ν-Support Vector Machine as Conditional Value-at-Risk Minimization

Akiko Takeda TAKEDA@AE.KEIO.AC.JP
Keio University, 3-14-1 Hiyoshi, Kouhoku, Yokohama, Kanagawa 223-8522, Japan

Masashi Sugiyama SUGI@CS.TITECH.AC.JP
Tokyo Institute of Technology, 2-12-1 O-okayama, Meguro-ku, Tokyo 152-8552, Japan

Abstract

The ν-support vector classification (ν-SVC) algorithm was shown to work well and provide intuitive interpretations, e.g., the parameter ν roughly specifies the fraction of support vectors. Although ν corresponds to a fraction, it cannot take the entire range between 0 and 1 in its original form. This problem was settled by a non-convex extension of ν-SVC and the extended method was experimentally shown to generalize better than original ν-SVC. However, its good generalization performance and convergence properties of the optimization algorithm have not been studied yet. In this paper, we provide new theoretical insights into these issues and propose a novel ν-SVC algorithm that has guaranteed generalization performance and convergence properties.

1. Introduction

Support vector classification (SVC) is one of the most successful classification algorithms in modern machine learning (Schölkopf & Smola, 2002). SVC finds a hyperplane that separates training samples in different classes with maximum margin (Boser et al., 1992). The maximum margin hyperplane was shown to minimize an upper bound of the generalization error according to the Vapnik-Chervonenkis theory (Vapnik, 1995). Thus the generalization performance of SVC is theoretically guaranteed.

SVC was extended to be able to deal with non-separable data by trading the margin size with the data separation error (Cortes & Vapnik, 1995). This soft-margin formulation is commonly referred to as C-SVC since the trade-off is controlled by the parameter C. C-SVC was shown to work very well in a wide range of real-world applications (Schölkopf & Smola, 2002).

An alternative formulation of the soft-margin idea is ν-SVC (Schölkopf et al., 2000)—instead of the parameter C, ν-SVC involves another trade-off parameter ν that roughly specifies the fraction of support vectors (or sparseness of the solution). Thus, the ν-SVC formulation provides us richer interpretation than the original C-SVC formulation, which would be potentially useful in real applications.

Since the parameter ν corresponds to a fraction, it should be able to be chosen between 0 and 1. However, it was shown that admissible values of ν are actually limited (Crisp & Burges, 2000; Chang & Lin, 2001). To cope with this problem, Perez-Cruz et al. (2003) introduced the notion of negative margins and proposed extended ν-SVC (Eν-SVC) which allows ν to take the entire range between 0 and 1. They also experimentally showed that the generalization performance of Eν-SVC is often better than that of original ν-SVC. Thus the extension contributes not only to elucidating the theoretical property of ν-SVC, but also to improving its generalization performance.

However, there remain two open issues in Eν-SVC. The first issue is that the reason why a high generalization performance can be obtained by Eν-SVC was not completely explained yet. The second issue is that the optimization problem involved in Eν-SVC is non-convex and theoretical convergence properties of the Eν-SVC optimization algorithm have not been studied yet. The purpose of this paper is to provide new theoretical insights into these two issues.

After reviewing existing SVC methods in Section 2, we

Appearing in *Proceedings of the 25th International Conference on Machine Learning*, Helsinki, Finland, 2008.

elucidate the generalization performance of Eν-SVC in Section 3. We first show that the Eν-SVC formulation could be interpreted as minimization of the *conditional value-at-risk* (CVaR), which is often used in finance (Rockafellar & Uryasev, 2002; Gotoh & Takeda, 2005). Then we give new generalization error bounds based on the CVaR risk measure. This theoretical result justifies the use of Eν-SVC.

In Section 4, we address non-convexity of the Eν-SVC optimization problem. We first give a new optimization algorithm that is guaranteed to converge to one of the local optima within a finite number of iterations. Based on this improved algorithm, we further show that the global solution can be actually obtained within finite iterations even though the optimization problem is non-convex.

Finally, in Section 5, we give concluding remarks and future prospects. Proofs of all theorems and lemmas are sketched in Appendix unless mentioned.

2. Support Vector Classification

In this section, we formulate the classification problem and briefly review support vector algorithms.

2.1. Classification Problem

Let us address the classification problem of learning a decision function h from $\mathcal{X}$ ($\subset \mathbb{R}^n$) to $\{\pm 1\}$ based on training samples $(\boldsymbol{x}_i, y_i)$ ($i \in M := \{1, ..., m\}$). We assume that the training samples are i.i.d. following the unknown probability distribution $P(\boldsymbol{x}, y)$ on $\mathcal{X} \times \{\pm 1\}$.

The goal of the classification task is to obtain a classifier h that minimizes the generalization error (or the risk):

$$R[h] := \int \frac{1}{2}|h(\boldsymbol{x}) - y| dP(\boldsymbol{x}, y),$$

which corresponds to the misclassification rate for unseen test samples.

For the sake of simplicity, we generally focus on linear classifiers, i.e.,

$$h(\boldsymbol{x}) = \mathrm{sign}(\langle \boldsymbol{w}, \boldsymbol{x} \rangle + b), \tag{1}$$

where $\boldsymbol{w}$ ($\in \mathbb{R}^n$) is a non-zero normal vector, b ($\in \mathbb{R}$) is a bias parameter, and $\mathrm{sign}(\xi) = 1$ if $\xi \geq 0$ and -1 otherwise.

Most of the discussions in this paper can be directly applicable to non-linear kernel classifiers (Schölkopf & Smola, 2002). Thus we may not lose generality by restricting ourselves to linear classifiers.

2.2. Support Vector Classification

The Vapnik-Chervonenkis theory (Vapnik, 1995) showed that a large margin classifier has a small generalization error. Motivated by this theoretical result, Boser et al. (1992) developed an algorithm for finding the hyperplane $(\boldsymbol{w}, b)$ with maximum margin:

$$\min_{\boldsymbol{w}, b} \frac{1}{2}\|\boldsymbol{w}\|^2 \quad \text{s.t. } y_i(\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle + b) \geq 1, \quad i \in M. \tag{2}$$

This is called (hard-margin) support vector classification (SVC) and valid when the training samples are linearly separable. In the following, we omit "$i \in M$" in the constraint for brevity.

