# Leave-one-out Bounds for Kernel Methods

Tong Zhang
IBM T.J. Watson Research Center
Yorktown Heights, NY 10598
tzhang@watson.ibm.com

**Abstract**

In this paper, we study leave-one-out style cross-validation bounds for kernel methods. The essential element in our analysis is a bound on the parameter estimation stability for regularized kernel formulations. Using this result, we derive bounds on expected leave-one-out cross-validation errors, which lead to expected generalization bounds for various kernel algorithms. In addition, we also obtain variance bounds for leave-one-out errors. We apply our analysis to some classification and regression problems, and compare them with previous results.

## 1 Introduction

Kernel methods such as Gaussian processes for regression and support vector machines for classification have become popular recently. Although in effect these methods may use infinite dimensional features in the corresponding reproducing kernel Hilbert spaces (RKHS), the kernel representation makes the computation feasible. An important aspect of kernel methods is their good generalization abilities despite of their large underlying feature spaces. This means that these learning methods can accurately predict outputs associated with previously unobserved data.

A popular method to study such generalization ability is the so-called Vapnik-Chervonenkis (VC) style analysis Vapnik (1998). This method depends on the uniform convergence of observed errors of the hypothesis family to their true errors. The rate of uniform convergence depends on an estimate of certain sample-dependent covering numbers (growth numbers) for the underlying hypothesis family. More recently, other related techniques from the empirical process theory have also been explored (for example, see van de Geer (2000); van der Vaart and Wellner (1996)). Although this framework is quite general and powerful, they also have various disadvantages. For example, the derived generalization bounds can be loose.

Because of various disadvantages of the VC-style analysis, other methods to estimate generalization performance have been introduced. One interesting idea is to bound the leave-one-out error of a learning algorithm. This is useful since if the training data are independently drawn from a fixed underlying distribution, then the expected leave-one-out

error equals the expected test error, which can be used to measure the generalization ability of the learning method.

Leave-one-out bounds have received much attention recently. For example, see Forster and Warmuth (2000); Jaakkola and Haussler (1999); Joachims (2000); Kearns and Ron (1999); Vapnik (1998) and references therein. Also in Jaakkola and Haussler (1999); Joachims (2000); Vapnik (1998), the leave-one-out analysis has already been employed to study the generalization ability of support vector classification.

In this paper, we extend their results by deriving a general leave-one-out bound for a class of convex dual kernel learning machines and apply it to classification and regression problems. We compare our bounds with some existing results. Part of this work has been presented at the computational learning theory conference in 2001 Zhang (2001). The current version is more detailed and contains a number of improved results.

We organize the paper as follows. In Section 2, we outline the relationship of the generalization ability of a learning algorithm and its leave-one-out error. The motivation of leave-one-out analysis is then presented in the context of previous work. In Section 3, we present the general kernel learning machine formulation and review the corresponding RKHS representation. We then derive a general leave-one-out bound for the estimated parameter that is the foundation of our analysis. This analysis is applied in Section 4 to formulations with bounded sub-gradients. Section 5 and Section 6 contain additional results of this analysis on some classification and regression problems. Concluding remarks are given in Section 7.

# 2 Leave-one-out analysis and related works

In supervised learning, we want to predict an unobserved output value $y$ based on an observed input vector $x$. This requires us to estimate a functional relationship $y \approx p(x)$ from a set of training examples. Usually the quality of the predictor $p(x)$ can be measured by a loss function $L(p(x), x, y)$. Our goal is to find $p(x)$ so that the expected true loss (risk) of $p$ defined below is as small as possible:

$$Q_L(p(\cdot)) = E_{x,y} L(p(x), x, y), \tag{1}$$

where we use $E_{x,y}$ to denote the expectation with respect to the true (but unknown) underlying distribution $D$.

We assume that the observed training data $D_n = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ are independently drawn from the unknown underlying distribution $D$. A learning algorithm is a procedure $\mathcal{A}$ that takes the training samples $D_n$ as the input and produces a predictor $\hat{p} = \mathcal{A}(D_n)$ as the output. The quality of a learning algorithm can be measured by the true risk $Q_L(\hat{p}(\cdot))$ of the learned predictor. In the literature, the value $Q_L(\hat{p}(\cdot))$ is often referred to as the generalization error (with respect to a loss function $L$). Clearly the generalization error $Q_L(\hat{p}(\cdot))$ is a random variable that depends on the training data $D_n$, and a fundamental problem in learning theory is to understand the behavior of this random variable for a learning algorithm. In this paper, we are mostly interested in the expected generalization error

$$Q_L(\mathcal{A}, n) = E_{D_n} Q_L(\hat{p}(\cdot)) = E_{D_n} Q_L(\mathcal{A}(D_n)), \tag{2}$$

where the expectation $E_{D_n}$ is with respect to the training data $D_n$. Clearly this is a very natural quantity that characterizes the performance of a learning method $\mathcal{A}$.

Consider $n+1$ samples $D_{n+1} = \{(x_1, y_1), \ldots, (x_{n+1}, y_{n+1})\}$. Let $D^{(i)}_{n+1}$ be the subset of $D_{n+1}$ with the $i$-th datum removed:

$$D^{(i)}_{n+1} = \{(x_1, y_1), \ldots, (x_{i-1}, y_{i-1}), (x_{i+1}, y_{i+1}), \ldots, (x_{n+1}, y_{n+1})\}.$$

Consider a learning algorithm $\mathcal{A}$, and let $\hat{p}^{(i)} = \mathcal{A}(D^{(i)}_{n+1})$ be the predictor obtained from this algorithm based on $D^{(i)}_{n+1}$. Then the leave-one-out error on sample $D_{n+1}$ can be defined as:

$$Z_L(\mathcal{A}, D_{n+1}) = \frac{1}{n+1}\sum_{i=1}^{n+1} L(\hat{p}^{(i)}(x_i), x_i, y_i).$$

Taking expectation with respect to the $n+1$ samples $D_{n+1}$, and observe that each sample $(x_i, y_i)$ is drawn from the same underlying distribution $D$, we obtain the following equality for the expected leave-one-out error:

$$\begin{aligned} Z_L(\mathcal{A}, n+1) =& E_{D_{n+1}} Z_L(\mathcal{A}, D_{n+1}) \\ =& \frac{1}{n+1}\sum_{i=1}^{n+1} E_{D_{n+1}} L(\hat{p}^{(i)}(x_i), x_i, y_i) \\ =& \frac{1}{n+1}\sum_{i=1}^{n+1} E_{D^{(i)}_{n+1}} Q_L(\hat{p}^{(i)}(\cdot)) = \frac{1}{n+1}\sum_{i=1}^{n+1} Q_L(\mathcal{A}, n) = Q_L(\mathcal{A}, n). \end{aligned}$$

That is, the expected generalization error equals the expected leave-one-out error (with one more training sample). Therefore estimates of the leave-one-out error of an algorithm can be used to bound its generalization performance. The purpose of this paper is to obtain such bounds for kernel learning methods.

In the standard machine learning analysis, one often assumes that the predictor $\hat{p}$ is taken from a hypothesis function class $C$ that models the relationship of the input $x$ and the output $y$. A frequently studied learning algorithm is the empirical risk minimization (ERM) method, where we find a predictor in $C$ that minimizes the empirical risk:

$$\hat{p} = \mathcal{A}(D_n) = \arg\inf_{p \in C} \frac{1}{n}\sum_{i=1}^{n} L(p(x_i), x_i, y_i). \tag{3}$$

This formulation is related to a different form of penalized learning formulation often used in practical computation. Assume $C$ belongs to a large function space $H$ (which is often dense in the set of continuous functions), and is defined by the numerical constraint: $C = \{p \in H : r(p) \leq c_0\}$, where $r : H \to R$ is a functional called regularization condition. By introducing a Lagrangian multiplier $\lambda_n$ ($\geq 0$), we can rewrite (3) as

$$\mathcal{A}(D_n) = \arg\inf_{p \in H} \left[\frac{1}{n}\sum_{i=1}^{n} L(p(x_i), x_i, y_i) + \lambda_n r(p)\right]. \tag{4}$$

For kernel methods, $H$ is a Hilbert function space (see Section 3) with norm $\|\cdot\|$, and $r(p) = \frac{1}{2}\|p(\cdot)\|^2$.

The purpose of this paper is to analyze kernel methods of the penalized form (4), and our bounds will be specified in terms of the regularization parameter $\lambda_n$. Note that from a theoretical point of view, (3) and (4) are not exactly the same since the theoretical analysis of (3) usually assumes that $C$ is fixed while the theoretical analysis of (4) usually assumes that $\lambda_n$ is fixed.

Kernel methods that are related to what we consider here have been studied both in non-parametric statistics and in machine learning. We shall discuss related theoretical results as well as the contribution of the current work.

One particular but very important kernel formulation is least squares regression where $L(p, x, y) = (p - y)^2$. In machine learning, this method is also referred to as a Gaussian process, which has a natural Bayesian interpretation. In statistics, this formulation often appears as smoothing splines, which are standard non-parametric regression models Wahba (1990).

Most theoretical results in non-parametric statistics are based on the non-penalized formulation (3). Denote by $\Delta Q_L(\mathcal{A}, n) = Q_L(\mathcal{A}, n) - \inf_{p \in C} Q_L(p)$, which measures how good a learning algorithm (that chooses predictors from $C$) performs when compared to the best possible predictor in $C$. One important concept is *minimax risk*: if we know that the unknown target function which minimizes the risk belongs to $C$, but the observation $y$ is corrupted with noise, then one can measure the quality of a learning algorithm using its worst case $\Delta Q_L(\mathcal{A}, n)$ value among all possible target functions in $C$; the minimax risk is then defined as the lowest possible worst case $\Delta Q_L(\mathcal{A}, n)$ value among all possible learning algorithms $\mathcal{A}$.

In non-parametric statistics one is specifically interested in how fast the minimax risk converges to zero as $n \to \infty$, and what algorithm achieves the fastest possible rate of convergence. For the least squares loss, it turns out that under mild regularity conditions, the optimal rate of convergence can be determined by the $L_2$-metric entropy $H(C, \epsilon)$ of the function class $C$ (cf. Yang and Barron (1999) and reference therein).[1] Consider an $L_2$ metric $d$ defined with respect to the input distribution as $d(p_1, p_2) = E_x^{1/2}(p_1(x) - p_2(x))^2$, then the $L_2$-metric entropy $H(C, \epsilon)$ is defined as the logarithm of the smallest number of $L_2$-balls (in this $d$-metric) with radius $\epsilon$ that can cover $C$.

Specifically, if the $L_2$-metric entropy of the function class is of the order $\epsilon^{-\rho}$ ($\rho > 0$), then the minimax risk of the optimal predictor converges to zero at a rate of the order $n^{-2/(2+\rho)}$. It is also known that if $\rho > 2$ then ERM over the whole function class $C$ as in (3) may converge slower than the optimal minimax rate (and thus will be a suboptimal learning method), see Birgé and Massart (1993). However when $0 < \rho < 2$, by using the now standard proof techniques (such as chaining and ratio uniform convergence inequalities) from the theory of empirical processes, it can be shown that the empirical risk minimization method (3) achieves the optimal minimax rate when some additional regularity conditions are satisfied.[2] The details are not important for this paper and we refer the interested readers to Birgé

[1] For simplicity, we shall assume that both $p(x)$ and $y$ are bounded in this discussion.

[2] For example this is true when the uniform (or bracketing) entropy number (see van der Vaart and Wellner (1996) Section 2.5 for definitions) has the same order of complexity as the $L_2$ metric entropy. The differences among these definitions are technical and irrelevant

and Massart (1993); van de Geer (2000); van der Vaart and Wellner (1996). Since it can be shown that the uniform metric entropy for kernel function classes do not grow faster than $\epsilon^{-2}$, at least for the least squares method, it is possible to use the standard empirical process theory to obtain convergence bounds for the ERM kernel formulation (3) that approximately matches the optimal minimax rate.[3]

For example, a family of well-studied kernel function classes are 1-dimensional smoothing splines on $[0,1]$. Here the function class $C$ is a subset of $r$-th differentiable functions ($r > 0.5$): $C = \{p : \int_0^1 [p^{(r)}(x)]^2 dx \leq c_0\}$, where $p^{(r)}$ denotes the $r$-th derivative of function $p$. It is known that the uniform entropy number is of the order $H(C,\epsilon) = O(\epsilon^{-1/r})$, leading to the minimax rate of the order $n^{-2r/(2r+1)}$ which can be achieved using the empirical risk minimization method.

In the learning theory literature, the empirical risk minimization method (3) for kernel methods has recently been studied by a number of authors (for example Chucker and Samle (2002); Evgeniou et al. (2000)). Many researchers are not fully aware of the recent developments in the statistical community mentioned above. Although these studies can lead to useful insights, the obtained bounds are not necessarily tight, and often do not match the correct minimax rates when applied to specific kernel formulations.

Although the empirical risk minimization method for (3) has been widely studied and the minimax issue is well-understood, the behavior of the regularized learning formulation (4) is less studied. From a technical point of view, the main difference is that in (4) we would like to specify our learning bounds in terms of $\lambda_n$ rather than through the restricted function class $C$. However, the transition from (3) to (4) is not straight-forward.

For the penalized formulation (4), it is natural to compare the expected generalization error $Q_L(\mathcal{A}, n)$ to the best possible risk $\inf_{p \in H} Q_L(p)$ using predictors from $H$. Since the predictor is taken from a very large function class $H$ (often dense in the set of continuous functions) with possibly $\infty$ metric entropy number, the entropy-dependent lower bound in the minimax analysis indicates that a direct empirical risk minimization over $H$ will not converge. The term $\lambda_n r(p)$ is needed to stabilize the estimation process but it introduces an explicit bias. Therefore bounds for (4) will contain both bias and learning complexity terms. From the theoretical point of view, it is important to study the trade-off between the bias and the learning complexity. For example, to obtain an estimator that is consistent, it is necessary to allow the bias to converge to zero by choosing $\lambda_n \to 0$ when $n \to \infty$. However, it is also necessary to make sure that $\lambda_n \to 0$ at a relatively slow pace so that the learning complexity vanishes in the limit (otherwise the method overfits the training data). Therefore an important aspect of the analysis is to understand the behavior of the learning complexity when $\lambda_n \to 0$.

