Consistency and Convergence Rates of One-Class SVM and Related Algorithms

Régis Vert
Laboratoire de Recherche en Informatique
Bât. 490, Université Paris-Sud
91405, Orsay Cedex, France
Regis.Vert@lri.fr

Jean-Philippe Vert
Ecole des Mines de Paris
35 rue Saint Honoré
77300 Fontainebleau, France
Jean-Philippe.Vert@ensmp.fr

June 1, 2005

Abstract

We determine the asymptotic limit of the function computed by support vector machines (SVM) and related algorithms that minimize a regularized empirical convex loss function in the reproducing kernel Hilbert space of the Gaussian RBF kernel, in the situation where the number of examples tends to infinity, the bandwidth of the Gaussian kernel tends to 0, and the regularization parameter is held fixed. Non-asymptotic convergence bounds to this limit in the L_2 sense are provided, together with upper bounds on the classification error that is shown to converge to the Bayes risk, therefore proving the Bayes-consistency of a variety of methods although the regularization term does not vanish. These results are particularly relevant to the one-class SVM, for which the regularization can not vanish by construction, and which is shown for the first time to be a consistent density level set estimator.

Keywords: Regularization, Gaussian kernel RKHS, One-class SVM, Convex loss functions, kernel density estimation.

1. Introduction

Given n i.i.d. copies $(X_1, Y_1), \ldots, (X_n, Y_n)$ of a random variable $(X, Y) \in \mathbb{R}^d \times \{-1, 1\}$, we study in this paper the limit and consistency of learning algorithms that solve the following problem:

$$\arg\min_{f \in \mathcal{H}_\sigma} \left\{ \frac{1}{n} \sum_{i=1}^{n} \phi\left(Y_i f(X_i)\right) + \lambda \| f \|^2_{\mathcal{H}_\sigma} \right\} , \tag{1}$$

where $\phi : \mathbb{R} \to \mathbb{R}$ is a convex loss function and $\mathcal{H}_\sigma$ is the reproducing kernel Hilbert space (RKHS) of the normalized Gaussian radial basis function kernel (denoted simply Gaussian kernel below):

$$k_\sigma(x, x') = \frac{1}{\left(\sqrt{2\pi}\sigma\right)^d} \exp\left(\frac{-\| x - x' \|^2}{2\sigma^2} \right) . \tag{2}$$

This framework encompasses in particular the classical support vector machine (SVM) (Boser et al., 1992) when $\phi(u) = \max(1-u, 0)$. Recent years have witnessed important the-

oretical advances aimed at understanding the behavior of such regularized algorithms when n tends to infinity and λ decreases to 0. In particular the consistency and convergence rates of the two-class SVM (see, e.g., Steinwart, 2002; Zhang, 2004; Steinwart and Scovel, 2004, and references therein) have been studied in detail, as well as the shape of the asymptotic decision function (Steinwart, 2003; Bartlett and Tewari, 2004). The case of more general convex loss functions has also attracted a lot of attention recently (Zhang, 2004; Lugosi and Vayatis, 2004; Bartlett et al., 2003), and been shown to provide under general assumptions consistent procedure for the classification error.

All results published so far, however, study the case where λ decreases as the number of points tends to infinity (or, equivalently, where $\lambda\sigma^{-d}$ converges to 0 if one uses the classical non-normalized version of the Gaussian kernel instead of (2)). Although it seems natural to reduce regularization as more and more training data are available — even more than natural, it is the spirit of regularization (Tikhonov and Arsenin, 1977; Silverman, 1982) —, there is at least one important situation where λ is typically held fixed: the one-class SVM (Schölkopf et al., 2001). In that case, the goal is to estimate an α-quantile, that is, a subset of the input space $\mathcal{X}$ of given probability α with minimum volume. The estimation is performed by thresholding the function output by the one-class SVM, that is, the SVM (1) with only positive examples; in that case λ is supposed to determine the quantile level[1]. Although it is known that the fraction of examples in the selected region converges to the desired quantile level α (Schölkopf et al., 2001), it is still an open question whether the region converges towards a quantile, that is, a region of minimum volume. Besides, most theoretical results about the consistency and convergence rates of two-class SVM with vanishing regularization constant do not translate to the one-class case, as we are precisely in the seldom situation where the SVM is used with a regularization term that does not vanish as the sample size increases.

The main contribution of this paper is to show that Bayes consistency for the classification error can be obtained for algorithms that solve (1) without decreasing λ, if instead the bandwidth σ of the Gaussian kernel decreases at a suitable rate. We prove upper bounds on the convergence rate of the classification error towards the Bayes risk for a variety of functions ϕ and of distributions P, in particular for SVM (Theorems 6). Moreover, we provide an explicit description of the function asymptotically output by the algorithms, and establish converge rates towards this limit for the L_2 norm (Theorem 7). In particular, we show that the decision function output by the one-class SVM converges towards the density to be estimated, truncated at the level 2λ (Theorem 8); we finally show (Theorem 9) that this implies the consistency of one-class SVM as a density level estimator for the excess-mass functional (Hartigan, 1987).

This paper is organized as follows. In Section 2, we set the framework of this study and state the main results. The rest of the paper is devoted to the proofs of these results. In Section 3, we provide a number of known and new properties of the Gaussian RKHS. Sections 4 is devoted to the proof of the main theorem that describes the speed of convergence of the regularized ϕ-risk of its empirical minimizer towards its minimum. This proof involves in particular a control of the sample error in this particular setting that is dealt with in Section 5. Section 6 relates the minimization of the regularized ϕ-risk to more

1. While the original formulation of the one-class SVM involves a parameter ν, there is asymptotically a one-to-one correspondance between λ and ν

classical measures of performance, in particular classification error and L_2 distance to the limit. These results are discussed in more detail in Section 7 for the case of the 1- and 2-SVM. Finally the proof of the consistency of the one-class SVM as a density level set estimator is postponed to Section 8.

2. Notations and Main Results

Let (X, Y) be a pair of random variables taking values in $\mathbb{R}^d \times \{-1, 1\}$, with distribution P. We assume throughout this paper that the marginal distribution of X is absolutely continuous with respect to Lebesgue measure with density $\rho : \mathbb{R}^d \to \mathbb{R}$, and that is has a compact support included in a compact set $\mathcal{X} \subset \mathbb{R}^d$. Let $\eta : \mathbb{R}^d \to [0, 1]$ denote a measurable version of the conditional distribution of $Y = 1$ given X, the so-called *regression function*.

The normalized Gaussian radial basis function (RBF) kernel k_σ with bandwidth parameter $\sigma > 0$ is defined for any $(x, x') \in \mathbb{R}^d \times \mathbb{R}^d$ by:

$$k_\sigma(x, x') = \frac{1}{\left(\sqrt{2\pi}\sigma\right)^d} \exp\left(\frac{-\| x - x' \|^2}{2\sigma^2}\right) ,$$

the corresponding reproducing kernel Hilbert space (RKHS) is denoted by $\mathcal{H}_\sigma$, with associted norm $\| \, . \, \|_{\mathcal{H}_\sigma}$. Moreover let $\kappa_\sigma = \| \, k_\sigma \, \|_{L_\infty} = 1/\left(\sqrt{2\pi}\sigma\right)^d$. Several useful properties of this kernel and its RKHS are gathered in Section 3.

Denoting by $\mathcal{M}$ the set of measurable real-valued functions on $\mathbb{R}^d$, we define several risks for functions $f \in \mathcal{M}$:

- The classification error rate, usually refered to as *(true) risk* of f, when Y is predicted by the sign of $f(X)$, is denoted by

  $$R\left(f\right) = P\left(\operatorname{sign}\left(f(X)\right) \neq Y\right) ,$$

  and the minimium achievable classification error rate over $\mathcal{M}$ is called the Bayes risk:

  $$R^* = \inf_{f \in \mathcal{M}} R(f).$$

- For a scalar $\lambda > 0$ fixed throughout this paper and a convex function $\phi : \mathbb{R} \to \mathbb{R}$, the ϕ-risk regularized by the RKHS norm is defined, for any $\sigma > 0$ and $f \in \mathcal{H}_\sigma$, by

  $$R_{\phi,\sigma}\left(f\right) = \mathbb{E}_P\left[\phi\left(Y f\left(X\right)\right)\right] + \lambda \| \, f \, \|^2_{\mathcal{H}_\sigma}$$

  Furthermore, for any real $r \geq 0$, we know that ϕ is Lipschitz on $[-r, r]$, and we denote by $L\left(r\right)$ the Lipschitz constant of the restriction of ϕ to the interval $[-r, r]$. For example, for the hinge loss $\phi(u) = \max(0, 1 - u)$ one can take $L(r) = 1$, and for the squared hinge loss $\phi(u) = \max(0, 1 - u)^2$ one can take $L(r) = 2(r + 1)$.

- Finally, the L_2-norm regularized ϕ-risk is, for any $f \in \mathcal{M}$:

  $$R_{\phi,0}\left(f\right) = \mathbb{E}_P\left[\phi\left(Y f\left(X\right)\right)\right] + \lambda \| \, f \, \|^2_{L_2}$$

  where,

  $$\| \, f \, \|^2_{L_2} = \int_{\mathbb{R}^d} f(x)^2 dx \in [0, +\infty].$$

Each of these risks has an empirical counterpart where the expectation with respect to P is replaced by an average over an i.i.d. sample $T = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$. In particular, the following empirical version of $R_{\phi,\sigma}$ will be used

$$\forall \sigma > 0, f \in \mathcal{H}_\sigma, \quad \widehat{R}_{\phi,\sigma}(f) = \frac{1}{n} \sum_{i=1}^{n} \phi(Y_i f(X_i)) + \lambda \| f \|^2_{\mathcal{H}_\sigma}$$

The main focus of this paper is the analysis of learning algorithms that minimize the empirical ϕ-risk regularized by the RKHS norm $\widehat{R}_{\phi,\sigma}$, and their limit as the number of points tends to infinity and the kernel width σ decreases to 0 at a suitable rate when n tends to ∞, λ being kept fixed. Roughly speaking, our main result shows that in this situation, if ϕ is a convex loss function, the minimization of $\widehat{R}_{\phi,\sigma}$ asymptotically amounts to minimizing $R_{\phi,0}$. This stems from the fact that the empirical average term in the definition of $\widehat{R}_{\phi,\sigma}$ converges to its corresponding expectation, while the norm in $\mathcal{H}_\sigma$ of a function f decreases to its L_2 norm when σ decreases to zero. To turn this intuition into a rigorous statement, we need a few more assumptions about the minimizer of $R_{\phi,0}$ and about P. First, we observe that the minimizer of $R_{\phi,0}$ is indeed well-defined and can often be explicitly computed (the following lemma is part of Theorem 28):

Lemma 1 *For any $x \in \mathbb{R}^d$, let*

$$f_{\phi,0}(x) = \arg\min_{\alpha \in \mathbb{R}} \left\{ \rho(x) \left[\eta(x)\phi(\alpha) + (1 - \eta(x))\phi(-\alpha) \right] + \lambda \alpha^2 \right\} .$$

Then $f_{\phi,0}$ is measurable and satisfies:

$$R_{\phi,0}(f_{\phi,0}) = \inf_{f \in \mathcal{M}} R_{\phi,0}(f)$$

Second, let us recall the notion of modulus of continuity (DeVore and Lorentz, 1993):

Definition 2 (Modulus of continuity) *Let f be a Lebesgue measurable function from $\mathbb{R}^d$ to $\mathbb{R}$. Then its modulus of continuity in the L_1-norm is defined for any $\delta \geq 0$ as follows*

$$\omega(f, \delta) = \sup_{0 \leq \| t \| \leq \delta} \| f(. + t) - f(.) \|_{L_1} , \tag{3}$$

where $\| t \|$ is the Euclidian norm of $t \in \mathbb{R}^d$.

Our main result can now be stated as follows:

Theorem 3 (Main Result) *Let $\sigma_1 > \sigma > 0$, $0 < p < 2$, $\delta > 0$, and let $\hat{f}_{\phi,\sigma}$ denote a minimizer of the $\widehat{R}_{\phi,\sigma}$ risk over $\mathcal{H}_\sigma$, where ϕ is assumed to be convex. Assume that the marginal density ρ is bounded, and let $M = \sup_{x \in \mathbb{R}^d} \rho(x)$. Then there exist constants $(K_i)_{i=1\ldots4}$ (depending only on p, δ, λ, d, and M) such that the following holds with probability greater*

than $1 - e^{-x}$ *over the draw of the training data*

$$
\begin{aligned}
R_{\phi,0}(\hat{f}_{\phi,\sigma}) - R^*_{\phi,0} \leq\ & K_1 L \left(\sqrt{\frac{\kappa_\sigma \phi(0)}{\lambda}} \right)^{\frac{4}{2+p}} \left(\frac{1}{\sigma} \right)^{\frac{[2+(2-p)(1+\delta)]d}{2+p}} \left(\frac{1}{n} \right)^{\frac{2}{2+p}} \\
& + K_2 L \left(\sqrt{\frac{\kappa_\sigma \phi(0)}{\lambda}} \right)^{2} \left(\frac{1}{\sigma} \right)^{d} \frac{x}{n} \\
& + K_3 \frac{\sigma^2}{\sigma_1^2} \\
& + K_4 \omega(f_{\phi,0}, \sigma_1) \ .
\end{aligned} \tag{4}
$$

The first two terms in r.h.s. of (4) bound the estimation error (also called sample error) associated with the Gaussian RKHS, which naturally tends to be small when the number of training data increases and when the RKHS is 'small', i.e., when σ is large. As is usually the case in such variance/bias splitings, the variance term here depends on the dimension d of the input space. Note that it is also parametrized by both p and δ. The third term measures the error due to penalizing the L_2-norm of a fixed function in $\mathcal{H}_{\sigma_1}$ by its $\| \, . \, \|_{\mathcal{H}_\sigma}$-norm, with $0 < \sigma < \sigma_1$. This is a price to pay to get a small estimation error. As for the fourth term, it is a bound on the approximation error of the Gaussian RKHS. Note that, once λ and σ have been fixed, σ_1 remains a free variable parameterizing the bound itself.

In order to highlight the type of convergence rates one can obtain from Theorem 3, let us assume that the ϕ loss function is Lipschitz on $\mathbb{R}$ (e.g., take the hinge loss), and suppose that for some $0 \leq \beta \leq 1$, $c_1 > 0$, and for any $h \geq 0$, the density function ρ satisfies the following inequality

$$
\omega(\rho, h) \leq c_1 h^\beta \ . \tag{5}
$$

Then we can optimize the right hand side of (4) w.r.t. σ_1, σ, p and δ by balancing the four terms. This eventually leads to:

$$
R_{\phi,0}\left(\hat{f}_{\phi,\sigma}\right) - R^*_{\phi,0} = O_P\left(\left(\frac{1}{n} \right)^{\frac{2\beta}{4\beta+(2+\beta)d} - \epsilon} \right) \ , \tag{6}
$$

for any $\epsilon > 0$. This rate is achieved by choosing

$$
\sigma_1 = \left(\frac{1}{n} \right)^{\frac{2}{4\beta+(2+\beta)d} - \frac{\epsilon}{\beta}} \ , \tag{7}
$$

$$
\sigma = \sigma_1^{\frac{2+\beta}{2}} = \left(\frac{1}{n} \right)^{\frac{2+\beta}{4\beta+(2+\beta)d} - \frac{\epsilon(2+\beta)}{2\beta}} \ , \tag{8}
$$

$p = 2$ and δ as small as possible (that is why an arbitray small quantity ϵ appears in the rate).

