ISSN 1342-2804

Research Reports on Mathematical and Computing Sciences

Support Vector Machine Based on Conditional Value-at-Risk Minimization

Akiko Takeda

March 2007,
August 2007 (revised), B–439

Department of
Mathematical and
Computing Sciences
Tokyo Institute of Technology

SERIES B: Operations Research

B-439 Support Vector Machine Based on Conditional Value-at-Risk Minimization

Akiko Takeda†

March 2007, August 2007 (revised)

Abstract. A binary linear classification method, CGS method, was recently proposed by Gotoh and Takeda. The classification model was developed by introducing a risk measure known as the conditional value-at-risk (β-CVaR). It is found to be equivalent to Extended ν-SVC of Perez-Cruz et al., and especially in the convex case, equivalent to ν-SVC of Schölkopf et al.

The aim of this paper is to propose β-SVM by extending CGS classification model, investigate the relation between β-SVM and (Extended) ν-SVMs, and discuss theoretical aspects, mainly generalization performance, of β-SVM. The formula of a generalization error bound includes β-CVaR or a related quantity. It implies that the minimum β-CVaR obtained via β-SVM plays an important role to control the generalization error of β-SVM. The viewpoint from CVaR minimization is useful to make sure of the validity of not only β-SVM but also ν-SVM. We furthermore show a numerical example of nonconvex β-SVR, an extension of ν-SVR.

Keywords: support vector machines, geometric margin, conditional value-at-risk, nonconvex quadratic programming, extended ν-SVM

1 Introduction

Support vector classifications (SVCs) are widely used as computationally powerful tools for binary classification. The binary classification problem in $\mathbb{R}^n$ finds a decision function $h : \chi \to \{\pm 1\}$ based on given training data

$$(\boldsymbol{x}_i, y_i) \in \chi \times \{\pm 1\}, \quad i \in M := \{1, ..., m\}.$$

The examples $\boldsymbol{x}_i$, $i \in M$, are taken from some nonempty set $\chi \subset \mathbb{R}^n$ and the labels y_i, $i \in M$ are from binary values: -1 or 1. We assume that the data were generated independently from some unknown probability distribution $P(\boldsymbol{x}, y)$. The goal of the learning process is to find a decision function h which classifies unseen data $(\boldsymbol{x}, y)$, generated from $P(\boldsymbol{x}, y)$, so that hopefully $h(\boldsymbol{x}) = y$ holds. Minimizing the empirical risk

$$R_{emp}[h] = \frac{1}{m} \sum_{i \in M} \frac{1}{2} |h(\boldsymbol{x}_i) - y_i|$$

does not necessarily lead to a small risk

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

It is difficult to find h which minimizes $R[h]$, since we do not know P.

In the case where the training data are linearly separable, the simplest kind of SVC, known as hard margin SVM, requires computing the distance from training data $(\boldsymbol{x}, y)$ to a hyperplane $\{\boldsymbol{x} \in \mathbb{R}^n : \langle \boldsymbol{w}, \boldsymbol{x} \rangle + b = 0\}$ by

$$g(\boldsymbol{w}, b; \boldsymbol{x}) := \frac{\langle \boldsymbol{w}, \boldsymbol{x} \rangle + b}{\|\boldsymbol{w}\|},$$

and finding a solution $(\boldsymbol{w}, b)$ which maximizes the *geometric margin*: $\min_{i \in M} y_i g(\boldsymbol{w}, b; \boldsymbol{x}_i)$. Namely, a decision function of hard margin SVM is formed with an optimal solution $(\boldsymbol{w}^*, b^*)$ of

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

as $h(\boldsymbol{x}) = \mathrm{sign}(\langle \boldsymbol{w}^*, \boldsymbol{x} \rangle + b^*)$. Here, $\mathrm{sign}(\xi)$ is a function such that $\mathrm{sign}(\xi) = 1$ if $\xi \geq 0$ and -1, otherwise. In practice, training data are usually not linearly separable, and there are various soft margin approaches which incorporate a trade-off between maximizing the geometric margin and minimizing a measure of classification error on the training data. A well-known realization of a soft margin is C-SVM, classic formulation proposed by [4]. Also, to resolve the difficulty in choosing the parameter value C a priori which measures the trade-off of two objectives, an alternative formulation ν-SVC was proposed by [13] and developed by [5, 6]. The model includes more meaningful parameter ν instead of C. Moreover, to extend the permissible range of ν up to $[0, 1)$, Extended ν-SVC was developed by Perez-Cruz et al. [11] for a nonlinearly separable dataset based on a reinterpretation on how maximum margin hyperplanes are constructed for a separable dataset.

As one of soft margin approaches in the nonlinearly separable case, Gotoh and Takeda [8] proposed the *conditional geometric score* (CGS) optimization method. It regards

$$f(\boldsymbol{w}, b; \boldsymbol{x}, y) = -y g(\boldsymbol{w}, b; \boldsymbol{x}) \tag{1}$$

as a cost function, and minimizes mean of the β-tail distribution of $f(\boldsymbol{w}, b; \boldsymbol{x}, y)$ with respect to decision variables $(\boldsymbol{w}, b)$. The β-tail expectation of $f(\boldsymbol{w}, b; \boldsymbol{x}, y)$ is known as *conditional value-at-risk* (β-CVaR) [12], and denoted by $\phi_\beta(\boldsymbol{w}, b)$. In CGS classification procedure, β-CVaR is minimized as $\min_{\boldsymbol{w}, b} \phi_\beta(\boldsymbol{w}, b)$. The CGS problem has no square regularization term $\|\boldsymbol{w}\|$ in the objective function. But it is transformed into a nonconvex problem with the constraint $\|\boldsymbol{w}\| = 1$. The constraint is expected to raise the generalization ability by avoiding over-fitting to training data as well as the generalization term in ν-SVC of [13].

We formulate a kernelized variant of CGS classification by adopting a kernel function in the CGS problem according to SVC, and call β-SVC for the CGS problem and the kernelized variant. β-SVC is keenly related to other existing classification methods such as ν-SVC and Extended ν-SVC (see Table 1). β-SVC is essentially the same as the formulation of Extended ν-SVC proposed by Perez-Cruz et al [11]. Depending on the given parameter β, β-SVC is formulated as a convex *quadratic programming* (QP) problem or a nonconvex QP problem. The decision function of convex β-SVC, obtained by solving a convex QP problem, is equivalent to that of ν-SVC [13, 5, 6]. We also propose β-SVR: a CVaR minimization problem for a regression problem. β-SVR considers $f(\boldsymbol{w}, b; \boldsymbol{x}, y) = |y - C g(\boldsymbol{w}, b; \boldsymbol{x})|$ as a cost function with parameter $C > 0$, and minimizes β-CVaR for the distribution of f. The formulation of β-SVR has no square regularization term $\|\boldsymbol{w}\|$ in the objective function, but

Table 1: β-SVMs are equivalent to existing SVMs

	convex case	nonconvex case
classification	ν-SVC [13]	Extended ν-SVC [11]
regression	ν-SVR [13]	–

has the constraint $\|\boldsymbol{w}\| = C$ to raise the generalization ability. We prove theoretically that ν-SVR of [13] and the convex case of β-SVR are essentially the same if we set a parameter of each model appropriately. Thus, β-SVR is regarded as an extension of ν-SVR, and has a possibility to find a good classifier.

The aim of this paper is to propose β-SVM based on CVaR minimization, investigate the relation between β-SVM and (Extended) ν-SVMs, and discuss theoretical aspects, mainly generalization performance, of β-SVM. We furthermore show a numerical example of nonconvex β-SVR, since there are no existing SVRs which correspond to the nonconvex model as far as we know. The viewpoint from CVaR minimization is useful to make sure of the validity of not only β-SVM but also ν-SVM. The combination of the theory of generalization performance developed by [1, 13, 16] and CVaR risk measure of [12] leads to an estimation of a generalization error bound for convex β-SVC. Indeed, with probability at least $1-\delta$ over m training data points, a decision function $h = \mathrm{sign}(\langle \boldsymbol{w}, \boldsymbol{v} \rangle + b)$ with $\|\boldsymbol{w}\| = 1$ and $\phi_\beta(\boldsymbol{w}, b) < 0$ has a probability of test error bounded according to

$$R[h] \quad \leq (1-\beta) + \sqrt{\frac{2}{m}\left(\frac{4c^2(1+B_R^2)^2}{\phi_\beta(\boldsymbol{w},b)^2}\log_2(2m) - 1 + \log\left(\frac{2}{\delta}\right)\right)},$$

where c is a constant and B_R is the radius of a ball that all the data points live in. The convex β-SVC can be formulated as $\min_{\boldsymbol{w},b} \phi_\beta(\boldsymbol{w}, b)$ subject to $\|\boldsymbol{w}\| = 1$ and $\phi_\beta(\boldsymbol{w}, b) < 0$. Therefore, we have the tightest upper bound for $R[h]$ at the optimal solution of convex β-SVC. Similarly, we have a generalization error bound for nonconvex β-SVC, which implies the validity of Extended ν-SVC as well as β-SVM.

This paper is organized as follows. In Section 2, CGS classification [8] and its nonlinear kernel-based variants are shown. Section 3 presents a generalization error bound for β-SVC. It indicates that an optimal solution of β-SVC plays an important role for minimizing the error bound. Section 4 considers a regression problem from the point of view of CVaR minimization. A convex case of a CVaR minimization problem, convex β-SVR, is found to be equivalent to ν-SVR if we set a parameter of each model appropriately. Section 5 shows a numerical example of nonconvex β-SVR.

2 β-SVC Based on CVaR Minimization

We present a linear classification method proposed by Gotoh and Takeda [8] for nonlinearly separable datasets. And then, we incorporate nonlinearity in the classification model by following SVC.

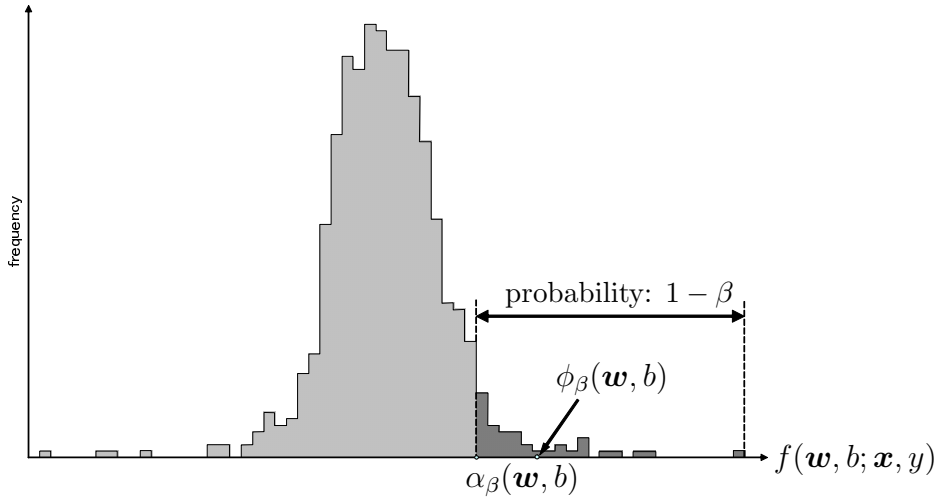

Figure 1: Illustration of the β-tail expectation of f

2.1 Linear Classification

Let us define the distribution function of f by

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

and a threshold α_β with a confidence level $\beta \in (0, 1)$ by

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

It is expected that f exceeds α_β only in $(1-\beta)\times 100\%$ as Figure 1 shows. The β-percentile α_β is known as the *value-at-risk* (β-VaR) in the context of financial risk management. It is typically used by security houses or investment banks to measure the market risk of their asset portfolios. We note that α_β is well-defined because $\Phi(\alpha\,|\boldsymbol{w}, b)$ is right continuous and nondecreasing with respect to α. Following the discussion of Rockafellar & Uryasev [12], we introduce the β-tail distribution function to focus on the tail part of $\Phi(\alpha\,|\boldsymbol{w}, b)$ as

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

Using the expectation operator $\boldsymbol{E}_\beta[\cdot]$ under the β-tail distribution Φ_β, let us define the β-tail expectation of f, known as β-CVaR, by $\phi_\beta(\boldsymbol{w}, b) := \boldsymbol{E}_\beta\,[\,f(\boldsymbol{w}, b; \boldsymbol{x}, y)\,]$. Denoting the expectation under the original distribution Φ by $\boldsymbol{E}[\cdot]$, Rockafellar & Uryasev [12] show that β-CVaR satisfies the following relation:

$$\begin{aligned} \boldsymbol{E}[\,f(\boldsymbol{w}, b; \boldsymbol{x}, y) \,|\, f(\boldsymbol{w}, b; \boldsymbol{x}, y) \geq \alpha_\beta(\boldsymbol{w}, b)\,] &\leq \phi_\beta(\boldsymbol{w}, b) \\ &\leq \boldsymbol{E}[\,f(\boldsymbol{w}, b; \boldsymbol{x}, y) \,|\, f(\boldsymbol{w}, b; \boldsymbol{x}, y) > \alpha_\beta(\boldsymbol{w}, b)\,]. \end{aligned}$$

The CGS classification model (or β-SVC with linear kernel) [8] minimizes β-CVaR $\phi_\beta(\boldsymbol{w}, b)$ for geometric score distribution $f(\boldsymbol{w}, b; \boldsymbol{x}_i, y_i)$, $i \in M$, of (1) to define $h(\boldsymbol{x}) = \mathrm{sign}(\langle \boldsymbol{w}^*, \boldsymbol{x}\rangle + b^*)$.

In the problem, a threshold is set on β-VaR of $f(\boldsymbol{w}, b; \boldsymbol{x}, y)$, and expected excess of $f(\boldsymbol{w}, b; \boldsymbol{x}, y)$ over β-VaR, which corresponds to β-CVaR, is regarded as the loss or risk. Minimizing β-CVaR $\phi_\beta(\boldsymbol{w}, b)$ is shown in [12] to be equivalent to minimize

$$F_\beta(\boldsymbol{w}, b, \alpha) := \alpha + \frac{1}{1-\beta} \boldsymbol{E}\left[\, [f(\boldsymbol{w}, b; \boldsymbol{x}, y) - \alpha]^+ \,\right], \tag{2}$$

where $[X]^+ := \max\{X, 0\}$, with respect to $(\boldsymbol{w}, b, \alpha)$. We have $\min_{\boldsymbol{w},b} \phi_\beta(\boldsymbol{w}, b) = \min_{\boldsymbol{w},b,\alpha} F_\beta(\boldsymbol{w}, b, \alpha)$. Moreover, an optimal solution $(\boldsymbol{w}^*, b^*, \alpha^*)$ of $\min_{\boldsymbol{w},b,\alpha} F_\beta(\boldsymbol{w}, b, \alpha)$ is also optimal in $\min_{\boldsymbol{w},b} \phi_\beta(\boldsymbol{w}, b)$, and the solution α^*, obtained as a by-product, is almost equal to β-VaR of $f(\boldsymbol{w}^*, b^*; \boldsymbol{x}, y)$. Namely, the solution α^* is guaranteed to be in the range $[\alpha_\beta(\boldsymbol{w}^*, b^*),\ \alpha^+_\beta(\boldsymbol{w}^*, b^*)]$, where $\alpha^+_\beta(\boldsymbol{w}, b)$ is *upper β-VaR* defined by

$$\alpha^+_\beta(\boldsymbol{w}, b) := \inf\{\, \alpha \,|\, \Phi(\alpha \,|\boldsymbol{w}, b) > \beta \,\}.$$

Note that $\alpha_\beta(\boldsymbol{w}^*, b^*)$ and $\alpha^+_\beta(\boldsymbol{w}^*, b^*)$ are the same except when $\Phi(\alpha \,|\boldsymbol{w}^*, b^*)$ is constant at level β over a certain α-interval. Thus, an optimal solution α^* approximately indicates the threshold α of $\Phi(\alpha \,|\boldsymbol{w}, b) = \beta$ where f exceeds α only in $(1-\beta) \times 100\%$.

