Journal of Machine Learning Research 24 (2023) 1-40 Submitted 9/22; Revised 6/23; Published 7/23

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

Oh-Ran Kwon kwon0085@umn.edu
Hui Zou zouxx019@umn.edu
School of Statistics
University of Minnesota
Minneapolis, MN 55455, USA

Editor: Aarti Singh

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, the non-negativity of the loss function in ERM is not necessary to be classification-calibrated and to produce a Bayes consistent classifier. We introduce the leaky hockey stick loss (LHS loss), the first negatively divergent margin-based loss function. We prove that the LHS loss is classification-calibrated. When the hinge loss is replaced with the LHS loss in the ERM approach for deriving the kernel support vector machine (SVM), the corresponding optimization problem has a well-defined solution named the kernel leaky hockey stick classifier (LHS classifier). Under mild regularity conditions, we prove that the kernel LHS classifier is Bayes risk consistent. In our theoretical analysis, we overcome multiple challenges caused by the negative divergence of the LHS 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 LHS classifier and compare it to the kernel SVM on simulated data and fifteen benchmark data sets. To conclude this work, we further present a class of negatively divergent margin-based loss functions that have similar theoretical properties to those of the LHS loss. Interestingly, the LHS loss can be viewed as a limiting case of this family of negatively divergent margin-based loss functions.

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 are generated from some underlying distribution $\text{pr}(\mathbf{X}, \mathbf{Y})$, and let $p(\mathbf{x}) = \text{pr}(\mathbf{Y} = 1 \mid \mathbf{X} = \mathbf{x})$. Under the standard 0-1 loss, the optimal classification rule is $\text{sgn}\,[p(\mathbf{x}) - 1/2]$ (a.k.a. Bayes rule). Throughout the paper, we assume that $p(\mathbf{X}) \neq 1/2$ 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.

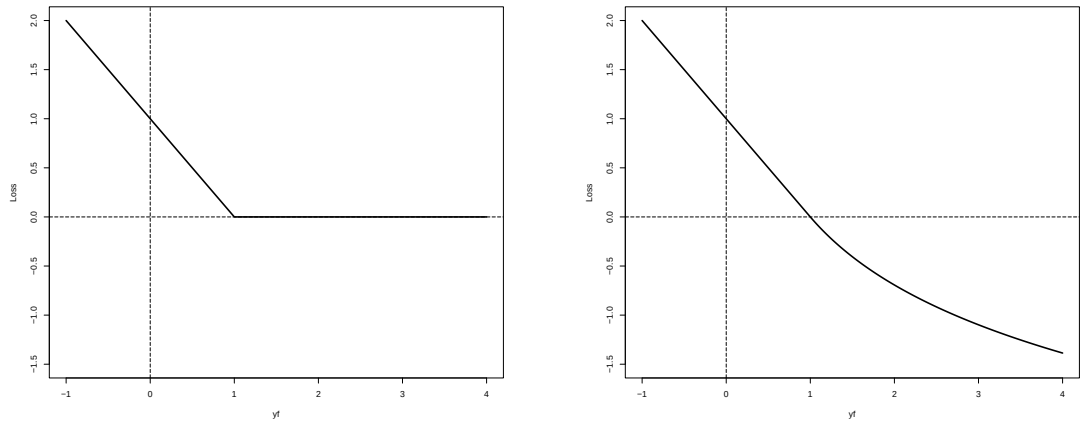
Extensive research has been devoted to classification, and many classification algorithms have been developed and widely used in practice, ranging from the classical methods such as discriminant analysis and logistic regression to modern techniques such as support vector machines (SVM) (Vapnik, 2013), boosting (Freund et al., 1996), random forests (Breiman, 2001), 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 $\sum_{i=1}^{n} L(y_i f(\mathbf{x}_i))/n$, where L is referred to as a margin-based loss function in the literature and $yf(\mathbf{x})$ is called the margin per training data point. A classification learning 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, boosting can be viewed as minimizing the empirical risk with an ℓ_1 penalty (Rosset et al., 2004). 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. In contrast, the loss function in 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. In contrast, the loss function in ERM can be far more flexible. For example, SVM uses the hinge loss, logistic regression uses the logit loss (or the binomial deviance loss), and AdBoost uses the exponential loss (Hastie et al., 2009; Friedman et al., 2000). All these loss functions are non-negative, which makes them qualified as loss functions in decision theory. An interesting question, which has not been asked before in the literature, is: Can we use a function having negative values in ERM for classification? In the ERM framework, a loss function 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 hockey stick loss (LHS loss). The expression of the LHS loss is given in (1) and the picture of this loss function is displayed in Figure 1. We use the word 'hockey stick' because our loss function resembles it and is continuously differentiable. We use the term 'leaky' because the right side diverges to negative infinity motivated by the name of leaky ReLU.

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

Given the training data, the empirical leaky hockey stick risk is $\sum_{i=1}^{n} L(y_i f(\mathbf{x}_i))/n$. If we use the notion from decision theory, the LHS 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 one, the LHS loss becomes negative, meaning that it actually gives a reward. The larger the margin, the bigger the reward. Intuitively, this sounds reasonable. Of course, we need to provide formal theoretical and numerical evidence to justify the use of the LHS loss in ERM, which is the main focus of this paper.



(a) Hinge loss

(b) Leaky hockey stick loss

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

There have been many 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 $\mathrm{E}[L(\mathbf{Y}f(\mathbf{X})) \mid \mathbf{X} = \mathbf{x}]$ 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. However, 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 $\mathrm{E}[L(\mathbf{Y}f(\mathbf{X})) \mid \mathbf{X} = \mathbf{x}]$ 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 LHS loss, our analysis reveals that the global minimizer of $\mathrm{E}[L(\mathbf{Y}f(\mathbf{X})) \mid \mathbf{X} = \mathbf{x}]$ 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 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 LHS loss is used in the ERM approach to derive a classifier, the resulting classifier is named the kernel leaky hockey stick classifier (LHS classifier). We establish the Bayes consistency of the kernel LHS classifier, which in turn offers the most important justification for the LHS loss. Therefore, we can claim that the LHS 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 to non-negative loss functions. Their analyses are not applicable in the case of LHS loss. We use new techniques to prove the Bayes consistency of the kernel LHS classifier.

For our theoretical analysis, we overcome multiple challenges caused by the negative divergence of the LHS loss. For non-negative loss functions, the existence of the global minimizer and minimum objective value of $E[L(\mathbf{Y}f(\mathbf{X}))]$ can be checked easily and is crucial to study Bayes risk consistency. They are used to approximate the misclassification error of a derived classifier. However, we find out that they may not exist for the LHS loss, which is technically non-trivial to address. Still, we provide the necessary and sufficient conditions for the existence through a quantity of $\mathbf{X}$ containing negligible information about the optimal decision boundary. This result tells that the general approach for studying Bayes risk consistency does not apply to the LHS loss. Instead, we use different quantities and techniques to approximate the classification error and get the Bayes risk consistency result for the kernel LHS classifier. Additionally, it is not obvious to prove the existence of the global solutions to the linear LHS classifier and the kernel LHS classifier different from using a non-negative loss function. We find a specific compact set and show that a global minimizer on the set is also an actual global minimizer.

The geometry of SVM 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 shared by a class of non-negative loss functions that vanishes to zero quickly enough, such as the hinge loss, the exponential loss, and the logit loss. Their result provides the unified margin maximization view of many popular classification algorithms. However, their theory does not cover the LHS loss because the LHS loss violates their conditions. Nevertheless, we show that the linear LHS classifier also has an interesting and new margin maximization view. This result suggests that the linear LHS classifier and the linear SVM are different, although their kernel versions approach the same limit (Bayes rule).

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

We have implemented the algorithms for the linear and kernel LHS classifiers in the `lhsc` package in `R`, which is available at `http://github.com/ohrankwon/lhsc`.

The remainder of the paper is organized as follows. In Section 2, we prove that the LHS loss function is classification-calibrated. In Section 3, we prove that the linear LHS classifier is well-defined given training data. We provide the dual problem of the linear LHS classifier. When the data is linearly separable, we show the margin-maximization picture of the linear LHS classifier and compare it with the margin-maximization picture of the linear SVM. In Section 4, we consider the kernel LHS classifier in a reproducing kernel Hilbert space (RKHS). We first prove that the kernel LHS classifier is well-defined on training data. We then derive an efficient algorithm to solve the kernel LHS classifier. In Section 5, we establish the Bayes risk consistency of the kernel LHS classifier. Section 6 contains the numerical results. Concluding remarks are given in Section 7 where we further provide a class of negatively divergent loss functions that are classification-calibrated and have Bayes consistency. Technical details and proofs are provided in Appendices.

2. 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 $\mathrm{E}[L(\mathbf{Y}f(\mathbf{X})) \mid \mathbf{X} = \mathbf{x}]$ 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. The 'if' statement follows from the fact that any $\bar{f}$ is also a population minimizer and the definition of Fisher consistency. The 'only if' statement comes from the fact that $\mathrm{E}[L(\mathbf{Y}\bar{f}(\mathbf{X})) \mid \mathbf{X}] = \mathrm{E}[L(\mathbf{Y}\check{f}(\mathbf{X})) \mid \mathbf{X}]$ almost surely and that for any $\delta_{\mathbf{X}}$ such that $\mathrm{sgn}(\delta_{\mathbf{X}}) \neq \mathrm{sgn}(p(\mathbf{X}) - 1/2)$, $\mathrm{E}[L(\mathbf{Y}\bar{f}(\mathbf{X})) \mid \mathbf{X}] < \mathrm{E}[L(\mathbf{Y}\delta_{\mathbf{X}}) \mid \mathbf{X}]$ almost surely by the definition of classification-calibration.

For a non-negative loss function L, $\mathrm{E}[L(\mathbf{Y}\bar{f}(\mathbf{X}))]$ is always finite. However, the LHS loss does not guarantee the finiteness of $\mathrm{E}[L(\mathbf{Y}\bar{f}(\mathbf{X}))]$ (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 LHS loss is classification-calibrated. The LHS loss is the first negatively divergent loss function proven to satisfy the classification-calibration property.

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

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

Also, $\mathrm{sgn}\{\bar{f}(\mathbf{x})\} = f^*(\mathbf{x})$, where $f^*(\mathbf{x}) = \mathrm{sgn}\{p(\mathbf{x}) - 1/2\}$ is the Bayes rule.

Remark 1 *In general, a global minimizer of $\mathrm{E}[L(\mathbf{Y}f(\mathbf{X})) \mid \mathbf{X} = \mathbf{x}]$ 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 $\mathrm{E}[L(\mathbf{Y}\alpha) \mid \mathbf{X} = \mathbf{x}] = L(-\alpha) \to -\infty$ as $\alpha \to -\infty$. If $p(\mathbf{x}) = 1$, then $\mathrm{E}[L(\mathbf{Y}f(\mathbf{X})) \mid \mathbf{X} = \mathbf{x}] = L(\alpha) \to -\infty$ as $\alpha \to \infty$.*

3. Linear Leaky Hockey Stick Classifier

3.1 Existence of a global solution

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

When L is a non-negative continuous loss function, there exists a global minimizer to the regularized ERM problem. Proving existence is considered well-understood. 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 LHS classifier. The loss term in the regularized ERM can diverge to negative infinity even though the LHS 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 LHS classifier 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 value on D^c.

Theorem 2 *(Existence of global solution) Let $(\mathbf{x}_i, y_i) \in \mathbb{R}^p \times \{-1, +1\}$ 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*

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

where L is the LHS loss and $\lambda > 0$ is a tuning parameter.

Remark 2 *(Uniqueness) The minimizer $\hat{\boldsymbol{\beta}}$ is uniquely determined because the LHS loss is convex and the ℓ_2 regularizer is strictly convex. However, $\hat{\beta}_0$ may not. Here is an illustrative example. Let the data be*

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

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 Constrained optimization formulation and its dual

We can reformulate the optimization problem (2) as a constrained optimization problem. The objective is to maximize a new criterion of margins under the constraints that all data points lie on the correct side of the hyperplane after being perturbed by slack variables η_is. The next lemma illustrates this result.

Lemma 1 *(Constrained optimization formulation) The linear LHS classifier classifier* (2) *can be equally written as* $\operatorname{sgn}(\hat{w}_0 + \mathbf{x}^T \hat{\mathbf{w}})$*, where $(\hat{w}_0, \hat{\mathbf{w}}^T)^T$ is the solution to*

$$\begin{aligned} \min_{w_0, \mathbf{w}} \quad & -\sum_{i=1}^{n} \log d_i + C \sum_{i=1}^{n} \eta_i \\ \textit{subject to} \quad & d_i = y_i(w_0 + x_i^T \mathbf{w}) + \eta_i \geq 0 \ \forall i, \\ & \eta_i \geq 0 \ \forall i, \ \textit{and} \ \mathbf{w}^T \mathbf{w} \leq 1, \end{aligned} \tag{3}$$

for some tuning parameter $C > 0$.

The dual form is presented in the next lemma.

Lemma 2 *(The dual problem) Let* $\mathbf{X}$ *be a* $n \times p$ *matrix with ith row* $\mathbf{x}_i^T$*,* $\mathbf{y}$ *be a* $n \times 1$ *vector with ith element* y_i*, and* $\mathbf{Y}$ *be a diagonal matrix with ith diagonal element* y_i*. The constrained optimization problem (3) has a dual which is*

$$\max_{\boldsymbol{\sigma} \in \mathbb{R}^n} \ \mathbf{1}^T \log \boldsymbol{\sigma} - \left\| \mathbf{X}^T \mathbf{Y} \boldsymbol{\sigma} \right\| \quad \textit{subject to} \ \ 0 < \boldsymbol{\sigma} \leq C\mathbf{1} \ \textit{and} \ \mathbf{y}^T \boldsymbol{\sigma} = 0, \tag{4}$$

where $\mathbf{1}$ *is a* $n \times 1$ *vector with all components being 1 and* $\log \boldsymbol{\sigma}$ *is a vector whose components are the* log *of those of* $\boldsymbol{\sigma}$*.*

The optimality conditions to be a primal and dual optimal solution pair are presented in (17) in Appendix A. It is straightforward to show the strong duality holds by checking Slater's condition. An optimal solution to (4) exists as long as there exist i and j with $y_i = +1$ and $y_j = -1$ (that is, there are both negative and positive class sample in the data), which is assumed in Appendix A.5. In Appendix B, we briefly give a geometric interpretation of the dual problem (4).

From the optimality conditions, if $\mathbf{X}^T \mathbf{Y} \boldsymbol{\sigma} \neq 0$, we can easily recover w_0 and $\mathbf{w}$ from the solution to the dual by setting $\mathbf{w} = \mathbf{X}^T \mathbf{Y} \boldsymbol{\sigma} / \left\| \mathbf{X}^T \mathbf{Y} \boldsymbol{\sigma} \right\|$ and $w_0 = y_i / \sigma_i - x_i^T w$ for some i with $0 < \sigma_i < C$. Unlike the SVM whose dual is a simple quadratic programming problem, the dual of the leaky hockey stick classifier given in (4) is much harder to solve than solving quadratic programming. As an alternative, we directly solve (2) in Theorem 2. See the details in section 4.2.

3.3 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 margin. Rosset et al. (2003) discussed a family of loss functions that share the same margin picture as that of SVM. Specifically, they investigated for which loss functions the solution of the regularized ERM,

$$\hat{\boldsymbol{\beta}}_\lambda = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^n L(y_i \mathbf{x}_i^T \boldsymbol{\beta}) + \lambda \|\boldsymbol{\beta}\|_q^q, \tag{5}$$

where $q \geq 1$, converges to the 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 $\hat{\boldsymbol{\beta}}_\lambda / \|\hat{\boldsymbol{\beta}}_\lambda\|_q$ is

$$\arg \max_{\|\mathbf{w}\|_q = 1} \min y_i \mathbf{w}^T \mathbf{x}_i. \tag{6}$$

The family of loss functions covers the hinge loss, the exponential loss, and the logit 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 LHS loss violates their sufficient condition, and we find that the LHS loss optimizes a different margin at the limit. Still, the convergent point finds the separating hyperplane that perfectly separates 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 $\min_i y_i \mathbf{x}_i^T \bar{\mathbf{w}} > 0$ with $\|\bar{\mathbf{w}}\|_q = 1$, $q \geq 1$. Let $\hat{\boldsymbol{\beta}}_\lambda$ be the solutions to (5) with the LHS loss and the ℓ_q regularizer. Then, as $\lambda \to 0$, any convergence point of $\hat{\boldsymbol{\beta}}_\lambda / \|\hat{\boldsymbol{\beta}}_\lambda\|_q$ 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

$$\hat{\boldsymbol{\beta}}_\lambda / \|\hat{\boldsymbol{\beta}}_\lambda\|_q \to \arg\max_{\|\mathbf{w}\|_q = 1} \prod_{i=1}^n y_i \mathbf{x}_i^T \mathbf{w} \mathbb{1}\{y_i \mathbf{x}_i^T \mathbf{w} \geq 0\}. \tag{7}$$

We visualize the two separating hyperplanes defined in (6) and (7) 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 $\mathbf{X} \sim \sum_{i=1}^n N(\mu_i^-, 0.6 \cdot \mathbf{I})/3$ if $y = -1$ and $\mathbf{X} \sim \sum_{i=1}^3 N(\mu_i^+, 0.6 \cdot \mathbf{I})/3$ 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 generate ten samples from each distribution and see the data are separable. Since the generating distribution is known, the optimal decision boundary (solid line in Figure 2) can be calculated exactly. Figure 2 (a)-(d) show the decision boundary of the new margin maximizer defined in (7) (long-dashed line), along with that of the standard margin maximizer defined in (6) (dashed line). The two boundaries are similar to each other in (a) and (b), while they are noticeably different in (c) and (d).

