Found. Comput. Math. (2002) 2:413–428

DOI: 10.1007/s102080010030

FOUNDATIONS OF COMPUTATIONAL MATHEMATICS
The Journal of the Society for the Foundations of Computational Mathematics

Best Choices for Regularization Parameters in Learning Theory: On the Bias–Variance Problem

Felipe Cucker[1] and Steve Smale[2]

[1]Department of Mathematics
City University of Hong Kong
83 Tat Chee Avenue
Kowloon, Hong Kong
macucker@math.cityu.edu.hk

[2]Department of Mathematics
University of California
Berkeley, CA, USA
smale@math.berkeley.edu

1. Introduction

The goal of learning theory (and a goal in some other contexts as well) is to find an approximation of a function $f_\rho : X \to Y$ known only through a set of pairs $\mathbf{z} = (x_i, y_i)_{i=1}^m$ drawn from an unknown probability measure ρ on $X \times Y$ (f_ρ is the "regression function" of ρ).

An approach championed by Poggio (see, e.g., [**5**]) with ideas going back to Ivanov [**7**] and Tikhonov [**13**] is to minimize

$$\frac{1}{m}\sum_{i=1}^{m}(f(x_i) - y_i)^2 + \gamma \|Af\|^2_{\mathcal{L}^2_\rho(X)},$$

where A is an operator and $\mathcal{L}^2_\rho(X)$ is the Hilbert space of square integrable functions on X with measure ρ_X on X defined via ρ. See [**4**] (in the sequel denoted

Date received: January 30, 2002. Final version received: April 20, 2002. Communicated by Michael Shub. Online publication: May 29, 2002.
AMS classification: 68T05, 62J02.

by [**CS**]*) for background on this and, even more importantly, for results used here.

This minimization is well-conditioned and solved by straightforward finite-dimensional least squares linear algebra (see Theorem 1 below) to yield $f_{\gamma,\mathbf{z}} : X \to Y$. The problem is posed: *How good an approximation is $f_{\gamma,\mathbf{z}}$ to f_ρ, or measure the error $\int_X (f_{\gamma,\mathbf{z}} - f_\rho)^2$?* and *What is the best choice of γ to minimize this error*?

Our goal in this paper is to give some answers to these questions.

Main Result. *We exhibit, for each $m \in \mathbb{N}$ and $\delta \in [0, 1)$, a function*

$$E_{m,\delta} = E : \mathbb{R}^+ \to \mathbb{R}$$

such that, for all $\gamma > 0$,

$$\int_X (f_{\gamma,\mathbf{z}} - f_\rho)^2 \leq E(\gamma)$$

with confidence $1 - \delta$. There is a unique minimizer of $E(\gamma)$ which is found by an easy algorithm to yield the "best" regularization parameter $\gamma = \gamma^*$.

The bound $E(\gamma)$ found is a natural one, a bound which flows from the hypotheses and thus yields a γ^* which could be useful in the algorithmics for $f_{\gamma,\mathbf{z}}$. Of course, γ^* depends on the number of examples m, confidence $1 - \delta$, as well as the operator A and a simple invariant of ρ.

2. RKHS and Regularization Parameters

Let X be a compact domain or a manifold in Euclidean space and let $Y = \mathbb{R}$ (one can extend all that follows to $Y = \mathbb{R}^k$ with $k \in \mathbb{N}$). Let ρ be a Borel probability measure on $Z = X \times Y$.

For every $x \in X$, let $\rho(y \mid x)$ be the conditional (with respect to x) probability measure on Y and let ρ_X be the marginal probability measure on X, i.e., the measure on X defined by $\rho_X(S) = \rho(\pi^{-1}(S))$ where $\pi : X \times Y \to X$ is the projection. Notice that ρ, $\rho(y \mid x)$, and ρ_X are related as follows. For every integrable function

* Corrections to [**CS**]:

(1) A regularity hypothesis on measure ρ_X on X requiring every open set on X to have positive measure is needed for our extension of Mercer Theorem and its applications. This is a mild hypothesis since open sets with zero measure can be deleted from X with no harm.

(2) In connection with the matrices associated to a Mercer kernel, the "positive definite" condition should be relaxed to "positive semidefinite."

$\varphi : X \times Y \to \mathbb{R}$ a version of Fubini's theorem states that

$$\int_{X\times Y} \varphi(x, y)\, d\rho = \int_X \left(\int_Y \varphi(x, y)\, d\rho(y \mid x) \right) d\rho_X.$$

This "breaking" of ρ into the measures $\rho(y \mid x)$ and ρ_X corresponds to looking at Z as a product of an input domain X and an output set Y.

Define $f_\rho : X \to Y$ by

$$f_\rho(x) = \int_Y y\, d\rho(y \mid x).$$

The function f_ρ is called the *regression function* of ρ. For each $x \in X$, $f_\rho(x)$ is the average of the y coordinate of $\{x\}\times Y$.

In what follows we assume that $f_\rho \in \mathcal{L}^2_\rho(X)$ is bounded. We also assume that

$$M_\rho = \inf\{\bar{M} \geq 0 \mid \{(x, y) \in Z \mid |y - f_\rho(x)| \geq \bar{M}\} \text{ has measure zero}\}$$

is finite. Note that this implies that

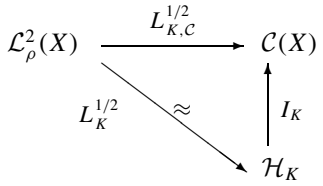
$$|y| \leq M = \max\{\|f_\rho\|_\infty + M_\rho, 1\}$$

almost surely.