Recently various stability based methods have been proposed to study the behavior of

to this paper. It is worth mentioning though, that for kernel methods, the uniform entropy number is bounded, and (by definition) has the same order of complexity as the worst case metric entropy.

[3]Although convergence rates for kernel methods obtained from empirical process techniques are often relatively tight in terms of their dependency on the sample size $n$, they almost always contain very bad constants. These bounds can also be loose relative to some other factors such as function class size and noise size etc.

the penalized learning formulation (4) Bousquet and Elisseeff (2002); Kutin and Niyogi (2002); Zhang (2002a). The underlying idea is closely related to the analysis of leave-one-out error presented in this paper. Exponential type probability bounds can be obtained in their analysis. Such exponential bounds give more detailed information than the expected generalization error bound defined in (2) which we are interested in here. However one pays a price with this generality since the expected generalization error implied from such an analysis will be worse than results obtained here using the leave-one-out analysis.

As an example, we shall still consider the least squares regression problem. It was pointed out in Zhang (2002a) that the bound given in Bousquet and Elisseeff (2002) was not tight, where it was shown that a learning complexity term of the order $O(1/\sqrt{\lambda_n^2 n})$ in Bousquet and Elisseeff (2002) can be improved to $O(1/\lambda_n^2 n)$. It was further pointed out in Zhang (2002a) that even the improved bound derived there (when averaged to obtain an expected generalization error bound) does not match the corresponding expected generalization bound using the leave-one-out analysis, where the learning complexity is of the order $O(1/\lambda_n n)$ (see Section 5). If the target function belongs to $H$, then we can choose $\lambda_n \sim n^{-1/2}$, and results in Section 5 imply that the expected generalization error $Q_L(A, n)$ converges to the risk of the target function $\inf_{q \in H} Q_L(p)$ at a rate of the order $O(n^{-1/2})$. This is clearly the best kernel independent bound (in terms of $n$) one can obtain since it matches the minimax rate of smoothing splines when the smoothing parameter $r \to 0.5$. As a comparison, even with the refined analysis in Zhang (2002a), one can only obtain a rate of the order $O(n^{-1/3})$ by setting $\lambda_n \sim n^{-1/3}$. Similar conclusion also holds for other loss functions. See Section 4 and 5 for more detailed comparisons.

One question is whether this discrepancy is due to the suboptimal proof-techniques used in the literature or it is due to something more fundamental. Although we do not have an affirmative answer to this question, we think that the $O(1/\lambda_n^2 n)$ behavior for the probability bounds might not be improvable. One may notice that the leave-one-out variance bounds given in Section 3 and 4 also have this $O(1/\lambda_n^2 n)$ behavior. In addition to this worse behavior in terms of the regularization parameter $\lambda_n$, probability bounds such as those in Bousquet and Elisseeff (2002); Kutin and Niyogi (2002); Zhang (2002a) often require the loss function to be bounded. Again similar conditions are needed in the leave-one-out variance analysis presented in Section 3 and 4. As a comparison, this restriction is not necessary in the leave-one-out expected generalization analysis.

Based on the above discussion, it is clear that the leave-one-out expected generalization error analysis is very useful for understanding the behavior of the penalized kernel learning formulation (4) which is widely used in practice. In fact it is possible to obtain tight expected generalization bounds both for the kernel dependent case (where we need to use the eigen-decomposition structure of the underlying kernel) and for the kernel independent case. However, in order to focus on the main underlying methodology, we shall only consider the kernel independent case in this paper.

# 3 Kernel learning machines and leave-one-out approximation bound

## 3.1 Kernel representation

The goal of many machine learning problems is to find a function that can predict the output variable $y$ based on the input variable $x$. Typically, one needs to restrict the size of the hypothesis function family so that a stable estimate within the function family can be obtained from a finite number of samples. Let the training samples be $(x_1, y_1), \ldots, (x_n, y_n)$. We assume that the hypothesis function family that predicts $y$ based on $x$ can be specified with the following kernel method:

$$p(\alpha, x) = \sum_{i=1}^{n} \alpha_i K(x_i, x), \tag{5}$$

where $\alpha = [\alpha_i]_{i=1,\ldots,n}$ is a parameter vector that needs to be estimated from the data. $K$ is a symmetric positive kernel. That is, $K(a, b) = K(b, a)$, and the $n \times n$ Gram matrix $G = [K(x_i, x_j)]_{i,j=1,\ldots,n}$ is always positive semi-definite.

**Definition 3.1** *Let $H_0 = \{\sum_{i=1}^{\ell} \alpha_i K(x_i, x) : \ell \in N, \alpha_i \in R\}$. $H_0$ is an inner product space with norm defined as*

$$\|\sum_i \alpha_i K(x_i, \cdot)\| = (\sum_{i,j} \alpha_i \alpha_j K(x_i, x_j))^{1/2}.$$

*Let $H$ be the closure of $H_0$ under the norm $\|\cdot\|$, which forms a Hilbert space, called the reproducing kernel Hilbert space of $K$.*

It is well-known and not difficult to check that the norm $\|\cdot\|$ in Definition 3.1 is well-defined, and it defines an inner product. Consider two functions in $H_0$: $p_1(x) = \sum_{i=1}^{n} \alpha_i K(x_i, x)$ and $p_2(x) = \sum_{j=1}^{m} \beta_j K(x'_j, x)$. The inner product can be expressed as:

$$p_1 \cdot p_2 = \sum_{i=1}^{n} \sum_{j=1}^{m} \alpha_i K(x_i, x'_j) \beta_j.$$

Now assume that $p_1(\cdot) \perp p_2(\cdot)$ where $p_2(x) = K(x'_j, x)$, then

$$0 = p_1 \cdot p_2 = \sum_{i=1}^{n} \alpha_i K(x_i, x'_j) = p_1(x'_j).$$

Clearly by taking limit, we know that this property also holds for all $p_1 \in H$. We thus obtain the following well known property for reproducing kernel Hilbert spaces:

**Proposition 3.1** *Let $p(x) \in H$. Consider the projection $p_0(x)$ of $p(x)$ onto the subspace spanned by functions $p_i(x) = K(x_i, x)$ $(i = 1, \ldots, n)$. Then $p_0(x_i) = p(x_i)$ for all $i = 1, \ldots, n$.*

In the recent machine learning literature, kernel representation (5) has frequently been discussed under the assumptions of the Mercer's theorem, which gives a feature space representation of $H$ (for example, see Cristianini and Shawe-Taylor (2000), chapter 3). Here we represent each input vector $x$ as a possibly infinite dimensional feature vector $[\phi_j(x)]_{j=1,\ldots}$. The kernel becomes $K(x_1, x_2) = \sum_{j=1}^{\infty} \phi_j(x_1)\phi_j(x_2)$ and a function $p \in H$ can be represented as $p(\cdot) = \sum_{j=1}^{\infty} w_j \phi_j(\cdot)$ where $\|p\| = (\sum_{j=1}^{\infty} w_j^2)^{1/2}$. Although this representation provides useful insights, it is not technically essential for the purpose of this paper. We take a more general approach that only relies on simple properties of a positive symmetric kernel function $K$ without considering the eigen-decomposition structure of any specific kernel.

In some practical kernel learning formulations, a bias term may be included in (5), where the corresponding function space $H'$ has the form $p(\alpha, x) + b = \sum_{i=1}^{n} \alpha_i K(x_i, x) + b$. It is well-known that $H'$ can be considered as the reproducing kernel Hilbert space with kernel $K'(x_1, x_2) = K(x_1, x_2) + 1$. It is easier to see this using the feature-space representation: $K'(x_1, x_2) = 1 + \sum_{i=1}^{\infty} \phi_i(x_1)\phi_i(x_2)$, and hence any function in the reproducing kernel Hilbert space of $K'$ has the form $p(\cdot) = \sum_{i=1}^{\infty} w_i \phi_i(\cdot) + b$, with its norm defined as $(\sum_{i=1}^{\infty} w_i^2 + 1)^{1/2}$. Therefore from the learning point of view, we don't lose any representation power by only considering a kernel representation without an explicit bias term. We shall mention that by treating $H'$ as the reproducing kernel Hilbert space of a different kernel, the resulting learning formulations may be slightly different from those in the literature where the bias $b$ is typically not included in the penalization term. The explicit bias formulation will significantly complicate the derivation in the leave-one-out analysis framework, though it is still possible to handle it. Since an explicit bias formulation does not have any advantage in approximation power, for simplicity and clarity, we shall focus on the representation (5) without the bias term.

Functions in $H$ can be used to approximate an arbitrary function $p(x)$. However only certain functions can be well approximated while others cannot. Therefore it is useful to define a metric that characterizes how well a certain function can be approximated. In particular, given a sequence of observations $X_n = \{x_1, \ldots, x_n\}$, we can approximate $p(x)$ by using functions in $H$ that match the values of $p(x)$ at $X_n$. There can be many possible such function interpolations (assume one exists). Therefore a natural choice is to select the minimum $\|\cdot\|$-norm interpolation of $p(\cdot)$. The corresponding norm can be used to measure the cost of interpolating $p(x)$ at $X_n$ using functions in $H$. This leads to the following definition:

**Definition 3.2** *Denote by $X_n$ a sequence of samples $x_1, \ldots, x_n$. We use $G(X_n)$ to denote the Gram matrix $[K(x_i, x_j)]_{i,j=1,\ldots,n}$.*

*For any function $p(x)$ and symmetric positive kernel $K$, we define*

$$\|p(X_n)\| = \inf \{+\infty\} \cup \{s \geq 0 : \forall \alpha, p(X_n)^T \alpha \leq s(\alpha^T G(X_n)\alpha)^{1/2}\},$$

*where $\alpha$ denotes an $n$-dimensional vector. We use the convention that $+\infty \times 0 = +\infty$ in the definition.*

*We also define $\|p(x)\|_{[n]} = \sup_{X_n} \|p(X_n)\|$, where $X_n$ consists of a sequence of $n$ samples.*

The following property is useful. It shows that $p(X_n)$ can be considered as an interpolation of $p(\bar{\alpha}, \cdot) \in H$, and $\|p(X_n)\| = \|p(\bar{\alpha}, \cdot)\|$. Proposition 3.3 implies that for any such

interpolation function $q \in H$, $\|p(X_n)\| \leq \|q\|$. Therefore the quantity $\|p(X_n)\|$ can be regarded as the norm corresponding to the minimum $\|\cdot\|$-norm interpolation of $p(\cdot)$ using functions in $H$. The proofs are left to Appendix A.

**Proposition 3.2** *Let $K$ be a symmetric positive kernel, and consider samples $X_n = \{x_1, \ldots, x_n\}$. For any function $p(x)$, the following two situations may happen:*

- *$p(X_n)$ is not in the range of the Gram matrix $G(X_n)$: $\|p(X_n)\| = +\infty$.*
- *$p(X_n)$ can be represented as $p(X_n) = G(X_n)\bar{\alpha}$: $\|p(X_n)\| = (\bar{\alpha}^T G(X_n)\bar{\alpha})^{1/2}$.*

*In particular, if $G(X_n)$ is non-singular, then $\|p(X_n)\| = (p(X_n)^T G(X_n)^{-1} p(X_n))^{1/2}$.*

**Proposition 3.3** *Let $K$ be a symmetric positive kernel, then $\forall p(x) \in H$,*

$$\|p(\cdot)\|_{[n]} \leq \|p(\cdot)\|.$$

*This implies that $\forall x$:*

$$|p(x)| \leq \|p(\cdot)\| K(x,x)^{1/2}.$$

We will later show that the quantity $\|p(\cdot)\|_{[n]}$ characterizes the learning property of using the kernel representation (5) to approximate a target function $p(x)$. In fact, our learning bounds containing this quantity directly yield general approximation bounds, as shown at the end of Section 5. This is also a new quantity that has not been considered in the traditional approximation literature. For example, the approximation property of kernel representation is typically studied using standard analytical techniques such as Fourier analysis (see Mhaskar et al. (1999); Wendland (1998) and references therein). Such results usually depend on many rather complicated quantities as well as the data dimensionality. Compared with these previous results, our bounds using $\|p(x)\|_{[n]}$ are simpler to express and are more generally applicable. However, we do not discuss specific approximation consequences of our analysis, but rather concentrate on the general learning aspect.

When the target function $p$ is not in $H$, we do not give conditions to bound $\|p\|_{[n]}$. This quantity can be bounded using estimates of $\|p(X_n)\|$, and such estimates may be obtained using techniques in approximation theory for analyzing the stability of $G(X_n)$. The stability of the system can be measured by the condition number of $G(X_n)$ (in 2-norm) or the 2-norm of $G(X_n)^{-1}$. See Narcowich et al. (1998) and references there-in for related analysis. The quantity $\|p(X_n)\|$ is clearly related to the 2-norm of $G(X_n)^{-1}$ (the latter gives an upper bound) which measures the stability. However, the stability concept as used in approximation theory only has numerical consequences but no consequence in approximation rate or learning accuracy. On the other hand, the quantity $\|p(x)\|_{[n]}$ determines the rate of approximation (or learning) when the kernel representation (5) is used. Although it is a well-known fact that the norm of $G(X_n)^{-1}$ degrades as the number of samples increases, one may introduce smoothness conditions on $p$ so that $\|p(x)\|_{[n]}$ behaves nicely for a wide range of function families.

## 3.2 Duality and leave-one-out approximation bound for kernel machines

We consider the following general formulation of dual kernel learning machines that can be used to estimate $p(\alpha, \cdot)$ in (5) from the training data:

$$\hat{\alpha} = \arg\min_{\alpha} \left[ \sum_{i=1}^{n} g(-\alpha_i, x_i, y_i) + \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j K(x_i, x_j) \right]. \tag{6}$$

We assume that $g(a, b, c)$ is a convex function of $a$. For simplicity, we require that a solution of (6) exists, but not necessarily unique.

In order to treat classification and regression under the same general framework, we shall consider convex functions from the general convex analysis point of view, as in Rockafellar (1970). Especially we allow a convex function to take the $+\infty$ value which is equivalent to a constraint.

