Available online at www.sciencedirect.com

SCIENCE DIRECT®

Neural Networks 18 (2005) 251–259

Neural Networks

www.elsevier.com/locate/neunet

An asymptotic statistical analysis of support vector machines with soft margins

Kazushi Ikeda*, Tsutomu Aoishi[1]

Graduate School of Informatics, Kyoto University, Yoshida-Honmachi, Sakyo, Kyoto 606-8501, Japan

Received 20 November 2003; revised 26 November 2004; accepted 26 November 2004

Abstract

The generalization properties of support vector machines (SVMs) are examined. From a geometrical point of view, the estimated parameter of an SVM is the one nearest the origin in the convex hull formed with given examples. Since introducing soft margins is equivalent to reducing the convex hull of the examples, an SVM with soft margins has a different learning curve from the original. In this paper we derive the asymptotic average generalization error of SVMs with soft margins in simple cases, that is, only when the dimension of inputs is one, and quantitatively show that soft margins increase the generalization error.

Keywords: Learning curve; Asymptotic generalization error; Soft margin; Reduced convex hull

1. Introduction

In recent years, support vector machines (SVMs) have attracted much attention as a new classification technique with good generalization ability in applications such as pattern classification (Cristianini & Shawe-Taylor, 2000; Schölkopf, Burges, & Smola, 1998; Smola, Bartlett, Schölkopf, & Schuurmans, 2000; Vapnik, 1995, 1998). The basic idea of SVMs consists of mapping input vectors into a high-dimensional feature space and separating the feature vectors linearly with the optimal hyperplane in terms of margins, i.e. the distances of given examples from a separating hyperplane.

To assure convergence in linearly inseparable cases and to avoid overfitting to noisy data or outliers in examples, soft margins have been introduced in SVMs, which make them less sensitive to given examples, using slack variables for margin-constraint violation (Cristianini & Shawe-Taylor, 2000; Vapnik, 1995, 1998).

The theoretical background for the generalization ability of SVMs has been presented mainly in a framework of probably approximately correct (PAC) learning (Cristianini & Shawe-Taylor, 2000; Vapnik, 1995, 1998) where a kind of complexity of a learning-machine class called the VC dimension plays an important role (Vapnik & Chervonenkis, 1971). Another criterion for measuring the generalization ability is the average generalization error (Amari, 1993; Amari, Fujita, & Shinomoto, 1992; Amari & Murata, 1993; Baum & Haussler, 1989; Ikeda & Amari, 1996; Opper & Haussler, 1991) called a learning curve. Studies of the learning curves of kernel methods, including SVMs, are still being developed both from a statistical mechanical approach (Dietrich, Opper, & Sompolinsky, 1999; Opper & Urbanczik, 2001; Risau-Gusman & Gordon, 2000) and an asymptotic statistical approach (Ikeda, 2003, 2004a,b).
The former approach takes noise into account in terms of a finite temperature (Opper & Urbanczik, 2001), not soft margins. So far, the latter approach has never considered soft margins. Intuitively speaking, introducing soft margins increases the generalization error if the given problem is linearly separable, although it is necessary in inseparable cases. In this paper, we review the geometrical meaning of soft margins and quantify the effects of soft margins on the asymptotic generalization ability in simple cases, that is, where the input space is one-dimensional. Although we analyze only noiseless and one-dimensional cases, the rather

* Corresponding author. Tel.: +81 75 753 5501; fax: +81 75 753 4755.
E-mail address: kazushi@i.kyoto-u.ac.jp (K. Ikeda).

[1] Present address: Tochigi R&D Center, Honda R&D Co., Ltd, Tochigi 321-3321, Japan.

0893-6080/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.neunet.2004.11.008

negative result shown later, that the generalization error goes up as the softness increases, is important since, in general, we do not know if the given data are noisy, nor do we know the degree of noise involved. This work gives practitioners who employ the soft-margin technique a warning about the risk in generalization performance.

2. Geometry of support vector machines

Here, we formulate SVMs and consider their geometrical meaning. Although an SVM non-linearly maps input vectors to the corresponding feature vectors, we regard the feature vectors as the input vectors and consider, for brevity, a homogeneous linear dichotomy called a Perceptron whose separating function is represented by $\mathbf{w}'\mathbf{x}$, where $'$ denotes the transpose. Note that an inhomogeneous linear dichotomy whose separating function is represented by $\mathbf{w}'\mathbf{x}+b$ is easily transformed to a homogeneous one $\tilde{\mathbf{w}}'\tilde{\mathbf{x}}$ using $\tilde{\mathbf{w}}'=(\tilde{\mathbf{w}}',b)$ and $\tilde{\mathbf{x}}'=(\mathbf{x}',1)$, which is referred to as lifting up (Fig. 1).

Suppose a set F_N of N examples $(\mathbf{x}_n, y_n)$, $n=1,2,\ldots,N$, is given. Then, since the margin of a separating hyperplane denoted by $\mathbf{w}$ is defined as the minimum distance between the examples and the hyperplane, it is expressed as $\min_n \mathbf{w}'\mathbf{f}_n/\|\mathbf{w}\|$, where $\mathbf{f}_n=y_n\mathbf{x}_n$. Note that introducing $\mathbf{f}_n$ can be regarded as making all of the examples positive (Fig. 1). The problem of finding $\hat{\mathbf{w}}$ that maximizes the margin is equivalent to the following optimization problem with linear inequalities,

$$\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{w}\|^2 \quad \text{s.t. } \mathbf{w}'\mathbf{f}_n \ge 1, \quad n=1,\ldots,N, \tag{1}$$

if the given examples are linearly separable, i.e. there exists a hyperplane that separates the examples correctly. The above problem can be rewritten as

$$\min_{\mathbf{w}} \max_{\boldsymbol{\alpha}} L(\mathbf{w},\boldsymbol{\alpha}), \tag{2}$$

$$L(\mathbf{w},\boldsymbol{\alpha}) = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{n=1}^{N}\alpha_n(\mathbf{w}'\mathbf{f}_n - 1) \tag{3}$$

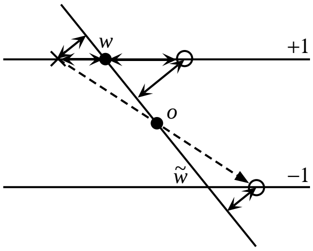

Fig. 1. Geometrical view of lifting up where the origin is denoted by O. Since the distances from $\tilde{\mathbf{w}}$ are proportional to those from $\mathbf{w}$, the lifting up does not change the problem of separating the examples at all. Neither does the transformation of a negative example shown by × to a positive one shown by ○.

