 γ 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 [19] that the quadratic penalty $\Lambda_n\|g\|^2_{\mathcal{H}}$ 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 $\Lambda_n\|g\|_{\mathcal{H}}$. 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.

4 Kernel projection machine and numerical results.

In this section, we will try to compare the finite dimensional projection approach to the classical SVM on real data. 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

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

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 (Ψ_i) are also unknown. A standard approach, the so-called *Nyström approximation* [22], 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 $\widehat{\Psi}_i$ of the operator

$$T_{k,n}f(x)=\frac{1}{n}\sum_{i=1}^{n}f(x_i)k(x,x_i) \tag{17}$$

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 [23]. 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 $K_{1,n}=\frac{1}{n}(k(X_i,X_j))_{1\leq i,j\leq n}$ to obtain the eigenfunctions of $T_{k,n}$. Indeed, for $j\geq 1$ such that $\widehat{\lambda}_j>0$,

$$\widehat{\Psi}_j(x)=\frac{1}{\sqrt{\widehat{\lambda}_j n}}\sum_{i=1}^{n}V_j^{(i)}k(X_i,x)\,, \tag{18}$$

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

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 $\widehat{\Psi}_i$ converges to Ψ_i as $n \to \infty$ and details of this convergence have been studied (see, e.g., [24–26]), 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 $(\widehat{\Psi}_i)$ are estimated using this second 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* $\mathcal{D} = (X_i, Y_i)_{i=1,\dots,n}$ *be an i.i.d. sample drawn from* P *and* $\mathcal{D}' = (X'_j)_{j=1,\dots,m}$ *be an indepent, unlabeled i.i.d. sample.*

Assume the unlabeled sample $\mathcal{D}'$ *is used to construct a family of functions* $(\widehat{\Psi}_i)$ *and put* $\widehat{S}_D = \operatorname{span}\left\{\widehat{\Psi}_1, \dots, \widehat{\Psi}_D\right\}$.

Follow the same estimation procedure as in Theorem 1 with the data-dependent subspace family $\widehat{S}_D$ *used in place of* S_D, *i.e. definitions* (4), (6), *giving rise to estimator* $\widetilde{f'}_{\widehat{D}}$.

Then, there exist universal constants C_1, C_2 *such that for any* $K > 1$, *if condition* (7) *on the penaly function holds, then*

$$\mathbb{E}_{\mathcal{D},\mathcal{D}'}\left[L_h(\widetilde{f'}_{\widehat{D}}, f^*)\right] \le \frac{K}{K-1}\mathbb{E}_{\mathcal{D}'}\left[\inf_{D\ge 1}\left(\inf_{f\in\widehat{S}_D} L_h(f, f^*) + 2\mathbb{E}_{\mathcal{D}}\left[\mathrm{pen}_n(D)\right]\right)\right] + \frac{C_2 K}{h_0 n}. \qquad (19)$$

Proof. Conditionally to the unlabeled sample $\mathcal{D}'$ the functions $(\widehat{\Psi}_i)$, and the models $\widehat{S}_D$ are fixed; therefore Theorem 1 applies. We then take the expectation of (8) with respect to $\mathcal{D}'$. □

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.

4.1 The kernel projection machine (KPM) algorithm.

Using the approximate eigenfunctions, the first step consists in computing the empirical minimizers over $\langle \mathbf{1}, \widehat{\Psi}_1, \cdots, \widehat{\Psi}_D\rangle$:

$$\widehat{f}_D = \arg\min_{f\in\langle\mathbf{1},\widehat{\Psi}_1,\cdots,\widehat{\Psi}_D\rangle} \sum_{i=1}^n (1 - Y_i f(X_i))_+ . \qquad (20)$$

Note that the constant function $\mathbf{1}$ 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 $(\widehat{\gamma}, \widehat{b})$ of $\widehat{f}$, of the form $\widehat{f}_D = \sum_{j=1}^D \frac{\widehat{\gamma}_j}{\sqrt{n\widehat{\lambda}_j}}\widehat{\phi}_j + b^*$, then equation (18) leads to $\widehat{\phi}_j(X_i) = \left(\sqrt{\frac{n}{\widehat{\lambda}_i}} K_{1,n} V_j\right)^{(i)} = \sqrt{n\widehat{\lambda}_j}V_j^{(i)}$, so that $(\widehat{\gamma}, \widehat{b})$ are given as the solutions of

$$(\widehat{\gamma}, \widehat{b}) = \arg\min_{\gamma\in\mathbb{R}^D, b\in\mathbb{R}} \sum_{i=1}^n \left(1 - Y_i\left(\sum_{j=1}^D \gamma_j V_j^{(i)} + b\right)\right)_+ .$$

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

1. given data $X_1, \dots, X_n \in \mathcal{X}$ and a positive kernel k defined on $\mathcal{X} \times \mathcal{X}$, compute the kernel matrix $K_{1,n}$ and its eigenvectors $V_1, \dots, V_n$ together with its eigenvalues in decreasing order $\widehat{\lambda}_1 \geq \widehat{\lambda}_2 \geq \ldots \geq \widehat{\lambda}_n$.