The problem of β-SVC, $\min_{\boldsymbol{w},b} \phi_\beta(\boldsymbol{w}, b)$, is reformulated as $\min_{\boldsymbol{w},b,\alpha} F_\beta(\boldsymbol{w}, b, \alpha)$, that is,

$$\begin{array}{ll} \min\limits_{\boldsymbol{w},b,\alpha,\boldsymbol{z}} & \alpha + \dfrac{1}{(1-\beta)m} \displaystyle\sum_{i \in M} z_i \\ \text{subject to} & z_i + y_i(\, \langle \boldsymbol{w}, \boldsymbol{x}_i \rangle + b \,) + \alpha \geq 0, \quad i \in M, \\ & z_i \geq 0, \quad i \in M, \\ & \boldsymbol{w}^\top \boldsymbol{w} = 1. \end{array} \tag{3}$$

Let the numbers of data with positive and negative labels be m_+ and m_-, respectively, and suppose that m_+ and m_- are positive. Then Problem (3) is proved to have an optimal solution when the parameter β is chosen so that

$$\beta_{min} := 1 - \frac{2\min\{m_+, m_-\}}{m} \leq \beta < 1.$$

Problem (3) is essentially the same as the formulation of Extended ν-SVC proposed by Perez-Cruz et al [11]:

$$\begin{array}{ll} \min\limits_{\boldsymbol{w},b,\rho,\boldsymbol{z}} & -m\nu\rho + \displaystyle\sum_{i \in M} z_i \\ \text{subject to} & z_i + y_i(\, \langle \boldsymbol{w}, \boldsymbol{x}_i \rangle + b \,) - \rho \geq 0, \quad i \in M, \\ & z_i \geq 0, \quad i \in M, \\ & \frac{1}{2}\boldsymbol{w}^\top \boldsymbol{w} = 1. \end{array}$$

Note that Extended ν-SVC with $\nu = 1 - \beta$ and β-SVC generate the same classifier. Extended ν-SVC was developed for a nonlinearly separable dataset based on a reinterpretation on how maximum margin hyperplanes are constructed for a separable dataset. Thus, the discussion on β-SVC makes it possible to interpret Extended ν-SVC from a different viewpoint.

It is important to note that the optimal value of (3) is nondecreasing with respect to β. Indeed, for arbitrary parameter $\beta_1 \leq \beta_2$ in $[\beta_{min}, 1)$ and their optimal solutions $(\boldsymbol{w}^*_{\beta_i}, b^*_{\beta_i}, \alpha^*_{\beta_i})$, $i = 1, 2$, of (3), we have

$$\min_{\boldsymbol{w},b,\alpha} F_{\beta_1}(\boldsymbol{w}, b, \alpha) \leq F_{\beta_1}(\boldsymbol{w}^*_{\beta_2}, b^*_{\beta_2}, \alpha^*_{\beta_2}) \leq \min_{\boldsymbol{w},b,\alpha} F_{\beta_2}(\boldsymbol{w}, b, \alpha).$$

When the training data are linearly separable, there exists $(\boldsymbol{w}, b)$ such that $y_i g(\boldsymbol{w}, b; \boldsymbol{x}_i) > 0$, *i.e.,* $f(\boldsymbol{w}, b; \boldsymbol{x}_i, y_i) < 0$ holds for all $i \in M$. Then, at an optimal solution $(\boldsymbol{w}^*_\beta, b^*_\beta)$ for any β, β-CVaR $\phi_\beta(\boldsymbol{w}^*_\beta, b^*_\beta)$ and β-VaR $\alpha_\beta(\boldsymbol{w}^*_\beta, b^*_\beta)$ obviously take negative values. In the nonlinearly separable case, however, β-CVaR and β-VaR possibly become positive especially for large β. Since the optimal value of (3), $\phi_\beta(\boldsymbol{w}^*_\beta, b^*_\beta) = F_\beta(\boldsymbol{w}^*_\beta, b^*_\beta, \alpha^*_\beta)$, is nondecreasing with respect to β, there may exist $\bar{\beta}$ which induces 0 optimal value in (3), that is, $\phi_{\bar{\beta}}(\boldsymbol{w}^*_{\bar{\beta}}, b^*_{\bar{\beta}}) = 0$, though it is difficult to find such $\bar{\beta}$ exactly. When $\phi_\beta(\boldsymbol{w}^*_\beta, b^*_\beta) < 0$ holds for all β or $\phi_\beta(\boldsymbol{w}^*_\beta, b^*_\beta) > 0$ does for all β, $\bar{\beta}$ can be set to $\bar{\beta} = 1$ or $\bar{\beta} = \beta_{min}$, respectively. With the use of $\bar{\beta}$, Problem (3) is classified into two cases: the convex case where the optimal value of (3) is negative for $\beta \in [\beta_{min}, \bar{\beta})$, and the nonconvex case where its optimal value is nonnegative for $\beta \in [\bar{\beta}, 1)$.

Problem (3) is not obviously convex. But when β is in the range $[\beta_{min}, \bar{\beta})$, it can be transformed into a convex problem:

$$\begin{array}{ll} \min\limits_{\boldsymbol{w},b,\alpha,\boldsymbol{z}} & \alpha + \dfrac{1}{(1-\beta)m} \displaystyle\sum_{i \in M} z_i \\ \text{subject to} & z_i + y_i(\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle + b) + \alpha \geq 0, \quad i \in M, \\ & z_i \geq 0, \quad i \in M, \\ & \boldsymbol{w}^\top \boldsymbol{w} \leq 1. \end{array} \tag{4}$$

The nonconvex constraint $\boldsymbol{w}^\top \boldsymbol{w} = 1$ of (3) is relaxed into a convex constraint $\boldsymbol{w}^\top \boldsymbol{w} \leq 1$, since $\boldsymbol{w}^{*\top} \boldsymbol{w}^* = 1$ is attained at an optimal solution $\boldsymbol{w}^*$ of Problem (3) with $\beta \in [\beta_{min}, \bar{\beta})$. Therefore, we can use an efficient solution method such as the interior-point method and Sequential Minimal Optimization (SMO) algorithm for the convex QP (4) with $\beta \in [\beta_{min}, \bar{\beta})$.

On the other hand, if $\boldsymbol{w}^\top \boldsymbol{w} = 1$ is replaced with $\boldsymbol{w}^\top \boldsymbol{w} \leq 1$ for β-SVC (3) with $\beta \in (\bar{\beta}, 1)$, one has the meaningless optimal solution $\boldsymbol{w} = \boldsymbol{0}$ and $b = 0$. Therefore, in this case, the nonconvex constraint $\boldsymbol{w}^\top \boldsymbol{w} = 1$ is essential. The nonconvex constraint $\boldsymbol{w}^\top \boldsymbol{w} = 1$ in Problem (3) can be replaced with $\boldsymbol{w}^\top \boldsymbol{w} \geq 1$ as far as β is in $[\bar{\beta}, 1)$. This type of a nonconvex problem, whose feasible region is the intersection of a polyhedral set with a concave inequality (say, $\boldsymbol{w}^\top \boldsymbol{w} \geq 1$), is often referred to a *linear reverse convex program* (LRCP), and several kinds of global optimization algorithms such as cutting methods are proposed for finding a global minimizer of an LRCP (see [9]). But they consume long computation time as the size of the problem becomes large. To save computation time, we can compromise with a local minimizer of (3) by applying a local search algorithm [8] or Extended ν-SVC algorithm [11].

2.2 Nonlinear Kernel-Based Classification

We consider incorporating a kernel function into β-SVC (3). At first, we consider the convex case of (3) which is transformed into a convex problem (4). Taking dual for Problem (4) and

incorporating kernels $k(\boldsymbol{x}_i, \boldsymbol{x}_j)$ to dot products $\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle$ in the objective function, we have a QP problem:

$$\begin{array}{ll} \max\limits_{\boldsymbol{\lambda}} & -\sum\limits_{i,j \in M} y_i y_j k(\boldsymbol{x}_i, \boldsymbol{x}_j) \lambda_i \lambda_j \\ \text{subject to} & \sum\limits_{i \in M} \lambda_i y_i = 0, \\ & \sum\limits_{i \in M} \lambda_i = 1, \\ & 0 \leq \lambda_i \leq \dfrac{1}{(1-\beta)m}, \quad i \in M. \end{array} \tag{5}$$

This is exactly ν-SVC with parameter $\nu = 1 - \beta$. See [8] for details on the derivation of the dual (5).

Note that when a kernel matrix, which consists of $k(\boldsymbol{x}_i, \boldsymbol{x}_j)$, $i, j \in M$, is positive definite as in radial basis functions (RBF), the optimal value of (5) is negative for any $\beta \in [\beta_{min}, 1)$. However, when a positive semi-definite kernel matrix is applied, (5) with β beyond $\bar{\beta}$ provides a solution with 0 optimal value. Chang and Lin [5] pointed out that ν-SVC has zero optimal value for any $[0, \nu_{min}]$, but the value ν_{min} corresponding to $1 - \bar{\beta}$ is not clarified. When the optimal value of (5) becomes 0, we see that the parameter β is in $[\bar{\beta}, 1)$. That is, the optimal solution of (5) is already irrelevant to β-CVaR, and the resulting decision function may be unreliable with respect to prediction accuracy. If such a case occurs, this is a clever way of switching the concerned problem (5) to a nonconvex one.

Next, the nonconvex case is considered. Similar to the convex case, we consider to incorporate a kernel function into (3). The kernel function is usually incorporated in the dual formulation of SVC, but the difficulty with incorporating a kernel function into (3) is caused by duality gap between Problem (3) and its dual problem. In other words, solving the dual problem of (3) never means solving Problem (3). That is why we draw an analogy between a convex problem (4) and its kernelized problem, and then, based on this analogy, we propose to incorporate a kernel function into a nonconvex problem (3).

To this end, we consider the dual problem of (5) with a positive (semi)-definite kernel matrix. Taking dual for (5) leads to

$$\begin{array}{ll} \min\limits_{\boldsymbol{w}, b, \alpha, \boldsymbol{z}} & \alpha + \dfrac{1}{(1-\beta)m} \sum\limits_{i \in M} z_i \\ \text{subject to} & z_i + y_i(\langle \boldsymbol{w}, \boldsymbol{v}_i \rangle + b) + \alpha \geq 0, \quad i \in M, \\ & z_i \geq 0, \quad i \in M, \\ & \boldsymbol{w}^\top \boldsymbol{w} \leq 1, \end{array} \tag{6}$$

where $\boldsymbol{v}_i$ is obtained from a kernel matrix by decomposing it as

$$\begin{bmatrix} k(\boldsymbol{x}_1, \boldsymbol{x}_1) & \cdots & k(\boldsymbol{x}_1, \boldsymbol{x}_m) \\ & \cdots & \\ k(\boldsymbol{x}_m, \boldsymbol{x}_1) & \cdots & k(\boldsymbol{x}_m, \boldsymbol{x}_m) \end{bmatrix} = \begin{bmatrix} \boldsymbol{v}_1^\top \\ \vdots \\ \boldsymbol{v}_m^\top \end{bmatrix} [\boldsymbol{v}_1 \ \cdots \ \boldsymbol{v}_m] .$$

The dimension of $\boldsymbol{v}_i$ depends on the rank r of a given kernel matrix. In popular kernel functions $k(\boldsymbol{x}_i, \boldsymbol{x}_j)$, the rank r becomes larger than or equal to the dimension n of original

data $\boldsymbol{x}_i$. Considering that the nonconvex β-SVC (3) replaces the convex constraint $\boldsymbol{w}^\top \boldsymbol{w} \leq 1$ of (4) with $\boldsymbol{w}^\top \boldsymbol{w} = 1$, we formulate nonconvex kernelized β-SVC with decision variables $\boldsymbol{w} \in \mathbb{R}^r$ $b \in \mathbb{R}$, $\alpha \in \mathbb{R}$ and $\boldsymbol{z} \in \mathbb{R}^m$ as

$$\begin{array}{ll} \min\limits_{\boldsymbol{w},b,\alpha,\boldsymbol{z}} & \alpha + \dfrac{1}{(1-\beta)m} \displaystyle\sum_{i \in M} z_i \\ \text{subject to} & z_i + y_i(\langle \boldsymbol{w}, \boldsymbol{v}_i \rangle + b) + \alpha \geq 0, \quad i \in M, \\ & z_i \geq 0, \quad i \in M, \\ & \boldsymbol{w}^\top \boldsymbol{w} = 1. \end{array} \tag{7}$$

Under a suitable constraint qualification, a local minimizer $(\boldsymbol{w}^*, b^*, \alpha^*, \boldsymbol{z}^*)$ of Problem (7) satisfies the Karush-Kuhn-Tucker (KKT) conditions: there is a vector $\boldsymbol{\lambda} \in \mathbb{R}^m$ and $\delta \in \mathbb{R}$ satisfying

$$\begin{cases} \delta \geq 0 \text{ (convex case)}, \quad \delta \leq 0 \text{ (nonconvex case)}, \\ 0 \leq \lambda_i \leq \dfrac{1}{(1-\beta)m}, \ i \in M, \\ \displaystyle\sum_{i \in M} \lambda_i y_i \boldsymbol{v}_i = \delta \boldsymbol{w}^*, \ \sum_{i \in M} \lambda_i y_i = 0, \ \sum_{i \in M} \lambda_i = 1, \\ \lambda_i \{ z_i^* + y_i(\langle \boldsymbol{w}^*, \boldsymbol{v}_i \rangle + b^*) + \alpha^* \} = 0, \quad i \in M, \\ z_i^* \left(\dfrac{1}{(1-\beta)m} - \lambda_i \right) = 0, \quad i \in M. \end{cases} \tag{8}$$

The vector $\boldsymbol{\lambda}$ of λ_i, $i \in M$, and δ are called KKT multipliers. The last two equations of (8) are called complementarity conditions. We also call a point $(\boldsymbol{w}, b, \alpha, \boldsymbol{z})$ satisfying KKT conditions as a *KKT point*.

