REVIEW Communicated by Tomaso Poggio

On Different Facets of Regularization Theory

Zhe Chen
zhechen@soma.crl.mcmaster.ca
Simon Haykin
haykin@mcmaster.ca
Adaptive Systems Lab, Communications Research Laboratory, McMaster University, Hamilton, Ontario, Canada L8S 4K1

No more things should be presumed to exist than are absolutely necessary.—William of Occam

Everything should be made as simple as possible, but not simpler.—Albert Einstein

This review provides a comprehensive understanding of regularization theory from different perspectives, emphasizing smoothness and simplicity principles. Using the tools of operator theory and Fourier analysis, it is shown that the solution of the classical Tikhonov regularization problem can be derived from the regularized functional defined by a linear differential (integral) operator in the spatial (Fourier) domain. State-of-the-art research relevant to the regularization theory is reviewed, covering Occam's razor, minimum length description, Bayesian theory, pruning algorithms, informational (entropy) theory, statistical learning theory, and equivalent regularization. The universal principle of regularization in terms of Kolmogorov complexity is discussed. Finally, some prospective studies on regularization theory and beyond are suggested.

1 Introduction

Most of the inverse problems posed in science and engineering areas are ill posed—computational vision (Poggio, Torre, & Koch, 1985; Bertero, Poggio, & Torre, 1988), system identification (Akaike, 1974; Johansen, 1997), nonlinear dynamic reconstruction (Haykin, 1999), and density estimation (Vapnik, 1998a), to name a few. In other words, given the available input data, the solution to the problem is nonunique (one-to-many) or unstable. The classical regularization techniques, developed by Tikhonov in the 1960s (Tikhonov & Arsenin, 1977), have been shown to be powerful in making the solution well posed and thus have been applied successfully in

model selection and complexity control. Regularization theory was introduced to the machine learning community (Poggio & Girosi, 1990a, 1990b; Barron, 1991). Poggio and Girosi (1990a, 1990b) showed that a regularization algorithm for learning is equivalent to a multilayer network with a kernel in the form of a radial basis function (RBF), resulting in an RBF network. The regularization solution was originally derived by a differential linear operator and its Green's function (Poggio & Girosi, 1990a, 1990b). A large class of generalized regularization networks (RNs) is reviewed in Girosi, Jones, and Poggio (1995). In the classical regularization theory, a recent trend in studying the smoothness of the functional is to put the functionals into the reproducing kernel Hilbert space (RKHS), which has been well developed in different areas (Aronszajn, 1950; Parzen, 1961, 1963; Yosida, 1978; Kailath, 1971; Wahba, 1990). By studying the properties of the functionals in the RKHS, many learning models, including smoothing splines (Kimeldorf & Wahba, 1970; Wahba, 1990), RNs (Girosi et al., 1995), support vector machines (SVMs; Vapnik, 1995, 1998a), and gaussian processes (MacKay, 1998; Williams, 1998a), can be related to each other. In this sense, regularization theory has gone beyond its original implication and expectation.

In this review, we emphasize the theory of operators, Green's function, and kernel functions. In particular, based on operator theory and Fourier analysis, we derive the spectral regularization framework (Chen & Haykin, 2001a, 2001b) along the lines of classical spatial regularization (Poggio & Girosi, 1990a; see also Haykin, 1999). As we show, the regularization approach is per se to expand the solution in terms of a set of Green's functions, which depend on the form of stabilizer in the context of differential or integral operators and their associated boundary conditions. With two principles of regularization, smoothness and simplicity, in mind, we provide an extensive overview of regularization theory. The main contributions of this article are to review the theory of regularization, examine the relations between regularization theory and other related theoretical work, and present some new perspectives.

The rest of the review is organized as follows. Section 2 briefly formulates the ill-posed problem and introduces regularization theory as the solution. Section 3 introduces the classical Tikhonov regularization theoretical framework with prerequisite materials on machine learning and operator theory. Following the theory of operator and Green's function, we derive the spectral regularization framework using Fourier analysis. Starting with Occam's razor and minimum length description (MDL) theory in section 4, we present different facets of regularization theory from sections 5 through 10. Various relevant research topics are reviewed and their connections to regularization theory highlighted. Finally, the universal principle of Kolmogorov complexity for regularization is explored in section 11, followed by the summary and comments in section 12.

2 Why Use Regularization?

A problem is said to be well posed in the Hadamard sense[1] if it satisfies the following three conditions (Tikhonov & Arsenin, 1977; Morozov, 1984; Haykin, 1999):

1. *Existence*. For every input vector $\mathbf{x} \in X$, there exists an output vector $y = F(\mathbf{x})$, where $y \in Y$.[2]
2. *Uniqueness*. For any pair of input vectors $\mathbf{x}, \mathbf{z} \in X$, it follows that $F(\mathbf{x}) = F(\mathbf{z})$ if and only if $\mathbf{x} = \mathbf{z}$.
3. *Continuity (stability)*. The mapping is continuous, that is, for any $\epsilon > 0$ there exists $\zeta = \zeta(\epsilon)$ such that the condition $d_X(\mathbf{x}, \mathbf{z}) < \zeta$ implies that $d_Y(F(\mathbf{x}), F(\mathbf{z})) < \epsilon$, where $d(\cdot, \cdot)$ represents the distance metric between two arguments in their respective spaces.

If any of these three conditions is not satisfied, the problem is said to be ill posed. In terms of operator language, we have

Definition 1 (Kress, 1989). *Let A: $X \to Y$ be an operator from a normed space X into a normed space Y. The equation $Ax = y$ is said to be well posed if A is bijective*[3] *and the inverse operator A^{-1}: $Y \to X$ is continuous. Otherwise the equation is called ill posed.*

According to definition 1, if A is not surjective, then $Ax = y$ is not solvable for all $y \in Y$ and thus violates the existence condition; if A is not injective, $Ax = y$ may have more than one solution and thus violates the uniqueness condition; if A^{-1} exists but is not continuous, then the solution x does not depend continuously on the observation y, which again violates the stability condition (Kress, 1989).

The following three equivalent conditions are usually used to describe the constraints of the solution of the inverse problem (Poggio & Koch, 1985):

- Find x that minimizes $\|Ax - y\|$ that satisfies $\|Px\| < c_1$, where c_1 is a positive scalar constant.
- Find x that minimizes $\|Px\|$ that satisfies $\|Ax - y\| < c_2$, where c_2 is a positive scalar constant.

[1] Hadamard sense is a special case of Tikhonov sense (Vapnik, 1998a).

[2] Throughout this review, X, Y represent the range or domain in specific functional space.

[3] The operator is bijective if it is injective and surjective. If for each $y \in A(X)$ there is at most (or at least) one element $x \in X$ with $Ax = y$, then A is said to be injective (or surjective).

- Find x that minimizes $\|Ax - y\|^2 + \lambda\|Px\|^2$, where λ is a regularization parameter ($\lambda = c_2/c_1$),

where x is the solution, y is the observation, A and P represent some linear operators, and $\|\cdot\|$ is some kind of norm operator depending on a specific physical scenario. The first condition is called the quasi-solution and the second is called the discrepancy principle, both of which belong to the constrained optimization problem. The third condition is an unconstrained optimization problem, which we focus on in this review.

Ill-posed problems are ubiquitous in the real world, since most inverse problems are subject to some physical constraints and have no unique solution. We briefly describe three examples relevant to the neural networks and signal processing communities.

2.1 Computational Vision. The early vision problem is often referred to the first processing stage in computational vision, which consists of decoding two-dimensional images in terms of properties of three-dimensional objects. Many computational vision problems, such as shape from shading, surface reconstruction, edge detection, and computation of optical flow, are generally ill posed (Poggio & Koch, 1985; Poggio et al., 1985; Bertero et al., 1988). In general, the solutions to these ill-posed problems can be formulated as (Poggio et al., 1985)

$$\arg\min_x \|Ax - y\|^2 + \lambda\|Px\|^2. \tag{2.1}$$

By invoking different regularization tools (Marroquin, Mitter, & Poggio, 1987; Poggio, Voorhees, & Yuille, 1988; Yuille & Grzywacz, 1988; Schnorr & Sprengel, 1994; Bernard, 1999), many computational vision problems haven been solved with great success.

2.2 Density Estimation. Density estimation is a fundamental problem in machine learning. Suppose the observed data are sampled by the density $p(\mathbf{x})$ from the cumulative distribution function $F(\mathbf{x})$, which is expressed by

$$\int_{-\infty}^{\mathbf{x}} p(\tau)\, d\tau = F(\mathbf{x}). \tag{2.2}$$

Now the problem is formulated as: Given some data $\mathbf{x}_i$ ($i = 1, \ldots, \ell$), how can $p(\mathbf{x})$ be estimated from a finite number of noisy observations? Empirically, one may estimate the distribution function by

$$F_\ell(\mathbf{x}) = \frac{1}{\ell}\sum_{i=1}^{\ell} \Theta(\mathbf{x} - \mathbf{x}_i), \tag{2.3}$$

where $\Theta(\cdot)$ is a step function; $\Theta(\mathbf{x}) = 1$ if all $x_n > 0$ and $\Theta(\cdot) = 0$ otherwise. The density is further estimated by solving

$$\int_{-\infty}^{\mathbf{x}} p(\tau)\, d\tau = F_\ell(\mathbf{x}). \tag{2.4}$$

Solving this inverse operator is generally a stochastic ill-posed problem, especially in the high-dimensional case (Vapnik, 1998a).

2.3 Dynamic Reconstruction. Nonlinear dynamic reconstruction (e.g., sea clutter dynamics) is a common problem in the physical world (Haykin, 1999). Without loss of generality, the nonlinear dynamics can be formulated by the following state-space model,

$$\mathbf{x}_{t+1} = F(\mathbf{x}_t) + \boldsymbol{\nu}_{1,t} \tag{2.5}$$

$$y_t = G(\mathbf{x}_t) + \nu_{2,t}, \tag{2.6}$$

where F is a nonlinear mapping vector-valued function, G is a scalar-valued function, and $\boldsymbol{\nu}_{1,t}$ and $\nu_{2,t}$ represent process noise and measurement noise contaminating the state variable $\mathbf{x}_t$ and the observable y_t, respectively. Given a time series of observable y_t, the problem is to reconstruct the dynamics described by F, which is generally ill posed in the following sense: (1) for some unknown reasons, the existence condition may be violated; (2) there may not be sufficient information in the observation for reconstructing the nonlinear dynamics uniquely, which thus violates the uniqueness condition; and (3) the unavoidable presence of noise adds uncertainty to the dynamic reconstruction—when the signal-to-noise ratio (SNR) is too low, the continuity condition is also violated.

One way to solve ill-posed problems is to make the problems well posed by incorporating prior knowledge into the solutions (Tikhonov & Arsenin, 1977; Morozov, 1984; Wahba, 1990, 1995; Vapnik, 1998a). The forms of prior knowledge vary and are problem dependent. The most popular and important prior knowledge is the smoothness prior,[4] which assumes that the functional mapping of the input to the output space is continuous and smooth. This is where regularization naturally comes in, arising from the well-known Tikhonov's regularization theory,[5] which we detail in the next section.

[4] Another methodology to embed prior knowledge is the theory of hints and virtual examples. See Abu-Mostafa (1995), and Niyogi, Girosi, and Poggio (1998).

[5] Many well-known statistical terminologies (for example, *ridge regression, penalized least-square estimate, penalized likelihood estimate, smoothing splines*, and *averaging kernel regression*) can be well formulated in Tikhonov's regularization framework.

3 Regularization Framework

3.1 Machine Learning. Consider the following machine learning problem: Given a set of observation data (learning examples) $\{(\mathbf{x}_i, y_i) \in \mathbb{R}^N \times \mathbb{R}\}_{i=1}^{\ell} \subset X \times Y$, the learning machine f is expected to find the solution to the inverse problem. In other words, it needs to approximate a real function in the hypothesis satisfying the constraints $f(\mathbf{x}_i) = y(\mathbf{x}_i) \equiv y_i$, where $y(\mathbf{x})$ is supposed to be a deterministic function in the target space. Hence, the learning can be viewed as a multivariate functional approximation problem (Poggio & Girosi, 1990a, 1990b). It can be also viewed as an interpolation problem (Powell, 1985; Micchelli, 1986; Broomhead & Lowe, 1988), where f is an interpolant parameterized by weights $\mathbf{w}$. Note that this problem is ill posed in that the approximants satisfying the constraints are not unique. To solve the ill-posed problem, we usually require some smoothness property of the solution f, and the regularization theory naturally comes in.

Statistically, the approximation accuracy is measured by the expectation of the approximation error. In the Hilbert space,[6] denoted as $\mathbb{H}$, the expected risk functional may be expressed as

$$\begin{aligned}\mathcal{R} &= \int_{X\times Y} L(\mathbf{x}, y)\, dP(\mathbf{x}, y) \\ &= \int_{X\times Y} L(\mathbf{x}, y) p(\mathbf{x}, y)\, d\mathbf{x}\, dy, \end{aligned} \tag{3.1}$$

where $L(\mathbf{x}, y)$ represents the loss functional. A common loss function is the mean squared error defined by L_2 norm. Suppose y is given by a nonlinear function $f(\mathbf{x})$ corrupted by additive white noise independent of $\mathbf{x}$: $y = f(\mathbf{x}) + \varepsilon$, where ε is bounded and follows an unknown probability metric $\mu(\varepsilon)$ (namely, the noise ε can have various probability density models, as we discuss in section 8). In that case, $p(\mathbf{x}, y) = p(\mathbf{x})p(y \mid \mathbf{x})$, and the conditional probability density $p(y \mid \mathbf{x})$ is represented by the metric function $\mu[y - f(\mathbf{x})]$. In particular, the expected risk functional with the L_2 norm is given by

$$\mathcal{R} = \int_{X\times Y} [y - f(\mathbf{x})]^2 p(\mathbf{x}, y)\, d\mathbf{x}\, dy. \tag{3.2}$$

In practice, the joint probability $p(\mathbf{x}, y)$ is unknown, and an estimate of $\mathcal{R}$ based on finite (say, ℓ) observations is used instead, with an empirical risk functional

$$\mathcal{R}_{emp} = \sum_{i=1}^{\ell} [y_i - f(\mathbf{x}_i)]^2, \tag{3.3}$$

[6] Hilbert space is defined as an inner product space that is complete in the norm induced by the inner product. The common measure space where the Lebesque square-integrable functions are defined is a special case of Hilbert space: L_2 space.

which introduces an estimate $\hat{y}(\mathbf{x})$ ($\hat{y} = y - \varepsilon = f(\mathbf{x})$). Quantitatively, suppose $A \leq L(\mathbf{x}, y) \leq B$ and $0 \leq \eta \leq 1$; for the loss taking the value η, the generalization error has the upper bound (Vapnik, 1995):

$$\mathcal{R} \leq \mathcal{R}_{emp} + (B - A)\sqrt{\frac{d_{VC}\log(1 + 2\ell/d_{VC}) - \log(\eta/4)}{\ell}} \tag{3.4}$$

with the probability $1 - \eta$, where d_{VC} is a nonnegative integer called the VC dimension, which is a capacity metric of the learning machine. The second term on the right-hand side of equation 3.4 determines the VC confidence.[7]

3.2 Tikhonov regularization. In the theory of regularization, the expected risk is decomposed into two parts—the empirical risk functional $\mathcal{R}_{emp}[f]$ and the regularizer risk functional $\mathcal{R}_{reg}[f]$:

$$\begin{aligned}\mathcal{R}[f] &= \mathcal{R}_{emp}[f] + \lambda\mathcal{R}_{reg}[f] \\ &= \frac{1}{2}\sum_{i=1}^{\ell}[y_i - f(\mathbf{x}_i)]^2 + \frac{1}{2}\lambda\|\mathbf{D}f\|^2,\end{aligned} \tag{3.5}$$

where $\|\cdot\|$ is the norm operator.[8] λ is a regularization parameter that controls the trade-off between the identity (goodness of fit) of data and the roughness of the solution. $\mathbf{D}$ is a linear differential operator, which is defined as the Fréchet differential of Tikhonov functional (Tikhonov & Arsenin, 1977; Haykin, 1999). Geometrically, $\mathbf{D}$ is interpreted as a local linear approximation of the curve (or manifold) in high-dimensional space. The smoothness prior implicated in $\mathbf{D}$ makes the solution stable and insensitive to noise.

Definition 2, Fréchet differential (Balakrishnan, 1976). *A function f mapping X to Y is said to be Fréchet differentiable at a point x, if for every h in X,*

$$\lim_{\epsilon \to 0} \frac{f(x + \epsilon h) - f(x)}{\epsilon} = df(x, h)$$

exists, and defines a linear bounded transformation (in h) mapping X into Y. $df(x, h) = F(x)h$ is the Fréchet differential; $F(x)$ is the Fréchet derivative.

[7] See Niyogi and Girosi (1996, 1999) for detailed discussions on the approximation error (bias) and estimation error (variance).

[8] It is usually referred to the L_2 norm in the Hilbert space unless stated otherwise. Also note that it can be defined in a particular form in the RKHS.

Since the Fréchet differential is regarded as the best local linear approximation of a functional, we have

$$d\mathcal{R}(f,h) = \frac{d}{d\beta}\mathcal{R}(f+\beta h)\Big|_{\beta=0}, \tag{3.6}$$

where $h(\mathbf{x})$ is a constant fixed function of $\mathbf{x}$. By using the Riesz representation Theorem (Yosida, 1978; Debnath & Mikusinski, 1999) and following the steps in Haykin (1999), we have

$$\begin{aligned} d\mathcal{R}_{emp}(f,h) &= \frac{d}{d\beta}\mathcal{R}_{emp}(f+\beta h)\Big|_{\beta=0} \\ &= -\sum_{i=1}^{\ell}[y_i - f(\mathbf{x}_i)]h(\mathbf{x}_i) \\ &= -\left\langle h, \sum_{i=1}^{\ell}(y_i - f(\mathbf{x}_i))\delta(\mathbf{x}-\mathbf{x}_i)\right\rangle, \end{aligned} \tag{3.7}$$

where $\delta(\cdot)$ is the Dirac delta function and $\langle\cdot,\cdot\rangle$ denotes the inner product of two functionals in the Hilbert space $\mathbb{H}$. Similarly, the Fréchet differential of the regularizing term $\mathcal{R}_{reg}$ is written as

$$\begin{aligned} d\mathcal{R}_{reg}(f,h) &= \frac{d}{d\beta}\mathcal{R}_{reg}(f+\beta h)\Big|_{\beta=0} \\ &= \int \mathbf{D}[f+\beta h]\mathbf{D}h\,d\mathbf{x}\Big|_{\beta=0} \\ &= \int \mathbf{D}f\mathbf{D}h\,d\mathbf{x} = \langle \mathbf{D}h, \mathbf{D}f\rangle. \end{aligned} \tag{3.8}$$

The above results are well known in the spatial domain. See Haykin (1999) for details.

3.3 Operator Theory and Green's Function. Instead of presenting a rigorous mathematical treatment on operator theory, we briefly introduce the basic concepts and theorems that are relevant to our purposes and sketch a picture of differential and integral operators and their associated Green's (kernel) functions. More information can be found in books on functional analysis (Balakrishnan, 1976; Yosida, 1978; Kress, 1989; Debnath & Mikusinski, 1999).

