Local algorithms for pattern recognition and dependencies estimation

V. Vapnik, L. Bottou

AT&T Bell Laboratories, Holmdel, NJ 07733

Abstract

In a previous publications (Bottou & Vapnik, 92, Vapnik, 92), we have described local learning algorithms, which result on performance improvements for real problems. We present here the theoretical framework on which these algorithms are based.

First, we present a new statement of certain learning problems, namely the Local Risk Minimization. We review the basic results of the uniform convergence theory of learning, and extend these results to local risk minimization.

We also extend the Structural Risk Minimization principle for both pattern recognition problems and regression problems. This extended induction principle is the basis for a new class of algorithms.

1 Introduction.

The concept of learning is wide enough to encompass several mathematical statements. The notions of *risk minimization* and of *loss function* (cf. Vapnik, 82), for instance, have unified several problems, such as pattern recognition, regression, and density estimation.

The classical analysis of learning deals with the modelization of an hypothetical truth, given a set of independent examples. In the classical statement of the pattern recognition problem, for instance, we select a classification function given a training set. This selection aims at keeping small the number of misclassifications observed when the selected function is applied to new patterns extracted from the same distribution as the training patterns.

In this statement, the underlying distribution of examples plays the role of an hypothetical truth, and the selected function models a large part of this truth, i.e. the dependence of the class on the input data.

We introduce in this paper a different statement of the learning problem. In this *local statement*, the selection of the classification function aims at reducing the probability of misclassification for a *given* test pattern. This process, of course, can be repeated for any particular test pattern. This is quite a different task: Instead of estimating a function, we estimate the value of a function on a given point (or in the vicinity of a given point).

The difference between these two statements can be illustrated by a practical example. A multilayer network illustrates the classical approach: The training procedure builds a model using a training set. This model then is used for all testing patterns. On the other hand, the nearest neighbor method is the simplest local algorithm: Given a testing point, we estimate its class by searching the closest pattern in the training set. This process must be repeated for each particular test pattern.

The statement of the problem defines the goal of the learning procedure. This goal is evaluated *a posteriori* by the performance of the system on some test data. The training data however does not even provide enough information to define this goal unambiguously. We must then rely on an *induction principle*, i.e. a heuristic method for "guessing" a general truth on the basis of a limited number of examples. Any learning algorithm assumes explicitly or implicitly some induction principle, which determines the elementary properties of the algorithm.

The simplest induction principle, namely the *principle of empirical risk minimization* (ERM) is also the most commonly used. According to this principle, we should choose the function which minimizes the number of errors on the training set.

The theory of empirical risk minimization was developed (Vapnik,82) in the 70's. In the case of pattern recognition, this theory provides a bound on the probability of errors, p, when the classification function is chosen among a set of functions of finite VC dimension. In the simplest case, with probability $1-\eta$ the following inequality is true (Vapnik,82, page 156, Theorem 6.7).

$$p \leq \nu + D, \tag{1}$$

where ν is the frequency of error on the training set, and

$$D = 2\sqrt{\frac{h(\ln\frac{2l}{h}+1)-\ln\frac{\eta}{9}}{l}}$$

is a confidence interval, which depends on the number of training examples l, on the VC-dimension h of the set of functions, and on the confidence η.

When the VC-dimension of the set of functions increases, the frequency of error on the training set decreases, but the width D of the confidence interval increases. This behavior leads to a new induction principle, namely the *principle of structural risk minimization* (SRM).

Consider a collection of subsets imbedded in the set of functions,

$$S_1 \subset S_2 \subset \cdots \subset S_n$$

where S_k is a subset of functions with VC-dimension h_k, and $h_k < h_{k+1}$.

For each subset S_k, a function f_k minimizes the frequency of error on the training set, and thus fulfils the inequality (1). Successive functions f_k yield a decreasing number of errors ν on the training set, but have increasingly wider confidence interval D. The principle of structural risk minimization consists in selecting the subset S_{k^*} and the function f_{k^*} which minimizes the right hand side of inequality (1), named the *guaranteed risk.*

The SRM principle requires the choice of a nested structure on the set of functions. An adequate structure can significantly improve the generalisation performance; a poor structure may have a limited negative impact.

In the local statement of the learning problem, we aim at selecting a valid function in the vicinity of a given test point x_0. On the basis of the training set, we will select a "width" for this vicinity, as well as a function for classification vectors in this vicinity.

To solve this problem, an extended SRM principle will be considered. We will minimize the guaranteed risk not only by selecting a subset S_{k^*} and a function $f_{k^*} \in S_{k^*}$, but also by selecting the width β of the vicinity of the point x_0. Using β as an additional parameter allows us to find a deeper minimum of the guaranteed risk as demonstrated on a practical application in (Bottou & Vapnik,92).

The paper is organized as follows: First, we state and discuss the problem of risk minimization and the problem of local risk minimization. In section 3, we derive a few useful bounds for uniform convergence of averages to their expectations. In sections 4 and 5 we derive bounds of the "local risk" for the problem for pattern recognition and the problem of regression. In section 6, we extend the structural risk minimization principle to local algorithms.

2 Global and Local Risk Minimization.

Like many illustrous scientists, we will assume, in this paper, that a metaphysical truth rules both our training examples and our testing cases. Like many illustrous statisticians, we will also assume that this truth can be represented, for our purposes, by an unknown probability distribution $F(x,y)$, defined on a space of input-output pairs $(x,y) \in R^n \times R^1$.

