PREDICTION ALGORITHMS: COMPLEXITY, CONCENTRATION AND CONVEXITY

Peter L. Bartlett

Division of Computer Science and Department of Statistics
University of California, Berkeley
bartlett@eecs.berkeley.edu

Abstract: In this paper, we review two families of algorithms used to estimate large-scale statistical models for prediction problems, kernel methods and boosting algorithms. We focus on the computational and statistical properties of prediction algorithms of this kind. Convexity plays an important role for these algorithms, since they exploit the computational advantages of convex optimization procedures. However, in addition to its computational advantages, the use of convexity in these methods also confers some attractive statistical properties. We present some recent results that show the advantages of convexity for estimation rates, the rates at which the prediction accuracies approach their optimal values. In addition, we present results that quantify the cost of using a convex loss function in place of the real loss function of interest.

Keywords: prediction problems; machine learning; statistical analysis; convex optimization

1. INTRODUCTION

We consider several recent machine learning methods that have proven successful in a variety of multivariate prediction problems. These methods include support vector machines and other kernel methods (Boser *et al.*, 1992; Cortes and Vapnik, 1995; Cristianini and Shawe-Taylor, 2000; Schölkopf and Smola, 2002), and boosting algorithms, such as AdaBoost (Freund and Schapire, 1997). Because these methods involve large-scale statistical models, computational efficiency is an important consideration, and the development of these methods has been influenced by the availability of efficient algorithms for convex optimization. At the heart of each of these methods is the optimization of some convex criterion over a convex set. In addition to its computational advantages, the use of convexity in models and prediction methods has statistical consequences. In this paper we examine some of these consequences.

Consider a multivariate prediction problem in which a covariate $X \in \mathcal{X}$ is associated with a response variable $Y \in \mathcal{Y}$. We quantify the cost of an incorrect prediction using a loss function, $l : \mathcal{Y}^2 \to \mathbb{R}^+$, so that $l(\hat{y}, y)$ is the loss incurred when $\hat{y}$ is predicted but the correct response is y. The aim is to choose a function $f : \mathcal{X} \to \mathcal{Y}$, typically from a particular class $\mathcal{F}$ of functions, so that the expectation of the loss, $\mathbf{E}l(f(X), Y)$, is small, where the expectation is with respect to some probability distribution over $\mathcal{X} \times \mathcal{Y}$. We assume that we do not know this probability distribution, but we get information about it in the form of (X, Y) pairs.

Three key issues arise in prediction problems of this kind:

approximation error — which depends on how well suited the model $\mathcal{F}$ is to the problem,

estimation error — which arises in estimating $f \in \mathcal{F}$ from limited data, and

computational complexity — since the choice of f usually involves a nontrivial optimization problem.

In examining these three issues, we shall focus on the central role played by the convexity of the loss function l, and convexity of the class of functions $\mathcal{F}$.

Our motivating examples of large-scale statistical models $\mathcal{F}$ arise from kernel methods, where $\mathcal{F}$ is a subset of a reproducing kernel Hilbert space of functions on $\mathcal{X}$, and boosting algorithms, where $\mathcal{F}$ is a subset of the linear span of a set of simpler functions. Section 2 describes these algorithms and presents some of their properties.

Both of these methods involve an infinite dimensional convex function class $\mathcal{F}$. As a result, the statistical issue is crucial, and careful regularization is essential to choose an appropriate model complexity, so as to balance the representational power of the model with the statistical cost of estimating it from data. Section 3 presents recent estimation error bounds that measure the complexity of a model in terms of so-called *Rademacher averages*. These results motivate regularization methods that are used to choose an appropriate model complexity.

Both boosting and kernel methods choose a function from a convex class to minimize a convex criterion, even when the performance criterion of interest involves the expectation of a non-convex loss function. Section 4 presents estimation error bounds for convex problems such as these. It turns out that the estimation error decreases to zero more rapidly with the sample size in these cases. These error bounds use a more refined notion of model complexity: a local version of the Rademacher averages. A converse result is also presented, which shows that, in some cases, these faster rates cannot be achieved when the class $\mathcal{F}$ is not convex.

Section 5 considers the effect of convexity on the approximation error. In many practical problems, such as pattern classification problems, the loss function of interest is not convex, but a convex upper bound is used as a surrogate, and it is important to understand how the former behaves when the latter is minimized.

Section 6 discusses the implications of these results for system identification, and presents some directions for future work.

2. PREDICTION ALGORITHMS

In this section we describe two families of prediction algorithms: kernel methods and boosting methods. We consider the following situation: we have a number of independent, identically distributed pairs,

$$(X, Y), (X_1, Y_1), \ldots, (X_n, Y_n)$$

from the product space, $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{Y}$ is a bounded subset of $\mathbb{R}$. (We shall consider extensions to dependent data in Section 6.) We have a bounded loss function, $l : \mathcal{Y}^2 \to [0, 1]$, and a prediction algorithm aims to choose a function $f : \mathcal{X} \to \mathbb{R}$ so that the risk, $R(f) = \mathbf{E}l(Y, f(X))$, is small. The algorithms that we consider choose a function f with a small empirical risk,

$$\mathbf{E}_n l(f(X), Y),$$

where we use $\mathbf{E}_n$ to denote the expectation under the empirical distribution, or sample average,

$$\mathbf{E}_n g(X, Y) = \frac{1}{n} \sum_{i=1}^{n} g(X_i, Y_i).$$

Often the class of functions from which f is chosen is so large that it is always possible to obtain a function with a small empirical risk. In such cases, it is essential to use some kind of regularization, to ensure that the empirical risk is a good approximation of the risk. For instance, if we define a norm on the space of functions, we might choose f to minimize the regularized empirical risk,

$$\mathbf{E}_n l(f(X), Y) + C\|f\|,$$

where C is a regularization coefficient that determines the trade-off between the complexity of f and its performance on the training data.

2.1 Kernel methods

Kernel methods use functions from a subset $\mathcal{F}$ of a *reproducing kernel Hilbert space* (RKHS). For more details on kernel methods, and especially support vector machines (Boser *et al.*, 1992; Cortes and Vapnik, 1995), a popular example of these methods, see (Cristianini and Shawe-Taylor, 2000; Schölkopf and Smola, 2002; Vapnik, 1998).

Recall that a RKHS $\mathcal{H}$ of real functions on $\mathcal{X}$ is a Hilbert space for which there is a reproducing kernel, $k : \mathcal{X}^2 \to \mathbb{R}$, which satisfies the following two properties:

(1) k has the reproducing property: for all $x \in \mathcal{X}$,

$$\langle k(x, \cdot), f \rangle = f(x)$$

(2) k spans $\mathcal{H}$, that is,

$$\operatorname{span}\left(\{k(x, \cdot) : x \in \mathcal{X}\}\right)$$

is dense in H.

