Huber-Norm Regularization for Linear Prediction Models

Oleksandr Zadorozhnyi[1], Gunthard Benecke[1], Stephan Mandt[2], Tobias Scheffer[1], Marius Kloft[3]

[1] University of Potsdam, Department of Computer Science
{zadorozh, gunthard.benecke, tobias.scheffer}@uni-potsdam.de
[2] Columbia University, Data Science Institute, Department of Computer Science
sm3976@columbia.edu
[3] Humboldt-Universität zu Berlin, Department of Computer Science
kloft@hu-berlin.de

Abstract. In order to avoid overfitting, it is common practice to regularize linear prediction models using squared or absolute-value norms of the model parameters. In our article we consider a new method of regularization: Huber-norm regularization imposes a combination of ℓ_1 and ℓ_2-norm regularization on the model parameters. We derive the dual optimization problem, prove an upper bound on the statistical risk of the model class by means of the Rademacher complexity and establish a simple type of oracle inequality on the optimality of the decision rule. Empirically, we observe that logistic regression with Huber-norm regularizer outperforms ℓ_1-norm, ℓ_2-norm, and elastic-net regularization for a wide range of benchmark data sets.

1 Introduction

Linear classification and regression models—such as the support vector machine (SVM) and logistic and linear regression—are widely used in machine learning, and regularized empirical-risk minimization is a standard approach to optimizing their parameters. To avoid overfitting, linear models are typically either densely or sparsely regularized. With an ℓ_2 regularizer, one obtains a dense weight vector in which all features contribute to the prediction task. For interpretability, one is often interested in a sparse solution in which many entries of the weight vector are zero. To this end, one may employ an ℓ_1 absolute value norm regularizer [25, 18]. While this type of regularization may lead to lower predictive accuracies than ℓ_2 regularization [10], the result focuses only on the most relevant features.

This paper promotes the idea of using a combination of both types of regularization, thus combining the best of both worlds. Instead of using just a single weight vector $\mathbf{w}$ that is either dense *or* sparse, we employ a sum of two weight vectors $\mathbf{w}+\mathbf{v}$. While $\mathbf{w}$ is ℓ_2 regularized and therefore dense, $\mathbf{v}$ is ℓ_1 regularized and therefore sparse. Having two different weight vectors with different regularizations allows linear models to more flexibly fit the data. It comes at a moderate computational cost, since the number of parameters is doubled.

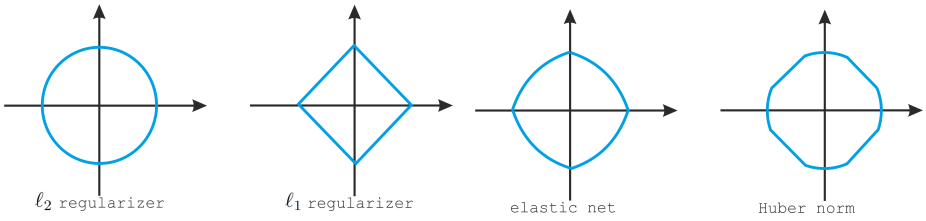

Fig. 1. Geometrical illustration of the proposed Huber-norm regularizer and comparison to common regularizers.

We first show that the proposed combination of two weight vectors is mathematically equivalent to imposing Huber-norm [7] regularization on the empirical risk of a linear model. This approach is known to be statistically more robust [8] in the sense that individual sparse weights do not necessarily involve a huge cost in the loss. This Huber norm involves quadratic costs near the origin and linear costs far away from the origin, this way penalizing outliers less severely. Because of this analogy, we call our method *Huber-norm regularization*. We derive uniform and data-dependent upper bounds on the statistical risk of the model class by means of the Rademacher complexity. We deduce a simple type of oracle inequality on the inference efficiency of the decision rule which measures the deviation of the model's risk from the lowest risk of any model in the class.

Our empirical studies show that *Huber-norm regularized logistic regression* outperforms ℓ_1- and ℓ_2-regularized as well as elastic-net-regularized logistic regression [26] in the majority of cases over a wide range of benchmark problems. To support this claim we provide evidence based on empirical studies on the UCI machine learning repository, where our method performs best among the compared methods on 23 out of 31 data sets. On particular data set—the well-known Iris data set—Huber-norm regularization leads to a prediction accuracy of 0.96 while the next-best method merely achieves 0.84.

Our paper is organized as follows. Section 2 reviews related work. In Section 3, we describe our model and its basic properties. We also prove the equivalence of the two weight vectors to Huber-norm regularization in the conventional setting. In Section 4 we then present the underlying theoretical foundations of our approach, where we prove an upper bound on the statistical risk. We present our experimental results in section 5 and conclude in Section 6.

2 Related Work

Comparisons between ℓ_1-norm and ℓ_2-norm SVMs are ubiquitous in the literature [25, 14, 13]. A robust alternative to the SVM based on the smooth ramp loss [23] requires the convex-concave procedure to convert this non-convex optimization problem into a convex one [24]. Another way of making the SVM robust [20] is based on the weighted LS-SVM that yields sparse results. Differ-

ent type of classification problems for the SVM (both convex and non-convex) are discussed by Hailong et al. [6] where the conjugate gradient approach is used to solve the optimization problem.

Our novel type of regularizer relates to the elastic-net regularizer [26] that simply amounts to taking the sum of an ℓ_1 and ℓ_2 regularizer. Our proposed regularizer is very different, as is evident from Figure 1. The plot shows contours of different regularizers in comparison. As a major difference between the elastic net and our approach, our regularizer grows asymptotically linearly for large weight vectors whereas the elastic net grows asymptotically quadratically. Lastly, our theoretical contributions are based on fundamental work by Vapnik [22].

The Huber norm [7] is frequently used as a loss function; it penalizes outliers asymptotically linearly which makes it more robust than the squared loss. The Huber norm is used as a regularization term of optimization problems in image super resolution [21] and other computer-graphics problems. The *inverse Huber function* [17] has been studied as a regularizer for regression problems. While the Huber norm penalizes large weights asymptotically linearly, the inverse Huber function imposes an asymptotically squared penalty on large weights.

3 Huber-Norm-Regularized Linear Models

In this section, after formally introducing the problem setting and optimization criterion, we show that this optimization criterion has an equivalent formulation in which the Huber norm becomes explicit. We derive the dual form and show how *Huber-norm regularization* for linear models can be implemented.

