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, & Kashiwagi, 1993). However, it is observed that an equivalent reg-
ularization 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 \( x \in X \), there exists an output vector \( y = F(x) \), where \( y \in Y \).

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

3. **Continuity.** The mapping is continuous, that is, for any \( \epsilon > 0 \) there exists \( d = d(\varepsilon) \) such that the condition \( d_X(x, z) < \delta \) implies that \( d_Y(F(x), F(z)) < \varepsilon \), 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

\[
\begin{align*}
    x_{t+1} &= F(x_t, w_t) \tag{2.2} \\
    y_t &= G(x_t) + m_t 	ag{2.3}
\end{align*}
\]

where \(F\) is nonlinear mapping function \(\mathbb{R}^N \rightarrow \mathbb{R}^N\), \(G\) is a scalar-valued function, and \(w_t\) and \(m_t\), represent the process noise and measurement noise contaminating the state variable \(x_t\) 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 (\(w_t\) 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 \(^\text{1}\), 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) $\{(x_i, y_i) \in \mathbb{R}^N \times \mathbb{R}\}_{i=1}^{\ell} \subset X \times Y$, 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 $f(x_i) = y_i$, where $y(x)$ 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 $w$. 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

$$
\mathcal{R} = \int_{X \times Y} L(x, y)dP(x, y) \\
= \int_{X \times Y} L(x, y)p(x, y)dxdy \tag{3.1}
$$

where $L(x, y)$ 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 $f(x)$ corrupted by additive white noise independent of $x$: $y = f(x) + \varepsilon$, where $\varepsilon$ is bounded and follows some unknown probability metric $\mu(\varepsilon)$. In that case, $p(x, y) = p(x)p(y|x)$ and $p(y|x)$ is represented by the metric function $\mu[y - f(x)]$. Hence, the expected cost functional with the $L_2$ norm is given by

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

In practice, the joint probability $p(x, y)$ is unknown, and an estimate of $\mathcal{R}$ based on

\(^{1}\) Another methodology to embed prior knowledge is the theory of hints, see (Abu-Mostafa, 1995) for more information.
on finite observations ($\ell$) is used instead, with an empirical risk functional

$$ R_{\text{emp}} = \sum_{i=1}^{\ell} [y_i - f(x_i)]^2, \quad (3.3) $$

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

$$ R \leq R_{\text{emp}} + \sqrt{\frac{d_{\text{VC}}(\log(2\ell \cdot d_{\text{VC}} + 1) - \log(\frac{\eta}{4})}{\ell}} \quad (3.4) $$

with the probability $1 - \eta$, where $d_{\text{VC}}$ 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) $R_{\text{emp}}$ and the regularizer risk $R_{\text{reg}}$:

$$ R[f] = R_{\text{emp}}[f] + \lambda R_{\text{reg}}[f] \quad (3.5) $$

$$ = \frac{1}{2} \sum_{i=1}^{\ell} [y_i - f(x_i)]^2 + \frac{1}{2} \lambda \|Df\|^2 \quad (3.6) $$

where $\|\cdot\|$ is the norm operator $^2$. $\lambda$ is the regularization parameter which controls the trade-off between the identity (goodness-of-fit) of data and the roughness of the solution. $D$ is a linear differential operator, which is defined as Fréchet differential of Tikhonov functional (Tikhonov & Arsenin, 1977; Poggio & Girosi, 1990a; Haykin, 1999). Geometrically, $D$ is interpreted as the local linear approximation of the curve in the space. The smoothness prior embedded in $D$ 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)

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

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

$$ dR_{\text{emp}}(f, h) = \frac{d}{d\beta} R_{\text{emp}}(f + \beta h) |_{\beta=0} $$

$$ = - \sum_{i=1}^{\ell} [y_i - f(x_i)] h(x_i) $$

$$ = - \left\langle h, \sum_{i=1}^{\ell} (y_i - f(x_i)) \delta(x - x_i) \right\rangle, \quad (3.8) $$.  

