On Different Facets of Regularization Theory

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

Dedicated to the memory of Claude E. Shannon (1916-2001)

This paper provides a new viewpoint on regularization theory from different perspectives. It is shown that the solution of regularization problems can be derived from the Fourier operator in transformation domain with equivalent form from linear differential operator in spatial domain. State-of-the-art research in regularization theory are reviewed, with extended discussions on Occam's razor, minimum length description, Bayesian framework, pruning algorithms, statistical physics, informational coding (entropy) theory, statistical learning theory, regularization networks, equivalent regularization and early stopping. The universal principle of regularization in terms of Kolmogorov complexity is also explored. Finally, some prospective studies on regularization theory are suggested.

"No more things should be presumed to exist than are absolutely necessary."
- W. Occam

1 Introduction

Most of the inverse problems encountered in science and engineering areas are ill-posed, e.g. computational vision (Poggio, Torre, & Koch, 1985), system identification (Akaike, 1974; Johansen, 1997), nonlinear dynamics reconstruction (Haykin, 1999), image restoration (Velipasaoglu, Sun, & Zhang et al., 2000), and density estimation (Vapnik, 1998). In other words, given the available input data, the solution to the problem is non-unique (one-to-many) or nonstable. Regularization techniques, developed by Tikhonov in 1960's (Tikhonov & Arsenin, 1977), have been shown to be powerful in making the solution well-posed and have been applied successfully in model selection and complexity control. Specially, regularization was introduced to the machine learning community (Poggio & Girosi, 1990a, 1990b; Barron, 1991), and it was shown (Poggio & Girosi, 1990a, 1990b) that regularization algorithm for learning is equivalent to multilayer network with the radial basis function, namely radial basis function (RBF) network. A large class of generalized regularization networks are reviewed in (Girosi, Jones, & Poggio, 1995). It is noted that the original regularization solution was derived from the spatial domain by differential linear operator and Green's function (Poggio & Girosi, 1990a, 1990b; Watanabe, Namatame, & Kashiwaghi, 1993). However, it is observed that an equivalent regularization framework can be derived from the Fourier domain, in some sense we call spectral regularization (Chen & Haykin, 2001a, 2001b; Chen, 2001a). At the time being, it seems necessary to retrospect some important results of regularization theory from some new perspectives. Hence, the intention of this paper is not only reviewing the regularization theory, but also examining the relationship of regularization theory and other theoretical studies as well as presenting some new results.

In this paper, we derive theoretically the spectral regularization framework following similar steps as in (Poggio & Girosi, 1990a; Haykin, 1999), and we also provide some insightful discussions on the various regularizers as well as their geometrical and physical interpretation. State-of-the-art regularization approaches are reviewed. The connection between regularization to Occam's razor, minimum length description (MDL) principle, Bayesian theory, information theory, statistical physics is examined. The regularization networks (RNs) and their links to statistical learning theory and support vector machines (SVMs). The equivalent regularization techniques, particularly early stopping in machine learning are briefly addressed. The relation of the Kolmogorov complexity principle to regularization is also discussed.

The rest of paper is organized as follows: Section 2 briefly formulates the ill-posed problem and introduce the necessity of regularization theory as the solution. Section 3 introduces the regularization theoretical framework in the language of machine learning and the traditional Tikhonov regularization theory. Following the theory of Green's function and Fourier analysis, we derive an equivalent theoretical framework for spectral (transformation) regularization, geometrical interpretation on spatial and spectral regularization is given. Starting with Occam's razor and MDL theory in section 4, we present different facets of regularization theory in literature from sections 5 to 10, the state-of-the-art research are briefly reviewed and their connection to regularization theory are highlighted, which cover Bayesian theory, information theory, statistical physics, statistical learning theory, pruning algorithms, and equivalent regularization techniques. Finally, the universal principle of Kolmogorov complexity for regularization is explored in section 11 followed by a summary and comments in section 12.

2 Why Regularization: The Solution to Ill-posed Problems

A problem is said to be well-posed in the Hadamard sense if it satisfies the following three conditions (Haykin & Principe, 1998; Dontchev & Zolezzi, 1993):

1. *Existence*. For every input vector $\mathbf{x} \in X$, there exists an output vector $y = F(\mathbf{x})$, where $y \in Y$.

2. *Uniqueness*. For any pair of input vectors $\mathbf{x}, \mathbf{z} \in X$, it follows $F(\mathbf{x}) = F(\mathbf{z})$ if and only if $\mathbf{x} = \mathbf{z}$.

3. *Continuity*. The mapping is continuous, that is, for any $\epsilon > 0$ there exists $d = d(\varepsilon)$ such that the condition $d_X(\mathbf{x}, \mathbf{z}) < \delta$ implies that $d_Y(F(\mathbf{x}), F(\mathbf{z})) < \varepsilon$, where $d(\cdot, \cdot)$ represents the distance metric between two arguments.

If any of above three conditions is *not* satisfied, the problem is said to be *ill-posed*. Unfortunately, most real-world problems are ill-posed, following are a few examples we want to mention in the real world.

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, e.g. shape from shading, surface reconstruction, edge detection, computation of optical flow, are generally ill-posed (Poggio, Torre, & Koch, 1985; Bertero, Poggio, & Torre, 1988). In general, the solution of this problem can be formulated by (Poggio, Torre, & Koch, 1985)

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

where the first term is aimed to satisfy the constraints $Ax = y$, the second term plays the role of stabilizing functional, and $\|\cdot\|$ is some kind of norm operator dependent on specific physical situation.

2. Dynamic reconstruction: Nonlinear dynamic reconstruction (e.g. sea clutter) problem is a difficult but fundamental problem in many areas (Haykin & Principe, 1998; Haykin, 1999). Generally, the nonlinear dynamics can be formulated by the following state-space models

$$\mathbf{x}_{t+1} = F(\mathbf{x}_t, \mathbf{w}_t) \tag{2.2}$$

$$y_t = G(\mathbf{x}_t) + m_t \tag{2.3}$$

where F is nonlinear mapping function $\mathbb{R}^N \to \mathbb{R}^N$, G is a scalar-valued function, and $\mathbf{w}_t$ and m_t, represent the process noise and measurement noise contaminating the state variable $\mathbf{x}_t$ and the observable y_t. Now 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: Firstly, for some unknown reasons the existence condition may be violated; secondly, there may not be sufficient information in the observation (time series) for reconstructing the nonlinear dynamics uniquely, which thus violates the uniqueness condition; thirdly, the unavoidable presence of noise ($\mathbf{w}_t$ as well as m_t) adds uncertainty to the dynamic reconstruction, when the signal-to-noise ratio (SNR) is too low, the continuity condition is also possibly violated.

3. Density estimation: Density estimation is a general and basic problem. Suppose the observed data are sampled by the density $p(x)$ from the distribution function $F(x)$, which is related to each other by

$$\int_{-\infty}^{x} p(\tau) d\tau = F(x). \tag{2.4}$$

Now the problem is formulated as: given some data x_i $(i = 1, \cdots, \ell)$, how to estimate $p(x)$ from a finite number of (noisy or noiseless) observations? Empirically, one may estimate the distribution function by

$$F_\ell(x) = \frac{1}{\ell} \sum_{i=1}^{\ell} \Theta(x - x_i) \tag{2.5}$$

and the density is further estimated by solving the equation

$$\int_{-\infty}^{x} p(\tau) d\tau = F_\ell(x), \tag{2.6}$$

which is generally an ill-posed problem by solving the inverse operator $Ax = y$, especially in high-dimensional case (Vapnik, 1998).

In order to handle the ill-posed problems mentioned and the others, one way to make the problems more well-posed is to incorporate some prior knowledge into the solution (Tikhonov & Arsenin, 1977; Wahba, 1990; Dontchev & Zolezzi, 1993; Vapnik, 1998; Haykin, 1999). The forms of prior knowledge vary and are problem-dependent, the most popular and important prior knowledge is the smoothness prior, which assumes that the the functional mapping from input space to output space is usually continuous and smooth (usually measured by the order of differentiability), that is why the regularization naturally comes in arising from the well-known Tikhonov's regularization theory [1], which we will describe in detail in the following.

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 $f(\mathbf{x}_i) = y(\mathbf{x}_i) = y_i$, where $y(\mathbf{x})$ is supposed to be a deterministic function in the hypothesis space. In another viewpoint, this is also a interpolation problem, the f is an interpolant parameterized by the weights $\mathbf{w}$. Note that this problem is ill-conditioned in that the approximants satisfying the constraints are not unique. To solve the ill-posed problem we usually require the solution to be smooth, that is why regularization comes in.

Statistically speaking, the approximation accuracy is measured by the expectation of the approximation error. In Hilbert space $\mathbb{H}$, 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, the 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 some unknown probability metric $\mu(\varepsilon)$. In that case, $p(\mathbf{x}, y) = p(\mathbf{x})p(y|\mathbf{x})$ and $p(y|\mathbf{x})$ is represented by the metric function $\mu[y - f(\mathbf{x})]$. Hence, the expected cost 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

[1] Another methodology to embed prior knowledge is the theory of hints, see (Abu-Mostafa, 1995) for more information.

on finite observations (ℓ) is used instead, with an empirical risk functional

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

which produces an estimate $\hat{y}(\mathbf{x})$ ($\hat{y} = y - \varepsilon = f(\mathbf{x})$). Quantitatively, for all $0 \leq \eta \leq 1$, for the loss taking the value η, the generalization error has the upper bound (Vapnik, 1998a)

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

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

3.2 Tikhonov regularization

In regularization theory, the expected risk is decomposed into two parts, empirical risk (L_2 norm) $\mathcal{R}_{emp}$ and the regularizer risk $\mathcal{R}_{reg}$:

$$\begin{aligned} \mathcal{R}[f] &= \mathcal{R}_{emp}[f] + \lambda \mathcal{R}_{reg}[f] && (3.5) \\ &= \frac{1}{2}\sum_{i=1}^{\ell}[y_i - f(\mathbf{x}_i)]^2 + \frac{1}{2}\lambda\|\mathbf{D}f\|^2 && (3.6) \end{aligned}$$

where $\|\cdot\|$ is the norm operator [2]. λ is the regularization parameter which 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 *Fréchet differential* of Tikhonov functional (Tikhonov & Arsenin, 1977; Poggio & Girosi, 1990a; Haykin, 1999). Geometrically, $\mathbf{D}$ is interpreted as the local linear approximation of the curve in the space. The smoothness prior embedded in $\mathbf{D}$ makes the solution more stable and smoother.

