From the Support Vector Machine to the Bounded Constraint Machine

Seo Young Park and Yufeng Liu[*]

The Support Vector Machine (SVM) has been successfully applied for classification problems in many different fields. It was originally proposed using the idea of searching for the maximum separation hyperplane. In this article, in contrast to the criterion of maximum separation, we explore alternative searching criteria which result in the new method, the Bounded Constraint Machine (BCM). Properties and performance of the BCM are explored. To connect the BCM with the SVM, we investigate the Balancing Support Vector Machine (BSVM), which can be viewed as a bridge from the SVM to the BCM. The BCM is shown to be an extreme case of the BSVM. Theoretical properties such as Fisher consistency and asymptotic distributions for coefficients are derived, and the entire solution path of the BSVM is developed. Our numerical results demonstrate how the BSVM and the BCM work compared to the SVM.

Keywords and phrases: Bayes rule, Classification, Consistency, Robustness, Support vector machine.

1. INTRODUCTION

The Support Vector Machine (SVM) has been popular due to its success in many applications [6, 19]. It was originally proposed using the criterion of searching for the optimal separating hyperplane. It is now well known that the SVM can be fit in the *loss + penalty* framework using the hinge loss [20]. In this regularization framework, *loss* measures goodness of fit and *penalty* reflects smoothness of the resulting model.

Despite its success, the SVM has some drawbacks. One known drawback is that the SVM classifier only depends on the set of support vectors (SVs), which include training data points that are correctly classified but relatively close to the boundary as well as those misclassified training points. Extreme outliers can have a relatively big impact on the resulting classifier. In the literature, there have been some attempts to modify the SVM to gain robustness to outliers [4, 15, 18, 22]. The idea is to truncate the unbounded hinge

loss function so that the effect of extreme outliers can be controlled. The corresponding optimization, however, involves challenging nonconvex minimization. Another drawback is that the standard SVM was originally designed for binary classification. Its extension to multicategory classification is nontrivial. Previous attempts include [5, 10, 19, 21]. Despite that these extensions seem natural and reasonable, not all of them are Fisher consistent [13].

Our motivation for this paper is to modify the criterion of the SVM. Instead of the maximum separation criterion whose solution only depends on a subset of the training data, we propose to use an alternative criterion so that all data points can influence the solution. One main advantage of using all data points for the classifier is that the resulting classifier may depend less heavily on a smaller subset and consequently can be more robust to outliers. More specifically, we propose the Bounded Constraint Machine (BCM), which minimizes the sum of the signed distance to the classification boundary subject to some constraints on the solution. Our focus in this paper is on binary classification. However, the BCM can be extended for multicategory classification directly with Fisher consistency.

To further study the relationship between the SVM and BCM, we investigate another method, the Balancing Support Vector Machine (BSVM). The BSVM can be viewed as a modification of the SVM with all training points influencing the resulting classifier. The BSVM is characterized using the parameter v with $v = 0$ corresponding to the SVM and $v = \infty$ corresponding to the BCM. As a result, the BSVM helps to build a continuous path from the SVM to BCM by changing the value of v. Along with the effect of v, the properties of the BSVM including Fisher consistency and asymptotic behaviors of the coefficients are investigated.

In practice, the performance of these methods may vary from problem to problem. Therefore, it may be desirable to treat v data dependent. To improve the computational efficiency, we establish the entire solution path with respect to the value of v, so that we can get the solution of the BSVM for every value of v efficiently.

The rest of the article is organized as follows. Section 2 briefly reviews the standard SVM and proposes the BCM. In Section 3, we investigate the BSVM and describe its behavior using the Lagrange dual problem. The effect of v is explored and we show how the BSVM builds a connection from the SVM to the BCM. Section 4 shows Fisher consistency of the BSVM and BCM, as well as some asymptotic

properties. Section 5 develops the regularized solution path with respect to v. Numerical results are reported in Section 6 and Section 7 gives some discussion. The proofs of our theorems are included in the Appendix.

2. THE SVM AND THE BCM

In standard binary classification, we want to build a classifier based on a training sample $\{(\boldsymbol{x}_i, y_i)|i = 1, 2, \ldots, n\}$, where $\boldsymbol{x}_i \in \mathcal{X} \subset \mathbb{R}^d$ is a vector of predictors, and $y_i \in \{+1, -1\}$ is its class membership. Typically it is assumed that the training data are distributed according to an unknown probability distribution $P(\boldsymbol{x}, y)$. The goal is to find a decision function $f(\boldsymbol{x})$ and its associated classifier sign$[f(\boldsymbol{x})]$ which minimizes the misclassification rate. In this paper, we focus on linear learning, which seeks a linear classifier $f(\boldsymbol{x}) = b + \boldsymbol{x}^T \boldsymbol{w}$. The same idea can be generalized to nonlinear learning through basis expansion or kernel trick.

Many well known classifiers can be formulated in a $loss+$ $penalty$ framework

$$(1) \qquad \min_f \sum_{i=1}^n l(y_i f(\boldsymbol{x}_i)) + \lambda J(f),$$

where $l(\cdot)$ is a loss function that measures goodness of fit, $J(f)$ is a penalty term that assesses generalization of the model, and λ is a tuning parameter which balances the tradeoff between those two [20]. One may formulate the optimization as $\min_f C \sum_{i=1}^n l(y_i f(\boldsymbol{x}_i)) + J(f)$, which is essentially the same as (1) with λ playing the same role as $1/C$. In this paper, we use both notations C and λ for convenience. Note that the loss function l here is a function of $yf(\boldsymbol{x})$, which shows 'correctness' of the classification for a particular observation $\boldsymbol{x}$. In particular, with the classification rule sign$[f(\boldsymbol{x})]$, positive $yf(\boldsymbol{x})$ implies correct classification and negative $yf(\boldsymbol{x})$ implies wrong classification. Moreover, we can think of the absolute value of $f(\boldsymbol{x})$ as our 'confidence' in class label prediction, considering the value of $f(\boldsymbol{x})$ close to zero indicates that $\boldsymbol{x}$ is near the decision boundary. Thus, large value of $yf(\boldsymbol{x})$ implies classification for $\boldsymbol{x}$ is correct, and as the value of $yf(\boldsymbol{x})$ goes to negative infinity, it means the classification with high confidence was wrong. Hence we generally want values of $yf(\boldsymbol{x})$ to be large, and it should be and usually is reflected in the shape of the loss function $l(yf(\boldsymbol{x}))$, which explains why many common loss functions are nonincreasing in $yf(\boldsymbol{x})$. Some typical examples include the hinge loss [20], the logistic loss [11], the exponential loss [7], and the ψ loss [18].

2.1 The Standard SVM

The SVM is a typical method of form (1). In particular, it employs the hinge loss function $l(yf(\boldsymbol{x})) = [1 - yf(\boldsymbol{x})]_+$, and the penalty term $J(f) = \frac{1}{2}\|\boldsymbol{w}\|^2$. Note that the value of the hinge loss $l(yf(\boldsymbol{x}))$ increases as $yf(\boldsymbol{x})$ becomes smaller and it stays at zero when $yf(\boldsymbol{x}) \geq 1$. That is, the SVM puts loss

on the misclassified data points but nothing on the correctly classified observations once $yf(\boldsymbol{x})$ becomes greater than 1. Hence the data points with $yf(\boldsymbol{x}) \geq 1$ have no influence on the SVM solution. To further explain, we express the dual problem

$$(2) \qquad \min_{\boldsymbol{\alpha}} \frac{1}{2} \sum_{i,j=1}^n y_i y_j \alpha_i \alpha_j \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle - \sum_{i=1}^n \alpha_i$$

$$\text{subject to} \quad \sum_{i=1}^n y_i \alpha_i = 0; 0 \leq \alpha_i \leq C, \forall i = 1, \ldots, n,$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product. Using the α_i obtained from (2), $\boldsymbol{w}$ can be calculated as $\sum_{i=1}^n \alpha_i y_i \boldsymbol{x}_i$, and b can be obtained by the KKT conditions. Thus the classification function can be written as $f(\boldsymbol{x}) = \sum_{i=1}^n \alpha_i y_i \langle \boldsymbol{x}_i, \boldsymbol{x} \rangle + b$. Furthermore, $\alpha_i > 0$ implies $y_i f(\boldsymbol{x}_i) \leq 1$ and actually that is the only case that $(\boldsymbol{x}_i, y_i)$ can affect the solution. On the other hand, when $\alpha_i = 0$, the observation $(\boldsymbol{x}_i, y_i)$ has no impact on the solution. A point $\boldsymbol{x}_i$ with $\alpha_i > 0$ is a SV, which is the observation satisfying $y_i f(\boldsymbol{x}_i) \leq 1$.

2.2 The BCM

Due to the design of the SVM, its solution only depends on the set of SVs. This helps to simplify the solution. However, if the training dataset is noisy with outliers, the solution can be deteriorated. To solve the problem, we propose a different optimization criterion. In particular, we propose to minimize the sum of signed distances to the boundary and solve the following problem

$$\min_f J(f) - C \sum_{i=1}^n y_i f(\boldsymbol{x}_i)$$
$$\text{subject to} \quad -1 \leq f(\boldsymbol{x}_i) \leq 1, \forall i = 1, \ldots, n.$$

That is, we try to maximize $\sum_{i=1}^n y_i f(\boldsymbol{x}_i)$, while forcing all the training data to stay between the hyperplanes $f(\boldsymbol{x}) = \pm 1$. One can view that the BCM uses the hinge loss of the SVM with $y_i f(\boldsymbol{x}_i) \in [-1, 1]$. With the constraints, the BCM makes use of all training points to obtain the resulting classifier.

One advantage of the BCM is that it can be extended to the multicategory case directly. Assume that we have a k-class problem with $y \in \{1, \ldots, k\}$. Let $\boldsymbol{f} = (f_1, \ldots, f_k)$ be the decision function vector with $\sum_{j=1}^k f_j = 0$. Then the multicategory BCM solves the following problem

$$(3) \qquad \min_{\boldsymbol{f}} \sum_{j=1}^k \|f_j\|^2 - C \sum_{i=1}^n f_{y_i}(\boldsymbol{x}_i)$$

$$\text{subject to} \sum_{j=1}^k f_j(\boldsymbol{x}_i) = 0; f_l(\boldsymbol{x}_i) \geq -1;$$

$$\forall i = 1, \ldots, n, l = 1, \ldots, k.$$

It can be shown that the multicategory BCM is Fisher consistent. However, we will focus on binary classification in this paper.
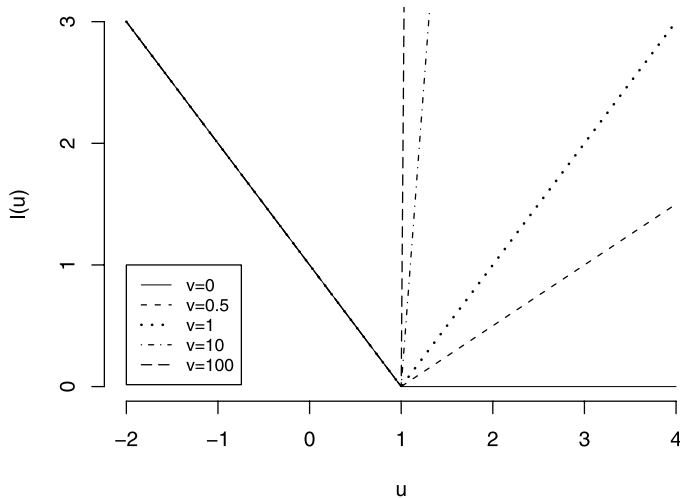
Figure 1. Plot of loss function $l(u) = g(u)$ for the BSVM with different values of v.