Recall, $\|f\|$ denotes, unless otherwise specified, the norm of f in $\mathcal{L}^2_\rho(X)$. Let K be a *Mercer kernel*. That is, $K : X\times X \to \mathbb{R}$ is continuous, symmetric, and K is *positive semidefinite*, i.e., for all finite sets $\{x_1, \ldots, x_k\} \subset X$ the $k\times k$ matrix $K[\mathbf{x}]$ whose (i, j) entry is $K(x_i, x_j)$ is positive semidefinite. Then (see Chapter III of [**CS**]) K determines a linear operator $L_K : \mathcal{L}^2_\rho(X) \to \mathcal{C}(X)$ given by

$$(L_K f)(x) = \int K(x, t) f(t)\, dt$$

which is well-defined, positive, and compact. In addition, there exists a Hilbert space $\mathcal{H}_K$ of continuous functions on X (called *reproducing kernel Hilbert space*, RKHS for short) associated to K and X and independent of ρ such that the linear map $L_K^{1/2}$ is a Hilbert isomorphism between $\mathcal{L}^2_\rho(X)$ and $\mathcal{H}_K$. Here $L_K^{1/2}$ denotes the square root of L_K, i.e., the only linear operator satisfying $L_K^{1/2} \circ L_K^{1/2} = L_K$. Thus, we have the following diagram:



where we write $L_{K,\mathcal{C}}$ to emphasize that the target is $\mathcal{C}(X)$ and I_K denotes the inclusion. If K is $\mathcal{C}^\infty$ then I_K is compact. In the sequel, K denotes a $\mathcal{C}^\infty$ Mercer kernel, and $\| \ \|_K$ denotes the norm in $\mathcal{H}_K$.

Let $\mathbf{z} = (z_1, \ldots, z_m)$ with $z_i = (x_i, y_i) \in X \times Y$ for $i = 1, \ldots, m$. We also write $\mathbf{x} = (x_1, \ldots, x_m)$, $\mathbf{y} = (y_1, \ldots, y_m)$. Note that since K is a Mercer kernel, $K[\mathbf{x}]$ is positive semidefinite.

For $\gamma > 0$, let $\Phi(\gamma)$ and $\Phi_{\mathbf{z}}(\gamma)$ be the problems, respectively,

$$\begin{array}{ll} \min & \displaystyle\int_X (f(x) - y)^2 + \gamma \|f\|_K^2, \\ \text{s.t.} & f \in \mathcal{H}_K, \end{array}$$

and

$$\begin{array}{ll} \min & \displaystyle\frac{1}{m}\sum_{i=1}^m (f(x_i) - y_i)^2 + \gamma \|f\|_K^2, \\ \text{s.t.} & f \in \mathcal{H}_K. \end{array}$$

For $x \in X$, let $K_x : X \to \mathbb{R}$ be given by $K_x(t) = K(x, t)$.

Theorem 1. *For all $\gamma > 0$, the minimizers f_γ and $f_{\gamma,\mathbf{z}}$ of $\Phi(\gamma)$ and $\Phi_{\mathbf{z}}(\gamma)$, respectively, exist and are unique. In addition,*

$$f_\gamma = (\mathrm{Id} + \gamma L_K^{-1})^{-1} f_\rho$$

and $f_{\gamma,\mathbf{z}}$ is given by

$$f_{\gamma,\mathbf{z}}(x) = \sum_{i=1}^m a_i K(x, x_i)$$

where $a = (a_1, \ldots, a_m)$ is the unique solution of the well-posed linear system in $\mathbb{R}^m$,

$$(\gamma m \,\mathrm{Id} + K[\mathbf{x}])a = \mathbf{y}.$$

Finally, for $f = \sum_{i=1}^m a_i K_{x_i}$ we have $\|f\|_K^2 = a^{\mathrm{T}} K[\mathbf{x}]a$.

Proof. See Propositions 7 and 8 and Theorem 2 in Chapter III of [**CS**] and its references, and [**5**] and its references. □

3. Estimating the Confidence

Define, for $f \in \mathcal{L}_\rho^2(X)$, its *error*

$$\mathcal{E}(f) = \int_Z (f(x) - y)^2$$

and, given a sample $\mathbf{z} \in Z^m$, its *empirical error*

$$\mathcal{E}_{\mathbf{z}}(f) = \frac{1}{m}\sum_{i=1}^{m}(f(x_i) - y_i)^2.$$

Note that from the equality $\mathcal{E}(f_{\gamma,\mathbf{z}}) = \mathcal{E}(f_{\gamma,\mathbf{z}}) - \mathcal{E}(f_\gamma) + \mathcal{E}(f_\gamma)$ we deduce that

$$\mathcal{E}(f_{\gamma,\mathbf{z}}) \leq |\mathcal{E}(f_{\gamma,\mathbf{z}}) - \mathcal{E}(f_\gamma)| + \mathcal{E}(f_\gamma).$$

We will call the first term in the right-hand side, the *sample error* (this use of this expression slightly differs from the one in [**CS**]) and the second, the *approximation error*. Note that the sample error is a random variable on the space Z^m. In this section we will bound the confidence for the sample error to be small enough. The main result is Theorem 2 below.

For $r > 0$ let $B_r = \{f \in \mathcal{H}_K \mid \|f\|_K \leq r\}$ and $\mathcal{H}(r) = \overline{I_K(B_r)}$. Notice that this is a compact subset of $\mathcal{C}(X)$ so that, for every η, the *covering number*

$$\mathcal{N}(\mathcal{H}(r), \eta) = \min\{s \in \mathbb{N} \mid \exists\ s \text{ closed balls of radius } \eta \text{ in } \mathcal{C}(X) \text{ covering } \mathcal{H}(r)\}$$

is finite. Also, let

$$\mathbf{C}_K = \max\left\{1,\ \sup_{x,t\in X}|K(x,t)|\right\}$$

and

$$R_\gamma = \frac{\sqrt{\mathbf{C}_K}\|f_\rho\|_\infty}{\gamma} \qquad \text{and} \qquad r_\gamma = \frac{\sqrt{\mathbf{C}_K}M}{\gamma}.$$

Theorem 2. *For all* $\gamma, \varepsilon > 0$,

$$\operatorname*{Prob}_{\mathbf{z}\in Z^m}\{|\mathcal{E}(f_\gamma) - \mathcal{E}(f_{\gamma,\mathbf{z}})| \leq \varepsilon\}$$
$$\geq 1 - 4\left[m + \mathcal{N}\left(\mathcal{H}(r_\gamma), \frac{\varepsilon\gamma}{32M(\gamma + \mathbf{C}_K)}\right)\right] e^{-\frac{m\varepsilon^2\gamma^4}{128M^4(\gamma+\mathbf{C}_K)^4}}.$$