2.3. C-Support Vector Classification

Cortes and Vapnik (1995) extended the SVC algorithm to non-separable cases and proposed trading the margin size with the data separation error (i.e., "soft-margin"):

$$\min_{\boldsymbol{w}, b, \boldsymbol{\xi}} \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{m} \xi_i$$
$$\text{s.t. } y_i(\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle + b) \geq 1 - \xi_i, \quad \xi_i \geq 0,$$

where C (> 0) controls the trade-off. This formulation is usually referred to as C-SVC, and was shown to work very well in various real-world applications (Schölkopf & Smola, 2002).

2.4. ν-Support Vector Classification

ν-SVC is another formulation of soft-margin SVC (Schölkopf et al., 2000):

$$\min_{\boldsymbol{w}, b, \boldsymbol{\xi}, \rho} \frac{1}{2}\|\boldsymbol{w}\|^2 - \nu\rho + \frac{1}{m}\sum_{i=1}^{m} \xi_i$$
$$\text{s.t. } y_i(\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle + b) \geq \rho - \xi_i, \quad \xi_i \geq 0, \quad \rho \geq 0,$$

where ν ($\in \mathbb{R}$) is the trade-off parameter.

Schölkopf et al. (2000) showed that if the ν-SVC solution yields $\rho > 0$, C-SVC with $C = 1/(m\rho)$ produces the same solution. Thus ν-SVC and C-SVC are equivalent. However, ν-SVC has additional intuitive interpretations, e.g., ν is an upper bound on the fraction of margin errors and a lower bound on the fraction of support vectors (i.e., sparseness of the solution). Thus, the ν-SVC formulation would be potentially more useful than the C-SVC formulation in real applications.

2.5. Eν-SVC

Although ν has an interpretation as a fraction, it cannot always take its full range between 0 and 1 (Crisp & Burges, 2000; Chang & Lin, 2001).

2.5.1. Admissible Range of ν

For an optimal solution $\{\alpha_i^C\}_{i=1}^m$ of dual C-SVC, let

$$\zeta(C) := \frac{1}{Cm}\sum_{i=1}^m \alpha_i^C,$$
$$\nu_{\min} := \lim_{C\to\infty} \zeta(C) \quad \text{and} \quad \nu_{\max} := \lim_{C\to 0} \zeta(C).$$

Then, Chang and Lin (2001) showed that for $\nu \in (\nu_{\min}, \nu_{\max}]$, the optimal solution set of ν-SVC is the same as that of C-SVC with some C (not necessarily unique). In addition, the optimal objective value of ν-SVC is strictly negative. However, for $\nu \in (\nu_{\max}, 1]$, ν-SVC is unbounded, i.e., there exists no solution; for $\nu \in [0, \nu_{\min}]$, ν-SVC is feasible with zero optimal objective value, i.e., we end up with just having a trivial solution ($\boldsymbol{w} = \boldsymbol{0}$ and $b = 0$).

2.5.2. Increasing Upper Admissible Range

It was shown by Crisp and Burges (2000) that

$$\nu_{\max} = 2\min(m_+, m_-)/m,$$

where m_+ and m_- are the number of positive and negative training samples. Thus, when the training samples are balanced (i.e., $m_+ = m_-$), $\nu_{\max} = 1$ and therefore ν can reach its upper limit 1. When the training samples are imbalanced (i.e., $m_+ \neq m_-$), Perez-Cruz et al. (2003) proposed modifying the optimization problem of ν-SVC as

$$\min_{\boldsymbol{w},b,\boldsymbol{\xi},\rho} \frac{1}{2}\|\boldsymbol{w}\|^2 - \nu\rho + \frac{1}{m_+}\sum_{i:y_i=1}\xi_i + \frac{1}{m_-}\sum_{i:y_i=-1}\xi_i$$
$$\text{s.t. } y_i(\langle \boldsymbol{w}, \boldsymbol{x}_i\rangle + b) \geq \rho - \xi_i, \quad \xi_i \geq 0, \quad \rho \geq 0,$$

i.e., the effect of positive and negative samples are balanced. Under this modified formulation, $\nu_{\max} = 1$ holds even when training samples are imbalanced.

For the sake of simplicity, we assume $m_+ = m_-$ in the rest of this paper; when $m_+ \neq m_-$, all the results can be simply extended in a similar way as above.

2.5.3. Decreasing Lower Admissible Range

When $\nu \in [0, \nu_{\min}]$, ν-SVC produces a trivial solution ($\boldsymbol{w} = \boldsymbol{0}$ and $b = 0$) as shown in Chang and Lin (2001). To prevent this, Perez-Cruz et al. (2003) proposed allowing the margin ρ to be negative and enforcing the norm of $\boldsymbol{w}$ to be unity:

$$\min_{\boldsymbol{w},b,\boldsymbol{\xi},\rho} -\nu\rho + \frac{1}{m}\sum_{i=1}^m \xi_i$$
$$\text{s.t. } y_i(\langle \boldsymbol{w}, \boldsymbol{x}_i\rangle + b) \geq \rho - \xi_i,\ \xi_i \geq 0,\ \|\boldsymbol{w}\|^2 = 1. \quad (3)$$

By this modification, a non-trivial solution can be obtained even for $\nu \in [0, \nu_{\min}]$. This modified formulation is called extended ν-SVC (Eν-SVC).

The Eν-SVC optimization problem is non-convex due to the equality constraint $\|\boldsymbol{w}\|^2 = 1$. Perez-Cruz et al. (2003) proposed the following iterative algorithm for computing a solution. First, for some initial $\widetilde{\boldsymbol{w}}$, solve the problem (3) with $\|\boldsymbol{w}\|^2 = 1$ replaced by $\langle \widetilde{\boldsymbol{w}}, \boldsymbol{w}\rangle = 1$. Then, using the optimal solution $\widehat{\boldsymbol{w}}$, update $\widetilde{\boldsymbol{w}}$ by

$$\widetilde{\boldsymbol{w}} \longleftarrow \gamma\widetilde{\boldsymbol{w}} + (1-\gamma)\widehat{\boldsymbol{w}} \quad (4)$$

for $\gamma = 9/10$, and iterate this procedure until convergence.

