EFFECTS OF NORMS ON LEARNING PROPERTIES OF SUPPORT VECTOR MACHINES

*Kazushi Ikeda**

Kyoto University
Graduate School of Informatics
Kyoto 606-8501 Japan

Noboru Murata

Waseda University
School of Science and Engineering
Shinjuku, Tokyo 169-8555 Japan

ABSTRACT

Support Vector Machines (SVMs) are known to have a high generalization ability, yet a heavy computational load since margin maximization results in a quadratic programming problem. It is known that this maximization task results in a pth-order programming problem if we employ the L_p norm instead of the L_2 norm. In this paper, we theoretically show the effects of p on the learning properties of SVMs by clarifying its geometrical meaning.

1. INTRODUCTION

In recent years, support vector machines (SVMs) have attracted much attention in the field of not only machine learning [1–4] but also signal processing [5]. The 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. distances of given examples from a separating hyperplane.

In the original SVMs, the distances of given examples from a separating hyperplane were evaluated in the 2-norm, that is, the Euclidean norm. We examine in this paper what happens if we employ the p-norm for an arbitrary $1 \le p \le \infty$. It is known that p affects the computational load of the problem [6, 7]. For example, when $p = 1$ or $p = \infty$, the problem of maximizing margins results in a linear programming problem, whereas a quadratic programming problem results when $p = 2$. However, it has not been clarified how the change of norm affects the learning properties of SVMs. An experimental result was reported in [6] in which the generalization error had very little dependency of p in computer simulations. We give a theoretical explanation on the above phenomena.

In this paper, we analyze the so-called linear kernels' case, that is, consider homogeneous linear dichotomies with an input vector $\boldsymbol{x} \in X$ whose corresponding output $y \in \{\pm 1\}$ is determined by $y = \mathrm{sgn}(\boldsymbol{w}'\boldsymbol{x})$ where $\boldsymbol{w}$ is the parameter vector, $'$ denotes the transpose and $\mathrm{sgn}(\cdot)$ outputs the sign of $\cdot$.

*This study is supported in part by a Grand-in-Aid for Scientific Research (14084210, 15700130) from the Japanese government.

In the following, we denote the nth of a given set of N examples by $(\boldsymbol{x}^{(n)}, y^{(n)})$ where $y^{(n)}$ is made with a fixed true parameter $\boldsymbol{w}^*$, that is,

$$y^{(n)} = \mathrm{sgn}(\boldsymbol{w}^{*\prime} x^{(n)}), \qquad n = 1, \ldots, N. \tag{1}$$

2. SUPPORT VECTOR MACHINES

An SVM chooses the separating hyperplane $\hat{\boldsymbol{w}}'\boldsymbol{x} = 0$ maximizing the margin which is defined as the minimum distance between examples and the hyperplane. Since the distance between an example $(\boldsymbol{x}^{(n)}, y^{(n)})$ and a hyperplane $\boldsymbol{w}'\boldsymbol{x} = 0$ is expressed as $\frac{\boldsymbol{w}'\boldsymbol{f}^{(n)}}{\|\boldsymbol{w}\|}$ where $\boldsymbol{f}^{(n)} = y^{(n)}\boldsymbol{x}^{(n)}$, $\boldsymbol{w}$ and $c\boldsymbol{w}$ have the same distance for $c > 0$ due to linearity. To absorb this ambiguity, an SVM sets the minimum distance to unity, that is, $\boldsymbol{w}'\boldsymbol{f}^{(n)} \ge 1$. Maximizing the margin, then, results in minimizing $\|\boldsymbol{w}\|$. Hence the problem of finding $\hat{\boldsymbol{w}}$ that maximizes the margin is equivalent to the following quadratic programming problem with linear inequalities,

$$\min_{\boldsymbol{w}} \frac{1}{2}\|\boldsymbol{w}\|^2 \qquad \text{s.t. } \boldsymbol{w}'\boldsymbol{f}^{(n)} \ge 1. \tag{2}$$

It is well known that problem (2) is equivalent to

$$\min_{\boldsymbol{\alpha}} \frac{1}{2}\|\boldsymbol{w}\|^2 - \sum_{n=1}^{N} \alpha_n \tag{3}$$

where $\alpha_n \ge 0$ are the Lagrangian multipliers, and

$$\boldsymbol{w} = \sum_{n=1}^{N} \alpha_n \boldsymbol{f}^{(n)}, \quad \alpha_n \ge 0, \tag{4}$$

which is also a quadratic programming problem with linear constraints. This is called the dual problem of (2). Note that the SVM solution $\hat{\boldsymbol{w}}$ necessarily has the form of (4).