2. for each dimension D such that $\widehat{\lambda}_D > 0$ solve the linear optimization problem

$$\left(\widehat{\gamma}, \widehat{b}\right) = \arg\min_{\gamma,b,\xi} \sum_{i=1}^{n} \xi_i$$

$$\text{under the constraints: } \forall i = 1 \ldots n \ : \qquad \xi_i \geq 0\,, \text{ and } Y_i \left(\sum_{j=1}^{D} \gamma_j V_j^{(i)} + b \right) \geq 1 - \xi_i\,. \quad (21)$$

   Next, put $\widehat{\alpha}_i = \displaystyle\sum_{j=1}^{D} \frac{\widehat{\gamma}_j}{n\widehat{\lambda}_j} V_j^{(i)}$, and finally $\widehat{f}_D = \sum_{i=1}^{n} \widehat{\alpha}_i k(x_i, .) + \widehat{b}\,.$

3. The last step is the model selection problem consisting in choosing the dimension D; for this step, we use the penalized clipped hinge loss as studied earlier:

$$\widehat{D} = \arg\min_{D \geq 1} \left(\frac{1}{n} \sum_{i=1}^{n} (1 - Y_i \widetilde{f}_D(X_i))_+ + \lambda D \right) , \qquad (22)$$

   where $\widetilde{f}_D = \operatorname{clip}(\widehat{f}_D)$.

4.2 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 λ 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.

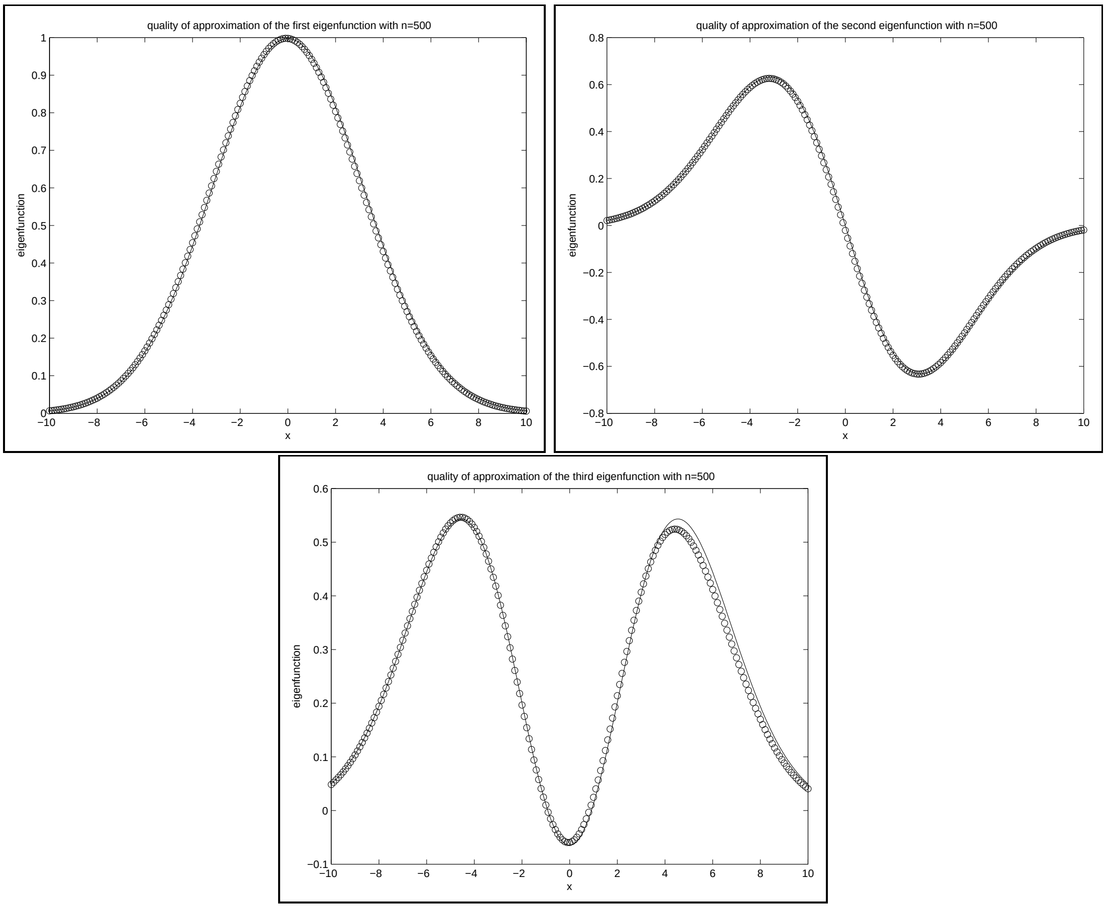
In our final experiments, we consider in turn a comparison for a realistic practical scenario: in this case, 5-fold cross-validation is used to select the regularization parameter for each method.

4.2.1 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 $a = \frac{1}{4}$ 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 [24]: 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.

Figure 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.



4.2.2 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 [27]: 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. [27] reports the results obtained by applying several state-of-art classification algorithms, 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 σ_G (kernel width) and C_G (SVM regulatization 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 σ_G fixed to the value reported in [27]. Without selecting the regularization parameter, the family of classifiers obtained by the KPM algorithm is $(\widetilde{f}_D)_{D\geq 1}$ and the family obtained by the SVM algorithm is $(\widehat{f}_C)_{C\in\mathcal{C}}$. The set $\mathcal{C}$ 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 $100\,C_G$ 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 1, 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.

Table 1: Best classifier in the family (nb. of wins)

	KPM	**SVM**
Banana	31	67
Breast Cancer	44	50
Diabetis	55	42
Flare Solar	19	63
German	43	49
Heart	27	64

Parameter selection by cross-validation. Table 2 presents results of SVM (resp. KPM) where the regularization parameters C (resp. λ) is chosen by 5 fold-cross-validation separately on each of the samples. The results are presented in the form { mean of the 100 test errors } ± { variance of the 100 test errors }.