Equivalently, we can define a reproducing kernel Hilbert space H as the completion of the span of a symmetric function k that satisfies a certain positive definiteness property. Thus, it is possible to work with a large, even infinite-dimensional, space of functions using only linear combinations of functions of the form $x \mapsto k(x', x)$.

Kernel methods typically choose a function $f \in \mathcal{H}$ so as to minimize a combination of the sample average of a loss function and a regularization term involving the squared norm, $\|f\|_{\mathcal{H}}^2$, of the function in the RKHS. For example, the kernel ridge regression approach is to minimize

$$\frac{1}{n}\sum_{i=1}^{n}(f(x_i) - y_i)^2 + C\|f\|_{\mathcal{H}}^2,$$

where C is a positive regularization coefficient. The following theorem shows that for any regularization term that is an increasing function of the norm, the minimizer has an expansion in terms of the data.

Theorem 1. Fix a kernel k and let $\mathcal{H}$ be the corresponding RKHS. For any function $L : \mathbb{R}^n \to \mathbb{R}$ and any increasing function $\Omega : \mathbb{R} \to \mathbb{R}$, define an objective function,

$$J(f) = L(f(x_1), \ldots, f(x_n)) + \Omega\left(\|f\|_{\mathcal{H}}^2\right). \quad (1)$$

If $J(f^*) = \min_{f \in \mathcal{H}} J(f)$, then for some real coefficients $\alpha_1, \ldots, \alpha_n$, we can write f as

$$f(\cdot) = \sum_{i=1}^{n} \alpha_i k(x_i, \cdot). \quad (2)$$

The proof is straightforward: for any $f \in \mathcal{H}$, let f_0 be the projection of f on to the subspace spanned by the $k(x_i, \cdot)$. Then the reproducing property implies

$$f(x_i) = \langle f, k(x_i, \cdot)\rangle = \langle f_0, k(x_i, \cdot)\rangle = f_0(x_i),$$

which, with Pythagorus' theorem, implies that

$$\begin{aligned} J(f) &= L(f_0(x_1), \ldots, f_0(x_n)) \\ &\quad + \Omega\left(\|f_0\|_{\mathcal{H}}^2 + \|f - f_0\|_{\mathcal{H}}^2\right) \\ &\geq J(f_0), \end{aligned}$$

with equality iff f lies in the span of the $k(x_i, \cdot)$. See (Kimeldorf and Wahba, 1971) for the first result of this form and (Schölkopf and Smola, 2002) for a review of subsequent refinements.

This result implies that an objective function that is of this form (a norm-regularized sample average of a loss function) can be expressed entirely in terms of the *Gram matrix* of inner products, $K \in \mathbb{R}^{n \times n}$, with

$$K_{ij} = k(x_i, x_j).$$

Specifically, for any f of the form (2), we have

$$\begin{aligned} (f(x_1), \ldots, f(x_n))' &= K\alpha, \\ \|f\|_{\mathcal{H}}^2 &= \alpha' K \alpha. \end{aligned}$$

where $\alpha = (\alpha_1, \ldots, \alpha_n)' \in \mathbb{R}^n$. Thus, for the optimization of a $J : \mathcal{H} \to \mathbb{R}$ of the form (1), it suffices to consider the dual optimization problem

$$J^*(\alpha) = L(K\alpha) + \Omega(\alpha' K \alpha).$$

This representation has a number of advantages. For typical choices of L and Ω, it is a finite-dimensional convex optimization problem, often a quadratic program. For many loss functions, the solution is sparse, that is, the number of non-zero α_i is small, which has advantages for the efficiency of evaluating the function f. It is only necessary to compute k on the finite sample, which avoids the need to work explicitly with the inner product space $\mathcal{H}$, which may be large. On the other hand, there can be advantages to using the primal form of the optimization problem (1); see, for example, (Suykens *et al.*, 2002).

The major design decision in algorithms of this kind is the choice of the kernel k. For vector data, a common choice is the gaussian radial basis kernel,

$$k(u, v) = \exp\left(-\|u - v\|^2/\sigma^2\right),$$

is commonly used. In this case, the corresponding RKHS is infinite dimensional. Often, information about the problem domain can be used to choose an inner product that is more appropriate for the problem. Kernel methods have been successfully applied to document discrimination using kernels that compare word frequency counts (Joachims, 1998). In this case, the dimensionality of the RKHS is not infinite, but it is very large; it is the number of distinct words. More sophisticated kernels have been defined for text and other sequence data, such as genome sequences (Watkins, 2000), for images (Romdhani *et al.*, 2001), and for data for which an accurate generative model can be estimated (Tsuda *et al.*, 2002).

2.2 Boosting algorithms

Boosting algorithms make predictions using functions from the linear span of a fixed class of functions. The first boosting algorithm arose in a constructive proof of the following surprising result about two-class pattern classification (Schapire, 1990).

Theorem 2. Consider a class $\mathcal{G}$ of functions mapping from $\mathcal{X}$ to $\{\pm 1\}$, a sequence

$$(x_1, y_1), \ldots, (x_n, y_n) \in \mathcal{X} \times \{\pm 1\}$$

where $y_i = f^*(x_i)$ for some $f^* \in \mathcal{G}$, and the discrete loss, $l(\hat{y}, y)$, defined as the indicator function of $\hat{y} \neq y$. If there is a polynomial-time algorithm (we call it the *base learning algorithm*) that can choose an $f \in \mathcal{G}$ that has empirical risk slightly better than random guessing,

$$\mathbf{E}_n l(f(X), Y) \le \frac{1}{2} - \gamma,$$

then there is a polynomial time 'boosting' algorithm that can choose a g that has empirical risk smaller than ϵ.

The boosting algorithm from the proof of this result repeatedly calls the learning algorithm for $\mathcal{F}$, and forms a thresholded linear combination of the functions that it returns. Although this result makes strong assumptions about the process generating the data, when a particular boosting algorithm—AdaBoost (Freund and Schapire, 1997)—was tried in practice, it gave significant performance improvements over the learning algorithm for the class $\mathcal{G}$ of basis functions (Drucker and Cortes, 1996; Freund and Schapire, 1996). Typical basis function classes used in practice include decision trees, regression trees, and neural networks.