To further understand the connection between the SVM and BCM, we propose the BSVM in Section 3 and use the BSVM as a bridge to connect the SVM and BCM.

3. THE BSVM: A BRIDGE BETWEEN THE SVM AND THE BCM

The SVM only uses the SV set to calculate its solution, while the BCM utilizes all training points. To connect these two, we study the BSVM using the following loss function

$$(4) \qquad g(u) = \begin{cases} 1 - u & \text{if } u \leq 1, \\ v(u-1) & \text{otherwise,} \end{cases}$$

where v is the slope of the loss function when $u \in (1, \infty)$, as shown in Fig. 1. Note that v determines how much the solution will rely on the data points with $yf(\boldsymbol{x}) \geq 1$, and the problem becomes equivalent to the SVM when $v = 0$. Here, we would like to acknowledge that the loss $g(u)$ was previously presented by Ming Yuan in the Statistical Learning Conference at Snowbird, UT in 2007. We use the BSVM as a bridge to connect the SVM with the proposed BCM.

Note that when $v = \infty$, the BSVM becomes equivalent to solving

$$(5) \qquad \begin{aligned} &\min_{(b,\boldsymbol{w})} J(f) - C \sum_{i=1}^{n} y_i f(\boldsymbol{x}_i) \\ &\text{subject to} \quad f(\boldsymbol{x}_i) \leq 1, \forall i = 1, \dots, n. \end{aligned}$$

Comparing to the BCM in (3), the only difference is that the BCM has the constraint $f(\boldsymbol{x}_i) \geq -1$ but the BSVM with $v = \infty$ does not. Typically this difference does not matter since the solution of (5) usually satisfies $f(\boldsymbol{x}_i) \geq -1$. The only case that the BCM actually works differently from the BSVM is when a data point moves far away from its own class, even further than the other class. This rarely happens in practice. Thus, the BSVM with $v = \infty$ can be viewed

as a good approximation of the BCM. Overall, the BSVM builds a continuum from the standard SVM ($v = 0$) to the BCM ($v = \infty$).

3.1 Interpretation of the BSVM

Since the loss $g(u)$ for the BSVM is not a decreasing function and it imposes big loss values even on the correctly classified data points as well as misclassified observations, it might seem counterintuitive. However, the increasing part with $y_i f(\boldsymbol{x}_i) > 1$ may help to bring the decision boundary towards the correctly classified points, which can be desirable in some situations. To understand the behavior of the BSVM further, we rewrite its primal problem as follows

$$\min_{(b,\boldsymbol{w})} \frac{1}{2}\|\boldsymbol{w}\|^2 + C \sum_{i=1}^{n} \xi_i$$
$$\text{subject to} \quad \xi_i \geq 1 - y_i f(\boldsymbol{x}_i); \xi_i \geq v(y_i f(\boldsymbol{x}_i) - 1),$$
$$\forall i = 1, \dots, n.$$

The corresponding Lagrange primal can be written as

$$(6)$$
$$\begin{aligned} L(\boldsymbol{w}, b, \boldsymbol{\alpha}) = &\frac{1}{2}\|\boldsymbol{w}\|^2 + C \sum_{i=1}^{n} \xi_i + \sum_{i=1}^{n} \gamma_i[1 - y_i f(\boldsymbol{x}_i) - \xi_i] \\ &+ \sum_{i=1}^{n} \delta_i[v y_i f(\boldsymbol{x}_i) - v - \xi_i]. \end{aligned}$$

Setting derivatives to zero gives

$$(7) \qquad \frac{\partial L}{\partial \boldsymbol{w}} = \boldsymbol{w} - \sum_{i=1}^{n} y_i \gamma_i \boldsymbol{x}_i + \sum_{i=1}^{n} v y_i \delta_i \boldsymbol{x}_i = \boldsymbol{0}$$

$$(8) \qquad \frac{\partial L}{\partial b} = -\sum_{i=1}^{n} y_i \gamma_i + v \sum_{i=1}^{n} y_i \delta_i = 0$$

$$(9) \qquad \frac{\partial L}{\partial \xi_i} = C - \gamma_i - \delta_i = 0,$$

and KKT conditions are

$$(10) \qquad \gamma_i(1 - y_i f(\boldsymbol{x}_i) - \xi_i) = 0$$
$$(11) \qquad \delta_i(v y_i f(\boldsymbol{x}_i) - v - \xi_i) = 0.$$

Then the corresponding dual problem becomes

$$(12)$$
$$\min_{\boldsymbol{\alpha}} \frac{1}{2} \sum_{i,j=1}^{n} y_i y_j \alpha_i \alpha_j \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle - \sum_{i=1}^{n} \alpha_i$$
$$\text{subject to} \quad \sum_{i=1}^{n} y_i \alpha_i = 0; -Cv \leq \alpha_i \leq C, \forall i = 1, \dots, n.$$

Once the solution of (12) is obtained, $\boldsymbol{w}$ can be calculated as $\sum_{i=1}^{n} \alpha_i y_i \boldsymbol{x}_i$ and b can be determined by KKT conditions. This problem is almost identical to the SVM problem. The difference is on the constraint. In particular, we
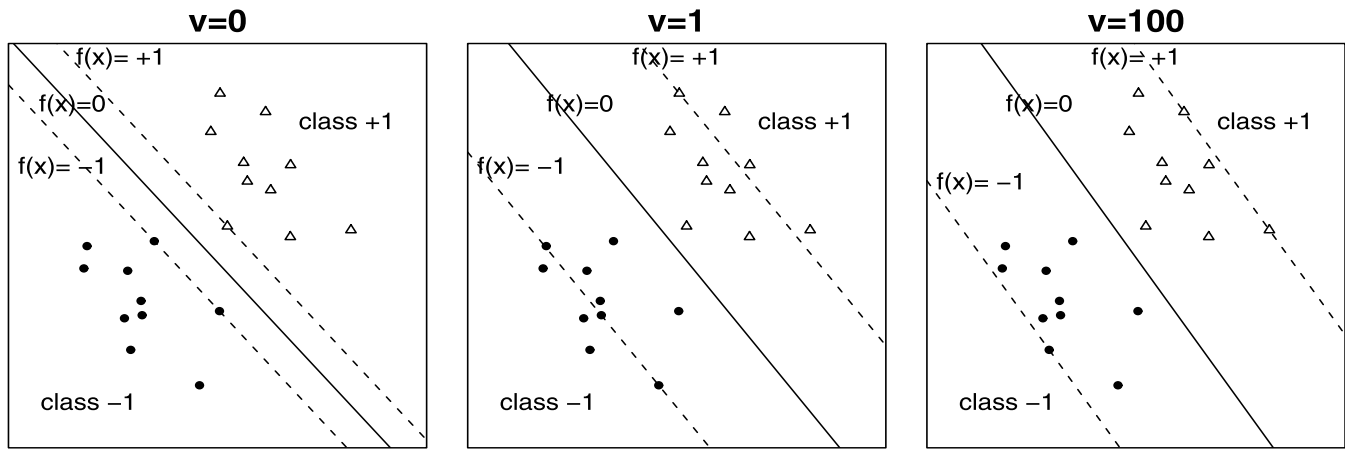
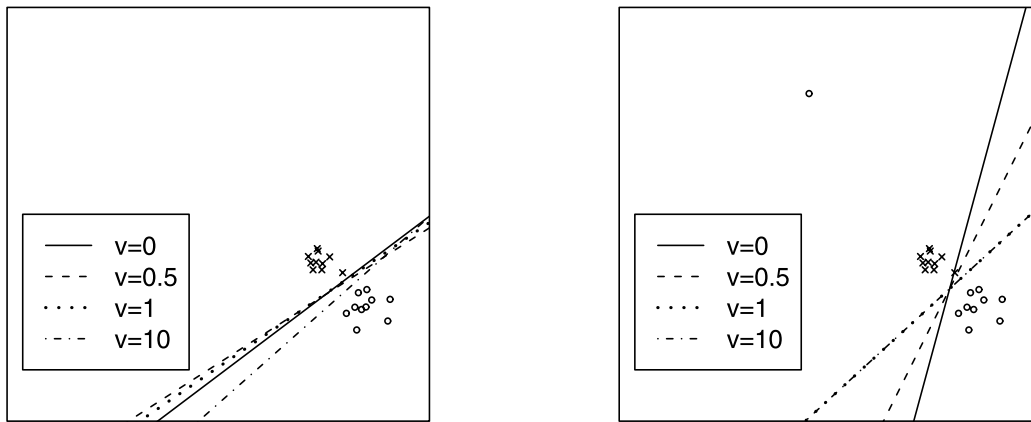Figure 2. Plots of the effect of different values of v on the BSVM.

Figure 3. A graphical illustration of the robustness of the BSVM: the decision boundary of the BSVM stays stable when there is an extreme outlier, while that of the SVM moves dramatically towards the outlier.

have $0 \leq \alpha_i \leq C$ for the SVM, but $-Cv \leq \alpha_i \leq C$ for the BSVM. This helps to explain the difference in behaviors between the SVM and the BSVM. In contrast to the SVM, the BSVM with $v > 0$ makes use of all data points to determine the solution. Points with $y_i f_i \leq 1$ may help to reduce the effect of outliers and consequently the BSVM classifier can be more robust against outliers.

3.2 Effect of v

In the separable case, the standard SVM, i.e. the BSVM with $v = 0$, finds the decision boundary which maximizes the distance from the decision boundary to the nearest data point, i.e., the distance between $f(\boldsymbol{x}) = \pm 1$ is maximized. Here, the soft margins $f(\boldsymbol{x}) = \pm 1$ are the hyperplanes that bound the data points of each class, so that the observations are forced to lie outside of the soft margins. The BSVM with $v > 0$ maximizes the distance between $f(\boldsymbol{x}) = \pm 1$ as well, but the observations are clustered around the hyperplanes $f(\boldsymbol{x}) = \pm 1$ without being forced to be outside of the margin lines. When $v = 1$, the BSVM minimizes $\sum_i |1 - y_i f(\boldsymbol{x}_i)|$,

resulting data points laid inside and outside of $f(\boldsymbol{x}) = \pm 1$ evenly as shown in the middle panel of the Fig. 2. As the value of v becomes high, the value of $v[y_i f(\boldsymbol{x}_i) - 1]_+$, which is the distance between the hyperplanes $f(\boldsymbol{x}) = \pm 1$ and the observations outside of them, becomes larger. Thus the hyperplanes $f(\boldsymbol{x}) = \pm 1$ move towards outside to reduce it. As v goes to infinity, the BSVM reduces to the BCM and the hyperplanes $f(\boldsymbol{x}) = \pm 1$ go far enough to bound all data points. The right panel of the Fig. 2 illustrates the behavior of the BCM with large v.