Let $(\boldsymbol{w}^*, b^*)$ be an optimal solution of (7) and $(\boldsymbol{\lambda}^*, \delta^*)$ be its KKT multipliers. Using the relation $\boldsymbol{w}^* = \frac{1}{\delta^*} \sum_{i \in M} \lambda_i^* y_i \boldsymbol{v}_i$, we can estimate the label of a new data point $\boldsymbol{x}$ as $h(\boldsymbol{x}) = \mathrm{sign}(\frac{1}{\delta^*} \sum_{i \in M} \lambda_i^* y_i k(\boldsymbol{x}, \boldsymbol{x}_i) + b^*)$.

The kernelization of nonconvex β-SVC (7) is natural in the sense that (7) is equivalent to ν-SVC if $\beta \leq \bar{\beta}$. The kernelization may be connected to one of kernelization approaches proposed for Extended ν-SVC in [11]: kernelization in the primal with the use of a kernel PCA decomposition. There are several issues that need further investigation: for example, which kind of a kernel function $k(\boldsymbol{x}, \boldsymbol{x})$ has a good fit in β-SVC (7). As stated in [11], the learning method can be used together with the new trend in the fields of machine learning in which the kernel matrix is learned with semi-definite programming [10]. If the optimal kernel matrix is not full-rank, nonconvex β-SVC (7) has a possibility to find a good classifier. Although there remains considerable room for improvement regarding the kernelization, it makes possible to evaluate the generalization error of nonconvex β-SVC.

3 Generalization Error Bound

We show that the parameter β of β-SVC bounds the fractions of margin errors and support vectors. That is, the well-known property of ν-SVC holds not only for convex β-SVC with

$\beta \in [\beta_{min}, \bar{\beta})$ but also for nonconvex β-SVC with $\beta \in [\bar{\beta}, 1)$. Using this property, we derive a generalization error bound from β-SVC in convex and nonconvex cases, and confirm that the classifier achieved via β-SVC is appropriate to enhance the generalization performance.

3.1 Parameter β Controlling Fractions of Support Vectors and Margin Errors

Considering the equivalence between convex β-SVC and ν-SVC, we see that the parameter $\beta \in [\beta_{min}, \bar{\beta})$ of β-SVC bounds the fractions of margin errors and support vectors. In the nonconvex case, margin errors and support vectors are defined with a KKT point $(\boldsymbol{w}^*, b^*, \alpha^*, \boldsymbol{z}^*)$ of (7) and its KKT multiplier vector $\boldsymbol{\lambda}^*$. Concretely, *margin errors* are the training data $(\boldsymbol{v}_i, y_i)$ lying outside of $y_i(\langle \boldsymbol{w}^*, \boldsymbol{v}_i \rangle + b^*) + \alpha^* \geq 0$, that is, data points with $z_i^* > 0$, and *support vectors* are data points with positive KKT multipliers $\lambda_i^* > 0$. Now, let Err denote the indices of margin errors, and SV does those of support vectors. In ν-SVC, margin errors and support vectors are defined with the use of an optimal solution of a convex QP problem equivalent to (6), but in β-SVC they are defined with a KKT point of a nonconvex problem (7). We can show the β property: $\frac{|Err|}{m} \leq 1 - \beta \leq \frac{|SV|}{m}$ for nonconvex β-SVC in the same manner as ν property for ν-SVC.

Remind that the dimension of $\boldsymbol{v}$ is r. We can show that the difference of two bounds of $1-\beta$ are $\frac{|SV|-|Err|}{m} \leq \frac{r+1}{m}$ at a KKT point with strict complementarity. The inclusion relation:

$$SV = \{\ i\ :\ \lambda_i^* > 0\} \quad \subseteq \{\ i\ :\ y_i(\langle \boldsymbol{w}^*, \boldsymbol{v}_i \rangle + b^*) + \alpha^* \leq 0\}, \tag{9}$$

holds with equality under the strict complementarity assumption. Then, we have $|SV| - |Err| = |\{\ i\ :\ y_i(\langle \boldsymbol{w}^*, \boldsymbol{v}_i \rangle + b^*) + \alpha^* = 0\ \}|$. According to the sign of $\eta_i := y_i(\langle \boldsymbol{w}^*, \boldsymbol{v}_i \rangle + b^*) + \alpha^*$, $i \in M$, we classify the indices in M into three groups:

$$\begin{aligned}
\mathcal{K}_- &:= \{i \in M : \eta_i < 0\} = \{i \in M : z_i^* > 0,\ z_i^* + y_i(\langle \boldsymbol{w}^*, \boldsymbol{v}_i \rangle + b^*) + \alpha^* = 0\},\\
\mathcal{K}_+ &:= \{i \in M : \eta_i > 0\} = \{i \in M : z_i^* = 0,\ z_i^* + y_i(\langle \boldsymbol{w}^*, \boldsymbol{v}_i \rangle + b^*) + \alpha^* > 0\},\\
\mathcal{K}_0 &:= \{i \in M : \eta_i = 0\} = \{i \in M : z_i^* = 0,\ z_i^* + y_i(\langle \boldsymbol{w}^*, \boldsymbol{v}_i \rangle + b^*) + \alpha^* = 0\}.
\end{aligned}$$

Note that $|\mathcal{K}_-| + |\mathcal{K}_+| + |\mathcal{K}_0| = m$. Also, Problem (7) has at most $r+m+1$ active inequality-constraints at a nondegenerate KKT point, since the problem has $r+m+2$ decision variables. Therefore, we see that $|\mathcal{K}_-| + |\mathcal{K}_+| + 2|\mathcal{K}_0| \leq r+m+1$, which implies $|SV| - |Err| = |\mathcal{K}_0| \leq r+1$.

Now we define index sets $Err(\boldsymbol{w}, b)$ and $SV_+(\boldsymbol{w}, b)$ for arbitrary $(\boldsymbol{w}, b)$ using β-VaR, $\alpha_\beta(\boldsymbol{w}, b)$, computed from the distribution of $f(\boldsymbol{w}, b; \boldsymbol{v}_i, y_i)$, $i \in M$, of (1). They play an important role in estimating a generalization error bound of β-SV classifier. Let $Err(\boldsymbol{w}, b)$ denote the indices of training data $(\boldsymbol{v}_i, y_i)$ satisfying $y_i(\langle \boldsymbol{w}, \boldsymbol{v}_i \rangle + b) + \alpha_\beta(\boldsymbol{w}, b) < 0$, and $SV_+(\boldsymbol{w}, b)$ does those of $(\boldsymbol{v}_i, y_i)$ with $y_i(\langle \boldsymbol{w}, \boldsymbol{v}_i \rangle + b) + \alpha_\beta(\boldsymbol{w}, b) \leq 0$. When $(\boldsymbol{w}, b)$ is set to a KKT point $(\boldsymbol{w}^*, b^*)$ of Problem (7), those sets include the indices of margin errors and support vectors, respectively. Indeed, between $Err(\boldsymbol{w}^*, b^*)$ and Err, and between $SV_+(\boldsymbol{w}^*, b^*)$ and SV, we have the following relations: $SV \subseteq SV_+(\boldsymbol{w}^*, b^*)$ and $Err \subseteq Err(\boldsymbol{w}^*, b^*)$. The paper [12] has shown that α^* is an optimal solution of $\min_\alpha F_\beta(\boldsymbol{w}^*, b^*, \alpha)$, and moreover,

$\alpha^* \in [\alpha_\beta(\boldsymbol{w}^*, b^*),\ \alpha_\beta^+(\boldsymbol{w}^*, b^*)]$. From the inequality $\alpha_\beta(\boldsymbol{w}^*, b^*) \leq \alpha^*$ and the inclusion relation (9) induced from a complementarity condition, we have $SV \subseteq SV_+(\boldsymbol{w}^*, b^*)$. Also, $Err \subseteq Err(\boldsymbol{w}^*, b^*)$ is derived, similarly.

Proposition 3.1 below implies that $1-\beta$ provides an upper and lower bound for the fraction of $Err(\boldsymbol{w}, b)$, defined by $\frac{|Err(\boldsymbol{w},b)|}{m}$, and the fraction of $SV_+(\boldsymbol{w}, b)$, $\frac{|SV_+(\boldsymbol{w},b)|}{m}$, respectively.

Proposition 3.1 *Let $(\boldsymbol{w}, b)$ be arbitrary with $\|\boldsymbol{w}\| = 1$. Then, $1-\beta$ is an upper bound on the fraction of $Err(\boldsymbol{w}, b)$, and also, a lower bound on the fraction of $SV_+(\boldsymbol{w}, b)$. The difference of these fractions is equal to a probability jump of $f(\boldsymbol{w}, b; \boldsymbol{v}_i, y_i)$, $i \in M$, at $\alpha_\beta(\boldsymbol{w}, b)$, i.e., $\frac{1}{m}|\{\ i\ :\ f(\boldsymbol{w}, b; \boldsymbol{v}_i, y_i) = \alpha_\beta(\boldsymbol{w}, b)\ \}|$.*

Proof: We apply the results of [12] to β-CVaR in the discrete distribution $\Phi(\cdot\,|\boldsymbol{w}, b)$ associated with $f(\boldsymbol{w}, b; \boldsymbol{v}_i, y_i)$, $i \in M$, of (1). An upper bound of β is provided in [12] as

$$\beta^+(\boldsymbol{w}, b) = \Phi(\alpha_\beta(\boldsymbol{w}, b)\,|\boldsymbol{w}, b) = P\{f(\boldsymbol{w}, b; \boldsymbol{v}, y) \leq \alpha_\beta(\boldsymbol{w}, b)\}.$$

Therefore, we have

$$\begin{aligned}\beta^+(\boldsymbol{w}, b) &= 1 - P\{f(\boldsymbol{w}, b; \boldsymbol{v}, y) > \alpha_\beta(\boldsymbol{w}, b)\}\\ &= 1 - \tfrac{1}{m}|\{\ i\ :\ y_i(\,\langle \boldsymbol{w}, \boldsymbol{v}_i\rangle + b\,) + \alpha_\beta(\boldsymbol{w}, b) < 0\ \}|\\ &= 1 - \tfrac{|Err(\boldsymbol{w},b)|}{m}.\end{aligned}$$

Next, we describe the left limit of $\Phi(\cdot\,|\boldsymbol{w}, b)$ at α as

$$\Phi(\alpha^-\,|\boldsymbol{w}, b) = P\{\,f(\boldsymbol{w}, b; \boldsymbol{v}, y) < \alpha\,\}.$$

A lower bound of β is shown in [12] as $\beta^-(\boldsymbol{w}, b) = \Phi(\alpha_\beta(\boldsymbol{w}, b)^-\,|\boldsymbol{w}, b)$, which is equivalent to

$$\begin{aligned}\beta^-(\boldsymbol{w}, b) &= P\{f(\boldsymbol{w}, b; \boldsymbol{v}, y) < \alpha_\beta(\boldsymbol{w}, b)\}\\ &= 1 - \tfrac{1}{m}|\{\ i\ :\ y_i(\,\langle \boldsymbol{w}, \boldsymbol{v}_i\rangle + b\,) + \alpha_\beta(\boldsymbol{w}, b) \leq 0\ \}|\\ &= 1 - \tfrac{|SV_+(\boldsymbol{w},b)|}{m}.\end{aligned}$$

The difference between $\beta^+(\boldsymbol{w}, b)$ and $\beta^-(\boldsymbol{w}, b)$ is equal to a jump of $\Phi(\cdot\,|\boldsymbol{w}, b)$ at $\alpha_\beta(\boldsymbol{w}, b)$, which is given as

$$\begin{aligned}\Phi(\alpha_\beta(\boldsymbol{w}, b)\,|\boldsymbol{w}, b) - \Phi(\alpha_\beta(\boldsymbol{w}, b)^-\,|\boldsymbol{w}, b) &= P\{f(\boldsymbol{w}, b; \boldsymbol{v}, y) = \alpha_\beta(\boldsymbol{w}, b)\}\\ &= \tfrac{1}{m}|\{\ i\ :\ y_i(\,\langle \boldsymbol{w}, \boldsymbol{v}_i\rangle + b\,) + \alpha_\beta(\boldsymbol{w}, b) = 0\ \}|.\end{aligned}$$

∎

3.2 Convex Case of β-SVC with $\beta \in [\beta_{min}, \bar{\beta})$

Problem (7) with parameter $\beta \in [\beta_{min}, \bar{\beta})$ results in Problem (6), which is proved to be equivalent to ν-SVC with $\nu = 1-\beta$. Indeed, the optimal hyperplanes $\{\boldsymbol{v}\ :\ \langle \boldsymbol{w}^*, \boldsymbol{v}\rangle + b^* = 0\}$ of β-SVC and ν-SVC are the same. There is a connection between β-SVC and C-SVC, which

is a classic formulation proposed by [4]. Let the optimal value of β-SVC (6) be ϕ^*_β and its optimal α be α^*. Note that ϕ^*_β and α^* are negative as far as β is in $[\beta_{min}, \bar{\beta})$. Then, β-SVC is equivalent to C-SVC with $C = \frac{1}{m\alpha^*\phi^*_\beta(1-\beta)}$. We confirm it by taking into account the relation between C-SVC and ν-SVC such as $C = \frac{1}{m\rho^*}$, where ρ^* is the margin of class separation in ν-SVC. The value ρ^* can be obtained by $\alpha^*\phi^*_\beta(1-\beta)$ with the use of an optimal solution α^* and the optimal value ϕ^*_β of β-SVC (6).

Now we consider the generalization performance of β-SVC (6). Let V denote the ball of radius $\tilde{B}_R$ in $\mathbb{R}^{r+1}$, *i.e.,* $V = \{\tilde{\boldsymbol{v}} \in \mathbb{R}^{r+1} \ : \ \|\tilde{\boldsymbol{v}}\| \leq \tilde{B}_R\}$, $\mathcal{F}$ be a class of real-valued functions on V defined by

$$\mathcal{F} = \{\tilde{\boldsymbol{v}} \mapsto \langle \tilde{\boldsymbol{w}}, \tilde{\boldsymbol{v}} \rangle \ : \ \|\tilde{\boldsymbol{w}}\| \leq 1, \ \tilde{\boldsymbol{v}} \in V\}, \tag{10}$$

and the margin of class separation be $\gamma > 0$. Then, a discussion on the generalization error of ν-SVC [13], which is based on the studies of [1, 16], leads to the following statement: there is a constant c such that with probability at least $1-\delta$, a decision function $h = \mathrm{sign}(g)$ with $g \in \mathcal{F}$ has test error $R[h]$ such as

$$R[h] \leq \frac{1}{m}|\{ \ i \ : \ y_i g(\boldsymbol{v}_i) < \gamma \ \}| + \sqrt{\frac{2}{m}\left(\frac{4c^2\tilde{B}_R^2}{\gamma^2}\log_2(2m) - 1 + \log\left(\frac{2}{\delta}\right)\right)}. \tag{11}$$

We apply these results to an inhomogeneous hyperplane $g(\boldsymbol{v}) = \langle \boldsymbol{w}, \boldsymbol{v} \rangle + b$, and obtain the following proposition and theorem. These statements imply the validity of β-SVC (6) in the sense that β-SVC minimizes an upper bound of test error.

