Leaky Hinge Loss: The First Negatively Divergent Margin-based Loss Function for Classification

Oh-Ran Kwon and Hui Zou*
School of Statistics, University of Minnesota

September, 2022

Abstract

Many modern classification algorithms are formulated through the regularized empirical risk minimization (ERM) framework where the risk is defined based on a loss function. We point out that although the loss function in decision theory is non-negative by definition, non-negativity of the loss function in ERM is not necessary in order to be classification-calibrate and to produce a Bayes consistent classifier. We introduce the leaky hinge loss, the first negatively divergent margin-based loss function. We prove that the leaky hinge loss is classification-calibrated. When the hinge loss is replaced with the leaky hinge loss in the EMR approach for deriving the kernel support vector machine (SVM), the corresponding optimization problem has a well-defined solution named the kernel leaky SVM. Under mild regularity conditions, we prove that the kernel leaky SVM is Bayes risk consistent. In our theoretical analysis, we overcome multiple challenges caused by the negative divergence of the leaky hinge loss that does not exist in the analysis of the usual kernel machines. For a numerical demonstration, we provide a computationally efficient algorithm to solve the kernel leaky SVM and compare it to the kernel SVM on simulated data and fifteen benchmark real datasets.

Keywords: Bayes risk consistency, Classification-calibrated, Loss function, Majorization minimization principle, Margin maximizing

1 Introduction

This paper concerns binary classification where the task is to predict an unobserved binary output value $y \in \{-1, 1\}$ based on an observed input vector $\mathbf{x} \in \mathbb{R}^p$. The classifier is a mapping from the input space $\mathcal{X}$ to $\{-1, 1\}$ via a classification function $\hat{f}$, and the predicted y value is $\text{sgn}[\hat{f}(\mathbf{x})]$. The decision boundary is the set $\{\mathbf{x} : \hat{f}(\mathbf{x}) = 0\}$. Suppose that data

*Corresponding author: Hui Zou, zouxx019@umn.edu. This work is supported in part by NSF grants 2015120 and 2220286.

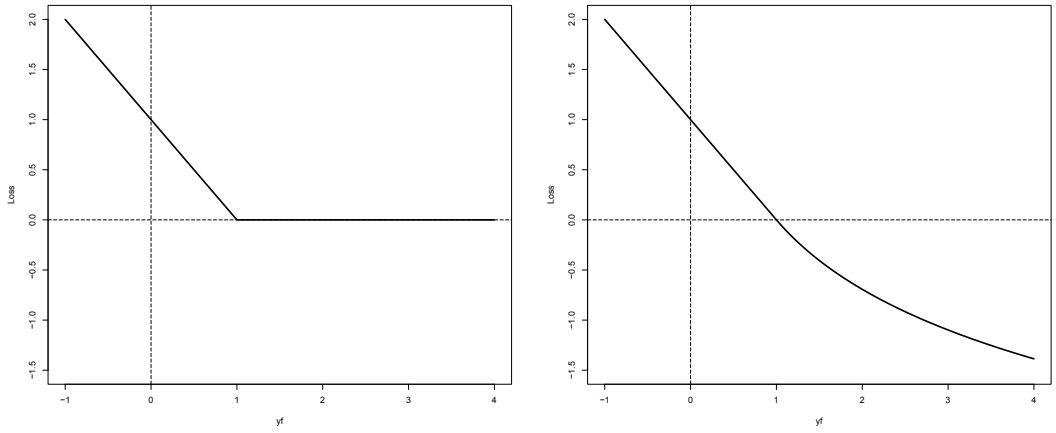
are generated from some underlying distribution $\mathrm{P}(\mathbf{X}, \mathbf{Y})$, and let $p(\mathbf{x}) = \mathrm{P}(\mathbf{Y} = 1|\mathbf{X} = \mathbf{x})$. Under the standard 0-1 loss, the optimal classification rule is $\operatorname{sgn}[p(\mathbf{x}) - 1/2]$ (a.k.a. Bayes rule). Throughout the paper, we assume that $p(\mathbf{X}) \neq \frac{1}{2}$ almost surely. The *optimal decision boundary* is $\{\mathbf{x} : p(\mathbf{x}) = 1/2\}$. Given training data, $\{(\mathbf{x}_i, y_i), i = 1, \ldots, n\}$, one aims to develop a classifier that mimics the Bayes rule as closely as possible. Extensive research has been devoted to classification, and many classification algorithms have been developed and widely used in practice, ranging from the classical methods such as discriminant analysis and logistic regression to modern techniques such as support vector machines (SVM) (Vapnik, 2013), boostings (Freund et al., 1996), random forests (Breiman, 2001), neural networks, and deep learning (Goodfellow et al., 2016).

Regularized empirical risk minimization (ERM) is a fundamental framework for designing a new classification algorithm and analyzing its statistical properties. The empirical risk is defined as $\frac{1}{n}\sum_{i=1}^{n} L(y_i f(\mathbf{x}_i))$, then a classification algorithm is derived by trying to minimize the empirical risk via a regularized method. Many classification algorithms such as Kernel SVM and 1-norm SVM (Zhu et al., 2003) can be cast in this framework. Also, boostings can be viewed as minimizing the empirical risk with an ℓ_1 penalty (Rosset et al., 2004). In the literature, $yf(\mathbf{x})$ is called the margin and L is referred to as a margin-based loss function. Obviously, the terms "risk" and "loss function" are borrowed from the statistical decision theory where a loss function is naturally non-negative. Note that the loss function in ERM is used to derive the classifier, while the loss function in the decision theory is used to measure the theoretical performance of a statistical method. In classification, the loss for measuring performance is usually the 0-1 loss as previously stated, whereas the loss function in ERM can be far more flexible. For example, the SVM uses the hinge loss, the logistic regression uses the logit loss, and AdBoost uses the exponential loss (Hastie et al., 2009; Friedman et al., 2000). Of course, all these loss functions are non-negative, which make them qualified as loss functions in decision theory. A really interesting question, which has not been asked before in the literature, is that can we use a function that has negative values in ERM for classification? In the ERM framework, a loss function being bounded from below is equivalent to being non-negative because we can vertically lift the loss function without changing the regularized ERM problem (as a constant does not affect the minimization). Thus, the real question is **whether we could use a negatively-divergent function in ERM for classification**.

In this paper, we provide an affirmative answer and the new function called the leaky hinge loss. The expression of the leaky hinge loss is given in (1) and the picture of this loss function is displayed in Figure 1.

$$L(yf) = \begin{cases} -\log yf, & yf > 1, \\ 1 - yf, & yf \leq 1, \end{cases} \tag{1}$$

Given the training data, the empirical leaky hinge risk is $\frac{1}{n}\sum_{i=1}^{n} L(y_i f(\mathbf{x}_i))$. If we use the notion from decision theory, the leaky hinge loss should not be called a loss function as its values diverge to negative infinity as the margin approaches positive infinity. Nevertheless, we still use loss in the name to follow the convention. A positive margin means the classification is correct. When the margin is larger than 1, the leaky hinge loss becomes negative, meaning that it actually gives a reward. The larger the margin, the bigger the reward. Intuitively, this



(a) Hinge loss (b) Leaky hinge loss

Figure 1: Plot of the hinge loss and the leaky hinge loss

sounds reasonable. Of course, we need to provide formal theoretical and numerical evidence for justifying the use of the leaky hinge loss in ERM, which is the main focus of this paper.