The idea toward the proof of Theorem 2 is to write

$$\mathcal{E}(f_\gamma) - \mathcal{E}(f_{\gamma,\mathbf{z}}) = \mathcal{E}(f_\gamma) - \mathcal{E}_{\mathbf{z}}(f_\gamma) + \mathcal{E}_{\mathbf{z}}(f_\gamma) - \mathcal{E}_{\mathbf{z}}(f_{\gamma,\mathbf{z}}) + \mathcal{E}_{\mathbf{z}}(f_{\gamma,\mathbf{z}}) - \mathcal{E}(f_{\gamma,\mathbf{z}})$$

from which it follows that

$$|\mathcal{E}(f_\gamma) - \mathcal{E}(f_{\gamma,\mathbf{z}})| \leq |\mathcal{E}(f_\gamma) - \mathcal{E}_{\mathbf{z}}(f_\gamma)| + |\mathcal{E}_{\mathbf{z}}(f_\gamma) - \mathcal{E}_{\mathbf{z}}(f_{\gamma,\mathbf{z}})| + |\mathcal{E}_{\mathbf{z}}(f_{\gamma,\mathbf{z}}) - \mathcal{E}(f_{\gamma,\mathbf{z}})|.$$

We first (see Proposition 1 below), bound (with high confidence) the first and last terms in the sum above. Toward this end, we give bounds on $\|f_\gamma\|_K$, $\|f_{\gamma,\mathbf{z}}\|_K$, $\|f_\gamma\|_\infty$, and $\|f_{\gamma,\mathbf{z}}\|_\infty$.

Lemma 1. *For all* $\gamma > 0$,

$$\|f_\gamma\|_K \leq R_\gamma.$$

Proof. Let $f_\rho = \sum c_i \varphi_i$. Then

$$f_\gamma = \sum_{i=1}^{\infty} \left(1 + \frac{\gamma}{\lambda_i}\right)^{-1} c_i \varphi_i = \sum_{i=1}^{\infty} \left(\frac{\lambda_i}{\gamma + \lambda_i}\right) c_i \varphi_i$$

and, therefore,

$$\begin{aligned} \|f_\gamma\|_K^2 &= \sum_{i=1}^{\infty} \frac{\lambda_i}{(\gamma + \lambda_i)^2} c_i^2 \\ &\leq \max_{i \geq 1} \frac{\lambda_i}{(\gamma + \lambda_i)^2} \sum_{i=1}^{\infty} c_i^2 \\ &\leq \frac{1}{\gamma^2} \max_{i \geq 1} \lambda_i \ \|f_\rho\|^2 \\ &\leq \frac{\mathbf{C}_K}{\gamma^2} \|f_\rho\|^2. \end{aligned}$$

□

Lemma 2. *For all* $\gamma > 0$,

$$\|f_{\gamma,\mathbf{z}}\|_K \leq r_\gamma.$$

Proof. Since $f_{\gamma,\mathbf{z}} = \sum a_i K_{x_i}$ we have $\|f_{\gamma,\mathbf{z}}\|_K^2 = a^{\mathrm{T}} K[\mathbf{x}]a$.

Also, since $a = (\gamma m \,\mathrm{Id} + K[\mathbf{x}])^{-1}\mathbf{y}$ it follows that

$$\|a\| \leq \|\mathbf{y}\| \|(\gamma m \,\mathrm{Id} + K[\mathbf{x}])^{-1}\| \leq \sqrt{m} M \frac{1}{\gamma m} = \frac{M}{\gamma \sqrt{m}},$$

where $\|a\|$ and $\|\mathbf{y}\|$ refer to the Euclidean norm in $\mathbb{R}^m$. Therefore

$$\|f_{\gamma,\mathbf{z}}\|_K^2 \leq \|a\|^2 \|K[\mathbf{x}]\| \leq \frac{M^2}{\gamma^2 m} \mathbf{C}_K m = \frac{M^2}{\gamma^2} \mathbf{C}_K,$$

where $\|K[\mathbf{x}]\|$ denotes the operator norm of $K[\mathbf{x}] : \mathbb{R}^m \to \mathbb{R}^m$ with respect to the Euclidean norm in both domain and target space and we have used that, since each entry of $K[\mathbf{x}]$ is bounded in absolute value by $\mathbf{C}_K$, $\|K[\mathbf{x}]\| \leq \mathbf{C}_K m$. □

Corollary 1. *For all* $\gamma > 0$, $\|f_\gamma\|_\infty \leq \mathbf{C}_K \|f_\rho\|_\infty / \gamma$ *and* $\|f_{\gamma,\mathbf{z}}\|_\infty \leq \mathbf{C}_K M/\gamma$.

Proof. By Theorem 2 in Chapter III of [**CS**], $\|I_K\| \leq \sqrt{\mathbf{C}_K}$. □

Remark 1. Note that for all $\gamma > 0$, $r_\gamma \geq R_\gamma$.

Proposition 1. *For all* $\varepsilon, \gamma > 0$,

(i)

$$\operatorname*{Prob}_{\mathbf{z}\in Z^m}\{|\mathcal{E}(f_\gamma) - \mathcal{E}_{\mathbf{z}}(f_\gamma)| \le \varepsilon\} \ge 1 - \mathcal{N}\left(\mathcal{H}(R_\gamma), \frac{\varepsilon\gamma}{8M(\gamma + \mathbf{C}_K)}\right) 2e^{-\frac{m\varepsilon^2\gamma^4}{8M^4(\gamma+\mathbf{C}_K)^4}}.$$

(ii)

$$\operatorname*{Prob}_{\mathbf{z}\in Z^m}\{|\mathcal{E}(f_{\gamma,\mathbf{z}}) - \mathcal{E}_{\mathbf{z}}(f_{\gamma,\mathbf{z}})| \le \varepsilon\} \ge 1 - \mathcal{N}\left(\mathcal{H}(r_\gamma), \frac{\varepsilon\gamma}{8M(\gamma + \mathbf{C}_K)}\right) 2e^{-\frac{m\varepsilon^2\gamma^4}{8M^4(\gamma+\mathbf{C}_K)^4}}.$$