As seen above, $\boldsymbol{x}^{(n)}$ and $y^{(n)}$ do not appear alone but necessarily in the form $y^{(n)}\boldsymbol{x}^{(n)} (= \boldsymbol{f}^{(n)})$. This means that the example $(\boldsymbol{x}^{(n)}, y^{(n)})$ is equivalent to $(\boldsymbol{f}^{(n)}, 1)$, and introducing $\boldsymbol{f}^{(n)}$, which is also called an example, can be regarded as making all the examples positive. It is easily shown that the problem of maximizing the margin is equivalent to finding the most distant hyperplane from the convex hull F_N of $D_N = \{\boldsymbol{f}^{(n)}\}_{n=1}^N$ where

$$F_N = \{\boldsymbol{f}|\boldsymbol{f} = \sum_{n=1}^{N} t_n \boldsymbol{f}^{(n)}, \sum_{n=1}^{N} t_n = 1, t_n \geq 0\}. \quad (5)$$

Although maximizing the margin is easy to understand intuitively, it is not applicable to a linearly inseparable set of examples, that is, there exists no parameter $\boldsymbol{w}$ such that $\boldsymbol{w}'\boldsymbol{f}^{(n)} \geq 0$ for all n. To assure convergence in linearly inseparable cases and to avoid overfitting to noisy data or outliers in examples, soft margins were introduced in SVMs [1]. They make the separating hyperplane less sensitive to given examples by using slack variables ξ_n, $\xi = 1, \ldots, N$, for margin-constraint violation and the problem is formulated as

$$\begin{aligned} &\min_{\boldsymbol{w},\boldsymbol{\xi}} \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{n=1}^{N}\xi_n \\ &\text{s.t. } \boldsymbol{w}'\boldsymbol{f}^{(n)} \geq 1 - \xi_n, \quad \xi_n \geq 0, \end{aligned} \quad (6)$$

where C is a given constant.

3. ν-SVM WITH A DIFFERENT NORM

Let us consider a rather general problem of (6):

$$\begin{aligned} &\min_{\boldsymbol{w},\boldsymbol{\xi},\beta} \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{n=1}^{N}\xi_n - \beta \\ &\text{s.t. } \boldsymbol{w}'\boldsymbol{f}^{(n)} \geq \beta - \xi_n, \quad \xi_n \geq 0. \end{aligned} \quad (7)$$

This is a variation of SVM called the ν-SVM [8]. This problem reduces to (6) if we fix β to unity and hence the solution of (6) is a suboptimal solution of (7). Its dual problem is written as

$$\begin{aligned} &\min_{\boldsymbol{w}} \frac{1}{2}\|\boldsymbol{w}\|^2 \\ &\text{s.t. } \boldsymbol{w} = \sum_{n=1}^{N}\alpha_n \boldsymbol{f}^{(n)}, 0 \leq \alpha_n \leq C, \quad \sum_{n=1}^{N}\alpha_n = 1. \end{aligned} \quad (8)$$

When $C \geq 1$, the solution $\hat{\boldsymbol{w}}$ of (8) corresponds to the point nearest the origin in the convex hull F_N of D_N. Hence its geometrical meaning in the parameter space becomes clear as the original SVM with hard margins shown in (2) can be understood intuitively in the input space. Other formulations, such as (3) and (6), do not have clear intuitive explanations.