Table 2: Test errors

	SVM	**KPM**
Banana ($\sigma = 0.7071$)	10.69 ± 0.67	10.91 ± 0.57
Breast Cancer($\sigma = 5$)	26.68 ± 5.23	28.73 ± 4.42
Diabetis ($\sigma = 3.1623$)	23.79 ± 2.01	23.77 ± 1.69
Flare Solar ($\sigma = 3.8730$)	32.62 ± 1.86	32.52 ± 1.78
German ($\sigma = 5.2440$)	23.79 ± 2.12	24.09 ± 2.38
Heart ($\sigma = 7.7460$)	16.23 ± 3.18	17.35 ± 3.54

These two tables highlight that the performances of KPM are comparable with the SVM. Considering table 1, the SVM appears to have a slight advantage over the KPM which also appears in table 2 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 σ_G is used for the KPM and the SVM, whereas it was originally tailored in [27] for good performance of the SVM only.

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

5 Conclusion and discussion.

5.1 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.

5.2 Comparison with other work.

We provided a detailed comparison of the theoretical bounds obtained for the KPM and of the bounds obtained in [19] for the SVM. It is more difficult to draw a meaningful comparison with other known bounds on the SVM; for example, in [17], 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 [15], estimators with a finite expansion on a fixed function basis are considered, which is related to the present setting. A ℓ_1-penalty in the coefficients is considered, and the procedure is shown to be adapative to the "Tsybakov noise exponent" parameter (a more general version of (5)). Here, we study a ℓ_0-penalty and our results are not adaptive to the noise margin parameter. However, [15] 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.

5.3 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 autoadjoint 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 [28] , 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 [29]), but has received renewed attention recently due in particular to the strong link of support vector machines with inverse problems, as studied notably in [18]. 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.

5.4 Ending remarks and future work.

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

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 several different kernels. Oracle inequalities can be obtained in this case using the same methodology (it suffices to change accordingly the weight x_D appearing in the proof of Theorem 1). To avoid additional technicalities, in this paper, only the simplest version involving one model for each dimension is stated. However, it is clear that an extension where several subspaces with the same dimension D are available is straightforward.

Taking into account precisely the variability in the data-dependent models is also an interesting topic. One possible lead is to use stability results on the estimated models (see, e.g., [26]); another is to extend the study of the method in a semi-supervised setting where unlabeled data is available, thereby developing the ideas of Section 4.

Appendix

A Risk bounds for the clipped finite dimensional approach.

In this appendix, Theorem 1 is proved considering a slightly more generic setting. The training data $((X_1, Y_1), \ldots, (X_n, Y_n))$ belong to $(\mathcal{X} \times \mathcal{Y})^n$ and we only assume that the loss function is Lipschitz.

A.1 Clipped empirical risk minimization on one model.

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

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

and

$$\widetilde{s}_D = \mathrm{clip}(\widehat{s}_D) \,.$$

The goal is to estimate the target function s minimizing the average loss 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 γ is:

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

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

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

We will prove two results; the first one (Theorem 5) provides a 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 $\Psi : [0, \infty) \to [0, \infty)$ is called *sub-root* if it is non-negative, non-decreasing and if $r \to \frac{\Psi(r)}{\sqrt{r}}$ 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 $\Psi \equiv 0$).

Let $\mathcal{F}$ be a set of functions. The following notation for the Rademacher average of $\mathcal{F}$ will be useful:

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

where $(\varepsilon_i)_{i=1 \cdots n}$ 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 concidered only with respect to ε: the variables $X_1, \ldots, X_n$ are "fixed".

We now recall the following result:

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

$$\widehat{\Psi}_n(r) \geq 2(10 \vee \kappa) \mathbb{E}_\epsilon \mathcal{R}_n \{f \in \mathrm{star}(\mathcal{F}); P_n f^2 \leq 2r\} + (2(10 \vee \kappa) + 11) \frac{\xi}{n} \,. \tag{23}$$

Then, for any $K > 1$ with probability at least $1 - 3e^{-\xi}$,

$$\forall f \in \mathcal{F}, Pf \leq \frac{K}{K-1} P_n f + \frac{6K}{\kappa} \widehat{r}^* + \frac{\xi(11 + 5\kappa K)}{n} \,.$$

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

$$\mathcal{F}_D = \{(x, y) \to \gamma(s(x), y) - \gamma(s^*(x), y), s \in \mathrm{clip}(S_D)\} \,. \tag{24}$$

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 A.4.

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 $\widehat{\psi}(r) = \mathbb{E}_\varepsilon \mathcal{R}_n \left\{ f \in \mathrm{star}(\mathcal{F}_D) \,, P_n f^2 \leq 2r \right\}$. Then, the following holds:*

$$\widehat{r}^*_{D,n} \leq A_1 \frac{D+1}{n} \left(\left(\log \frac{n}{D} \right)_+ + 1 \right) \,. \tag{25}$$

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) γ *satisfies assumption* (**A**)*.*

(iii) $\forall s \in \mathrm{clip(S_D)}$, $\|\gamma(s) - \gamma(s^*)\|_2^2 \leq \kappa L(s, s^*)$ *.*

Then, for all $K > 1$ *, the following inequality holds:*

$$\mathbb{E}\left[L(\widetilde{s}_D, s^*)\right] \leq \frac{K}{K-1}\left(\inf_{t \in S_D} L(t, s^*) + C_3 K \frac{(10 \vee \kappa)^2}{\kappa} \frac{D}{n} \log n\right) + C_4 \frac{K(\kappa \vee 10)}{n} ,$$

where C_3 *and* C_4 *are numerical constants.*

Proof of Theorem 5 . Let s_D be an arbitraty element of S_D . First, note that

$$P_n(\gamma(\widetilde{s}_D) - \gamma(s)) \leq P_n(\gamma(\widehat{s}_D) - \gamma(s)) \leq P_n(\gamma(s_D) - \gamma(s)) , \tag{26}$$

where the first inequality follows from assumption (**A2**) and the last from the definition of the empirical minimizer. Thus,

$$L(\widetilde{s}_D, s^*) = P(\gamma(\widetilde{s}_D) - \gamma(s^*) \leq (P - P_n)(\gamma(\widetilde{s}_D) - \gamma(s^*)) + P_n(\gamma(s_D) - \gamma(s^*)) . \tag{27}$$

