The Annals of Statistics
2018, Vol. 46, No. 6B, 3362–3389
https://doi.org/10.1214/17-AOS1661

ASSESSING ROBUSTNESS OF CLASSIFICATION USING AN ANGULAR BREAKDOWN POINT

BY JUNLONG ZHAO[1], GUAN YU[2] AND YUFENG LIU[2]

Beijing Normal University, State University of New York at Buffalo and University of North Carolina at Chapel Hill

Robustness is a desirable property for many statistical techniques. As an important measure of robustness, the breakdown point has been widely used for regression problems and many other settings. Despite the existing development, we observe that the standard breakdown point criterion is not directly applicable for many classification problems. In this paper, we propose a new breakdown point criterion, namely angular breakdown point, to better quantify the robustness of different classification methods. Using this new breakdown point criterion, we study the robustness of binary large margin classification techniques, although the idea is applicable to general classification methods. Both bounded and unbounded loss functions with linear and kernel learning are considered. These studies provide useful insights on the robustness of different classification methods. Numerical results further confirm our theoretical findings.

1. Introduction. Classification problems are commonly seen in practice. There are numerous classification methods available in the literature; see Hastie, Tibshirani and Friedman (2009) for a comprehensive review. Among the existing methods, large margin classification techniques, such as the Support Vector Machine (SVM) [Vapnik (1998)], have been extensively studied in recent years. Let $\mathcal{X}$ denote the domain of the p-dimensional vector of input variables X, and $\mathcal{Y}$ denote the class label set that equals $\{-1, 1\}$ for binary classification. Assume that the training data $\{(X_i, Y_i), 1 \le i \le n\}$ are i.i.d. copies of $(X, Y) \in \mathcal{X} \times \mathcal{Y}$ with the unknown distribution P.

Many classification methods can be formulated as solving an optimization problem. For binary classification, we aim to estimate a function $f(x) : R^p \to R$ and use $\operatorname{sign}(f(x))$ as the classification rule. Typically, large margin techniques can be fit in the general regularization framework which minimizes the objective function $n^{-1}\sum_{i=1}^n \ell(Y_i f(X_i)) + \lambda J(f)$, where $\ell(u)$ is the loss function,

Received April 2016; revised November 2017.

[1] Supported in part by the National Science Foundation of China Grants (No. 11471030, 61472475), the Fundamental Research Funds for the Central Universities and KLAS 130026507.

[2] Supported in part by US NIH Grants P01 CA-142538, R01 GM-126550 and NSF Grants DMS-1407241 and IIS-1632951.

MSC2010 subject classifications. Primary 62H30; secondary 62G35.

Key words and phrases. Breakdown point, classification, loss function, reproducing kernel Hilbert spaces, robustness.

$Y_i \in \{1, -1\}$, $J(f)$ is the penalty function on f and λ is the tuning parameter. Many loss functions have been proposed and studied in the literature. In particular, the 0–1 loss $\ell(u) = \mathbf{1}(u \le 0)$ is the theoretical loss corresponding to the misclassification error directly. Due to the difficulty of minimizing the objective function with the 0–1 loss, one often uses surrogate loss functions in practice. The hinge loss $\ell(u) = (1 - u)_+$ for the SVM [Vapnik (1998)], the exponential loss $\ell(u) = \exp(-u)$ for the AdaBoost [Freund and Schapire (1997)] and the deviance loss $\ell(u) = \log(1 + \exp(-u))$ for the penalized logistic regression [Lin et al. (2000)] are commonly used.

In practice, outliers are often encountered, which can greatly reduce the effectiveness of various methods. Therefore, robustness is a very important consideration in statistical modeling. For classification problems, it has been observed numerically that classifiers with the unbounded loss functions can be sensitive to outliers. For example, Biggio, Nelson and Laskov (2012) showed that a specifically selected outlier can significantly reduce the classification accuracy of SVM. To overcome this problem, various techniques have been developed using bounded loss functions to reach robustness [Liu and Shen (2006), Shen et al. (2003), Wu and Liu (2007)]. Several other authors also proposed various robust variants of SVM, such as Krause and Singer (2004), Xu, Crammer and Schuurmans (2006).

Robust classifiers are desirable for classification problems with potential outliers. However, a systematic comparison of different classification methods in terms of robustness is nontrivial. In the literature, there exist several robustness measures such as qualitative robustness [Hable and Christmann (2011), Hampel (1971)], influence function [Hampel (1974)] and breakdown point [Hampel (1971)]. As pointed by Hable and Christmann (2011), qualitative robustness mainly concerns the equicontinuity of the estimator. Influence function describes the effects of small deviations (the local stability of a statistical procedure) whereas the breakdown point takes into account the global reliability and describes the effects of large deviations [Ronchetti (1997)]. Despite commonly used in various settings, these robustness measures are not sufficient for classification. For the qualitative robustness, in recent years, Hable and Christmann (2011) considered the qualitative robustness of SVM. As to the influence function, Christmann and Steinwart (2004) considered the influence function in binary classification with convex losses. They showed that the influence function exists under some conditions, for example, $\ell(u)$ is twice continuously differentiable and either $\mathcal{X}$ or the kernel function $K(x, x)$ is bounded. For more general classification settings such as $\ell(u)$ being not differentiable or nonconvex, not much work has been developed on the influence function.

For the breakdown point, since the introduction of this concept, it has been extended for various settings [Donoho and Huber (1983), Genton and Lucas (2003), Hubert, Rousseeuw and Van Aelst (2008), Sakata and White (1995), Stromberg and Ruppert (1992)]. Among these works, the finite sample breakdown point

[Donoho and Huber (1983)] is simple and has been widely used. Besides this popular criterion, Genton and Lucas (2003) introduced a more general definition of breakdown point for different settings, such as times series, nonlinear regression, etc. According to Genton and Lucas (2003), an estimator breaks down if the remaining uncontaminated observations have no effect on the estimator any more. Despite the progress in different areas, the research for finite sample breakdown point in classification is limited. Kanamori, Fujiwara and Takeda (2014) developed a robust variant of the ν-SVM method [Scholkopf et al. (2000)] and considered the finite sample breakdown point of their method. In general, the breakdown point for classification problems has not yet been studied systematically.

To better understand the robustness of different classification methods, we consider the criterion of breakdown point in this paper. As will be shown in Section 2, the finite sample breakdown point, which is widely used in regression and other settings, is not suitable for classification problems in many cases. For classification, in contrast to the regression setting, the key effect of outliers is to change the classification boundary rather than the norm of coefficients in the classification function. Motivated from this, we propose a new criterion, namely angular breakdown point, to measure robustness of classification methods. The proposed angular breakdown point, as an extension of the finite-sample breakdown point to classification problems, is also a measurement on global reliability. We demonstrate that the proposed angular breakdown point provides new useful insights on robustness which cannot be obtained via the existing robustness measures. The angular breakdown point is studied for classification problems with bounded or unbounded loss functions. Our theoretical and numerical studies illustrate the robustness properties of different loss functions for both linear and kernel-based binary large margin classifiers. These results shed some lights on the potential advantages of bounded loss functions over unbounded ones.

The rest of this paper is organized as follows. In Section 2, we show the motivation and definition of angular breakdown point. In Section 3, we study the effect of outliers on linear classification, and the theoretical properties of angular breakdown point for binary classification with linear learning, where both bounded and unbounded loss functions are studied. In Section 4, the angular breakdown point for binary kernel learning with bounded or unbounded loss functions is considered. The simulation results and real data analysis are presented in Sections 5 and 6, respectively. In Section 7, we conclude this paper and discuss some potential applications of our proposed angular breakdown point criterion in data analysis. Selected proofs are shown in the Appendix. Other proofs are given in the online Supplementary Material [Zhao, Yu and Liu (2018)].

2. Motivation and definition of angular breakdown point.

Let $\mathbb{Z}_n = \{z_i : z_i = (x_i, y_i), i = 1, \dots, n\}$ denote n i.i.d. samples of $Z = (X, Y)$. For motivation, we first consider linear classification with a classification function $f(x) = b + \beta^T x$

where $\beta \in R^p$ and $b \in R$. For a large margin classification method with a loss function $\ell(u)$, let $\tilde{\beta}_0 = (b_0, \beta_0^T)^T$ be the population optimizer, that is,

$$\tilde{\beta}_0 = \arg \min_{b \in R, \beta \in R^p} E_Z[\ell((b + \beta^T X)Y)].$$

In practice, we estimate b_0 and β_0 by

$$(\hat{b}, \hat{\beta}) = \arg\min_{b,\beta} \frac{1}{n} \sum_{i=1}^{n} \ell(y_i(b + \beta^T x_i)) + \lambda J(f), \tag{2.1}$$

where λ is a tuning parameter and $J(f)$ is a regularization term.

2.1. *Motivation for angular breakdown point.* One of the most popular measures for the robustness of an estimator is the replacement finite-sample breakdown point (FBP) [Donoho and Huber (1983)]. If we use FBP to measure the robustness of $\hat{\beta}$, the breakdown point is defined as

$$\varepsilon^*(\hat{\beta}, \mathbb{Z}_n) = \min\left\{\frac{m}{n} : \sup_{\tilde{\mathbb{Z}}_n} \|\hat{\beta}(\tilde{\mathbb{Z}}_n) - \hat{\beta}(\mathbb{Z}_n)\| = \infty\right\}, \tag{2.2}$$

where $\tilde{\mathbb{Z}}_n$ is the contaminated sample obtained by replacing m of the original n observations $\mathbb{Z}_n$ with arbitrary values, $\hat{\beta}(\tilde{\mathbb{Z}}_n)$ is the estimate of β using the contaminated sample $\tilde{\mathbb{Z}}_n$ and $\| \cdot \|$ is the l_2 norm.

Although FBP is very effective for regression problems [Sakata and White (1995), Stromberg and Ruppert (1992)], the definition of the breakdown point in (2.2) is not suitable for classification problems. For binary classification, when a large margin classifier as in (2.1) is used, a new observation x is classified according to $\mathrm{sign}(\hat{b} + \hat{\beta}^T x)$. In contrast to regression, the scale of $\hat{\beta}$ (i.e., $\|\hat{\beta}\|$) does not directly reflect the classification performance. Compared with $\|\hat{\beta}\|$, the direction of $\hat{\beta}$ (i.e., $\hat{\beta}/\|\hat{\beta}\|$) plays a key role in classification. Even if $\|\hat{\beta}(\tilde{\mathbb{Z}}_n)\|$ is very large, the decision boundary acquired by $\hat{\beta}(\tilde{\mathbb{Z}}_n)$ can be close to the true boundary and, therefore, the classification performance can still be excellent. Another major drawback of FBP for classification is that $\|\hat{\beta}(\tilde{\mathbb{Z}}_n) - \hat{\beta}(\mathbb{Z}_n)\| = \infty$ is often unattainable for classification (see Section 3.1). In fact, for a large margin classifier, the main effect of outliers is to change the direction or angle of the estimate rather than the norm. As a result, an alternative criterion to quantify the breakdown point for classification problems is needed. We illustrate this by the following toy example.

Toy example. Consider a linear classification problem. We assume that the covariate vector $X|Y$ follows the normal distribution $N(\mathrm{sign}(Y)u_0, I_2)$, where $Y \in \{1, -1\}$ and $u_0 = (1, 0)$. We set the sample size for each group to be 100. For the positive class (i.e., $Y = 1$), we replace one observation by an outlier generated from

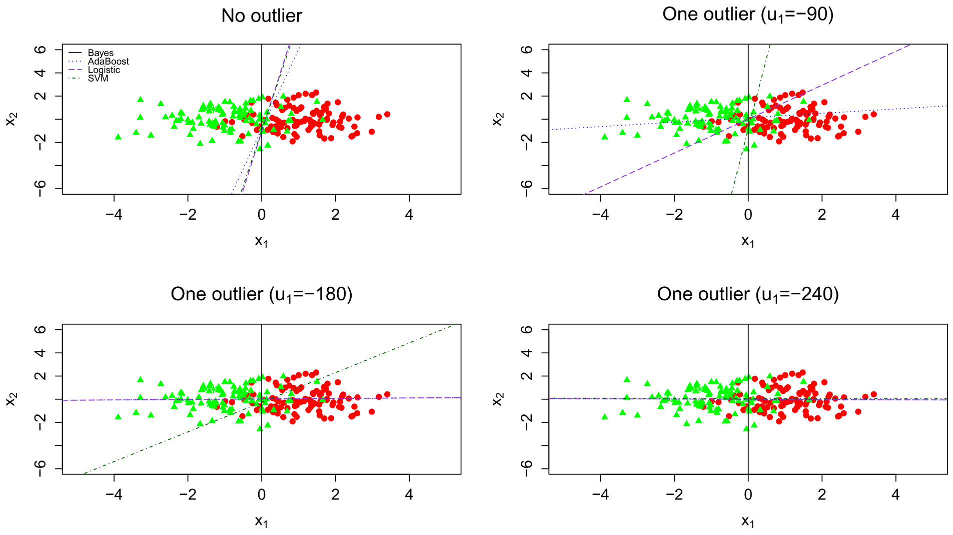

FIG. 1. *Illustration of the effect of one outlier for the toy example. As the outlier gets more extreme, the estimated decision boundaries become near orthogonal to the Bayes boundary.*

the normal distribution $N((u_1, 0)^T, I_2)$ with $u_1 \in \{-90, -180, -240\}$. For this example, three loss functions are considered: the exponential loss $\ell(u) = \exp(-u)$ used in AdaBoost, the deviance loss $\ell(u) = \log(1 + \exp(-u))$ for logistic regression and the hinge loss $\ell(u) = (1 - u)_+$ for the SVM. We set the tuning parameter $\lambda = 0$ for both AdaBoost and logistic regression. For the SVM, we set the tuning parameter $\lambda = 1/200$. Denote $\hat{\beta}_{\text{ada}}$, $\hat{\beta}_{\log}$ and $\hat{\beta}_{\text{svm}}$ as the estimates obtained by these three methods.