Theorem 3 shows that, when ϕ is convex, minimizing the $\widehat{R}_{\phi,\sigma}$ risk for well-chosen width σ is a an algorithm consistant for the $R_{\phi,0}$-risk. In order to relate this consistency with more traditional measures of performance of learning algorithms, the next theorem shows that under a simple additionnal condition on ϕ, $R_{\phi,0}$-risk-consistency implies Bayes consistency:

Theorem 4 *If ϕ is convex, differentiable at 0, with $\phi'(0) < 0$, then for every sequence of functions $(f_i)_{i\geq 1} \in \mathcal{M}$,*

$$\lim_{n\to+\infty} R_{\phi,0}(f_i) = R^*_{\phi,0} \implies \lim_{n\to+\infty} R(f_i) = R^*$$

This theorem results from a more general quantitative analysis of the relationship between the excess $R_{\phi,0}$-risk and the excess R-risk (see Theorem 30). In order to state a refined version of it in the particular case of the support vector machine algorithm, we first need to introduce the notion of *low density exponent*:

Definition 5 *We say that a distribution P with ρ as marginal density of X w.r.t. Lebesgue measure has a* low density exponent $\gamma \geq 0$ *if there exists $(c_2, \epsilon_0) \in (0, +\infty)^2$ such that*

$$\forall \epsilon \in [0, \epsilon_0], \quad P\left(\left\{x \in \mathbb{R}^d : \rho(x) \leq \epsilon\right\}\right) \leq c_2 \epsilon^\gamma.$$

We are now in position to state a quantitative relationship between the excess $R_{\phi,0}$-risk and the excess R-risk in the case of support vector machines:

Theorem 6 *Let $\phi_1(\alpha) = \max(1-\alpha, 0)$ be the hinge loss function, and $\phi_2(\alpha) = \max(1-\alpha, 0)^2$, be the squared hinge loss function. Then for any distribution P with low density exponent γ, there exist constant $(K_1, K_2, r_1, r_2) \in (0, +\infty)^4$ such that for any $f \in \mathcal{M}$ with an excess $R_{\phi_1,0}$-risk upper bounded by r_1 the following holds:*

$$R(f) - R^* \leq K_1 \left(R_{\phi_1,0}(f) - R^*_{\phi_1,0}\right)^{\frac{\gamma}{2\gamma+1}},$$

and if the excess regularized $R_{\phi_2,0}$-risk upper bounded by r_2 the following holds:

$$R(f) - R^* \leq K_2 \left(R_{\phi_2,2}(f) - R^*_{\phi_2,2}\right)^{\frac{\gamma}{2\gamma+1}},$$

We note that Theorem 32 generalizes this result to any loss function through the introduction of variational arguments, in the spirit of Bartlett et al. (2003). Hence the consistency of SVM is prooved, together with upper bounds on the convergence rates, for the first time in a situation where the effect of regularization does not vanish asymptotically.

Another consequence of the $R_{\phi,0}$-consistency of an algorithm is the L_2convergence of the function output by the algorithm to the minimizer of the $R_{\phi,0}$-risk:

Lemma 7 *For any $f \in \mathcal{M}$, the following holds:*

$$\| f - f_{\phi,0} \|^2_{L_2} \leq \frac{1}{\lambda}\left(R_{\phi,0}(f) - R^*_{\phi,0}\right).$$

This result is particularly relevant to study algorithms whose objective are not binary classification. Consider for example the 1-class SVM algorithm, which served as the initial motivation for this paper. Then we can state the following

Theorem 8 *Let ρ_λ denote the density truncated as follows:*

$$\rho_\lambda(x) = \begin{cases} \frac{\rho(x)}{2\lambda} & \text{if } \rho(x) \leq 2\lambda, \\ 1 & \text{otherwise.} \end{cases} \tag{9}$$

Let $\hat{f}_\sigma$ denote the function output by the one-class SVM. Then, under the general conditions of Theorem 3,

$$\lim_{n\to+\infty} \| \hat{f}_\sigma - \rho_\lambda \|_{L_2} = 0 \ ,$$

for σ choosen as in Equation (8).

A very interesting by-product of this theorem is the consistency of the one-class SVM algorithm for density level set estimation, which to the best of our knowledge has not been stated so far:

Theorem 9 *Let $0 < \mu < 2\lambda < M$, let C_μ be the level set of the density function ρ at level μ, and $\widehat{C}_\mu$ be the level set of $2\lambda\hat{f}_\sigma$ at level μ, where $\hat{f}_\sigma$ is still the function outptut by the one-class SVM. For any distribution Q, for any subset C of $\mathbb{R}^d$, define the excess-mass of C with respect to Q as follows:*

$$H_Q(C) = Q(C) - \mu Leb(C) \ .$$

Then, under the general assumptions of Theorem 3, we have

$$\lim_{n\to+\infty} H_P(C_\mu) - H_P\left(\widehat{C}_\mu\right) = 0 \ ,$$

for σ choosen as in Equation (8).

The excess-mass functional was first introduced by Hartigan (1987) to assess the quality of density level set estimators. It is maximized by the true density level set C_μ and acts as a risk functional in the one-class framework. The proof of Theorem 9 is based on the following general result: if $\hat{\rho}$ is a density estimator converging to the true density ρ in the L_2 sense, then for any fixed $0 < \mu < \sup_{\mathbb{R}^d}\{\rho\}$, the excess mass of the level set of $\hat{\rho}$ at level μ converges to the excess mass of C_μ. In other words, as is the case in the classification framework, plug-in rules built on L_2-consistent density estimators are consistent with respect to the excess mass.

3. Some Properties of the Gaussian kernel and its RKHS

This section presents known and new results about the Gaussian kernel and its associated RKHS, that are useful for the proofs of our results. They concern the explicit description of the RKHS norm in terms of Fourier transforms, its relation with the L_2-norm, and some approximation properties of convolutions with the Gaussian kernel. They make use of basic properties of Fourier transforms which we now recall.
For any f in $L_1(\mathbb{R}^d)$, its Fourier transform $\mathcal{F}[f] : \mathbb{R}^d \to \mathbb{R}$ is defined by

$$\mathcal{F}[f](\omega) = \int_{\mathbb{R}^d} e^{-i<x,\omega>} f(x)dx \ .$$

If in addition $\mathcal{F}[f] \in L_1(\mathbb{R}^d)$, f can be recovered from $\mathcal{F}[f]$ by the inverse Fourier formula:

$$f(x) = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \mathcal{F}[f](\omega) e^{i<x,\omega>} d\omega \ .$$

Finally Parseval's equality relates the L_2-norm of a function and its Fourier transform if $f \in L_1(\mathbb{R}^d) \cap L_2(\mathbb{R}^d)$ and $\mathcal{F}[f] \in L_2(\mathbb{R}^d)$:

$$\| f \|_{L_2}^2 = \frac{1}{(2\pi)^d} \| \mathcal{F}[f] \|_{L_2}^2 . \tag{10}$$

3.1 Fourier Representation of the Gaussian RKHS

The normalized Gaussian radial basis function kernel k_σ with bandwidth parameter $\sigma > 0$, more simply refered to as Gaussian kernel in the rest of this paper, is defined on $\mathbb{R}^d \times \mathbb{R}^d$ by

$$k_\sigma(x, x') = \frac{1}{\left(\sqrt{2\pi}\sigma\right)^d} \exp\left(\frac{-\| x - x' \|^2}{2\sigma^2}\right) . \tag{11}$$

The normalizing constant

$$\kappa_\sigma = \| k_\sigma \|_{L_\infty} = \left(\sqrt{2\pi}\sigma\right)^{-d} , \tag{12}$$

ensures that the kernel integrates to 1 for any $\sigma > 0$. For any $u \in \mathbb{R}^d$, the expression $k_\sigma(u)$ denotes $k_\sigma(0, u)$, with Fourier transform known to be:

$$\mathcal{F}[k_\sigma](\omega) = e^{\frac{-\sigma^2 \| \omega \|^2}{2}} . \tag{13}$$

Let $\mathcal{H}_\sigma$ denote the RKHS associated with the gaussian kernel k_σ. The general study of translation invariant kernels provides an accurate characterization of their associated RKHS in terms of the their Fourier transform (see, e.g., Matache and Matache, 2002). In the case of the Gaussian kernel, the following holds :

Lemma 10 (Characterization of $\mathcal{H}_\sigma$) *Let $\mathcal{C}_0(\mathbb{R}^d)$ denote the set of continuous functions on $\mathbb{R}^d$ that vanish at infinity. The set*

$$\mathcal{H}_\sigma = \left\{ f \in \mathcal{C}_0(\mathbb{R}^d) \ : \ f \in L_1(\mathbb{R}^d) \text{ and } \int_{\mathbb{R}^d} |\mathcal{F}[f](\omega)|^2 e^{\frac{\sigma^2 \| \omega \|^2}{2}} d\omega < \infty \right\} \tag{14}$$

is the RKHS associated with the gaussian kernel k_σ, and the associated dot product is given for any $f, g \in \mathcal{H}_\sigma$ by

$$< f, g >_{\mathcal{H}_\sigma} = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \mathcal{F}[f](\omega) \mathcal{F}[g](\omega)^* e^{\frac{\sigma^2 \| \omega \|^2}{2}} d\omega , \tag{15}$$

where a^* denotes the conjugate of a complex number a. In particular the associated norm is given for any $f \in \mathcal{H}_\sigma$ by

$$\| f \|_{\mathcal{H}_\sigma}^2 = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} |\mathcal{F}[f](\omega)|^2 e^{\frac{\sigma^2 \| \omega \|^2}{2}} d\omega . \tag{16}$$

This lemma readily implies several basic facts about Gaussian RKHS and their associated norms summarized in the next lemma. In particular, it shows that the family $(\mathcal{H}_\sigma)_{\sigma > 0}$ forms a nested collection of models, and that for any fixed function, the RKHS norm decreases to the L_2-norm as the kernel bandwidth decreases to 0:

Lemma 11 *The following statements hold:*

1. *For any* $0 < \tau < \sigma$,

$$\mathcal{H}_\sigma \subset \mathcal{H}_\tau \subset L_2(\mathbb{R}^d) \; . \tag{17}$$

Moreover, for any $f \in \mathcal{H}_\sigma$,

$$\| f \|_{\mathcal{H}_\sigma} \geq \| f \|_{\mathcal{H}_\tau} \geq \| f \|_{L_2} \tag{18}$$

and

$$0 \leq \| f \|^2_{\mathcal{H}_\tau} - \| f \|^2_{L_2} \leq \frac{\tau^2}{\sigma^2} \left(\| f \|^2_{\mathcal{H}_\sigma} - \| f \|^2_{L_2} \right) \; . \tag{19}$$

2. *For any* $\sigma > 0$ *and* $f \in \mathcal{H}_\sigma$,

$$\lim_{\tau \to 0} \| f \|_{\mathcal{H}_\tau} = \| f \|_{L_2} \; . \tag{20}$$

3. *For any* $\sigma > 0$ *and* $f \in \mathcal{H}_\sigma$,

$$\| f \|_{L_\infty} \leq \sqrt{\kappa_\sigma} \| f \|_{\mathcal{H}_\sigma} \; . \tag{21}$$

Proof Equations (17) and (18) are direct consequences of the characterization of the Gaussian RKHS (16) and of the observation that

$$0 < \tau < \sigma \implies e^{\frac{\sigma^2 \| \omega \|^2}{2}} \geq e^{\frac{\tau^2 \| \omega \|^2}{2}} \geq 1.$$

In order to prove (19), we derive from (16) and Parseval's equality (10):

$$\| f \|^2_{\mathcal{H}_\tau} - \| f \|^2_{L_2} = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} |\mathcal{F}[f](\omega)|^2 \left[e^{\frac{\tau^2 \| \omega \|^2}{2}} - 1 \right] d\omega \; . \tag{22}$$

For any $0 \leq u \leq v$, we have $(e^u - 1)/u \leq (e^v - 1)/v$ by convexity of e^u, and therefore:

$$\| f \|^2_{\mathcal{H}_\tau} - \| f \|^2_{L_2} \leq \frac{1}{(2\pi)^d} \frac{\tau^2}{\sigma^2} \int_{\mathbb{R}^d} |\mathcal{F}[f](\omega)|^2 \left[e^{\frac{\sigma^2 \| \omega \|^2}{2}} - 1 \right] d\omega, \tag{23}$$

which leads to (19). Equation (20) is now a direct consequence of (19). Finally, (21) is a classical bound derived from the observation that, for any $x \in \mathbb{R}^d$,

$$\begin{aligned} | f(x) | &= | < f, k_\sigma >_{\mathcal{H}_\sigma} | \\ &\leq \| f \|_{\mathcal{H}_\sigma} \| k_\sigma \|_{\mathcal{H}_\sigma} \\ &= \sqrt{\kappa_\sigma} \| f \|_{\mathcal{H}_\sigma} \; . \end{aligned}$$

■

3.2 Links with the Non-Normalized Gaussian Kernel

It is common in the machine learning literature to work with a non-normalized version of the Gaussian RBF kernel, namely the kernel:

$$\tilde{k}_\sigma(x, x') = \exp\left(\frac{-\| x - x' \|^2}{2\sigma^2}\right) . \tag{24}$$

From the relation $k_\sigma = \kappa_\sigma \tilde{k}_\sigma$ (remember that κ_σ is defined in Equation 12), we deduce from the general theory of RKHS that $\mathcal{H}_\sigma = \tilde{\mathcal{H}}_\sigma$ and

$$\forall f \in \mathcal{H}_\sigma, \quad \| f \|_{\tilde{\mathcal{H}}_\sigma} = \sqrt{\kappa_\sigma} \| f \|_{\mathcal{H}_\sigma} . \tag{25}$$

As a result, all statements about k_σ and its RKHS easily translate into statements about $\tilde{k}_\sigma$ and its RKHS. For example, (18) shows that, for any $0 < \tau < \sigma$ and $f \in \tilde{\mathcal{H}}_\sigma$,

$$\| f \|_{\tilde{\mathcal{H}}_\sigma} \geq \sqrt{\frac{\kappa_\sigma}{\kappa_\tau}} \| f \|_{\tilde{\mathcal{H}}_\tau} = \left(\frac{\tau}{\sigma}\right)^{\frac{d}{2}} \| f \|_{\tilde{\mathcal{H}}_\tau} ,$$

a result that was shown recently (Steinwart et al., 2004, Corollary 3.12).