using the Lagrangian multipliers $\alpha_n \ge 0$, $n=1,\ldots,N$, where $\boldsymbol{\alpha}=(\alpha_1,\ldots,\alpha_N)'$. Hence, by differentiating $L(\mathbf{w},\boldsymbol{\alpha})$ by $\mathbf{w}$ and $\boldsymbol{\alpha}$, the condition is derived under which $\mathbf{w}$ is a saddle point of $L(\mathbf{w},\boldsymbol{\alpha})$:

$$\frac{\partial L}{\partial \mathbf{w}} = \mathbf{w} - \sum_{n=1}^{N}\alpha_n\mathbf{f}_n = \mathbf{0}. \tag{4}$$

This means that the solution of the problem (1) is of the form

$$\hat{\mathbf{w}} = \sum_{n=1}^{N}\alpha_n\mathbf{f}_n, \quad 0 \le \alpha_n, \tag{5}$$

and, in addition, problem (1) is equivalent to

$$\min_{\boldsymbol{\alpha}}\left[\frac{1}{2}\|\mathbf{w}\|^2 - \sum_{n=1}^{N}\alpha_n\right], \tag{6}$$

which is a quadratic programming problem with linear constraints. This is called the dual problem of (1).

Let us consider a rather general problem

$$\min_{\mathbf{w},\beta}\left[\frac{1}{2}\|\mathbf{w}\|^2 - \beta\right] \quad \text{s.t. } \mathbf{w}'\mathbf{f}_n \ge \beta. \tag{7}$$

This is equivalent to what is called the ν-SVM (Schölkopf, Smola, Williamson, & Bartlett, 2000) without soft margins. It is obvious that this problem reduces to (1) if we fix β to unity and hence the solution of (1) is a suboptimal solution of (7). The dual problem of (7) is written as

$$\min_{\boldsymbol{\alpha}}\frac{1}{2}\|\mathbf{w}\|^2 \quad \text{s.t. } \mathbf{w}=\sum_{n=1}^{N}\alpha_n\mathbf{f}_n, \quad 0\le\alpha_n\le 1, \quad \sum_{n=1}^{N}\alpha_n = 1. \tag{8}$$

From a geometrical point of view, (8) means that the solution $\hat{\mathbf{w}}$ is the point nearest the origin in the convex hull of F_N, i.e. the smallest convex set that contains all points in F_N, where F_N is the set of vectors $\mathbf{f}_n$, $n=1,\ldots,N$.

Note that (Bennett & Bredensteiner, 2000) has considered two convex hulls in an affine space that consist of positive and negative examples, whereas we consider only a single convex hull of all of the given examples.

3. Geometry of SVMs with soft margins

When the example set F_N is not linearly separable, the margin cannot be positive and no $\hat{\mathbf{w}}$ exists that satisfies (1). This leads to the optimal $\boldsymbol{\alpha}$ diverging in (6) and the optimal $\mathbf{w}$ being $\mathbf{0}$ in (7) and (8). Hence, SVMs do not work properly in this situation. To cope with this limitation, slack variables ξ_n, $n=1,\ldots,N$, have been introduced that allow the margin constraints to be violated in the following way,

$$\min_{\mathbf{w},\xi}\left[\frac{1}{2}\|\mathbf{w}\|^2+C\sum_{n=1}^{N}\xi_n\right]\quad \text{s.t. } \mathbf{w}'\mathbf{f}_n\ge 1-\xi_n,\quad \xi_n\ge 0, \tag{9}$$

where C is a given constant and $\boldsymbol{\xi}=(\xi_1,\ldots,\xi_n)'$. The ν-SVM with soft margins is formulated in the same way as

$$\min_{\mathbf{w},\xi,\beta}\left[\frac{1}{2}\|\mathbf{w}\|^2+C\sum_{n=1}^{N}\xi_n-\beta\right]\quad \text{s.t. } \mathbf{w}'\mathbf{f}_n\ge \beta-\xi_n,\quad \xi_n\ge 0. \tag{10}$$

Note that this is of a slightly different form to the original ν-SVM introduced by Schölkopf et al. (2000). However, their equivalence can easily be proven taking into account $1/\nu=CN$. The dual problem of (10) is written as

$$\min_{\mathbf{w}}\frac{1}{2}\|\mathbf{w}\|^2\quad \text{s.t. } \mathbf{w}=\sum_{n=1}^{N}\alpha_n\mathbf{f}_n,\quad 0\le\alpha_n\le C,\quad \sum_{n=1}^{N}\alpha_n=1. \tag{11}$$

This means that $\hat{\mathbf{w}}$ has the same form as (8) but has different constraints, $0\le\alpha_n\le C$, which represent the so-called reduced convex hull introduced in Bennett and Bredensteiner (2000). We only consider the problem of the point in the reduced convex hull nearest the origin in the following sections.

Let us consider here a variant of the introduction of soft margins which employs the $L2$-norm instead of the $L1$-norm in (10), i.e.

$$\min_{\mathbf{w},\xi,\beta}\left[\frac{1}{2}\|\mathbf{w}\|^2+\frac{C}{2}\sum_{n=1}^{N}\xi_n^2-\beta\right]\quad \text{s.t. } \mathbf{w}'\mathbf{f}_n\ge \beta-\xi_n,\quad \xi_n\ge 0, \tag{12}$$

and its dual problem

$$\min_{\boldsymbol{\alpha},\xi}\left[\frac{1}{2}\|\mathbf{w}\|^2+\frac{C}{2}\sum_{n=1}^{N}\xi_n^2\right] \tag{13}$$

$$\text{s.t. } \mathbf{w}=\sum_{n=1}^{N}\alpha_n\mathbf{f}_n,\quad 0\le\alpha_n\le C\xi_n,\quad \sum_{n=1}^{N}\alpha_n=1.$$

Since it is clear that the optimal value of ξ_n is α_n/C, (13) can be rewritten as

$$\min_{\boldsymbol{\alpha}}\left[\frac{1}{2}\|\mathbf{w}\|^2+\frac{1}{2C}\sum_{n=1}^{N}\alpha_n^2\right] \tag{14}$$

$$\text{s.t. } \mathbf{w}=\sum_{n=1}^{N}\alpha_n\mathbf{f}_n,\quad 0\le\alpha_n,\quad \sum_{n=1}^{N}\alpha_n=1.$$

This problem is essentially the same as (8), the only difference being their kernel functions. That is, (14) employs

$$K(\mathbf{x}_n,\mathbf{x}_m)=\mathbf{x}_n'\mathbf{x}_m+\delta_{nm}/C$$

as a kernel function instead of the inner product $\mathbf{x}_n'\mathbf{x}_m$ in (8) where δ_{nm} is the Cronecker delta. Hence, we can rewrite (14) in the form of (8) by a certain coordinate transformation. So, we consider only (10) and (11) in the following.