Figure 1 shows the decision boundaries of the Bayes classifier, AdaBoost, logistic regression and SVM for four cases. When there is no outlier, the Bayes decision boundary is $x_1 = 0$. The decision boundaries of AdaBoost, logistic regression and SVM are close to the optimal Bayes boundary. As the effect of the outlier increases (i.e., u_1 decreases), the decision boundaries of these three methods change significantly and the corresponding classification errors increase. For the case with $u_1 = -240$, their decision boundaries are almost orthogonal to the optimal Bayes decision boundary. For that case, the classification errors of AdaBoost, logistic regression and SVM are 0.45, 0.45 and 0.445, respectively. Although these three methods tend to have very poor classification performance as the effect of the outlier increases, we observe that $\|\hat{\beta}_{\text{ada}}\|$, $\|\hat{\beta}_{\log}\|$ and $\|\hat{\beta}_{\text{svm}}\|$ are always bounded. Therefore, the definition of the breakdown point in (2.2) is not effective for this problem. In addition, we check the inner product between these estimates ($\hat{\beta}_{\text{ada}}$, $\hat{\beta}_{\log}$ and $\hat{\beta}_{\text{svm}}$) and the theoretical best coefficient vector $\beta_0 = (1, 0)^T$. We found that the inner products decrease dramatically as the effect of the outlier increases. In the case with $u_1 = -240$, all inner products are negative, which indicates that the angles between the estimates ($\hat{\beta}_{\mathrm{ada}}$, $\hat{\beta}_{\log}$ and $\hat{\beta}_{\mathrm{svm}}$) and β_0 are larger than $\pi/2$. In this case, classification by these methods is completely failed due to one extreme outlier.

2.2. *Definition of angular breakdown point.* Motivated by the above toy example and the effect of outliers on classification which will be theoretically studied in Section 3.1, we propose the following novel angular breakdown point to quantity the robustness of large margin classification methods.

DEFINITION 1 (Population anglular breakdown point). The angular breakdown point for large margin classification is defined by

$$\varepsilon(\beta_0, \mathbb{Z}_n) = \min\left\{\frac{m}{n} : \hat{\beta}(\tilde{\mathbb{Z}}_n) \in S_0^-\right\},$$

where $S_0^- = \{\beta : \beta^T \beta_0 \leq 0\}$.

As a remark, we note that the angular breakdown point represents the minimum fraction of outliers needed such that the angle between the estimated coefficient $\hat{\beta}(\tilde{\mathbb{Z}}_n)$ and the true coefficient β_0 is at least $\pi/2$, the case when the classification method can be equivalent to the random guessing or have low discriminating power depending on the distribution of the uncontaminated sample $\mathbb{Z}_n$. In practice, the true coefficient β_0 is unknown and the angular breakdown point in Definition 1 is intractable computationally. To assess the robustness of the estimate, we define the following sample angular breakdown point, considering the difference between estimates with and without outliers. This is similar to the traditional breakdown point. Without loss of generality, we assume that the estimate of β using the original sample $\mathbb{Z}_n$, denoted as $\hat{\beta}(\mathbb{Z}_n)$, is nonzero throughout this paper.

DEFINITION 1′ (Sample angular breakdown point). The sample angular breakdown point for large margin classification is defined by

$$\varepsilon(\hat{\beta}, \mathbb{Z}_n) = \min\left\{\frac{m}{n} : \hat{\beta}(\tilde{\mathbb{Z}}_n) \in \hat{S}_0^-\right\} \qquad \text{where } \hat{S}_0^- = \{\beta : \beta^T \hat{\beta}(\mathbb{Z}_n) \leq 0\}.$$

As we will see below, the sample angular breakdown point generally has the same properties as those of the population angular breakdown point. In Sections 3 and 4, we study the theoretical properties of our proposed angular breakdown point.

3. Angular breakdown point in linear classification. In this section, we first study the effect of outliers on linear classification theoretically. Then we study the theoretical properties of the proposed angular breakdown point for binary classification with linear learning, where both bounded and unbounded loss functions are studied. Before proceeding further, we introduce some notation. Let $\mathbb{Z}_{n-m} = \{z_i = (x_i, y_i), i = 1, \dots, n-m\}$ and $\mathbb{Z}_m^o = \{z_i^o = (x_i^o, y_i^o), i = 1, \dots, m\}$ denote the $n-m$ uncontaminated and m contaminated observations, respectively, with $\tilde{\mathbb{Z}}_n = \mathbb{Z}_{n-m} \cup \mathbb{Z}_m^o$ representing the whole sample. Denote $\tilde{\beta} = (b, \beta^T)^T$. Then the objective function for the linear binary classification with sample $\tilde{\mathbb{Z}}_n$ can be formulated as

$$
\begin{aligned}
L_{\lambda,n}(\tilde{\beta}, \tilde{\mathbb{Z}}_n) &= \left[\frac{1}{n}\sum_{i=1}^{n-m} \ell\bigl(y_i(b + \beta^T x_i)\bigr) + \lambda J(\beta)\right] + \frac{1}{n}\sum_{i=1}^{m} \ell\bigl(y_i^o(b + \beta^T x_i^o)\bigr) \\
&:= G_{\lambda,n}(\tilde{\beta}, \mathbb{Z}_{n-m}) + F_n(\tilde{\beta}, \mathbb{Z}_m^o),
\end{aligned} \tag{3.1}
$$

where $G_{\lambda,n}(\tilde{\beta}, \mathbb{Z}_{n-m})$ and $F_n(\tilde{\beta}, \mathbb{Z}_m^o)$ are two terms only involving the uncontaminated and contaminated observations, respectively. We assume that the penalty function $J(\beta)$ satisfies conditions: (1) $J(\beta) \geq 0$ and $J(\beta) = J(-\beta)$; (2) $J(\beta) = 0$ if and only if $\beta = 0$ and (3) $J(\beta) \to \infty$, as $\|\beta\| \to \infty$.

3.1. *Effect of outliers on linear classification.* To follow up the toy example, we now theoretically study the effect of outliers on linear classification. To this end, we need to introduce linearly separable datasets. A dataset $D = \{(x_i, y_i), i = 1, \dots, n\} \subseteq \mathcal{X} \times \{-1, 1\}$ of binary classification is linearly separable, if there exists a hyperplane $\alpha^T x + a = 0$ for some $\alpha \in R^p$ and $a \in R$ such that $(\alpha^T x_i + a)y_i > 0$ for any $(x_i, y_i) \in D$.

PROPOSITION 1. *Suppose that the nonnegative loss function $\ell(u)$ satisfies the following conditions*: (i) $\ell(0) < \infty$; (ii) $\lim_{u\to-\infty} \ell(u) = \infty$. *Then the following two conclusions hold*: (1) *For the original observations $\mathbb{Z}_n$ and any contaminated observations $\mathbb{Z}_m^o$, we have $\|\hat{\beta}(\mathbb{Z}_n)\| < \infty$ and $\|\hat{\beta}(\tilde{\mathbb{Z}}_n)\| < \infty$ for any $\lambda > 0$*; (2) *If neither $\mathbb{Z}_n$ nor $\tilde{\mathbb{Z}}_n$ is linearly separable, we have $\|\hat{\beta}(\mathbb{Z}_n)\| < \infty$ and $\|\hat{\beta}(\tilde{\mathbb{Z}}_n)\| < \infty$ for $\lambda = 0$. Therefore, $\hat{\beta}(\tilde{\mathbb{Z}}_n)$ does not break down in terms of the breakdown point defined in* (2.2).

In general, $\tilde{\mathbb{Z}}_n$ is not linearly separable when there exists outliers. Furthermore, the commonly used methods such as the SVM, penalized logistic regression, AdaBoost and the least square loss all satisfy the assumptions in Proposition 1. The estimates of these methods will not break down in terms of the breakdown point defined in (2.2). However, as shown in the toy example, these methods can be sensitive to outliers and break down if an extreme outlier exists. Thus, even if we move the outliers arbitrarily, the norm of $\hat{\beta}$ can be still finite and, therefore, the traditional breakdown point (2.2) will not be effective for classification problems. In addition, we point out that the general definition of breakdown point proposed by Genton and Lucas (2003) can be also ineffective here. According to the general definition, an estimator breaks down if the uncontaminated sample does not affect the estimator any more. Since (3.3) below shows that the estimates of SVM, penalized logistic regression, AdaBoost and the least square loss are always affected by the remaining uncontaminated sample, these methods cannot be viewed as a breakdown. However, we can see from Figure 1 that the classification boundaries of these methods are badly affected and the corresponding classification errors are close to 0.5, the case of random guessing.

From both Proposition 1 and the toy example, we see that $\|\hat{\beta}(\tilde{\mathbb{Z}}_n) - \hat{\beta}(\mathbb{Z}_n)\| = \infty$ is not attainable in general, and thus the traditional breakdown point in (2.2) is not applicable for classification problems. Given $\|\hat{\beta}(\tilde{Z}_n)\| < \infty$, since there are at least two observations z_{i_1}, z_{i_2} such that $y_{i_1} = 1$ and $y_{i_2} = -1$, one can check that $|\hat{b}| < \infty$ under the condition of Proposition 1. Therefore, without loss of generality, we assume that the minimization of the objection function in (3.1) is taken over the set $\Delta_{BL} = \{(b, \beta), |b| < \infty, \beta \in R^p\}$ to simplify the analysis.

To further illustrate the effect of outliers, we first consider the case with a single outlier (i.e., $m = 1$) denoted by $z_1^o = (x_1^o, y_1^o)$ with $y_1^o \in \{1, -1\}$. Then $\tilde{\mathbb{Z}}_n = \mathbb{Z}_{n-1} \cup \{z_1^o\}$ and

$$
\begin{aligned}
L_{\lambda,n}(\tilde{\beta}, \tilde{\mathbb{Z}}_n) &= \left[\lambda J(\beta) + \frac{1}{n}\sum_{i=1}^{n-1} \ell\big((b + \beta^T x_i)y_i\big)\right] + \frac{1}{n}\ell\big((b + \beta^T x_1^o)y_1^o\big) \\
&:= G_{\lambda,n}(\tilde{\beta}, \mathbb{Z}_{n-1}) + F_n\big(\tilde{\beta}, \mathbb{Z}_1^o\big),
\end{aligned} \tag{3.2}
$$

where $\mathbb{Z}_1^o = z_1^o$. Denote the minimizer of (3.2) by

$$
\big(\hat{b}, \hat{\beta}(\tilde{\mathbb{Z}}_n)\big) = \arg \min_{(b,\beta)\in\Delta_{BL}} L_{\lambda,n}(\tilde{\beta}, \tilde{\mathbb{Z}}_n).
$$

Assume that $\ell(u)$ is a nonnegative, unbounded and continuous decreasing function with $\lim_{u\to-\infty} \ell(u) = \infty$. To better understand the effect of this outlier, we set $\|x_1^o\| \to \infty$. Note that for any β with $\beta^T x_1^o y_1^o / \|x_1^o\| < 0$, we have $\ell((b + \beta^T x_1^o)y_1^o) \to \infty$ as $\|x_1^o\| \to \infty$ for any bounded b. As a result, the minimizer $(\hat{b}, \hat{\beta}(\tilde{\mathbb{Z}}_n))$ of $L_{\lambda,n}(\tilde{\beta}, \tilde{\mathbb{Z}}_n)$ must satisfy $\hat{\beta}(\tilde{\mathbb{Z}}_n) \in S^+_{z_1^o} := \{\beta : \beta^T \bar{x}_1^o y_1^o \geq 0\}$, where $\bar{x}_1^o = x_1^o / \|x_1^o\|$. Therefore, the effect of the outlier z_1^o is equivalent to imposing a constraint on the feasible solution. Specifically, it can be rewritten as

$$
\min_{(b,\beta)\in\Delta_{BL}} G_{\lambda,n}(\tilde{\beta}, \mathbb{Z}_{n-1}) \qquad \text{s.t. } \beta \in S^+_{z_1^o}. \tag{3.3}
$$

To further study (3.3), we observe that the set $S^+_{z_1^o}$ is a cone, that is, if $\beta \in S^+_{z_1^o}$, then $c\beta \in S^+_{z_1^o}$ for any constant $c \geq 0$. For $\lambda > 0$, one can see that $\|\hat{\beta}_\lambda(\tilde{\mathbb{Z}}_n)\|$ is still finite, based on the fact that $G_{\lambda,n}(0, \mathbb{Z}_{n-1}) < \infty$ and $G_{\lambda,n}(\tilde{\beta}, \mathbb{Z}_{n-1}) = \infty$ with $\|\beta\| = \infty$. For $\lambda = 0$, we have the same conclusion by Proposition 1.