Since v decides how much the decision boundary depends on the correctly classified observations, performance of the BSVM is affected by the value of v. The BSVM with big value of v tends to depend on the correctly classified data, which makes it less sensitive against outliers. The BCM can be viewed as the most extreme case with $v = \infty$. The toy example in Fig. 3 illustrates this behavior. When there is no outlier as shown on the left panel, the SVM and the BSVM with different values of v perform similarly. However, when an observation moves far away from its own class, the de-
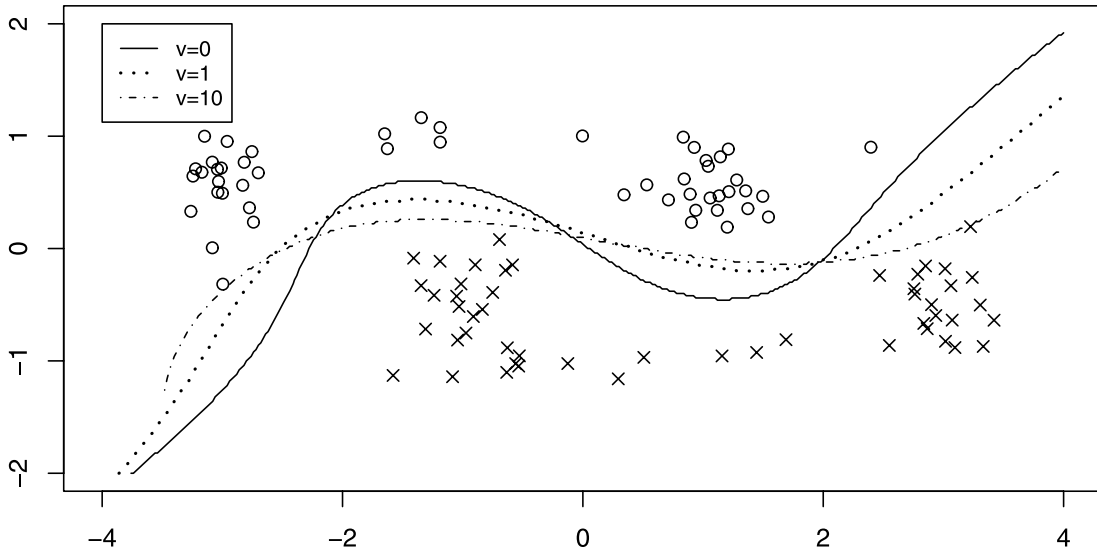
Figure 4. A graphical comparison of the SVM vs. BSVM: the decision boundary of the SVM reflects the wavy shaped structure of the data near the border, while that of the BSVM is flattened by the observations far from the boundary.

cision boundary of the SVM moves towards the outlier, resulting in a data point misclassified. In contrast, the BSVM with large v is more stable because the effect of the outlier is greatly reduced by the correctly classified data. Therefore, correctly classified data in the BSVM help to robustify the decision boundary so that a small number of outliers can not cause a drastic change on the decision boundary.

It is worthwhile to point out that the RSVM [22] can also deliver robust classifiers. It achieves robustness via removing potential outliers from the set of SVs for the standard SVM. Consequently, the RSVM gains robustness by using a smaller but more robust set of observations. In contrast, the BSVM tries to reduce the impact of outliers by making use of more data points. Both methods are reasonable, however, they use different philosophies in using the training data to obtain robustness.

As a remark, we note that the BSVM may not always produce better results than that of the SVM. It can be suboptimal in a situation as the toy example shown in Fig. 4. The true boundary is wavy shaped, but the observations far away from the boundary are aligned in parallel. The SVM works fairly well, but the decision boundary of the BSVM becomes flat as the value of the v goes large due to the influences of the data points far from the boundary. Hence, choice of v should be made carefully based on the characteristic of the problem.

4. PROPERTIES OF THE BSVM AND THE BCM

4.1 Fisher consistency

In this section, we discuss Fisher consistency of the BSVM and the BCM. Fisher consistency, also known as classification-calibration [2], requires that the population minimizer of a loss function has the same sign as $P(x) - 1/2$ in the binary case [12]. This is a desirable property for classification. The following theorem establishes Fisher consistency of the BSVM.

Theorem 1. *The minimizer f^* of $E[g(Yf(\boldsymbol{X}))|\boldsymbol{X} = \boldsymbol{x}]$ is $sign[P(\boldsymbol{x}) - 1/2]$.*

For the BCM, we consider the multicategory case due to its simple extension. In the multicategory case, Fisher consistency requires that $\operatorname{argmax}_j f_j^* = \operatorname{argmax}_j P_j$, where $\boldsymbol{f}^*(\boldsymbol{x}) = (f_1^*(\boldsymbol{x}), \ldots, f_k^*(\boldsymbol{x}))$ denotes the minimizer of expected value of the loss function. The following theorem shows Fisher consistency for the multicategory BCM.

Theorem 2. *The minimizer $\boldsymbol{f}^*$ of $E[-f_Y(\boldsymbol{X})]$, subject to $\sum_j^k f_j(\boldsymbol{x}) = 0$ and $f_l(\boldsymbol{x}) \geq -1$ for $\forall l$, satisfies the following: $f_j^*(\boldsymbol{x}) = k - 1$ if $j = \operatorname{argmax}_j P_j(\boldsymbol{x})$ and -1 otherwise.*

4.2 Asymptotic study of the BSVM

In this section, we study asymptotic distributions of the coefficients in the BSVM. [9] established Bahadur type representation [1, 3] of the classical SVM coefficients to study their asymptotic behavior. This representation allows us to see how the margin lines of the SVM and the underlying probability distribution of observations affects asymptotic behaviors of the coefficients. This idea can be generalized to the BSVM with some modifications on the Bahadur representation of the coefficients and regularity conditions to adopt the loss function of the BSVM. We show that the coefficients of the BSVM have asymptotic normality, as that of the standard SVM.

First, we introduce new notations for convenience. Let $\boldsymbol{\beta} = (\beta_0, \boldsymbol{\beta}_+)$ denote $(b, \boldsymbol{w})$ which is the coefficients in

the BSVM. Let $\tilde{\boldsymbol{x}} = (1, \boldsymbol{x}^T)^T = (1, x_1, \ldots, x_d)^T = (\tilde{x}_0, \tilde{x}_1, \ldots, \tilde{x}_d)^T$ and denote the linear decision function for given $\boldsymbol{X} = \boldsymbol{x}$ as $f(\boldsymbol{x}; \boldsymbol{\beta}) = \boldsymbol{x}^T \boldsymbol{\beta} = \beta_0 + \boldsymbol{x}^T \boldsymbol{\beta}_+$. Let $\pi_+ = P(Y = 1) > 0$ and $\pi_- = P(Y = -1) > 0$, with $\pi_+ + \pi_- = 1$. Let h_+ and h_- be the density functions of $\boldsymbol{X}$ given $Y = 1$ and -1, respectively. Denote the objective function of the BSVM

$$(13) \qquad q_{\lambda,n}(\boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^{n} g(y_i f(\boldsymbol{x}_i; \boldsymbol{\beta})) + \frac{\lambda}{2} \|\boldsymbol{\beta}_+\|.$$

The population version of (13) without the penalty term is denoted by

$$(14) \qquad Q(\boldsymbol{\beta}) = E[g(Y f(\boldsymbol{X}; \boldsymbol{\beta}))]$$

and the minimizers of (13) and (14) are denoted by $\hat{\boldsymbol{\beta}}_{\lambda,n}$ and $\boldsymbol{\beta}^*$ respectively. Let the indicator function be $\psi(z) = I_{\{z \geq 0\}}$ for $z \in \mathbb{R}$ and denote the $(d+1)$-dimensional vector $S(\boldsymbol{\beta}) = E[-\psi(1 - Y f(\boldsymbol{X}; \boldsymbol{\beta}))Y\tilde{\boldsymbol{X}} + v\psi(Y f(\boldsymbol{X}; \boldsymbol{\beta}) - 1)Y\tilde{\boldsymbol{X}}]$ and the $(d+1) \times (d+1)$ matrix $H(\boldsymbol{\beta}) = (1 + v)E[\delta(1 - Y f(\boldsymbol{X}; \boldsymbol{\beta}))\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^T]$, where δ is the Dirac delta function. One can show that $S(\boldsymbol{\beta})$ and $H(\boldsymbol{\beta})$ are the gradient and Hessian matrix of $Q(\boldsymbol{\beta})$, respectively.

Now we state the regularity conditions for the asymptotic results. Here, $C_1, C_2, \ldots$ are positive constants which do not depend on n.

A1 The densities h_+ and h_- are continuous and have finite second moments.

A2 There exists $B(\boldsymbol{x}_0, r_0)$, a ball centered at $\boldsymbol{x}_0$ with radius $r_0 > 0$ such that $\pi_+ h_+(\boldsymbol{x}) + \pi_- h_-(\boldsymbol{x}) > C_1$ for every $\boldsymbol{x} \in B(\boldsymbol{x}_0, r_0)$.

A3 For some $1 \leq i^* \leq d$,

$$\pi_+ \left\{ \int_{\mathcal{X}} (I_{\{x_{i^*} \leq F_{i^*}^+\}} - vI_{\{x_{i^*} > F_{i^*}^+\}}) x_{i^*} h_+(\boldsymbol{x}) d\boldsymbol{x} \right\}$$
$$> \pi_- \left\{ \int_{\mathcal{X}} (I_{\{x_{i^*} \geq G_{i^*}^-\}} - vI_{\{x_{i^*} < G_{i^*}^-\}}) x_{i^*} h_-(\boldsymbol{x}) d\boldsymbol{x} \right\}$$

or

$$\pi_+ \left\{ \int_{\mathcal{X}} (I_{\{x_{i^*} \geq F_{i^*}^-\}} - vI_{\{x_{i^*} < F_{i^*}^-\}}) x_{i^*} h_+(\boldsymbol{x}) d\boldsymbol{x} \right\}$$
$$< \pi_- \left\{ \int_{\mathcal{X}} (I_{\{x_{i^*} \leq G_{i^*}^+\}} - vI_{\{x_{i^*} > G_{i^*}^+\}}) x_{i^*} h_-(\boldsymbol{x}) d\boldsymbol{x} \right\}$$

for $F_{i^*}^+, G_{i^*}^+, F_{i^*}^-, G_{i^*}^- \in [-\infty, \infty]$ such that

$$\int_{\mathcal{X}} I_{\{x_{i^*} \leq F_{i^*}^+\}} h_+(\boldsymbol{x}) d\boldsymbol{x} = \min\left\{1, \frac{\frac{\pi_-}{\pi_+} + v}{1 + v}\right\},$$

$$\int_{\mathcal{X}} I_{\{x_{i^*} \leq G_{i^*}^+\}} h_-(\boldsymbol{x}) d\boldsymbol{x} = \min\left\{1, \frac{\frac{\pi_+}{\pi_-} + v}{1 + v}\right\},$$

$$\int_{\mathcal{X}} I_{\{x_{i^*} \geq F_{i^*}^-\}} h_+(\boldsymbol{x}) d\boldsymbol{x} = \min\left\{1, \frac{\frac{\pi_-}{\pi_+} + v}{1 + v}\right\},$$

$$\int_{\mathcal{X}} I_{\{x_{i^*} \geq G_{i^*}^-\}} h_-(\boldsymbol{x}) d\boldsymbol{x} = \min\left\{1, \frac{\frac{\pi_+}{\pi_-} + v}{1 + v}\right\}.$$

A4 For an orthogonal transformation A_{j^*} that maps $\boldsymbol{\beta}_+^*/\|\boldsymbol{\beta}_+^*\|$ to the j^*-th unit vector e_{j^*} for some $1 \leq j^* \leq d$, there exist rectangles

$$\mathcal{D}^+ = \{\boldsymbol{x} \in M^+ : l_i \leq (A_{j^*}\boldsymbol{x})_i \leq v_i, l_i < v_i \text{ for } i \neq j^*\}$$

and

$$\mathcal{D}^- = \{\boldsymbol{x} \in M^- : l_i \leq (A_{j^*}\boldsymbol{x})_i \leq v_i, l_i < v_i \text{ for } i \neq j^*\}$$

such that $h_+(\boldsymbol{x}) \geq C_2 > 0$ on $\mathcal{D}^+$, and $h_-(\boldsymbol{x}) \geq C_3 > 0$ on $\mathcal{D}^-$, where $M^+ = \{\boldsymbol{x} \in \mathcal{X} | \beta_0^* + \boldsymbol{x}^T \boldsymbol{\beta}_+^* = 1\}$ and $M^- = \{\boldsymbol{x} \in \mathcal{X} | \beta_0^* + \boldsymbol{x}^T \boldsymbol{\beta}_+^* = -1\}$.