4. Kernel Leaky Hockey Stick Classifier

4.1 Formulation

In practice, the linear LHS classifier could be restrictive when the Bayes rule is highly nonlinear. To obtain a nonlinear classifier boundary with the LHS 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 the kernel LHS classifier as the classifier $\text{sgn}\{\hat{\alpha}_0 + \hat{h}(\mathbf{x})\}$, where $(\hat{\alpha}_0, \hat{h})$ is the solution to

$$\min_{\substack{\alpha_0 \in \mathbb{R} \\ h \in \mathcal{H}_K}} \left(\sum_{i=1}^n L\left(y_i(\alpha_0 + h(\mathbf{x}_i))\right)/n + \lambda \|h\|_{\mathcal{H}_K}^2 \right). \tag{8}$$

While (8) 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,

$$\hat{h}(\mathbf{x}) = \sum_{i=1}^n \hat{\alpha}_i K(\mathbf{x}, \mathbf{x}_i), \text{ and thus } ||\hat{h}||_{\mathcal{H}_K}^2 = \sum_{i=1}^n \sum_{j=1}^n K(\mathbf{x}_i, \mathbf{x}_j) \hat{\alpha}_i \hat{\alpha}_j. \tag{9}$$

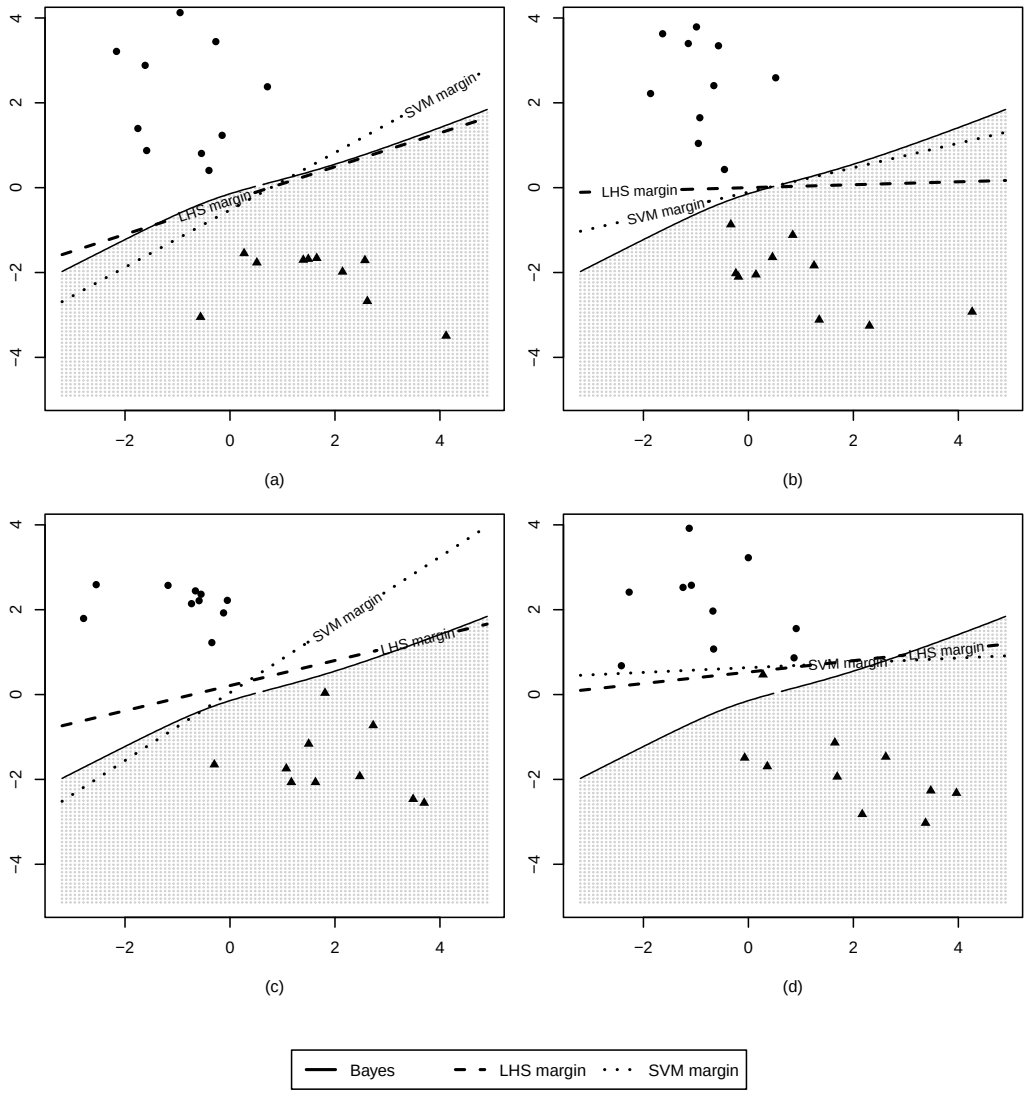


Figure 2: Decision boundaries of the optimal Bayes rule along with separating hyperplanes defined in (6) and (7) which are labelled as SVM margin and LHS classifier margin, respectively.

We note that the representer theorem holds irrespective of whether a loss function is non-negative or negatively divergent. In light of (9), (8) reduces to

$$\min_{\alpha_0,\boldsymbol{\alpha}} \mathcal{L}_{\mathbf{K}}(\alpha_0,\boldsymbol{\alpha}) = \min_{\alpha_0,\boldsymbol{\alpha}} \left[\sum_{i=1}^{n} L\left(y_i(\alpha_0 + \mathbf{K}_i^T\boldsymbol{\alpha})\right)/n + \lambda \boldsymbol{\alpha}^T \mathbf{K} \boldsymbol{\alpha} \right], \tag{10}$$

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 3 *(Existence of global solution and uniqueness) Let $(\mathbf{x}_i, y_i) \in \mathbb{R}^p \times \{-1,+1\}$ 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 LHS classifier (10) *exists. The proof is omitted since it can be easily deduced from Theorem 2.* $\hat{h}$ *is uniquely determined because the formulation* (8) *is strictly convex in* h*, but* $\hat{\alpha}_0$ *may not. Again, this is also the case for the kernel SVM.*

4.2 Algorithm

There are well-known optimization methods for the standard kernel SVM and the ℓ_2-regularized kernel logistic regression (KLR). The kernel SVM can be reformulated as a quadratic programming problem, and its dual problem can be solved efficiently by sequential minimal optimization (SMO). For KLR, the Newton-Raphson iteration method is commonly used, which leads to iterative weighted least squares. However, these methods are insufficient when the loss is switched to the LHS loss. We consider the dual problem for our method in Lemma 2, but we find that it is not easy to solve, unlike the dual of the kernel SVM. Therefore, we focus on solving (10). On the other hand, the Newton-Raphson algorithm is not applicable here because the LHS loss is not twice differentiable. In this subsection, we develop an algorithm to solve the kernel LHS classifier based on the Majorization-minimization (MM) principle (Hunter and Lange, 2004) thanks to a quadratic upper bound of the LHS loss introduced later in Lemma 3. Our algorithm iteratively updates the solution with an explicit and consistent step size for each tuning parameter λ. We further introduce an efficient computational technique to search solutions for a sequence of λs.

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 $\mathcal{Q}_{\mathbf{K}}(\boldsymbol{\theta} \mid \tilde{\boldsymbol{\theta}})$ which majorizes the objective function $\mathcal{L}_{\mathbf{K}}(\boldsymbol{\theta})$ by a quadratic function (Böhning and Lindsay, 1988).

Lemma 3 *The LHS loss* L *has a quadratic upper bound,*

$$L(u) \le L(\tilde{u}) + L'(\tilde{u})(u - \tilde{u}) + (u - \tilde{u})^2/2, \quad u, \tilde{u} \in \mathbb{R},$$

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

Let $\tilde{\mathbf{z}}$ be an $n \times 1$ vector with ith element $y_i L'\{y_i(\tilde{\alpha}_0 + \mathbf{K}_i^T \tilde{\boldsymbol{\alpha}})\}/n$. By Lemma 3, we have

$$\begin{aligned}
\mathcal{L}_{\mathbf{K}}(\boldsymbol{\theta}) &= \sum_{i=1}^n L\{y_i(\alpha_0 + \mathbf{K}_i^T \boldsymbol{\alpha})\}/n + \lambda \boldsymbol{\alpha}^T \mathbf{K} \boldsymbol{\alpha} \\
&\le \sum_{i=1}^n L\{y_i(\tilde{\alpha}_0 + \mathbf{K}_i^T \tilde{\boldsymbol{\alpha}})\}/n + \lambda \boldsymbol{\alpha}^T \mathbf{K} \boldsymbol{\alpha} \\
&\quad + \begin{pmatrix} \mathbf{1}^T \tilde{\mathbf{z}} \\ \mathbf{K}\tilde{\mathbf{z}} \end{pmatrix}^T \begin{pmatrix} \alpha_0 - \tilde{\alpha}_0 \\ \boldsymbol{\alpha} - \tilde{\boldsymbol{\alpha}} \end{pmatrix} + \frac{1}{2n} \begin{pmatrix} \alpha_0 - \tilde{\alpha}_0 \\ \boldsymbol{\alpha} - \tilde{\boldsymbol{\alpha}} \end{pmatrix}^T \begin{pmatrix} n & \mathbf{1}^T \mathbf{K} \\ \mathbf{K1} & \mathbf{KK} \end{pmatrix} \begin{pmatrix} \alpha_0 - \tilde{\alpha}_0 \\ \boldsymbol{\alpha} - \tilde{\boldsymbol{\alpha}} \end{pmatrix} \\
&= \sum_{i=1}^n L\{y_i(\tilde{\alpha}_0 + \mathbf{K}_i^T \tilde{\boldsymbol{\alpha}})\}/n + \lambda \tilde{\boldsymbol{\alpha}}^T \mathbf{K} \tilde{\boldsymbol{\alpha}} \\
&\quad + \tilde{\gamma}_{\mathbf{K}}^T \begin{pmatrix} \alpha_0 - \tilde{\alpha}_0 \\ \boldsymbol{\alpha} - \tilde{\boldsymbol{\alpha}} \end{pmatrix} + \frac{1}{2n} \begin{pmatrix} \alpha_0 - \tilde{\alpha}_0 \\ \boldsymbol{\alpha} - \tilde{\boldsymbol{\alpha}} \end{pmatrix}^T \mathbf{P}_{\mathbf{K},\lambda} \begin{pmatrix} \alpha_0 - \tilde{\alpha}_0 \\ \boldsymbol{\alpha} - \tilde{\boldsymbol{\alpha}} \end{pmatrix} = \mathcal{Q}_{\mathbf{K}}(\boldsymbol{\theta} \mid \tilde{\boldsymbol{\theta}}),
\end{aligned}$$

where

$$\tilde{\gamma}_{\mathbf{K}} = \begin{pmatrix} \mathbf{1}^T \tilde{\mathbf{z}} \\ \mathbf{K}\tilde{\mathbf{z}} + 2\lambda \mathbf{K} \tilde{\boldsymbol{\alpha}} \end{pmatrix} \quad \text{and} \quad \mathbf{P}_{\mathbf{K},\lambda} = \begin{pmatrix} n & \mathbf{1}^T \mathbf{K} \\ \mathbf{K1} & \mathbf{KK} + 2n\lambda \mathbf{K} \end{pmatrix}.$$

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

$$\begin{pmatrix} \alpha_0 \\ \boldsymbol{\alpha} \end{pmatrix} = \arg\min_{\alpha_0, \boldsymbol{\alpha}} \mathcal{Q}_{\mathbf{K}}(\boldsymbol{\theta} \mid \boldsymbol{\theta}_m) = \begin{pmatrix} \tilde{\alpha}_0 \\ \tilde{\boldsymbol{\alpha}} \end{pmatrix} - n\mathbf{P}_{\mathbf{K},\lambda}^{-1}\tilde{\gamma}_{\mathbf{K}}. \tag{11}$$

In practice, λ is unknown, and we rely on cross-validation to select the best λ. 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 LHS classifier 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 that only needs to invert a matrix once. Compute the eigen decomposition $\mathbf{K} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^T$ and invert $\mathbf{P}_{\mathbf{K},\lambda}$ blockwise as follows.

$$\begin{aligned} \mathbf{P}_{\mathbf{K},\lambda}^{-1} &= \begin{pmatrix} n & \mathbf{1}^T\mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^T \\ \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^T\mathbf{1} & \mathbf{U}\boldsymbol{\Pi}_{\mathbf{K},\lambda}\mathbf{U}^T \end{pmatrix}^{-1} \\ &= g_{\mathbf{K}} \begin{pmatrix} 1 \\ -\mathbf{v}_{\mathbf{K}} \end{pmatrix} \begin{pmatrix} 1 & -\mathbf{v}_{\mathbf{K}}^T \end{pmatrix} + \begin{pmatrix} 0 & \mathbf{0}^T \\ \mathbf{0} & \mathbf{U}\boldsymbol{\Pi}_{\mathbf{K},\lambda}^{-1}\mathbf{U}^T \end{pmatrix}, \end{aligned} \tag{12}$$

where $\boldsymbol{\Pi}_{\mathbf{K},\lambda} = \boldsymbol{\Lambda}^2 + 2n\lambda\boldsymbol{\Lambda}$, $g_{\mathbf{K}} = 1/(n - \mathbf{1}^T\mathbf{U}\boldsymbol{\Lambda}\boldsymbol{\Pi}_{\mathbf{K},\lambda}^{-1}\boldsymbol{\Lambda}\mathbf{U}^T\mathbf{1})$, and $\mathbf{v}_{\mathbf{K}} = \mathbf{U}\boldsymbol{\Lambda}\boldsymbol{\Pi}_{\mathbf{K},\lambda}^{-1}\mathbf{U}^T\mathbf{1}$. Replacing $\mathbf{P}_{\mathbf{K},\lambda}^{-1}$ with (12), we see that the right-hand side of (11) becomes

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

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

Remark 4 *We defer the algorithm of the linear LHS classifier to Appendix C as we take a similar procedure. The code for the linear and kernel LHS classifier is available from the authors upon request.*

5. Bayes Risk Consistency

In this section, we establish the Bayes risk consistency of the kernel LHS classifier, which, in our view, provides the most important justification for this new loss function.

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

$$\hat{f}_n = \operatorname*{argmin}_{f \in \mathcal{H}_K} \left[\sum_{i=1}^{n} L\left(y_i f(\mathbf{x}_i)\right)/n + \lambda_n ||f||_{\mathcal{H}_K}^2 \right], \tag{13}$$

and f^* be the Bayes rule. Let the expected misclassification rate of a classification function $\hat{f}$ be denoted as $R(\hat{f}) = \mathrm{pr}[\mathbf{Y} \neq \mathrm{sgn}\{\hat{f}(\mathbf{X})\}]$. We say the kernel LHS classifier 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 $\mathrm{E}[L(\mathbf{Y}f(\mathbf{X})) \mid \mathbf{X} = \mathbf{x}]$.

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 resulting from using a finite sample size on kernel methods. It allows us 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 does not apply to the LHS 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 LHS loss is more involved. To study when $\mathrm{E}[L(\mathbf{Y}\bar{f}(\mathbf{X}))]$ is finite and not finite, we introduce $g(\delta) = \mathrm{pr}(p(\mathbf{X})(1-p(\mathbf{X})) \leq \delta/2(1-\delta/2))$. Intuitively, for a small δ, $g(\delta)$ can be understood as the probability of the random variable $\mathbf{X}$ having a negligible amount of information about the optimal decision boundary.

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

Assumption 1 is mild as it is satisfied as long as the set of points at which the cumulative distribution functions of $\mathbf{X}|\mathbf{Y}=1$ and $\mathbf{X}|\mathbf{Y}=0$ are discontinuous is countable. Assumption 1 may not satisfy otherwise. An example that the assumption is violated is when $\mathrm{pr}(\mathbf{X} = \mathrm{x}|\mathbf{Y}=0) \propto 1/\mathrm{x}^2$ and $\mathrm{pr}(\mathbf{X} = \mathrm{x}|\mathbf{Y}=1) \propto 1/\mathrm{x}^3$ with x taking a value in $\mathbb{N}$.

Theorem 4 *Consider the underlying distribution satisfying Assumption 1. Let L be the LHS 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*

$$\int_0^{\delta'} \delta^{-1} g(\delta) d\delta < \infty.$$

It turns out that $\mathrm{E}[L(\mathbf{Y}\bar{f}(\mathbf{X}))]$ is finite if and only if $g(\delta)$ satisfies the certain condition, as shown in Theorem 4. Unlike a non-negative loss function, $\mathrm{E}[L(\mathbf{Y}\bar{f}(\mathbf{X}))]$ is not always finite for the LHS loss. One example that $\mathrm{E}[L(\mathbf{Y}\bar{f}(\mathbf{X}))]$ does not exist is when $\mathrm{P}(\{\mathbf{X} : p(\mathbf{X}) = 0 \text{ or } 1\}) > 0$, that is, when there is a non-zero probability of $\mathbf{X}$ having no label noise. $\mathrm{E}[L(\mathbf{Y}\bar{f}(\mathbf{X}))]$ exists when $g(\delta)$ is bounded above by δ up to a constant as $\delta \to 0$.

Remark 5 *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 $\hat{f}_n(\mathbf{x}) = -\sum_{i=1}^n L'(y_i\hat{f}_n(\mathbf{x}_i))y_i K(\mathbf{x}_i,\mathbf{x})/(2n\lambda_n)$, assuming that $\mathbf{K}$ is invertible. It indicates $|\hat{f}_n| \leq \frac{1}{2\lambda_n}B$, and thus $\mathrm{E}[L(\mathbf{Y}\hat{f}(\mathbf{X}))]$ is finite.*

This result motivates us to devise a different upper bound by considering 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 4 *Let f^* be the Bayes rule and $\hat{f}_n$ be the equation (13). Then for any $0 < \delta \leq 1$,*

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

where L is the LHS 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 a 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 Assumption 1 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 LHS classifier and the standard SVM in terms of classification accuracy. Such a comparison will directly show the impact of using the LHS 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 LHS classifier is similar to or better than the kernel SVM in terms of classification accuracy, which implies that the LHS classifier is a worthy new classifier in that it gives a reasonable performance, we can claim that it is valid to use a negatively divergent loss to get a classification learning algorithm. In practice, we recommend trying both methods and using cross-validation to pick the better one for a given data set.