Proof. We use Theorem B of [**CS**], but proved with Hoeffding's inequality instead of Bernstein's. This yields, for a compact subset $\mathcal{H}$ of $\mathcal{C}(X)$ such that $|f(x) - y| \le M$ almost everywhere for all $f \in \mathcal{H}$, the uniform estimate

$$\operatorname*{Prob}_{\mathbf{z}\in Z^m}\left\{\sup_{f\in\mathcal{H}} |\mathcal{E}(f) - \mathcal{E}_{\mathbf{z}}(f)| \le \varepsilon\right\} \ge 1 - \mathcal{N}\left(\mathcal{H}, \frac{\varepsilon}{8M}\right) 2e^{-\frac{m\varepsilon^2}{8M^4}}.$$

For (i), use this estimate applied to $\mathcal{H} = \mathcal{H}(R_\gamma)$, and

$$M = \|f_\gamma\|_\infty + M_\rho + \|f_\rho\|_\infty \le \frac{\mathbf{C}_K\|f_\rho\|_\infty}{\gamma} + M \le \frac{M(\gamma + \mathbf{C}_K)}{\gamma}.$$

and note that

$$\operatorname*{Prob}_{\mathbf{z}\in Z^m}\{|\mathcal{E}(f_\gamma) - \mathcal{E}_{\mathbf{z}}(f_\gamma)| \le \varepsilon\} \ge \operatorname*{Prob}_{\mathbf{z}\in Z^m}\left\{\sup_{f\in\mathcal{H}(R_\gamma)} |\mathcal{E}(f) - \mathcal{E}_{\mathbf{z}}(f)| \le \varepsilon\right\}.$$

A similar proof, now with $\mathcal{H} = \mathcal{H}(r_\gamma)$, and

$$M = \|f_\gamma\|_\infty + M_\rho + \|f_\rho\|_\infty \le \frac{\mathbf{C}_K M}{\gamma} + M = \frac{M(\gamma + \mathbf{C}_K)}{\gamma},$$

yields (ii). □

We now proceed with the middle term $|\mathcal{E}_{\mathbf{z}}(f_\gamma) - \mathcal{E}_{\mathbf{z}}(f_{\gamma,\mathbf{z}})|$.

In what follows, for $f : X \to \mathbb{R}$ and $\mathbf{x} \in X^m$, we denote by $f[\mathbf{x}]$ the point $(f(x_1), \ldots, f(x_m)) \in \mathbb{R}^m$. Also, if $v \in \mathbb{R}^m$, we denote $\|v\|_{\max} = \max\{|v_1|, \ldots, |v_m|\}$.

Proposition 2. *For all* $\gamma, \varepsilon > 0$,

$$\operatorname*{Prob}_{\mathbf{z}\in Z^m}\{\|f_\gamma[\mathbf{x}] - f_{\gamma,\mathbf{z}}[\mathbf{x}]\|_{\max} \le 2\varepsilon\} \ge 1 - 4me^{-\frac{m\varepsilon^2\gamma^4}{2\mathbf{C}_K^2M^2(\gamma+\mathbf{C}_K)^2}}.$$

Proof of Theorem 2. Recall,

$$|\mathcal{E}(f_\gamma) - \mathcal{E}(f_{\gamma,\mathbf{z}})| \le |\mathcal{E}(f_\gamma) - \mathcal{E}_\mathbf{z}(f_\gamma)| + |\mathcal{E}_\mathbf{z}(f_\gamma) - \mathcal{E}_\mathbf{z}(f_{\gamma,\mathbf{z}})| + |\mathcal{E}_\mathbf{z}(f_{\gamma,\mathbf{z}}) - \mathcal{E}(f_{\gamma,\mathbf{z}})|.$$

The first and last terms are each bounded by ε with probabilities at least

$$1 - \mathcal{N}\left(\mathcal{H}(r_\gamma), \frac{\varepsilon\gamma}{8M(\gamma + \mathbf{C}_K)}\right) 2e^{-\frac{m\varepsilon^2\gamma^4}{8M^4(\gamma+\mathbf{C}_K)^4}}$$

by Proposition 1 and the fact that $r_\gamma \ge R_\gamma$. For the middle term note that

$$\begin{aligned} |\mathcal{E}_\mathbf{z}(f_\gamma) - \mathcal{E}_\mathbf{z}(f_{\gamma,\mathbf{z}})| &= \frac{1}{m}\left|\sum_{i=1}^m (f_\gamma(x_i) - y_i) - \sum_{i=1}^m (f_{\gamma,\mathbf{z}}(x_i) - y_i)\right| \\ &\le \frac{1}{m}\sum_{i=1}^m |f_\gamma(x_i) - f_{\gamma,\mathbf{z}}(x_i)| \\ &\le \|f_\gamma[\mathbf{x}] - f_{\gamma,\mathbf{z}}[\mathbf{x}]\|_{\max}. \end{aligned}$$

Now apply Proposition 2 to bound this term by 2ε with probability at least

$$1 - 4me^{-\frac{m\varepsilon^2\gamma^4}{2\mathbf{C}_K^2 M^2(\gamma+\mathbf{C}_K)^2}}$$

and the conclusion follows by noting that $2\mathbf{C}_K^2 M^2(\gamma + \mathbf{C}_K)^2 \le 8M^4(\gamma + \mathbf{C}_K)^4$ and by replacing ε by $\varepsilon/4$. □

It only remains to prove Proposition 2. Toward this end, recall, Hoeffding's inequality states that if ξ is a random variable on a probability space Z bounded almost everywhere by M with mean μ, then

$$\operatorname*{Prob}_{\mathbf{z}\in Z^m}\left\{\left|\frac{1}{m}\sum_{i=1}^m \xi(z_i) - \mu\right| \le \varepsilon\right\} \ge 1 - 2e^{-\frac{m\varepsilon^2}{2M}}.$$

Lemma 3. *For all $\gamma, \varepsilon > 0$ and all $t \in X$,*

$$\operatorname*{Prob}_{\mathbf{z}\in Z^m}\left\{\left|\frac{1}{m\gamma}\sum_{i=1}^m K(x_i, t)(f_\rho(x_i) - y_i)\right| \le \varepsilon\right\} \ge 1 - 2e^{-\frac{m\varepsilon^2\gamma^2}{2(\mathbf{C}_K M_\rho)^2}}.$$