We can understand algorithms such as AdaBoost as a particular example of the greedy convex optimization approach described in Figure 1. In this algorithm, $\mathcal{G}$ is a class of $[-1, 1]$-valued functions on $\mathcal{X}$ (the basis functions from Theorem 2) and $A : \mathrm{span}(\mathcal{G}) \to \mathbb{R}$ is a convex objective function. Figure 2 shows an analogous algorithm for convex combinations, that is, functions from the convex hull of $\mathcal{G}$.

Fig. 1. Greedy boosting algorithm for linear combinations.

set $f_0 = 0$
for $t = 1, \ldots, T$
  choose $\alpha_t \in \Lambda \subseteq \mathbb{R}$ and $g_t \in \mathcal{G}$ to minimize

$$A\left(f_{t-1} + \alpha_t g_t\right).$$

  set $f_t = f_{t-1} + \alpha_t g_t$
return f_T.

Fig. 2. Greedy boosting algorithm for convex combinations.

set $f_0 = 0$
for $t = 1, \ldots, T$
  choose $\alpha_t \in [0, 1]$ and $g_t \in \mathcal{G}$ to minimize

$$A\left((1 - \alpha_t) f_{t-1} + \alpha_t g_t\right).$$

  set $f_t = (1 - \alpha_t) f_{t-1} + \alpha_t g_t$
return f_T.

The Adaboost algorithm is a special case of the algorithm of Figure 1, where the objective function A is defined by

$$A(f) = \mathbf{E}_n \exp(-Y f(X)),$$

where $Y \in \{\pm 1\}$ is the binary response variable. This approach has been extended to a variety of related objective functions, such as logistic regression (Friedman *et al.*, 2000). With the freedom of choosing an element f of the (possibly infinite-dimensional) space of functions spanned by $\mathcal{G}$, regularization is essential. In some cases, allowing the algorithm to run for only a small number of iterations is a successful regularization strategy. This is the approach typically used with Adaboost. With some assumptions on the size of the steps α_t taken by the algorithm, we can view early stopping as constraining the one norm of the parameters. Of course, restricting the functions f_t to be convex combinations, as in the algorithm of Figure 2, constrains this norm more directly.

For a quadratic objective function,

$$A(f) = \mathbf{E}_n (Y - f(X))^2,$$

the algorithm of Figure 2 is equivalent to the greedy optimization approach proposed in (Jones, 1997) and applied to neural network parameter optimization in (Barron, 1993; Lee *et al.*, 1996). Indeed, a neural network with a one norm constraint on its parameters is an example of the class of functions used by the algorithm of Figure 2. Related algorithms have been studied in (Friedman *et al.*, 2000; Mason *et al.*, 2000; Zhang, 2003; Rätsch *et al.*, 2001).

The following theorem, from (Zhang, 2003), shows that the boosting algorithm of Figure 2 converges to an optimal solution.

Theorem 3. Suppose that, for some finite M, the convex loss function $l : \mathcal{Y}^2 \to \mathbb{R}$ is such that for all $y' \in \mathcal{Y}$, $\partial^2/\partial y^2 l(y, y') \le M$. Let $A(f) = \mathbf{E}_n l(f(X), Y)$. Then the algorithm of Figure 2 produces a function f_T satisfying

$$A(f_T) \le \inf_{f \in \mathrm{co}(\mathcal{G})} A(f) + \frac{2M}{(T + 2)}.$$

Thus, if the empirical risk, $\mathbf{E}_n l_g$, can be efficiently optimized over the class $\mathcal{G}$, then the empirical risk can be efficiently optimized over the convex hull of $\mathcal{G}$.

There are analogous results for the convergence rates of the boosting algorithms of Figure 1 (Zhang and Yu, 2003).

3. COMPLEXITY, CONCENTRATION, AND ESTIMATION ERROR

Suppose a prediction algorithm produces a function f from some class $\mathcal{F}$, with the aim of minimizing the risk, $\mathbf{E} l(Y, f(X))$. The optimal choice is an $f \in \mathcal{F}$ that achieves the minimal risk,

$$\inf_{f \in \mathcal{F}} \mathbf{E} l(Y, f(X)).$$

Since we only receive information about the stochastic process that generates the data from the finite sample, our choice $\hat{f} \in \mathcal{F}$ suffers an estimation error,

$$\mathbf{E} l(Y, \hat{f}(X)) - \inf_{f \in \mathcal{F}} \mathbf{E} l(Y, f(X)).$$

In this section, we present bounds on the estimation error for prediction algorithms. Such bounds motivate regularization schemes, since they show how the estimation error depends on the complexity of the function class. For instance, we shall see that the regularization term that is commonly used in kernel methods, involving the RKHS norm, can be motivated by upper bounds on the estimation error. It is possible to show that a bound on estimation error can be used to define a *complexity regularization scheme*, that is, an algorithm for choosing the model complexity by penalizing complex models. For instance, Theorems 5 and 7 below show that, for any λ, with high probability over the sample, every f in an RKHS $\mathcal{H}$ satisfying $\|f\|_{\mathcal{H}} \leq \lambda$ has

$$\mathbf{E} f \leq \mathbf{E}_n f + \lambda c_n,$$

where c_n is a constant that depends on n, $\mathcal{H}$, and properties of the probability distribution P. This motivates a complexity regularization scheme in which we choose $f \in \mathcal{H}$ to minimize the upper bound,

$$\mathbf{E}_n f + c_n \|f\|_{\mathcal{H}}.$$

This has the form of the objective function in Theorem 1. It is possible to give performance guarantees for (slightly modified versions of) these complexity regularization schemes; see (Bartlett *et al.*, 2002) for details.

The bounds in this section are based on results describing the convergence of sample averages to expectations, uniformly over the class $\mathcal{F}$. We use a measure of complexity of a function class known as the *Rademacher averages.*

Definition 4. Suppose that $\mathcal{F}$ is a class of real-valued functions defined on $\mathcal{X}$, and P is a probability distribution on $\mathcal{X}$. Then the *Rademacher averages of* $\mathcal{F}$ are

$$R_n(\mathcal{F}) = \mathbf{E} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i f(X_i).$$

where $X_1, \ldots, X_n \in \mathcal{X}$ are independently chosen according to P, and $\epsilon_1, \ldots, \epsilon_n$ are independent uniform $\{\pm 1\}$-valued random variables (Rademacher random variables). The *empirical Rademacher averages of* $\mathcal{F}$ are

$$\hat{R}_n(\mathcal{F}) = \mathbf{E} \left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i f(X_i) \,\middle|\, X_1, \ldots, X_n \right].$$

The Rademacher averages are a natural notion of the complexity of the class $\mathcal{F}$ with respect to the probability distribution P. The ϵ_i define a random vector in the n-cube. If $R_n(\mathcal{F})$ is large, then for many such random vectors, there is a function that defines a vector $(f(X_1), \ldots, f(X_n))$ with large inner product with the random direction. In that case, we should not expect a small empirical average of an $f \in \mathcal{F}$ to imply a small expectation. The converse is also true.