3.3.1 Operator. Roughly speaking, an operator is a kind of correspondence relating one function to another in terms of a differential or integral equation (and, thus, the differential or integral operator arises from

the correspondence). An operator can be linear or nonlinear, finite dimensional or infinite dimensional. We discuss only linear operators here. Given a bounded linear operator $A\colon X \to Y$, its adjoint operator is defined as $\tilde{A}\colon Y \to X$, that is, $\langle Ax, y\rangle = \langle x, \tilde{A}y\rangle$ $(x \in X, y \in Y)$ and $\|A\| = \|\tilde{A}\|$. It is called adjoint because it describes the inverse correspondence between two functions in the operator, with exchanging range and domain. The role of the adjoint operator is similar to the (conjugate) transpose of a (complex) matrix. An operator is a self-adjoint operator if it is equal to its adjoint operator, in other words, $\langle Ax, x'\rangle = \langle x', \tilde{A}x\rangle$ $(x, x' \in X)$. Analogous to matrix computation, we can imagine functionals as infinite-dimensional vectors in the sense of generalized functions, and linear operators can be viewed as infinite-dimensional matrices. Solving the linear differential (integral) operator is in essence solving a differential (integral) equation. Many concepts familiar in linear algebra and matrix theory (e.g., norm, determinant, condition number, trace) can be extended to operator theory. The spectral theory is used to define, in operator language, the properties similar to eigenvectors and eigenvalues in matrix computation. Given a differential or integral operator, we may define the eigenvectors as follows: Find the value ξ for which there exists a nonzero function f satisfying the equation $Af = \xi f$, where the vector function f satisfying this equation is called the eigenvector of the operator A and ξ is the corresponding eigenvalue.

Definition 3. *Given a positive-definite kernel function K, an integral operator $\mathbf{T}$ is defined by $\mathbf{T}f = f(s) = \int K(s,x)f(x)\,dx$; its adjoint operator $\tilde{\mathbf{T}}$ is defined by $\tilde{\mathbf{T}}f = f(x) = \int K(x,s)f(s)\,ds$. $\mathbf{T}$ is self-adjoint if and only if $K(s,x) = \overline{K(x,s)}$ for all s and x, where the overline denotes complex conjugate.*

Remarks.

- An integral operator with a symmetric kernel $K(s,x)$ is self-adjoint. Note that definition 3 is valid only for homogeneous integral equations.[9] The integral equation given in the definition 3 is sometimes called a Fredholm integral equation of the first kind (Kress, 1989).
- From definition 3, one may have

$$\tilde{\mathbf{T}}\mathbf{T}f = g\colon\ g(s) = \int K'(s,x)f(x)\,dx \tag{3.9}$$

$$K'(s,x) = \int K(s,t)\overline{K(x,t)}\,dt. \tag{3.10}$$

[9] For the inhomogeneous integral equation, we have $f(s) = \int K(s,x)f(x)\,dx + h(s)$, where $h(s)$ is an even function. It is also called a Fredholm integral equation of the second kind.

Denoting $\mathbf{K} = \tilde{\mathbf{T}}\mathbf{T}$, we know, by definition, that $\mathbf{K}$ is still an integral operator with the associated kernel function K', and $\mathbf{K}$ is also a self-adjoint operator; likewise, $\mathbf{L} = \tilde{\mathbf{D}}\mathbf{D}$ is still a differential operator.[10]

- The integral operator essentially calculates $\mathbb{E}[f(x)]$ (where $\mathbb{E}[\cdot]$) represents mathematical expectation), provided $K(s, x)$ takes the form of a Backus-Gilbert averaging kernel: $K(s, x) = \sum_{i=1}^{m} s_i(x)k_i(s)$, where $k_i(\cdot)$ is a known smooth function (O'Sullivan, 1986). In particular, $K(s, x)$ can be a covariance kernel if $\mathbf{T}$ is *nuclear* (Balakrishnan, 1976).

3.3.2 Norm. The norm operator essentially determines the smoothness of a functional, depending on the functional space where the inner product is defined. For example, a general norm in the Hilbert space $\mathbb{H}^m$ is defined by

$$\begin{aligned}\|f\|_{\mathbb{H}^m} &= \left[\int_{\mathbb{R}} \left(|f|^2 + \left|\frac{\partial f}{\partial x}\right|^2 + \cdots + \left|\frac{\partial^m f}{\partial x^m}\right|^2\right) dx\right]^{1/2} \\ &= \left[\sum_{k=0}^{m} \left\|\frac{\partial^k f}{\partial x^k}\right\|_2^2\right]^{1/2};\end{aligned}$$

when $m = 0$, it reduces to the L_2 norm. If the norm operator is defined in the Sobolev space $\mathbb{W}_p^m$ (Adams, 1975),[11] the norm is given by

$$\|f\|_{\mathbb{W}_p^m} = \left[\int_{\mathbb{R}} \left(|f|^p + \left|\frac{\partial f}{\partial x}\right|^p + \cdots + \left|\frac{\partial^m f}{\partial x^m}\right|^p\right) dx\right]^{1/p}. \tag{3.11}$$

In other words, for $0 \le p \le \infty$, $m = \{0, 1, \ldots\}$, the functional f is said to belong to Sobolev space $\mathbb{W}_p^m$ if it is m-times differentiable with an associated norm $\|f\|_{\mathbb{W}_p^m} = \|f\|_p + \|f^{(m)}\|_p$. Based on different norm operators, different smoothness functionals can be defined. For example, the smoothing splines arise from the case of $m = 2$, $p = 2$ (Wahba, 1990). In the RKHS (Aronsajn, 1950), the Hilbert norm is defined by[12]

$$\|f\|_{\mathbb{H}^m} \equiv \|f\|_K = \int_{\mathbb{R}^N} \frac{|\mathcal{F}(\mathbf{s})|^2}{\mathcal{K}(\mathbf{s})} d\mathbf{s}, \tag{3.12}$$

where K is the (unique) kernel function associated to the RKHS, and $\mathcal{K}$ is the Fourier transform of K. For this reason, RKHS is sometimes called a proper functional Hilbert space.

[10] It can be understood by observing $\mathbf{L}(u, v) = \tilde{\mathbf{D}}(\mathbf{D}u, v) = \tilde{\mathbf{D}}(u, \mathbf{D}v)$ and $\mathbf{K}(u, v) = \tilde{\mathbf{T}}(\mathbf{T}u, v) = \tilde{\mathbf{T}}(u, \mathbf{T}v)$.

[11] Hilbert space is a special case of Sobolev space, $\mathbb{H}^m = \mathbb{W}_2^m$, and L_2 space is $\mathbb{W}_2^0$.

[12] For a discussion of iterative evaluation of the RKHS norm, see Kailath (1971).

3.3.3 Green's Function. It is well known in functional analysis (Yosida, 1978) that given a positive integral operator $\mathbf{T}$, we can always find a (pseudo-)differential operator,[13] $\mathbf{D}$, as its inverse. The operator $\mathbf{D}$ corresponds to the inner product of the RKHS with a reproducing kernel K associated to $\mathbf{T}$, where the kernel K is called Green's function of the differential operator $\mathbf{D}$. If the operator $\mathbf{D}$ is a certain form of pseudodifferential operator, the functional space induced by the kernel[14] is a Sobolev space (Adams, 1975).

In analogy to the inverse of a matrix, Green's function represents the inverse of a (sufficiently regular) linear differential operator (Lanczos, 1961). For a large class of problems, it appears in the form of a kernel function that depends on the position of two points in the given domain. Green's function can be defined as the solution of a certain differential equation that has the Dirac delta function on the driving force. The reciprocity theorem makes it possible to define the Green's function in terms of either a differential operator or its adjoint operator (see Lanczos, 1961, chap. 5 for details). Mathematically, we have the following definition:

Definition 4 (Courant & Hilbert, 1970). *Given a linear differential operator* $\mathbf{L}$, *the function* $G(\mathbf{x}, \boldsymbol{\xi})$ *is said to be the Green's function for* $\mathbf{L}$ *if it has the following properties:*

- *For a fixed* $\boldsymbol{\xi}$, $G(\mathbf{x}, \boldsymbol{\xi})$ *is a function of* $\mathbf{x}$ *and satisfies the given boundary conditions.*
- *Except at the point* $\mathbf{x} = \boldsymbol{\xi}$, *the derivatives of* $G(\mathbf{x}, \boldsymbol{\xi})$ *with respect to* $\mathbf{x}$ *are all continuous; the number of derivatives is determined by the order of operator* $\mathbf{L}$.
- *With* $G(\mathbf{x}, \boldsymbol{\xi})$ *considered as a function of* $\mathbf{x}$, *it satisfies the partial differential equation* $\mathbf{L}G(\mathbf{x}, \boldsymbol{\xi}) = \delta(\mathbf{x} - \boldsymbol{\xi})$.

Denoting $\varphi(\mathbf{x})$ as a continuous function of $\mathbf{x} \in \mathbb{R}^N$, it follows that the function

$$F(\mathbf{x}) = \int_{\mathbb{R}^N} G(\mathbf{x}, \boldsymbol{\xi})\varphi(\boldsymbol{\xi})\, d\boldsymbol{\xi} \tag{3.13}$$

is the solution of the differential equation

$$\mathbf{L}F(\mathbf{x}) = \varphi(\mathbf{x}), \tag{3.14}$$

[13] The pseudodifferential operator differs from the differential operator in that it may contain an infinite sum of differential operators and its Fourier transform is not necessarily a polynomial.

[14] There exists a one-to-one correspondence between the operator and the kernel according to Schwartz's kernel theorem.

where $G(\mathbf{x}, \boldsymbol{\xi})$ is the Green's function for the linear differential operator $\mathbf{L}$. The proof can be found in Courant and Hilbert (1970) and Haykin (1999).

3.4 Fourier Analysis and Spectral Regularization. Studying regularization theory in the Fourier domain dates back to Kimeldorf and Wahba (1970, 1971), Duchon (1977), Micchelli (1986), Wahba (1990), and Girosi (1992). It has been also treated fairly well in the kernel learning framework (e.g., Schölkopf & Smola, 2002). The common approach in the literature is to define the smoothness of a stabilizer in some functional space, for example, the seminorms in the Banach space (Duchon, 1977; Micchelli, 1986) or the Hilbert norm in the Sobolev space (Kimeldorf & Wahba, 1970). (See Girosi, 1992, 1993; Girosi et al., 1995, for details.)

In what follows, we establish the spectral regularization framework in the Fourier domain in a slightly different way. Starting with the definition of Fourier operator and Parseval theorem, we define the spectral operator and further discuss the regularization scheme in terms of operators and integral equations. To derive the regularization solution, we make use of operator theory and Green's function to establish the spatial and spectral regularization frameworks, with discussions relating to other relevant work. We also point out important theorems and give another proof of the regularization solution by using the Riesz representation theorem.

3.4.1 Spectral Operator. Essentially, the spectral operator is an integral operator given by definition 3:

$$\mathbf{T}f = \int_{\mathbb{R}^N} K(\mathbf{s}, \mathbf{x}) f(\mathbf{x})\, d\mathbf{x}.$$

In the case of the Fourier operator, $K(\mathbf{s}, \mathbf{x}) = \exp(-j\langle \mathbf{s}, \mathbf{x}\rangle)$ where $j \equiv \sqrt{-1}$. Formally, we have:

Definition 5. *For any functional $f(\mathbf{x}) \in \mathbb{H}$, the Fourier operator $\mathbf{T}$ is defined by $\mathbf{T}f = \mathcal{F}$: $\mathcal{F}(\mathbf{s}) = \int_{-\infty}^{+\infty} f(\mathbf{x}) \exp(-j\mathbf{x}\mathbf{s})\, d\mathbf{x}$; $\mathcal{F}(\mathbf{s}) \in \mathbb{H}$.*[15]

Theorem 1, Plancherel identity (Yosida, 1978). *Given two functionals $f, g \in \mathbb{H}$ and their corresponding Fourier transform $\mathcal{F}$ and $\mathcal{G}$, the Plancherel identity (Parseval theorem) states that $\langle f(\mathbf{x}), g(\mathbf{x})\rangle = \frac{1}{2\pi}\langle \mathcal{F}(\mathbf{s}), \mathcal{G}(\mathbf{s})\rangle$.*[16] *Written in the operator form, it is expressed by $\langle f, g\rangle = \langle \mathbf{T}f, \mathbf{T}g\rangle$.*

[15] The range and domain of Fourier operator are both in the Hilbert space.

[16] For simplicity of notation, we henceforth take all the constants that appear in the definitions of (inverse) Fourier transform to be 1. Note that $\langle \mathcal{F}(\mathbf{s}), \mathcal{G}(\mathbf{s})\rangle = \int \mathcal{F}(\mathbf{s})\overline{\mathcal{G}(\mathbf{s})}\, d\mathbf{s}$.

Remarks.

- If we define the differential operator $\mathbf{D}$ as

$$\mathbf{D} = \sum_{-\infty}^{\infty} \frac{(-1)^n}{n!} \frac{d^n}{dx^n}, \tag{3.15}$$

then its corresponding spectral operator is

$$\mathbf{T_D} = \sum_{-\infty}^{\infty} \frac{(-1)^n (js)^n}{n!} = \exp(-js). \tag{3.16}$$

where $\mathbf{T_D}$ denotes the Fourier operator functioning on a differential operator $\mathbf{D}$ (with respect to f), namely, $\mathbf{T_D} f = \mathbf{T}(\mathbf{D}f) = \mathbf{D}(\mathbf{T}f) = \mathbf{D}\mathcal{F}$.

- Recalling equation 3.10, for the Fourier operator $\mathbf{T}$, the kernel function associated with the operator $\mathbf{K}$ is given by

$$K'(\mathbf{x}, \mathbf{x}_i) = \int \exp(j\mathbf{s}\mathbf{x}) \exp(-j\mathbf{s}\mathbf{x}_i)\, d\mathbf{s} = \delta(\mathbf{x} - \mathbf{x}_i),$$

and it follows that

$$\begin{aligned} \mathbf{K}f &= \int K'(\mathbf{x}, \mathbf{x}_i) f(\mathbf{s})\, d\mathbf{s} \\ &= \int \delta(\mathbf{x} - \mathbf{x}_i) f(\mathbf{s})\, d\mathbf{s} = f(\mathbf{x} - \mathbf{x}_i). \end{aligned} \tag{3.17}$$

Example 1, Dirichlet kernel (Lanczos, 1961; Vapnik, 1995, 1998a).

$$K(\theta) = \frac{\sin(n + \frac{1}{2})\theta}{2\pi \sin\frac{\theta}{2}}. \tag{3.18}$$

From equation 3.18, the truncated Fourier series is written as

$$f_n(x) = \int_{-\pi}^{\pi} f(s) K_n(s, x)\, ds \tag{3.19}$$

where

$$\begin{aligned} K_n(s, x) &= \frac{1}{\pi} \sum_{k=0}^{n} (\cos ks \cos kx + \sin ks \sin kx) \\ &= \frac{1}{\pi} \sum_{k=0}^{n} \cos k(s - x) = K_n(s - x), \end{aligned} \tag{3.20}$$

where $K_n(s-x)$ defines an integral operator in the sense that

$$f(x) = \int_{-\pi}^{\pi} f(s)K(s-x)\,ds. \tag{3.21}$$

In particular, the Dirac delta function $\delta(s,x)$ has the series 3.19 as its Fourier expansion when $n \to \infty$ (see appendix A for the proof). Hence, the Dirichlet kernel corresponds to a truncated Fourier expansion mapping (Burges, 1999).

Example 2, Fejér kernel (Lanczos, 1961; Vapnik, 1998a).

$$\Phi_n(\theta) = \frac{\sin^2 \frac{n}{2}\theta}{2\pi n \sin^2 \frac{\theta}{2}},$$

which is the arithmetic mean of the Dirichlet kernel. In other words, the Fourier coefficients of the Fejér kernel are the weighted version of those of the Dirichlet kernel dependent on n: $a'_k = (1-k/n)a_k$, $b'_k = (1-k/n)b_k$.

Other examples, such as periodic gaussian kernel, B-spline kernel, Laplacian kernel, and regularized Fourier expansion kernel, can be found in Vapnik (1998a, 1998b), Smola, Schölkopf, & Müller (1998), and Schölkopf & Smola (2002). It is noted that the Dirichlet kernel, Fejér kernel, and B-spline kernel are interpolation kernels; translationally invariant kernels (e.g., gaussian) are convolution kernels; and polynomial kernels and Fourier kernel are dot product kernels.

3.4.2 Regularization Scheme. Consider the operator equation $A\varphi = f$ (A: $X \to Y$); the regularization scheme is to find an approximated solution φ^ϵ related to φ, such that $\|\varphi^\epsilon - \varphi\| \le \epsilon$, where ϵ is a small, positive value. Generally, it consists of finding a linear operator R_λ: $Y \to X(\lambda > 0)$ with the property of pointwise convergence $\lim_{\lambda\to 0} R_\lambda A\varphi = \varphi$ for all $\varphi \in X$, or, equivalently, $R_\lambda f \to A^{-1}f$ as $\lambda \to 0$. However, as shown in Kress (1989, theorem 15.6), for the regularization scheme, the operator R_λ cannot be uniformly bounded with respect to λ, and the operator $R_\lambda A$ cannot be norm convergent as $\lambda \to 0$. In particular, one has the approximation error,

$$\|\varphi_\lambda^\epsilon - \varphi\| \le \|R_\lambda A\varphi - \varphi\| + \epsilon\|R_\lambda\|,$$

which states that the error consists of two parts: the first term of the right-hand side is due to the approximation error between R_λ and A^{-1}, and the second term reflects the influence of incorrect data. In general, the first term decreases as $\lambda \to 0$, whereas the second term increases as $\lambda \to 0$. The regularization parameter, λ, controls the trade-off between accuracy (the first term) and stability (the second term). The regularization approach

amounts to finding an appropriate λ to achieve the minimum error of the regularized risk functional.

Definition 6 (Kress, 1989). *The regularization scheme R_λ, namely, the choice of regularization parameter $\lambda = \lambda(\epsilon)$, is called regular if for all $f \in A(X)$ and all $f^\epsilon \in Y$ with $\|f^\epsilon - f\| \leq \epsilon$, there holds $R_{\lambda(\epsilon)} f^\epsilon \to A^{-1} f$, $\epsilon \to 0$.*

By using the tools of eigendecomposition (ED) and singular value decomposition (SVD) in spectral theory, we have the following theorems:

Theorem 2 (Kress, 1989). *For a bounded linear operator, there holds*

$$A(X)^\perp = N(\tilde{A}) \quad \text{and} \quad N(\tilde{A})^\perp = \overline{A(X)},$$

where $A(X)^\perp$ means for all $\varphi \in X$ and $g \in A(X)^\perp$, $\langle A\varphi, g\rangle = 0$; $N(\tilde{A})$ denotes the null space of $\tilde{A}$, in the sense that $\tilde{A}g = 0$ for all g.

Theorem 3 (Kress, 1989). *Let X be a Hilbert space, and let $A\colon X \to X$ be a (nonzero) self-adjoint compact operator. Then all eigenvalues of A are real. All eigenspaces $N(\xi I - A)$ for nonzero eigenvalues ξ have finite dimension, and eigenspaces associated with different eigenvalues are orthogonal. Suppose the eigenvalues are ordered such that $|\xi_1| \geq |\xi_2| \geq \cdots$, and denote by $P_n\colon X \to N(\xi_n I - A)$ the orthogonal projection onto the eigenspace for the eigenvalue ξ_n; then there holds*

$$A = \sum_{n=1}^{\infty} \xi_n P_n$$

in the sense of norm convergence. Let $Q\colon X \to N(A)$ denote the orthogonal projection onto the null space $N(A)$; then there holds

$$\varphi = \sum_{n=1}^{\infty} P_n \varphi + Q\varphi$$

for all $\varphi \in X$.