Consider a convex function $p(u) : R^d \to R^*$, where $R$ is the real line, and $R^*$ denotes the extended real line $R \cup \{+\infty\}$. However, we assume that convex functions do not achieve $-\infty$. We also assume that any convex function $p(u)$ in this paper contains at least one point $u_0$ such that $p(u_0) < +\infty$. Convex functions that satisfy these conditions are called *proper* convex functions. This definition is very general: virtually all practically interesting convex functions are proper. We only consider *closed* convex functions. That is, $\forall u$, $p(u) = \lim_{\epsilon \to 0^+} \inf\{p(v) : \|v - u\| \leq \epsilon\}$. This condition essentially means that the convex set above the graph of $u$: $\{(u, y) : y \geq p(u)\}$ is closed.

In this paper, we use $\nabla p(u)$ to denote a *subgradient* of a convex function $p$ at $u$, which is a vector that satisfies the following condition:

$$\forall u', \quad p(u') \geq p(u) + \nabla p(u)^T (u' - u).$$

The set of all subgradients of $p$ at $u$ is called the *subdifferential* of $p$ at $u$ and is denoted by $\partial p(u)$.

Denote by $L_n(\alpha)$ the objective function in (6):

$$L_n(\alpha) = \sum_{i=1}^{n} g(-\alpha_i, x_i, y_i) + \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j K(x_i, x_j).$$

In this paper, we assume that (6) has a solution with a finite objective value. However, we do not assume that the solution is unique. The following proposition gives a sufficient condition for the existence of such a solution. All examples given in this paper will satisfy the condition.

**Proposition 3.4** *Assume that for all $(x_i, y_i)$, $g(\alpha_i, x_i, y_i)$ is a (proper) closed convex function in $\alpha_i$. If $\lim_{\alpha_i \to \pm\infty} g(\alpha_i, x_i, y_i) = +\infty$ for all $i$, then (6) has a solution with finite value.*

*Proof.* Given $\bar{\alpha}$ such that $g(\bar{\alpha}_i, x_i, y_i) < +\infty$, the region $\Gamma_{\bar{\alpha}} = \{\alpha \in R^n : L_n(\alpha) \leq L_n(\bar{\alpha})\}$ is bounded in $R^n$. Since solving (6) in $R^n$ is equivalent to solving (6) in $\Gamma_{\bar{\alpha}}$, we can find

a sequence of $\alpha^j \in \Gamma_{\bar{\alpha}}$ such that $L_n(\alpha^j) \to \inf_\alpha L_n(\alpha)$. Since $L_n$ is the sum of closed convex functions, it is also closed. By definition, let $\hat{\alpha}$ be a limiting point of $\alpha^j$, we have $L_n(\hat{\alpha}) = \inf_\alpha L_n(\alpha)$. $\Box$

In the following, we introduce a convex duality of (6), which becomes useful in our later discussions. For function $g(u, b, c)$ in (6), we define its dual with respect to the first parameter as

$$f(v, b, c) = n \cdot \sup_u \left[uv - g(u, b, c)\right].$$

Given $f$, $g$ can be obtained as:

$$g(u, b, c) = \sup_v \left[uv - \frac{1}{n} f(v, b, c)\right]. \tag{7}$$

We consider the following primal learning formulation in the corresponding reproducing kernel Hilbert space of $K$:

$$\hat{p}(\cdot) = \arg\min_{p(\cdot)\in H} \left[\frac{1}{n}\sum_{i=1}^{n} f(p(x_i), x_i, y_i) + \frac{1}{2}\|p(\cdot)\|^2\right]. \tag{8}$$

Note that this formulation has a penalized empirical risk minimization form as in (4). We want to prove that (8) and (6) are equivalent. This is given by the following strong duality theorem (its special situation appeared in Jaakkola and Haussler (1999)). The proof is left to Appendix B.

**Theorem 3.1** *Any solution of (8) can be written as $\hat{p}(x) = \sum_{i=1}^n \hat{\alpha}_i K(x_i, x)$ for some $\hat{\alpha}$. For any solution $\hat{\alpha}$ of (6), the function $p(\hat{\alpha}, x) = \sum_{i=1}^n \hat{\alpha}_i K(x_i, x)$ is a solution of the primal optimization problem (8). The converse is also true if the Gram matrix $G(X_n)$ is non-singular.*

Note that the proof of Theorem 3.1 also implies that even when $G(X_n)$ is singular, a solution $p(x)$ of (8) can still be written as $p(\hat{\alpha}, x)$ where $\hat{\alpha}$ is a solution of (6). In addition, the sum of the optimal values of (8) and (6) is zero (for example, see Zhang (2002b)). However this fact is not important in this paper.

We may now introduce the following leave-one-out approximation bound for kernel methods, which forms the foundation of our analysis. The proof is given in Appendix C.

**Lemma 3.1** *Consider training set $D_n = \{(x_1, y_1), \ldots, (x_n, y_n)\}$. Let $\hat{\alpha}$ be the solution of (6), and let $\hat{\alpha}^{[k]}$ be the solution of (6) with the $k$-th datum $(x_k, y_k)$ removed from the training set $D_n$, then*

$$\|p(\hat{\alpha}, \cdot) - p(\hat{\alpha}^{[k]}, \cdot)\| \leq |\hat{\alpha}_k| K(x_k, x_k)^{1/2}.$$

## 3.3 Leave-one-out error and variance for kernel learning machines

As shown in Section 2, a useful aspect of leave-one-out analysis is that the expected leave-one-out error of a learning algorithm equals its expected generalization error. However the training error of a learner may not be closely related to its generalization error. It is thus interesting to compare the leave-one-out error to the training error. If the two are close, then we know that the expected generalization error is close to the expected training error. If the learner is obtained by approximately minimizing its training error, then we know that the expected generalization error is also approximately minimized. The goal of this section is to bound the leave-one-out error of a kernel learning machine in terms of its training error.

We use notations introduced in Section 2. Here we use $\mathcal{A}_K$ to denote the dual kernel learning method (6) (or equivalently, its primal learning formulation (8). $Z_L(\mathcal{A}_K, D_n)$ is used to denote the leave-one-out error with respect to the training samples $D_n$ of size $n$.

**Definition 3.3** *Let $L(p, x, y)$ be an arbitrary loss function, we define*

$$\Delta L_\delta(p, x, y) = \sup_{|t|\leq\delta} L(p + t, x, y) - L(p, x, y),$$
$$\Delta_{||} L_\delta(p, x, y) = \sup_{|t|\leq\delta} |L(p + t, x, y) - L(p, x, y)|.$$

Using Lemma 3.1, we can easily obtain the following general leave-one-out bound. We will study the consequence of this bound in regression and classification in subsequent sections.

**Theorem 3.2** *Under the assumptions of Lemma 3.1. Let $L(p, x, y)$ be an arbitrary loss function, and $Z_L(\mathcal{A}_K, D_n)$ be the corresponding leave-one-out cross-validation error with respect to $L(p, x, y)$:*

$$Z_L(\mathcal{A}_K, D_n) = \frac{1}{n}\sum_{k=1}^{n} L(p(\hat{\alpha}^{[k]}, x_k), x_k, y_k).$$

*Then*

$$n\, Z_L(\mathcal{A}_K, D_n) \leq \sum_{k=1}^{n} L(p(\hat{\alpha}, x_k), x_k, y_k) + \sum_{k=1}^{n} \Delta L_{\delta_k}(p(\hat{\alpha}, x_k), x_k, y_k), \tag{9}$$

*where $\delta_k = |\hat{\alpha}_k| K(x_k, x_k)$.*

*Proof.* Using Lemma 3.1 and Proposition 3.3, we obtain $|p(\hat{\alpha}^{[k]}, x_k) - p(\hat{\alpha}, x_k)| \leq \delta_k$. This implies that

$$L(p(\hat{\alpha}^{[k]}, x_k), x_k, y_k) \leq L(p(\hat{\alpha}, x_k), x_k, y_k) + \Delta L_{\delta_k}(p(\hat{\alpha}, x_k), x_k, y_k).$$

Summing over $k$, we obtain the theorem. □

Although the expected leave-one-out error is the expected generalization error, the former may not be a stable estimator of the latter. To see this, we consider the following example: assume that the input is a 0-1 valued binary random variable with probability of 0.7 for

0 and 0.3 for 1, and the output is always 1. Given $n$ such input-output pairs, the leaner predicts 1 if an even number of inputs are ones, and predicts 0 otherwise. Clearly depending on whether the number of observed ones are even or odd, the leave-one-out classification error rate for the learner will be approximately 0.7 or 0.3. The expected classification error of this learner will be approximately 0.5. This implies that the leave-one-out classification error is not a stable estimator even when $n \to +\infty$.

It is thus useful to bound the variance of leave-one-out estimate. In order to do so for kernel learning machines, we use the following modified version of Efron-Stein style concentration inequality Efron and Stein (1981); Steele (1986). A self-contained proof is given in Appendix D.

**Lemma 3.2** *Let $Z$ be an arbitrary function of the training data: $Z = u(z_1, \ldots, z_n)$, and let $Z^{(i)}$ be a fixed function of the training data with $z_i$ excluded ($i = 1, \ldots, n$): $Z^{(i)} = u^{(i)}(z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_n)$, where $z_i = (x_i, y_i)$. Then the variance of $Z$ can be bounded as:*

$$\mathrm{Var}(Z) \leq E \sum_{i=1}^{n} (Z - Z^{(i)})^2,$$

*where the expectation (and variance) is with respect to the training data.*

Note that the choice of $u^{(i)}$ that minimizes the right hand side of Lemma 3.2 is the expectation of $u$ with respect to the $i$-th component: $E_{z_i} u(z_1, \ldots, z_n)$.

Clearly, if we let $Z_L(\mathcal{A}_K, D_n)$ be the leave-one-out error, and $Z_L^{(i)}(\mathcal{A}_K, D_n)$ be the leave-one-out error with the $i$-th datum removed from the training data, then the variance of $Z_L(\mathcal{A}_K, D_n)$ can be bounded by the right hand side of Lemma 3.2. For kernel machines, the latter can be bounded using the following result:

**Theorem 3.3** *Let $Z_L(\mathcal{A}_K, D_n)$ be the leave-one-out error on training set $D_n$ as in Theorem 3.2, and let $Z_L^{(i)}(\mathcal{A}_K, D_n) = Z_L(\mathcal{A}_K, D_n^{(i)})$ be the corresponding leave-one-out error with the $i$-th datum removed from the training data. We have the following inequality*

$$\begin{aligned}
&\sum_{i=1}^{n} \left[ n\, Z_L(\mathcal{A}_K, D_n) - (n-1)\, Z_L^{(i)}(\mathcal{A}_K, D_n) \right]^2 \\
\leq &(2n-1) \left[ \frac{1}{n} \sum_{i=1}^{n} L(p(\hat{\alpha}^{[i]}, x_i), x_i, y_i)^2 + \sum_{i,j:i \neq j} \Delta_{||} L_{\delta_{i,j}}(p(\hat{\alpha}^{[i]}, x_i), x_i, y_i)^2 \right],
\end{aligned}$$

*where $\delta_{i,j} = |\hat{\alpha}_j^{[i]}| K(x_i, x_i)^{1/2} K(x_j, x_j)^{1/2}$ for all $i \neq j$.*

*Proof.* Let $Z = n Z_L(\mathcal{A}_K, D_n)$ and $Z^{(i)} = (n-1)\, Z_L^{(i)}(\mathcal{A}_K, D_n)$. For all $i \neq j$, denote by $\hat{\alpha}^{[i,j]}$ the solution of (6) with both the $i$-th and the $j$-th data points removed from the training set. Given any $j$, using Lemma 3.1 and Proposition 3.3, we have for all $i \neq j$: $|p(\hat{\alpha}^{[i]}, x_i) - p(\hat{\alpha}^{[i,j]}, x_i)| \leq \delta_{i,j}$. Therefore

$$|L(p(\hat{\alpha}^{[i]}, x_i), x_i, y_i) - L(p(\hat{\alpha}^{[i,j]}, x_i), x_i, y_i)| \leq \Delta_{||} L_{\delta_{i,j}}(p(\hat{\alpha}^{[i]}, x_i), x_i, y_i).$$

Summing over $i(i \neq j)$, we obtain:

$$|(Z - L(p(\hat{\alpha}^{[j]}, x_j), x_j, y_j)) - Z^{(j)}| \leq \sum_{i:i\neq j} \Delta_{||} L_{\delta_{i,j}}(p(\hat{\alpha}^{[i]}, x_i), x_i, y_i).$$

We thus obtain:

$$\begin{aligned}
&(Z - Z^{(j)})^2 \\
\leq &\left[ n \cdot \frac{1}{n} |L(p(\hat{\alpha}^{[j]}, x_j), x_j, y_j)| + \sum_{i:i\neq j} \Delta_{||} L_{\delta_{i,j}}(p(\hat{\alpha}^{[i]}, x_i), x_i, y_i) \right]^2 \\
\leq &(2n-1) \left[ n \cdot (\frac{1}{n} L(p(\hat{\alpha}^{[j]}, x_j), x_j, y_j))^2 + \sum_{i:i\neq j} \Delta_{||} L_{\delta_{i,j}}(p(\hat{\alpha}^{[i]}, x_i), x_i, y_i)^2 \right].
\end{aligned}$$

Now the theorem can be obtained by summing over $j$. □

# 4 Leave-one-out analysis for bounded sub-gradient formulations

It is clear that Theorem 3.2 can be used to bound leave-one-out errors in terms of training errors. If each $\hat{\alpha}_k$ is small ($k = 1, \ldots, n$), and the loss is continuous, then the corresponding leave-one-out error is not much larger than the training error. Since the expected leave-one-out error is the expected generalization error, we know that the expected generalization error is not much larger than the expected error on the training set. Similarly one can obtain a bound on the variance of leave-one-out error using Theorem 3.3.

In order to apply (9), we need to estimate $\hat{\alpha}$. Although this quantity is available if we solve the kernel formulation (6) based on the training data, it is also very useful to estimate the quantity for all possible training data based on properties of the underlying learning formulation. Such an estimate can be used to derive a bound for the expected generalization error. This section considers a relatively simple situation. From (25), we know that if $\sup |\nabla_1 f(p, x, y)|$ is bounded, then each component $\hat{\alpha}$ is of the order $O(1/n)$. Similarly, if $L(p, x, y)$ is Lipschitz, then $L(p(\hat{\alpha}^{[k]}, x), x, y) - L(p(\hat{\alpha}, x), x, y)$ can be bounded accordingly.