$^2$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 \(H\). Similarly, the Fréchet differential of the regularizing term \(R_{\text{reg}}\) is led to
\[
dR_{\text{reg}}(f, h) = \frac{d}{d\beta} R_{\text{reg}}(f + \beta h) \big|_{\beta=0} = \int D[f + \beta h] D h \, dx \big|_{\beta=0} = \int D[f] D h \, dx = \langle D h, D f \rangle. \tag{3.9}
\]
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 problems, 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 \(L\), the function \(G(x, \xi)\) is said to be Green’s function if it has the following properties (Courant & Hilbert, 1970; Haykin, 1999):

- For a fixed \(\xi\), \(G(x, \xi)\) is a function of \(x\) and satisfies the given boundary condition.
- Except at the point \(x = \xi\), the derivatives of \(G(x, \xi)\) with respect to \(x\) are all continuous, the number of derivatives is determined by the order of operator \(L\).
- With \(G(x, \xi)\), considered as a function of \(x\), it satisfies the partial differential equation
  \[
  LG(x, \xi) = \delta(x - \xi) \tag{3.10}
  \]
  where \(\delta\) 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 \(\varphi(x)\) a continuous function of \(x \in \mathbb{R}^N\), then the function
\[
F(x) = \int_{\mathbb{R}^N} G(x, \xi) \varphi(\xi) d\xi \tag{3.11}
\]
is the solution of the differential equation
\[
LF(x) = \varphi(x) \tag{3.12}
\]
where \(G(x, \xi)\) is the Green’s function for the linear differential operator \(L\), 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 $T$ in functional space, for all functionals $f \in H$, we always have $Tf \in H$. (It can be proven that Fourier operator $T$ is a linear integral operator.)

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

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

$$
Tf(s) = \int_{\mathbb{R}^N} f(x)K(s, x)dx
$$

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

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

then the corresponding transformation operator is

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

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

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

From 3.16, the truncated Fourier series are written by

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

where

$$
K_n(s, x) = \frac{1}{\pi} \sum_{k=0}^{n-\infty} (\cos ks \cos nx + \sin ks \sin nx)
= \frac{1}{\pi} \sum_{k=0}^{n} \cos k(s - x) = K_n(s - x)
$$

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

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

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

$$\langle Dh, Df \rangle = \langle T_Dh, T_Df \rangle$$  \hfill (3.21)

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

$$dR_{reg}(f, h) = \int DfDh \, dx = \langle Dh, Df \rangle = \int TfTf ds = \langle Th, Tf \rangle.$$  \hfill (3.22)

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

$$\int_{\mathbb{R}^N} u(x)Dv(x)dx = \int_{\mathbb{R}^N} v(x)\tilde{D}u(x)dx$$  \hfill (3.23)

and

$$\int_{\Omega} u(s)Tv(s)ds = \int_{\Omega} v(s)\tilde{T}u(s)ds$$  \hfill (3.24)

where $\Omega$ is the support in frequency domain. Equations 3.23 and 3.24 are called Green's identity (Lanczos, 1961; Poggio & Girosi, 1990a). Viewing $D$ and $T$ 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 $D(u, v) = (Du)v + (Dv)u$ and $T(u, v) = T(Tu, v) = T(u, Tv)$, where $D$ and $T$ are differential and integral operators, respectively.
Using Green’s identity, we further obtain the equivalent form of 3.22 from 3.23 by setting \( u(x) = Df(x) \) and \( Dv(x) = Dh(x) \)

\[
d_R \rho_{reg}(f, h) = \int h(x) \tilde{D}Df(x)dx = \left\langle h, \tilde{D}Df \right\rangle(x),
\]

and another form from 3.24 by setting \( u(s) = Tf(s) \) and \( Tv(s) = Th(s) \)

\[
d_R \rho_{reg}(f, h) = \int h(s) \tilde{T}Tf(s)ds = \left\langle h, \tilde{T}Tf \right\rangle(s).
\]

And the Fréchet differential

\[
d_R(f, h) = d_R \rho_{emp}(f, h) + \lambda d_R \rho_{reg}(f, h) = 0
\]
can be rewritten from 3.8 by virtue of 3.25 and 3.26 in the following forms

\[
d_R(f, h) = \left\langle h(x), \left[ \tilde{D}Df(x) - \frac{1}{\lambda} \sum_{i=1}^{\ell} (y_i - f) \delta(x - x_i) \right] \right\rangle,
\]

or

\[
d_R(f, h) = \left\langle h(s), \left[ \tilde{T}Tf(s) - \mathcal{F}\left\{ \frac{1}{\lambda} \sum_{i=1}^{\ell} (y_i - f) \delta(x - x_i) \right\} \right] \right\rangle,
\]