3.3 Convolution with the Gaussian kernel

Besides its positive definiteness, the Gaussian kernel is commonly used as a kernel for function approximation through convolution. Recall that the convolution between two functions $f, g \in L_1\left(\mathbb{R}^d\right)$ is the function $f * g \in L_1\left(\mathbb{R}^d\right)$ defined by

$$f * g\,(x) = \int_{\mathbb{R}^d} f\,(x - u)\, g\,(u)\, du$$

and that it satisfies

$$\mathcal{F}\,[f * g] = \mathcal{F}\,[f]\,\mathcal{F}\,[g] \ . \tag{26}$$

The convolution with a Gaussian RBF kernel is a technically convenient tool to map any square integrable function to a Gaussian RKHS, as the following lemma shows:

Lemma 12 *For any $\sigma > 0$ and any $f \in L_1\left(\mathbb{R}^d\right) \cap L_2\left(\mathbb{R}^d\right)$,*

$$k_\sigma * f \in \mathcal{H}_{\sqrt{2}\sigma}$$

and

$$\| k_\sigma * f \|_{\mathcal{H}_{\sqrt{2}\sigma}} = \| f \|_{L_2} . \tag{27}$$

Proof Using (16), then (26) and (13), followed by Parseval's inequality (10), we compute:

$$\begin{aligned}
\| k_\sigma * f \|^2_{\mathcal{H}_{\sqrt{2}\sigma}} &= \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} |\mathcal{F}\,[k_\sigma * f]\,(\omega)|^2 e^{\sigma^2 \| \omega \|^2} d\omega \\
&= \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} |\mathcal{F}\,[f]\,(\omega)|^2 e^{-\sigma^2 \| \omega \|^2} e^{\sigma^2 \| \omega \|^2} d\omega \\
&= \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} |\mathcal{F}\,[f]\,(\omega)|^2 d\omega \\
&= \| f \|^2_{L_2} .
\end{aligned}$$

■

The following technical lemma, used in the proof of the main theorem below, upper bounds the difference between a RKHS norm and a L_2 norm of a function smoothed by convolution:

Lemma 13 *For any $\sigma, \tau > 0$ that satisfy $0 < \sigma < \sqrt{2}\tau$, and for any $f \in L_1\left(\mathbb{R}^d\right) \cap L_2\left(\mathbb{R}^d\right)$,*

$$k_\tau * f \in \mathcal{H}_\sigma$$

and

$$\| k_\tau * f \|^2_{\mathcal{H}_\sigma} - \| k_\tau * f \|^2_{L_2} \leq \frac{\sigma^2}{2\tau^2} \| f \|^2_{L_2} . \tag{28}$$

Proof Because $0 < \sigma < \sqrt{2}\tau$, Lemma 12 and (17) imply

$$k_\tau * f \in \mathcal{H}_{\sqrt{2}\tau} \subset \mathcal{H}_\sigma ,$$

and, using (19) and (27),

$$\begin{aligned} \| k_\tau * f \|^2_{\mathcal{H}_\sigma} - \| k_\tau * f \|^2_{L_2} &\leq \frac{\sigma^2}{2\tau^2} \left(\| k_\tau * f \|^2_{\mathcal{H}_{\sqrt{2}\tau}} - \| k_\tau * f \|^2_{L_2} \right) \\ &\leq \frac{\sigma^2}{2\tau^2} \| k_\tau * f \|^2_{\mathcal{H}_{\sqrt{2}\tau}} \\ &= \frac{\sigma^2}{2\tau^2} \| f \|^2_{L_2} . \end{aligned}$$

■

A final result we need is an estimate of the approximation properties of convolution with the Gaussian kernel. Convolving a function with a Gaussian kernel with decreasing bandwidth is known to provide an approximation of the original function under general conditions. For example, the assumption $f \in L_1(\mathbb{R}^d)$ is sufficient to show that $\| k_\sigma * f - f \|_{L_1}$ goes to zero when σ goes to zero (see, for example Devroye and Lugosi, 2000). We provide below a more quantitative estimate for the rate of this convergence under some assumption on the modulus of continuity of f (see Definition 2), using methods from DeVore and Lorentz (1993, Section 7.2).

Lemma 14 *Let f be a bounded function in $L_1(\mathbb{R}^d)$. Then for all $\sigma > 0$, the following holds:*

$$\| k_\sigma * f - f \|_{L_1} \leq (1 + \sqrt{d})\omega(f, \sigma) ,$$

where $\omega(f, .)$ denotes the modulus of continuity of f in the L_1 norm.

Proof Using the fact that k_σ is normalized, then Fubini's theorem and then the definition of ω, the following can be derived

$$\begin{aligned}
\| k_\sigma * f - f \|_{L_1} &= \int_{\mathbb{R}^d} \left| \int_{\mathbb{R}^d} k_\sigma(t)[f(x+t) - f(x)]dt \right| dx \\
&\leq \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} k_\sigma(t) \left| f(x+t) - f(x) \right| dt dx \\
&= \int_{\mathbb{R}^d} k_\sigma(t) \left[\int_{\mathbb{R}^d} \left| f(x+t) - f(x) \right| dx \right] dt \\
&\leq \int_{\mathbb{R}^d} k_\sigma(t) \| f(.+t) - f(.) \|_{L_1} dt \\
&\leq \int_{\mathbb{R}^d} k_\sigma(t) \omega(f, \| t \|) dt \; .
\end{aligned}$$

Now, using the subadditivity property of ω (DeVore and Lorentz, 1993, Section 2.6), the following inequality can be derived for any non-negative λ and δ

$$\omega(f, \lambda\delta) \leq (1+\lambda)\omega(f, \delta) \; .$$

Applying this and also Hölder's inequality leads to

$$\begin{aligned}
\| k_\sigma * f - f \|_{L_1} &\leq \int_{\mathbb{R}^d} \left(1 + \frac{\| t \|}{\sigma} \right) \omega(f, \sigma) k_\sigma(t) dt \\
&= \omega(f, \sigma) \left[1 + \frac{1}{\sigma} \int_{\mathbb{R}^d} \| t \| k_\sigma(t) dt \right] \\
&\leq \omega(f, \sigma) \left[1 + \frac{1}{\sigma} \left(\int_{\mathbb{R}^d} \| t \|^2 \frac{1}{(\sqrt{2\pi}\sigma)^d} e^{\frac{-\| t \|^2}{2\sigma^2}} dt \right)^{\frac{1}{2}} \right] \\
&= \omega(f, \sigma) \left[1 + \frac{1}{\sigma} \left(\sum_{i=1}^{d} \int_{\mathbb{R}^d} t_i^2 \frac{1}{(\sqrt{2\pi}\sigma)^d} e^{\frac{-\| t \|^2}{2\sigma^2}} dt \right)^{\frac{1}{2}} \right] \\
&= \omega(f, \sigma) \left[1 + \frac{1}{\sigma} \left(\sum_{i=1}^{d} \int_{\mathbb{R}^d} t_i^2 \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-t_i^2}{2\sigma^2}} dt_i \right)^{\frac{1}{2}} \right] \\
&= \omega(f, \sigma) \left[1 + \frac{1}{\sigma} \sqrt{d} \left(\int_{\mathbb{R}^d} u^2 \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-u^2}{2\sigma^2}} du \right)^{\frac{1}{2}} \right] \; .
\end{aligned}$$

The integral term is exactly the variance of a gaussian random variable, namely σ^2. Hence we end up with

$$\| k_\sigma * f - f \|_{L_1} \leq (1 + \sqrt{d})\omega(f, \sigma) \; .$$

■

4. Proof of Theorem 3

The proof of Theorem 3 is based on the following decomposition of the excess $R_{\phi,0}$-risk for the minimizer of the $\widehat{R}_{\phi,\sigma}$-risk:

Lemma 15 *For any $0 < \sigma < \sqrt{2}\sigma_1$ and any sample $(x_i, y_i)_{i=1,\dots,n}$, the minimizer $\hat{f}_{\phi,\sigma}$ of $\widehat{R}_{\phi,\sigma}$ satisfies:*

$$\begin{aligned} R_{\phi,0}(\hat{f}_{\phi,\sigma}) - R^*_{\phi,0} \leq & \left[R_{\phi,\sigma}(\hat{f}_{\phi,\sigma}) - R^*_{\phi,\sigma}\right] \\ & + \left[R_{\phi,\sigma}(k_{\sigma_1} * f_{\phi,0}) - R_{\phi,0}(k_{\sigma_1} * f_{\phi,0})\right] \\ & + \left[R_{\phi,0}(k_{\sigma_1} * f_{\phi,0}) - R^*_{\phi,0}\right] \end{aligned} \tag{29}$$

Proof The excess $R_{\phi,0}$ risk decomposes as follows:

$$\begin{aligned} R_{\phi,0}(\hat{f}_{\phi,\sigma}) - R^*_{\phi,0} \quad = \quad & \left[R_{\phi,0}\left(\hat{f}_{\phi,\sigma}\right) - R_{\phi,\sigma}\left(\hat{f}_{\phi,\sigma}\right)\right] \\ & + \left[R_{\phi,\sigma}(\hat{f}_{\phi,\sigma}) - R^*_{\phi,\sigma}\right] \\ & + \left[R^*_{\phi,\sigma} - R_{\phi,\sigma}(k_{\sigma_1} * f_{\phi,0})\right] \\ & + \left[R_{\phi,\sigma}(k_{\sigma_1} * f_{\phi,0}) - R_{\phi,0}(k_{\sigma_1} * f_{\phi,0})\right] \\ & + \left[R_{\phi,0}(k_{\sigma_1} * f_{\phi,0}) - R^*_{\phi,0}\right] \ . \end{aligned}$$

Note that by Lemma 12, $k_{\sigma_1} * f_{\phi,0} \in \mathcal{H}_{\sqrt{2}\sigma_1} \subset \mathcal{H}_\sigma \subset L_2\left(\mathbb{R}^d\right)$ which justifies the introduction of $R_{\phi,\sigma}(k_{\sigma_1} * f_{\phi,0})$ and $R_{\phi,0}(k_{\sigma_1} * f_{\phi,0})$. Now, by definition of the different risks using Equation 18, we have

$$R_{\phi,0}\left(\hat{f}_{\phi,\sigma}\right) - R_{\phi,\sigma}\left(\hat{f}_{\phi,\sigma}\right) = \lambda\left(\| \hat{f}_{\phi,\sigma} \|^2_{L_2} - \| \hat{f}_{\phi,\sigma} \|^2_{\mathcal{H}_\sigma}\right) \leq 0,$$

and

$$R^*_{\phi,\sigma} - R_{\phi,\sigma}(k_{\sigma_1} * f_{\phi,0}) \leq 0.$$

■

Hence, controlling $R_{\phi,0}(\hat{f}_{\phi,\sigma}) - R^*_{\phi,0}$ boils down to controlling each of the three terms arising from the previous split:

- The first term in (29) is usually referred to as the sample error or estimation error. The control of such quantities has been the topic of much research recently, including for example Tsybakov (1997); Mammen and Tsybakov (1999); Massart (2000); Bartlett et al. (2005); Koltchinskii (2003); Steinwart and Scovel (2004). Using estimates of local Rademacher complexities through covering numbers for the Gaussian RKHS due to Steinwart and Scovel (2004), we prove below the following result

  Lemma 16 *For any $\sigma > 0$ small enough, let $\hat{f}_{\phi,\sigma}$ be the minimizer of the $\widehat{R}_{\phi,\sigma}$-risk on a sample of size n, where ϕ is a convex loss function. For any $0 < p \leq 2$, $\delta > 0$,*

and $x \geq 1$, the following holds with probability at least $1 - e^x$ over the draw of the sample:

$$R_{\phi,\sigma}(\hat{f}_{\phi,\sigma}) - R_{\phi,\sigma}(f_{\phi,\sigma}) \leq K_1 L \left(\sqrt{\frac{\kappa_\sigma \phi(0)}{\lambda}}\right)^{\frac{4}{2+p}} \left(\frac{1}{\sigma}\right)^{\frac{[2+(2-p)(1+\delta)]d}{2+p}} \left(\frac{1}{n}\right)^{\frac{2}{2+p}} + K_2 L \left(\sqrt{\frac{\kappa_\sigma \phi(0)}{\lambda}}\right)^2 \left(\frac{1}{\sigma}\right)^d \frac{x}{n},$$

where K_1 and K_2 are positive constants depending neither on σ, nor on n.

- The second term in (29) can be upper bounded by

$$\frac{\phi(0)\,\sigma^2}{2\lambda\sigma_1^2}.$$

Indeed, using Lemma 13, we have

$$\begin{aligned} R_{\phi,\sigma}(k_{\sigma_1} * f_{\phi,0}) - R_{\phi,0}(k_{\sigma_1} * f_{\phi,0}) &= \| k_{\sigma_1} * f_{\phi,0} \|_{\mathcal{H}_\sigma}^2 - \| k_{\sigma_1} * f_{\phi,0} \|_{L_2}^2 \\ &\leq \frac{\sigma^2}{2\sigma_1^2} \| f_{\phi,0} \|_{L_2}^2. \end{aligned}$$

Since $f_{\phi,0}$ minimizes $R_{\phi,0}$, we have $R_{\phi,0}(f_{\phi,0}) \leq R_{\phi,0}(0)$, which leads to $\| f_{\phi,0} \|_{L_2}^2 \leq \phi(0)/\lambda$. Eventually, we have

$$R_{\phi,\sigma}(k_{\sigma_1} * f_{\phi,0}) - R_{\phi,0}(k_{\sigma_1} * f_{\phi,0}) \leq \frac{\phi(0)\,\sigma^2}{2\lambda\sigma_1^2}.$$

- The third term in (29) can be upper bounded by

$$\left(2\lambda \| f_{\phi,0} \|_{L_\infty} + L(\| f_{\phi,0} \|_{L_\infty}) M\right) \left(1 + \sqrt{d}\right) \omega(f_{\phi,0}, \sigma_1).$$

Indeed,

$$\begin{aligned} & R_{\phi,0}(k_{\sigma_1} * f_{\phi,0}) - R_{\phi,0}(f_{\phi,0}) \\ = \;& \lambda \left[\| k_{\sigma_1} * f_{\phi,0} \|_{L_2}^2 - \| f_{\phi,0} \|_{L_2}^2 \right] + \left[\mathbb{E}_P \left[\phi\left(Y(k_{\sigma_1} * f_{\phi,0})(X)\right) \right] - \mathbb{E}_P \left[\phi\left(Y f_{\phi,0}(X)\right) \right] \right] \\ = \;& \lambda < k_{\sigma_1} * f_{\phi,0} - f_{\phi,0}, k_{\sigma_1} * f_{\phi,0} + f_{\phi,0} >_{L_2} + \mathbb{E}_P \left[\phi\left(Y(k_{\sigma_1} * f_{\phi,0})(X)\right) - \phi\left(Y f_{\phi,0}(X)\right) \right]. \end{aligned}$$

Now, since $\| k_{\sigma_1} * f_{\phi,0} \|_{L_\infty} \leq \| f_{\phi,0} \|_{L_\infty}$, then using Lemma 14, we obtain:

$$\begin{aligned} R_{\phi,0}(k_{\sigma_1} * f_{\phi,0}) - R_{\phi,0}(f_{\phi,0}) &\leq 2\lambda \| f_{\phi,0} \|_{L_\infty} \| k_{\sigma_1} * f_{\phi,0} - f_{\phi,0} \|_{L_1} \\ &\quad + L(\| f_{\phi,0} \|_{L_\infty}) \mathbb{E}_P \left[|(k_{\sigma_1} * f_{\phi,0})(X) - f_{\phi,0}(X)| \right] \\ &\leq \left(2\lambda \| f_{\phi,0} \|_{L_\infty} + L(\| f_{\phi,0} \|_{L_\infty}) M\right) \| k_{\sigma_1} * f_{\phi,0} - f_{\phi,0} \|_{L_1} \\ &\leq \left(2\lambda \| f_{\phi,0} \|_{L_\infty} + L(\| f_{\phi,0} \|_{L_\infty}) M\right) \left(1 + \sqrt{d}\right) \omega(f_{\phi,0}, \sigma_1), \end{aligned}$$

where $M = \sup_{x \in \mathbb{R}^d} p(x)$ is supposed to be finite.