In the classical statement of *global risk minimization*, a parameter $\alpha \in \Lambda$ defines a model $x \longrightarrow f(x,\alpha)$ of the output y. A loss function, $Q(y, f(x,\alpha))$, measurable with respect to $dF(x,y)$, quantifies the quality of the estimate $f(x,\alpha)$ for the outputs y. We wish then to minimize the global risk functional

$$R(\alpha) = \int Q(y, f(x,\alpha))dF(x,y) \tag{2}$$

over all functions $\{f(x,\alpha), \alpha \in \Lambda\}$, when the distribution $F(x,y)$ is unknown, but when a random independent sample of size l

$$x_1, y_1; \cdots ; x_l, y_l \tag{3}$$

is given.

Let us introduce the statement of *local risk minimization*, in the vicinity of a given point x_0. In this statement, we aim at modeling the truth in a small neighborhood around x_0.

A nonnegative function $K(x, x_0, \beta)$ embodies the notion of vicinity. This function depends on the point x_0, on a "locality" parameter $\beta \in [0, \infty[$, and satisfies:

$$\begin{array}{ll} i) & 0 \leq K(x, x_0, \beta) \leq 1, \\ ii) & K(x_0, x_0, \beta) = 1, \end{array}$$

For example, both the "hard threshold" locality function (4) and the "normal" locality function (5) meet these conditions.

$$K_1(x, x_0, \beta) \;\; = \;\; \begin{cases} 1 & \text{if } |x - x_0| \leq \beta/2 \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

$$K_2(x, x_0, \beta) \;=\; \exp\left\{-\frac{(x-x_0)^2}{\beta}\right\} \tag{5}$$

Let us define the norm of locality function as

$$||K(x_0, \beta)|| = \int K(x, x_0, \beta) dP(x, y)$$

Let us consider again a parametric function $f(x, \alpha)$, and a measurable loss function, $Q(y, f(x, \alpha))$. We want to minimize the local risk functional

$$R(\alpha, \beta, x_0) = \int Q(y, f(x, \alpha)) \frac{K(x, x_0, \beta)}{||K(x_0, \beta)||} dF(x, y) \tag{6}$$

over the parameters α and β, when the distribution function $F(x, y)$ is unknown, but when a random independent sample $x_1, y_1, \ldots, x_l, y_l$ is given.

In most cases, the knowledge of the distribution $F(x, y)$ would make this problem trivial. For example, if the locality function is either the hard threshold locality function (4), or the normal locality function (5), we would select $\beta = 0$, and adjust α to get the "right" value for $f(x_0, \alpha)$. The true distribution, $F(x, y)$, however, is unknown. Selecting a non trivial value for the locality parameter β might reduce the generalization error induced by the unavoidable inaccuracy of parameter α. A new induction principle has been developed to take advantage of this fact. It is described in section 6.

Let us apply the statement of local risk minimization to the problems of pattern recognition and regression.

In the case of pattern recognition, the outputs y take only two values 0 or 1 and $\{f(x, \alpha),\ \alpha \in \Lambda\}$ is a set of indicator functions. The simplest loss function

$$Q(y, f(x, \alpha)) = \begin{cases} 0 & \text{if } y = f(x, \alpha) \\ 1 & \text{if } y \neq f(x, \alpha) \end{cases}$$

merely indicates the presence or absence of classification error. The risk functional (2) then measures the probability of classification error for the function $f(x, \alpha)$. The global pattern recognition problem consists in selecting, on the basis of the training set, a function $f(x, \alpha^*)$ which guarantees a small probability of classification error.

Now, let us consider the local risk functional (6), using the hard threshold locality function (4). This functional measures the conditional probability of classification error knowing $||x - x_0|| \leq \beta^*/2$.

The local pattern recognition problem consists in selecting, on the basis of the training set, a value for the locality parameter β^* and a function $f(x, \alpha^*)$ which guarantee a small probability of classification error in $||x - x_0|| \leq \frac{\beta^*}{2}$.

In the case of the regression problem, the outputs y are real values, and $\{f(x, \alpha), \alpha \in \Lambda\}$ is a set of real functions. We will consider a quadratic loss function,

$$Q(y, f(x, \alpha)) = (y - f(x, \alpha))^2 \tag{7}$$

The minimum of the global risk functional (2) is achieved by the closest function of the class $\{f(x, \alpha), \alpha \in \Lambda\}$ to the regression function

$$r(x) = E(y \mid x) = \int y \; dF(y \mid x) \tag{8}$$

using a quadratic metric

$$\rho^2(f_1, f_2) = \int (f_1(x) - f_2(x))^2 dF(x)$$

The minimum of the local risk functional (6), using the locality function (4) is achieved by the closest function of the class $\{f(x, \alpha), \alpha \in \Lambda\}$ to the regression function, $r(x)$ using a metric

$$\rho^2_{x_0}(f_1, f_2) = \int (f_1(x) - f_2(x))^2 \ dF(x\ , x_0, \beta)), \tag{9}$$

where

$$dF(x, x_0, \beta) = \frac{K(x, x_0, \beta)}{||K(x, x_0, \beta)||} dF(x).$$

The local regression problem consists, on the basis of the training set, in selecting a value for the locality parameter β^* and a function $f(x, \alpha^*)$ which guarantees a small conditional quadratic error.