6.1 Simulation

We first compare the LHS classifier 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 both have half of their components as 1s and the other half as -1s. We draw μ_k^+ and μ_k^-, $k = 1, \ldots, K$ from

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

Given μ_k^+ and μ_k^-, let $(\mathbf{X}, \mathbf{Y})$ be a random pair such that $\text{pr}(\mathbf{Y} = -1) = \text{pr}(\mathbf{Y} = +1) = 1/2$ and $\mathbf{X} \in \mathbb{R}^p$ with

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

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

Sampling from the above distribution can also be illustrated using label flipping. Firstly, generate samples with the positive class from $\sum_{k=1}^{10} N(m_k^+, \sigma^2\mathbf{I}_{p\times p})/K$ with each m_k^+ drawn from $\text{N}(\boldsymbol{\mu}^+, \mathbf{I}_{p\times p})$. Each sample would be from one of $N(m_k, \sigma^2), k = 1, \ldots K$, with an equal probability. Next, if the sample is from $N(m_k^+, \sigma^2)$ for some $k > 2K/3$, flip the label into the

n	Kernel	Misclassification rate (%) LHS Classifier	SVM
Example 1: $K = 3$, $p = 2$, Bayes error: 11.13%			
50	Linear	34.63 (0.05)	34.35 (0.05)
	Gaussian	17.11 (0.04)	17.29 (0.04)
90	Linear	33.16 (0.05)	32.80 (0.05)
	Gaussian	14.40 (0.04)	15.13 (0.04)
200	Linear	31.62 (0.05)	31.43 (0.05)
	Gaussian	12.80 (0.03)	13.21 (0.03)
900	Linear	30.65 (0.05)	30.72 (0.05)
	Gaussian	11.47 (0.03)	11.50 (0.03)
Example 2: $K = 10$, $p = 2$, Bayes error: 11.51%			
50	Linear	40.37 (0.05)	40.78 (0.05)
	Gaussian	23.84 (0.04)	24.13 (0.04)
90	Linear	39.72 (0.05)	40.46 (0.05)
	Gaussian	18.79 (0.04)	19.62 (0.04)
200	Linear	38.00 (0.05)	38.58 (0.05)
	Gaussian	15.45 (0.04)	16.38 (0.04)
900	Linear	37.12 (0.05)	38.11 (0.05)
	Gaussian	12.66 (0.03)	13.00 (0.03)
Example 3: $K = 3$, $p = 30$, Bayes error: 13.48%			
50	Linear	32.87 (0.05)	32.95 (0.05)
	Gaussian	31.36 (0.05)	30.97 (0.05)
90	Linear	28.80 (0.05)	29.17 (0.05)
	Gaussian	26.39 (0.04)	26.70 (0.04)
200	Linear	25.26 (0.04)	25.68 (0.04)
	Gaussian	22.61 (0.04)	22.79 (0.04)
900	Linear	22.42 (0.04)	22.50 (0.04)
	Gaussian	18.07 (0.04)	18.02 (0.04)

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

negative class. Similarly, draw negative class samples from $\sum_{k=1}^{10} N(m_k^-, \sigma^2 \mathbf{I}_{p\times p})/K$ with each m_k^- drawn from $\mathrm{N}(\boldsymbol{\mu}^-, \mathbf{I}_{p\times p})$, and then flip the label if the sample is from $N(\mu_k^-, \sigma^2)$ for some $k > 2K/3$.

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 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 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 LHS classifier are comparable in general. The Gaussian LHS classifier slightly outperforms the Gaussian SVM in Example 1, and the linear LHS classifier consistently outperforms the linear SVM in Example 2 and Example 3. Second, the misclassification rates of both the Gaussian LHS classifier 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 examples

We examine the performance of the LHS classifier compared to SVM on 15 data sets from the University of California at Irvine Machine Learning Repository (Dua and Graff, 2017). These data sets 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 a five-fold cross-validation for selecting an optimal λ from 100 λ-values. The remaining 1/3 of observations is set as the test set for calculating the misclassification rate. We repeat this process 100 times and report the average misclassification rates with the corresponding standard errors in Table 2. For both algorithms, the maximum iteration number is set to $10,000$.

For these real data example, the computational time averaged by three repetitions is also compared. The timing is based on a standard PC with 2.9 GHz Dual-Core Intel Core i5 processor and 8GB of memory.

When comparing the linear LHS classifier and the linear SVM, we observe that the linear LHS classifier outperforms the linear SVM on ten data sets. When comparing the kernel classifiers, the Gaussian LHS classifier outperforms the kernel SVM on six data sets.

Regarding computational time, the LHS classifier is advantageous compared to SVM. The linear LHS classifier is much faster than the linear SVM on all data sets. The Gaussian LHS classifier is also much faster than the kernel SVM on thirteen out of fifteen data sets. In the cases where the Gaussian LHS classifier is slower may be because a substantial amount of time is required when λ is very close to zero.

7. Concluding Remarks

In this paper, we have introduced the first negatively divergent loss function named the LHS loss for margin-based classification. Despite some technical difficulties brought by the negative divergence of the loss function, we have proved the classification-calibration property of the LHS loss and established the Bayes risk consistency of the leaky kernel

Dataset	n	p	Kernel	LHS Classifier		SVM	
				Error (%)	Time (s)	Error (%)	Time (s)
Arrhythmia	68	233	Linear	21.39 (0.84)	0.48	21.39 (0.85)	10.08
			Gaussian	20.13 (0.82)	2.18	21.78 (0.85)	10.23
Australian	690	14	Linear	13.59 (0.23)	0.35	13.62 (0.23)	344.94
			Gaussian	13.89 (0.23)	2.34	13.61 (0.23)	11.96
Banknote	1372	4	Linear	1.07 (0.05)	6.84	1.08 (0.05)	10.41
			Gaussian	0.44 (0.03)	25.49	0.01 (0.00)	11.68
Biodeg	1055	41	Linear	13.56 (0.18)	6.13	13.23 (0.18)	652.33
			Gaussian	12.23 (0.17)	23.87	12.28 (0.17)	26.03
Bupa	345	6	Linear	31.59 (0.77)	0.06	31.79 (0.77)	18.35
			Gaussian	31.16 (0.77)	0.38	31.34 (0.77)	4.26
Chess	3196	36	Linear	2.82 (0.05)	41.41	3.25 (0.05)	69.64
			Gaussian	1.70 (0.04)	239.25	0.93 (0.03)	152.53
cle:Heart	297	13	Linear	16.15 (0.37)	0.11	15.85 (0.37)	3.31
			Gaussian	16.47 (0.37)	0.36	16.31 (0.37)	3.92
Hepatitis	80	19	Linear	14.44 (0.67)	0.16	16.19 (0.70)	3.21
			Gaussian	13.59 (0.65)	2.84	13.89 (0.66)	3.09
Hungarian	261	10	Linear	17.69 (0.41)	0.08	18.52 (0.41)	3.35
			Gaussian	18.69 (0.42)	0.26	17.90 (0.41)	3.74
LSVT	126	310	Linear	14.31 (0.53)	0.60	16.69 (0.57)	20.86
			Gaussian	15.10 (0.55)	15.62	16.07 (0.56)	27.86
Musk	475	166	Linear	17.08 (0.30)	2.30	16.85 (0.30)	26.82
			Gaussian	10.48 (0.24)	4.11	8.29 (0.22)	62.61
Parkinsons	195	22	Linear	15.25 (0.40)	0.16	14.14 (0.43)	9.40
			Gaussian	11.68 (0.40)	1.72	9.06 (0.35)	5.85
Sonar	208	60	Linear	23.39 (0.51)	0.61	25.10 (0.52)	8.73
			Gaussian	17.36 (0.45)	0.91	15.65 (0.43)	8.09
Spectf	80	22	Linear	30.07 (0.87)	0.13	31.19 (0.88)	5.00
			Gaussian	26.74 (0.84)	0.14	28.04 (0.85)	7.79
Vertebral	310	6	Linear	14.88 (0.35)	0.14	15.18 (0.35)	14.38
			Gaussian	16.81 (0.37)	1.14	15.83 (0.36)	6.69

Table 2: Misclassification rates averaged by 100 runs and computation time averaged by 3 runs on 15 data sets from the University of California at Irvine Machine Learning Repository. The standard error is given in parentheses. The computational time includes tuning the parameters.

SVM. We further have provided numeric evidence to show that the linear and kernel LHS classifier is at least as competitive as the usual linear and kernel SVM. All of these provide a 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).

At a high level, this paper aims to change the conventional view of the loss function used for deriving a learning algorithm in classification. Before this paper, the loss function is usually borrowed from the loss function in decision theory which should be non-negative. We carefully examine the LHS loss and find that the non-negativity is unnecessary to get to the learning algorithm. Although the LHS classifier is as competitive as the SVM in the settings we analyzed, this is only used to justify the validity of our approach. We hope this paper will stimulate more interest in the study of negatively divergent loss functions in machine learning.

Before ending, we provide some further discussions on the proposed method.

7.1 Non-likelihood method

Owing to its margin-maximization view, the SVM was largely regarded as a non-likelihood based method, in contrast to the logistic regression. Later, Franc et al. (2011) interpreted the linear SVM as the maximum likelihood estimator (MLE) of an appropriate likelihood function. They considered a class of the probability density function (pdf) in the form:

$$p(\mathbf{x}, y; \tau, \mathbf{w}) = C(\tau) \cdot \exp\left(-\frac{1}{2} L\left(y\mathbf{x}^T(\tau\mathbf{w})\right)\right) \cdot h(\mathbf{x}), \quad (\mathbf{x}, y) \in \mathbb{R}^p \times \{-1, +1\}, \tag{14}$$

where L is the hinge loss, $\tau > 0$, $\mathbf{w}$ is the unit vector satisfying $\mathbf{w}^T\mathbf{w} = 1$, $C(\tau)$ is a normalization constant, and $h(\mathbf{x}) \geq 0$ is a piece-wise continuous and integrable function which ensures that $C(\tau)$ does not depend on $\mathbf{w}$. Note that $p(\mathbf{x}, y; \tau, \mathbf{w})$ is a well-defined pdf following from the non-negativity of the hinge loss. Then, they showed that the normalized solution to (5) with the hinge loss and the ℓ_2 regularizer is equivalent to the MLE of $\mathbf{w}$ from (14) for some $\tau > 0$. So, by (14), the SVM cannot be completely separated from the likelihood approach.

If we try to replace the hinge loss with the leaky hockey stick loss in (14), we do not end up with a proper distribution function, as $\exp\left(-\frac{1}{2}L\left(y\mathbf{x}^T(\tau\mathbf{w})\right)\right)$ is not bounded and thereby not integrable. Therefore, the leaky hockey stick classifier is even farther away from the likelihood approach than the SVM.

7.2 A family of negatively divergent loss functions

The leaky hockey stick (LHS) loss is not the only negatively divergent margin-based loss function for classification. It is the first one we discovered and studied. During the revision, we find a family of negatively divergent loss functions which share similar properties to the LHS loss.

We define a negatively divergent loss functions L_r as follows: with $r \in (1, \infty)$,

$$L_r(yf) = \begin{cases} r\left(1 - (yf)^{1/r}\right), & yf > 1, \\ 1 - yf, & yf \leq 1. \end{cases} \tag{15}$$

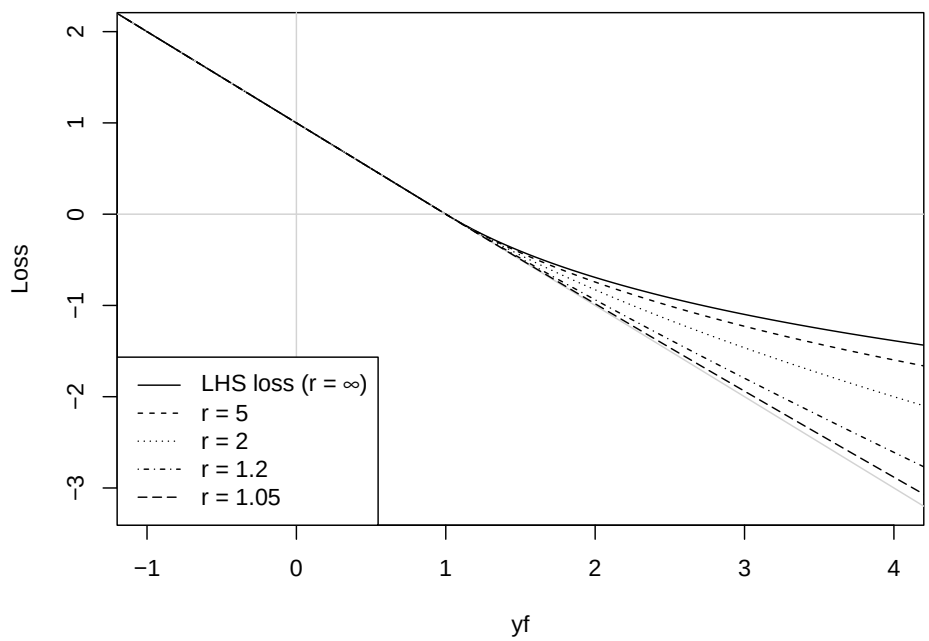


Figure 3: Plot of other negatively divergent margin-based loss functions. The gray diagonal reference line is when $y = 1 - x$.

Figure 3 illustrates (15) for various r values. Note the LHS loss is (15) with r diverging to ∞.

The next theorem shows that the loss function (15) is classification calibrated, saying that minimizing the loss (15) results in the Bayes rule in population.

Theorem 6 *(Classification-calibration) For any* $\mathbf{x}$ *such that* $p(\mathbf{x}) \in (0,1)$*, a global minimizer of* $\mathrm{E}[L_r(\mathbf{Y}f(\mathbf{X})) \mid \mathbf{X} = \mathbf{x}]$ *uniquely exists and is*

$$\begin{aligned} &-\{(1-p(\mathbf{x}))/p(\mathbf{x})\}^{r/(r-1)}, && \textit{if } \ p(\mathbf{x}) < 1/2, \\ \textit{and } &+\{p(\mathbf{x})/(1-p(\mathbf{x}))\}^{r/(r-1)}, && \textit{if } \ p(\mathbf{x}) > 1/2. \end{aligned}$$

Note that $\mathrm{sgn}\{\bar{f}(\mathbf{x})\} = f^*(\mathbf{x})$*, where* f^* *is the Bayes rule.*

We define $\hat{f}_n$ as the classification function based on another negatively divergent loss function L (15) with sample size n,

$$\hat{f}_n = \underset{f \in \mathcal{H}_K}{\mathrm{argmin}} \left[\sum_{i=1}^{n} L_r\left(y_i f(\mathbf{x}_i)\right)/n + \lambda_n ||f||^2_{\mathcal{H}_K} \right]. \tag{16}$$

Theorem 7 *(Bayes risk consistency) Assume Assumption 1 and that* $g(\delta_n)^{(r-1)/r}\delta^{2/(1-r)} \to 0$ *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, where* $\hat{f}_n$ *is defined in* (16) *and* f^* *is the Bayes rule.*

Acknowledgments

We would like to thank the anonymous reviewers and Action Editor for their constructive and useful insights and suggestions, which have improved the quality of the article. Zou's research is supported in part by NSF DMS 1915842 and 2015120.

Appendix A. Proofs

A.1 Proof of Theorem 1

Fix $\mathbf{x}$ and let $\alpha = f(\mathbf{x})$. We have

$$\begin{aligned}\mathrm{E}[L(\mathbf{Y}\alpha)|\mathbf{X}=\mathbf{x}] &= p(\mathbf{x})L(\alpha) + \{1-p(\mathbf{x})\}L(-\alpha)\\ &= \begin{cases} p(\mathbf{x})(-\log\alpha) + (1-p(\mathbf{x}))(1+\alpha), & \text{if } \alpha > 1,\\ p(\mathbf{x})(1-\alpha) + \{1-p(\mathbf{x})\}(1+\alpha), & \text{if } -1 \le \alpha \le 1,\\ p(\mathbf{x})(1-\alpha) + (1-p(\mathbf{x}))(-\log(-\alpha)), & \text{if } \alpha < -1.\end{cases}\end{aligned}$$

If $0 < p(\mathbf{x}) < 1$, the global minimizer exists because $\mathrm{E}[L(\mathbf{Y}\alpha)|\mathbf{X}=\mathbf{x}]$ is coercive:

$$\mathrm{E}[L(\mathbf{Y}\alpha)|\mathbf{X}=\mathbf{x}] \to \infty \quad \text{as } |\alpha| \to \infty.$$

The derivative of $\mathrm{E}[L(\mathbf{Y}\alpha)|\mathbf{X}=\mathbf{x}]$ with respect to α is

$$\frac{\partial}{\partial\alpha}\mathrm{E}[L(\mathbf{Y}\alpha)|\mathbf{X}=\mathbf{x}] = \begin{cases} -p(\mathbf{x})\frac{1}{\alpha} + \{1-p(\mathbf{x})\}, & \text{if } \alpha > 1,\\ -p(\mathbf{x}) + \{1-p(\mathbf{x})\}, & \text{if } -1 \le \alpha \le 1,\\ -p(\mathbf{x}) - \{1-p(\mathbf{x})\}\frac{1}{\alpha}, & \text{if } \alpha < -1.\end{cases}$$

$\frac{\partial}{\partial\alpha}\mathrm{E}[L(\mathbf{Y}\alpha)|\mathbf{X}=\mathbf{x}] = 0$ holds if and only if

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

α is the unique minimizer because $\mathrm{E}[L(\mathbf{Y}\alpha)|\mathbf{X}=\mathbf{x}]$ is convex.