In the case of an orthonormal basis, $\langle \varphi_n, \varphi_k\rangle = \delta_{n,k}$, we have the following expansion representation:

$$A\varphi = \sum_{n=1}^{\infty} \xi_n \langle \varphi, \varphi_n\rangle \varphi_n,$$

$$\varphi = \sum_{n=1}^{\infty} \langle \varphi, \varphi_n\rangle \varphi_n + Q\varphi.$$

Theorem 4 (Kress, 1989). *Let X and Y be Hilbert spaces. Let A: $X \to Y$ be a linear compact operator and $\tilde{A}$: $Y \to X$ be its adjoint. Let σ_n denote the singular values of A, which are the square roots of the eigenvalues of the self-adjoint compact operator $\tilde{A}A$: $X \to X$. Let $\{\sigma_n\}$ be an ordered sequence of nonzero singular values of A according to the dimension of the null spaces $N(\sigma_n^2 I - \tilde{A}A)$. Then there exist orthonormal sequences $\{\varphi_n\}$ in X and $\{g_n\}$ in Y such that*

$$A\varphi_n = \sigma_n g_n, \quad \tilde{A}g_n = \sigma_n \varphi_n$$

for all integers n. For each $\varphi \in X$, there holds the SVD

$$\varphi = \sum_{n=1}^{\infty} \langle \varphi, \varphi_n \rangle \varphi_n + Q\varphi$$

with the orthogonal projection Q: $X \to N(A)$ and

$$A\varphi = \sum_{n=1}^{\infty} \sigma_n \langle \varphi, \varphi_n \rangle g_n.$$

Theorem 5, Picard theorem (Kress, 1989). *Let A: $X \to Y$ be a linear compact operator with singular system $(\sigma_n, \varphi_n, g_n)$. The Fredholm integral equation of the first kind $A\varphi = f$ is solvable if and only if $f \in N(\tilde{A})^{\perp}$ and*

$$\sum_{n=1}^{\infty} \frac{1}{\sigma_n^2} |\langle f, g_n \rangle|^2 < \infty.$$

Then a solution is given by

$$\varphi = \sum_{n=1}^{\infty} \frac{1}{\sigma_n} \langle f, g_n \rangle \varphi_n.$$

The Picard theorem essentially describes the ill-posed nature of the integral equation $A\varphi = f$. The perturbation ratio $\|\varphi^{\epsilon}\| / \|f^{\epsilon}\| = \epsilon / \sigma_n$ determines the degree of ill posedness, the more quickly σ decays and the more severe the ill posedness is.

For more examples of regularization in terms of operators and Fredholm integral equations of the first kind, see Kress (1989).

3.4.3 Regularization Solution. By virtue of theorem 1, it follows that

$$\langle \mathbf{D}h, \mathbf{D}f \rangle = \langle \mathbf{T_D}h, \mathbf{T_D}f \rangle. \tag{3.22}$$

For ease of notation, we henceforth simply denote $\mathbf{T_D}$ as $\mathbf{T}$. From equation 3.22, equation 3.8 is rewritten as

$$d\mathcal{R}_{reg}(f,h) = \int \mathbf{D}f\mathbf{D}h\,d\mathbf{x} = \langle \mathbf{D}h, \mathbf{D}f\rangle$$
$$= \int \mathbf{T}f\overline{\mathbf{T}h}\,d\mathbf{s} = \int \overline{\mathbf{T}f}\mathbf{T}h\,d\mathbf{s} = \langle \mathbf{T}h, \mathbf{T}f\rangle. \tag{3.23}$$

For any pair of functions $u(\mathbf{x})$ and $v(\mathbf{x})$, given a linear differential operator $\mathbf{D}$ and its associated Fourier operator $\mathbf{T}$ (i.e., $\mathbf{T_D}$), their adjoint operators, $\tilde{\mathbf{D}}$ and $\tilde{\mathbf{T}}$, are uniquely determined to satisfy the boundary conditions

$$\int_{\mathbb{R}^N} u(\mathbf{x})\mathbf{D}v(\mathbf{x})\,d\mathbf{x} = \int_{\mathbb{R}^N} v(\mathbf{x})\tilde{\mathbf{D}}u(\mathbf{x})\,d\mathbf{x} \tag{3.24}$$

and

$$\int_{\Omega} u(\mathbf{s})\overline{\mathbf{T}v(\mathbf{s})}\,d\mathbf{s} = \int_{\Omega} v(\mathbf{s})\tilde{\mathbf{T}}u(\mathbf{s})\,d\mathbf{s}, \tag{3.25}$$

where Ω represents spectrum support in the frequency domain. Equation 3.24 is called Green's identity, which describes the bilinear identity of matrix calculus into the realm of function space (Lanczos, 1961). Equation 3.25 follows from the fact that $\langle u(\mathbf{x}), \mathbf{D}v(\mathbf{x})\rangle = \langle \tilde{\mathbf{D}}u(\mathbf{x}), v(\mathbf{x})\rangle$, $\langle u(\mathbf{x}), \mathbf{D}v(\mathbf{x})\rangle = \langle u(\mathbf{s}), \mathbf{T}v(\mathbf{s})\rangle$.

Using Green's identity and setting $u(\mathbf{x}) = \mathbf{D}f(\mathbf{x})$ and $\mathbf{D}v(\mathbf{x}) = \mathbf{D}h(\mathbf{x})$, we further obtain an equivalent form of equation 3.23 in the light of equation 3.24, as shown by

$$d\mathcal{R}_{reg}(f,h) = \int h(\mathbf{x})\tilde{\mathbf{D}}\mathbf{D}f(\mathbf{x})\,d\mathbf{x} = \langle h, \tilde{\mathbf{D}}\mathbf{D}f\rangle(\mathbf{x}). \tag{3.26}$$

On the other hand, we also obtain another form in the light of equation 3.25 by setting $u(\mathbf{s}) = \mathbf{T}f(\mathbf{s})$ and $\mathbf{T}v(\mathbf{s}) = \mathbf{T}h(\mathbf{s})$, as given by

$$d\mathcal{R}_{reg}(f,h) = \int h(\mathbf{s})\tilde{\mathbf{T}}\mathbf{T}f(\mathbf{s})\,d\mathbf{s} = \langle h, \tilde{\mathbf{T}}\mathbf{T}f\rangle(\mathbf{s}). \tag{3.27}$$

Thus, the condition that the Fréchet differential being zero

$$d\mathcal{R}(f,h) = d\mathcal{R}_{emp}(f,h) + \lambda d\mathcal{R}_{reg}(f,h) = 0$$

can be rewritten, by virtue of equations 3.8, 3.26, and 3.27, in the following form,

$$d\mathcal{R}(f,h) = \left\langle h(\mathbf{x}), \left[\tilde{\mathbf{D}}\mathbf{D}f(\mathbf{x}) - \frac{1}{\lambda}\sum_{i}^{\ell}(y_i - f)\delta(\mathbf{x}-\mathbf{x}_i)\right]\right\rangle, \tag{3.28}$$

or

$$d\mathcal{R}(f,h) = \left\langle h(\mathbf{s}), \left[\tilde{\mathbf{T}}\tilde{\mathbf{T}} f(\mathbf{s}) - \mathcal{F}\left\{ \frac{1}{\lambda} \sum_{i}^{\ell} (y_i - f)\delta(\mathbf{x} - \mathbf{x}_i) \right\} \right] \right\rangle, \tag{3.29}$$

where $\mathcal{F}\{\cdot\}$ denotes taking the Fourier transform. The necessary condition for $f(\mathbf{x})$ being an extremum of $\mathcal{R}[f]$ is that $d\mathcal{R}(f,h) = 0$ for all $h \in \mathbb{H}$, or, equivalently, the following conditions are satisfied in the distribution sense:

$$\tilde{\mathbf{D}}\mathbf{D} f_\lambda(\mathbf{x}) = \frac{1}{\lambda} \sum_{i=1}^{\ell} [y_i - f(\mathbf{x}_i)]\delta(\mathbf{x} - \mathbf{x}_i) \tag{3.30}$$

and

$$\begin{aligned} \tilde{\mathbf{T}}\tilde{\mathbf{T}} f_\lambda(\mathbf{s}) &= \mathcal{F}\left\{ \frac{1}{\lambda} \sum_{i=1}^{\ell} [y_i - f(\mathbf{x}_i)]\delta(\mathbf{x} - \mathbf{x}_i) \right\} \\ &= \frac{1}{\lambda} \sum_{i=1}^{\ell} [y_i - f(\mathbf{x}_i)] \exp(-j\mathbf{x}_i\mathbf{s}). \end{aligned} \tag{3.31}$$

Equation 3.30 is the Euler-Lagrange equation of the Tikhonov functional $\mathcal{R}[f]$, and equation 3.31 is its Fourier counterpart. Denoting $\mathbf{L} = \tilde{\mathbf{D}}\mathbf{D}$ and $\mathbf{K} = \tilde{\mathbf{T}}\tilde{\mathbf{T}}$ as earlier, $G(\mathbf{x}, \boldsymbol{\xi})$ is the Green's function for the linear differential operator $\mathbf{L}$. Recalling definition 4,

$$\mathbf{L}G(\mathbf{x}, \boldsymbol{\xi}) = \delta(\mathbf{x} - \boldsymbol{\xi}), \tag{3.32}$$

we may derive its counterpart in the frequency domain

$$\mathbf{K}G(\mathbf{s}, \boldsymbol{\xi}) = \exp(-j\mathbf{s}\boldsymbol{\xi}). \tag{3.33}$$

Recalling equation 3.13, it follows that

$$f(\mathbf{x}) = \int_{\mathbb{R}^N} G(\mathbf{x}, \boldsymbol{\xi})\varphi(\boldsymbol{\xi})\, d\boldsymbol{\xi} \tag{3.34}$$

is the solution of the following differential equation and integral equation,

$$\mathbf{L}f(\mathbf{x}) = \varphi(\mathbf{x}), \tag{3.35}$$

$$\mathbf{K}f(\mathbf{s}) = \Phi(\mathbf{s}), \tag{3.36}$$

where $\Phi(\mathbf{s})$ is the Fourier transform of $\varphi(\mathbf{x})$. In the light of equations 3.32 through 3.34, we may derive the solution of the regularization problem as

$$f_\lambda(\mathbf{x}) = \sum_{i=1}^{\ell} w_i G(\mathbf{x}, \mathbf{x}_i), \tag{3.37}$$

where $w_i = [y_i - f(\mathbf{x}_i)]/\lambda$, and $G(\mathbf{x}, \mathbf{x}_i)$ is a positive-definite Green's function for all i (see appendix B for an outline of the proof).

Remarks.

- Provided the operator $\mathbf{L} = \tilde{\mathbf{D}}\mathbf{D}$ is defined by the Laplace operator (Yuille & Grzywacz, 1988; Poggio & Girosi, 1990a; Haykin, 1999),[17]

$$\mathbf{L} = \sum_{n=0}^{\infty} \frac{(-1)^n}{n!2^n} \nabla^{2n}, \tag{3.38}$$

  where

$$\nabla^2 = \sum_{i=1}^{\ell} \sum_{j=1}^{\ell} \frac{\partial^2}{\partial x_i \partial x_j},$$

  then the corresponding operator $\mathbf{K} = \tilde{\mathbf{T}}\mathbf{T}$ in the spectral domain reads

$$\mathbf{K} = \sum_{n=0}^{\infty} \frac{(-1)^{2n} s^{2n}}{n!2^n} = \exp\left(\frac{s^2}{2}\right). \tag{3.39}$$

  By noting that

$$\mathbf{L}G(\mathbf{x}) = \delta(\mathbf{x}), \tag{3.40}$$

$$\mathbf{K}G(\mathbf{s}) = 1, \tag{3.41}$$

  from the property of Green's function, it further follows that

$$\mathcal{G}(\mathbf{s}) = \exp\left(-\frac{\|\mathbf{s}\|^2}{2}\right), \tag{3.42}$$

$$G(\mathbf{x}) = \exp\left(-\frac{\|\mathbf{x}\|^2}{2}\right), \tag{3.43}$$

  where $G(\mathbf{x}) \leftrightarrow \mathcal{G}(\mathbf{s})$ is a Fourier transform pair.

- Note that the solution to the Tikhonov regularization problem, given by equation 3.37, is incomplete in the sense that it only represents the solution modulo a term that lies in the null space of the operator $\mathbf{D}$ (Poggio & Girosi, 1990a). In general, the solution to the Tikhonov

[17] This can be also regarded as an RKHS norm of the Lévy's n-dimensional ($n \to \infty$) Brownian motion (Kailath, 1971).

regularization problem is given by (Kimeldorf & Wahba, 1970; Poggio & Girosi, 1990a; Girosi et al., 1995)

$$f_\lambda(\mathbf{x}) = \sum_{i=1}^{\ell} w_i G(\mathbf{x}, \mathbf{x}_i) + \beta(\mathbf{x}), \tag{3.44}$$

where $\beta(\mathbf{x})$ is a term that lies in the null space of $\mathbf{D}$ (the simplest case is $\beta(\mathbf{x}) = const.$),[18] which satisfies the orthogonal condition $\sum_{i=1}^{\ell} w_i \beta(\mathbf{x}_i) = 0$. Hence, the functional space of the solution f_λ is an RKHS of the direct sum of two orthogonal RKHS.[19] Equation 3.44 can be understood using an analogy of solving a matrix equation $Ax = b$. The general solution of the equation is given by $x = (A^\dagger + Z)b = ((A^T A)^{-1} A^T + Z)b$ († represents the Moore-Penrose pseudoinverse), where Z accounts for the orthogonal (null) space where $A^\dagger$ lies. Hence, the regularization solution 3.37 is somehow similar to the minimum-norm solution of the matrix equation, as given by $x = A^\dagger b = (A^T A)^{-1} A^T b$.

- The essence of regularization is to find a proper subspace, namely, the eigenspace of $\mathbf{L}f$ or $\mathbf{K}f$, within which the operator behaves like a "well-posed" operator (Lanczos, 1961). The solution in the subspace is unique.
- Another viewpoint to look at the regularization solution is the following: Rewriting equation 3.5 in a matrix form,

$$\begin{aligned} \mathcal{R}[f] &= \frac{1}{2}\|\mathbf{y} - \mathbf{f}\|^2 + \frac{1}{2}\lambda(\mathbf{D}f)^T(\mathbf{D}f) \\ &= \frac{1}{2}(\mathbf{y} - \mathbf{f})^T(\mathbf{y} - \mathbf{f}) + \frac{1}{2}\lambda\tilde{\mathbf{D}}\mathbf{D}f \\ &= \frac{1}{2}(\mathbf{y} - \mathbf{f})^T(\mathbf{y} - \mathbf{f}) + \frac{1}{2}\lambda\mathbf{f}^T K\mathbf{f}, \end{aligned} \tag{3.45}$$

and taking the derivative of of risk function with respect to $\mathbf{f}$ and setting it to zero, we have $\mathbf{f} = (\mathbf{I} + \lambda K)^{-1}\mathbf{y}$, which gives rise to a smoothing matrix[20] $\mathbf{S} = (\mathbf{I} + \lambda K)^{-1}$, where K is a quadratic symmetric penalty matrix associated with $\mathbf{L}$ (Hastie, 1996). By taking the ED of $\mathbf{S}$, say $\mathbf{S} = \mathbf{U}_{\mathbf{\Sigma}}\mathbf{U}^T$, $\mathbf{U}^T\mathbf{y}$ expresses $\mathbf{y}$ in terms of the orthonormal basis defined by the column vectors of $\mathbf{U}$, which is quite similar to the Fourier expansion (Hastie, 1996).

[18] We notice that for $\beta(\mathbf{x})$, $\mathbf{D}\beta(\mathbf{x}) = 0$ and $\mathbf{T}_\mathbf{D}\beta(\mathbf{s}) = 0$.

[19] For a good mathematical treatment on the RKHS in the context of regularization theory, see Wahba (1990, 1999), Corradi and White (1995), Girosi (1998), Vapnik (1998a), and Evgeniou, Pontil, and Poggio (2000).

[20] When $\lambda = 0$, the smoothing matrix reduces to an identity matrix, and the learning problem reduces to a pure interpolation problem in the noiseless situation.

- Observing equation 3.37, one may replace $\sum$ by $\int$, $G(\mathbf{x}, \mathbf{x}_i)$ by $K(\mathbf{x}, \mathbf{x}_i)$, and w_i by $[y_i - f(\mathbf{x}_i)]/\lambda$. Suppose that K is a translationally invariant-reproducing kernel that is positive definite and satisfies $K(\mathbf{x}, \mathbf{x}_i) = K(\mathbf{x} - \mathbf{x}_i)$. Then the approximated function can be expressed by the convolution between the observation and a kernel:

$$\begin{aligned} f(\mathbf{x}) &= \frac{1}{\lambda} \int (y_i - f(\mathbf{x}_i)) K(\mathbf{x} - \mathbf{x}_i)\, d\mathbf{x}_i \\ &= -\frac{1}{\lambda} \int f(\mathbf{x}_i) K(\mathbf{x} - \mathbf{x}_i)\, d\mathbf{x}_i + \frac{1}{\lambda} \int y_i K(\mathbf{x} - \mathbf{x}_i)\, d\mathbf{x}_i, \end{aligned}$$

  where the first term has exactly the same form as the integral operator, and the second term is similar to kernel regression. Intuitively, $f(\mathbf{x})$ can be reconstructed by the data sample smoothed by an averaged kernel, which accounts for a convolutional window (the so-called Parzen window).

- In RBF (regularization) networks, the number of RBF (i.e., Green's function) units, m, is exactly equal to the observation number ℓ. In practice, we may choose $m < \ell$, which corresponds to the so-called hyperbasis function networks or generalized RBF networks. From the matrix equation viewpoint, there are ℓ equations associated with ℓ observation pairs but m unknowns; hence, it is an overdetermined situation (Golub & Van Loan, 1996). In order to find the minimum norm solution, we might use SVD to obtain $\hat{\mathbf{f}} = \sum_{i=1}^{m} \frac{(\mathbf{v}_i^T \mathbf{G}^T \mathbf{y})}{\sigma_i^2} \mathbf{v}_i$ (assuming the rank of $\mathbf{G}$ is m), where $\mathbf{v}_i$ are the vectors of the unitary matrix $\mathbf{V}$ and σ_i are the singular values. For more discussion on their optimization, see Moody & Darken (1989), Girosi (1992), and Haykin (1999).

The preceding derivation states that the solution of the classical Tikhonov regularization is independent of the domain where the regularizer (stabilizer) is defined, and the regularization is equivalent to the expansion of the solution in terms of a linear superposition of ℓ Green's functions centered at the specific observation. Note that some invariance properties are implicitly embedded in $\mathbf{T}$,[21] which implies that the derived Green's function should be translationally and rotationally invariant. In other words, $G(\mathbf{x}, \mathbf{x}_i)$ is inherently an RBF with the radially symmetric and shift-invariant form (Poggio & Girosi, 1990a),

$$G(\mathbf{x}, \mathbf{x}_i) = G(|\mathbf{x} - \mathbf{x}_i|). \tag{3.46}$$

More generally, it can be the hyperbasis function (HyperBF) (Poggio & Girosi, 1990a) with the form

$$G(\mathbf{x}, \mathbf{x}_i) = G(|\mathbf{x} - \mathbf{x}_i|^T \Sigma^{-1} |\mathbf{x} - \mathbf{x}_i|), \tag{3.47}$$

[21] Fourier transform is invariant to shift, rotation, and starting point.

where Σ is a positive-definite covariance matrix; or it can be the RBF with the weighted norm (so-called Mahalanobis distance metric),

$$G(\mathbf{x}, \mathbf{x}_i) = G(\|\mathbf{x} - \mathbf{x}_i\|_{\mathbf{Q}}) = G(|\mathbf{x} - \mathbf{x}_i|^T \mathbf{Q}^T \mathbf{Q} |\mathbf{x} - \mathbf{x}_i|). \tag{3.48}$$

3.4.4 Spectral Regularization. Thus far, we have been able to establish a functionally equivalent form of spectral regularizer, being the counterpart of equation 3.5,

$$\begin{aligned}\mathcal{R}[f] &= \mathcal{R}_{emp}[f] + \lambda \mathcal{R}_{reg}[\mathcal{F}] \\ &= \frac{1}{2}\sum_{i=1}^{\ell}[y_i - f(\mathbf{x}_i)]^2 + \frac{1}{2}\lambda \|\mathbf{D}\mathcal{F}\|^2 \\ &= \frac{1}{2}\sum_{i=1}^{\ell}[y_i - f(\mathbf{x}_i)]^2 + \frac{1}{2}\lambda \|\mathbf{T}f\|^2. \end{aligned} \tag{3.49}$$

Written in a matrix form, the spectral regularizer is given by

$$\|\mathbf{T}f\|^2 = (\mathbf{T}f)^T(\mathbf{T}f) = \tilde{\mathbf{T}}\mathbf{T}f = \mathbf{f}^T K \mathbf{f},$$

which again relates to the same smoothing matrix as equation 3.45.