3.1 Problem Setting and Preliminaries

We consider the standard supervised prediction setup, where we are given a training sample $S = \{\mathbf{x}_i, y_i\}_{i=1}^n$ from a space $\mathcal{X} \times \mathcal{Y}$ with $\mathcal{X} = \mathbb{R}^d$. We aim at finding a linear function f that predicts well. A common way to achieve this is to first define a loss function $\ell : \mathbb{R} \times \mathcal{Y} \to \mathbb{R}_+ \cup \{0\}$ that measures the deviation of the prediction $f(\mathbf{x})$ from the correct value y, such as the logistic loss $\ell(f(\mathbf{x}), y) := \log(1 + \exp(-yf(\mathbf{x})))$ or hinge loss $\ell(f(\mathbf{x}), y) := \max(0, 1 - yf(\mathbf{x}))$. The empirical risk is then the averaged loss over the training sample, $\hat{L}_n(f) = \frac{1}{n}\sum_{i=1}^n \ell(f(\mathbf{x}_i), y_i)$ of f.

In this paper we consider methods that employ linear prediction functions $f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x}$. To avoid overfitting, one usually uses a regularizer such as the ℓ_1 regularizer $R_1(\mathbf{w}) = \|\mathbf{w}\|_1$, the ℓ_2 regularizer $R_2(\mathbf{w}) = \|\mathbf{w}\|_2^2$, or the elastic-net regularizer $R_{en}(\mathbf{w}) = \|\mathbf{w}\|_1 + \|\mathbf{w}\|_2^2$. This results in the *regularized empirical risk minimization* or short *reg-ERM problem:*

$$\min_{\mathbf{w}} \lambda R(\mathbf{w}) + \frac{1}{n}\sum_{i=1}^n \ell(y_i, \mathbf{w}^\top \mathbf{x}_i).$$

The ℓ_1 and elastic-net regularizers produce sparse, ℓ_2-norm regularizer dense weight vectors. Hence, depending on the problem, the regularizer can be chosen to match the underlying sparsity of the problem.

3.2 Linear Models with Sums of Dense and Sparse Weights

Using ℓ_1-, ℓ_2-, or elastic-net-regularized ERM either produces dense or sparse solutions. In this paper, we argue it can be beneficial to produce dense solutions with pronounced feature weights as in ℓ_1-norm regularized methods. We propose to consider linear models of the form $f(\mathbf{x}) := (\mathbf{v} + \mathbf{w})^\top \mathbf{x}$ (for notational convenience, we disregard constant offsets and assume that the first element of each $\mathbf{x}$ is a constant 1) and the regularizer $R_H(\mathbf{v}, \mathbf{w}) = \lambda \|\mathbf{v}\|_1 + \mu \|\mathbf{w}\|_2^2$, hence resulting in the following optimization problem.

Optimization Problem 1 (Sums of dense and sparse weights) *Given $\lambda, \mu > 0$ and loss function $\ell(t, y)$, solve:*

$$(\hat{\mathbf{w}}, \hat{\mathbf{v}}) = \arg\min_{\mathbf{v},\mathbf{w}} G(\mathbf{w}, \mathbf{v}, S)$$

$$\text{with } G(\mathbf{w}, \mathbf{v}, S) = \lambda ||\mathbf{v}||_1 + \mu \|\mathbf{w}\|_2^2 + \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, (\mathbf{w} + \mathbf{v})^\top \mathbf{x}_i), \tag{1}$$

where $||\cdot||_2$ and $||\cdot||_1$ denote standard ℓ_2-norm and ℓ_1-norm correspondingly.

For reasons that will become clear in the section below we call the method *Huber-regularized empirical risk minimization* or short *Huber-regERM*. Note that by letting $\lambda \to \infty$, we obtain the classic ℓ_2-norm regularization, while letting $\mu \to \infty$ leads to ℓ_1-norm regularization. Thus these methods are obtained as limit cases of our method. *Elastic-net-regularization* is not a special case of this framework, but it could be obtained by enforcing an additional constraint $\mathbf{v} = \mathbf{w}$.

3.3 Geometry of the Huber Norm

The following geometrical interpretation lets us compare linear models with sums of dense and sparse weights to the ℓ_1, ℓ_2, and elastic-net regularizers. We prove that Problem 1 is equivalent to the following problem.

Optimization Problem 2 (Equivalent Huber-Norm Problem) *Optimization Problem 1 can equivalently be formulated as:*

$$\hat{\mathbf{z}} = \arg\min_{\mathbf{z}} R_H(\mathbf{z}) + \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, \mathbf{z}^\top \mathbf{x}_i) \tag{2}$$

$$\text{where } R_H(\mathbf{z}) = \textstyle\sum_{i=1}^{d} r_H(z_i), \text{ and } r_H(z_i) = \begin{cases} \lambda \left(|z_i| - \frac{\lambda}{4\mu} \right) & \text{if } |z_i| \geq \frac{\lambda}{2\mu} \\ \mu z_i^2, & \text{otherwise} \end{cases}.$$

Note that $R_H(\mathbf{z})$ is the Huber norm of $\mathbf{z}$. While the Huber norm is often used as a robust loss function that is less sensitive to outliers, Optimization Problem 2 employs the Huber norm as regularizer. Intuitively, this results in a regularization scheme that is less sensitive to individual features which have a stong impact on

f than ℓ_2 regularization. Figure 1 illustrates isotropic lines for the Huber-norm regularizer and known regularizers for $\lambda = \mu = 1$. The Huber norm is composed of linear and squared segments. While it does not encourage sparsity as the ℓ_1 regularizer does, it encourages that most attributes only have a small impact on the decision function.

Proof (Equivalence of Optimization Problems 1 and 2). Let $\mathbf{z} = \mathbf{w}+\mathbf{v}$. Problem 1 can then be formulated as

$$\begin{aligned}\min_{\mathbf{w},\mathbf{v}} G(\mathbf{w},\mathbf{v},S) &= \min_{\mathbf{z},\mathbf{v}} \lambda||\mathbf{v}||_1 + \mu\left\|\mathbf{z}-\mathbf{v}\right\|_2^2 + \frac{1}{n}\sum_{i=1}^{n}\ell(y_i,\mathbf{z}^\top\mathbf{x}_i) \\ &= \min_{\mathbf{z}}\left(\mu\min_{\mathbf{v}}\left(\frac{\lambda}{\mu}||\mathbf{v}||_1 + \left\|\mathbf{z}-\mathbf{v}\right\|_2^2\right) + \frac{1}{n}\sum_{i=1}^{n}\ell(y_i,\mathbf{z}^\top\mathbf{x}_i)\right). \quad (3)\end{aligned}$$

Let us define $R(\mathbf{v},\mathbf{z}) := \overline{c}||\mathbf{v}||_1 + \left\|\mathbf{z}-\mathbf{v}\right\|_2^2$ where $\overline{c} := \frac{\lambda}{\mu}$. It remains to be shown that $\min_{\mathbf{v}} R(\mathbf{v},\mathbf{z})$ is a Huber-norm regularizer.