Theorem 5. For $\mathcal{F} \subseteq [0,1]^{\mathcal{X}}$ and $\delta > 0$, with probability at least $1 - \delta$ over i.i.d. sequences $X_1, \ldots, X_n \in \mathcal{X}$, we have

$$\sup_{f \in \mathcal{F}} \left(\mathbf{E} f - \frac{1}{n} \sum_{i=1}^{n} f(X_i) \right) \leq 2 R_n(\mathcal{F}) + \sqrt{\frac{\log 1/\delta}{2n}}.$$

Thus, if $\hat{f} \in \mathcal{F}$ minimizes $\mathbf{E}_n f$, then with probability at least $1 - \delta$,

$$\mathbf{E} \hat{f} \leq \inf_{f \in \mathcal{F}} \mathbf{E} f + 2 R_n(\mathcal{F}) + c \sqrt{\frac{\log 1/\delta}{n}}.$$

The proof is in two steps. The first step uses the bounded differences concentration inequality (McDiarmid, 1989), to show that

$$Z = \sup_{f \in \mathcal{F}} \left(\mathbf{E} f - \frac{1}{n} \sum_{i=1}^{n} f(X_i) \right)$$

is with high probability close to its expectation. (The conditions of this concentration result require that the random variable Z changes only slightly whenever an X_i changes. This is clearly satisfied here, because of the boundedness of the range of functions in $\mathcal{F}$.) The second step uses an easy symmetrization argument to show that

$$\mathbf{E} Z \leq 2 R_n(\mathcal{F}).$$

(There is also a partial converse to this result, which shows that the inequality cannot be loose unless the class $\mathcal{F}$ is small.)

It is easy to see that we can accurately estimate $R_n(\mathcal{F})$ from the data: it is the expectation of a random variable that also satisfies the conditions of the bounded differences concentration inequality. This implies that, with probability at least $1 - \delta$ over the random X_i and ϵ_i,

$$\mathbf{E} \hat{f} \leq \inf_{f \in \mathcal{F}} \mathbf{E} f + \sup_{f \in \mathcal{F}} \frac{2}{n} \sum_{i=1}^{n} \epsilon_i f(X_i) + c \sqrt{\frac{\log 1/\delta}{n}}$$

for some universal constant c.

In our application, we will apply these results to the loss functions $l_f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ defined by

$$l_f(x, y) = l(f(x), y).$$

Suppose that the function $\alpha \mapsto l(\alpha, y)$ is L-Lipschitz for all y, that is, there is an L such that for all $y \in \mathcal{Y}$ and all α_1, α_2 in the range of functions in $\mathcal{F}$, we have

$$|l(\alpha_1, y) - l(\alpha_2, y)| \leq L|\alpha_1 - \alpha_2|.$$

Then we can relate the Rademacher averages for the class of loss functions,

$$l_{\mathcal{F}} = \{l_f : f \in \mathcal{F}\},$$

to the Rademacher averages for the class $\mathcal{F}$. Indeed, the Ledoux-Talagrand contraction inequality (Ledoux and Talagrand, 1991) implies that

$$R_n(l_{\mathcal{F}}) \leq L R_n(\mathcal{F}).$$

The Rademacher averages can be bounded in terms of other notions of complexity, such as covering numbers and combinatorial dimensions (see, for example, (Anthony and Bartlett, 1999; Vidyasagar, 1997)), which allows us to use Theorem 5 to obtain many other error estimates, in most cases in an improved form; see, for example, (Bartlett *et al.*, 2002; Bartlett and Mendelson, 2002; Koltchinskii and Panchenko, 2002; Mendelson, 2002*b*; Dudley, 1999).

For the prediction algorithms described in Section 2, we can compute the Rademacher averages directly. The following theorem is straightforward. It shows that the Rademacher averages of the scaled convex hull of a class $\mathcal{G}$ are equal to the scaled Rademacher averages of $\mathcal{G}$.

Theorem 6. For a class $\mathcal{F} \subseteq \mathbb{R}^{\mathcal{X}}$,

(1) For $\lambda \in \mathbb{R}$, if $\lambda\mathcal{F} = \{\lambda f : f \in \mathcal{F}\}$, then $R_n(\lambda\mathcal{F}) = |\lambda| R_n(\mathcal{F})$.
(2) $R_n(\mathrm{co}(\mathcal{F})) = R_n(\mathcal{F})$.
(3) Hence, for $\mathcal{G} = \lambda\,\mathrm{co}(\mathcal{F})$),

$$R_n(\mathcal{G}) = |\lambda| R_n(\mathcal{F}).$$

For kernel classes, the empirical Rademacher averages can be estimated by the sum of the eigenvalues of the Gram matrix, which is readily computable from the data. There is a similar result for the Rademacher averages, in terms of the eigenvalues of a certain integral operator.

Theorem 7. For a kernel k and the corresponding RKHS $\mathcal{H}$, define $\mathcal{F} = \{f \in \mathcal{H} : \|f\|_{\mathcal{H}} \leq \lambda\}$. Then we have

$$\hat{R}_n(\mathcal{F}) \leq \lambda \frac{\sqrt{\mathrm{trace}(K)}}{n} = \lambda \sqrt{\frac{\bar{R}^2}{n}},$$

where $K_{ij} = k(X_i, X_j)$ is the $n \times n$ Gram matrix, and so its trace is n times the average squared norm of the point evaluation functionals in $\mathcal{H}$,

$$\bar{R}^2 = \frac{1}{n}\sum_{i=1}^{n} \|k(X_i, \cdot)\|_{\mathcal{H}}^2.$$

Furthermore,

$$R_n(\mathcal{F}) \leq \lambda \sqrt{\frac{\sum_{j=1}^{\infty} \theta_j}{n}},$$

where $\theta_1, \theta_2, \ldots$ are the non-increasing eigenvalues of the integral operator $T_k : L_2(P) \to L_2(P)$ defined by

$$T_k f = \int_{\mathcal{X}} k(\cdot, x) f(x) dP(x).$$

We can estimate the Rademacher averages for certain convex neural network classes. For instance, consider the class $\mathcal{F}_B$ of two layer networks defined on $\mathcal{X} = \mathbb{R}^d$, with a bound on the one-norm of the parameters, that is, functions in $\mathcal{F}_B$ are of the form