4. Reduced convex hull

Fig. 2 shows examples of $\hat{\mathbf{w}}$ when $C=1$ and $C<1$. Suppose that C is the inverse of a natural number M. Then, the reduced convex hull of F_N with $C=1/M$ is equivalent to the convex hull of the set $\tilde{F}_N$ of ${}_NC_M$ vertices, each of which is a center of gravity of M distinct vectors in F_N. In fact, it is easily shown that any point in the reduced convex hull of F_N is written as a non-negatively weighted sum of vectors in $\tilde{F}_N$ (Ikeda & Aoishi, 2002). This means that soft margins reflect the ditribution of examples in the convex hull of F_N which is neglected in hard margins' case. For example, a large amount of area is reduced where examples are sparsely distributed (right area in Fig. 2), while less where examples are dense (left area in Fig. 2).

Since the center of M examples has a different distribution from the given examples, introducing soft margins in some sense means changing the distribution of the examples. Intuitively, averaging the data increases the signal-to-noise ratio if each input of the examples, $\mathbf{x}_n$, has an independent additive noise. In such a case, the separating hyperplane with hard margins is too sensitive to the noise and is expected to have a larger generalization error. Hence, the use of soft margins may improve the performance. In the noiseless case, however, averaging the data decreases the probability that points in the neighborhood of the separating hyperplane appear. This means that the learning machine loses information to some extent about the correct boundary and has a larger generalization error since the probability that a test input is chosen in the neighborhood of the separating hyperplane is unchanged.

In the following, we derive the learning curves of SVMs with soft margins in simple cases, providing the input space is one-dimensional and that the given data are noiseless and linearly separable. The derivation is performed using the fact that an SVM with soft margins given F_N is equivalent to an SVM with hard margins given the centers of M vectors in F_N as examples.

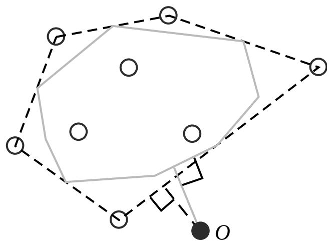

Fig. 2. Points nearest the origin O of the convex hull (the dashed line, $C=1$) and the reduced convex hull (the gray solid line, $C=1/2$) of examples shown by ○'s.

5. Average generalization error of SVMs

Let us fix the true parameter vector $\mathbf{w}^* = (0,\dots,0,1)' \in S^m$ and assume that N inputs, $\mathbf{x}_n$, $n=1,\dots,N$, are independently uniformly chosen from S^m, where S^m is an m-dimensional unit sphere. Then, the vectors $\mathbf{f}_n = y_n\mathbf{x}_n$, $n=1,\dots,N$, are uniformly distributed in S^m_+ where

$$y_n = \mathrm{sgn}(\mathbf{w}^{*'}, \mathbf{x}_n), \tag{15}$$

$$\mathrm{sgn}(s) = \begin{cases} +1, & \text{if} \quad s \ge 0, \\ -1, & \text{otherwise,} \end{cases} \tag{16}$$

$$S^m_+ = \{\mathbf{f} | \mathbf{w}^{*'}\mathbf{f} \ge 0, \mathbf{f} \in S^m\}. \tag{17}$$

In this case, the probability that an estimate $\hat{\mathbf{w}}$ mispredicts the output of a new input $\mathbf{x}$ is written as θ/π where θ is the angle between $\hat{\mathbf{w}}$ and $\mathbf{w}^*$ (Fig. 3). In this paper, we define the average generalization error as the probability that an estimate $\hat{\mathbf{w}}$ mispredicts the output of a new input averaged over the given examples, which is often termed the prediction error. In the following subsections, we derive the average generalization error of SVMs for $m=1$ in the asymptotic limit of $N \to \infty$.

5.1. Hard margins' case

We first consider the case of hard margins, $M=1$, where the nearest point in the convex hull of examples is the midpoint of the two examples nearest to both endpoints of the semicircle S^1_+. Let the two examples be denoted by $\mathbf{f}_L$ and $\mathbf{f}_R$, and their angles with the endpoints by θ_R and θ_L, respectively. Then, the SVM solution $\hat{\mathbf{w}}$ is written as $(\mathbf{f}_L + \mathbf{f}_R)/2$ and its angle θ with $\mathbf{w}^*$ becomes $\theta = |\theta_L - \theta_R|/2$, as shown in Fig. 4.

Since the examples are independently chosen, the probability that the angle Θ of the nearest point with an endpoint is less than θ_L is written as

$$\mathrm{Prob}[\Theta \le \theta_L] = 1 - \left(1 - \frac{\theta_L}{\pi}\right)^N,$$

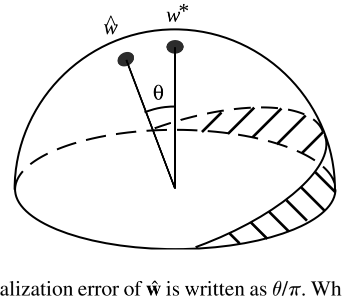

Fig. 3. The generalization error of $\hat{\mathbf{w}}$ is written as θ/π. When a new input $\mathbf{f}$ is chosen from the shadowed area in the input space S^m_+, the estimate $\hat{\mathbf{w}}$ mispredicts the output of $\mathbf{f}$.