There have been a lot of studies on the choice of a loss function in ERM for classification. Lin (2004) proposed *Fisher consistency* which requires any global minimizer of $\mathrm{E}[L(\mathbf{Y}f(\mathbf{X}))]$ has the same sign as the Bayes rule almost surely. Bartlett et al. (2006) proposed a more-refined *classification-calibration* condition, requiring that a global minimizer of $\mathrm{E}[L(\mathbf{Y}f(\mathbf{X}))|\mathbf{X}=\mathbf{x}]$ has the same sign as the Bayes rule almost surely. The two conditions usually coincide for non-negative loss functions, and the two are used as the same condition. It is often easy to verify these conditions for non-negative convex loss functions, but non-negative non-convex loss functions can also satisfy these conditions. A famous example is ψ-learning loss (Shen et al., 2003). There is no result on whether a negatively-divergent L can be classification-calibrated or Fisher consistent. In order to answer this question, we first need to check whether a global minimizer of $\mathrm{E}[L(\mathbf{Y}f(\mathbf{X}))]$ or $\mathrm{E}[L(\mathbf{Y}f(\mathbf{X}))|\mathbf{X}=\mathbf{x}]$ is well-defined in the sense that the minimizer is finite-valued and the minimum objective is finite-valued. This technical issue is non-trivial to address when L is negatively-divergent. For the leaky hinge loss, our analysis reveals that the global minimizer of $\mathrm{E}[L(\mathbf{Y}f(\mathbf{X}))|\mathbf{X}=\mathbf{x}]$ is always well-defined and has the same sign as the Bayes rule, but the global minimizer of $\mathrm{E}[L(\mathbf{Y}f(\mathbf{X}))]$ may not be well-defined unless some further conditions are imposed.

The fundamental justification for a loss function in ERM for classification is the Bayes consistency, that is, the expected misclassification rate of the resulting classifier converges to that of the Bayes rule as the sample size increases. When the leaky hinge loss is used in the EMR approach to derive a classifier, the resulting classifier is named the kernel leaky SVM. We establish the Bayes consistency of the kernel leaky SVM, which in turns offers the most important justification of the leaky hinge loss. Therefore, we can claim that the leaky hinge loss is the first negatively divergent margin loss function for classification. Bayes consistency has been established for some kernel machines (Zhang, 2004; Steinwart, 2005), but their studies are limited on non-negative loss functions. Specifically, they analyzed the expected misclassification rate of the derived classifier by some quantities associated with

the loss function, but those quantities cannot be applicable in the case of the leaky hinge loss. We use some new techniques to prove the Bayes consistency of the kernel leaky SVM.

The geometry of SVMs is best described in the linear space where its margin maximization interpretation is clearly shown. Rosset et al. (2003) showed that this geometric interpretation is in fact shared by a class of non-negative loss function that vanishes to zero quickly enough, such as the hinge loss, the exponential loss and the binomial deviance loss (or the logit loss). Their result provides the unified margin maximization view of many popular classification algorithms. However, their theory does not cover the leaky hinge loss because the leaky hinge loss violates their conditions. Nevertheless, we show that the linear leak SVM also has an interesting and new margin maximization view. This result suggests that the linear leaky SVM and the linear SVM can be very different, although their kernel versions approach to the same limit (Bayes rule).

For a numerical demonstration, we develop an efficient algorithm to solve the leaky SVMs. This allows us to compare the leaky SVM to the standard SVM. We do the extensive comparison of the kernel leaky SVM and the kernel SVM using simulated data and 15 benchmark datasets from Dua and Graff (2017). The linear leaky SVM outperforms the linear SVM on 10 out 15 benchmark datasets, and the kernel leaky SVM and the kernel SVM have more similar performances.

The remainder of the paper is organized as follows. In Section 2, we prove that the leaky hinge loss function is classification-calibrated. In Section 3, we prove that the linear leaky SVM is well-defined given a training data. When the data is linearly separable, we show the margin-maximization picture of the linear leaky SVM and compare it with the margin-maximization picture of the linear SVM. In Section 4 we consider the kernel leaky SVM in a reproducing kernel Hilbert space (RKHS). We first prove that the kernel leaky SVM is well-defined on a training data. We then derive an efficient algorithm to solve the kernel leaky SVM. In Section 5, we establish the Bayes risk consistency of the kernel leaky SVM. Section 6 contains the numerical results. Technical details and proofs are provided in a supplementary file to this paper.

2 The Classification-calibration Property

Lin (2004) proposed *Fisher consistency* as a necessary condition on a loss function. It is defined to be that any global minimizer $\check{f}$ (if it exists) of $\mathrm{E}[L(\mathbf{Y}f(\mathbf{X}))]$, generally referred to as a population minimizer, has the same sign function as the Bayes rule almost surely. Later, Bartlett et al. (2006) defined *classification-calibration*. A loss function L is called classification-calibrated if any global minimizer $\bar{f}$ (if it exists) of $\mathrm{E}[L(\mathbf{Y}f(\mathbf{X}))|\mathbf{X}=\mathbf{x}]$ has the same sign as the Bayes rule almost surely. For example, the hinge loss is classification-calibrated, which can be directly shown by Theorem 2 of Bartlett et al. (2006).

An unspoken assumption in those two definitions is the existence of a global minimizer. When $\bar{f}$ exists (which can be easily checked) and $|\mathrm{E}[L(\mathbf{Y}\bar{f}(\mathbf{X}))]| < \infty$, then a population minimizer $\check{f}$ exists. We further see that loss function L is classification-calibrated if and only if it is Fisher consistent. If statement follows from the fact that any $\bar{f}$ is also a population minimizer and the definition of Fisher consistency. Only if statement comes from the fact that $\mathrm{E}[L(\mathbf{Y}\bar{f}(\mathbf{X}))|\mathbf{X}] = \mathrm{E}[L(\mathbf{Y}\check{f}(\mathbf{X}))|\mathbf{X}]$ almost surely and that for any $\delta_{\mathbf{X}}$ such

that $\text{sgn}(\delta_{\mathbf{X}}) \neq \text{sgn}(p(\mathbf{X}) - 1/2)$, $\text{E}[L(\mathbf{Y}\bar{f}(\mathbf{X}))|\mathbf{X}] < \text{E}[L(\mathbf{Y} \cdot \delta_{\mathbf{X}})|\mathbf{X}]$ almost surely by the definition of classification-calibration.

In the case of non-negative loss function L, $\text{E}[L(\mathbf{Y}\bar{f}(\mathbf{X}))]$ is always finite. However, the leaky hinge loss does not guarantee the finitness of $\text{E}[L(\mathbf{Y}\bar{f}(\mathbf{X}))]$ (see Theorem 4 for further information), thereby not ensuring the existence of a global minimizer that Fisher consistency implicitly assumes. Instead, we verify that the leaky hinge loss is classification-calibrated. The leaky hinge loss is the first negatively divergent loss function satisfying the classification-calibration property.

