How to Design a Regularization Term for Improving Generalization

Akiko NAKASHIMA and Hidemitsu OGAWA
Department of Computer Science,
Tokyo Institute of Technology,
2-12-1, O-okayama, Meguro-ku, Tokyo, 152-8552, Japan.
email: akko@cs.titech.ac.jp

Abstract

In supervised learning, the regularization method is often used for improving the level of generalization. In this paper we give a necessary and sufficient condition of an optimal regularization term, i.e., a regularization operator and parameter. The optimality is discussed based on the projection learning criterion in which the minimization of a generalization error is explicitly considered. We suggest how to design the optimal regularization term so as to satisfy the obtained condition.

1 Introduction

The error back-propagation method is used for training a feedforward neural network [1]. However, this method does not always provide a high level of generalization. Since it is an algorithm to minimize the training error, the trained network tends to overfit noisy training examples. In order to improve the level of generalization, the regularization method is often used [2]. In this method, the sum of training error and a regularization term is minimized.

For implementation of regularization learning, we need to determine a type of the regularization term and a value of the regularization parameter in advance. In conventional cases, first, a type of the regularization term is determined, and then a value of the regularization parameter is decided. As a type of the regularization term, smoothness of the learning result is often considered. The smoothness has a variety of mathematical representations such as the first- or the second-order derivative of the learning result, the number of hidden units, magnitude of weights in a neural network. A value of the regularization parameter is decided by using the cross-validation method [3], bootstrapping [4], the Bayesian method [5], and so on.

In this paper, the generalization capability of regularization learning is discussed based on the projection learning criterion [6] in which the minimization of a generalization error is explicitly considered. We derive a condition for regularization learning to provide the same level of generalization as projection learning. We suggest how to design a regularization term, i.e., a type of the regularization term and a value of the regularization parameter so as to satisfy the obtained condition.

2 Supervised learning as an inverse problem

In this section, we shall present the basic framework of the supervised learning problem according to [7]. Let us begin by considering a three-layer feedforward neural network whose numbers of input, hidden, and output units are L, N, and 1, respectively. Let x be the L dimensional vector consisting of L inputs to the input layer units. The input-output relation of the network can be regarded as a function of L variables. It is denoted by $f_0(x)$. Let $f(x)$ be a real valued target function. A set of M training examples is denoted by $\{x_m, y_m\}_{m=1}^M$, where x_m is a given input vector and y_m is the corresponding output written by

$$y_m = f(x_m) + n_m \tag{1}$$

with additive noise n_m. $\{x_m\}_{m=1}^M$ and $\{y_m\}_{m=1}^M$ are referred to as a *training set* and a *set of teacher signals*, respectively. The *supervised learning problem* is to obtain the best approximation $f_0(x)$ to the target function $f(x)$ from $\{x_m, y_m\}_{m=1}^M$.

For the training set $\{x_m\}_{m=1}^M$, the corresponding true outputs $\{f(x_m)\}_{m=1}^M$ are uniquely determined from f. Hence, we can introduce the *sampling operator* A which maps f to the vector consisting of

0-7803-5871-6/99/$10.00 © 1999 IEEE

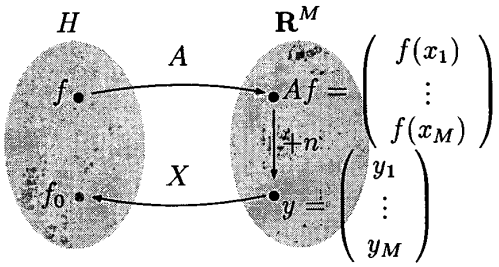

Figure 1: Formalization of learning problem. The supervised learning problem is formalized as an inverse problem of obtaining a learning operator $A^{(J)}$ satisfying a learning criterion J.