Note that **A1** is needed to guarantee that $S(\boldsymbol{\beta})$ and $H(\boldsymbol{\beta})$ are well-defined and continuous in $\boldsymbol{\beta}$. If **A1** is met, the condition that $h_+(bx_0) > 0$ or $h_-(bx_0) > 0$ for some $\boldsymbol{x}_0$ implies **A2**. **A3** is the condition to ensure that $\boldsymbol{\beta}_+^* \neq \boldsymbol{0}$, and if $\pi_+ = \pi_-$, then it simply means that the mean vectors of the conditional class distributions are different. **A4** ensures the positive-definiteness of $H(\boldsymbol{\beta})$ around $\boldsymbol{\beta}^*$. This condition is easily satisfied when the supports of h_+ and h_- are convex. Assuming these regularity conditions, we have a Bahadur-type representation of $\hat{\boldsymbol{\beta}}_{\lambda,n}$ as shown in Theorem 3. This induces the asymptotic normality of $\hat{\boldsymbol{\beta}}_{\lambda,n}$ (Theorem 4).

Theorem 3. *Suppose* **A1**–**A4** *are satisfied. Then, for* $\lambda = o(n^{-1/2})$,

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{\lambda,n} - \boldsymbol{\beta}^*) = -\frac{1}{\sqrt{n}}H(\boldsymbol{\beta}^*)^{-1}$$
$$\times \sum_{i=1}^{n} (I_{\{y_i f(\boldsymbol{X}_i; \boldsymbol{\beta}^*) \leq 1\}} - vI_{\{y_i f(\boldsymbol{X}_i; \boldsymbol{\beta}^*) > 1\}}) y_i \tilde{\boldsymbol{X}}_i + o_{\mathbb{P}}(1).$$

Theorem 4. *Suppose* **A1**–**A4** *are satisfied. Then, for* $\lambda = o(n^{-1/2})$,

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{\lambda,n} - \boldsymbol{\beta}^*) \to N(0, H(\boldsymbol{\beta}^*)^{-1}G(\boldsymbol{\beta}^*)H(\boldsymbol{\beta}^*)^{-1})$$

in distribution as $n \to \infty$*, where*

$$G(\boldsymbol{\beta}) = E[(I_{\{y_i f(\boldsymbol{X}_i; \boldsymbol{\beta}^*) \leq 1\}} + v^2 I_{\{y_i f(\boldsymbol{X}_i; \boldsymbol{\beta}^*) > 1\}})\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^T].$$

This result can be used for building a confidence bound for $\boldsymbol{\beta}$ or $f(\boldsymbol{x}; \boldsymbol{\beta})$ for a specific $\boldsymbol{x}$. The proofs are given in the Appendix.

To illustrate the asymptotic results, we introduce a simple toy example as follows. Let the one-dimensional explanatory variable x follows $N(1, 1)$ if it belongs to class 1, and otherwise it follows $N(-1, 1)$. Then it can be shown that $\beta_0^* = 0$ and $\beta_+^* = 1$, which gives
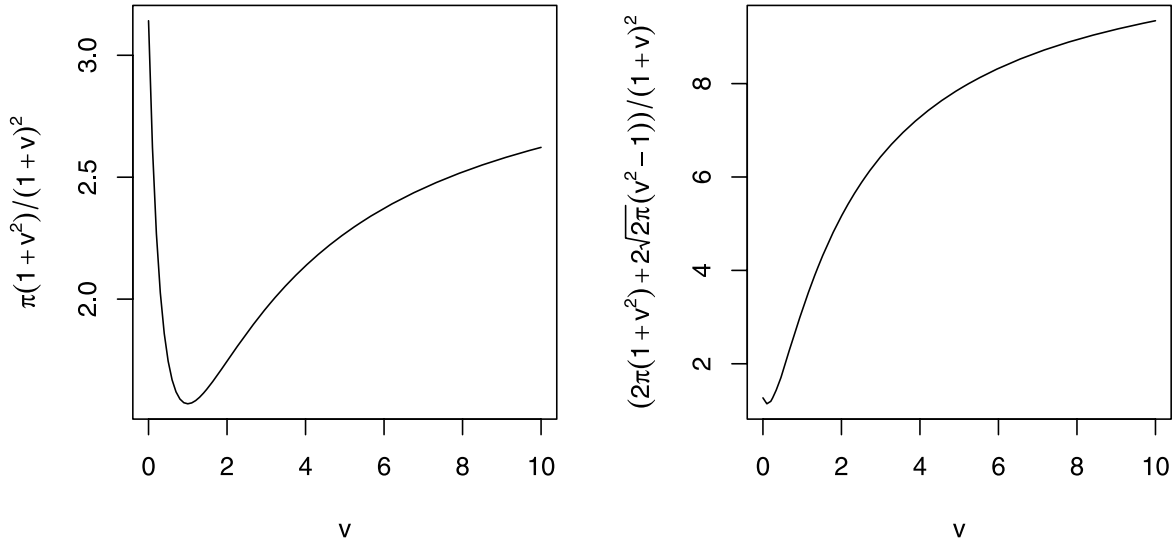
Figure 5. Plots of the asymptotic variances in (15).

$$H(\beta^*) = (1+v) \begin{pmatrix} (2\pi)^{-1/2} & 0 \\ 0 & (2\pi)^{-1/2} \end{pmatrix},$$

and

$$G(\beta^*) = \begin{pmatrix} \frac{1}{2}(1+v^2) & 0 \\ 0 & (1+v^2) + \sqrt{\frac{2}{\pi}}(v^2-1) \end{pmatrix}.$$

Thus, by Theorem 3, we have

$$(15) \qquad \sqrt{n} \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_+ \end{pmatrix} \to N \left(\begin{pmatrix} 0 \\ 1 \end{pmatrix}, \Sigma \right),$$

where

$$\Sigma = \frac{1}{(1+v)^2} \begin{pmatrix} \pi(1+v^2) & 0 \\ 0 & 2\pi(1+v^2) + 2\sqrt{2\pi}(v^2-1) \end{pmatrix}.$$

The asymptotic variances of coefficients shown in (15) depend on v. As shown in Fig. 5, the variances of both coefficients decrease as v increases for a while, then increase in v. Thus in this example the middle range values of v give smaller asymptotic variances.

5. REGULARIZED SOLUTION PATH OF THE BSVM WITH RESPECT TO v

In this section, we discuss how to obtain the entire solution path efficiently with respect to v. Using this path, we can compare the performances of the BSVM with different values of v without additional computational burden. [8] established the entire regularization path for the SVM for every value of λ. In the BSVM procedure, we have two parameters to choose, λ and v, and here we derive an algorithm that fits the BSVM with respect to v for a fixed λ.

We first categorize the observations according to their relative positions to the hyperplane $f(\boldsymbol{x}) = \pm 1$. In particular, let $\mathcal{E} = \{i : y_i f(\boldsymbol{x}_i) = 1\}$, $\mathcal{L} = \{i : y_i f(\boldsymbol{x}_i) < 1\}$, and $\mathcal{R} = \{i : y_i f(\boldsymbol{x}_i) > 1\}$. From (9)–(11), notice that

(16)   For any $i \in \mathcal{L}$, $\gamma_i = C, \delta_i = 0$, thus $\alpha_i = C$

(17)   For any $i \in \mathcal{R}$, $\gamma_i = 0, \delta_i = C$, thus $\alpha_i = -Cv$

(18)   For any $i \in \mathcal{E}$, α_i can be any number in $[-Cv, C]$.

For a fixed C, we start with a sufficiently large v which induces $y_i f(\boldsymbol{x}_i) \le 1, \forall i = 1, \dots, n$, and go down to a smaller v. As the value of v decreases, the memberships of $\mathcal{E}, \mathcal{L}$, and $\mathcal{R}$ change. We say that an *event* occurred when any point changes its membership. There are three kinds of events:

E1. A point from $\mathcal{L}$ has just entered $\mathcal{E}$.
E2. A point from $\mathcal{R}$ has just entered $\mathcal{E}$.
E3. One or more points from $\mathcal{E}$ have entered either $\mathcal{L}$ or $\mathcal{R}$.

Once an event occurs, the sets $\mathcal{E}$, $\mathcal{L}$, and $\mathcal{R}$ will stay stable for a while until the next event occurs. This is because, for an observation to pass through $\mathcal{E}$, its α_i must change from C to $-Cv$ or vice versa. Therefore, we denote by v_1 our starting point, and let $v_2 > v_3 > \cdots$ be the values of v at which each of the events occurs.

Given v_l, we next study how to obtain v_{l+1}, and establish paths of α_i for $v \in [v_l, v_{l+1}]$. Let $\tau_i = \alpha_i/v = (\gamma_i - v\delta_i)/v$ for $i = 1, \dots, n$ and $\tau_0 = b/v$. We use superscript or subscript l to denote anything given $v = v_l$. For now, we assume $\mathcal{E}^l \ne \emptyset$.

For $v_l > v > v_{l+1}$, we have

(19)

$$
\begin{aligned}
f(\boldsymbol{x}) &= f(\boldsymbol{x}) - \frac{v}{v_l}f^l(\boldsymbol{x}) + \frac{v}{v_l}f^l(\boldsymbol{x}) \\
&= v\left[\sum_{j=1}^{n}\tau_j y_j \boldsymbol{x}_j^T \boldsymbol{x} + \tau_0 - \tau_j^l y_j \boldsymbol{x}_j^T \boldsymbol{x} - \tau_0^l + \frac{1}{v_l}f^l(\boldsymbol{x})\right] \\
&= v\left[\sum_{j=1}^{n}(\tau_j - \tau_j^l)y_j \boldsymbol{x}_j^T \boldsymbol{x} + (\tau_0 - \tau_0^l) + \frac{1}{v_l}f^l(\boldsymbol{x})\right] \\
&= v\left[C\left(\frac{1}{v} - \frac{1}{v_l}\right)\sum_{j\in\mathcal{L}^l}y_j \boldsymbol{x}_j^T \boldsymbol{x} + \sum_{j\in\mathcal{E}^l}(\tau_j - \tau_j^l)y_j \boldsymbol{x}_j^T \boldsymbol{x}\right. \\
&\quad \left. + (\tau_0 - \tau_0^l) + \frac{1}{v_l}f^l(\boldsymbol{x})\right].
\end{aligned}
$$

The last equality in (19) follows from the fact that $\tau_j - \tau_j^l = C(\frac{1}{v} - \frac{1}{v_l})$ for $j \in \mathcal{L}^l$ and $\tau_j - \tau_j^l = 0$ for $j \in \mathcal{R}^l$. Thus, for $i \in \mathcal{E}^l$,

(20)

$$
\begin{aligned}
\frac{1}{v} &= \frac{1}{v}y_i f(\boldsymbol{x}_i) \\
&= C\left(\frac{1}{v} - \frac{1}{v_l}\right)\sum_{j\in\mathcal{L}^l}y_i y_j \boldsymbol{x}_j^T \boldsymbol{x}_i + \sum_{j\in\mathcal{E}^l}(\tau_j - \tau_j^l)y_i y_j \boldsymbol{x}_j^T \boldsymbol{x}_i \\
&\quad + y_i(\tau_0 - \tau_0^l) + \frac{1}{v_l}.
\end{aligned}
$$

Writing $\kappa_j = \tau_j - \tau_j^l$ for $j \in \{0\} \cup \mathcal{E}^l$, we have

$$
\sum_{j\in\mathcal{E}^l}\kappa_j y_i y_j \boldsymbol{x}_j^T \boldsymbol{x}_i + y_i\kappa_0 = \left(\frac{1}{v} - \frac{1}{v_l}\right)\left[1 - C\sum_{j\in\mathcal{L}^l}y_i y_j \boldsymbol{x}_j^T \boldsymbol{x}_i\right].
$$

Let m be the number of points in $\mathcal{E}^l$. We can rewrite (21) in a matrix form

$$
\boldsymbol{K}_l\boldsymbol{\kappa} + \kappa_0 \boldsymbol{y}_l = \left(\frac{1}{v} - \frac{1}{v_l}\right)\boldsymbol{d}_l,
$$

where $\boldsymbol{K}_l$ is the $m \times m$ matrix with ij-th entry $y_i y_j \boldsymbol{x}_j^T \boldsymbol{x}_i$ for $i,j \in \mathcal{E}^l$, and $\boldsymbol{\kappa}$, $\boldsymbol{y}_l$, and $\boldsymbol{d}_l$ are the $m \times 1$ matrices with i-th entry κ_i, y_i, and $1 - C\sum_{j\in\mathcal{L}^l}y_i y_j \boldsymbol{x}_j^T \boldsymbol{x}_i$ for $i \in \mathcal{E}^l$, respectively.