3 Theory of Uniform Convergence.

For simplicity, we will refer to the pairs (x, y) as a vector z, and we will denote the loss function $Q(y, f(x, \alpha))$ as $Q(z, \alpha)$. The notation $F(z)$ denotes a probability distribution on the pairs (x, y).

First, we will review uniform convergence results for the global risk functional

$$R(\alpha) = \int Q(z, \alpha) dF(z) \tag{10}$$

These results are then extended to the the local risk functional, using a transformation of the probability distribution $F(z)$. The global risk can be made local by "oversampling" the probability distribution around the point x_0.

We already have stressed the fact that optimizing (10) is generally impossible, unless we know $F(z)$ exactly. If our knowledge of $F(z)$ is limited to a random independent sample

$$z_1, \cdots, z_l, \tag{11}$$

we must rely on an induction principle, like the empirical risk minimization (ERM), or the structural risk minimization (SRM). A good induction principle should provide a way to select a value α_l^* which guarantees a small value for the risk $R(\alpha_l^*)$. More precisely, two questions should be answered:

1. When is the method of empirical risk minimization consistent? In other words, does the generalization risk $R(\alpha_l^*)$ converge to the minimum of the risk functional $R(\alpha)$ when the size l of the training set increases?

2. How fast is this convergence? In general, the number of training examples is limited, and the answer to this question is of crucial practical importance.

Many induction principles, including SRM and ERM, rely on the empirical risk functional

$$E_l(\alpha) = \frac{1}{l} \sum_{i=1}^{l} Q(z_i, \alpha) \tag{12}$$

which estimates the risk $R(\alpha)$ using the training set (11). In these cases, the answers to the two questions stated above depend on the quality of the estimation (12). More precisely,

1. Does the empirical risk functional $E_l(\alpha)$ converge to the risk functional $R(\alpha)$ when the size of the training set increases, *uniformly* over the set of functions $\{Q(z, \alpha), \alpha \in \Lambda\}$? The uniform convergence takes place if

$$\forall \epsilon > 0 \quad \lim_{l \to \infty} \quad P\{\sup_{\alpha \in \Lambda} \ |R(\alpha) - E_l(\alpha)| \ > \ \epsilon\} = 0$$

2. What is the rate of this convergence?

The theory of uniform convergence of empirical risk to actual risk developed in the 70's and 80's (cf. Vapnik, 82), contains a necessary and sufficient condition for uniform convergence, and provides bounds on the rate of uniform convergence. These bounds do not depend on the distribution function $F(z)$; they are based on a measure of the capacity (VC-dimension) of the set of functions $\{Q(z, \alpha)), \ \alpha \in \Lambda\}$.

Definition 1. The VC-dimension of the set of indicator functions $\{Q(z, \alpha), \alpha \in \Lambda\}$ is the maximum number h of vectors $z_1, ..., z_h$ that functions of the set $\{Q(z, \alpha), \ \alpha \in \Lambda\}$ can separate into two classes in all 2^h possible ways.

Definition 2. The VC-dimension of the set of real functions $\{Q(z, \alpha), \ \alpha \in \Lambda\}$ is the defined as the VC-dimension of the following set of indicator functions

$$\Psi(z, \alpha, \delta) = \theta(Q(z, \alpha) + \delta), \quad \alpha \in \Lambda, \quad \delta \in (-\infty, \infty)$$

where

$$\theta(u) = \begin{cases} 1 & \text{if } u > 0 \\ 0 & \text{otherwise.} \end{cases}$$

Two theorems are valid for a set of indicator loss functions. We assume here that the loss functions $Q(z, \alpha), \ \alpha \in \Lambda$ are set indicator functions defined in z-space.

Theorem 1. Let the set of indicator functions $\{Q(z, \alpha), \ \alpha \in \Lambda\}$ have VC-dimension h. Then the following inequality is true:

$$P\{\sup_{\alpha \in \Lambda} |R(\alpha) - E_l(\alpha)| > \epsilon\} \ < \ 9 \left(\frac{2le}{h}\right)^h \exp\left\{-\frac{\epsilon^2 l}{4}\right\} \tag{13}$$

This theorem is proven in (Vapnik,82, page 170, Theorem A.2). The quantity $(2l)^h/h!$ has been bounded by the more convenient quantity $(2le/h)^h$, using Stirling's formula.

Bound (13), however, is limited by the behavior of the absolute difference between the risk and the empirical risk when the risk $R(\alpha)$ is close to 1/2. Theorem 2 provides a bound on the "relative" difference between the risk and the empirical risk.

Theorem 2. Let the set of indicator functions $\{Q(z,\alpha),\ \alpha \in \Lambda\}$ have VC-dimension h. Then the following inequality is true:

$$P\{\sup_{\alpha\in\Lambda} \frac{R(\alpha)-E_l(\alpha)}{\sqrt{R(\alpha)}} > \epsilon\} \ < \ 12\left(\frac{2le}{h}\right)^h \exp\{-\frac{\epsilon^2 l}{4}\} \tag{14}$$

This theorem is proven in (Vapnik,82, page 176, Theorem A.3.). Again, the quantity $(2l)^h/h!$ has been bounded by the more convenient quantity $(2le/h)^h$, using Stirling's formula.