In order to control $(P - P_n)(\gamma(\widetilde{s}_D) - \gamma(s^*))$, we apply as announced Theorem 3 to the class of functions $\mathcal{F}_D = \{\gamma(s) - \gamma(s^*), t \in \mathrm{clip}(S_D)\}$. 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) ensures that $Pf^2 \leq \kappa Pf$ for all $f \in \mathcal{F}_D$, which was 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 $\widehat{\psi}(r) = \mathbb{E}_\varepsilon \mathcal{R}_n \left\{f \in \mathrm{star}(\mathcal{F}_D) \, , P_n f^2 \leq 2r\right\}$. In sight of (23) , we need a bound on the fixed point $\widetilde{r}^*_{D,n}$ of an affine tranform $a\widehat{\psi}(x) + b$ of $\widehat{\psi}$ (with $a = 2(10 \vee \kappa)$, $b = (2(10 \vee \kappa) + 11)\frac{\xi}{n}$). Elementary arguments not reproduced here (see [30], Lemma 4.10) show that

$$\widetilde{r}^*_{D,n} \leq 4a^2 r^*_{D,n} + 2b \leq 16(10 \vee \kappa)^2 A_1 \frac{D+1}{n}\left(\left(\log \frac{n}{D}\right)_+ + 1\right) + \frac{2\xi}{n}(2(10 \vee \kappa) + 11) .$$

Consequently, Theorem 3 implies that $\forall K > 1$, with probability at least $1 - 3e^{-\xi}$, $\forall f \in \mathcal{F}_D$,

$$(P - P_n)f \leq \frac{1}{K-1} P_n f + \frac{6K}{\kappa} \widetilde{r}^*_{D,n} + \frac{\xi(11 + 5\kappa K)}{n} .$$

Since the bound is available simultaneously for all functions in $\mathcal{F}_D$, we can apply it to the random function $f = \gamma(\widetilde{s}_D) - \gamma(s^*) \in \mathcal{F}_D$. This yields that, with probability at least $1 - 3e^{-\xi}$,

$$(P - P_n)(\gamma(\widetilde{s}_D) - \gamma(s^*)) \leq \frac{1}{K-1} P_n(\gamma(\widetilde{s}_D) - \gamma(s^*)) + \frac{6K}{\kappa} \widetilde{r}^*_{D,n} + \frac{\xi(11 + 5\kappa K)}{n} . \tag{28}$$

Using again inequality (26), we get:

$$(P - P_n)(\gamma(\widetilde{s}_D) - \gamma(s^*)) \leq \frac{1}{K-1} P_n(\gamma(s_D) - \gamma(s^*)) + \frac{6K}{\kappa} \widetilde{r}^*_{D,n} + \frac{\xi(11 + 5\kappa K)}{n} .$$

We now plug this inequality into (27) to obtain

$$L(\widetilde{s}_D, s^*) \leq \frac{K}{K-1} P_n(\gamma(s_D) - \gamma(s^*)) + \frac{6K}{\kappa} \widetilde{r}^*_{D,n} + \frac{\xi(11 + 5\kappa K)}{n} .$$

This concludes the proof of Theorem 5 by integrating with respect to the sample, then taking the infimum over $s_D \in S_D$. □

A.2 Model selection by penalization.

We now present a relatively general result about penalized minimization of the clipped empirical loss 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 $\dim(S_D) \leq D$. Assume the following:*

(i) The target function is bounded by 1: $|s^*(x)| \leq 1$.

*(ii) γ satisfies assumption (**A**).*

(iii) $\forall s \in \mathrm{clip(S_D)}$, $\|\gamma(s) - \gamma(s^*)\|_2^2 \leq \kappa L(s, s^*)$.

Let $K > 1$. Choosing the dimension with the following penalized criterion

$$\widehat{D} = \operatorname*{arg\,min}_{D\geq 1} \left(\frac{1}{n} \sum_{i=1}^{n} \gamma(\widetilde{s}_D, (X_i, Y_i)) + \mathrm{pen}_n(D) \right) ,$$

with a possibly data dependent penalty function pen_n such that

$$\forall\, D \geq 1 \,, \mathrm{pen}_n(D) \geq C_5 K \frac{(\kappa \vee 10)^2}{\kappa} \frac{D}{n} \log n \,, \tag{29}$$

the following inequality holds

$$\mathbb{E}[L(\widetilde{s}_{\widehat{D}}, s^*)] \leq \frac{K}{K-1} \left(\inf_{D\geq 1} \left(\inf_{s\in S_D} L(s, s^*) + \mathbb{E}[\mathrm{pen}_n(D)] \right) \right) + \frac{C_6 K(\kappa \vee 10)}{n} ,$$

where C_5 and C_6 are numerical constants.