Perez-Cruz et al. (2003) experimentally showed that the generalization performance of Eν-SVC with $\nu \in [0, \nu_{\min}]$ is often better than that with $\nu \in (\nu_{\min}, \nu_{\max}]$, implying that E$\nu$-SVC is a promising classification algorithm. However, it is not clear how the notion of negative margins influences on the generalization performance and how fast the above iterative algorithm converges. The goal of this paper is to give new theoretical insights into these issues.

3. Justification of the Eν-SVC Criterion

In this section, we give a new interpretation of Eν-SVC and theoretically explain why it works well.

3.1. New Interpretation of Eν-SVC as CVaR minimization

Let $f(\boldsymbol{w}, b; \boldsymbol{x}, y)$ be the *margin error* for a sample $(\boldsymbol{x}, y)$:

$$f(\boldsymbol{w}, b; \boldsymbol{x}, y) := -\frac{y(\langle \boldsymbol{w}, \boldsymbol{x}\rangle + b)}{\|\boldsymbol{w}\|}.$$

Let us consider the distribution of margin errors over all training samples:

$$\Phi(\alpha|\boldsymbol{w}, b) := P\{(\boldsymbol{x}_i, y_i) \mid f(\boldsymbol{w}, b; \boldsymbol{x}_i, y_i) \leq \alpha\}.$$

For $\beta \in [0, 1)$, let $\alpha_\beta(\boldsymbol{w}, b)$ be the 100β-percentile of the margin error distribution:

$$\alpha_\beta(\boldsymbol{w}, b) := \min\{\alpha \mid \Phi(\alpha|\boldsymbol{w}, b) \geq \beta\}.$$

Thus only the fraction $(1-\beta)$ of the margin error $f(\boldsymbol{w}, b; \boldsymbol{x}_i, y_i)$ exceeds the threshold $\alpha_\beta(\boldsymbol{w}, b)$ (see Figure 1). $\alpha_\beta(\boldsymbol{w}, b)$ is commonly referred to as the *value-at-risk* (VaR) in finance and is often used by security houses or investment banks to measure the market risk of their asset portfolios (Rockafellar & Uryasev, 2002; Gotoh & Takeda, 2005).

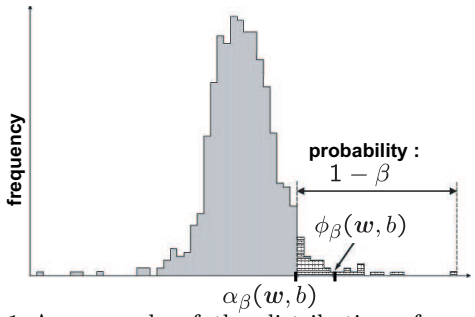

Figure 1. An example of the distribution of margin errors $f(\boldsymbol{w},b;\boldsymbol{x}_i,y_i)$ over all training samples. $\alpha_\beta(\boldsymbol{w},b)$ is the 100β-percentile called the value-at-risk (VaR), and the mean $\phi_\beta(\boldsymbol{w},b)$ of the β-tail distribution is called the conditional VaR (CVaR).

Let us consider the β-tail distribution of $f(\boldsymbol{w},b;\boldsymbol{x}_i,y_i)$:

$$\Phi_\beta(\alpha|\boldsymbol{w},b) := \begin{cases} 0 & \text{for } \alpha < \alpha_\beta(\boldsymbol{w},b), \\ \frac{\Phi(\alpha|\boldsymbol{w},b)-\beta}{1-\beta} & \text{for } \alpha \geq \alpha_\beta(\boldsymbol{w},b). \end{cases}$$

Let $\phi_\beta(\boldsymbol{w},b)$ be the mean of the β-tail distribution of $f(\boldsymbol{w},b;\boldsymbol{x}_i,y_i)$ (see Figure 1 again):

$$\phi_\beta(\boldsymbol{w},b) := \boldsymbol{E}_{\Phi_\beta}[f(\boldsymbol{w},b;\boldsymbol{x}_i,y_i)],$$

where $\boldsymbol{E}_{\Phi_\beta}$ denotes the expectation over the distribution Φ_β. $\phi_\beta(\boldsymbol{w},b)$ is called the *conditional VaR* (CVaR). By definition, the CVaR is always larger than or equal to the VaR:

$$\phi_\beta(\boldsymbol{w},b) \geq \alpha_\beta(\boldsymbol{w},b). \tag{5}$$

Let us consider the problem of minimizing the CVaR $\phi_\beta(\boldsymbol{w},b)$ (which we refer to as minCVaR):

$$\min_{\boldsymbol{w},b} \phi_\beta(\boldsymbol{w},b). \tag{6}$$

Then we have the following theorem.

Theorem 1 *The solution of the minCVaR problem* (6) *is equivalent to the solution of the Eν-SVC problem* (3) *with*

$$\nu = 1-\beta.$$

Theorem 1 shows that Eν-SVC actually minimizes the CVaR $\phi_{1-\nu}(\boldsymbol{w},b)$. Thus, E$\nu$-SVC could be interpreted as minimizing the mean margin error over a set of "bad" training samples. In contrast, the hard-margin SVC problem (2) can be equivalently expressed in terms of the margin error as

$$\min_{\boldsymbol{w},b} \max_{i\in M} f(\boldsymbol{w},b;\boldsymbol{x}_i,y_i).$$

Thus hard-margin SVC minimizes the margin error of the single "worst" training sample. This analysis shows that Eν-SVC can be regarded as an extension of hard-margin SVC to be less sensitive to an outlier (i.e., the single "worst" training sample).

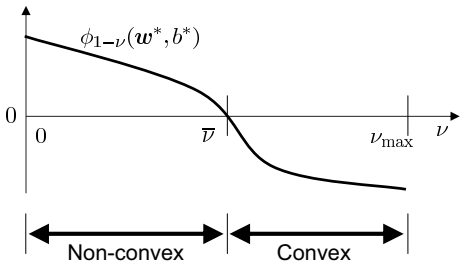

Figure 2. A profile of the CVaR $\phi_{1-\nu}(\boldsymbol{w}^*,b^*)$ as a function of ν. As shown in Section 4, the Eν-SVC optimization problem can be cast as a convex problem if $\nu \in (\overline{\nu},\nu_{\max}]$, while it is essentially non-convex if $\nu \in (0,\overline{\nu})$.