From (8), we have $\sum_{j=1}^{n}\tau_i y_i = 0$. Thus,

(21) $\quad 0 = \sum_{j=1}^{n}(\tau_j - \tau_j^l)y_j = \sum_{j\in\mathcal{E}^l}\kappa_j y_j + C\left(\frac{1}{v} - \frac{1}{v_l}\right)\sum_{j\in\mathcal{L}^l}y_j.$

Using the matrix form, we have

(22) $\qquad \boldsymbol{y}_l^T \boldsymbol{\kappa} = -C\left(\frac{1}{v} - \frac{1}{v_l}\right)\sum_{j\in\mathcal{L}^l}y_j.$

Combining (21) and (22), we have the linear equations

$$
\boldsymbol{A}_l\boldsymbol{\kappa}^* = \left(\frac{1}{v} - \frac{1}{v_l}\right)\boldsymbol{d}_l^*,
$$

where

$$
\boldsymbol{A}_l = \begin{pmatrix} 0 & \boldsymbol{y}_l^T \\ \boldsymbol{y}_l & \boldsymbol{K}_l \end{pmatrix}, \quad \boldsymbol{\kappa}^* = \begin{pmatrix} \kappa_0 \\ \boldsymbol{\kappa} \end{pmatrix}, \quad \boldsymbol{d}_l^* = \begin{pmatrix} -C\sum_{j\in\mathcal{L}^l}y_j \\ \boldsymbol{d}_l \end{pmatrix}.
$$

Define $\boldsymbol{s}_l = \boldsymbol{A}_l^{-1}\boldsymbol{d}_l^*$, and denote its entries by s_j for $j \in \mathcal{E}^l$, then we have

(23) $\qquad \boldsymbol{\kappa}^* = \left(\frac{1}{v} - \frac{1}{v_l}\right)\boldsymbol{s}_l \quad \text{for } j \in \{0\} \cup \mathcal{E}^l,$

which implies

(24) $\qquad \alpha_j = \left(\frac{\alpha_j^l - s_j^l}{v_l}\right)v + s_j^l \quad \text{for } j \in \mathcal{E}^l$

(25) $\qquad b = \left(\frac{b^l - s_0^l}{v_l}\right)v + s_0^l.$

Hence, α_j and b are piecewise linear in v.
Combining (19) and (23) gives

(26)

$$
\begin{aligned}
f(\boldsymbol{x}) &= \frac{v}{v_l}f^l(\boldsymbol{x}) + vC\left(\frac{1}{v} - \frac{1}{v_l}\right)\sum_{j\in\mathcal{L}^l}y_j \boldsymbol{x}_j^T \boldsymbol{x} \\
&\quad + \sum_{j\in\mathcal{E}^l}s_j^l y_j \boldsymbol{x}_j^T \boldsymbol{x} + b_0^l - \frac{v}{v_l}\left[\sum_{j\in\mathcal{E}^l}s_j^l y_j \boldsymbol{x}_j^T \boldsymbol{x} + b_0^l\right].
\end{aligned}
$$

Writing $h^l(\boldsymbol{x}) = \sum_{j\in\mathcal{E}^l}s_j^l y_j \boldsymbol{x}_j^T \boldsymbol{x} + b_0^l$, we have

(27)

$$
f(\boldsymbol{x}) = \frac{v}{v_l}\left[f^l(\boldsymbol{x}) - h^l(\boldsymbol{x})\right] + h^l(\boldsymbol{x}) + vC\left(\frac{1}{v} - \frac{1}{v_l}\right)\sum_{j\in\mathcal{L}^l}y_j \boldsymbol{x}_j^T \boldsymbol{x}.
$$

The path (24)–(27) continues until one of the following occurs.

P1. One of the observations in $\mathcal{L}^l$ or $\mathcal{R}^l$ attains $y_i f(\boldsymbol{x}_i) = 1$.
P2. One of the α_i for $i \in \mathcal{E}^l$ reaches a boundary ($-Cv$ or C).

Note that **P1** implies the event **E1** or **E2**, and **P2** precedes **E3** or they coincide. Hence, we can obtain v_{l+1} by choosing the largest $v < v_l$ for which any of **P1** or **P2** occurs. Since $f(\boldsymbol{x}_i) = 1/y_i = y_i$ when **P1** happens, from (27), we have

(28) $\qquad v_l y_i = v[f^l(\boldsymbol{x}) - h^l(\boldsymbol{x})] + v_l h^l(\boldsymbol{x})$
$$
+ v_l C\sum_{j\in\mathcal{L}^l}y_j \boldsymbol{x}_j^T \boldsymbol{x} - vC\sum_{j\in\mathcal{L}^l}y_j \boldsymbol{x}_j^T \boldsymbol{x}.
$$

Thus, v for which **P1** happens is

(29) $\qquad v = \dfrac{v_l y_i - v_l h^l(\boldsymbol{x}) - v_l C\sum_{j\in\mathcal{L}^l}y_j \boldsymbol{x}_j^T \boldsymbol{x}}{f^l(\boldsymbol{x}) - h^l(\boldsymbol{x}) - C\sum_{j\in\mathcal{L}^l}y_j \boldsymbol{x}_j^T \boldsymbol{x}}.$

Furthermore, for **P2** to happen, either $\alpha_i = -Cv$ or $\alpha_i = C$ should happen. From (24), this implies

$$
(30) \qquad v = \frac{v_l s_i^l}{s_i^l - Cv_l - \alpha_i^l}
$$

or

$$
(31) \qquad v = \frac{v_l(C - s_i^l)}{a_i^l - s_i^l}.
$$

Hence, given v_l, we compute (29), (30), and (31), then set the largest v among the ones smaller than v_l as v_{l+1}. For $v \in (v_{l+1}, v_l)$, the solutions are calculated by (24), (25), and (27). We repeat this procedure until v runs all the way down to zero to obtain the whole solution path for every value of v.

So far we assume $\mathcal{E}$ is nonempty. It is a reasonable assumption since we can force $\mathcal{E}$ to be nonempty, by selecting a good b. This is possible because b is not uniquely determined when $\mathcal{E}$ is empty. More specifically, suppose $\mathcal{E} = \emptyset$ for $v \in [v_0 - \epsilon, v_0]$, with $\epsilon > 0$. By (8), (16), and (17), we have

$$
0 = \sum_{i=1}^{n} (\gamma_i - v\delta_i) y_i = c \sum_{i \in \mathcal{L}} y_i - Cv \sum_{i \in \mathcal{R}} y_i,
$$

for $v \in [v_0 - \epsilon, v_0]$. Thus, we have

$$
\sum_{i \in \mathcal{L}} y_i = \sum_{i \in \mathcal{R}} y_i = 0.
$$

Now consider the objective function. Solving (1) with $g(u)$ in (4) is equivalent to minimizing

$$
\begin{aligned}
(32) \quad & \frac{1}{2}\|\boldsymbol{w}\|^2 + C\left[\sum_{i \in \mathcal{L}}(1 - y_i f(\boldsymbol{x}_i)) + \sum_{i \in \mathcal{R}} v(y_i f(\boldsymbol{x}_i) - 1) \right] \\
& = \frac{1}{2}\|\boldsymbol{w}\|^2 + C\left[c_L - vc_R - \sum_{i \in \mathcal{L}} y_i \boldsymbol{x}_i^T \boldsymbol{w} + v \sum_{i \in \mathcal{R}} y_i \boldsymbol{x}_i^T \boldsymbol{w} \right. \\
& \left. + \left(-\sum_{i \in \mathcal{L}} y_i + v \sum_{i \in \mathcal{R}} y_i \right) b \right],
\end{aligned}
$$

where c_L and c_R are the number of entries in $\mathcal{L}$ and $\mathcal{R}$, respectively. Note that b in (33) vanishes because $-\sum_{i \in \mathcal{L}} y_i + v \sum_{i \in \mathcal{R}} y_i = 0$. Hence, given $\boldsymbol{w}$, minimizer b could be any value in the set B, where

$$
\begin{aligned}
(33) \quad B = \Bigg\{ b \in \mathbb{R} : & \frac{1}{2}\|\boldsymbol{w}\|^2 + C \sum_{i=1}^{n} g(y_i f(\boldsymbol{x}_i)) = \frac{1}{2}\|\boldsymbol{w}\|^2 \\
& + \left[\sum_{i \in \mathcal{L}}(1 - y_i f(\boldsymbol{x}_i)) + \sum_{i \in \mathcal{R}} v(y_i f(\boldsymbol{x}_i) - 1) \right] \Bigg\},
\end{aligned}
$$

that is, b can take any value unless it moves any points from $\mathcal{L}$ to $\mathcal{R}$, or vice versa. Hence, we can take any b satisfying

$$
\begin{aligned}
y_i f(\boldsymbol{x}_i) &\leq 1 \quad \text{for } i \in \mathcal{L} \\
y_i f(\boldsymbol{x}_i) &\geq 1 \quad \text{for } i \in \mathcal{R},
\end{aligned}
$$

which is equivalent to

$$
\begin{aligned}
b &\leq 1 - \boldsymbol{x}_i^T \boldsymbol{w} \quad \text{for } i \in \mathcal{L}_+ \\
b &\geq -1 - \boldsymbol{x}_i^T \boldsymbol{w} \quad \text{for } i \in \mathcal{L}_- \\
b &\geq 1 - \boldsymbol{x}_i^T \boldsymbol{w} \quad \text{for } i \in \mathcal{R}_+ \\
b &\leq -1 - \boldsymbol{x}_i^T \boldsymbol{w} \quad \text{for } i \in \mathcal{R}_-,
\end{aligned}
$$

where $\mathcal{L}_+ = \mathcal{L} \cap \{i : y_i = 1\}$, $\mathcal{L}_- = \mathcal{L} \cap \{i : y_i = -1\}$, $\mathcal{R}_+ = \mathcal{R} \cap \{i : y_i = 1\}$, and $\mathcal{R}_- = \mathcal{R} \cap \{i : y_i = -1\}$. Letting

$$
\begin{aligned}
i_{L+} &= \arg\max_{i \in \mathcal{L}_+} \boldsymbol{x}_i^T \boldsymbol{w} \\
i_{L-} &= \arg\min_{i \in \mathcal{L}_-} \boldsymbol{x}_i^T \boldsymbol{w} \\
i_{R+} &= \arg\min_{i \in \mathcal{R}_+} \boldsymbol{x}_i^T \boldsymbol{w} \\
i_{R-} &= \arg\max_{i \in \mathcal{R}_-} \boldsymbol{x}_i^T \boldsymbol{w},
\end{aligned}
$$

we have

$$
\begin{aligned}
\max\{-1 - \boldsymbol{x}_{i_{L-}}^T \boldsymbol{w}, 1 - \boldsymbol{x}_{i_{R+}}^T \boldsymbol{w}\} &\leq b \\
&\leq \min\{1 - \boldsymbol{x}_{i_{L+}}^T \boldsymbol{w}, -1 - \boldsymbol{x}_{i_{R-}}^T \boldsymbol{w}\}.
\end{aligned}
$$

Without loss of generality, we can assume $1 - \boldsymbol{x}_{i_{L+}}^T \boldsymbol{w} \leq -1 - \boldsymbol{x}_{i_{R-}}^T \boldsymbol{w}$. Then take $b = 1 - \boldsymbol{x}_{i_{L+}}^T \boldsymbol{w}$. This b belongs to B and we have $i_{L+} \in \mathcal{E}$. Consequently, we choose b that induces $\mathcal{E} \neq \emptyset$. Hence the case of empty $\mathcal{E}$ is resolved.