Now, Theorem 3 is proved by plugging the last three bounds in Equation 29.

5. Proof of Lemma 16 (sample error)

In order to upper bound the sample error, it is useful to work with a set of functions as "small" as possible, in a meaning made rigorous below. Although we study algorithms that work on the whole RKHS $\mathcal{H}_\sigma$ *a priori*, let us first show that we can drastically "downsize" it.

Indeed, recall that the marginal distribution of P in X is assumed to have a support included in a compact $\mathcal{X} \subset \mathbb{R}^d$. The restriction of k_σ to $\mathcal{X}$, denoted $k_\sigma^{\mathcal{X}}$, is a positive definite kernel on $\mathcal{X}$ (Aronszajn, 1950) with RKHS defined by:

$$\mathcal{H}_\sigma^{\mathcal{X}} = \left\{ f_{/\mathcal{X}} : f \in \mathcal{H}_\sigma \right\} \ , \tag{30}$$

where $f_{/\mathcal{X}}$ denotes the restriction of f to $\mathcal{X}$, and RKHS norm:

$$\forall f^{\mathcal{X}} \in \mathcal{H}_\sigma^{\mathcal{X}}, \quad \| f \|_{\mathcal{H}_\sigma^{\mathcal{X}}} = \inf \left\{ \| f \|_{\mathcal{H}_\sigma} : f \in \mathcal{H}_\sigma \text{ and } f_{/\mathcal{X}} = f^{\mathcal{X}} \right\} \ . \tag{31}$$

For any $f^{\mathcal{X}} \in \mathcal{H}_\sigma^{\mathcal{X}}$ consider the following risks:

$$R_{\phi,\sigma}^{\mathcal{X}}(f^{\mathcal{X}}) = \mathbb{E}_{P_{/\mathcal{X}}} \left[\phi \left(Y f^{\mathcal{X}}(X) \right) \right] + \lambda \| f^{\mathcal{X}} \|_{\mathcal{H}_\sigma^{\mathcal{X}}}^2 \ ,$$
$$\widehat{R}_{\phi,\sigma}^{\mathcal{X}}(f^{\mathcal{X}}) = \frac{1}{n} \sum_{i=1}^{n} \phi \left(Y_i f^{\mathcal{X}}(X_i) \right) + \lambda \| f^{\mathcal{X}} \|_{\mathcal{H}_\sigma^{\mathcal{X}}}^2 \ .$$

We first show that the sample error is the same in $\mathcal{H}_\sigma$ and $\mathcal{H}_\sigma^{\mathcal{X}}$:

Lemma 17 *Let $f_{\phi,\sigma}^{\mathcal{X}}$ and $\hat{f}_{\phi,\sigma}^{\mathcal{X}}$ be respectively the minimizers of $R_{\phi,\sigma}^{\mathcal{X}}$ and $\widehat{R}_{\phi,\sigma}^{\mathcal{X}}$. Then it holds almost surely that*

$$R_{\phi,\sigma} \left(f_{\phi,\sigma} \right) = R_{\phi,\sigma}^{\mathcal{X}} \left(f_{\phi,\sigma}^{\mathcal{X}} \right) \ ,$$
$$R_{\phi,\sigma} \left(\hat{f}_{\phi,\sigma} \right) = R_{\phi,\sigma}^{\mathcal{X}} \left(\hat{f}_{\phi,\sigma}^{\mathcal{X}} \right) \ .$$

From Lemma 17 we deduce that a.s.,

$$R_{\phi,\sigma} \left(\hat{f}_{\phi,\sigma} \right) - R_{\phi,\sigma} \left(f_{\phi,\sigma} \right) = R_{\phi,\sigma}^{\mathcal{X}} \left(\hat{f}_{\phi,\sigma}^{\mathcal{X}} \right) - R_{\phi,\sigma}^{\mathcal{X}} \left(f_{\phi,\sigma}^{\mathcal{X}} \right) \ , \tag{32}$$

In order to upper bound this term, we use concentration inequalities based on local Rademacher complexities (Bartlett et al., 2003, 2005; Steinwart and Scovel, 2004). In this approach, a crucial role is played by the covering number of a functional class $\mathcal{F}$ under the empirical L_2-norm. Remember that for a given sample $T = (X_1, X_2, \ldots, X_n)$ and $\epsilon > 0$, an ϵ-cover for the empirical L_2 norm, if it exists, is a family of function $(f_i)_{i \in I}$ such that:

$$\forall f \in \mathcal{F}, \exists i \in I, \quad \left(\frac{1}{n} \sum_{j=1}^{n} \left(f \left(X_j \right) - f_i \left(X_j \right) \right)^2 \right)^{\frac{1}{2}} \leq \epsilon \ .$$

The covering number $\mathcal{N} \left(\mathcal{F}, \epsilon, L_2(T) \right)$ is then defined as the smallest cardinal of an ϵ-cover.

We can now mention the following result, adapted to our notations and setting, that exactly fits our need.

Theorem 18 (Steinwart et al., 2005, Theorem 5.8.) *For $\sigma > 0$, let $\mathcal{F}_\sigma$ be a convex subset of $\mathcal{H}_\sigma^{\mathcal{X}}$ and let ϕ be a convex loss function. Define $\mathcal{G}_\sigma$ as follows:*

$$\mathcal{G}_\sigma := \left\{ g_f(x,y) = \phi(yf(x)) + \lambda \| f \|_{\mathcal{H}_\sigma^{\mathcal{X}}}^2 - \phi(y f_{\phi,\sigma}^{\mathcal{X}}(x)) - \lambda \| f_{\phi,\sigma}^{\mathcal{X}} \|_{\mathcal{H}_\sigma^{\mathcal{X}}}^2 \ : \ f \in \mathcal{F}_\sigma \right\} . \quad (33)$$

where $f_{\phi,\sigma}^{\mathcal{X}}$ minimizes $R_{\phi,\sigma}^{\mathcal{X}}$ over $\mathcal{F}_\sigma$. Suppose that there are constants $c \geq 0$ and $B > 0$ such that, for all $g \in \mathcal{G}_\sigma$,

$$\mathbb{E}_P \left[g^2 \right] \leq c \mathbb{E}_P \left[g \right] ,$$

and

$$\| g \|_{L_\infty} \leq B .$$

Furthermore, assume that there are constants $a \geq 1$ and $0 < p < 2$ with

$$\sup_{T \in Z^n} \log \mathcal{N} \left(B^{-1} \mathcal{G}_\sigma, \epsilon, L_2(T) \right) \leq a \epsilon^{-p} \quad (34)$$

for all $\epsilon > 0$. Then there exists a constant $c_p > 0$ depending only on p such that for all $n \geq 1$ and all $x \geq 1$ we have

$$Pr^* \left(T \in Z^n \ : \ R_{\phi,\sigma}^{\mathcal{X}}(\hat{f}_{\phi,\sigma}^{\mathcal{X}}) > R_{\phi,\sigma}^{\mathcal{X}}(f_{\phi,\sigma}^{\mathcal{X}}) + c_p \epsilon(n,a,B,c,x) \right) \leq e^{-x} , \quad (35)$$

where

$$\epsilon(n,a,B,c,p,x) = \left(B + B^{\frac{2p}{2+p}} c^{\frac{2-p}{2+p}} \right) \left(\frac{a}{n} \right)^{\frac{2}{2+p}} + (B+c) \frac{x}{n} .$$

From the inequalities $\| f_{\phi,\sigma}^{\mathcal{X}} \|_{\mathcal{H}_\sigma^{\mathcal{X}}}^2 \leq \phi(0)/\lambda$ and $\| \hat{f}_{\phi,\sigma}^{\mathcal{X}} \|_{\mathcal{H}_\sigma^{\mathcal{X}}}^2 \leq \phi(0)/\lambda$, we see that it is enough to take

$$\mathcal{F}_\sigma = \sqrt{\frac{\phi(0)}{\lambda}} \mathcal{B}_\sigma^{\mathcal{X}} , \quad (36)$$

where $\mathcal{B}_\sigma^{\mathcal{X}}$ is the unit ball of $\mathcal{H}_\sigma^{\mathcal{X}}$, to derive a control of (32) from Theorem 18. In order to apply this theorem we now provide uniform upper bounds over $\mathcal{G}_\sigma$ for the variance of g and its uniform norm, as well as an upper bound on the covering number of $\mathcal{G}_\sigma$.

Lemma 19 *For all $\sigma > 0$, for all $g \in \mathcal{G}_\sigma$,*

$$\mathbb{E}_P \left[g^2 \right] \leq \left(L \left(\sqrt{\frac{\kappa_\sigma \phi(0)}{\lambda}} \right) \sqrt{\kappa_\sigma} + 2 \sqrt{\lambda \phi(0)} \right)^2 \frac{2}{\lambda} \mathbb{E}_P \left[g \right] . \quad (37)$$

Lemma 20 *For all $\sigma > 0$, for all $g \in \mathcal{G}_\sigma$,*

$$\| g \|_{L_\infty} \leq 2L \left(\sqrt{\frac{\kappa_\sigma \phi(0)}{\lambda}} \right) \sqrt{\frac{\kappa_\sigma \phi(0)}{\lambda}} + \phi(0) . \quad (38)$$

Lemma 21 *Let*

$$B = 2L \left(\sqrt{\frac{\kappa_\sigma \phi(0)}{\lambda}} \right) \sqrt{\frac{\kappa_\sigma \phi(0)}{\lambda}} + \phi(0) . \quad (39)$$

For all $\sigma > 0$, $0 < p \leq 2$, $\delta > 0$, $\epsilon > 0$, the following holds:

$$\log \mathcal{N}(B^{-1} \mathcal{G}_\sigma, \epsilon, L_2(T)) \leq c_2 \sigma^{-((1-p/2)(1+\delta))d} \epsilon^{-p} , \quad (40)$$

where c_1 and c_2 are constants that depend neither on σ, nor on ϵ (but they depend on p, δ, d and λ).

Combining now the results of Lemmas 19, 20 and 21 allows to apply Theorem 18 with $\mathcal{F}_\sigma$ defined by (36), any $p \in [0,2]$, and the following parameters:

$$\begin{aligned} c &= \left(L \left(\sqrt{\frac{\kappa_\sigma \phi(0)}{\lambda}} \right) \sqrt{\kappa_\sigma} + 2\sqrt{\lambda \phi(0)} \right)^2 \frac{2}{\lambda} \,, \\ \alpha &= 1 \,, \\ B &= 2L \left(\sqrt{\frac{\kappa_\sigma \phi(0)}{\lambda}} \right) \sqrt{\frac{\kappa_\sigma \phi(0)}{\lambda}} + \phi(0) \,, \\ a &= c_2 \sigma^{-((1-p/2)(1+\delta))d} \,, \end{aligned}$$

from which we deduce Lemma 16.

Proof of Lemma 17 Because the support of P is included in $\mathcal{X}$, the following trivial equality holds:

$$\forall f \in \mathcal{H}_\sigma, \quad \mathbb{E}_P \left[\phi \left(Y f \left(X \right) \right) \right] = \mathbb{E}_{P_{/\mathcal{X}}} \left[\phi \left(Y f_{/\mathcal{X}} \left(X \right) \right) \right] \,. \tag{41}$$

Using first the definition of the restricted RKHS (30), then (41) and (31), we obtain

$$\begin{aligned} R^{\mathcal{X}}_{\phi,\sigma} \left(f^{\mathcal{X}}_{\phi,\sigma} \right) &= \inf_{f^{\mathcal{X}} \in \mathcal{H}^{\mathcal{X}}_\sigma} \mathbb{E}_{P_{/\mathcal{X}}} \left[\phi \left(Y f^{\mathcal{X}} \left(X \right) \right) \right] + \lambda \| f^{\mathcal{X}} \|_{\mathcal{H}^{\mathcal{X}}_\sigma} \\ &= \inf_{f \in \mathcal{H}_\sigma} \mathbb{E}_{P_{/\mathcal{X}}} \left[\phi \left(Y f_{/\mathcal{X}} \left(X \right) \right) \right] + \lambda \| f_{/\mathcal{X}} \|_{\mathcal{H}^{\mathcal{X}}_\sigma} \\ &= \inf_{f \in \mathcal{H}_\sigma} \mathbb{E}_P \left[\phi \left(Y f \left(X \right) \right) \right] + \lambda \| f \|_{\mathcal{H}_\sigma} \\ &= R_{\phi,\sigma} \left(f_{\phi,\sigma} \right) \,. \end{aligned} \tag{42}$$

The same line or proof easily leads to $R_{\phi,\sigma} \left(\hat{f}_{\phi,\sigma} \right) = R^{\mathcal{X}}_{\phi,\sigma} \left(\hat{f}^{\mathcal{X}}_{\phi,\sigma} \right)$, after observing that with probability 1, $X_i \in \mathcal{X}$ for $i = 1, \ldots, n$, and therefore:

$$\forall f \in \mathcal{H}_\sigma, \quad \frac{1}{n} \sum_{i=1}^n \phi \left(Y_i f(X_i) \right) = \frac{1}{n} \sum_{i=1}^n \phi \left(Y_i f_{/\mathcal{X}}(X_i) \right) \,.$$

■

Proof of Lemma 19 We prove the uniform upper bound on the variances of the excess-loss functions in terms of their expectation, using an approach similar to but slightly simpler than Bartlett et al. (2003, Lemma 15) and Steinwart and Scovel (2004, Proposition 6.1). First we observe, using (21) and the fact that $\mathcal{F}_\sigma \subset \sqrt{\phi(0)/\lambda} B_\sigma$, that for any $f \in \mathcal{F}_\sigma$,

$$\begin{aligned} \| f \|_{L_\infty} &\leq \sqrt{\kappa_\sigma} \| f \|_{\mathcal{H}_\sigma} \\ &\leq \sqrt{\frac{\kappa_\sigma \phi(0)}{\lambda}}. \end{aligned}$$