3.2. Justification of Eν-SVC

We have shown the equivalence between Eν-SVC and minCVaR. Here we derive new bounds of the generalization error based on the notion of CVaR and try to justify the use of Eν-SVC.

When training samples are linearly separable, the margin error $f(\boldsymbol{w},b;\boldsymbol{x}_i,y_i)$ is negative for all samples. Then, at the optimal solution $(\boldsymbol{w}^*,b^*)$, the CVaR $\phi_{1-\nu}(\boldsymbol{w}^*,b^*)$ is always negative. However, in non-separable cases, $\phi_{1-\nu}(\boldsymbol{w}^*,b^*)$ could be positive particularly when ν is close to 0. Regarding the CVaR, we have the following lemma.

Lemma 2 *$\phi_{1-\nu}(\boldsymbol{w}^*,b^*)$ is continuous with respect to ν and is strictly decreasing when ν is increased.*

Let $\overline{\nu}$ be such that

$$\phi_{1-\overline{\nu}}(\boldsymbol{w}^*,b^*) = 0$$

if such $\overline{\nu}$ exists; we set $\overline{\nu} = \nu_{\max}$ if $\phi_{1-\nu}(\boldsymbol{w}^*,b^*) > 0$ for all ν and we set $\overline{\nu} = 0$ if $\phi_{1-\nu}(\boldsymbol{w}^*,b^*) < 0$ for all ν. Then we have the following relation (see Figure 2):

$$\phi_{1-\nu}(\boldsymbol{w}^*,b^*) < 0 \text{ for } \nu \in (\overline{\nu},\nu_{\max}],$$
$$\phi_{1-\nu}(\boldsymbol{w}^*,b^*) > 0 \text{ for } \nu \in (0,\overline{\nu}).$$

Below, we analyze the generalization error of Eν-SVC depending on the value of ν.

3.2.1. Justification When $\nu \in (\overline{\nu},\nu_{\max}]$

Theorem 3 *Let $\nu \in (\overline{\nu},\nu_{\max}]$. Suppose that support $\mathcal{X}$ is in a ball of radius R around the origin. Then, for all $(\boldsymbol{w},b)$ such that $\|\boldsymbol{w}\| = 1$ and $\phi_{1-\nu}(\boldsymbol{w},b) < 0$, there exists a positive constant c such that the following bound hold with probability at least $1-\delta$:*

$$R[h] \leq \nu + G(\alpha_{1-\nu}(\boldsymbol{w},b)), \tag{7}$$

where

$$G(\gamma) = \sqrt{\frac{2}{m}\left(\frac{4c^2(R^2+1)^2}{\gamma^2}\log_2(2m) - 1 + \log\frac{2}{\delta}\right)}.$$

The generalization error bound in (7) is furthermore upper-bounded as

$$\nu + G(\alpha_{1-\nu}(\boldsymbol{w}, b)) \leq \nu + G(\phi_{1-\nu}(\boldsymbol{w}, b)).$$

$G(\gamma)$ is monotone decreasing as $|\gamma|$ increases. Thus, the above theorem shows that when $\phi_{1-\nu}(\boldsymbol{w}, b) < 0$, the upper bound $\nu + G(\phi_{1-\nu}(\boldsymbol{w}, b))$ is lowered if the CVaR $\phi_{1-\nu}(\boldsymbol{w}, b)$ is reduced. Since Eν-SVC minimizes $\phi_{1-\nu}(\boldsymbol{w}, b)$ (see Theorem 1), the upper bound of the generalization error is also minimized.

3.2.2. Justification When $\nu \in (0, \overline{\nu}]$

Our discussion below depends on the sign of $\alpha_{1-\nu}(\boldsymbol{w}, b)$. When $\alpha_{1-\nu}(\boldsymbol{w}, b) < 0$, we have the following theorem.

Theorem 4 *Let $\nu \in (0, \overline{\nu}]$. Then, for all $(\boldsymbol{w}, b)$ such that $\|\boldsymbol{w}\| = 1$ and $\alpha_{1-\nu}(\boldsymbol{w}, b) < 0$, there exists a positive constant c such that the following bound holds with probability at least $1 - \delta$:*

$$R[h] \leq \nu + G(\alpha_{1-\nu}(\boldsymbol{w}, b)).$$

A proof of the above theorem is omitted since the proof follows a similar line to the proof of Theorem 3. This theorem shows that when $\alpha_{1-\nu}(\boldsymbol{w}, b) < 0$, the upper bound $\nu + G(\alpha_{1-\nu}(\boldsymbol{w}, b))$ is lowered if $\alpha_{1-\nu}(\boldsymbol{w}, b)$ is reduced. On the other hand, Eq.(5) shows that the VaR $\alpha_{1-\nu}(\boldsymbol{w}, b)$ is upper-bounded by the CVaR $\phi_{1-\nu}(\boldsymbol{w}, b)$. Therefore, minimizing $\phi_{1-\nu}(\boldsymbol{w}, b)$ by Eν-SVC may have an effect of lowering the upper bound of the generalization error.

When $\alpha_{1-\nu}(\boldsymbol{w}, b) > 0$, we have the following theorem.

Theorem 5 *Let $\nu \in (0, \overline{\nu}]$. Then, for all $(\boldsymbol{w}, b)$ such that $\|\boldsymbol{w}\| = 1$ and $\alpha_{1-\nu}(\boldsymbol{w}, b) > 0$, there exists a positive constant c such that the following bound hold with probability at least $1 - \delta$:*

$$R[h] \geq \nu - G(\alpha_{1-\nu}(\boldsymbol{w}, b)).$$

Moreover, the lower bound of $R[h]$ is bounded from above as

$$\nu - G(\alpha_{1-\nu}(\boldsymbol{w}, b)) \leq \nu - G(\phi_{1-\nu}(\boldsymbol{w}, b)).$$

A proof of the above theorem is also omitted since the proof resembles to Theorem 3. Theorem 5 implies that the lower bound $\nu - G(\alpha_{1-\nu}(\boldsymbol{w}, b))$ of the generalization error is upper-bounded by $\nu - G(\phi_{1-\nu}(\boldsymbol{w}, b))$. On the other hand, Eq.(5) and $\alpha_{1-\nu}(\boldsymbol{w}, b) > 0$ yields $\phi_{1-\nu}(\boldsymbol{w}, b) > 0$. Thus minimizing $\phi_{1-\nu}(\boldsymbol{w}, b)$ by Eν-SVC may contribute to lowering the lower bound $\nu - G(\alpha_{1-\nu}(\boldsymbol{w}, b))$ of the generalization error.