A.2 Proof of Theorem 2

Let $(\hat{\beta}_0, \hat{\boldsymbol{\beta}})$ be any minimizer if it exists. Without loss of generality, let's say that $p_+ \ge p_-$ where $p_+ = \#\{i : y_i = +1\}/n > 0$ and $p_- = \#\{i : y_i = -1\}/n > 0$. Assume that,

$$\left\|(\hat{\beta}_0 \;\; \hat{\boldsymbol{\beta}}^T)^T\right\|^2 > C_1^2 + C_2^2,$$

where,

$$C_1 = \left(\frac{p_+}{p_-} + \frac{1}{\sqrt{p_-}} + \frac{\max_i ||\mathbf{x}_i||}{2\sqrt{\lambda p_-}}\right)^2 > 1,$$

$$\text{and} \;\; C_2 = \frac{\max_i ||\mathbf{x}_i||}{\lambda} + \sqrt{\frac{C_1}{\lambda}}.$$

Since $|L(a) - L(b)| \leq |a - b|$ for any $a, b \in \mathbb{R}$,

$$\begin{aligned} L\left(y_i(\beta_0 + \mathbf{x}_i^T \boldsymbol{\beta})\right) &\geq L(y_i\beta_0) - |y_i \mathbf{x}_i^T \boldsymbol{\beta}| \\ &\geq L(y_i\beta_0) - ||\mathbf{x}_i|| \cdot ||\boldsymbol{\beta}|| \\ &\geq L(y_i\beta_0) - \max_i ||\mathbf{x}_i|| \cdot ||\boldsymbol{\beta}||. \end{aligned}$$

Then, we have

$$\begin{aligned} \mathcal{L}(\beta_0, \boldsymbol{\beta}) &\geq \frac{1}{n}\sum_{i=1}^n L(y_i\beta_0) - \max_i ||\mathbf{x}_i|| \cdot ||\boldsymbol{\beta}|| + \lambda ||\boldsymbol{\beta}||^2 \\ &= \frac{1}{n}\sum_{i=1}^n L(y_i\beta_0) + \lambda \left(\|\boldsymbol{\beta}\| - \frac{\max_i ||\mathbf{x}_i||}{2\lambda} \right)^2 - \lambda \left(\frac{\max_i ||\mathbf{x}_i||}{2\lambda} \right)^2. \end{aligned}$$

(Case 1) Suppose that $\hat{\beta}_0^2 \leq C_1^2$. It implies that $||\hat{\boldsymbol{\beta}}||^2 > C_2^2$ and we obtain that,

$$\begin{aligned} \mathcal{L}(\hat{\beta}_0, \hat{\boldsymbol{\beta}}) &\geq \frac{1}{n}\sum_{i=1}^n L(y_i\hat{\beta}_0) + \lambda \left(||\hat{\boldsymbol{\beta}}|| - \frac{\max_i ||\mathbf{x}_i||}{2\lambda} \right)^2 - \lambda \left(\frac{\max_i ||\mathbf{x}_i||}{2\lambda} \right)^2 \\ &\geq \frac{1}{n}\sum_{i=1}^n (1 - y_i\hat{\beta}_0) + \lambda \left(||\hat{\boldsymbol{\beta}}|| - \frac{\max_i ||\mathbf{x}_i||}{2\lambda} \right)^2 - \lambda \left(\frac{\max_i ||\mathbf{x}_i||}{2\lambda} \right)^2 \\ &> (1 - |\hat{\beta}_0|) + \lambda \left(C_2 - \frac{\max_i ||\mathbf{x}_i||}{2\lambda} \right)^2 - \lambda \left(\frac{\max_i ||\mathbf{x}_i||}{2\lambda} \right)^2 \\ &\geq 1 - C_1 + \lambda \left(\frac{\max_i ||\mathbf{x}_i||}{2\lambda} + \sqrt{\frac{C_1}{\lambda}} \right)^2 - \lambda \left(\frac{\max_i ||\mathbf{x}_i||}{2\lambda} \right)^2 \\ &> 1 = \mathcal{L}(0, \mathbf{0}), \end{aligned}$$

which is a contradiction. The third inequality is because $||\hat{\boldsymbol{\beta}}|| > C_2 > \frac{\max_i ||\mathbf{x}_i||^2}{2\lambda}$.

(Case 2) Suppose $\hat{\beta}_0^2 > C_1^2$ (> 1). $\log|\hat{\beta}_0|$ is bounded by $\log|\hat{\beta}_0| \leq \frac{|\hat{\beta}_0|-1}{|\hat{\beta}_0|^{1/2}}$. Then we obtain that,

$$\begin{aligned}
\mathcal{L}(\hat{\beta}_0, \hat{\boldsymbol{\beta}}) &\geq \frac{1}{n}\sum_{i=1}^{n} L(y_i\hat{\beta}_0) + \lambda\left(||\hat{\boldsymbol{\beta}}|| - \frac{\max_i ||\mathbf{x}_i||}{2\lambda}\right)^2 - \lambda\left(\frac{\max_i ||\mathbf{x}_i||}{2\lambda}\right)^2 \\
&\geq p_+ \cdot (-\log|\hat{\beta}_0|) + p_- \cdot (1 + |\hat{\beta}_0|) - \lambda\left(\frac{\max_i ||\mathbf{x}_i||}{2\lambda}\right)^2 \\
&\geq p_+ \cdot \left(-\frac{|\hat{\beta}_0| - 1}{|\hat{\beta}_0|^{1/2}}\right) + p_- \cdot (1 + |\hat{\beta}_0|) - \lambda\left(\frac{\max_i ||\mathbf{x}_i||}{2\lambda}\right)^2 \\
&\geq -p_+|\hat{\beta}_0|^{1/2} + p_-|\hat{\beta}_0| - \lambda\left(\frac{\max_i ||\mathbf{x}_i||}{2\lambda}\right)^2 \\
&= p_-\left(|\hat{\beta}_0|^{1/2} - \frac{1}{2}\frac{p_+}{p_-}\right)^2 - p_-\left(\frac{1}{2}\frac{p_+}{p_-}\right)^2 - \frac{\max_i ||\mathbf{x}_i||^2}{4\lambda} \\
&> p_-\left(C_1^{1/2} - \frac{1}{2}\frac{p_+}{p_-}\right)^2 - p_-\left(\frac{1}{2}\frac{p_+}{p_-}\right)^2 - \frac{\max_i ||\mathbf{x}_i||^2}{4\lambda} \\
&= p_-\left(\frac{1}{2}\frac{p_+}{p_-} + \frac{1}{\sqrt{p_-}} + \frac{\max_i ||\mathbf{x}_i||}{2\sqrt{\lambda p_-}}\right)^2 - p_-\left(\frac{1}{2}\frac{p_+}{p_-}\right)^2 - \frac{\max_i ||\mathbf{x}_i||^2}{4\lambda} \\
&> 1 = \mathcal{L}(0, \mathbf{0}),
\end{aligned}$$

which is a contradiction.

Therefore, we can conclude that,

$$\left\|(\hat{\beta}_0 \;\; \hat{\boldsymbol{\beta}}^T)^T\right\|^2 \leq C_1^2 + C_2^2.$$

Eventually, (2) is equivalent to minimizing a continuous function over a non-empty compact region. Therefore, there exists a global minimizer.

A.3 Proof of Lemma 1

We start from the formulation (3), and show that this can be equally written as $\mathrm{sgn}(\hat{\beta}_0 + \mathbf{x}^T\hat{\boldsymbol{\beta}})$, where $(\hat{\beta}_0, \hat{\boldsymbol{\beta}}^T)^T$ is the solution to (2).

Let $G(\eta_i) = -\log\big(y_i(w_0 + x_i^T\mathbf{w}) + \eta_i\big) + C\eta_i$. Then the objective function of (3) can be written as $\sum_{i=1}^n G(\eta_i)$. We minimize it over η_i. Since

$$\begin{aligned}
G'(\eta_i) &= -\frac{1}{y_i(w_0 + x_i^T\mathbf{w}) + \eta_i} + C = 0 \;\Rightarrow\; \frac{1}{y_i(w_0 + x_i^T\mathbf{w}) + \eta_i} = \frac{1}{C} \\
\text{and}\;\; G''(\eta_i) &= \frac{1}{\big(y_i(w_0 + x_i^T\mathbf{w}) + \eta_i\big)^2} > 0,
\end{aligned}$$

if $y_i(w_0 + x_i^T\mathbf{w}) > \frac{1}{C}$, $\eta_i^* = 0$ is the minimizer and otherwise, $\eta_i^* = \frac{1}{C} - y_i(w_0 + x_i^T\mathbf{w})$ is the minimizer. By plugging in the minimizer η_i^* into $\sum_{i=1}^n G(\eta_i)$, (3) can be written as

$$\min_{w_0,\mathbf{w}} \sum_{i=1}^n \tilde{L}\left(y_i(w_0 + x_i^T\mathbf{w})\right) \quad \text{subject to} \quad \mathbf{w}^T\mathbf{w} \le 1,$$
$$\text{where} \quad \tilde{L}(v) = \begin{cases} -\log v & v > 1/C \\ \log C + 1 - Cv & v \le 1/C. \end{cases}$$

To simplify, let

$$L(u) = \tilde{L}(u/C) - \log C = \begin{cases} -\log u & u > 1 \\ 1 - u & u \le 1. \end{cases}$$

By setting $\beta_0 = C \cdot w_0$ and $\boldsymbol{\beta} = C \cdot \mathbf{w}$, we have

$$\begin{aligned} \sum_{i=1}^n \tilde{L}\left(y_i(w_0 + x_i^T\boldsymbol{w})\right) &= \sum_{i=1}^n L\left(C \cdot y_i(w_0 + x_i^T\boldsymbol{w})\right) + n\log C \\ &= \sum_{i=1}^n L\left(y_i(\beta_0 + x_i^T\boldsymbol{\beta})\right) + n\log C \end{aligned}$$

and

$$\mathbf{w}^T\mathbf{w} \le 1 \;\Leftrightarrow\; \boldsymbol{\beta}^T\boldsymbol{\beta} \le C^2,$$

which proves Lemma 1.

A.4 Proof of Lemma 2

The Lagrangian function of (3) is

$$\begin{aligned} &L(w_0, \mathbf{w}, \eta_i, \alpha_i, \gamma_i, \lambda) \\ &= -\sum_{i=1}^n \log d_i + C\sum_{i=1}^n \eta_i - \sum_{i=1}^n \alpha_i d_i - \sum_{i=1}^n \gamma_i\eta_i + \lambda(\mathbf{w}^T\mathbf{w} - 1) \\ &= -\sum_{i=1}^n \log\left(y_i(w_0 + \mathbf{x}_i^T\mathbf{w}) + \eta_i\right) + C\sum_{i=1}^n \eta_i - \sum_{i=1}^n \alpha_i\left(y_i(w_0 + \mathbf{x}_i^T\mathbf{w}) + \eta_i\right) - \sum_{i=1}^n \gamma_i\eta_i + \lambda(\mathbf{w}^T\mathbf{w} - 1), \end{aligned}$$

with Lagrange multipliers $\alpha_i \ge 0$, $\gamma_i \ge 0$, and $\lambda \ge 0$. The optimality conditions say that $(w_0, \mathbf{w}, \eta_i, \alpha_i, \gamma_i, \lambda)$ is an optimal solution-Lagrange multiplier pair if and only if the following four groups of conditions hold:

(Lagrangian optimality)

$$\begin{aligned} &\frac{\partial L}{\partial w_0} = -\sum_{i=1}^n \frac{y_i}{d_i} - \sum_{i=1}^n \alpha_i y_i = 0, \\ &\frac{\partial L}{\partial \mathbf{w}} = -\sum_{i=1}^n \frac{y_i\mathbf{x}_i}{d_i} - \sum_{i=1}^n \alpha_i y_i x_i + 2\lambda\mathbf{w} = 0, \\ &\text{and} \quad \frac{\partial L}{\partial \eta_i} = -\frac{1}{d_i} + C - \alpha_i - \gamma_i = 0. \end{aligned}$$

(Dual feasibility) $\alpha_i \geq 0, \gamma_i \geq 0,$ and $\lambda \geq 0$.
(Complementary slackness) $\alpha_i d_i = \alpha_i \left(y_i(w_0 + x_i^T w) + \eta_i\right) = 0,$ and $\gamma_i \eta_i = 0$.
(Primal feasibility) $d_i = y_i(w_0 + x_i^T w) + \eta_i \geq 0,$ $\eta_i \geq 0,$ and $w^T w = 1$.

The Complementary slackness implies that $\alpha_i = 0$ for all i. The Lagrangian optimality can be re-written as

$$\sum_{i=1}^{n} \frac{y_i}{d_i} = 0,\ 2\lambda \mathbf{w} = \sum_{i=1}^{n} \frac{y_i \mathbf{x}_i}{d_i},\ \text{and } C - \gamma_i = \frac{1}{d_i},$$

and then we get

$$C \sum_{i=1}^{n} \eta_i = n - 2\lambda \mathbf{w}^T \mathbf{w},\ \text{and } \mathbf{w}^T \mathbf{w} = \frac{1}{4\lambda^2} \left(\sum_{i=1}^{n} \sum_{j=1}^{n} (C - \gamma_i)(C - \gamma_j) y_i y_j x_i^T x_j \right).$$

Then, the dual problem becomes

$$\begin{aligned} \max_{\gamma_i} \ & \sum_{i=1}^{n} \log(C - \gamma_i) + n - \sqrt{\sum_{i=1}^{n} \sum_{j=1}^{n} (C - \gamma_i)(C - \gamma_j) y_i y_j x_i^T x_j} \\ \text{subject to } \ & 0 \leq \gamma_i < C,\ \text{and } \sum_{i=1}^{n} y_i (C - \gamma_i) = 0. \end{aligned}$$

By setting $\sigma_i = C - \gamma_i$, the Lemma 2 holds.

To organize, the optimality conditions to be a primal and dual optimal solution pair are the following:

$$\begin{aligned} & \sigma_i d_i = 1,\ \mathbf{Y}^T \boldsymbol{\sigma} = 0, \\ & 0 < \sigma_i \leq C,\ \eta_i \geq 0,\ (C - \sigma_i)\eta_i = 0,\ \mathbf{w}^T \mathbf{w} \leq 1, \\ & \text{Either } \mathbf{X}^T \mathbf{Y} \boldsymbol{\sigma} = 0 \text{ or } \mathbf{w} = \mathbf{X}^T \mathbf{Y} \boldsymbol{\sigma} / \left\| \mathbf{X}^T \mathbf{Y} \boldsymbol{\sigma} \right\| \\ & , \end{aligned} \tag{17}$$

where σ_i is the ith element of $\boldsymbol{\sigma}$.

A.5 Proof of the existence of the solution to (4)

Assume that there exist i, j with $y_i = +1$ and $y_j = -1$. Let $D(\boldsymbol{\sigma}) = \mathbf{1}^T \log \boldsymbol{\sigma} - \left\| \mathbf{X}^T \mathbf{Y} \boldsymbol{\sigma} \right\|$ and $\mathcal{F}_{\boldsymbol{\sigma}} = \{\boldsymbol{\sigma} \in \mathbb{R}^n | 0 < \boldsymbol{\sigma} \leq C\mathbf{1} \text{ and } \mathbf{y}^T \boldsymbol{\sigma} = 0\}$. As $\mathcal{F}_{\boldsymbol{\sigma}} \neq \emptyset$, there exists $\tilde{\boldsymbol{\sigma}} \in \mathcal{F}_{\boldsymbol{\sigma}}$. Consider the set $S = \{\boldsymbol{\sigma} \in \mathcal{F}_{\boldsymbol{\sigma}} | D(\boldsymbol{\sigma}) \geq D(\tilde{\boldsymbol{\sigma}})\}$, and note that $S \subseteq \tilde{S} = \{\boldsymbol{\sigma} \in \mathcal{F}_{\boldsymbol{\sigma}} | \mathbf{1}^T \log \boldsymbol{\sigma} \geq D(\tilde{\boldsymbol{\sigma}})\}$. For any $\boldsymbol{\sigma} \in \tilde{S}$, $\sigma_j C^{n-1} \geq \Pi_{i=1}^{n} \sigma_i \geq \exp(D(\tilde{\boldsymbol{\sigma}}))$ for all j, and thereby $\sigma_j \geq \exp(D(\tilde{\boldsymbol{\sigma}}))/C^{n-1}$ for all j. By Weierstrass's theorem, the problem $\min\{D(\boldsymbol{\sigma}) | \boldsymbol{\sigma} \in \mathcal{F}_{\boldsymbol{\sigma}} \cap \{\boldsymbol{\sigma} | \boldsymbol{\sigma} \geq \exp(D(\tilde{\boldsymbol{\sigma}}))/C^{n-1} \cdot \mathbf{1}\}\}$ has at least one global solution. The set of global solutions to the problem $\max\{D(\boldsymbol{\sigma}) | \boldsymbol{\sigma} \in \mathcal{F}_{\boldsymbol{\sigma}} \cap \{\boldsymbol{\sigma} | \boldsymbol{\sigma} \geq \exp(D(\tilde{\boldsymbol{\sigma}}))/C^{n-1} \cdot \mathbf{1}\}\}$ is a global solution to $\max\{D(\boldsymbol{\sigma}) | \boldsymbol{\sigma} \in \mathcal{F}_{\boldsymbol{\sigma}}\}$.

A.6 Lemma 5 and its proof

We introduce Lemma 5 which is used to prove Lemma 6 and Theorem 3. Lemma 5 says that as λ goes to 0, the size of $\hat{\boldsymbol{\beta}}_\lambda$ must diverge to minimize the regularized problem (5) since the LHS loss is a strictly decreasing function.

Lemma 5 *Assume the data* $\{\mathbf{x}_i, y_i\}_{i=1}^n$ *is separable, i.e.,* $\exists \bar{\mathbf{w}}$ *s.t.* $\bar{m} := \min_i y_i \mathbf{x}_i^T \bar{\mathbf{w}} > 0$ *with* $\|\bar{\mathbf{w}}\|_q = 1$*. Then* $\|\hat{\boldsymbol{\beta}}_\lambda\|_q \to \infty$ *as* $\lambda \to 0$*.*

Proof Fix $\lambda_0 > 0$. For any $\epsilon > 1$,

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^n L(y_i \boldsymbol{\beta}^T \mathbf{x}_i) + \frac{\lambda_0}{\epsilon} \|\boldsymbol{\beta}\|_q^q \leq \sum_{i=1}^n L(y_i (\epsilon^{1/q} \bar{\mathbf{w}})^T \mathbf{x}_i) + \lambda_0 \|\bar{\mathbf{w}}\|_q^q.$$

This indicates that as $\epsilon \to \infty$,

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^n L(y_i \boldsymbol{\beta}^T \mathbf{x}_i) + \frac{\lambda_0}{\epsilon} \|\boldsymbol{\beta}\|_q^q \to -\infty.$$

In other words, as $\lambda \to 0$,

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^n L(y_i \boldsymbol{\beta}^T \mathbf{x}_i) + \lambda \|\boldsymbol{\beta}\|_q^q \to -\infty,$$

which implies that $\sum_{i=1}^n L(y_i \hat{\boldsymbol{\beta}}_\lambda^T \mathbf{x}_i) \to -\infty$ as $\lambda \to 0$. Therefore $\|\hat{\boldsymbol{\beta}}_\lambda\|_q$ must diverge. ■

A.7 Lemma 6 and its proof

It is not trivial whether the LHS loss prefers a separating decision boundary over a non-separating one because it has a negatively divergent property. Lemma 6 shows that the linear decision boundary defined by $\hat{\mathbf{w}}_\lambda$ (or $\hat{\boldsymbol{\beta}}_\lambda$) perfectly separates training data into two classes when λ is small.