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

\[
\tilde{D}Df_\lambda(x) = \frac{1}{\lambda} \sum_{i=1}^{\ell} (y_i - f(x_i)) \delta(x - x_i),
\]

and

\[
\tilde{T}Tf_\lambda(s) = \mathcal{F}\left\{ \frac{1}{\lambda} \sum_{i=1}^{\ell} (y_i - f(x_i)) \delta(x - x_i) \right\}.
\]

Equations 3.29 is the Euler-Lagrange equation of Tikhonov functional \( R[f] \) and 3.30 is its Fourier counterpart. Denoting \( L = \tilde{D}D \) and \( K = \tilde{T}T \), \( G(x, \xi) \) 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)

\[
L G(x, \xi) = \delta(x - \xi)
\]

and its counterpart in frequency domain

\[
K G(s, \xi) = \exp(-j s \xi),
\]

similar to 3.11, it can be proven

\[
f(x) = \int_{\mathbb{R}^N} G(x, \xi) \varphi(\xi)d\xi
\]
is the solution of the following differential and integral equations

\[
\begin{align*}
L f(x) &= \varphi(x), \\ K f(s) &= \Phi(s),
\end{align*}
\]  

(3.34) (3.35)

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

\[
f_\lambda(x) = \sum_{i=1}^\ell w_i G(x, x_i)
\]  

(3.36)

where \( w_i = \left[ y_i - f(x_i) \right] / \lambda \), and \( G(x, x_i) \) 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)

\[
L = \sum_{n=0}^\infty \frac{(-1)^n}{n!2^n} \nabla^{2n}
\]  

(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

\[
K = \sum_{n=0}^\infty \frac{(-1)^{2n} s^{2n}}{n!2^n} = \exp\left(\frac{s^2}{2}\right),
\]  

(3.38)

and we further have

\[
\begin{align*}
L G(x) &= \delta(x), \\ K G(s) &= 1.
\end{align*}
\]  

(3.39) (3.40)

- In general, the solution to the regularization problem is

\[
f_\lambda(x) = \sum_{i=1}^{\ell} w_i G(x, x_i) + \beta(x)
\]  

(3.41)

where \( \beta(x) \) is a term that lies in the null space of regularized functional, which satisfies the orthogonal condition \( \sum_{i=1}^{\ell} w_i \beta(x_i) = 0 \), hence the functional space of solution \( f_\lambda \) 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 $Tf$ (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 state 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 $D$ or $T$ \(^5\). Note that some invariance properties are implicitly embedded in $T$ (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 \textit{translationally and rotationally invariant}, in other words, $G(x, x_j)$ is inherently a radial basis function (RBF) with the radially symmetric and shift invariant form (Poggio & Girosi, 1990a; Haykin, 1999)

\begin{equation}
G(x, x_j) = G(\|x - x_j\|),
\end{equation}

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

\begin{equation}
G(x, x_j) = G(|x - x_j|^T \Sigma^{-1} |x - x_j|),
\end{equation}

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

\begin{eqnarray}
G(s) &=& \exp(-\|s\|^2/2), \\
G(x) &=& \exp(-\|x\|^2/2),
\end{eqnarray}

where $G(x) \leftrightarrow G(s)$ 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 \textit{periodic translationally invariant} Gaussian kernel can be

\begin{equation}
G(x, x_j) = \exp(-j(x - x_j)) \exp(-\|x - x_j\|^2) \\
= \cos(x - x_j) \exp(-\|x - x_j\|^2) \\
- j \sin(x - x_j) \exp(-\|x - x_i\|^2)
\end{equation}

The previous equation reminds us $G(x, x_j)$ 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{equation}
R[f] = R_{\text{emp}}[f] + \lambda R_{\text{reg}}[F]
\end{equation}

\begin{equation}
= \frac{1}{2} \sum_{i=1}^{\ell} (y_i - f(x_i))^2 + \frac{1}{2} \lambda \|Tf\|^2
\end{equation}