By using a property of the Fourier transform, when $f(\mathbf{x}) \leftrightarrow \mathcal{F}(\mathbf{s})$, one obtains $\frac{\partial^m f(\mathbf{x})}{\partial \mathbf{x}^m} \leftrightarrow (j\mathbf{s})^m \mathcal{F}(\mathbf{s})$. Intuitively, we may think of the differential operator $\|\mathbf{D}f\|$ as taking m-order derivative in the Hilbert space $\mathbb{H}^m$; thus, the corresponding $\|\mathbf{T}f\|^2$ has the form of seminorms (Micchelli, 1986):

$$\|\mathbf{T}f\|^2 = \int |\mathcal{F}(\mathbf{s})|^2 \times \|\mathbf{s}\|^{2m} d\mathbf{s}. \tag{3.50}$$

If the norm operator is defined in the RKHS associated with a kernel K and its Fourier transform $\mathcal{K}(\mathbf{s}) = \|\mathbf{s}\|^{-2m}$,[22] the seminorm reduces to the form

$$\|\mathbf{T}f\|^2 = \int_{\mathbb{R}^N} \frac{|\mathcal{F}(\mathbf{s})|^2}{\mathcal{K}(\mathbf{s})} d\mathbf{s}, \tag{3.51}$$

which defines a semi-Hilbert space (Micchelli, 1986). Equation 3.51 has the same form as the spectral penalty stabilizer given in Wahba (1990) and Girosi et al. (1995). $\mathcal{K}(\mathbf{s})$ can be viewed as a low-pass filter, which plays a role of smoothing in the sense that the solution f can be viewed as the convolution of the observation with a smoothing filter,

$$f(\mathbf{x}) = y(\mathbf{x}) * B(\mathbf{x}), \tag{3.52}$$

[22] The smoothness order m has to be large enough to guarantee K to be continuous and decay fast to zero as $s \to \infty$, for example, $m > N/2$.

where $*$ denotes the operation of convolution product and $B(\mathbf{x})$ is the inverse Fourier transform of $\mathcal{B}(\mathbf{s})$, given by Poggio et al. (1988) and Girosi et al. (1995):

$$\mathcal{B}(\mathbf{s}) = \frac{\mathcal{K}(\mathbf{s})}{\lambda + \mathcal{K}(\mathbf{s})}. \tag{3.53}$$

When $\lambda = 0$, $f(\mathbf{x}) = y$, it follows that $\mathcal{B}(\mathbf{s}) = 1$, $B(\mathbf{x}) = \delta(\mathbf{x})$, which corresponds to an interpolation problem in the noiseless situation. Many choices of kernel K can be found in Girosi et al. (1995).

3.4.5 Important Theorems. We briefly mention four important theorems in the literature and indicate their relations to the theory of regularization:

- *Riesz representation theorem* (Yosida, 1978; Debnath & Mikusinski, 1999): Any bounded linear functional on a Hilbert space has a unique representer in that space. This is useful in describing the dual space to any space that contains the compactly supported continuous functions as a dense subspace. Green's function is developed from this theorem (Lanczos, 1961). The norm of the bounded linear operator is the same as the norm of its representer. Based on this theorem, we are able to derive another proof of the regularization solution (see appendix C).
- *Bochner's theorem* (Bochner, 1955; see also Wahba, 1990; Vapnik, 1998a): Among the continuous functions on $\mathbb{R}^N$, the positive definite functions are those functions that are the Fourier transforms of finite measures. Hence, the Fourier transform of a positive measure constitutes a Hilbert-Schmidt kernel.
- *Representer theorem* (Kimeldorf & Wahba, 1970; see also Wahba, 1999): Minimizing the risk functional with spectral stabilizer 3.51 with interpolation constraints is well posed, and the solution is generally given by equation 3.44. The proof was first established for the squared loss function and later extended to the general pointwise loss functions. See Schölkopf and Smola (2002, chap. 4) for many variations of this theorem.
- *Sobolev embedding theorem* (see e.g., Yosida, 1978): This describes the relation between the Hilbert space $\mathbb{H}^m$ and a class of smooth functions whose derivatives exist pointwise. It is essentially related to the generalized function with infinite differentiability in the distribution sense. See Watanabe, Namatame, and Kashiwaghi (1993), and Zhu, Williams, Rohwer, and Morciniec (1998) for discussion.

3.4.6 Interpretation. Geometrically, the smoothness of a functional is measured by the order of differentiability, tangent distance, or curvature in the spatial domain; in the frequency domain, the smoothness is seen

from its power spectral density (which is the Fourier transform of the correlation function). When the solution is smooth, the spectral components are concentrated in the low-frequency bands. In a biological point of view, regularization is essentially the expansion of a set of Green's functions taking the form of the RBF, which behaves like a lookup table mapping similar to the working mechanism of the brain. For the case of a gaussian RBF, its factorizable property lends itself to a physiologically convincing support (Poggio & Girosi, 1990a, 1990b); and the covariance matrix Σ in equation 3.47 defines the properties (e.g., shape, size, orientation) of the receptive field (Xu, Krzyzak, & Yullie, 1994; Haykin, 1999). Regularization based on the entropy principle also finds some plausible biological supports, which we detail in section 6. Finally, the regularization problem has a very nice probabilistic interpretation (Poggio & Girosi, 1990a; Girosi et al., 1995; Evgeniou et al., 2000), on which we will give detailed discussions (in sections 5 and 8) in a Bayesian framework.

3.5 Transformation Regularization: Numerical Aspects. Thus far in section 3, we have discussed regularization in the continuous domain. In practice, we are more concerned about regularization in the standpoint of numerical calculation (Hansen, 1992, 1998). This approach is another kind of transformation regularization (or subspace regularization) in the sense that regularization is taken in a subspace domain by matrix decomposition (e.g., ED, SVD, or QR decomposition). Due to the analogy between operators and matrices, we can rewrite the risk functional in a matrix form,

$$\mathcal{R} = \|\mathbf{y} - \mathbf{G}\mathbf{w}\|^2 + \lambda\|\mathbf{P}\mathbf{w}\|^2, \tag{3.54}$$

where $\mathbf{w} = [w_1, \ldots, w_\ell]^T$, $\mathbf{G} = G(\mathbf{x}, \mathbf{x}_i)$ is a radial basis matrix, and $\mathbf{y} = [y_1, \ldots, y_\ell]^T$. $\mathbf{P}$ is a user-designed matrix for the stabilizer. Since $\mathbf{G}$ is usually ill conditioned,[23] one might use matrix decomposition or factorization to alleviate this situation. Taking the SVD of $\mathbf{G}$,

$$\mathbf{G} = \mathbf{U}\Sigma\mathbf{Z}^T, \tag{3.55}$$

where $\mathbf{U}$ and $\mathbf{Z}$ are left and right singular vectors of $\mathbf{G}$, Σ is a diagonal matrix with the singular values σ_i $(i = 1, \ldots, \ell)$ in the main diagonal. In the case of zero-order Tikhonov regularization (where $\mathbf{P}$ is an identity matrix $\mathbf{I}$), suppose $\mathbf{w} = \mathbf{Z}\boldsymbol{\alpha}$, we obtain

$$(\Sigma^T\Sigma + \lambda\mathbf{I})\boldsymbol{\alpha} = \Sigma^T\mathbf{d}, \tag{3.56}$$

[23] By ill conditioned, we mean that the conditional number (defined as the ratio of the largest eigenvalue to the smallest eigenvalue) of the matrix is very large, and the matrix is rank deficient or almost rank deficient.

where $\mathbf{d} = \mathbf{U}^T\mathbf{y}$ is a vector of Fourier coefficients (Hansen, 1998).[24] Furthermore, applying the SVD to $\mathbf{P}$, suppose $\mathbf{P} = \mathbf{VSZ}^T$. Then the spectral coefficients corresponding to the zero-order and nonzero-order[25] of the Tikhonov regularization can be described, respectively, as

$$c_i = \frac{\sigma_i^2}{\sigma_i^2 + \lambda}, \quad \alpha_i = \frac{c_i d_i}{\sigma_i} \tag{3.57}$$

and

$$c_i = \frac{\sigma_i^2}{\sigma_i^2 + \lambda s_i^2}, \quad \alpha_i = \frac{c_i d_i}{\sigma_i}, \tag{3.58}$$

where s_i are defined as the diagonal elements of the diagonal matrix $\mathbf{S}$. Furthermore, in the case of nonzero-order Tikhonov regularization, provided $\mathbf{w} = \mathbf{Z}_{\boldsymbol{\alpha}}$, we obtain

$$(\boldsymbol{\Sigma}^T\boldsymbol{\Sigma} + \lambda\mathbf{S}^T\mathbf{S})_{\boldsymbol{\alpha}} = \boldsymbol{\Sigma}^T\mathbf{d}. \tag{3.59}$$

The solution to equation 3.54 when $\mathbf{P} = \mathbf{I}$ can be computed explicitly as $\mathbf{w} = (\mathbf{G} + \lambda\mathbf{I})^{\dagger}\mathbf{y}$, which has the same form as in the continuous case (see appendix C). More generally, we have the relationship

$$\mathbf{Z}^T\mathbf{w} = \boldsymbol{\Sigma}_\lambda^{\dagger}\mathbf{d} = \boldsymbol{\Sigma}_\lambda\mathbf{U}^T\mathbf{y}, \tag{3.60}$$

in which

$$\boldsymbol{\Sigma}_\lambda = \begin{cases} diag\left\{\frac{\sigma_i^2}{\sigma_i^2+\lambda}\right\}, & \mathbf{P} = \mathbf{I} \\ diag\left\{\frac{\gamma_i(\gamma_i^2+1)^{1/2}}{\gamma_i^2+\lambda}\right\}, & \mathbf{P} \neq \mathbf{I} \end{cases}, \tag{3.61}$$

where $\gamma_i = (\sigma_i^2 + s_i^2)^{1/2}$ are the generalized singular values of the $(\boldsymbol{\Sigma}, \mathbf{S})$ (see appendix D for a definition).

The discrete Picard condition (Hansen, 1998) states that a necessary condition for obtaining a good regularized solution is that the magnitude of the Fourier coefficients $|d_i|$ must decay to zero faster than the singular values σ_i. By reweighting a posteriori the generalized singular values s_i according to their contributions (through calculating the geometric mean of $|d_i|$), the

[24] Note that it coincides with the remark following equation 3.45. In a loose sense, Fourier transform can be viewed as diagonalizing the regularization operator in the continuous domain, a fact that corresponds to applying ED or SVD to the regularizer matrix in the discrete domain.

[25] For instance, provided $\mathbf{D} = \nabla$ is a first-order gradient operator, we can define $\mathbf{P} = \mathbf{J}(\mathbf{w})\mathbf{w}^{-1}$, where $\mathbf{J}(\mathbf{w})$ is a Jacobian matrix; provided $\mathbf{D} = \nabla^2$ is a second-order Laplace operator, we can define $\mathbf{P} = \mathbf{H}(\mathbf{w})\mathbf{w}^{-1}$, where $\mathbf{H}(\mathbf{w})$ is a Hessian matrix.

weight coefficients can be devised to be the reciprocal of the energy spectrum of the data (Velipasaoglu, Sun, Zhang, Berrier, & Khoury, 2000). Since the ill-conditioned matrix $\mathbf{G}$ usually has a wide range of singular values,[26] the singular values corresponding to the components with high energy and high SNR are supposed to be penalized less and those corresponding to the components with low energy and low SNR are penalized more (Velipasaoglu et al., 2000). In addition, for discussions on regularization in the context of optimization, see Dontchev and Zolezzi (1993).

3.6 Choosing the Regularization Parameter. In order to obtain a stable and convergent regularized solution, the choice of regularization parameter is very important. There are several approaches for determining the regularization parameter in practice.

According to MacKay (1992), regularization parameter λ can be estimated by a Bayesian method within the second evidence framework.[27] Given data $\mathcal{D}$ and model $\mathcal{M}$, the regularization parameter can be estimated by

$$p(\lambda \mid \mathcal{D}, \mathcal{M}) \propto p(\mathcal{D} \mid \lambda, \mathcal{M}) p(\lambda \mid \mathcal{M}). \tag{3.62}$$

The regularization parameter can be also estimated by the average squared error and generalized cross-validation (GCV) approaches (Wahba, 1990, 1995; Haykin, 1999; Yee & Haykin, 2001). The advantage of the GCV estimate over the average squared error and ordinary cross-validation approaches lies in the fact that it does not need any prior knowledge of the noise variance and treats all observation data equally in the estimate. Yee (1998) and Yee and Haykin (2001) proved that the regularized strict interpolation radial basis function network (SIRBFN) is asymptotically equivalent to the Nadaraya-Watson regression estimate with mean-square consistency, in which the optimal λ is allowed to vary with the new observations.

The optimal adaptive regularization parameter and its sufficient convergence condition was studied in Leung and Chow (1999), and it is shown that the choice of λ should be

$$\lambda \geq \frac{-\|\nabla \mathcal{R}_{emp}(\mathbf{w})\|^2}{\langle \nabla \mathcal{R}_{emp}(\mathbf{w}), \nabla \mathcal{R}_{reg}(\mathbf{w}) \rangle} \tag{3.63}$$

and

$$\lambda \geq \frac{\langle \nabla \mathcal{R}_{emp}(\mathbf{w}), \nabla \mathcal{R}_{reg}(\mathbf{w}) \rangle}{-\|\nabla \mathcal{R}_{reg}(\mathbf{w})\|^2} \tag{3.64}$$

in order to guarantee the convergence of $\mathcal{R}_{emp}(\mathbf{w})$ and $\mathcal{R}_{reg}(\mathbf{w})$, respectively.

[26] Alternatively, its singularity can be observed using QR decomposition, which is discussed in section 9.

[27] The first evidence framework is to estimate the posterior probability of the weights while supposing λ is known. See section 5 for details.

The discussion so far assumes that only a single regularization parameter is used in the regularized functional; however, this is not a strict requirement. In fact, multiple regularization parameters are sometimes necessary in practical scenarios, depending on the prior knowledge underlying the data. In general, the regularized functional can be defined by

$$\begin{aligned}\mathcal{R}_{reg}[f] &= \frac{1}{|\mathbf{S}|}\int_{\mathbb{R}^N}\left(\frac{\partial f}{\partial \mathbf{x}}\right)^T \mathbf{S}\left(\frac{\partial f}{\partial \mathbf{x}}\right) d\mathbf{x} \\ &= \frac{1}{|\mathbf{S}|}\int_{\mathbb{R}^N}\sum_k s_k\left(\frac{\partial f}{\partial x_k}\right)^2 dx_k, \end{aligned} \tag{3.65}$$

where $\mathbf{S}$ is a diagonal matrix (i.e., with radial symmetry constraint) with nonnegative elements $\mathrm{diag}\{s_1, s_2, \ldots, s_N\}$, $|\mathbf{S}|$ denotes the determinant of $\mathbf{S}$. One possible choice of s_k $(k = 1, \ldots, N)$ can be assigned to be proportional to the variances of x_k along different coordinates. See Girosi (1992) for further practical examples and discussions.

4 Occam's Razor and MDL Principle

Occam's razor is a principle that favors the shortest (simplest) hypothesis that can well explain a given set of observations. In the artificial intelligence and neural network communities, it is commonly used for complexity control in data modeling, which is essentially connected to regularization theory. MacKay (1992) provided a descriptive discussion on Occam's razor from a Bayesian perspective (see section 5 for discussion). The evidence can be viewed as the product of a best-fit likelihood and Occam's factor. In a regression problem, Occam's razor is anticipated to find the simplest smoothlike model that can well approximate or interpolate the observation data.

The minimum description length (MDL) principle is based on an information-theoretic analysis of the concepts of complexity and randomness. It was proposed as a computable approximation of Kolmogorov complexity (Rissaen, 1978; Cherkassky & Mulier, 1998). The idea of MDL is to view machine learning as a process of encoding the information carried by the observations into the model. The code length is interpreted as a characterization of the data related to the generalization ability of the code. Specifically, the observation $\mathcal{D} = \{\mathbf{x}_i, y_i\}$ is assumed to be drawn independently from an unknown distribution, and the learning problem is formulated as the dependency estimation of y on $\mathbf{x}$. A metric measuring the complexity of the data length is given by Rissaen, 1978,

$$\mathcal{R} = L(\mathcal{D} \mid \mathcal{M}) + L(\mathcal{M}), \tag{4.1}$$

where the first term measures a code length (log likelihood) of the difference between the data and the model output, and the second term measures the

code length of the model, which relates to the number of lookup table (say m) of a codebook, that is, $L(\mathcal{M}) = \lceil \log_2 m \rceil$ (Cherkassky & Mulier, 1998). The MDL principle is close in spirit to Akaike's information-theoretic criterion (Akaike, 1974).

MDL is closely related to the regularization principle (MacKay, 1992; Hinton & van Camp, 1993; Rohwer & van der Rest, 1996; Cherkassky & Mulier, 1998). Hinton and van Camp (1993) found that the regularization of weight decay and soft weight sharing are indeed the vindication of MDL. Basically, a neural network acts like an encoder-decoder. The data and weights constitute an information flow, which is transferred through the channel, that is, the hidden layers (see Figure 1 for illustration). Regularization attempts to keep the weights simple by penalizing the information they carry. The amount of information in the weights can be controlled by adding some noise with a specific density, and the noise level (i.e., regularization parameter) can be adjusted during the learning process to optimize the trade-off between the empirical error (misfit of data) and the amount of information in the weights (regularizer). Suppose the approximation error is zero-mean gaussian distributed with standard deviation σ_i and quantization width μ; one then obtains (Hinton & van Camp, 1993)

$$-\log_2 p(\varepsilon_i) \propto -\log_2 \mu + \log_2 \sigma + \frac{\varepsilon_i^2}{2\sigma^2}, \tag{4.2}$$

and the misfit of data is measured by the empirical risk,

$$\mathcal{R}_{emp} = k\ell + \frac{\ell}{2} \log_2 \left(\frac{1}{\ell} \sum_i \varepsilon_i^2 \right), \tag{4.3}$$

where k is a constant dependent on μ. Hence, minimizing the squared error (the second term of the right-hand side of 4.3) is equivalent to the MDL principle. Assuming the weights are gaussian distributed with zero-mean and standard deviation σ_w, the regularizer $L(\mathcal{M})$ of weight decay can be written by an MDL metric,

$$\mathcal{R}_{reg} = \frac{1}{2\sigma_w^2} \sum_i w_i^2, \tag{4.4}$$

in which σ_w^2 plays the role of regularization parameter. In the noisy weight case, MDL corresponds to introducing a high variance to $L(\mathcal{D} \mid \mathcal{M})$; consequently, the misfit of data is less reliable. More generally, the weights may be assumed to have a mixture density $p(\mathbf{w}) = \sum_i \pi_i p(w_i)$. For a detailed discussion, see Hinton and van Camp (1993), Bishop (1995a), and Cherkassky and Mulier (1998).[28]

[28] For a detailed discussion of MDL and the structural risk minimization (SRM) principle, as well as the shortcoming of MDL, see Vapnik (1998a).