When $\|x_1^o\|$ is large, as shown in (3.3), the main effect of the contaminated observation (x_1^o, y_1^o) for large margin classifiers is to impose a constraint on the feasible solution, equivalently, to change the direction of $\hat{\beta}(\tilde{\mathbb{Z}}_n)$ rather than its norm. When $m = 1$, $\hat{\beta}(\tilde{\mathbb{Z}}_n)$ belongs to the feasible set $S^+_{z^o_1}$ controlled by the outlier (x^o_i, y^o_i), and it also depends on the uncontaminated data set $\mathbb{Z}_{n-1}$. Since it is difficult to measure the exact deviation of $\hat{\beta}(\tilde{\mathbb{Z}}_n)$ from the theoretical optimizer $\beta_0 \in R^p$, we consider the worst outlier by maximizing the minimum angle between β_0 and $S^+_{z^o_1}$, that is,

$$\max_{z^o_1} \min_{\beta \in S^+_{z^o_1}} \angle(\beta_0, \beta). \tag{3.4}$$

Note that (3.4) is equivalent to

$$\min_{z^o_1} \max_{\beta \in S^+_{z^o_1}} \beta_0^T \beta / \big(\|\beta_0\| \|\beta\|\big). \tag{3.5}$$

We define $\beta_0^T \beta / (\|\beta_0\| \|\beta\|) = 0$, if $\beta = 0$ or $\beta_0 = 0$. When $z^o_1 = (x^o_1, y^o_1)$ satisfies $x^o_1 y^o_1 = -c_1 \cdot \beta_0$ for any $c_1 > 0$, one can show that (3.4) equals to the optimal value $\pi/2$, and equivalently (3.5) equals to 0. The assumption that $\|x^o_1\| \to \infty$ is satisfied by setting $c_1 \to \infty$. For this worst outlier, since $\hat{\beta}(\tilde{\mathbb{Z}}_n) \in S^+_{z^o_1}$, we have $(\hat{\beta}(\tilde{\mathbb{Z}}_n))^T \beta_0 \leq 0$, that is, $\angle(\beta_0, \hat{\beta}(\tilde{\mathbb{Z}}_n)) \geq \pi/2$.

In general, if there are m outliers $\mathbb{Z}^o_m$ such that $\|x^o_i\| \to \infty$ for any $i \in \{1, 2, \ldots, m\}$, we define $S^+_{\mathbb{Z}^o_m} = \bigcap_{i=1}^m S^+_{z^o_i}$, where $S^+_{z^o_i}$'s are similarly defined as $S^+_{z^o_1}$. The optimal solution $\hat{\beta}(\tilde{\mathbb{Z}}_n)$ is constrained in $S^+_{\mathbb{Z}^o_m}$, and it is also affected by the uncontaminated sample $\mathbb{Z}_{n-m}$. Thus, it is reasonable to consider the worst $\bar{\mathbb{Z}}^o_m$ defined as

$$\bar{\mathbb{Z}}^o_m = \arg\min_{\mathbb{Z}^o_m} \Big[\sup_{\beta \in S^+_{\mathbb{Z}^o_m}} \beta_0^T \beta / \big(\|\beta_0\| \|\beta\|\big) \Big] := \arg\min_{\mathbb{Z}^o_m} A(\mathbb{Z}^o_m). \tag{3.6}$$

We can check that the optimal solution in (3.6) is achieved at $\bar{\mathbb{Z}}^o_m := \{z^o_i = (x^o_i, y^o_i) : x^o_i y^o_i = -c_i \beta_0, 0 < c_i \to \infty, 1 \leq i \leq m\}$ and $A(\bar{\mathbb{Z}}^o_m) \leq 0$. Therefore, for any possible $\hat{\beta}(\tilde{\mathbb{Z}}_n)$, we have $(\hat{\beta}(\tilde{\mathbb{Z}}_n))^T \beta_0 \leq 0$, that is, the angle between $\hat{\beta}(\tilde{\mathbb{Z}}_n)$ and the true coefficient β_0 is at least $\pi/2$.

In summary, as shown in the above theoretical study, for binary linear classification, the main effect of outliers for large margin classifiers is to impose a constraint on the feasible solution, equivalently, to change the direction of $\hat{\beta}(\tilde{\mathbb{Z}}_n)$ rather than its norm.

3.2. *Large margin classifiers with unbounded loss functions.*

In this section, we evaluate the angular breakdown point for different loss functions. We make the following assumption:

(A1) Suppose that $\ell(u)$ is a decreasing and continuous function with $\lim_{u \to \infty} \ell(u) = 0$ and $\lim_{u \to -\infty} \ell(u) = C_l \leq \infty$.

The assumption (A1) is a very weak assumption, which covers many commonly used loss functions such as the hinge loss for the SVM, the deviance loss for logistic regression and the exponential loss for AdaBoost. For an unbounded loss with $C_l = \infty$, we have the following conclusion.

THEOREM 1. (i) *Assume that* $\ell(u)$ *satisfies* (A1) *with* $C_l = \infty$. *Then the population angular breakdown point* $\varepsilon(\beta_0, \mathbb{Z}_n) = 1/n$ *for binary classification and the same is true for the sample angular breakdown point in Definition* 1′. (ii) *The same conclusion holds for the square loss* $(1 - y(b + \beta^T x))^2$.

Theorem 1 indicates that for linear binary classification, the angular breakdown point for methods with an unbound loss is $1/n$, that is, a single outlier is sufficient to result in angular breakdown. In fact, from the proof of Theorem 1, we can check that the conclusion of Theorem 1 still holds, if the condition $\lim_{u\to\infty} \ell(u) = 0$ in (A1) is relaxed to be $\lim_{u\to\infty} \ell(u) = C_r < \infty$.

3.3. *Large margin classifiers with bounded loss functions.* In this section, we study the proposed angular breakdown point for a bounded loss with $C_l < \infty$ in binary classification. This includes the sigmoid loss function $\ell(u) = (1 + e^u)^{-1}$ [Mason et al. (2000)], the ψ-loss used in ψ-learning [Shen et al. (2003)] and the truncated hinge loss in the robust SVM [Wu and Liu (2007)]. Recall that the objective function for the linear binary classification with sample $\tilde{\mathbb{Z}}_n$ can be formulated as

$$L_{\lambda,n}(\tilde{\beta}, \tilde{\mathbb{Z}}_n) = G_{\lambda,n}(\tilde{\beta}, \mathbb{Z}_{n-m}) + F_n(\tilde{\beta}, \mathbb{Z}^o_m).$$

When a bounded loss $\ell(\cdot)$ is used, a nonconvex optimization is encountered and the global minimizer may not be achieved. The breakdown point of global minimizer is analyzed in this paper. However, the breakdown point of local minimizer is still unclear and deserves further study. We would like to point out that asymptotically, as $n \to \infty$ and $m/n \to 0$, bounded loss functions are more robust than unbounded ones. In fact, due to $\ell(u) \le C_l < \infty$, we have $L_{\lambda,n}(\tilde{\beta}, \tilde{\mathbb{Z}}_n) - G_{\lambda,n}(\tilde{\beta}, \mathbb{Z}_{n-m}) \to 0$ as $n \to \infty$. Therefore, the effect of outliers disappears when $C_l < \infty$ as $n \to \infty$. Note that m can be finite or increase to infinity with a lower order than n.

We now focus on the finite sample analysis of the angular breakdown point for bounded loss functions. Recall that $(\hat{b}, \hat{\beta}(\tilde{\mathbb{Z}}_n)) = \arg\min_{\tilde{\beta}\in\Delta_{BL}} L_{\lambda,n}(\tilde{\beta}, \tilde{\mathbb{Z}}_n)$. Let $I(\cdot)$ be the indicator function and

$$G^u_{\lambda,n}(\tilde{\beta}, \mathbb{Z}_{n-m}) = G_{\lambda,n}(\tilde{\beta}, \mathbb{Z}_{n-m}) + \frac{m}{n} C_l,$$

$$G^l_{\lambda,n}(\tilde{\beta}, \mathbb{Z}_{n-m}) = G_{\lambda,n}(\tilde{\beta}, \mathbb{Z}_{n-m}) + \frac{m}{n} \ell(0) I(\tilde{\beta} = 0).$$

Note that $0 \le \ell(u) \le C_l$. The terms $\frac{m}{n} C_l$ and $\frac{m}{n} \ell(0) I(\tilde{\beta} = 0)$ are the upper and lower bounds of $F_n(\tilde{\beta}, \mathbb{Z}^o_m)$, indicating the largest and smallest effects of outliers, respectively. Therefore, $G^u_{\lambda,n}(\tilde{\beta}, \mathbb{Z}_{n-m})$ and $G^l_{\lambda,n}(\tilde{\beta}, \mathbb{Z}_{n-m})$ are the upper and lower bounds of $L_{\lambda,n}(\tilde{\beta}, \tilde{\mathbb{Z}}_n)$, respectively. One can see that $G^u_{\lambda,n}(\tilde{\beta}, \mathbb{Z}_{n-m})$ is a nondecreasing function of m. Moreover, we can check that $G^l_{\lambda,n}(\tilde{\beta}, \mathbb{Z}_{n-m})$ is a nonincreasing function of m, by using the fact that $G_{\lambda,n}(0, \mathbb{Z}_{n-m}) = \frac{n-m}{n}\ell(0)$ and checking the case of $\tilde{\beta} = 0$ and $\tilde{\beta} \neq 0$ separately.

Recall that $\Delta_{BL} = \{(b, \beta) : |b| < \infty, \beta \in R^p\}$ and $S_0^- = \{\beta : \beta^T\beta_0 \leq 0\}$. We define $S_0^+ = \{\beta : \beta^T\beta_0 > 0\}$, $\Delta^+_{BL} = \{(b, \beta) : \beta \in S_0^+, |b| < \infty\}$, $\Delta^-_{BL} = \{(b, \beta) : \beta \in S_0^-, |b| < \infty\}$. Theorem 2 below studies the angular breakdown point for bounded loss functions.

THEOREM 2. *Assume that $\beta_0 \neq 0$ and the loss function $\ell(u)$ satisfies* (A1) *with $\ell(0) \leq C_l < \infty$. Then the following two statements are equivalent*:

(1) *$\hat{\beta}(\tilde{\mathbb{Z}}_n)$ does not break down in terms of the proposed population angular breakdown point when there are m arbitrary outliers in the training sample of size n*;

(2) $\min_{\tilde{\beta}\in\Delta^+_{BL}} G^u_{\lambda,n}(\tilde{\beta}, \mathbb{Z}_{n-m}) < \min_{\tilde{\beta}\in\Delta^-_{BL}} G^l_{\lambda,n}(\tilde{\beta}, \mathbb{Z}_{n-m})$.

Furthermore, the corresponding population angular breakdown point is m_0/n, where m_0 is the smallest value of m such that (2) *fails. The same is true for the sample angular breakdown point by replacing β_0 with $\hat{\beta}(\mathbb{Z}_n)$ in the associated notation.*

From Theorem 2, we can conclude that the loss function with a smaller C_l leads to a more robust classifier. For the case with $m = 0$, the equivalence between the two statements (1) and (2) in Theorem 2 can be shown directly. We also note that m_0 defined in Theorem 2 always exists. First, one can show that $G^u_{\lambda,n}(\tilde{\beta}, \mathbb{Z}_{n-m})$ is a nondecreasing function of m, while $G^l_{\lambda,n}(\tilde{\beta}, \mathbb{Z}_{n-m})$ is a nonincreasing function of m. Furthermore, given n, as $m \to n$, we have $G^u_{\lambda,n}(\tilde{\beta}, \mathbb{Z}_{n-m}) \to \lambda J(\beta) + C_l$ and $G^l_{\lambda,n}(\tilde{\beta}, \mathbb{Z}_{n-m}) \to \lambda J(\beta) + \ell(0)I(\beta = 0)$. Note that $\ell(0) < C_l$, $J(\beta) = J(-\beta)$ and $-\beta \in S_0^-$ if $\beta \in S_0^+$. Thus, when m is large enough, the inequality (2) in Theorem 2 fails. Therefore, m_0 always exists. In the following Proposition 2, we derive a lower bound of m_0. We assume that the estimate based on the sample $\mathbb{Z}_n$ without any outlier does not break down in terms of our definition of angular breakdown.