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

\(^5\)Girosi, Jones and Poggio (1995, Appendix A) also provided a different proof of solution to regularization problem from the smoothness functional.
$G(s)$ is a low-pass filter with property $G(s)|_{s \to \infty} = 0$ (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

$$
\mathcal{R} = \|y - Gw\|^2 + \lambda \|Pw\|^2
$$

where $w = [w_1, \ldots, w_\ell]^T$, $G = G(x, x_i)$ is a radial basis matrix and $y = [y_1, \ldots, y_\ell]^T$. $P$ is a user-designed matrix for regularizer. Since $G$ is usually ill-conditioned, we always do some matrix decomposition for reducing redundancy. Taking the singular value decomposition (SVD) of $G$

$$
G = U \Sigma Z^T
$$

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

$$
(\Sigma^T \Sigma + \lambda I)\alpha = \Sigma^T d
$$

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

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

and

$$
c_i = \frac{\sigma_i^2}{\sigma_i^2 + \lambda s_i^2}, \quad \alpha_i = \frac{c_i d_i}{\sigma_i},
$$
respectively, where $s_i$ are defined as the diagonal elements of the diagonal matrix $S$ by SVD of $P = VSZ^T$ (viewing the operator as a matrix). In nonzero-order Tikhonov regularization case, provided $w = Z\alpha$ we have

$$\Sigma^T \Sigma + \lambda S^T S \alpha = \Sigma^T d.$$  \hspace{1cm} (3.54)

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

$$Z^T w = \Sigma^T_{\lambda} d = \Sigma_{\lambda} U^T y, \hspace{1cm} (3.55)$$

in which

$$\Sigma_{\lambda} = \begin{cases} \text{diag} \left\{ \frac{\sigma_i^2}{\gamma_i^2 + \lambda} \right\}, & P = I \\ \text{diag} \left\{ \frac{\gamma_i (\gamma_i^2 + 1)^{1/2}}{\gamma_i^2 + \lambda} \right\}, & P \neq I \end{cases} \hspace{1cm} (3.56)$$

where $\gamma_i = (\sigma_i^2 + s_i^2)^{1/2}$ are the generalized singular values of the $(\Sigma, S)$ (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 $\sigma_i$. 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 $G$ 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 $\lambda$ can be estimated using Bayesian method with second evidence framework (the first evidence framework is estimating the posterior probability of weights while supposing $\lambda$ is known, see section 5 for details). Given the data $D$ and model $M$, regularization parameter can be estimated by

$$p(\lambda|D, M) \propto p(D|\lambda, M) p(\lambda|M). \hspace{1cm} (3.57)$$

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 \( \lambda \) 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 \( \lambda \) should be

\[
\lambda \geq \frac{-\|\nabla R_{\text{emp}}(w)\|^2}{\langle \nabla R_{\text{emp}}(w), \nabla R_{\text{reg}}(w) \rangle},
\]

and

\[
\lambda \geq \frac{\langle \nabla R_{\text{emp}}(w), \nabla R_{\text{reg}}(w) \rangle}{-\|\nabla R_{\text{reg}}(w)\|^2},
\]

in order to guarantee the convergence of \( R_{\text{emp}}(w) \) and \( R_{\text{reg}}(w) \), 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 (Rissanen, 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 \( \mathcal{D} = \{x_i, y_i\} \) is viewed as being drawn independently from some unknown distribution, the learning problem is formulated as the dependency estimation of the \( y \) upon \( x \). Naturally, a metric measuring the complexity of the data length is given (Rissanen, 1978):

\[
R = L(\mathcal{D}|\mathcal{M}) + L(\mathcal{M})
\]

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 $\varepsilon$ is Gaussian distributed with quantization width $\mu$, one will have (Hinton & van Camp, 1993)

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

(4.2)

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

$$
R_{\text{emp}} = k\ell + \frac{\ell}{2} \log_2 \left( \frac{1}{\ell} \sum_i \varepsilon_i^2 \right)
$$

(4.3)

where $k$ is a constant depending on $\mu$. 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 $L(M)$ of weight decay can be also written by MDL principle

$$
R_{\text{reg}} = \frac{1}{2\sigma_w^2} \sum_i w_i^2
$$

(4.4)

in which the $\sigma_w$ controlling the noise level acts like the regularization parameter.