For some problems, it is useful to limit the range of $p(\hat{\alpha}, \cdot)$ and $p(\hat{\alpha}^{[k]}, \cdot)$. Let $p$ be a measurable function, and let $\kappa_p = (2 \sup_{x,y} f(p(x), x, y) + \|p(\cdot)\|^2_{[n]})^{1/2}$. Note that we allow $\kappa_p$ to be $+\infty$. Now if $f$ is non-negative, then by comparing the primal empirical risks, we obtain $\|p(\hat{\alpha}, \cdot)\| \leq \kappa_p$, $\|p(\hat{\alpha}^{[i]}, \cdot)\| \leq \kappa_p$ for all $i$, and $\|p(\hat{\alpha}^{[i,j]}, \cdot)\| \leq \kappa_p$ for all $i \neq j$. Thus $|p(\hat{\alpha}, x)|, |p(\hat{\alpha}^{[i]}, x)|, |p(\hat{\alpha}^{[i,j]}, x)| \leq \kappa_p K(x, x)^{1/2}$. Based on this observation, we obtain the following result:

**Theorem 4.1** *Under the assumptions of Theorem 3.3. Let*

$$\begin{aligned}
\kappa \geq &\inf_{p(\cdot)} \left[ 2 \sup_{x,y} f(p(x), x, y) + \|p(\cdot)\|^2_{[n]} \right]^{1/2}, \\
M \geq &\sup_x K(x, x)^{1/2},
\end{aligned}$$

*where we allow both quantities to be $+\infty$. We further assume that $f(p,x,y)$ is non-negative when $\kappa M < +\infty$. Let*

$$\begin{aligned} M_f &= \sup_{|p|\leq \kappa M, x, y} \nabla_1 f(p,x,y), \\ M_L &= \sup_{|p_1|,|p_2|\leq \kappa M, x, y} |L(p_1,x,y) - L(p_2,x,y)|/|p_1 - p_2|, \\ U_L &= \sup_{|p|\leq \kappa M, x, y} |L(p(x),x,y)|. \end{aligned}$$

*Given training data $D_n$, let $A_X = \frac{M_L M_f}{n}\sum_{k=1}^n K(x_k,x_k)$. We have the following inequalities:*

$$Z_L(\mathcal{A}_K, D_n) \leq \frac{1}{n}\sum_{k=1}^n L(p(\hat{\alpha}, x_k), x_k, y_k) + \frac{A_X}{n}. \tag{10}$$

$$\sum_{i=1}^n \left[ Z_L(\mathcal{A}_K, D_n) - \frac{n-1}{n} Z_L^{(i)}(\mathcal{A}_K, D_n) \right]^2 \leq \frac{2}{n}\left[ U_L^2 + A_X^2 - \frac{M_L^2 M_f^2}{n^2}\sum_{k=1}^n K(x_k,x_k)^2 \right]. \tag{11}$$

*Proof.* The definition of $\kappa$ implies that $\|p(\alpha,\cdot)\| \leq \kappa$ for all $p(\alpha,\cdot)$ that solves (8) with $m$-points removed from the training set ($m = 0, 1, 2$). Hence $|p(\alpha, x)| \leq \kappa M$. This implies that we can replace $L(p,x,y)$ by $L(\max(\min(p,\kappa M), -\kappa M), x, y)$ in the proof. This new $L$ has a global Lipschitz constant of $M_L$. Therefore $\Delta_{||} L_\delta(p,x,y) \leq M_L \delta$. Using (25) we also obtain $|\hat{\alpha}_k| \leq \frac{M_f}{n}$ for all training data. Let $\delta_k = |\hat{\alpha}_k| K(x_k,x_k)$, then

$$\Delta L_{\delta_k}(p(\hat{\alpha}, x_k), x_k, y_k) \leq \frac{M_f M_L}{n} K(x_k, x_k).$$

Sum over $k$ and apply Theorem 3.2, we obtain (10).

Similarly, let $\delta_{i,j} = |\hat{\alpha}_j^{[i]}| K(x_i,x_i)^{1/2} K(x_j,x_j)^{1/2}$. We have

$$\Delta_{||} L_{\delta_{i,j}}(p(\hat{\alpha}^{[i]}, x_i), x_j, y_j) \leq \frac{M_f M_L}{n} K(x_i,x_i)^{1/2} K(x_j,x_j)^{1/2}.$$

Summing over $i, j (i \neq j)$, we obtain

$$\sum_{i,j:i\neq j} \Delta_{||} L_{\delta_{i,j}}(p(\hat{\alpha}^{[i]}, x_i), x_j, y_j)^2 \leq A_X^2 - \frac{M_L^2 M_f^2}{n^2}\sum_{k=1}^n K(x_k,x_k)^2.$$

Using Theorem 3.3, we obtain (11). □

Note that if $\kappa M = +\infty$, then the condition $|p| \leq +\infty$ in the definitions of $M_f$, $M_L$ and $U_L$ can be interpreted as $|p| < +\infty$.

**Corollary 4.1** *Under the assumptions of Theorem 4.1. If $E\, K(x,x) < +\infty$, then*

$$\begin{aligned} E\, Z_L(\mathcal{A}_K, D_n) \leq & E\, \frac{1}{n}\sum_{k=1}^n L(p(\hat{\alpha}, x_k), x_k, y_k) + \frac{M_f M_L}{n} E\, K(x,x), \\ \mathrm{Var}(Z_L(\mathcal{A}_K, D_n)) \leq & \frac{2}{n}[U_L^2 + (M_f M_L)^2 (E\, K(x,x))^2], \end{aligned}$$

*where the expectation (variance) is with respect to the training data $D_n$.*

*Proof.* Taking expectation with respect to the training samples in (10) leads to the first inequality. Now, note that

$$\begin{aligned} E\, A_X^2 =& (M_L M_f)^2 \left[ (E\, K(x,x))^2 + \frac{1}{n} \mathrm{Var} K(x,x) \right] \\ \leq& (M_L M_f)^2 \left[ (E\, K(x,x))^2 + \frac{1}{n} E\, K(x,x)^2 \right]. \end{aligned}$$

Taking expectation over the training samples in (11) and recall Lemma 3.2, we obtain the second inequality. □

We have shown that if both $M_f$ and $M_L$ are bounded, then the expected generalization error is at most $O(1/n)$ more than that of the expected training error. Furthermore, if $L$ is bounded, then the variance of the leave-one-out estimate is $O(1/n)$. Let $L'(p,x,y) \geq L(p,x,y)$ be an upper bound of the loss which is convex in $p$. In the following, we choose $f$ such that

$$f(p,x,y) = c_n L'(p,x,y),$$

where $c_n$ is a regularization parameter. Note that this is equivalent to the penalized formulation (4) with $\lambda_n = 1/c_n$. Since $p(\hat{\alpha}, \cdot)$ solves (8), we can obtain from Corollary 4.1

$$E\, Z_{L'}(\mathcal{A}_K, D_n) \leq \inf_{p(\cdot)} \left[ E\, L'(p,x,y) + \frac{1}{2c_n} \|p(\cdot)\|_{[n]}^2 \right] + \frac{c_n M_{L'}^2}{n} E\, K(x,x). \tag{12}$$

Clearly this implies that if we choose $c_n = o(n)$ such that $c_n \to \infty$, then as $n \to \infty$ the expected generalization error with respect to the $L'$-loss will be no more than the best approachable $L'$-loss in $H$: $\inf_{p \in H} E\, L'(p,x,y)$.

Consider a non-negative loss $L'$. If we let $L = \min(L', U)$ for some $U \geq 0$, then it is clear that in addition to (12) we also have

$$\mathrm{Var}(Z_L(\mathcal{A}_K, D_n)) \leq \frac{2}{n} \left[ U^2 + c_n^2 M_{L'}^2 (E\, K(x,x))^2 \right]. \tag{13}$$

This shows that if we want the variance of leave-one-out error to approach zero as $n \to \infty$, $c_n$ has to be chosen such that $c_n = o(\sqrt{n})$. Clearly this condition is more restrictive than the requirement of $c_n = o(n)$ in (12). This also implies that if we choose a large $c_n$, then the variance of the leave-one-out error can be large even when the expected generalization of the estimated predictor using kernel learning is not much worse than that of the optimal predictor.

Analysis in this section can be useful for certain robust regression and classification formulations. For regression, we consider the following two scenarios:

- Absolute deviation: $L(p,x,y) = \min(L'(p,y), U)$ where

  $$L'(p,y) = |p - y|.$$

  Let $M_y = \sup_y |y|$, then we may set $\kappa = (2c_n M_y)^{1/2}$, and have

  $$M_f \leq c_n, \quad M_{L'} \leq 1, \quad U_L \leq \min((2c_n M_y)^{1/2} M + M_y, U).$$

- Huber's robust loss: $L(p,x,y) = \min(L'(p,y), U)$ where

$$L'(p,y) = \begin{cases} |p-y| - 1 & |p-y| > 2, \\ \frac{1}{4}(p-y)^2 & |p-y| \le 2. \end{cases}$$

  Let $M_y = \sup_y |y|$, then we may set $\kappa = (2c_n M_y)^{1/2}$, and have

$$M_f \le c_n, \quad M_{L'} \le 1, \quad U_L \le \min((2c_n M_y)^{1/2} M + M_y, U).$$

We can also consider binary classification problems with labels $y = \pm 1$. The decision rule is to predict $y$ as 1 if $p(x) \ge 0$ and $-1$ if $p(x) < 0$. We define the classification error function of this prediction as:

$$I(p,y) = \begin{cases} 0 & py > 0, \\ 1 & py \le 0. \end{cases}$$

Our goal is thus to produce a predictor $p(x)$ such that $p(x)$ and $y$ has the same sign. We consider the following two widely used formulations. Instead of using classification error as the loss, we consider losses that are upper bounds of classification error (up to a scale factor).

- Logistic regression: $L(p,x,y) = \min(L'(p,y), U)$ where

$$L'(p,y) = \ln(1 + \exp(-py)).$$

  We may set $\kappa = (2c_n \ln 2)^{1/2}$, and have

$$M_f \le c_n, \quad M_{L'} \le 1, \quad U_L \le \min((2c_n \ln 2)^{1/2} M + \ln 2, U).$$

- SVM loss: $L(p,x,y) = \min(L'(p,y), U)$ where

$$L'(p,y) = \max(0, 1 - py).$$

  We may set $\kappa = (2c_n)^{1/2}$, and have

$$M_f \le c_n, \quad M_{L'} \le 1, \quad U_L \le \min((2c_n)^{1/2} M + 1, U).$$

- Modified Huber's loss: $L(p,x,y) = \min(L'(p,y), U)$ where

$$L'(p,y) = \begin{cases} -py & py < -1, \\ \frac{1}{4}(1-py)^2 & py \in [-1,1], \\ 0 & py > 1. \end{cases}$$

  We may set $\kappa = (c_n/2)^{1/2}$, and have

$$M_f \le c_n, \quad M_L \le 1, \quad U_L \le \min((c_n/2)^{1/2} M + 1/2, U).$$

We'd like to mention that the modified Huber's loss has certain advantages over the standard SVM loss, and hence is interesting by itself Zhang (2002c). However the specific theoretical motivation is not important for the purpose of this paper. Clearly for all of the above cases $M_{L'} = 1$. The analysis in this section can thus be applied. From (12) and (13), we obtain the following result:

**Corollary 4.2** *Consider absolute deviation or Huber's robust loss for regression, and logistic regression, SVM, or Modified Huber's loss for classification. If we choose $f$ as $f(p, x, y) = c_n L'(p, y)$ and solve the corresponding learning problem (6) with $g(p, x, y)$ given by (7). Then the following leave-one-out bounds are valid:*

$$E\, Z_{L'}(\mathcal{A}_K, D_n) \leq \inf_{p(\cdot)} \left[ E\, L'(p(x), y) + \frac{1}{2c_n} \|p(\cdot)\|^2_{[n]} \right] + \frac{c_n}{n} E\, K(x, x),$$
$$\mathrm{Var}(Z_L(\mathcal{A}_K, D_n)) \leq \frac{2}{n} \left[ U_L^2 + c_n^2 (E\, K(x, x))^2 \right],$$

*where $p(\cdot)$ is an arbitrary measurable function.*

If we restrict $p(\cdot)$ to be in $H$, then $\|p(\cdot)\|^2_{[n]}$ can be replaced by $\|p\|^2$. Since we can choose $\kappa = \sqrt{2c_n L'(0, 0, 1)}$ for the above classification formulations, we can let $L(p, x, y) = L'(p, y)$ and set $U_L = L'(-\sqrt{2c_n L'(0, 0, 1) \sup_x K(x, x)}, 0, 1)$ in the above variance bound.

Note that $c_n$ in the above lemma is equivalent to $1/\lambda_n$ under the penalized formulation (4). Therefore using notations from Section 2, the expected generalization error can be bounded as:

$$Q_{L'}(\mathcal{A}_K, n-1) \leq \inf_{p \in H} \left[ Q_{L'}(p) + \frac{\lambda_n}{2} \|p\|^2 \right] + O\left( \frac{1}{\lambda_n n} \right). \tag{14}$$

The second term on the right hand side is the bias term mentioned in Section 2, and the third term is the learning complexity term. As a comparison, the leave-one-out variance term is of the order $O(\frac{U_L^2}{n} + \frac{1}{\lambda_n^2 n})$, which is usually much larger since typically we require that $\lambda_n \to 0$ as $n \to \infty$.