4. New Optimization Algorithm

As reviewed in Section 2.5, Eν-SVC involves a non-convex optimization problem. In this section, we give a new efficient optimization procedure for Eν-SVC. Our proposed procedure involves two optimization algorithms depending on the value of ν. We first describe the two algorithms and then show how these two algorithms are chosen for practical use.

4.1. Optimization When $\nu \in (\overline{\nu}, \nu_{\max}]$

Lemma 6 *When $\nu \in (\overline{\nu}, \nu_{\max}]$, the E$\nu$-SVC problem (3) is equivalent to*

$$\begin{aligned} &\min_{\boldsymbol{w}, b, \boldsymbol{\xi}, \rho} \; -\nu\rho + \frac{1}{m}\sum_{i=1}^{m} \xi_i \\ &\text{s.t. } y_i(\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle + b) \geq \rho - \xi_i, \; \xi_i \geq 0, \; \|\boldsymbol{w}\|^2 \leq 1. \end{aligned} \quad (8)$$

This lemma shows that the equality constraint $\|\boldsymbol{w}\|^2 = 1$ in the original problem (3) can be replaced by $\|\boldsymbol{w}\|^2 \leq 1$ without changing the solution. Due to the convexity of $\|\boldsymbol{w}\|^2 \leq 1$, the above optimization problem is convex and therefore we can easily obtain the global solution by a standard optimization software.

4.2. Optimization When $\nu \in (0, \overline{\nu}]$

If $\nu \in (0, \overline{\nu}]$, the E$\nu$-SVC optimization problem is essentially non-convex and therefore we need a more elaborate algorithm.

4.2.1. Local Optimum Search

Here, we propose the following iterative algorithm for finding a local optimum.

Algorithm 7 (The Eν-SVC local optimum search algorithm for $\nu \in (0, \overline{\nu}]$)

Step 1: Initialize $\widetilde{\boldsymbol{w}}$.
Step 2: Solve the following linear program:

$$\begin{aligned} &\min_{\boldsymbol{w}, b, \boldsymbol{\xi}, \rho} \; -\nu\rho + \frac{1}{m}\sum_{i=1}^{m} \xi_i \\ &\text{s.t. } y_i(\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle + b) \geq \rho - \xi_i, \; \xi_i \geq 0, \; \langle \widetilde{\boldsymbol{w}}, \boldsymbol{w} \rangle = 1, \end{aligned} \quad (9)$$

and let the optimal solution be $(\widehat{\boldsymbol{w}}, \widehat{b}, \widehat{\boldsymbol{\xi}}, \widehat{\rho})$.
Step 3: If $\widetilde{\boldsymbol{w}} = \widehat{\boldsymbol{w}}$, terminate and output $\widetilde{\boldsymbol{w}}$. Otherwise, update $\widetilde{\boldsymbol{w}}$ by $\widetilde{\boldsymbol{w}} \longleftarrow \widehat{\boldsymbol{w}}/\|\widehat{\boldsymbol{w}}\|$.
Step 4: Repeat Steps 2–3.

The linear program (9) is the same as the one proposed by Perez-Cruz et al. (2003), i.e., the equality constrained $\|\boldsymbol{w}\|^2 = 1$ of the original problem (3) is

replaced by $\langle \widetilde{\boldsymbol{w}}, \boldsymbol{w} \rangle = 1$. The updating rule of $\widetilde{\boldsymbol{w}}$ in Step 3 is different from the one proposed by Perez-Cruz et al. (2003) (cf. Eq.(4)).

We define a "corner" (or "0-dimensional face") of Eν-SVC (3) as the intersection of an edge of the polyhedral cone formed by linear constraints of (3) and $\|\boldsymbol{w}\|^2 = 1$. Under the new update rule, the algorithm visits a corner of Eν-SVC (3) in each iteration. Since Eν-SVC has finite corners, we can show that Algorithm 7 with the new update rule terminates in a finite number of iterations, i.e., less than or equal to the number of corners of Eν-SVC.

Theorem 8 *Algorithm 7 terminates within a finite number of iterations of Steps 2–3. Furthermore, a solution of the modified Eν-SVC algorithm is a local minimizer if it is unique and non-degenerate.*

4.2.2. Global Optimum Search

Next, we show that the global solution can be actually obtained within finite iterations, despite the non-convexity of the optimization problem.

A naive approach to searching for the global solution is to run the local optimum search algorithm many times with different initial values and choose the best local solution. However, there is no guarantee that this naive approach can find the global solution. Below, we give a more systematic way to find the global solution based on the following lemma.

Lemma 9 *When $\nu \in (0, \overline{\nu}]$, the E$\nu$-SVC problem* (3) *is equivalent to*

$$\min_{\boldsymbol{w}, b, \boldsymbol{\xi}, \rho} \; -\nu\rho + \frac{1}{m}\sum_{i=1}^{m} \xi_i$$
$$\text{s.t. } y_i(\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle + b) \geq \rho - \xi_i,\ \xi_i \geq 0,\ \|\boldsymbol{w}\|^2 \geq 1. \quad (10)$$

Lemma 9 could be proved in a similar way as Lemma 6, so we omit the proof. This lemma shows that the equality constraint $\|\boldsymbol{w}\|^2 = 1$ in the original Eν-SVC problem (3) can be replaced by $\|\boldsymbol{w}\|^2 \geq 1$ without changing the solution if $\nu \in (0, \overline{\nu}]$.

The problem (10) is called a *linear reverse convex program* (LRCP), which is a class of non-convex problems consisting of linear constraints and one concave inequality ($\|\boldsymbol{w}\|^2 \geq 1$ in the current case). The feasible set of the problem (10) consists of a finite number of *faces*. For LRCPs, Horst and Tuy (1995) showed that the local optimal solutions correspond to 0-dimensional faces (or corners). This implies that all the local optimal solutions of the Eν-SVC problem (10) can be traced by checking all the faces.