Lemma 6 *Assume the data* $\{\mathbf{x}_i, y_i\}_{i=1}^n$ *are separable, i.e.,* $\exists \bar{\mathbf{w}}$ *s.t.* $\bar{m} = \min_i y_i \mathbf{x}_i^T \bar{\mathbf{w}} > 0$ *with* $\|\bar{\mathbf{w}}\|_q = 1$*. Then, there exists* δ_0 *such that for any* $\lambda < \delta_0$*,* $1 \leq \min y_i \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_\lambda$*.*

Proof Let $1/q + 1/r = 1$. By Lemma 5, there exists δ_0 such that for any $\lambda < \delta_0$,

$$\|\hat{\boldsymbol{\beta}}_\lambda\|_q \geq \max\left\{\frac{1}{\bar{m}}, \frac{(\max_i \|\mathbf{x}_i\|_r)^{n-1}}{\bar{m}^n}, \frac{1}{\max_i \|\mathbf{x}_i\|_r}\right\}.$$

Suppose by contradiction that there exists λ_0 such that $\lambda_0 < \delta_0$ but $1 > y_k \mathbf{x}_k^T \hat{\boldsymbol{\beta}}_{\lambda_0}$ for some k. Then,

$$\begin{aligned}
\sum_{i=1}^{n} L\left(y_i \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_{\lambda_0}\right) > & \sum_{i=1, i\neq k}^{n} L\left(y_i \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_{\lambda_0}\right) \\
\geq & \sum_{i=1, i\neq k}^{n} -\log\left(y_i \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_{\lambda_0}\right) \mathbb{1}_{\{y_i \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_{\lambda_0} > 1\}} \\
\geq & \sum_{i=1, i\neq k}^{n} -\log\left(\max_i \|\mathbf{x}_i\|_r \|\hat{\boldsymbol{\beta}}_{\lambda_0}\|_q\right) \mathbb{1}_{\{y_i \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_{\lambda_0} > 1\}} \\
\geq & -(n-1)\log(\max_i \|\mathbf{x}_i\|_r) - (n-1)\log(\|\hat{\boldsymbol{\beta}}_{\lambda_0}\|_q) \\
\geq & -n\log(\bar{m}) - n\log(\|\hat{\boldsymbol{\beta}}_{\lambda_0}\|_q) = -n\log(\min_i y_i \mathbf{x}_i^T \bar{\mathbf{w}} \cdot \|\hat{\boldsymbol{\beta}}_{\lambda_0}\|_q) \\
\geq & \sum_{i=1}^{n} L\left(y_i \mathbf{x}_i^T \bar{\mathbf{w}} \|\hat{\boldsymbol{\beta}}_{\lambda_0}\|_q\right),
\end{aligned}$$

where the third inequality is by Hölder's inequality, the fourth inequality is because $\max_i \|\mathbf{x}_i\|_r \|\hat{\boldsymbol{\beta}}_{\lambda_0}\|_q \geq 1$, and the fifth inequality is because $\bar{m}^n \|\hat{\boldsymbol{\beta}}_{\lambda_0}\|_q \geq (\max_i \|\mathbf{x}_i\|_r)^{n-1}$. This contradicts with

$$\sum_{i=1}^{n} L\left(y_i \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_{\lambda_0}\right) \leq \sum_{i=1}^{n} L\left(y_i \mathbf{x}_i^T \bar{\mathbf{w}} \|\hat{\boldsymbol{\beta}}_{\lambda_0}\|_q\right).$$

■

A.8 Proof of Theorem 3

Let $\mathbf{w}^*$ be a convergence point of $\frac{\hat{\boldsymbol{\beta}}_\lambda}{\|\hat{\boldsymbol{\beta}}_\lambda\|_q}$ as $\lambda \to 0$, with $\|\mathbf{w}^*\|_q = 1$. Then there is a sequence $\{\lambda_i\}$ such that $\frac{\hat{\boldsymbol{\beta}}(\lambda_j)}{\|\hat{\boldsymbol{\beta}}(\lambda_j)\|_q} \to \mathbf{w}^*$ as $\lambda_i \to 0$. Assume by contradiction that $\tilde{\mathbf{w}}$ is $\|\tilde{\mathbf{w}}\|_q = 1$ and

$$0 \leq m^* = \prod_{i=1}^{n} y_i \mathbf{x}_i^T \mathbf{w}^* \mathbb{1}_{\{y_i \mathbf{x}_i^T \mathbf{w}^* > 0\}} < \tilde{m} = \prod_{i=1}^{n} y_i \mathbf{x}_i^T \tilde{\mathbf{w}} \mathbb{1}_{\{y_i \mathbf{x}_i^T \tilde{\mathbf{w}} > 0\}}.$$

By the continuity of the product of l_q margins in $\mathbf{w}$, there exist $\epsilon > 0$ and $\delta > 0$, such that

$$\prod_{i=1}^{n} y_i \mathbf{w}^T \mathbf{x}_i \mathbb{1}_{\{y_i \mathbf{x}_i^T \mathbf{w} > 0\}} < \tilde{m} - \epsilon, \ \forall \mathbf{w} \in N_{\mathbf{w}^*} := \{\mathbf{w} : \|\mathbf{w}\|_q = 1, \ \|\mathbf{w} - \mathbf{w}^*\|_q < \delta\}.$$

There exists j such that $\frac{\hat{\boldsymbol{\beta}}_{\lambda_j}}{\|\hat{\boldsymbol{\beta}}_{\lambda_j}\|_q} \in N_{\mathbf{w}^*}$ and $\lambda_j < \delta_0$ where δ_0 is from Lemma 6. Then we have

$$\sum_{i=1}^n L\left(y_i \mathbf{x}_i^T \tilde{\mathbf{w}} \|\hat{\boldsymbol{\beta}}_{\lambda_j}\|_q\right) \leq -n \log(\|\hat{\boldsymbol{\beta}}_{\lambda_j}\|_q) - \sum_{i=1}^n \log(y_i \mathbf{x}_i^T \tilde{\mathbf{w}})$$
$$< -n \log(\|\hat{\boldsymbol{\beta}}_{\lambda_j}\|_q) - \sum_{i=1}^n \log(y_i \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_{\lambda_j} / \|\hat{\boldsymbol{\beta}}_{\lambda_j}\|_q) = \sum_{i=1}^n L\left(y_i \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_{\lambda_j}\right),$$

where the second inequality is because $\frac{\hat{\boldsymbol{\beta}}_{\lambda_j}}{\|\hat{\boldsymbol{\beta}}_{\lambda_j}\|_q} \in N_{\mathbf{w}^*}$ and $1 \leq \min y_i \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_{\lambda_j}$ by Lemma 6. Therefore $\mathbf{w}^*$ is not a convergence point of $\frac{\hat{\boldsymbol{\beta}}_\lambda}{\|\hat{\boldsymbol{\beta}}_\lambda\|_q}$, which is a contradiction.

Since $\|\frac{\hat{\boldsymbol{\beta}}_\lambda}{\|\hat{\boldsymbol{\beta}}_\lambda\|_q}\|_q = 1$, the convergence points exist. If the product of l_q margins-maximizing separating hyper-plane is unique, then we can conclude that

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

A.9 Proof of Lemma 3

Let $v(a) = \frac{a^2}{2} - L(a)$. v is strictly convex by its second order condition since,

$$0 < v''(a) = 1 - L''(a) = \begin{cases} 1 & \text{if } a \leq 1 \\ 1 - \frac{1}{a^2} & \text{if } a > 1. \end{cases}$$

Then its first order condition,

$$v(u) > v(\tilde{u}) + v'(\tilde{u})(u - \tilde{u}) \quad \forall u, \tilde{u} \in \mathbb{R}, u \neq \tilde{u},$$

directly implies Lemma 3.

A.10 Proof of Theorem 4

Given $\mathbf{X} = \mathbf{x}$, we have that

$$\begin{aligned}
\mathrm{E}[L(\mathbf{Y}\bar{f}(\mathbf{x}))|\mathbf{X} = \mathbf{x}] &= p(\mathbf{x}) L\{\bar{f}(\mathbf{x})\} + \{1 - p(\mathbf{x})\} L\{-\bar{f}(\mathbf{x})\} \\
&= \begin{cases} p(\mathbf{x}) L\left(-\frac{1-p(\mathbf{x})}{p(\mathbf{x})}\right) + \{1 - p(\mathbf{x})\} L\left(\frac{1-p(\mathbf{x})}{p(\mathbf{x})}\right) & \text{if } p(\mathbf{x}) \leq \frac{1}{2} \\ p(\mathbf{x}) L\left(\frac{p(\mathbf{x})}{1-p(\mathbf{x})}\right) + \{1 - p(\mathbf{x})\} L\left(-\frac{p(\mathbf{x})}{1-p(\mathbf{x})}\right) & \text{if } p(\mathbf{x}) > \frac{1}{2} \end{cases} \\
&= \begin{cases} p(\mathbf{x}) \left(1 + \frac{1-p(\mathbf{x})}{p(\mathbf{x})}\right) + \{1 - p(\mathbf{x})\} \left(-\log \frac{1-p(\mathbf{x})}{p(\mathbf{x})}\right) & \text{if } p(\mathbf{x}) \leq \frac{1}{2} \\ p(\mathbf{x}) \left(-\log \frac{p(\mathbf{x})}{1-p(\mathbf{x})}\right) + \{1 - p(\mathbf{x})\} \left(1 + \frac{p(\mathbf{x})}{1-p(\mathbf{x})}\right) & \text{if } p(\mathbf{x}) > \frac{1}{2} \end{cases} \\
&= \begin{cases} 1 + \{1 - p(\mathbf{x})\} \left(-\log \frac{1-p(\mathbf{x})}{p(\mathbf{x})}\right) & \text{if } p(\mathbf{x}) \leq \frac{1}{2} \\ p(\mathbf{x}) \left(-\log \frac{p(\mathbf{x})}{1-p(\mathbf{x})}\right) + 1 & \text{if } p(\mathbf{x}) > \frac{1}{2} \end{cases} \\
&= 1 - \frac{1}{2} \cdot (1 + |2p(\mathbf{x}) - 1|) \cdot \log\left(\frac{\frac{1}{2}(1 + |2p(\mathbf{x}) - 1|)}{\frac{1}{2}(1 - |2p(\mathbf{x}) - 1|)}\right).
\end{aligned}$$

The last equality is because

$$\frac{1}{2}\{1-|2p(\mathbf{x})-1|\} = \begin{cases} p(\mathbf{x}) & \text{if } p(\mathbf{x}) \le \frac{1}{2} \\ 1-p(\mathbf{x}) & \text{if } p(\mathbf{x}) > \frac{1}{2}, \end{cases}$$

$$\text{and} \quad \frac{1}{2}\{1+|2p(\mathbf{x})-1|\} = \begin{cases} 1-p(\mathbf{x}) & \text{if } p(\mathbf{x}) \le \frac{1}{2} \\ p(\mathbf{x}) & \text{if } p(\mathbf{x}) > \frac{1}{2}. \end{cases}$$

Let $a(\mathbf{x}) = |2p(\mathbf{x})-1|$. We see that

$$\begin{aligned} \mathrm{E}[L(\mathbf{Y}\bar{f}(\mathbf{X}))] &= \mathrm{E}[p(\mathbf{X})L(\bar{f}(\mathbf{X})) + (1-p(\mathbf{X}))L(-\bar{f}(\mathbf{X}))] \\ &= \mathrm{E}\left[1 - \frac{1}{2}(1+a(\mathbf{X}))\log\frac{1+a(\mathbf{X})}{1-a(\mathbf{X})}\right]. \end{aligned}$$

Since $0 \le a(\mathbf{x}) \le 1$ and $1 \le \frac{1+a(\mathbf{x})}{1-a(\mathbf{x})}$ for any $\mathbf{x}$,

$$1 - E\left[\log\frac{1+a(\mathbf{X})}{1-a(\mathbf{X})}\right] \le E[\mathbf{Y}\bar{f}(\mathbf{X})] \le 1 - \frac{1}{2}E\left[\log\frac{1+a(\mathbf{X})}{1-a(\mathbf{X})}\right],$$

which indicates that $E[L(\mathbf{Y}\bar{f}(\mathbf{X}))]$ is finite if and only if $E\left[\log\frac{1+a(\mathbf{X})}{1-a(\mathbf{X})}\right]$ is finite.

Note that $g(\delta) = \mathrm{P}(|2p(\mathbf{X})-1| \ge 1-\delta) = \mathrm{P}(p(\mathbf{X}) \le \delta/2 \text{ or } p(\mathbf{X}) \ge 1-\delta/2)$. Because,

$$\begin{aligned} \mathrm{E}\left[\log\frac{1+a(\mathbf{X})}{1-a(\mathbf{X})}\right] &= \int_0^\infty \mathrm{P}\left(\log\frac{1+a(\mathbf{X})}{1-a(\mathbf{X})} \ge t\right)dt = \int_0^\infty \mathrm{P}\left(a(\mathbf{X}) \ge 1 - \frac{2}{1+e^t}\right)dt \\ &= \int_0^\infty g\left(\frac{2}{1+e^t}\right)dt = \int_0^{\log\frac{2-\delta'}{\delta'}} g\left(\frac{2}{1+e^t}\right)dt + \int_{\log\frac{2-\delta'}{\delta'}}^\infty g\left(\frac{2}{1+e^t}\right)dt, \end{aligned}$$

and $g(\delta) \le 1$, $\mathrm{E}\left[\log\frac{1+a(\mathbf{X})}{1-a(\mathbf{X})}\right]$ is finite if and only if $\int_{\log\frac{2-\delta'}{\delta'}}^\infty g\left(\frac{2}{1+e^t}\right)dt$ is finite.

Also, because,

$$\int_0^{\delta'} g(u)\frac{1}{u}du \le \int_{\log\frac{2-\delta'}{\delta'}}^\infty g\left(\frac{2}{1+e^t}\right)dt = \int_0^{\delta'} g(u)\frac{2}{u(2-u)}du \le \frac{2}{2-\delta'}\int_0^{\delta'} g(u)\frac{1}{u}du,$$

$\int_{\log\frac{2-\delta'}{\delta'}}^\infty g\left(\frac{2}{1+e^t}\right)dt$ is finite if and only if $\int_0^{\delta'} g(u)\frac{1}{u}du$ is finite.

A.11 Proof of Lemma 4

Note that $g(\delta) = \mathrm{P}(|2p(\mathbf{X})-1| \ge 1-\delta) = \mathrm{P}(p(\mathbf{X}) \le \delta/2 \text{ or } p(\mathbf{X}) \ge 1-\delta/2)$. As $R(f) = \mathrm{E}_{\{\mathbf{X}:f(\mathbf{X})\ge 0\}}[1-p(\mathbf{X})] + \mathrm{E}_{\{\mathbf{X}:f(\mathbf{X})\le 0\}}[p(\mathbf{X})]$, we have that

$$\begin{aligned} &R(\hat{f}_n) - R(f^*) \\ &= \mathrm{E}_{\{\mathbf{X}:\hat{f}_n(\mathbf{X})\ge 0,\ f^*(\mathbf{X})<0\}}[1-2p(\mathbf{X})] + \mathrm{E}_{\{\mathbf{X}:\hat{f}_n(\mathbf{X})<0,\ f^*(\mathbf{X})\ge 0\}}[2p(\mathbf{X})-1] \\ &= \mathrm{E}_{\{\mathbf{X}:\hat{f}_n(\mathbf{X})f^*(\mathbf{X})\le 0\}}|2p(\mathbf{X})-1| \\ &= \mathrm{E}_{\{\mathbf{X}:\hat{f}_n(\mathbf{X})f^*(\mathbf{X})\le 0, |2p(\mathbf{X})-1|\ge 1-\delta\}}|2p(\mathbf{X})-1| + \mathrm{E}_{\{\mathbf{X}:\hat{f}_n(\mathbf{X})f^*(\mathbf{X})\le 0, |2p(\mathbf{X})-1|<1-\delta\}}|2p(\mathbf{X})-1| \\ &\le \mathrm{P}(\hat{f}_n(\mathbf{X})f^*(\mathbf{X}) \le 0, |2p(\mathbf{X})-1| \ge 1-\delta) + \mathrm{E}_{\{\mathbf{X}:\hat{f}_n(\mathbf{X})f^*(\mathbf{X})\le 0,\ |2p(\mathbf{X})-1|<1-\delta\}}|2p(\mathbf{X})-1|. \end{aligned}$$

Let $\zeta\{\bar{f}, \mathbf{x}\} = \mathrm{E}[L(\mathbf{Y}\bar{f}(\mathbf{x}))|\mathbf{X} = \mathbf{x}]$. Referring to the proof of Theorem 4, we have that

$$\zeta\{\bar{f}, \mathbf{x}\} = 1 - \frac{1}{2} \cdot (1 + |2p(\mathbf{x}) - 1|) \cdot \log \left(\frac{\frac{1}{2}(1 + |2p(\mathbf{x}) - 1|)}{\frac{1}{2}(1 - |2p(\mathbf{x}) - 1|)} \right).$$

We define $\gamma(a)$ for $a \in [0, 1)$ as

$$\gamma(a) = 1 - \left[1 - \frac{1}{2} \cdot (1 + a) \cdot \log \left(\frac{\frac{1}{2}(1 + a)}{\frac{1}{2}(1 - a)} \right) \right].$$

Notice that $\gamma(0) = 0$ and

$$\gamma'(a) = \frac{1}{2} + \frac{1}{2} \left(\frac{1 + a}{1 - a} + \log \frac{1 + a}{1 - a} \right) \geq 1.$$

This implies that $\gamma(a) \geq a$ for $a \in [0, 1)$. Thus, for any $\mathbf{x}$ such that $|2p(\mathbf{x}) - 1| < 1$, if we let $a = |2p(\mathbf{x}) - 1|$, we have

$$1 - \zeta\{\bar{f}, \mathbf{x}\} = \gamma(|2p(\mathbf{x}) - 1|) \geq |2p(\mathbf{x}) - 1|.$$

Therefore,

$$\begin{aligned} R(\hat{f}_n) - R(f^*) \leq\ & \mathrm{P}(\hat{f}_n(\mathbf{X})f^*(\mathbf{X}) \leq 0, |2p(\mathbf{X}) - 1| \geq 1 - \delta) \\ & + E_{\{\mathbf{X}: \hat{f}_n(\mathbf{X})f^*(\mathbf{X}) \leq 0,\ |2p(\mathbf{X}) - 1| < 1 - \delta\}}[1 - \zeta\{\bar{f}, \mathbf{X}\}]. \end{aligned}$$

The convexity of L and $\hat{f}_n(\mathbf{x})f^*(\mathbf{x}) \leq 0$ imply that

$$\zeta(\hat{f}_n, \mathbf{x}) \geq L\{(2p(\mathbf{x}) - 1)\hat{f}_n(\mathbf{x})\} \geq L(0) = 1.$$

We conclude that

$$\begin{aligned} R(\hat{f}_n) - R(f^*) \leq\ & \mathrm{P}(\hat{f}_n(\mathbf{X})f^*(\mathbf{X}) \leq 0, |2p(\mathbf{X}) - 1| \geq 1 - \delta) \\ & + E_{\{\mathbf{X}: \hat{f}_n(\mathbf{X})f^*(\mathbf{X}) \leq 0,\ |2p(\mathbf{X}) - 1| < 1 - \delta\}}[\zeta\{\hat{f}_n, \mathbf{X}\} - \zeta\{\bar{f}, \mathbf{X}\}] \\ \leq\ & \mathrm{P}(\hat{f}_n(\mathbf{X})f^*(\mathbf{X}) \leq 0, |2p(\mathbf{X}) - 1| \geq 1 - \delta) \\ & + E_{\{\mathbf{X}: |2p(\mathbf{X}) - 1| < 1 - \delta\}}[\zeta\{\hat{f}_n, \mathbf{X}\} - \zeta\{\bar{f}, \mathbf{X}\}] \ \text{ (by the definition of } \bar{f}) \\ =\ & \mathrm{P}(\hat{f}_n(\mathbf{X})f^*(\mathbf{X}) \leq 0, |2p(\mathbf{X}) - 1| \geq 1 - \delta) \\ & + E_{\{\mathbf{X}: |2p(\mathbf{X}) - 1| < 1 - \delta\}}[L\{\mathbf{Y}\hat{f}_n(\mathbf{X})\} - L\{\mathbf{Y}\bar{f}(\mathbf{X})\}]. \end{aligned}$$

A.12 Lemma 7 and its proof

We introduce a lemma before proving Theorem 5.