Both Theorem 1 and 2 can be generalized to the case of uniformly bounded loss functions. We assume now that the loss functions $Q(z,\alpha)$ are nonnegative, and satisfy the condition

$$0 \le Q(z,\alpha) \le B,\ \alpha \in \Lambda \tag{15}$$

Theorem 3. Let the uniformly bounded set of real functions (15) have VC-dimension h. Then the following bound is true:

$$P\{\sup_{\alpha\in\Lambda} |R(\alpha)-E_l(\alpha)| > \epsilon\} \ < \ 9\left(\frac{2le}{h}\right)^h \exp\{-\frac{\epsilon^2 l}{4B^2}\} \tag{16}$$

This theorem (16) is proved in Appendix 1.

Theorem 4. Let a uniformly bounded set of functions (15) have VC-dimension h. Then the following bound is valid:

$$P\{\sup_{\alpha\in\Lambda} \frac{R(\alpha)-E_l(\alpha)}{\sqrt{R(\alpha)}} > \epsilon\} \ < \ 12\left(\frac{2le}{h}\right)^h \exp\{-\frac{\epsilon^2 l}{4B}\} \tag{17}$$

This theorem (17) is proved in Appendix 2.

Finally, we need a bound on the rate of uniform convergence for a set of unbounded real functions $\{Q(z,\alpha),\ \alpha \in \Lambda\}$. Such a bound requires some restriction on the large deviations of the set of loss functions. This is also true for the classical bounds. Although the law of large numbers says that the average of random values converges to their mathematical expectation, the rate of convergence could be slow.

Next example shows that even in the case when the set of functions contains only one function it is impossible to bound the rate of convergence without additional information.

Consider a random variable ζ which takes two values: 0 with probability $1-\epsilon$ and $1/\epsilon^2$ with probability ϵ. The expectation of this random variable is

$$E(\zeta) = (1-\epsilon)0 + \frac{\epsilon}{\epsilon^2} = \frac{1}{\epsilon}$$

The empirical average is null if all l observations are 0. The probability of this event is

$$P(0) = (1-\epsilon)^l$$

For a small ϵ, the expectation $E(\zeta)$ is large, but the empirical average is null with a high probability.

In Theorems 1 to 4, we have assumed a uniform bound (1 or B) on the losses $Q(z,\alpha)$. This bound forbids large deviations.

We consider now the case of nonnegative, unbounded losses, which satisfy the following mild restriction:

$$\sup_{\alpha\in\Lambda} \frac{\sqrt{\int Q^2(z,\alpha)dF(z)}}{\int Q(z,\alpha)dF(z)} \ < \ \tau. \tag{18}$$

This condition reflects a restriction on the "tails" of the distribution of the losses $Q(z,\alpha)$. For instance, let $Q(z,\alpha)$ be a quadratic loss $(y-f(x,\alpha))^2$. If the random variable

$$\zeta_\alpha = y - f(x,\alpha)$$

is distributed according to normal law, the ratio of moments in condition (18) is equal to $\sqrt{3}$ (independ of values of parameters). If the random variable ζ_α is distributed according to Laplace law, this ratio is equal to $\sqrt{2}$ (also independ of value of parameters).

The following result has been proved in (Vapnik,82, page 202, 3rd equation).

Theorem 5. Let $\{Q(z,\alpha),\ \alpha\in\Lambda\}$ be a set of nonnegative real functions with VC-dimension h. Then the following bound is true:

$$P\{\sup_{\alpha\in\Lambda} \frac{R(\alpha)-E_l(\alpha)}{\sqrt{\int Q^2(z,\alpha)dF(z)}} > \epsilon a(\epsilon)\} \ < \ 12\left(\frac{2le}{h}\right)^h \exp\{-\frac{\epsilon^2 l}{4}\} \tag{19}$$

where

$$a(\epsilon) = \sqrt{1-\frac{1}{2}\ln\epsilon}$$

In this formulation again, $(2l)^h/h!$ has been bounded by $(2le/h)^h$. We obtain a uniform bound for the relative difference between the risk and the empirical risk by applying condition (18) to this result.

$$P\{\sup_{\alpha\in\Lambda} \frac{R(\alpha)-E_l(\alpha)}{R(\alpha)} > \tau\epsilon a(\epsilon)\} \ < \ 12\left(\frac{2le}{h}\right)^h \exp\{-\frac{\epsilon^2 l}{4}.\} \tag{20}$$

Let us extend this inequality to the case of local algorithms. First, for any fixed value of α and β, note that the local risk functional

$$R(\alpha,\beta,x_0) \ = \ \int Q(z,\alpha)\frac{K(x,x_0,\beta)}{||K(x_0,\beta)||}dF(z) \ = \ \int Q(z,\alpha)dF(z,\beta)$$

is equal to the expectation of the loss function $Q(z,\alpha)$ with respect to a new distribution function $F(z,\beta)$ defined by

$$\forall\Omega\subset R^{n+1}, \quad \int_\Omega dF(z,\beta) \ = \ \int_\Omega \frac{K(x,x_0,\beta)}{||K(x_0,\beta)||}dF(z).$$

We will consider the set of functions $Q(z,\alpha),\ \alpha \in \Lambda$ and the set of probability distribution functions $F(z,\beta),\ \beta \in [0,\infty]$ which satisfy the following inequality .

$$\sup_{\alpha,\beta} \frac{\sqrt{\int Q^2(z,\alpha)dF(z,\beta)}}{\int Q(z,\alpha)dF(z,\beta)} < \tau. \tag{21}$$