PROPOSITION 2. *Under the assumption of Theorem 2, suppose that the estimate $\hat{\tilde{\beta}}(\mathbb{Z}_n) = (\hat{b}, \hat{\beta}^T)^T$ based on $\mathbb{Z}_n$ does not break down, that is, $\delta_L(\mathbb{Z}_n) = \min_{\tilde{\beta}\in\Delta^-_{BL}} L_{\lambda,n}(\tilde{\beta}, \mathbb{Z}_n) - \min_{\tilde{\beta}\in\Delta^+_{BL}} L_{\lambda,n}(\tilde{\beta}, \mathbb{Z}_n) > \varepsilon_1$, where ε_1 is a positive constant. Then $m_0 \geq n\varepsilon_1/(4C_l)$. The same is true for the sample angular breakdown point by replacing β_0 with $\hat{\beta}(\mathbb{Z}_n)$ in the associated notation.*

From Proposition 2, we observe that the lower bound of m_0 is related to $\delta_L(\mathbb{Z}_n)/C_l$ up to a constant. Note that $\min_{\tilde{\beta}\in\Delta^+_{BL}} L_{\lambda,n}(\tilde{\beta}, \mathbb{Z}_n) = \min_{\tilde{\beta}\in\Delta_{BL}} L_{\lambda,n}(\tilde{\beta},$

$\mathbb{Z}_n$) which is the global minimum. The term $\delta_L(\mathbb{Z}_n)$, as the gap between the global minimum and the local minimum obtained on the region of breakdown, can be viewed as a measure of the performance of the loss $\ell(\cdot)$ on the data set $\mathbb{Z}_n$. A loss function that is not very flat tends to have a larger value of $\delta_L(\mathbb{Z}_n)$, and consequently deliver a larger lower bound of m_0. A loss function with a smaller upper bound C_l also tends to deliver a larger lower bound of m_0. However, a loss function with small C_l may have a small value of $\delta_L(\mathbb{Z}_n)$. We cannot conclude that a loss function with a smaller upper bound is more robust.

Note that the assumption $\delta_L(\mathbb{Z}_n) > \varepsilon_1$ is weak. We can show that it holds with probability tending to 1. Let $B(\tilde{\beta}_0, \eta)$ be a neighborhood of $\tilde{\beta}_0$ with radius $\eta > 0$. We assume the following identification condition that for any $\eta_0 > 0$, there exists $\varepsilon_0 > 0$, such that

$$\inf_{\tilde{\beta} \notin B(\tilde{\beta}_0, \eta_0)} E\ell(\tilde{\beta}^T \tilde{X} Y) > E\ell(\tilde{\beta}_0^T \tilde{X} Y) + \varepsilon_0. \tag{3.7}$$

Since $\beta_0 \neq 0$ and β_0 is an inner point of S_0^+, there exists some constant $\tilde{\varepsilon}_0 > 0$ such that $\inf_{\tilde{\beta} \in \Delta_{BL}^-} E\ell(\tilde{\beta}^T \tilde{X} Y) > E\ell(\tilde{\beta}_0^T \tilde{X} Y) + \tilde{\varepsilon}_0 = \inf_{\beta \in \Delta_{BL}^+} E\ell(\tilde{\beta}^T \tilde{X} Y) + \tilde{\varepsilon}_0$. Under conditions of standard learning theory, we have

$$\min_{\tilde{\beta} \in \Delta_{BL}^+} L_{\lambda,n}(\tilde{\beta}, \mathbb{Z}_n) \to_p \inf_{\tilde{\beta} \in \Delta_{BL}^+} E\ell(\tilde{\beta}^T \tilde{X} Y),$$

$$\min_{\tilde{\beta} \in \Delta_{BL}^-} L_{\lambda,n}(\tilde{\beta}, \mathbb{Z}_n) \to_p \inf_{\tilde{\beta} \in \Delta_{BL}^-} E\ell(\tilde{\beta}^T \tilde{X} Y).$$

Therefore, as $n \to \infty$, the assumption on $\delta_L(\mathbb{Z}_n)$ in Proposition 2 holds with probability tending to 1. It is also important to note that m_0 is generally larger than 1 as shown in our numerical studies in Section 5. In contrast, as we have seen in Section 3.2, the angular breakdown point for unbounded loss functions is $1/n$.

For example, for AdaBoost, logistic regression and linear SVM, $C_l = \infty$, and consequently the lower bound of m_0 is 0. This means that a single outlier can make these methods break down. On the other hand, for the sigmoid loss, $C_l = 1$, and consequently, $m_0 \geq n\delta_L(\mathbb{Z}_n)/4$, where $\delta_L(\mathbb{Z}_n)$ converges to a positive number under mild conditions. Next, we briefly discuss the evaluation of m_0. Since the population angular breakdown point has the same lower bound as the sample angular breakdown point, we can evaluate m_0 by checking the sample angular breakdown point numerically. For the sample angular breakdown point, to evaluate m_0, we only need to calculate $\tilde{\delta}_L(\mathbb{Z}_n)$, where

$$\tilde{\delta}_L(\mathbb{Z}_n) = \min_{\tilde{\beta} \in \hat{\Delta}_{BL}^-} L_{\lambda,n}(\tilde{\beta}, \mathbb{Z}_n) - \min_{\tilde{\beta} \in \hat{\Delta}_{BL}^+} L_{\lambda,n}(\tilde{\beta}, \mathbb{Z}_n),$$

with $\hat{\Delta}_{BL}^- = \{(b, \beta) : |b| < \infty, \beta^T \hat{\beta}(\mathbb{Z}_n) \leq 0\}$ and $\hat{\Delta}_{BL}^+ = \{(b, \beta) : |b| < \infty, \beta^T \times \hat{\beta}(\mathbb{Z}_n) > 0\}$. To calculate $\tilde{\delta}_L(\mathbb{Z}_n)$, we first obtain the global minimizer $\hat{\tilde{\beta}}$ by solving the optimization problem $\min_{\tilde{\beta}} L_{\lambda,n}(\tilde{\beta}, \mathbb{Z}_n)$. Since the global minimizer

$\hat{\hat{\beta}} = (\hat{b}, \hat{\beta}(\mathbb{Z}_n))^T \in \hat{\Delta}^+_{BL}$, we have $\min_{\tilde{\beta}\in\hat{\Delta}^+_{BL}} L_{\lambda,n}(\tilde{\beta}, \mathbb{Z}_n) = L_{\lambda,n}(\hat{\hat{\beta}}, \mathbb{Z}_n)$. Thus, we only need to compute $\min_{\tilde{\beta}\in\hat{\Delta}^-_{BL}} L_{\lambda,n}(\tilde{\beta}, \mathbb{Z}_n)$. When a nonconvex loss is used, some algorithms such as Difference of Convex functions (DC) algorithm [Horst and Thoai (1999), Wu and Liu (2007)] can be used. Then we can calculate $\tilde{\delta}_L(\mathbb{Z}_n)$ and the lower bound of m_0 is $n\tilde{\delta}_L(\mathbb{Z}_n)/(4C_l)$.

4. Angular breakdown point for classification with kernel. We only focus on the discussion on linear classification so far. Next, we generalize the definition of angular breakdown point for kernel classification. Both unbounded and bounded loss functions are considered. Kernel methods have been widely used for nonlinear learning; see Schölkopf and Smola (2002) for a comprehensive review.

Suppose that the uncontaminated observations are from $\mathcal{X} \times \mathcal{Y}$ and contaminated ones are from $\mathcal{X}^o \times \mathcal{Y}$, where $\mathcal{Y}$ equals $\{1, -1\}$. Here, $\mathcal{X}$ and $\mathcal{X}^o$ can be the same or different. Let $\mathcal{X}_0 = \mathcal{X} \cup \mathcal{X}^o$, and $\mathcal{H}$ be the Reproducing Kernel Hilbert Space (RKHS) [Wahba (1990)] with the kernel $K(x, x)$, where $x \in \mathcal{X}_0$. Denote (b_0, f_0) be the true parameters associated with the loss $\ell(u)$, that is, $(b_0, f_0) = \arg\min_{b\in R, f\in\mathcal{H}} E(\ell(Y[b + f(X)]))$. Without loss of generality, we assume $|b_0| \le M_0 < \infty$. Consequently,

$$(b_0, f_0) = \arg \min_{(b,f)\in\Delta_{BK}} E\big(\ell\big(Y\big[b + f(X)\big]\big)\big),$$

where $\Delta_{BK} = \{(b, f) : |b| \le M_0 < \infty, f \in \mathcal{H}\}$. Suppose that we have m outliers $\{(x_i^o, y_i^o), i = 1, \dots, m\}$. By the reproducing property of the kernel, taking the feature map $\phi(x) = K(x, \cdot)$, the corresponding objective function can be written as follows:

$$\begin{aligned}
L_{\lambda,n}(b, f, \tilde{\mathbb{Z}}_n) &= \lambda \|f\|^2_{\mathcal{H}} + \frac{1}{n}\Bigg[\sum_{i=1}^{n-m} \ell\big(y_i\big(b + f(x_i)\big)\big) + \sum_{i=1}^{m} \ell\big(y_i^o\big(b + f\big(x_i^o\big)\big)\big)\Bigg] \\
&= \Bigg[\lambda \|f\|^2_{\mathcal{H}} + \frac{1}{n}\sum_{i=1}^{n-m} \ell\big(y_i\big[b + \langle f, \phi(x_i)\rangle_{\mathcal{H}}\big]\big)\Bigg] \qquad (4.1) \\
&\quad + \frac{1}{n}\sum_{i=1}^{m} \ell\big(y_i^o\big[b + \langle f, \phi(x_i^o)\rangle_{\mathcal{H}}\big]\big) \\
&:= G_{\lambda,n}(b, f, \mathbb{Z}_{n-m}) + F_n\big(b, f, \mathbb{Z}^o_m\big).
\end{aligned}$$

Similar to the linear case, when the dataset is not linearly separable in the original feature space, we only need to minimize the object function over Δ_{BK} with M_0 being sufficiently large. Let $(\hat{b}, \hat{f}_\lambda) = \arg\min_{(b,f)\in\Delta_{BK}} L_{\lambda,n}(b, f, \tilde{\mathbb{Z}}_n)$. By the representer theorem [Kimeldorf and Wahba (1970)], we have

$$\hat{f}_\lambda(\cdot) = \sum_{i=1}^{n-m} \hat{\alpha}_i K(x_i, \cdot) + \sum_{i=1}^{m} \hat{\alpha}_i^o K(x_i^o, \cdot) = \sum_{i=1}^{n-m} \hat{\alpha}_i \phi(x_i) + \sum_{i=1}^{m} \hat{\alpha}_i^o \phi(x_i^o),$$

for some constants $\hat{\alpha}_i$ and $\hat{\alpha}_i^o$. Furthermore, by the reproducing property of the kernel, we have $f_0(x) = \langle f_0(\cdot), \phi(x)\rangle_{\mathcal{H}}$. Thus,

$$\langle \hat{f}_\lambda, f_0\rangle_{\mathcal{H}} = \sum_{i=1}^{n-m} \hat{\alpha}_i f_0(x_i) + \sum_{i=1}^{m} \hat{\alpha}_i f_0(x_i^o).$$

Similar to the linear case, the angular breakdown point for kernel learning can be defined as follows.

DEFINITION 2. The population angular breakdown point for large margin classifiers with kernel is defined as

$$\varepsilon_{\mathcal{H}}(f_0, \mathbb{Z}_n) = \min\left\{\frac{m}{n} : \langle \hat{f}_\lambda, f_0\rangle_{\mathcal{H}} \le 0\right\}.$$

As in the linear case, outliers affect the angle between $\hat{f}_\lambda$ and f_0 in the feature space. For the unbounded kernel (e.g., the polynomial kernel with $\mathcal{X}_0 = R^p$), as $\|\phi(x_i^o)\|_{\mathcal{H}} \to \infty$, then the effect of outliers is similar to the linear case, by imposing the constraint $y_i^o\langle f, \phi(x_i^o)\rangle \ge 0$ on the direction of f in the feature space.

In practice, since f_0 is always unknown, we can calculate the sample angular breakdown point for large margin classifiers with kernel defined as follows.

DEFINITION 2′. The sample angular breakdown point for large margin classifiers with kernel is defined as

$$\varepsilon_{\mathcal{H}}(\hat{f}, \mathbb{Z}_n) = \min\left\{\frac{m}{n} : \langle \hat{f}_\lambda, \hat{f}_{\mathbb{Z}_n}\rangle_{\mathcal{H}} \le 0\right\},$$

where $\hat{f}_\lambda$ is defined above and $\hat{f}_{\mathbb{Z}_n}$ is the estimates of f_0 using the observations $\mathbb{Z}_n$.