$$x \mapsto \sum_{i=1}^{k} w_i \sigma\left(v_i^T x\right),$$

for any $k \geq 1$, $w \in \mathbb{R}^k$, and $v_i \in \mathbb{R}^d$ satisfying $\|w\|_1 \leq B$ and $\|v_i\|_1 \leq B$. Here, $\sigma : \mathbb{R} \to [-1, 1]$ is a monotonic nonlinearity. This class is infinite-dimensional. However, we can write

$$\mathcal{F}_B = B\,\mathrm{co}\left(\sigma\left(B\,\mathrm{co}(\mathcal{G})\right)\right),$$

where $\mathcal{G} = \{(x_1, \ldots, x_d) \mapsto x_j : 1 \leq j \leq d\}$ and we use $\sigma(A)$ to denote $\{x \mapsto \sigma(f(x)) : f \in A\}$. Thus, using the properties of the Rademacher averages we have already encountered, together with an estimate for the Rademacher averages for a finite set (Massart, 2000), we have the following result

Theorem 8. For the neural network class $\mathcal{F}_B$ defined above,

$$R_n(\mathcal{F}_B) \leq B^2 \sqrt{\frac{2 \log d}{n}},$$

where B is the bound on the parameters and d is the dimensionality of the input space.

4. CONVEXITY AND ESTIMATION ERROR

The approach in the previous section relied on convergence of empirical averages to expectations, uniformly over a class $\mathcal{F}$. This uniform convergence property suffices to ensure that the expected loss (risk) of the empirical risk minimizer $\hat{f}$ is small. However, it is not necessary for the deviations between empirical averages and expectations to be small uniformly over the class. In particular, if the worst-case difference between empirical averages and expectations occurs far from $\hat{f}$, these estimates might be too conservative. Indeed, in many problems it is possible to do better. Rather than the loss class $l_{\mathcal{F}}$, we consider the class of excess loss,

$$\begin{aligned}\mathcal{L}_{\mathcal{F}} &= \{l_f - l_{f^*} : f \in \mathcal{F}\} \\ &= \{(x,y) \mapsto l(f(x),y) - l(f^*(x),y) : f \in \mathcal{F}\},\end{aligned}$$

where $f^* \in \mathcal{F}$ is the minimizer of the expected loss, $f^* = \arg\min_{f \in \mathcal{F}} \mathbf{E} l_f$. Then if $\mathcal{F}$ is convex and l is strictly convex, we can obtain faster convergence rates. The crucial property of the loss function is its *modulus of convexity.*

Definition 9. (Modulus of convexity). For a convex function $f : \mathbb{R} \to \mathbb{R}$, the *modulus of convexity* of f is the function $\delta : [0,\infty) \to [0,\infty]$ satisfying

$$\delta(\epsilon) = \inf\left\{ \frac{f(x_1)+f(x_2)}{2} - f\left(\frac{x_1+x_2}{2}\right) \right. \\ \left. : x_1, x_2 \in S,\, |x_1, x_2| \geq \epsilon \right\}.$$

In the following theorem, we use a more refined measure of the complexity of the excess loss class $\mathcal{L}_{\mathcal{F}}$, which we call the *local Rademacher averages.* These are the Rademacher averages of a small subset,

$$R_n\left(\{g \in \mathcal{L}_{\mathcal{F}} : \mathbf{E}g \leq \epsilon\}\right).$$

Theorem 10. There is a constant K for which the following holds. Consider a loss function $l : \mathcal{Y}^2 \to [0, B]$. Suppose that, for all y, the function $l(\cdot, y)$ is L-Lipschitz and strictly convex, with modulus of convexity $\delta(\epsilon) \geq c\epsilon^r$. Define $\beta = \min(1, 2/r)$. Fix a convex class $\mathcal{F}$ of real functions on $\mathcal{X}$. For i.i.d. data $(X_1, Y_1), \ldots, (X_n, Y_n)$, let $\hat{f} \in \mathcal{F}$ be the minimizer of the empirical risk, $\mathbf{E}_n l(f(X), Y)$. Then with probability at least $1 - e^{-x}$,

$$\mathbf{E} l_{\hat{f}} \leq \mathbf{E} l_{f^*} + \epsilon,$$

where

$$\begin{aligned}\epsilon &= K \max\left\{ \epsilon^*, \left(\frac{c_r L^2 x}{n}\right)^{1/(2-\beta)}, \frac{BLx}{n} \right\}, \\ \epsilon^* &\geq R_n\left(\{g \in \mathcal{L}_{\mathcal{F}} : \mathbf{E}g \leq \epsilon^*\}\right), \\ c_r &= \begin{cases} (2c)^{-2/r} & \text{if } r \geq 2, \\ (2c)^{-1} B^{2-r} & \text{otherwise.} \end{cases}\end{aligned}$$

The theorem is from (Bartlett *et al.*, 2003*b*). It simplifies and generalizes results from (Mendelson, 2002*b*), which built on ideas of (Lee *et al.*, 1996). The proof is based on the observation that for a strictly convex loss the excess loss function has well-behaved variance: if l is roughly quadratic, then for any probability distribution we have

$$\mathbf{E}\left(l_f - l_{f^*}\right)^2 \leq c\mathbf{E}\left(l_f - l_{f^*}\right),$$

for some universal constant c. Having control of the variance implies that, even though the fluctuations of $\mathbf{E}_n l_f$ might be large, the fluctuations of $\mathbf{E}_n l_f$ and $\mathbf{E}_n l_{f^*}$ are strongly correlated when f is not too far from f^*, as it is when f is the empirical minimizer. This fact, together with some recent concentration inequalities (see (Bousquet, 2002) and the references cited there), implies the theorem.

For the prediction algorithms described in Section 2, we can also compute the local Rademacher averages. In the following theorem, we consider the example of scaled convex hulls, a class used by boosting algorithms. The assumption that $\mathcal{G}$ is finite dimensional can be replaced by an assumption on its covering numbers or a suitable combinatorial dimension; see (Bartlett *et al.*, 2003*b*) for details.