Simplifying $R = \mathbf{v}^\top\mathbf{v} - 2\mathbf{v}^\top\left(\mathbf{z} - \frac{\overline{c}}{2}\operatorname{sgn}(\mathbf{v})\right) + \mathbf{z}^\top\mathbf{z}$, we find

$$\min_{\mathbf{v}} R = \min_{\mathbf{v}(v_1,\ldots,v_d)}\left(\sum_{i=1}^{d} v_i^2 - 2v_i(z_i - \frac{\overline{c}}{2}\operatorname{sgn}(v_i))\right) + \sum_{i=1}^{d} z_i^2. \quad (4)$$

For each $i \in \{1,...,d\}$ we minimize $R_i := v_i^2 - 2v_i(z_i - \frac{\overline{c}}{2}\operatorname{sgn}(v_i))$ with respect to v_i. This is equivalent to:

$$\begin{cases}\min\limits_{v_i} v_i^2 - 2(z_i - \frac{\overline{c}}{2})v_i & \text{if } v_i > 0 \\ \min\limits_{v_i} v_i^2 - 2(z_i + \frac{\overline{c}}{2})v_i & \text{if } v_i \leq 0.\end{cases}$$

We can minimize each of these two quadratic terms analytically:

$$\begin{cases}-(z_i - \frac{\overline{c}}{2})^2 & \text{if } z_i \in \mathcal{A} := \{z \in \mathbb{R} : |z| \geq \frac{\overline{c}}{2}\} \\ 0 & \text{if } z_i \in \mathcal{A}_c := \{z \in \mathbb{R} : |z| < \frac{\overline{c}}{2}\}.\end{cases}$$

This means, that for Equation 4 we have explicitly:

$$\min_{\mathbf{v}} R = \sum_{i=1}^{d}\left(z_i^2 - \left(z_i - \frac{\overline{c}}{2}\right)^2 \mathbb{I}_{z_i\in\mathcal{A}}\right) = \sum_{i=1}^{d}\left(z_i^2\mathbb{I}_{z_i\in\mathcal{A}_C} + \overline{c}\left(|z_i| - \frac{\overline{c}}{4}\right)\mathbb{I}_{z_i\in\mathcal{A}}\right).$$

This is exactly the Huber-norm regularizer $R_H(\mathbf{z})$ of Optimization Problem 2. □

3.4 Dual Problem

In order to classify a training point, we need to compute the scalar product $(\mathbf{w}+\mathbf{v})^\top\mathbf{x}$ which may be expensive when the dimension of vectors $\mathbf{w},\mathbf{v}$ is large.

One possible solution to overcome this consists in considering a weighted sum of constraints together with an objective function computed on the training sample. This leads to a dual approach. Steinwart [19] gives a general overview of dual optimization problems for SVMs using ℓ_2- and ℓ_1-norm regularizers. The dual form of the optimization problem depends on the loss function. We complete Steinwart's overview by deriving the dual form of the *Huber-norm regularized SVM* in the following.

Optimization Problem 3 (Dual Huber-Norm SVM Problem)
Optimization Problem 1 with hinge-loss loss function (Huber-Norm SVM) has an equivalent dual form which can be formulated as follows:

$$\max_{\alpha \in \mathbb{R}^n} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \mathbf{x}_i^\top \mathbf{x_j}$$
$$s.t.\ \alpha \in [0, C]^n \wedge ||\mathbf{X}^\top \alpha||_\infty \leq \frac{\lambda}{2\mu}, \tag{5}$$

where $C := \frac{1}{2n\mu}$ *and* $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_n) \in \mathbb{R}^{n \times d}$.

Proof. The Lagrangian $L(\mathbf{w}, \mathbf{v}, \xi, \alpha, \eta)$ that corresponds to Equation 1 is given as follows:

$$L(\mathbf{w}, \mathbf{v}, \xi, \alpha, \eta) := C \sum_{i=1}^{n} \xi_i + \frac{\lambda}{2\mu} ||\mathbf{v}||_1 + \frac{1}{2} ||\mathbf{w}||_2^2 +$$
$$\sum_{i=1}^{n} \alpha_i \left(1 - y_i (\mathbf{w}^\top + \mathbf{v}^\top) \mathbf{x}_i - \xi_i\right) - \sum_{i=1}^{n} \eta_i \xi_i, \tag{6}$$

where $\alpha = (\alpha_1, ..., \alpha_n) \in [0, \infty)^n$ and $\eta = (\eta_1, ..., \eta_n) \in [0, \infty)^n$. So the dual problem [3] can be written as:

$$\max_{\alpha, \eta} \inf_{\mathbf{w}, \mathbf{v}, \xi} L(\mathbf{w}, \mathbf{v}, \xi, \alpha, \eta). \tag{7}$$

Grouping the terms in the Lagrangian gives us:

$$L(\mathbf{w}, \mathbf{v}, \xi, \alpha, \eta) = \sum_{i=1}^{n} (C - \alpha_i - \eta_i) \xi_i + \frac{\lambda}{2\mu} ||\mathbf{v}||_1$$
$$- \sum_{i=1}^{n} \alpha_i y_i \mathbf{v}^\top \mathbf{x}_i + \frac{1}{2} ||\mathbf{w}||_2^2 - \sum_{i=1}^{n} \alpha_i y_i \mathbf{w}^\top \mathbf{x}_i + \sum_{i=1}^{n} \alpha_i.$$