Without loss of generality, we assume that $\hat{f}_{\mathbb{Z}_n} \ne 0$ throughout this paper. Similar to the linear case, the sample angular breakdown point generally has the same properties as those of the angular breakdown point. Theorem 3 below studies the angular breakdown point for kernel learning with unbounded loss functions.

THEOREM 3. *Suppose that* (A1) *holds with $C_l = \infty$ and that $\mathcal{H}$ is a RKHS with the finite dimension d. Furthermore, assume that* (1) *the kernel function $K(x, y)$ is continuous*; (2) *for any $M > 0$, there exists d points $x_1, \dots, x_d \in \mathcal{X}_0$ such that $K(x_i, x_i) > M$*; *and* (3) $\{K(\cdot, x_i)\}_{i=1}^d$ *are linearly independent. Then we have $\varepsilon_{\mathcal{H}}(f_0, \mathbb{Z}_n) \le d/n$. The same conclusion holds for the sample angular breakdown point in Definition 2′.*

Theorem 3 shows the breakdown point of kernel learning with some assumptions on the kernel. Note that the RKHS associated with the polynomial kernel

$K(x,x) = (x^T x + c)^m$ has a finite dimension for any constant c. Therefore, if $\mathcal{X} = R^p$, there exists $x_1, \ldots, x_d$ in $\{x : x \in R^p, \|x\| > c_M\}$ for a large c_M such that the assumptions of Theorem 3 hold.

In fact, the assumption that the RKHS has a finite dimension is not necessary. Theorem 3 can be generalized in the following Proposition 3 without the assumption on the dimension of RKHS.

PROPOSITION 3. *Assume that* (A1) *holds with* $C_l = \infty$. *Given an increasing series* $\{M_i\}$ *with* $M_i \to \infty$, *denote by* I_i *the set with the smallest cardinality such that* (1) $\min_{x_t \in \mathcal{X}_0, t \in I_i} K(x_t, x_t) > M_i$, (2) $f_0 \in \operatorname{span}\{\phi(x_t), x_t \in \mathcal{X}_0, t \in I_i\}$. *Define* $I_0 = \limsup_i |I_i|$. *Then* $\varepsilon_{\mathcal{H}}(f_0, \mathbb{Z}_n) \le |I_0|/n$. *The same is true for the sample angular breakdown point in Definition* 2′ *by replacing* f_0 *with* $\hat{f}_{\mathbb{Z}_n}$ *in* (2).

Theorem 3 and Proposition 3 provide upper bounds for the angular breakdown point of kernel learning using unbounded loss functions and unbounded kernels. For a bounded kernel, such as Gaussian kernel $K(x,y) = \exp(-\|x-y\|^2/(2\sigma^2))$, the feature map $\phi(x) = K(x, \cdot)$ has a bounded norm and consequently the effect of outliers is limited compared to the case of an unbounded kernel. Therefore, bounded kernels should be more robust than unbounded ones. This is confirmed by the lower bound on the breakdown point shown in Theorem 4 below.

To make the conditions in Proposition 3 more clear, we will show the following Proposition 4 which can be considered as a special case of Proposition 3. Let $\{e_i(x), i = 1, 2, \ldots\}$ denote the orthogonal basis of $\mathcal{H}$ obtained from the spectral decomposition. Denote the kernel function $K(x,y) = \sum_{i=1}^{\infty} e_i(x) e_i(y)$ and the feature map $\phi : x \mapsto (e_1(x), e_2(x), \ldots)^T$. Suppose that f_0 belongs to a finite subspace of $\mathcal{H}$. Without loss of generality, we assume that $f_0 \in \operatorname{span}\{e_i(x), i = 1, \ldots, d\}$. For any positive integer m, denote $\operatorname{span}\{e_1(x), \ldots, e_m(x)\}$ as $\mathcal{H}_m$, which is the subspace of $\mathcal{H}$ spanned by only the first m basis. The kernel function associated with $\mathcal{H}_m$ is denoted as $K_m(x,y) = \sum_{i=1}^{m} e_i(x) e_i(y) = \langle \phi_m(x), \phi_m(y) \rangle_{\mathcal{H}_m}$, where the feature map is $\phi_m(x) = (e_1(x), \ldots, e_m(x))^T$.

PROPOSITION 4. *Assume that* $f_0 \in \mathcal{H}_d$, $K(x,y)$ *is continuous and* (A1) *holds with* $C_l = \infty$. *Suppose that there exists some positive integer* m *with* $m \ge d$ *such that the following conditions hold*: (i) *For any* $M > 0$, *there exist* m *points* $x_1, \ldots, x_m$, *such that* $K_m(x_i, x_i) > M$, $1 \le i \le m$; (ii) $\{\phi_m(x_i), 1 \le i \le m\}$ *is linearly independent. Then* $\varepsilon_{\mathcal{H}}(f_0, \mathbb{Z}_n) \le m/n$.

Note that the RKHS $\mathcal{H}$ is considered to be of infinite dimension in Proposition 4 while the function f_0 is assumed to belong to a finite dimensional subspace. Moreover, the assumption that $f_0 \in \mathcal{H}_d$ is only for convenience. Specifically, f_0 can be in any subspace $\tilde{\mathcal{H}}$ of finite dimensions, and $\mathcal{H}_m$ can be a subspace containing $\tilde{\mathcal{H}}$ such that the assumptions of Proposition 4 hold. Conditions (i) holds when $K_m(x,y)$ is unbounded and $\mathcal{X}_0 = R^p$. Since $d \le m$, it is obvious that

$f_0 \in \mathcal{H}_d \subseteq \mathcal{H}_m$ which is of finite dimensions. The proof of Proposition 4 is very similar to that of Theorem 3. From the proof of Theorem 3, we can check that $f_0(\cdot) \in \operatorname{span}\{\phi_m(x_i), 1 \le i \le m\}$. Thus, the conditions in Proposition 4 are special cases of those of Proposition 3.

Next, we derive the lower bound of the angular breakdown point. We consider two cases: the case with both the loss function and the kernel being unbounded, and the case either the loss or the kernel function being bounded. Let $\sup_{x \in \mathcal{X}_0} K(x,x) \le C_K \le \infty$. Suppose that we replace m observations in $\mathbb{Z}_n$ by m outliers $(x_i^o, y_i^o) \in \mathcal{X}_0 \times \mathcal{Y}$, $i = 1, \dots, m$. Recalling (4.1), we have

$$L_{\lambda,n}(b, f, \tilde{\mathbb{Z}}_n) = G_{\lambda,n}(b, f, \mathbb{Z}_{n-m}) + F_n(b, f, \mathbb{Z}_m^o). \tag{4.2}$$

Recall that $|b| \le M_0 < \infty$ in Δ_{BK}. For the optimal solution $\hat{f}_\lambda$, we can check that $\|\hat{f}_\lambda\|_{\mathcal{H}} < \sqrt{\ell(0)/\lambda}$. Let $G_1^o = \ell(-M_0 - \sqrt{C_K \ell(0)/\lambda})$, which is the upper bound of $\ell(y[b + \langle f, \phi(x)\rangle_{\mathcal{H}}])$ for any $\|f\|_{\mathcal{H}} < \sqrt{\ell(0)/\lambda}$. Let

$$\delta_\lambda(\mathbb{Z}_n) = \inf_{(b,f)\in T_\lambda^-} G_{\lambda,n}(b, f, \mathbb{Z}_n) - \inf_{(b,f)\in T_\lambda^+} G_{\lambda,n}(b, f, \mathbb{Z}_n),$$

where $T_\lambda^+ = \{(b, f) : \langle f, f_0\rangle_{\mathcal{H}} > 0, \|f\|_{\mathcal{H}} \le \sqrt{\ell(0)/\lambda}, |b| \le M_0 < \infty\}$ and $T_\lambda^- = \{(b, f) : \langle f, f_0\rangle_{\mathcal{H}} \le 0, \|f\|_{\mathcal{H}} \le \sqrt{\ell(0)/\lambda}, |b| \le M_0 < \infty\}$. Note that even there is no outlier, it is still possible that the estimate breaks down in the finite sample case due to the limited sample size. In this case, the definition of breakdown point is meaningless. Similar to Proposition 2, we make the following assumption to avoid this trivial case:

(A2) Suppose that the estimate $(\hat{b}, \hat{f})$ based on $\mathbb{Z}_n$ does not break down in terms of our definition of angular breakdown, that is, $\delta_\lambda(\mathbb{Z}_n) \ge \varepsilon_2$, where ε_2 is a positive constant.

The assumption (A2) is similar to the assumption on $\delta_L(\mathbb{Z}_n)$ in Proposition 2. As the discussion after Proposition 2, the assumption (A2) is also weak. Under a identification condition similar to (3.7) and other mild conditions, we can prove that the probability $P(\delta_\lambda(\mathbb{Z}_n) > \varepsilon_2)$ tends to 1 as $n \to \infty$. Theorem 4 below derives a lower bound of the angular breakdown point for the kernel classification.

THEOREM 4. *Suppose* (A1) *and* (A2) *hold. Then*

$$\varepsilon_{\mathcal{H}}(f_0, \mathbb{Z}_n) \ge m_1/n,$$

where m_1 is the smallest integer larger than $n\varepsilon_2/(4\min\{G_1^o, C_l\})$ and ε_2 is defined in (A2). *The same conclusion holds for the sample angular breakdown point defined in Definition 2′, by replacing f_0 with $\hat{f}_{\mathbb{Z}_n}$ in the definition of $\delta_\lambda(\mathbb{Z}_n)$ in* (A2).

The lower bound of the angular breakdown point shown in Theorem 4 is related to $\delta_\lambda(\mathbb{Z}_n)$ up to a constant. A larger value of $\delta_\lambda(\mathbb{Z}_n)$ delivers a larger lower bound of the proposed angular breakdown point. Similar to the linear case, the term $\delta_\lambda(\mathbb{Z}_n)$ can be viewed as a measure of the performance of the loss $\ell(u)$ on the data set $\mathbb{Z}_n$. The lower bound in Theorem 4 can be applied to two cases. For the case with unbounded kernel and unbounded loss functions with $C_K = \infty$ and $C_l = \infty$, we have $\varepsilon_{\mathcal{H}}(f_0, \mathbb{Z}_n) \geq 1/n$. For the case with at least one of C_K and C_l being bounded, we have $\min\{G_1^o, C_l\} < \infty$ and $m_1 \to \infty$ as $n \to \infty$. Combined with the results in Theorem 3, we can conclude that binary large margin classifiers with either bounded kernel or bounded loss functions tend to be more robust than those with unbounded kernel and loss functions.

Interestingly, in contrast to the essential role of bounded loss functions for linear learning, robustness can be achieved for kernel learning through either bounded kernel or bounded loss functions. Therefore, when an unbounded loss function is used, such as exponential loss or hinge loss, methods using the polynomial kernel $K(x,x) = (x^T x + c)^m$ with $\mathcal{X} = R^p$, are less robust than methods using the Gaussian kernel $K(x,y) = \exp(-\|x-y\|^2/(2\sigma^2))$ in terms of angular breakdown point. Hable and Christmann (2011) showed that the SVM is finite sample qualitative robustness, when the kernel is bounded and continuous, and the loss function satisfies uniform Lipschitz condition. This result matches with our result about the role of a bounded kernel in achieving robustness.

5. Simulation. In this section, we perform simulation studies to compare the robustness of large margin classifiers using different loss functions. The bounded sigmoid loss and three unbounded losses, including the exponential loss for the AdaBoost, the deviance loss for the penalized logistic regression and the hinge loss for the SVM, are considered. We study four examples, including two linear classification examples and two kernel classification examples.

EXAMPLE 1 (Linear classification). Consider two classes with the class label $Y \in \{1, -1\}$ and the covariate vector $X|Y$ follows the normal distribution $N(\text{sign}(Y)u_1, I_{10})$, where $u_1 = (1,1,0,\dots,0)^T \in R^{10}$. Only the first two covariates are useful to distinguish two classes. For each class, we generate 50 training samples to fit the models, 50 tuning samples to choose the best tuning parameters and 2500 test samples to evaluate different methods.