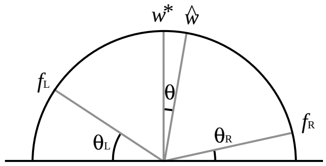

Fig. 4. SVM solution in the case of $m=1$.

and thus the density function is

$$p(\theta_L) = \frac{N}{\pi}\left(1 - \frac{\theta_L}{\pi}\right)^{N-1}.$$

In the same way, when θ_L is fixed, the conditional probability density function of θ_R, that is $p(\theta_R|\theta_L)$ is written as

$$p(\theta_R|\theta_L) = \frac{N-1}{\pi - \theta_L}\left(1 - \frac{\theta_R}{\pi - \theta_L}\right)^{N-2}.$$

Hence, the average generalization error $\epsilon_g(N)$ is written as

$$\begin{aligned}\epsilon_g(N) &= \left\langle \frac{|\theta_R - \theta_L|}{2\pi} \right\rangle \\ &= \frac{1}{2\pi}\int_0^\pi \int_0^{\pi - \theta_L} |\theta_L - \theta_R| p(\theta_R|\theta_L)\mathrm{d}\theta_R p(\theta_L)\mathrm{d}\theta_L \\ &= \frac{1}{2\pi}\int_0^\pi \int_0^{\pi - \theta_L} |\theta_L - \theta_R| \frac{N}{\pi} \\ &\quad \times \frac{N-1}{\pi}\left(1 - \frac{\theta_L + \theta_R}{\pi}\right)^{N-2} \mathrm{d}\theta_R \mathrm{d}\theta_L. \end{aligned} \tag{18}$$

By setting $s = \theta_L + \theta_R$ and $t = \theta_L - \theta_R$, (18) is calculated as

$$\begin{aligned}\epsilon_g(N) &= \frac{1}{2\pi}\int_0^\pi \int_{-s}^{s} |t| \frac{N}{\pi}\frac{N-1}{\pi}\left(1 - \frac{s}{\pi}\right)^{N-2} \frac{1}{2}\mathrm{d}t\mathrm{d}s \\ &= \frac{1}{2(N+1)}.\end{aligned}$$

5.2. Soft margins' case

In the case of a fixed $M \ge 2$, we consider the center of gravity of the nearest M examples from each endpoint. If we know the distributions of the angles θ_R and θ_L of $\mathbf{f}_L$ and $\mathbf{f}_R$ with endpoints, where $\mathbf{f}_L$ and $\mathbf{f}_R$ are the centers of the M examples nearest each of the endpoints, we can derive the average generalization error since the SVM solution $\hat{\mathbf{w}}$ is written as $(\mathbf{f}_L + \mathbf{f}_R)/2$ and its angle θ with $\mathbf{w}^*$ is $\theta = |\theta_L - \theta_R|/2$ as seen in the preceding subsection. Hence, we will consider the distributions first.

Let us denote by θ_l the angle between the $(l-1)$st and lth nearest examples to an endpoint named L (Fig. 5). Then, the conditional probability density function of each angle at

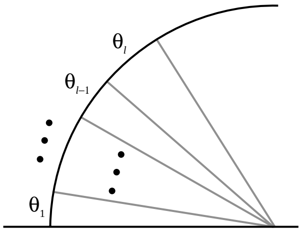

Fig. 5. Distribution of angles. θ_l is the angle between the $(l-1)$st and lth nearest examples.

the endpoint is written as

$$p(\theta_l|\theta_1,\ldots,\theta_{l-1})$$

$$=\frac{N-(l-1)}{\pi-\sum_{j<l}\theta_j}\left(1-\frac{\theta_l}{\pi-\sum_{j<l}\theta_j}\right)^{N-l}, \tag{19}$$

and that of the angle θ_{Rl} between the $(l-1)$st and lth nearest examples to the other endpoint named R

$$p(\theta_{Rl}|\theta_{R1},\ldots,\theta_{Rl-1},\theta_1,\ldots,\theta_l)$$

$$=\frac{N-2l+1}{\pi-\sum_{j<l}\theta_{Rj}-\sum_{j<l}\theta_j}\left(1-\frac{\theta_{Rl}}{\pi-\sum_{j<l}\theta_{Rj}-\sum_{j<l}\theta_j}\right)^{N-2l} \tag{20}$$

in the same way as in the case of hard margins. Let us show for the sake of clarification the case where $M=2$. Since the midpoint of the first and second nearest examples to an endpoint is a nearest center, its angle θ_L to the endpoint and its density $p(\theta_L)$ are written using (19) as

$$\theta_L=\frac{\theta_1+(\theta_1+\theta_2)}{2},$$

$$p(\theta_L)=\int_0^{2\theta_L}p_1\left(\theta_L-\frac{\eta}{2}\right)p_2\left(\eta|\theta_L-\frac{\eta}{2}\right)\mathrm{d}\eta$$

$$=\int_0^{2\theta_L}\frac{N}{\pi}\frac{N-1}{\pi}\left(1-\frac{\theta_L}{\pi}-\frac{\eta}{2\pi}\right)^{N-2}\mathrm{d}\eta$$

$$=\frac{2N}{\pi}\left(1-\frac{\theta_L}{\pi}\right)^{N-1}-\frac{2N}{\pi}\left(1-\frac{2\theta_L}{\pi}\right)^{N-1},$$

respectively, when $\theta_L\le\pi/2$, where p_1 and p_2 denote the density functions of θ_1 and θ_2, that is $p(\theta_1)$ and $p(\theta_2|\theta_1)$. When $\theta_L>\pi/2$

$$p(\theta_L)=\int_0^{\pi}\frac{N}{\pi}\frac{N-1}{\pi}\left(1-\frac{\theta_L}{\pi}-\frac{\eta}{2\pi}\right)^{N-2}\mathrm{d}\eta$$

$$=\frac{2N}{\pi}\left(1-\frac{\theta_L}{\pi}\right)^{N-1}-\frac{2N}{\pi}\left(1-\frac{\theta_L+1/2}{\pi}\right)^{N-1}.$$

For an arbitrary M, since the center of gravity of the M nearest examples to the endpoint L is a nearest point in the reduced convex hull, the angle θ_L satisfies

$$\theta_L=\frac{\theta_1+(\theta_1+\theta_2)+\cdots+\sum_{i=1}^{M}\theta_i}{M}, \tag{21}$$

$$\theta_L=\sum_{i=1}^{M}\frac{M+1-i}{M}\theta_i, \tag{22}$$

$$\theta_L=\sum_{i=1}^{M}\frac{i}{M}\theta_{M+1-i}, \tag{23}$$

and hence its density $p(\theta_L)$ is written as

$$p(\theta_L)=\frac{N}{\pi(M-1)!}$$

$$\times\sum_{i=1}^{M}i^{M-1}(-1)^{M-i}{}_MC_i\left(1-\frac{M\theta_L}{\pi i}\right)^{N-1}, \tag{24}$$

when $\theta_L<\pi/M$. See Appendix A for proof. Note that

$$\mathrm{Prob}\left[\theta_L>\frac{\pi}{M}\right]=o\left(\frac{1}{N}\right).$$

See Appendix B for proof.