Now, considering the infimum with respect to $\mathbf{v}$ and $\mathbf{w}$ separately, and using the definition of a conjugate function [3, 19] we obtain:

$$\inf_{\mathbf{v}} \frac{\lambda}{2\mu} ||\mathbf{v}||_1 - \sum_{i=1}^{n} \alpha_i y_i \mathbf{v}^\top \mathbf{x}_i = -\sup_{\mathbf{v}} \frac{\lambda}{2\mu} ||\mathbf{v}||_1 + \mathbf{v}^\top \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i$$
$$= \begin{cases} 0, & \text{when } |\mathbf{X}^\top \alpha||_\infty \leq \frac{\lambda}{2\mu} \\ -\infty, & \text{otherwise,} \end{cases} \tag{8}$$

where $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_n) \in \mathbb{R}^{n \times d}$ is the data matrix whose rows $\mathbf{x}_i^\top$ are the instances and $\|\cdot\|_\infty$ -supremum norm in $\mathbb{R}^d$. Analogously, for $\mathbf{w}$ we have:

$$\begin{aligned}\inf_{\mathbf{w}} \frac{1}{2}||\mathbf{w}||_2^2 - \sum_{i=1}^{n} \alpha_i y_i \mathbf{w}^\top \mathbf{x}_i &= -\sup_{\mathbf{w}} -\frac{1}{2}||\mathbf{w}||_2^2 + \mathbf{w}^\top \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i \\ &= \frac{1}{2}\left(\sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i\right)^\top \left(\sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i\right). \end{aligned} \tag{9}$$

Finally, computing the gradient with respect to ξ gives that for each $i \in \{1, ...n\}$:

$$C - \eta_i - \alpha_i = 0 \Leftrightarrow \alpha_i = C - \eta_i. \tag{10}$$

Now, for fixed λ, μ, and $\mathbf{X}$, define $P = \{\alpha | \alpha \in [0, C]^n \wedge ||\mathbf{X}^\top \alpha||_\infty \leq \frac{\lambda}{2\mu}\}$,where $\mathbf{y} = (y_1, ..., y_n) \in \mathbb{R}^n$. Substituting Equations 8, 9, and 10 into Equation 7 gives the following dual problem:

$$\max_{\alpha \in P} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \mathbf{x}_i^\top \mathbf{x}_j, \tag{11}$$

which is a quadratic optimization problem within set P and can be solved with known methods.

□

By close inspection of Equation 11, we observe that our dual optimization problem closely resembles the one for *SVM* using ℓ_2 regularization, but with a difference in the form of the domain P of the optimization problem.

3.5 Algorithm & Implementation

Algorithm 1 implements *Huber-regularized empirical risk minimization for linear models*. The algorithm works by alternatingly minimizing the occurring ℓ_1-norm and ℓ_2-norm regularized minimization problems, respectively. For each step of optimization procedure we use gradient descent, assuming that the other vector is constant. The gradient of the ℓ_1 norm of $\mathbf{v}$ is not defined for $\mathbf{v} = \mathbf{0}$; here, we use subgradients [3].

4 Theoretical Analysis

In this section we present a theoretical analysis of the proposed *Huber-norm regularizer for linear models*. We obtain bounds on the statistical risk based on the established framework of Rademacher complexities [2, 16] and, consequently, on the norms of the vectors $\mathbf{v}, \mathbf{w}$ and number of training samples n [2].

Algorithm 1 Optimization Procedure

Input: $S = \{\mathbf{x}_i, y_i\}_{i=1}^n$
$\mathbf{w} = \mathbf{0}$, $\mathbf{v} = \mathbf{0}$.
repeat
    solve $\hat{\mathbf{w}} := \arg\min_{\mathbf{w}} G(\mathbf{w}, \mathbf{v}, S)$ by gradient descent,
    solve $\hat{\mathbf{v}} := \arg\min_{\mathbf{v}} G(\hat{\mathbf{w}}, \mathbf{v}, S)$ by gradient descent,
    let $\mathbf{w}, \mathbf{v} = (\hat{\mathbf{w}}, \hat{\mathbf{v}})$.
until convergence.
Output: $\mathbf{w}, \mathbf{v}$

4.1 Preliminaries and Aim

Let $S = \{\mathbf{x_i}, y_i\}_{i=1}^n$ be a sample of n training points that are independently drawn from one and the same distribution $P_{X,Y}$ over $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{X} = \mathbb{R}^d$; let the output space $\mathcal{Y}$ be discrete for classification and continuous for regression. In this theoretical analysis, we study the *Huber-regERM* model class

$$\mathcal{F} := \{f : \mathbf{x} \mapsto (\mathbf{w} + \mathbf{v})^\top \mathbf{x} : \mathbb{R}^d \to \mathbb{R} | \, \|\mathbf{w}\|_2 \leq W, \|\mathbf{v}\|_2 \leq V\}, \tag{12}$$

where W and V are initially unknown constants. Loss function $\ell : \mathbb{R} \times \mathcal{Y} \to \mathbb{R}_+ \cup \{0\}$ may be any convex loss function that is $\mathcal{L}$-Lipschitz continuous and absolutely bounded by constant $B \in \mathbb{R}$. The aim of our theoretical analysis is to obtain bounds on the deviation of the *risk* $L(f) = E_{P_{X,Y}}[\ell(f(\mathbf{x}), y)]$ of the model $f \in \mathcal{F}$ from *empirical risk* $\hat{L}_n(f) = \frac{1}{n}\sum_{i=1}^n \ell(f(\mathbf{x_i}), y_i)$.

Let $\{\sigma_i\}_{i=1}^n$ be independent Rademacher random variables, meaning that each of them is uniformly distributed over $\{-1, +1\}$. Denote by Σ the joint uniform distribution of $\sigma_1, \ldots, \sigma_n$. Then the *empirical Rademacher complexity* is defined as

$$\hat{\mathfrak{R}}_S(\ell \circ \mathcal{F}) := E_\Sigma \left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \sigma_i \ell(f(\mathbf{x}_i), y_i)) \right], \tag{13}$$

and the (theoretical) *Rademacher complexity* [2, 16] is defined as $\mathfrak{R}_n(\ell \circ \mathcal{F}) := E_S[\hat{\mathfrak{R}}_S(\ell \circ \mathcal{F})]$. Here, the expectation is taken under the distribution of the sample S. It has been shown [2, 16] that when ℓ is $\mathcal{L}$-Lipschitz continuous in the second argument, then with probability at least $1 - \delta$, for all $f \in \mathcal{F}$:

$$L(f) \leq \hat{L}_n(f) + 2\mathcal{L} E_S \left[\hat{\mathfrak{R}}_S(\mathcal{F}) \right] + B \sqrt{\frac{\log \delta^{-1}}{2n}}. \tag{14}$$

4.2 Bounds on the Risk of Huber-regularized Linear Models

Our main theoretical contributions are bounds on statistical risk based on data-dependent and uniform upper bounds on the Rademacher complexity of the model class $\mathcal{F}$ defined by Equation 12.