Theorem 1. *(Classification-calibration) Let L be the leaky hinge loss. For any $\mathbf{x}$ such that $p(\mathbf{x}) \in (0, 1)$, a global minimizer of $\text{E}[L(\mathbf{Y}f(\mathbf{X}))|\mathbf{X} = \mathbf{x}]$ uniquely exists and is,*

$$\bar{f}(\mathbf{x}) = \begin{cases} -\frac{1-p(\mathbf{x})}{p(\mathbf{x})}, & \text{if } p(\mathbf{x}) < \frac{1}{2}, \\ +\frac{p(\mathbf{x})}{1-p(\mathbf{x}),} & \text{if } p(\mathbf{x}) > \frac{1}{2}. \end{cases}$$

Also, $\text{sgn}[\bar{f}(\mathbf{x})] = f^*(\mathbf{x})$, where $f^*(\mathbf{x}) = \text{sgn}\left[p(\mathbf{x}) - \frac{1}{2}\right]$ is the Bayes rule.

Remark A. In general, a global minimizer of $\text{E}[L(\mathbf{Y}f(\mathbf{X}))|\mathbf{X} = \mathbf{x}]$ is allowed to take values $\pm\infty$ because what matters is only the sign of the minimizer (Lin, 2004). If we allow the minimum objective to take values $\pm\infty$, Theorem 1 can be extended to include $p(\mathbf{x})$ equals to 0 and 1. It is because if $p(\mathbf{x}) = 0$, then $\text{E}[L(\mathbf{Y}\alpha)|\mathbf{X} = \mathbf{x}] = L(-\alpha) \to -\infty$ as $\alpha \to -\infty$. If $p(\mathbf{x}) = 1$, then $\text{E}[L(\mathbf{Y}f(\mathbf{X}))|\mathbf{X} = \mathbf{x}] = L(\alpha) \to -\infty$ as $\alpha \to \infty$.

3 The Linear Leaky SVM

Theorem 1 offers a justification for the leaky hinge loss with an infinite amount data. In applications, the data size is always finite. Thus, we need to further study the properties of classifiers using the leaky hinge loss. We first examine the linear leaky SVM to understand the characteristics of the leaky hinge loss.

3.1 Existence of the global solution

When L is a non-negative continuous loss function, the existence of a global minimizer to the regularized ERM problem is obvious. It is because the sublevel set of the objective function $\mathcal{L}$ of the regularized ERM problem, $\{\boldsymbol{\beta} : \mathcal{L}(\boldsymbol{\beta}) \leq c\}$, is compact for any $c \in \mathbb{R}$, and the global minimizer of the continuous objective function on a compact set must exist by the extreme value theorem.

In contrast, it is not trivial to show the existence of a global minimizer of the linear leaky SVM because the loss term in the regularized ERM can diverge to negative infinity even if the leaky hinge loss is convex and training data have finite samples. To show the existence, we find a specific compact set D of $\boldsymbol{\beta}$, so that the objective function of the linear leaky SVM can have a global minimizer on D. Then, we show that the objective value at the global minimizer on D is always smaller than any objective values on D^c.

Theorem 2. *(Existence of global solution) Let* $(\mathbf{x}_i, y_i) \in \mathbf{R}^d \times \{-1, +1\}$ *for* $i = 1, \ldots, n$ *be training data. Suppose there exist* i, j *with* $y_i = +1$ *and* $y_j = -1$*. Then, there exists a global solution to,*

$$\min \mathcal{L}(\beta_0, \boldsymbol{\beta}) = \min_{\beta_0, \boldsymbol{\beta}} \left[\frac{1}{n} \sum_{i=1}^{n} L\left(y_i(\beta_0 + \mathbf{x}_i^T \boldsymbol{\beta})\right) + \lambda \boldsymbol{\beta}^T \boldsymbol{\beta} \right], \tag{2}$$

where L *is the leaky hinge loss and* $\lambda > 0$ *is a tuning parameter.*

Remark B. (Uniqueness) The minimizer $\hat{\boldsymbol{\beta}}$ is uniquely determined because the leaky hinge loss is convex and the ℓ_2 regularizer is strictly convex. However, $\hat{\beta}_0$ is not. Here is an illustrative example. Let the data be

$$\begin{aligned} &y_1 = -1, \quad y_2 = -1, \quad y_3 = 1, \quad y_4 = 1, \\ &x_1 = 1, \quad x_2 = -1, \quad x_3 = 1, \quad x_4 = -1. \end{aligned}$$

Then both $(0, 0)$ and $(1, 0)$ are global minimizers. Non-uniqueness of the intercept term can also occur in the linear SVM.

3.2 A geometric picture

The standard SVM has a well-known geometric interpretation when the training data are linearly separable, i.e., when there exists $\bar{\mathbf{w}}$ such that $\min_i y_i \mathbf{x}_i^T \bar{\mathbf{w}} > 0$. In such a separable case, it finds a decision boundary that maximizes the minimal margin. Rosset et al. (2003) discussed a family of loss functions that shares the same margin picture as that of SVM. Specifically, they investigated for which loss functions, the solution of the regularized ERM,

$$\hat{\boldsymbol{\beta}}_\lambda = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} L(y_i \mathbf{x}_i^T \boldsymbol{\beta}) + \lambda \|\boldsymbol{\beta}\|_q^q, \tag{3}$$

where $q \geq 1$, converges to the minimal margin maximizer as the regularizer disappears. They found that if a loss function is non-negative and vanishes quickly enough to 0, then as $\lambda \to 0$, every convergent point of $\frac{\hat{\boldsymbol{\beta}}_\lambda}{\|\hat{\boldsymbol{\beta}}_\lambda\|_q}$ is

$$\arg \max_{\|\mathbf{w}\|_q = 1} \min y_i \mathbf{w}^T \mathbf{x}_i. \tag{4}$$

The family of loss functions covers the hinge loss, the exponential loss, and the binomial deviance loss. Their result provides a unified view of popular classification algorithms in that they converge to the same solution provided the same regularizer. For example, Boosting, 1-norm SVM (Zhu et al., 2003), and ℓ_1 penalized logistic regression give the same classifier at the limit.

Interestingly, the leaky hinge loss violates their sufficient condition, and we find that the leaky hinge loss optimizes a different margin at the limit. Still, the convergent point finds the separating hyperplane that can perfectly separate the data.

Theorem 3. *Assume training data,* $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$*, are separable, i.e., there exists* $\bar{\mathbf{w}}$ *such that* $\bar{m} = \min_i y_i \mathbf{x}_i^T \bar{\mathbf{w}} > 0$ *with* $\|\bar{\mathbf{w}}\|_q = 1$*,* $q \geq 1$*. Let* $\hat{\boldsymbol{\beta}}_\lambda$ *be the solutions to* (3) *with the leaky*

hinge loss and the ℓ_q regularizer. Then, as $\lambda \to 0$, any convergence point of $\frac{\hat{\boldsymbol{\beta}}_\lambda}{\|\hat{\boldsymbol{\beta}}_\lambda\|_q}$ maximizes the product of the positive part of margins,

$$\prod_{i=1}^{n} y_i \mathbf{x}_i^T \mathbf{w} \mathbb{1}\{y_i \mathbf{x}_i^T \mathbf{w} \geq 0\},$$

where $\mathbb{1}(\cdot)$ is the indicator function. If the maximizer is unique, we can conclude that

$$\frac{\hat{\boldsymbol{\beta}}_\lambda}{\|\hat{\boldsymbol{\beta}}_\lambda\|_q} \to \arg\max_{\|\mathbf{w}\|_q=1} \prod_{i=1}^{n} y_i \mathbf{x}_i^T \mathbf{w} \mathbb{1}\{y_i \mathbf{x}_i^T \mathbf{w} \geq 0\}. \tag{5}$$