As mentioned in Section 2, this order $O(\frac{1}{\lambda_n^2 n})$ learning complexity behavior is also presented in the algorithmic stability based approach such as Bousquet and Elisseeff (2002). For example, typical bounds obtained there (such as those for SVM and least squares) in Bousquet and Elisseeff (2002) are of the following form (when adapted to our notations): with probability of $1 - \eta$ over training data $D_n$, we have

$$Q_L(\mathcal{A}_K, D_n) \leq \inf_{p \in H} \left[ \frac{1}{n} \sum_{i=1}^{n} L'(p(x_i), x_i, y_i) + \frac{\lambda_n}{2} \|p\|^2 \right] + O\left( \frac{1}{\lambda_n n} \right) + O\left( \sqrt{\frac{-\ln(\eta)}{\lambda_n^2 n}} \right),$$

where the $O(\cdot)$ notation also depends on $U_L$. Taking expectation with respect to the training data, we obtain the following expected generalization bound from their analysis:

$$Q_L(\mathcal{A}_K, n) \leq \inf_{p \in H} \left[ Q_{L'}(p) + \frac{\lambda_n}{2} \|p\|^2 \right] + O\left( \frac{1}{\lambda_n n} \right) + O\left( \sqrt{\frac{1}{\lambda_n^2 n}} \right). \tag{15}$$

Clearly similar to our leave-one-out variance bound, the $O(1/\lambda_n^2 n)$ learning complexity behavior is also present in their analysis.

Consider now for example that the value $\inf_{p\in H} Q_{L'}(p)$ can be achieved at $p^* \in H$: $Q_{L'}(p^*) = \inf_{p\in H} Q_{L'}(p)$. In this case, our analysis implies $Q_{L'}(\mathcal{A}_K, n) \leq Q_{L'}(p^*) + O(\lambda_n) + O(\frac{1}{\lambda_n n})$. Now by choosing $\lambda_n = O(1/\sqrt{n})$, we obtain $Q_{L'}(\mathcal{A}_K, n) \leq Q_{L'}(p^*) + O(1/\sqrt{n})$. Note that the leave-one-out variance bound is of the order $O(\frac{U_L^2}{n} + 1)$, which does not even converge to zero as $n \to \infty$. The analysis in Bousquet and Elisseeff (2002) has the same issue. The best convergence rate from (15) is obtained by setting $\lambda_n = n^{-1/4}$, and the resulting bound becomes $Q_{L'}(\mathcal{A}_K, n) \leq Q_{L'}(p^*) + O(n^{-1/4})$. If we consider the effect of $U_L$ dependency in the $O(\cdot)$ notation, the best possible rate of convergence from their analysis can be even slower.

We are not trying to criticize the algorithmic stability based analysis such as Bousquet and Elisseeff (2002). As we mentioned earlier, such analysis leads to probability bounds that provide more detailed information than expected generalization bounds presented in this paper. However, it is important to see that their analysis does not lead to the correct expected generalization error bounds, and we do not know an easy fix. In this regard, the leave-one-out analysis presented in this paper is useful since it does not suffer from this problem. However a disadvantage of the leave-one-out analysis is that it only leads to expected generalization bounds.

Another interesting classification formulation is exponential loss where we let $f(p, x, y) = c_n \exp(-py)$. This function is used in boosting but can also be combined with kernel methods.

**Corollary 4.3** *Consider loss function $L(p, x, y) = \exp(-py)$ and choose the corresponding $f$ in (8) as $f(p, x, y) = c_n \exp(-py)$. Let $M = \sup_x K(x, x)^{1/2}$, then the leave-one-out error satisfies:*

$$E\, Z_L(\mathcal{A}_K, D_n) \leq \inf_{p(\cdot)} \left[ E\, L(p(x), x, y) + \frac{1}{2c_n} \|p(\cdot)\|_{[n]}^2 \right] + \frac{c_n \exp(\sqrt{8c_n} M)}{n} E\, K(x, x),$$

$$\mathrm{Var}(Z_L(\mathcal{A}_K, D_n)) \leq \frac{2}{n} \left[ \exp(\sqrt{8c_n} M) + c_n^2 \exp(\sqrt{32c_n} M) (E\, K(x, x))^2 \right],$$

*where $p(\cdot)$ is an arbitrary measurable function.*

*Proof.* Consider $p = 0$, which leads to a choice of $\kappa = \sqrt{2c_n}$. We can now apply Corollary 4.1 with $M_f = c_n \exp(\sqrt{2c_n} M)$, and $M_L = U_L = \exp(\sqrt{2c_n} M)$. □

Although the analysis given in this section appears to be relatively general in the sense that it can be used to study any kernel learning problems with bounded $M_L$ and $M_f$. However for many problems (such as least squares regression), $M_L$ and $M_f$ may depend on the regularization parameter $c_n$. In such cases, the analysis given in this section can be suboptimal. It is thus necessary to estimate the behavior of $\hat{\alpha}$ using more elaborated methods. This in general requires us to obtain a metric that measures the size of $\hat{\alpha}$ based on some observable quantities. Although any quantity can be used, we will specifically consider bounding $\hat{\alpha}$ through the optimal training risk of the primal objective function (8). The advantage of using this quantity is that it is directly optimized by the kernel learning algorithm. Therefore we can bound its value using an arbitrary function $p \in H$ (the idea has already been used in this section). We will apply this method to specific learning problems in regression and

classification in the next two sections. Although for those problems variance estimates can also be obtained, they will be of rather complicated forms. Therefore for simplicity we do not include the corresponding variance calculations.

# 5 Leave-one-out analysis for some regression formulations

In this section, we study some regression formulations. We define

$$D_\epsilon(p, y) = \max(|p - y| - \epsilon, 0)$$

for all $\epsilon \geq 0$. In Theorem 3.2, let

$$L(p, x, y) = D_\epsilon(p, y)^s, \qquad (1 \leq s < +\infty)$$

which is a generalized version of Vapnik's $\epsilon$-insensitive regression loss Vapnik (1998). For regression problems, our goal is to find a function $p(x)$ to minimize the expected loss

$$Q(p(\cdot)) = E_{(x,y)} D_\epsilon(p(x), y)^s,$$

where $E$ denotes the expectation with respect to an unknown distribution. The training samples $(x_1, y_1), \ldots, (x_n, y_n)$ are independently drawn from the same underlying distribution.

Since we assume that $s \in [1, +\infty)$, $L$ is convex. We may choose $f$ in (8) such that

$$f(p, x, y) = \frac{c_n}{s} D_\epsilon(p, y)^s, \tag{16}$$

where $c_n > 0$ is a regularization parameter.

The purpose of the $\epsilon$ parameter in the above formulation is to obtain a sparse representation of $\hat{\alpha}$. In fact, this can be easily seen from (25) since the formula shows that a component $\hat{\alpha}_k = 0$ as long as $|p(\hat{\alpha}, x) - y| < \epsilon$. In this case, $g$ in (6) becomes

$$g(-\alpha, x, y) = \frac{(c_n/n)^{-t/s}}{t} |\alpha|^t - \alpha y + \epsilon |\alpha|, \tag{17}$$

where $1/s + 1/t = 1$.

For the above loss function, the following leave-one-out cross-validation bound can be directly obtained from Theorem 3.2:

**Lemma 5.1** *Under the conditions of Lemma 3.1, the following bound on leave-one-out error is valid for all $s \geq 1$:*

$$\left[\sum_{k=1}^n D_\epsilon(p(\hat{\alpha}^{[k]}, x_k), y_k)^s\right]^{1/s} \leq \left[\sum_{k=1}^n D_\epsilon(p(\hat{\alpha}, x_k), y_k)^s\right]^{1/s} + \left[\sum_{k=1}^n |\hat{\alpha}_k K(x_k, x_k)|^s\right]^{1/s}.$$

*Proof.* Note that $\forall \delta > 0$:

$$\sup_{|\Delta p| \leq \delta} D_\epsilon(p + \Delta p, y) \leq D_\epsilon(p, y) + \delta.$$

Let $L(p, x, y) = D_\epsilon(p, y)^s$. The right-hand-side of equation (9) can be bounded as

$$\sum_{k=1}^{n} (D_\epsilon(p(\hat{\alpha}, x_k), y_k) + |\hat{\alpha}_k| K(x_k, x_k))^s.$$

The lemma then follows from the Minkowski inequality. □

The generic leave-one-out regression bound in the above lemma can be computed from the training data. In order to derive results for the expected generalization error, we shall further investigate the behavior of this bound. Intuitively, analysis given in Section 4 suggests that $\hat{\alpha}_k$ is often of the order $O(1/n)$. Therefore when $n \to \infty$, the normalized second term on the right hand side (by multiplying the normalization factor $n^{-1/s}$) converges to zero at the rate $O(n^{-1/s})$. Roughly speaking this leads to the correct convergence rate of $n^{-1/2}$ (as mentioned in Section 2) for the least squares formulation where $s = 2$.

Similar to the analysis in Section 4, in order to obtain exact generalization bounds, we shall start with an estimate of $\hat{\alpha}$. With $g$ in (6) given by (17), we obtain from equation (25):

$$|\hat{\alpha}_i| = \frac{c_n}{n} \max(|p(\hat{\alpha}, x_i) - y_i| - \epsilon, 0)^{s-1} \qquad (i = 1, \ldots, n).$$

Therefore $\forall k$:

$$|\hat{\alpha}_k|^t = \frac{c_n^t}{n^t} D_\epsilon(p(\hat{\alpha}, x_k), y_k)^s.$$

Summing over $k$, we obtain:

$$\sum_{k=1}^{n} |\hat{\alpha}_k|^t \leq \frac{c_n^t}{n^t} \sum_{k=1}^{n} D_\epsilon(p(\hat{\alpha}, x_k), y_k)^s.$$

Clearly the left-hand side of the above inequality measures the size of $\hat{\alpha}$ using its $t$-norm, and the right hand side gives a bound in terms of the training error. Therefore we have been able to bound the size of $\hat{\alpha}$ using training error, as suggested at the end of Section 4.

To proceed with the analysis, we consider another quantity $s'$ such that $1/s' = 1/t + 1/t'$, where $s', t' \geq 1$. Using Jensen's inequality, some simple algebra, and Hölder's inequality, we

can obtain the following:

$$\begin{aligned}
&\frac{1}{n^{\max(1/s-1/s',0)}}\left[\sum_{k=1}^{n}|\hat{\alpha}_k K(x_k,x_k)|^s\right]^{1/s}\\
\leq&\left[\sum_{k=1}^{n}|\hat{\alpha}_k K(x_k,x_k)|^{s'}\right]^{1/s'}\\
\leq&\left[\sum_{k=1}^{n}|\hat{\alpha}_k|^t\right]^{1/t}\left[\sum_{k=1}^{n}K(x_k,x_k)^{t'}\right]^{1/t'}\\
\leq&\frac{c_n}{n}\left[\sum_{k=1}^{n}D_\epsilon(p(\hat{\alpha},x_k),y_k)^s\right]^{1/t}\left[\sum_{k=1}^{n}K(x_k,x_k)^{t'}\right]^{1/t'}.
\end{aligned}$$

Substituting into Lemma 5.1, we obtain the following leave-one-out bound:

**Lemma 5.2** *Under the conditions of Lemma 3.1 with $g$ given by (17). Consider $s,t,s',t'\geq 1$: $1/s+1/t=1$ and $1/s'=1/t+1/t'$. Then*

$$\begin{aligned}
\left[\sum_{k=1}^{n}D_\epsilon(p(\hat{\alpha}^{[k]},x_k),y_k)^s\right]^{1/s}\leq&\left[\sum_{k=1}^{n}D_\epsilon(p(\hat{\alpha},x_k),y_k)^s\right]^{1/s}+\\
&c_n n^{\max(1/s-1/s',0)-1}\left[\sum_{k=1}^{n}D_\epsilon(p(\hat{\alpha},x_k),y_k)^s\right]^{1/t}\left[\sum_{k=1}^{n}K(x_k,x_k)^{t'}\right]^{1/t'}.
\end{aligned}$$

Note that in the above lemma, we have successfully removed the dependency of the leave-one-out bound on $\hat{\alpha}$. Next, we would like to investigate the behavior of the expected leave-one-out error, which gives the expected generalization error. In the following, for any random variable $\xi$, we use the more compact notation $E^{1/s}\xi$ to denote $(E\xi)^{1/s}$.

**Theorem 5.1** *Under the conditions of Lemma 5.2. Assume further that $1/s=1/u+1/v$ where $u,v\geq 1$. Denote by $X_n$ the training set $\{(x_1,y_1),\ldots,(x_n,y_n)\}$. Let*

$$\bar{Q}(X_n)=\frac{1}{n}\sum_{k=1}^{n}D_\epsilon(p(\hat{\alpha},x_k),y_k)^s$$

*be the observed average training error. Then the expected leave-one-out error can be bounded as*

$$E^{1/s}\,Q(p(\hat{\alpha}^{[k]},\cdot))\leq E^{1/s}\,\bar{Q}(X_n)+\frac{c_n}{n^{\min(1/s,1/s')+1/s}}E^{1/u}\,\bar{Q}(X_n)^{u/t}E^{1/v}\left[\sum_{k=1}^{n}K(x_k,x_k)^{t'}\right]^{v/t'},$$

*where $E$ denotes the expectation over $n$ random training samples $X_n$.*

*Proof.* Let $d_n = c_n n^{\max(1/s-1/s',0)-1}$. Using Lemma 5.2 and the Minkowski inequality, we have

$$\begin{aligned}
&\left[E\sum_{k=1}^{n} D_\epsilon(p(\hat{\alpha}^{[k]},x_k),y_k)^s\right]^{1/s} \\
&\leq \left[E\sum_{k=1}^{n} D_\epsilon(p(\hat{\alpha},x_k),y_k)^s\right]^{1/s} + d_n E^{\frac{1}{s}}\left[\left[\sum_{k=1}^{n} D_\epsilon(p(\hat{\alpha},x_k),y_k)^s\right]^{\frac{s}{t}}\left[\sum_{k=1}^{n} K(x_k,x_k)^{t'}\right]^{\frac{s}{t'}}\right] \\
&\leq \left[E\sum_{k=1}^{n} D_\epsilon(p(\hat{\alpha},x_k),y_k)^s\right]^{1/s} + d_n E^{\frac{1}{u}}\left[\sum_{k=1}^{n} D_\epsilon(p(\hat{\alpha},x_k),y_k)^s\right]^{\frac{u}{t}} E^{\frac{1}{v}}\left[\sum_{k=1}^{n} K(x_k,x_k)^{t'}\right]^{\frac{v}{t'}},
\end{aligned}$$

where the second inequality follows from the Hölder's inequality. Also note that

$$E\sum_{k=1}^{n} D_\epsilon(p(\hat{\alpha}^{[k]},x_k),y_k)^s = nE\, D_\epsilon(p(\hat{\alpha}^{[k]},x_k),y_k)^s = nE\, Q(p(\hat{\alpha}^{[k]},\cdot)).$$

We thus obtain the theorem. □

The next step is to bound the training error using the error of an arbitrary function $p \in H$. This gives the following result. Note that the bound can be improved using moment inequalities for sum of independent variables. However this introduces another level of complexity. For our purpose, we will use plain Jensen's inequality for simplicity.