As a result, for any $(x,y) \in \mathcal{X} \times \{-1,+1\}$,

$$\begin{aligned} |\, g_f\,(x,y)\,| &\leq |\,\phi\,(yf(x)) - \phi\,(yf_{\phi,\sigma}(x))\,| + \lambda \left|\, \| f \|^2_{\mathcal{H}_\sigma} - \| f_{\phi,\sigma} \|^2_{\mathcal{H}_\sigma} \,\right| \\ &\leq L\left(\sqrt{\frac{\kappa_\sigma \phi(0)}{\lambda}}\right) |\, f(x) - f_{\phi,\sigma}(x)\,| + \lambda \| f - f_{\phi,\sigma} \|_{\mathcal{H}_\sigma} \| f + f_{\phi,\sigma} \|_{\mathcal{H}_\sigma} \\ &\leq \left(L\left(\sqrt{\frac{\kappa_\sigma \phi(0)}{\lambda}}\right)\sqrt{\kappa_\sigma} + 2\sqrt{\lambda \phi(0)} \right) \| f - f_{\phi,\sigma} \|_{\mathcal{H}_\sigma} \,. \end{aligned} \tag{43}$$

Taking the square on both sides of this inequality and averaging with respect to P leads to:

$$\forall f \in \mathcal{F}_\sigma, \quad \mathbb{E}_P\left[g_f^2\right] \leq \left(L\left(\sqrt{\frac{\kappa_\sigma \phi(0)}{\lambda}}\right)\sqrt{\kappa_\sigma} + 2\sqrt{\lambda \phi(0)} \right)^2 \| f - f_{\phi,\sigma} \|^2_{\mathcal{H}_\sigma} \,. \tag{44}$$

On the other hand, we deduce from the convexity of ϕ that for any $(x,y) \in \mathcal{X} \times \{-1,+1\}$ and any $f \in \mathcal{F}_\sigma$:

$$\begin{aligned} &\frac{\phi(yf(x)) + \lambda \| f \|^2_{\mathcal{H}_\sigma} + \phi(yf_{\phi,\sigma}(x)) + \lambda \| f_{\phi,\sigma} \|^2_{\mathcal{H}_\sigma}}{2} \\ &\qquad \geq \phi\left(\frac{yf(x) + yf_{\phi,\sigma}(x)}{2}\right) + \lambda \frac{\| f \|^2_{\mathcal{H}_\sigma} + \| f_{\phi,\sigma} \|^2_{\mathcal{H}_\sigma}}{2} \\ &\qquad = \phi\left(y \frac{f + f_{\phi,\sigma}}{2}(x)\right) + \lambda \| \frac{f + f_{\phi,\sigma}}{2} \|^2_{\mathcal{H}_\sigma} + \lambda \| \frac{f - f_{\phi,\sigma}}{2} \|^2_{\mathcal{H}_\sigma} \,. \end{aligned}$$

Averaging this inequality with respect to P rewrites:

$$\begin{aligned} \frac{R_{\phi,\sigma}(f) + R_{\phi,\sigma}\,(f_{\phi,\sigma})}{2} &\geq R_{\phi,\sigma}\left(\frac{f + f_{\phi,\sigma}}{2}\right) + \lambda \| \frac{f - f_{\phi,\sigma}}{2} \|^2_{\mathcal{H}_\sigma} \\ &\geq R_{\phi,\sigma}\,(f_{\phi,\sigma}) + \lambda \| \frac{f - f_{\phi,\sigma}}{2} \|^2_{\mathcal{H}_\sigma} \,, \end{aligned}$$

where the second inequality is due to the definition of $f_{\phi,\sigma}$ as a minimizer of $R_{\phi,\sigma}$. Therefore we get, for any $f \in \mathcal{F}_\sigma$,

$$\begin{aligned} \mathbb{E}_P\,[g_f] &= R_{\phi,\sigma}(f) + R_{\phi,\sigma}\,(f_{\phi,\sigma}) \\ &\geq \frac{\lambda}{2} \| f - f_{\phi,\sigma} \|^2_{\mathcal{H}_\sigma} \,. \end{aligned} \tag{45}$$

Combining (44) and (45) finishes the proof of Lemma 19 ■

Proof of Lemma 20 Following a path similar to (43), we can write for any $f \in \mathcal{F}_\sigma$ and any $(x,y) \in \mathcal{X} \times \{-1,+1\}$:

$$\begin{aligned} |\, g_f\,(x,y)\,| &\leq |\,\phi\,(yf(x)) - \phi\,(yf_{\phi,\sigma}(x))\,| + \lambda \left|\, \| f \|^2_{\mathcal{H}_\sigma} - \| f_{\phi,\sigma} \|^2_{\mathcal{H}_\sigma} \,\right| \\ &\leq L\left(\sqrt{\frac{\kappa_\sigma \phi(0)}{\lambda}}\right) |\, f(x) - f_{\phi,\sigma}(x)\,| + \lambda \frac{\phi(0)}{\lambda} \\ &\leq 2L\left(\sqrt{\frac{\kappa_\sigma \phi(0)}{\lambda}}\right)\sqrt{\frac{\kappa_\sigma \phi(0)}{\lambda}} + \phi(0) \,. \end{aligned}$$

■

Proof of Lemma 21 Let us introduce the notations $l_\phi \circ f(x,y) = \phi\left(y(f(x)\right)$ and $L_{\phi,\sigma} \circ f = l_\phi \circ f(x,y) + \lambda \| f \|^2_{\mathcal{H}^{\mathcal{X}}_\sigma}$, for $f \in \mathcal{H}^{\mathcal{X}}_\sigma$ and $(x,y) \in \mathcal{X} \times \{-1,1\}$. We can then rewrite (33) as:

$$\mathcal{G}_\sigma = \left\{ L_{\phi,\sigma} \circ f - L_{\phi,\sigma} \circ f^{\mathcal{X}}_{\phi,\sigma} \ : \ f \in \mathcal{F}_\sigma \right\} \ .$$

The covering number of a set does not change when the set is translated by a constant, therefore:

$$\mathcal{N}\left(B^{-1}\mathcal{G}_\sigma, \epsilon, L_2(T)\right) = \mathcal{N}\left(B^{-1}L_{\phi,\sigma} \circ \mathcal{F}_\sigma, \epsilon, L_2(T)\right) \ .$$

Denoting now $[a,b]$ the set of constant functions with values between a and b, we deduce, from the fact that $\lambda \| f \|^2_{\mathcal{H}^{\mathcal{X}}_\sigma} \leq \phi(0)$ for $f \in \mathcal{F}_\sigma$, that

$$B^{-1}L_{\phi,\sigma} \circ \mathcal{F}_\sigma \subset B^{-1}l_\phi \circ \mathcal{F}_\sigma + \left[0, B^{-1}\phi(0)\right] \ .$$

Using the sub-additivity of the entropy we therefore get:

$$\log \mathcal{N}\left(B^{-1}\mathcal{G}_\sigma, 2\epsilon, L_2(T)\right) \leq \log \mathcal{N}\left(B^{-1}l_\phi \circ \mathcal{F}_\sigma, \epsilon, L_2(T)\right) + \log \mathcal{N}\left(\left[0, B^{-1}\phi(0)\right], \epsilon, L_2(T)\right) \ . \tag{46}$$

In order to upper bound the first term in the r.h.s. of (46), we observe that for any $f \in \mathcal{F}_\sigma$ and $x \in \mathcal{X}$,

$$| f(x) | \leq \sqrt{\kappa_\sigma} \| f \|_{\mathcal{H}^{\mathcal{X}}_\sigma} \leq \sqrt{\frac{\phi(0)\kappa_\sigma}{\lambda}} \ ,$$

and therefore a simple computation shows that, if $u(x,y) = B^{-1}\phi(yf(x))$ and $u'(x,y) = B^{-1}\phi(yf'(x))$ are two elements of $B^{-1}l_\phi \circ \mathcal{F}_\sigma$ (with $f, f' \in \mathcal{F}_\sigma$), then for any sample T:

$$\| u - u' \|_{L_2(T)} \leq B^{-1}L \left(\sqrt{\frac{\phi(0)\kappa_\sigma}{\lambda}} \right) \| f - f' \|_{L_2(T)} \ .$$

and therefore

$$\begin{aligned} \log \mathcal{N}\left(B^{-1}l_\phi \circ \mathcal{F}_\sigma, \epsilon, L_2(T)\right) &\leq \log \mathcal{N}\left(\mathcal{F}_\sigma, B\epsilon L \left(\sqrt{\frac{\phi(0)\kappa_\sigma}{\lambda}} \right)^{-1}, L_2(T)\right) \\ &\leq \log \mathcal{N}\left(\mathcal{B}^{\mathcal{X}}_\sigma, B\epsilon L \left(\sqrt{\frac{\phi(0)\kappa_\sigma}{\lambda}} \right)^{-1} \sqrt{\frac{\lambda}{\phi(0)}}, L_2(T)\right) \ . \end{aligned} \tag{47}$$

Recalling the definition of B in (39), we obtain:

$$B\epsilon L \left(\sqrt{\frac{\phi(0)\kappa_\sigma}{\lambda}} \right)^{-1} \sqrt{\frac{\lambda}{\phi(0)}} \geq 2\epsilon\sqrt{\kappa_\sigma} \ ,$$

and therefore

$$\log \mathcal{N}\left(B^{-1}l_\phi \circ \mathcal{F}_\sigma, \epsilon, L_2(T)\right) \leq \log \mathcal{N}\left(\mathcal{B}^{\mathcal{X}}_\sigma, 2\epsilon\sqrt{\kappa_\sigma}, L_2(T)\right) \ .$$

The second term in the r.h.s. of (46) is easily upper bounded by:

$$\log \mathcal{N}\left(\left[0, B^{-1}\phi(0)\right], \epsilon, L_2(T)\right) \leq \log\left(\frac{\phi(0)}{B\epsilon}\right) ,$$

and we finally get:

$$\log \mathcal{N}\left(B^{-1}\mathcal{G}_\sigma, 2\epsilon, L_2(T)\right) \leq \log \mathcal{N}\left(\mathcal{B}_\sigma^{\mathcal{X}}, 2\epsilon\sqrt{\kappa_\sigma}, L_2(T)\right) + \log\left(\frac{\phi(0)}{B\epsilon}\right) . \tag{48}$$

We now need to upper bound the covering number of the unit ball in the RKHS. We make use of the following result, proved by Steinwart and Scovel (2004, Theorem 2.1): if $\tilde{\mathcal{B}}_\sigma^{\mathcal{X}}$ denotes the unit ball of the RKHS associated with the non-normalized Gaussian kernel (24) on a compact set, then for all $0 < p \leq 2$ and all $\delta > 0$ there exists a constant $c_{p,\delta,d}$ independent of σ such that for all $\tilde{\epsilon} > 0$ we have:

$$\log \mathcal{N}\left(\tilde{\mathcal{B}}_\sigma^{\mathcal{X}}, \tilde{\epsilon}, L_2(T)\right) \leq c_{p,\delta,d}\sigma^{(1-p/2)(1+\delta)d}\epsilon^{-p} . \tag{49}$$

Now, using (25), we observe that

$$\mathcal{B}_\sigma^{\mathcal{X}} = \sqrt{\kappa_\sigma}\tilde{\mathcal{B}}_\sigma^{\mathcal{X}} ,$$

and therefore:

$$\begin{aligned} \log \mathcal{N}\left(\mathcal{B}_\sigma^{\mathcal{X}}, 2\epsilon\sqrt{\kappa_\sigma}, L_2(T)\right) &= \log \mathcal{N}\left(\sqrt{\kappa_\sigma}\tilde{\mathcal{B}}_\sigma^{\mathcal{X}}, 2\epsilon\sqrt{\kappa_\sigma}, L_2(T)\right) \\ &= \log \mathcal{N}\left(\tilde{\mathcal{B}}_\sigma^{\mathcal{X}}, 2\epsilon, L_2(T)\right) . \end{aligned} \tag{50}$$

Plugging (49) into (50), and (50) into (48) finally leads to the announced result, after observing that the second term in the r.h.s. of (48) becomes negligible compared to the first one and can therefore be hidden in the constant for ϵ small enough. ■

6. Some properties of the L_2-norm-regularized ϕ-risk (proofs of Theorem 30 and Lemma 7)

In this section we investigate the conditions on the loss function ϕ under which the Bayes consistency of the minimization of the regularized ϕ-risk holds. In the spirit of Bartlett et al. (2003), we introduce a notion of classification-calibration for regularized loss functions ϕ, and upper bound the excess risk of any classifier f in terms of its excess of regularized ϕ-risk. We also upper-bound the L_2-distance between f and $f_{\phi,0}$ in terms of the excess of regularized ϕ-risk of f, which is useful to proove Bayes consistency in the one-class setting.

6.1 Classification Calibration

In the classical setting, Bartlett et al. (2003, Definition 1) introduce the following notion of classification-calibrated loss functions:

Definition 22 *For any $(\eta, \alpha) \in [0,1] \times \mathbb{R}$, let the generic conditional ϕ-risk be defined by:*

$$C_\eta(\alpha) = \eta\phi(\alpha) + (1-\eta)\,\phi(-\alpha).$$

The loss function ϕ is said to be classification-calibrated if, for any $\eta \in [0,1]\backslash\{1/2\}$:

$$\inf_{\alpha\in\mathbb{R}:\alpha(2\eta-1)\leq 0} C_\eta(\alpha) > \inf_{\alpha\in\mathbb{R}} C_\eta(\alpha)$$

The importance here is in the *strict* inequality, which implies in particular that if the global infimum of C_η is reached at some point α, then $\alpha > 0$ (resp. $\alpha < 0$) if $\eta > 1/2$ (resp. $\eta < 1/2$). This condition, that generalizes the requirement that the minimizer of $C_\eta(\alpha)$ has the correct sign, is a minimal condition that can be viewed as a pointwise form of Fisher consistency for classification. In our case, noting that for any $f \in \mathcal{M}$, the L_2-regularized ϕ-risk can be rewritten as follows:

$$R_{\phi,0}(f) = \int_{\mathbb{R}^d} \left\{ [\eta(x)\phi(f(x)) + (1-\eta(x))\,\phi(-f(x))]\, p(x) + \lambda f(x)^2 \right\} dx\ ,$$

we introduce the regularized generic conditional ϕ-risk:

$$\forall(\eta,\rho,\alpha) \in [0,1]\times(0,+\infty)\times\mathbb{R}, \quad C_{\eta,\rho}\left(\alpha\right) = C_\eta\left(\alpha\right) + \frac{\lambda\alpha^2}{\rho}\ ,$$

as well as the related weighted regularized generic conditional ϕ-risk:

$$\forall(\eta,\rho,\alpha) \in [0,1]\times[0,+\infty)\times\mathbb{R}, \quad G_{\eta,\rho}\left(\alpha\right) = \rho C_\eta\left(\alpha\right) + \lambda\alpha^2\ .$$

This leads to the following notion of classification-calibration:

Definition 23 *We say that ϕ is classification calibrated for the regularized risk, or* R-classification-calibrated*, if for any $(\eta,\rho) \in [0,1]\backslash\{1/2\}\times(0,+\infty)$*

$$\inf_{\alpha\in\mathbb{R}:\alpha(2\eta-1)\leq 0} C_{\eta,\rho}(\alpha) > \inf_{\alpha\in\mathbb{R}} C_{\eta,\rho}(\alpha)$$

The following result clarifies the relationship between the properties of classification-calibration and R-classification-calibration.