We visualize the two different separating hyperplanes defined in (4) and (5) by using data generated from the following model. Suppose that $\mathbf{X} \in \mathbb{R}^2$ in each class is from the mixture of three Gaussian distributions that $\mathbf{X} \sim \frac{1}{3}\sum_{i=1}^{n} N(\mu_i^-, 0.6 \cdot \mathbf{I})$ if $y = -1$ and $\mathbf{X} \sim \frac{1}{3}\sum_{i=1}^{3} N(\mu_i^+, 0.6 \cdot \mathbf{I})$ if $y = +1$, where $\mathbf{I}$ is an identity matrix. We randomly generate $\mu_i^-, i = 1, 2, 3$, from $N\left((1.8, -1.8)^T, \mathbf{I}\right)$ and $\mu_i^+, i = 1, 2, 3$, from $N\left((-1.8, 1.8)^T, \mathbf{I}\right)$. In each plot in Figure 2, we see the separable data of 20 drawn from the distribution. Since the generating distribution is known for each class, the optimal decision boundary (solid line in Figure 2) can be calculated exactly. Figure 2 (a)-(d) show the decision boundary from the new margin maximizer defined in (5) (long-dashed line), along with that from the standard margin maximizer defined in (4) (dashed line). The two boundaries are similar to each other in (a) and (b) while they are noticeably different in (c) and (d).

4 The Kernel Leaky SVM

4.1 Formulation

The linear leaky SVM could be restrictive in practice when the Bayes rule is highly nonlinear. In order to obtain a nonlinear classifier boundary with the leaky hinge loss, we consider a nonparametric approach in a reproducing kernel Hilbert space (RKHS) by following the statistical derivation of the kernel SVM (Hastie et al., 2009).

Let $\mathcal{H}_K$ be the RKHS generated by a positive definite kernel K. We define *kernel leaky SVM* as the classifier $\text{sgn}\{\hat{\alpha}_0 + \hat{h}(\mathbf{x})\}$ where $(\hat{\alpha}_0, \hat{h})$ is the solution to

$$\min_{\substack{\alpha_0 \in \mathbb{R} \\ h \in \mathcal{H}_K}} \left(\frac{1}{n} \sum_{i=1}^{n} L\left(y_i(\alpha_0 + h(\mathbf{x}_i))\right) + \lambda \|h\|_{\mathcal{H}_K}^2 \right). \tag{6}$$

While (6) is defined over an infinite dimensional space, it can be shown by the representer theorem (Wahba, 1990) that the solution is finite dimensional and has the form,

$$\hat{h}(\mathbf{x}) = \sum_{i=1}^{n} \hat{\alpha}_i K(\mathbf{x}, \mathbf{x}_i), \text{ and thus } ||\hat{h}||_{\mathcal{H}_K}^2 = \sum_{i=1}^{n}\sum_{j=1}^{n} K(\mathbf{x}_i, \mathbf{x}_j)\hat{\alpha}_i\hat{\alpha}_j. \tag{7}$$

We note that the representer theorem holds irrespective of whether a loss function is non-negative or negatively divergent.

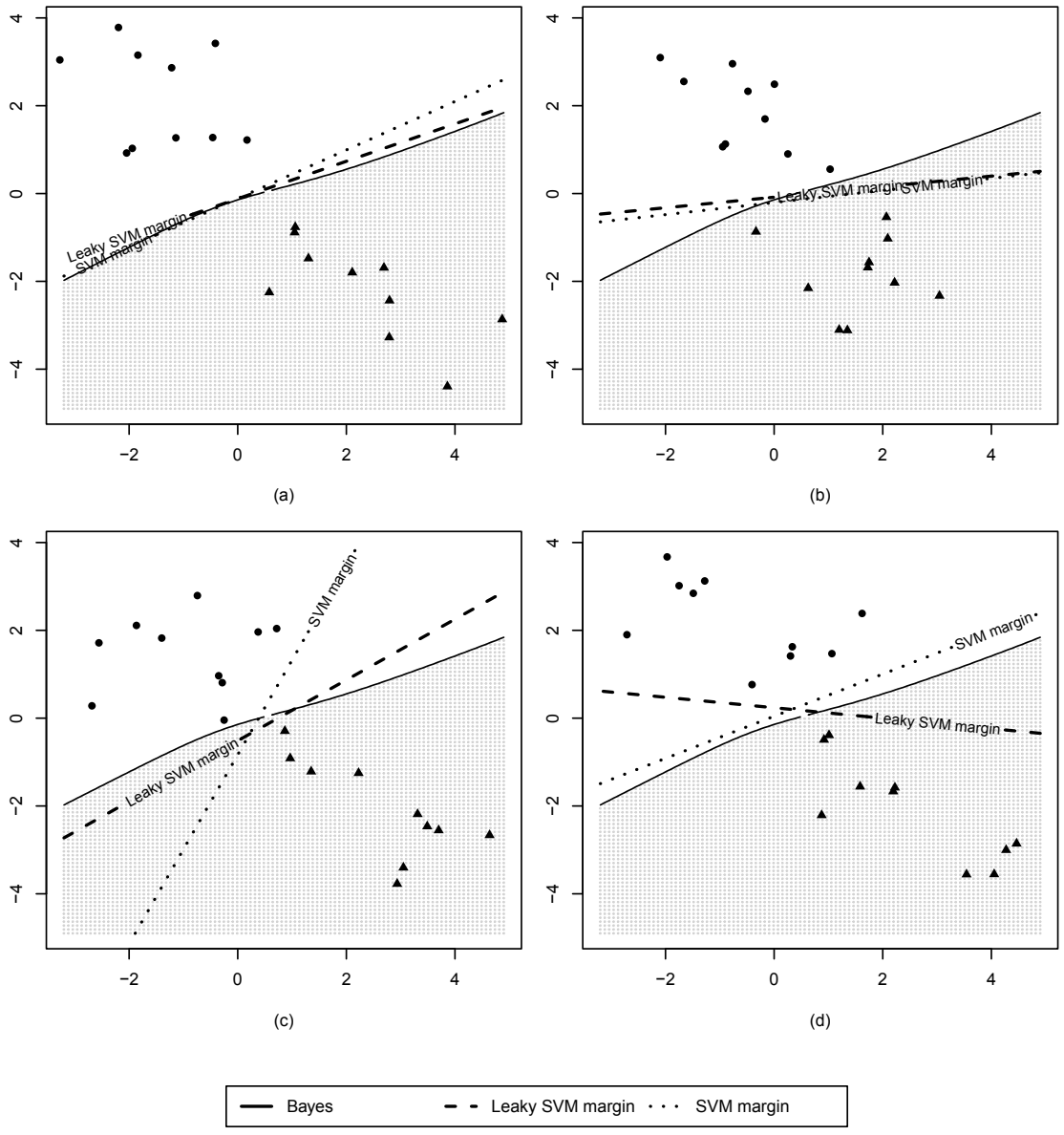


Figure 2: Decision boundaries of the optimal Bayes rule along with separating hyperplanes defined in (4) and (5) which are labelled as SVM margin and leaky SVM margin, respectively.