When $C < 1$, we can consider the reduced convex hull of D_N instead of F_N. See [9–11] for the meaning and characteristics of reduced convex hulls. We discuss the ν-SVM that employs the p-norm instead of 2-norm, which is defined as

$$\|\boldsymbol{w}\|_p = \begin{cases} \left(\sum_{i=1}^{M}|w_i|^p\right)^{1/p} & 1 \leq p < \infty \\ \max_{1\leq i\leq M}|w_i| & p = \infty \end{cases} \quad (9)$$

where w_i denotes the ith element of $\boldsymbol{w}$. We use the term ν-SVM$^{(p)}$ to refer to it in this paper. In [6], they applied this idea to the original SVM described as (6) for $p = 1$ and $p = \infty$ and showed that in that case, maximizing the margin resulted in a linear programming problem. See also [7] for the relationship between the norm and the computational complexity. Note that, for $p \in (1, \infty)$, the q-norm satisfying

$$1/p + 1/q = 1 \quad (10)$$

is called the dual norm of the p-norm. In the case of $p = 1$, the ∞-norm is the dual norm of the 1-norm and vice versa. They satisfy Hölder's inequality,

$$\boldsymbol{w}'\boldsymbol{v} \leq \|\boldsymbol{w}\|_p\|\boldsymbol{v}\|_q. \quad (11)$$

The distance of an example $\boldsymbol{x}^{(n)}$ from a hyperplane $\boldsymbol{w}'\boldsymbol{x} = 0$ in the p-norm is defined as

$$\min_{\boldsymbol{x}} \|\boldsymbol{x}^{(n)} - \boldsymbol{x}\|_p \qquad \text{s.t. } \boldsymbol{w}'\boldsymbol{x} = 0. \quad (12)$$

Hence the minimizer denoted by $\hat{\boldsymbol{x}}$ is calculated using the Lagrangian function

$$L(\boldsymbol{x}, \lambda) = \|\boldsymbol{x}^{(n)} - \boldsymbol{x}\|_p^p + \lambda\boldsymbol{w}'\boldsymbol{x} \quad (13)$$

where λ is the Lagrangian multiplier and the distance is written as $\frac{|\boldsymbol{w}'\boldsymbol{x}^{(n)}|}{\|\boldsymbol{w}\|_q}$ where q satisfies (10). Hence, the ν-SVM$^{(p)}$ is formulated as

$$\begin{aligned} &\min_{\boldsymbol{w},\boldsymbol{\xi},\beta} \frac{1}{q}\|\boldsymbol{w}\|_q^q + C\sum_{n=1}^{N}\xi_n - \beta \\ &\text{s.t. } \boldsymbol{w}'\boldsymbol{f}^{(n)} \geq \beta - \xi_n, \quad \xi_n \geq 0 \end{aligned} \quad (14)$$

for $1 < p < \infty$. It is easily shown that the dual problem of (14) is described as

$$\begin{aligned} &\min_{\boldsymbol{v}} \frac{1}{p}\|\boldsymbol{v}\|_p^p \\ &\text{s.t. } \boldsymbol{v} = \sum_{n=1}^{N}\alpha_n \boldsymbol{f}^{(n)}, 0 \leq \alpha_n \leq C, \quad \sum_{n=1}^{N}\alpha_n = 1, \end{aligned} \quad (15)$$

which means that the ν-SVM$^{(p)}$ is a problem of finding the point in a reduced convex hull nearest the origin in the p-norm. The solution of (14) denoted by $\hat{\boldsymbol{w}}$ is given by

$$\hat{w}_i = \text{sgn}(\hat{v}_i)|\hat{v}_i|^{p-1} \quad (16)$$

where $\hat{\boldsymbol{v}}$ denotes the solution of (15).

4. NEAREST POINT AND ESTIMATOR

As shown in the preceding section, the estimated parameter $\hat{\boldsymbol{w}}$ by ν-SVM$^{(p)}$ is determined through the nearest point $\hat{\boldsymbol{v}}$ in the p-norm. Here we discuss the geometrical relationship between $\hat{\boldsymbol{w}}$ and $\hat{\boldsymbol{v}}$, more specifically, its dependency on the norm. For brevity, we set $C = 1$ and consider only the ν-SVM$^{(p)}$ with hard margins, since the case of the ν-SVM$^{(p)}$ with soft margins is straightforward by considering the reduced convex hull.