In the same way, the density $p(\theta_R)$ of the other nearest center is calculated as

$$p(\theta_R)=\frac{N}{\pi(M-1)!}$$

$$\times\sum_{i=1}^{M}i^{M-1}(-1)^{M-i}{}_MC_i\left(1-\frac{M\theta_R}{\pi i}\right)^{N-1}, \tag{25}$$

when $\theta_R<\pi/M$. In the following, we use the approximation that

$$p(\theta_L,\theta_R)=p(\theta_L)p(\theta_R), \tag{26}$$

although θ_L and θ_R do not satisfy this since they are made from the same example set. In fact, the density $p(\theta_R|\theta_1,\theta_2,\ldots,\theta_M)$ can be derived in the same way as (19) and (24) as

$$p(\theta_R|\theta_1,\theta_2,\ldots,\theta_M)$$

$$=\frac{N-M}{\Pi_R(M-1)!}\sum_{i=1}^{M}i^{M-1}(-1)^{M-i}{}_MC_i\left(1-\frac{M\theta_R}{\Pi_R i}\right)^{N-1} \tag{27}$$

where $\Pi_R=\pi-\sum_{j\le M}\theta_j$ approaches π as the number N of examples increases, because the probability decreases exponentially that θ_l, $l=1,\ldots,M$ takes a large value from (19) and Appendix B. This means that the approximation above holds true asymptotically in the limit of $N\to\infty$.

Using the distribution (26), the asymptotic average generalization error for an arbitrary M can be derived as follows. Since the average prediction error $\epsilon_g(N)$ is written as

$$\epsilon_g(N)=\left\langle\frac{|\theta_R-\theta_L|}{2\pi}\right\rangle,$$

it has the following upper and lower bounds as

$$\frac{1}{2\pi}\int_0^{\delta/M}\int_0^{\delta/M}|\theta_L-\theta_R|p(\theta_L)p(\theta_R)\mathrm{d}\theta_L\mathrm{d}\theta_R \le \epsilon_\mathrm{g}(N)$$

$$\le \frac{1}{2\pi}\int_0^{\delta/M}\int_0^{\delta/M}|\theta_L-\theta_R|p(\theta_L)p(\theta_R)\mathrm{d}\theta_L\mathrm{d}\theta_R + \pi\ \mathrm{Prob}\left[\theta_L>\frac{\delta}{M}\right]+\pi\ \mathrm{Prob}\left[\theta_R>\frac{\delta}{M}\right]$$

for an arbitrary $0<\delta\le\pi$, which means that $\epsilon_\mathrm{g}(N)$ converges to the left-hand side as N increases because it is of the order $O(1/N)$, as shown below. From (24)–(26),

$$\frac{1}{2\pi}\int_0^{\delta/M}\int_0^{\delta/M}|\theta_L-\theta_R|p(\theta_R)p(\theta_L)\mathrm{d}\theta_R\mathrm{d}\theta_L$$
$$=\frac{1}{2\pi}\frac{N^2M^2}{\pi^2(M!)^2}\sum_{i=1}^M\sum_{j=1}^M i^{M-1}j^{M-1}(-1)^{i+j}{}_MC_{iM}C_j$$
$$\times\int_0^{\delta/M}\int_0^{\delta/M}|\theta_L-\theta_R|\left(1-\frac{M\theta_R}{\pi i}\right)^{N-1}$$
$$\times\left(1-\frac{M\theta_L}{\pi j}\right)^{N-1}\mathrm{d}\theta_R\mathrm{d}\theta_L. \tag{28}$$

In the following, we use the approximation that

$$\left(1-\frac{M\theta_R}{\pi i}\right)\left(1-\frac{M\theta_L}{\pi j}\right)=\left(1-\frac{M\theta_R}{\pi i}-\frac{M\theta_L}{\pi j}\right), \tag{29}$$

where $\theta_R\theta_L$ is neglected since it is less than δ^2/M^2 and δ can be arbitrary small. Then, the integral of (28) is calculated as

$$\int_0^{\delta/M}\int_0^{\delta/M}|\theta_L-\theta_R|\left(1-\frac{M\theta_R}{\pi i}-\frac{M\theta_L}{\pi j}\right)^{N-1}\mathrm{d}\theta_R\mathrm{d}\theta_L$$
$$=2\int_0^{\delta/M}\int_{\theta_L}^{\delta/M}(\theta_R-\theta_L)\left(1-\frac{M\theta_R}{\pi i}-\frac{M\theta_L}{\pi j}\right)^{N-1}\mathrm{d}\theta_R\mathrm{d}\theta_L$$
$$=2\int_0^{\delta/M}\frac{\pi^2i^2}{M^2N(N+1)}\left(1-\frac{(i+j)M\theta_L}{\pi ij}\right)^{N+1}\mathrm{d}\theta_L+o\left(\frac{1}{N}\right)$$
$$=\frac{2\pi^3i^3j}{M^3N(N+1)(N+2)(i+j)}+o\left(\frac{1}{N}\right).$$

Hence

$$\epsilon_\mathrm{g}(N)=\frac{1}{NM(M!)^2}\sum_{i,j=1}^M\frac{(-1)^{i+j}i^{M+2}j^M{}_MC_{iM}{}^MC_j}{i+j}+o\left(\frac{1}{N}\right). \tag{30}$$

For example, $\epsilon_\mathrm{g}(N)=(7/12N)+o(1/N)$ for $M=2$ and $\epsilon_\mathrm{g}(N)=(239/360N)+o(1/N)$ for $M=3$.

5.3. Asymptotic analysis

Suppose $1\ll M\ll N$, where M is so large that $1/M$ can be neglected compared to $O(1)$ and N is so large that the approximation (26) still holds true even when M is large. Then, an asymptotic analysis on M can be given as follows. From (19) and (20), the joint distribution of $\theta_1,\ldots,\theta_M,\theta_{R1},\ldots,\theta_{RM}$ is written as

$$p(\theta_1,\ldots,\theta_M,\theta_{R1},\ldots,\theta_{Rl})$$
$$=\frac{N!}{\pi^{2M}(N-2M)!}\left(1-\frac{\sum_{j\le M}\theta_j+\sum_{j\le M}\theta_{Rj}}{\pi}\right)^{N-2M}. \tag{31}$$