In light of (7), (6) reduces to,

$$\min_{\alpha_0,\boldsymbol{\alpha}} \mathcal{L}_{\mathbf{K}}(\alpha_0,\boldsymbol{\alpha}) = \min_{\alpha_0,\boldsymbol{\alpha}} \left[\frac{1}{n}\sum_{i=1}^{n} L\left(y_i(\alpha_0 + \mathbf{K}_i^T\boldsymbol{\alpha})\right) + \lambda \boldsymbol{\alpha}^T \mathbf{K} \boldsymbol{\alpha} \right], \tag{8}$$

where $\mathbf{K}$ is the kernel matrix that $[\mathbf{K}]_{ij} = K(\mathbf{x}_i, \mathbf{x}_j)$ and $\mathbf{K}_i$ is the ith column of $\mathbf{K}$.

Remark C (Existence of global solution and uniqueness) Let $(\mathbf{x}_i, y_i) \in \mathbf{R}^d \times \{-1, +1\}$ for $i = 1, \ldots, n$ be training data. Suppose there exist i, j with $y_i = +1$ and $y_j = -1$. Then a global solution to the kernel leaky SVM (8) exists. The proof is omitted since it can be

easily deduced from Proposition 2. $\hat{h}$ is uniquely determined because the formulation (6) is strictly convex in h, but $\hat{\alpha}_0$ may not. Again, this also occurs in the case of the kernel SVM.

4.2 Algorithm

In this subsection we derive an algorithm based on the Majorization-minimization (MM) principle (Hunter and Lange, 2004) to efficiently compute solutions of the kernel leaky SVM for a sequence of tuning parameter λ. To simplify the notation, let $\boldsymbol{\theta} = (\alpha_0, \boldsymbol{\alpha}^T)^T$. Let $\tilde{\boldsymbol{\theta}} = (\tilde{\alpha}_0, \tilde{\boldsymbol{\alpha}}^T)^T$ be the current value. First, we construct $\mathcal{Q}_{\mathbf{K}}(\boldsymbol{\theta}|\tilde{\boldsymbol{\theta}})$ which majorizes the objective function $\mathcal{L}_{\mathbf{K}}(\boldsymbol{\theta})$ by a quadratic function (Böhning and Lindsay, 1988).

Lemma 1. *The leaky hinge loss L has a quadratic upper bound,*

$$L(u) \le L(\tilde{u}) + L'(\tilde{u})(u - \tilde{u}) + \frac{1}{2}(u - \tilde{u})^2, \quad u, \tilde{u} \in \mathbf{R},$$

and the equality holds only when $u = \tilde{u}$.

Let $\tilde{\mathbf{z}}$ be an $n \times 1$ vector with ith element $y_i V'\{y_i(\tilde{\alpha}_0 - \mathbf{K}_i\tilde{\boldsymbol{\alpha}})\}/n$. By Lemma 1, we have

$$\begin{aligned}
\mathcal{L}_{\mathbf{K}}(\boldsymbol{\theta}) &= \frac{1}{n}\sum_{i=1}^{n} L\left(y_i(\alpha_0 + \mathbf{K}_i^T\boldsymbol{\alpha})\right) + \lambda\boldsymbol{\alpha}^T\mathbf{K}\boldsymbol{\alpha} \\
&\le \frac{1}{n}\sum_{i=1}^{n} L\left(y_i(\tilde{\alpha}_0 + \mathbf{K}_i\tilde{\boldsymbol{\alpha}})\right) + \lambda\tilde{\boldsymbol{\alpha}}^T\mathbf{K}\tilde{\boldsymbol{\alpha}} \\
&\quad + \tilde{\gamma}_{\mathbf{K}}^T\begin{pmatrix}\alpha_0 - \tilde{\alpha}_0 \\ \boldsymbol{\alpha} - \tilde{\boldsymbol{\alpha}}\end{pmatrix} + \frac{1}{2n}\begin{pmatrix}\alpha_0 - \tilde{\alpha}_0 \\ \boldsymbol{\alpha} - \tilde{\boldsymbol{\alpha}}\end{pmatrix}^T\mathbf{P}_{\mathbf{K},\lambda}\begin{pmatrix}\alpha_0 - \tilde{\alpha}_0 \\ \boldsymbol{\alpha} - \tilde{\boldsymbol{\alpha}}\end{pmatrix} \quad = \mathcal{Q}_{\mathbf{K}}(\boldsymbol{\theta}|\tilde{\boldsymbol{\theta}}),
\end{aligned}$$

where

$$\tilde{\gamma}_{\mathbf{K}} = \begin{pmatrix}\mathbf{1}^T\tilde{\mathbf{z}} \\ \mathbf{K}\tilde{\mathbf{z}} + 2\lambda\mathbf{K}\tilde{\boldsymbol{\alpha}}\end{pmatrix} \quad \text{and} \quad \mathbf{P}_{\mathbf{K},\lambda} = \begin{pmatrix}n & \mathbf{1}^T\mathbf{K} \\ \mathbf{K}\mathbf{1} & \mathbf{K}\mathbf{K} + 2n\lambda\mathbf{K}\end{pmatrix}.$$

The equality holds only if $\boldsymbol{\theta} = \tilde{\boldsymbol{\theta}}$. Second, we update $\boldsymbol{\theta}$ by the minimizer of,

$$\begin{pmatrix}\alpha_0 \\ \boldsymbol{\alpha}\end{pmatrix} = \arg\min_{\alpha_0,\boldsymbol{\alpha}} \mathcal{Q}_{\mathbf{K}}(\boldsymbol{\theta}|\boldsymbol{\theta}_m) = \begin{pmatrix}\tilde{\alpha}_0 \\ \tilde{\boldsymbol{\alpha}}\end{pmatrix} - n\mathbf{P}_{\mathbf{K},\lambda}^{-1}\tilde{\gamma}_{\mathbf{K}}. \tag{9}$$

In practice, λ is unknown and we would rely on cross validation; from a sequence of λ values such that $\lambda_1, \ldots, \lambda_M$, we choose the optimal value which minimizes the cross validation error. The kernel leaky SVM would be computed on a sequence of λ values and, of course, $\mathbf{P}_{\mathbf{K},\lambda}^{-1}$ has to be repeatedly evaluated for each λ. Unfortunately, inverting a matrix M times would be expensive, as the inversion of a $n \times n$ matrix costs $O(n^3)$ operations.