Let us define the unnormalized local risk, $\mathcal{R}(\alpha,\beta,x_0)$, and the unnormalized empirical local risk, $\mathcal{E}_l(\alpha,\beta,x_0)$,as follows:

$$\begin{aligned} \mathcal{R}(\alpha,\beta,x_0) &= \int Q(z,\alpha)K(x,x_0,\beta)dF(z), \\ \mathcal{E}_l(\alpha,\beta,x_0) &= \frac{1}{l}\sum_{i=1}^{l} Q(z_i,\alpha)K(x_i,x_0,\beta). \end{aligned}$$

We will show that under condition (21) the following inequality is true

$$P\{\sup_{\alpha\in\Lambda} \frac{\mathcal{R}(\alpha,\beta,x_0)-\mathcal{E}_l(\alpha,\beta,x_0)}{\mathcal{R}(\alpha,\beta,x_0)} > \frac{\tau\epsilon a(\epsilon)}{\sqrt{||K(x_0,\beta)||}}\} \ < \ 12\left(\frac{2le}{h_1}\right)^{h_1}\exp\{-\frac{\epsilon^2 l}{4}\} \tag{22}$$

where h_1 is VC-dimension of the set of functions

$$\{Q(z,\alpha)K(x,x_0,\beta), \quad \alpha\in\Lambda,\ \beta\in[0,\infty[\}$$

To prove this inequality, we note that theorem 5 implies the following inequality.

$$P\{\sup_{\alpha\in\Lambda} \frac{\mathcal{R}(\alpha,\beta,x_0)-\mathcal{E}_l(\alpha,\beta,x_0)}{\sqrt{\int Q^2(z,\alpha)K^2(z,x_0,\beta)dF(z)}} > \epsilon a(\epsilon)\} \ < \ 12\left(\frac{2le}{h_1}\right)^{h_1}\exp\{-\frac{\epsilon^2 l}{4}\} \tag{23}$$

Moreover, since $0 \le K(x,x_0,\beta) \le 1$, we have

$$\sqrt{\int Q^2(z,\alpha)K^2(x,x_0,\beta)dF(z)} \le \sqrt{\int Q^2(z,\alpha)K(x,x_0,\beta)dF(z)} \tag{24}$$

$$\le \ \sqrt{\int Q^2(z,\alpha)||K(x_0,\beta)||dF(z,\beta)}$$

and according to (21), the following inequality is true for any $\beta \in [0,\infty[$.

$$\sqrt{\int Q^2(z,\alpha)dF(z,\beta)} \le \tau\int Q(z,\alpha)dF(z,\beta) = \tau\frac{\mathcal{R}(\alpha,\beta,x_0)}{||K(x_0,\beta||} \tag{25}$$

Inequality (22) is derived from inequalities (23), (24) and (25).

4 Bounds for the Local Risk in Pattern Recognition.

In this section, we apply the previous results to the problem of pattern recognition. Consider the set of integrands of the unnormalized local risk functional, $\mathcal{R}(\alpha, \beta, x_0)$:

$$\{ \;\; Q(z,\alpha)K(x,x_0,\beta), \quad \alpha \in \Lambda, \;\; \beta \in [0,\infty] \;\; \} \tag{26}$$

where $Q(z,\alpha)$ is an indicator function and $K(x,x_0,\beta)$ a nonnegative real function.

Let h_1 be the VC-dimension of the set of indicator loss functions $\{Q(z,\alpha), \;\; \alpha \in \Lambda\}$. Let h_2 be the VC-dimension of the set of nonnegative real functions $\{K(x,x_0,\beta), \;\; \beta \in [0,\infty[\}$. Since $Q(z,\alpha)$ takes only the values 0 or 1, the following equality is true for any nonnegative real function $r(z,\beta)$.

$$\theta\{Q(z,\alpha)r(z,\beta)+c\} = Q(z,\alpha)\theta\{r(z,\beta)+c\} \quad \alpha \in \Lambda, \;\; \beta \in [0,\infty[$$

Moreover, it is known that VC-dimension of the product of two sets of indicator functions does not exceed the sum of the VC-dimension of each set of indicator functions. Therefore, the definition of the VC-dimension of a set of real function implies that the VC-dimension of the set of functions

$$\{ \;\; Q(z,\alpha)K(x,x_0,\beta), \quad \alpha \in \Lambda, \;\; \beta \in [0,\infty[\;\; \}$$

does not exceed $h_1 + h_2$. Let us apply Theorem 4 to this set of functions.

$$P\{ \sup_{\alpha\in\Lambda, \;\; \beta\in[0,\infty[} \frac{\mathcal{R}(\alpha,\beta,x_0) - \mathcal{E}_l(\alpha,\beta,x_0)}{\sqrt{\mathcal{R}(\alpha,\beta,x_0)}} > \epsilon\} \; < \; 12 \left(\frac{2le}{h_1+h_2}\right)^{h_1+h_2} \exp\{-\frac{\epsilon^2 l}{4}\}.$$

Let $\eta/2$ denote the right hand side of this inequality. By solving the equation

$$12 \left(\frac{2le}{h_1+h_2}\right)^{h_1+h_2} \exp\{-\frac{\epsilon^2 l}{4}\} \; = \; \eta/2$$

and replacing the result into our inequality, we obtain an equivalent formulation: With probability $1-\eta/2$, the following inequality is true for all functions in $\{Q(z,\alpha), \;\; \alpha \in \Lambda, \;\; \beta \in [0,\infty[\}$.