Since *Fréchet differential* is regarded as the best local linear approximation of a functional, it can be defined as (Tikhonov & Arsenin, 1977)

$$d\mathcal{R}(f,h) = \frac{d}{d\beta}\mathcal{R}(f + \beta h)|_{\beta=0}, \tag{3.7}$$

where $h(\mathbf{x})$ is a constant fixed function of $\mathbf{x}$. Following the steps in (Haykin, 1999), we have

$$\begin{aligned} d\mathcal{R}_{emp}(f,h) &= \frac{d}{d\beta}\mathcal{R}_{emp}(f + \beta h) \,|_{\beta=0} \\ &= -\sum_{i=1}^{\ell}[y_i - f(\mathbf{x}_i)]h(\mathbf{x}_i) \\ &= -\Big\langle h, \sum_{i=1}^{\ell}(y_i - f(\mathbf{x}_i))\delta(\mathbf{x} - \mathbf{x}_i)\Big\rangle, && (3.8) \end{aligned}$$

[2] It is usually referred to L_2 norm in Hilbert space if not stated otherwise. Also note that it can be defined in a particular form in Sobolev space and Besov space (Chen & Haykin, 2001c).

where $< \cdot, \cdot >$ denotes the inner product of two functions in the Hilbert space $\mathbb{H}$. Similarly, the *Fréchet differential* of the regularizing term $\mathcal{R}_{reg}$ is led to

$$\begin{aligned} d\mathcal{R}_{reg}(f,h) &= \frac{d}{d\beta}\mathcal{R}_{reg}(f+\beta h)\mid_{\beta=0} \\ &= \int \mathbf{D}[f+\beta h]\mathbf{D}h\, d\mathbf{x}\mid_{\beta=0} \\ &= \int \mathbf{D}f\mathbf{D}h\, d\mathbf{x} = \big\langle \mathbf{D}h, \mathbf{D}f \big\rangle. \end{aligned} \tag{3.9}$$

The above results are well-known for the spatial domain, for the details of proof, see (Haykin, 1999).

3.3 Green's function

In analogy to the inverse of a matrix, Green's functions represent the inverse of a (sufficiently regular) differential operator (Lanczos, 1961). For a large class of problem, it appears in the form of a kernel function which 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 which has Dirac delta function on the right side, the *Reciprocity Theorem* makes it possible to define Green's function either in terms of the adjoint or given differential operator (see Lanczos, 1961, chapter 5 for detail).

Given a linear differential operator $\mathbf{L}$, the function $G(\mathbf{x}, \boldsymbol{\xi})$ is said to be Green's function if it has the following properties (Courant & Hilbert, 1970; Haykin, 1999):

- For a fixed ξ, $G(\mathbf{x}, \boldsymbol{\xi})$ is a function of $\mathbf{x}$ and satisfies the given boundary condition.
- 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

$$\mathbf{L}G(\mathbf{x}, \boldsymbol{\xi}) = \delta(\mathbf{x} - \boldsymbol{\xi}) \tag{3.10}$$

  where δ is Dirac delta function.

Hence, the Green's function plays a role for the linear differential operator that is similar to that for the inverse matrix for a matrix equation (Haykin, 1999).

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

$$F(\mathbf{x}) = \int_{\mathbb{R}^N} G(\mathbf{x}, \boldsymbol{\xi})\varphi(\boldsymbol{\xi})d\boldsymbol{\xi} \tag{3.11}$$

is the solution of the differential equation

$$\mathbf{L}F(\mathbf{x}) = \varphi(\mathbf{x}) \tag{3.12}$$

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

3.4 Fourier analysis and spectral regularization

In what follows, we will prove from transformation (spectral) domain of the equivalent regularizer, which firstly starts with a definition and a theorem:

Definition 1 *Given a Fourier operator* $\mathbf{T}$ *in functional space, for all functionals* $f \in \mathbb{H}$*, we always have* $\mathbf{T}f \in \mathbb{H}$. (It can be proven that Fourier operator $\mathbf{T}$ is a linear integral operator.)

Theorem 1 Plancherel Identity. *Given a Fourier transform of some functional in* $\mathbb{H}$*, the Plancherel Identity (Parseval theorem) states that* $\langle f, g \rangle = \frac{1}{2\pi}\langle \mathcal{F}(\mathbf{s}), \mathcal{G}(\mathbf{s}) \rangle$*, where* $\mathcal{F}$ *and* $\mathcal{G}$ *are Fourier transform of functionals of* $f(\mathbf{x})$ *and* $g(\mathbf{x})$ *respectively. In operator form, it can be written as* $\langle f, g \rangle = \langle \mathbf{T}f, \mathbf{T}g \rangle$ [3] *.*

Remarks: Basically, the transformation operator is an integral operator with the form

$$\mathbf{T}f(\mathbf{s}) = \int_{\mathbb{R}^N} f(\mathbf{x}) K(\mathbf{s}, \mathbf{x}) d\mathbf{x} \tag{3.13}$$

in which $K(\mathbf{s}, \mathbf{x})$ is a generic kernel function. In the case of Fourier operator, $K(\mathbf{s}, \mathbf{x}) = \exp(-j < \mathbf{s}, \mathbf{x} >)$ where $j = \sqrt{-1}$. If we define the differential operator $\mathbf{D}$ as

$$\mathbf{D} = \sum_{-\infty}^{\infty} \frac{(-1)^n}{n!} \frac{d^n}{dx^n}, \tag{3.14}$$

then the corresponding transformation operator is

$$\mathbf{T} = \sum_{-\infty}^{\infty} \frac{(-1)^n (js)^n}{n!} = \exp(-js). \tag{3.15}$$

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

$$K(\theta) = \frac{\sin(n + \frac{1}{2})\theta}{2\pi \sin \frac{\theta}{2}}. \tag{3.16}$$

From 3.16, the truncated Fourier series are written by

$$f_n(x) = \int_{-\pi}^{\pi} f(s) K_n(s, x) ds \tag{3.17}$$

where

$$\begin{aligned} K_n(s, x) &= \frac{1}{\pi} \sum_{k=0}^{n \to \infty} (\cos ns \cos nx + \sin ns \sin nx) \\ &= \frac{1}{\pi} \sum_{k=0}^{n} \cos k(s - x) = K_n(s - x) \end{aligned} \tag{3.18}$$

[3] For the simplicity of notation, we henceforth take all the constants that appear in the definitions of (inverse) Fourier transform to be 1.

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

$$f(x) = \int_{-\pi}^{\pi} f(s)K(s-x)ds. \tag{3.19}$$

In particular, it is found that the Dirac-delta function $\delta(s,x)$ has the series 3.17 as its Fourier expansion (see Appendix A for proof).

Example 2 Fejér kernel (Lanczos, 1961)

$$\Phi_n(\theta) = \frac{\sin^2 \frac{n}{2}\theta}{2\pi n \sin^2 \frac{\theta}{2}}, \tag{3.20}$$

which is the arithmetic mean of Dirichlet kernel. In other words, its Fourier coefficients are the weighted version of those of Dirichlet kernel dependent of n:

$$a'_k = (1-k/n)a_k, \quad b'_k = (1-k/n)b_k.$$

Other examples, such as *periodic Gaussian kernel*, *B-spline kernel*, and *regularized Fourier expansion kernel* are given in (Vapnik, 1998a, 1998b; Smola, Scholkopf, & Muller, 1998). It is noted that the Dirichlet kernel, Fejér kernel, B-spline kernel are interpolation-based kernels, whereas translationally invariant kernels (e.g. Fourier, Gaussian) are convolution-based kernels.

By the virtue of Theorem 1, it follows that

$$\langle \mathbf{D}h, \mathbf{D}f \rangle = \langle \mathbf{T_D}h, \mathbf{T_D}f \rangle \tag{3.21}$$

where $\mathbf{T_D}$ denotes the Fourier operator of differential operator $\mathbf{D}$. For ease of notation, we henceforth simply write $\mathbf{T_D}$ as T. Thus 3.9 is rewritten from 3.21

$$\begin{aligned} d\mathcal{R}_{reg}(f,h) &= \int \mathbf{D}f\mathbf{D}h \, d\mathbf{x} = \langle \mathbf{D}h, \mathbf{D}f \rangle \\ &= \int \mathbf{T}f\mathbf{T}h d\mathbf{s} = \langle \mathbf{T}h, \mathbf{T}f \rangle. \end{aligned} \tag{3.22}$$

For any pair of functions $u(\mathbf{x})$ and $v(\mathbf{x})$ and their corresponding Fourier pair $u(\mathbf{s})$ and $v(\mathbf{s})$, given the linear differential operator $\mathbf{D}$ and Fourier operator $\mathbf{T}$, their *self-adjoint* operators, $\tilde{\mathbf{D}}$ and $\tilde{\mathbf{T}}$, are uniquely determined to satisfy boundary conditions (Lanczos, 1961)

$$\int_{\mathbb{R}^N} u(\mathbf{x})\mathbf{D}v(\mathbf{x})d\mathbf{x} = \int_{\mathbb{R}^N} v(\mathbf{x})\tilde{\mathbf{D}}u(\mathbf{x})d\mathbf{x} \tag{3.23}$$

and

$$\int_{\Omega} u(\mathbf{s})\mathbf{T}v(\mathbf{s})d\mathbf{s} = \int_{\Omega} v(\mathbf{s})\tilde{\mathbf{T}}u(\mathbf{s})d\mathbf{s} \tag{3.24}$$

where Ω is the support in frequency domain. Equations 3.23 and 3.24 are called *Green's identity* (Lanczos, 1961; Poggio & Girosi, 1990a). Viewing $\mathbf{D}$ and $\mathbf{T}$ as matrices, their adjoint operators can be interpreted as the matrix transpose [4] (Lanczos, 1961; Haykin, 1999).

[4] Equations 3.23 and 3.24 are easily understood by observing $\mathbf{D}(u,v) = (\mathbf{D}u)v + (\mathbf{D}v)u$ and $\mathbf{T}(u,v) = \mathbf{T}(\mathbf{T}u,v) = \mathbf{T}(u,\mathbf{T}v)$, where $\mathbf{D}$ and $\mathbf{T}$ are differential and integral operators, respectively.