Theorem 1 (Uniform risk bound for Huber regularization) *Let $\mathcal{F}$ be defined by Equation 12, let ℓ be a $\mathcal{L}$-Lipschitz continuous loss function, and let R be a constant such that $|\ell(t,y)| \leq R$ for all $t \in \mathbb{R}$ and $y \in \mathcal{Y}$. Let the ℓ_2 norm of all instances is bounded by $\|\mathbf{x}\|_2 \leq R_{\mathbf{x}}$ with probability 1 by some $R_{\mathbf{x}}$. Then, for every $\delta \in (0,1)$, with probability at least $1-\delta$ the following holds for all $f \in \mathcal{F}$:*

$$L(f) \leq \hat{L}_n(f) + 2\mathcal{L}\sqrt{\frac{2(W^2+V^2)}{n}}R_{\mathbf{x}} + R\sqrt{\frac{\log \delta^{-1}}{2n}} \tag{15}$$

where $W = \sqrt{\frac{R}{\mu}}, V = \frac{R}{\lambda}$

Instead of relying on a uniform bound $R_{\mathbf{x}}$ on the data $\mathbf{x}_i$, we can give the following data-dependent bound on the risk.

Proposition 1 (Data-dependent risk bound for Huber regularization) *Let $\mathcal{F}$ be defined by Equation 12, and let ℓ be a $\mathcal{L}$-Lipschitz continuous loss function. Then, for every $\delta \in (0,1)$, with probability at least $1-\delta$ the following holds for all $f \in \mathcal{F}$, where W, V, and R as defined as in Theorem 1:*

$$L(f) \leq \hat{L}_n(f) + 2\mathcal{L}\frac{\sqrt{2(W^2+V^2)\sum_{i=1}^{n}\|\mathbf{x}_i\|^2}}{n} + (2\mathcal{L}+1)R\sqrt{\frac{\log(\frac{2}{\delta})}{2n}}. \tag{16}$$

4.3 Lemmata and Auxilary Results

The risk bounds are based on the following three lemmas.

Lemma 1 *For the functional class $\mathcal{F}$ of Equation 12, the following data-dependend bound on the empirical Rademacher complexity holds:*

$$\hat{\mathfrak{R}}_S(\mathcal{F}) \leq \frac{\sqrt{2(W^2+V^2)\sum_{i=1}^{n}\|\mathbf{x}_i\|^2}}{n}. \tag{17}$$

Lemma 2 *For the functional class $\mathcal{F}$ of Equation 12, the (theoretical) Rademacher complexity is bounded as follows:*

$$\mathfrak{R}_n(\mathcal{F}) = E_S[\hat{\mathfrak{R}}_S(\mathcal{F})] \leq \sqrt{\frac{2(W^2+V^2)}{n}}R_{\mathbf{x}}. \tag{18}$$

where $R_{\mathbf{x}}$ is a constant such that $||\mathbf{x}||_2 \leq R_{\mathbf{x}}$ almost surely under P_X.

Lemma 3 *Let $(\hat{\mathbf{w}}, \hat{\mathbf{v}}) = \arg\min_{\mathbf{v},\mathbf{w}} G(\mathbf{w}, \mathbf{v}, S)$. Then $\|\hat{\mathbf{w}}\|_2 \leq \sqrt{\frac{R}{\mu}}$, $\|\hat{\mathbf{v}}\|_2 \leq \frac{R}{\lambda}$, where R as in Theorem 1.*

Proof (Lemma 1). Following the ideas presented by Mohri [16], we rewrite the empirical Rademacher complexity using the Cauchy-Schwartz inequality:

$$\begin{aligned}\hat{\mathfrak{R}}_S(\mathcal{F}) &= \frac{1}{n} E_\sigma \left[\sup_{\|\mathbf{w}\|_2 \leq W, \|\mathbf{v}\|_2 \leq V} \sum_{i=1}^{n} \left(\sigma_i (\mathbf{w} + \mathbf{v})^\top \mathbf{x}_i \right) \right] \\ &= \frac{1}{n} E_\sigma \left[\sup_{\|\mathbf{w}\|_2 \leq W, \|\mathbf{v}\|_2 \leq V} (\mathbf{w} + \mathbf{v})^\top \sum_{i=1}^{n} \sigma_i \mathbf{x}_i \right] \\ &\leq \frac{1}{n} E_\sigma \left[\sup_{\|\mathbf{w}\|_2 \leq W, \|\mathbf{v}\|_2 \leq V} \|\mathbf{w} + \mathbf{v}\|_2 \left\| \sum_{i=1}^{n} \sigma_i \mathbf{x}_i \right\|_2 \right]. \end{aligned} \tag{19}$$

Using the inequality $\sum_{i=1}^{d} (w_i + v_i)^2 \leq 2 \sum_{i=1}^{d} (w_i^2 + v_i^2)$ for the right-hand side of Equation 19, according to the restrictions on the norms of $\mathbf{w}, \mathbf{v}$ we get:

$$E_\sigma \left[\sup_{\|\mathbf{w}\|_2 \leq W, \|\mathbf{v}\|_2 \leq V} \|\mathbf{w} + \mathbf{v}\|_2 \left\| \sum_{i=1}^{n} \sigma_i \mathbf{x}_i \right\|_2 \right] \leq \sqrt{2(W^2 + V^2)} E_\sigma \left[\left\| \sum_{i=1}^{n} \sigma_i \mathbf{x}_i \right\|_2 \right] \tag{20}$$

and because of Jensen's inequality for $E_\sigma [\|\cdot\|]$, linearity of expectation and independence of σ_i, σ_j for $j \neq i$ we obtain:

$$\begin{aligned} E_\sigma \left[\left\| \sum_{i=1}^{n} \sigma_i \mathbf{x}_i \right\|_2 \right] &\leq \sqrt{E_\sigma \left[\sum_{i,j=1}^{n} \sigma_i \sigma_j \mathbf{x_i}^t \mathbf{x}_j \right]} \\ &= \sqrt{\sum_{i=1}^{n} E_\sigma \left[\|\mathbf{x}_i\|_2^2 \right]} = \sqrt{\sum_{i=1}^{n} \|\mathbf{x}_i\|_2^2}. \end{aligned} \tag{21}$$

Uniting the results of Inequality 20 and Equation 21 in Equation 19 we get the statement of Lemma 1.

□

Proof (Lemma 2). Using Lemma 1 and the assumption that the $\mathbf{x}_i$ are uniformly bounded by constant $R_\mathbf{x}$ we obtain:

$$\hat{\mathfrak{R}}_S(\mathcal{F}) \leq \sqrt{\frac{2(W^2 + V^2)}{n}} R_\mathbf{x}. \tag{22}$$

Equation 22 no longer depends on the sample, and therefore Lemma 2 follows.