$$\mathcal{R}(\alpha,\beta,x_0) \; \leq \; \mathcal{E}_l(\alpha,\beta,x_0) + \vartheta \left(1 + \sqrt{1 + \frac{4}{\vartheta}\mathcal{E}_l(\alpha,\beta,x_0)}\right) \tag{27}$$

where

$$\vartheta = \epsilon^2 = 2\frac{(h_1+h_2) \; (\ln\frac{2l}{h_1+h_2}+1) - \ln\frac{\eta}{24}}{l} \tag{28}$$

By dividing both sides of inequality (27) by $||K(x_0,\beta)||$, we obtain

$$R(\alpha,\beta,x_0) \; \leq \; \frac{1}{||K(x_0,\beta)||} \left(\mathcal{E}_l(\alpha,\beta,x_0) + \vartheta \left(1 + \sqrt{1 + \frac{4}{\vartheta}\mathcal{E}_l(\alpha,\beta,x_0)}\right)\right). \tag{29}$$

The value of $||K(x_0,\beta)||$ in the right hand side of inequality (29) depends on the distribution function $F(z)$. A lower bound for the value $||K(x_0,\beta)||$ is obtained by using the empirical functional:

$$\frac{1}{l}\sum_{i=1}^{l} K(x_i,x_0,\beta)$$

where $z_i = (x_i, y_i)$ are the elements of the training set (11). Applying Theorem 3 to the set of uniformly bounded functions $\{K(x,x_0,\beta),\ \beta \in [0,\infty[\}$ result in

$$P\{\sup_{\beta\in[0,\infty[}\left|\ ||K(x_0,\beta)|| - \frac{1}{l}\sum_{i=1}^{l} K(x_i,x_0,\beta)\ \right| > \epsilon\} \ < \ 9\left(\frac{2le}{h_2}\right)^{h_2} \exp\{-\frac{\epsilon^2 l}{4}.\}$$

In other words, the following inequality is simultaneously true for all $\beta \in [0,\infty[$, with probability $1-\eta/2$:

$$||K(x_0,\beta)|| > \left(\frac{1}{l}\sum_{i=1}^{l} K(x_i,x_0,\beta) - 2\sqrt{\frac{h_2(\ln\frac{2l}{h_2}+1) - \ln\frac{\eta}{18}}{l}}\right)_+ \qquad (30)$$

where $(u)_+ \ = \ \mathrm{Max}\{0,u\}$. Let us define $\mathcal{K}(x_0,\beta)$ as the right hand side of inequality (30).

$$\mathcal{K}(x_0,\beta) = \left(\frac{1}{l}\sum_{i=1}^{l} K(x_i,x_0,\beta) - 2\sqrt{\frac{h_2(\ln\frac{2l}{h_2}+1) - \ln\frac{\eta}{18}}{l}}\right)_+ \qquad (31)$$

By combining inequalities (29) and (30), we obtain the following theorem, which provides a bound for the local risk functional in the case of pattern recognition.

Theorem 6. Let the VC-dimension of the set of indicator functions $\{Q(z,\alpha),\ \alpha\in\Lambda\}$ be h_1. Let the VC-dimension of the set of real functions $\{K(x,x_0,\beta),\ \ \beta\in[0,\infty[\}$ be h_2. The following equality is simultaneousely fulfilled for all $\alpha\in\Lambda$ and $\beta\in[0,\infty[$, with probability $1-\eta$:

$$R(\alpha,\beta,x_0) \le \frac{1}{\mathcal{K}(x_0,\beta)}\left(\mathcal{E}_l(\alpha,\beta,x_0) + \vartheta\left(1+\sqrt{1+\frac{4}{\vartheta}\mathcal{E}_l(\alpha,\beta,x_0)}\right)\right) \qquad (32)$$

where

$$\vartheta = \frac{(h_1+h_2)(\ln\frac{2l}{h_1+h_2}+1) - \ln\frac{\eta}{24}}{l}$$

As expected, the VC-dimension h_1 and h_2 affect the quantity ϵ, which controls the second term of the sum. The VC-dimension h_2 of the set of locality functions $\{K(x,x_0,\beta),\ \beta\in[0,\infty[\}$, however, also affects the first term of the sum, which is the empirical estimate of the local risk functional.

Therefore, it seems extremely advisable to use *Monotonic Radial Basis Functions* for defining the vicinity of a point x_0. In fact, the VC-dimension of the set of Radial Basis Functions

$$\{K(x,x_0,\beta) = K_\beta(||x-x_0||)\}$$

where the $K_\beta(r)$ are the monotonically decreasing functions of r, is equal to 1.

5 Bounds of the Local Risk in Regression Estimation.

In this section we apply the results presented in section 3 to the problem of local regression estimation. The loss functions $Q(z,\alpha)$ are now real functions.

In the case of pattern recognition, the loss functions were indicator functions. In this case, we have proved that the VC-dimension of the set $\{Q(z,\alpha)K(x,x_0,\beta),\ \alpha\in\Lambda,\ \ \beta\in[0,\infty[\}$ does not exceed the sum of the VC-dimensions of the sets of functions $\{Q(z,\alpha),\ \alpha\in\Lambda\}$ and $\{K(x,x_0,\beta),\ \ \beta\in[0,\infty[\}$.