4.1. Estimated Parameter for a Hyperplane

Let us consider the point $\hat{\boldsymbol{v}}$ in a hyperplane nearest the origin in the p-norm. Without loss of generality, we assume that the hyperplane is expressed as $\boldsymbol{e}'\boldsymbol{v} = d(> 0)$ where $\boldsymbol{e}$ is the normal vector of the hyperplane and its elements e_i are positive. Then, the problem (15) is written as

$$\min_{\boldsymbol{v}} \frac{1}{p}\|\boldsymbol{v}\|_p^p \quad \text{s.t. } \boldsymbol{e}'\boldsymbol{v} = d, \quad \boldsymbol{e} \geq 0, \quad \boldsymbol{v} \geq 0 \tag{17}$$

since all elements of the solution $\hat{\boldsymbol{v}}$ are positive. This is proven using the fact that

$$\sum_i e_i|v_i| \geq \sum_i e_i v_i = d \tag{18}$$

whereas $(v_1, v_2, \ldots, v_M)$ and $(|v_1|, |v_2|, \ldots, |v_M|)$ have the same p-norm. The Lagrangian function of (17) is written as

$$L(\boldsymbol{v}, \lambda) = \frac{1}{p}\|\boldsymbol{v}\|_p^p - \lambda\left(\boldsymbol{e}'\boldsymbol{v} - d\right) \tag{19}$$

where λ is the Lagrangian multiplier, and the optimal $\boldsymbol{v}$ satisfies

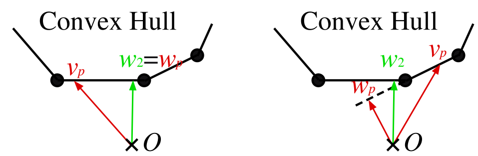
$$\frac{\partial L}{\partial v_i} = v_i^{p-1} - \lambda e_i = 0. \tag{20}$$

Hence, from (16),

$$\lambda = \frac{v_i^{p-1}}{e_i} = \frac{w_i}{e_i} \tag{21}$$

and $\hat{\boldsymbol{w}}$ is parallel to $\boldsymbol{e}$. This means that $\hat{\boldsymbol{w}}$ and $\boldsymbol{e}$ have the same output for any input due to linearity of dichotomy. It is worth emphasizing that the output by $\hat{\boldsymbol{w}}$ does not depend on p although $\hat{\boldsymbol{v}}$ varies according to p.

We show an example of $M = 2$ for clarity. In Fig. 1, $\boldsymbol{v}_p$ and $\boldsymbol{w}_p$ denote respectively the point nearest the origin in the p-norm and the estimated parameter induced by $\boldsymbol{v}_p$ and normalized to $\|\boldsymbol{w}_p\|_2 = 1$, that is, $\boldsymbol{w}_p = \hat{\boldsymbol{w}}/\|\hat{\boldsymbol{w}}\|_2$ in the p-norm. When $\boldsymbol{v}_2$ and $\boldsymbol{v}_p$ lie on the same hyperplane as case (a), $\boldsymbol{w}_2 = \boldsymbol{w}_p$ holds from (21) and the estimated parameter does not depend on p. However, $\boldsymbol{w}_p$ is in general different from $\boldsymbol{w}_2$ as case (b).



(a) (b)

Fig. 1. Geometrical View of $\hat{\boldsymbol{v}}$ and $\hat{\boldsymbol{w}}$.

4.2. Dependency of Estimator on the Norm

The nearest point may exist at an edge or a vertex of the convex hull of examples. Suppose the point $\hat{\boldsymbol{v}}$ in an affine space is written as the intersection of K hyperplanes

$$\boldsymbol{e}^{(k)\prime}\boldsymbol{v} = d^{(k)}(> 0), k = 1, \ldots, K(\leq M), \tag{22}$$

nearest the origin in the p-norm. Then $\hat{\boldsymbol{w}}$ is written as

$$\hat{\boldsymbol{w}} = \sum_{k=1}^{K} \lambda_k \boldsymbol{e}^{(k)} \tag{23}$$

from (16) where $\lambda_k \geq 0$ are the Lagrangian multipliers. That is, $\hat{\boldsymbol{w}}$ is located in the convex cone of $\boldsymbol{e}^{(k)}$, depending on p.