Proof. Let s_D be a fixed element of S_D. The definition of $\widehat{s}_{\widehat{D}}$ leads to the following chain of inequalities: $\forall D \geq 1$,

$$P_n\gamma(\widetilde{s}_{\widehat{D}}) + \mathrm{pen}_n(\widehat{D}) \leq P_n\gamma(\widetilde{s}_D) + \mathrm{pen}_n(D) \leq P_n\gamma(\widehat{s}_D) + \mathrm{pen}_n(D) \leq P_n\gamma(s_D) + \mathrm{pen}_n(D) \,, \tag{30}$$

where the second inequality is due to assumption **(A2)**. Thus, $\forall D \geq 1$,

$$\begin{aligned} L(\widetilde{s}_{\widehat{D}}, s^*) &= (P - P_n)(\gamma(\widetilde{s}_{\widehat{D}}) - \gamma(s^*)) + P_n(\gamma(\widetilde{s}_{\widehat{D}}) - \gamma(s^*)) \\ &\leq (P - P_n)(\gamma(\widetilde{s}_{\widehat{D}}) - \gamma(s^*)) + P_n(\gamma(s_D) - \gamma(s^*)) + \mathrm{pen}_n(D) - \mathrm{pen}_n(\widehat{D}) \,. \quad (31) \end{aligned}$$

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 (28) in the proof of Theorem 5, obtaining that $\forall K > 1$, with probability at least $1 - 3e^{-\xi - x_{D'}}$,

$$(P - P_n)(\gamma(\widetilde{s}_{D'}) - \gamma(s^*)) \leq \frac{1}{K-1} P_n(\gamma(\widetilde{s}_{D'}) - \gamma(s^*)) + C\frac{K(10\vee\kappa)^2}{\kappa}\frac{D'}{n}\log(n) + C'\frac{K(\xi + x_{D'})(10\vee\kappa)}{n} ,$$

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

$$(P - P_n)(\gamma(\widetilde{s}_{\widehat{D}}) - \gamma(s^*)) \leq \frac{1}{K-1} P_n(\gamma(\widetilde{s}_{\widehat{D}}) - \gamma(s^*)) + C\frac{K(10\vee\kappa)^2}{\kappa}\frac{\widehat{D}}{n}\log(n) + C'\frac{K(\xi + x_{\widehat{D}})(10\vee\kappa)}{n} .$$

Plugging this inequality into (31) and using again (30) leads to:

$$L(\widetilde{s}_{\widehat{D}}, s^*) \leq C\frac{K(10\vee\kappa)^2}{\kappa}\frac{\widehat{D}}{n}\log(n) + C'\frac{K(\xi + x_{\widehat{D}})(10\vee\kappa)}{n} + \frac{K}{K-1}\left(P_n(\gamma(s_D) - \gamma(s^*)) + \mathrm{pen}_n(D) - \mathrm{pen}_n(\widehat{D})\right) .$$

Choosing $x_D = 2\log(D+1)$, condition (29) entails

$$L(\widetilde{s}_{\widehat{D}}, s^*) \leq C' \frac{K\xi(10 \vee \kappa)}{n} + \frac{K}{K-1}\left(P_n(\gamma(s_D) - \gamma(s^*)) + \mathrm{pen}_n(D)\right) .$$

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

$$L(\widetilde{s}_{\widehat{D}}, s^*) \leq \frac{K}{K-1}\mathbb{E}\left[\inf_{D\geq 1}\left(P_n(\gamma(s_D) - \gamma(s^*)) + \mathrm{pen}_n(D)\right)\right] + C' \frac{3K(10 \vee \kappa)}{n} ,$$

and finally:

$$L(\widetilde{s}_{\widehat{D}}, s^*) \leq \frac{K}{K-1}\inf_{D\geq 1}\left(P(\gamma(s_D) - \gamma(s^*)) + \mathbb{E}[\mathrm{pen}_n(D)]\right) + C'(M,B)\frac{3K(10 \vee \kappa)}{n} .$$

This concludes the proof of Theorem 6. □

A.3 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 [19], but we provide a proof here for completeness.) This concludes the proof of Theorem 1.

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

$$\|\gamma_h(f) - \gamma_h(f^*)\|_2^2 \leq \frac{1}{h_0} L_h(f, f^*) .$$

Proof of Lemma 7. We can write explicilty

$$L_h(f, f^*) = \int p(x)[(1 - f(x))_+ - (1 - f^*(x))_+] + (1 - p(x))[(1 + f(x))_+ - (1 + f^*(x))_+] \, dP(x) ,$$

and

$$\|\gamma_h(f) - \gamma_h(f^*)\|_2^2 = \int p(x)[(1-f(x))_+ - (1-f^*(x))_+]^2 + (1-p(x))[(1+f(x))_+ - (1+f^*(x))_+]^2 \, dP(x) .$$

Without loss of generality, we suppose that $f^*(x) = 1$ i.e. $p(x) \geq \frac{1}{2}$. Since $-1 \leq f(x) \leq 1$, $p(x)[(1 - f(x))_+ - (1 - f^*(x))_+]^2 + (1 - p(x))[(1 + f(x))_+ - (1 + f^*(x))_+]^2 = (1 - f(x))^2$. It yields $\frac{(f(x)-1)^2}{p(x)[(1-f(x))_+)]+(1-p(x))[(1+f(x))_+-2]} = \frac{1-f(x)}{2p(x)-1}$. We therefore obtain

$$\frac{p(x)[(1 - f(x))_+)]^2 + (1 - p(x))[(1 + f(x))_+ - 2]^2}{p(x)[(1 - f(x))_+)] + (1 - p(x))[(1 + f(x))_+ - 2]} \leq \frac{1}{h_0} ,$$

and the statement follows by integrating with respect to x. □

A.4 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. [9]. 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 ν be a probability measure on $\mathcal{X}$. Let ε such that $0 < \varepsilon < \frac{A}{4}$. Then the following holds :*

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

where $V(\mathcal{G})$ is the VC-dimension of all subgraphs of functions 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 $\operatorname{star}(\mathcal{F}_D)$. Note that since this set contains the null function and has diameter bounded by 1, any covering number for $\varepsilon > 1$ is equal to 1. In what follows we therefore assume $\varepsilon \leq 1$.