In the noisy weight case, MDL corresponds to introducing high variance in $L(D|M)$, i.e. the misfit of data is less reliable. More generally, the weights may be assumed to be a mixture density $p(w) = \sum_i \pi_i p(w_i)$, 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 $D$, what we care about is finding the most probable model for the observation data, following Bayes formula, the posterior probability of model $M$ is estimated by $p(M|D) = p(D|M)p(M)/p(D)$, where the denominator represents the evidence as a normalizing constant. Maximizing $p(M|D)$ is further equivalent to minimization of

$$
- \log p(M|D) \propto -\log p(D|M) - \log p(M)
$$

(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 $M$.

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

$$p(w|D,M) = \frac{p(D|w,M)p(w|M)}{p(D|M)} \quad (5.2)$$

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

$$p(D|w,M) = \prod_i p(x_i, y_i|w, M)$$

$$= \prod_i p(y_i|x_i, w, M)p(x_i) \quad (5.3)$$

and 5.2 is rewritten by

$$p(w|D,M) \propto p(w|M) \prod_i p(x_i, y_i|w, M). \quad (5.4)$$

The priors $p(w|M)$ is characterized as

$$p(w|M) = \frac{\exp(-\lambda R_{reg}(w))}{Z_w(\lambda)} \quad (5.5)$$

where $Z_w(\lambda) = \int d\mathbf{w} \exp(-\lambda R_{reg}(\mathbf{w}))$. The choices of $\lambda$ and $R_{reg}(\mathbf{w})$ are often built on some assumption of $p(\mathbf{w})$. For instance, when $\mathbf{w}$ is the Gaussian prior as $p(\mathbf{w}) \propto \exp(-\lambda^2 \|\mathbf{w}\|^2)$, it may lead to maximum a posteriori (MAP) estimate:

$$-\log p(w|D,M) = -\frac{1}{2} \|\mathbf{w}\|^2 - \sum_i \log p(x_i) - \sum_i \log p(y_i|x_i, w) \quad (5.6)$$

when $p(y_i|x_i, w)$ is usually measured by $L_2$ metric, 5.6 becomes

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

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

- weight decay (Hinton, 1989): it was shown from the Bayesian perspective (Brumtine & Weigend, 1991; MacKay, 1992) that weight decay 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.
As summarized in Table 1, most regularizers correspond to the weight priors with different \textit{a priori} probability density functions (PDF)\textsuperscript{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 \( p(w) \propto \cosh^{-1/\beta}(\beta w) \), \( p(w) \) approximates Laplace prior and we have \( \frac{\partial}{\partial w_j} \sum_j |w_j| \approx \tanh(\beta w_j) \). In above sense, the regularizer is sometimes written as \( \|Dw\|^2 \) in place of \( \|Df\|^2 \).

\begin{table}[h]
\centering
\begin{tabular}{|c|c|c|}
\hline
\textbf{regularizer} & \textbf{PDF prior} & \textbf{comment} \\
\hline
constant & uniform distribution & uniform prior \\
\hline
\( w^2 \) & Gaussian distribution \( N(0,1) \) & weight decay \\
\hline
\( \frac{w^2}{1+w^2} \) & uniform + Gaussian & weight elimination \\
\hline
\( \ln(1+w^2) \) & Cauchy distribution & Cauchy prior \\
\hline
\( |w| \) & Laplace distribution & Laplace prior \\
\hline
\( \ln(\cosh(w)) \) & supergaussian distribution & supergaussian prior \\
\hline
\end{tabular}
\end{table}

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 \( f(x) \leftrightarrow F(s) \) is a Fourier pair, from