In summary, one can get the entire solution path for the BSVM with respect to v as follows:

Step 1. Start with a sufficiently large v_0 and let $v_l = v_0$.
Step 2. For v_l, obtain the solution of the BSVM. If $\mathcal{E}^l$ is empty, choose b as either upper or lower bound of (34) so that $\mathcal{E}^l$ becomes nonempty.
Step 3. Calculate (29), (30), and (31), then set the minimum of them as v_{l+1}, at which the next event happens.
Step 4. For $v \in (v_{l+1}, v_l)$, compute the path using (27).
Step 5. If $v_{l+1} \leq 0$, then set $v_{l+1} = 0$ and obtain the solution of the BSVM for $v_{l+1} = 0$ and stop. Otherwise, then set $v_l = v_{l+1}$ and go to **Step 2**.

6. NUMERICAL RESULTS

In this section, numerical studies are carried out to examine the performance of the BSVM, BCM, and the RSVM [22]. We note that the RSVM with truncation location at 0 is equivalent to ψ-learning [16].

6.1 Simulation

In two simulated data sets, we generate training, tuning, and testing sets with sample sizes 100, 100, and 10^6, respectively. For each value of $v = 0, 0.1, 0.2, 0.5, 1, 2, 5, 10, 50$, the tuning parameter λ is chosen by a grid search based on the tuning error. The misclassification rate is calculated based on the test set to evaluate the performance. For comparison, we also include the misclassification rate when both v and λ are tuned. Each procedure is repeated for 100 times and the corresponding mean performance is reported.
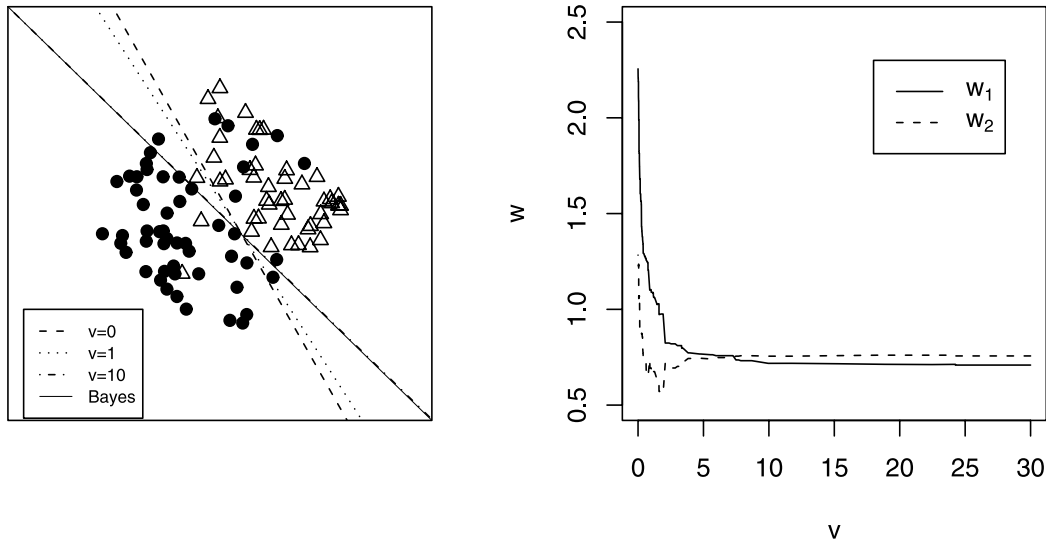
Figure 6. Left: Plot of different classification boundaries in Example 1. Right: Illustration of the solution path of w with respect to v in Example 1.

Table 1. Testing errors of the simulated Example 1

Method		Data contamination rates		
		0%	5%	10%
BSVM	$v = 0$	0.0150(0.0101)	0.0730(0.0156)	0.1289(0.0212)
(with	$v = 0.1$	0.0239(0.0165)	0.0747(0.0169)	0.1295(0.0191)
tuning set)	$v = 0.2$	0.0247(0.0162)	0.0753(0.0163)	0.1283(0.0183)
	$v = 0.5$	0.0243(0.0147)	0.0729(0.0138)	0.1254(0.0161)
	$v = 1$	0.0222(0.0128)	0.0707(0.0130)	0.1224(0.0148)
	$v = 2$	0.0186(0.0113)	0.0673(0.0107)	0.1176(0.0107)
	$v = 5$	0.0137(0.0080)	0.0620(0.0087)	0.1112(0.0072)
	$v = 10$	0.0107(0.0069)	0.0593(0.0066)	0.1091(0.0069)
	$v = 50$	0.0100(0.0073)	0.0586(0.0059)	0.1080(0.0062)
BSVM (both λ and v tuned)		0.0107(0.0069)	0.0586(0.0059)	0.1113(0.0075)
BCM		0.0095(0.0066)	0.0576(0.0053)	0.1079(0.0062)
RSVM	$s = -1$	0.0150(0.0103)	0.0649(0.0099)	0.1169(0.0136)
	$s = 0$	0.0161(0.0110)	0.0700(0.0136)	0.1225(0.0154)
Bayes Error		0.00	0.05	0.10

Example 1. The data are generated as follows. First, (x_1, x_2) is sampled from a square $\{(x_1, x_2) : -\sqrt{2} < x_1 + x_2 < \sqrt{2}, -\sqrt{2} < x_1 - x_2 < \sqrt{2}\}$. Then, set $y = 1$ if $x_1 + x_2 > 0$ and $y = -1$ otherwise. To illustrate the effect of outliers, we randomly flip the class membership of 0%, 5%, and 10% of data. A typical example of training data set and the resulting BSVM boundaries are plotted in left panel of Fig. 6. The corresponding solution paths of w are provided in the right panel of Fig. 6. Interestingly, the solution doesn't change once the value v gets sufficiently large. Note that performance of the RSVM is pretty good as well especially when there are outliers, but the BSVM with larger v works better.

Test error results are summarized in Table 1. Regarding to the effect of v, a larger v produces better results. This is not surprising because of the data structure of this example. Because the data points are aligned quite parallel to the true boundary, the observations far from the boundary reflects the overall structure of the data, resulting in favor to the BSVM with high v which uses information from those data far from the boundary. The BSVM with both λ and v tuned gives reasonable performance, which is close to the result of a large v. As the limit of the BSVM, the BCM gives the best performance in this example.

Example 2. We generate equal numbers of data points for class 1 and class -1. For class 1, 40%, 40%, and 20% of the observations are generated from $N((1, 0.5)^T, \sigma^2 I)$, $N((-3, 0.5)^T, \sigma^2 I)$, and $N((0, 1)^T, \Sigma)$, respectively, where I is 2×2 identity matrix and $\Sigma = \text{diag}((4\sigma)^2, (\sigma/3)^2)$. For

		Standard deviation	
Method		$\sigma = 0.3$	$\sigma = 0.5$
BSVM	$v = 0$	0.0052(0.0046)	0.0574(0.0177)
(with	$v = 0.1$	0.0055(0.0048)	0.0695(0.0212)
tuning set)	$v = 0.2$	0.0060(0.0054)	0.0749(0.0197)
	$v = 0.5$	0.0083(0.0059)	0.0857(0.0176)
	$v = 1$	0.0107(0.0060)	0.0954(0.0148)
	$v = 2$	0.0150(0.0075)	0.1073(0.0163)
	$v = 5$	0.0233(0.0100)	0.1164(0.0128)
	$v = 10$	0.0265(0.0108)	0.1212(0.0131)
	$v = 50$	0.0288(0.0097)	0.1231(0.0139)
BSVM (both λ and v tuned)		0.0060(0.0054)	0.0574(0.0177)
BCM		0.0267(0.0114)	0.1214(0.0174)
RSVM	$s = -1$	0.0052(0.0045)	0.0528(0.0126)
	$s = 0$	0.0039(0.0018)	0.0517(0.0121)
Bayes Error		0.000159	0.022104

Table 2. Testing errors of the simulated Example 2

Table 3. Testing errors of the real data example in Section 6.2.

Method		Testing errors
BSVM	$v = 0$	0.0203(0.0170)
	$v = 0.1$	0.0174(0.0178)
	$v = 0.2$	0.0145(0.0181)
	$v = 0.5$	0.0145(0.0181)
	$v = 1$	0.0145(0.0181)
	$v = 2$	0.0145(0.0181)
	$v = 5$	0.0145(0.0181)
	$v = 10$	0.0145(0.0181)
	$v = 50$	0.0145(0.0181)
BSVM (both λ and v tuned)		0.0174(0.0178)
BCM		0.0145(0.0181)
RSVM	$s = -1$	0.0203(0.0170)
	$s = 0$	0.0203(0.0170)

class 2, 40%, 40%, and 20% of the observations are generated from $N((3, -0.5)^T, \sigma^2 I)$, $N((-1, -0.5)^T, \sigma^2 I)$, and $N((0, -1)^T, \Sigma)$. We use two different values of σ, 0.3 and 0.5, and a typical example of the data when $\sigma = 0.3$ is plotted in Fig. 4. As shown in Table 2, the results are opposite to Example 1. The smaller values of v yield better results. This is not surprising considering the nature of this dataset. Since the information about observations near the boundary is critical for classification in this dataset, it is better to use more information about those observations. If we use large v, the data far from the boundary pull the decision boundary towards them, delivering a flat decision boundary which does not reflect well the data structure around the boundary. Notice that the standard SVM (BSVM with $v = 0$), the BSVM with both λ and v tuned, and the RSVM work reasonably well for this example.

6.2 Real data

In this section, we apply the BSVM and BCM to the lung cancer data described in [14]. In this data set, there are 12,625 genes with 17 normal subjects and 188 lung cancer patients. We first filter the genes using the ratio of the sample standard deviation and sample mean of each gene to obtain 316 genes. Then we standardize the genes so that each gene has sample mean 0 and sample standard deviation 1. We randomly divide subjects into three groups of training, tuning, and testing sets with the sample sizes 68, 68, and 69 respectively, and we build a model for each value of λ using the data in training set. Then λ is selected based on its performance on the tuning set by a grid search. Using the model with the selected λ, the misclassification rate on the testing set is calculated. This whole procedure is repeated for 10 times.

The results are reported in Table 3. As shown in the table, the BSVM with a large v and BCM perform slightly better than the standard SVM, while the RSVM does not improve the standard SVM. This may be due to the nature of this data set. However, the difference is not significantly large.

7. DISCUSSION

In this article, we propose the BCM as an alternative classifier to the SVM. To connect the BCM with the SVM, we study the BSVM which builds a continuous path between them. Moreover, we have shown Fisher consistency and asymptotic distributions of the solution of the BSVM. For computational implementation, we derive the entire solution path of the BSVM with respect to v.

We have shown several numerical examples to illustrate the effect of v. Our results indicate that the choice of v is indeed important for the performance of the BSVM. Although one may treat v as a tuning parameter, it will be more desirable to have a more efficient approach to select v. One possibility is to derive the GACV curve with respect to v and choose the value of v which minimizes the GACV.