$\{f(x_m)\}_{m=1}^M$:

$$Af = (f(x_1), f(x_2), \cdots, f(x_M))^T, \tag{2}$$

where T denotes the transpose of the vector. Let y and n be the M-dimensional vectors consisting of elements $\{y_m\}_{m=1}^M$ and $\{n_m\}_{m=1}^M$, respectively. y is referred to as a *teacher vector*. From Eqs.(1) and (2), y is expressed by

$$y = Af + n. \tag{3}$$

Assume that f belongs to a real reproducing kernel Hilbert space (RKHS) denoted by H. In practice, the space H is determined by, for example, a model selection technique. Let D be the domain of f, which is a subset of the L-dimensional Euclidean space $\mathbf{R}^L$. The reproducing kernel, $K(x, x')$ of H, is a bivariate function defined on $D \times D$ [8]. For any f in H and x' in D, it holds that

$$\langle f(\cdot), K(\cdot, x')\rangle = f(x'), \tag{4}$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product in H. Due to the property (4), the sampling operator A is written in the form

$$A = \sum_{m=1}^M e_m \otimes \overline{K(x, x_m)}, \tag{5}$$

where $\{e_m\}_{m=1}^M$ is the so-called standard basis in $\mathbf{R}^M$, The notation $(\cdot \otimes \bar{\cdot})$ is the Neumann-Schatten product defined by

$$(e_m \otimes \overline{g})f = \langle f, g\rangle e_m. \tag{6}$$

Note that A is a linear operator even when we are concerned with the nonlinear function f.

Now the supervised learning problem is the problem of obtaining the best approximation, say f_0, to f from y. This can be considered as an *inverse problem* of obtaining an operator X which provides f_0 from y:

$$f_0 = Xy. \tag{7}$$

X is called a *learning operator*. It can be optimized based on different learning criteria. We denote a criterion by J in general, and the operator X satisfying J by $A^{(J)}$. In this paper, X is a linear operator, which does not mean that f and f_0 are linear functions of x. The above formalization of the learning problem is illustrated in Fig.1.

3 Regularization Learning

For training a neural network, the training error given by

$$\sum_{m=1}^M (f_0(x_m) - y_m)^2 \tag{8}$$

is often minimized by using the error back-propagation algorithm [1]. This criterion does not require that the error for novel inputs is minimized. Therefore, even if the training error is minimized, a high level of generalization is not always achieved. Moreover, the training error is measured between outputs of the neural network and noisy teacher signals. Thus, the learning result tends to overfit the noisy training examples. This phenomenon is called *overfitting*. It is one of serious problems in the learning problem [9]. The criterion which requires minimization of Eq.(8) is called the *rote memorization learning criterion*. Let S_t be a subspace in $\mathbf{R}^M$ defined as

$$S_t = \mathcal{R}(A) + \mathcal{R}(Q), \tag{9}$$

where $\mathcal{R}(A)$ is the range of A, and Q is the correlation matrix of noise defined by

$$Q = \underset{n}{E}(nn^T). \tag{10}$$

Since noisy teacher vector y lies in S_t, the rote memorization learning criterion is written as follows:

Definition 1 (*Rote memorization learning*) If an operator X minimizes the functional

$$J_{RM}[X] = \|Af_0 - y\|^2 \tag{11}$$

for any y in S_t, then X is called a *rote memorization learning* (RML) *operator* and denoted by $A^{(RM)}$, where $\|\cdot\|$ is the norm in $\mathbf{R}^M$.

Let $\mathcal{R}(A)^\perp$ be the orthogonal complement of $\mathcal{R}(A)$ and $P_{\mathcal{R}(A)^\perp}$ be the orthogonal projection matrix onto $\mathcal{R}(A)^\perp$. $P_{\mathcal{R}(A)^\perp}$ is given by

$$P_{\mathcal{R}(A)^\perp} = I_M - AA^\dagger, \tag{12}$$

where I_M is the identity matrix in $\mathbf{R}^M$ and $A^\dagger$ is the Moore-Penrose generalized inverse of A [10]. The orthogonal projection matrix onto S_t denoted by P_{S_t} is given by

$$P_{S_t} = AA^\dagger + P_{\mathcal{R}(A)^\perp} Q (P_{\mathcal{R}(A)^\perp} Q)^\dagger. \tag{13}$$

A general form of the RML operator is given by

$$A^{(RM)} = A^\dagger + Y - A^\dagger A Y P_{S_t}, \tag{14}$$

where Y is an arbitrary operator from $\mathbf{R}^M$ to H.