\textsuperscript{7}The complexity of weights is evaluated by the negative logarithm of probability density function at weight values.
Fourier transform’s property, we have $\frac{\partial f(x)}{\partial x} \leftrightarrow s\mathcal{F}(s)$. According to 3.6 and 3.48, regularization constitutes to minimization of the regularizer:

$$\min \{ \| Df \|^2 \} \leftrightarrow \min \{ \| Tf \|^2 \}.$$  

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

$$\| Tf \|^2 = \| \sum_i w_i s_i G(s_i) \|^2$$

$$\leq \sum_i |w_i|^2 \sum_i |s_i|^2 \sum_i \| G(s_i) \|^2 \leq \left( \sum_i |s_i|^2 + \sum_i |w_i|^2 + \sum_i \| G(s_i) \|^2 \right)^2.$$  

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

$$\min \{ \| Tf \|^2 \} \leftrightarrow \min \left\{ \sum_i |s_i|^2 + |w_i|^2 + \| G(s_i) \|^2 \right\}$$  

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

$$\log |G(s_i)|^2 = \log |G_i(x)|^2.$$  

\(^8\)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 $G(x_i, x_j)$ 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 $G_i$ ($i = 1, \cdots, \ell$) be the $i$th vector of matrix $G$, $G_{ij}$ be $j$th ($j = 1, \cdots, m$) component of vector $G_i$, normalizing every component $G_{ij}$ by its row vector $^9$

$$P_{ij} = \frac{|G_{ij}|^2}{\|G_i\|^2},$$

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

$$\sum_{j=1}^{m} P_{ij} = 1.$$  \hspace{1cm} (6.5)

And the information-theoretic entropy is defined by

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

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

$$\hat{H} = \frac{H}{\ell \log m}. \hspace{1cm} (6.7)$$

Hence the connection between 6.4 and 6.7 is bridged by

$$\min \left\{ \sum |G_x|^2 \right\} \leftrightarrow \min \left\{ \sum \log |G_x|^2 \right\} \leftrightarrow \min \{ \hat{H} \}.$$  

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 (Kamiura, 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 (Kamiura, 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):

$$I(A, B) = H(B) - H(B|A).$$

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 $E$ is defined as the difference between expected energy and entropy:

$$E = \sum_i p_i E_i - (-T \sum_i p_i \log p_i)$$

$$= \sum_i p_i E_i + T \sum_i p_i \log p_i$$

(7.1)

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

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

(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 $\lambda$ 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 $E$ 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

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

(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):

$$H = H_{\text{imported}} + H_{\text{generated}},$$

(7.4)

where $H_{\text{imported}}$ and $H_{\text{generated}}$ denote the imported and generated entropy of the open system respectively. Although $H_{\text{generated}}$ always increases, the decrease speed of $H_{\text{imported}}$ is always faster than the increase speed of $H_{\text{generated}}$, 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 $H_{\text{imported}}$ is information-theoretic entropy which we denote as $H_{\text{info}}$ and $H_{\text{generated}}$ is physical-thermodynamic entropy, the former deals with signal whereas the latter deals with substance (see Stubbs, 1991 for detailed discussions), the decrease of $H_{\text{imported}}$ corresponds to the regularized functional, viz. MinEnt regularization discussed in section 6. Hence the regularized cost functional can be written as

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

(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)

$$\mathcal{R} = E[(y - f(x))^2|x]$$

$$= E \left[ (y - E[y|x] + E[y|x] - f(x))^2 | x \right]$$

$$= E \left[ (y - E[y|x])^2 | x \right] + (E[y|x] - f(x))^2$$

(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 $\lambda$ 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 $x_i$ 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 $f(x)$ in terms of some type of semidefinite positive kernel function (not necessarily satisfying Mercel condition)

$$ f(x) = \sum_i \alpha_i K(x_i, x) + b $$

when $K(x_i, x_j) = < Df(x_i, \cdot), Df(x_j, \cdot) >$, the regularization network is particularly equivalent to SVM. Furthermore, as we know from the Green’s function

$$ \tilde{D}Df(x_i, x) = \delta_{x_i}(x) = \delta(x - x_i) $$

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

$$ G(x_i, x_j) = < DG(x_i, \cdot), DG(x_j, \cdot) > $$
$$ = < \Phi(x_i), \Phi(x_j) > $$

with $\Phi : x_i \rightarrow DG(x_i, \cdot)$.