Lemma 24 *For any function $\phi : \mathbb{R} \to [0,+\infty)$, $\phi(x)$ is R-classification-calibrated if and only if for any $t > 0$, $\phi(x) + tx^2$ is classification-calibrated.*

Proof For any $\phi : \mathbb{R} \to [0,+\infty)$ and $\rho > 0$, let $\phi'(x) = \phi(x) + \lambda x^2/\rho$ and C'_η the corresponding generic conditional ϕ'-risk. Then one easily gets, for any $\alpha \in \mathbb{R}$

$$C'_\eta\left(\alpha\right) = C_{\eta,\rho}\left(\alpha\right).$$

As a result, ϕ is R-classification-calibrated if and only if, for any ρ, ϕ' is classification-calibrated, which proves the lemma. ■

Remark 25 *Classification-calibration and R-classification-calibration are two different properties related to each other by Lemma 24, but none of them implies the other one. For example, it can be shown that* $\phi(x) = 1$ *on* $(-\infty, -2]$, $\phi(x) = 2$ *on* $[-1, 1]$, $\phi(x) = 0$ *on* $[2, +\infty)$, *and* ϕ *continous linear on* $[-2, -1]$ *and* $[1, 2]$, *defines a classification-calibrated function which is not R-classification-calibrated. Conversely, the function* $\phi(x) = e^x$ *for* $x \leq 0$ *and* $\phi(x) = e^{-2x}$ *for* $x \geq 0$ *can be shown to be R-classification-calibrated, but not classification-calibrated.*

6.2 Classification-calibration of convex loss functions

The following lemma states the equivalence between classification calibration and R-classification calibration for convex loss functions, and it gives a simple characterization of this property.

Lemma 26 *For a convex function* $\phi : \mathbb{R} \to [0, +\infty)$, *the following properties are equivalent:*

1. ϕ *is classification-calibrated,*
2. ϕ *is R-classification-calibrated,*
3. ϕ *is differentiable at* 0 *and* $\phi'(0) < 0$.

Proof The equivalence of the first and the third properties is shown in Bartlett et al. (2003, Theorem 6). From this and lemma 24, we deduce that ϕ is R-classification-calibrated iff $\phi(x) + tx^2$ is classification-calibrated for any $t > 0$, iff $\phi(x) + tx^2$ is differentiable at 0 with negative derivative (for any $t > 0$, iff $\phi(x)$ is differentiable at 0 with negative derivative. This proves the equivalence between the second and third properties. ■

6.3 Some properties of the minimizer of the $R_{\phi,0}$-risk

When ϕ is convex, the function $C_\eta(\alpha)$ is a convex function of α (as a convex combination of convex functions), and therefore $G_{\eta,\rho}(\alpha)$ is strictly convex and diverges to $+\infty$ in $-\infty$ and $+\infty$; as a result, for any $(\eta, \rho) \in [0, 1] \times [0, +\infty)$, there exists a unique $\alpha(\eta, \rho)$ that minimizes $G_{\eta,\rho}$ on $\mathbb{R}$. It satisties the following inequality:

Lemma 27 *If* $\phi : \mathbb{R} \to [0, +\infty)$ *is a convex function, then for any* $(\eta, \rho) \in [0, 1] \times [0, +\infty)$ *and any* $\alpha \in \mathbb{R}$,

$$G_{\eta,\rho}(\alpha) - G_{\eta,\rho}(\alpha(\eta, \rho)) \geq \lambda(\alpha - \alpha(\eta, \rho))^2 \ . \tag{51}$$

Proof For any $(\eta, \rho) \in [0, 1] \times [0, +\infty)$, the function $G_{\eta,\rho}(\alpha)$ is the sum of the convex function $\rho C_\eta(\alpha)$ and of the strictly convex function $\lambda\alpha^2$. Let us denote by $C_\eta^+(\alpha)$ the right-hand derivative of C_η at the point α (which is well defined for convex functions). The right-hand derivative of a convex function being non-negative at a minimum, we have (denoting $\alpha_* = \alpha(\eta, \rho)$):

$$\rho C_\eta^+(\alpha_*) + 2\lambda\alpha_* \geq 0 \ . \tag{52}$$

Now, for any $\alpha > \alpha_*$, we have by convexity of C_η:

$$C_\eta(\alpha) \geq C_\eta(\alpha_*) + (\alpha - \alpha_*) C_\eta^+(\alpha_*) \ . \tag{53}$$

Moreover, by direct calculation we get:

$$\lambda\alpha^2 = \lambda\alpha_*^2 + 2\lambda\alpha_*\left(\alpha - \alpha_*\right) + \lambda\left(\alpha - \alpha_*\right)^2 \ . \tag{54}$$

Mutliplying (53) by ρ, adding (54) and plugging (52) into the result leads to:

$$G_{\eta,\rho}\left(\alpha\right) - G_{\eta,\rho}\left(\alpha_*\right) \geq \lambda\left(\alpha - \alpha_*\right)^2 .$$

By symetry, this inequality is also valid for $\alpha \leq \alpha_*$. ■

From this result we obtain the following characterization and properties of the minimizer of the $R_{\phi,0}$-risk:

Theorem 28 *If $\phi : \mathbb{R} \to [0, +\infty)$ is a convex function, then the function $f_{\phi,0} : \mathbb{R}^d \to \mathbb{R}$ defined for any $x \in \mathbb{R}^d$ by*

$$f_{\phi,0}(x) = \alpha(\eta(x), \rho(x))$$

satisfies:

1. *$f_{\phi,0}$ is measurable.*
2. *$f_{\phi,0}$ minimizes the $R_{\phi,0}$-risk:*

$$R_{\phi,0}\left(f_{\phi,0}\right) = \inf_{f \in \mathcal{M}} R_{\phi,0}(f) \ .$$

3. *For any $f \in \mathcal{M}$, the following holds:*

$$\| f - f_{\phi,0} \|_{L_2}^2 \leq \frac{1}{\lambda}\left(R_{\phi,2}(f) - R_{\phi,2}^*\right).$$

Proof From Lemma 27 and the fact that for any $\alpha \in \mathbb{R}$, the mapping $(\eta, \rho) \in [0,1] \times [0, +\infty) \mapsto G_{\eta,\rho}(\alpha)$ is continuous, we can deduce that the mapping $(\eta, \rho) \in [0,1] \times [0, +\infty) \mapsto \alpha\left(\eta, \rho\right)$ is continuous. Indeed, fix $(\eta_0, \rho_0) \in [0,1] \times [0, +\infty)$ and the corresponding $\alpha_0 = \alpha\left(\eta_0, \rho_0\right)$; for any $\epsilon > 0$, the mapping $(\eta, \rho) \mapsto G_{\eta,\rho}(\alpha)$ is absolutely continuous on the compact $[\alpha_0 - \epsilon, \alpha_0 + \epsilon]$, and therefore there exists a neighborhood B of (η_0, ρ_0) such that for any $(\eta, \rho) \in B$,

$$\sup_{\alpha \in [\alpha_0 - \epsilon, \alpha_0 + \epsilon]} \left| G_{\eta,\rho}\left(\alpha\right) - G_{\eta_0,\rho_0}\left(\alpha\right) \right| < \frac{\lambda\epsilon^2}{3}.$$

In particular, this implies that for any $(\eta, \rho) \in B$,

$$G_{\eta,\rho}\left(\alpha_0 + \epsilon\right) > G_{\eta,\rho}\left(\alpha_0\right) + \frac{\lambda\epsilon^2}{3} \quad \text{and} \quad G_{\eta,\rho}\left(\alpha_0 - \epsilon\right) > G_{\eta,\rho}\left(\alpha_0\right) + \frac{\lambda\epsilon^2}{3}$$

which implies, by convexity of $G_{\eta,\rho}$, that $\alpha\left(\eta, \rho\right) \in [\alpha_0 - \epsilon, \alpha_0 + \epsilon]$ as soon as $(\eta, \rho) \in B$, therefore proving the continuity of $(\eta, \rho) \to \alpha\left(\eta, \rho\right)$. As a result, $f_{\phi,0}$ is measurable as a continuous fonction of two measurable functions η and ρ.

Now, we have by construction, for any $f \in \mathcal{M}$:

$$\forall x \in \mathbb{R}^d, \quad G_{\eta(x),\rho(x)}\left(f_{\phi,0}(x)\right) \leq G_{\eta(x),\rho(x)}\left(f(x)\right)$$

which after integration leads to:

$$R_{\phi,0}(f_{\phi,0}) \leq R_{\phi,0}(f) \ ,$$

proving the second statement of the theorem.

Finally, for any $f \in \mathcal{M}$, rewriting (51) with $\alpha = f(x)$, $\rho = \rho(x)$ and $\eta = \eta(x)$ shows that:

$$\forall x \in \mathbb{R}^d, \quad G_{\eta(x),\rho(x)}\left(f(x)\right) - G_{\eta(x),\rho(x)}\left(f_{\phi,0}(x)\right) \leq \lambda \left(f(x) - f_{\phi,0}(x)\right)^2 \ ,$$

which proves the third statement of Theorem 28. ■

6.4 Relating the $R_{\phi,0}$-risk with the classification error rate

In the "classical" setting (with a regularization parameter converging to 0), the idea of relating the convexified risk to the true risk (more simply called risk) has recently gained a lot of interest. Zhang (2004) and Lugosi and Vayatis (2004) upper bound the excess-risk by some function of the excess ϕ-risk to prove consistency of various algorithms (and obtain upper bounds for the rates of convergence of the risk to the Bayes risk). These ideas were then generalized by Bartlett et al. (2003), which we now adapt to our framework.

Let us define, for any $(\eta, \rho) \in [0,1] \times (0, +\infty)$,

$$M\left(\eta, \rho\right) = \min_{\alpha \in \mathbb{R}} C_{\eta,\rho}(\alpha) = C_{\eta,\rho}\left(\alpha(\eta,\rho)\right),$$

and for any $\rho > 0$ the function ψ_ρ defined for all θ in $[0,1]$ by

$$\psi_\rho\left(\theta\right) := \phi(0) - M\left(\frac{1+\theta}{2}, \rho\right).$$

The following lemma summarizes a few properties of M and ψ_ρ. Explicit computations for some classifical loss functions are performed in Section 7.

Lemma 29 *If $\phi : \mathbb{R} \to [0, +\infty)$ is a convex function, then for any $\rho > 0$, the following properties hold:*

1. *the function $\eta \mapsto M(\eta, \rho)$ is symetric around $1/2$, concave, and continuous on $[0,1]$;*
2. *ψ_ρ is convex, continuous on $[0,1]$, nonnegative, increasing, and $\psi(0) = 0$;*
3. *if $0 < \rho < \tau$, then $\psi_\rho \leq \psi_\tau$ on $[0,1]$;*
4. *ϕ is R-classification-calibrated if and only if $\psi_\rho(\theta) > 0$ for $\theta \in (0,1]$.*

Proof For any $\rho > 0$, let

$$\phi_\rho(x) = \phi(x) + \frac{\lambda x^2}{\rho}.$$

The corresponding generic conditional ϕ_ρ-risk C'_η satisfies

$$C'_\eta\left(\alpha\right) = C_{\eta,\rho}\left(\alpha\right)$$

ϕ_ρ being convex, the first two points are direct consequences of Bartlett et al. (2003, Lemma 5 & 7).

To prove the third point, it suffices to observe that for $0 < \rho \leq \tau$ we have for any $(\eta, \alpha) \in [0,1] \times \mathbb{R}$:

$$C_{\eta,\rho}(\alpha) - C_{\eta,\tau}(\alpha) = \lambda \alpha^2 \left(\frac{1}{\rho} - \frac{1}{\tau} \right) \geq 0,$$

which implies, by taking the minimum in α:

$$M(\eta, \rho) \geq M(\eta, \tau),$$

and therefore, for $\theta \in [0,1]$

$$\psi_\rho(\theta) \leq \psi_\tau(\theta).$$

Finally, by lemma 26, ϕ is R-classification-calibrated iff ϕ_ρ is classification-calibrated (because both properties are equivalent to saying that ϕ is differentiable at 0 and $\phi'(0) < 0$), iff $\psi_\rho(\theta) > 0$ for $\theta \in (0,1]$ by Bartlett et al. (2003, Lemma 4). ■

We are now in position to state a first result to relate the excess $R_{\phi,2}$-risk to the excess-risk. The dependence on $\rho(x)$ generates difficulties compared with the "classical" setting, which forces us to separate the low density regions from the rest in the analysis.