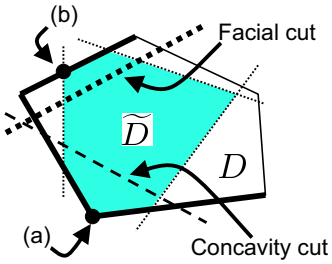

Figure 3. A 0-dimensional face (a) and three proper faces (bold solid lines) of D are identified in $\widetilde{D}$. If the corner (a) is found in Step 2, a concavity cut is constructed. If the corner (b) is found, a facial cut is constructed. If these two cuts are added to $\widetilde{D}$, the remaining area includes no face of D.

Let D be the feasible set of Eν-SVC (3). Below, we summarize the Eν-SVC training algorithm based on the *cutting plane method*, which is an efficient method of tracing faces.

Algorithm 10 (The Eν-SVC global optimum search algorithm for $\nu \in (0, \overline{\nu}]$)

Step 1: $\widetilde{D} \longleftarrow D$.
Step 2: Find a local solution by Algorithm 7.
Step 3: Identify a face of D in $\widetilde{D}$ that corresponds the local solution.
Step 4a: If the face is a corner, construct a "concavity cut".
Step 4b: If the face is a proper face, construct a "facial cut".
Step 5: Add the cut to the problem (9) and $\widetilde{D}$.
Step 6: Repeat Steps 2–5 until $\widetilde{D}$ includes no face of D.
Step 7: Output the best local optimal solution as the global solution.

If the local solution obtained in Step 2 is a corner of D (i.e., the local solution is not on any cutting plane as (a) in Figure 3), a *concavity cut* (Horst & Tuy, 1995) is constructed. The concavity cut has a role of removing the local solution, i.e., a 0-dimensional face of D and its neighborhood. Otherwise, a *facial cut* (Majthay & Whinston, 1974) is constructed to eliminate the proper face (see (b) in Figure 3).

Since the total number of distinct faces of D is finite in the current setting and a facial cut or a concavity cut eliminates at least one face at a time, Algorithm 10 is guaranteed to terminate within finite iterations (precisely, less than or equal to the number of all dimensional faces of Eν-SVC). Furthermore, since the addition of a concavity cut or a facial cut does not remove local solutions which are better than the best local solution found so far, Algorithm 10 is guaranteed to

trace all *sufficient* local solutions. Thus we can always find a global solution within finite iterations by Algorithm 10. A more detailed discussion on the concavity cut and the facial cut is shown in Horst and Tuy (1995) and Majthay and Whinston (1974), respectively.

4.3. Choice of Two Algorithms

We have two convergent algorithms when $\nu \in (\overline{\nu}, \nu_{\max}]$ and $\nu \in (0, \overline{\nu}]$. Thus, choosing a suitable algorithm depending on the value of ν would be an ideal procedure. However, the value of the threshold $\overline{\nu}$ is difficult to explicitly compute since it is defined via the optimal value $\phi_{1-\overline{\nu}}(\boldsymbol{w}^*, b^*)$ (see Figure 2). Therefore, it is not straightforward to choose a suitable algorithm for a given ν.

When we use Eν-SVC in practice, we usually compute the solutions for several different values of ν and choose the most promising one based on, e.g., cross-validation. In such scenarios, we can properly switch two algorithms without explicitly knowing the value of $\overline{\nu}$—our key idea is that the solution of the problem (8) is non-trivial (i.e., $\boldsymbol{w} \neq \boldsymbol{0}$) if and only if $\nu \in (\overline{\nu}, \nu_{\max}]$. Thus if the solutions are computed from large ν to small ν, the switching point can be identified by checking the triviality of the solution. The proposed algorithm is summarized as follows.

Algorithm 11 (The Eν-SVC algorithm for ($\nu_{\max} \geq$) $\nu_1 > \nu_2 > \cdots > \nu_k > 0$)

Step 1: $i \longleftarrow 1$.
Step 2: Compute $(\boldsymbol{w}^*, b^*)$ for ν_i by solving (8).
Step 3a: If $\boldsymbol{w}^* \neq \boldsymbol{0}$, accept $(\boldsymbol{w}^*, b^*)$ as the solution for ν_i, increment i, and go to Step 2.
Step 3b: If $\boldsymbol{w}^* = \boldsymbol{0}$, reject $(\boldsymbol{w}^*, b^*)$.
Step 4: Compute $(\boldsymbol{w}^*, b^*)$ for ν_i by Algorithm 10.
Step 5: Accept $(\boldsymbol{w}^*, b^*)$ as the solution for ν_i, increment i, and go to Step 4 unless $i > k$.

5. Conclusions

We characterized the generalization error of Eν-SVC in terms of the *conditional value-at-risk* (CVaR, see Figure 1) and showed that a good generalization performance is expected by Eν-SVC. We then derived a globally convergent optimization algorithm even though the optimization problem involved in Eν-SVC is non-convex.

We introduced the threshold $\overline{\nu}$ based on the sign of the CVaR (see Figure 2). We can check that the problem (8) is equivalent to ν-SVC in the sense that they share the same negative optimal value in $(\overline{\nu}, \nu_{\max}]$ and $(\nu_{\min}, \nu_{\max}]$, respectively (Gotoh & Takeda, 2005). On the other hand, the problem (8) and ν-SVC have the zero optimal value in $(0, \overline{\nu}]$ and $[0, \nu_{\min}]$, respectively. Thus, although the definitions of $\overline{\nu}$ and $\nu_{\min}$ are different, they would be essentially the same. We will study the relation between $\overline{\nu}$ and $\nu_{\min}$ in more detail in the future work.

References

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. *COLT* (pp. 144–152). ACM Press.

Chang, C.-C., & Lin, C.-J. (2001). Training ν-support vector classifiers: Theory and algorithms. *Neural Computation*, *13*, 2119–2147.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*, 273–297.

Crisp, D. J., & Burges, C. J. C. (2000). A geometric interpretation of ν-SVM classifiers. *NIPS 12* (pp. 244–250). MIT Press.

Gotoh, J., & Takeda, A. (2005). A linear classification model based on conditional geometric score. *Pacific Journal of Optimization*, *1*, 277–296.