Using Green's identity, we further obtain the equivalent form of 3.22 from 3.23 by setting $u(\mathbf{x}) = \mathbf{D}f(\mathbf{x})$ and $\mathbf{D}v(\mathbf{x}) = \mathbf{D}h(\mathbf{x})$

$$d\mathcal{R}_{reg}(f,h) = \int h(\mathbf{x})\tilde{\mathbf{D}}\mathbf{D}f(\mathbf{x})d\mathbf{x} = \langle h, \tilde{\mathbf{D}}\mathbf{D}f\rangle(\mathbf{x}), \tag{3.25}$$

and another form from 3.24 by setting $u(\mathbf{s}) = \mathbf{T}f(\mathbf{s})$ and $\mathbf{T}v(\mathbf{s}) = \mathbf{T}h(\mathbf{s})$

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

And the Fr*é*chet differential

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

can be rewritten from 3.8 by virtue of 3.25 and 3.26 in the following forms

$$d\mathcal{R}(f,h) = \left\langle h(\mathbf{x}), \left[\tilde{\mathbf{D}}\mathbf{D}f(\mathbf{x}) - \frac{1}{\lambda}\sum_i^{\ell}(y_i - f)\delta(\mathbf{x}-\mathbf{x}_i)\right]\right\rangle, \tag{3.27}$$

or

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

where $\mathcal{F}(\cdot)$ denotes the Fourier transform. The necessary condition for $f(\mathbf{x})$ be the relative extremum of $\mathcal{R}$ is $d\mathcal{R} = 0$ for all $h \in \mathbb{H}$, hence from 3.27 and 3.28 we have

$$\tilde{\mathbf{D}}\mathbf{D}f_\lambda(\mathbf{x}) = \frac{1}{\lambda}\sum_{i=1}^{\ell}[y_i - f(\mathbf{x}_i)]\delta(\mathbf{x}-\mathbf{x}_i), \tag{3.29}$$

and

$$\tilde{\mathbf{T}}\mathbf{T}f_\lambda(\mathbf{s}) = \mathcal{F}\left\{\frac{1}{\lambda}\sum_{i=1}^{\ell}[y_i - f(\mathbf{x}_i)]\delta(\mathbf{x}-\mathbf{x}_i)\right\}. \tag{3.30}$$

Equations 3.29 is the *Euler-Lagrange equation* of Tikhonov functional $\mathcal{R}[f]$ and 3.30 is its Fourier counterpart. Denoting $\mathbf{L} = \tilde{\mathbf{D}}\mathbf{D}$ and $\mathbf{K} = \tilde{\mathbf{T}}\mathbf{T}$, $G(\mathbf{x}, \xi)$ be the Green's function of the linear differential operator, whose role is similar to the inverse matrix for a matrix equation (Haykin, 1999). From the properties of Green's function in spatial domain (Lanczos, 1961; Poggio & Girosi, 1990a; Haykin, 1999)

$$\mathbf{L}G(\mathbf{x}, \boldsymbol{\xi}) = \delta(\mathbf{x} - \boldsymbol{\xi}) \tag{3.31}$$

and its counterpart in frequency domain

$$\mathbf{K}G(\mathbf{s}, \boldsymbol{\xi}) = \exp(-j\mathbf{s}\boldsymbol{\xi}), \tag{3.32}$$

similar to 3.11, it can be proven

$$f(\mathbf{x}) = \int_{\mathbb{R}^N} G(\mathbf{x}, \xi)\varphi(\boldsymbol{\xi})d\boldsymbol{\xi} \tag{3.33}$$

is the solution of the following differential and integral equations

$$\begin{aligned} \mathbf{L}f(\mathbf{x}) &= \varphi(\mathbf{x}), && (3.34) \\ \mathbf{K}f(\mathbf{s}) &= \Phi(\mathbf{s}), && (3.35) \end{aligned}$$

where $\Phi(\mathbf{s})$ is the Fourier transform of $\varphi(\mathbf{x})$. From 3.31-3.33, we may derive the solution of regularization problem

$$f_\lambda(\mathbf{x}) = \sum_{i=1}^{\ell} w_i G(\mathbf{x}, \mathbf{x}_i) \tag{3.36}$$

where $w_i = [y_i - f(\mathbf{x}_i)]/\lambda$, and $G(\mathbf{x}, \mathbf{x}_i)$ is a positive-definite Green's function (the proof is given in Appendix B).

Remarks:

- Provided the self-adjoint operator in spatial domain is defined by (Poggio & Girosi, 1990a, 1990b; Haykin, 1999)

$$\mathbf{L} = \sum_{n=0}^{\infty} \frac{(-1)^n}{n!2^n} \nabla^{2n} \tag{3.37}$$

  where

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

  then the corresponding self-adjoint operator in spectral domain is

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

  and we further have

$$\begin{aligned} \mathbf{L}G(\mathbf{x}) &= \delta(\mathbf{x}), && (3.39) \\ \mathbf{K}G(\mathbf{s}) &= 1. && (3.40) \end{aligned}$$

- In general, the solution to the regularization problem is

$$f_\lambda(\mathbf{x}) = \sum_{i=1}^{\ell} w_i G(\mathbf{x}, \mathbf{x}_i) + \beta(\mathbf{x}) \tag{3.41}$$

  where $\beta(\mathbf{x})$ is a term that lies in the null space of regularized functional, which satisfies the orthogonal condition $\sum_{i=1}^{\ell} w_i \beta(\mathbf{x}_i) = 0$, hence the functional space of solution f_λ is a reproducing kernel Hilbert space (RKHS) of the direct sum of two orthogonal RKHS. For more mathematical treatment on RKHS, see (Wahba, 1990; Girosi, 1998; Vapnik, 1998 and the references therein).

- The spirit of regularization technique is actually finding a proper subspace, i.e. eigen-space of the operator $\mathbf{T}f$ (dependent on kernel function K), within the subspace the operator behaves like a "well-posed" operator (Lanczos, 1961). Also, the solution in the subspace is unique.

The above derivations states that the solution of regularization method is independent of the domain of regularizer, regularization is equivalent to the expansion of the solution in terms of a set of Green's function depending on $\mathbf{D}$ or $\mathbf{T}$ [5]. Note that some invariance properties are implicitly embedded in $\mathbf{T}$ (it is well known that the Fourier operator is invariant to shift, rotation, and starting point), this prior further implies that the derived Green's function should be *translationally and rotationally invariant*, in other words, $G(\mathbf{x}, \mathbf{x}_i)$ is inherently a radial basis function (RBF) with the radially symmetric and shift invariant form (Poggio & Girosi, 1990a; Haykin, 1999)

$$G(\mathbf{x}, \mathbf{x}_i) = G(\|\mathbf{x} - \mathbf{x}_i\|), \tag{3.42}$$

or more generally, it can be the generalized RBF (GRBF)

$$G(\mathbf{x}, \mathbf{x}_i) = G(|\mathbf{x} - \mathbf{x}_i|^T \Sigma^{-1} |\mathbf{x} - \mathbf{x}_i|), \tag{3.43}$$

where Σ is usually the covariance matrix. In particular, given the assumptions of 3.37 and 3.38, one may derive that

$$\begin{aligned} \mathcal{G}(\mathbf{s}) &= \exp(-\frac{\|\mathbf{s}\|^2}{2}), & (3.44) \\ G(\mathbf{x}) &= \exp(-\frac{\|\mathbf{x}\|^2}{2}), & (3.45) \end{aligned}$$

where $G(\mathbf{x}) \leftrightarrow \mathcal{G}(\mathbf{s})$ is a Fourier transform pair. The choices of Green's functions can be extended to the reproducing kernel functions in RKHS and a wide class of regularized RBF networks, which is called regularized networks (RNs) (Girosi, Jones, & Poggio, 1995). For example, the *periodic translationally invariant* Gaussian kernel can be

$$\begin{aligned} G(\mathbf{x}, \mathbf{x}_i) &= \exp(-j(\mathbf{x} - \mathbf{x}_i)) \exp(-\|\mathbf{x} - \mathbf{x}_i\|^2) \\ &= \cos(\mathbf{x} - \mathbf{x}_i) \exp(-\|\mathbf{x} - \mathbf{x}_i\|^2) \\ &\quad -j \sin(\mathbf{x} - \mathbf{x}_i) \exp(-\|\mathbf{x} - \mathbf{x}_i\|^2) \end{aligned} \tag{3.46}$$

The previous equation reminds us $G(\mathbf{x}, \mathbf{x}_i)$ is somehow similar to the Gabor function in time-frequency analysis (Strichartz, 1994).

So far, we may write a transformation version of spectral regularizer, being the counterparts of 3.5 and 3.6

$$\begin{aligned} \mathcal{R}[f] &= \mathcal{R}_{emp}[f] + \lambda \mathcal{R}_{reg}[\mathcal{F}] & (3.47) \\ &= \frac{1}{2} \sum_{i=1}^{\ell} [y_i - f(\mathbf{x}_i)]^2 + \frac{1}{2} \lambda \|\mathbf{T}f\|^2 & (3.48) \end{aligned}$$

where $f(\mathbf{x}) \leftrightarrow \mathcal{F}(\mathbf{s})$ is denoted by a Fourier pair. Note that 3.48 is also in line with the smoothness functional of spectral penalty $\int_{\mathbb{R}^N} |\mathcal{F}(\mathbf{s})|^2 / \mathcal{G}(\mathbf{s}) d\mathbf{s}$, where

[5] Girosi, Jones and Poggio (1995, Appendix A) also provided a different proof of solution to regularization problem from the smoothness functional.

$\mathcal{G}(\mathbf{s})$ is a low-pass filter with property $\mathcal{G}(\mathbf{s})|_{|\mathbf{s}|\to\infty} = 0$ (Girosi, Jones & Poggio, 1995).

The geometrical interpretation of Tikhonov (spatial) regularization is explicit: the smoothness is measured by its differentiability, tangent distance, or curvature; whereas the interpretation of spectral regularization can be viewed from its power spectrum: when the reconstructed functional is smooth, the spectral component is concentrated on the low frequency, hence much penalization will be put on the high frequency. In addition, the physical and biological interpretation of spectral reconstruction have been partly discussed earlier (Chen, 2001a).

3.5 Transformation regularization: numerical aspect

In the preceding part we have discussed the regularization in continuous case, however, in practice, we care more about the regularization problem in the standpoint of numerical calculation (Hansen, 1998), which we will discussed in the following. Basically, this approach is another kind of *transformation regularization*, different from the spectral regularization discussed above, the regularization is taken in the transformation domain (or subspace) by matrix decomposition (SVD, PCA or QR decomposition) instead of in the frequency domain (usually by taking kernel convolution operation). Consider the common spatial regularization, rewriting 3.36 as a matrix form in terms of cost functional

$$\mathcal{R} = \|\mathbf{y} - \mathbf{G}\mathbf{w}\|^2 + \lambda\|\mathbf{P}\mathbf{w}\|^2 \tag{3.49}$$

where $\mathbf{w} = [w_i, \cdots, w_\ell]^T$, $\mathbf{G} = G(\mathbf{x}, \mathbf{x}_i)$ is a radial basis matrix and $\mathbf{y} = [y_1, \cdots, y_\ell]^T$. $\mathbf{P}$ is a user-designed matrix for regularizer. Since $\mathbf{G}$ is usually ill-conditioned, we always do some matrix decomposition for reducing redundancy. Taking the singular value decomposition (SVD) of $\mathbf{G}$

$$\mathbf{G} = \mathbf{U}\mathbf{\Sigma}\mathbf{Z}^T \tag{3.50}$$

where $\mathbf{U}$ and $\mathbf{Z}$ are left and right singular vectors of $\mathbf{G}$, $\mathbf{\Sigma}$ is a diagonal matrix with singular value σ_i ($i = 1, \cdots, \ell$) on the diagonal. In the case of zero-order Tikhonov regularization (i.e. $\mathbf{P} = \mathbf{I}$ as identity matrix), suppose $\mathbf{w} = \mathbf{Z}\boldsymbol{\alpha}$, we obtain

$$(\mathbf{\Sigma}^T\mathbf{\Sigma} + \lambda\mathbf{I})\boldsymbol{\alpha} = \mathbf{\Sigma}^T\mathbf{d} \tag{3.51}$$

where $\mathbf{d} = \mathbf{U}^T\mathbf{y}$ is a vector of Fourier coefficients (Hansen, 1998). Furthermore, the spectral coefficients corresponding to zero-order and nonzero-order (e.g. $\mathbf{D} = \nabla$ as first-order gradient operator and $\mathbf{P} = \mathbf{J}(\mathbf{w})\mathbf{w}^{-1}$ where $\mathbf{J}(\mathbf{w})$ is Jacobian matrix, or $\mathbf{D} = \nabla^2$ as second-order Laplace operator and $\mathbf{P} = \mathbf{H}(\mathbf{w})\mathbf{w}^{-1}$ where $\mathbf{H}(\mathbf{w})$ is Hessian matrix) of Tikhonov regularization can be described as

$$c_i = \frac{\sigma_i^2}{\sigma_i^2 + \lambda}, \quad \alpha_i = \frac{c_i d_i}{\sigma_i}, \tag{3.52}$$

and

$$c_i = \frac{\sigma_i^2}{\sigma_i^2 + \lambda s_i^2}, \quad \alpha_i = \frac{c_i d_i}{\sigma_i}, \tag{3.53}$$

respectively, where s_i are defined as the diagonal elements of the diagonal matrix $\mathbf{S}$ by SVD of $\mathbf{P} = \mathbf{VSZ}^T$ (viewing the operator as a matrix). In nonzero-order Tikhonov regularization case, provided $\mathbf{w} = \mathbf{Z}\boldsymbol{\alpha}$ we have

$$(\boldsymbol{\Sigma}^T\boldsymbol{\Sigma} + \lambda\mathbf{S}^T\mathbf{S})\boldsymbol{\alpha} = \boldsymbol{\Sigma}^T\mathbf{d}. \tag{3.54}$$