Theorem 11. For the class $\mathcal{F} = \lambda\,\mathrm{co}(\mathcal{G}))$, where $\mathcal{G}$ is a subset of a d-dimensional subspace, and a loss function l for which the function $l(\cdot, y)$ is uniformly Lipschitz and strictly convex, with modulus of convexity $\delta(\epsilon) \geq c\epsilon^2$, we have

$$\begin{aligned}&R_n\left(\{g \in \mathcal{L}_{\mathcal{F}} : \mathbf{E}g \leq \epsilon\}\right) \\ &\leq c_1 \lambda \left(n^{-(d+2)/(2d+2)} + n^{-1/2}\epsilon^{1/(d+2)}\right).\end{aligned}$$

The result involving Rademacher averages of the full class, Theorem 6, cannot give rates better than $n^{-1/2}$, whereas the rate above is always faster, particularly if d is small.

Similarly, for kernel classes we get improved rates. This theorem combines results from (Bartlett *et al.*, 2003*a*) and (Mendelson, 2002*a*).

Theorem 12. For a kernel k and the corresponding RKHS $\mathcal{H}$, define $\mathcal{F} = \{f \in \mathcal{H} : \|f\|_{\mathcal{H}} \leq \lambda\}$. Suppose l is a loss function for which the function $l(\cdot, y)$ is uniformly Lipschitz and strictly convex, with modulus of convexity $\delta(\epsilon) \geq c\epsilon^2$. Then we have

$$\begin{aligned}&R_n\left(\{g \in \mathcal{L}_{\mathcal{F}} : \mathbf{E}g \leq \epsilon\}\right) \\ &\leq c_1 \lambda \left(\frac{1}{n} \sum_{j=1}^{\infty} \min\{c_2\epsilon, \theta_i\}\right)^{1/2},\end{aligned}$$

where $\theta_1, \theta_2, \ldots$ are the non-increasing eigenvalues of the integral operator $T_k : L_2(P) \to L_2(P)$ defined by

$$T_k f = \int_{\mathcal{X}} k(\cdot, x) f(x) dP(x).$$

Since the trace of T_k is always at least as large as the sum in this expression, the bound is no worse than that involving the Rademacher averages of the full class, Theorem 7. If the eigenvalues decrease rapidly, the estimation error rate we obtain by exploiting the convexity information can be considerably better. For instance, if they decrease exponentially quickly and the loss function has a quadratic modulus of convexity, then Theorems 10 and 12 imply that, with high probability,

$$\mathbf{E}l_{\hat{f}} \leq \inf_{f:\|f\|\leq\lambda} \mathbf{E}l_f + \lambda^2 c_n,$$

where c_n is a constant that depends on the sample size n and the eigenvalues of T_k. Bounds of this kind motivate a complexity regularization scheme in which we choose $f \in \mathcal{H}$ to minimize

$$\mathbf{E}_n f + c_n \|f\|_H^2.$$

Again, this has the form of the objective function in Theorem 1.

In the previous section, an easy concentration argument showed that we could estimate the Rademacher averages of the full class from the empirical Rademacher averages. For the local Rademacher averages considered here, something similar is also possible. This is more complicated, in part because the function class

$$\{l_f - l_{f^*} : f \in \mathcal{F}, \mathbf{E}\left(l_f - l_{f^*}\right) \leq \epsilon\}$$

is defined in terms of the risk minimizer f^* and the probability distribution, which are both unknown. It turns out that for a convex class $\mathcal{F}$ and strictly convex loss l, we can replace these local Rademacher averages with the following empirical local Rademacher averages.

$$\hat{\xi}_{\mathcal{F}}(\epsilon) = \hat{R}_n\left(\left\{f \in \mathcal{F} : \mathbf{E}_n\left(f - \hat{f}\right)^2 \leq c\epsilon\right\}\right),$$

where $\hat{f}$ is the minimizer of the empirical risk, $\mathbf{E}_n l_f$. In the case of kernel classes, this allows us to replace the eigenvalues of the integral operator T_k with the eigenvalues of the Gram matrix, since

$$\hat{\xi}_{\mathcal{F}}(\epsilon) \leq c\left(\frac{1}{n}\sum_{i=1}^{n} \min\left\{\frac{\lambda_i}{n}, \epsilon\right\}\right)^{1/2},$$

where $\lambda_1, \cdots, \lambda_n$ are the eigenvalues of the Gram matrix $K_{ij} = k(X_i, X_j)$. See (Bartlett *et al.*, 2003*a*) for details.

These results show that, if the loss function is strictly convex and the class $\mathcal{F}$ is convex, then the estimation error decreases rapidly to zero as a function of the sample size n, in some cases considerably faster than $n^{-1/2}$. There is a converse result for the condition on the convexity of the class $\mathcal{F}$. The following theorem, from (Lee *et al.*, 1998), shows that for a nonconvex class with quadratic loss, the minimax convergence rate of the estimation error

$$\mathbf{E}l(\hat{f}(X), Y) - \inf_{f\in\mathcal{F}} \mathbf{E}l(f(X), Y)$$

cannot be faster than $n^{-1/2}$.

Theorem 13. Fix a function class $\mathcal{F}$ and the quadratic loss function, $l(\hat{y}, y) = (\hat{y} - y)^2$. For $\epsilon, \delta, B > 0$, define $n(\epsilon, \delta, B)$ as the smallest integer satisfying the following condition. For every sample size $n \geq n(\epsilon, \delta, B)$,

$$\inf_{\hat{f}} \sup_{P} P^n \Big\{D_n : \mathbf{E}l(\hat{f}(D_n, X), Y) > \inf_{f\in\mathcal{F}} \mathbf{E}l(f(X), Y) + \epsilon\Big\} \leq \delta,$$

where the infimum is over prediction algorithms $\hat{f}$ that take as input a random X and an i.i.d. data sequence $D_n = (X_1, Y_1, \ldots, X_n, Y_n)$ chosen according to the probability distribution P, and the supremum is over all probability distributions P on $\mathcal{X} \times [-B, B]$.

If, for some probability distribution P_X on $\mathcal{X}$, the closure of $\mathcal{F}$ in $L_2(P_X)$ is not convex, then for sufficiently large B, there is a constant c for which

$$n(\epsilon, \delta, B) \geq c\frac{\ln(1/\delta)}{\epsilon^2}.$$

5. CONVEXITY AND APPROXIMATION

It is clear from the statistical and computational results of the previous sections that if we have a convex loss function and a nonconvex function class $\mathcal{F}$, then working with the convex hull of $\mathcal{F}$ does not add significant computational cost, and can reduce the estimation error. Clearly, it increases the function class, so it can only help the approximation properties. That is, if the loss function is convex, it often makes sense to ensure that the function class is convex.