Proposition 3.2 *Suppose that all the data points $\boldsymbol{v} = \Phi(\boldsymbol{x})$ in feature space live in a ball of radius B_R centered at the origin. Let $\beta \in [\beta_{min}, \bar{\beta})$, $(\boldsymbol{w}, b)$ be an arbitrary point satisfying $\|\boldsymbol{w}\| = 1$ and $\phi_\beta(\boldsymbol{w}, b) < 0$. Then, with probability at least $1-\delta$ over the training set, a decision function $h = sign(\langle \boldsymbol{w}, \boldsymbol{v} \rangle + b)$ has a probability of test error bounded according to*

$$R[h] \ \leq (1-\beta) + \sqrt{\frac{2}{m}\left(\frac{4c^2(1+B_R^2)^2}{\phi_\beta(\boldsymbol{w}, b)^2}\log_2(2m) - 1 + \log\left(\frac{2}{\delta}\right)\right)}. \tag{12}$$

Proof: To evaluate test error of $g(\boldsymbol{v}) = \langle \boldsymbol{w}, \boldsymbol{v} \rangle + b$, vectors $\boldsymbol{w}$ and $\boldsymbol{v}$ are lifted up as $\tilde{\boldsymbol{w}}^\top = (\boldsymbol{w}^\top \ b)$ and $\tilde{\boldsymbol{v}}^\top = (\boldsymbol{v}^\top \ 1)$, respectively. By rescaling $\tilde{\boldsymbol{w}}$ to length 1, we regard $\langle \boldsymbol{w}, \boldsymbol{v} \rangle + b$ as an element of $\mathcal{F}$. For the margin of class separation, $-\alpha_\beta(\boldsymbol{w}, b) > 0$, the rescaling operation yields $-\alpha_\beta(\boldsymbol{w}, b)/\sqrt{1+b^2}$ as γ of (11). Now we consider an upper bound for $|b|$. Note that any $\boldsymbol{v}$ in $\|\boldsymbol{v}\| \leq B_R$ satisfies $\langle \boldsymbol{w}, \boldsymbol{v} \rangle - B_R \leq 0$ and $\langle \boldsymbol{w}, \boldsymbol{v} \rangle + B_R \geq 0$ for all $\boldsymbol{w}$ in a unit ball. This implies that reasonable b is in the bounded interval $[-B_R, B_R]$. Therefore, we provide a bound for b as $|b| \leq B_R$, and using $Err(\boldsymbol{w}, b) = \{i : y_i(\ \langle \boldsymbol{w}, \boldsymbol{v}_i \rangle + b\ ) < -\alpha_\beta(\boldsymbol{w}, b)\}$, rewrite (11) as

$$\begin{aligned} R[h] \ &\leq \ \frac{|Err(\boldsymbol{w}, b)|}{m} + \sqrt{\frac{2}{m}\left(\frac{4c^2(1+B_R^2)^2}{\alpha_\beta(\boldsymbol{w}, b)^2}\log_2(2m) - 1 + \log\left(\frac{2}{\delta}\right)\right)}, \\ &\leq \ (1-\beta) + \sqrt{\frac{2}{m}\left(\frac{4c^2(1+B_R^2)^2}{\alpha_\beta(\boldsymbol{w}, b)^2}\log_2(2m) - 1 + \log\left(\frac{2}{\delta}\right)\right)}. \end{aligned} \tag{13}$$

Proposition 3.1 ensures that $|Err(\boldsymbol{w}, b)|/m$ is no more than $1 - \beta$. Furthermore, from the relation $\alpha_\beta(\boldsymbol{w}, b) \leq \phi_\beta(\boldsymbol{w}, b)$ for any $(\boldsymbol{w}, b)$, the desired result (12) follows. ∎

It is possible to take account of the class of functions including "$+b$": $\mathcal{F}^+ = \{f + b \ : f \in \mathcal{F}, b \in \mathbb{R}\}$ according to [16]. The resulting upper bound for $R[h]$ includes $\phi_\beta(\boldsymbol{w}, b)^2$ or $\alpha_\beta(\boldsymbol{w}, b)^2$ in the denominator as well as the above bounds of $R[h]$.

Theorem 3.3 *Let $\beta \in [\beta_{min}, \bar{\beta})$. An optimal solution $(\boldsymbol{w}, b)$ of β-SVC (6) minimizes a bound of test error in (12).*

Proof: To make the bound of $R[h]$ small with fixed β, it may be natural to find a solution $(\boldsymbol{w}, b)$ which maximizes $\phi_\beta(\boldsymbol{w}, b)^2$ in the right-hand side of (12), that is, minimizes $\phi_\beta(\boldsymbol{w}, b)$. Note that β-SVC (6) is equivalent to the problem:

$$\begin{array}{ll} \min\limits_{\boldsymbol{w}, b} & \phi_\beta(\boldsymbol{w}, b) \\ \text{subject to} & \|\boldsymbol{w}\| = 1, \quad \phi_\beta(\boldsymbol{w}, b) < 0, \end{array} \tag{14}$$

as far as $\beta \in [\beta_{min}, \bar{\beta})$. The constraints for $(\boldsymbol{w}, b)$ in the above problem coincide with the conditions in Proposition 3.2, and hence, inequality (12) holds for decision functions $h(\boldsymbol{x}) = \mathrm{sign}(\langle \boldsymbol{w}, \Phi(\boldsymbol{x}) \rangle + b)$ constructed with feasible solutions $(\boldsymbol{w}, b)$ of β-SVC (14). Among such feasible solutions, an optimal solution of β-SVC achieves a minimized bound of test error in (12). ∎

β-SVC (6) finds a function of $\mathcal{F}$ with the minimum generalization error bound. For the function $h(\boldsymbol{x}) = \mathrm{sign}(\langle \boldsymbol{w}^*, \Phi(\boldsymbol{x}) \rangle + b^*)$ constructed with an optimal solution $(\boldsymbol{w}^*, b^*)$ of (6), test error $R[h]$ is ensured to be no more than

$$(1 - \beta) + \sqrt{\frac{2}{m} \left(\frac{4c^2 (1 + B_R^2)^2}{\phi_\beta(\boldsymbol{w}^*, b^*)^2} \log_2(2m) - 1 + \log \left(\frac{2}{\delta} \right) \right)}$$

with probability at least $1 - \delta$. We see that as parameter β becomes large in $[\beta_{min}, \bar{\beta})$, the first term does small while the second one does large with $\phi_\beta(\boldsymbol{w}^*, b^*)^2 \approx 0$. From the above observation, we confirm that the hyperplane achieved via β-SVC (6) is appropriate to enhance the generalization performance.

3.3 Nonconvex Case of β-SVC with $\beta \in [\bar{\beta}, 1)$

This section deals with the nonconvex case of β-SVC whose optimal value is $\phi_\beta(\boldsymbol{w}^*, b^*) \geq 0$. The formula (11) of a generalization error bound requires positive value for γ. To regard $\alpha_\beta(\boldsymbol{w}^*, b^*)$ as γ, we exclude the special case of $\alpha_\beta(\boldsymbol{w}^*, b^*) = 0$, and divide the concerned β-SVC into two classes where either $\alpha_\beta(\boldsymbol{w}^*, b^*) < 0 \leq \phi_\beta(\boldsymbol{w}^*, b^*)$ or $0 < \alpha_\beta(\boldsymbol{w}^*, b^*) \leq \phi_\beta(\boldsymbol{w}^*, b^*)$ holds.

Firstly, suppose that β-SVC provides an optimal solution $(\boldsymbol{w}^*, b^*)$ such that $\alpha_\beta(\boldsymbol{w}^*, b^*) < 0$. Then, for arbitrary $(\boldsymbol{w}, b)$ satisfying $\|\boldsymbol{w}\| = 1$ and $\alpha_\beta(\boldsymbol{w}, b) < 0$, we estimate a generalization

error bound of (13) by following the proof of Proposition 3.2. Thus, minimizing $\alpha_\beta(\boldsymbol{w}, b)$ with respect to $(\boldsymbol{w}, b)$ with $\|\boldsymbol{w}\| = 1$ leads to minimizing the generalization error bound.

Secondly, we consider β-SVC whose optimal solution satisfies $0 < \alpha_\beta(\boldsymbol{w}^*, b^*)$. Considering $(\boldsymbol{w}, b)$ which satisfies $\|\boldsymbol{w}\| = 1$ and $\phi_\beta(\boldsymbol{w}, b) > 0$ as a feasible solution of β-SVC, we define a decision function $h = \mathrm{sign}(g)$ with $g(\boldsymbol{v}) = \langle \boldsymbol{w}, \boldsymbol{v} \rangle + b$ and estimate a probability of test error of the decision function. In this case, $-\alpha_\beta(\boldsymbol{w}, b)$ used in the definition of $Err(\boldsymbol{w}, b)$ is negative, though the margin γ in (11) is required to be positive. To resolve this issue, we prepare a decision function $\bar{h} = -\mathrm{sign}(g)$. Note that decisions via the function $\bar{h}$ are opposed to those of h. Applying (11) to the function $\bar{h}$, we have

$$\begin{aligned} R[\bar{h}] &= 1 - R[h] \\ &\leq \frac{1}{m} |\{\ i\ : -y_i(\langle \boldsymbol{w}, \boldsymbol{v} \rangle + b) < \alpha_\beta(\boldsymbol{w}, b)\ \}| + \sqrt{\frac{2}{m}\left(\frac{4c^2(1+B_R^2)^2}{\alpha_\beta(\boldsymbol{w}, b)^2}\log_2(2m) - 1 + \log\left(\frac{2}{\delta}\right)\right)}, \end{aligned}$$

where $\alpha_\beta(\boldsymbol{w}, b)/\sqrt{1+b^2} > 0$ corresponds to γ in (11). Since

$$1 - \frac{1}{m}|\{\ i\ :\ -y_i(\langle \boldsymbol{w}, \boldsymbol{v} \rangle + b) < \alpha_\beta(\boldsymbol{w}, b)\ \}| = \frac{|SV_+(\boldsymbol{w}, b)|}{m} \geq 1 - \beta,$$

we have

$$R[h] \geq (1-\beta) - \sqrt{\frac{2}{m}\left(\frac{4c^2(1+B_R^2)^2}{\alpha_\beta(\boldsymbol{w}, b)^2}\log_2(2m) - 1 + \log\left(\frac{2}{\delta}\right)\right)}.$$

To depress the lower bound of $R[h]$, it is reasonable to find a solution $(\boldsymbol{w}, b)$ which minimizes $\alpha_\beta(\boldsymbol{w}, b)$ under the constraints $\|\boldsymbol{w}\| = 1$ as far as $\alpha_\beta(\boldsymbol{w}, b) > 0$.

From the above discussion, we see that minimizing a generalization error bound is closely related to minimizing β-VaR, $\alpha_\beta(\boldsymbol{w}, b)$, in any β-SVC problem with $\alpha_\beta(\boldsymbol{w}^*, b^*) < 0$ or $\alpha_\beta(\boldsymbol{w}^*, b^*) > 0$. VaR is, however, known to be difficult to optimize. Even when calculated using scenarios, VaR is nonconvex and nonsmooth as a function of positions. By contrast, minimizing CVaR is easy compared to minimizing VaR. Since CVaR is greater than or equal to VaR, small CVaR ensures small VaR. Therefore, we expect that nonconvex β-SVC also yields a reasonable decision function which depresses an upper bound or a lower bound for the generalization error.

4 CVaR Minimization for a Regression Problem

In the previous sections, we have shown that the β-SVC induced from minimizing β-CVaR is deeply related to (Extended) ν-SVC. Now we introduce β-CVaR into regression and consider the relation between the resulting β-CVaR minimization, β-SVR, and ν-SVR [13].

The main issue of a linear regression problem is to estimate a linear function of an input vector $\boldsymbol{x}$, $h(\boldsymbol{x}) = \langle \boldsymbol{w}^*, \boldsymbol{x} \rangle + b^*$, so that, for a new data point $\bar{\boldsymbol{x}}$, the function value $h(\bar{\boldsymbol{x}})$ predicts well its actual response value. The parameter values $(\boldsymbol{w}^*, b^*)$ are selected based on the set

of m training data, $(\boldsymbol{x}_i, y_i), i \in M$, which consists of input examples $\boldsymbol{x}_i \in \mathbb{R}^n$ and response values $y_i \in \mathbb{R}$, $i \in M$. Now, regarding

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

as a cost function for a given parameter $C > 0$, we consider the distribution of $f(\boldsymbol{w}, b; \boldsymbol{x}_i, y_i)$, $\forall i \in M$. Then, the CVaR minimization problem for a regression problem, β-SVR, is formulated as

$$\min_{\boldsymbol{w}, b, \alpha} \quad \alpha + \frac{1}{1-\beta}\boldsymbol{E}\left[\, [f(\boldsymbol{w}, b; \boldsymbol{x}, y) - \alpha]^+ \,\right], \tag{15}$$

according to (2). We define a regressor $h(\boldsymbol{x}) = \langle \frac{C\boldsymbol{w}^*}{\|\boldsymbol{w}^*\|}, \boldsymbol{x} \rangle + \frac{Cb^*}{\|\boldsymbol{w}^*\|}$ by using the resulting optimal solution $(\boldsymbol{w}^*, b^*)$. The function $h(\boldsymbol{x})$ has the normal vector $\frac{C\boldsymbol{w}^*}{\|\boldsymbol{w}^*\|}$ whose norm is restricted to C.

The following proposition exhibits an equivalent problem to β-SVR (15). The problem is also regarded as a CVaR minimization problem for a cost function $f(\boldsymbol{w}, b; \boldsymbol{x}, y) = |y - \langle \boldsymbol{w}, \boldsymbol{x} \rangle - b|$ under an additional restriction $\|\boldsymbol{w}\| = C$.