The solution of 3.49 when $\mathbf{P} = \mathbf{I}$ can be computed explicitly by $\mathbf{w} = (G + \lambda\mathbf{I})^\dagger\mathbf{y}$, where $\dagger$ is Moore-Penrose pseduoinverse. More generally, we have the relationship

$$\mathbf{Z}^T\mathbf{w} = \boldsymbol{\Sigma}_\lambda^\dagger\mathbf{d} = \boldsymbol{\Sigma}_\lambda\mathbf{U}^T\mathbf{y}, \tag{3.55}$$

in which

$$\boldsymbol{\Sigma}_\lambda = \begin{cases} diag\left\{\frac{\sigma_i^2}{\sigma_i^2+\lambda}\right\}, & \mathbf{P} = \mathbf{I} \\ diag\left\{\frac{\gamma_i(\gamma_i^2+1)^{1/2}}{\gamma_i^2+\lambda}\right\}, & \mathbf{P} \neq \mathbf{I} \end{cases} \tag{3.56}$$

where $\gamma_i = (\sigma_i^2 + s_i^2)^{1/2}$ are the *generalized singular values* of the $(\boldsymbol{\Sigma}, \mathbf{S})$ (see Appendix C for 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 value σ_i. By reweighting the generalized singular values s_i *a posteriori* according to their contribution (through calculating the geometric mean of $|d_i|$), the weight coefficients can be devised to shape as the reciprocal of the energy spectrum of the data (Velipasaoglu, Sun, & Zhang, et al, 2000). Since the ill-posed matrix $\mathbf{G}$ usually has a wide range of singular values [6], the singular values corresponding to components with high energy and high SNR are supposed to penalize less, and those corresponding to components with low energy and low SNR are penalized more (Velipasaoglu, Sun, & Zhang, et al, 2000).

3.6 Choosing regularization parameter

In order to obtain an stable and convergent regularized solution, the choice of regularization parameter is critically important, there are several ways for choosing regularization parameter in practice.

According to (MacKay, 1992), regularization parameter λ can be estimated using Bayesian method with second evidence framework (the first evidence framework is estimating the posterior probability of weights while supposing λ is known, see section 5 for details). Given the data $\mathcal{D}$ and model $\mathcal{M}$, regularization parameter can be estimated by

$$p(\lambda|\mathcal{D}, \mathcal{M}) \propto p(\mathcal{D}|\lambda, \mathcal{M})p(\lambda|\mathcal{M}). \tag{3.57}$$

Regularization parameter can be also estimated by average-squared error or generalized cross-validation (GCV) approaches (Wahba, 1990; Haykin, 1999; Yee, 1998; Yee & Haykin, 2001). The advantage of GCV estimate over average-square error and ordinary cross-validation approaches lies that it is no need

[6] It can be alternatively used with QR decomposition to observe its singularity, which will be discussed in section 9.

of any prior knowledge of noise variance and it treats the observation data equally in the estimate. In (Yee, 1998; Yee & Haykin, 2001), it is shown that the regularized strict interpolation radial basis function network (SIRBFN) is asymptotically equivalent to *Nadaraya-Watson regression estimate* with mean-square consistent, the regularization parameter λ is allowed to vary with new observation.

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

$$\lambda \geq \frac{-\|\nabla \mathcal{R}_{emp}(\mathbf{w})\|^2}{< \nabla \mathcal{R}_{emp}(\mathbf{w}), \nabla \mathcal{R}_{reg}(\mathbf{w}) >}, \tag{3.58}$$

and

$$\lambda \geq \frac{< \nabla \mathcal{R}_{emp}(\mathbf{w}), \nabla \mathcal{R}_{reg}(\mathbf{w}) >}{-\|\nabla \mathcal{R}_{reg}(\mathbf{w})\|^2}, \tag{3.59}$$

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

4 Occam's Razor and MDL Principle

Philosophically speaking, Occam's razor is the principle which favors the shortest hypothesis that can explain well the observation. In AI and machine learning community, it is commonly used in the complexity control for data modelling, and naturally connected to regularization theory. MacKay (1992) provided a descriptive discussions on Occam's razor from Bayesian framework (see section 5 for discussion). The evidence can be approximately viewed as the product of a best fit likelihood and the Occam's factor. In functional approximation (regression) problem, Occam's razor expects to find the simplest and smooth-like model that can approximate or interpolate well the given observation data.

Minimum description length (MDL) principle is based on the information-theoretic analysis of the randomness concept (Rissaen, 1978; Haykin, 1999; Cherkassky & Mulier, 1998). The idea of MDL is to view the machine learning as a process of encoding information of observations in the model. The code length is viewed 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 viewed as being drawn independently from some unknown distribution, the learning problem is formulated as the dependency estimation of the y upon $\mathbf{x}$. Naturally, a metric measuring the complexity of the data length is given (Rissaen, 1978):

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

MDL principle is closely related to the regularization (e.g. Hinton & van Camp, 1993; Rohwer & van der Rest, 1996; MacKay, 1992; Cherkassky & Mulier, 1998). Hinton and van Camp (1993) found that the cases of weight decay and soft weight-sharing are vindication of the MDL approach. Basically, neural networks work like an encoder-decoder, the data and weights acting like information flow are transferred in the channel: hidden layers (see Figure 1 for illustration). Regularization is trying to keep the weight simple by penalizing the information they carry. The amount of information in the weights can be controlled by adding some noise with specific density, and the noise level can be adjusted during the learning process to optimize the trade-off between empirical error (misfit of data) and the amount of information in the weights (regularizer). Suppose the approximation error ε is Gaussian distributed with quantization width μ, one will have (Hinton & van Camp, 1993)

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

and the misfit of data is measured by the empirical risk (Hinton & van Camp, 1993)

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

where k is a constant depending on μ. Hence minimizing the squared error (the second term of right side of 4.3) is equivalent to MDL principle. Assuming the weights are white Gaussian distribution, the regularizer $L(\mathcal{M})$ of *weight decay* can be also written by MDL principle

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

in which the σ_w controlling the noise level acts like the regularization parameter. In the noisy weight case, MDL corresponds to introducing high variance in $L(\mathcal{D}|\mathcal{M})$, i.e. the misfit of data is less reliable. More generally, the weights may be assumed to be a mixture density $p(\mathbf{w}) = \sum_i \pi_i p(w_i)$, a detailed discussion is referred to (Hinton & van Camp, 1993; Bishop, 1995; Cherkassky & Mulier, 1998).

5 Bayesian Theory

Bayesian theory is an efficient approach dealing with the prior knowledge and is naturally connected to the choice of regularization operator (MacKay, 1992). This is not surprising since regularization theory has a good Bayesian interpretation (Poggio & Girosi, 1990a, 1990b). Many efforts have been put the Bayesian framework in the machine learning and model control (Mackay, 1992; Bruntine & Weigend, 1991; Williams, 1994).

Given the data $\mathcal{D}$, what we care about is finding the most probable model for the observation data, following Bayes formula, the *posterior* probability of model $\mathcal{M}$ is estimated by $p(\mathcal{M}|\mathcal{D}) = p(\mathcal{D}|\mathcal{M})p(\mathcal{M})/p(\mathcal{D})$, where the denominator represents the *evidence* as a normalizing constant. Maximizing $p(\mathcal{M}|\mathcal{D})$ is further equivalent to minimization of

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

where the first term corresponds to the MDL described in the previous subsection, and the second term represents the minimal code length for model $\mathcal{M}$.

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

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

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

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

and 5.2 is rewritten by

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

The priors $p(\mathbf{w}|\mathcal{M})$ is characterized as

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

where $Z_{\mathbf{w}}(\lambda) = \int d\mathbf{w} \exp(-\lambda \mathcal{R}_{reg}(\mathbf{w}))$. The choices of λ and $\mathcal{R}_{reg}(\mathbf{w})$ are often built 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 may lead to *maximum a posteriori* (MAP) estimate:

$$\begin{aligned} -\log p(\mathbf{w}|\mathcal{D},\mathcal{M}) = \quad & -\tfrac{\lambda}{2}\|\mathbf{w}\|^2 - \textstyle\sum_i \log p(\mathbf{x}_i) \\ & -\textstyle\sum_i \log p(y_i|\mathbf{x}_i,\mathbf{w}) \end{aligned} \tag{5.6}$$

when $p(y_i|\mathbf{x}_i,\mathbf{w})$ is usually measured by L_2 metric, 5.6 becomes

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

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

- weight decay (Hinton, 1989): it was shown from the Bayesian perspective (Bruntine & Weigend, 1991; MacKay, 1992) that weight dacay is equivalent to maximum likelihood estimate (MLE) under the Gaussian assumption. Weight decay is equivalent to the well-studied ridge regression in statistics (Wahba, 1990), which is a version of zero-order Tikhonov regularization.
- weight elimination (Weigend, Rumelhart, & Humberman, 1991): it was interpreted as negative log-likelihood prior probability of the weights. The weights are assumed to be a mixture of uniform and Gaussian-like distributions (Weigend, Rumelhart, & Humberman, 1991).
- approximate smoother (Moody & Rognvaldsson, 1997): it is shown (Chen, 2001a) that approximate smoother is an approximated form of the generalized spatial regularizer.