**Corollary 5.1** *Under the conditions of Theorem 5.1. Let*

$$Q_{q,n} = \inf_{p(\cdot)} E^{1/q}\left[D_\epsilon(p(x),y) + \frac{s}{2c_n}\|p(\cdot)\|^2_{[n]}\right]^q,$$

*where $p$ denotes an arbitrary measurable function. Let*

$$\ell_n = 1 + \min(1/s, 1/s') - \max(1/t, 1/u) - \max(1/t', 1/v).$$

*The expected generalization error is bounded by*

$$E^{1/s}\, Q(p(\hat{\alpha}^{[k]},\cdot)) \leq Q_{1,n}^{1/s} + \frac{c_n}{n^{\ell_n}} Q_{u/t,n}^{1/t}\, E^{1/v} K(x,x)^v,$$

*where $E$ denotes the expectation over $n$ random training samples.*

*Proof.* We would like to bound the three terms on the right hand side of Theorem 5.1 separately. For the first term,

$$E\,\bar{Q}(X_n) \leq \inf_p\; E\left[\frac{1}{n}\sum_{k=1}^{n} D_\epsilon(p(x_k),y_k)^s + \frac{s}{2c_n}\|p(X_n)\|^2\right] = Q_{1,n}.$$

Now for any measurable function $p(\cdot)$, using Jensen's inequality to bound the second term, we obtain

$$\begin{aligned}
E^{1/u}\,\bar{Q}(X_n)^{u/t} \leq & E^{1/u}\left[\frac{1}{n}\sum_{k=1}^{n} D_\epsilon(p(x_k),y_k)^s + \frac{s}{2c_n}\|p(X_n)\|^2\right]^{u/t} \\
\leq & E^{1/u}\, n^{\max(-1,-u/t)} \sum_{k=1}^{n}\left[D_\epsilon(p(x_k),y_k)^s + \frac{s}{2c_n}\|p(X_n)\|^2\right]^{u/t} \\
= & n^{\max(0,1/u-1/t)} E^{1/u}\left[D_\epsilon(p(x),y)^s + \frac{s}{2c_n}\|p\|^2_{[n]}\right]^{u/t}.
\end{aligned}$$

Similarly, using Jensen's inequality, we have

$$\begin{aligned}
E^{1/v}\left[\sum_{k=1}^{n} K(x_k,x_k)^{t'}\right]^{v/t'} \leq & E^{1/v}\, n^{\max(v/t'-1,0)} \sum_{k=1}^{n} K(x_k,x_k)^v \\
\leq & n^{\max(1/t',1/v)} E^{1/v}\, K(x,x)^v.
\end{aligned}$$

Applying the above bounds to Theorem 5.1, we obtain the desired result. □

Corollary 5.1 can be applied with any choice of $(s',t')$ and $(u,v)$. We may make the following specific choice that maximizes the rate $\ell_n$:

- $s \in [1,2]$: $s'=s$, $t'=st/(t-s)$, $u=t$ and $v=t'$. We have

$$E^{1/s}\, Q(p(\hat{\alpha}^{[k]},\cdot)) \leq Q_{1,n}^{1/s} + \frac{c_n}{n} Q_{1,n}^{1/t}\, E^{(t-s)/st} K(x,x)^{st/(t-s)}.$$

- $s \geq 2$: $s'=t$, $t'=+\infty$, $u=s$ and $v=+\infty$. We have

$$E^{1/s}\, Q(p(\hat{\alpha}^{[k]},\cdot)) \leq Q_{1,n}^{1/s} + \frac{c_n}{n^{2/s}} Q_{s/t,n}^{1/t} \sup_x K(x,x).$$

Clearly if we let $s=1$ and $t=+\infty$, then we are able to reproduce the expected generalization error bound for absolute deviation in Corollary 4.2. With fixed $c_n$, it is easy to see that when $s>2$, the asymptotic convergence rate is $O(n^{-2/s})$, which decreases as $s$ increases. When $s<2$, we are more tolerant to large input $K(x,x)$ since the bound depends on $E^{(t-s)/st}K(x,x)^{st/(t-s)}$, which becomes less sensitive to large $K(x,x)$ as $s$ decreases. In addition, for smaller $s$, $t$ is larger, and hence the factor $Q_{1,n}^{1/t}$ (for $s\in[1,2]$, or $Q_{s/t,n}^{1/t}$ for $s\geq 2$) has a smaller impact on the bound. This means even when $Q_{1,n}$ is relatively large, we can still expect a fast convergence. All these observations imply that the formulation is more robust when $s$ is smaller. Note that regression using the Huber's loss has the same behavior as that of $s=1$. This gives a theoretical justification on the robustness of Huber's loss function.

For $s\geq 2$, we require $\sup_x K(x,x)$ to be bounded. It is clear that by choosing $v<+\infty$, we can relax this condition to the boundedness of $EK(x,x)^v$. However, asymptotic convergence rate reduces to $O(n^{-2/s+1/v})$. It is also interesting to observe that the convergence decreases

to $O(1)$ as $s \to \infty$. This is not very surprising. We may consider the scenario that $s = \infty$. Clearly the risk function $Q$ becomes the essential upper bound of $|p(x) - y|$: $Q(p) = \inf\{a : P(|p(x) - y| > a) = 0\}$, which cannot be reliably estimated with a finite number of samples. This is another way to see that large $s$ is very sensitive to outliers.

We shall now consider the least squares regression formulation where $s = 2$. Assume that $K(x, x)$ is bounded. Let $\lambda_n = 2/c_n$ in (4), Corollary 5.1 (with the above mentioned parameter choices of $s' = s, t' = st/(t - s), u = t$ and $v = t'$) implies that the expected generalization error with $n - 1$ training data can be bounded as:

$$E\, Q(p(\hat{\alpha}^{[k]}, \cdot)) \leq \left(1 + \frac{2 \sup_x K(x, x)}{\lambda_n n}\right)^2 \inf_{p \in H} \left[Q(p) + \frac{\lambda_n}{2}\|p\|^2\right].$$

Similarly to (14), the above bound implies a $O(\frac{1}{\lambda_n n})$ learning complexity. As mentioned in Section 2, if $\inf_{p \in H} Q(p)$ can be achieved by $p^* \in H$, then we can let $\lambda_n = O(1/\sqrt{n})$ to obtain the correct convergence rate (to the minimum value $Q(p^*)$) of the order $O(1/\sqrt{n})$. Similar to the discussion in Section 4, our result compares favorable to the least squares bound of the form (15) in Bousquet and Elisseeff (2002), which leads to the best achievable rate of the order $O(n^{-1/4})$. One also note that our results can be further improved when the problem is noise-free: $Q(p^*) = 0$. Here we may let $\lambda_n = O(1/n)$ to obtain an expected generalization error of the order $O(1/n)$. The best achievable rate in Bousquet and Elisseeff (2002) based on (15) is still no better than $O(n^{-1/4})$ under this scenario.

Next we would like to consider a choice of $c_n$ so that the expected generalization error converges to the best approachable error when $n \to \infty$. For simplicity, assume $1 < s \leq 2$, $s' = s$, and $E^{(t-s)/st} K(x, x)^{st/(t-s)}$ is finite. We also restrict $p$ such that $p \in H$ in the estimate of $Q_{1,n}$. This gives the following results:

- $\inf_{p \in H} E_{x,y}\, D_\epsilon(p(x), y)^s = 0$: if we pick $c_n \to \infty$ such that $c_n = O(n)$, then
$$\lim_{n \to \infty} E\; E_{x,y}\, D_\epsilon(p(\hat{\alpha}, x), y)^s = 0.$$
- $\inf_{p \in H} E_{x,y}\, D_\epsilon(p(x), y)^s > 0$: we pick $c_n \to \infty$ such that $c_n = o(n)$, then
$$\lim_{n \to \infty} E\; E_{x,y}\, D_\epsilon(p(\hat{\alpha}, x), y)^s = \inf_{p \in H} E_{x,y}\, D_\epsilon(p(\hat{\alpha}, x), y)^s.$$

In the above, the first is the noiseless case, and the second is the noisy case. However, in both cases, we do not need to assume that there is a target function $p \in H$ that achieves the minimum of $\inf_{p \in H} E_{x,y} |p(x) - y|^s$. For example, we may have a function $\tilde{p}(x)$ such that $E_{x,y} |\tilde{p}(x) - y|^s = \inf_{p \in H} E_{x,y} |p(x) - y|^s$, but $\tilde{p}(x)$ does not belong to $H$. The function may lie in the closure of $H$ under the $L_s$-topology. For some kernel functions, this closure of $H$ may contain all continuous functions (and $L_s$). In this case, Corollary 5.1 implies that when $n \to \infty$, the kernel learning formulation (6) is able to learn all continuous target functions even under observation noise. We call this property universal learning.

The leave-one-out analysis can also be used to derive approximation bounds for kernel methods. For function approximation, we are interested in the best possible $L_s$ approximation error of an arbitrary function using $n$ terms of kernel expression:

$$A_n(p_*) = \inf_{\alpha} E_x^{1/s} |p(\alpha, x) - p_*(x)|^s,$$

where $p_*$ is the target function, and $p(\alpha, x)$ is given by (5). We may produce an approximation by solving (8) with $f(p, x, y) = c_n|p(\alpha, x) - p_*(x)|^s$. It is clear that $A_n(p_*) \leq E^{1/s}E_x\,|p(\hat{\alpha}^{[k]}, x) - p_*(x)|^s$. Applying Corollary 5.1, we obtain

$$\begin{aligned} A_{n-1}(p_*) \leq & \inf_{c_n>0}\left[(\frac{s}{2c_n}\|p_*\|^2_{[n]})^{1/s} + \frac{c_n}{n^{\ell_n}}(\frac{s}{2c_n}\|p_*\|^2_{[n]})^{1/t}\; E^{1/v}K(x,x)^v\right] \\ = & \sqrt{\frac{s\|p_*\|^2_{[n]}}{2n^{\ell_n}}E^{1/v}K(x,x)^v}. \end{aligned}$$

Again by maximizing the rate $\ell_n$, we obtain the following approximation bounds:

- $s \in [1, 2]$: $s' = s$, $t' = st/(t-s)$, $u = t$ and $v = t'$. We have

$$A_{n-1}(p_*) \leq \sqrt{\frac{s\|p_*\|^2_{[n]}}{2n}E^{(t-s)/st}K(x,x)^{st/(t-s)}}.$$

- $s \geq 2$: $s' = t$, $t' = +\infty$, $u = s$ and $v = +\infty$. We have

$$A_{n-1}(p_*) \leq \sqrt{\frac{s\|p_*\|^2_{[n]}}{2n^{2/s}}\sup_x K(x,x)}.$$

This shows that the minimum-norm interpolation quantity $\|p_*\|_{[n]}$ characterizes the rate of approximation using kernel expression (5).

# 6 Leave-one-out analysis for some binary classification formulations

In this section, we study some binary classification formulations using ideas similar to that given in Section 5. The following result is a direct consequence of Theorem 3.2. A similar bound can be found in Jaakkola and Haussler (1999).

**Lemma 6.1** *Under the conditions of Lemma 3.1, the following bound on leave-one-out classification error is valid:*

$$\sum_{k=1}^n I(p(\hat{\alpha}^{[k]}, x_k), y_k) \leq \sum_{k=1}^n I(p(\hat{\alpha}, x_k) - |\hat{\alpha}_k|y_k K(x_k, x_k), y_k).$$

The above bound can be regarded as a margin style error bound. Although useful computationally, it does not provide useful theoretical insights. In this section, we show how to estimate the right hand side of the above bound using techniques similar to our analysis of regression problems. One difficulty is the non-convexity of the classification error function. In practice, people use convex upper bounds of $I(p, y)$ to remedy the problem. There are

many possible choices. In this section, we will consider powers of $D_+(p,y) = \max(0, 1-py)$, which are related to support vector machines. Our goal is to find a function $p(x)$ to minimize the expected loss $Q(p(\cdot)) = E_{x,y} f(p(x), x, y)$ where $E$ denotes the expectation with respect to an unknown distribution. The training samples $(x_1, y_1), \ldots, (x_n, y_n)$ are independently drawn from the same underlying distribution.

The following lemma plays the same role of Lemma 5.1.