Proposition 4.1 *The CVaR minimization problem β-SVR (15) is equivalent to*

$$\begin{array}{lll} \min\limits_{\boldsymbol{w}, b, \alpha, \boldsymbol{z}} & \alpha + \dfrac{1}{(1-\beta)m}\displaystyle\sum_{i \in M} z_i & \\ \textit{subject to} & z_i - |y_i - \langle \boldsymbol{w}, \boldsymbol{x}_i \rangle - b| + \alpha \geq 0, \quad i \in M, \\ & z_i \geq 0, \quad i \in M, \\ & \boldsymbol{w}^\top \boldsymbol{w} = C^2. \end{array} \tag{16}$$

in the following sense:

- *An optimal solution $(\bar{\boldsymbol{w}}, \bar{b}, \bar{\alpha}, \bar{\boldsymbol{z}})$ of (16) provides that of (15) by $(\eta\bar{\boldsymbol{w}}, \eta\bar{b}, \bar{\alpha})$, $\eta > 0$.*
- *An optimal solution $(\boldsymbol{w}^*, b^*, \alpha^*)$ of (15) provides that of (16) by $(C\boldsymbol{w}^*/\|\boldsymbol{w}^*\|, Cb^*/\|\boldsymbol{w}^*\|, \alpha^*, \boldsymbol{z}^*)$, where $z_i^* = [\ |y_i - C\frac{\langle \boldsymbol{w}^*, \boldsymbol{x}_i \rangle + b^*}{\|\boldsymbol{w}^*\|}| - \alpha^*]^+$.*

Proof: Only the first statement is shown, since the second one can be proved in the similar way. Suppose on the contrary that $(\eta\bar{\boldsymbol{w}}, \eta\bar{b}, \bar{\alpha})$ is not an optimal solution of (15), that is,

$$\begin{array}{rl} & \bar{\alpha} + \frac{1}{(1-\beta)m}\sum_{i \in M}\left[\ \left| y_i - C\frac{\langle \bar{\boldsymbol{w}}, \boldsymbol{x}_i \rangle + \bar{b}}{\|\bar{\boldsymbol{w}}\|} \right| - \bar{\alpha}\right]^+ \\ > & \alpha^* + \frac{1}{(1-\beta)m}\sum_{i \in M}\left[\ \left| y_i - C\frac{\langle \boldsymbol{w}^*, \boldsymbol{x}_i \rangle + b^*}{\|\boldsymbol{w}^*\|} \right| - \alpha^*\right]^+ \end{array}$$

This strict inequality implies that a feasible solution $(\frac{C\boldsymbol{w}^*}{\|\boldsymbol{w}^*\|}, \frac{Cb^*}{\|\boldsymbol{w}^*\|}, \alpha^*, \boldsymbol{z}^*)$ of (16) yields less objective function value than the optimal solution $(\bar{\boldsymbol{w}}, \bar{b}, \bar{\alpha}, \bar{\boldsymbol{z}})$ of (16). It is a contradiction to the optimality of $(\bar{\boldsymbol{w}}, \bar{b}, \bar{\alpha}, \bar{\boldsymbol{z}})$. ∎

Note that a constraint $\alpha \geq 0$ is redundant for (16), though ν-SVR includes it. Since $f(\boldsymbol{w}, b; \boldsymbol{x}, y)$ is always nonnegative, β-VaR must be nonnegative. And hence, $\bar{\alpha} \geq 0$ holds at

the optimality in (16) without the constraint $\alpha \geq 0$. From the same reason, the constraint can be deleted from the formulation of ν-SVR.

The proposition implies that the regressor $h(\boldsymbol{x}) = \langle \bar{\boldsymbol{w}}, \boldsymbol{x} \rangle + \bar{b}$, which is constructed with an optimal solution $(\bar{\boldsymbol{w}}, \bar{b}, \bar{\alpha}, \bar{\boldsymbol{z}})$ of (16), is equal to $h(\boldsymbol{x}) = \langle \frac{C\boldsymbol{w}^*}{\|\boldsymbol{w}^*\|}, \boldsymbol{x} \rangle + \frac{Cb^*}{\|\boldsymbol{w}^*\|}$ of Problem (15). Therefore, we consider Problems (15) and (16) as β-SVR. For any positive value of C, β-SVR (16) has an optimal solution because of the feasibility and boundedness of (16). Problem (16) is a nonconvex QP problem with a nonconvex constraint $\boldsymbol{w}^\top \boldsymbol{w} = C^2$, and difficult to find a global optimal solution in general. Similar to β-SVC, β-SVR of (16) is converted to a convex problem or nonconvex one depending on the parameter C.

4.1 Convex Case of β-SVR with $C \leq C_\beta$

Let $\widetilde{\boldsymbol{w}}_\beta$ be an optimal solution of a *linear programming* (LP) problem, constructed by deleting a nonconvex quadratic constraint $\boldsymbol{w}^\top \boldsymbol{w} = C^2$ from (16). Then, the threshold $C_\beta := \|\widetilde{\boldsymbol{w}}_\beta\|$ divides β-SVR (16) into two cases: for any $\beta \in (0, 1)$, β-SVR (16) with $C \leq C_\beta$ is transformed into a convex problem and β-SVR with $C > C_\beta$ is essentially nonconvex one.

Proposition 4.2 *Let $\beta \in (0, 1)$. Suppose that the LP, which is constructed by deleting a nonconvex quadratic constraint $\boldsymbol{w}^\top \boldsymbol{w} = C^2$ from (16), has a unique optimal solution $(\widetilde{\boldsymbol{w}}_\beta, \widetilde{b}_\beta, \widetilde{\alpha}_\beta, \widetilde{\boldsymbol{z}}_\beta)$. When $C \leq C_\beta := \|\widetilde{\boldsymbol{w}}_\beta\|$, β-SVR (16) is reformulated as a convex problem:*

$$\begin{array}{ll} \min\limits_{\boldsymbol{w}, b, \alpha, \boldsymbol{z}} & \alpha + \dfrac{1}{(1-\beta)m} \displaystyle\sum_{i \in M} z_i \\ \textit{subject to} & z_i - |y_i - \langle \boldsymbol{w}, \boldsymbol{x}_i \rangle - b| + \alpha \geq 0, \quad i \in M, \\ & z_i \geq 0, \quad i \in M, \\ & \boldsymbol{w}^\top \boldsymbol{w} \leq C^2. \end{array} \tag{17}$$

Proof: When $C = C_\beta$, the optimal solution of (17) is $\widetilde{\boldsymbol{w}}_\beta$. It certainly satisfies $\boldsymbol{w}^\top \boldsymbol{w} = C^2$, and thus, convex β-SVR (17) yields an optimal solution of β-SVR (16).
Now we consider $C < C_\beta$. Let $(\boldsymbol{w}^*, b^*, \alpha^*, \boldsymbol{z}^*)$ be an optimal solution of (17). If $\boldsymbol{w}^{*\top} \boldsymbol{w}^* < C^2$ holds at optimality in the case of $C < C_\beta$, there exists a feasible solution

$$\begin{pmatrix} \bar{\boldsymbol{w}} \\ \bar{b} \\ \bar{\alpha} \\ \bar{\boldsymbol{z}} \end{pmatrix} = t \begin{pmatrix} \boldsymbol{w}^* \\ b^* \\ \alpha^* \\ \boldsymbol{z}^* \end{pmatrix} + (1-t) \begin{pmatrix} \widetilde{\boldsymbol{w}}_\beta \\ \widetilde{b}_\beta \\ \widetilde{\alpha}_\beta \\ \widetilde{\boldsymbol{z}}_\beta \end{pmatrix}, \quad 0 < t < 1,$$

satisfying $\bar{\boldsymbol{w}}^\top \bar{\boldsymbol{w}} = C^2$. The existence of the feasible solution contradicts the optimality of (17), since

$$\bar{\alpha} + \frac{1}{(1-\beta)m} \sum_{i \in M} \bar{z}_i < \alpha^* + \frac{1}{(1-\beta)m} \sum_{i \in M} z_i^*$$

holds. ∎

The convex case of β-SVR (17) is closely related to ν-SVR [13]. To show the relation, we investigate Problem (17) from a dual viewpoint. For the constraints of (17), we introduce KKT multipliers $\lambda_i^{(1)}, \lambda_i^{(2)}, \eta \geq 0$, and obtain KKT conditions. Suppose that $\lambda_i^{(1)}$ corresponds to the constraint $z_i - (y_i - \langle \boldsymbol{w}, \boldsymbol{x}_i \rangle - b) + \alpha \geq 0$ and $\lambda_i^{(2)}$ does to $z_i + (y_i - \langle \boldsymbol{w}, \boldsymbol{x}_i \rangle - b) + \alpha \geq 0$. Then, the KKT conditions of (17) include

$$\eta \boldsymbol{w}^* = \sum_{i \in M} (\lambda_i^{(1)} - \lambda_i^{(2)}) \boldsymbol{x}_i.$$

The multiplier η corresponds to the constraint $\boldsymbol{w}^\top \boldsymbol{w} \leq C^2$. The Lagrange dual function results in

$$\mathcal{L}(\boldsymbol{\lambda}^{(1)}, \boldsymbol{\lambda}^{(2)}, \eta) = \sum_{i \in M} (\lambda_i^{(1)} - \lambda_i^{(2)}) y_i - \sum_{i \in M} (\lambda_i^{(1)} - \lambda_i^{(2)}) \langle \boldsymbol{w}^*, \boldsymbol{x}_i \rangle + \frac{1}{2} \eta (\boldsymbol{w}^{*\top} \boldsymbol{w}^* - C^2),$$

which is restated in the case of $\eta > 0$ by

$$\sum_{i \in M} (\lambda_i^{(1)} - \lambda_i^{(2)}) y_i - \frac{1}{2\eta} \sum_{i,j \in M} (\lambda_i^{(1)} - \lambda_i^{(2)})(\lambda_j^{(1)} - \lambda_j^{(2)}) \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle - \frac{1}{2} \eta C^2 \tag{18}$$

and in the case of $\eta = 0$ by

$$\sum_{i \in M} (\lambda_i^{(1)} - \lambda_i^{(2)}) y_i.$$

Minimizing the function (18) with respect to η leads to the optimal solution:

$$\eta^* = \frac{1}{C} \sqrt{\sum_{i,j \in M} (\lambda_i^{(1)} - \lambda_i^{(2)})(\lambda_j^{(1)} - \lambda_j^{(2)}) \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle}, \tag{19}$$

and thus, the dual problem of (17) is described as

$$\begin{array}{ll} \max\limits_{\boldsymbol{\lambda}^{(1)}, \boldsymbol{\lambda}^{(2)}} & \sum\limits_{i \in M} (\lambda_i^{(1)} - \lambda_i^{(2)}) y_i - C \sqrt{\sum\limits_{i,j \in M} (\lambda_i^{(1)} - \lambda_i^{(2)})(\lambda_j^{(1)} - \lambda_j^{(2)}) \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle} \\ \text{subject to} & \sum_{i \in M} (\lambda_i^{(1)} - \lambda_i^{(2)}) = 0, \\ & \sum_{i \in M} (\lambda_i^{(1)} + \lambda_i^{(2)}) = 1, \\ & 0 \leq \lambda_i^{(1,2)} \leq \frac{1}{(1-\beta)m}, \ i \in M. \end{array} \tag{20}$$

Note that ν-SVR of [13] is given as

$$\begin{array}{ll} \min\limits_{\boldsymbol{w}, b, \alpha, \boldsymbol{z}} & \frac{1}{2} \|\boldsymbol{w}\|^2 + \widehat{C} \left(\nu \alpha + \frac{1}{m} \sum\limits_{i \in M} z_i \right) \\ \text{subject to} & z_i - |y_i - \langle \boldsymbol{w}, \boldsymbol{x}_i \rangle - b| + \alpha \geq 0, \quad i \in M, \\ & z_i \geq 0, \quad i \in M, \end{array} \tag{21}$$

without a redundant constraint $\alpha \geq 0$. Then, its dual problem is formulated as

$$\begin{array}{ll} \max\limits_{\boldsymbol{\lambda}^{(1)},\boldsymbol{\lambda}^{(2)}} & \sum\limits_{i\in M}(\lambda_i^{(1)}-\lambda_i^{(2)})y_i - \dfrac{\widehat{C}\nu}{2}\sum\limits_{i,j\in M}(\lambda_i^{(1)}-\lambda_i^{(2)})(\lambda_j^{(1)}-\lambda_j^{(2)})\langle \boldsymbol{x}_i,\boldsymbol{x}_j\rangle \\ \text{subject to} & \sum_{i\in M}(\lambda_i^{(1)}-\lambda_i^{(2)})=0, \\ & \sum_{i\in M}(\lambda_i^{(1)}+\lambda_i^{(2)})=1, \\ & 0\leq \lambda_i^{(1,2)}\leq \frac{1}{\nu m},\ i\in M. \end{array} \tag{22}$$

Problem (22) seems to be different from the standard description of the dual problem of ν-SVR, but by scaling variables as $(\boldsymbol{\lambda}^{(1)},\boldsymbol{\lambda}^{(2)})=\frac{1}{\widehat{C}\nu}(\boldsymbol{\mu}^{(1)},\boldsymbol{\mu}^{(2)})$ and multiplying a constant $(\widehat{C}\nu)$ to the objective function in (22), we have the standard description of ν-SVR with variables $(\boldsymbol{\mu}^{(1)},\boldsymbol{\mu}^{(2)})$. Therefore, between the optimal solutions of the primal and dual problems, the relation $\bar{\boldsymbol{w}}=\sum_{i\in M}(\bar{\mu}_i^{(1)}-\bar{\mu}_i^{(2)})\boldsymbol{x}_i=(\widehat{C}\nu)\sum_{i\in M}(\bar{\lambda}_i^{(1)}-\bar{\lambda}_i^{(2)})\boldsymbol{x}_i$ holds.

Theorem 4.3 *Let $\nu=1-\beta\in(0,1)$, and $(\bar{\boldsymbol{\lambda}}^{(1)},\bar{\boldsymbol{\lambda}}^{(2)})$ be an optimal solution of ν-SVR (22) with parameter $\widehat{C}>0$. Then, the solution is also optimal in convex β-SVR (20) with the parameter*

$$C=(\widehat{C}\nu)\sqrt{\sum_{i,j\in M}(\bar{\lambda}_i^{(1)}-\bar{\lambda}_i^{(2)})(\bar{\lambda}_j^{(1)}-\bar{\lambda}_j^{(2)})\langle \boldsymbol{x}_i,\boldsymbol{x}_j\rangle}.$$