We further introduce a computational technique only need to invert a matrix once. Compute the eigen decomposition $\mathbf{K} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^T$ and inverted $\mathbf{P}_{\mathbf{K},\lambda}$ blockwise as follows.

$$\begin{aligned}
\mathbf{P}_{\mathbf{K},\lambda}^{-1} &= \begin{pmatrix}n & \mathbf{1}^T\mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^T \\ \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^T\mathbf{1} & \mathbf{U}\boldsymbol{\Pi}_{\mathbf{K},\lambda}\mathbf{U}^T\end{pmatrix}^{-1} \\
&= g_{\mathbf{K}}\begin{pmatrix}1 \\ -\mathbf{v}_{\mathbf{K}}\end{pmatrix}\begin{pmatrix}1 & -\mathbf{v}_{\mathbf{K}}^T\end{pmatrix} + \begin{pmatrix}0 & \mathbf{0}^T \\ \mathbf{0} & \mathbf{U}\boldsymbol{\Pi}_{\mathbf{K},\lambda}^{-1}\mathbf{U}^T\end{pmatrix},
\end{aligned} \tag{10}$$

where $\mathbf{\Pi}_{\mathbf{K},\lambda} = \mathbf{\Lambda}^2 + 2n\lambda\mathbf{\Lambda}$, $g_{\mathbf{K}} = 1/(n - \mathbf{1}^T\mathbf{U}\mathbf{\Lambda}\mathbf{\Pi}_{\mathbf{K},\lambda}^{-1}\mathbf{\Lambda}\mathbf{U}^T\mathbf{1})$, and $\mathbf{v}_{\mathbf{K}} = \mathbf{U}\mathbf{\Lambda}\mathbf{\Pi}_{\mathbf{K},\lambda}^{-1}\mathbf{U}^T\mathbf{1}$.

Replacing $\mathbf{P}_{\mathbf{K},\lambda}^{-1}$ with (10), we see that the right hand side of (9) becomes

$$\begin{pmatrix}\alpha_0\\ \boldsymbol{\alpha}\end{pmatrix} = \begin{pmatrix}\tilde{\alpha}_0\\ \tilde{\boldsymbol{\alpha}}\end{pmatrix} - n\left\{g_{\mathbf{K}}\left(\mathbf{1}^T\tilde{\mathbf{z}} - \mathbf{v}_{\mathbf{K}}^T\mathbf{K}(\tilde{\mathbf{z}} + 2\lambda\tilde{\boldsymbol{\alpha}})\right)\begin{pmatrix}1\\ -\mathbf{v}_{\mathbf{K}}\end{pmatrix} + \begin{pmatrix}0\\ \mathbf{U}\mathbf{\Pi}_{\mathbf{K},\lambda}^{-1}\mathbf{\Lambda}\mathbf{U}^T(\tilde{\mathbf{z}} + 2\lambda\tilde{\boldsymbol{\alpha}})\end{pmatrix}\right\}$$

and the operation cost is reduced to $O(n^2)$.

Remark D. We defer the algorithm of linear leaky SVM to Section B of Supplementary Material in the supplementary file as we take a similar procedure. The code for the linear and kernel leaky SVM is available from the authors upon request.

5 Bayes Risk Consistency

In this section we establish the Bayes risk consistency of the kernel leaky SVM which, in our views, provides the most important justification of this new loss function.

Let $\hat{f}_n$ be a classification function of the kernel leaky SVM with sample size n,

$$\hat{f}_n = \underset{f\in\mathcal{H}_K}{\operatorname{argmin}}\left[\frac{1}{n}\sum_{i=1}^{n} L\left(y_i f(\mathbf{x}_i)\right) + \lambda_n||f||^2_{\mathcal{H}_K}\right], \tag{11}$$

and f^* is the Bayes rule. Let the expected misclassification rate of a classification function $\hat{f}$ be denoted as $R(\hat{f}) = \mathrm{P}[\mathbf{Y} \neq \mathrm{sgn}\{\hat{f}(\mathbf{X})\}]$. We say the kernel leaky SVM is Bayes risk consistent if $R(\hat{f}_n) \to R(f^*)$ in probability.

When a loss function L is non-negative and classification-calibrated, Bartlett et al. (2006) showed that for any measureable function $\hat{f}$, $R(\hat{f}) - R(f^*)$ can be bounded in terms of $\mathrm{E}[L(\mathbf{Y}\hat{f}(\mathbf{X}))] - \mathrm{E}[L(\mathbf{Y}\bar{f}(\mathbf{X}))]$, where $\bar{f}(\mathbf{x})$ is a global minimizer of $\mathrm{E}[L(\mathbf{Y}f(\mathbf{X}))|\mathbf{X} = \mathbf{x}]$. It implies that if we obtain $\hat{f}_n$ such that $\mathrm{E}[L(\mathbf{Y}\hat{f}_n(\mathbf{X}))] - \mathrm{E}[L(\mathbf{Y}\bar{f}(\mathbf{X}))]$ is small, then the misclassification rate of $\hat{f}_n$ is close to that of the Bayes rule. It extends Zhang (2004)'s result under weaker conditions. Zhang (2004) gave a comparable result for a convex loss function satisfying certain conditions, and used the leave-one-out analysis to obtain estimation error resulted from using a finite sample size on kernel methods. It allows to establish the Bayes risk consistency of a class of kernel machines equipped with the hinge loss, logistic regression loss, and exponential loss.

This general approach is not applicable to the leaky hinge loss because it implicitly assumes $\mathrm{E}[L(\mathbf{Y}\bar{f}(\mathbf{X}))]$ is finite, which is ensured when the loss function is non-negative. The analysis for the leaky hinge loss is more involved. We introduce $g(\delta) = \mathrm{P}(p(\mathbf{X})\cdot(1-p(\mathbf{X})) \leq \frac{\delta}{2}(1-\frac{\delta}{2}))$. Intuitively, for a small δ, $g(\delta)$ can be understood as the probability of the random variable $\mathbf{X}$ having negligible amount of information about the optimal decision boundary. It turns out that $\mathrm{E}[L(\mathbf{Y}\bar{f}(\mathbf{X}))]$ is finite when $g(\delta)$ is bounded above by δ up to a constant as $\delta \to 0$.

Assumption 1. *There exists δ' such that g is continuous on $(0, \delta')$.*

Theorem 4. *Consider the underlying distribution satisfying Assumption 1. Let L be the leaky hinge loss and $\bar{f}(\mathbf{x})$ is defined in Theorem 1. $\mathrm{E}[L\left(\mathbf{Y}\bar{f}(\mathbf{X})\right)]$ is finite if and only if $\int_0^{\delta'} \frac{g(\delta)}{\delta} d\delta < \infty$.*

Remark E. It is also necessary to prove $\mathrm{E}[L(\mathbf{Y}\hat{f}_n(\mathbf{X}))]$ is finite before further theoretical discussion. Define $B = sup_{\mathbf{x},\mathbf{y}} K(\mathbf{x},\mathbf{y}) < \infty$ for any $\mathbf{x},\mathbf{y}$. By the representer theorem and the first-order optimality condition, we see that

$$\hat{f}_n(\mathbf{x}) = -\frac{1}{2n\lambda_n}\sum_{i=1}^{n} L'(y_i\hat{f}_n(\mathbf{x}_i))y_i K(\mathbf{x}_i,\mathbf{x}).$$

It indicates $|\hat{f}_n| \leq \frac{1}{2\lambda_n}B$, and thus $\mathrm{E}[L(\mathbf{Y}\hat{f}_n(\mathbf{X}))]$ is finite.

This result motivates us to devise a different upper bound by taking account of the amount of available information in $\mathbf{x}$. We adopt the same bound when the information in $\mathbf{x}$ is relatively large, and attempt a different bound otherwise.