This is no longer true in the case of real loss functions. For example, let $\{Q(z,\alpha),\ \alpha\in\Lambda\}$ be the set of monotonically increasing functions, and $\{K(x,x_0,\beta),\ \beta\in[0,\infty[\}$ be the set of monotonically decreasing functions. Although the VC-dimension of both sets is 1, the VC-dimension of the product of these sets is infinite.

In order to apply the uniform convergence results, we will assume that the VC-dimension h_1 of the set of function $\{Q(z,\alpha)K(x,x_0,\beta),\ \alpha\in\Lambda,\ \ \beta\in[0,\infty[\}$ is finite. We also assume that the functions $Q(z,\alpha)$ are nonnegative, and satisfy condition (21).

From inequality (22) we derive the following inequality, which is simultaneously valid for all $\alpha\in\Lambda,\ \beta\in[0,\infty[$, with probability $1-\eta/2$.

$$R(\alpha,\beta,x_0)\leq\frac{\mathcal{E}_l(\alpha,\beta,x_0)}{||K(x_0,\beta)||\left(1-\tau\epsilon\sqrt{1-\frac{\ln\epsilon^2||K(x_0,\beta)||}{4}}\right)_+}\tag{33}$$

where

$$\epsilon=2\sqrt{\frac{h_1(\ln\frac{2l}{h_1}+1)-\ln\frac{\eta}{24}}{l\ ||K(x_0,\beta)||}}\tag{34}$$

In section 4, we have proved that inequality (31) is true. Using (31) and (33), we obtain the following result:

Theorem 7. Let the VC-dimension of the set of nonnegative real functions

$$\{Q(z,\alpha)K(x,x_0,\beta),\ \alpha\in\Lambda,\ \beta\in[0,\infty[\}$$

be h_1. Let the VC-dimension of the set of locality functions

$$\{K(x,x_0,\beta),\ \ \beta\in[0,\infty[\}$$

be h_2. The following inequality is simultaneously valid for all $\alpha\in\Lambda,\ \beta\in[0,\infty[$, with probability $1-\eta$,

$$R(\alpha,\beta,x_0)\leq\frac{\mathcal{E}_l(\alpha,\beta,x_0)}{\mathcal{K}(x_0,\beta)\left(1-\tau\epsilon\sqrt{1-\frac{\ln\epsilon^2\mathcal{K}(x_0,\beta)}{4}}\right)_+}\tag{35}$$

where

$$\epsilon=2\sqrt{\frac{h_1(\ln\frac{2l}{h_1}+1)-\ln\frac{\eta}{24}}{l\ \mathcal{K}(x_0,\beta)}}$$

and $\mathcal{K}(x_0,\beta)$ is defined in (31).

This result provides a bound on the local risk functional for the case of regression estimation.

6 Local Structural Risk Minimization.

We can now formulate the principle of local structural risk minimization, using the bounds provided by theorem 6 and 7. In this section, the Local Structural Risk Minimization (LSRM) principle is formulate for pattern recognition. The regresion case is essentially similar.

Let us consider a nested structure on the set of indicator functions $\{Q(z,\alpha), \quad \alpha \in \Lambda\}$,

$$S_1 \subset S_2 \subset, \cdots, \subset S_n = \{Q(z,\alpha), \quad \alpha \in \Lambda\}. \tag{36}$$

Let the VC-dimension of each subset S_p be h_p, with

$$h_1 < h_2 < \cdots < h_n.$$

We have proved, in section 4, that the VC-dimension of the set of functions

$$\{Q(z,\alpha)K(x,x_0,\beta), \ \alpha \in \Lambda_p, \ \beta \in [0,\infty[\}$$

is smaller than $h_p + h^*$, where h^* denotes the VC-dimension of the set of real functions $\{K(x,x_0,\beta), \beta \in [0,\infty[\}$.

According to theorem 6, the following inequality is simultaneously valid for all element S_p of the structure, with probability $1 - n\eta$:

$$R(\alpha,\beta,x_0) \le \frac{1}{\mathcal{K}(x_0,\beta)}\left(\mathcal{E}_l(\alpha,\beta,x_0) + \vartheta\left(1 + \sqrt{1 + \frac{4}{\vartheta}\mathcal{E}_l(\alpha,\beta,x_0)}\right)\right) \tag{37}$$

where

$$\begin{aligned}
\vartheta &= \frac{(h_p + h^*)(\ln \frac{2l}{h_p+h^*} + 1) - \ln \frac{\eta}{24}}{l}, \\
\mathcal{E}_l(\alpha,\beta,x_0) &= \frac{1}{l}\sum_{i=1}^{l} Q(z_i,\alpha)K(x_i,x_0,\beta), \\
\mathcal{K}(x_0,\beta) &= \left(\frac{1}{l}\sum_{i=1}^{l} K(x_i,x_0,\beta) - 2\sqrt{\frac{h^*(\ln \frac{2l}{h^*} + 1) - \ln \frac{\eta}{18}}{l}}\right)_+ .
\end{aligned}$$

Principle. The Local Structural Risk Minimization Principle consists in choosing the element of structure S_p and the parameters $\alpha \in \Lambda_p$ and $\beta \in [0,\infty[$ which minimize the guaranteed risk as defined by the right hand side of inequality (37).

The various constants in bound (37) are the result of technical properties of the bounding derivations. The "proven" value is irrelevant to practical problems. Therefore, it is advisable to design experiments to measure these constants, and to use these measured values instead of using the "proven" values.