**Lemma 6.2** *Under the conditions of Lemma 3.1, the following bound on the leave-one-out error is valid for all $s \geq 1$:*

$$\left[\sum_{k=1}^{n} D_+(p(\hat{\alpha}^{[k]}, x_k), y_k)^s\right]^{1/s} \leq \left[\sum_{k=1}^{n} D_+(p(\hat{\alpha}, x_k), y_k)^s\right]^{1/s} + \left[\sum_{k=1}^{n} |\hat{\alpha}_k K(x_k, x_k)|^s\right]^{1/s}.$$

*Proof.* Note that $D_+(p+\Delta p, y) \leq D_+(p,y) + |\Delta p|$. The rest of the proof is the same as that of Lemma 5.1. □

To further investigate theoretical properties of support vector machine type losses, we need to bound the right hand side of Lemma 6.2 in a way similar to the regression analysis. In particular, we consider the following formulation for solving the classification problem, with $f$ in (8) chosen as

$$f(p,x,y) = \frac{c_n}{s} D_+(p(x), y)^s, \tag{18}$$

where $c_n > 0$ is a regularization parameter, and $1 \leq s < \infty$. In this case, $g$ in (6) becomes

$$g(-\alpha, x, y) = \frac{(c_n/n)^{-t/s}}{t}|\alpha|^t - \alpha y, \quad (\alpha y \geq 0) \tag{19}$$

where $1/s + 1/t = 1$. Note that if $s = 1$, then $t = +\infty$. Equation (19) can be equivalently written as:

$$g(\alpha, x, y) = \begin{cases} -\alpha y & \text{if } \alpha y \in [0, c_n/n], \\ +\infty & \text{otherwise.} \end{cases} \tag{20}$$

With $g$ in (6) given by (19), we obtain from equation (25):

$$\hat{\alpha}_k y_k = \frac{c_n}{n} \max(1 - p(\hat{\alpha}, x_k) y_k, 0)^{s-1} \qquad (k = 1, \ldots, n). \tag{21}$$

Note that if $s = 1$, then the right-hand-side of equation (21) is not uniquely defined at $p(\hat{\alpha}, x_i) y_i = 1$. The equation becomes $\hat{\alpha}_i y_i \in [0, c_n/n]$ in this case. Similar to the regression analysis in Section 5, we obtain $\forall k$:

$$|\hat{\alpha}_k|^t = \frac{c_n^t}{n^t} D_+(p(\hat{\alpha}, x_k), y_k)^s.$$

Summing over $k$, we obtain

$$\sum_{k=1}^{n} |\hat{\alpha}_k|^t \leq \frac{c_n^t}{n^t} \sum_{k=1}^{n} D_+(p(\hat{\alpha}, x_k), y_k)^s.$$

Clearly we can use the same derivation as that in Section 5, and simply replace the symbol $D_\epsilon$ by $D_+$. This leads to the following counterpart of Lemma 5.2:

**Lemma 6.3** *Under the conditions of Lemma 3.1 with $g$ given by (19). Consider $s, t, s', t' \geq 1$: $1/s + 1/t = 1$ and $1/s' = 1/t + 1/t'$. Then*

$$\left[\sum_{k=1}^{n} D_+(p(\hat{\alpha}^{[k]}, x_k), y_k)^s\right]^{1/s} \leq \left[\sum_{k=1}^{n} D_+(p(\hat{\alpha}, x_k), y_k)^s\right]^{1/s} + c_n n^{\max(1/s-1/s',0)-1} \left[\sum_{k=1}^{n} D_+(p(\hat{\alpha}, x_k), y_k)^s\right]^{1/t} \left[\sum_{k=1}^{n} K(x_k, x_k)^{t'}\right]^{1/t'}.$$

Using the same proof of Theorem 5.1, we obtain the following bound on expected generalization error.

**Theorem 6.1** *Under the conditions of Lemma 6.3. Assume further that $1/s = 1/u + 1/v$ where $u, v \geq 1$. Denote by $X_n$ the training set $\{(x_1, y_1), \ldots, (x_n, y_n)\}$. Let*

$$Q(p(\cdot)) = E_{x,y} D_+(p(\hat{\alpha}^{[k]}, x), y)^s$$

*be the expected true error of a predictor $p(\cdot)$, and let*

$$\bar{Q}(X_n) = \frac{1}{n}\sum_{k=1}^{n} D_+(p(\hat{\alpha}, x_k), y_k)^s$$

*be the observed training error. Then the expected leave-one-out error can be bounded as*

$$E^{1/s}\, Q(p(\hat{\alpha}^{[k]}, \cdot)) \leq E^{1/s}\, \bar{Q}(X_n) + \frac{c_n}{n^{\min(1/s, 1/s')+1/s}} E^{1/u}\, \bar{Q}(X_n)^{u/t} E^{1/v} \left[\sum_{k=1}^{n} K(x_k, x_k)^{t'}\right]^{v/t'},$$

*where $E$ denotes the expectation over $n$ random training samples $X_n$.*

The proof of the following result is the same as that of Corollary 5.1.

**Corollary 6.1** *Under the conditions of Theorem 6.1. Let*

$$Q_{q,n} = \inf_{p(\cdot)} E^{1/q} \left[D_+(p(x), y) + \frac{s}{2c_n}\|p(\cdot)\|_{[n]}^2\right]^q,$$

*where $p$ denotes an arbitrary measurable function. Let*

$$\ell_n = 1 + \min(1/s, 1/s') - \max(1/t, 1/u) - \max(1/t', 1/v).$$

*The expected generalization error is bounded by*

$$E^{1/s}\, Q(p(\hat{\alpha}^{[k]}, \cdot)) \leq Q_{1,n}^{1/s} + \frac{c_n}{n^{\ell_n}} Q_{u/t,n}^{1/t}\, E^{1/v} K(x, x)^v,$$

*where $E$ denotes the expectation over $n$ random training samples.*

Similar to the regression analysis, we obtain the following bounds from Corollary 6.1:

- $s \in [1,2]$: $s' = s$, $t' = st/(t-s)$, $u = t$ and $v = t'$. We have

$$E^{1/s}\, Q(p(\hat{\alpha}^{[k]}, \cdot)) \leq Q_{1,n}^{1/s} + \frac{c_n}{n} Q_{1,n}^{1/t}\, E^{(t-s)/st} K(x,x)^{st/(t-s)}.$$

- $s \geq 2$: $s' = t$, $t' = +\infty$, $u = s$ and $v = +\infty$. We have

$$E^{1/s}\, Q(p(\hat{\alpha}^{[k]}, \cdot)) \leq Q_{1,n}^{1/s} + \frac{c_n}{n^{2/s}} Q_{s/t,n}^{1/t} \sup_x K(x,x).$$

Formulations with $s \geq 2$ can be very sensitive to outliers. However they can still be useful when the data is nearly separable. In the general case, if we want to choose $c_n$ so that the expected generalization error converges to the optimal error, then $c_n^{s/2} = o(n)$. This can be naturally compared to the exponential loss formulation in Corollary 4.3 where we require $c_n \exp(\sqrt{8c_n}M) = o(n)$. This shows that we require $c_n$ to grow slowly when we use a loss function that heavily penalizes outliers.

The case of $s = 1$ corresponds to the support vector machine formulation. In this case if there exists some $p \in H$ such that $E\, D_+(p(x), y)^s \approx 0$ and $K(x,x)$ is bounded, then we can use another method to estimate the size of $\alpha$ which leads to a better bound. We multiply (21) by $1 - p(\hat{\alpha}, x_k)y_k$ and sum over $k$ to obtain

$$\begin{aligned}
\sum_{k=1}^n |\hat{\alpha}_k| &= \sum_{k=1}^n \hat{\alpha}_k y_k (1 - p(\hat{\alpha}, x_k) y_k) + \|p(\hat{\alpha}, \cdot)\|^2 \\
&= \sum_{k=1}^n \frac{c_n}{n} \max(1 - p(\hat{\alpha}, x_k) y_k, 0)^{s-1} (1 - p(\hat{\alpha}, x_k) y_k) + \|p(\hat{\alpha}, \cdot)\|^2 \\
&\leq \sum_{k=1}^n \frac{c_n}{n} D_+(p(\hat{\alpha}, x_k), y_k)^s + \|p(\hat{\alpha}, \cdot)\|^2.
\end{aligned}$$

This implies

$$\sum_{k=1}^n |\hat{\alpha}_k K(x_k, x_k)| \leq \left[ \sum_{k=1}^n \frac{c_n}{n} D_+(p(\hat{\alpha}, x_k), y_k)^s + \|p(\hat{\alpha}, \cdot)\|^2 \right] \sup_k K(x_k, x_k). \tag{22}$$

This estimate of $\hat{\alpha}$ can be compared to the corresponding estimate used in Corollary 6.1 (and Corollary 4.2) with $s = 1$: there we used an estimate of $\sup_k |\hat{\alpha}_k| \leq c_n/n$, which leads to a bound $\sum_{k=1}^n |\hat{\alpha}_k K(x_k, x_k)| \leq c_n/n \sum_{k=1}^n K(x_k, x_k)$. If $c_n$ is large, then the estimate given in (22) is better when $\sum_{k=1}^n \frac{1}{n} D_+(p(\hat{\alpha}, x_k), y_k)^s$ is close to zero and $\sup_x K(x,x)$ is well bounded. Using (22), we obtain the following result from Lemma 6.2:

**Theorem 6.2** *Under the conditions of Lemma 3.1 with g given by (20). Let* $M = \sup_k K(x_k, x_k)^{1/2}$, *then*

$$\sum_{k=1}^n D_+(p(\hat{\alpha}^{[k]}, x_k), y_k) \leq (1 + \frac{c_n M^2}{n}) \sum_{k=1}^n D_+(p(\hat{\alpha}, x_k), y_k) + \|p(\hat{\alpha}, \cdot)\|^2 M^2.$$

Taking expectation over the training data, we obtain

**Corollary 6.2** *Under the conditions of Lemma 3.1 with g given by (20). Let* $M = \sup_x K(x,x)^{1/2}$, *then*

$$E\, D_+(p(\hat{\alpha}^{[k]}, x_k), y_k) \leq \frac{\max(n, c_n M^2) + c_n M^2}{n} \inf_{p(\cdot)} \left[ E_{x,y} D_+(p(x), y) + \frac{1}{2c_n} \|p\|^2_{[n]} \right],$$

*where* $E$ *denotes the expectation over* $n$ *random training samples.* $p$ *denotes an arbitrary measurable function.*

If the problem is separable, Theorem 6.2 implies a leave-one-out bound of

$$\sum_{k=1}^{n} D_+(p(\hat{\alpha}^{[k]}, x_k), y_k) \leq \frac{n + 2c_n \sup_k K(x_k, x_k)}{2c_n} \|p(X_n)\|^2$$

for all function $p$ such that $p(x_k)y_k \geq 1$ $(k = 1, \ldots, n)$. We can let $c_n \to +\infty$, then the formulation becomes the optimal margin hyperplane (separable SVM) method. In this case the above analysis implies the following leave-one-out classification error bound:

$$\sum_{k=1}^{n} I(p(\hat{\alpha}^{[k]}, x_k), y_k) \leq \|p(\hat{\alpha}, \cdot)\|^2 \sup_k K(x_k, x_k).$$

This bound is identical to Vapnik's bound for optimal margin hyperplane method in Vapnik (1998). It shows that the optimal margin method can find a decision function that has classification error approaches zero as the sample size $n \to +\infty$ when the problem can be separated by a function in $H$ with a large margin. In addition to this result of Vapnik, a bound similar to Theorem 6.2 has also appeared in Joachims (2000).

We can further consider the case that the problem is separable but not by any function in $H$ (or not by a large margin). As we have pointed out in Section 5, there may be a function $p \notin H$ (but say, $p$ belongs to $H$'s closure under the $L_1$ topology) so that $p(x)y > 0$ almost everywhere. We can assume in this case that

$$\inf_{p \in H} E_{x,y} \max(1 - p(x)y, 0) = 0.$$

It is clear from Corollary 6.1 that the optimal margin hyperplane method may not estimate a classifier $p(\hat{\alpha}, \cdot)$ that has expected classification error approaches zero when $n \to +\infty$. That is, the method is not universal even for separable problems. This is not surprising since when $n \to \infty$, $\|p(\hat{\alpha}, \cdot)\|$ may also go to $+\infty$ at the same rate of $n$. On the other hand, by Corollary 6.1, with regularization parameter $c_n = O(n)$, $\lim_{n \to +\infty} E_{x,y} I(p(\hat{\alpha}, x), y) = 0$. This means that the soft-margin SVM with fixed $C = c_n/n$ is a universal learning method for separable problems. A similar conclusion can be drawn from Corollary 6.1 for $1 < s \leq 2$.

Finally we shall point out that although our end goal is the classification loss $E_{x,y} I(p(x), y)$, it is desirable to minimize a convex risk in the kernel formulation, which not only is computationally more efficient, but also reduces the variance associated with the estimation. Results in this section show that $E_{x,y} D_+(p(\hat{\alpha}, x), y)^s$ achieves the best possible approachable value

in $H$ as $n \to +\infty$. If $H$ is dense in the set of continuous functions in a compact set under the uniform norm topology, then it can be shown (for example, see Zhang (2002c)) that a function that approximately minimizes $E_{x,y}D_+(p(\hat{\alpha},x),y)^s$ will also approximately minimize $E_{x,y}I(p(x),y)$ among all Borel measurable functions. This implies that results obtained in this section can also be used to establish universal learning results for non-separable problems, though we only minimize a convex upper bound of the classification error function instead of the classification loss itself. This connection indicates that it is important to study the expected generalization error with respect to the $D_+(p(\hat{\alpha},x),y)^s$ loss even if our purpose is to minimize the classification error.

# 7 Summary

In this paper, we derived a general leave-one-out approximation bound for kernel methods. The approximation bound leads to a very general leave-one-out cross-validation bound for an arbitrary loss function. In addition, we have also studied variance bounds for leave-one-out estimates. Our bounds depend very weakly on properties of the underlying kernel. For example, they do not depend on the eigen-decomposition structure of the kernel. On one hand, this means that bounds we have obtained are very general. On the other hand, this also indicates that it might be possible to improve our analysis for specific kernels by taking theirs structures into consideration.

We have applied the derived bound to some regression and classification problems, which demonstrated the power of leave-one-out analysis. In fact, to the best of our knowledge, the expected generalization results obtained from our analysis are the best available bounds for general kernel learning machines. In addition, our bounds reflect both learning and approximation aspects of the underlying problems. Based on these results, we are able to demonstrate universal learning properties of certain kernel methods. Our analysis also suggests that even for noiseless problems in regression, or separable problems for classification, it is still helpful to use penalty type regularization. This is because in this case we can choose the regularization parameter as $c_n = O(n)$ to obtain universal learning methods.

In this paper, we show that the minimal interpolation norm $\|p(x)\|_{[n]}$ of a function $p(x)$ determines the rate of learning $p(x)$ with the corresponding $\|\cdot\|$ kernel function. However, we do not investigate the behavior of $\|p(x)\|_{[n]}$ for any specific function class and any specific kernel formulation. It will be interesting to study such issues, which may lead to useful insights into various kernel formulations.

# A Properties of kernel representation

We prove Proposition 3.2 and Proposition 3.3.

*Proof* of Proposition 3.2. First we consider the situation that $p(X_n)$ can be represented as

$$p(X_n) = G(X_n)\bar{\alpha}.$$

Define $s = (\bar{\alpha}^T G(X_n)\bar{\alpha})^{1/2}$, then $\forall \alpha$,

$$p(X_n)^T \alpha = \bar{\alpha}^T G(X_n)\alpha \leq s(\alpha^T G(X_n)\alpha)^{1/2}.$$

The equality can be achieved at $\alpha = \bar{\alpha}$. This shows that $s = \|p(X_n)\|$.