In order to improve generalization, the regularization method is often appl ied. In this method, some functional of f_0 is added to Eq.(11). It is expressed by $\|Lf_0\|^2$, where $\|\cdot\|$ is the norm in a RKHS H', and L is an operator with closed range from H to H'. Note that H' is not limited in function spaces. It can be a vector space.

Definition 2 (*Regularization learning*) Let L be any fixed closed range operator from H to H' and α be any fixed positive constant. If an operator X minimizes the functional

$$J_R[X] = \|Af_0 - y\|^2 + \alpha \|Lf_0\|^2 \tag{15}$$

for any y in S_t, then X is called a *regularization learning* (RL) *operator* and denoted by $A^{(R)}$.

L and α are called a regularization operator and a regularization parameter, respectively. Let S be the subspace in H defined as

$$S = \mathcal{R}(A^*) + \mathcal{R}(L^*). \tag{16}$$

Let $\mathcal{N}(A)$ be the null space of A. The orthogonal projection operator onto $\mathcal{N}(A)$ is denoted by $P_{\mathcal{N}(A)}$. It is given by

$$P_{\mathcal{N}(A)} = I - A^\dagger A, \tag{17}$$

where I is the identity operator on H. The orthogonal projection operator onto S denoted by P_S is given by

$$P_S = A^\dagger A + P_{\mathcal{N}(A)} L^* (P_{\mathcal{N}(A)} L^*)^\dagger. \tag{18}$$

A general form of the RL operator is given by

$$A^{(R)} = (A^*A + \alpha L^* L)^\dagger A^* + Y - P_S Y P_{S_t}. \tag{19}$$

Since the learing result depends on L and α, it is important to determine an adequate L and α in advance so that a high level of generalization can be achieved. Therefore, we will show how to design an adequate L and α to improve generalization capability.

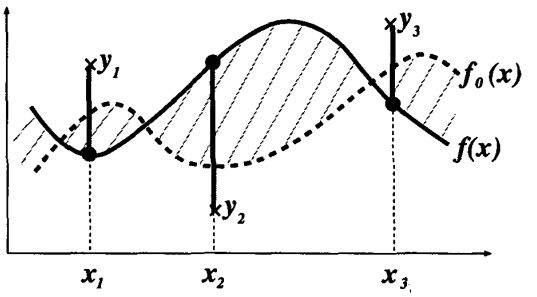

Figure 2: Generalization error. The solid and dotted lines are a target function $f(x)$ and a learning result $f_0(x)$, respectively. The generalization error is measured by the value corresponding to the shaded region.

4 Conditions for improving generalization capability

Generalization capability is estimated by generalization error which is measured in various ways. For example,

$$\int (f_0(x) - f(x))^2 dx \tag{20}$$

may be considered, which corresponds to the shaded region in Fig.2. Since f_0 depends on the nature of noise, Eq.(20) is averaged over the noise ensemble n:

$$\underset{n}{E} \int (f_0(x) - f(x))^2 dx, \tag{21}$$

where E_n denotes the expectation over the noise ensemble $\{n\}$. More generally, Eq.(21) is expressed as

$$\underset{n}{E} \|f_0 - f\|^2, \tag{22}$$

where $\|\cdot\|$ is the norm in H. In this paper, we use Eq.(22) as the generalization error.

Considering the minimization of the generalization error in Eq.(22), projection learning was proposed [6]. When the number of training examples M is less than the dimension of H, it is impossible to reconstruct exactly the target function f from the given training examples. Hence, we search for the best approximation to f, i.e. f_0, in the subspace $\mathcal{R}(X)$. It is called a search space. The best approximation to f in $\mathcal{R}(X)$ is the orthogonal projection of f onto $\mathcal{R}(X)$. From Eqs.(7) and (3), we have

$$f_0 = XAf + Xn. \tag{23}$$

The first term XAf in the right-hand side of Eq.(23) is a signal component of f_0. It is independent of n in y. Hence, it is required that the signal component of f_0 agrees with the best approximation to f in $\mathcal{R}(X)$. That is

$$XAf = P_{\mathcal{R}(X)}f, \tag{24}$$

where $P_{\mathcal{R}(X)}$ is the orthogonal projection operator onto $\mathcal{R}(X)$. Substituting Eqs.(23) and (24) into eq.(22), we have

$$\underset{n}{E}\|f_0 - f\|^2 = \|(I - P_{\mathcal{R}(X)})f\|^2 + E_n\|Xn\|^2. \tag{25}$$

As the search space $\mathcal{R}(X)$ becomes larger, the first term in Eq.(25) becomes smaller. Hence, we require that X has the largest range of all operators satisfying Eq.(24). It is known that such a largest range is $\mathcal{R}(A^*)$ [6]. Under this requirements, the second term in Eq.(25) should be minimized. This idea leads us to the following definition.