We approximate the above to an exponential distribution, that is

$$p(\theta_1,\ldots,\theta_M,\theta_{R1},\ldots,\theta_{Rl})$$
$$=\frac{N^{2M}}{\pi^{2M}}\exp\left[-\frac{N}{\pi}\left(\sum_{i=1}^M\theta_i+\sum_{j=1}^M\theta_{Rj}\right)\right], \tag{32}$$

which is based on neglecting the second- and higher-order terms of θ_l and θ_{Rl}, as is done in (29). From (32), the moment generating function $\phi(t)$ of $(\theta_L-\theta_R)/M^{1/2}$ is written as

$$\phi(t)=\left\langle\exp\left(t\frac{\theta_L-\theta_R}{M^{1/2}}\right)\right\rangle \tag{33}$$

$$=\prod_{i=1}^M\left\langle\exp\left(\frac{M+1-i}{M^{3/2}}t\theta_i\right)\right\rangle\times\prod_{j=1}^M\left\langle\exp\left(-\frac{M+1-j}{M^{3/2}}t\theta_{Rj}\right)\right\rangle \tag{34}$$

$$=\prod_{i=1}^M\frac{1}{1-\frac{\pi i}{NM^{3/2}}t}\prod_{j=1}^M\frac{1}{1+\frac{\pi j}{NM^{3/2}}t} \tag{35}$$

$$=\prod_{i=1}^M\frac{1}{1-\frac{\pi^2i^2}{N^2M^3}t^2}. \tag{36}$$

Hence the cumulant generating function $\psi(t)=\log\phi(t)$ is written as

$$\psi(t)=-\sum_{i=1}^M\log\left(1-\frac{\pi^2i^2}{N^2M^3}t^2\right) \tag{37}$$

$$=t^2\sum_{i=1}^M\frac{\pi^2i^2}{N^2M^3}+O\left(\frac{1}{N^2M}\right) \tag{38}$$

$$=\frac{\pi^2}{3N^2}t^2+O\left(\frac{1}{N^2M}\right). \tag{39}$$

This means that $(\theta_L-\theta_R)$ asymptotically obeys a normal distribution with mean zero and variance $(2\pi^2M/3N^2)$ when $M\to\infty$ and $N\to\infty$. Therefore, the average of the absolute value $|\theta_L-\theta_R|$ is asymptotically

$$\langle|\theta_L-\theta_R|\rangle=\frac{2\sqrt{\pi}M^{1/2}}{\sqrt{3}N}+o\left(\frac{M^{1/2}}{N}\right). \tag{40}$$

Hence, the average generalization error for a large M is

$$\epsilon_g(N) = \frac{M^{1/2}}{\sqrt{3\pi}N} + o\left(\frac{M^{1/2}}{N}\right). \tag{41}$$

6. Computer simulations

To confirm the validity of the theoretical analysis given above, some computer simulations were carried out. In each of the experiments below, the average generalization error was calculated as θ/π where θ is the angle between the true parameter $\mathbf{w}^* = (0, 1)'$ and the SVM solution $\hat{\mathbf{w}}$.

Fig. 6 shows the average generalization error versus the number of examples. The stars, crosses and circles show the experimental average generalization errors averaged over 100 trials for $M=1$, 3 and 5, respectively. The solid, dashed and dotted lines represent the theoretical average generalization errors for $M=1$, 3 and 5, respectively. The theoretical values agree well with the experimental data and the validity of the analysis is confirmed.

To see the relationship between soft margins and the average generalization errors more clearly, we plot the coefficient of the average generalization errors versus M in Fig. 7. Here, the crosses represent the experimental results with 1000 examples averaged over 3000 trials whereas the solid curve represents the theoretical value shown in (30). They agree well and the validity of the analysis is confirmed again.

Fig. 8 shows the coefficient of the average generalization errors versus M for large M's where the crosses represent the experimental results with 5000 examples averaged over 1000 trials, and the solid line represents the theoretical value shown in (39). Once again, the results validate the asymptotic analysis.

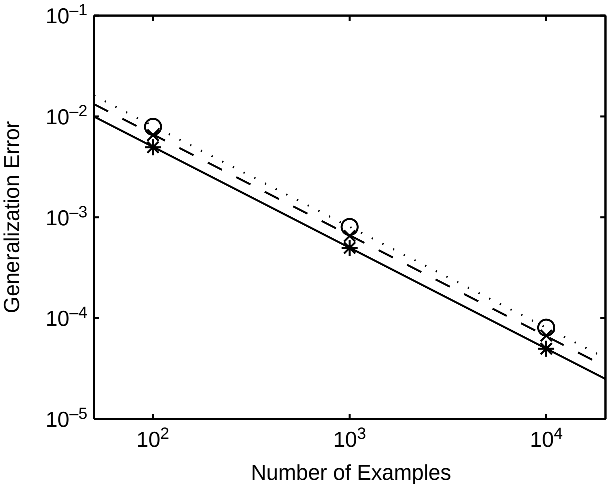

Fig. 6. Learning curve of SVMs with soft margins in the one-dimensional case.

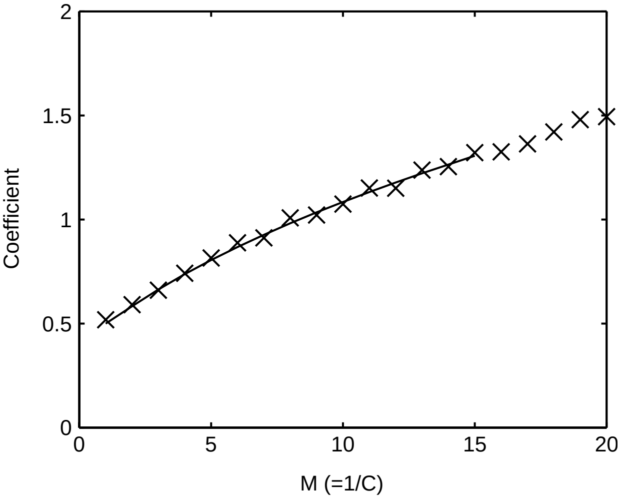

Fig. 7. Coefficient of the average generalization error versus the parameter $M=1/C$.

7. Conclusions and discussions

The effects of soft margins on the generalization ability of SVMs have been examined. We derived the asymptotic average generalization errors in the simple noiseless case of $m=1$ under the assumption that the parameter C, which represents the 'softness' of margins, is the reciprocal of a positive integer M. The results show that soft margins increase the generalization errors. Although we analyze only noiseless and one-dimensional cases, the results contribute to the knowledge of practitioners using the soft-margin technique since its risk in generalization performance has until now been unknown.