Theorem 30 *Suppose ϕ is a convex classification-calibrated function, and for any $\epsilon > 0$, let*

$$A_\epsilon = \left\{ x \in \mathbb{R}^d : \rho(x) \leq \epsilon \right\}.$$

For any $f \in \mathcal{M}$ the following holds:

$$R(f) - R^* \leq \inf_{\epsilon > 0} \left\{ P(A_\epsilon) + \psi_\epsilon^{-1} \left(R_{\phi,0}(f) - R_{\phi,0}^* \right) \right\} \tag{55}$$

Proof First note that for convex classification-calibrated functions, ψ_ϵ is continuous and strictly increasing on $[0,1]$, and is therefore invertible, which justifies the use of ψ_ϵ^{-1} in (55). Fix now a function $f \in \mathcal{M}$, and let $U(x) = 1$ if $f(x)(2\eta(x) - 1) < 0$, 0 otherwise (U is the indicator function of the set where f and the optimal classifier disagree). For any $\epsilon > 0$, if

we define $B_\epsilon = \mathbb{R}^d \backslash A_\epsilon$, we can compute:

$$\begin{aligned}
R_{\phi,0}(f) - R^*_{\phi,0} &= \int_{\mathbb{R}^d} \left[C_{\eta(x),\rho(x)}(f(x)) - M(\eta(x),\rho(x)) \right] \rho(x) dx \\
&\geq \int_{\mathbb{R}^d} \left[C_{\eta(x),\rho(x)}(f(x)) - M(\eta(x),\rho(x)) \right] U(x)\rho(x) dx \\
&\geq \int_{\mathbb{R}^d} \left[\phi(0) - M(\eta(x),\rho(x)) \right] U(x)\rho(x) dx \\
&= \int_{\mathbb{R}^d} \psi_{\rho(x)}(|\, 2\eta(x) - 1 \,|)\, U(x)\rho(x) dx \\
&\geq \int_{B_\epsilon} \psi_{\rho(x)}(|\, 2\eta(x) - 1 \,|)\, U(x)\rho(x) dx \\
&\geq \int_{B_\epsilon} \psi_{\epsilon}(|\, 2\eta(x) - 1 \,|)\, U(x)\rho(x) dx \\
&= \int_{B_\epsilon} \psi_{\epsilon}(U(x)\,|\, 2\eta(x) - 1 \,|)\, \rho(x) dx \\
&= P(B_\epsilon) \int_{B_\epsilon} \psi_{\epsilon}(U(x)\,|\, 2\eta(x) - 1 \,|)\, \frac{\rho(x)}{P(B_\epsilon)} dx \\
&\geq P(B_\epsilon)\, \psi_\epsilon \left(\frac{1}{P(B_\epsilon)} \int_{B_\epsilon} |\, 2\eta(x) - 1 \,|\, U(x)\rho(x) dx \right) \\
&\geq \psi_\epsilon \left(\int_{B_\epsilon} |\, 2\eta(x) - 1 \,|\, U(x)\rho(x) dx \right) \\
&= \psi_\epsilon \left(\int_{R^d} |\, 2\eta(x) - 1 \,|\, U(x)\rho(x) dx - \int_{A_\epsilon} |\, 2\eta(x) - 1 \,|\, U(x)\rho(x) dx \right) \\
&\geq \psi_\epsilon \left(\int_{R^d} |\, 2\eta(x) - 1 \,|\, U(x)\rho(x) dx - P(A_\epsilon) \right) \\
&= \psi_\epsilon \left(R(f) - R^* - P(A_\epsilon) \right) ,
\end{aligned}$$

where the successive (in)equalities are respectively justified by: (i) the definition of $R_{\phi,0}$ and the second point of Theorem 28; (ii) the fact that $U \leq 1$; (iii) the fact that when f and $2\eta - 1$ have different signs, then $C_{\eta,\rho}(f) \geq C_{\eta,\rho}(0) = \phi(0)$; (iv) the definition of ψ_ρ; (v) the obvious fact that $B_\epsilon \subset \mathbb{R}^d$; (vi) the observation that, by definition, ρ is larger than ϵ on B_ϵ, and the third point of Lemma 29; (vii) the fact that $\psi_\epsilon(0) = 0$ and $U(x) \in \{0, 1\}$; (viii) a simple division and multiplication by $P(B_\epsilon) > 0$; (ix) Jensen's inequality; (x) the convexity of ψ_ϵ and the facts that $\psi(0) = 0$ and $P(B_\epsilon) < 1$; (xi) the fact that $B_\epsilon = \mathbb{R}^d \backslash A_\epsilon$; (xii) the upper bound $|\, 2\eta(x) - 1 \,| U(x) \leq 1$ and the fact that ψ_ϵ is increasing; and (xiii) a classical inequality that can be found, e.g., in Devroye et al. (1996, Theorem 2.2). Composing each side by the strictly increasing function ψ_ϵ^{-1} leads to the announced result. ■

Theorem 4 is just a corollary of the previous Theorem:

Corollary 31 *If ϕ is a convex (R-)classification-calibrated loss function, then for any probability P whose marginal in X is absolutely continuous with respect to the Lebesgue measure, and every sequence of functions $f_i \in \mathcal{M}$, $\lim_{n\to+\infty} R_{\phi,0}(f_i) = R^*_{\phi,0}$ implies $lim_{n\to+\infty} R(f_i) = R^*$.*

Proof For any $\delta > 0$, choosing ϵ small enough to ensure $P(A_\epsilon) < \delta/2$, and $N \in \mathbb{N}$ such that for any $n > N$,

$$R_{\phi,0}(f_n) - R^*_{\phi,0} < \psi_\epsilon\left(\frac{\delta}{2}\right)$$

ensures, by Theorem 30, that for any $n > N$, $R(f_n) - R* < \delta$. ■

This important result shows that any consistency result for the regularized ϕ-risk implies consistency for the true risk, that is, convergence to the Bayes risk. Besides, convergence rates for the regularized ϕ-risk towards its minimum translate into convergence rates for the risk towards the Bayes risk thanks to (55).

6.5 Refinments under Low noise Assumption

When the distribution P satisfies a low noise assumption as defined in section 2, we have the following result:

Theorem 32 *Let ϕ be a convex loss function such that there exist $(\kappa, \beta, \nu) \in (0, +\infty)^3$ satisfying:*

$$\forall (\epsilon, u) \in (0, +\infty) \times \mathbb{R}, \quad \psi_\epsilon^{-1}(u) \leq \kappa u^\beta \epsilon^{-\nu}.$$

Then for any distribution P with low density exponent γ, there exist constant $(K, r) \in (0, +\infty)$ such that for any $f \in \mathcal{M}$ with an excess regularized ϕ-risk upper bounded by r the following holds:

$$R(f) - R^* \leq K\left(R_{\phi,0}(f) - R^*_{\phi,0}\right)^{\frac{\beta\gamma}{\gamma-\nu}}.$$

Proof Let $(c_2, \epsilon_0) \in (0, +\infty)^2$ such that

$$\forall \epsilon \in [0, \epsilon_0], \quad P(A_\epsilon) \leq c_2 \epsilon^\gamma, \tag{56}$$

and define

$$r = \epsilon_0^{\frac{\gamma+\nu}{\beta}} \kappa^{-\frac{1}{\beta}} c_2^{\frac{1}{\beta}}. \tag{57}$$

Given any function $f \in \mathcal{M}$ such that $\delta = R_{\phi,0}(f) - R^*_{\phi,0} \leq r$, let

$$\epsilon = \kappa^{\frac{1}{\gamma+\nu}} c_2^{-\frac{1}{\gamma+\nu}} \delta^{\frac{\beta}{\gamma+\nu}}. \tag{58}$$

Because $\delta \leq r$, we can upper bound ϵ by:

$$\begin{aligned} \epsilon &\leq \kappa^{\frac{1}{\gamma+\nu}} c_2^{-\frac{1}{\gamma+\nu}} r^{\frac{\beta}{\gamma+\nu}} \\ &= \epsilon_0. \end{aligned}$$

This implies, by (56), that

$$\begin{aligned} P(A_\epsilon) &\leq c_2 \epsilon^\gamma \\ &\leq \kappa^{\frac{\gamma}{\gamma+\nu}} c_2^{\frac{\nu}{\gamma+\nu}} \delta^{\frac{\beta\gamma}{\gamma+\nu}}. \end{aligned} \tag{59}$$

On the other hand,

$$\begin{aligned}\psi_\epsilon^{-1}(\delta) &\leq \kappa\delta^\beta\epsilon^{-\nu} \\ &= \kappa^{\frac{\gamma}{\gamma+\nu}} c_2^{\frac{\nu}{\gamma+\nu}} \delta^{\frac{\beta\gamma}{\gamma+\nu}}.\end{aligned} \tag{60}$$

Combining theorem 30 with (59) and (60) leads to the result claimed with the constant r defined in (57) and

$$K = 2\kappa^{\frac{\gamma}{\gamma+\nu}} c_2^{\frac{\nu}{\gamma+\nu}}.$$

■

7. The case of SVM

7.1 1-SVM

Let $\phi(\alpha) = \max(1-\alpha, 0)$. Then we easily obtain, for any $(\eta, \rho) \in [-1,1] \times (0, +\infty)$:

$$C_{\eta,\rho}(\alpha) = \begin{cases} \eta(1-\alpha) + \lambda\alpha^2/\rho & \text{if } \alpha \in (-\infty, -1] \\ \eta(1-\alpha) + (1-\eta)(1+\alpha) + \lambda\alpha^2/\rho & \text{if } \alpha \in [-1, 1] \\ (1-\eta)(1+\alpha) + \lambda\alpha^2/\rho & \text{if } \alpha \in [1, +\infty). \end{cases}$$

This shows that $C_{\eta,\rho}$ is strictly deacreasing on $(-\infty, -1]$ and strictly increasing on $[1, +\infty)$; as a result it reaches its minimum on $[-1, 1]$. Its derivative on this interval is equal to:

$$\forall \alpha \in (-1, 1), \quad C'_{\eta,\rho}(\alpha) = \frac{2\lambda\alpha}{\rho} + 1 - 2\eta.$$

This shows that $C_{\eta,\rho}$ reaches its minimum at the point:

$$\alpha(\eta, \rho) = \begin{cases} -1 & \text{if } \eta \leq 1/2 - \lambda/\rho \\ (\eta - 1/2)\rho/\lambda & \text{if } \eta \in [1/2 - \lambda/\rho, 1/2 + \lambda/\rho] \\ 1 & \text{if } \eta \geq 1/2 - \lambda/\rho \end{cases}$$

and that the value of this minimum is equal to:

$$M(\eta, \rho) = \begin{cases} 2\eta + \lambda/\rho & \text{if } \eta \leq 1/2 - \lambda/\rho \\ 1 - \rho(\eta - 1/2)^2/\lambda & \text{if } \eta \in [1/2 - \lambda/\rho, 1/2 + \lambda/\rho] \\ 2(1-\eta) + \lambda/\rho & \text{if } \eta \geq 1/2 + \lambda/\rho \end{cases}$$

From this we deduce that for all $(\rho, \theta) \in (0, +\infty) \times [-1, 1]$:

$$\psi_\rho(\theta) = \begin{cases} \rho\theta^2/(4\lambda) & \text{if } 0 \leq \theta \leq 2\lambda/\rho, \\ \theta - \lambda/\rho & \text{if } 2\lambda/\rho \leq \theta \leq 1 \end{cases}$$

whose inverse function is

$$\psi_\rho^{-1}(u) = \begin{cases} \sqrt{4\lambda u/\rho} & \text{if } 0 \leq u \leq \lambda/\rho, \\ u + \lambda/\rho & \text{if } u \geq \lambda/\rho. \end{cases} \tag{61}$$

7.2 2-SVM

Let $\phi(\alpha) = \max(1-\alpha, 0)^2$. Then we obtain, for any $(\eta, \rho) \in [-1,1] \times (0, +\infty)$:

$$C_{\eta,\rho}(\alpha) = \begin{cases} \eta(1-\alpha)^2 + \lambda\alpha^2/\rho & \text{if } \alpha \in (-\infty, -1] \\ \eta(1-\alpha)^2 + (1-\eta)(1+\alpha)^2 + \lambda\alpha^2/\rho & \text{if } \alpha \in [-1, 1] \\ (1-\eta)(1+\alpha)^2 + \lambda\alpha^2/\rho & \text{if } \alpha \in [1, +\infty). \end{cases}$$

This shows that $C_{\eta,\rho}$ is strictly deacreasing on $(-\infty, -1]$ and strictly increasing on $[1, +\infty)$; as a result it reaches its minimum on $[-1, 1]$. Its derivative on this interval is equal to:

$$\forall \alpha \in (-1, 1), \quad C'_{\eta,\rho}(\alpha) = 2\left(1 + \frac{\lambda}{\rho}\right)\alpha + 1 - 2\eta.$$

This shows that $C_{\eta,\rho}$ reaches its minimum at the point:

$$\alpha(\eta, \rho) = (2\eta - 1)\frac{\rho}{\lambda + \rho}.$$

and that the value of this minimum is equal to:

$$M(\eta, \rho) = 1 - (2\eta - 1)^2 \frac{\rho}{\lambda + \rho}.$$

From this we deduce that for all $(\rho, \theta) \in (0, +\infty) \times [-1, 1]$:

$$\psi_\rho(\theta) = \frac{\rho}{\lambda + \rho}\theta^2$$

whose inverse function is

$$\psi_\rho^{-1}(u) = \sqrt{\left(1 + \frac{\lambda}{\rho}\right)u}. \tag{62}$$

Remark 33 *The minimum of $C_{\eta,\rho}$ begin reached on $(-1, 1)$ for any $(\eta, \rho) \in [0, 1] \times (0, +\infty)$, the result would be identical for any convex loss function ϕ' that is equal to $(1-\alpha)^2$ on $(-\infty, 1)$. Indeed, the corresponding function $C'_{\eta,\rho}$ would coincide with $C_{\eta,\rho}$ on $(-1, 1)$ and would be no smaller than $C_{\eta,\rho}$ outside of this interval; it would therefore have the same minimal value reached at the same point, and consequently the same function M and ψ. This is for example the case with the loss function used in LS-SVM, $\phi'(\alpha) = (1-\alpha)^2$*

We can now summarize the upper bounds on the excess-risk obtained for $1 - class$ and $2 - class$ SVM.

Theorem 34 *Let $\phi_1(\alpha) = \max(1-\alpha, 0)$ and $\phi_2(\alpha) = \max(1-\alpha, 0)^2$. Then for any distribution P with low density exponent γ, there exist constant $(K_1, K_2, r_1, r_2) \in (0, +\infty)^4$ such that for any $f \in \mathcal{M}$ with an excess regularized ϕ_1-risk upper bounded by r_1 the following holds:*

$$R(f) - R^* \leq K_1 \left(R_{\phi_1,0}(f) - R^*_{\phi_1,0}\right)^{\frac{\gamma}{2\gamma+1}},$$

and if the excess regularized ϕ_2-risk upper bounded by r_2 the following holds:

$$R(f) - R^* \leq K_2 \left(R_{\phi_2,0}(f) - R^*_{\phi_2,0}\right)^{\frac{\gamma}{2\gamma+1}},$$

Proof Starting with $\phi_1(\alpha) = \max(1-\alpha, 0)$, let us follow the proof of theorem 32 by taking $\beta = \nu = 1/2$ and $\kappa = 2\sqrt{\lambda}$. For r defined as in (57), let us choose

$$r_1 = \min\left(r, \left(\frac{c_2\lambda^{\gamma+\nu}}{\kappa 2^{\frac{\gamma+\nu}{\beta}}}\right)^{\frac{1}{\beta+\gamma+\nu}}\right).$$

For a function $f \in \mathcal{M}$, choosing ϵ as in (58), $\delta \leq r_1$ implies

$$\begin{aligned}\delta &\leq \left(\frac{c_2\lambda^{\gamma+\nu}}{\kappa 2^{\frac{\gamma+\nu}{\beta}}}\right)^{\frac{1}{\beta+\gamma+\nu}} \\ &= \left(\epsilon^{-(\gamma+\nu)} 2^{-\frac{\gamma+\nu}{\beta}} \lambda^{\gamma+\nu}\delta^{\beta}\right)^{\frac{1}{\beta+\gamma+\nu}}\end{aligned}$$

and therefore:

$$\delta 2^{-\frac{1}{\beta}} \leq \frac{\lambda}{\epsilon}.$$

This ensures by (61) that for $u = \delta 2^{-\frac{1}{\beta}}$, one indeed has

$$\psi_\rho^{-1}(u) = \kappa u^{\beta}\epsilon^{-\nu},$$

which allows the rest of the proof, in particular (57), to be valid. This proves the result for ϕ_1, with

$$K_1 = 2 \times 2^{\frac{2\gamma}{2\gamma+1}} \lambda^{\frac{\gamma}{2\gamma+1}} c_2^{\frac{1}{2\gamma+1}}.$$