Proof: Let $(\boldsymbol{\lambda}^{(1)*},\boldsymbol{\lambda}^{(2)*})$ be an optimal solution of (20). For simplicity of notation, we use $\varphi_\nu(\boldsymbol{\lambda}^{(1)},\boldsymbol{\lambda}^{(2)})$ as the objective function of (22) and $\psi_\beta(\boldsymbol{\lambda}^{(1)},\boldsymbol{\lambda}^{(2)})$ as that of (20). Then, $\varphi_\nu(\bar{\boldsymbol{\lambda}}^{(1)},\bar{\boldsymbol{\lambda}}^{(2)})$ indicates the optimal value of (22) and $\psi_\beta(\boldsymbol{\lambda}^{(1)*},\boldsymbol{\lambda}^{(2)*})$ does that of (20). Since $(\bar{\boldsymbol{\lambda}}^{(1)},\bar{\boldsymbol{\lambda}}^{(2)})$ is a feasible solution for the problem (20), we have

$$\begin{array}{rl} & \psi_\beta(\boldsymbol{\lambda}^{(1)*},\boldsymbol{\lambda}^{(2)*}) \\ \geq & \sum\limits_{i\in M}(\bar{\lambda}_i^{(1)}-\bar{\lambda}_i^{(2)})y_i-(\widehat{C}\nu)\sum\limits_{i,j\in M}(\bar{\lambda}_i^{(1)}-\bar{\lambda}_i^{(2)})(\bar{\lambda}_j^{(1)}-\bar{\lambda}_j^{(2)})\langle \boldsymbol{x}_i,\boldsymbol{x}_j\rangle, \\ = & \varphi_\nu(\bar{\boldsymbol{\lambda}}^{(1)},\bar{\boldsymbol{\lambda}}^{(2)})-\dfrac{1}{2\widehat{C}\nu}C^2, \end{array} \tag{23}$$

which implies that $\psi_\beta(\boldsymbol{\lambda}^{(1)*},\boldsymbol{\lambda}^{(2)*})\geq\varphi_\nu(\boldsymbol{\lambda}^{(1)},\boldsymbol{\lambda}^{(2)})-\frac{1}{2\widehat{C}\nu}C^2$ holds for any feasible solution $(\boldsymbol{\lambda}^{(1)},\boldsymbol{\lambda}^{(2)})$ of ν-SVR (22). Moreover, when $(\bar{\boldsymbol{\lambda}}^{(1)},\bar{\boldsymbol{\lambda}}^{(2)})=(\boldsymbol{\lambda}^{(1)*},\boldsymbol{\lambda}^{(2)*})$, the equality holds in (23). Indeed, with the use of $C=(\widehat{C}\nu)\|\sum_{i\in M}(\bar{\lambda}_i^{(1)}-\bar{\lambda}_i^{(2)})\boldsymbol{x}_i\|$, $\psi_\beta(\boldsymbol{\lambda}^{(1)*},\boldsymbol{\lambda}^{(2)*})$ is described as

$$\begin{array}{r} \sum_{i\in M}(\lambda_i^{(1)*}-\lambda_i^{(2)*})y_i-(\widehat{C}\nu)\sqrt{\sum_{i,j\in M}(\bar{\lambda}_i^{(1)}-\bar{\lambda}_i^{(2)})(\bar{\lambda}_j^{(1)}-\bar{\lambda}_j^{(2)})\langle \boldsymbol{x}_i,\boldsymbol{x}_j\rangle} \\ \times\sqrt{\sum_{i,j\in M}(\lambda_i^{(1)*}-\lambda_i^{(2)*})(\lambda_j^{(1)*}-\lambda_j^{(2)*})\langle \boldsymbol{x}_i,\boldsymbol{x}_j\rangle}, \end{array}$$

which is equal to $\varphi_\nu(\boldsymbol{\lambda}^{(1)*},\boldsymbol{\lambda}^{(2)*})-\frac{1}{2\widehat{C}\nu}C^2$ under the condition that $(\boldsymbol{\lambda}^{(1)*},\boldsymbol{\lambda}^{(2)*})=(\bar{\boldsymbol{\lambda}}^{(1)},\bar{\boldsymbol{\lambda}}^{(2)})$. From above, we see that $(\boldsymbol{\lambda}^{(1)*},\boldsymbol{\lambda}^{(2)*})=(\bar{\boldsymbol{\lambda}}^{(1)},\bar{\boldsymbol{\lambda}}^{(2)})$ is an optimal solution of Problems (20) and (22). ∎

The above theorem is paraphrased from the primal point of view: the optimal solution $(\bar{\boldsymbol{w}}, \bar{b}, \bar{\alpha}, \bar{\boldsymbol{z}})$ of ν-SVR (21) is optimal in convex β-SVR (17) under the parameter $C = \|\bar{\boldsymbol{w}}\|$. Note that it is also optimal in the nonconvex β-SVR (16) because of $\bar{\boldsymbol{w}}^\top \bar{\boldsymbol{w}} = C^2$. Since (19) implies $\eta^* = \frac{1}{\widehat{C}\nu}$, the optimal solution $\boldsymbol{w}^* = \frac{1}{\eta^*} \sum_{i \in M} (\lambda_i^{(1)*} - \lambda_i^{(2)*}) \boldsymbol{x}_i$ of (17) is equal to $\bar{\boldsymbol{w}}$. Then, $\boldsymbol{w}^{*\top} \boldsymbol{w}^* = \bar{\boldsymbol{w}}^\top \bar{\boldsymbol{w}} = C^2$ holds.

The next theorem shows how to set the parameter $\widehat{C}$ of ν-SVR (22) so that (22) yields the same optimal solution as (20). We can prove it in the similar manner with the proof of Theorem 4.3.

Theorem 4.4 *Let $\nu = 1 - \beta \in (0, 1)$, $(\boldsymbol{\lambda}^{(1)*}, \boldsymbol{\lambda}^{(2)*})$ be an optimal solution of convex β-SVR (20) with parameter $C > 0$, and $\eta^* = \frac{1}{C} \sqrt{\sum_{i,j \in M} (\lambda_i^{(1)*} - \lambda_i^{(2)*})(\lambda_j^{(1)*} - \lambda_j^{(2)*}) \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle}$. When $\eta^* > 0$, the solution $(\boldsymbol{\lambda}^{(1)*}, \boldsymbol{\lambda}^{(2)*})$ is also optimal in ν-SVR (22) with parameter $\widehat{C} = \frac{1}{\nu \eta^*}$.*

From these two theorems, ν-SVR as well as β-SVR can be interpreted in the viewpoint of minimizing β-CVaR for the distribution of $f(\boldsymbol{w}, b; \boldsymbol{x}, y) = \left| y - C \frac{\langle \boldsymbol{w}, \boldsymbol{x} \rangle + b}{\|\boldsymbol{w}\|} \right|$. Since convex β-SVR (20) and ν-SVR (22) are essentially the same, it may not worth solving convex β-SVR. But we can solve it as a second order cone programming problem, and also generalize convex β-SVR (20) to a nonlinear variant by substituting a kernel $k(\boldsymbol{x}_i, \boldsymbol{x}_j)$ for the dot product $\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle$.

4.2 Nonconvex Case of β-SVR with $C > C_\beta$

When C exceeds the threshold C_β, convex β-SVR (17) yields an optimal solution $\boldsymbol{w}^*$ which satisfies $\boldsymbol{w}^{*\top} \boldsymbol{w}^* < C^2$. The solution is no longer an optimal solution of β-SVR (16). Since Problem (16) with $C > C_\beta$ is essentially a nonconvex problem, we need to deal with the nonconvex problem with a nonconvex quadratic constraint $\boldsymbol{w}^\top \boldsymbol{w} = C^2$.

Problem (16) of nonconvex β-SVR is referred to an LRCP as well as (7) of β-SVC. For the LRCP (7) of β-SVC, a local search algorithm [8] or Extended ν-SVC algorithm [11] finds a local minimizer of (7). To apply those local search algorithms to nonconvex β-SVR (16), however, some modifications are necessary in their algorithms. In this paper, we avoid the discussion on efficient algorithms for β-SVR (16), and show a numerical example obtained by applying a sequential quadratic programming (SQP) method, one of the most successful general methods for local optimization of nonlinear constrained optimization problems, to (16). To investigate nonconvex β-SVR experimentally, an efficient algorithm specialized for an LRCP of β-SVR is necessary.

ν-SVR can not deal with the parameter value $\widehat{C}$ corresponding to $C > C_\beta$ of nonconvex β-SVR. Hence, there is a chance to find a good regressor with β-SVR. Indeed, for the toy example shown in Section 5, β-SVR with a linear kernel provides a good regressor. When RBF kernel is applied to β-SVR, however, the threshold C_β tends to be large, and the best performance be achieved with convex case of β-SVR with $C \leq C_\beta$.

4.3 Generalization Performance of β-SVR

It is possible to extend the discussion on the generalization performance in classification to regression, according to [2, 7]. For the purpose, we consider the high-dimensional feature space where $\Phi(\boldsymbol{x}_i) = \boldsymbol{v}_i \in \mathbb{R}^r$, $i \in M$, exist. For a KKT point of nonconvex β-SVR (16), we can show the β property: $\frac{|Err|}{m} \leq 1-\beta \leq \frac{|SV|}{m}$, and estimate the difference of those bounds as $\frac{|SV|-|Err|}{m} \leq \frac{r+1}{m}$.

For CVaR minimization (15) or equivalently (16) in regression, a similar statement to Proposition 3.1 is shown: suppose that $\alpha_\beta(\boldsymbol{w}, b)$ is β-VaR for the distribution of $f(\boldsymbol{w}, b; \boldsymbol{v}_i, y_i)$, $i \in M$, and define $Err(\boldsymbol{w}, b) = \{\ i\ :\ -|y_i - \langle \boldsymbol{w}, \boldsymbol{v}_i\rangle - b| + \alpha_\beta(\boldsymbol{w}, b) < 0\}$ and $SV_+(\boldsymbol{w}, b) = \{\ i\ :\ -|y_i - \langle \boldsymbol{w}, \boldsymbol{v}_i\rangle - b| + \alpha_\beta(\boldsymbol{w}, b) \leq 0\}$. Then, for arbitrary $(\boldsymbol{w}, b)$ satisfying $\|\boldsymbol{w}\| = C$, the inequalities $\frac{|Err(\boldsymbol{w},b)|}{m} \leq 1-\beta \leq \frac{|SV_+(\boldsymbol{w},b)|}{m}$ hold.

Proposition 4.5 *Suppose that all the data points $\boldsymbol{v} = \Phi(\boldsymbol{x})$ in feature space live in a ball of radius B_R centered at the origin, and all y satisfy $|y| \leq y_{max}$. Let $\beta \in (0,1)$, $C > 0$, $(\boldsymbol{w}, b)$ be arbitrary with $\|\boldsymbol{w}\| = C$, and θ be a threshold for test error such as $\theta > \phi_\beta(\boldsymbol{w}, b)$. Then, with probability at least $1-\delta$ over the training set, the function $g(\boldsymbol{v}) = \langle \boldsymbol{w}, \boldsymbol{v}\rangle + b$ has a probability $R_\theta[g]$ of $|y - g(\boldsymbol{v})| > \theta$ for any $(\boldsymbol{v}, y)$ bounded by*

$$R_\theta[g] \ \leq (1-\beta) + \sqrt{\frac{2}{m}\left(\frac{4c^2\tilde{B}_R^2\{C^2 + (y_{max} + CB_R)^2 + 1\}}{(\theta - \phi_\beta(\boldsymbol{w}, b))^2}\log_2(2m) - 1 + \log\left(\frac{2}{\delta}\right)\right)}, \tag{24}$$

where $\tilde{B}_R = \sqrt{B_R^2 + (y_{max} + \theta)^2 + 1}$ is a constant.

Proof: The statement is proved by the inequality (11) of the generalization error for classification. To make the CVaR minimization problem (16) of regression fit to the setting of (11), we consider the distribution of $\theta - |y_i - g(\boldsymbol{v}_i)|$, $i \in M$ and the threshold $\theta - \alpha_\beta(\boldsymbol{w}, b)$ for training error. Then, as the fraction of margin errors, we have

$$\frac{1}{m}|\{\ i\ :\ \theta - |y_i - \langle \boldsymbol{w}, \boldsymbol{v}_i\rangle - b| < \theta - \alpha_\beta(\boldsymbol{w}, b)\}| = \frac{|Err(\boldsymbol{w}, b)|}{m} \leq 1-\beta.$$

The function $\theta - |y - \langle \boldsymbol{w}, \boldsymbol{v}\rangle - b|$ is described as $\langle \tilde{\boldsymbol{w}}, \tilde{\boldsymbol{v}}^{(1)}\rangle$ or $\langle \tilde{\boldsymbol{w}}, \tilde{\boldsymbol{v}}^{(2)}\rangle$ using

$$\tilde{\boldsymbol{w}} = \begin{pmatrix} \boldsymbol{w} \\ b \\ 1 \end{pmatrix}, \quad \tilde{\boldsymbol{v}}^{(1)} = \begin{pmatrix} \boldsymbol{v} \\ 1 \\ \theta - y \end{pmatrix}, \quad \tilde{\boldsymbol{v}}^{(2)} = \begin{pmatrix} -\boldsymbol{v} \\ -1 \\ \theta + y \end{pmatrix}.$$

The function is scaled due to the condition $\|\tilde{\boldsymbol{w}}\| \leq 1$ of (10), and then, the margin results in

$$\gamma = \frac{\theta - \alpha_\beta(\boldsymbol{w}, b)}{\sqrt{C^2 + (y_{max} + CB_R)^2 + 1}}.$$

As shown in the proof of Proposition 3.2, the reasonable b is in a bounded interval

$$\left[\min_{(\boldsymbol{w}:\|\boldsymbol{w}\|=C, \boldsymbol{v}, y)} y - \langle \boldsymbol{w}, \boldsymbol{v}\rangle, \quad \max_{(\boldsymbol{w}:\|\boldsymbol{w}\|=C, \boldsymbol{v}, y)} y - \langle \boldsymbol{w}, \boldsymbol{v}\rangle\right].$$

The interval is included in the range of b given by $|b| \leq y_{max} + CB_R$. Moreover, the radius of a ball including all $\tilde{\boldsymbol{v}}^{(1)}$ and $\tilde{\boldsymbol{v}}^{(2)}$ is computed as $\tilde{B}_R = \sqrt{B_R^2 + (y_{max} + \theta)^2 + 1}$. As a result, a generalization error bound is obtained as

$$(1-\beta) + \sqrt{\frac{2}{m}\left(\frac{4c^2\tilde{B}_R^2\{C^2 + (y_{max} + CB_R)^2 + 1\}}{(\theta - \alpha_\beta(\boldsymbol{w}, b))^2}\log_2(2m) - 1 + \log\left(\frac{2}{\delta}\right)\right)},$$

which is also bounded by (24) because of $\alpha_\beta(\boldsymbol{w}, b) \leq \phi_\beta(\boldsymbol{w}, b)$. ∎

Now we state without proof that an optimal solution of β-SVR depresses an upper bound of the generalization error.

Theorem 4.6 *Let $\beta \in (0, 1)$. An optimal solution $(\boldsymbol{w}, b)$ of β-SVR minimizes a bound of test error in (24).*

The problem of β-SVR minimizes $\phi_\beta(\boldsymbol{w}, b)$ under the constraint $\|\boldsymbol{w}\| = C$. Therefore, with an optimal solution of β-SVR, the bound of (24) is minimized. Note that the bound (24) includes two parameters C and β. Is there any nice choice for the values of C and β to minimize the bound? For small β, the optimal value of β-SVR, $\phi_\beta(\boldsymbol{w}^*, b^*)$, becomes small, and thus, the second term in (24) is also small. However, the training error term $(1 - \beta)$ becomes large. It is necessary to adjust the parameter β to achieve a highly accurate regressor. It is the same with parameter C. As far as $C \in (0, C_\beta]$, an optimal solution of β-SVR is achievable via convex β-SVR (17). For larger C, the feasible region in (17) becomes larger, and as a result, the optimal value of (17) becomes small. The small optimal value $\phi_\beta(\boldsymbol{w}^*, b^*)$ makes the denominator in (24) large, while larger C makes the numerator large. Therefore, it is necessary to find fitting parameter values for β and C.

5 Numerical Results

The classification models of β-SVC and Extended ν-SVC of [11] are essentially the same when linear kernel is applied. In the first subsection, we show the results of β-SVC with polynomial kernel for one benchmark problem of classification. We need more investigation to choose a kernel function such that the performance of the estimator on unseen data is high, and hence, just show an example which implies the effectiveness of kernelized β-SVC. As stated before, it may be better to use β-SVC after the kernel matrix is learned with semi-definite programming.