Table 1: Regularizers and weight priors.

regularizer	**PDF prior**	**comment**
constant	uniform distribution	uniform prior
w^2	Gaussian distribution $\mathcal{N}(0,1)$	weight decay
$\frac{w^2}{1+w^2}$	uniform + Gaussian	weight elimination
$\ln(1+w^2)$	Cauchy distribution	Cauchy prior
$\lvert w\rvert$	Laplace distribution	Laplace prior
$\ln(\cosh(w))$	supergaussian distribution	supergaussian prior

As summarized in Table 1, most regularizers correspond to the weight priors with different *a priori* probability density functions (PDF)[7]. It should be pointed out that (i) weight decay is a special case of weight elimination; (ii) the role of weight elimination is similar to Cauchy prior; (iii) under the approximation $p(\mathbf{w}) \propto \cosh^{-1/\beta}(\beta\mathbf{w})$, $p(\mathbf{w})$ approximates Laplace prior and we have $\frac{\partial}{\partial \mathbf{w}_j}\sum_j |\mathbf{w}_j| \approx \tanh(\beta\mathbf{w}_j)$. In above sense, the regularizer is sometimes written as $\|\mathbf{D}\mathbf{w}\|^2$ in place of $\|\mathbf{D}f\|^2$.

6 Shannon's Information Theory

According to information theory (Shannon, 1948), the efficiency of coding is measured by its entropy, the less entropy, the more efficient the encoder (Cover & Thomas, 1991). Suppose the learning machine (neural network) as an encoder, the information flow is transmitted through the channel, in which the noisy information is nonlinearly filtered, a counterpart illustration is shown in Figure 1. It is expected the information be encoded more efficiently, namely, via using less look-up tables (hidden basis functions) or less codes (connection weights). The entropy reduction in coding the information has been supported by neurobiological observation (Chen, 2001a).

MinEnt regularization may find reasonable theoretical support and convincing interpretation in the regression and classification problems as well. In functional approximation, the information in the data are expected to concentrate on the as few hidden units as possible, that is well-known sparse representation (it is well studied and observed the sensory information in the visual cortex are sparsely coded (Daugman, 1989; Atick, 1992; Olshausen & Field, 1996)). In some sense, the *maximum energy concentration* corresponds to *minimum Shannon entropy* (Coifman & Wickerhauser, 1992). On the other hand, in pattern recognition, people expect one or few hidden units correspond to specific pattern (or specific pattern are coded by specific hidden units), in other words, the knowledge learned by the machines (neural networks) are not *uniformly* distributed all the hidden units and connections.

In some sense, the hidden layer in the neural network acts like a Fourier expander, with the hidden units as the main spectral components (a sketchy discussion is given in Appendix D; see Chen, 2001a; Menon, Mehrotra, Mohan, & Ranka, 1996 for more details). suppose $f(x) \leftrightarrow \mathcal{F}(s)$ is a Fourier pair, from

[7]The complexity of weights is evaluated by the negative logarithm of probability density function at weight values.

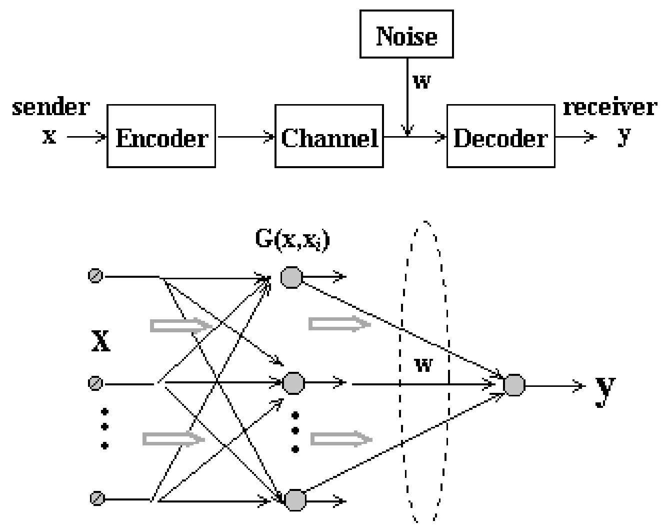

Figure 1: A schematic illustrations of neural network as an encoder-decoder.

Fourier transform's property, we have $\frac{\partial f(x)}{\partial x} \leftrightarrow s\mathcal{F}(s)$ [8]. According to 3.6 and 3.48, regularization constitutes to minimization of the regularizer:

$$\min \{\|\mathbf{D}f\|^2\} \leftrightarrow \min \{\|\mathbf{T}f\|^2\}.$$

Consider the first-order derivative of $\mathbf{D}f$, suppose $f = \sum_i w_i G(\mathbf{x}, \mathbf{x}_i)$ (for simplicity we only discuss the scalar case in frequency domain), then we have

$$\|\mathbf{T}f\|^2 = \|\sum_i w_i s_i \mathcal{G}(s_i)\|^2 \tag{6.1}$$

$$\leq \sum_i |w_i|^2 \sum_i |s_i|^2 \sum_i |\mathcal{G}(s_i)|^2 \leq \left(\frac{\sum_i |s_i|^2 + \sum_i |w_i|^2 + \sum_i |\mathcal{G}(s_i)|^2}{3} \right)^3 \tag{6.2}$$

The first inequality arises from Cauchy's inequality, the second one arises from the inequality of arithmetic and geometric means. Therefore, we further have an alternative form

$$\min \{\|\mathbf{T}f\|^2\} \leftrightarrow \min \left\{ \sum_i |s_i|^2 + |w_i|^2 + |\mathcal{G}(s_i)|^2 \right\} \tag{6.3}$$

where the first term is nonnegative, the second term corresponds to the weight decay. Hence, minimizing $\|\mathbf{T}f\|$ is functionally equivalent to minimizing the sum of the second and the third terms in 6.3. Now, observing the third term $|\mathcal{G}(s_i)|^2 = |G_i(x)|^2$ (by Parseval theorem), one may also have

$$\log |\mathcal{G}(s_i)|^2 = \log |G_i(x)|^2. \tag{6.4}$$

[8]Here we neglect the constant involved.

In some sense, the network output can be viewed as reconstructing the spectral components developed in the hidden units where the nonlinear activation functions (e.g. tanh or Gaussian) act as an integral operator (Appendix D, for more detailed treatment, see Chen, 2001b). Recalling $\mathbf{G}(\mathbf{x}_i, \mathbf{x}_j)$ is a $\ell \times \ell$ symmetric matrix, but it should be pointed out the symmetry is not necessary and our statement holds for any $\ell \times m$ ($\ell \neq m$) asymmetric matrix (which corresponds to generalized regularization networks). For the purpose of expression clarity, we denote $\mathbf{G}_i$ ($i = 1, \cdots, \ell$) be the ith vector of matrix $\mathbf{G}$, G_{ij} be jth ($j = 1, \cdots, m$) component of vector $\mathbf{G}_i$, normalizing every component G_{ij} by its row vector [9]

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

and we finally obtain a symmetric probability matrix $\mathbf{P}$ with $P_{ij} = P_{ji}$, which satisfy the relationship

$$\sum_{j=1}^{m} P_{ij} = 1. \tag{6.5}$$

And the information-theoretic entropy is defined by

$$H = \sum_{i=1}^{\ell} H_i = -\sum_{i=1}^{\ell} \sum_{j=1}^{m} P_{ij} \log P_{ij}. \tag{6.6}$$

Specifically, when $P_{i1} = \cdots = P_{im} = \frac{1}{m}$, H_i obtains the maximum value of $\log m$ by Lagrange multipliers, and $H_{max} = \ell \ln m$. Therefore, the *normalized entropy* is computed as

$$\widehat{H} = \frac{H}{\ell \log m}. \tag{6.7}$$

Hence the connection between 6.4 and 6.7 is bridged by

$$\min \left\{ \sum |G_{\mathbf{x}}|^2 \right\} \leftrightarrow \min \left\{ \sum \log |G_{\mathbf{x}}|^2 \right\} \leftrightarrow \min \{\widehat{H}\}.$$

This is an information-theoretic regularizer which we call MinEnt regularization (Chen, 2001a), which is also closely related to maximization of collective information principle proposed earlier (Kamimura, 1997), and decorrelation of the information by the nonlinear filter of hidden layer (Deco, Finnoff, & Zimmermann, 1995), though the starting points of same conclusion are different. The spectral regularization and MinEnt regularization implementation in the case of multilayer perceptrons with hyperbolic tangent (tanh) activation function are analyzed in detail in (Chen, 2001a).

MinEnt regularization principle is also connected to the well-studied the Infomax principle in machine learning, in supervised learning (Kamimura, 1997) as well as in unsupervised learning (e.g. see Becker, 1996). Supposing the input units be represented by A, hidden units be represented by B, then the mutual

[9] If kernel function is normalized RBF, this step is not necessary.

information $I(A, B)$ between A and B can be represented by their conditional entropy (Cover & Thomas, 1991; Haykin, 1999):

$$I(A, B) = H(B) - H(B|A). \tag{6.8}$$

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

On the other hand, it is necessary to clarify some confusion between the MinEnt regularization and MaxEnt principle in statistic physics, which will be discussed below. According to thermodynamics, the *close* system is always increasing the entropy. But the system (learning machine) discussed here is an *open* system, it absorbs some information (negative entropy) from outer environment thus minimizes the entropy. MaxEnt principle can be used in the blind separation and also be used for regularization, which is called MaxEnt regularization (e.g. Hansen, 1998), but these subjects are beyond the reach of current paper.

7 Statistical Physics

Considering a physical system at the T temperature environment, according to the theory of thermodynamics and statistical physics (Haken, 2000), the *Helmholtz free energy* $\mathcal{E}$ is defined as the difference between expected energy and entropy:

$$\begin{aligned} \mathcal{E} &= \sum_i p_i E_i - (-T \sum_i p_i \log p_i) \\ &= \sum_i p_i E_i + T \sum_i p_i \log p_i \end{aligned} \tag{7.1}$$

where E_i is the energy of a state i, p_i is the probability of the state i. Minimizing $\mathcal{E}$ we obtain the Boltzmann distribution whose probability is exponentially related to E_i (Haken, 2000; Hinton & van Camp, 1993)

$$p_i = \frac{\exp(-\beta E_i)}{\sum_i \exp(-\beta E_i)} = \exp(-\lambda - \beta E_i) \tag{7.2}$$

where $1/\beta$ has a physical interpretation of the multiplication product of absolute temperature T and the Boltzmann's constant. In regularization language, p_i measures the prior distribution, Lagrange multiplier λ plays the role of regularization parameter (with physical meaning of forces).

Physically speaking, a *closed* system with minimum free energy is meant that the system reaches an thermal equilibrium state, which is governed by the Gibbs distribution. Intuitively, free energy $\mathcal{E}$ can be naturally viewed as the empirical cost functional [10], which is anticipated to minimize the misfit part of data (empirical risk) and maximize the thermodynamic entropy, the latter of which corresponds to the principle of maximum entropy (MaxEnt) in the nature

[10] Not surprisingly, free energy minimization principle can be applied in machine learning, like Boltzmann machine and Helmholtz machine.

(Second Law of Thermodynamic states that the open systems always increase entropy). In the minimum point, the free energy is with the form of function

$$\mathcal{E} = -\log \sum_i \exp(-\beta E_i), \tag{7.3}$$

where $\sum_i \exp(-\beta E_i)$ corresponds to the *partition function* in statistical physics (Haken, 2000; Hinton & van Camp, 1993).