Definition 3 (*Projection learning*)[6] If X minimizes the functional

$$J_P[X] = \underset{n}{E}\|Xn\|^2 \tag{26}$$

under the constraint

$$XA = P_{\mathcal{R}(A^*)}, \tag{27}$$

then X is called a *projection learning* (PL) *operator* and denoted by $A^{(P)}$.

A general form of the PL operator is given by

$$A^{(P)} = V^\dagger A^* U^\dagger + Y(I_M - P_{S_t}), \tag{28}$$

where $U = AA^* + Q$, $V = A^*U^\dagger A$, and Y is an arbitrary operator from $\mathbf{R}^M$ to H. In PL the signal component XAf of f_0 agrees with the best approximation to f in $\mathcal{R}(A^*)$. Hence, when the number of training examples is large enough for $\mathcal{R}(A^*)$ to agree with H, XAf agrees exactly with the target function f.

If all RL operators satisfy the PL criterion, RL can provides the same level of generalization as PL. In order to obtain the condition for the RL operators to satisfy the PL criterion, the concept of forward admissibility [7] is useful. Let a set of all $A^{(J)}$ be denoted by $A\{J\}$.

Definition 4 (*Forward admissibility*) If all RL operators satisfy the PL criterion, i.e., if it holds that

$$A\{J_R\} \subset A\{J_P\}, \tag{29}$$

then it is said that PL *always admits*, or *admits* for short, RL. In this case, it is also said that *forward admissibility* holds.

We have the following theorem.

Theorem 1 (*Forward admissibility*) PL admits RL if and only if

$$\mathcal{N}(L) = \mathcal{R}(A^*) \tag{30}$$

and

$$Q\mathcal{R}(A)^\perp \subset \mathcal{R}(A)^\perp \tag{31}$$

hold or

$$A = 0 \quad \text{and} \quad Q = 0 \tag{32}$$

hold.

Eq.(32) means that the teacher vector y is the zero vector because of Eq.(3). Therefore, Eqs.(30) and (31) are essential for forward admissibility. From these conditions, we suggest how to design an adequate regularization operator L and a parameter α in the next section.

5 Design of an optimal regularization term

Eq.(30) is expressed in terms of A and L, while Eq.(31) is written by A and Q. Hence, whenever Eq.(31) is satisfied, RL can achieve the same level of generalization as PL by designing a regularization operator L so as to satisfy Eq.(30). The operator L satisfying Eq.(30) is constructed as follows.

Theorem 2 Eq.(30) holds if and only if L is given by

$$L = TP_{\mathcal{N}(A)}, \tag{33}$$

where T is an arbitrary closed range operator from H to H' satisfying

$$\mathcal{N}(T) \cap \mathcal{N}(A) = \{0\}. \tag{34}$$

So far, the gradient operator, Laplacian, or the identity operator on H has been used for the regularization operator L [2]. These operators does not always satisfy Eq.(30) as shown in Section 7. That is, smoothing the learning result f_0 itself is not substantial for RL to provide the same level of generalization as PL. However, if we use these operators as T in Eq.(33), we can construct the regularization operator satisfying Eq.(30). Eq.(33) shows that smoothing the $\mathcal{N}(A)$-component of f_0 is useful.