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

$$\mathcal{N}\left(\varepsilon, \operatorname{star}(\mathcal{F}), L_2(P_n)\right) \leq \left(\left\lceil \frac{2}{\varepsilon} \right\rceil + 1 \right) \mathcal{N}\left(\frac{\varepsilon}{2}, \mathcal{F}, L_2(P_n) \right) . \qquad (32)$$

Moreover,

$$\begin{aligned} \mathcal{N}\left(\frac{\varepsilon}{2}, \mathcal{F}, L_2(P_n)\right) &\leq \mathcal{M}\left(\frac{\varepsilon}{2}, \mathcal{F}, L_2(P_n)\right) \\ &\leq \mathcal{M}\left(\frac{\varepsilon}{2}, \{t - s, t \in \operatorname{clip}(S_D)\}, L_2(P_n)\right) \\ &= \mathcal{M}\left(\frac{\varepsilon}{2}, \{t + 1, t \in \operatorname{clip}(S_D)\}, L_2(P_n)\right) , \end{aligned}$$

where the second inequality holds because γ 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 $0 < \varepsilon < 1$,

$$\begin{aligned} \mathcal{M}\left(\frac{\varepsilon}{2}, \{t + 1, t \in \operatorname{clip}(S_D)\}, L_2(P_n)\right) &\leq 3 \left(\frac{32e}{\varepsilon^2} \log\left(\frac{48e}{\varepsilon^2} \right) \right)^{V(\operatorname{clip}(S_D)+1)} \\ &\leq \left(\frac{11}{\varepsilon} \right)^{4V(\operatorname{clip}(S_D)+1)} . \end{aligned}$$

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

Moreover, $V(\operatorname{clip}(S_D) + 1) = V(\operatorname{clip}(S_D)) \leq V(S_D) \leq D + 1$. 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 [31]). Finally, the middle inequality comes from the observation that a discrete set shattered by the subgraph class of $\operatorname{clip}(\mathcal{G})$ is also shattered by the subgraph class of $\mathcal{G}$. Namely, let $(z_i, t_i)_{1 \leq i \leq K}$ be a finite set shattered by the subgraph class of $\operatorname{clip}(\mathcal{G})$. 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 $(\operatorname{clip}(g_j))_{1 \leq j \leq 2^K}$ is a witness of the shattering in $\operatorname{clip}(\mathcal{G})$, then so is (g_j) in $\mathcal{G}$.

Now, gathering inequality (32) and the previous ones, we get for $0 < \varepsilon < 1$,

$$\begin{aligned} \log \mathcal{N}(\varepsilon, \operatorname{star}(\mathcal{F}), L_2(P_n)) &\leq \log\left(\frac{4}{\varepsilon}\right) + 4(D+1)\log\left(\frac{20}{\varepsilon}\right) \\ &\leq 5(D+1)\log\left(\frac{11}{\varepsilon}\right) . \end{aligned}$$

We now use Dudley's entropy integral (also known as chaining technique, see [32]) theorem, which states that

$$\mathbb{E}_\varepsilon \mathcal{R}_n(\mathcal{G}) \leq \frac{4\sqrt{2}}{\sqrt{n}} \int_0^\infty \sqrt{\log \mathcal{N}(u, \mathcal{G}, L_2(P_n))} du\,.$$

Therefore, using the above upper bound on the covering numbers and standard calculus,

$$\begin{aligned}\mathbb{E}_\epsilon \mathcal{R}_n \left(f \in \operatorname{star}(\mathcal{F}),\, P_n f^2 \leq 2r\right) &\leq \frac{4\sqrt{2}}{\sqrt{n}} \int_0^{\sqrt{2r}\wedge 1} \sqrt{\log \mathcal{N}(\varepsilon, \operatorname{star}(\mathcal{F}), L_2(P_n))} du \\ &\leq \frac{4\sqrt{10}}{\sqrt{n}} \sqrt{D+1} \int_0^{\sqrt{2r}\wedge 1} \sqrt{\log\left(\frac{20}{u}\right)} du \\ &\leq 22 \left(rn^{-1}(D+1)\log\left(\frac{11}{\sqrt{2r}\wedge 1}\right)\right)^{\frac{1}{2}} := \widehat{\phi}(r)\,.\end{aligned}$$

We easily check that $\widehat{\phi}$ is a sub-root function. We now want to upper bound the fixed point of $\widehat{\phi}$. Let us denote $\widetilde{r}^*_{D,n} = A_1 \frac{D+1}{n}\left(\left(\log \frac{n}{D}\right)_+ + 1\right)$; in order (25) to hold, A_1 has to be chosen such that

$$\widehat{\phi}\left(\widetilde{r}^*_{D,n}\right) \geq \widetilde{r}^*_{D,n}\,. \tag{33}$$

If $A_1 \geq 11^2/(2e)$, then it can be checked that

$$\log\left(\frac{11}{\sqrt{2\widetilde{r}^*_{D,n}} \wedge 1}\right) = \log\left(\frac{11}{\sqrt{2\widetilde{r}^*_{D,n}}}\right) \vee \log(11) \leq \log(11)\left(\left(\log\frac{n}{D}\right)_+ + 1\right)\,,$$

and therefore

$$\widehat{\phi}\left(A_1 \frac{D+1}{n}\left(\left(\log\frac{n}{D}\right)_+ + 1\right)\right) \leq C\frac{D+1}{n}\left(\left(\log\frac{n}{D}\right)_+ + 1\right)\,,$$

where $C = 22\sqrt{A_1 \log(11)}$.We obtain (33) by noting that $C \leq A_1$ if $A_1 \geq 1200$.