Horst, R., & Tuy, H. (1995). *Global optimization: Deterministic approaches*. Berlin: Springer-Verlag.

Majthay, A., & Whinston, A. (1974). Quasi-concave minimization subject to linear constraints. *Discrete Mathematics*, *9*, 35–59.

Perez-Cruz, F., Weston, J., Hermann, D. J. L., & Schölkopf, B. (2003). Extension of the ν-SVM range for classification. *Advances in Learning Theory: Methods, Models and Applications 190* (pp. 179–196). Amsterdam: IOS Press.

Rockafellar, R. T., & Uryasev, S. (2002). Conditional value-at-risk for general loss distributions. *Journal of Banking & Finance*, *26*, 1443–1472.

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

Schölkopf, B., & Smola, A. J. (2002). *Learning with kernels*. Cambridge, MA: MIT Press.

Vapnik, V. N. (1995). *The nature of statistical learning theory*. Berlin: Springer-Verlag.

A. Sketch of Proof of Theorem 1

Let $(\boldsymbol{w}^*, b^*, \alpha^*)$ be the optimal solution of

$$\min_{\boldsymbol{w}, b, \alpha} F_\beta(\boldsymbol{w}, b, \alpha), \tag{11}$$

where, for $[X]^+ := \max\{X, 0\}$,

$$F_\beta(\boldsymbol{w}, b, \alpha) := \alpha + \frac{\sum_{i \in M} [f(\boldsymbol{w}, b; \boldsymbol{x}_i, y_i) - \alpha]^+}{(1-\beta)m}. \tag{12}$$

Then Rockafellar and Uryasev (2002) showed that

$$F_\beta(\boldsymbol{w}^*, b^*, \alpha^*) = \phi_\beta(\boldsymbol{w}^*, b^*) = \min_{\boldsymbol{w}, b} \phi_\beta(\boldsymbol{w}, b), \qquad (13)$$

i.e, the problems (6) and (11) are equivalent.

Introducing slack variables ξ_i, imposing $\|\boldsymbol{w}\|^2 = 1$ (which does not change the solution essentially; only the scale is changed), and letting $\nu = 1 - \beta$ and $\rho = -\alpha$ in Eq.(11), we establish the theorem.

B. Sketch of Proof of Lemma 2

Since Eq.(11) only involves continuous functions, continuity of $F_\beta(\boldsymbol{w}^*, b^*, \alpha^*)$ with respect to β is clear. From Eq.(13), $\phi_\beta(\boldsymbol{w}^*, b^*)$ is also continuous. Let $(\boldsymbol{w}^*_{\beta_i}, b^*_{\beta_i}, \alpha^*_{\beta_i})$ be the optimal solutions of Eq.(11) for $0 < \beta_1 < \beta_2 < 1$. Then we have

$$\begin{aligned}\phi_{\beta_1}(\boldsymbol{w}^*_{\beta_1}, b^*_{\beta_1}) &= F_{\beta_1}(\boldsymbol{w}^*_{\beta_1}, b^*_{\beta_1}, \alpha^*_{\beta_1}) \leq F_{\beta_1}(\boldsymbol{w}^*_{\beta_2}, b^*_{\beta_2}, \alpha^*_{\beta_2}) \\ &< F_{\beta_2}(\boldsymbol{w}^*_{\beta_2}, b^*_{\beta_2}, \alpha^*_{\beta_2}) = \phi_{\beta_2}(\boldsymbol{w}^*_{\beta_2}, b^*_{\beta_2}),\end{aligned}$$

where the first inequality is due to optimality of $(\boldsymbol{w}^*_{\beta_1}, b^*_{\beta_1}, \alpha^*_{\beta_1})$ and the second strict inequality is clear from Eq.(12). Thus $\phi_\beta(\boldsymbol{w}^*, b^*)$ is strictly increasing with respect to β, implying that $\phi_{1-\nu}(\boldsymbol{w}^*, b^*)$ is strictly decreasing with respect to ν.

C. Sketch of Proof of Theorem 3

For a homogeneous classifier $h(\widetilde{\boldsymbol{x}}) = \mathrm{sign}(\langle \widetilde{\boldsymbol{w}}, \widetilde{\boldsymbol{x}} \rangle)$, the following lemma holds:

Lemma 12 *(Schölkopf et al., 2000) Suppose that support $\mathcal{X}$ of $\widetilde{\boldsymbol{x}}$ is in a ball of radius $\widetilde{R}$ around the origin. Then, for all $\widetilde{\boldsymbol{w}}$ such that $\|\widetilde{\boldsymbol{w}}\| = 1$, there exists a positive constant c such that the following bound holds with probability at least $1 - \delta$:*

$$\begin{aligned}R[h] \leq & \frac{|\{i \mid y_i \langle \widetilde{\boldsymbol{w}}, \widetilde{\boldsymbol{x}}_i \rangle < \widetilde{\gamma}\}|}{m} \\ & + \sqrt{\frac{2}{m}\left(\frac{4c^2\widetilde{R}^2}{\widetilde{\gamma}^2}\log_2(2m) - 1 + \log\frac{2}{\delta}\right)}.\end{aligned}$$

Let $\widetilde{\boldsymbol{w}} = \frac{(\boldsymbol{w}^\top, b)^\top}{\sqrt{1+b^2}}$ and $\widetilde{\boldsymbol{x}} = (\boldsymbol{x}^\top, 1)^\top$. Then our classifier (1) can be regarded as homogeneous. The assumption that all the data points $\boldsymbol{x}$ live in a centered ball of radius R implies that all the data points $\widetilde{\boldsymbol{x}}$ live in a centered ball of radius

$$\widetilde{R} = \sqrt{R^2 + 1}.$$

The assumption $\|\boldsymbol{w}\| = 1$ implies $\|\widetilde{\boldsymbol{w}}\| = 1$. Then we can apply Lemma 12 to the current setting. The condition $y_i \langle \widetilde{\boldsymbol{w}}, \widetilde{\boldsymbol{x}}_i \rangle < \widetilde{\gamma}$ results in

$$y_i(\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle + b) < \widetilde{\gamma}\sqrt{1+b^2} := \gamma.$$

When all the data points $\boldsymbol{x}$ live in a centered ball of radius R, we can assume without loss of generality that $|b| \leq R$. Then we have

$$\frac{1}{\widetilde{\gamma}^2} = \frac{1+b^2}{\gamma^2} \leq \frac{1+R^2}{\gamma^2}.$$

Now let us set

$$\gamma = -\alpha_{1-\nu}(\boldsymbol{w}, b).$$

Then we can show that

$$\tfrac{1}{m}|\{i \mid y_i(\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle + b) < -\alpha_{1-\nu}(\boldsymbol{w}, b)\}| \leq \nu.$$

We omit its proof due to lack of space. Then we obtain the upper bound $\nu + G(\alpha_{1-\nu}(\boldsymbol{w}, b))$; the upper bound $\nu + G(\phi_{1-\nu}(\boldsymbol{w}, b))$ is clear from Eq.(5).