Proof. Consider the random variable

$$z \mapsto \frac{1}{\gamma}K(x, t)(f_\rho(x) - y).$$

It is almost everywhere bounded by $(1/\gamma)\mathbf{C}_K M_\rho$. Its mean is 0 since, by Fubini's theorem,

$$\int_Z \frac{1}{\gamma} K(x,t)(f_\rho(x) - y) = \int_X \frac{1}{\gamma} K(x,t) \left(\int_Y f_\rho(x) - y \, d\rho(y \mid x) \right) d\rho_X$$

and the inner integral is 0 by definition of f_ρ. Now apply Hoeffding's inequality. □

Lemma 4. *For all $\gamma, \varepsilon > 0$ and all $t \in X$,*

$$\operatorname*{Prob}_{\mathbf{z} \in Z^m} \left\{ \left| f_\gamma(t) - \frac{1}{m\gamma} \sum_{i=1}^m K(x_i, t)(f_\rho(x_i) - f_\gamma(x_i)) \right| \le \varepsilon \right\}$$
$$\ge 1 - 2e^{-\frac{m\varepsilon^2\gamma^2}{2\mathbf{C}_K^2(\|f_\rho\|_\infty + \sqrt{\mathbf{C}_K} R_\gamma)^2}}.$$

Proof. By Theorem 1,

$$\begin{aligned} f_\gamma = (\mathrm{Id} + \gamma L_K^{-1})^{-1} f_\rho \quad &\Rightarrow \quad f_\gamma + \gamma L_K^{-1} f_\gamma = f_\rho \\ &\Rightarrow \quad L_K f_\gamma + \gamma f_\gamma = L_K f_\rho \\ &\Rightarrow \quad f_\gamma = \frac{1}{\gamma} L_K (f_\rho - f_\gamma) \\ &\Rightarrow \quad f_\gamma(t) = \int_X \left(\frac{1}{\gamma} K(x,t)(f_\rho(x) - f_\gamma(x)) \right) d\rho_X. \end{aligned}$$

The function inside the last integral can thus be considered a random variable on X with mean $f_\gamma(t)$. It is bounded by $(\mathbf{C}_K/\gamma)(\|f_\rho\|_\infty + \sqrt{\mathbf{C}_K} R_\gamma)$. Again, apply Hoeffding's inequality. □

Lemma 5. *For all $\gamma, \varepsilon > 0$,*

$$\operatorname*{Prob}_{\mathbf{z} \in Z^m} \left\{ \left\| \left(\mathrm{Id} + \frac{K[\mathbf{x}]}{\gamma m} \right) f_\gamma[\mathbf{x}] - \frac{K[\mathbf{x}]\mathbf{y}}{\gamma m} \right\|_{\max} \le 2\varepsilon \right\} \ge 1 - 4me^{-\frac{m\varepsilon^2\gamma^4}{2\mathbf{C}_K^2 M^2(\gamma + \mathbf{C}_K)^2}}.$$

Proof. From Lemmas 3 and 4 it follows that, with a probability at least

$$1 - 2\left(e^{-\frac{m\varepsilon^2\gamma^2}{2(\mathbf{C}_K M_\rho)^2}} + e^{-\frac{m\varepsilon^2\gamma^2}{2\mathbf{C}_K^2(\|f_\rho\|_\infty + \sqrt{\mathbf{C}_K} R_\gamma)^2}} \right)$$

for every $t \in X$,

$$\left| f_\gamma(t) + \frac{1}{m\gamma} \sum_{i=1}^m K(x_i, t)(f_\gamma(x_i) - y_i) \right| \le 2\varepsilon.$$

Note that, since

$$\max\{M_\rho, \|f_\rho\|_\infty + \sqrt{\mathbf{C}_K} R_\gamma\} \leq M + \sqrt{\mathbf{C}_K} r_\gamma = \frac{M(\gamma + \mathbf{C}_K)}{\gamma}$$

the confidence above is at least

$$1 - 4e^{-\frac{m\varepsilon^2\gamma^4}{2\mathbf{C}_K^2 M^2(\gamma+\mathbf{C}_K)^2}}.$$

Applying this to $t = x_1, \ldots, x_m$ and writing the m resulting inequalities in matrix form we obtain that, with confidence at least the one in the statement,

$$\left\| f_\gamma[\mathbf{x}] + \frac{1}{m\gamma} K[\mathbf{x}] f_\gamma[\mathbf{x}] - \frac{1}{m\gamma} K[\mathbf{x}][\mathbf{y}] \right\|_{\max} \leq 2\varepsilon. \qquad \square$$

Lemma 6. *For all* $\gamma, \varepsilon > 0$,

$$\left(\mathrm{Id} + \frac{K[\mathbf{x}]}{\gamma m}\right) f_{\gamma,\mathbf{z}}(x) = \frac{K[\mathbf{x}]\mathbf{y}}{\gamma m}.$$

Proof. In Proposition 8, Chapter III of [**CS**] it is shown that

$$\begin{aligned} f_{\gamma,\mathbf{z}}(t) &= \sum_{i=1}^m \tfrac{y_i - f_{\gamma,\mathbf{z}}(x_i)}{\gamma m} K(x_i, t) \\ \Rightarrow \quad \gamma m f_{\gamma,\mathbf{z}}(t) &= \sum_{i=1}^m (y_i - f_{\gamma,\mathbf{z}}(x_i)) K(x_i, t) \\ \Rightarrow \quad \gamma m f_{\gamma,\mathbf{z}}(t) &+ \sum_{i=1}^m f_{\gamma,\mathbf{z}}(x_i) K(x_i, t) = \sum_{i=1}^m y_i K(x_i, t). \end{aligned}$$

Applying this equality for $t = x_1, \ldots, x_m$ and writing the resulting m equalities in matrix form we obtain

$$\gamma m f_{\gamma,\mathbf{z}}[\mathbf{x}] + f_{\gamma,\mathbf{z}}[\mathbf{x}] K[\mathbf{x}] = K[\mathbf{x}]\mathbf{y}$$

from which the statement follows. $\square$

Proof of Proposition 2. From Lemmas 5 and 6 it follows that

$$\left\| \left(\mathrm{Id} + \frac{K[\mathbf{x}]}{\gamma m}\right) f_\gamma[\mathbf{x}] - \left(\mathrm{Id} + \frac{K[\mathbf{x}]}{\gamma m}\right) f_{\gamma,\mathbf{z}}[\mathbf{x}] \right\|_{\max} \leq 2\varepsilon,$$

i.e.,

$$\left\| \left(\mathrm{Id} + \frac{K[\mathbf{x}]}{\gamma m}\right) (f_\gamma[\mathbf{x}] - f_{\gamma,\mathbf{z}}[\mathbf{x}]) \right\|_{\max} \leq 2\varepsilon$$

with the stated confidence. The result now follows since $K[\mathbf{x}]/\gamma m$ is positive definite and therefore $\|(\mathrm{Id} + K[\mathbf{x}]/\gamma m)^{-1}\| \geq 1$. $\square$