As shown in Section 3, for linear binary classification, the angular breakdown point for a method with an unbounded loss is $1/n$, that is, a single outlier is sufficient to result in angular breakdown. To show the rational of our proposed angular breakdown criterion and verify this theoretical result, we replace the first observation in the positive class by one outlier generated independently from the distribution $N(u_0, I_p)$, where $u_0 = [(1-r)_+ - r]u_1$ and r is a parameter controlling the severity of that particular outlier. We study 11 different cases with $r = 0, 25, 50, 75, \dots, 250$. Clearly, when $r = 0$, there is no outlier in the training dataset. For each case, we repeat the simulation 100 times. To evaluate different methods, we compare their average testing errors and the proportions of angular

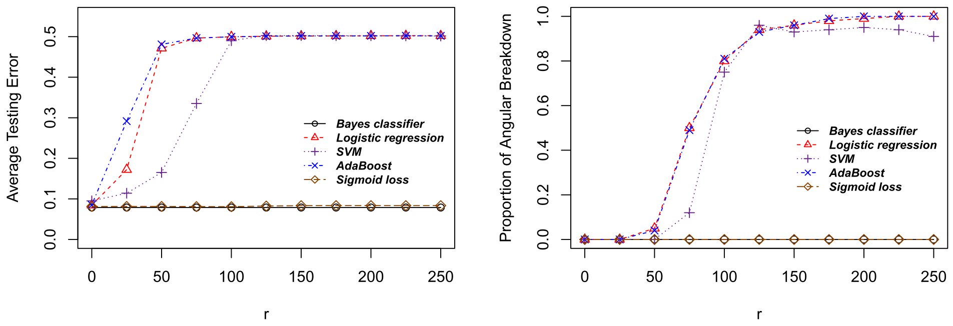

FIG. 2. *Performance comparison of the average testing errors (left panel) and the proportions of angular breakdown (right panel) of Example 1.*

breakdown (PAB). The definition of angular breakdown point for a given method needs β_0 as the minimizer of the corresponding expected loss. The exact calculation of β_0 can be difficult depending on the form of the loss function. For simplicity, in this simulation example, we use the direction of the Bayes decision boundary, that is, $\beta_{\text{Bayes}} = (1/\sqrt{2}, 1/\sqrt{2}, 0, 0, \dots, 0)^T$ as the true β_0 to check whether each method breaks down according to Definition 1. Thus, for a given method, the corresponding PAB can be calculated by $PAB = \frac{1}{100}\sum_{k=1}^{100} I(\hat{\beta}_k^T \beta_{\text{Bayes}})$, where $\hat{\beta}_k$'s are estimates of β_0, and $I(x)$ is the indicator function which equals to 1 if $x \le 0$, and 0 otherwise.

Figure 2 shows the performance comparison of the average testing errors and the proportions of angular breakdown for Example 1. As $r = 0$, there is no outlier in the training dataset. The performance of different methods using either bounded or unbounded loss functions are very similar to the performance of the Bayes classifier. In this case, none of these methods breaks down according to our proposed angular breakdown criterion. As the effect of the outlier increases (r increases), both the average testing errors and the proportions of angular breakdown of the methods using unbounded loss functions increase rapidly. Among the three methods with unbounded loss functions, as expected, the exponential loss for AdaBoost has the worst performance, followed by the penalized logistic regression, and then the SVM. The hinge loss for the SVM appears to be the most robust one among those three unbounded loss functions. When the outlier is severe enough, such as $r = 250$, the performance of the penalized logistic regression, AdaBoost and SVM are almost the same as random guessing. All of these three methods break down in most simulation replications. In contrast, the method using the bounded sigmoid loss is very robust. As r increases from 0 to 250, its average testing errors are always very close to the Bayes error and it never breaks down. This verifies our theoretical result that for the linear binary classification, a single outlier is sufficient to result in angular breakdown for the methods using unbounded loss functions while the methods using bounded loss are more robust. Thus, the pro-

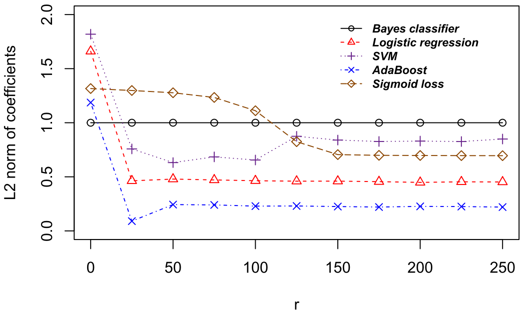

FIG. 3. *The l_2 norms of the estimated coefficient vectors.*

posed angular breakdown point is an effective measure to quantify the robustness of classification methods.

To examine the performance of the traditional breakdown point (2.2), we also calculate the l_2 norms of the estimated coefficient vectors. As shown in Figure 3, all of these methods have bounded l_2 norms, even for the methods using the unbounded loss functions. Thus, none of these methods breaks down according to the traditional breakdown criterion. This further confirms that the traditional criterion of breakdown is not suitable for classification problems.

EXAMPLE 2 (High dimensional linear classification). For this high dimensional linear classification example, the dimension $p = 100$. The other settings are the same as those of Example 1. Only the first two dimensions are useful to distinguish two classes. The direction of the Bayes decision boundary $\beta_{\text{Bayes}} = (1/\sqrt{2}, 1/\sqrt{2}, 0, 0, \ldots, 0)^T$. For each class, we generate 80 training samples to fit the models, 80 tuning samples to choose the best tuning parameters and 2500 test samples to evaluate different methods.

Figure 4 shows the performance comparison of the average testing errors and the proportions of angular breakdown for Example 2. It also indicates that a sin-

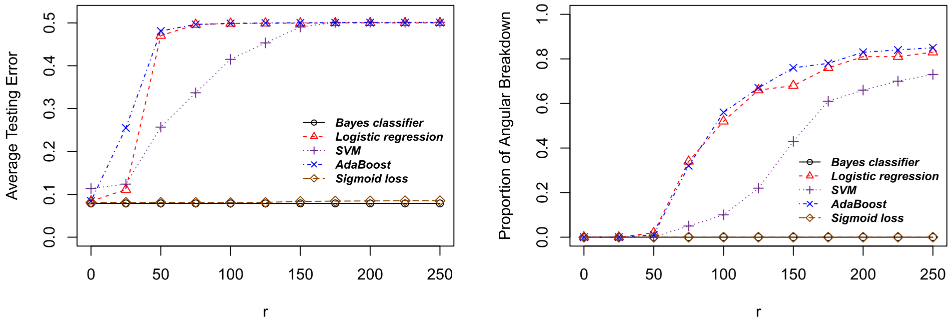

FIG. 4. *Performance comparison of the average testing errors (left panel) and the proportions of angular breakdown (right panel) of Example 2.*

gle outlier is sufficient to result in angular breakdown for the methods using unbounded loss functions while the methods using bounded loss functions are more robust. This example also demonstrates that our proposed angular breakdown point criterion can be also used for high dimensional data.

EXAMPLE 3 (Kernel classification). Consider two classes with the class label $Y \in \{1, -1\}$. Let the covariate vector $X = (R\cos(2\pi\theta), R\sin(2\pi\theta))$, where $\theta \sim$ Uniform$(0, 1)$ and

$$R|Y = +1 \sim \text{the distribution with the density } \frac{2}{\pi}\exp\left(-\frac{R^2}{\pi}\right)I(R > 0);$$

$$R|Y = -1 \sim \text{the distribution with the density } \frac{12}{\pi}\exp\left(-\frac{36R^2}{\pi}\right)I(R > 0).$$

For each class, we generate 50 training samples, 50 tuning samples and 2500 test samples. Furthermore, we replace the first observation in the negative class by one outlier $(R_0\cos(\pi/4), R_0\sin(\pi/4))$, where R_0 follows the half-normal distribution with the density function $\frac{2\theta_0}{\pi}\exp(-\frac{R_0^2\theta_0^2}{\pi})I(R_0 > 0)$. We create 10 different values of $\theta_0 \in [10^{-2}, 10^{-5}]$ which are placed evenly on the logarithmic scale. We also consider the case with no outlier in the training dataset. For each case, we repeat the simulation 100 times. In each simulation, we use the polynomial kernel $K(s,t) = (s^T t + 1)^2$ for all methods. In our definition of the angular breakdown point for the kernel case, we need to know f_0. In general, f_0 depends on different loss functions and is difficult to compute. For this example, we replace f_0 by $f_{\text{Bayes}} = x_1^2 + x_2^2 - \pi\log(6)/35$ which corresponds to the Bayes classifier. Thus, for a given method, PAB for this example can be obtained by $PAB = \frac{1}{100}\sum_{k=1}^{100} I(\langle \hat{f}_k, f_{\text{Bayes}}\rangle_{\mathcal{H}})$, where $\hat{f}_k$'s are estimates of f_0.

Figure 5 displays the results for Example 3. When there is no outlier in the training dataset (the first case in the plot), the performance of all methods using either

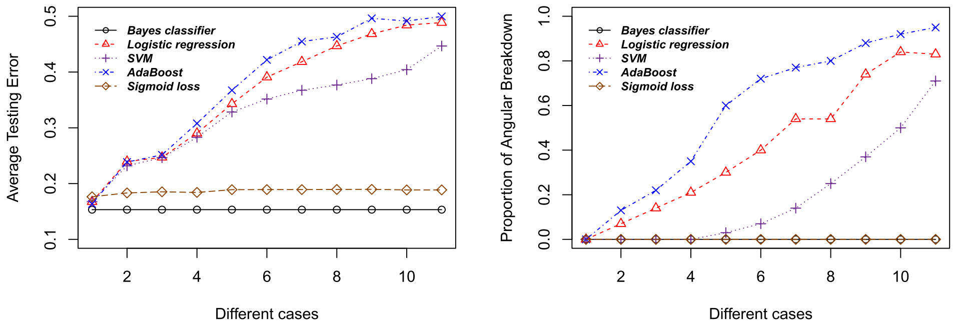

FIG. 5. *Performance comparison of the average testing errors (left panel) and the proportions of angular breakdown (right panel) of Example 3. The values on the X axes represent the case indices. As the case index increases, the effect of the outlier becomes more extreme (the value of θ_0 decreases).*

bounded or unbounded loss functions are very similar to the performance of the Bayes classifier. None of methods breaks down according to our proposed angular breakdown criterion. As the magnitude of the outlier increases (θ_0 decreases), for the methods using unbounded loss functions, both the average testing errors and the proportions of angular breakdown increase. Among the three methods using the unbounded loss functions, the AdaBoost method obtains the worst performance while the SVM performs best. This is caused by the difference between the exponential loss of the AdaBoost and the hinge loss of the SVM, in the sense that the exponential loss assigns much larger loss values for misclassified points than the corresponding loss values of the hinge loss. For this kernel example, compared with the penalized logistic regression, SVM and AdaBoost, the method using the bounded sigmoid loss is very robust and delivers the best performance. The average testing errors corresponding to the sigmoid loss are always very close to the Bayes error, and the method never breaks down in all these cases.

EXAMPLE 4 (Kernel classification with noisy features). Consider two classes with the class label $Y \in \{1, -1\}$ and the covariate vector $X = (R\cos(2\pi\theta), R\sin(2\pi\theta), X_3, X_4, \dots, X_{10})$, where we use the same method shown in Example 3 to generate R and θ. The noisy features $X_3, X_4, \dots, X_{10}$ follow a normal distribution with mean 0 and standard deviation 0.25. For each class, we generate 100 training samples, 100 tuning samples and 2500 test samples. For each experiment, we replace the first observation in the negative class by an outlier $(R_0\cos(\pi/4), R_0\sin(\pi/4), O_3, O_4, \dots, O_{10})$, where we use the same method shown in Example 3 to generate R_0. The features $O_3, O_4, \dots, O_{10}$ are generated from a normal distribution with mean 0 and standard deviation 0.25. The polynomial kernel $K(s,t) = (s^T t + 1)^2$ is used for all large margin classification methods.

For this example, we also have $f_{\text{Bayes}} = x_1^2 + x_2^2 - \pi\log(6)/35$. Figure 6 shows the performance comparison of the average testing errors and the proportions of

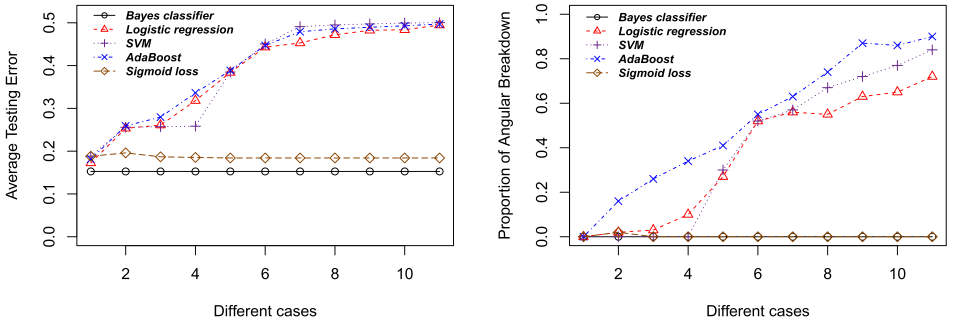

FIG. 6. *Performance comparison of the average testing errors (left panel) and the proportions of angular breakdown (right panel) of Example 4. The values on the X axes represent the case indices. As the case index increases, the effect of the outlier becomes more extreme (the value of θ_0 decreases).*

angular breakdown. It indicates that the method using the bounded sigmoid loss delivers much lower misclassification errors than the other methods using unbounded loss functions. The proportion of angular breakdown of the sigmoid method is much smaller than the PAB of the other methods such as AdaBoost, SVM and penalized logistic regression. These results indicate that for the kernel classification, the method using bounded loss can be also more robust than the method using unbound loss. They also demonstrate that the proposed angular breakdown point is an effective measure to quantify the robustness of kernel classification methods.