5 Bayesian Theory

Bayesian theory is an efficient approach to deal with prior information. It lends itself naturally to a choice of the regularization operator (MacKay, 1992). This is not surprising since regularization theory has a nice Bayesian interpretation (Poggio & Girosi, 1990a; Evgeniou et al., 2000). Much effort has been spent to apply the Bayesian framework to machine learning and complexity control (MacKay, 1992; Bruntine & Weigend, 1991; Williams, 1994).

Given the data $\mathcal{D}$, what one cares about is finding out the most probable model underlying the observation data. Following the Bayes formula, the posterior probability of model $\mathcal{M}$ is estimated by $p(\mathcal{M} \mid \mathcal{D}) = p(\mathcal{D} \mid \mathcal{M})p(\mathcal{M})/p(\mathcal{D})$, where the denominator is a normalizing constant representing the evidence. Maximizing $p(\mathcal{M} \mid \mathcal{D})$ is equivalent to minimizing $-\log p(\mathcal{M} \mid \mathcal{D})$, as shown by

$$-\log p(\mathcal{M} \mid \mathcal{D}) \propto -\log p(\mathcal{D} \mid \mathcal{M}) - \log p(\mathcal{M}), \tag{5.1}$$

where the first term of the right-hand side corresponds to the MDL described in the previous section and the second term represents the minimal code length for the model $\mathcal{M}$.

Once $\mathcal{M}$ is known, one can apply Bayes theorem to estimate its parameters. The posterior probability of the weights $\mathbf{w}$, given the data $\mathcal{D}$ and model $\mathcal{M}$, is estimated by

$$p(\mathbf{w} \mid \mathcal{D}, \mathcal{M}) = \frac{p(\mathcal{D} \mid \mathbf{w}, \mathcal{M})p(\mathbf{w} \mid \mathcal{M})}{p(\mathcal{D} \mid \mathcal{M})}, \tag{5.2}$$

where $p(\mathcal{D} \mid \mathbf{w}, \mathcal{M})$ represents the likelihood and $p(\mathbf{w} \mid \mathcal{M})$ is the prior probability given the model. Assuming the training patterns are identically and independently distributed, we obtain

$$\begin{aligned} p(\mathcal{D} \mid \mathbf{w}, \mathcal{M}) &= \prod_i p(\mathbf{x}_i, y_i \mid \mathbf{w}, \mathcal{M}) \\ &= \prod_i p(y_i \mid \mathbf{x}_i, \mathbf{w}, \mathcal{M})p(\mathbf{x}_i). \end{aligned} \tag{5.3}$$

Ignoring the normalizing denominator, equation 5.2 is further rewritten as

$$p(\mathbf{w} \mid \mathcal{D}, \mathcal{M}) \propto p(\mathbf{w} \mid \mathcal{M}) \prod_i p(\mathbf{x}_i, y_i \mid \mathbf{w}, \mathcal{M}). \tag{5.4}$$

Furthermore, the prior $p(\mathbf{w} \mid \mathcal{M})$ is characterized as

$$p(\mathbf{w} \mid \mathcal{M}) = \frac{\exp(-\lambda \mathcal{R}_{reg}(\mathbf{w}))}{Z_{\mathbf{w}}(\lambda)}, \tag{5.5}$$

where $Z_{\mathbf{w}}(\lambda) = \int d\mathbf{w} \exp(-\lambda \mathcal{R}_{reg}(\mathbf{w}))$. The choices of λ and $\mathcal{R}_{reg}(\mathbf{w})$ usually depend on some assumption of $p(\mathbf{w})$. For instance, when $\mathbf{w}$ is the gaussian prior as $p(\mathbf{w}) \propto \exp(-\frac{\lambda}{2}\|\mathbf{w}\|^2)$, it leads to the maximum a posteriori (MAP) estimate:

$$\log p(\mathbf{w} \mid \mathcal{D}, \mathcal{M}) = -\frac{\lambda}{2}\|\mathbf{w}\|^2 + \sum_i \log p(\mathbf{x}_i) + \sum_i \log p(y_i \mid \mathbf{x}_i, \mathbf{w}). \quad (5.6)$$

When $p(y_i \mid \mathbf{x}_i, \mathbf{w})$ is measured by the L_2 metric under gaussian noise model assumption, equation 5.6 reads

$$-\log p(\mathbf{w} \mid \mathcal{D}, \mathcal{M}) \propto \frac{\lambda}{2}\|\mathbf{w}\|^2 + \sum_i [y_i - f(\mathbf{x}_i, \mathbf{w})]^2.$$

In particular, remarks on several popular regularization techniques are in order.

Remarks.

- Weight decay (Hinton, 1989): Weight decay is equivalent to the well-known ridge regression in statistics (Wahba, 1990), which is a version of zero-order Tikhonov regularization.
- Weight elimination (Weigend, Rumelhart, & Humberman, 1991): Weight elimination can be interpreted as the negative log-likelihood prior probability of the weights. The weights are assumed to be a mixture of uniform and gaussian-like distributions.
- Approximate smoother (Moody & Rögnvaldsson, 1997): In the case of sigmoid networks with tanh nonlinearity, by letting $\mathcal{R}_{reg} = \|\frac{\partial^k f(\mathbf{x})}{\partial \mathbf{x}^k}\|^2$, it can be shown that it reduces to some sort of approximate smoother $\sum_j v_j^2 \|\mathbf{w}_j\|^p$ when $k = 1$.

As summarized in Table 1, most regularizers correspond to weight priors with different probability density functions.[29] It is noteworthy to point out that (1) weight decay is a special case of weight elimination, (2) the role of weight elimination is similar to the Cauchy prior, and (3) in the case of $p(\mathbf{w}) \propto \cosh^{-1/\beta}(\beta \mathbf{w})$ ($p(\mathbf{w})$ approximates to the Laplace prior as $\beta \to \infty$), we have $\frac{\partial}{\partial \mathbf{w}_j} \sum_j |\mathbf{w}_j| \approx \tanh(\beta \mathbf{w}_j)$. In the sense of weight priors, the regularizer is sometimes written as $\|\mathbf{P}\mathbf{w}\|^2$ in place of $\|\mathbf{D}f\|^2$ (we discuss the functional priors later in section 8).

[29] The complexity of weights is evaluated by the negative logarithm of probability density function of the weights.

Table 1: Regularizers and Weight Priors.

Regularizer	Distribution	Comment
Constant	Uniform distribution	Uniform prior
w^2	Gaussian distribution	Weight decay
$\frac{w^2}{1+w^2}$	Uniform + gaussian	Weight elimination
$\log(1 + w^2)$	Cauchy distribution	Cauchy prior
$\|w\|$	Laplace distribution	Laplace prior
$\log(\cosh(w))$	Supergaussian distribution	Supergaussian prior

6 Shannon's Information Theory

According to information theory (Shannon, 1948), the efficiency of coding is measured by its entropy: the less the entropy, the more efficient the encoder (Cover & Thomas, 1991). Imagine a learning machine (neural network) as an encoder; the information flow is transmitted through the channel, in which the noisy information is nonlinearly filtered. A schematic illustration is shown in Figure 1. Naturally, it is anticipated that the information is encoded as efficiently as possible, that is, by using fewer look-up tables (basis functions) or fewer codes (connection weights). The entropy reduction in coding the information has been validated by neurobiological observations (Daugman, 1989). Intuitively, we may think that the minimum entropy (MinEnt) can be used as a criterion for regularization.

Actually, MinEnt regularization may find plausible theoretical support and interpretation in classification (Watanabe, 1981) and regression problems. In functional approximation, information in the data is expected to concentrate on as few hidden units as possible, which refers to sparse coding or sparse representation.[30] In other words, maximum energy concentration corresponds to minimum Shannon entropy (Coifman & Wickerhauser, 1992). On the other hand, in pattern recognition, one might anticipate that only a few hidden units correspond to a specific pattern (or specific pattern is coded by specific hidden units). In other words, the knowledge learned by the neural network is not uniformly distributed in the hidden units and its connections. Based on this principle, we may derive a MinEnt regularizer for the translationally invariant RBF network as follows.

Recalling $\mathbf{G}(\mathbf{x}_i, \mathbf{x}_j)$ is an $\ell \times \ell$ symmetric matrix, it is noted that the symmetry is not a necessary condition and our statement holds for any $\ell \times m$ ($\ell \neq m$) asymmetric matrix (which corresponds to the generalized RBF network). For the purpose of expression clarity, we denote $\mathbf{G}_i$ ($i = 1, \ldots, \ell$) as

[30] It was well studied and observed that the sensory information in the visual cortex is sparsely coded (Barlow, 1989; Atick, 1992; Olshausen & Field, 1996; Harpur & Prager, 1996).

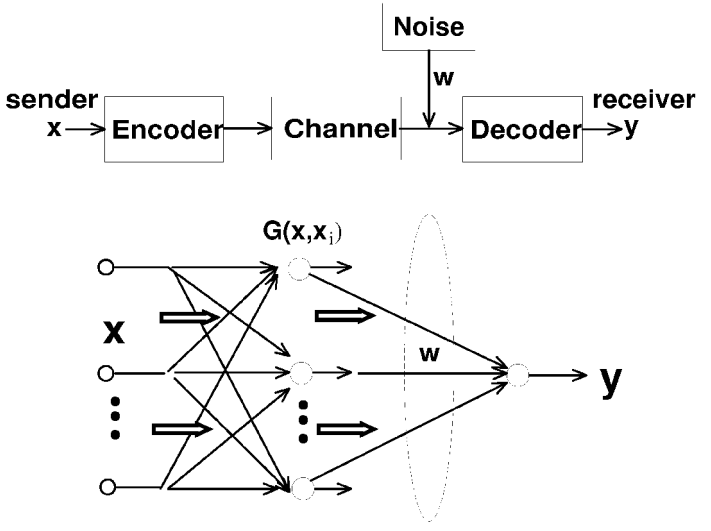

Figure 1: A schematic illustration of a neural network as an encoder-decoder.

the ith vector of matrix $\mathbf{G}$ and G_{ij} as the jth ($j = 1, \ldots, m$) component of vector $\mathbf{G}_i$. In order to measure the coding efficiency, we normalize every component G_{ij} by its row vector,[31]

$$P_{ij} = \frac{|G_{ij}|^2}{\|\mathbf{G}_i\|^2}, \tag{6.1}$$

and consequently obtain a (symmetric) probability matrix $\mathbf{P}$ with elements P_{ij}, which satisfy the relationship

$$\sum_{j=1}^{m} P_{ij} = 1. \tag{6.2}$$

Thus, the information-theoretic entropy[32] is defined by

$$H = \sum_{i=1}^{\ell} H_i = -\sum_{i=1}^{\ell} \sum_{j=1}^{m} P_{ij} \log P_{ij}. \tag{6.3}$$

[31] If the kernel function is a normalized RBF, this step is not necessary.

[32] One can, alternatively, use the generalized Renyi entropy instead of the Shannon entropy.

Specifically, when $P_{i1} = \cdots = P_{im} = \frac{1}{m}$, H_i obtains the maximum value of $\log m$, and consequently $H_{\max} = \ell \log m$. Therefore, the normalized entropy can be computed as

$$\hat{H} = \frac{H}{\ell \log m}. \tag{6.4}$$

By setting $\mathcal{R}_{reg} = \hat{H}$, we obtain an information-theoretic regularizer that relates to the simplicity principle of coding, given a fixed m.[33] It is also closely related to the maximization of collective information principle (Kamimura, 1997) and the decorrelation of the information by the nonlinear filter of hidden layer (Deco, Finnoff, & Zimmermann, 1995).

The MinEnt regularization principle is also connected to the well-studied Infomax principle in supervised learning (Kamimura, 1997) as well as in unsupervised learning (Becker, 1996). Suppose the input units are represented by A and hidden units by B; then the mutual information $I(A, B)$ between A and B can be represented by their conditional entropy (Cover & Thomas, 1991; Haykin, 1999):

$$I(A, B) = H(B) - H(B \mid A). \tag{6.5}$$

Minimizing the conditional entropy $H(B \mid A)$, the uncertainty of hidden units given the input data results equivalently in maximizing the mutual information between input and hidden layers.

7 Statistical Learning Theory

Rewriting the expected risk $\mathcal{R}$ in an explicit form, one can decompose it into two parts (O'Sullivan, 1986; Geman, Bienenstock, & Doursat, 1992),[34]

$$\begin{aligned}\mathcal{R} &= \mathbb{E}[(y - f(\mathbf{x}))^2 \mid \mathbf{x}] \\ &= \mathbb{E}[(y - \mathbb{E}[y \mid \mathbf{x}] + \mathbb{E}[y \mid \mathbf{x}] - f(\mathbf{x}))^2 | \mathbf{x}] \\ &= \mathbb{E}[(y - \mathbb{E}[y \mid \mathbf{x}])^2 | \mathbf{x}] + (\mathbb{E}[y \mid \mathbf{x}] - f(\mathbf{x}))^2,\end{aligned} \tag{7.1}$$

which is the well-known bias-variance dilemma in statistics (Geman et al., 1992; Wolpert, 1997; Breiman, 1998). The first term on the right-hand side of equation 7.1 is the bias of the approximation, and the second term measures the variance of the solution. The regularization coefficient λ in equation 3.5 or 3.49 controls the trade-off between the two terms in the error decomposition.

[33] The redundant hidden nodes can be pruned based on this principle.

[34] The error decomposition of the Kullback-Leibler divergence risk functional was discussed in Heskes (1998).

In recent decades a new statistical learning framework has been formalized in terms of structural risk minimization (SRM). Based on this paradigm, support vector machines (SVMs) have been developed for a wide class of learning problems (Vapnik, 1995, 1998a, 1998b; see also Schölkopf & Smola, 2002). SVM can be regarded as a type of regularization network with the same form of solution as equation 3.37 but trained with a different loss function, and consequently with a different solution (Girosi, 1998; Evgeniou et al., 2000). In the SVM, the ϵ-insensitive loss function is used and trained by quadratic programming; some of the weights w_i are zero, and the $\mathbf{x}_i$ corresponding to the nonzero w_i are called support vectors. The solution found by the SVM is usually sparse (Vapnik, 1995; Poggio & Girosi, 1998). By choosing a specific kernel function, the mapping from data space to feature space corresponds to a regularization operator, which explains the reason that SVMs exhibit good generalization performance in practice.

In particular, writing $f(\mathbf{x})$ in terms of a positive semidefinite kernel function (not necessarily satisfying Mercer condition),[35] we obtain

$$f(\mathbf{x}) = \sum_i \alpha_i K(\mathbf{x}_i, \mathbf{x}) + b. \tag{7.2}$$

When $K(\mathbf{x}_i, \mathbf{x}_j) = \langle \mathbf{D}K(\mathbf{x}_i, \cdot), \mathbf{D}K(\mathbf{x}_j, \cdot) \rangle$, regularization network (RN) is equivalent to the SVM (Girosi, 1998; Smola et al., 1998). Recalling the Green's function $\tilde{\mathbf{D}}\mathbf{D}G(\mathbf{x}_i, \mathbf{x}) = \delta_{\mathbf{x}_i}(\mathbf{x}) = \delta(\mathbf{x} - \mathbf{x}_i)$, minimizing the risk functional, we obtain

$$G(\mathbf{x}_i, \mathbf{x}_j) = \langle \mathbf{D}G(\mathbf{x}_i, \cdot), \mathbf{D}G(\mathbf{x}_j, \cdot) \rangle = \langle \Phi(\mathbf{x}_i), \Phi(\mathbf{x}_j) \rangle \tag{7.3}$$

with Φ: $\mathbf{x}_i \rightarrow \mathbf{D}G(\mathbf{x}_i, \cdot)$. Hence, the Green's function is actually the kernel function induced by the Hilbert norm of the RKHS. Further discussions on the links between SVM and RNs can be found in Girosi (1998), Smola et al. (1998), and Evgeniou et al. (2000).

Note that in the SVM, the dimensionality of the kernel (Gram) matrix is the same as the number of observations, $\mathcal{L}$, which is in line with the RBF network; similar to the generalized RBF network in which the number of hidden units is less than $\mathcal{L}$, SVM can be also trained with a reduced set. We recommend that readers consult Schölkopf and Smola (2002) for an exhaustive treatment of regularization theory in the context of kernel learning and SVMs.

8 Bayesian Interpretation: Revisited

Neal (1996) and Williams (1998b) found that a large class of neural networks with certain weight priors will converge to a gaussian process (GP) prior

[35] An insightful discussion on the bias b in terms of the conditional positive-definite kernel was given in Poggio, Mukherjee, Rifkin, Rakhlin, and Verri (2001).

over the functions, in the limit of an infinite number of hidden units. This fact motivated researchers to consider the change from priors on weights to priors on functions: instead of imposing the prior on the weight (parameter) space, one can directly impose the prior on the functional space.[36] A common way is to define the prior in the RKHS (Wahba, 1990; Evgeniou et al., 2000). Using Bayes formula $p(f \mid \mathcal{D}) \propto p(\mathcal{D} \mid f)p(f)$, we have

$$p(\mathcal{D} \mid f) \propto \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{\ell}(y_i - f(\mathbf{x}_i))^2\right), \tag{8.1}$$

where σ^2 is the variance of the gaussian noise, and the prior probability $p(f)$ is given by

$$p(f) \propto \exp(-\|f\|_K^2/2s^2), \tag{8.2}$$

where the stabilizer $\|f\|_K^2$ is a norm defined in the RKHS associated with the kernel K. In particular, Parzen (1961, 1963) showed that the choice of the RKHS is equivalent to the choice of a zero-mean stochastic process with covariance kernel K (which is assumed to be symmetric positive definite), that is, $\mathbb{E}[f(\mathbf{x})f(\mathbf{y})] = s^2K(\mathbf{x}, \mathbf{y})$. And the regularization parameter is shown to be the SNR (i.e., the variance ratio σ^2/s^2; Bernard, 1999).

Hence, for the RNs or SVMs, choosing a kernel K is equivalent to assuming a gaussian prior on the functional f with normalized covariance equal to K (Papageorgiou, Girosi, & Poggio, 1998; Evgeniou et al., 2000). Usually, K is chosen to be positive definite and stationary (which corresponds to the shift-invariant property in the RBF). Also, choosing the covariance kernel is inherently related to finding the correlation function of the gaussian process (Wahba, 1990; Schölkopf & Smola, 2002). Generally, if the function is represented by the mixtures of gaussian processes, the kernel should also be a mixture of covariance kernels.

It is noteworthy to point out that the Bayesian interpretation discussed above applies not only to the classical regularization functional 3.5 with the square loss function but also to a generic loss functional (Evgeniou et al., 2000),

$$\mathcal{R}[f] = \sum_{i=1}^{\ell} L(y_i - f(\mathbf{x}_i)) + \lambda\|f\|_K^2, \tag{8.3}$$

where L is any monotonically increasing loss function. When $p(f)$ is assumed to be gaussian, the kernel is essentially the correlation function

[36] One can also define the prior in the functional space and further project it to the weight space (Zhu & Rohwer, 1996).