4. Estimating the Sample Error

The expression $|\mathcal{E}(f_\gamma) - \mathcal{E}(f_{\gamma,\mathbf{z}})|$ is called the *sample error* (of $f_{\gamma,\mathbf{z}}$). In the previous section we estimated the confidence of obtaining a small sample error when the sample size m and an error bound ε are given. In this section we will fix a confidence $1-\delta$ and a sample size m and obtain bounds for the sample error.

Lemma 7. *Let $c_1, c_2 > 0$ and $s > q > 0$. Then the equation*

$$x^s - c_1 x^q - c_2 = 0$$

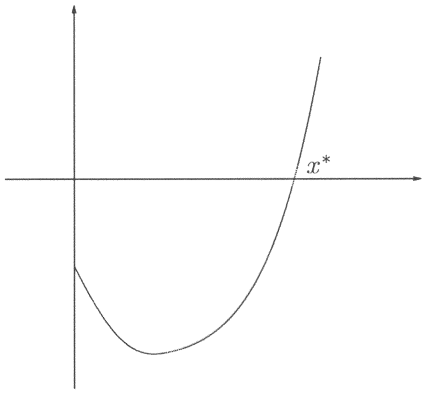
has a unique positive zero x^*. In addition,

$$x^* \le \max\{(2c_1)^{1/(s-q)}, (2c_2)^{1/s}\}.$$

Proof. Let $\varphi(x) = x^s - c_1 x^q - c_2$. Then, taking the derivative with respect to x, $\varphi'(x) = sx^{s-1} - qc_1 x^{q-1} = x^{q-1}(sx^{s-q} - qc_1)$. Thus

$$\varphi'(x) = 0 \quad \Leftrightarrow \quad x^{s-q} = \frac{qc_1}{s}$$

and this derivative has a unique positive zero. The first statement follows since $\varphi(0) < 0$, $\varphi'(0^+) \le 0$, and $\varphi(x) \to +\infty$ when $x \to +\infty$.



The second statement is a well-known bound (see [**10, Theorem 4.2(iv)**]). □

Remark 2. Note that, given c_1, c_2, s, and t, one can efficiently compute (a good approximation of) x^* using algorithms such as Newton's method.

By Theorem 2, the sample error ε satisfies, with confidence $1-\delta$,

$$4m\mathcal{N}\left(\mathcal{H}(r_\gamma), \frac{\varepsilon\gamma}{32M(\gamma + \mathbf{C}_K)}\right) e^{-\frac{m\varepsilon^2\gamma^4}{128M^4(\gamma+\mathbf{C}_K)^4}} \ge \delta,$$

i.e.,

$$\frac{m\varepsilon^2\gamma^4}{128M^4(\gamma+\mathbf{C}_K)^4} - \ln\left(\frac{4m}{\delta}\right) - \ln\mathcal{N}\left(\mathcal{H}(r_\gamma), \frac{\varepsilon\gamma}{32M(\gamma+\mathbf{C}_K)}\right) \le 0. \quad (1)$$

Now we recall (see Section 6 in Chapter I of [**CS**]) that, for every $t < 2$, there exists a constant C_t independent of ε and γ, such that

$$\begin{aligned}\ln\mathcal{N}\left(\mathcal{H}(r_\gamma), \frac{\varepsilon\gamma}{32M(\gamma+\mathbf{C}_K)}\right) &\le \left(\frac{r_\gamma C_t 32M(\gamma+\mathbf{C}_K)}{\varepsilon\gamma}\right)^t \\ &\le \left(\frac{32C_tM^2(\gamma+\mathbf{C}_K)^2}{\varepsilon\gamma^2}\right)^t\end{aligned}$$

(a different bound appears in [**15**]). Note that in the last inequality we replaced r_γ by its definition and used that $\sqrt{\mathbf{C}_K} \le (\mathbf{C}_K + \gamma)$. Using this bound for the covering number, inequality (1) becomes

$$\frac{m\varepsilon^2\gamma^4}{128M^4(\gamma+\mathbf{C}_K)^4} - \ln\left(\frac{4m}{\delta}\right) - \left(\frac{32C_tM^2(\gamma+\mathbf{C}_K)^2}{\varepsilon\gamma^2}\right)^t \le 0.$$

Write

$$v = \frac{\varepsilon\gamma^2}{32M^2(\gamma+\mathbf{C}_K)^2}.$$

Then the inequality above takes the form

$$c_0v^2 - c_1 - c_2v^{-t} \le 0, \quad (2)$$

where $c_0 = m/4$, $c_1 = \ln(4m/\delta)$, $c_2 = C_t^t$, the tth power of C_t. Note that one could fix, for example, $t = 1$.

Now take the equality in (2) to obtain the equation

$$\varphi(v) = v^{t+2} - \frac{c_1}{c_0}v^t - \frac{c_2}{c_0} = 0$$

and note that this equation has only one positive zero by Lemma 7. Let $v^*(m,\delta)$ be this solution. Then, also by Lemma 7,

$$v^*(m,\delta) \le \left\{\left(\frac{8\ln(4m)-\ln(\delta)}{m}\right)^{1/2}, \left(\frac{8C_t^t}{m}\right)^{t+2}\right\} \quad (3)$$

and

$$\varepsilon = \frac{32M^2(\gamma+\mathbf{C}_K)^2}{\gamma^2}v^*(m,\delta)$$

and we can conclude stating the following result.