Lemma 2. *Let f^* be the Bayes rule and $\hat{f}_n$ be the equation (11). Then for any $0 < \delta \leq 1$,*

$$\begin{aligned} R(\hat{f}_n) - R(f^*) \leq\ & \mathrm{P}[\mathrm{sgn}\{f^*(\mathbf{X})\} \neq \mathrm{sgn}\{\hat{f}_n(\mathbf{X})\} \text{ and } p(\mathbf{X})(1-p(\mathbf{X})) \leq 2/\delta \cdot (1-\delta/2)] \\ & + \mathrm{E}_{\{\mathbf{X}:p(\mathbf{X})(1-p(\mathbf{X}))>2/\delta\cdot(1-\delta/2)\}}\left[L(\mathbf{Y}\hat{f}_n(\mathbf{X})) - L(\mathbf{Y}\bar{f}(\mathbf{X}))\right], \end{aligned}$$

where L is the leaky hinge loss function.

If we can find a sequence of δ_n such that the bound in the above lemma converges to 0, the Bayes risk consistency can be shown. We consider the kernel K that is *universal* (Steinwart, 2001) so that the corresponding RKHS can be rich enough. Let $C(\mathcal{X})$ be the space of continuous bounded functions on compact domain $\mathcal{X}$. A continuous kernel K on a $\mathcal{X}$ is defined as universal if $\mathcal{H}_K$ is dense in $C(\mathcal{X})$, i.e., for every function $g \in C(\mathcal{X})$ and every $\varepsilon > 0$, there exists a function $f \in \mathcal{H}_K$ with $\|f - g\|_\infty \leq \varepsilon$. For example, the gaussian kernel is universal.

Theorem 5. *(Bayes risk consistency) Assume that Assumption 1 holds and that $-log(\delta)g(\delta) \to 0$ as $\delta \to 0$. Suppose that the input space $\mathcal{X}$ is compact and $\mathcal{H}_K$ is the RKHS induced by a universal kernel K on $\mathcal{X}$. If $0 < \inf_{\mathbf{x},\mathbf{y}\in\mathcal{X}} K(\mathbf{x},\mathbf{y}) < \sup_{\mathbf{x},\mathbf{y}\in\mathcal{X}} K(\mathbf{x},\mathbf{y}) < \infty$, and as $n \to \infty$, $\lambda_n \to 0$ and $\lambda_{n+1}^{-1} - \lambda_n^{-1} \to 0$, then $R(\hat{f}_n) - R(f^*) \to 0$ in probability.*

6 Numeric Examples

This section compares the leaky SVM and the standard SVM in terms of classification accuracy. Such a comparison will directly show the impact of using the leaky hinge loss. Also, the SVM is one of the best classification algorithms in an extensive numerical study conducted by Fernández-Delgado et al. (2014). Therefore, as long as the leaky SVM is better than or similar to the kernel SVM in terms of classification accuracy, we can claim that the leaky SVM is a worthy new classifier.

6.1 Simulation

We first compare the leaky SVM to the SVM on simulated data. A mixture gaussian model is used for the simulation. Let $\boldsymbol{\mu}^+ = (1,\ldots,1,-1,\ldots,-1)^T \in \mathbb{R}^p$ and $\boldsymbol{\mu}^- =$

$(-1, \ldots, -1, 1, \ldots, 1)^T \in \mathbb{R}^p$ where the both have the half of components as 1s and the other half as -1s. We draw μ_k^+ and μ_k^-, $k = 1, \ldots, K$ from

$$\mu_k^+ \sim N\left(\boldsymbol{\mu}^+, \mathbf{I}_{p\times p}\right) \text{ if } k \leq \frac{2}{3}K, \ \mu_k^+ \sim N\left(\boldsymbol{\mu}^-, \mathbf{I}_{p\times p}\right) \text{ if } k > \frac{2}{3}K,$$
$$\text{and } \ \mu_k^- \sim N\left(\boldsymbol{\mu}^-, \mathbf{I}_{p\times p}\right) \text{ if } k \leq \frac{2}{3}K, \ \mu_k^- \sim N\left(\boldsymbol{\mu}^+, \mathbf{I}_{p\times p}\right) \text{ if } k > \frac{2}{3}K.$$

Given μ_k^+ and μ_k^-, let $(\mathbf{X}, \mathbf{Y})$ be a random pair such that $\mathrm{P}(\mathbf{Y} = -1) = \mathrm{P}(\mathbf{Y} = +1) = 0.5$ and $\mathbf{X} \in \mathbb{R}^p$ with

$$\mathbf{X}|(\mathbf{Y} = -1) \sim \sum_{k=1}^{K} \frac{1}{K} N(\mu_k^-, \ \sigma^2 \mathbf{I}_{p\times p}), \text{ and } \mathbf{X}|(\mathbf{Y} = +1) \sim \sum_{k=1}^{K} \frac{1}{K} N(\mu_k^+, \ \sigma^2 \mathbf{I}_{p\times p}),$$

so that the model can have a highly nonlinear optimal decision boundary.

We consider $K = 3$, $p = 2$, $\sigma = 1/\sqrt{10}$ with the Bayes error 11.13% in Example 1; $K = 10$, $p = 2$, $\sigma = 1/\sqrt{50}$ with the Bayes error 11.51% in Example 2; and $K = 3$, $p = 10$, $\sigma = 1/\sqrt{10}$ with the Bayes error 13.48% in Example 3. We vary sample size $n = 50, 90, 200, 900$. We consider both linear classifiers and Gaussian kernel classifiers and we select the best λ among 100 λ-values by five-fold cross-validation. We compute the SVM classifier by the R package `kernlab`. The simulations are repeated for 100 times under the above setting. We summarize the average of misclassification rates with the corresponding standard error in Table 1.

We have several observations from Table 1. First, the SVM and the leaky SVM are comparable to each other in general. The Gaussian leaky SVM slightly outperforms the Gaussian SVM in Example 1, and the linear leaky SVM consistently outperforms the linear SVM in Example 2 and Example 3. Second, the misclassification rates of both Gaussian leaky SVM and Guassin SVM get closer to the Bayes error rate as the sample size increases, although the convergence is relatively slower for more complicated models.

6.2 Real Data Example

We examine the performance of the leaky SVM compared to the SVM on 15 datasets from University of California at Irvines Machine Learning Repository (Dua and Graff, 2017). These datasets have various combinations of sample size and dimension. We randomly sample 2/3 observations as the training set to fit and tune each model with five-fold cross-validation for selecting an optimal λ from 100 λ-values. The remaining 1/3 observations is set as the test set for calculating the misclassification rate. We repeat this process 100 times and report the average misclassifcation rates with the corresponding standard errors in Table 2.

When comparing the linear leaky SVM and the linear SVM, we observe that the linear leaky SVM outperforms the linear SVM on 10 datasets. When comparing the kernel classifers, the Gaussian leaky SVM outperforms the kernel SVM on 6 datasets.