Acknowedgements

We thank the members of the Neural Network research group at Bell Labs, Holmdel, for usefull discussions. S. Solla and C. Cortes provided help to render this article more clear.

Appendix 1: Proof of Theorem 3.

Using Lebesgue's sums, we can write:

$$\begin{aligned} K &= \sup_{\alpha\in\Lambda} |R(\alpha) - E_l(\alpha)| \\ &= \sup_{\alpha\in\Lambda} \left| \int Q(z,\alpha) dP(z,\alpha) - \frac{1}{l}\sum_{i=1}^{l} Q(z_i,\alpha) \right| \\ &= \lim_{N\to\infty} \frac{B}{N} \sup_{\alpha\in\Lambda} \left| \sum_{n=o}^{N-1} P\{Q(z,\alpha) > \frac{Bn}{N}\} - \sum_{n=o}^{N-1} v\{Q(z,\alpha) > \frac{Bn}{N}\} \right| \end{aligned}$$

where $v\{Q(z,\alpha) > \frac{Bn}{N}\}$ denotes the frequency of the event $\{Q(z,\alpha) > \frac{Bn}{N}\}$, obtained on the basis of the sample $z_1, \cdots, z_l$. Then

$$\begin{aligned} K &\leq \lim_{N\to\infty} \frac{B}{N} \sup_{\alpha\in\Lambda} \sum_{n=o}^{N-1} \left| P\{Q(z,\alpha) > \frac{Bn}{N}\} - v\{Q(z,\alpha) > \frac{Bn}{N}\} \right| \\ &\leq B \sup_{\alpha\in\Lambda,\ \beta\in[0,B]} |P\{Q(z,\alpha) > \beta\} - v\{Q(z,\alpha) > \beta\}| \\ &= B \sup_{\alpha\in\Lambda,\ \beta\in[0,B]} \left| \int \theta\{Q(z,\alpha) - \beta\} dF(z) - \sum_{i=1}^{l} \theta\{Q(z_i,\alpha) - \beta\} \right| \end{aligned}$$

Using Theorem 1 and this inequality, we obtain

$$\begin{aligned} & P\{\sup_{\alpha\in\Lambda} \left| \int Q(z,\alpha) dF(z) - \frac{1}{l}\sum_{i=1}^{l} Q(z_i,\alpha) \right| > \epsilon\} \\ \leq\ & P\{ \sup_{\alpha\in\Lambda,\ \beta\in[0,B]} \left| \int \theta\{Q(z,\alpha) - \beta\} dF(z) - \sum_{i=1}^{l} \theta\{Q(z_i,\alpha) - \beta\} \right| > \frac{\epsilon}{B}\} \\ <\ & 9 \left(\frac{2le}{h}\right)^h \exp\{-\frac{\epsilon^2 l}{4}\}, \end{aligned}$$

where h is the VC-dimension of the set of indicator functions

$$\{\ \theta(Q(z,\alpha) - \beta),\ \alpha \in \Lambda, \beta \in [0,B]\ \}$$

According to definition 2, this quantity is the VC-dimension of the set of real loss functions $\{Q(z,\alpha),\ \alpha \in \Lambda\}$. Theorem 3 is thus proven.

Appendix 2: Proof of Theorem 4.

Again, consider a set real functions $\{Q(z,\alpha), \alpha \in \Lambda\}$ of VC-dimension h, and assume $0 < Q(z,\alpha) < B$. The following result is proven in (Vapnik,82, page 197, Lemma).

$$P\{\sup_{\alpha\in\Lambda} \frac{R(\alpha) - E_l(\alpha)}{\int_0^B \sqrt{P\{Q(z,\alpha) > \gamma\}} d\gamma} > \epsilon\} \ <\ 12 \frac{(2l)^h}{h!} \exp\{-\frac{\epsilon^2 l}{4}\}. \tag{38}$$

Using Cauchy inequality, we can write

$$\int_0^B 1\,\sqrt{P\{Q(z,\alpha)>\gamma\}}\;d\gamma \;\leq\; \sqrt{\int_0^B 1^2 d\gamma\,\int_0^B P\{Q(z,\alpha)>\gamma\}d\gamma} \;=\; \sqrt{BR(\alpha)}.$$

We replace this result in inequality (46); we bound $(2l)^h/h!$ by the more convenient expression $(2le/h)^h$; and obtain:

$$P\{\sup_{\alpha\in\Lambda}\frac{R(\alpha)-E_l(\alpha)}{\sqrt{\int Q(z,\alpha)dF(z)}} > \epsilon\sqrt{B}\} < 12\left(\frac{2le}{h}\right)^h \exp\{-\frac{\epsilon^2 l}{4}\}$$

Theorem 4 is thus proven.

References.

- [Bottou & Vapnik, 92] L. Bottou, V. Vapnik, *Local Learning Algorithm*, Neural Computation, *In press.*
- [Vapnik, 82] V. Vapnik. *Estimation of Dependencies Based on Empirical Data*, Springer-Verlag, N.Y., 1982.
- [Vapnik, 92] V. Vapnik. *Principles of Risk Minimization for Learning Theory* In David S Touretszky,editor, *Neural Information Proceeding Sistem*, voluem 4. Morgan Kaufmann Publishers, San Mateo, CA, 1992. *In press.*