6. Real data analysis.

In this section, we study the robustness of different loss functions using the Wisconsin Diagnostic Breast Cancer (WDBC) data. The goal of the corresponding breast cancer study is to use a digitized image of a fine needle aspirate of a breast mass to diagnose the corresponding breast cancer status. More details on the WDBC data are provided at the UCI Machine Learning Repository (https://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/wdbc.names). The dataset has 569 subjects. For each subject, there are 30 real-valued input features and a binary response of diagnosis indicating either malignant or benign of the corresponding cancer. For this analysis, we first scale the data by standardizing each input feature to have mean zero and standard deviation one. We randomly split the whole dataset into training, tuning and test sets of sizes 100 (50 in each class), 100 (50 in each class) and 369, respectively.

To study the robustness of different losses, we randomly select m training samples and flip their labels [from malignant (benign) to benign (malignant)]. We study 11 cases with $m = 0, 5, 10, \dots, 50$. For each case, we use linear learning and repeat the simulation 100 times. When $m = 0$, there is no outlier in the training dataset. The estimate of β_0 obtained by each method in this case is used to calculate the proportion of angular breakdown according to Definition $1'$. We report both the average testing errors and the proportions of angular breakdown of different methods.

Figure 7 provides the performance comparison of the average testing errors and the proportions of angular breakdown using the WDBC data. When there is no outlier in the training data, all these four methods obtain excellent performance. However, as the number of outliers increases, the average testing errors of the penalized logistic regression, AdaBoost and SVM increase significantly while the average testing error of the method using the sigmoid loss does not increase much. For the worst case with 50 outliers, the performance of the method using the sigmoid loss is still reasonable while the average testing errors of the penalized logistic regression, AdaBoost and SVM are 0.463, 0.469 and 0.44, respectively. In addition, as the number of outliers increases, the proportions of angular breakdown of the three methods with unbounded loss functions also increase significantly. Compared with the penalized logistic regression and AdaBoost, the SVM is more

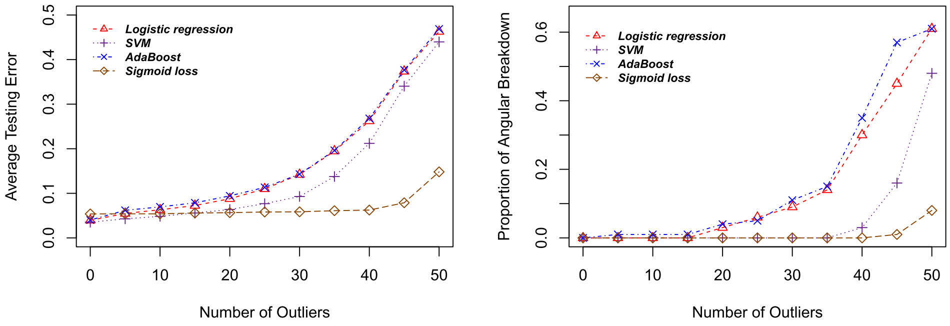

FIG. 7. *Performance comparison of the average testing errors* (*left panel*) *and the proportions of angular breakdown* (*right panel*) *using the WDBC data.*

robust in terms of the proportion of angular breakdown. Among these four methods, the method using the bounded sigmoid loss delivers the lowest proportion of angular breakdown. For the worst case with 50 outliers, the corresponding PAB is only 0.08 while the PAB's of the penalized logistic regression, AdaBoost, and SVM are 0.61, 0.61 and 0.48, respectively. Based on this real data analysis, we can conclude that the method using the bounded loss function is more robust than those methods with unbounded loss functions. The proposed angular breakdown criterion works well for classification.

7. Discussion. Robustness is a very important consideration for statistical modeling. Most existing robustness measures such as breakdown point focus on regression problems. New criteria tailored for classification are greatly needed. For classification, we are more concerned about the classification boundary than the parameter estimation of the corresponding classification function. Motivated by this observation, we propose the novel angular breakdown point criterion to quantify robustness of different classification methods. Our theoretical and numerical studies indicate that large margin classification methods with bounded loss functions tend to be more robust than those with unbounded ones. This is consistent with existing studies on robust classification in the literature. In this paper, we focus on the angular breakdown point of binary large margin classification methods. The proposed angular breakdown point can be generalized to multicategory classification, which can be explored for future work. In addition, our theoretical results can help us design new classification methods that have high angular breakdown point. As shown in Proposition 2 and the following discussion, the lower bound of the angular breakdown point depends on C_l [the upper bound of the loss function $\ell(u)$] and $\delta_L(\mathbb{Z}_n)$ (the gap between the global minimum and the local minimum in the breakdown region). If the upper bound C_l is small and the gap $\delta_L(\mathbb{Z}_n)$ is large, the classification method will have a high angular breakdown point. In practice, we can use bounded loss (e.g., the sigmoid loss studied in this paper) to control C_l. In addition, we can design the loss function to be not very flat in order to obtain a large value of $\delta_L(\mathbb{Z}_n)$. We need to consider these two aspects to design the loss function with a high angular breakdown point.

Let us consider the following example about the clipped loss $\ell_c(y, x^T\beta) = \min(1, \ell(y, x^T\beta))$. This loss function is bounded and, therefore, could deliver robust classification performance. However, since it is nonconvex, it is not easy to solve the optimization problem. To solve this issue, Yu et al. (2010) proposed the ρ relaxed loss function $\ell_\rho(y, x^T\beta) = \rho\ell(y, x^T\beta) + 1 - \rho$, which is a convex relaxation of the clipped loss. It has been shown that $\ell_c(y, x^T\beta) = \min_{0\le\rho\le 1} \ell_\rho(y, x^T\beta)$. The parameter ρ controls the degree of the convexity and the upper bound of the loss function. Motivated by the idea of Yu et al. (2010), we may design a new loss function $\tilde{\ell}_\rho(u)$, where ρ controls both the convexity and the upper bound of the loss function. The tuning parameter ρ can be chosen by maximizing the lower bound of the angular breakdown point. In this case, as shown in Proposition 2, the lower bound of the angular breakdown point of the method using the loss function $\tilde{\ell}_\rho$ will be a function of ρ. Considering the family of loss functions $\{\tilde{\ell}_\rho : 0 \le \rho \le 1\}$, in order to obtain a good loss function with a high angular breakdown point, we can find the parameter ρ which maximizes the lower bound of the sample angular breakdown point, that is,

$$\rho^* = \arg \max_{\rho\in[0,1]} \frac{n}{4C_l} \cdot \Bigl(\min_{\tilde{\beta}\in\hat{\Delta}_{BL}^-} L_{\lambda,n}(\tilde{\beta}, \mathbb{Z}_n) - \min_{\tilde{\beta}\in\hat{\Delta}_{BL}^+} L_{\lambda,n}(\tilde{\beta}, \mathbb{Z}_n)\Bigr),$$

where $\hat{\Delta}_{BL}^- = \{(b, \beta) : |b| < \infty, \beta^T\hat{\beta}(\mathbb{Z}_n) \le 0\}$ and $\hat{\Delta}_{BL}^+ = \{(b, \beta) : |b| < \infty, \beta^T\hat{\beta}(\mathbb{Z}_n) > 0\}$. Therefore, if we consider a family of loss functions with some parameters, our criterion is useful to choose these parameters that lead to a high angular breakdown point.

APPENDIX: SELECTED PROOFS

A.1. Proof of Theorem 1.

PROOF. We only prove the conclusions on the population angular breakdown point. The conclusion on the sample angular breakdown point can be proved by the same argument, where β_0 is replaced by $\hat{\beta}(\mathbb{Z}_n)$ in the proof.

(i) According to the analysis in Section 3.1, we consider one outlier $(x_1^o, y_1^o) \in R^p \times \{1, -1\}$, such that $x_1^o y_1^o = -c \cdot \beta_0$ with the positive number $c \to \infty$. According to the definition of $S_{z_1^o}^+$, we have $\hat{\beta}^T(-\beta_0) \ge 0$, that is, $\hat{\beta} \in S_0^-$. Consequently, the angular breakdown point is $1/n$ for any unbounded loss $\ell(u)$ satisfying (A1).

(ii) Now we consider the square loss function $\ell(z, f) = (1 - yf(x))^2$ with $f(x) = b + \beta^T x$ and $z = (x, y)$. Recall that, for the uncontaminated observations $\{z_i = (x_i, y_i), i = 1, \dots, n-1\}$, $\|x_i\|$ is assumed to be bounded. Taking the outlier

$z^o_{1c} = (x^o_{1c}, y^o_{1c}) = (-c\beta_0, 1)$ with $c \to \infty$, we have

$$\begin{aligned} L_{\lambda,n}(\tilde{\beta}, \tilde{\mathbb{Z}}_n) &= \left[\lambda J(\beta) + \frac{1}{n}\sum_{i=1}^{n-1}\bigl(1 - y_i\bigl(b + \beta^T x_i\bigr)\bigr)^2\right] \\ &\quad + \frac{1}{n}\bigl(1 - y^o_{1c}\bigl(b + \beta^T x^o_{1c}\bigr)\bigr)^2 \\ &:= G_{\lambda,n}(\tilde{\beta}, \mathbb{Z}_{n-1}) + F_n\bigl(\tilde{\beta}, z^o_{1c}\bigr). \end{aligned} \tag{A.1}$$

Denote $\hat{\tilde{\beta}}$ be the minimizer of $L_{\lambda,n}(\tilde{\beta}, \tilde{\mathbb{Z}}_n)$ over $\tilde{\beta} = (b, \beta) \in \Delta_{BL} = \{(b, \beta), |b| < \infty, \beta \in R^p\}$. Regardless $\lambda = 0$ or not, since $L_{\lambda,n}(\tilde{\beta}, \tilde{\mathbb{Z}}_n)|_{\tilde{\beta}=0} = 1$, it is easy to see that $\hat{\tilde{\beta}} = (\hat{b}, \hat{\beta}^T)^T$ has a bounded norm. Thus, the minimization of $L_{\lambda,n}(\tilde{\beta}, \tilde{\mathbb{Z}}_n)$ can be taken over the set $\Delta^{(1)}_{BL} := \{(b, \beta) : |b| < M, \|\beta\| < M\}$ for some M being sufficiently large. Therefore, given $\mathbb{Z}_{n-1}$, $G_{\lambda,n}(\tilde{\beta}, \mathbb{Z}_{n-1})$ is bounded on the set $\Delta^{(1)}_{BL}$.

Let $(\tilde{x}^o_{1c}, y^o_{1c}) = ((1, x^o_{1c})^T, y^o_{1c})$. For any small $\varepsilon > 0$, define $E_{c,\varepsilon} = \{\tilde{\beta} : \|\tilde{\beta}\| \geq \varepsilon, |\tilde{\beta}^T \tilde{x}^o_{1c}|/\|\tilde{\beta}\|\|\tilde{x}^o_{1c}\| > \varepsilon\}$ and $D^+_{c,\varepsilon} := \{\tilde{\beta} : \operatorname{sign}(\tilde{\beta}^T \tilde{x}^o_{1c} y^o_{1c}) = 1\} \cap E_{c,\varepsilon}$. Note that $\|\tilde{x}^o_{1c}\| \geq c\|\beta_0\| \to \infty$, as $c \to \infty$. Thus, as $c \to \infty$, for any $\tilde{\beta} \in D^+_{c,\varepsilon}$, we have

$$\begin{aligned} (1 + \tilde{\beta}^T \tilde{x}^o_{1c} y^o_{1c})^2 - (1 - \tilde{\beta}^T \tilde{x}^o_{1c} y^o_{1c})^2 &= 2\|\tilde{\beta}\|\|\tilde{x}^o_{1c}\| \frac{\tilde{\beta}^T \tilde{x}^o_{1c}}{\|\tilde{\beta}\|\|\tilde{x}^o_{1c}\|} \to \infty, \\ \min\{(1 + \tilde{\beta}^T \tilde{x}^o_{1c} y^o_{1c})^2, (1 - \tilde{\beta}^T \tilde{x}^o_{1c} y^o_{1c})^2\} &\to \infty. \end{aligned} \tag{A.2}$$

Define $D^-_{c,\varepsilon} = -D^+_{c,\varepsilon}$. Then for any $\tilde{\beta} \in D^+_{c,\varepsilon}$, we have $-\tilde{\beta} \in D^-_{c,\varepsilon}$. The first part of (A.2) indicates that for any $\tilde{\beta} \in D^+_{c,\varepsilon}$, we have $F_n(\tilde{\beta}, z^o_{1c}) < F_n(-\tilde{\beta}, z^o_{1c})$. That is, the minimum of $F_n(\tilde{\beta}, z^o_{1c})$ is achieved in the set $D^+_{c,\varepsilon} \cup D^0_c \cup E^c_{c,\varepsilon}$, where $D^0_c = \{\tilde{\beta} : \tilde{\beta}^T \tilde{x}_{1c} y^o_{1c} = 0\}$ and $E^c_{c,\varepsilon}$ denotes the complement of $E_{c,\varepsilon}$. The second part of (A.2) shows that the optimal value $\hat{\tilde{\beta}}$ is achieved only in the set $D^0_c \cup E^c_{c,\varepsilon}$.