Lemma 7 *Suppose the input space $\mathcal{X}$ is compact and $\mathcal{H}_K$ is the RKHS induced by a universal kernel K on $\mathcal{X}$. For any $\delta \in (0, 1]$ and $\epsilon > 0$, there exists $\bar{f}_\epsilon \in \mathcal{H}_K$ such that $\sup_{\mathbf{x} \in \mathcal{X}} |\bar{f}_\epsilon(\mathbf{x})| \leq \frac{2-\delta}{\delta} + \frac{\epsilon}{2}$ and*

$$\left| \mathrm{E}\left[\mathbb{1}\{|2p(\mathbf{X}) - 1| < 1 - \delta\} \left\{ L(\mathbf{Y}\bar{f}_\epsilon(\mathbf{X})) - L(\mathbf{Y}\bar{f}(\mathbf{X})) \right\} \right] \right| < \epsilon.$$

Furthermore, there exists a continuous function τ_ϵ such that

$$\sup_{\mathbf{x}} |\bar{f}_\epsilon(\mathbf{x}) - \tau_\epsilon(\mathbf{x})| < \epsilon/2 \text{ and } \mathrm{P}\{\mathrm{sgn}(\tau_\epsilon(\mathbf{X})) \neq \mathrm{sgn}(\bar{f}(\mathbf{X}))\} \leq \epsilon/4 \cdot \delta/(2-\delta).$$

Proof Let $f_\delta(\mathbf{x})$ be a truncated function of $\bar{f}$:

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

By Lusin's theorem, there exists a continuous function $\varrho(\mathbf{x})$ such that $\mathrm{P}\{\varrho(\mathbf{X}) \neq f_\delta(\mathbf{X})\} \leq \epsilon/4 \cdot \delta/(2-\delta)$. Let

$$\tau(\mathbf{x}) = \begin{cases} \varrho(\mathbf{x}), & \text{if } |\varrho(\mathbf{x})| \leq \frac{2-\delta}{\delta}, \\ \frac{2-\delta}{\delta}\frac{\varrho(\mathbf{x})}{|\varrho(\mathbf{x})|}, & \text{if } |\varrho(\mathbf{x})| > \frac{2-\delta}{\delta}. \end{cases}$$

Since $\sup_{\mathbf{x}} |f_\delta(\mathbf{x})| \leq \frac{2-\delta}{\delta}$, $\mathrm{P}\{\tau(\mathbf{X}) \neq f_\delta(\mathbf{X})\} \leq \epsilon/4 \cdot \delta/(2-\delta)$. Hence,

$$\begin{aligned} &|\mathrm{E}\left[\mathbb{1}\left\{|2p(\mathbf{X})-1| < 1-\delta\right\}\left\{L(\mathbf{Y}\tau(\mathbf{X})) - L(\mathbf{Y}f_\delta(\mathbf{X}))\right\}\right]| \\ &\leq E_{\{\mathbf{X}:\tau(\mathbf{X})\neq f_\delta(\mathbf{X})\}}|f_\delta(\mathbf{X}) - \tau(\mathbf{X})| \\ &\leq 2\frac{2-\delta}{\delta}\frac{\epsilon}{4}\frac{\delta}{2-\delta} = \epsilon/2, \end{aligned}$$

where the first inequality comes from the fact that

$$|L(u_1) - L(u_2)| \leq |u_1 - u_2|, \ \forall u_1, u_2 \in \mathbb{R}.$$

By the definition of a universal kernel and the continuity of a function $\tau(\mathbf{x})$, there exists a function $\bar{f}_\epsilon \in \mathcal{H}_K$ such that

$$\sup_{\mathbf{x}} |\bar{f}_\epsilon(\mathbf{x}) - \tau(\mathbf{x})| < \epsilon/2.$$

Combining the above together, we obtain that

$$\begin{aligned} &\left|\mathrm{E}\left[\mathbb{1}\left\{|2p(\mathbf{X})-1| < 1-\delta\right\}\left\{L(\mathbf{Y}\bar{f}_\epsilon(\mathbf{X})) - L(\mathbf{Y}\bar{f}(\mathbf{X}))\right\}\right]\right| \\ =&\left|\mathrm{E}\left[\mathbb{1}\left\{|2p(\mathbf{X})-1| < 1-\delta\right\}\left\{L(\mathbf{Y}\bar{f}_\epsilon(\mathbf{X})) - L(\mathbf{Y}f_\delta(\mathbf{X}))\right\}\right]\right| \\ \leq&\left|\mathrm{E}\left[\mathbb{1}\left\{|2p(\mathbf{X})-1| < 1-\delta\right\}\left\{L(\mathbf{Y}\bar{f}_\epsilon(\mathbf{X})) - L(\mathbf{Y}\tau(\mathbf{X}))\right\}\right]\right| \\ &+ |\mathrm{E}\left[\mathbb{1}\left\{|2p(\mathbf{X})-1| < 1-\delta\right\}\left\{L(\mathbf{Y}\tau(\mathbf{X})) - L(\mathbf{Y}f_\delta(\mathbf{X}))\right\}\right]| \leq \epsilon. \end{aligned}$$

■

A.13 Proof of Theorem 5

Let $B := \sup_{\mathbf{x},\mathbf{y}\in\mathcal{X}} K(\mathbf{x},\mathbf{y})$ and $C := \inf_{\mathbf{x},\mathbf{y}\in\mathcal{X}} K(\mathbf{x},\mathbf{y})$. Set $\delta_n = \inf\{\delta : \lambda_n \le \frac{C\delta^2 g(\delta)}{(2-\delta)^2}\}$. Note that as $\lambda_n \to 0$, $\delta_n \to 0$. Also, by the Stolz-Cesàro theorem, if $\lambda_{n+1}^{-1} - \lambda_n^{-1} \to 0$, $(n\lambda_n)^{-1} \to 0$. Let N be such that $\delta_N < \delta'$ and $B \cdot (2\lambda_N)^{-1} \ge 1$.

Fix any $n > N$. By Lemma 4,

$$\begin{aligned} R(\hat{f}_n) - R(f^*) \le\ & \mathrm{P}[\mathrm{sgn}\{f^*(\mathbf{X})\} \ne \mathrm{sgn}\{\hat{f}_n(\mathbf{X})\}, |2p(\mathbf{X}) - 1| \ge 1 - \delta_n] \\ & + \mathrm{E}\left[\mathbb{1}\left\{|2p(\mathbf{X}) - 1| < 1 - \delta_n\right\}\left(L(\mathbf{Y}\hat{f}_n(\mathbf{X})) - L(\mathbf{Y}\bar{f}(\mathbf{X}))\right)\right] \\ \le\ & g(\delta_n) + \mathrm{E}\left[\mathbb{1}\left\{|2p(\mathbf{X}) - 1| < 1 - \delta_n\right\}\left(L(\mathbf{Y}\hat{f}_n(\mathbf{X})) - L(\mathbf{Y}\bar{f}(\mathbf{X}))\right)\right], \end{aligned}$$

where the last inequality is by the definition of g.

Let $\mathbf{T}_n = \{(\mathbf{X}_k, \mathbf{Y}_k)\}_{k=1}^n$, where each random pair has the same distribution to that of $(\mathbf{X}, \mathbf{Y})$. To prove the theorem, it is enough to show that

$$\lim_{n\to\infty} \mathrm{E}_{\mathbf{T}_{n-1}}\left[\mathrm{E}\left[\mathbb{1}\left\{|2p(\mathbf{X}) - 1| < 1 - \delta_n\right\}\left(L(\mathbf{Y}\hat{f}_{n-1}(\mathbf{X})) - L(\mathbf{Y}\bar{f}(\mathbf{X}))\right)\right]\right] = 0,$$

by Markov's inequality and the assumption implying that $g(\delta_n) \to 0$ as $\delta_n \to 0$.

We define $\hat{f}^{[k]}$ as the solution of (13) where the k-th datum is excluded from the training data,

$$\hat{f}^{[k]} = \operatorname*{argmin}_{f\in\mathcal{H}_K}\left[\frac{1}{n-1}\sum_{i=1, i\ne k}^{n} L\left(y_i f(\mathbf{x}_i)\right) + \lambda_{n-1}||f||_{\mathcal{H}_K}^2\right].$$

We have

$$\begin{aligned} 0 &\le \frac{1}{n-1}\sum_{i=1,i\ne k}^{n} L\left(y_i\hat{f}_n(\mathbf{x}_i)\right) + \lambda_{n-1}||\hat{f}_n||_{\mathcal{H}_K}^2 - \frac{1}{n-1}\sum_{i=1,i\ne k}^{n} L\left(y_i\hat{f}^{[k]}(\mathbf{x}_i)\right) - \lambda_{n-1}||\hat{f}^{[k]}||_{\mathcal{H}_K}^2 \\ &\le -\frac{1}{n-1}\sum_{i=1,i\ne k}^{n} L'\left(y_i\hat{f}_n(\mathbf{x}_i)\right) y_i\left\{\hat{f}^{[k]}(\mathbf{x}_i) - \hat{f}_n(\mathbf{x}_i)\right\} + \lambda_{n-1}||\hat{f}_n||_{\mathcal{H}_K}^2 - \lambda_{n-1}||\hat{f}^{[k]}||_{\mathcal{H}_K}^2 \\ &= -\frac{1}{n-1}\sum_{i=1,i\ne k}^{n} L'\left(y_i\hat{f}_n(\mathbf{x}_i)\right) y_i\left\langle K(\mathbf{x}_i,\cdot), \hat{f}^{[k]} - \hat{f}_n\right\rangle + \lambda_{n-1}||\hat{f}_n||_{\mathcal{H}_K}^2 - \lambda_{n-1}||\hat{f}^{[k]}||_{\mathcal{H}_K}^2 \\ &= -\frac{1}{n-1}\sum_{i=1,i\ne k}^{n} L'\left(y_i\hat{f}_n(\mathbf{x}_i)\right) y_i\left\langle K(\mathbf{x}_i,\cdot), \hat{f}^{[k]} - \hat{f}_n\right\rangle - 2\lambda_{n-1}\left\langle \hat{f}_n, \hat{f}^{[k]} - \hat{f}_n\right\rangle - \lambda_{n-1}||\hat{f}^{[k]} - \hat{f}_n||_{\mathcal{H}_K}^2, \end{aligned}$$

where the first inequality is by the definition of $\hat{f}^{[k]}$, the second inequality is by the convexity of L, and the third equality is by the reproducing property.

By the KKT condition and the representer theorem,

$$\hat{f}_n(\mathbf{x}) = -\frac{1}{2n\lambda_n}\sum_{i=1}^{n} L'\left(y_i\hat{f}_n(\mathbf{x}_i)\right) y_i K(\mathbf{x}_i, \mathbf{x}), \tag{18}$$

assuming that $\mathbf{K}$ is invertible. Thus, we further have

$$\begin{aligned}
\lambda_{n-1}||\hat{f}^{[k]} - \hat{f}_n||^2_{\mathcal{H}_K} &\leq -\frac{1}{n-1}\sum_{i=1, i\neq k}^{n} L'\left(y_i \hat{f}_n(\mathbf{x}_i)\right) y_i \left\langle K(\mathbf{x}_i, \cdot), \hat{f}^{[k]} - \hat{f}_n \right\rangle \\
&\quad + \frac{\lambda_{n-1}}{n\lambda_n}\sum_{i=1}^{n} L'\left(y_i \hat{f}_n(\mathbf{x}_i)\right) y_i \left\langle K(\mathbf{x}_i, \cdot), \hat{f}^{[k]} - \hat{f}_n \right\rangle \\
&\leq \left|\frac{\lambda_{n-1}}{n\lambda_n} - \frac{1}{n-1}\right| \sum_{i=1, i\neq k}^{n} \left|L'\left(y_i \hat{f}_n(\mathbf{x}_i)\right)\right| \left|\left\langle K(\mathbf{x}_i, \cdot), \hat{f}^{[k]} - \hat{f}_n \right\rangle\right| \\
&\quad + \frac{\lambda_{n-1}}{n\lambda_n} \left|L'\left(y_k \hat{f}_n(\mathbf{x}_k)\right)\right| \left|\left\langle K(\mathbf{x}_k, \cdot), \hat{f}^{[k]} - \hat{f}_n \right\rangle\right| \\
&\leq \left|\frac{\lambda_{n-1}}{n\lambda_n} - \frac{1}{n-1}\right| \sum_{i=1, i\neq k}^{n} \|K(\mathbf{x}_i, \cdot)\|_{\mathcal{H}_K} \|\hat{f}^{[k]} - \hat{f}_n\|_{\mathcal{H}_K} \\
&\quad + \frac{\lambda_{n-1}}{n\lambda_n} \|K(\mathbf{x}_k, \cdot)\|_{\mathcal{H}_K} \|\hat{f}^{[k]} - \hat{f}_n\|_{\mathcal{H}_K} \\
&\leq \left(\left|\frac{n-1}{n}\frac{\lambda_{n-1}}{\lambda_n} - 1\right| + \frac{\lambda_{n-1}}{n\lambda_n}\right) \sqrt{\sup K(\mathbf{x}, \mathbf{x})} \|\hat{f}^{[k]} - \hat{f}_n\|_{\mathcal{H}_K},
\end{aligned}$$

where the third inequality is by the Cauchy Schwarz inequality. It gives us

$$\begin{aligned}
||\hat{f}^{[k]} - \hat{f}_n||_{\mathcal{H}_K} &\leq \left(\left|\frac{n-1}{n\lambda_n} - \frac{1}{\lambda_{n-1}}\right| + \frac{1}{n\lambda_n}\right)\sqrt{B} \\
&\leq \left(\left|\frac{1}{\lambda_n} - \frac{1}{\lambda_{n-1}}\right| + \frac{2}{n\lambda_n}\right)\sqrt{B}.
\end{aligned}$$

Then, we have

$$\begin{aligned}
L\left(y_k \hat{f}^{[k]}(\mathbf{x}_k)\right) - L\left(y_k \hat{f}_n(\mathbf{x}_k)\right) &\leq \left|\hat{f}^{[k]}(\mathbf{x}_k) - \hat{f}_n(\mathbf{x}_k)\right| \\
&\leq \left|\left\langle K(\mathbf{x}_k, \cdot), \hat{f}^{[k]} - \hat{f}_n \right\rangle_{\mathcal{H}_K}\right| \\
&\leq \sqrt{B} \cdot ||\hat{f}^{[k]} - \hat{f}_n||_{\mathcal{H}_K} \\
&\leq B \cdot \left(\left|\frac{1}{\lambda_n} - \frac{1}{\lambda_{n-1}}\right| + \frac{2}{n\lambda_n}\right),
\end{aligned}$$

for $k = 1, \ldots, n$, where the first inequality is by the Lipschitz continuity of the LHS loss function and the second inequality is by reproducing property.

We see that

$$
\begin{aligned}
& \mathrm{E}_{\mathbf{T}_{n-1}}\left[\mathrm{E}\left[\mathbb{1}\left\{|2p(\mathbf{X})-1|<1-\delta_n\right\} L\left(\mathbf{Y}\hat{f}_{n-1}(\mathbf{X})\right)\right]\right] \\
= & \frac{1}{n}\sum_{k=1}^{n}\mathrm{E}_{\mathbf{T}_n}\left[\mathbb{1}\left\{|2p(\mathbf{X}_k)-1|<1-\delta_n\right\} L\left(\mathbf{Y}_k\hat{f}^{[k]}(\mathbf{X}_k)\right)\right] \\
= & \mathrm{E}_{\mathbf{T}_n}\left[\frac{1}{n}\sum_{k=1}^{n}\mathbb{1}\left\{|2p(\mathbf{X}_k)-1|<1-\delta_n\right\} L\left(\mathbf{Y}_k\hat{f}^{[k]}(\mathbf{X}_k)\right)\right] \\
\leq & \mathrm{E}_{\mathbf{T}_n}\left[\frac{1}{n}\sum_{k=1}^{n}\mathbb{1}\left\{|2p(\mathbf{X}_k)-1|<1-\delta_n\right\}\left\{L\left(\mathbf{Y}_k\hat{f}_n(\mathbf{X}_k)\right)+B\cdot\left(\left|\frac{1}{\lambda_n}-\frac{1}{\lambda_{n-1}}\right|+\frac{2}{n\lambda_n}\right)\right\}\right] \\
& +\mathrm{E}_{\mathbf{T}_n}\left[\frac{1}{n}\sum_{k=1}^{n}\mathbb{1}\left\{|2p(\mathbf{X}_k)-1|\geq 1-\delta_n\right\} L\left(\mathbf{Y}_k\hat{f}_n(\mathbf{X}_k)\right)\right] \\
& \quad -\mathrm{E}_{\mathbf{T}_n}\left[\frac{1}{n}\sum_{k=1}^{n}\mathbb{1}\left\{|2p(\mathbf{X}_k)-1|\geq 1-\delta_n\right\} L\left(\mathbf{Y}_k\hat{f}_n(\mathbf{X}_k)\right)\right],
\end{aligned}
$$

where the first and the second inequalities are because $(\mathbf{X},\mathbf{Y})$ and $(\mathbf{X}_k,\mathbf{Y}_k)$ have the same distribution for any $k=1,\ldots,n$.