However, the neural network is somehow an *open* system in the sense that it exchanges the energy and information with the outer environment (human being). In this case, the entropy of the open system should be described by Prigogine (Prigogine, 1980; Stubbs, 1991):

$$H = H_{imported} + H_{generated}, \tag{7.4}$$

where $H_{imported}$ and $H_{generated}$ denote the imported and generated entropy of the open system respectively. Although $H_{generated}$ always increases, the decrease speed of $H_{imported}$ is always faster than the increase speed of $H_{generated}$, then H always decreases. Equation 7.4 states that (Stubbs, 1991), by import of energy or low entropy (information) the open systems can obtain lower entropy and greater complexity or order, which happens in the living mechanisms and human brain. We argue that $H_{imported}$ is information-theoretic entropy which we denote as H_{info} and $H_{generated}$ is physical-thermodynamic entropy, the former deals with signal whereas the latter deals with substance (see Stubbs, 1991 for detailed discussions), the decrease of $H_{imported}$ corresponds to the regularized functional, viz. MinEnt regularization discussed in section 6. Hence the regularized cost functional can be written as

$$\mathcal{R} = \mathcal{E} + H_{info}. \tag{7.5}$$

An insightful discussion on neural nets, information entropy and thermodynamic entropy was found in (Stubbs, 1991) and some references therein. It should be additionally noted that we have drawn the same conclusion about entropy reduction in neural network learning as Stubbs (1991) found earlier although our discussions are rooted from different perspectives.

8 Statistical Learning Theory

Rewriting the expected error $\mathcal{R}$ in an explicit form, we can decompose it into two parts (Geman, Bienenstock, & Doursat, 1992; Wolpert, 1997; Breiman, 1998)

$$\begin{aligned}
\mathcal{R} &= E[(y - f(\mathbf{x}))^2|\mathbf{x}] \\
&= E\Big[(y - E[y|\mathbf{x}] + E[y|\mathbf{x}] - f(\mathbf{x}))^2|\mathbf{x}\Big] \\
&= E\Big[(y - E[y|\mathbf{x}])^2|\mathbf{x}\Big] + \big(E[y|\mathbf{x}] - f(\mathbf{x})\big)^2
\end{aligned} \tag{8.1}$$

where $E[\cdot]$ denotes expectation operator. Equation 8.1 is actually the well-known bias-variance dilemma in statistics (Geman, Bienenstock, & Doursat, 1992), the first term is the bias of approximation and second term measures the variance of the solution, the regularization coefficient λ in 3.5 or 3.47 controls the trade-off of two terms in the expectation.

In recent years, a new statistical learning framework are formalized in the *structural risk minimization* (SRM), based on which support vector machines (SVMs) are built for a general learning problem (pattern recognition, functional approximation, and density estimation) (Vapnik, 1998a, 1998b). SVMs can be also regarded as some type of regularization network with exactly the same solution f in 3.36 but trained in a different way and therefore with different values of weights w_i (Girosi, 1998; Evgeniou, Pontil, & Poggio, 2000). In SVM, some of weights w_i are zero and the $\mathbf{x}_i$ corresponding to nonzero w_i are called support vectors, thus the solution found by SVMs is usually a sparse representation (Poggio & Girosi, 1998). Choosing specific kernel functions, the mapping from original data space to feature space corresponds to the regularization operators, that is why SVMs always exhibit good generalization capability in practice. Some insightful discussions on the links between SVM and regularization networks is found in (Girosi, 1998; Smola, Scholkopf, & Muller, 1998; Evgeniou, Pontil, & Poggio, 2000).

Writing $f(\mathbf{x})$ in terms of some type of semidefinite positive kernel function (not necessarily satisfying Mercel condition)

$$f(\mathbf{x}) = \sum_i \alpha_i K(\mathbf{x}_i, \mathbf{x}) + b \tag{8.2}$$

when $K(\mathbf{x}_i, \mathbf{x}_j) = < \mathbf{D}K(\mathbf{x}_i, \cdot), \mathbf{D}K(\mathbf{x}_j, \cdot) >$, the regularization network is particularly equivalent to SVM. Furthermore, as we know from the Green's function

$$\tilde{\mathbf{D}}\mathbf{D}G(\mathbf{x}_i, \mathbf{x}) = \delta_{\mathbf{x}_i}(\mathbf{x}) = \delta(\mathbf{x} - \mathbf{x}_i) \tag{8.3}$$

using G to minimize the risk functional of 3.5, we have

$$\begin{aligned} G(\mathbf{x}_i, \mathbf{x}_j) &= < \mathbf{D}G(\mathbf{x}_i, \cdot), \mathbf{D}G(\mathbf{x}_j, \cdot) > \\ &= < \Phi(\mathbf{x}_i), \Phi(\mathbf{x}_j) > \end{aligned} \tag{8.4}$$

with $\Phi : \mathbf{x}_i \to \mathbf{D}G(\mathbf{x}_i, \cdot)$.

In Fourier (transformation) domain, the regularization operator may be written by (Girosi, Jones, & Poggio, 1995; Smola, Scholkopf, & Muller, 1998):

$$< \mathbf{D}f, \mathbf{D}h > = \frac{1}{(2\pi)^{N/2}} \int_\Omega \frac{\mathcal{F}(\mathbf{s})\mathcal{H}^*(\mathbf{s})}{\mathcal{G}(\mathbf{s})} d\mathbf{s} \tag{8.5}$$

where $*$ denotes complex conjugate in complex function case, or equivalently we may write in the form of Hilbert norm

$$\|\mathbf{D}f\|^2 = \frac{1}{(2\pi)^{N/2}} \int_\Omega \frac{\|\mathcal{F}(\mathbf{s})\|^2}{\mathcal{G}(\mathbf{s})} d\mathbf{s} = \|f\|_{\mathbb{H}^m} \tag{8.6}$$

with $\mathcal{G}(\mathbf{s}) = \mathcal{G}(-\mathbf{s})$, $\mathcal{G}(\mathbf{s})|_{|\mathbf{s}| \to \infty} = 0$, and $\Omega := \text{supp}[\mathcal{G}(\mathbf{s})]$. $\mathcal{G}(\mathbf{s})$ describes the filter property of $\tilde{\mathbf{D}}\mathbf{D}$. In the case of spectral regularization operators 3.47, it can be shown that the equation

$$\begin{aligned} G(\mathbf{x}, \mathbf{x}_i) &= \frac{1}{(2\pi)^{N/2}} \int_{\mathbb{R}^N} \exp(j\mathbf{s}(\mathbf{x} - \mathbf{x}_i))\mathcal{G}(\mathbf{s}) d\mathbf{s} \\ &= \frac{1}{(2\pi)^{N/2}} \int_{\mathbb{R}^N} \exp(j\mathbf{s}\mathbf{x})\mathcal{G}(\mathbf{s}) \exp(-j\mathbf{s}\mathbf{x}_i) d\mathbf{s} \\ &= G(\mathbf{x} - \mathbf{x}_i) \end{aligned} \tag{8.7}$$

is a translationally invariant Green's function, and it is a special case of *Bochner's Theorem*, which states that the Fourier transform of a positive measure constitutes a positive *Hilbert-Schmidt* kernel (Smola, Scholkopf, & Muller, 1998).

9 Pruning algorithms

Pruning algorithms in connectionist community are an efficient way to enhance the generalization (Reed, 1993; Haykin, 1999; Cherkassky & Mulier, 1998), it can be viewed as some sort of regularization method where the complexity terms are measured by some regularizer. Pruning can be connection pruning or node pruning. Basically, there are four kinds of pruning algorithms developed from different perspectives, the first two kinds are mainly concerned about connection pruning and the last two are concerned about node pruning:

- Penalty function (Setiono, 1997) or regularizer based pruning approaches, e.g. weight decay (Hinton, 1989), weight elimination (Weigend, Rumelhart, & Humberman, 1991), Laplace prior (Williams, 1994), or MinEnt regularization (Chen, 2001a). Soft weight-sharing (Nowlan & Hinton, 1992) is another kind of pruning algorithm which is supposed the weights are represented by a mixture of Gaussian and the weights are expected to share the same value.
- Second-order (Hessian) information based pruning approaches, e.g. optimal brain damage (OBD) (LeCun, Denker, & Solla, 1990) and optimal brain surgeon (OBS) (Hassibi, Stock, & Wolff, 1992).
- Information-theoretic criteria based pruning scheme (Deco, Finnoff, & Zimmermann, 1995; Kamimura, 1997).
- Matrix decomposition 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), contribution analysis (Sanger, 1989).

The matrix decomposition-based pruning schemes are based on the observation of ill-posedness of the matrix $\mathbf{G}$. For instance, taking QR decomposition $\mathbf{Gw} = QR$, we may obtain the 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 $\widehat{\mathbf{G}} = \mathbf{G}L_1R_1^{-1}$ ($\mathbf{w} = [L_1 \ \ L_2]$), and the new weight vector $\hat{\mathbf{w}}$ and $\hat{\mathbf{G}}$ are calculated by

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

10 Equivalent Regularization

In machine learning community, there are various approaches dealing with equivalent regularization. For instance, early stopping (Bishop, 1995a; Haykin, 1999;

Cherkassky & Mulier, 1998), *Tangent distance* and *Tangent Prop* (Simard, LeCun, Denker & Victorri, 1998), 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 kernel regression estimator as target, choosing the variance of noise is equivalent to choosing the bandwidth of kernel of regression estimator. For a detailed discussion on training with noise, see (An, 1996; Bishop, 1995a, 1995b; Reed, Marks, & Oh, 1995). An insightful equivalence discussion between gain scaling, learning rate and scaling weight magnitude is found in (Thimm, Moerland, & Fiesler, 1996).

Many regularization techniques correspond to the structural learning principle. The *structural learning* here is meant the learning process and its determinants (e.g. hypothesis, environment, data, parameters and implementation) are 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 the terms of implementation of learning.

Early stopping is referred to the training network is expected to stop before going to minimum while observing the generalization error of independent validation set begin to increase. In the case of quadratic cost function of $\mathcal{R}_{emp}$, early stopping is similar to weight-decay regularization, the product of iteration index t and the learning rate η plays the role of regularization parameter λ, in the sense that the components of weight vector parallel to the eigenvectors of the Hessian satisfy (Bishop, 1995a)

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

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

where $\mathbf{w}^*$ denotes the desired minimum point in weight space, σ_i is 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 starting points (Cherkasskay & Mulier, 1998).

11 Kolmogorov Complexity: A Universal Principle for Regularization?

As discussed above, regularization theory can be built from many principles to measure the model complexity and control the generalization ability. However, it should be noted that neither of the principles (e.g. MDL, Bayes, entropy) has the universality, in the sense that those principles cannot be applied to any arbitrary areas.

Can we find a universal principle for regularization and machine learning? That questions naturally led the authors to think about an old but recently resurgent and popular theory in computational learning theory (partly due to recent outstanding work by mathematician George Chaitin), Kolmogorov complexity (or algorithmic complexity or Kolmogorov-Chaitin complexity).