$\mathbb{E}[f(\mathbf{x}_i)f(\mathbf{x}_j)]$ by viewing f as a stochastic process. In the case of zero-mean gaussian processes, we have the covariance kernel,

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-|\mathbf{x}_i - \mathbf{x}_j|^2). \tag{8.4}$$

In the nongaussian situation, L is not a square loss function (Girosi, Poggio, & Caprile, 1991). In particular, for the SVM, it was shown (Evengiou et al., 2000) that the ϵ-insensitive loss function can be interpreted by a nongaussian noise model with superposition of gaussian processes with different variances and means,[37]

$$\exp(-|x|_\epsilon) = \int_{-\infty}^{\infty} dt \int_0^{\infty} d\beta \lambda(t)\mu(\beta)\sqrt{\beta}\exp(-\beta(x-t)^2), \tag{8.5}$$

where $\beta = 1/2\sigma^2$ and

$$\lambda_\epsilon(t) = \frac{1}{2(\epsilon+1)}(\chi_{[-\epsilon,\epsilon]}(t) + \delta(t-\epsilon) + \delta(t+\epsilon)), \tag{8.6}$$

$$\mu(\beta) \propto \beta^2 \exp(-1/4\beta), \tag{8.7}$$

where $\chi_{[-\epsilon,\epsilon]}(t)$ is 1 for $t \in [-\epsilon, \epsilon]$ and 0 otherwise.

L can be also chosen as Huber's loss function (Huber, 1981):

$$L(\xi) = \begin{cases} \frac{1}{2}|\xi|^2 & |\xi| < c \\ c|\xi| - \frac{c^2}{2} & |\xi| \geq c \end{cases}, \tag{8.8}$$

which was used to model the ϵ-contaminated noise density: $p(\xi) = (1-\epsilon)g(\xi) + \epsilon h(\xi)$ (where $g(\xi)$ is a fixed density and $h(\xi)$ is an arbitrary density, both of which are assumed to be symmetric with respect to origin, $0 < \epsilon < 1$). Provided $g(\xi)$ is chosen to be a gaussian density $g(\xi) = \frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{\xi^2}{2\sigma^2})$, the robust noise density is derived as

$$p(\xi) = \begin{cases} \frac{1-\epsilon}{\sqrt{2\pi}\sigma}\exp\left(-\frac{\xi^2}{2\sigma^2}\right), & |\xi| < c\sigma \\ \frac{1-\epsilon}{\sqrt{2\pi}\sigma}\exp\left(\frac{c^2}{2\sigma^2} - \frac{c}{\sigma}|\xi|\right), & |\xi| \geq c\sigma \end{cases}, \tag{8.9}$$

where c is determined from the normalization condition (Vapnik, 1998a).

Table 2 lists some loss functions and their associated additive noise probability density models, which are commonly used in regression. Albeit we discuss the learning problem in the regression framework, the discussion is also valid for classification problems. In addition, the loss function can

Table 2: Loss Functions and Associated Noise Density Models.

	Loss Function $L(\xi)$	**Noise Density Model** $p(\xi)$
Gaussian	$\frac{1}{2}\xi^2$	$\frac{1}{\sqrt{2\pi}}\exp(-\lvert\xi\rvert^2/2)$
Laplacian	$\lvert\xi\rvert$	$\frac{1}{2}\exp(-\lvert\xi\rvert)$
ϵ-insensitive	$\lvert\xi\rvert_\epsilon$	$\frac{1}{2(\epsilon+1)}\exp(-\lvert\xi\rvert_\epsilon)$
Huber's loss	$\begin{cases}\frac{1}{2}\xi^2 & \lvert\xi\rvert < c \\ c\lvert\xi\rvert - \frac{c^2}{2} & \text{otherwise}\end{cases}$	$\propto \begin{cases}\exp(-\frac{\xi^2}{2}) & \lvert\xi\rvert < c \\ \exp(\frac{c^2}{2} - c\lvert\xi\rvert) & \text{otherwise}\end{cases}$
Talvar	$\begin{cases}\xi^2/2 & \lvert\xi\rvert < c \\ c^2/2 & \text{otherwise}\end{cases}$	$\propto \begin{cases}\exp(-\xi^2/2c) & \lvert\xi\rvert < c \\ \exp(-c^2/2) & \text{otherwise}\end{cases}$
Cauchy	$\frac{1}{2}\log(1+\xi^2)$	$\frac{1}{\pi}\frac{1}{1+\xi^2}$
Hyperbolic cosine	$\log(\cosh(\xi))$	$\frac{1}{\pi\cosh(\xi)}$
L_p norm	$\frac{1}{r}\lvert\xi\rvert^r$	$\frac{r}{2\Gamma(1/r)}\exp(-\lvert\xi\rvert^r)$

include higher-order cumulants to enhance robustness in the nongaussian noise situation (Leung & Chow, 1997, 1999, 2001).

In the light of the discussions, studying the priors on the functional provides much freedom for incorporating of prior knowledge to the learning problem. In particular, the problem is in essence to design a kernel that may explain well the observation data underlying the functional. For example, if one knows that a functional might not be described by a gaussian process (e.g., Laplacian process or others), one can design a particular kernel (Laplacian kernel or other localized nonstationary kernels) without worrying about the smoothness property. See Schölkopf and Smola (2002) and Genton (2000) for more discussion.

9 Pruning Algorithms

Pruning is an efficient way to improve the generalization ability of neural networks (Reed, 1993; Cherkassky & Mulier, 1998; Haykin, 1999). It can be viewed as a sort of regularization approach where the complexity term is measured by a stabilizer. Pruning can be either connections pruning or nodes pruning. Roughly, there are four kinds of pruning algorithms developed from different perspectives, with the first two kinds mainly concerned with connections pruning and the last two concerned with nodes pruning:

- Penalty function (Setiono, 1997) or stabilizer-based pruning approaches, such as weight decay (Hinton, 1989), weight elimination (Weigend et al., 1991), and Laplace prior (Williams, 1994). Soft weight

[37] A limitation of Bayesian interpretation of SVM was also discussed in Evengiou et al. (2000).

sharing (Nowlan & Hinton, 1992) is another kind of pruning algorithm that supposes the weights are represented by a mixture of gaussians and the weights are supposed to share the same value.

- Second-order (Hessian) information-based pruning approaches, such as optimal brain damage (OBD; LeCun, Denker, & Solla, 1990) and optimal brain surgeon (OBS; Hassibi, Stock, & Wolff, 1992).
- Information-theoretic criteria-based pruning schemes (Deco et al., 1995; Kamimura, 1997).
- Matrix decomposition (or factorization)–based pruning methods, such as principal component analysis (PCA; Levin, Leen, & Moody, 1994), SVD (Kanjilal & Banerjee, 1995), QR decomposition (Jou, You & Chang, 1994), discriminant component pruning analysis (DCP; Koene & Takane, 1999), and contribution analysis (Sanger, 1989).

The matrix decomposition–based pruning schemes are based on the observation of the ill-conditioned matrix $\mathbf{G}$. For instance, taking the QR decomposition, $\mathbf{Gw} = QR$, we may obtain a new expression after pruning some hidden nodes (see appendix E for derivation):

$$\mathbf{y} = \mathbf{Gw} \rightarrow \hat{\mathbf{y}} = \hat{\mathbf{G}}\hat{\mathbf{w}}, \tag{9.1}$$

where $\hat{\mathbf{G}} = \mathbf{G}L_1R_1^{-1}$ ($\mathbf{w} = [L_1 \quad L_2]$), and $\hat{\mathbf{w}}$ is calculated by

$$\hat{\mathbf{w}}^T = \mathbf{w}^T L_1 + \mathbf{w}L_2(R_1^{-1}R_2)^T. \tag{9.2}$$

10 Equivalent Regularization

In the machine learning community, there are various approaches implementing the implicit regularization (neither the explicit form of equation 3.5 nor equation 3.49), which we call equivalent regularization. To name a few, these approaches include early stopping (Bishop, 1995a; Cherkassky & Mulier, 1998), use of momentum (Haykin, 1999), incorporation of invariance (Abu-Mostafa, 1995; Leen, 1995; Niyogi et al., 1998), tangent distance and tangent prop (Simard, Victorri, LeCun, & Denker, 1992; Simard, LeCun, Denker, & Victorri, 1998; Vapnik, 1998a), smoothing regularizer (Wu & Moody, 1996), flat minima (Hochreiter & Schmidhuber, 1997), sigmoid gain scaling, target smoothing (Reed, Marks, & Oh, 1995), and training with noise (An, 1996; Bishop, 1995a, 1995b). Particularly, training with noise is an approximation to training with a kernel regression estimator; choosing the variance of the noise is equivalent to choosing the bandwidth of the kernel of the regression estimator. For a detailed discussion on training with noise, see An (1996), Bishop (1995a, 1995b), and Reed et al. (1995). An equivalence discussion between gain scaling, learning rate, and scaling weight magnitude is given in Thimm, Moerland, & Fiesler (1996).

Many regularization techniques correspond to the structural learning principle. With structural learning, the learning process (in terms of hypothesis, data space, parameters, or implementation) is controlled and performed in a nested structure:

$$S_0 \subset S_1 \subset \cdots S_m \subset \cdots.$$

The capacity control and generalization performance are guaranteed and improved by constraining the learning process in the specific structure, which can be viewed as an implicit regularization (Cherkasskay & Mulier, 1998). Early stopping is an example of structural learning in terms of implementation of a learning process.

By early stopping, it is said that the training of a network is stopped before it goes to the minimum error, while observing the error on an independent validation set begin to increase. In the case of quadratic risk function $\mathcal{R}_{emp}$, early stopping is similar to weight-decay regularization; the product of the iteration index t and the learning rate η plays the role of regularization parameter λ, in the sense that the components of weight vectors parallel to the eigenvectors of the Hessian satisfy (Bishop, 1995a)

$$\mathbf{w}_i^{(t)} \simeq \mathbf{w}^*, \xi_i \gg (\eta t)^{-1} \tag{10.1}$$

$$|\mathbf{w}_i^{(t)}| \ll |\mathbf{w}^*|, \xi_i \ll (\eta t)^{-1}, \tag{10.2}$$

where $\mathbf{w}^*$ denotes the desired minimum point where $\mathcal{R}_{emp}$ achieves in the weight space and ξ_i represent the eigenvalues of the Hessian matrix $\mathbf{H}(\mathbf{w})$. In this sense, early stopping can be interpreted as an implicit regularization where a penalty is defined on a searching path in the parametric space. The solutions are penalized according to the number of gradient descent steps taken along the path from the starting point (MacKay, 1992; Cherkasskay & Mulier, 1998).

11 Kolmogorov Complexity: A Universal Principle for Regularization?

Regularization theory can be established from many principles (e.g., MDL, Bayes, entropy), with many visible successes in model selection, complexity control, and generalization improvement. The weakness of these principles is that none of them has universality, in the sense that they cannot be applied to an arbitrary area under an arbitrary situation. Is it possible to find a universal principle for regularization theory in machine learning? This question leads us to think in terms of a well-known concept in the computational learning community: Kolmogorov complexity (or algorithmic complexity).

Kolmogorov complexity theory, motivated by the Turing machine, was first studied by Solomonoff and Kolmogorov. The main thrust of Kolmo-

gorov complexity lies in its universality, which is dedicated to constructing a universal learning method based on the universal coding method (Schmidhuber, 1994). According to Kolmogorov complexity theory, any complexity can be measured by the length of the shortest program for a universal Turing machine, which correctly reproduces the observation data in a look-up table. The core of Kolmogorov complexity theory contains three parts: complexity, randomness, and information.

It is interesting to make a comparison between these three concepts and machine learning. In particular, the MDL principle is well connected to the complexity; the randomness is inherently related to the well-known no-free-lunch (NFL) theorems (Wolpert, 1996), such as NFL for cross-validation (Zhu, 1996; Goutte, 1997), optimization (Wolpert & Macready, 1997), noise prediction (Magdon-Ismail, 2000), and early stopping (Cataltepe, Abu-Mostafa, & Magon-Ismail, 1999). All NFL theorems basically state that no learning algorithms can be universally good; some algorithms that perform well will perform poorly in other situations. The information is related to the entropy theory, which measures the uncertainty in a probabilistic sense. From the Bayesian viewpoint, Kolmogorov complexity can be understood by dealing with a universal prior (so-called Solomonoff-Levin distribution), which measures the prior probability of guessing a halting program that computes the bitstrings on a universal Turing machine. Since the Kolmogorov complexity and the universal prior cannot be computed, some generalized complexity concepts for the purpose of computability were developed (e.g., Levin complexity). An extended discussion on this subject, however, is beyond the scope of this review.[38] Some theoretical studies of the relations between MDL, Bayes theory and Kolmogorov complexity can be found in Vitanyi and Li (2000).[39]

Can the Kolmogorov complexity be a universal principle for regularization theory? In other words, can it cover different principles such as MDL, Bayes, and beyond? There seems to exist an underlying relationship between Kolmogorov complexity theory and regularization theory (especially in the machine learning area). However, for the time being, a solid theoretical justification is missing. Although some seminal work has been reported (Pearlmutter & Rosenfeld, 1991; Schmidhuber, 1994), the studies and results were empirical, and theoretic verification was still in an early stage. The question remains unanswered and needs investigation.

12 Summary and Beyond

This review presents a comprehensive understanding of regularization theory from different viewpoints. In particular, the spectral regularization

[38] For a descriptive and detailed treatment on the Kolmogrov complexity, see Cover and Thomas (1991) and a special issue on the Kolmogrov complexity in *Computer* (vol. 42, no. 4, 1999).

[39] See NIPS2001 workshop for more information.

framework is derived in the vein of the Fourier operator and Plancherel identity. State-of-the-art research on various regularization techniques is reviewed, and many related topics in machine learning are addressed and explored.

The contents of this review cover such topics as functional analysis, operator theory, machine learning, statistics, statistical learning, Bayesian inference, information theory, computational learning theory, matrix theory, numerical analysis, and optimization. Nevertheless, they are closely related to regularization theory. Roughly, Occam's razor, MDL, and MinEnt are the principles of implementing the regularization; Bayesian theory, information theory, and statistical learning theory can be formulated in the theoretic framework level, which establishes the mathematical foundation for the regularization principle, whereas pruning algorithms, equivalent regularization approaches, RNs, SVMs, and GP belong to the application or implementation level. A schematic relationship of the topics in this review is illustrated in Figure 2.

To this end, we conclude with some comments on possible future investigations related to the topics discussed in this review:

- It has been shown that there exists a close relationship between regularization theory and sparse representation (Poggio & Girosi, 1998), SVMs (Girosi, 1998; Smola et al., 1998; Evgeniou et al., 2000; Schölkopf & Smola, 2002), gaussian processes (MacKay, 1998; Williams, 1998b; Zhu et al., 1998), independent component analysis (Hochreiter & Schmidhuber, 1998; Evgeniou et al., 2000; Bach & Jordan, 2001), wavelet approximation (Bernard, 1999), matching pursuit (Mallat & Zhang, 1993), and basis pursuit (Chen, Donoho, & Saunders, 1998). Further efforts will be to put many machine learning problems into a general framework and discuss their properties.
- Prospective studies of regularized networks are devoted to build an approximation framework in hybrid functional spaces. Many encouraging results have been attained in the RKHS. One can extend the framework to other functional spaces, such as generalized Fock space (van Wyk & Durrani, 2000) or Besov space. The idea behind the hybrid approximation is to find an overcomplete representation of an unknown function by means of a direct sum of some possibly overlapping functional spaces, which results in a very sparse representation of the function of interest. The SRM principle (Vapnik, 1995, 1998a) seems to be a solid framework and a powerful mathematical tool for this goal. In addition, the algorithmic implementation of regularization remains an important issue. Fast algorithms beyond the quadratic programming in the kernel learning community are anticipated. A recent new direction is the on-line or incremental learning algorithms developed for GNs, SVMs, or GP (Zhu & Rohwer, 1996; De Nicolao & Ferrari-Trecate, 2001; Cauwenberghs & Poggio, 2001; Csató & Opper, 2002).

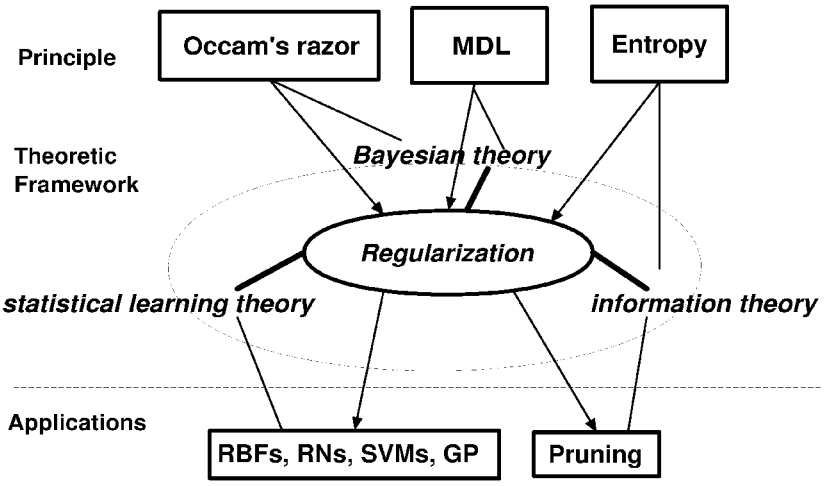

Figure 2: Schematic illustration of the theory of regularization. The dashed lines distinguish three levels: principle, theory, and application; the undirectional solid lines represent some conceptual links; and the directional arrows represent the routes from principle to theory and from theory to applications.

- A wide class of RNs, including RBF (Powell, 1985; Micchelli, 1986; Broomhead & Lowe, 1988; Girosi, 1992, 1993), hyperBF (Poggio & Girosi, 1990a), smoothing splines (Silverman, 1984; Wahba, 1990), and generalized additive models (Hastie & Tibshirani, 1990), can be formalized mathematically based on regularization theory (Girosi et al., 1995). Since the connection between RNs and SVMs was theoretically established, one has much freedom to choose the regularization operator and the associated basis functions or kernels (Genton, 2000). For example, it can be extended to wavelet networks (Koulouris, Bakshi, & Stephanopoulous, 1995; Mukherjee & Nayar, 1996; Bernard, 1999; Gao, Harris, & Gunn, 2001). Besides, theoretic studies on the generalization bounds of the RNs or kernel machines are always important (Xu et al., 1994; Corradi & White, 1995; Krzyzak & Linder, 1996; Freeman & Saad, 1995; Niyogi & Girosi, 1996, 1999; Vapnik, 1998a; Williamson, Smola, & Schölkopf, 2001; Cucker & Smale, 2002), for which we are left with many open problems.

- Most well-established smoothing operators in the literature so far are defined in either the spatial domain (e.g., smoothing splines) or frequency domain (e.g., RKHS norm), as reviewed in this article. One can, however, consider the space-frequency smoothing operator, which es-

sentially corresponds to a localized, nonstationary (or locally stationary) kernel in the general case.[40] Careful design of the operator may achieve multiresolution smoothing or approximation, which is the spirit of structural learning. Besides, the operator and the associated kernels can be time varying, which allows on-line estimation from the data.

- Canu and Elisseeff (1999) reported that if the Radon-Nikodym derivative instead of Fréchet derivative is used in the regularized functional, the solution to the regularization problem gives rise to a sigmoid-shaped network, which partially answers the unanswered question posed by Girosi et al. (1995): Can the sigmoid-like neural network be derived from regularization theory?

- A wide class of neural networks and stochastic models form a curved exponential family of parameterized neuromanifolds (Amari & Nagaoka, 2000). It will be possible to study the intrinsic relationship between differential geometry and regularization theory in terms of choice of kernels (Burges, 1999). Naturally, the smoothness of the nonlinear feature mapping in terms of kernel (which is connected to the covariance property of the functional of interest) is measured by the curvature of the corresponding hypersurface. The higher the curvature is, the less smooth the parameterized model is, and thus the poorer generalization performance is anticipated (Zhu & Rohwer, 1995). It is also possible to incorporate some invariance to implement the equivalent regularization (Simard et al., 1998; Burges, 1999; Schölkopf & Smola, 2002).