This can intuitively be interpreted as the fact that soft margins average the given data and decrease the probability that points lie in the neighborhood of the separating hyperplane, whereas the probability that a test input is

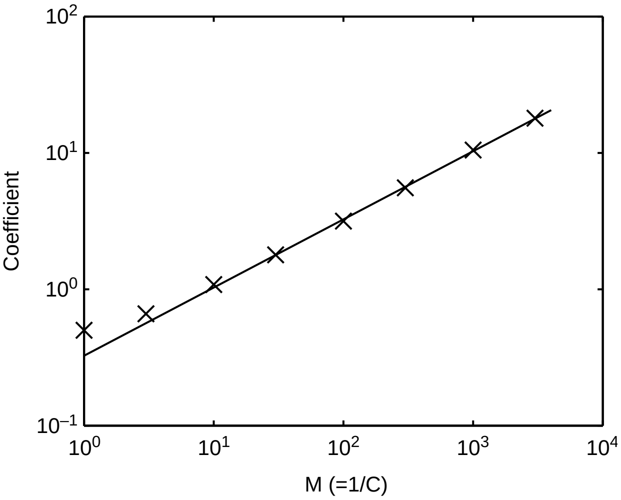

Fig. 8. Coefficient of the average generalization error versus the parameter $M=1/C \gg 1$.

chosen from this area remains constant. However, in general, the process of averaging increases the signal-to-noise ratio and improves the robustness against additive noise. Hence, when we have some knowledge about the noise in given data, it may be possible to choose a better parameter C. This work presents some fundamental research in this area. The analysis for more general cases including noisy data is the subject of future work.

Acknowledgements

The authors would like to thank Professor Hideaki Sakai for his valuable suggestions. This study is supported in part by a Grant-in-Aid for Scientific Research (14084210, 15700130) from the Ministry of Education, Culture, Sports, Science and Technology of Japan.

Appendix A. Density function of θ_L

Let $f_k(d)$ be the conditional probability density function of $d_k = \sum_{i=1}^{k}(i/M)\theta_{M+1-i}$ where N examples are uniformly chosen from S_+^1 and $\theta_1, \theta_2, \ldots, \theta_{M-k-1}$ are given. We will show that

$$f_k(d) = \frac{M(N-M+k)}{\Pi k!} \times \sum_{i=1}^{k} i^{k-1}(-1)^{k-i}{}_kC_i\left(1-\frac{Md}{\Pi i}\right)^{N-M+k-1}, \quad \text{(A1)}$$

by mathematical induction where $\Pi = \pi - \sum_{j<M-k}\theta_j$. If (A1) holds true for k, $f_{k+1}(d)$ can be written as

$$f_{k+1}(d) = \int_0^{Md/(k+1)} f_k\left(d - \frac{k+1}{M}\eta\right)p_{M-k}(\eta)\mathrm{d}\eta,$$

where p_{k+1} is the density function of θ_{k+1}, that is

$$p(\theta_{M-k}|\theta_1, \ldots, \theta_{M-k-1}) = \frac{N-M+k+1}{\Pi}\left(1-\frac{\theta_{M-k}}{\Pi}\right)^{N-M+k}. \quad \text{(A2)}$$

Hence

$$\begin{aligned} f_{k+1}(d) &= \int_0^{Md/(k+1)} \frac{M(N-M+k)}{(\Pi-\eta)k!} \\ &\times \sum_{i=1}^{k} i^{k-1}(-1)^{k-i}{}_kC_i\left(1-\frac{M(d-(k+1)\eta/M)}{(\Pi-\eta)i}\right)^{N-M+k-1} \\ &\times \frac{N-M+k+1}{\Pi}\left(1-\frac{\eta}{\Pi}\right)^{N-M+k}\mathrm{d}\eta \end{aligned} \quad \text{(A3)}$$

$$\begin{aligned} &= \frac{M}{k!}\sum_{i=1}^{k} i^{k-1}(-1)^{k-i}{}_kC_i\int_0^{Md/(k+1)}\frac{N-M+k+1}{\Pi} \\ &\times \frac{N-M+k}{\Pi}\left(1-\frac{Md-(k+1-i)\eta}{\Pi i}\right)^{N-M+k-1}\mathrm{d}\eta \end{aligned} \quad \text{(A4)}$$

$$\begin{aligned} &= \frac{M}{k!}\sum_{i=1}^{k} i^{k-1}(-1)^{k-i}{}_kC_i\frac{N-M+k+1}{\Pi} \\ &\times \frac{i}{k+1-i}\left[\left(1-\frac{Md}{\Pi(k+1)}\right)^{N-M+k} \right. \\ &\left. -\left(1-\frac{Md}{\Pi i}\right)^{N-M+k}\right] \end{aligned} \quad \text{(A5)}$$

$$\begin{aligned} &= \frac{M(N-M+k+1)}{\Pi(k+1)!}\sum_{i=1}^{k} i^{k}(-1)^{k+1-i} \\ &\times {}_{k+1}C_i \times\left[\left(1-\frac{Md}{\Pi i}\right)^{N-M+k} - \left(1-\frac{Md}{\Pi(k+1)}\right)^{N-M+k}\right] \end{aligned} \quad \text{(A6)}$$

$$\begin{aligned} &= \frac{M(N-M+k+1)}{\Pi(k+1)!} \\ &\times \sum_{i=1}^{k} i^{k}(-1)^{k+1-i}{}_{k+1}C_i^{k+1k+1}\left(1-\frac{Md}{\Pi i}\right)^{N-M+k} \\ &- \frac{M(N-M+k+1)}{\Pi(k+1)!}\left(1-\frac{Md}{\Pi(k+1)}\right)^{N-M+k} \\ &\times \sum_{i=1}^{k} i^{k}(-1)^{k+1-i}{}_{k+1}C_i^{k+1k+1} \end{aligned} \quad \text{(A7)}$$

$$\begin{aligned} &= \frac{M(N-M+k+1)}{\Pi(k+1)!} \\ &\times \sum_{i=1}^{k+1} i^{k}(-1)^{k+1-i}{}_{k+1}C_i\left(1-\frac{Md}{\Pi i}\right)^{N-M+k} \end{aligned} \quad \text{(A8)}$$

where the last equality is obtained using the fact that the summation of the second term in (A7) is equal to $-(k+1)^k$. Hence

$$p(\theta_L) = f_M(\theta_L) = \frac{N}{\pi(M-1)!}\sum_{i=1}^{M} i^{M-1}(-1)^{M-i}{}_MC_i\left(1-\frac{M\theta_L}{\pi i}\right)^{N-1}, \quad \text{(A9)}$$