From Lemma 7, there is $\bar{f}_{\delta_n}\in\mathcal{H}_K$ such that $\sup_{\mathbf{x}\in\mathcal{X}}|\bar{f}_{\delta_n}(\mathbf{x})|\leq\frac{2-\delta_n}{\delta_n}+\frac{\delta_n}{2}$ and

$$
\left|\mathrm{E}\left[\mathbb{1}\left\{|2p(\mathbf{X})-1|<1-\delta_n\right\}\left\{L(\mathbf{Y}\bar{f}_{\delta_n}(\mathbf{X}))-L(\mathbf{Y}\bar{f}(\mathbf{X}))\right\}\right]\right|<\delta_n.
$$

Also, there exists a function τ_{δ_n} such that

$$
\sup_{\mathbf{x}}|\bar{f}_{\delta_n}(\mathbf{x})-\tau_{\delta_n}(\mathbf{x})|<\delta_n/2 \textit{ and } \mathrm{P}\{\mathrm{sgn}(\tau_{\delta_n}(\mathbf{X}))\neq\mathrm{sgn}(\bar{f}(\mathbf{X}))\}\leq\delta_n/4\cdot\delta_n/(2-\delta_n).
$$

By the definition of $\hat{f}_n$, we have

$$
\frac{1}{n}\sum_{i=1}^{n}L\left(y_i\hat{f}_n(\mathbf{x}_i)\right)+\lambda_n||\hat{f}_n||^2_{\mathcal{H}_K}\leq\frac{1}{n}\sum_{i=1}^{n}L\left(y_i\bar{f}_{\delta_n}(\mathbf{x}_i)\right)+\lambda_n||\bar{f}_{\delta_n}||^2_{\mathcal{H}_K}.
$$

We further have that,

$$
\begin{aligned}
& \mathrm{E}_{\mathbf{T}_{n-1}}\left[\mathrm{E}\left[\mathbb{1}\left\{|2p(\mathbf{X})-1|<1-\delta_n\right\} L\left(\mathbf{Y}\hat{f}_{n-1}(\mathbf{X})\right)\right]\right] \\
\leq\ & \mathrm{E}_{\mathbf{T}_n}\left[\frac{1}{n}\sum_{k=1}^{n}\mathbb{1}\left\{|2p(\mathbf{X}_k)-1|<1-\delta_n\right\}\left\{L\left(\mathbf{Y}_k\bar{f}_{\delta_n}(\mathbf{X}_k)\right)+B\cdot\left(\left|\frac{1}{\lambda_n}-\frac{1}{\lambda_{n-1}}\right|+\frac{2}{n\lambda_n}\right)\right\}\right] \\
& + \mathrm{E}_{\mathbf{T}_n}\left[\frac{1}{n}\sum_{k=1}^{n}\mathbb{1}\left\{|2p(\mathbf{X}_k)-1|\geq 1-\delta_n\right\}\left\{L\left(\mathbf{Y}_k\bar{f}_{\delta_n}(\mathbf{X}_k)\right)-L\left(\mathbf{Y}_k\tau_{\delta_n}(\mathbf{X}_k)\right)+L\left(\mathbf{Y}_k\tau_{\delta_n}(\mathbf{X}_k)\right)\right\}\right] \\
& \quad + \lambda_n\|\bar{f}_{\delta_n}\|_{\mathcal{H}_K}^2-\mathrm{E}_{\mathbf{T}_n}\left[\frac{1}{n}\sum_{k=1}^{n}\mathbb{1}\left\{|2p(\mathbf{X}_k)-1|\geq 1-\delta_n\right\} L\left(\mathbf{Y}_k\hat{f}_n(\mathbf{X}_k)\right)\right] \\
\leq\ & \mathrm{E}_{\mathbf{T}_n}\left[\frac{1}{n}\sum_{k=1}^{n}\mathbb{1}\left\{|2p(\mathbf{X}_k)-1|<1-\delta_n\right\} L\left(\mathbf{Y}_k\bar{f}_{\delta_n}(\mathbf{X}_k)\right)\right]+B\cdot\left(\left|\frac{1}{\lambda_n}-\frac{1}{\lambda_{n-1}}\right|+\frac{2}{n\lambda_n}\right) \\
& + \mathrm{E}_{\mathbf{T}_n}\left[\frac{1}{n}\sum_{k=1}^{n}\mathbb{1}\left\{|2p(\mathbf{X}_k)-1|\geq 1-\delta_n\right\}\frac{\delta_n}{2}\right] \\
& \quad + \mathrm{E}_{\mathbf{T}_n}\left[\frac{1}{n}\sum_{k=1}^{n}\mathbb{1}\left\{|2p(\mathbf{X}_k)-1|\geq 1-\delta_n\right\} L\left(\mathbf{Y}_k\tau_{\delta_n}(\mathbf{X}_k)\right)\right] \\
& \qquad + \frac{\lambda_n}{C}\left(\frac{2-\delta_n}{\delta_n}+\frac{\delta_n}{2}\right)^2+\mathrm{E}_{\mathbf{T}_n}\left[\frac{1}{n}\sum_{k=1}^{n}\mathbb{1}\left\{|2p(\mathbf{X}_k)-1|\geq 1-\delta_n\right\}\log\left(\frac{B}{2\lambda_n}\right)\right] \\
\leq\ & \mathrm{E}_{\mathbf{T}_{n-1}}\left[\mathrm{E}\left[\mathbb{1}\left\{|2p(\mathbf{X})-1|<1-\delta_n\right\} L\left(\mathbf{Y}\bar{f}_{\delta_n}(\mathbf{X})\right)\right]\right]+B\cdot\left(\left|\frac{1}{\lambda_n}-\frac{1}{\lambda_{n-1}}\right|+\frac{2}{n\lambda_n}\right) \\
& + g(\delta_n)\delta_n/2+g(\delta_n)+\frac{\delta_n^2}{4(2-\delta_n)}\left(1+\frac{\delta_n}{2}+\frac{2-\delta_n}{\delta_n}+\frac{\delta_n}{2}\right) \\
& \quad + \frac{\lambda_n}{C}\left(\frac{2-\delta_n}{\delta_n}+\frac{\delta_n}{2}\right)^2+g(\delta_n)\log\left(\frac{B}{2\lambda_n}\right) \\
\leq\ & \mathrm{E}_{\mathbf{T}_{n-1}}\left[\mathrm{E}\left[\mathbb{1}\left\{|2p(\mathbf{X})-1|<1-\delta_n\right\} L\left(\mathbf{Y}\bar{f}(\mathbf{X})\right)\right]\right]+\delta_n+B\cdot\left(\left|\frac{1}{\lambda_n}-\frac{1}{\lambda_{n-1}}\right|+\frac{2}{n\lambda_n}\right) \\
& + g(\delta_n)\delta_n/2+g(\delta_n)+\frac{\delta_n^2}{4(2-\delta_n)}\left(1+\frac{\delta_n}{2}+\frac{2-\delta_n}{\delta_n}+\frac{\delta_n}{2}\right) \\
& \quad + \frac{\lambda_n}{C}\left\{\left(\frac{2-\delta_n}{\delta_n}\right)^2+2-\delta_n+\frac{{\delta_n}^2}{4}\right\}+g(\delta_n)\log(B/2)-g(\delta_n)\log(\lambda_n) \\
=\ & \mathrm{E}_{\mathbf{T}_{n-1}}\left[\mathrm{E}\left[\mathbb{1}\left\{|2p(\mathbf{X})-1|<1-\delta_n\right\} L\left(\mathbf{Y}\bar{f}(\mathbf{X})\right)\right]\right]+\delta_n+B\cdot\left(\left|\frac{1}{\lambda_n}-\frac{1}{\lambda_{n-1}}\right|+\frac{2}{n\lambda_n}\right) \\
& + g(\delta_n)\delta_n/2+g(\delta_n)+\frac{\delta_n^2}{4(2-\delta_n)}\left(1+\frac{\delta_n}{2}+\frac{2-\delta_n}{\delta_n}+\frac{\delta_n}{2}\right) \\
& \quad + g(\delta_n)+\frac{g(\delta_n)\delta_n^2}{(2-\delta_n)^2}\left\{2-\delta_n+\frac{{\delta_n}^2}{4}\right\}+g(\delta_n)\log(B/2) \\
& \qquad - g(\delta_n)\left\{\log(g(\delta_n))+\log(C)+2\log(\delta_n)-2\log(2-\delta_n)\right\},
\end{aligned}
$$

where the second inequality is because $|\hat{f}_n(\mathbf{x})| \leq \frac{B}{2\lambda_n}$ from (18) and $\|\bar{f}_{\delta_n}\|^2_{\mathcal{H}_K} \leq \frac{1}{C}\left(\frac{2-\delta_n}{\delta_n} + \frac{\delta_n}{2}\right)^2$ by Theorem 3.11 of Paulsen and Raghupathi (2016). The last equality is because $\lambda_n = \frac{C\delta_n^2 g(\delta_n)}{(2-\delta_n)^2}$ for a large enough n such that $\delta_n < \delta'$ following from Assumption 1 and the definition of δ_n.

Hence, for any $n > N$,

$$
\begin{aligned}
0 &\leq \mathrm{E}_{\mathbf{T}_{n-1}}\left[\mathrm{E}\left[\mathbb{1}\left\{|2p(\mathbf{X})-1| < 1-\delta_n\right\}\left\{L(\mathbf{Y}\hat{f}_{n-1}(\mathbf{X})) - L\left(\mathbf{Y}\bar{f}(\mathbf{X})\right)\right\}\right]\right] \\
&\leq \delta_n + B\cdot\left(\left|\frac{1}{\lambda_n} - \frac{1}{\lambda_{n-1}}\right| + \frac{2}{n\lambda_n}\right) + g(\delta_n)\delta_n/2 + g(\delta_n) + \frac{\delta_n^2}{4(2-\delta_n)}\left(1+\delta_n+\frac{2-\delta_n}{\delta_n}\right) \\
&\quad + g(\delta_n) + \frac{g(\delta_n)\delta_n^2}{(2-\delta_n)^2}\left\{2-\delta_n+\frac{{\delta_n}^2}{4}\right\} + g(\delta_n)\log(B/2) \\
&\quad - g(\delta_n)\left\{\log(g(\delta_n)) + \log(C) + 2\log(\delta_n) - 2\log(2-\delta_n)\right\}.
\end{aligned}
$$

As the right-hand side goes to 0 as $n \to \infty$, we prove the theorem.

A.14 Proof of Theorem 6

Fix $\mathbf{x}$ and let $\alpha = f(\mathbf{x})$. We have

$$
\begin{aligned}
\mathrm{E}[L_r(\mathbf{Y}\alpha)|\mathbf{X}=\mathbf{x}] &= p(\mathbf{x})L_r(\alpha) + \{1-p(\mathbf{x})\}L_r(-\alpha) \\
&= \begin{cases} p(\mathbf{x})r\left(1-\alpha^{1/r}\right) + (1-p(\mathbf{x}))(1+\alpha), & \text{if } \alpha > 1, \\ p(\mathbf{x})(1-\alpha) + (1-p(\mathbf{x}))(1+\alpha), & \text{if } -1 \leq \alpha \leq 1, \\ p(\mathbf{x})(1-\alpha) + (1-p(\mathbf{x}))r\left(1-(-\alpha)^{1/r}\right), & \text{if } \alpha < -1. \end{cases}
\end{aligned}
$$

If $0 < p(\mathbf{x}) < 1$, the global minimizer exists because $\mathrm{E}[L_r(\mathbf{Y}\alpha)|\mathbf{X}=\mathbf{x}]$ is coercive:

$$
\mathrm{E}[L_r(\mathbf{Y}\alpha)|\mathbf{X}=\mathbf{x}] \to \infty \quad \text{as } |\alpha| \to \infty.
$$

$\frac{\partial}{\partial\alpha}\mathrm{E}[L_r(\mathbf{Y}\alpha)|\mathbf{X}=\mathbf{x}] = 0$ holds if and only if

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

α is the unique minimizer because $\mathrm{E}[L_r(\mathbf{Y}\alpha)|\mathbf{X}=\mathbf{x}]$ is convex.

A.15 Proof of Theorem 7

Given $\mathbf{X} = \mathbf{x}$, let $\zeta\{\bar{f}, \mathbf{x}\} = \mathrm{E}[L_r(\mathbf{Y}\bar{f}(\mathbf{x}))|\mathbf{X} = \mathbf{x}]$. We see that

$$
\begin{aligned}
\zeta\{\bar{f}, \mathbf{x}\} &= p(\mathbf{x})L_r\{\bar{f}(\mathbf{x})\} + \{1 - p(\mathbf{x})\}L_r\{-\bar{f}(\mathbf{x})\} \\
&= \begin{cases} p(\mathbf{x})L_r\left(-\left(\frac{1-p(\mathbf{x})}{p(\mathbf{x})}\right)^{\frac{r}{r-1}}\right) + \{1 - p(\mathbf{x})\}L_r\left(\left(\frac{1-p(\mathbf{x})}{p(\mathbf{x})}\right)^{\frac{r}{r-1}}\right) & \text{if } p(\mathbf{x}) \le \frac{1}{2} \\ p(\mathbf{x})L_r\left(\left(\frac{p(\mathbf{x})}{1-p(\mathbf{x})}\right)^{\frac{r}{r-1}}\right) + \{1 - p(\mathbf{x})\}L_r\left(-\left(\frac{p(\mathbf{x})}{1-p(\mathbf{x})}\right)^{\frac{r}{r-1}}\right) & \text{if } p(\mathbf{x}) > \frac{1}{2} \end{cases} \\
&= \begin{cases} p(\mathbf{x})\left(1 + \left(\frac{1-p(\mathbf{x})}{p(\mathbf{x})}\right)^{\frac{r}{r-1}}\right) + \{1 - p(\mathbf{x})\}\left(r - r\left(\frac{1-p(\mathbf{x})}{p(\mathbf{x})}\right)^{\frac{1}{r-1}}\right) & \text{if } p(\mathbf{x}) \le \frac{1}{2} \\ p(\mathbf{x})\left(r - r\left(\frac{p(\mathbf{x})}{1-p(\mathbf{x})}\right)^{\frac{1}{r-1}}\right) + \{1 - p(\mathbf{x})\}\left(1 + \left(\frac{p(\mathbf{x})}{1-p(\mathbf{x})}\right)^{\frac{r}{r-1}}\right) & \text{if } p(\mathbf{x}) > \frac{1}{2} \end{cases} \\
&= \begin{cases} p(\mathbf{x}) - (r-1)\left(\frac{(1-p(\mathbf{x}))^r}{p(\mathbf{x})}\right)^{1/(r-1)} + r\{1 - p(\mathbf{x})\} & \text{if } p(\mathbf{x}) \le \frac{1}{2} \\ \{1 - p(\mathbf{x})\} - (r-1)\left(\frac{p(\mathbf{x})^r}{1-p(\mathbf{x})}\right)^{1/(r-1)} + rp(\mathbf{x}) & \text{if } p(\mathbf{x}) > \frac{1}{2} \end{cases} \\
&= \frac{1}{2}\{1 - |2p(\mathbf{x}) - 1|\} - \frac{r-1}{2}\frac{\{1 + |2p(\mathbf{x}) - 1|\}^{r/(r-1)}}{\{1 - |2p(\mathbf{x}) - 1|\}^{1/(r-1)}} + \frac{r}{2}\{1 + |2p(\mathbf{x}) - 1|\} \\
&= -\frac{r-1}{2}\frac{\{1 + |2p(\mathbf{x}) - 1|\}^{r/(r-1)}}{\{1 - |2p(\mathbf{x}) - 1|\}^{1/(r-1)}} + 1 + \frac{r-1}{2}\{1 + |2p(\mathbf{x}) - 1|\} \\
&= 1 - \frac{r-1}{2}\{1 + |2p(\mathbf{x}) - 1|\}\left[\left(\frac{1 + |2p(\mathbf{x}) - 1|}{1 - |2p(\mathbf{x}) - 1|}\right)^{1/(r-1)} - 1\right].
\end{aligned}
$$

We further see that

$$1 - \zeta\{\bar{f}, \mathbf{x}\} \ge |2p(\mathbf{x}) - 1|.$$

Following the similar procedure of the proof of Lemma 4, for any $0 < \delta \le 1$, we have that

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

Therefore, to prove the theorem, it is enough to show that there exists a sequence of δ_n such that

$$\lim_{n\to\infty} \mathrm{E}_{\mathbf{T}_{n-1}}\left[\mathrm{E}\left[\mathbb{1}\{|2p(\mathbf{X}) - 1| < 1 - \delta_n\}\left(L_r(\mathbf{Y}\hat{f}_{n-1}(\mathbf{X})) - L_r(\mathbf{Y}\bar{f}(\mathbf{X}))\right)\right]\right] = 0.$$

Here, $\mathbf{T}_n = \{(\mathbf{X}_k, \mathbf{Y}_k)\}_{k=1}^n$ where each random pair has the same distribution as that of $(\mathbf{X}, \mathbf{Y})$.

Let $B := \sup_{\mathbf{x},\mathbf{y}\in\mathcal{X}} K(\mathbf{x}, \mathbf{y})$ and $C := \inf_{\mathbf{x},\mathbf{y}\in\mathcal{X}} K(\mathbf{x}, \mathbf{y})$. We set $\delta_n = \inf\{\delta : \lambda_n \le Cg(\delta)\,(\delta/(2 - \delta))^{2r/(r-1)}\}$. Note that as $\lambda_n \to 0$, $\delta_n \to 0$. Also, by the Stolz-Cesàro theorem, if $\lambda_{n+1}^{-1} - \lambda_n^{-1} \to 0$, $(n\lambda_n)^{-1} \to 0$. Let N be such that $\delta_N < \delta'$ and $B \cdot (2\lambda_N)^{-1} \ge 1$.