- Another area not covered in this review is the nonclassical Tikhonov regularization, which involves the nonconvex risk functional. Without the nice quadratic property, it is still possible to use variational methods (e.g., mean-field approximation method) or stochastic sampling methods (e.g., Gibbs sampling, Markov chain Monte Carlo) to handle regularized problems. Some studies in this direction can be found in Geman and Geman (1984) and Lemm (1996, 1998).

Appendix A: Proof of Dirichlet Kernel

For the purpose of self-containing the article, the proof of Dirichlet kernel given in Lanczos (1961) is presented here. Observe that the Dirac delta

[40] Choosing a localized kernel with compact support is essentially related to finding an operator with composition of some band-limiting and time-limiting operators.

function $\delta(s, x)$ satisfies the following conditions:

$$\int_{-\pi}^{\pi} \delta(s, x) \cos kx \, dx = \cos ks, \tag{A.1}$$

$$\int_{-\pi}^{\pi} \delta(s, x) \sin kx \, dx = \sin ks. \tag{A.2}$$

Consider a symmetric, translationally invariant function $G(s, x) = G(x, s) = G(s - x) \equiv G(\theta) = G(-\theta)$, which is zero everywhere except in the interval $|\theta| \leq \epsilon$, where ϵ is a small, positive value. The expansion coefficients a_k and b_k of this function are

$$\begin{aligned} a_k &= \frac{1}{\pi} \int_{-\epsilon}^{\epsilon} \cos k(s + \theta) g(\theta) \, d\theta \\ &= \frac{1}{\pi} \cos k\xi \int_{-\epsilon}^{\epsilon} G(\theta) \, d\theta, \end{aligned} \tag{A.3}$$

$$\begin{aligned} b_k &= \frac{1}{\pi} \int_{-\epsilon}^{\epsilon} \sin k(s + \theta) g(\theta) \, d\theta \\ &= \frac{1}{\pi} \sin k\xi \int_{-\epsilon}^{\epsilon} G(\theta) \, d\theta, \end{aligned} \tag{A.4}$$

where $\xi \in (s - \epsilon, s + \epsilon)$. Provided $\int_{-\epsilon}^{\epsilon} G(\theta) \, d\theta = 1$, letting $\epsilon \to 0$, then the point $\xi \to s$, and one may obtain the desired expansion coefficients in equations A.1 and A.2. If we replace $K(s - x)$ by the Fourier expansion coefficients of the Dirac function, then comparing equation 3.21, the Dirichlet kernel $K_n(s, x)$ acts like a Dirac delta function:

$$\int_{-\pi}^{\pi} f(s) \delta(s, x) \, ds = f(x). \tag{A.5}$$

Appendix B: Proof of Regularization Solution

The proof of regularization solution was partly given in Haykin (1999) and is rewritten here for completeness. By virtue of equation 3.13, applying $\mathbf{L}$ to the function $f(\mathbf{x})$, we obtain

$$\begin{aligned} \mathbf{L}f(\mathbf{x}) &= \mathbf{L} \int_{\mathbb{R}^N} G(\mathbf{x}, \boldsymbol{\xi}) \varphi(\boldsymbol{\xi}) \, d\xi \\ &= \int_{\mathbb{R}^N} \mathbf{L}G(\mathbf{x}, \boldsymbol{\xi}) \varphi(\boldsymbol{\xi}) \, d\boldsymbol{\xi} \\ &= \int_{\mathbb{R}^N} \delta(\mathbf{x} - \boldsymbol{\xi}) \varphi(\boldsymbol{\xi}) \, d\boldsymbol{\xi} \\ &= \varphi(\mathbf{x}). \end{aligned} \tag{B.1}$$

Similarly, applying $\mathbf{K}$ to the function $f(\mathbf{x})$ yields

$$\begin{aligned}\mathbf{K}f(\mathbf{s}) &= \mathbf{K}\int_{\mathbb{R}^N} G(\mathbf{x},\boldsymbol{\xi})\varphi(\boldsymbol{\xi})\,d\boldsymbol{\xi}\\ &= \int_{\mathbb{R}^N} \mathbf{K}G(\mathbf{s},\boldsymbol{\xi})\varphi(\boldsymbol{\xi})\,d\boldsymbol{\xi}\\ &= \int_{\mathbb{R}^N} \exp(-j\mathbf{s}\boldsymbol{\xi})\varphi(\boldsymbol{\xi})\,d\boldsymbol{\xi}\\ &= \Phi(\mathbf{s}).\end{aligned} \tag{B.2}$$

The solution of the regularization problem is further derived by setting

$$\varphi(\boldsymbol{\xi}) = \frac{1}{\lambda}\sum_{i=1}^{\ell}[y_i - f(\mathbf{x}_i)]\delta(\boldsymbol{\xi}-\mathbf{x}_i), \tag{B.3}$$

$$\Phi(\boldsymbol{\omega}) = \mathcal{F}\{\varphi(\boldsymbol{\xi})\} = \frac{1}{\lambda}\sum_{i=1}^{\ell}[y_i - f(\mathbf{x}_i)]\exp(-j\mathbf{x}_i\boldsymbol{\omega}). \tag{B.4}$$

In the spatial domain, we have

$$\begin{aligned}f_\lambda(\mathbf{x}) &= \int_{\mathbb{R}^N} G(\mathbf{x},\boldsymbol{\xi})\left\{\frac{1}{\lambda}\sum_{i=1}^{\ell}[y_i - f(\mathbf{x}_i)]\delta(\boldsymbol{\xi}-\mathbf{x}_i)\right\}d\boldsymbol{\xi}\\ &= \frac{1}{\lambda}\sum_{i=1}^{\ell}[y_i - f(\mathbf{x}_i)]\int_{\mathbb{R}^N} G(\mathbf{x},\boldsymbol{\xi})\delta(\boldsymbol{\xi}-\mathbf{x}_i)\,d\boldsymbol{\xi}\\ &= \sum_{i=1}^{\ell} w_i G(\mathbf{x},\mathbf{x}_i),\end{aligned} \tag{B.5}$$

where $w_i = [y_i - f(\mathbf{x}_i)]/\lambda$. Equivalently, in the frequency domain, we have

$$\begin{aligned}f_\lambda(\mathbf{x}) &= \int_{\mathbb{R}^N} \mathcal{F}\{G(\mathbf{x},\boldsymbol{\xi})\}\overline{\mathcal{F}\left\{\frac{1}{\lambda}\sum_{i=1}^{\ell}[y_i - f(\mathbf{x}_i)]\delta(\boldsymbol{\xi}-\mathbf{x}_i)\right\}}\,d\boldsymbol{\omega}\\ &= \int_{\mathbb{R}^N} \mathcal{G}(\mathbf{x},\boldsymbol{\omega})\frac{1}{\lambda}\sum_{i=1}^{\ell}[y_i - f(\mathbf{x}_i)]\overline{\exp(-j\mathbf{x}_i\boldsymbol{\omega})}\,d\boldsymbol{\omega}\\ &= \frac{1}{\lambda}\sum_{i=1}^{\ell}[y_i - f(\mathbf{x}_i)]\int_{\mathbb{R}^N} \mathcal{G}(\mathbf{x},\boldsymbol{\omega})\exp(j\mathbf{x}_i\boldsymbol{\omega})\,d\boldsymbol{\omega}\\ &= \sum_{i=1}^{\ell} w_i G(\mathbf{x},\mathbf{x}_i),\end{aligned} \tag{B.6}$$

which is identical to equation B.5. The second line in the above equation uses equation B.4, and the first line follows from the Parseval theorem.

Appendix C: Another Proof of Regularization Solution

The proof is slightly modified from the proof given in Bernard (1999).

From Riesz's representation theorem, we know that $\forall \mathbf{x} \in \mathbb{R}^N$; there exists a basis function $\phi_{\mathbf{x}} \in \mathbb{H}$ such that $f(\mathbf{x}) = \langle f, \phi_{\mathbf{x}} \rangle$ for all $f \in \mathbb{H}$. By defining an interpolation kernel,

$$K(\mathbf{x}, \mathbf{s}) = \langle \phi_{\mathbf{x}}, \phi_{\mathbf{s}} \rangle, \tag{C.1}$$

with the property $K(\mathbf{x}, \mathbf{s}) = \overline{K(\mathbf{s}, \mathbf{x})}$, the solution to the regularization problem can be written by the basis function $\phi_{\mathbf{x}}$:

$$\begin{aligned} \mathcal{R}[f] &= \frac{1}{2}\sum_{i=1}^{\ell}[y_i - f(\mathbf{x}_i)]^2 + \frac{1}{2}\lambda\|f\|_{\mathbb{H}}^2 \\ &= \frac{1}{2}\sum_{i=1}^{\ell}[y_i - f(\mathbf{x}_i)]^2 + \frac{1}{2}\lambda\langle f, f\rangle. \end{aligned} \tag{C.2}$$

Differentiating equation C.2 with respect to f and setting to be zero, we obtain

$$\begin{aligned} d\mathcal{R}[f] &= \lambda\langle f, df\rangle + \sum_{i=1}^{\ell}\langle df, \phi_{\mathbf{x}_i}\rangle\langle f, \phi_{\mathbf{x}_i}\rangle - \sum_{i=1}^{\ell} y_i\langle df, \phi_{\mathbf{x}_i}\rangle \\ &= \left\langle \lambda f + \sum_{i=1}^{\ell}\langle f, \phi_{\mathbf{x}_i}\rangle\phi_{\mathbf{x}_i} - \sum_{i=1}^{\ell} y_i\phi_{\mathbf{x}_i} \,\middle|\, df \right\rangle = 0. \end{aligned} \tag{C.3}$$

It further follows that

$$\lambda f = \sum_{i=1}^{\ell}(y_i - \langle f, \phi_{\mathbf{x}_i}\rangle)\phi_{\mathbf{x}_i}.$$

Hence, the solution can be represented by a linear combination of basis functions $\phi_{\mathbf{x}_i}$:

$$f(\mathbf{x}) = \sum_{i=1}^{\ell} c_i\phi_{\mathbf{x}_i} = \sum_{i=1}^{\ell} c_i K(\mathbf{x}, \mathbf{x}_i).$$

Denote $\mathbf{f} = [\langle f, \phi_{\mathbf{x}_1}\rangle, \ldots, \langle f, \phi_{\mathbf{x}_\ell}\rangle]^T$, $\mathbf{c} = [c_1, \ldots, c_\ell]^T$, $\mathbf{y} = [y_1, \ldots, y_\ell]^T$, K as an ℓ-by-ℓ matrix; writing in a matrix form, we have $K\mathbf{c} = \mathbf{f}$, $\lambda\mathbf{f} = K(\mathbf{y} - \mathbf{f})$, and $\mathbf{c} = (K + \lambda\mathbf{I})^{-1}\mathbf{y}$.

Appendix D: GSVD

The generalized singular value decomposition (GSVD) is defined as

$$[\mathbf{U}, \mathbf{V}, \mathbf{Z}, \mathbf{\Sigma}, \mathbf{S}] = \mathrm{GSVD}(\mathbf{A}, \mathbf{B}),$$

where $\mathbf{A}$, $\mathbf{B}$ are $m \times p$ and $n \times p$ matrices, respectively; $\mathbf{U}_{m\times m}$, $\mathbf{V}_{n\times n}$ are the unitary matrices; matrix $\mathbf{Z}_{p\times q}$ ($q = \min\{m + n, p\}$) is usually (but not necessarily) square; and $\mathbf{\Sigma}$ and $\mathbf{S}$ are diagonal matrices. All satisfy the following relationship:

$$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{Z}^T, \quad \mathbf{B} = \mathbf{V}\mathbf{S}\mathbf{Z}^T, \quad \mathbf{\Sigma}^T\mathbf{\Sigma} + S^T\mathbf{S} = \mathbf{I}.$$

Suppose σ_i and s_i are the on-diagonal singular values in the singular matrices $\mathbf{\Sigma}$ and $\mathbf{S}$, respectively; the generalized singular values are defined by $\gamma_i = (\sigma_i^2 + s_i^2)^{1/2}$.

Appendix E: QR Decomposition

Applying the QR decomposition to the matrix $\mathbf{G}$ (Golub & Van Loan, 1996), one has

$$\mathbf{G}L = QR,$$

where $\mathbf{G}$ is an $\ell \times m$ input matrix, L is an $m \times m$ transposition matrix, Q is an $\ell \times \ell$ full-rank matrix, and R is an $\ell \times m$ upper triangle matrix. In particular, R and Q are expressed by

$$R = \begin{bmatrix} R_1 & R_2 \\ \mathbf{0} & \mathbf{0} \end{bmatrix}, Q = \begin{bmatrix} Q_1 & Q_2 \end{bmatrix};$$

henceforth,

$$\mathbf{G}L = \begin{bmatrix} \mathbf{G}L_1 & \mathbf{G}L_2 \end{bmatrix},$$

and it further follows that

$$Q_1 = \mathbf{G}L_1R_1^{-1}, \mathbf{G}L_2 = Q_1R_2 = \mathbf{G}L_1R_1^{-1}R_2,$$

where L_1 is an $m \times r$ matrix, L_2 is an $m \times (m-r)$ matrix, Q_1 is an $\ell \times r$ matrix, Q_2 is an $\ell \times (m-r)$ matrix, R_1 is a $r \times r$ matrix, and R_2 is a $r \times (m-r)$ matrix.

Suppose the number of hidden nodes is pruned from m to r ($m > r$): $\mathbf{G}_{\ell\times m} \rightarrow \hat{\mathbf{G}}_{\ell\times r}$, that is, there are $(m - r)$ redundant hidden nodes to be

deleted. Denoting the new weight vector by $\hat{\mathbf{w}}$, the new expression for the network is written by $\hat{\mathbf{y}} = \hat{\mathbf{G}}\hat{\mathbf{w}}$, where $\hat{\mathbf{G}} = Q_1 = \mathbf{G}L_1R_1^{-1}$, and $\hat{\mathbf{w}}_{r\times 1}$ is calculated as

$$\hat{\mathbf{w}}^T = \mathbf{w}^T L_1 + \mathbf{w}L_2(R_1^{-1}R_2)^T.$$

Acknowledgments

The work is supported by the Natural Sciences and Engineering Research Council (NSERC) of Canada. Z. C. is partly supported by Clifton W. Sherman Scholarship. This work was initially motivated by and benefited a lot from the NIPS community. We thank Rambrandt Bakker for some helpful feedbacks on reading the manuscript and the anonymous reviewers for many critical comments on the earlier draft.

References

Abu-Mostafa, Y. S. (1995). Hints. *Neural Computation, 7*, 639–671.

Adams, R. A. (1975). *Sobolev spaces*. San Diego, CA: Academic Press.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*, 716–723.

Amari, S., & Nagaoka, H. (2000). *Methods of information geometry*. New York: AMS and Oxford University Press.

An, G. (1996). The effects of adding noise during backpropagation training on a generalization performance. *Neural Computation, 8*, 643–674.

Aronszajn, N. (1950). Theory of reproducing kernels. *Transactions of the American Mathematical Society, 68*, 337–404.

Atick, J. J. (1992). Could information theory provide an ecological theory of sensory processing? *International Journal of Neural Systems, 3*, 213–251.

Bach, F. R., & Jordan, M. I. (2001). *Kernel independent component analysis* (Tech. Rep.). Berkeley: Division of Computer Science, University of California, Berkeley.

Balakrishnan, A. V. (1976). *Applied functional analysis*. New York: Springer-Verlag.

Barlow, H. B. (1989). Unsupervised learning. *Neural Computation, 1*, 311–325.

Barron, A. R. (1991). Complexity regularization with application to artificial neural networks. In G. Roussas (Ed.), *Nonparametric functional estimation and related topics* (pp. 561–576). Dordrecht: Kluwer.

Becker, S. (1996). Mutual information maximization: Models of cortical self-organization. *Network: Computation in Neural Systems, 7*, 7–31.

Bernard, C. (1999). *Wavelets and ill-posed problems: Optical flow and data interpolation*. Unpublished doctoral dissertation, Ecole Polytechnique, France. Avaiable on-line at: http://www.cmap.polytechnique.fr/~bernard/these/table-en.pdf.

Bertero, M., Poggio, T., & Torre, V. (1988). Ill-posed problems in early vision. *Proceedings of the IEEE, 76*(8), 869–889.

Bishop, C. (1995a). *Neural networks for pattern recognition*. New York: Oxford University Press.

Bishop, C. (1995b). Training with noise is equivalent to Tikhonov regularization. *Neural Computation, 7*, 108–116.

Bochner, S. (1955). *Harmonic analysis and the theory of probability*. Berkeley: University of California Press.

Breiman, L. (1998). Bias-variance, regularization, instability and stabilization. In C. Bishop (Ed.), *Neural networks and machine learning* (pp. 27–56). New York: Springer-Verlag.

Broomhead, D. S., & Lowe, D. (1988). Multivariate functional interpolation and adaptive networks. *Complex Systems, 2*, 321–355.

Bruntine, W. L., & Weigend, A. S. (1991). Bayesian back-propagation. *Complex Systems, 5*, 603–643.

Burges, C. (1999). Geometry and invariance in kernel based methods. In B. Schölkopf, C. Burges, & A. Smola (Eds.), *Advances in kernel methods—Support vector learning* (pp. 89–116). Cambridge, MA: MIT Press.

Canu, S., & Elisseeff, A. (1999). *Regularization, kernels and sigmoid net*. Unpublished manuscript. Available on-line at: http://psichaud.insa-rouen.fr/~scanu/.

Cataltepe, Z., Abu-Mostafa, Y. S., & Magon-Ismail, M. (1999). No free lunch for early stopping. *Neural Computation, 11*, 995–1001.

Cauwenberghs, G., & Poggio, T. (2001). Incremental and decremental support vector machine learning. In T. K. Leen, T. G. Dietterich, & V. Tresp (Eds.), *Advances in neural information processing systems, 13* (pp. 409–415). Cambridge, MA: MIT Press.

Chen, S., Donoho, D., & Saunders, M. A. (1998). Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing, 20*(1), 33–61.

Chen, Z., & Haykin, S. (2001a). A new view on regularization theory. In *Proc. IEEE Int. Conf. System, Man and Cybernetics* (pp. 1642–1647). Tucson, AZ.

Chen, Z., & Haykin, S. (2001b). *On different facets of regularization theory* (Tech. Rep.). Hamilton, ON: Communications Research Laboratory, McMaster University.

Cherkassky, V., & Mulier, F. (1998). *Learning from data: Concepts, theory and methods*. New York: Wiley.

Coifman, R. R., & Wickerhauser, M. V. (1992). Entropy-based algorithms for best-basis selection. *IEEE Transactions on Information Theory, 38*, 713–718.

Corradi, V., & White, H. (1995). Regularized neural networks: Some convergence rate results. *Neural Computation, 7*, 1225–1244.

Courant, R., & Hilbert, D. (1970). *Methods of mathematical physics* (Vol. 1). New York: Wiley.

Cover, T., & Thomas, J. A. (1991). *The elements of information theory*. New York: Wiley.

Csató, L., & Opper, M. (2002). Sparse on-line gaussian process. *Neural Computation, 14*, 641–668.

Cucker, F., & Smale, S. (2002). On the mathematical foundations of learning. *Bulletin of the American Mathematical Society, 39*, 1–49.

Daugman, J. G. (1989). Entropy reduction and decorrelation in visual coding by oriented neural receptive fields. *IEEE Transactions on Biomedical Engineering, 36*(1), 107–114.

Debnath, L., & Mikusinski, P. (1999). *Introduction to Hilbert spaces with applications* (2nd ed.). San Diego, CA: Academic Press.

Deco, G., Finnoff, W., & Zimmermann, H. G. (1995). Unsupervised mutual information criterion for elimination of overtraining in supervised multilayer networks. *Neural Computation, 7*, 86–107.