Theorem 3. *Given $m \ge 1$ and $0 < \delta \le 1$, for all $\gamma > 0$, the expression*

$$\mathcal{S}(\gamma) = \frac{32M^2(\gamma+\mathbf{C}_K)^2}{\gamma^2}v^*(m,\delta)$$

bounds the sample error with confidence at least $1-\delta$. □

5. Choosing the Optimal γ

We now focus on the approximation error $\mathcal{E}(f_\gamma)$. To do so we first apply part (1) of Theorem 3, Chapter II in [**CS**] with $H = \mathcal{L}^2_\rho(X)$, $s = 1$, $A = L_K^{1/2}$, and $a = f_\rho$, and use that $\|L_K^{-1/2} f\| = \|f\|_K$ to obtain that, for $0 < \theta < 1$ (e.g., for $\theta = \frac{1}{2}$),

$$\min_{f \in \mathcal{L}^2_\rho(X)} (\|f - f_\rho\|^2 + \gamma \|f\|_K^2) \leq \gamma^\theta \|L_K^{-\theta/2} f_\rho\|^2.$$

Since the minimum above is attained at f_γ we deduce

$$\|f_\gamma - f_\rho\|^2 + \gamma \|f_\gamma\|_K^2 \leq \gamma^\theta \|L_K^{-\theta/2} f_\rho\|^2.$$

A basic result in [**CS**, Proposition 1, Chapter I] states that, for all $f \in \mathcal{L}^2_\rho(X)$,

$$\mathcal{E}(f) = \int_X (f - f_\rho)^2 + \sigma_\rho^2, \tag{4}$$

where σ_ρ^2 is a nonnegative quantity depending only on ρ. Therefore the approximation error $\mathcal{E}(f_\gamma)$ is bounded by $\mathcal{A}(\gamma) + \sigma_\rho^2$ where

$$\mathcal{A}(\gamma) = \gamma^\theta \|L_K^{-\theta/2} f_\rho\|^2.$$

Proof of the Main Result. Let

$$E(\gamma) = \mathcal{A}(\gamma) + \mathcal{S}(\gamma).$$

Recall

$$\mathcal{E}(f_{\gamma,\mathbf{z}}) \leq \left|\mathcal{E}(f_\gamma) - \mathcal{E}(f_{\gamma,\mathbf{z}})\right| + \mathcal{E}(f_\gamma).$$

Then the error $\mathcal{E}(f_{\gamma,\mathbf{z}})$ satisfies the bound

$$\mathcal{E}(f_{\gamma,\mathbf{z}}) \leq E(\gamma) + \sigma_\rho^2$$

and, therefore, subtracting σ_ρ^2 from both sides and using (4) for $f = f_{\gamma,\mathbf{z}}$,

$$\int_X (f_{\gamma,\mathbf{z}} - f_\rho)^2 \leq E(\gamma).$$

This proves the first part of the Main Result. Note that this is actually a family of bounds parametrized by $t < 2$ and $0 < \theta < 1$ and depends on m, δ, K, and f_ρ.

For a point $\gamma > 0$ to be a minimum of $E(\gamma) = \mathcal{S}(\gamma) + \mathcal{A}(\gamma)$ it is necessary that $\mathcal{S}'(\gamma) + \mathcal{A}'(\gamma) = 0$. Taking derivatives, we get

$$\mathcal{S}'(\gamma) = -64M^2 v^*(m,\delta)\mathbf{C}_K \frac{\gamma + \mathbf{C}_K}{\gamma^3}$$

and

$$\mathcal{A}'(\gamma) = \theta\gamma^{\theta-1}\|L_K^{-\theta/2} f_\rho\|^2.$$

Thus the solutions of $\mathcal{A}'(\gamma) + \mathcal{S}'(\gamma) = 0$ are those of

$$\theta\gamma^{\theta+2}\|L_K^{-\theta/2} f_\rho\|^2 - 64M^2 v^*(m,\delta)\mathbf{C}_K(\gamma + \mathbf{C}_K) = 0,$$

i.e., those of

$$\gamma^{\theta+2} - \frac{64M^2 v^*(m,\delta)\mathbf{C}_K}{\theta\|L_K^{-\theta/2} f_\rho\|^2}\gamma - \frac{64M^2 v^*(m,\delta)\mathbf{C}_K^2}{\theta\|L_K^{-\theta/2} f_\rho\|^2} = 0. \tag{5}$$

Using again Lemma 7, we obtain a unique solution γ^* which is a minimizer of E since $E(\gamma) \to \infty$ as $\gamma \to 0$ or $\gamma \to \infty$. This finishes the proof of the Main Result. □

Corollary 2. *For every* $0 < \delta \le 1$,

$$\lim_{m\to\infty} E(\gamma^*) = 0.$$

Proof. The bound (3) shows that $v^*(m,\delta) \to 0$ when $m \to \infty$. Now, equation (5) shows that γ^* is the only positive root of

$$\gamma^{\theta+2} - Qv^*(m,\delta)\gamma - Qv^*(m,\delta)\mathbf{C}_K = 0, \tag{6}$$

where $Q = 64M^2\mathbf{C}_K/\theta\|L_K^{-\theta/2} f_\rho\|^2$. Then, by Lemma 7,

$$\gamma^* \le \max\{(2Qv^*(m,\delta))^{1/(\theta+1)}, (2Qv^*(m,\delta)\mathbf{C}_K)^{1/(\theta+2)}\}$$

from which it follows that $\gamma^* \to 0$ when $m \to \infty$. Note that this implies that

$$\lim_{m\to\infty} \mathcal{A}(\gamma^*) = \lim_{m\to\infty} (\gamma^*)^\theta \|L_K^{-\theta/2} f_\rho\|^2 = 0.$$

Finally, it follows from equation (6) that

$$(\gamma^*)^\theta - (Q\gamma^* + Q\mathbf{C}_K)\left[\frac{v^*(m,\delta)}{(\gamma^*)^2}\right] = 0$$

and, therefore, that $[v^*(m,\delta)/(\gamma^*)^2] \to 0$ when $m \to \infty$. This, together with Theorem 3, shows that $\lim_{m\to\infty} \mathcal{S}(\gamma^*) = 0$. □

6. Final Remarks

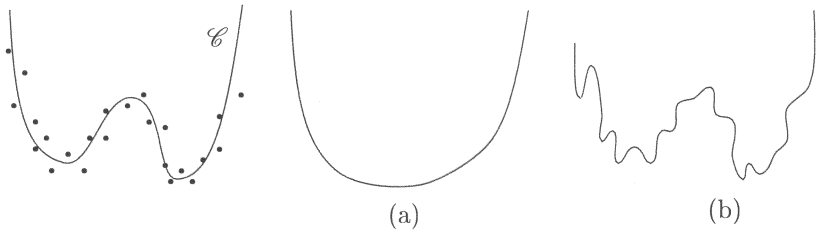
(1) This paper can be seen as a solution of one instance of the *Bias–Variance problem*. Roughly speaking, the "bias" of a solution f coincides with our approx-

imation error, and its "variance" with the sample error. Quoting [**3**],

> A model which is too simple, or too inflexible, will have a large bias, while one which has too much flexibility in relation to the particular data set will have a large variance. Bias and variance are complementary quantities, and the best generalization [i.e., the smallest error] is obtained when we have the best compromise between the conflicting requirements of small bias and small variance.