□

Naturally, one may not have any *a-priori* knowledge about the constants W and V that restrict the possible values of $\mathbf{w}$ and $\mathbf{v}$ in Inequality 18. Despite that, for a given optimization problem that includes the current class of models, one can apply certain arguments from which one can infer bounds for W and V. Lemma 3 gives us such bounds for Optimization Problem 1.

Proof (Lemma 3). When $(\hat{\mathbf{w}}, \hat{\mathbf{v}})$ is a solution of optimization problem (1), then

$$G(\hat{\mathbf{w}}, \hat{\mathbf{v}}, S) \leq G(\mathbf{0}, \mathbf{0}, S) \leq R$$

This implies that the optimal solution necessarily satisfies the following condition: $\lambda \|\mathbf{v}\|_1 + \mu \|\mathbf{w}\|_2 \leq R$. As far as $\|\mathbf{v}\|_1 \geq \|\mathbf{v}\|_2$ we have that in order to be an optimal solution $\hat{\mathbf{v}}$ should satisfy following constraint: $\|\mathbf{v}\|_2 \leq \frac{R}{\lambda}$. For $\hat{\mathbf{w}}$ we obtain straightforward necessary condition, that $\|\mathbf{w}\|_2^2 \leq \frac{R}{\mu}$ which implies the claim of Lemma 3.

□

Lemma 3 implies that the norms of the vectors $\mathbf{v}$ and $\mathbf{w}$ of a solution of Optimization Problem 1 necessary have to lie within balls with radius $W := \sqrt{\frac{R}{\mu}}$ for $\mathbf{w}$ and of radius $V := \frac{R}{\lambda}$ for $\mathbf{v}$, centered in the origin.

4.4 Proof of the Huber-Norm Risk Bounds

We are now equipped to prove Theorem 1.

Proof (Theorem 1). Lemma 2 gives us a bound on the Rademacher complexity of the functional class of Equation 12, and Lemma 3 gives us necessary constraints on the norms W and V. Inserting both into Inequality 14, we obtain Theorem 1.

□

Proof (Proposition 1). Lemma 1 gives us a data-dependent bound on the empirical Rademacher complexity of the functional class of Equation 12. Adapting Inequality (3.14) from theorem 3.1 in Mohri et al. [16] for our needs, we have with probability at least $1 - \frac{\delta}{2}$:

$$\mathfrak{R}_n(\mathcal{F}) \leq \hat{\mathfrak{R}}_S(\mathcal{F}) + R\sqrt{\frac{\log(\frac{2}{\delta})}{2n}}. \tag{23}$$

Using the union bound for Inequality 14 (with $\frac{\delta}{2}$ instead of δ and constant R from Theorem 1) and Inequality 23, we get with probability $1 - \delta$:

$$L(f) \leq \hat{L}_n(f) + 2\mathcal{L}\hat{\mathfrak{R}}_S(\mathcal{F}) + 2\mathcal{L}R\sqrt{\frac{\log(\frac{2}{\delta})}{2n}} + R\sqrt{\frac{\log(\frac{2}{\delta})}{2n}}. \tag{24}$$

Together with Lemma 1 this yields the claim of Proposition 1.

□

4.5 Corollaries

In practice, we will be interested in obtaining upper bounds for concrete loss functions such as the hinge loss $\ell(t, y) = \max(0, 1 - yt)$ or logistic loss $\ell(y, t) = \log(1 + \exp(-yt))$ in case of two-class classification problems. Since these loss functions are 1-Lipschitz [19], Theorem 1 produces therefore following corollaries.

Corollary 1. *For Optimization Problem 1 under the assumptions of Theorem 1 with loss-function* $\ell(y,t) = \max(0, 1-yt), t \in \mathbb{R}, y \in \{-1,1\}$ *one obtains that, with probability at least* $1-\delta$ *for all* $f \in \mathcal{F}$*:*

$$L(f) \leq \hat{L}_n(f) + 2\sqrt{\frac{2(W^2+V^2)}{n}}R_{\mathbf{x}} + B\sqrt{\frac{\log \delta^{-1}}{2n}} \tag{25}$$

where $W = \sqrt{\frac{1}{\mu}}$, $V = \frac{1}{\lambda}$, $B = 1 + \sqrt{2(W^2+V^2)}R_{\mathbf{x}}$.

Corollary 2. *For Optimization Problem 1 under the assumptions of Theorem 1 with loss-function* $\ell(y,t) = \log(1+\exp(-yt)), t \in \mathbb{R}, y \in \{-1,1\}$ *one obtains that with probability at least* $1-\delta$ *for all* $f \in \mathcal{F}$*:*

$$L(f) \leq \hat{L}_n(f) + 2\sqrt{\frac{2(W^2+V^2)}{n}}R_{\mathbf{x}} + B^l\sqrt{\frac{\log \delta^{-1}}{2n}} \tag{26}$$

where $W = \sqrt{\frac{\log 2}{\mu}}$, $V = \frac{\log 2}{\lambda}$, $B^l := \frac{\exp(R_{\mathbf{x}}\sqrt{2(W^2+V^2)})}{\sqrt{\exp(R_{\mathbf{x}}\sqrt{2(W^2+V^2)})+1}}$.

Proof. For the hinge loss, under the conditions of Theorem 1, we have that for any $\mathbf{x} \in \mathbb{R}^d$, s.t, $\|\mathbf{x}\|_2 \leq R_{\mathbf{x}}$ the loss is bounded by $1+|(\mathbf{w}+\mathbf{v})^\top \mathbf{x}|$, which is upper-bounded by $1 + \sqrt{2(W^2+V^2)}R_{\mathbf{x}}$ as the combination of bounds on $\|\mathbf{w}+\mathbf{v}\|_2$ and $\|\mathbf{x}\|_2$. So, $|\ell(t,y)| \leq B := 1 + \sqrt{2(W^2+V^2)}R_{\mathbf{x}}$. Then the conclusion follows by applying Theorem 1. The proof for the logistic loss is analogous.

□

4.6 Discussion of Results

We will now compare the generalization performance of the developed *Huber-norm regularizer* with the performance of known regularizers.