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

$$ < Df, Dh >= \frac{1}{(2\pi)^{N/2}} \int_{\Omega} \frac{F(s)H^*(s)}{G(s)} ds $$

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

$$ ||Df||^2 = \frac{1}{(2\pi)^N/2} \int_{\Omega} \frac{||F(s)||^2}{G(s)} ds = ||f||_{H^m} $$

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

$$ G(x, x_i) = \frac{1}{(2\pi)^{N/2}} \int_{R^N} \exp(js(x - x_i))G(s)ds $$
$$ = \frac{1}{(2\pi)^{N/2}} \int_{R^N} \exp(jsx)G(s)\exp(-jsx_i)ds $$
$$ = G(x - x_i) $$

(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 $G$. For instance, taking QR decomposition $Gw = QR$, we may obtain the new expression after pruning some hidden nodes (see Appendix E for derivation)

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

(9.1)

where $\hat{G} = GL_1R_1^{-1} (w = [L_1 \ L_2])$, and the new weight vector $\hat{w}$ and $\hat{G}$ are calculated by

$$\hat{w}^T = w^T L_1 + wL_2(R_1^{-1}R_2)^T.$$

(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 $R_{emp}$, early stopping is similar to weight-decay regularization, the product of iteration index $t$ and the learning rate $\eta$ plays the role of regularization parameter $\lambda$, in the sense that the components of weight vector parallel to the eigenvectors of the Hessian satisfy (Bishop, 1995a)

$$w_i^{(t)} \simeq w^*, \quad \sigma_i \gg (\eta t)^{-1}$$

$$|w_i^{(t)}| \ll |w^*|, \quad \sigma_i \ll (\eta t)^{-1}$$

where $w^*$ denotes the desired minimum point in weight space, $\sigma_i$ is the eigenvalues of the Hessian matrix $H(w)$. 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.
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 prospective 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
electrifying 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 frame-
work (Vapnik, 1998a) seems to be a feasible framework and mathematical
tool for this goal. In addition, the algorithm implementation of regulariza-
tion 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)\(^\text{11}\). 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 deriva-
tive 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 \(\epsilon\) 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

$$
a_k = \frac{1}{\pi} \int_{-\epsilon}^{\epsilon} \cos k(s + \theta) g(\theta) d\theta
= \frac{1}{\pi} \cos k\xi \int_{-\epsilon}^{\epsilon} G(\theta) d\theta,
$$

(A.3)

$$
b_k = \frac{1}{\pi} \int_{-\epsilon}^{\epsilon} \sin k(s + \theta) g(\theta) d\theta
= \frac{1}{\pi} \sin k\xi \int_{-\epsilon}^{\epsilon} G(\theta) d\theta,
$$

(A.4)

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

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

(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 $L$ to function $f(x)$, we have

$$
L f(x) = L \int_{\mathbb{R}^N} G(x, \xi) \varphi(\xi) d\xi
= \int_{\mathbb{R}^N} L G(x, \xi) \varphi(\xi) d\xi
= \int_{\mathbb{R}^N} \delta(x - \xi) \varphi(\xi) d\xi
= \varphi(x).
$$

(B.1)

Similarly, applying $K$ to function $f(x)$

$$
K f(s) = K \int_{\mathbb{R}^N} G(x, \xi) \varphi(\xi) d\xi
= \int_{\mathbb{R}^N} K G(s, \xi) \varphi(\xi) d\xi
= \int_{\mathbb{R}^N} \exp(-js\xi) \varphi(\xi) d\xi
= \Phi(s).
$$

(B.2)
The solution of regularization problem is further derived by setting

\[ \varphi(\xi) = \frac{1}{\lambda} \sum_{i=1}^{\ell} [y_i - f(x_i)] \delta(\xi - x_i), \]  

(B.3)

\[ \Phi(\omega) = \mathcal{F}\{\varphi(\xi)\} = \mathcal{F}\left\{ \frac{1}{\lambda} \sum_{i=1}^{\ell} [y_i - f(x_i)] \delta(\xi - x_i) \right\} \]

\[ = \frac{1}{\lambda} \sum_{i=1}^{\ell} [y_i - f(x_i)] \exp(-j\omega x_i), \]  