Kolmogorov complexity theory, motivated by the Turing machine proposed by Computer scientist Alan Turing, was firstly studied by Solomonoff and Kolmogorov. The main thrust of Kolmogorov complexity lies in its *universality*, it is dedicated to construct universal learning methods based on universal coding methods (Schmidhuber, 1994). According to Kolmogorov complexity theory, any complexity can be measured by the length of the shortest program for a universal Turing machine that correctly reproduces the observation data like a look-up table. The Kolmogorov complexity theory mainly contain three parts: *complexity, randomness* and *information*. Mirroring the Kolmogorov complexity theory to machine learning theory, *complexity* closely connects to MDL or minimum message length (MML) principle in regularization theory; the *randomness* may also find the counterpart in learning theory, the well-known *no-free-lunch* theorems (e.g. for cross-validation, noise prediction, optimization, early stopping, bootstrapping) (see e.g. Wolpert & Macready, 1997; Goutte, 1997), all of which basically state that no learning algorithms can be universally good, some algorithms that perform exceptionally will comparably poorly in other situations, that reflects the key point of randomness; and the *information* is more connected to the entropy theory in machine learning.

In the Bayesian viewpoint, Kolmogorov complexity is dealing with the *universal prior* (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 universal prior are incomputable, some generalized complexity concepts for the purpose of computability was developed (e.g. Levin complexity). Due to the constraints of space, the extended discussions are beyond this paper. For some descriptive and detailed treatment on Kolmogrov complexity, the interested readers are encouraged to refer to (Cover & Thomas, 1991) and the special issue of Kolmogrov complexity at *Computer* journal (vol. 42, no. 4, 1999).

Can the Kolmogorov complexity be a universal principle for regularization theory? Some seminal works have been reported (Pearlmutter & Rosenfeld, 1991; Schmidhuber, 1994), but the studies were still limited to some toy problems, hence the question remains unanswered which needs to be further studied.

12 Summary and Beyond

This paper presents a unifying viewpoint on regularization theoretic framework from spatial and spectral (transformation) perspectives. The same results are obtained by using the Fourier (integral) operator instead of differential operator in the regularizer term. The state-of-the-art researches on regularization theories and techniques are thoroughly reviewed, many interested issues in machine learning community are addressed, The connections between the regularization theory and MDL principle, Bayesian theory, statistical learning theory, pruning algorithms are also discussed. Generalized regularization networks and equivalent regularization are examined and practical issues are further explored. Finally, the likelihood of a universal principle of Kolmogorov complexity for regularization is tentatively explored.

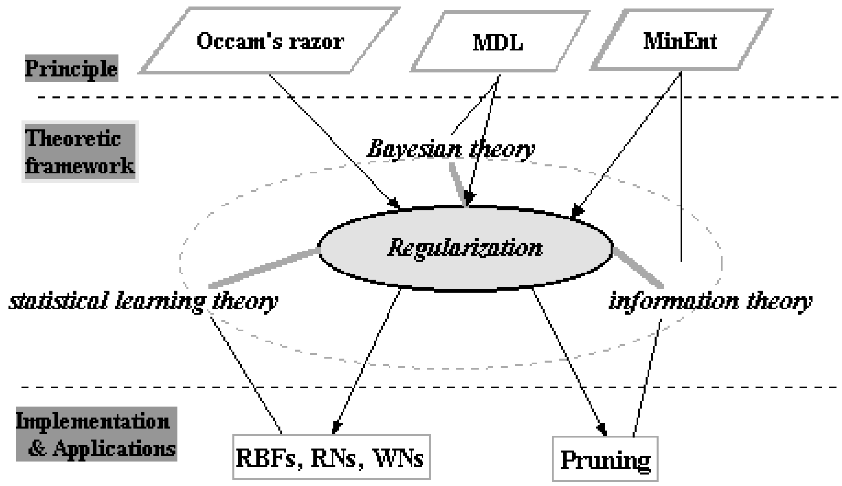

Figure 2: A schematic illustrations of unified regularization theory. The dashed line represents the classification of three levels: principle, tool and application levels; The undirectional solid line represents some conceptual link; and the directional arrow represents the route from principle to implementation, from theory to applications.

Although the contents of this paper are varied, they are closely related to the core of regularization and are sequentially discussed according to their relationships to regularization theory. Roughly speaking, Occam's razor, MDL, MinEnt are the principles of implementing regularization, all of which belong to the principle level; Bayesian theory, information theory, and statistical learning theory are related to regularization, but they belong to the theoretic framework level in the sense that they can be used as tools for dealing with principles to regularization (e.g. Bayesian for MDL, statistical physics for information entropy); whereas the pruning algorithms, equivalent regularization (early stopping), RBFs, RNs and GRNs belong to application level, in the sense that they are the direct results derived from regularization theory. A schematic relationship of the topics of this paper is illustrated in Figure 2.

To this end, the author would like to provide some personal comments beyond the topics in the current paper, hopefully the following prospectives may be some motivating points to study the regularization theory in machine learning:

- It has been interestingly shown that there are close relationship between regularization theory and sparse representation (Poggio & Girosi, 1998), SVMs (Girosi, 1998; Smola, Scholkopf, & Muller, 1998; Evgeniou, Pontil, & Poggio, 2000), independent component analysis, blind separation (Hochreiter & Schmidhuber, 1998), wavelet approximation, matching pursuit (Mallat & Zhang, 1993; Bernard, 1999), further efforts will be putting the machine learning problem to a more general framework and discussing their properties, which is still under investigation.

- The prospective studies of generalized regularized networks are devoted to build the approximation framework in the hybrid functional space, many encouraging results have been attained in reproducing kernel Hilbert space (RKHS), generalized Fock space (van Wyk & Durrani, 2000), Sobolev space and Besov space (a sketchy introduction on these functional spaces is given in Chen & Haykin, 2001b). The idea behind hybrid approximation is to find an overcomplete representation of an unknown function by means of direct sum of possibly overlapping function spaces, which results in a very sparse representation of the function of interest, SRM-SVM framework (Vapnik, 1998a) seems to be a feasible framework and mathematical tool for this goal. In addition, the algorithm implementation of regularization remains an important issue (that naturally connects to Kolmogorov complexity). The fast algorithms beyond the quadratic programming in SVM theory are favorably expected.

- A big class of regularization networks (RNs) and generalized regularization networks (GRNs) are formalized mathematically based on regularization theory (Girosi, Jones, & Poggio, 1995). Since the connection between RNs and SVMs was found (Girosi, 1998; Smola, Scholkopf, & Muller, 1998; Evgeniou, Pontil, & Poggio, 2000), we have much freedom to choose the basis (kernel) functions for GRNs or SVMs, and it can be extended to the wavelet networks (WNs) (Bernard, 1999)[11]. In addition, the theoretic study of the generalized regularized networks is also an important issue (Corradi & White, 1995; Niyogi & Girosi, 1996, 1999).

- It is reported (Canu & Elisseeff, 1999) 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 the sigmoid-shape network, which partly answers the unanswered question posed in (Girosi, Jones, & Poggio, 1995): *Can the sigmoid-like neural network be derived from the regularization theory?*

Appendix

A Proof of Dirichlet kernel

For the purpose of self-containing of this paper, the proof of Dirichlet kernel given in (Lanczos, 1961) is rewritten as follows. Observing Dirac-delta function $\delta(s,x)$ satisfies the following conditions

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

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

for a symmetric, translationally invariant function $G(s,x) = G(x,s) = G(s-x) = G(\theta) = G(-\theta)$ is zero everywhere except in the interval $|\theta| \leq \epsilon$, where ϵ is

[11] An discussion of connection of regularization theory to wavelet approximation and WNs is given in (Chen & Haykin, 2001b).

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 $\int_{-\epsilon}^{\epsilon}G(\theta)d\theta = 1$, and $\epsilon \to 0$, the point $\xi \to s$ and one may obtain the desired expansion coefficients (A.1)(A.2). Comparing 3.18, the Dirichlet kernel $K_n(s,x)$ acts like a Dirac-delta function

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

if one replaces $K(s-x)$ by the Fourier expansion coefficients of Dirac function, the proof is completed.

□

B The proof of regularization solution

The proof of regularization solution was partly given in (Poggio & Girosi, 1990a; Haykin, 1999) and is rewritten here for completeness. In virtue of 3.33, applying $\mathbf{L}$ to function $f(\mathbf{x})$, we have

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

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

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

The solution of regularization problem is further derived by setting

$$\begin{aligned}\varphi(\boldsymbol{\xi}) &= \frac{1}{\lambda}\sum_{i=1}^{\ell}[y_i - f(\mathbf{x}_i)]\delta(\boldsymbol{\xi} - \mathbf{x}_i), && \text{(B.3)}\\ \Phi(\boldsymbol{\omega}) &= \mathcal{F}\{\varphi(\boldsymbol{\xi})\} = \mathcal{F}\bigg\{\frac{1}{\lambda}\sum_{i=1}^{\ell}[y_i - f(\mathbf{x}_i)]\delta(\boldsymbol{\xi} - \mathbf{x}_i)\bigg\}\\ &= \frac{1}{\lambda}\sum_{i=1}^{\ell}[y_i - f(\mathbf{x}_i)]\exp(-j\mathbf{x}_i\boldsymbol{\omega}), && \text{(B.4)}\end{aligned}$$

then in spatial domain, we have

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

and equivalently in frequency domain

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

where $w_i = [y_i - f(\mathbf{x}_i)]/\lambda$. The first equality in previous equation follows from integral theorem in Fourier transform: given two real-valued function $f(t)$ and $g(t)$, $\mathcal{F}(\omega)$ and $\mathcal{G}(\omega)$ are their Fourier transform, $\int f(t)g(t)dt = \int \mathcal{F}(\omega)\mathcal{G}^*(\omega)d\omega$ where $*$ is complex conjugate. So far the proof is completed.

$\square$

C 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 $\mathbf{A}, \mathbf{B}$ is $m \times p$ and $n \times p$ matrix 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 (not necessarily) square, $\mathbf{\Sigma}$ and $\mathbf{S}$ are diagonal matrices, all of which satisfy

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

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

D Spectral reconstruction

Observing 3.37, one may replace $\sum$ by $\int$, $G(\mathbf{x}, \mathbf{x}_i)$ by $K(\mathbf{x}, \mathbf{x}_i)$, w_i by $[y_i - f(\mathbf{x}_i)]/\lambda$, suppose the kernel function is a translationally invariant reproducing kernel in RKHS (see e.g. Girosi, 1998), i.e. it is positive definite and satisfies $K(\mathbf{x}, \mathbf{x}_i) = K(\mathbf{x} - \mathbf{x}_i)$, thus the approximated function can be expressed by a convolution of observation and a moving kernel

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

From RKHS theory, (D.1) can be written as

$$f(\mathbf{x}) = \sum_{i=1}^{\ell} c_i \Phi(\mathbf{x})\Phi(\mathbf{x}_i). \tag{D.2}$$

In above sense, one can imagine the functional $f(\mathbf{x})$ is doing Fourier (frequency) analysis (or time-frequency analysis, depending on K), where the kernel function $K(\mathbf{x}, \mathbf{x}_i)$ accounts for a convolutional window, the metric function μ measures the distance between the synthesis function $f(\mathbf{x})$ and the desired value $y(\mathbf{x})$. Moving the window along the temporal domain, we may expect the scaled shifted window to fit the signal of interest (Chen, 2001a, 2001b).