De Nicolao, G., & Ferrari-Trecate, G. (2001). Regularization networks: Fast weight calculation via Kalman filtering. *IEEE Transactions on Neural Networks, 12*(2), 228–235.

Dontchev, A. L., & Zolezzi, T. (1993). *Well-posed optimization problems*. Berlin: Springer-Verlag.

Duchon, J. (1977). Spline minimizing rotation-invariant semi-norms in Sobolev spaces. In W. Schempp & K. Zeller (Eds.), *Constructive theory of functions of several variables* (pp. 85–100). Berlin: Springer-Verlag.

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

Freeman, A., & Saad, D. (1995). Learning and generalization in radial basis function network. *Neural Computation, 7*, 1000–1020.

Gao, J. B., Harris, C. J., & Gunn, S. R. (2001). On a class of support vector kernels based on frames in function Hilbert space. *Neural Computation, 13*, 1975–1994.

Geman, S., Bienenstock, E., & Doursat, R. (1992). Neural networks and the bias-variance dilemma. *Neural Computation, 4*, 1–58.

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distribution, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 6*, 721–741.

Genton, M. G. (2000). Classes of kernels for machine learning: A statistics perspective. *Journal of Machine Learning Research, 2*, 299–312.

Girosi, F. (1992). Some extensions of radial basis functions and their applications in artificial intelligence. *International Journal of Computer and Mathematics with Applications, 24*(12), 61–80.

Girosi, F. (1993). Regularization theory, radial basis functions and networks. In V. Cherkassky, J. H. Friedman, & H. Wechsler (Eds.), *From statistics to neural networks* (pp. 134–165). Berlin: Springer-Verlag.

Girosi, F. (1998). An equivalence between sparse approximation and support vector machines. *Neural Computation, 10*, 1455–1480.

Girosi, F., Jones, M., & Poggio, T. (1995). Regularization theory and neural network architecture. *Neural Computation, 7*, 219–269.

Girosi, F., Poggio, T., & Caprile, B. (1991). Extensions of a theory of networks for approximation and learning: Outliers and negative examples. In D. Touretzky & R. Lippmann (Eds.), *Advances in neural information processing systems, 3* (pp. 750–756). San Mateo, CA: Morgan Kaufmann.

Golub, G. H., & Van Loan, C. F. (1996). *Matrix computations* (3rd ed.). Baltimore, MD: John Hopkins University Press.

Goutte, C. (1997). Note on free lunches and cross-validation. *Neural Computation, 9*, 1245–1249.

Hansen, P. C. (1992). Analysis of discrete ill-posed problems by means of the *L*-curve. *SIAM Review, 34*(4), 561–580.

Hansen, P. C. (1998). *Rank deficient and discrete ill-posed problems: Numerical aspects of linear inversion*. Philadelphia: SIAM.

Harpur, G. F., & Prager, R. W. (1996). Development of low entropy coding in a recurrent network. *Network: Computation in Neural Systems, 7*, 277–284.

Hassibi, B., Stock, D. G., & Wolff, G. J. (1992). Optimal brain surgeon and general network pruning. In *Proc. Int. Conf. Neural Networks* (pp. 293–299), San Francisco.

Hastie, T. (1996). Pseudosplines. *Journal of Royal Statistical Society B, 58*, 379–396.

Hastie, T., & Tibshirani, R. (1990). *Generalized additive models*. London: Chapman and Hall.

Haykin, S. (1999). *Neural networks: A comprehensive foundation* (2nd ed.). Upper Saddle River, NJ: Prentice Hall.

Heskes, T. (1998). Bias/variance decomposition of likelihood-based estimator. *Neural Computation, 10*, 1425–1433.

Hinton, G. E. (1989). Connectionist learning procedure. *Artificial Intelligence, 40*, 185–234.

Hinton, G. E., & van Camp, D. (1993). Keeping neural networks simple by minimizing the description length of the weights. *Proc. Sixth ACM Conf. Computational Learning Theory* (pp. 5–13). San Cruz, CA:

Hochreiter, S., & Schmidhuber, J. (1997). Flat minima. *Neural Computation, 9*, 1–42.

Hochreiter, S., & Schmidhuber, J. (1998). Source separation as a by-product of regularization. In M. S. Kearns, S. Solla, & D. Cohn (Eds.), *Advances in neural information processing systems, 11* (pp. 459–465). Cambridge, MA: MIT Press.

Huber, P. (1981). *Rubust statistics*. New York: Wiley.

Johansen, T. A. (1997). On Tikhonov regularization, bias and variance in nonlinear system identification. *Automatica, 30*(3), 441–446.

Jou, I. C., You, S. S., & Chang, L. W. (1994). Analysis of hidden nodes for multilayer perceptron neural networks. *Pattern Recognition, 27*(6), 859–864.

Kailath, T. (1971). RKHS approach to detection and estimation problems—Part I: deterministic signals in gaussian noise. *IEEE Transactions on Information Theory, 17*(5), 530–549.

Kamimura, R. (1997). Information controller to maximize and minimize information. *Neural Computation, 9*, 1357–1380.

Kanjilal, P. P., & Banerjee, D. N. (1995). On the application of orthogonal transformation for the design and analysis of feedforward networks. *IEEE Transactions on Neural Networks, 6*(5), 1061–1070.

Kimeldorf, G. S., & Wahba, G. (1970). A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. *Annals of Mathematical Statistics, 41*(2), 495–502.

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

Koene, R., & Takane, Y. (1999). Discriminant component pruning: Regularization and interpretation of multilayered backpropagation networks. *Neural Computation, 11*, 783–802.

Koulouris, A., Bakshi, B. R., & Stephanopoulous, G. (1995). Empirical learning through neural networks: The wave-net solution. In G. Stephanopoulous & C. Han (Eds.), *Intelligent systems in process engineering* (pp. 437–484). San Diego, CA: Academic Press.

Kress, R. (1989). *Linear integral equations*. Berlin: Springer-Verlag.

Krzyzak, A., & Linder, T. (1996). Radial basis function networks and complexity regularization in function learning. In M. Mozer, M. Jordan, & T. Petsche (Eds.), *Advances in neural information processing systems, 9* (pp. 197–203). Cambridge, MA: MIT Press.

Lanczos, C. (1961). *Linear differential operators*. New York: D. Van Nostrand.

LeCun, Y., Denker, J. S., & Solla, S. A. (1990). Optimal brain damage. In D. Touretzky (Ed.), *Advances in neural information processing systems, 2* (pp. 598–605), CA: Morgan Kaufmann.

Leen, T. K. (1995). From data distributions to regularization in invariant learning. *Neural Computation, 7*, 974–981.

Lemm, J. C. (1996). *Prior information and generalized questions* (AI Memo 1598). Cambridge, MA: MIT.

Lemm, J. C. (1998). *How to implement a priori information: A statistical mechanics approach* (Tech. Rep. No. MS-TPI-98-12). Institute for Theoretical Physics, Münster University. Available on-line at: http://pauli.uni-muenster.de/~lemm.

Leung, C.-T., & Chow, T. (1997). A novel noise robust fourth-order cumulants cost function. *Neurocomputing, 16*(2), 139–147.

Leung, C.-T., & Chow, T. (1999). Adaptive regularization parameter selection method for enhancing generalization capability of neural networks. *Artificial Intelligence, 107*, 347–356.

Leung, C.-T., & Chow, T. (2001). Least third-order cumulant method with adaptive regularization parameter selection for neural networks. *Artificial Intelligence, 127*(2), 169–197.

Levin, A., Leen, T., & Moody, J. (1994). Fast pruning using principal components. In J. D. Cowan, G. Tessauro, & J. Alspetor (Eds.), *Advances in neural information processing systems, 6* (pp. 35–42). Cambridge, MA: MIT Press.

MacKay, D. J. C. (1992). *Bayesian methods for adaptive models*. Unpublished doctoral dissertation, California Institute of Technology. Available on-line at: http://wol.ra.phy.cam.ac.uk/mackay/.

MacKay, D. J. C. (1998). Introduction to gaussian processes. In C. Bishop (Ed.), *Neural networks and machine learning* (pp. 134–165). Berlin: Springer-Verlag.

Magdon-Ismail, M. (2000). No free lunch for noise prediction. *Neural Computation, 12*, 547–564.

Mallat, S., & Zhang, Z. (1993). Matching pursuit in a time-frequency dictionary. *IEEE Transactions on Signal Processing, 41*, 3397–3415.

Marroquin, J. L., Mitter, S., & Poggio, S. (1987). Probabilistic solution of ill-posed problems in computational vision. *Journal of American Statistical Association, 82*, 76–89.

Micchelli, C. A. (1986). Interpolation of scattered data: Distance matrices and conditionally positive definite functions. *Constructive Approximation, 2*, 11–22.

Moody, J., & Darken, C. J. (1989). Fast learning in networks of locally-tuned processing units. *Neural Computation, 1*, 281–294.

Moody, J., & Rögnvaldsson, T. (1997). Smoothing regularizers for projective basis function networks. In M. Mozer, M. Jordan, & T. Petsche (Eds.), *Advances in neural information processing systems, 9* (pp. 585–591), Cambridge, MA: MIT Press.

Morozov, V. A. (1984). *Methods for solving incorrectly posed problems*. New York: Springer-Verlag.

Mukherjee, S. & Nayar, S. K. (1996). Automatic generalization of RBF networks using wavelets. *Pattern Recognition, 29*(8), 1369–1383.

Neal, R. (1996). *Bayesian learning for neural networks*. Berlin: Springer-Verlag.

Niyogi, P., & Girosi, F. (1996). On the relationship between generalization error, hypothesis complexity, and sample complexity for radial basis function. *Neural Computation, 8*, 819–842.

Niyogi, P., & Girosi, F. (1999). Generalization bounds for function approximation from scattered noisy data. *Advances in Computational Mathematics, 10*, 51–80.

Niyogi, P., Girosi, F., & Poggio, T. (1998). Incorporating prior information in machine learning by creating virtual examples. *Proceedings of the IEEE, 86*(11), 2196–2208.

Nowlan, S. J., & Hinton, G. E. (1992). Simplifying neural networks by soft weight-sharing. *Neural Computation, 4*, 473–493.

Olshausen, B. A., & Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature, 381*, 607–609.

O'Sullivan, F. (1986). A statistical perspective on ill posed inverse problems (with discussions). *Statistical Science, 1*, 502–527.

Papageorgiou, C., Girosi, F., & Poggio, F. (1998). *Sparse correlation kernel analysis and reconstruction* (AI Memo 1635). Cambridge, MA: MIT.

Parzen, E. (1961). An approach to time series analysis. *Annals of Mathematical Statistics, 32*, 951–989.

Parzen, E. (1963). Probability density functionals and reproducing kernel Hilbert spaces. In M. Rosenblatt (Ed.), *Proc. Symposium of Time Series Analysis* (pp. 155–169). New York: Wiley.

Pearlmutter, B. A., & Rosenfeld, R. (1991). Chaitin-Kolmogorov complexity and generalization in neural networks. In D. Touretzky & R. Lippmann (Eds.), *Advances in neural information processing systems, 3* (pp. 925–931). San Mateo, CA: Morgan Kaufmann.

Poggio, T., & Girosi, F. (1990a). Networks for approximation and learning. *Proceedings of the IEEE, 78*(10), 1481–1497.

Poggio, T., & Girosi, F. (1990b). Regularization algorithms for learning that are equivalent to multilayer networks. *Science, 247*, 978–982.

Poggio, T., & Girosi, F. (1998). A sparse representation for function approximation. *Neural Computation, 10*, 1445–1454.

Poggio, T., & Koch, C. (1985). Ill-posed problems in early vision: From computational theory to analogue networks. *Proceedings of the Royal Society of London, B, 226*, 303–323.

Poggio, T., Mukherjee, S., Rifkin, R., Rakhlin, A., & Verri, A. (2001). *b* (CBCL Paper 198/AI Memo 2001-011). Cambridge, MA: MIT.

Poggio, T., Torre, V., & Koch, C. (1985). Computational vision and regularization theory. *Nature, 317*, 314–319.

Poggio, T., Voorhees, H., & Yuille, A. (1988). A regularized solution to edge detection. *Journal of Complexity, 4*, 106–123.

Powell, M. J. D. (1985). Radial basis functions for multivariable interpolation: A review. In *IMA Conference on Algorithms for the Approximation of Functions and Data* (pp. 143–167). RMCS, Shrivenham, England.

Reed, R. (1993). Pruning algorithms—A review. *IEEE Transactions on Neural Networks, 4*, 740–747.

Reed, R., Marks, R. J., & Oh, S. (1995). Similarities of error regularization, sigmoid gain scaling, target smoothing, and training with jitter. *IEEE Transactions on Neural Networks, 6*, 529–538.

Rissaen, J. (1978). Modeling by shortest data description. *Automatica, 14*, 465–471.

Rohwer, R., & van der Rest, J. C. (1996). Minimum description length, regularization, and multimodal data. *Neural Computation, 8*, 595–609.

Sanger, D. (1989). Contribution analysis: A technique for assigning responsibilities to hidden units in connectionist networks. *Connection Science, 1*, 115–138.

Schmidhuber, J. (1994). *Discovering problem solutions with low Kolmogorov complexity and high generalization capability* (Tech. Rep. FKI-194-94). Munich: Technische Universitat München. Available on-line at: http://papa.informatik.tu-muenchen.de/mitarbeiter/schmidhu.html.

Schnorr, C., & Sprengel, R. (1994). A nonlinear regularization approach to early vision. *Biological Cybernetics, 72*, 141–149.

Schölkopf, B., & Smola, A. (2002). *Learning with kernels: Support vector machines, regularization, optimization and beyond*. Cambridge, MA: MIT Press.

Setiono, R. (1997). A penalty function approach for pruning feedforward neural networks. *Neural Computation, 9*, 185–204.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal, 27*, 379–423, 623–656.

Silverman, B. W. (1984). Spline smoothing: The equivalent variable kernel method. *Annals of Statistics, 12*, 898–916.

Simard, P., LeCun, Y., Denker, J., & Victorri, B. (1998). Transformation invariance in pattern recognition: Tangent distance and tangent propagation. In G. B. Orr, & K-R. Müller (Eds.), *Neural networks: Tricks of the trade*. New York: Springer-Verlag.

Simard, P., Victorri, B., LeCun, Y., & Denker, J. (1992). Tangent prop—A formalism for specifying selected invariances in an adaptive network. In J. E. Moody, S. J. Hanson, & R. P. Lippmann (Eds.), *Advances in neural information processing systems, 4* (pp. 895–903). Cambridge, MA: MIT Press.

Smola, A., Schölkopf, B., & Müller, K.-R. (1998). The connection between regularization operators and support vector kernels. *Neural Networks, 11*, 637–649.

Thimm, G., Moerland, P., & Fiesler, E. (1996). The interchangeability of learning rate and gain in backpropagation neural networks. *Neural Computation, 8*, 451–460.

Tikhonov, A. N., & Arsenin, V. Y. (1977). *Solution of ill-posed problems*. Washington, DC: V. H. Winston.

van Wyk, M. A., & Durrani, T. S. (2000). A framework for multiscale and hybrid RKHS-based approximators. *IEEE Transactions on Signal Processing, 48*(12), 3559–3568.

Vapnik, V. (1995). *The nature of statistical learning theory*. New York: Springer-Verlag.

Vapnik, V. (1998a). *Statistical learning theory*. New York: Wiley.

Vapnik, V. (1998b). The support vector method of function estimation. In C. Bishop (Ed.) *Neural networks and machine learning* (pp. 239–268). New York: Springer-Verlag.

Velipasaoglu, E. O., Sun, H., & Zhang, F., Berrier, K. L., & Khoury, D. S. (2000). Spatial regularization of the electrocardiographic inverse problem and its application to endocardial mapping. *IEEE Transactions on Biomedical Engineering, 47*(3), 327–337.

Vitanyi, P. M. B., & Li, M. (2000). Minimum description length induction, Bayesianism, and Kolmogorov complexity. *IEEE Transactions on Information Theory, 46*(2), 446–464.

Wahba, G. (1990). *Spline models for observational data*. Philadephia: SIAM.

Wahba, G. (1995). Generalization and regularization in nonlinear system. In M. Arbib (Ed.), *Handbook of brain theory and neural networks* (pp. 426–432). Cambridge, MA: MIT Press.

Wahba, G. (1999). Support vector machines, reproducing kernel Hilbert spaces, and randomized GACV. In B. Schölkopf, C. Burges, & A. Smola (Eds.), *Advances in kernel methods—Support vector learning* (pp. 69–87). Cambridge, MA: MIT Press.

Watanabe, S. (1981). Pattern recognition as a quest for minimum entropy. *Pattern Recognition, 13*(5), 381–387.

Watanabe, K., Namatame, A., & Kashiwaghi, E. (1993). A mathematical foundation on Poggio's regularization theory. In *Proc. Int. Joint Conf. Neural Networks* (pp. 1717–1722). Nagoya, Japan.

Weigend, A., Rumelhart, D., & Humberman, B. A. (1991). Generalization by weight-elimination applied to currency exchange rate prediction. In *Proc. Int. Joint Conf. Neural Networks* (pp. 2374–2379). Singapore.

Williams, C. K. I. (1998a). Prediction with gaussian processes: From linear regression to linear prediction and beyond. In M. Jordan (Ed.), *Learning in graphical models*. Cambridge, MA: MIT Press.

Williams, C. K. I. (1998b). Computation with infinite neural networks. *Neural Computation, 10*, 1203–1216.

Williams, P. M. (1994). Bayesian regularization and pruning using a Laplace prior. *Neural Computation, 6*, 117–143.

Williamson, R. C., Smola, A., & Schölkopf, B. (2001). Generalization performance of regularization networks and support vector machines via entropy numbers of compact operators. *IEEE Transactions on Information Theory, 47*(6), 2516–2532.

Wolpert, D. H. (1996). The lack of a priori distinction between learning algorithms. *Neural Computation, 8*, 1341–1390.

Wolpert, D. H. (1997). On bias plus variance. *Neural Computation, 9*, 1211–1243.

Wolpert, D. H., & Macready, W. G. (1997). On free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation, 1*, 67–82.

Wu, L., & Moody, J. (1996). A smoothing regularizer for feedforward and recurrent neural network. *Neural Computation, 8*, 461–489.

Xu, L., Krzyzak, A., & Yullie, A. (1994). On radial basis function nets and kernel regression: Statistical consistency, convergence rates, and receptive field size. *Neural Networks, 7*(4), 609–628.

Yee, P. (1998). *Regularized radial basis function networks: Theory and applications to probability estimation, classification, and time series prediction*. Unpublished doctoral dissertation, McMaster University.

Yee, P., & Haykin, S. (2001). *Regularized radial basis function networks: Theory and applications*. New York: Wiley.

Yosida, K. (1978). *Functional analysis* (5th ed.). New York: Springer-Verlag.

Yuille, A., & Grzywacz, N. (1988). The motion coherence theory. In *Proc. Int. Conf. Computer Vision* (pp. 344–354). Washington, DC.

Zhu, H. (1996). No free lunch for cross validation. *Neural Computation, 8*, 1421–1426.

Zhu, H., & Rohwer, R. (1995). *Information geometric measurement of generalisation* (Tech. Rep., NCRG4350). Aston University. Available on-line at: http://www.ncrg.aston.ac.uk/Papers/.

Zhu, H., & Rohwer, R. (1996). Bayesian regression filters and the issue of priors. *Neural Computing and Applications, 4*, 130–142.

Zhu, H., Williams, C. K. I., Rohwer, R., & Morciniec, M. (1998). Gaussian regression and optimal finite dimensional linear models. In C. Bishop (Ed.), *Neural networks and machine learning* (pp. 167–184). New York: Springer-Verlag.

Received September 25, 2001; accepted May 28, 2002.