The BCM has a nice interpretation and performs well in many situations. However, its linear loss function may emphasize too much on the correctly classified observations comparing to wrongly classified observations. Hence, one can consider to modify the loss function form of the BCM to reduce the loss imposed on correctly classified data. Further investigation is necessary.

APPENDIX

Proof of Theorem 1

Let $f = f(\boldsymbol{x})$, $p = P(\boldsymbol{x})$, and $A(f) = E[g(Yf(\boldsymbol{X})|\boldsymbol{X} = \boldsymbol{x}]$. First, we show that the minimizer f^* of $A(f)$ is on $[-1, 1]$. When $f > 1$, $A(f) = pv(f-1) + (1-p)(1+f) > 2(1-p) = A(1)$. Similarly, when $f < -1$, $A(f) = p(1-f) + (1-p)v(-f-1) > 2p = A(-1)$. Thus, $f^* \in [-1, 1]$. For $f \in [-1, 1]$, $A(f) = p(1-f) + (1-p)(1+f) = (1-2p)f + 1$. Hence $f = 1$ minimizes $A(f)$ if $p > 1/2$, and otherwise, $f = -1$

minimizes $A(f)$. Therefore, $\operatorname{argmin}_f A(f) = \operatorname{sign}[p - 1/2]$. This completes the proof. $\qquad\square$

Proof of Theorem 2

It is easy to see that $f_l \le k - 1$ for $l = 1, \ldots, k$. Thus, one can show that the problem reduces to

$$
(34) \qquad \max_{\boldsymbol{f}} \sum_{l=1}^{k} P_l(\boldsymbol{x}) f_l(\boldsymbol{x})
$$

$$
\text{s.t.} \qquad \sum_{l=1}^{k} f_l(\boldsymbol{x}) = 0; -1 \le f_l(\boldsymbol{x}) \le k - 1, \forall l.
$$

Thus, the solution satisfies $f_j^*(\boldsymbol{x}) = k - 1$ if $j = \operatorname{argmax}_j P_j(\boldsymbol{x})$ and -1 otherwise. $\qquad\square$

Proofs of Theorem 3 and Theorem 4

First we give several lemmas that we need to prove the theorems. We remove the proofs of the lemmas to save space.

Lemma 1 guarantees that there is a finite minimizer of $Q(\boldsymbol{\beta})$. Lemmas 2 and 3 establishes $s(\boldsymbol{\beta})$ and $H(\boldsymbol{\beta})$, which are considered first and second derivatives of $Q(\boldsymbol{\beta})$, respectively.

Lemma 1. *Suppose that **A1** and **A2** are satisfied. Then $Q(\boldsymbol{\beta}) \to \infty$ as $\|\boldsymbol{\beta}\| \to \infty$ and the minimizer $\boldsymbol{\beta}^*$ exists.*

Lemma 2. *Suppose that **A1** is satisfied. If $\boldsymbol{\beta}_+ \ne \mathbf{0}$, then*

$$
\frac{\partial Q(\boldsymbol{\beta})}{\partial \beta_j} = S(\boldsymbol{\beta})_j
$$

for $j = 0, \ldots, d$.

Lemma 3. *Suppose that **A1** is satisfied. If $\boldsymbol{\beta}_+ \ne \mathbf{0}$, then*

$$
\frac{\partial^2 Q(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_j \partial \boldsymbol{\beta}_k} = H(\boldsymbol{\beta})_{jk}
$$

for $j, k = 0, \ldots, d$.

Lemma 4. *Suppose that **A1** and **A3** are satisfied. Then $\boldsymbol{\beta}_+^* \ne \mathbf{0}$.*

The following lemma establishes the lower bound of $H(\boldsymbol{\beta}^*)$.

Lemma 5. *Suppose **A1**, **A3**, and **A4** are met. Then,*

$$
\boldsymbol{\beta}^T H(\boldsymbol{\beta}^*) \boldsymbol{\beta} \ge (1 + v) C_4 \|\boldsymbol{\beta}\|^2,
$$

where C_4 may depend on $\boldsymbol{\beta}^*$.

Lemma 6. *Assume **A1**–**A4** are satisfied. Then $Q(\boldsymbol{\beta})$ has a unique minimizer.*

With the lemmas in place, we are now ready to prove Theorems 2 and 3. For $\boldsymbol{\theta} = (\theta_0, \boldsymbol{\theta}_+)^T \in \mathbb{R}^{d+1}$, define

$$
\Lambda_n(\boldsymbol{\theta}) = n \left(q_{\lambda,n} \left(\boldsymbol{\beta}^* + \frac{\boldsymbol{\theta}}{\sqrt{n}} \right) - q_{\lambda,n}(\boldsymbol{\beta}^*) \right)
$$

and

$$
\Gamma_n(\boldsymbol{\theta}) = E \Lambda_n(\boldsymbol{\theta}).
$$

By Taylor series expansion,

$$
\Gamma_n(\boldsymbol{\theta}) = n \left(Q \left(\boldsymbol{\beta}^* + \frac{\boldsymbol{\theta}}{\sqrt{n}} \right) - Q(\boldsymbol{\beta}^*) \right)
$$
$$
+ \frac{\lambda}{2} \left(\|\boldsymbol{\theta}_+\|^2 + 2\sqrt{n} \boldsymbol{\beta}_+^{*T} \boldsymbol{\theta}_+ \right)
$$
$$
= \frac{1}{2} \boldsymbol{\theta}^T H(\tilde{\boldsymbol{\beta}}) \boldsymbol{\theta} + \frac{\lambda}{2} \left(\|\boldsymbol{\theta}_+\|^2 + 2\sqrt{n} \boldsymbol{\beta}_+^{*T} \boldsymbol{\theta}_+ \right),
$$

where $\tilde{\boldsymbol{\beta}} = \boldsymbol{\beta}^* + (t/\sqrt{n}) \boldsymbol{\theta}$ for some $0 < t < 1$. Define $D_{jk}(\boldsymbol{\alpha}) = H(\boldsymbol{\beta}^* + \boldsymbol{\alpha})_{jk} + H(\boldsymbol{\beta}^*)_{jk}$ for $0 \le j, k \le d$. Because $H(\boldsymbol{\beta})$ is continuous in $\boldsymbol{\beta}$, there exists $\delta_1 > 0$ such that $\|\boldsymbol{\alpha}\| < \delta_1$ implies $|D_{jk}(\boldsymbol{\alpha})| < \epsilon_1$ for any $\epsilon_1 > 0$ and $0 \le j, k \le d$. Then, for sufficiently large n such that $\|(t/\sqrt{n})\boldsymbol{\theta}\| < \delta_1$, we have

$$
\left| \boldsymbol{\theta}^T \left(H(\tilde{\boldsymbol{\beta}}) - H(\boldsymbol{\beta}^*) \right) \boldsymbol{\theta} \right| \le \sum_{j,k} |\theta_j| |\theta_k| \left| D_{j,k} \left(\frac{t}{\sqrt{n}} \theta \right) \right|
$$
$$
\le \epsilon_1 \sum_{j,k} |\theta_j| |\theta_k|
$$
$$
\le 2\epsilon_1 \|\boldsymbol{\theta}\|^2,
$$

resulting

$$
\frac{1}{2} \boldsymbol{\theta}^T H(\tilde{\boldsymbol{\beta}}) \boldsymbol{\theta} = \frac{1}{2} \boldsymbol{\theta}^T H(\boldsymbol{\beta}^*) \boldsymbol{\theta} + o(1).
$$

Considering $\lambda = o(n^{-1/2})$, we have

$$
\Gamma_n(\boldsymbol{\theta}) = \frac{1}{2} \boldsymbol{\theta}^T H(\boldsymbol{\beta}^*) \boldsymbol{\theta} + o(1).
$$

Now, let

$$
\boldsymbol{W}_n = \sum_{i=1}^{n} \left(-\psi(1 - Y_i f(\boldsymbol{X}_i; \boldsymbol{\beta}^*)) Y_i \tilde{\boldsymbol{X}}_i \right.
$$
$$
\left. + v\psi(Y f(\boldsymbol{X}; \boldsymbol{\beta}) - 1) Y_i \tilde{\boldsymbol{X}}_i \right).
$$

Observe that $E(\boldsymbol{W}_n) = S(\boldsymbol{\beta}^*) = 0$ and $E(\boldsymbol{W}_n \boldsymbol{W}_n^T) = \sum_{i=1}^{n} E[(\psi(1 - Y_i f(\boldsymbol{X}_i; \boldsymbol{\beta}^*)) + v^2 \psi(Y_i f(\boldsymbol{X}_i; \boldsymbol{\beta}^*) - 1)) \tilde{\boldsymbol{X}}_i \tilde{\boldsymbol{X}}_i^T]$. Hence, by central limit theorem, we have

$$
\frac{1}{\sqrt{n}} \boldsymbol{W}_n \to N(0, n G(\boldsymbol{\beta}^*))
$$

in distribution.

Now, we define

$$
R_{i,n}(\boldsymbol{\theta}) = g(Y_i f(\boldsymbol{X}_i; \boldsymbol{\beta}^* + \boldsymbol{\theta}/\sqrt{n})) - g(Y_i f(\boldsymbol{X}_i; \boldsymbol{\beta}^*))
$$
$$
+ \psi(1 - Y_i f(\boldsymbol{X}_i; \boldsymbol{\beta}^*)) Y_i f(\boldsymbol{X}_i; \boldsymbol{\theta}/\sqrt{n})
$$
$$
- v\psi(Y_i f(\boldsymbol{X}_i; \boldsymbol{\beta}^*) - 1) Y_i f(\boldsymbol{X}_i; \boldsymbol{\theta}/\sqrt{n}),
$$

which gives

$$\Lambda_n(\boldsymbol{\theta}) = \Gamma_n(\boldsymbol{\theta}) + \boldsymbol{W}_n^T\boldsymbol{\theta}/\sqrt{n} + \sum_{i=1}^n (R_{i,n}(\boldsymbol{\theta}) - ER_{i,n}(\boldsymbol{\theta})).$$

If we let $z = Y_i f(\boldsymbol{X}_i; \boldsymbol{\beta}^* + \boldsymbol{\theta}/\sqrt{n})$ and $a = Y_i f(\boldsymbol{X}_i; \boldsymbol{\beta}^*)$, we can write

$$
\begin{aligned}
R_{i,n}(\boldsymbol{\theta}) &= g(z) - g(a) + I\{a \le 1\}(z-a) - vI\{a > 1\}(z-a) \\
&= I\{z \le 1\}(1-z) + I\{z > 1\}v(z-1) \\
&\quad - I\{a \le 1\}(1-a) - I\{a > 1\}v(a-1) \\
&\quad + I\{a \le 1\}(z-a) - I\{a > 1\}v(z-a) \\
&= I\{z \le 1\}(1-z) + I\{z > 1\}v(z-1) \\
&\quad + I\{a \le 1\}(z-1) - I\{a > 1\}v(z-1) \\
&= [I\{z \le 1\} - I\{a \le 1\}](1-z) \\
&\quad + [I\{z > 1\} - I\{a > 1\}]v(z-1) \\
&\le (a-z)I\{z \le 1, a > 1\} + v(z-a)I\{z > 1, a \le 1\} \\
&\le \max\{1, v\}|z-a|I\{|1-a| \le |z-a|\}.
\end{aligned}
$$