For $\phi_2(\alpha) = \max(1-\alpha, 0)^2$ we can observe from (62) that, for any $\epsilon \in (0, \epsilon_0]$,

$$\psi_\epsilon^{-1}(u) \leq \sqrt{(\lambda+\epsilon_0)\frac{u}{\rho}}.$$

and the proof of theorem 32 leads to the claimed result with $r_2 = r$ defined in (57), and

$$K_2 = 2 \times (\lambda+\epsilon_0)^{\frac{\gamma}{2\gamma+1}} c_2^{\frac{1}{2\gamma+1}}.$$

■

Remark 35 *We note here that ϵ can be chosen as small as possible in order to move the constant K_2 as close as possible to its lower bound:*

$$\bar{K}_2 = 2 \times \lambda^{\frac{\gamma}{2\gamma+1}} c_2^{\frac{1}{2\gamma+1}}.$$

but the counterpart of decreasing K_2 is to decrease r_2 too, by (57). We also notice the constant corresponding to the 1-SVM loss function is larger than that of the 2-SVM loss function, by a factor of up to $2^{\frac{2\gamma}{2\gamma+1}}$

8. Consistency of One-class SVM for Density Level Set Estimation

In this section we focus on the one-class case: η is identically equal to 1, and P is just considered as a distribution on $\mathbb{R}^d$. The aim is to estimate a density level set of level μ, for some $\mu > 0$:

$$C_\mu := \left\{ x \in \mathbb{R}^d \ : \ \rho(x) \geq \mu \right\} \tag{63}$$

The estimator that is considered here is the plug-in density level set estimator associated with $\hat{f}_\sigma$, denoted by $\widehat{C}_\mu$:

$$\widehat{C}_\mu := \left\{ x \in \mathbb{R}^d \ : \ 2\lambda \hat{f}_\sigma(x) \geq \mu \right\} \tag{64}$$

Recall that the asymptotic behaviour of $\hat{f}_\sigma$ in the one-class case is given in Theorem 8: $\hat{f}_\sigma$ converge to ρ_λ, which is proportionnal to the density ρ truncated at level 2λ. Taking into account the behaviour of ρ_λ, we only consider the situation where $0 < \mu < 2\lambda < \sup(\rho) = M$. ρ is still assumed to have a compact support $S \subset \mathcal{X}$. To assess the quality of $\widehat{C}_\mu$, we use the so-called *excess mass* functional, first introduced by Hartigan (1987), which is defined for any subset C of $\mathbb{R}^d$ as follows:

$$H_P(C) := P(C) - \mu \mathrm{Leb}(C) \ . \tag{65}$$

Note that H_P is defined with respect to both P and μ, and that it is maximized by C_μ. Hence, the quality of an estimate $\widehat{C}$ depends here on how its excess mass is close to this of C_μ.
The following lemma relates the $L2$ convergence of a density estimator to the consistence of the associated plug-in density level set estimator, with respect to the excess mass criterion:

Lemma 36 *Let P be a probability distribution on $\mathbb{R}^d$ with compact support $S \subset \mathcal{X}$. Assume that P is absolutely continuous with respect to the Lebesgue's measure, and let ρ denote its associated density function. Consider a density estimate $\hat{\rho}$ defined on $\mathbb{R}^d$. Then the following holds*

$$H_P(C_\mu) - H_P(\widehat{C}) \leq K_5 \| \hat{\rho} - \rho \|_{L_2} \ , \tag{66}$$

where $\widehat{C}$ is the level set of $\hat{\rho}$ at level μ, and K_5 is a positive constant depending only on μ and on ρ.

Proof To proof the lemma, it is convenient to first build an artificial classification problem using the density function ρ and the desired level μ, then to relate the excess-risk involved in this classification problem to the excess-mass involved in the original one-class problem. Let us consider the following joint distribution Q defined by its marginal density function

$$q(x) = \begin{cases} m\rho(x) + (1-m)\frac{1}{\mathrm{Leb}(S)} & \text{if } x \in S \ , \\ 0 & \text{otherwise} \ , \end{cases} \tag{67}$$

and by its regression function

$$\eta'(x) = \frac{m\rho(x)}{m\rho(x) + (1-m)\frac{1}{\mathrm{Leb}(S)}} \ , \tag{68}$$

where m is chosen such that

$$\eta'(x) = \frac{\rho(x)}{\rho(x) + \mu} \ , \tag{69}$$

that is

$$m = \frac{1}{1 + \mu \mathrm{Leb}(S)} \ . \tag{70}$$

In words, in the above artificial classification problem, the initial distribution P stands for the marginal distribution of the positive class, and the negative class is generated by the uniform distribution over the support of P. The mixture coefficient m is determined by the initialy desired density level μ. The corresponding Bayes classifier, which is the plug-in rule associated with η', is denoted by h^*.
Furthermore let us define $\hat{\eta}' = \hat{\rho}/\left(\hat{\rho} + \mu\right)$, which stands for an estimate of η' in our artificial classification problem, and $\hat{h}$ as the plug-in classifier associated with $\hat{\eta}'$: $\hat{h} = \mathrm{sign}\left(2\hat{\eta}' - 1\right)$. Then it is straightforward that h^* is the indicator function of C_μ, and that $\hat{h}$ is the indicator function of $\widehat{C}$. Moreover

$$R(\hat{h}) - R(h^*) = m\left(H_P(C^*) - H_P(\widehat{C})\right) \ .$$

Indeed,

$$\begin{aligned} R(\hat{h}) &= Q(\hat{h}(X) \neq Y) \\ &= Q(Y=-1)Q(\hat{h}(X)=1|Y=-1) + Q(Y=1)Q(\hat{h}(X)=-1|Y=1) \\ &= (1-m)\frac{\mathrm{Leb}\left(\widehat{C}\right)}{\mathrm{Leb}\left(S\right)} + m(1 - P(\widehat{C})) \ , \end{aligned}$$

and, similarly,

$$R(h^*) = (1-m)\frac{\mathrm{Leb}\left(C_\mu\right)}{\mathrm{Leb}\left(S\right)} + m(1 - P(C_\mu)) \ , \tag{71}$$

which prooves the claim.
Now, the following can be derived, starting from an equality that can be found in Devroye et al. (1996):

$$\begin{aligned} R(\hat{h}) - R(h^*) &= 2\mathbb{E}_Q\left[\left|\eta' - \frac{1}{2}\right| 1_{\hat{h} \neq h^*}\right] \\ &\leq 2\mathbb{E}_Q\left[\left|\eta' - \hat{\eta}'\right|^2\right]^{1/2} \\ &= 2\mu\left(\int_{\mathbb{R}^d}\left(\frac{|\hat{\rho}(x) - \rho(x)|}{(\hat{\rho}(x)+\mu)(\rho(x)+\mu)}\right)^2 q(x)dx\right)^{1/2} \\ &\leq 2\mu\sqrt{A}\left(\int_{\mathbb{R}^d}\left(\frac{|\hat{\rho}(x) - \rho(x)|}{(\hat{\rho}(x)+\mu)(\rho(x)+\mu)}\right)^2 dx\right)^{1/2} \\ &\leq 2\frac{\sqrt{A}}{\mu}\| \hat{\rho} - \rho \|_{L_2} \ , \end{aligned}$$

where A is a positive uniform upper bound on $q(x)$. Combining the previous equality with the last inequality concludes the proof. ■

We could just directly apply this lemma to $\hat{f}_\sigma$, ρ_λ and the distribution associated with ρ_λ, but this would not give the consistency of $\hat{f}_\sigma$ with respect to the excess mass H_P. The following lemma implies that the plug-in density level set estimator at level $0 < \mu < 2\lambda$ based on the one-class SVM estimator is indeed consistent with respect to the excess mass defined with P.

Theorem 37 *Let $\hat{f}$ be a squared integrable function that estimates ρ_λ (as defined in Equation 9). Let $0 < \mu < 2\lambda$. Let $\widehat{C}$ denote the level set of $2\lambda\hat{f}$ at level μ. Then*

$$H_P(C_\mu) - H_P(\widehat{C}) \leq K_6 \| \hat{f} - \rho_\lambda \|_{L_2} \tag{72}$$

where $K_6 > 0$ depends neither on σ, nor on n.

Proof Let us introduce the following density estimator:

$$\hat{\rho} = 2\lambda\hat{f}_\sigma + \tilde{\rho}_\lambda \ , \tag{73}$$

where the function $\tilde{\rho}_\lambda$ is defined as follows:

$$\tilde{\rho}_\lambda = \begin{cases} \rho(x) - 2\lambda & \text{if } \rho(x) \geq 2\lambda \ , \\ 0 & \text{otherwise,} \end{cases} \tag{74}$$

and let $\tilde{C}$ denote its associated plug-in density level set estimate at level μ. It can be checked that $\hat{\rho} - \rho = 2\lambda\left(\hat{f} - \rho_\lambda\right)$, implying that

$$\| \hat{\rho} - \rho \|_{L_2} = 2\lambda \| \hat{f} - \rho_\lambda \|_{L_2} \ . \tag{75}$$

Hence, using Lemma 36, we have

$$H_P\left(C_\mu\right) - H_P\left(\tilde{C}\right) \leq 2\lambda c \| \hat{f}_\sigma - \rho_\lambda \|_{L_2} \ , \tag{76}$$

leading to

$$H_P(C_\mu) - H_P(\widehat{C}) \leq 2\lambda c \| \hat{f}_\sigma - \rho_\lambda \|_{L_2} + \left| H_P(\widehat{C}) - H_P(\tilde{C}) \right| \ . \tag{77}$$

The last thing to do is to bound $\left| H_P(\widehat{C}) - H_P(\tilde{C}) \right|$. Since P has a bounded density w.r.t. the Lebesgue's measure,

$$\left| H_P(\widehat{C}) - H_P(\tilde{C}) \right| \leq (\mu + M)\text{Leb}\left(\widehat{C}\Delta\widehat{C}_\mu\right) \ . \tag{78}$$

By construction, if $C_{2\lambda}$ denotes the density level set at level 2λ, we have $\widehat{C} \cap \overline{C_{2\lambda}} = \tilde{C} \cap \overline{C_{2\lambda}}$ and $2\lambda\hat{f} \geq \mu \implies \hat{\rho} \geq \mu$. Hence

$$\begin{aligned}
\mathrm{Leb}\left(\widehat{C}\Delta\widehat{C}_\mu\right) &= \int_{C_{2\lambda}} 1_{\{2\lambda\hat{f}<\mu \,\wedge\, \hat{\rho}\geq\mu\}} \\
&\leq \int_{C_{2\lambda}} 1_{\{2\lambda\hat{f}<\mu\}} \\
&\leq \int_{C_{2\lambda}} \frac{2\lambda - 2\lambda\hat{f}}{2\lambda-\mu} 1_{\{2\lambda\hat{f}<\mu\}} \\
&\leq \frac{1}{2\lambda-\mu}\left(\int_{C_{2\lambda}} \left(2\lambda\rho_\lambda - 2\lambda\hat{f}\right)^2\right)^{1/2} \\
&\leq \frac{2\lambda}{2\lambda-\mu} \| \hat{f} - \rho_\lambda \|_{L_2} .
\end{aligned}$$

This concludes the proof. ■

Acknowledgments

The authors are grateful to Stéphane Boucheron, Pascal Massart and Ingo Steinwart for fruitful discussions and advices. They also want to thank an anonymous reviewer for pointing out mistakes in a previous version.

References

N. Aronszajn. Theory of reproducing kernels. *Trans. Am. Math. Soc.*, 68:337 – 404, 1950.

P. L. Bartlett, O. Bousquet, and S. Mendelson. Local rademacher complexities. *Annals of Statistics*, 2005. To appear.

P. L. Bartlett and A. Tewari. Sparseness vs estimating conditional probabilities: Some asymptotic results. In *Lecture Notes in Computer Science*, volume 3120, pages 564–578. Springer, 2004.

P.I. Bartlett, M.I. Jordan, and J.D. McAuliffe. Convexity, classification and risk bounds. Technical Report 638, UC Berkeley Statistics, 2003.

B. E. Boser, I. M. Guyon, and V. N. Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the 5th annual ACM workshop on Computational Learning Theory*, pages 144–152. ACM Press, 1992. URL `http://www.clopinet.com/isabelle/Papers/colt92.ps.Z`.

R. A. DeVore and G. G. Lorentz. *Constructive Approximation*. Springer Grundlehren der Mathematischen Wissenschaften. Springer Verlag, 1993.

L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*, volume 31 of *Applications of Mathematics*. Springer, 1996.

L. Devroye and G. Lugosi. *Combinatorial Methods in Density Estimation*. Springer Series in Statistics. Springer, 2000.

J. A. Hartigan. Estimation of a convex density contour in two dimensions. *J. Amer. Statist. Assoc.*, 82(397):267–270, 1987.

V. Koltchinskii. Localized rademacher complexities. Manuscript, september 2003.

G. Lugosi and N. Vayatis. On the Bayes-risk consistency of regularized boosting methods. *Ann. Stat.*, 32:30–55, 2004.

E. Mammen and A. Tsybakov. Smooth discrimination analysis. *Ann. Stat.*, 27(6):1808–1829, 1999.

P. Massart. Some applications of concentration inequalities to statistics. *Ann. Fac. Sc. Toulouse*, IX(2):245–303, 2000.

M. T. Matache and V. Matache. Hilbert spaces induced by Toeplitz covariance kernels. In *Lecture Notes in Control and Information Sciences*, volume 280, pages 319–334. Springer, Jan 2002.

B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson. Estimating the support of a high-dimensional distribution. *Neural Comput.*, 13:1443–1471, 2001.

B. W. Silverman. On the estimation of a probability density function by the maximum penalized likelihood method. *Ann. Stat.*, 10:795–810, 1982.

I. Steinwart, , D. Hush, and C. Scovel. An explicit description of the reproducing kernel Hilbert spaces of Gaussian RBF kernels. Technical Report LA-UR 04-8274, Los Alamos National Laboratory, 2004.

I. Steinwart. Support vector machines are universally consistent. *J. Complexity*, 18:768–791, 2002.

I. Steinwart. Sparseness of support vector machines. *J. Mach. Learn. Res.*, 4:1071–1105, 2003.

I. Steinwart and C. Scovel. Fast rates for support vector machines using gaussian kernels. Technical report, Los Alamos National Laboratory, 2004. submitted to Annals of Statistics.

A.N. Tikhonov and V.Y. Arsenin. *Solutions of ill-posed problems*. W.H. Winston, Washington, D.C., 1977.

A. B. Tsybakov. On nonparametric estimation of density level sets. *Ann. Stat.*, 25:948–969, June 1997.

T. Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *Ann. Stat.*, 32:56–134, 2004.