Not only the regularization operator L but also the regularization parameter α needs to be determined for RL. In the conventional methods, first the regularization operator is determined, then the parameter α is adjusted by using some resampling method such as cross-validation. On the other

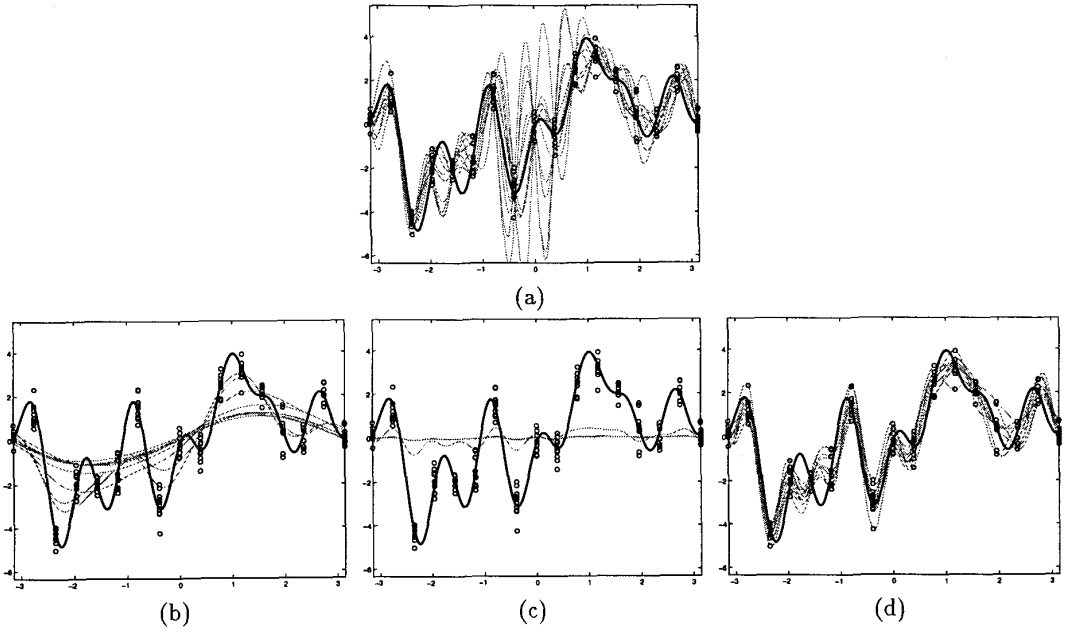

Figure 3: Simulation results. The circles denote the training examples. The black line shows the target function. The gray lines stand for the functions obtained by (a) RML, (b) not admitted RL using the second-order derivative, (c) not admitted RL using the norm of the learning result, and (d) admitted RL.

hand, if the regularization operator L is designed by Eq.(33), it is not necessary to adjust the parameter α. Since Eq.(30) and (31) do not include the regularization parameter α, we can use any positive value for α.

6 Effectiveness of regularization

In order to clarify the effect of adding the regularization term to the training error as shown in Eq.(15), let us compare the conditions of RL and RML to be admitted by PL. PL admits RML if and only if

$$\mathcal{N}(A) = \{0\} \tag{35}$$

and Eq.(31) hold or Eq.(32) holds [11]. Eq.(32) has no meaning in the learning problem as mentioned in Section 4. Hence, Eqs.(35) and (31) are essential. Since Eq.(31) is a common condition to RL and RML, we shall discuss the case that the condition (31) is satisfied.

If Eq.(35) holds, then the regularization operator L satisfying Eq.(30) is zero because of Eq.(33). That is, the RL criterion in Eq.(15) is reduced to the RML criterion in Eq.(11). Hence, we do not need to add any regularization term.

If Eq.(35) does not hold, we can not expect RML to achieve the same level of generalization as PL. Even in this case, we can expect RL to achieve it if we use the regularization operator designed by Eq.(33).

Since A is determined from the training set $\{x_m\}_{m=1}^M$ as Eq.(5), whether Eq.(35) holds or not depends on the training set. For example, in order that Eq.(35) may hold, the number of training examples must be greater than or equal to the dimension of H. Therefore, if the number of training examples is less than the dimension of H, RML can not achieve the same level of generalization as PL. It is always the case for infinite dimensional H. Even in this case, however, RL can achieve it if the regularization operator L is designed by Eq.(33). Table 1 shows which learning method can achieve the same level of generalization as PL.

Table 1: Learning methods admitted by PL under the condition $Q\mathcal{R}(A)^{\perp} \subset \mathcal{R}(A)$.