Thus, we have

$$|R_{i,n}(\boldsymbol{\theta})|$$
$$\le \max\{1, v\}(|f(\boldsymbol{X}_i; \boldsymbol{\theta})|/\sqrt{n})I_{\{|1-Y_if(\boldsymbol{X}_i;\boldsymbol{\beta}^*)| \le |f(\boldsymbol{X};\boldsymbol{\theta})|/\sqrt{n}\}},$$

resulting

$$
\begin{aligned}
\sum_{i=1}^n &E|R_{i,n}(\boldsymbol{\theta}) - ER_{i,n}(\boldsymbol{\theta})|^2 \\
&= \sum_{i=1}^n [E(R_{i,n}(\boldsymbol{\theta}))^2 - (ER_{i,n}(\boldsymbol{\theta}))^2] \\
&\le \sum_{i=1}^n E(R_{i,n}(\boldsymbol{\theta}))^2 \\
&\le \sum_{i=1}^n E\left[\max\{1, v^2\}\left|\frac{f(\boldsymbol{X}_i; \boldsymbol{\theta})}{\sqrt{n}}\right|^2\right. \\
&\qquad \times \left. I_{\{|1-Y_if(\boldsymbol{X}_i;\boldsymbol{\beta}^*)| \le |f(\boldsymbol{X};\boldsymbol{\theta})|/\sqrt{n}\}}\right] \\
&\le \max\{1, v^2\}\|\boldsymbol{\theta}\|^2 E\left[(1 + \|\boldsymbol{X}\|^2)\right. \\
&\qquad \times \left. I_{\{|1-Y_if(\boldsymbol{X}_i;\boldsymbol{\beta}^*)| \le \sqrt{1+\|\boldsymbol{X}\|^2}\|\boldsymbol{\theta}\|/\sqrt{n}\}}\right].
\end{aligned}
$$

Note that **A1** implies that $E(\|\boldsymbol{X}\|^2) < \infty$. Thus, for any $\epsilon > 0$, there exists C_5 such that $E[(1 + \|\boldsymbol{X}\|^2)I_{\{\|\boldsymbol{X}\| > C_5\}}] < \epsilon/2$. Observe

$$
\begin{aligned}
E&\left[(1 + \|\boldsymbol{X}\|^2)I_{\{|1-Y_if(\boldsymbol{X}_i;\boldsymbol{\beta}^*)| \le \sqrt{1+\|\boldsymbol{X}\|^2}\|\boldsymbol{\theta}\|/\sqrt{n}\}}\right] \\
&\le E\left[(1 + \|\boldsymbol{X}\|^2)I_{\{\|\boldsymbol{X}\| > C_5\}}\right] \\
&\quad + (1 + c_5^2)P\left(|1 - Y_if(\boldsymbol{X}_i;\boldsymbol{\beta}^*)| \le \sqrt{1+C_5^2}\|\boldsymbol{\theta}\|/\sqrt{n}\right).
\end{aligned}
$$

The second term $(1 + c_5^2)P(|1 - Y_if(\boldsymbol{X}_i;\boldsymbol{\beta}^*)| \le \sqrt{1+C_5^2}\|\boldsymbol{\theta}\|/\sqrt{n})$ goes to zero as $n \to \infty$ because of **A1**. Thus, we have $\sum_{i=1}^n E|R_{i,n}(\boldsymbol{\theta}) - ER_{i,n}(\boldsymbol{\theta})|^2 \to 0$ as $n \to \infty$. Hence, we can write

$$\Lambda_n(\boldsymbol{\theta}) = \Gamma_n(\boldsymbol{\theta}) + \boldsymbol{W}_n^T\boldsymbol{\theta}/\sqrt{n} + o_P(1).$$

Now, we define $\boldsymbol{\eta}_n(\boldsymbol{\theta}) = -H(\boldsymbol{\beta}^*)^{-1}\boldsymbol{W}_n/\sqrt{n}$. Using Convexity Lemma in [17], we have

$$\Lambda_n(\boldsymbol{\theta}) = \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\eta}_n)^T H(\boldsymbol{\beta}^*)(\boldsymbol{\theta} - \boldsymbol{\eta}_n) - \frac{1}{2}\boldsymbol{\eta}^T H(\boldsymbol{\beta}^*)\boldsymbol{\eta} + r_n(\boldsymbol{\theta}),$$

where, for each compact set $K \in \mathbb{R}$,

$$\sup_{\theta \in K} |r_n(\boldsymbol{\theta})| \to 0$$

in probability. Since $\boldsymbol{\eta}_n$ converges in distribution, there exists a compact set K which contains B_ϵ, where B_ϵ is a closed ball with center $\boldsymbol{\eta}_n$ and radius ϵ with probability arbitrarily close to one. This gives

$$(35) \qquad \Delta_n = \sup_{\theta \in B_\epsilon} |r_n(\boldsymbol{\theta})| \to 0$$

in probability. Now consider the outside of the ball B_ϵ. Writing $\boldsymbol{\theta} = \boldsymbol{\eta}_n + \gamma\boldsymbol{u}$ and $\boldsymbol{\theta}^* = \boldsymbol{\eta}_n + \epsilon\boldsymbol{u}$ with $\gamma > \epsilon$ and a unit vector $\boldsymbol{u}$, Lemma 6 and convexity of Λ_n gives

$$
\begin{aligned}
\frac{\epsilon}{\gamma}&\Lambda_n(\boldsymbol{\theta}) + \left(1 - \frac{\epsilon}{\gamma}\right)\Lambda_n(\boldsymbol{\eta}_n) \\
&\ge \Lambda_n(\boldsymbol{\theta}^*) \\
&\ge \frac{1}{2}(\boldsymbol{\theta}^* - \boldsymbol{\eta}_n)^T H(\boldsymbol{\beta}^*)(\boldsymbol{\theta}^* - \boldsymbol{\eta}_n) - \frac{1}{2}\boldsymbol{\eta}^T H(\boldsymbol{\beta}^*)\boldsymbol{\eta} - \Delta_n \\
&\ge \frac{C_4}{2}\epsilon^2 + \Lambda_n(\boldsymbol{\eta}_n) - 2\Delta_n.
\end{aligned}
$$

Thus, we have

$$\frac{\epsilon}{\gamma}(\Lambda_n(\boldsymbol{\theta}) - \Lambda_n(\boldsymbol{\eta}_n)) \ge \frac{C_4}{2}\epsilon^2 - 2\Delta_n,$$

finally giving

$$\inf_{\|\boldsymbol{\theta}-\boldsymbol{\eta}_n\|>\epsilon} \Lambda_n(\boldsymbol{\theta}) \ge \Lambda_n(\boldsymbol{\eta}_n) + \left(\frac{C_4}{2}\epsilon^2 - 2\Delta_n\right).$$

By (35), we can take Δ_n so that $\frac{C_4}{2}\epsilon^2 - 2\Delta_n > 0$ with probability tending to one. Therefore, the minimum of Λ_n cannot occur at any $\boldsymbol{\theta}$ with $\|\boldsymbol{\theta} - \boldsymbol{\eta}_n\| > \epsilon$. Note that the minimizer of Λ_n is $\sqrt{n}(\hat{\boldsymbol{\beta}}_{\lambda,n} - \boldsymbol{\beta}^*)$. Hence we have

$$P(\|\sqrt{n}(\hat{\boldsymbol{\beta}}_{\lambda,n} - \boldsymbol{\beta}^*) - \boldsymbol{\eta}_n\| > \epsilon) \to 0$$

resulting

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{\lambda,n} - \boldsymbol{\beta}^*) \to \boldsymbol{\eta}_n$$

in probability. This completes the proof.

REFERENCES

[1] BAHADUR, R. (1966). A note on quantiles in large samples. *Annals of Mathematical Statistics* **37** 577–580. MR0189095

[2] BARTLETT, P., JORDAN, M., and MCAULIFFE, J. (2006). Convexity, classification, and risk bounds. *Journal of the American Statistical Association* **101** 138–156. MR2268032

[3] CHAUDHURI, P. (1991). Nonparametric estimates of regression quantiles and their local bahadur representation. *The Annals of Statistics* **19** 760–777. MR1105843

[4] COLLOBERT, R., SINZ, F., WESTON, J., and BOTTOU, L. (2006). Large scale transductive svms. *Journal of Machine Learning Research* **7** 1687–1712. MR2274421

[5] CRAMMER, K. and SINGER, Y. (2001). On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of Machine Learning Research* **2** 265–292.

[6] CRISTIANINI, N. and SHAWE-TAYLOR, J. (2000). *An Introduction to Support Vector Machines*. Cambridge University Press.

[7] FRIEDMAN, J., HASTIE, T., and TIBSHIRANI, R. (2000). Additive logistic regression: A statistical view of boosting. *The Annals of Statistics* **28** 337–407. MR1790002

[8] HASTIE, T., ROSSET, S., TIBSHIRANI, R., and ZHU, J. (2004). The entire regularization path for the support vector machine. *Journal of Machine Learning Research* **5(Oct)** 1391–1415. MR2248021

[9] KOO, J.-Y., LEE, Y., KIM, Y., and PARK, C. (2008). A bahadur representation of the linear support vector machine. *Journal of Machine Learning Research* **9** 1343–1368. MR2426045

[10] LEE, Y., LIN, Y., and WAHBA, G. (2004). Multicategory support vector machines, theory, and application to the classification of microarray data and satellite radiance data. *Journal of the American Statistical Association* **99** 67–81. MR2054287

[11] LIN, X., WAHBA, G., XIANG, D., GAO, F., KLEIN, R., and KLEIN, B. (2000). Smoothing spline anova models for large data sets with bernoulli observations and the randomized gacv. *The Annals of Statistics* **28** 1570–1600. MR1835032

[12] LIN, Y. (2004). A note on margin-based loss functions in classification. *Stat. and Prob. Letters* **68** 73–82. MR2064687

[13] LIU, Y. (2007). Fisher consistency of multicategory support vector machines. *Eleventh International Conference on Artificial Intelligence and Statistics* 289–296.

[14] LIU, Y., HAYES, D. N., NOBEL, A., and MARRON, J. S. (2008). Statistical significance of clustering for high dimension low sample size data. *Journal of the American Statistical Association* **103** 1281–1293.

[15] LIU, Y. and SHEN, X. (2006). Multicategory ψ-learning. *Journal of the American Statistical Association* **101** 500–509. MR2256170

[16] LIU, Y., SHEN, X., and DOSS, H. (2005). Multicategory ψ-learning and support vector machine: computational tools. *Journal of Comput. and Graphical Statistics* **14** 219–236. MR2137899

[17] POLLARD, D. (1991). Asymptotics for least absolute deviation regression estimators. *Econometric Theory* **7** 186–199. MR1128411

[18] SHEN, X., TSENG, G., ZHANG, X., and WONG, W. (2003). On ψ-learning. *Journal of the American Statistical Association* **98** 724–734. MR2011686

[19] VAPNIK, V. (1998). *Statistical Learning Theory*. Wiley. MR1641250

[20] WAHBA, G. (1999). Support vector machines, reproducing kernel hilbert spaces and the randomized GACV. In *Advances in Kernel Methods Support Vector Learning*. MIT Press, 69–88.

[21] WESTON, J. and WATKINS, C. (1999). Support vector machines for multi-class pattern recognition. In *Proceedings of the 7th European Symposium on Artificial Neural Networks (ESANN-99)*, M. Verleysen, Ed. Bruges, Belgium, 219–224.

[22] WU, Y. and LIU, Y. (2007). Robust truncated hinge loss support vector machines. *Journal of the American Statistical Association* **102** 974–983. MR2411659

Seo Young Park
Department of Statistics and Operations Research
CB3260, University of North Carolina, Chapel Hill, NC 27599
E-mail address: seoyoung@email.unc.edu

Yufeng Liu
Department of Statistics and Operations Research
Carolina Center for Genome Sciences
CB3260, University of North Carolina, Chapel Hill, NC 27599
E-mail address: yfliu@email.unc.edu