5. ANGLES OF NEAREST POINTS AND ESTIMATORS

To see how much $\boldsymbol{w}_2$ and $\boldsymbol{w}_p$ differ in general, we consider the cosine of the angle θ between $\boldsymbol{v}_2$ and $\boldsymbol{v}_p$,

$$\cos\theta = \frac{\boldsymbol{v}_2'\boldsymbol{v}_p}{\|\boldsymbol{v}_2\|_2\|\boldsymbol{v}_p\|_2}. \tag{24}$$

Then, we can prove the following theorem.

Theorem 1 *For an arbitrary $p \in (1, \infty)$, $\cos\theta$ has the lower bound*

$$\cos\theta \geq M^{-\left|\frac{p-2}{2p}\right|} \tag{25}$$

where M is the dimension of the input space.

The above theorem states that the angle between the nearest points in the L_2 and L_p norms has a lower bound $1/\sqrt{M}$ irrespective of N and p.

Proof Sketch Since $\boldsymbol{v}_2$ and $\boldsymbol{v}_p$ are respectively the nearests points in the convex hull in the L_2 and L_p norms,

$$\|\boldsymbol{v}_p\|_2 \geq \|\boldsymbol{v}_2\|_2 \tag{26}$$

$$\|\boldsymbol{v}_2\|_p \geq \|\boldsymbol{v}_p\|_p \tag{27}$$

hold from the definition and

$$(\boldsymbol{v}_p - \boldsymbol{v}_2)'\boldsymbol{v}_2 \geq 0 \tag{28}$$

holds from the convexity. Therefore, when $p < 2$, the denominator of (24) satisfies

$$\|\boldsymbol{v}_2\|_2\|\boldsymbol{v}_p\|_2 \le \|\boldsymbol{v}_2\|_2\|\boldsymbol{v}_p\|_p \le \|\boldsymbol{v}_2\|_2\|\boldsymbol{v}_2\|_p \tag{29}$$

from $\|\cdot\|_2 \le \|\cdot\|_p$ and (27). Hence $\cos\theta$ is bounded as

$$\cos\theta \ge \frac{\|\boldsymbol{v}_2\|_2}{\|\boldsymbol{v}_2\|_p} \ge M^{\frac{p-2}{2p}} \tag{30}$$

from (28) since $\|\boldsymbol{v}_2\|_p^p$ has the maximum $M^{\frac{2-p}{2}}$ subject to $\|\boldsymbol{v}_2\|_2 = 1$.

When $p > 2$, in a similar way, the denominator of (24) satisfies

$$\|\boldsymbol{v}_2\|_2\|\boldsymbol{v}_p\|_2 \le \|\boldsymbol{v}_2\|_2\|\boldsymbol{v}_2\|_p M^{\frac{p-2}{2p}} \le \|\boldsymbol{v}_2\|_2^2 M^{\frac{p-2}{2p}} \tag{31}$$

since $\|\boldsymbol{v}_p\|_2$ has the maximum $\|\boldsymbol{v}_2\|_p M^{\frac{p-2}{2}}$ subject to (27) and from $\|\cdot\|_p \le \|\cdot\|_2$. Hence $\cos\theta$ is bounded as

$$\cos\theta \ge M^{\frac{2-p}{2p}}. \qquad \square \tag{32}$$

Similarly, we can show that the cosine of the angle η between $\boldsymbol{w}_2$ and $\boldsymbol{w}_p$ has the same lower bound.

Theorem 2 *For an arbitrary $p \in (1, \infty)$, $\cos\eta$ has the lower bound*

$$\cos\eta = \frac{\boldsymbol{w}_2'\boldsymbol{w}_p}{\|\boldsymbol{w}_2\|_2\|\boldsymbol{w}_p\|_2} \ge M^{-\left|\frac{p-2}{2p}\right|} \tag{33}$$

where M is the dimension of the input space.