B Proof of Lemma 2.

Let $\beta^*_i = \langle f^*, \phi_i \rangle_{L_2(P)}$. Obviously $\inf_{g\in S_D} \|g - f^*\|_{Q,2}$ is attained for $g = \sum_{i=1}^D \beta^*_i \phi_i$, therefore

$$\inf_{D\geq 1}\left(\inf_{g\in S_D} \|g-f^*\|_{Q,2} + \frac{D}{n}\right) = \inf_{D\geq 1}\left(\left(\sum_{i>D}\beta_i^{*2}\right)^{\frac{1}{2}} + \frac{D}{n}\right) = \mathcal{O}\left(\inf_{D\geq 1}\left(D^{\frac{1}{2}-\alpha} + \frac{D}{n}\right)\right) = \mathcal{O}\left(n^{-\frac{2\alpha-1}{2\alpha+1}}\right)\,,$$

where we have used $\beta^*_i = \mathcal{O}(j^{-\alpha})$ and taken $D = \lfloor n^{\frac{2}{2\alpha-1}} \rfloor$.

For the second inequality of the lemma, putting $\Lambda_n = n^{-\frac{2\gamma}{2\gamma+1}}$, we have

$$\inf_{g\in\mathcal{H}}\left(d(g,f^*) + \Lambda_n \|g\|^2_{\mathcal{H}}\right) = \inf_{(\beta_i)}\left(\mu_\beta + \Lambda_n \sum_{i\geq 1}\frac{\beta_i^2}{\lambda_i}\right)\,,$$

where $\mu_\beta = \sqrt{\sum_{i\geq 1}(\beta_i - \beta^*_i)^2}$. The infimum point (β°_i) is such that every partial derivative with respect to β_i cancels, so that

$$\beta^\circ_i = \frac{\beta^*_i \lambda_i}{\lambda_i + 2\Lambda_n \mu_{\beta^\circ}}\,,$$

and

$$\mu_{\beta^\circ} = 2\sqrt{\sum_{i\geq 1}\left(\frac{\beta_i^*\Lambda_n\mu_{\beta^\circ}}{\lambda_i + 2\Lambda_n\mu_{\beta^\circ}}\right)^2} \geq \frac{1}{2}\sqrt{\sum_{i\geq 1, \lambda_i\leq 2\Lambda_n\mu_{\beta^\circ}} \beta_i^{*2}}\,.$$

Since $\beta_i^* = \mathcal{O}(j^{-\alpha})$, this leads to

$$\mu_{\beta^\circ} \geq \mathcal{O}\left(n^{-\frac{2\alpha-1}{4\gamma+2}}\mu_{\beta^\circ}^{\frac{2\alpha-1}{4\gamma}}\right)\,.$$

Finally solving this inequality for $2\alpha - 1 < 4\gamma$ entails

$$\mu_{\beta^\circ} \geq \mathcal{O}\left(n^{\frac{-4(2\alpha-1)\gamma}{(4\gamma+2)(4\gamma-2\alpha+1)}}\right)\,,$$

so that

$$\inf_{g\in\mathcal{H}}\left(\|g - f^*\|_2 + \Lambda_n\|g\|_{\mathcal{H}}^2\right) \geq \mathcal{O}\left(n^{\frac{-4(2\alpha-1)\gamma}{2(2\gamma+1)(4\gamma-2\alpha+1)}}\right)\,.$$

This concludes the proof of the lemma. □

C 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. $e^{2\lambda x-\lambda^2} = \sum_{n\geq 0} H_n(x)\frac{\lambda^n}{n!}$. 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 ([33] and [34]). *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* $k(x,y) = e^{-b(x-y)^2}$.

$$\begin{array}{llll} T_k : & L_2(\mu) & \to & L_2(\mu) \\ & f & \to & \int_{\mathbb{R}} f(x)e^{-b(x-y)^2}d\mu(x) \end{array}$$

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

$$\Psi_j(x) = \frac{(4c)^{\frac{1}{4}}e^{-(c-a)x^2}H_{j-1}(\sqrt{2c}x)}{(2^{j-1}(j-1)!)^{\frac{1}{2}}}$$

where $c = \sqrt{a^2 + 2ab}$ *and* $A = a + b + c$ *and* $j \geq 1$.

Acknowledgements

The authors are extremely grateful to P. Massart for giving us the original idea of the algorithm as well as for invaluable comments and to R. Vert for his precious help.

References

[1] V. N. Vapnik, *Statistical Learning Theory.* New York: Wiley, 1998.

[2] G. Wahba, *Spline Models for Observational Data*, ser. CBMS-NSF Regional Conference Series in Applied Mathematics. Philadelphia, Pennsylvania: Society for Industrial and Applied Mathematics, 1990, vol. 59.

[3] A. Barron, L. Birgé, and P. Massart, “Risk bounds for model selection via penalization,” *Proba. Theory Relat. Fields*, vol. 113, pp. 301–413, 1999.

[4] Y. Baraud, “Model selection for regression on a random design,” *ESAIM Probab. Statist. 6 127–146*, 2002.

[5] P. Massart, *Concentration Inequalities and Model Selection.* Springer-Verlag, 2003, probability summer school, Saint Flour 2003 (to appear), available at http://www.math.u-psud.fr/ massart/stf2003_massart.pdf.

[6] ——, “Some applications of concentration inequalities to statistics,” *Annales de la Faculté des Sciences de Toulouse*, vol. IX, pp. 245–303, 2000.