When $\hat{\tilde{\beta}} \in D^0_c$, it follows that $\hat{b} - c\hat{\beta}^T \beta_0 = 0$. Since $c \to \infty$, $\beta_0 \neq 0$ and $|\hat{b}| < M$, we know that $\hat{\beta}^T \beta_0 \to 0$, as $c \to \infty$. Therefore, $\hat{\beta} \in S^-_0 = \{\beta : \beta^T \beta_0 \leq 0\}$. When $\hat{\tilde{\beta}} \in E^c_{c,\varepsilon}$, then $\hat{\tilde{\beta}} \in \{\tilde{\beta} : \|\tilde{\beta}\| < \varepsilon\} \cup \{\tilde{\beta} : |\tilde{\beta}^T \tilde{x}^o_{1c}|/\|\tilde{\beta}\|\|\tilde{x}^o_{1c}\| < \varepsilon\} := A_1 \cup A_2$, where A_1 and A_2 are defined accordingly. When $\hat{\tilde{\beta}} \in A_1$, since ε can be arbitrarily small, letting $\varepsilon \to 0$, we have $\|\hat{\beta}\| \to 0$. Next, we consider the case of $\hat{\tilde{\beta}} \in A_2$. Since $\tilde{x}^o_{1c} = (1, -c\beta^T_0)^T$, it follows that $\tilde{x}^o_{1c}/\|\tilde{x}^o_{1c}\| \to (0, -\beta^T_0)^T/\|\beta_0\|$, as $c \to \infty$. Letting $\varepsilon \to 0$, we have $|\hat{\beta}^T \beta_0|/\|\hat{\beta}\|\|\beta_0\| \to 0$. Moreover, it is easy to see that $\|\hat{\beta}\| \leq \lambda^{-1}\ell(0)$. Therefore, $\hat{\beta} \in S^-_0 = \{\beta : \beta^T \beta_0 \leq 0\}$. This completes the proof. □

A.2. Proof of Theorem 3.

PROOF. We only prove the conclusions on the population angular breakdown point. The conclusion on the sample angular breakdown point can be proved by the same argument, where f_0 is replaced by $\hat{f}_{\mathbb{Z}_n}$ in the proof.

Denote the orthogonal basis of $\mathcal{H}$ by $e_i(x)$, $i = 1, \dots, d$. Then $K(x, y) = \sum_{i=1}^{d} e_i(x)e_i(y)$ with the feature map $\phi : x \mapsto (e_1(x), \dots, e_d(x))$. Since $f \in \mathcal{H}$, we have $f(x) = \sum_{i=1}^{d} \alpha_i e_i(x)$ for some constant α_i. Therefore, we have $f(\cdot) = [\alpha_1, \dots, \alpha_d]^T$. By the assumption, for any series $\{M_c, c = 1, 2, \dots, \}$ with $M_c \to \infty$, we have d different elements $\{x_{c,i}, i = 1, \dots, d\} \in \mathcal{X}$, $c = 1, 2, \dots,$ such that $\phi(x_{c,i})$, $i = 1, \dots, d$ are linear independent. Consequently, $\operatorname{span}\{\phi(x_{c,i}), i = 1, \dots, d\} = R^d$. Recall that $f(\cdot) \in R^d$. Without loss of generality, we assume that there exists constants $\gamma_{c,1}, \dots, \gamma_{c,d}$ being positive, such that

$$
\begin{aligned}
f_0(\cdot) &= \sum_{1 \le i \le d_0} \gamma_{c,i} \frac{\phi(x_{c,i})}{\sqrt{K(x_{c,i}, x_{c,i})}} - \sum_{i \ge d_0+1} \gamma_{c,i} \frac{\phi(x_{c,i})}{\sqrt{K(x_{c,i}, x_{c,i})}} \\
&:= \sum_{1 \le i \le d_0} \gamma_{c,i} \tilde{\phi}(x_{c,i}) - \sum_{d_0+1 \le i \le d} \gamma_{c,i} \tilde{\phi}(x_{c,i}).
\end{aligned} \tag{A.3}
$$

Then we take the outlier $(x^o_{c,i}, y^o_{c,i})$, $i = 1, \dots, d$, such that $x^o_{c,i} = x_{c,i}$ and $y^o_{c,i} = -1$ for $1 \le i \le d_0$ and $y^o_{c,i} = 1$ as $d_0 + 1 \le i \le d$. Therefore, for any $f \in \mathcal{H}$ and any $1 \le i \le d_0$,

$$\langle f, y^o_{c,i}\phi(x^o_{c,i})\rangle = -\langle f, \phi(x_{c,i})\rangle = -\sqrt{K(x_{c,i}, x_{c,i})}\langle f, \tilde{\phi}(x_{c,i})\rangle$$

and for $d_0 + 1 \le i \le d$,

$$\langle f, y^o_{c,i}\phi(x^o_{c,i})\rangle = \langle f, \phi(x_{c,i})\rangle = \sqrt{K(x_{c,i}, x_{c,i})}\langle f, \tilde{\phi}(x_{c,i})\rangle.$$

Let $S_c := \{f : \langle f, \tilde{\phi}(x_{c,i})\rangle > 0, \text{for some } 1 \le i \le d_0\} \cup \{f : \langle f, \tilde{\phi}(x_{c,i})\rangle < 0 \text{ for some } d_0 + 1 \le i \in d\}$. For any $\varepsilon > 0$ and any $f \in S_c$ with $\|f\|_{\mathcal{H}} > \varepsilon$, due to $K(x_{c,i}, x_{c,i}) \to \infty$, as $c \to \infty$, we have $\min_{1 \le i \le d}\langle f, y^o_{c,i}\phi(x^o_{c,i})\rangle \to -\infty$. Consequently, $\max_{1 \le i \le d} \ell(\eta + \langle f, y^o_{c,i}\phi(x^o_{c,i})\rangle) \to \infty$ according to assumption (A1) and $(\eta, f) \in \Delta_{BK}$. Therefore, as ε being sufficiently small and c being sufficiently large, it follows that the optimal solution $\hat{f}_\lambda \in S^c_c = \{f : \langle f, \tilde{\phi}(x_{c,i})\rangle \le 0, \text{for all } 1 \le i \le d_0\} \cap \{f : \langle f, \tilde{\phi}(x_{c,i})\rangle \ge 0 \text{ for all } d_0 + 1 \le i \in d\}$. Consequently, by (A.3), we know $\langle \hat{f}_\lambda, f_0\rangle \le 0$. □

SUPPLEMENTARY MATERIAL

Supplement to "Assessing robustness of classification using an angular breakdown point" (DOI: 10.1214/17-AOS1661SUPP; .pdf). The supplementary material contains the remaining proof of the theoretical results.

REFERENCES

BIGGIO, B., NELSON, B. and LASKOV, P. (2012). Poisoning attacks against support vector machines. In *Proceedings of the* 29*th International Conference on Machine Learning* (J. Langford and J. Pineau, eds.) 1807–1814. ACM, New York.

CHRISTMANN, A. and STEINWART, I. (2004). On robustness properties of convex risk minimization methods for pattern recognition. *J. Mach. Learn. Res.* **5** 1007–1034.

DONOHO, D. and HUBER, P. J. (1983). The notion of breakdown point. In *A Festschrift for Erich L. Lehmann* 157–184. Wadsworth, Belmont, CA. MR0689745

FREUND, Y. and SCHAPIRE, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. System Sci.* **55** 119–139. MR1473055

GENTON, M. G. and LUCAS, A. (2003). Comprehensive definitions of breakdown points for independent and dependent observations. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **65** 81–94. MR1959094

HABLE, R. and CHRISTMANN, A. (2011). On qualitative robustness of support vector machines. *J. Multivariate Anal.* **102** 993–1007. MR2793871

HAMPEL, F. R. (1971). A general qualitative definition of robustness. *Ann. Math. Stat.* **42** 1887–1896. MR0301858

HAMPEL, F. R. (1974). The influence curve and its role in robust estimation. *J. Amer. Statist. Assoc.* **69** 383–393.

HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2009). *The Elements of Statistical Learning*: *Data Mining*, *Inference*, *and Prediction*, 2nd ed. Springer, New York. MR2722294

HORST, R. and THOAI, N. V. (1999). DC programming: Overview. *J. Optim. Theory Appl.* **103** 1–43.

HUBERT, M., ROUSSEEUW, P. J. and VAN AELST, S. (2008). High-breakdown robust multivariate methods. *Statist. Sci.* 92–119.

KANAMORI, T., FUJIWARA, S. and TAKEDA, A. (2014). Breakdown point of robust support vector machine. Preprint. Available at arXiv:1409.0934.

KIMELDORF, G. S. and WAHBA, G. (1970). A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. *Ann. Math. Stat.* **41** 495–502. MR0254999

KRAUSE, N. and SINGER, Y. (2004). Leveraging the margin more carefully. In *Proceedings of the Twenty-First International Conference on Machine Learning* 63. ACM, New York.

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. *Ann. Statist.* **28** 1570–1600. MR1835032

LIU, Y. and SHEN, X. (2006). Multicategory ψ-learning. *J. Amer. Statist. Assoc.* **101** 500–509.

MASON, L., BAXTER, J., BARTLETT, P. and FREAN, M. (2000). Boosting algorithms as gradient descent. In *In Advances in Neural Information Processing Systems* **12** 512–518. MIT Press, Cambridge.

RONCHETTI, E. (1997). Robust inference by influence functions. *J. Statist. Plann. Inference* **57** 59–72.

SAKATA, S. and WHITE, H. (1995). An alternative definition of finite-sample breakdown point with applications to regression model estimators. *J. Amer. Statist. Assoc.* **90** 1099–1106. MR1354027

SCHÖLKOPF, B. and SMOLA, A. J. (2002). *Learning with Kernels*: *Support Vector Machines*, *Regularization*, *Optimization*, *and Beyond*. MIT Press, Cambridge, MA.

SCHOLKOPF, B., SMOLA, A. J., WILLIAMSON, R. C. and BARTLETT, P. L. (2000). New support vector algorithms. *Neural Comput.* **12** 1207–1245.

SHEN, X., TSENG, G. C., ZHANG, X. and WONG, W. H. (2003). On ψ-learning. *J. Amer. Statist. Assoc.* **98** 724–734.

STROMBERG, A. J. and RUPPERT, D. (1992). Breakdown in nonlinear regression. *J. Amer. Statist. Assoc.* **87** 991–997.

VAPNIK, V. N. (1998). *Statistical Learning Theory*. Wiley, New York. MR1641250

WAHBA, G. (1990). *Spline Models for Observational Data* **59**. SIAM, Philadelphia, PA.

WU, Y. and LIU, Y. (2007). Robust truncated hinge loss support vector machines. *J. Amer. Statist. Assoc.* **102** 974–983. MR2411659

XU, L., CRAMMER, K. and SCHUURMANS, D. (2006). Robust support vector machine training via convex outlier ablation. In *American Association for Artificial Intelligence* **6** 536–542.

YU, Y., YANG, M., XU, L., WHITE, M. and SCHUURMANS, D. (2010). Relaxed clipping: A global training method for robust regression and classification. In *Advances in Neural Information Processing Systems* 2532–2540.

ZHAO, J., YU, G. and LIU, Y. (2018). Supplement to "Assessing robustness of classification using an angular breakdown point." DOI:10.1214/17-AOS1661SUPP.

J. ZHAO
DEPARTMENT OF STATISTICS
BEIJING NORMAL UNIVERSITY
XINWAI DISTRICT NO.19
BEIJING, 100875
P.R. CHINA
E-MAIL: zjlczh@126.com

G. YU
DEPARTMENT OF BIOSTATISTICS
STATE UNIVERSITY OF NEW YORK
AT BUFFALO
BUFFALO, NEW YORK 14214-3000
USA
E-MAIL: guanyu@buffalo.edu

Y. LIU
DEPARTMENT OF STATISTICS AND OPERATIONS
RESEARCH
DEPARTMENT OF GENETICS
DEPARTMENT OF BIOSTATISTICS
CAROLINA CENTER FOR GENOME SCIENCES
LINEBERGER COMPREHENSIVE CANCER CENTER
UNIVERSITY OF NORTH CAROLINA AT CHAPEL HILL
CHAPEL HILL, NORTH CAROLINA 27599-3260
USA
E-MAIL: yfliu@email.unc.edu