E Proof of QR decomposition

Apply QR decomposition to matrix $\mathbf{G}$,

$$\mathbf{G}L = QR \tag{E.1}$$

where $\mathbf{G}$ is $\ell \times m$ input matrix, L is $m \times m$ transposition matrix, Q is $\ell \times \ell$ full-rank matrix, R is $\ell \times m$ upper triangle matrix, both of which are expressed by

$$R = \begin{bmatrix} R_1 & R_2 \\ 0 & 0 \end{bmatrix}, \ Q = \begin{bmatrix} Q_1 & Q_2 \end{bmatrix} \tag{E.2}$$

henceforth,

$$\mathbf{G}L = \begin{bmatrix} \mathbf{G}L_1 & \mathbf{G}L_2 \end{bmatrix} \tag{E.3}$$

and it may further follow 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 $m \times r$ matrix, L_2 is $m \times (m-r)$ matrix, Q_1 is $\ell \times r$ matrix, Q_2 is $\ell \times (m-r)$ matrix, R_1 is $r \times r$ matrix, R_2 is $r \times (m-r)$ matrix.

Suppose the hidden nodes are pruned from m to r $(m < r)$: $\mathbf{G}_{\ell\times m} \to \hat{\mathbf{G}}_{\ell\times r}$, i.e. there are $(m-r)$ redundant hidden nodes. Denoting the new weight matrix be $\widehat{\mathbf{w}}$, thus the new expression for the network is rewritten as

$$\mathbf{y} = \mathbf{G}\mathbf{w} \to \hat{\mathbf{y}} = \hat{\mathbf{G}}\hat{\mathbf{w}} \tag{E.4}$$

where

$$\widehat{\mathbf{G}} = Q_1 = \mathbf{G}L_1R_1^{-1}. \tag{E.5}$$

And the new weight vector $\hat{\mathbf{w}}_{r\times 1}$ can be calculated as

$$\widehat{\mathbf{w}}^T = \mathbf{w}^T L_1 + \mathbf{w}L_2(R_1^{-1}R_2)^T. \tag{E.6}$$

□

Acknowledgements

The work is supported by NSERC of Canada. The first author is partly supported by Clifton W. Sherman Scholarship in McMaster University.

References

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

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

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

[4] Atick, J.J. (1992). Could information theory provide an ecological theory of sensory processing? *International Journal of Neural Systems, 3*, 23-251.

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

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

[7] Bernard, C. (1999). Wavelets and ill-posed problems: optical flow and data interpolation. Ph.D. thesis, Ecole Polytechnique, France.

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

[9] Bishop, C. (1993). Curvature-driven smoothing: a learning algorithm for feedforward networks. *IEEE Transactions on Neural Networks, 4*, 882-884.

[10] Bishop, C. (1995a). *Neural networks for pattern recognition.* Oxford University Press.

[11] Bishop, C. (1995b). Training with noise is equivalent to Tikohonv regularization. *Neural Computation, 7*, 108-116.

[12] Breiman, L. (1998). Bias-variance, regularization, instability and stablization. In: Bishop, C. (ed.) *Neural networks and machine learning*, NATO ISI series F: Computer and Systems Sciences, vol. 168, 27-56.

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

[14] Canu, S. & Elisseeff, A. (1999). Regularization, kernels and sigmoid net. unpublished manuscript. Avaiable on line `http://psichaud.insa-rouen.fr/~scanu/`

[15] Chen, Z. & Haykin, S. (2001a). A new view on regularization theory. *Proc. IEEE Int. Conf. System, Man and Cybernetics*, 1642-1647 Tucson, Arizona, USA.

[16] Chen, Z. & Haykin, S. (2001b). A unifying view on regularization theory. Technical Report. Communications Research Laboratory, McMaster University. Available on line `http://soma.crl.mcmaster.ca/~zhechen/tr-regu.ps`.

[17] Chen, Z. (2001a). Spectral regularization and MinEnt regularization. submitted to *ICASSP2002*, available on line `http://soma.crl.mcmaster.ca/~zhechen/regu.ps`.

[18] Chen, Z. (2001b). Supervised learning as a spectral reconstruction problem. to be submitted.

[19] Cherkassky, V. & Mulier, F. (1998). *Learning from data: concepts, theory and methods*, Wiley.

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

[21] Corradi, V. & White, H. (1995). Regularized neural networks: some convergence rate results. *Neural Computation, 7*, 1225-1244.

[22] Courant, R. & Hilbert, D. (1970). *Methods of mathematical physics*, vol. 1 & 2, Wiley.

[23] Cover, T. & Thomas, J.A. (1991). *The Elements of information theory*, Wiley.

[24] 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.

[25] 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.

[26] Dontchev, A.L. & Zolezzi, T. (1993). *Well-posed optimization problems*, Lecture Notes in Mathematics, 1543, Berlin: Springer-Verlag.

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

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

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

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

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

[32] Haken, H. (2000). *Information and self-organization.* (2nd enlarged edition), Springer.

[33] Haykin, S. (1999). *Neural networks: a comprehensive foundation* (2nd edition), Prentice-Hall.

[34] Haykin, S.& Principe, J. (1998). Making sense of a complex world. *IEEE Signal Processing Magazine, 15*(3), May, 66-81.

[35] Hansen, P.C. (1998). *Rank deficient and discrete ill-posed problems: numerical aspects of linear inversion.* SIAM, Philadephia, PA.

[36] Hassibi, B., Stock, D.G., & Wolff, G.J. (1992). Optimal brain surgeon and general network pruning. *Proc. Int. Conf. Neural Networks, ICNN'92*, 293-299, San Francisco, CA.

[37] 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*, San Cruz.

[38] Hinton, G.E. (1989). Connectionist learning procedure. *Artificial Intelligence, 40*, 185-234.

[39] Hochreiter, S., & Schmidhuber, J. (1997). Flat minima. *Neural Computation, 9*, 1-42.

[40] Hochreiter, S., & Schmidhuber, J. (1998). Source separation as by-product of regularization. *Advances in Neural Information Processing Systems, NIPS 11*, 459, MIT Press.

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

[42] 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.

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

[44] 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.

[45] Koene, R. & Takane, Y. (1999). Discriminant component pruning: regularization and interpretation of multilayered backpropagation networks. *Neural Computation, 11*, 783-802.

[46] Lanczos, C. (1961). *Linear differential operator*. D. Van Nostrand Company Ltd.

[47] LeCun, Y., Denker, J.S., & Solla, S.A. (1990). Optimal brain damage. *Advances in Neural Information Processing Systems, NIPS 2*, 598-605.

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

[49] Levin, A., Leen, T., & Moody, J. (1994). Fast pruning using principal components. *Advances in Neural Information Processing Systems, NIPS 6*, 35-42.

[50] MacKay, D.J.C. (1992). Bayesian methods for adaptive models. Ph.D. thesis, Deptartment of Computation and Neural Systems, Caltech.

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

[52] Menon, A., Mehrotra, K., Mohan, C. K., & Ranka, S. (1996). Characterization of a class of sigmoid functions with applications to neural networks. *Neural Networks, 9*(5), 819-835.

[53] Moody, J. & Rognvaldsson, T. (1997). Smoothing regularizers for projective basis function networks. *Advances in Neural Information Processing Systems, NIPS 9*, 585-591.

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

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

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

[57] 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.

[58] Pearlmutter, B.A, & Rosenfeld, R. (1991). Chaitin-Kolmogorov Complexity and Generalization in Neural Networks. *Advances in Neural Information Processing Systems, NIPS 3*, 925-931.

[59] Poggio, T., Torre, V. & Koch, C. (1985). Computational vision and regularization theory. *Nature, 317*, September, 314-319.

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

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

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

[63] Prigogine, I. (1980). *From being to becoming*, W.H. Freeman.

[64] Reed, R. (1993). Pruning algorithms - a review. *IEEE Transactions on Neural Networks, 4*, 740-747.

[65] 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.

[66] Rissaen, J. (1978). Modeling by shortest data description. *Automatica, 14*, 465-471.

[67] Rohwer, R. & van der Rest, J.C. (1996). Minimum description length, regularization, and multimodal data. *Neural Computation, 5*, 595-609.

[68] Sanger, D. (1989). Contribution analysis: a technique for assigning responsibilities to hidden units in connectionist networks. *Connection Science, 1*, 115-138.

[69] Schmidhuber, J. (1994). Discovering problem solutions with low Kolmogorov complexity and high generalization capability. Technical report FKI-194-94, Technische Universitat Munchen, Germany. Avaiable at `http://papa.informatik.tu-muenchen.de/mitarbeiter/schmidhu.html`

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

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

[72] Simard, P., LeCun, Y., Denker, J. & Victorri, B. (1998). Transformation invariance in pattern recognition: tangent distance and tangent propagation. In Orr, G.B., & Muller, K-R. (eds.) *Neural networks: tricks of the trade*, Lecture Notes in Computer Science, vol. 1524, Springer.

[73] Smola, A., Scholkopf, B., & Muller, K.-R. (1998). The connection between regularization operators and support vector kernels. *Neural Networks, 11*, 637-649.

[74] Strichartz, R. (1994). *A Guide to distribution theory and Fourier transforms*, CRC Press.

[75] Stubbs, D. F., 1991. Entropy and neural nets. *Neurocomputer and attention II: connectionism and neurocomputer*, 685-694, Manchester University Press.

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

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

[78] 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.

[79] Vapnik, V. (1998a). *Statistical learning theory*, Wiley.

[80] Vapnik, V. (1998b). The support vector method of function estimation. In: Bishop, C. (ed.) *Neural networks and machine learning*, NATO ISI series F: Computer and Systems Sciences, vol. 168, 239-268.

[81] Velipasaoglu, E.O., Sun, H., & Zhang, F., et al. (2000). Spatial regularization of the electrocardiographic inverse problem and its application to endocardial mapping. *IEEE Transactions on Biomedical Engineering, 47*(3), 327-337.

[82] Weigend, A., Rumelhart, D., & Humberman, B.A. (1991). Generalization by weight-elimination applied to currency exchange rate prediction. *Proc. Int. Joint Conf. Neural Networks, IJCNN'91-Singapore*, 2374-2379.

[83] Wahba, G. (1990). *Spline models for observational data*. SIAM, Philadephia.

[84] Watanabe, K., Namatame, A., & Kashiwaghi, E. (1993). A mathematical foundation on Poggio's regularization theory. *Proc. Int. Conf. Neural Networks, ICNN'93*, 1717-1722.

[85] Williams, P.M. (1994). Bayesian regularization and pruning using a Laplace prior. *Neural Computation, 7*, 117-143.

[86] Wolpert, D.H. (1997). On bias plus variance. *Neural Computation, 9*, 1211-1243.

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

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

[89] Yee, P. (1998). Regularized radial basis function networks: theory and applications to probability estimation, classification, and time series prediction. Ph.D. thesis, Department of Electrical and Computer Engineering, McMaster University.

[90] Yee, P. & Haykin, S. (2001). *Regularized radial basis function networks: theory and applications*, Wiley.