[7] P. Bartlett, O. Bousquet, and S. Mendelson, “Local Rademacher complexities,” *Annals of Statistics*, vol. 33, no. 4, pp. 1497–1537, 2005.

[8] Y. Lin, “Support Vector Machines and the Bayes rule in classification,” *Data Mining and Knowledge Discovery, 6, 259-275*, 2002.

[9] L. Györfi, M. Kohler, A. Krzyżak, and H. Walk, *A Distribution–Free Theory of Nonparametric Regression*, ser. Springer Series in Statistics. New York: Springer Verlag, 2002.

[10] P. L. Bartlett, “The sample complexity of pattern classification with neural networks: the size of the weights is more important than the size of the network,” *IEEE Transactions on Information Theory*, vol. 44, no. 2, pp. 525–536, 1998.

[11] D.-R. Chen, Q. Wu, Y. Ying, and D.-X. Zhou, “Support vector machine soft margin classifiers: Error analysis,” *Journal of Machine Learning Research*, vol. 5, pp. 1143–1175, 2004.

[12] E. Nedelec and P. Massart, “Risk bounds for statistical learning,” *Annals of Statistics*, To appear.

[13] E. Mammen and A. B. Tsybakov, “Asymptotical minimax recovery of sets with smooth boundaries,” *The Annals of Statistics*, vol. 23, no. 2, pp. 502–524, 1995.

[14] A. Tsybakov, “Optimal aggregation of classifiers in statistical learning,” *Annals of Statistics, 32 (1)*, 2004.

[15] B. Tarigan and S. Van de Geer, “Adaptivity of support vector machines with ℓ_1 penalty,” University of Leiden, Tech. Rep. Technical Report MI 2004-14, 2004, (to appear in *Bernoulli*).

[16] J.-Y. Audibert, “Model selection type aggregation with better variance control,” *Technical report CERTIS R.R. 06-20, Ecole Nationale des Ponts et Chaussees*, 2006.

[17] I. Steinwart and C. Scovel, “Fast rates for support vector machines using Gaussian kernels,” Los Alamos National Laboratory, Tech. Rep. LA-UR-04-8796, 2004, (submitted for publication).

[18] E. De Vito, L. Rosasco, A. Caponnetto, U. De Giovannini, and F. Odone, “Learning from examples as an inverse problem,” *Journal of Machine Learning Research*, vol. 6, pp. 883–904, 2005.

[19] G. Blanchard, O. Bousquet, and P. Massart, “Statistical performance of Support Vector Machines,” Laboratoire de mathématiques, Université Paris-Sud, Tech. Rep., 2004, under submission, manuscript available at http://ida.first.fraunhofer.de/˜blanchard/publi/BlaBouMas04.ps.gz.

[20] T. Evgeniou, M. Pontil, and T. Poggio, “Regularization networks and Support Vector Machines,” in *Advances in Large Margin Classifiers*, A. J. Smola, P. L. Bartlett, B. Schölkopf, and D. Schuurmans, Eds. Cambridge, MA: MIT Press, 2000, pp. 171–203.

[21] A. J. Smola and B. Schölkopf, “On a kernel-based method for pattern recognition, regression, approximation and operator inversion,” *Algorithmica*, vol. 22, pp. 211–231, 1998.

[22] Baker, *The numerical treatment of integral equations.* Oxford: Clarendon Press, 1977.

[23] B. Schölkopf, A. J. Smola, and K.-R. Müller, “Nonlinear component analysis as a kernel eigenvalue problem,” Max-Planck-Institut für biologische Kybernetik, Tech. Rep. 44, 1996.

[24] V. Koltchinskii, “Asymptotics of spectral projections of some random matrices approximating integral operators,” *Progress in Probability*, vol. 43, pp. 191–227, 1998.

[25] J. Dauxois, A. Pousse, and Y. Romain, “Asymptotic theory for the Principal Component Analysis of a vector random function: some applications to statistical inference,” *Journal of multivariate analysis*, vol. 12, pp. 136–154, 1982.

[26] G. Blanchard and L. Zwald, “On the convergence of eigenspaces in Kernel Principal Component Analysis,” in *Proceedings of the 19th. Neural Information Processing System (NIPS 2005).* MIT Press, 2005.

[27] G. Raetsch, “http://ida.first.gmd.de/˜raetsch/data/benchmarks.htm,” 1999, benchmark repository used in several Boosting, KFD and SVM papers.

[28] N. Bissantz, T. Hohage, A. Munk, and F. Ruymgaart, “Convergence rates of general regularization methods for statistical inverse problems and applications,” Georg-August Universität Göttingen, Tech. Rep. 2006-02, 2006.

[29] V. Vapnik, *Estimation of Dependences Based on Empirical Data*, ser. Springer Series in Statistics. Springer Verlag, 1982.

[30] O. Bousquet, “Concentration inequalities and empirical processes theory applied to the analysis of learning algorithms,” Ph.D. dissertation, Ecole Polytechnique, 2002.

[31] A. W. van der Vaart and J. A. Wellner, *Weak Convergence and Empirical Processes.* Springer, 1996.

[32] R. M. Dudley, *Uniform Central Limit Theorems*, ser. Cambridge Studies in advanced mathematics. Cambridge, U.K.: Cambridge University Press, 1999.

[33] H. Zhu, C. Williams, R. Rohwer, and M. Morciniec, “Gaussian regression and optimal finite dimensional linear models,” *Neural networks and machine learning*, 1998.

[34] C. K. I. Williams and M. Seeger, “The effect of the input density distribution on kernel-based classifiers,” in *Proceedings of the 17th International Conference on Machine Learning*, P. Langley, Ed. San Francisco, California: Morgan Kaufmann, 2000, pp. 1159–1166.