In the next subsection, we deal with regression problems: a toy example and one benchmark problem. As far as the parameter C of β-SVR is in the range $(0, C_\beta]$, we obtain identical prediction results from β-SVR and ν-SVR by selecting the corresponding parameter values for C of β-SVR and $\widehat{C}$ of ν-SVR. Using a toy example, we investigate nonconvex β-SVR with $C > C_\beta$. And then, we restrict ourselves to illustrating the corresponding parameters C and $\widehat{C}$ for a benchmark problem.

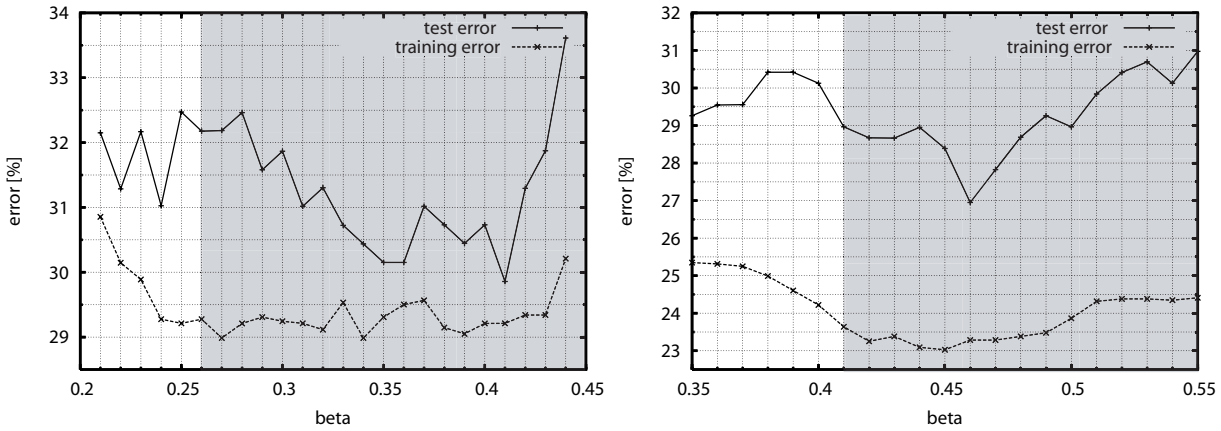

Figure 2: Training and test error rates of liver-disorders with a linear kernel (left) or a polynomial kernel (right).

5.1 Classification

We test the performance of β-SVC on liver-disorders dataset. The dataset is available from the UCI Machine Learning Data Repository [3]. It is shown in [11] that Extended ν-SVC with linear kernel nicely worked for the dataset. Here we show that β-SVC also works nicely when applying a polynomial kernel.

We used SeDuMi [14] software to solve a convex QP problem of convex β-SVC, and LPs successively induced from nonconvex β-SVC in Extended ν-SVC algorithm [11]. The SeDuMi solver is a MATLAB implemented interior-point method for optimization over symmetric cones. All computations were conducted on an Opteron 850 (2.4GHz) with 8GB of physical memory.

The liver-disorders dataset has 345 examples with 6 attributes. We performed ten-fold cross-validation for the dataset and checked the test error rates. As the polynomial kernel, $k(\boldsymbol{x}_i, \boldsymbol{x}_j) = (1/n \times \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle)^2$ is used. For liver-disorders, β-SVC achieved small test error with nonconvex β value in the both cases: linear and polynomial kernels. Both figures in Figure 2 show training error and test error rates obtained by β-SVC with different β. Each plot indicates the average error rate among 10 trials. In the linear kernel case, β-SVC is reformulated as a convex problem up to $\beta = 0.25$, while in the polynomial kernel case, up to $\beta = 0.40$. Error rates with $\beta \geq 0.26$ (linear kernel) and with $\beta \geq 0.41$ (polynomial kernel) are measured with local optimal solutions of β-SVC. Figure 2 (left) implies that β-SVC with $\beta = 0.41$ is proper to predict labels of new liver-disorders data points, though a nonconvex QP problem of β-SVC need to be solved. For β-SVC with a polynomial kernel, test error is minimized with $\beta = 0.46$.

5.2 Regression Estimation

We used SeDuMi [14] software to solve a convex QP problem of convex β-SVR, and the constrained optimization function "fmincon" in the MATLAB optimization toolbox for nonconvex β-SVR. This particular command uses a SQP algorithm.

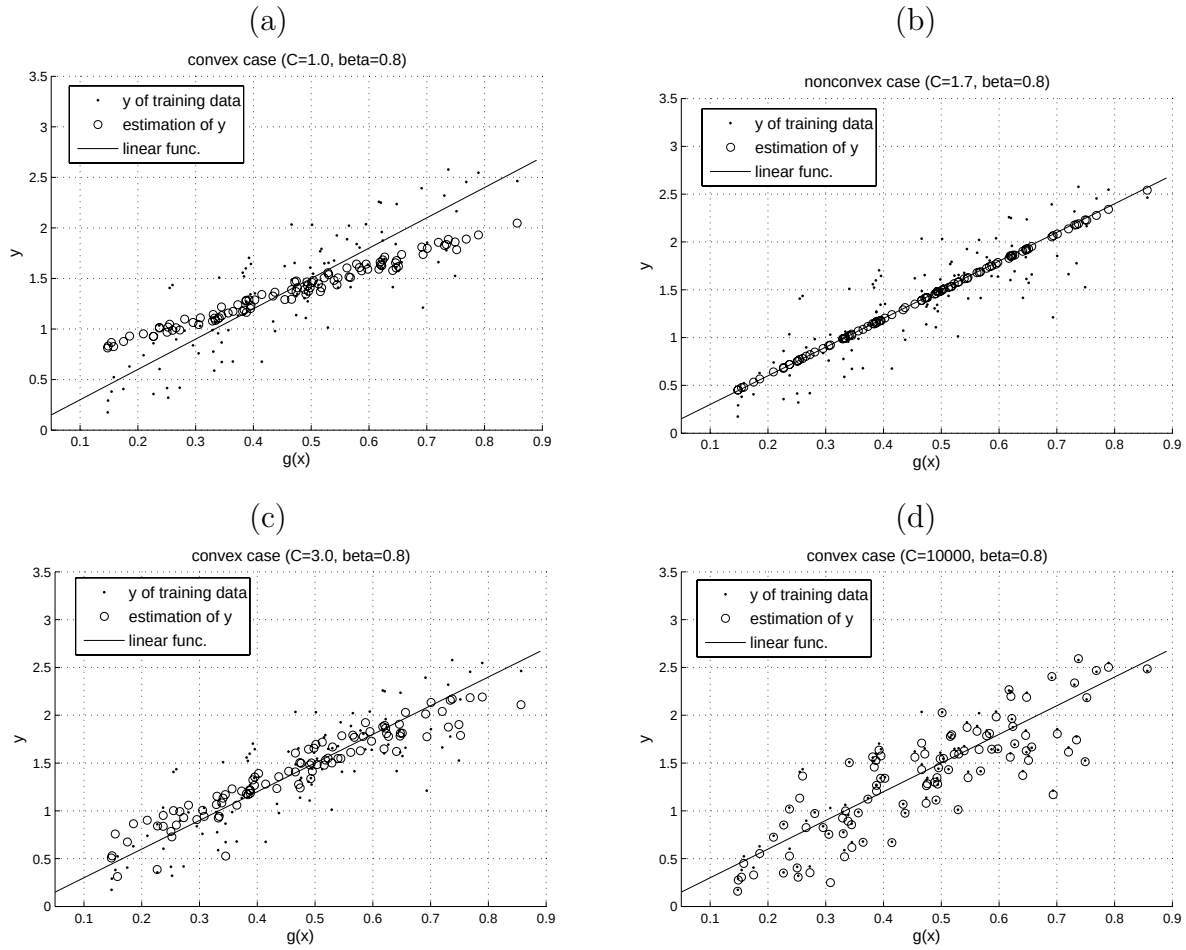

Figure 3: The solid line represents the linear function $y = \boldsymbol{e}^\top \boldsymbol{x}$ without noise, and each dot does y_i of a training data point $\boldsymbol{x}_i$, $i = 1, \ldots, m$. Each circle shows the predicted value of y_i based on β-SVR with $\beta = 0.8$. (a) shows the prediction based on convex β-SVR ($C = 1.0 \leq C_\beta$) with linear kernel. (b) shows the prediction based on nonconvex β-SVR ($C = 1.7 > C_\beta$) with linear kernel. (c) and (d) show the prediction based on convex β-SVR with RBF kernel. The parameter $C = 3.0 < C_\beta$ was used in (c), and $C = 10000 < C_\beta$ in (d).

Toy Example: We start with a toy example in order to check the performance of nonconvex β-SVR. The task is to estimate a noisy linear function given m training data points $(\boldsymbol{x}_i, y_i)$, $i \in M$. The dataset was generated as $y_i = \boldsymbol{e}^\top \boldsymbol{x}_i + v_i$ with an all-one vector $\boldsymbol{e}$. Here v_i was drawn from a Gaussian with zero mean and variance σ^2, and $\boldsymbol{x}_i$ was uniformly from $[0, 1]^n$. A toy example was randomly generated with $m = 100$, $n = 3$ and $\sigma = 0.1$. We solved β-SVR with the linear kernel or RBF kernel $k(\boldsymbol{x}, \boldsymbol{y}) = \exp(-\|\boldsymbol{x} - \boldsymbol{y}\|^2)$ for the toy example, and plotted the estimation of y_i for each training data point $\boldsymbol{x}_i$ in Figure 3.

Figures 3 (a)-(b) show the results of β-SVR with linear kernel, and Figures (c) -(d) does those with RBF kernel. The toy example had $C_\beta = 1.49$ at $\beta = 0.8$ when linear kernel was applied, while $C_\beta = 32024$ at $\beta = 0.8$ for RBF kernel. Since C_β is the threshold to distinguish convexity and nonconvexity of β-SVR, Figures (a), (c) and (d) show the prediction of y at

point $\boldsymbol{x}$ based on convex β-SVR, and only the prediction in Figure (b) was on nonconvex β-SVR. The horizontal axis shows the values of $g(\boldsymbol{x}) = \boldsymbol{e}^\top \boldsymbol{x}/n$. The linear function without noise, $\boldsymbol{e}^\top \boldsymbol{x} = ng(\boldsymbol{x})$, is shown as "linear func." in these figures. Note that nonconvex β-SVR of (b) gives the predicted value of y_i, $i \in M$, sufficiently close to "linear func." at $\boldsymbol{x}_i$. Indeed, for a regression estimate f of the nonconvex β-SVR, the test error is computed with respect to the linear function without noise as

$$\text{risk} = 1/m' \sum_{i=1}^{m'} |f(\bar{\boldsymbol{x}}_i) - \boldsymbol{e}^\top \bar{\boldsymbol{x}}_i| = 0.000$$

using $m' = 1000$ test data points $\bar{\boldsymbol{x}}_i$, $i = 1, \ldots, m'$, while the test error of the convex β-SVR in (a) is risk $= 0.055$. The nonconvex β-SVR with linear kernel nicely worked in the example, but the nonconvex β-SVR with RBF kernel did not. When RBF kernel was applied, a small test error (risk $= 0.034$) was achieved in (c) with $C = 3.0$. The β-SVR of (d) was still convex with $C = 10000$, but the large parameter value C led to overfitting.

In Figures 4 (e)-(f), we see the change in MSE of β-SVR with linear/RBF kernel for different values of the parameter C. The parameter β is fixed at $\beta = 0.8$. The "test MSE" curve in (e) implies that very small mean squared errors (MSE) is attained at nonconvex β-SVR with $C = 1.7$, where the predicted values of y are reported in Figure 3 (b). In the problem setting, nonconvex β-SVR predicts the response value y well. From Figure 4 (f), we see that β-SVR with RBF kernel resulted in overfitting to the training data as the parameter C becomes large. For the toy example, it is sufficient to consider the convex case of β-SVR when applying RBF kernel.

We furthermore investigate the results of nonconvex β-SVR, shown in Figure 3 (b). The case $\beta = 0.8$ of Figure 5 is corresponding to the example of convex β-SVR with $C = 1.0$ (Figure 3 (a)) or nonconvex β-SVR with $C = 1.7$ (Figure 3 (b)). As the curves of (g) show, there were not large differences in training MSE and test MSE with respect to β in this example. The VaR and CVaR curves in Figure (h) show optimal solutions α^* and the optimal values ϕ^*_β of β-SVR (16), respectively. The VaR, α^*, represents an automatically adjusted radius of insensitivity (tube width ϵ in ϵ-SVR). The larger α^* allows less points to lie outside the tube. From Figure (h), we confirm the relation of $\alpha^* \leq \phi^*_\beta$ and the nondecreasingness of ϕ^*_β with respect to β. The VaR coincides with CVaR with sufficiently large β. Figure (i) implies that C_β, the threshold to distinguish convexity and nonconvexity of β-SVR, changes according to β. In the example, at $\beta = 0.2$ or 0.3, β-SVR with $C = 1.7$ can be solved as a convex QP problem because of $C = 1.7 \leq C_\beta$. Let Err and SV be the sets of margin errors and support vectors, respectively, defined with a KKT point of nonconvex β-SVR (16). Then, we see from Figure (j) that $\frac{|Err|}{m} \leq 1 - \beta \leq \frac{|SV|}{m}$ holds. The difference of bounds satisfies $\frac{|SV|-|Err|}{m} \leq \frac{n+1}{m} = 0.04$ at any β.

Boston Housing Benchmark: The convex β-SVR with the parameter $C \leq C_\beta$ attains identical performances to ν-SVR, if parameters C and β of β-SVR are set properly. That is, by selecting C of β-SVR and $\widehat{C}$ of ν-SVR according to Theorem 4.3 or 4.4, we obtain the same regressor via convex β-SVR and ν-SVR with $\nu = 1 - \beta$. Now we focus on Boston housing benchmark problem from the UCI Machine Learning Data Repository [3], and demonstrate the correspondence between those parameters in convex β-SVR and ν-SVR.