Is the same true of the loss function? That is, if the function class is convex, should we replace a non-convex loss function with a convex one? In some cases, this gives significant computational advantages. One important example arises in pattern classification problems. Consider two-class pattern classification, which is a prediction problem of the kind we have been considering where the response variable Y takes values in the binary set $\{\pm 1\}$ and the loss function is the indicator function of $\mathrm{sign}(f(x)) \neq y$. For most interesting function classes, the problem of approximately minimizing risk in a minimax setting is provably hard (Arora *et al.*, 1997, for example). However, if we replace this nonconvex loss function by a surrogate of the form $\tilde{l}(f(x), y) = \phi(yf(x))$, where ϕ is convex, the optimization problem can often be solved efficiently. This is the approach used by support vector machines (kernel methods for pattern classification), AdaBoost, and logistic regression. In such cases, it is important to understand the impact of this choice on the statistical properties of the method. In particular, we are interested in how the risk,

$$\begin{aligned} R(f) &= \Pr(Y \neq \mathrm{sign}(f(X))) \\ &= \mathbf{E}l(f(X), Y), \end{aligned}$$

is related to the ϕ-risk,

$$R_\phi(f) = \mathbf{E}\phi(Yf(X)),$$

since the latter is the quantity that these methods minimize. The following theorem, from (Bartlett *et al.*, 2003*b*), considers the excess risk, that is, how much the risk of a function exceeds the optimal or Bayes risk,

$$R^* = \inf_f \Pr(Y \neq \mathrm{sign}(f(X))),$$

where the infimum is over all measurable f, and the excess ϕ-risk, that is, how much the ϕ-risk exceeds the optimal ϕ-risk,

$$R^*_\phi = \inf_g \mathbf{E}\phi(Yg(X)).$$

We first define a functional transform of the loss function.

Definition 14. We define the *ψ-transform* of a loss function as follows. Given $\phi : \mathbb{R} \to [0,\infty)$, define the function $\psi : [0,1] \to [0,\infty)$ by $\psi = \tilde{\psi}^{**}$, where

$$\tilde{\psi}(\theta) = \inf_{\alpha:\alpha\leq 0}\left(\frac{1+\theta}{2}\phi(\alpha) + \frac{1-\theta}{2}\phi(-\alpha)\right) - \inf_{\alpha\in\mathbb{R}}\left(\frac{1+\theta}{2}\phi(\alpha) + \frac{1-\theta}{2}\phi(-\alpha)\right).$$

and $g^{**} : [0,1] \to \mathbb{R}$ is the Fenchel-Legendre biconjugate of $g : [0,1] \to \mathbb{R}$, which is defined by

$$\mathrm{epi}\ g^{**} = \overline{\mathrm{co}}\ \mathrm{epi}\ g.$$

Here, $\overline{\mathrm{co}}\ S$ is the closure of the convex hull of the set S, and epi g is the epigraph of the function g, that is, the set $\{(x,t) : x \in [0,1], g(x) \leq t\}$.

Theorem 15. For any nonnegative loss function ϕ defined on $\mathbb{R}$, any measurable $f : \mathcal{X} \to \mathbb{R}$ and any probability distribution on $\mathcal{X} \times \{\pm 1\}$, we have

$$\psi(R(f) - R^*) \leq R_\phi(f) - R^*_\phi.$$

This bound shows that minimizing the ϕ-risk leads to minimal risk provided that the nonnegative function ψ is zero only at zero. It turns out that for convex ϕ this is the case if and only if ϕ is differentiable at zero and has $\phi'(0) < 0$. It can be shown that the bound cannot be improved anywhere, in general (Bartlett *et al.*, 2003*b*). Thus, it quantifies the cost of replacing the loss function of interest with a convex surrogate. It would be interesting to extend this result to other nonconvex loss functions; the definition of ψ is specific to the discrete loss l.

6. SYSTEM IDENTIFICATION, DEPENDENT DATA, AND OPEN PROBLEMS

We can express certain system identification problems as prediction problems of the kind studied here. Define $\mathcal{X}$ as the space of input sequences and $\mathcal{Y}$ as the space of system outputs (for instance, $\mathcal{Y} \subseteq \mathbb{R}$). Consider a class $\mathcal{F}$ of functions mapping from $\mathcal{X}$ to $\mathcal{Y}$, the family of models that we wish to identify.

Most of the results presented in this paper are based on the assumption that the pairs

$$(X_1, Y_1), \ldots, (X_n, Y_n), (X, Y)$$

are independent and identically distributed. In a system identification problem, this is not a reasonable assumption: the input sequences are of the form $X_t = (u_t, u_{t-1}, u_{t-2}, \ldots)$ for some infinite sequence (u_t). However, if the sequences (u_t) and (y_t) and the model class $\mathcal{F}$ satisfy some reasonable conditions, it is not difficult to extend the results in this paper to this setting. In particular, suppose that

(1) (u_t) and (y_t) are stationary stochastic processes satisfying a suitable mixing condition, so that (X_t, Y_t) and (X_{t+s}, Y_{t+s}) become 'closer to independent' in some sense as s increases, and
(2) models in $\mathcal{F}$ uniformly satisfy a fading memory property, which implies that for all input sequences,

$$f(u_t, u_{t-1}, u_{t-2}, \ldots)$$

and

$$f(u_t, \ldots, u_{t-s})$$

become close as s increases.

Then it is possible to relate the behavior of an empirical risk minimization algorithm in this setting to its behavior with i.i.d. data. In this way, many of the results presented above, including Theorems 5 and 10, can be extended to the system identification setting. In particular, this would show how rapidly the expected loss of the empirical minimizer converges to the minimal value on the model class $\mathcal{F}$. For details on this and related approaches see, for example, (Nobel and Dembo, 1993; Yu, 1994; Weyer, 2000; Vidyasagar and Karandikar, 2001; Karandikar and Vidyasagar, 2002)

On the other hand, it is not as simple to extend data-dependent error estimates to this setting, such as the bounds in terms of the empirical Rademacher averages, and their consequences in terms of eigenvalues of the kernel matrices. These results are appealing because they use the data to measure the complexity of the function class. It would be interesting to develop versions of these results that apply to mixing stochastic processes.

REFERENCES

Anthony, Martin and Peter L. Bartlett (1999). *Neural Network Learning: Theoretical Foun-*

dations. Cambridge University Press. Cambridge, UK.

Arora, S., L. Babai, J. Stern and Z. Sweedyk (1997). The hardness of approximate optima in lattices, codes, and systems of linear equations. *Journal of Computer and System Sciences* **54**, 317–331.

Barron, Andrew R. (1993). Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory* **39**, 930–945.