D. Sketch of Proof of Lemma 6

Since the difference between the problems (3) and (8) is only the norm constraint of $\boldsymbol{w}$, it is enough to show that for $\nu \in (\overline{\nu}, \nu_{\max}]$, $\|\boldsymbol{w}^*\|^2 = 1$ holds at the optimal solution $(\boldsymbol{w}^*, b^*, \boldsymbol{\xi}^*, \rho^*)$ of the problem (8). For such ν, $\phi_{1-\nu}(\boldsymbol{w}^*, b^*) < 0$ holds, i.e., the optimal value of Eν-SVC is negative. If we suppose $\|\boldsymbol{w}^*\|^2 < 1$, another feasible solution $(\boldsymbol{w}^*, b^*, \boldsymbol{\xi}^*, \rho^*)/\|\boldsymbol{w}^*\|$ achieves a smaller optimal value than $(\boldsymbol{w}^*, b^*, \boldsymbol{\xi}^*, \rho^*)$. This contradicts to the optimality of (8), and hence $\|\boldsymbol{w}^*\|^2 = 1$ is proved.

E. Sketch of Proof of Theorem 8

Let $(\widehat{\boldsymbol{w}}_k, \widehat{b}_k, \widehat{\boldsymbol{\xi}}_k, \widehat{\rho}_k)$ be an optimal solution of the linear program (9) in the k-th iteration. Then, a feasible solution of Eν-SVC (3) is given by

$$(\widetilde{\boldsymbol{w}}_k, \widetilde{b}_k, \widetilde{\boldsymbol{\xi}}_k, \widetilde{\rho}_k) = (\widehat{\boldsymbol{w}}_k, \widehat{b}_k, \widehat{\boldsymbol{\xi}}_k, \widehat{\rho}_k)/\|\widehat{\boldsymbol{w}}_k\|.$$

Since $(\widehat{\boldsymbol{w}}_k, \widehat{b}_k, \widehat{\boldsymbol{\xi}}_k, \widehat{\rho}_k)$ is at a corner of the feasible set of the linear program (9), $(\widetilde{\boldsymbol{w}}_k, \widetilde{b}_k, \widetilde{\boldsymbol{\xi}}_k, \widetilde{\rho}_k)$ is also a corner of the feasible set of Eν-SVC (3).

Let $q(\cdot)$ be the objective function of Eν-SVC (3), which is also the objective function of the linear program (9). Then we have

$$q(\widetilde{\boldsymbol{\xi}}_{k-1}, \widetilde{\rho}_{k-1}) > q(\widehat{\boldsymbol{\xi}}_k, \widehat{\rho}_k) \geq q(\widetilde{\boldsymbol{\xi}}_k, \widetilde{\rho}_k) = q(\widehat{\boldsymbol{\xi}}_k, \widehat{\rho}_k)/\|\widehat{\boldsymbol{w}}_k\|,$$

where the first inequality comes from the optimality of $(\widehat{\boldsymbol{\xi}}_k, \widehat{\rho}_k)$ of the linear program (9). The second inequality comes from $\|\widehat{\boldsymbol{w}}_k\| > 1$, which is ensured by $\langle \widetilde{\boldsymbol{w}}_{k-1}, \widehat{\boldsymbol{w}}_k \rangle = 1$. Thus the algorithm finds a distinct corner of Eν-SVC (3) in each iteration. Since the number of corners of Eν-SVC (3) is finite, the algorithm terminates within finite iterations.

Let $\Delta\boldsymbol{d} = (\Delta\boldsymbol{w}^\top\ \Delta b^\top\ \Delta\rho^\top\ \Delta\boldsymbol{\xi}^\top)^\top$ be a perturbation from the solution $\boldsymbol{d}^* = (\boldsymbol{w}^*, b^*, \rho^*, \boldsymbol{\xi}^*)$ of Algorithm 7. Note that $\boldsymbol{d}^*$ is an optimal solution of the linear program (9) with $\widetilde{\boldsymbol{w}} = \boldsymbol{w}^*$. Using the Karush-Kuhn-Tucker (KKT) optimality conditions, we can express the increase Δq of the objective value as

$$\begin{aligned}\Delta q &:= -\nu\Delta\rho + \frac{1}{m}\sum_{i \in M}\Delta\xi_i \\ &= \Delta\boldsymbol{d}^\top \left(\begin{array}{ccc|c} y_1\boldsymbol{x}_1 & \dots & y_m\boldsymbol{x}_m & \boldsymbol{O} \\ y_1 & \dots & y_m & \boldsymbol{0} \\ 1 & \dots & 1 & \boldsymbol{0} \\ & \boldsymbol{I} & & \boldsymbol{I} \end{array}\right) \begin{pmatrix} \boldsymbol{\lambda}^* \\ \boldsymbol{\mu}^* \end{pmatrix} - \delta^*\Delta\boldsymbol{w}^\top\boldsymbol{w}^*,\end{aligned}$$

where $\boldsymbol{\lambda}^* \in \mathbb{R}^m_+$, $\boldsymbol{\mu}^* \in \mathbb{R}^m_+$, and $\delta^* \leq 0$ are KKT multipliers. If $\Delta\boldsymbol{d}$ is a feasible perturbation (i.e., $\boldsymbol{d}^* + \Delta\boldsymbol{d}$ is feasible), we can show that $\Delta q > 0$ (we omit its proof due to lack of space), which implies that $\boldsymbol{d}^*$ is locally optimal.