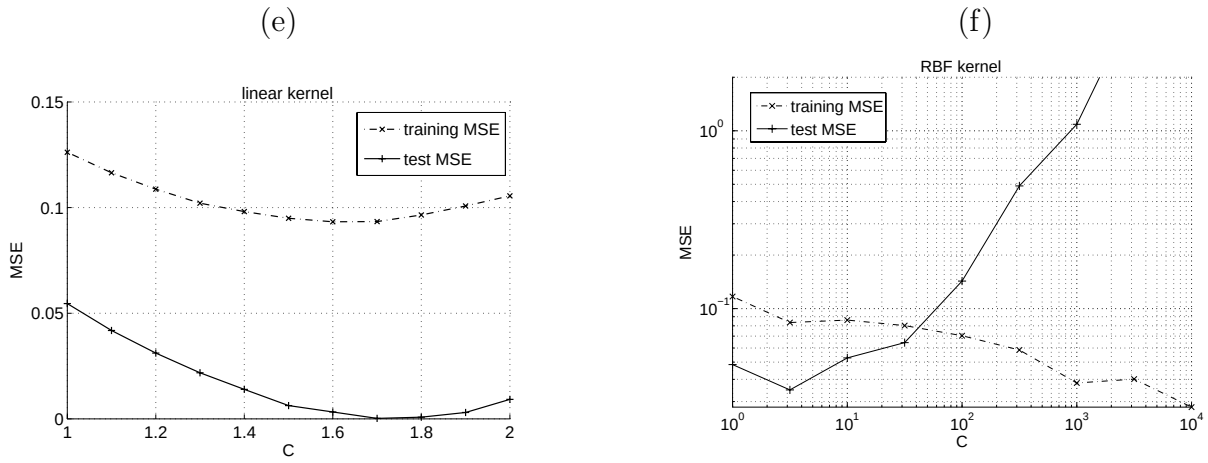

Figure 4: β is fixed at 0.8. (e) shows test MSE (solid line) and training MSE (dashed line) of β-SVR with linear kernel. As far as $C \leq C_\beta = 1.49$, β-SVR results in a convex problem. β-SVR with $C > C_\beta$ is nonconvex. (f) also shows test MSE and training MSE of β-SVR with RBF kernel in log scale. β-SVR with $C \leq C_\beta = 32024$ results in a convex problem.

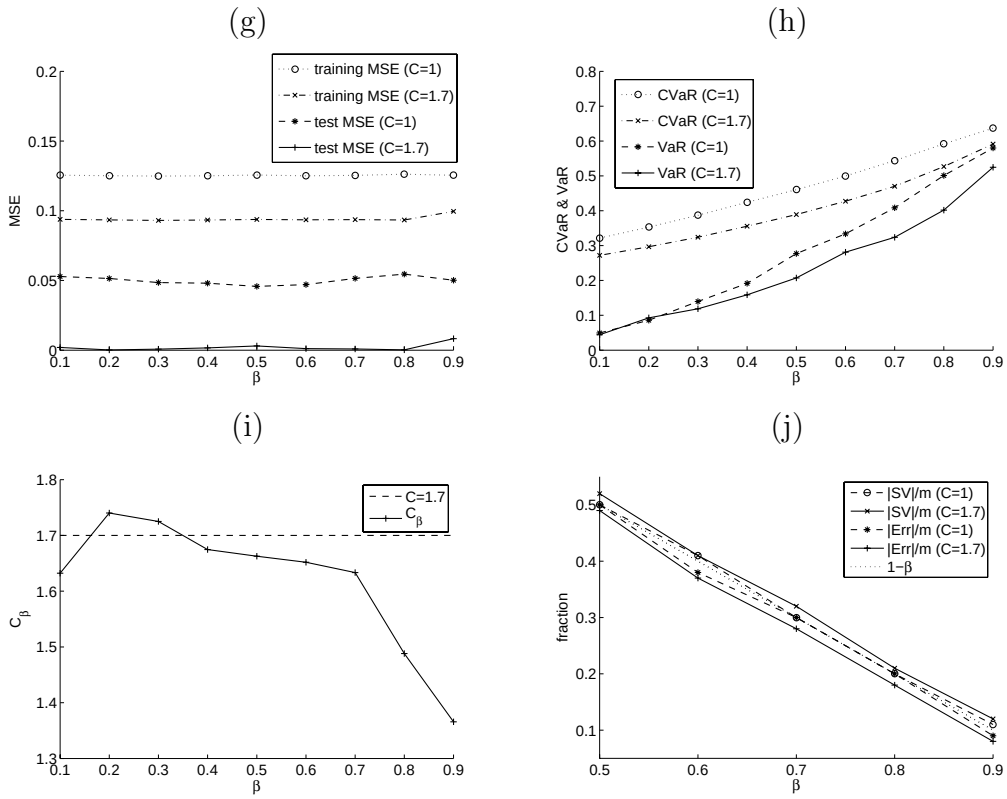

Figure 5: All results were obtained by β-SVR with linear kernel. (g) shows test MSE and training MSE. (h) shows VaR and CVaR. (i) shows threshold C_β which distinguishes convexity and nonconvexity of β-SVR. (j) shows the fraction of margin errors (Err) and that of support vectors (SV).

Table 2: Results for the Boston housing benchmark via β-SVR with $C = 200$

β	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
MSE	10.21	10.36	10.20	9.98	9.62	9.24	8.88	9.39
ϕ^*_β	1.29	1.46	1.65	1.86	2.11	2.41	2.81	3.47
α^*	0.00	0.21	0.45	0.72	1.01	1.40	1.84	2.53

We mostly follow the numerical experiment setting shown in [13]. The data were scaled linearly such that the values of each attribute lie between -1 and 1. The dataset consists of 506 examples in $n = 13$ dimensions. Those examples were split into a training set of $m = 481$ (or $m = 480$) examples and a test set of 25 (or 26, respectively) examples due to 20-fold cross-validation. We used the RBF kernel $k(\boldsymbol{x}, \boldsymbol{y}) = \exp(-\|\boldsymbol{w} - \boldsymbol{y}\|^2/(2\sigma^2))$ with $2\sigma^2 = 3.9$. The parameter $\widehat{C} = 500 \times m$ was used for ν-SVR (21) in [13], and thus, we select C of β-SVR so that the corresponding $\widehat{C}$ is close to $500 \times m$.

C_β is a threshold such that nonconvex β-SVR (16) can be transformed into convex one (17) as far as C is less than C_β. The dataset has the almost same value $C_\beta = 6757$ for any $\beta \in \{0.1, 0.2, \ldots, 0.9\}$. Since numerical experiments were performed with the choice from $C = 200$ to $C = 300$, the optimal regressor of β-SVR was computed via an optimal solution of convex β-SVR (17). Table 2 shows the average of MSE of β-SVR with $C = 200$ over 20 trials. The values ϕ^*_β and α^* denote the average optimal value and the average optimal solution of convex β-SVR (17), respectively. The value α^* corresponds to β-VaR or β-percentile of the distribution $f(\boldsymbol{w}^*, b^*, \boldsymbol{x}, y)$ of β-SVR, and ϕ^*_β corresponds to β-CVaR, mean excess loss, for the distribution.

With the different parameter C, MSE of β-SVR has changed as Figure 6 (left) shows. Among these parameter choice $C = 200$, $C = 250$ and $C = 300$, β-SVR with $C = 200$ seems to be good to achieve small MSE for almost all β, that is, to obtain stable prediction results for any β. ν-SVR with $\widehat{C} = 500 \times 481$ also achieves small MSE at $\beta = 1 - \nu = 0.8$, but the change in MSE with regarding to β is larger than β-SVR with $C = 200$. It may be easy to select the parameter C of β-SVR to attain small MSE, compared to $\widehat{C}$ of ν-SVR. But the difference between β-SVR and ν-SVR only lies in the different parameter setting. If we select the corresponding parameters for C and $\widehat{C}$, we obtain identical prediction results. The correspondence for the parameters is shown in Figure 6 (right). Following Theorem 4.4, $\widehat{C}$ of ν-SVR is computed by

$$\widehat{C} = \frac{C}{\nu\sqrt{\sum_{i,j\in M}(\lambda_i^{(1)*} - \lambda_i^{(2)*})(\lambda_j^{(1)*} - \lambda_j^{(2)*})k(\boldsymbol{x}_i, \boldsymbol{x}_j)}} \tag{25}$$

with the use of an optimal solution $(\boldsymbol{\lambda}^{(1)*}, \boldsymbol{\lambda}^{(2)*})$ of β-SVR (20). Note that the denominator in $\widehat{C}$ is positive because of positive definiteness of RBF kernel matrix. The right figure shows that as β increases, β-SVR with smaller C has the corresponding $\widehat{C}$ of (25), close to fixed $\widehat{C} = 500 \times m$ of ν-SVR (21). This observation implies that MSE of ν-SVR becomes close to that of β-SVR with smaller C as β becomes larger.

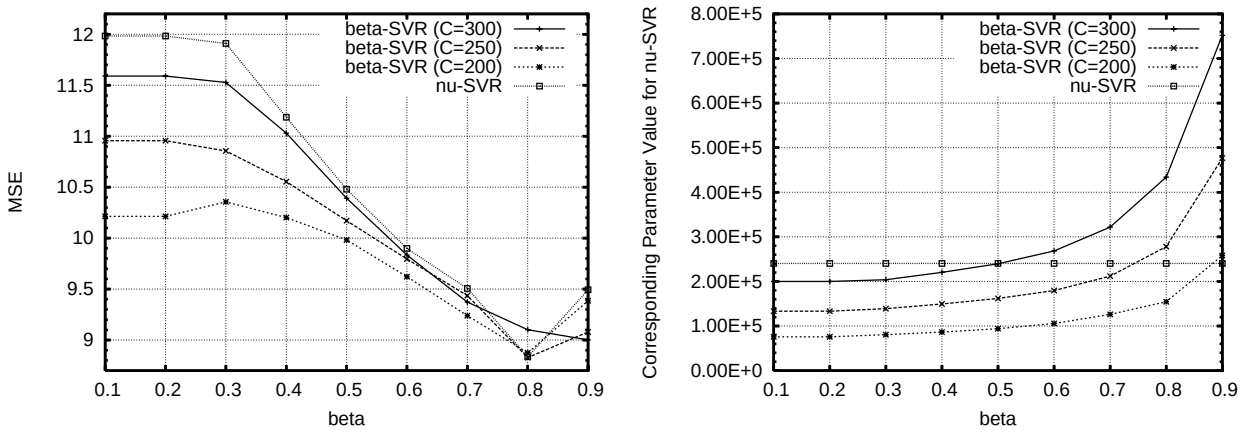

Figure 6: Left figure: average of mean squared errors (MSE) of test points over 20 trials for β-SVR with $C = 200$, $C = 250$ and $C = 300$, and ν-SVR with $\nu = 1 - \beta$ and $\widehat{C} = 500 \times 481$. Right figure: relation between the parameter $\widehat{C}$ of ν-SVR and C of β-SVR.

6 Concluding Remarks

We have proposed β-SVM based on CVaR minimization for classification and regression, and investigated the properties of β-SVM theoretically. β-SVM is closely connected with (Extended) ν-SVM. Indeed, in classification, CVaR minimization for the margin distribution leads to β-SVC, equivalent to Extended ν-SVC [11]. β-SVR model, CVaR minimization for a regression problem, is also essentially equivalent to ν-SVR, when the parameter C is sufficiently small. The paper gave a new viewpoint concerned with CVaR minimization for ν-SVM. The combination of the theory of generalization performance and CVaR risk measure makes it possible to estimate a generalization error bound for β-SVM. The formula of the generalization error bound includes β-CVaR or β-VaR, and thus, the minimum β-CVaR obtained via β-SVM plays an important role to control the generalization error of β-SVM. Therefore, the generalization analysis implies the validity of (Extended) ν-SVM as well as β-SVM.

A CVaR minimization problem can be constructed from arbitrary $f(\boldsymbol{w}, b; \boldsymbol{x}, y)$. The CVaR minimization problem for the distribution $f(\boldsymbol{w}, b; \boldsymbol{x}_i, y_i) = \frac{-y_i(\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle + b)}{\|\boldsymbol{w}\|}$, $i \in M$, is reformulated as β-SVC (3), and that for $f(\boldsymbol{w}, b; \boldsymbol{x}_i, y_i) = \left| y_i - C\frac{\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle + b}{\|\boldsymbol{w}\|} \right|$, $i \in M$, as β-SVR (16). As far as the parameters β of β-SVC and C of β-SVR are in some ranges, those nonconvex problems are transformed into convex QP problems, which are equivalent to ν-SVC and ν-SVR, respectively. In a CVaR minimization problem with any function $f(\boldsymbol{w}, b; \boldsymbol{x}, y)$, the parameter β controls the fractions of support vectors and margin errors defined by the resulting optimal solution. We have the possibility to improve classifier or regressor by using another appropriate function $f(\boldsymbol{w}, b; \boldsymbol{x}, y)$. Furthermore, a classifier or regressor constructed with a global optimal solution of β-SVM may enhance the prediction accuracy in comparison with a local optimal solution. Another possible research direction is to incorporate a kernel matrix learning method in β-SVM. If the optimal kernel matrix is not full-rank, nonconvex β-SVC (7) has a possibility to find a good classifier.

References

[1] P. L. Bartlett, The Sample Complexity of Pattern Classification with Neural Networks: The Size of the Weights is More Important than the Size of the Network, *IEEE Transactions on Information Theory* 44 (1998) 525-536.

[2] J. Bi and K.P. Bennett, A Geometric approach to support vector regression, *Neurocomputing* 55 (2003) 79-108.

[3] C.L. Blake and C.J. Merz, UCI Repository of machine learning databases [`http://www.ics.uci.edu/~mlearn/MLRepository.html`]. Irvine, CA: University of California, Department of Information and Computer Science, 1998.

[4] C. Cortes and V. Vapnik, Support vector networks, *Machine Learning* 20 (1995) 273-297.

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

[6] P.H. Chen, C.J. Lin and B. Schölkopf, A tutorial on nu-support vector machines, *Applied Stochastic Models in Business and Industry* 21 (2005) 111-136.

[7] N. Cristianini and J. Shawe-Taylor, An Introduction to Support Vector Machines and other kernel-based learning methods, Cambridge University Press, 2000.

[8] J. Gotoh and A. Takeda, A Linear Classification Model Based on Conditional Geometric Score, *Pacific Journal of Optimization* 1 (2005) 277-296.

[9] R. Horst and H. Tuy, Global Optimization: Deterministic Approaches, 3rd edition, Springer-Verlag, Berlin, 1995.

[10] G.R.G. Lanckriet, N. Cristianini, P. Bartlett, L. El Ghaoui and M.I. Jordan, Learning the Kernel Matrix with Semidefinite Programming *Journal of Machine Learning Research* 5 (2004) 27-72.

[11] F. Perez-Cruz, J. Weston, D. J. L. Hermann and B. Schölkopf, Extension of the ν-SVM Range for Classification, *Advances in Learning Theory: Methods, Models and Applications 190*, (Eds.) J.A.K. Suykens, G. Horvath, S. Basu, C. Micchelli and J. Vandewalle, IOS Press, Amsterdam (2003) 179-196.

[12] R.T. Rockafellar and S. Uryasev, Conditional value-at-risk for general loss distributions, *Journal of Banking and Finance* 26 (2002) 1443-1471.

[13] B. Schölkopf, A. J. Smola, R. C. Williamson and P. L. Bartlett, New support vector algorithms, *Neural Computation* 12 (2000) 1207-1245.

[14] J. F. Sturm, "Using SeDuMi 1.02, a Matlab Toolbox for Optimization over Symmetric Cones," *Optimization Methods and Software* 11-12 (1999) 625-653.

[15] A. Takeda and H. Nishino, On measuring the inefficiency with the inner-product norm in date envelopment analysis, *European J. Oper. Res.* 133 (2001) 377-393.

[16] R.C. Williamson, A. J.Smola and B. Schölkopf, Generalization Performance of Regularization Networks and Support Vector Machines via Entropy Numbers of Compact Operators, *IEEE Transactions on Information Theory* 47(2001) 2516-2532.