Comparison to ℓ_1 and ℓ_2-Norm Regularization. The optimization problems of the ℓ_2-norm and ℓ_1-norm empirical risk minimization are

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \mu \|\mathbf{w}\|_2^2 + \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, \mathbf{w}^T\mathbf{x}_i) \text{ and} \tag{27}$$

$$\hat{\mathbf{v}} = \arg\min_{\mathbf{v}} \lambda||\mathbf{v}||_1 + \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, \mathbf{v}^T\mathbf{x}_i), \tag{28}$$

respectively. Theoretical upper bounds on the statistical risk for both Equations 27 and 28 result from Mohri [16] for the Rademacher complexity of linear models. In these cases, the upper bound on the Rademacher complexity is also of the order of $\sqrt{\frac{1}{n}}$ and depends as well on the bounds on norms of the vectors W, V (for each case separately) and on the bounds on the data.

Comparison to Elastic Net. The optimization problem of the *empirical risk minimization with elastic-net regularizer* is

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \lambda||\mathbf{w}||_1 + \mu||\mathbf{w}||_2^2 + \frac{1}{n}\sum\nolimits_{i=1}^{n} \ell(y_i, \mathbf{w}^\top \mathbf{x}_i) \tag{29}$$

with $\ell(y, 0) = 1$ [26]. From a similar argumentation as in Theorem 1 [11, 16] one can infer that upper bounds on the Rademacher complexity for this procedure will also be of order $\mathcal{O}(\sqrt{\frac{W^2 R_{\mathbf{x}}^2}{n}})$, where now $W = \sqrt{\frac{1}{\lambda+\mu}}$ and $R_{\mathbf{x}}$ as before.

Oracle Inequality. We will relate the generalization performance of the model to the performance of the best possible model in that class—which is unknown in practice—using an oracle-type inequality [4, 12]. As a corollary of Theorem 1, we can obtain an oracle-type inequality in high probability for $\mathcal{F}$:

$$G(\hat{\mathbf{w}}, \hat{\mathbf{v}}, S) \leq \arg\min_{(\mathbf{w},\mathbf{v})} G(\mathbf{w}, \mathbf{v}, S) + 2\Delta,$$

where Δ is the parameter that defines the complexity of $(\hat{\mathbf{w}}, \hat{\mathbf{v}}) \in \mathcal{F}$ and is given explicitly in the following Proposition 2 that follows from Theorem 1.

Proposition 2 *Let all conditions of Theorem 1 hold, let* $(\hat{\mathbf{w}}, \hat{\mathbf{v}}) = \arg\min_{\mathbf{w},\mathbf{v}} G(\mathbf{w}, \mathbf{v}, S)$*, and let* W*,* V*,* $R_{\mathbf{x}}$*, and* R *be defined in Theorem 1. Then with probability at least* $1 - \delta$*:*

$$G(\hat{\mathbf{w}}, \hat{\mathbf{v}}, S) - \arg\min_{\mathbf{w},\mathbf{v}} G(\mathbf{w}, \mathbf{v}, S) \leq 2\mathcal{L}\sqrt{\frac{2(W^2 + V^2)}{n}} R_{\mathbf{x}} + R\sqrt{\frac{\log \delta^{-1}}{2n}}. \tag{30}$$

Tightness Comparison. Comparing the order of our upper risk bound with classical results for empirical risk minimization problems [1], [5] one can see that our bound is tight, and of order $\sqrt{\frac{1}{n}}$.

5 Experiments

This section compares *logistic regression with Huber-norm regularization* to *logistic regression with* ℓ_1, with ℓ_2, and with *elastic-net regularization*.

5.1 Experimental Setting

We conduct experiments on benchmark problems from the UCI repository [15]. In order to avoid a possible selection bias, we select the 31 first (in alphabetical order) classification problems that use matrix data format. We skip trivial problems for which all models achieve perfect accuracy. We transform categorical features into binary values using one-hot coding. For multi-class problems, we removed classes that have fewer instances than the number of cross-validation folds. All features are centered and scaled to unit variance. Missing values are

Table 1. Accuracies and standard errors for UCI data sets

Data Set	ℓ_1 *regularization*	*Elastic-net reg.*	ℓ_2 *reg.*	*Huber reg.*
abalone	0.236 ± 0.008	0.236 ± 0.008	0.238 ± 0.015	$\mathbf{0.262 \pm 0.016}*$
arrhythmia	0.687 ± 0.044	0.683 ± 0.049	0.634 ± 0.053	$\mathbf{0.722 \pm 0.033}*$
audiology	0.576 ± 0.071	0.688 ± 0.045	0.738 ± 0.044	$\mathbf{0.748 \pm 0.055}$
balance-scale	0.907 ± 0.016	0.910 ± 0.015	0.910 ± 0.015	$\mathbf{0.957 \pm 0.019}*$
bank	0.899 ± 0.001	0.899 ± 0.000	0.899 ± 0.000	$\mathbf{0.901 \pm 0.001}*$
banknote	0.977 ± 0.011	0.976 ± 0.011	0.977 ± 0.011	$\mathbf{0.991 \pm 0.004}$
blood	0.770 ± 0.010	0.769 ± 0.010	0.771 ± 0.013	$\mathbf{0.774 \pm 0.012}$
breast-canc	0.689 ± 0.029	0.692 ± 0.039	0.696 ± 0.052	$\mathbf{0.710 \pm 0.065}$
breast-canc-wisc	0.963 ± 0.017	0.970 ± 0.009	0.953 ± 0.013	$\mathbf{0.973 \pm 0.012}$
breast-canc-wisc-dia	0.952 ± 0.032	0.952 ± 0.031	0.959 ± 0.019	$\mathbf{0.977 \pm 0.017}$
breast-tissue	0.879 ± 0.083	0.878 ± 0.060	0.878 ± 0.060	$\mathbf{0.907 \pm 0.044}$
car	0.841 ± 0.012	0.842 ± 0.011	0.841 ± 0.010	$\mathbf{0.896 \pm 0.005}*$
climate-model	0.915 ± 0.004	0.915 ± 0.004	0.915 ± 0.004	$\mathbf{0.955 \pm 0.018}*$
congress-voting	$\mathbf{0.956 \pm 0.029}$	0.956 ± 0.029	0.954 ± 0.025	0.954 ± 0.032
conn-sonar	0.746 ± 0.050	0.760 ± 0.071	0.736 ± 0.036	$\mathbf{0.770 \pm 0.036}$
contraceptive	0.506 ± 0.041	0.505 ± 0.041	0.508 ± 0.042	$\mathbf{0.512 \pm 0.035}$
credit-approval	0.851 ± 0.012	0.855 ± 0.018	0.859 ± 0.015	$\mathbf{0.862 \pm 0.009}$
cylinder-bands	0.746 ± 0.014	0.780 ± 0.020	$\mathbf{0.802 \pm 0.025}$	0.798 ± 0.016
dermatology	$\mathbf{0.975 \pm 0.027}$	0.965 ± 0.031	0.970 ± 0.025	0.970 ± 0.026
echocardiogram	0.757 ± 0.058	0.770 ± 0.075	0.784 ± 0.119	$\mathbf{0.797 \pm 0.106}$
ecloi	0.840 ± 0.034	0.840 ± 0.034	0.837 ± 0.032	$\mathbf{0.871 \pm 0.070}$
first-order	0.822 ± 0.001	$\mathbf{0.822 \pm 0.001}$	0.822 ± 0.002	0.821 ± 0.001
flags	0.675 ± 0.046	$\mathbf{0.691 \pm 0.029}$	0.659 ± 0.032	0.670 ± 0.022
glass	0.588 ± 0.042	0.583 ± 0.033	0.592 ± 0.056	$\mathbf{0.603 \pm 0.052}$
haberman-survival	$\mathbf{0.735 \pm 0.005}$	0.726 ± 0.019	0.684 ± 0.114	0.709 ± 0.038
hepatitis	0.800 ± 0.075	0.806 ± 0.067	0.806 ± 0.077	$\mathbf{0.815 \pm 0.111}$
horse-colic	0.831 ± 0.025	$\mathbf{0.848 \pm 0.041}$	0.826 ± 0.025	0.845 ± 0.024
image-segmentation	0.829 ± 0.010	0.833 ± 0.011	0.846 ± 0.007	$\mathbf{0.865 \pm 0.127}$
ionosphere	$\mathbf{0.880 \pm 0.034}$	0.866 ± 0.012	0.878 ± 0.038	0.878 ± 0.042
iris	0.840 ± 0.060	0.833 ± 0.047	0.833 ± 0.071	$\mathbf{0.960 \pm 0.015}*$
leaf	0.644 ± 0.048	0.665 ± 0.065	0.675 ± 0.036	$\mathbf{0.834 \pm 0.036}*$