As described in Section 3, Chapter II in [**CS**], the bias–variance problem amounts to the choice of a compact subspace $\mathcal{H}$ of $\mathcal{C}(X)$ over which $\mathcal{E}_{\mathbf{z}}$ is minimized. A too-small space $\mathcal{H}$ will yield a large bias while one too large will yield a large variance. Several parameters (radius of balls, dimension, etc.) determine the "size" of $\mathcal{H}$ and different instances of the bias–variance problem are obtained by fixing all of them except one and minimizing the error over this nonfixed parameter. Our solution considers the ball of radius $r = \|f_{\gamma,\mathbf{z}}\|_K$ in $\mathcal{H}_K$ and $\mathcal{H} = \overline{I_K(B_r)}$ (a space over which $f_{\gamma,\mathbf{z}}$ minimizes $\mathcal{E}_{\mathbf{z}}$). The number r is our replacement of the VC-dimension. Since γ is inversely proportional to r, large γ corresponds to large bias or approximation error and small γ to large variance or sample error.

Failing to find a good compromise between bias and variance leads to what is called *underfitting* (large bias) or *overfitting* (large variance). As an example, consider the curve $\mathcal{C}$ in the figure below with the set of sample points and assume we want to approximate that curve with a polynomial of degree d (the parameter d determines, in our case, the dimension of $\mathcal{H}$). If d is too small, say $d = 2$, we obtain a curve as in (a) in the figure, which necessarily "underfits" the data points. If d is too large, we can tightly fit the data points but this "overfitting" yields a curve as in (b).



For more on the bias–variance problem, see [**3**], the above-mentioned section in [**CS**], [**6**], and the references in these papers. Note, however, that the literature on this problem is vast and we have only touched on it.

(2) One could interpret the main estimates in this paper in terms of algorithms for approximating solutions of integral equations by Monte Carlo methods. But for most algorithms in the theory of integral equations the points $x_i, i = 1, \ldots, m$, are not randomly choosen but taken, for example, as a set of lattice points of a domain $X \subset \mathbb{R}^n$ (this would correspond to *active learning* in the learning theory literature). Now one might take ρ_X as Lebesgue measure and the x_i from a uniform

grid of points. The theory in our previous paper [**CS**] should permit modifications to deal with this situation and our main result here as well.

(3) Our work can be interpreted in the area of statistics known as regularized nonparametric least squares regression. A general reference for this area is the book by Sara van de Geer [**14**]. Besides the references in this book, the papers [**2**], [**1**], [**8**], [**9**] are also related to our work. A result somewhat similar in spirit to our main result appears in [**11**], [**12**]. Here a function $E(m, n)$ is exhibited bounding the error in terms of the number m of examples and the number n of basis functions in a space of Gaussian radial basis functions and it is shown that, for each $m \in \mathbb{N}$, $E(m, n)$ has a unique minimizer n^*.

Acknowledgments

This work has been substantially funded by a grant from the Research Grants Council of the Hong Kong SAR (project number CityU 1002/99P). Also, the second named author expresses his appreciation to City University of Hong Kong for supporting his visit there in December 2001, when this paper was written.

References

[1] A. R. Barron, Approximation and estimation bounds for artificial neural networks, *Machine Learning* **14** (1994), 115–133.
[2] A. R. Barron, L. Birgé, and P. Massart, Risk bounds for model selection via penalisation, *Probab. Theory Related Fields* **113** (1995), 301–403.
[3] C. M. Bishop, *Neural Networks for Pattern Recognition*, Cambridge University Press, Cambridge, 1995.
[4] F. Cucker and S. Smale, On the mathematical foundations of learning, *Bull. Amer. Math. Soc.* **39** (2002), 1–49.
[5] T. Evgeniou, M. Pontil, and T. Poggio, Regularization networks and support vector machines, *Adv. in Comput. Math.* **13** (2000), 1–50.
[6] G. Golub, M. Heat, and G. Wahba, Generalized cross-validation as a method for choosing a good ridge parameter, *Technometrics* **21** (1979), 215–223.
[7] V. V. Ivanov, *The Theory of Approximate Methods and Their Application to the Numerical Solution of Singular Integral Equations*, Nordhoff International, 1976.
[8] O. V. Lepskii, Asymptotically minimax adaptive estimation. I: Upper bounds, optimally adaptive estimates, *Theory Probab. Appl.* **36** (1991), 682–697.
[9] O. V. Lepskii, Asymptotically minimax adaptive estimation. II: Schemes without optimal adaptation, adaptive estimators, *Theory Probab. Appl.* **37** (1992), 433–448.
[10] M. Mignotte, *Mathematics for Computer Algebra*, Springer-Verlag, New York, 1992.
[11] P. Niyogi and F. Girosi, On the relationship between generalization error, hypothesis complexity and sample complexity for radial basis functions, *Neural Comput.* **8** (1996), 819–842.
[12] P. Niyogi and F. Girosi, Generalization bounds for function approximation from scattered noisy data, *Adv. in Comput. Math.* **50** (1999), 51–80.
[13] A. N. Tikhonov and V. Y. Arsenin, *Solutions of Ill-Posed Problems*, W. H. Winston, 1977.
[14] S. van de Geer, *Empirical Processes in M-Estimation*, Cambridge University Press, Cambridge, 2000.
[15] D.-X. Zhou, The covering numbers in learning theory, *J. Complexity* 2001 (to appear).