Now assume that $p(X_n)$ is not in the range of $G(X_n)$. Using elementary linear algebra, we can write

$$p(X_n) = G(X_n)\bar{\alpha} + \beta,$$

where $\beta \neq 0$ is in the null space of $G(X_n)$: $G(X_n)\beta = 0$. This implies that $\beta^T p(X_n) = \beta^T G(X_n)\bar{\alpha} + \beta^T\beta > 0$, and $(\beta^T G(X_n)\beta)^{1/2} = 0$. From Definition 3.2, we must have $\|p(X_n)\| = +\infty$. □

*Proof* of Proposition 3.3. To prove the first part, we consider the orthogonal projection $p_0$ of $p$ onto the subspace spanned by function $p_i(x) = K(x_i, x)$. Proposition 3.1 implies that $p_0(x_i) = p(x_i)$ for all $i = 1, \ldots, n$. Let $X_n = \{x_1, \ldots, x_n\}$, then Proposition 3.2 implies that $\|p(X_n)\| = \|p_0\| \leq \|p\|$.

To prove the second part, denote $X_1 = \{x_1\}$. If $K(x_1, x_1) = 0$, then let $p_0(\cdot)$ be the orthogonal projection of $p(\cdot)$ onto the subspace spanned by function $K(x_1, \cdot)$. Proposition 3.1 implies that $p(x_1) = p_0(x_1) = 0$. If $K(x_1, x_1) > 0$, then $\|p(X_1)\| = |p(x_1)|/K(x_1, x_1)^{1/2}$. Therefore using the first part, we have $|p(x_1)|/K(x_1, x_1)^{1/2} \leq \|p(\cdot)\|$. □

# B Proof of Theorem 3.1

To prove the theorem, we first note that a solution of (8) must be a solution of

$$\hat{p}(\cdot) = \arg \min_{p(\cdot) \in H_{X_n}} \left[ \frac{1}{n} \sum_{i=1}^{n} f(p(x_i), x_i, y_i) + \frac{1}{2}\|p(\cdot)\|^2 \right], \tag{23}$$

where $H_{X_n}$ is the subspace of $H$ spanned by $K(x_i, \cdot)$ $(i = 1, \ldots, n)$. This is because for any $p \in H$, by Proposition 3.1, let $p_{X_n}$ be the orthogonal projection of $p$ onto the subspace $X_n$, then $p(x_i) = p_{X_n}(x_i)$ for all $i$ and $\|p_{X_n}\| \leq \|p\|$, with the equality holds only when $p \in H_{X_n}$. Therefore $p_{X_n}$ has a smaller primal objective value than $p$ if $p \notin H_{X_n}$. This shows that any solution of (8) can be written as $\hat{p}(x) = \sum_{i=1}^{n} \hat{\alpha}_i K(x_i, x)$ for some $\hat{\alpha}$.

Now, let $\hat{\alpha}$ be a solution of (6). Since $\hat{\alpha}$ achieves the minimum of $L_n(\alpha)$, we have (see page 264 in Rockafellar (1970)) $0 \in \partial_{\alpha_i} L_n(\hat{\alpha})$ for all $i$ (the sub-differential is with respect to each $\alpha_i$). By Theorem 23.8 in Rockafellar (1970), for each $i$, we can find a subgradient of $g(\alpha_i, x_i, y_i)$ at $\hat{\alpha}_i$ with respect to $\alpha_i$ such that the following first order condition holds:

$$-\nabla_1 g(-\hat{\alpha}_i, x_i, y_i) + \sum_{j=1}^{n} \hat{\alpha}_j K(x_i, x_j) = 0 \qquad (i = 1, \ldots, n), \tag{24}$$

where $\nabla_1 g$ denotes a subgradient of $g$ with respect to the first component. By the relationship of duality and subgradient in Rockafellar (1970), Section 23, we can rewrite (24) as:

$$\hat{\alpha}_i + \frac{1}{n}\nabla_1 f(p(\hat{\alpha}, x_i), x_i, y_i) = 0 \qquad (i = 1, \ldots, n). \tag{25}$$

Where $\nabla_1 f(v, b, c)$ denotes a subgradient of $f$ with respect to $v$. Now multiply the two sides by $K(x_i, x_\ell)$, and sum over $i$, we have

$$\sum_{i=1}^{n} \hat{\alpha}_i K(x_i, x_\ell) + \frac{1}{n} \sum_{i=1}^{n} \nabla_1 f(p(\hat{\alpha}, x_i), x_i, y_i) K(x_i, x_\ell) = 0 \qquad (\ell = 1, \ldots, n). \tag{26}$$

For any $\alpha$, we multiply (26) by $\alpha_\ell - \hat{\alpha}_\ell$, and sum over $\ell$ to obtain:

$$\sum_{i=1}^{n} \hat{\alpha}_i (p(\alpha, x_i) - p(\hat{\alpha}, x_i)) + \frac{1}{n} \sum_{i=1}^{n} \nabla_1 f(p(\hat{\alpha}, x_i), x_i, y_i)(p(\alpha, x_i) - p(\hat{\alpha}, x_i)) = 0. \tag{27}$$

Using the definition of subgradient, (27) implies

$$\frac{1}{2}(\|p(\alpha, \cdot)\|^2 - \|p(\hat{\alpha}, \cdot)\|^2) + \frac{1}{n} \sum_{i=1}^{n} (f(p(\alpha, x_i), x_i, y_i) - f(p(\hat{\alpha}, x_i), x_i, y_i)) \geq 0.$$

That is, $p(\hat{\alpha}, \cdot)$ achieves the minimum of (23).

Since the above steps can be reversed when the Gram matrix $G(X_n)$ is non-singular, the converse is also true.

# C Proof of Lemma 3.1

For notational simplicity, we assume $k = n$. Using (24), we have for all $i \leq n - 1$:

$$-\nabla_1 g(-\hat{\alpha}_i, x_i, y_i)(\hat{\alpha}_i^{[k]} - \hat{\alpha}_i) + \sum_{j=1}^{n} \hat{\alpha}_j K(x_i, x_j)(\hat{\alpha}_i^{[k]} - \hat{\alpha}_i) = 0.$$

By the definition of subgradient, we have

$$-\nabla_1 g(-\hat{\alpha}_i, x_i, y_i)(\hat{\alpha}_i^{[k]} - \hat{\alpha}_i) \leq g(-\hat{\alpha}_i^{[k]}, x_i, y_i) - g(-\hat{\alpha}_i, x_i, y_i),$$

which can now be equivalently written as:

$$g(-\hat{\alpha}_i, x_i, y_i) - \sum_{j=1}^{n} \hat{\alpha}_j K(x_i, x_j)(\hat{\alpha}_i^{[k]} - \hat{\alpha}_i) \leq g(-\hat{\alpha}_i^{[k]}, x_i, y_i).$$

Summing over $i$, we have

$$\begin{aligned}
&\sum_{i=1}^{n-1} \left[ g(-\hat{\alpha}_i, x_i, y_i) - \sum_{j=1}^{n} \hat{\alpha}_j K(x_i, x_j)(\hat{\alpha}_i^{[k]} - \hat{\alpha}_i) \right] + \frac{1}{2} \sum_{i=1}^{n-1} \sum_{j=1}^{n-1} \hat{\alpha}_i^{[k]} \hat{\alpha}_j^{[k]} K(x_i, x_j) \\
\leq &\sum_{i=1}^{n-1} g(-\hat{\alpha}_i^{[k]}, x_i, y_i) + \frac{1}{2} \sum_{i=1}^{n-1} \sum_{j=1}^{n-1} \hat{\alpha}_i^{[k]} \hat{\alpha}_j^{[k]} K(x_i, x_j) \\
\leq &\sum_{i=1}^{n-1} g(-\hat{\alpha}_i, x_i, y_i) + \frac{1}{2} \sum_{i=1}^{n-1} \sum_{j=1}^{n-1} \hat{\alpha}_i \hat{\alpha}_j K(x_i, x_j).
\end{aligned}$$

The second inequality above follows from the definition of $\hat{\alpha}^{[k]}$. Rearrange the above inequality, and denote $\hat{\alpha}_n^{[k]} = 0$, we obtain:

$$\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} K(x_i, x_j)(\hat{\alpha}_i^{[k]} - \hat{\alpha}_i)(\hat{\alpha}_j^{[k]} - \hat{\alpha}_j) \leq \frac{1}{2}\hat{\alpha}_n^2 K(x_n, x_n).$$

# D Variance style concentration inequality

We prove Lemma 3.2. Given $\{z_1, \ldots, z_n\}$, we denote by $E_i$ expectation with respect to the variables $z_{i+1}, \ldots, z_n$, conditioned on $z_1, \ldots, z_i$. Now, we have

$$Z = E_0 Z + \sum_{i=1}^{n} (E_i Z - E_{i-1} Z). \tag{28}$$

Let $Z_i = E_i Z - E_{i-1} Z$, then $Z_i$ is a function of $\{z_1, \ldots, z_i\}$ and $E_{i-1} Z_i = 0$. Obviously for $j < i$, we have

$$E\, Z_i Z_j = E\, E_{i-1}\, Z_i Z_j = E\, [Z_j E_{i-1} Z_i] = 0.$$

This implies that $E\,[\sum_{i=1}^n Z_i]^2 = E\,\sum_{i=1}^n Z_i^2$. Note that by definition $E_0 Z = EZ$, we thus have from (28) that

$$\mathrm{Var}(Z) = E\,[\sum_{i=1}^{n} Z_i]^2 = E\sum_{i=1}^{n} (E_i Z - E_{i-1} Z)^2.$$

Now use $E_{z_k}$ to denote the expectation over $z_k$, conditioned on all other variables, we obtain from Jensen's inequality:

$$\begin{aligned} E(E_i Z - E_{i-1} Z)^2 \leq & E\, E_i (Z - E_{z_i} Z)^2 \\ \leq & E[E_{z_i}(Z - E_{z_i} Z)^2 + ((Z^{(i)})^2 - E_{z_i} Z)^2] \\ = & E(Z - Z^{(i)})^2. \end{aligned}$$

This shows that

$$\mathrm{Var}(Z) = E\sum_{i=1}^{n} (E_i Z - E_{i-1} Z)^2 \leq E\,\sum_{i=1}^{n} (Z - Z^{(i)})^2.$$

# References


Birgé, L. and Massart, P. (1993). Rates of convergence for minimum contrast estimators. *Probab. Theory Related Fields*, 97(1-2):113–150.

Bousquet, O. and Elisseeff, A. (2002). Stability and generalization. *Journal of Machine Learning Research*, 2:499–526.

Chucker, F. and Samle, S. (2002). On the mathematical foundations of learning. *Bulletin (New Series) of the American Mathematical Society*, 39(1):1–49.

Cristianini, N. and Shawe-Taylor, J. (2000). *An Introduction to Support Vector Machines and other Kernel-based Learning Methods*. Cambridge University Press.

Efron, B. and Stein, C. (1981). The jackknife estimate of variance. *Ann. Statist.*, 9(3):586–596.

Evgeniou, T., Pontil, M., and Poggio, T. (2000). Regularization networks and support vector machines. *Advances in Computational Mathematics*, 13(1):1–50.

Forster, J. and Warmuth, M. (2000). Relative expected instantaneous loss bounds. In *COLT 00*, pages 90–99.

Jaakkola, T. and Haussler, D. (1999). Probabilistic kernel regression models. In *Proceedings of the 1999 Conference on AI and Statistics*.

Joachims, T. (2000). Estimating the generalization performance of an svm efficiently. In *Proceedings of the Seventeenth International Conference on Machine Learning (ICML 00)*, pages 431–438.

Kearns, M. and Ron, D. (1999). Algorithmic stability and sanity-check bounds for leave-one-out cross-validation. *Neural Computation*, 11(6):1427–1453.

Kutin, S. and Niyogi, P. (2002). Almost-everywhere algorithmic stability and generalization error. Technical Report TR-2002-03, Computer Science Department, The University of Chicago.

Mhaskar, H. N., Narcowich, F. J., and Ward, J. D. (1999). Approximation properties of zonal function networks using scattered data on the sphere. *Adv. Comput. Math.*, 11(2-3):121–137.

Narcowich, F. J., Sivakumar, N., and Ward, J. D. (1998). Stability results for scattered-data interpolation on Euclidean spheres. *Adv. Comput. Math.*, 8(3):137–163.

Rockafellar, R. T. (1970). *Convex analysis*. Princeton University Press, Princeton, NJ.

Steele, J. M. (1986). An Efron-Stein inequality for nonsymmetric statistics. *Ann. Statist.*, 14(2):753–758.

van de Geer, S. (2000). *Empirical Processes in M-estimation*. Cambridge University Press.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak convergence and empirical processes*. Springer Series in Statistics. Springer-Verlag, New York.

Vapnik, V. (1998). *Statistical learning theory*. John Wiley & Sons, New York.

Wahba, G. (1990). *Spline Models for Observational Data*. CBMS-NSF Regional Conference series in applied mathematics. SIAM.

Wendland, H. (1998). Error estimates for interpolation by compactly supported radial basis functions of minimal degree. *J. Approx. Theory*, 93(2):258–272.

Yang, Y. and Barron, A. (1999). Information-theoretic determination of minimax rates of convergence. *The Annals of Statistics*, 27:1564–1599.

Zhang, T. (2001). A leave-one-out cross validation bound for kernel methods with applications in learning. In *14th Annual Conference on Computational Learning Theory*, pages 427–443.

Zhang, T. (2002a). Generalization performance of some learning problems in Hilbert functional spaces. In Dietterich, T. G., Becker, S., and Ghahramani, Z., editors, *Advances in Neural Information Processing Systems 14*, Cambridge, MA. MIT Press.

Zhang, T. (2002b). On the dual formulation of regularized linear systems. *Machine Learning*, 46:91–129.

Zhang, T. (2002c). Statistical behavior and consistency of classification methods based on convex risk minimization. *The Annals of Statitics*. to appear.