(B.4)

then in spatial domain, we have

\[ f_\lambda(x) = \int_{\mathbb{R}^N} G(x, \xi) \left\{ \frac{1}{\lambda} \sum_{i=1}^{\ell} [y_i - f(x_i)] \delta(\xi - x_i) \right\} d\xi \]

\[ = \frac{1}{\lambda} \sum_{i=1}^{\ell} [y_i - f(x_i)] \int_{\mathbb{R}^N} G(x, \xi) \delta(\xi - x_i) d\xi \]

\[ = \frac{1}{\lambda} \sum_{i=1}^{\ell} [y_i - f(x_i)] G(x, x_i) \]

\[ = \sum_{i=1}^{\ell} w_i G(x, x_i), \]  

(B.5)

and equivalently in frequency domain

\[ f_\lambda(x) = \int_{\mathbb{R}^N} \mathcal{F}\{G(x, \xi)\} \mathcal{F}\left\{ \frac{1}{\lambda} \sum_{i=1}^{\ell} [y_i - f(x_i)] \delta(\xi - x_i) \right\} d\omega \]

\[ = \frac{1}{\lambda} \sum_{i=1}^{\ell} [y_i - f(x_i)] \int_{\mathbb{R}^N} G(x, \omega) \exp(jx_i \omega) d\omega \]

\[ = \frac{1}{\lambda} \sum_{i=1}^{\ell} [y_i - f(x_i)] G(x, x_i) \]

\[ = \sum_{i=1}^{\ell} w_i G(x, x_i), \]  

(B.6)

where \( w_i = [y_i - f(x_i)]/\lambda \). 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

\[ [U, V, Z, \Sigma, S] = \text{GSVD}(A, B) \]
where \( A, B \) is \( m \times p \) and \( n \times p \) matrix respectively; \( U_{m \times m}, V_{n \times n} \) are the unitary matrices, matrix \( Z_{p \times q} \) \((q = \min\{m + n, p\})\) is usually (not necessarily) square, \( \Sigma \) and \( S \) are diagonal matrices, all of which satisfy

\[
A = U\Sigma Z^T, \quad B = VSZ^T, \quad \Sigma^T \Sigma + S^T S = I.
\]

Suppose the on-diagonal singular values in the singular matrices \( \Sigma \) and \( S \) as \( \sigma_i \) 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 \( \int \), \( G(x, x_i) \) by \( K(x, x_i) \), \( w_i \) by \( [y_i - f(x_i)]/\lambda \), 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 \( K(x, x_i) = K(x - x_i) \), thus the approximated function can be expressed by a convolution of observation and a moving kernel

\[
f(x) = \int (y_i - f(x_i))K(x - x_i)dx_i
\]

\[
= -\frac{1}{\lambda} \int f(x_i)K(x - x_i)dx_i + \text{constant.} \quad (D.1)
\]

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

\[
f(x) = \sum_{i=1}^{\ell} c_i \Phi(x) \Phi(x_i). \quad (D.2)
\]

In above sense, one can imagine the functional \( f(x) \) is doing Fourier (frequency) analysis (or time-frequency analysis, depending on \( K \)), where the kernel function \( K(x, x_i) \) accounts for a convolutional window, the metric function \( \mu \) measures the distance between the synthesis function \( f(x) \) and the desired value \( y(x) \). 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 \( G \),

\[
GL = QR \quad (E.1)
\]

where \( G \) 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

\[
R = \begin{bmatrix} R_1 & R_2 \\ 0 & 0 \end{bmatrix}, \quad Q = \begin{bmatrix} Q_1 & Q_2 \end{bmatrix}
\]

(E.2)

henceforth,

\[
GL = \begin{bmatrix} GL_1 & GL_2 \end{bmatrix} \quad (E.3)
\]

and it may further follow that

\[
Q_1 = GL_1 R_1^{-1}, \quad GL_2 = Q_1 R_2 = GL_1 R_1^{-1} R_2,
\]

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

\[
y = Gw \rightarrow \hat{\gamma} = \hat{G}\hat{w}
\]

where

\[
\hat{G} = Q_1 = GL_1R_1^{-1}.
\]

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

\[
\hat{w}^T = w^T L_1 + wL_2(R_1^{-1}R_2)^T.
\]

Acknowledgements

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

References


Regularization theory