Bartlett, P. L. and S. Mendelson (2002). Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research* **3**, 463–482.

Bartlett, P. L., O. Bousquet and S. Mendelson (2003*a*). Local Rademacher complexities. Technical report. University of California at Berkeley.

Bartlett, P. L., S. Boucheron and G. Lugosi (2002). Model selection and error estimation. *Machine Learning* **48**, 85–113.

Bartlett, P., M. I. Jordan and J. D. McAuliffe (2003*b*). Convexity, classification, and risk bounds. Technical Report 638. Department of Statistics, University of California at Berkeley.

Boser, B. E., I. M. Guyon and V. N. Vapnik (1992). A training algorithm for optimal margin classifiers. In: *Proceedings of the 5th Annual Workshop on Computational Learning Theory*. ACM Press, New York, NY. pp. 144–152.

Bousquet, O. (2002). A Bennett concentration inequality and its application to suprema of empirical processes. *C. R. Acad. Sci. Paris Ser. I* **334**, 495–500.

Cortes, C. and V.N. Vapnik (1995). Support-vector networks. *Machine Learning* **20**, 273–297.

Cristianini, Nello and John Shawe-Taylor (2000). *An Introduction to Support Vector Methods*. Cambridge University Press.

Drucker, Harris and Corinna Cortes (1996). Boosting decision trees. In: *Advances in Neural Information Processing Systems*. Vol. 8. pp. 479–485.

Dudley, R. M. (1999). *Uniform Central Limit Theorems*. Cambridge University Press.

Freund, Y. and R.E. Schapire (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences* **55**(1), 119–139.

Freund, Yoav and Robert E. Schapire (1996). Experiments with a new boosting algorithm. In: *Proceedings of the 13th International Conference on Machine Learning*. Morgan Kaufmann. pp. 148–156.

Friedman, J., T. Hastie and R. Tibshirani (2000). Additive logistic regression: a statistical view of boosting. *Annals of Statistics* **28**, 337–374.

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. In: *Proceedings of the European Conference on Machine Learning*. Springer. pp. 137–142.

Jones, Lee K. (1997). The computational intractability of training sigmoidal neural networks. *IEEE Transactions on Information Theory* **43**(1), 167–173.

Karandikar, Rajeeva L. and M. Vidyasagar (2002). Rates of uniform convergence of empirical means with mixing processes. *Statistics and Probability Letters* **58**(3), 297–307.

Kimeldorf, George S. and Grace Wahba (1971). Some results on Tchebycheffian spline functions. *Journal of Mathematical Analysis and Applications* **33**(1), 82–95.

Koltchinskii, V. and D. Panchenko (2002). Empirical margin distributions and bounding the generalization error of combined classifiers. *Annals of Statistics* **30**(1), 1–50.

Ledoux, M. and M. Talagrand (1991). *Probability in Banach Spaces: Isoperimetry and Processes*. Springer.

Lee, W. S., P. L. Bartlett and R. C. Williamson (1996). Efficient agnostic learning of neural networks with bounded fan-in. *IEEE Transactions on Information Theory* **42**(6), 2118–2132.

Lee, Wee Sun, Peter L. Bartlett and Robert C. Williamson (1998). The importance of convexity in learning with squared loss. *IEEE Transactions on Information Theory* **44**(5), 1974–1980.

Mason, L., J. Baxter, P.L. Bartlett and M. Frean (2000). Functional gradient techniques for combining hypotheses. In: *Advances in Large Margin Classifiers* (A.J. Smola, P.L. Bartlett, B. Schölkopf and D. Schuurmans, Eds.). pp. 221–247. MIT Press. Cambridge, MA.

Massart, P. (2000). Some applications of concentration inequalities to statistics. *Annales de la Faculté des Sciences de Toulouse* **IX**, 245–303.

McDiarmid, C. (1989). On the method of bounded differences. In: *Surveys in Combinatorics 1989*. pp. 148–188. Cambridge University Press.

Mendelson, S. (2002*a*). Geometric parameters of kernel machines. In: *Proceedings of the 15th Annual Conference on Computational Learning Theory*. pp. 29–43.

Mendelson, S. (2002*b*). Improving the sample complexity using global data. *IEEE Transactions on Information Theory* **48**(7), 1977–1991.

Nobel, A. B. and A. Dembo (1993). A note on uniform laws of averages for dependent processes. *Statistics and Probability Letters* **17**, 169–172.

Rätsch, G., S. Mika and M. K. Warmuth (2001). On the convergence of leveraging. Technical Report NC-TR-01-098. NeuroCOLT2, Royal Holloway College.

Romdhani, S., B. Schölkopf, P. Torr and A. Blake (2001). Computationally efficient face detection. In: *Proceedings of the International Conference on Computer Vision*. pp. 695–700.

Schapire, Robert E. (1990). The strength of weak learnability. *Machine Learning* **5**(2), 197–227.

Schölkopf, B. and A. Smola (2002). *Learning with Kernels*. MIT Press. Cambridge, MA.

Suykens, J. A. K., T. Van Gestel, J. De Brabanter, B. De Moor and J. Vandewalle (2002). *Least Squares Support Vector Machines*. World Scientific.

Tsuda, K., T. Kin and K. Asai (2002). Marginalized kernels for biological sequences. *Bioinformatics* **18(Suppl. 1)**, S268–S275.

Vapnik, V. N. (1998). *Statistical Learning Theory*. Wiley and Sons. New York.

Vidyasagar, M. (1997). *A Theory of Learning and Generalization, with Applications to Neural Networks and Control Systems*. Springer. New York.

Vidyasagar, M. and R. L. Karandikar (2001). A learning theory approach to system identification and stochastic adaptive control. In: *IFAC Symposium on Adaptation and Learning*.

Watkins, C. (2000). Dynamic alignment kernels. In: *Advances in Large Margin Classifiers* (A.J. Smola, P.L. Bartlett, B. Schölkopf and D. Schuurmans, Eds.). pp. 39–50. MIT Press. Cambridge, MA.

Weyer, E. (2000). Finite sample properties of system identification of ARX models under mixing conditions. *Automatica* **36**(9), 1291–1299.

Yu, B. (1994). Rates of convergence for empirical processes of stationary mixing sequences. *Annals of Probability* **22**, 94–116.

Zhang, T. (2003). Sequential greedy approximation for certain convex optimization problems. *IEEE Transactions on Information Theory* **49**(3), 682–691.

Zhang, T. and B. Yu (2003). Boosting with early stopping: convergence and consistency. Technical Report 635. Department of Statistics, UC Berkeley.