	$\mathcal{N}(A) \neq \{0\}$	$\mathcal{N}(A) = \{0\}$
$\dim(H) < \infty$	RL	RML
$\dim(H) = \infty$	RL	—

7 Simulation

We show that RL admitted by PL can improve the generalization capability. Let H be the 21-dimensional function space spanned by

$$\{1, \sqrt{2}\sin nx, \sqrt{2}\cos nx\}_{n=1}^{10} \tag{36}$$

with the inner product defined as

$$\langle f, g\rangle = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x)g(x)dx \quad : f, g \in H. \tag{37}$$

The reproducing kernel of H is given by

$$K(x, x') = \begin{cases} \frac{\sin\frac{21(x-x')}{2}}{\sin\frac{(x-x')}{2}} & : x \neq x' \\ 21 & : x = x' \end{cases} \tag{38}$$

A training set $\{x_m\}_{m=1}^{17}$ is fixed as

$$x_m = -\pi + 2\pi(m-1)/16 \quad 1 \leq m \leq 17$$

Ten sets of training examples are sampled. Noise is generated from the normal distribution with zero-mean and $Q = 0.25I_{17}$. In this case, Eq.(31) holds, but Eq.(35) does not hold.

The learning results are shown in Fig.3. The black line shows the target function. The gray lines stand for the functions obtained by (a) RML, (b) and (c) not-admitted RL, and (d) admitted RL. In each of the graphs, ten learning results obtained from ten sets of training examples are shown. The regularization operator in (b) is $L = \frac{d^2}{dx^2}$. In (c), $L = I$. In (d), L is designed by Eq.(33) with $T = \frac{d^2}{dx^2}$. It is given by

$$L = \frac{d^2}{dx^2} - \sum_{m=1}^{17}\sum_{n=1}^{17}(K^{\dagger})_{mn}\frac{d^2}{dx^2}K(x, x_m)\otimes\overline{K(x, x_n)}, \tag{39}$$

where K is the $M \times M$ matrix whose mn-th element is given by $K(x_m, x_n)$. The regularization parameter is adjusted by the cross validation method in (b) and (c). In (d) α is determined arbitrarily.

In (a), the obtained functions are too rough. It is because the condition (35) is not satisfied. In (b) and (c), the obtained functions are too smooth and far from the target function. It shows that the cross validation method does not work well with the small number of training examples. In (d), the obtained functions are close to the target function on average. That is, we can improve generalization by admitted RL.

8 Conclusions

In this paper, we gave a necessary and sufficient condition for regularization learning to provide the same level of generalization as projection learning. We suggested how to design the regularization term so as to satisfy the obtained condition. The effectiveness of adding the regularization term is theoretically clarified.

Acknowledgement

One of the authors, A. Nakashima was supported by Research Fellowships of the Japan Society for the Promotion Science for Young Scientists.

References

[1] Rumelhart, D.E., Hinton, G.E., & Williams, R.J.(1986). Learning representations by back-propagating errors. *Nature, 323(6088)*, 533–536.

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

[3] Wahba, G.(1990). *Spline models for observational data.* Philadelphia: Society for Industrial and Applied Mathematics.

[4] Efron, B.(1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics, 7(1)*, 1-26.

[5] MacKay, D.J.C.(1992). Bayesian interpolation. *Neural Computation, 4(3)*, 415-447.

[6] Ogawa, H.(1987). Projection filter regularization of ill-conditioned problem. *Proc. of SPIE, Inverse Problems in Optics*, 808, 189–196.

[7] Ogawa, H.(1992). Neural network learning, generalization and over-learning. *Proc. of ICIIPS'92 (Beijing), 2*, 1–6.

[8] Bergman, S.(1970). *The kernel function and conformal mapping.* 2nd Edition, New York: American Mathematical Society.

[9] Bishop, C.M.(1995). *Neural networks for pattern recognition.* Oxford: Oxford Univ. Press.

[10] Groetsch, C.W.(1977). *Generalized inverses of linear operators.* New Yourk: Marcel Dekker.

[11] Hirabayashi, A., Ogawa, H, & Yamashita, Y. (1999). Admissibility of memorization learning with respect to projection learning in the presence of noise. *Trans. of the IEICE, E82-D(2)*, 488–496.