Proof Sketch Without loss of generality, we can assume that all the elements of $\boldsymbol{v}_p$ is non-negative. From (16), we consider the angle η between $\boldsymbol{v}_2$ and $\boldsymbol{v}_p^{p-1}$, which are respectively parallel to $\boldsymbol{w}_2$ and $\boldsymbol{w}_p$, where

$$\boldsymbol{v}_p^{p-1} = ((v_p)_1^{p-1}, (v_p)_2^{p-1}, \ldots, (v_p)_p^{p-1})'. \tag{34}$$

Here, we use the inequality

$$(\boldsymbol{v}_2 - \boldsymbol{v}_p)'\nabla L(\boldsymbol{v}_p) \ge 0 \tag{35}$$

from the convexity, instead of (28), where $\nabla L(\boldsymbol{v})$ is the gradient of $L(\boldsymbol{v}) = \sum_{i=1}^{M} v_i^p$ at the point $\boldsymbol{v}$ and is parallel to $\boldsymbol{v}_p^{p-1}$. The key inequality is $\|\cdot\|_{2p-2} \le \|\cdot\|_p$ when $p > 2$. Using this and others, we can show

$$\cos\eta = \frac{\boldsymbol{v}_2'\boldsymbol{v}_p^{p-1}}{\|\boldsymbol{v}_2\|_2\|\boldsymbol{v}_p^{p-1}\|_2} \ge \frac{\boldsymbol{v}_p'\boldsymbol{v}_p^{p-1}}{\|\boldsymbol{v}_p\|_2\|\boldsymbol{v}_p^{p-1}\|_2} \tag{36}$$

$$\ge \frac{\|\boldsymbol{v}_p\|_p}{\|\boldsymbol{v}_p\|_2}, \tag{37}$$

where (37) has a lower bound $M^{\frac{2-p}{2p}}$. The case of $p < 2$ is straightforward. $\square$

Note that the right-hand side of (36) is equal to the cosine of the angle between the nearest point $\boldsymbol{v}_p$ in the L_p norm and the corresponding estimator $\boldsymbol{w}_p$. Hence it also has the same lower bound.

6. CONCLUSIONS

In order to reduce the heavy computational load of SVMs, the idea of employing the 1- or ∞-norms instead of the 2-norm has been proposed in the literature. However, it was not clear how the change of norm affects the properties of SVMs. We gave a geometrical view for a generalized method adopting the p-norm for measuring the distance between examples and a separating hyperplane. More specifically, we showed the relationship of the estimator in the p-norm to the convex hull of examples and derived the bounds of angles of nearest points and estimators. This becomes possible by analyzing the ν-SVM instead of the original SVM. More rigorous analysis on the generalization ability is future work.

7. REFERENCES

[1] V. N. Vapnik, *The Nature of Statistical Learning Theory*, Springer-Verlag, 1995.

[2] B. Schölkopf et al., *Advances in Kernel Methods: Support Vector Learning*, Cambridge Univ. Press, 1998.

[3] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines*, Cambridge Univ. Press, 2000.

[4] K. Ikeda, "An asymptotic statistical theory of polynomial kernel methods," *Neural Comp.*, **16**, 1705–1719, 2004.

[5] F. Pérez-Cruz and O. Bousquet, "Kernel methods and their potential use in signal processing," *IEEE Signal Proc. Magazine*, **21**/3, 57–65, 2004.

[6] J. P. Pedroso and N. Murata, "Support vector machines with different norms: Motivation, formulations, and results," *Pattern Recog. Lett.*, **22**, 1263–1272, 2001.

[7] O. L. Mangasarian, "Arbitrary-norm separating plane," *Operations Research Letters*, **24**, 15–23, 1999.

[8] B. Schölkopf et al., "New support vector algorithms," *Neural Comp.*, **12**, 1207–1245, 2000.

[9] K. P. Bennett and E. J. Bredensteiner, "Duality and geometry in SVM classifiers," *Proc. ICML*, 57–64, 2000.

[10] D. Crisp and C. Burges, "A geometric interpretation of ν-SVM classifiers," In *NIPS*, **12**, 2000.

[11] K. Ikeda and T. Aoishi, "An asymptotic statistical analysis of support vector machines with soft margins," *Neural Networks*, in press.