Table 1: Misclassification rates, averaged by 100 runs, under mixture gaussian distributed data. The standard error is given in parentheses.

n	Kernel	Misclassification rate (%)	
		Leaky SVM	SVM
Example 1: $K = 3$, $p = 2$, Bayes error: 11.13%			
50	Linear	34.63 (.05)	34.35 (.05)
	Gaussian	17.11 (.04)	17.29 (.04)
90	Linear	33.16 (.05)	32.80 (.05)
	Gaussian	14.40 (.04)	15.13 (.04)
200	Linear	31.62 (.05)	31.43 (.05)
	Gaussian	12.80 (.03)	13.21 (.03)
900	Linear	30.65 (.05)	30.72 (.05)
	Gaussian	11.47 (.03)	11.50 (.03)
Example 2: $K = 10$, $p = 2$, Bayes error: 11.51%			
50	Linear	40.37 (.05)	40.78 (.05)
	Gaussian	23.84 (.04)	24.13 (.04)
90	Linear	39.72 (.05)	40.46 (.05)
	Gaussian	18.79 (.04)	19.62 (.04)
200	Linear	38.00 (.05)	38.58 (.05)
	Gaussian	15.45 (.04)	16.38 (.04)
900	Linear	37.12 (.05)	38.11 (.05)
	Gaussian	12.66 (.03)	13.00 (.03)
Example 3: $K = 3$, $p = 30$, Bayes error: 13.48%			
50	Linear	32.87 (.05)	32.95 (.05)
	Gaussian	31.36 (.05)	30.97 (.05)
90	Linear	28.80 (.05)	29.17 (.05)
	Gaussian	26.39 (.04)	26.70 (.04)
200	Linear	25.26 (.04)	25.68 (.04)
	Gaussian	22.61 (.04)	22.79 (.04)
900	Linear	22.42 (.04)	22.50 (.04)
	Gaussian	18.07 (.04)	18.02 (.04)

7 Summary

In this paper we have introduced the first negatively divergent loss function named the leaky hinge loss for margin-based classification. Despite some technical difficulties brought by the negatively divergence of the loss function, we have proved the classification-calibration property of the leaky hinge loss and established the Bayes risk consistency of the leaky kernel SVM. We have further provided numeric evidence to show that the linear and kernel leaky SVM is at least as competitive as the usual linear and kernel SVM. All of these provide a

Table 2: Misclassification rates, averaged by 100 runs, on 15 datasets from University of California at Irvines. The standard error is given in parentheses.

Dataset	n	p	Kernel	Misclassification rate (%)	
				Leaky SVM	SVM
Arrhythmia	68	233	Linear	21.39 (.84)	21.39 (.85)
			Gaussian	20.13 (.82)	21.78 (.85)
Australian	690	14	Linear	13.59 (.23)	13.62 (.23)
			Gaussian	13.89 (.23)	13.61 (.23)
Banknote	1372	4	Linear	1.07 (.05)	1.08 (.05)
			Gaussian	0.23 (.02)	0.01 (.00)
Biodeg	1055	41	Linear	13.56 (.18)	13.23 (.18)
			Gaussian	12.22 (.17)	12.28 (.17)
Bupa	345	6	Linear	31.59 (.77)	31.79 (.77)
			Gaussian	31.16 (.77)	31.34 (.77)
Chess	3196	36	Linear	2.82 (.05)	3.25 (.05)
			Gaussian	1.67 (.04)	0.93 (.03)
cle:Heart	297	13	Linear	16.15 (.37)	15.85 (.37)
			Gaussian	16.47 (.37)	16.31 (.37)
Hepatitis	80	19	Linear	14.44 (.67)	16.19 (.70)
			Gaussian	13.59 (.65)	13.89 (.66)
Hungarian	261	10	Linear	17.69 (.41)	18.52 (.41)
			Gaussian	18.69 (.42)	17.90 (.41)
LSVT	126	310	Linear	14.31 (.53)	16.69 (.57)
			Gaussian	15.10 (.55)	16.07 (.56)
Musk	475	166	Linear	17.08 (.30)	16.85 (.30)
			Gaussian	10.27 (.24)	8.29 (.22)
Parkinsons	195	22	Linear	15.25 (.40)	14.14 (.43)
			Gaussian	10.54 (.38)	8.97 (.35)
Sonar	208	60	Linear	23.39 (.51)	25.10 (.52)
			Gaussian	17.91 (.46)	15.65 (.43)
Spectf	80	22	Linear	30.07 (.87)	31.19 (.88)
			Gaussian	26.74 (.84)	28.04 (.85)
Vertebral	310	6	Linear	14.88 (.35)	15.18 (.35)
			Gaussian	16.81 (.37)	15.83 (.36)

full justification for using such a loss function for margin-based classification. A by-product of our theory offers a complementary result to Rosset et al. (2003).

The leaky hinge loss is not the only negatively divergent loss function for margin-based classification, because we have found a second loss function with similar properties. We opt to focus on the leaky hinge loss because it is directly linked to the hinge loss and it is the first

one that we found. We conjecture that there should be infinite many negatively-divergent loss functions.

The main purpose of this work is to carefully examine the roles of the loss function in classification. Although the leaky SVM is as competitive as the SVM, this is only used to justify the validity of our approach. We hope this paper will stimulate more interests in the study of loss functions in machine learning.

References

Bartlett, P. L., Jordan, M. I., and McAuliffe, J. D. (2006), "Convexity, classification, and risk bounds," *Journal of the American Statistical Association*, 101, 138–156.

Böhning, D. and Lindsay, B. G. (1988), "Monotonicity of quadratic-approximation algorithms," *Annals of the Institute of Statistical Mathematics*, 40, 641–663.

Breiman, L. (2001), "Random forests," *Machine learning*, 45, 5–32.

Dua, D. and Graff, C. (2017), "UCI Machine Learning Repository," .

Fernández-Delgado, M., Cernadas, E., Barro, S., and Amorim, D. (2014), "Do we need hundreds of classifiers to solve real world classification problems?" *The Journal of Machine Learning Research*, 15, 3133–3181.

Freund, Y., Schapire, R. E., et al. (1996), "Experiments with a new boosting algorithm," in *icml*, Citeseer, vol. 96, pp. 148–156.

Friedman, J., Hastie, T., and Tibshirani, R. (2000), "Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors)," *The annals of statistics*, 28, 337–407.

Goodfellow, I., Bengio, Y., and Courville, A. (2016), *Deep learning*, MIT press.

Hastie, T., Tibshirani, R., and Friedman, J. (2009), *The elements of statistical learning: data mining, inference and prediction*, Springer, 2nd ed.

Hunter, D. R. and Lange, K. (2004), "A tutorial on MM algorithms," *The American Statistician*, 58, 30–37.

Lin, Y. (2004), "A note on margin-based loss functions in classification," *Statistics & probability letters*, 68, 73–82.

Rosset, S., Zhu, J., and Hastie, T. (2003), "Margin Maximizing Loss Functions." in *NIPS*, pp. 1237–1244.

— (2004), "Boosting as a regularized path to a maximum margin classifier," *The Journal of Machine Learning Research*, 5, 941–973.

Shen, X., Tseng, G. C., Zhang, X., and Wong, W. H. (2003), "On ψ-learning," *Journal of the American Statistical Association*, 98, 724–734.

Steinwart, I. (2001), "On the influence of the kernel on the consistency of support vector machines," *Journal of machine learning research*, 2, 67–93.

— (2005), "Consistency of support vector machines and other regularized kernel classifiers," *IEEE transactions on information theory*, 51, 128–142.

Vapnik, V. (2013), *The nature of statistical learning theory*, Springer science & business media.

Wahba, G. (1990), *Spline models for observational data*, vol. 59, Siam.

Zhang, T. (2004), "Statistical behavior and consistency of classification methods based on convex risk minimization," *The Annals of Statistics*, 32, 56–85.

Zhu, J., Rosset, S., Tibshirani, R., and Hastie, T. J. (2003), "1-norm support vector machines," in *Advances in neural information processing systems*, Citeseer, p. None.