Appendix B. Distribution of θ_L for large θ_L

From (23)

$$\theta_L \le \theta_1 + \frac{M-1}{M}\Sigma_2^M$$

necessarily holds where $\sum_{n_1}^{n_2} = \sum_{j=n_1}^{n_2} \theta_j$ and thus

$$\mathrm{Prob}\left[\theta_L > \frac{\delta}{M}\right] \le \mathrm{Prob}\left[\theta_1 + \frac{M-1}{M}\Sigma_2^M > \frac{\delta}{M}\right]$$

for an arbitrary $0 < \delta \le \pi$. We will show in the following that $\mathrm{Prob}[\theta_L > (\delta/M)]$ approaches null faster than $O(1/N)$ as N increases. Using the joint probability density function

$$p(\theta_1, \theta_2, \ldots, \theta_M) = \frac{N}{\pi} \cdots \frac{N-M+1}{\pi}\left(1 - \frac{\Sigma_1^M}{\pi}\right)^{N-M},$$

derived from (19)

$$\begin{aligned}
\mathrm{Prob}&\left[\theta_1 + \frac{M-1}{M}\Sigma_2^M \le \frac{\delta}{M}\right] \\
&= \int_0^{\delta/M} \int_0^{(\delta/M-1)-(M/M-1)\theta_1} \int_0^{(\delta/M-1)-(M/M-1)\theta_1-\Sigma_2^2} \cdots \\
&\quad \times \int_0^{(\delta/M-1)-M(M-1)\theta_1-\Sigma_2^{M-1}} p(\theta_1, \theta_2, \ldots, \theta_M)\mathrm{d}\theta_M \cdots \mathrm{d}\theta_1 \\
&= \int_0^{\delta/M} \cdots \int_0^{(\delta/M-1)-(M/M-1)\theta_1-\Sigma_2^{M-2}} \frac{N}{\pi} \cdots \\
&\quad \times \frac{N-M+2}{\pi}\left(1 - \frac{\Sigma_1^{M-1}}{\pi}\right)^{N-M+1} \mathrm{d}\theta_{M-1} \cdots \mathrm{d}\theta_1 \\
&\quad - \int_0^{\delta/M} \cdots \int_0^{(\delta/M-1)-(M/M-1)\theta_1-\Sigma_2^{M-2}} \frac{N}{\pi} \cdots \\
&\quad \times \frac{N-M+2}{\pi}\left(1 - \frac{\delta-\theta_1}{\pi(M-1)}\right)^{N-M+1} \mathrm{d}\theta_{M-1} \cdots \mathrm{d}\theta_1.
\end{aligned}$$

Since the second term has an upper bound

$$N^{M-1}\left(1 - \frac{\delta}{\pi M}\right)^{N-M+1},$$

which is $o(1/N)$, it can be neglected in asymptotic analyses of $O(1/N)$. By recursively integrating the first term

$$\begin{aligned}
\mathrm{Prob}&\left[\theta_1 + \frac{M-1}{M}\Sigma_2^M \le \frac{\delta}{M}\right] \\
&= \int_0^{\delta/M} \frac{N}{\pi}\left(1 - \frac{\theta_1}{\pi}\right)^{N-1} \mathrm{d}\theta_1 + o\left(\frac{1}{N}\right) = 1 + o\left(\frac{1}{N}\right).
\end{aligned}$$

Hence

$$\mathrm{Prob}\left[\theta_L > \frac{\delta}{M}\right] \le o\left(\frac{1}{N}\right).$$

References

Amari, S. (1993). A universal theorem on learning curves. *Neural Networks*, *6*, 161–166.

Amari, S., Fujita, N., & Shinomoto, S. (1992). Four types of learning curves. *Neural Computation*, *4*, 605–618.

Amari, S., & Murata, N. (1993). Statistical theory of learning curves under entropic loss criterion. *Neural Computation*, *5*, 140–153.

Baum, E. B., & Haussler, D. (1989). What size net gives valid generalization? *Neural Computation*, *1*, 151–160.

Bennett, K. P., & Bredensteiner, E. J. (2000). Duality and geometry in SVM classifiers. In P. Langley (Ed.), *Proceedings of the seventeenth international conference on machine learning* (pp. 57–64). San Francisco: Morgan Kaufmann, 57–64.

Cristianini, N., & Shawe-Taylor, J. (2000). *An introduction to support vector machines*. Cambridge, UK: Cambridge University Press.

Dietrich, R., Opper, M., & Sompolinsky, H. (1999). Statistical mechanics of support vector networks. *Physical Review Letters*, *82*(14), 2975–2978.

Ikeda, K. (2003). Generalization error analysis for polynomial kernel methods—algebraic geometrical approach. In O. Kaynak, et al. (Ed.), *Artificial neural networks and neural information processing—ICANN/ICONIP 2003* (pp. 201–208). New York: Springer, 201–208.

Ikeda, K. (2004a). Geometry and learning curves of kernel methods with polynomial kernels. *Systems and Computers in Japan*, *35*(7), 41–48.

Ikeda, K. (2004b). An asymptotic statistical theory of polynomial kernel methods. *Neural Computation*, *16*(8), 1705–1719.

Ikeda, K., & Amari, S. (1996). Geometry of admissible parameter region in neural learning. *IEICE Transactions on Fundamentals*, *E79*, 409–414.

Ikeda, K., & Aoishi, T. (2002). Perceptron learning admissible to noise. *Proceedings of forum on information technology, Tokyo, H-9* (in Japanese).

Opper, M., & Haussler, D. (1991). Calculation of the learning curve of bayes optimal classification on algorithm for learning a perceptron with noise. In L. G. Valiant, & M. K. Warmuth (Eds.), *Proceedings of the fourth annual workshop on computational learning theory* (pp. 75–87). San Francisco: Morgan Kaufmann, 75–87.

Opper, M., & Urbanczik, R. (2001). Universal learning curves of support vector machines. *Physical Review Letters*, *86*(19), 4410–4413.

Risau-Gusman, S., & Gordon, M. B. (2000). Generalization properties of finite-size polynomial support vector machines. *Physical Review E*, *62*, 7092–7099.

Schölkopf, B., Burges, C., & Smola, A. J. (1998). *Advances in kernel methods: Support vector learning*. Cambridge, UK: Cambridge University Press.

Schölkopf, B., Smola, A., Williamson, R., & Bartlett, P. L. (2000). New support vector algorithms. *Neural Computation*, *12*(5), 1207–1245.

Smola, A. J., Bartlett, P. L., Schölkopf, B., & Schuurmans, D. (2000). *Advances in large margin classifiers*. Cambridge, MA: MIT Press.

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

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

Vapnik, V. N., & Chervonenkis, A. Y. (1971). On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, *16*, 264–280.