Following the same procedure in the proof of Theorem 5, thanks to the fact that L is convex and Lipschitz with the Lipschitz constant being 1, we arrive at

$$\begin{aligned}
&\mathrm{E}_{\mathbf{T}_{n-1}}\left[\mathrm{E}\left[\mathbb{1}\left\{|2p(\mathbf{X})-1|<1-\delta_n\right\} L_r\left(\mathbf{Y}\hat{f}_{n-1}(\mathbf{X})\right)\right]\right] \\
\leq\ & \mathrm{E}_{\mathbf{T}_n}\left[\frac{1}{n}\sum_{k=1}^{n}\mathbb{1}\left\{|2p(\mathbf{X}_k)-1|<1-\delta_n\right\}\left\{L_r\left(\mathbf{Y}_k\hat{f}_n(\mathbf{X}_k)\right)+B\cdot\left(\left|\frac{1}{\lambda_n}-\frac{1}{\lambda_{n-1}}\right|+\frac{2}{n\lambda_n}\right)\right\}\right] \\
&+\mathrm{E}_{\mathbf{T}_n}\left[\frac{1}{n}\sum_{k=1}^{n}\mathbb{1}\left\{|2p(\mathbf{X}_k)-1|\geq 1-\delta_n\right\} L_r\left(\mathbf{Y}_k\hat{f}_n(\mathbf{X}_k)\right)\right] \\
&\quad-\mathrm{E}_{\mathbf{T}_n}\left[\frac{1}{n}\sum_{k=1}^{n}\mathbb{1}\left\{|2p(\mathbf{X}_k)-1|\geq 1-\delta_n\right\} L_r\left(\mathbf{Y}_k\hat{f}_n(\mathbf{X}_k)\right)\right].
\end{aligned}$$

Also, following the similar logic to the proof of Lemma 7, as L_r is Lipschitz with the Lipschitz constant being 1, for any $\delta\in(0,1]$, we see that there exists $\bar{f}_\delta\in\mathcal{H}_K$ such that $\sup_{\mathbf{x}\in\mathcal{X}}|\bar{f}_\delta(\mathbf{x})|\leq\left(\frac{2-\delta}{\delta}\right)^{r/(r-1)}+\frac{\delta}{2}$ and,

$$\left|\mathrm{E}\left[\mathbb{1}\left\{|2p(\mathbf{X})-1|<1-\delta\right\}\left\{L_r(\mathbf{Y}\bar{f}_\delta(\mathbf{X}))-L_r(\mathbf{Y}\bar{f}(\mathbf{X}))\right\}\right]\right|<\delta.$$

Also, there exists a continuous function τ_δ such that

$$\sup_{\mathbf{x}}|\bar{f}_\delta(\mathbf{x})-\tau_\delta(\mathbf{x})|<\delta/2 \textit{ and } \mathrm{P}\{\mathrm{sgn}(\tau_\delta(\mathbf{X}))\neq\mathrm{sgn}(\bar{f}(\mathbf{X}))\}\leq\delta/4\cdot(\delta/(2-\delta))^{\frac{r}{r-1}}.$$

Therefore, we further have that

$$\begin{aligned}
&\mathrm{E}_{\mathbf{T}_{n-1}}\left[\mathrm{E}\left[\mathbb{1}\left\{|2p(\mathbf{X})-1|<1-\delta_n\right\} L_r\left(\mathbf{Y}\hat{f}_{n-1}(\mathbf{X})\right)\right]\right] \\
\leq\ & \mathrm{E}_{\mathbf{T}_n}\left[\frac{1}{n}\sum_{k=1}^{n}\mathbb{1}\left\{|2p(\mathbf{X}_k)-1|<1-\delta_n\right\}\left\{L_r\left(\mathbf{Y}_k\bar{f}_{\delta_n}(\mathbf{X}_k)\right)+B\cdot\left(\left|\frac{1}{\lambda_n}-\frac{1}{\lambda_{n-1}}\right|+\frac{2}{n\lambda_n}\right)\right\}\right] \\
&+\mathrm{E}_{\mathbf{T}_n}\left[\frac{1}{n}\sum_{k=1}^{n}\mathbb{1}\left\{|2p(\mathbf{X}_k)-1|\geq 1-\delta_n\right\}\left\{L_r\left(\mathbf{Y}_k\bar{f}_{\delta_n}(\mathbf{X}_k)\right)-L_r\left(\mathbf{Y}_k\tau_{\delta_n}(\mathbf{X}_k)\right)+L_r\left(\mathbf{Y}_k\tau_{\delta_n}(\mathbf{X}_k)\right)\right\}\right] \\
&\quad+\lambda_n\|\bar{f}_{\delta_n}\|^2_{\mathcal{H}_K}-\mathrm{E}_{\mathbf{T}_n}\left[\frac{1}{n}\sum_{k=1}^{n}\mathbb{1}\left\{|2p(\mathbf{X}_k)-1|\geq 1-\delta_n\right\} L_r\left(\mathbf{Y}_k\hat{f}_n(\mathbf{X}_k)\right)\right].
\end{aligned}$$

The first term of the right-hand side is less than or equal to

$$\mathrm{E}_{\mathbf{T}_{n-1}}\left[\mathrm{E}\left[\mathbb{1}\left\{|2p(\mathbf{X})-1|<1-\delta_n\right\} L_r\left(\mathbf{Y}\bar{f}(\mathbf{X})\right)\right]\right]+\delta_n+B\cdot\left(\left|\frac{1}{\lambda_n}-\frac{1}{\lambda_{n-1}}\right|+\frac{2}{n\lambda_n}\right).$$

The second term is less than or equal to

$$\begin{aligned}
&g(\delta_n)\frac{\delta_n}{2}+\mathrm{E}_{\mathbf{T}_n}\left[\frac{1}{n}\sum_{k=1}^{n}\mathbb{1}\left\{|2p(\mathbf{X}_k)-1|\geq 1-\delta_n\right\} L_r\left(\mathbf{Y}_k\tau_{\delta_n}(\mathbf{X}_k)\right)\right] \\
\leq\ & g(\delta_n)\delta_n/2+g(\delta_n)+\frac{\delta_n\delta_n^{r/(r-1)}}{4(2-\delta_n)^{r/(r-1)}}\left(1+\frac{\delta_n}{2}+\left(\frac{2-\delta_n}{\delta_n}\right)^{r/(r-1)}+\frac{\delta_n}{2}\right) \\
=\ & g(\delta_n)\delta_n/2+g(\delta_n)+\frac{\delta_n\delta_n^{r/(r-1)}}{4(2-\delta_n)^{r/(r-1)}}(1+\delta_n)+\frac{\delta_n}{4}.
\end{aligned}$$

For $n > N$, the sum of the third and the fourth terms is less than or equal to

$$\begin{aligned}
&\frac{\lambda_n}{C}\left(\left(\frac{2-\delta_n}{\delta_n}\right)^{r/(r-1)}+\frac{\delta_n}{2}\right)^2+g(\delta_n)\left(r\left(\frac{B}{2\lambda_n}\right)^{1/r}-r\right)\\
\leq\ & g(\delta_n)\left(\frac{\delta_n}{2-\delta_n}\right)^{2r/(r-1)}\left(\left(\frac{2-\delta_n}{\delta_n}\right)^{r/(r-1)}+\frac{\delta_n}{2}\right)^2\\
&+g(\delta_n)r(B/2)^{1/r}\left(Cg(\delta_n)\right)^{-1/r}\left(\frac{\delta_n}{2-\delta_n}\right)^{-2/(r-1)}-g(\delta_n)r,
\end{aligned}$$

where the first line is because $\|\bar{f}_{\delta_n}\|^2_{\mathcal{H}_K} \leq \frac{1}{C}\left(\left(\frac{2-\delta_n}{\delta_n}\right)^{r/(r-1)}+\frac{\delta_n}{2}\right)^2$ by Theorem 3.11 of Paulsen and Raghupathi (2016).

Combining the above together, we prove the theorem.

Appendix B. Geometric interpretation to the dual problem

To give a geometric interpretation to the dual problem (4), we assume that C is large enough so that the optimal solution to $\boldsymbol{\sigma}$ is $\boldsymbol{\sigma} \leq C\mathbf{1}$. The constraint in (4) implies that $\sum_{i=1}^n \mathbb{1}_{\{y_i=1\}}\sigma_i = \sum_{i=1}^n \mathbb{1}_{\{y_i=-1\}}\sigma_i$. This enables us to write σ_i as $\omega\tilde{\sigma}_i$ where ω is positive and $\sum_{i=1}^n \mathbb{1}_{\{y_i=1\}}\tilde{\sigma}_i = \sum_{i=1}^n \mathbb{1}_{\{y_i=-1\}}\tilde{\sigma}_i = 1$. Let $\tilde{\boldsymbol{\sigma}}$ be a $n \times 1$ vector with ith element $\tilde{\sigma}_i$. We rewrite (4) as

$$\begin{aligned}
&\max_{\omega,\tilde{\boldsymbol{\sigma}}} \quad n\log\omega + \log\tilde{\boldsymbol{\sigma}} - \omega\left\|\mathbf{X}^T\mathbf{Y}\tilde{\boldsymbol{\sigma}}\right\|\\
&\text{subject to} \quad \sum_{i=1}^n \mathbb{1}_{\{y_i=1\}}\tilde{\sigma}_i = \sum_{i=1}^n \mathbb{1}_{\{y_i=-1\}}\tilde{\sigma}_i = 1,\ \tilde{\sigma}_i > 0 \text{ for all } i, \text{ and } \omega > 0.
\end{aligned}$$

If we maximize over ω for fixed $\tilde{\boldsymbol{\sigma}}$ by substituing $\omega = n/\left\|\mathbf{X}^T\mathbf{Y}\tilde{\boldsymbol{\sigma}}\right\|$, it becomes

$$\begin{aligned}
&\max_{\tilde{\boldsymbol{\sigma}}} \quad \log\left(\prod_{i=1}^n \tilde{\sigma}_i\right)^{\frac{1}{n}} - \log\left\|\mathbf{X}^T\mathbf{Y}\tilde{\boldsymbol{\sigma}}\right\|\\
&\text{subject to} \quad \sum_{i=1}^n \mathbb{1}_{\{y_i=1\}}\tilde{\sigma}_i = \sum_{i=1}^n \mathbb{1}_{\{y_i=-1\}}\tilde{\sigma}_i = 1, \text{ and } \tilde{\sigma}_i > 0 \text{ for all } i.
\end{aligned}$$

It is minimizing the (log scaled) distance between points in the two convex hulls of $\{\mathbf{x}_i : y_i = 1\}$ and $\{\mathbf{x}_i : y_i = -1\}$ if noticing $\mathbf{X}^T\mathbf{Y}\tilde{\boldsymbol{\sigma}} = \sum_{i=1}^n \mathbb{1}_{\{y_i=1\}}\tilde{\sigma}_i\mathbf{x}_i - \sum_{i=1}^n \mathbb{1}_{\{y_i=-1\}}\tilde{\sigma}_i\mathbf{x}_i$. At the same time it is maximizing the (log scaled) geometric mean of $\tilde{\sigma}_i$.

Appendix C. Algorithm for the linear LHS classifier

Let $\boldsymbol{\theta} = (\beta_0, \boldsymbol{\beta}^T)^T$ and $\tilde{\boldsymbol{\theta}} = (\tilde{\beta}_0, \tilde{\boldsymbol{\beta}}^T)^T$ be the current value. Let $\mathbf{X}$ be a $n \times p$ matrix with ith row $\mathbf{x}_i$ and $\tilde{\mathbf{z}}$ be a $n \times 1$ vector with ith element $y_i L'\{y_i(\tilde{\beta}_0 - x_i^T\tilde{\boldsymbol{\beta}})\}/n$. By Lemma 3,

we have,

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}) &= \frac{1}{n}\sum_{i=1}^{n} L\left(y_i(\beta_0 + \mathbf{x}_i^T\boldsymbol{\beta})\right) + \lambda\boldsymbol{\beta}^T\boldsymbol{\beta} \\ &\leq \mathcal{Q}(\boldsymbol{\theta}|\tilde{\boldsymbol{\theta}}) = \frac{1}{n}\sum_{i=1}^{n} L\left(y_i(\tilde{\beta}_0 + \mathbf{x}_i^T\tilde{\boldsymbol{\beta}})\right) + \lambda\tilde{\boldsymbol{\beta}}^T\tilde{\boldsymbol{\beta}} \\ &\qquad + \tilde{\gamma}^T\begin{pmatrix}\beta_0 - \tilde{\beta}_0 \\ \boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}\end{pmatrix} + \frac{1}{2n}\begin{pmatrix}\beta_0 - \tilde{\beta}_0 \\ \boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}\end{pmatrix}^T \mathbf{P}_\lambda \begin{pmatrix}\beta_0 - \tilde{\beta}_0 \\ \boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}\end{pmatrix},\end{aligned}$$

where,

$$\tilde{\gamma} = \begin{pmatrix}\mathbf{1}^T\tilde{\mathbf{z}} \\ \mathbf{X}^T\tilde{\mathbf{z}} + 2\lambda\tilde{\boldsymbol{\beta}}\end{pmatrix} \quad \text{and} \quad \mathbf{P}_\lambda = \begin{pmatrix} n & \mathbf{1}^T\mathbf{X} \\ \mathbf{X}^T\mathbf{1} & \mathbf{X}^T\mathbf{X} + 2n\lambda\mathbf{I}_{p\times p}\end{pmatrix}.$$

Then we update $\boldsymbol{\theta}$ by the minimizer of $\mathcal{Q}(\boldsymbol{\theta}|\tilde{\boldsymbol{\theta}})$:

$$\begin{pmatrix}\beta_0 \\ \boldsymbol{\beta}\end{pmatrix} = \arg\min_{\beta_0,\boldsymbol{\beta}} \mathcal{Q}(\boldsymbol{\theta}|\boldsymbol{\theta}_m) = \begin{pmatrix}\tilde{\beta}_0 \\ \tilde{\boldsymbol{\beta}}\end{pmatrix} - n\mathbf{P}_\lambda^{-1}\tilde{\gamma}. \tag{19}$$

Here we introduce an efficient way to implement the algorithm that $O(p^3)$ operations appear only once. Let,

$$\mathbf{P}_0 = \begin{pmatrix} n & \mathbf{1}^T\mathbf{X} \\ \mathbf{X}^T\mathbf{1} & \mathbf{X}^T\mathbf{X}\end{pmatrix} \quad \text{and} \quad \mathbf{Q}_\lambda = \mathbf{P}_0 + 2n\lambda\mathbf{I}.$$

By eigen decomposition $\mathbf{P}_0 = \mathbf{U}\boldsymbol{\Pi}\mathbf{U}^T$ where $\mathbf{U} = [\ \mathbf{u}_1 \mid \mathbf{u}_2 \mid \ldots \mid \mathbf{u}_{p+1}\ ]$ is the square matrix whose ith column is the eigenvector of $\mathbf{P}_0$ and $\boldsymbol{\Pi}$ is a diagonal matrix whose ith diagonal element is the ith eigenvalue of $\mathbf{P}_0$. Then $\mathbf{Q}_\lambda = \mathbf{U}\boldsymbol{\Pi}_\lambda\mathbf{U}^T$ where $\boldsymbol{\Pi}_\lambda$ is a diagonal matrix with $(\boldsymbol{\Pi}_\lambda)_{ii} = d_i + 2n\lambda$.

$\mathbf{P}_\lambda$ can be partitioned into two matrices such that $\mathbf{P}_\lambda = \mathbf{Q}_\lambda + (\mathbf{P}_\lambda - \mathbf{Q}_\lambda)$ and by the Sherman-Morrison formula,

$$\mathbf{P}_\lambda^{-1} = \left\{\mathbf{Q}_\lambda + \begin{pmatrix}-2n\lambda & \mathbf{0}^T \\ \mathbf{0} & \mathbf{0}_{p\times p}\end{pmatrix}\right\}^{-1} = \mathbf{Q}_\lambda^{-1} + g\cdot\mathbf{v}\mathbf{v}^T = \mathbf{U}\boldsymbol{\Pi}_\lambda^{-\mathbf{1}}\mathbf{U}^T + g\cdot\mathbf{v}\mathbf{v}^T \tag{20}$$

where $\mathbf{v} = \mathbf{U}\boldsymbol{\Pi}_\lambda^{-\mathbf{1}}\mathbf{u}_1$ is the first column of $\mathbf{Q}_\lambda^{-1}$ and $g = 2n\lambda/(1 - 2n\lambda\boldsymbol{e}_1^T\mathbf{v})$ with $\boldsymbol{e}_1^T = (1, 0, \ldots, 0)$. Replacing $\mathbf{P}_\lambda^{-1}$ with (20), we see that the right hand side of (19) becomes,

$$\begin{pmatrix}\tilde{\beta}_0 \\ \tilde{\boldsymbol{\beta}}\end{pmatrix} - n\mathbf{U}\boldsymbol{\Pi}_\lambda^{-\mathbf{1}}\mathbf{U}^T\tilde{\gamma} - ng\cdot\mathbf{v}\mathbf{v}^T\tilde{\gamma}.$$

References

Peter L Bartlett, Michael I Jordan, and Jon D McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

Dankmar Böhning and Bruce G Lindsay. Monotonicity of quadratic-approximation algorithms. *Annals of the Institute of Statistical Mathematics*, 40(4):641–663, 1988.

Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL `http://archive.ics.uci.edu/ml`.

Manuel Fernández-Delgado, Eva Cernadas, Senén Barro, and Dinani Amorim. Do we need hundreds of classifiers to solve real world classification problems? *The Journal of Machine Learning Research*, 15(1):3133–3181, 2014.

Vojtech Franc, Alexander Zien, and Bernhard Schölkopf. Support vector machines as probabilistic models. In *ICML*, 2011.

Yoav Freund, Robert E Schapire, et al. Experiments with a new boosting algorithm. In *icml*, volume 96, pages 148–156. Citeseer, 1996.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *The annals of statistics*, 28(2):337–407, 2000.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: data mining, inference and prediction*. Springer, 2 edition, 2009. URL `http://www-stat.stanford.edu/~tibs/ElemStatLearn/`.

David R Hunter and Kenneth Lange. A tutorial on mm algorithms. *The American Statistician*, 58(1):30–37, 2004.

Yi Lin. A note on margin-based loss functions in classification. *Statistics & probability letters*, 68(1):73–82, 2004.

Vern I Paulsen and Mrinal Raghupathi. *An introduction to the theory of reproducing kernel Hilbert spaces*, volume 152. Cambridge university press, 2016.

Saharon Rosset, Ji Zhu, and Trevor Hastie. Margin maximizing loss functions. In *NIPS*, pages 1237–1244, 2003.

Saharon Rosset, Ji Zhu, and Trevor Hastie. Boosting as a regularized path to a maximum margin classifier. *The Journal of Machine Learning Research*, 5:941–973, 2004.

Xiaotong Shen, George C Tseng, Xuegong Zhang, and Wing Hung Wong. On ψ-learning. *Journal of the American Statistical Association*, 98(463):724–734, 2003.

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

Ingo Steinwart. Consistency of support vector machines and other regularized kernel classifiers. *IEEE transactions on information theory*, 51(1):128–142, 2005.

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

Grace Wahba. *Spline models for observational data*, volume 59. Siam, 1990.

Tong Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *The Annals of Statistics*, 32(1):56–85, 2004.

Ji Zhu, Saharon Rosset, Robert Tibshirani, and Trevor J Hastie. 1-norm support vector machines. In *Advances in neural information processing systems*, page None. Citeseer, 2003.