filled in using mean imputation for continuous values and are represented as a separate one-hot coded attribute for categorical values.

We run nested stratified cross validation with an outer loop of five folds. Regularization parameters $[\lambda, \mu]$ are tuned by an inner loop of three-fold cross validation on the training portion over the grid of $[10^{-5}, ..., 10^{3}] \times [10^{-3}, ..., 10^{4}]$.

5.2 Results

Table 1 shows the accuracies of different regularizers. For each problem the highest empirical accuracy is typeset in bold face; asterisks mark models that are significantly better than the best of the other three models, based on a paired t test with $p < 0.05$. *logistic regression with Huber-norm regularization*

achieves the highest empirical accuracy for 23 out of 31 problems; its accuracy is significantly higher than the accuracy of any other model for 8 problems. No reference methods outperform *Huber-norm regularization* significantly.

The UCI repository reflects a certain distribution $P(S)$ over data sets. We state the null hypothesis A that the probability of *Huber-norm regularization* outperforming all three reference methods on a randomly drawn problem under $P(S)$ does not exceed 0.5, and the null hypothesis B that the probability of *Huber-norm regularization* outperforming all three reference methods on a randomly drawn problem under $P(S)$ is below 0.5. We count each cross-validation fold of each UCI data set as a single observation of a binary random variable and determine the binomial likelihood of observing the outcomes which are reflected in Table 1. *Logistic regression with Huber-norm regularization* achieves a higher empirical accuracy than all three baselines in 86 out of 155 cross-validation folds, and an equally high accuracy as the best baseline in an additional 24 cases. We can therefore reject the null hypothesis A at $p = 0.09$ and null hypothesis B even at $p < 0.001$. We conclude that for the distribution of UCI problems, the *Huber-norm regularization* is the best-performing regularizer among the ℓ_1, ℓ_2, *elastic-net* and *Huber regularization*.

6 Conclusions

We proposed a new way of regularizing linear prediction models based on a combination of dense and sparse weight vectors. In more detail, we employ a linear weight vector that is the sum of two terms, $\mathbf{w} + \mathbf{v}$, where $\mathbf{w}$ is ℓ_2 regularized and $\mathbf{v}$ is ℓ_1 regularized. This results in an effective *Huber-norm regularizer* for $\mathbf{w} + \mathbf{v}$, which is very different from an elastic net. Starting with theoretical considerations, we first derived bounds on the statistical risk based on the framework of Rademacher complexities. In our subsequent experimental study, our algorithm showed higher predictive accuracies on a majority of UCI data sets, where we compared against ℓ_1, ℓ_2, and elastic-net regularization. In future work, we would like to study extensions to non-linear kernel functions and multiple kernels [9].

Acknowledgments

MK acknowledges support from the German Research Foundation (DFG) award KL 2698/2-1 and from the Federal Ministry of Science and Education (BMBF) award 031L0023A. The authors would like to thank Christoph Lippert, Gilles Blanchard, Florian Wenzel, and Shinichi Nakajima for helpful discussions.

References

1. Bartlett, P.L., Bousquet, O., Mendelson, S.: Local rademacher complexities. Annals of Statistics pp. 1497–1537 (2005)
2. Bartlett, P.L., Mendelson, S.: Rademacher and Gaussian complexities: Risk bounds and structural results. In: Proceedings of the International Conference on Computational Learning Theory. pp. 224–240. Springer (2001)

3. Boyd, S., Vandenberghe, L.: Convex Optimization. Cambridge University Press (2004)
4. Clarke, B., Fokoué, E., Zhang, H.: Principles and Theory for Data Mining and Machine Learning. Springer Verlag (2009)
5. Devroye, Luc and Lugosi, Gabor: Lower bounds in pattern recognition and learning. Pattern Recognition 28(7) (1995)
6. Hailong, H., Haien, L., Jianwei, L.: P-norm regularized SVM classifier by non-convex conjugate gradient algorithm. In: Proceedings of the Chinese Control and Decision Conference. vol. 3, pp. 2685–2690. IEEE (2013)
7. Huber, P.: Robust estimation of a location parameter. Annals of Mathematical Statistics 53, 73–101 (1964)
8. Huber, P.J.: Robust statistics. Springer (2011)
9. Kloft, M., Brefeld, U., Sonnenburg, S., Zien, A.: lp-Norm Multiple Kernel Learning. Journal of Machine Learning Research 12, 953–997 (2011)
10. Kloft, M., Brefeld, U., Laskov, P., Müller, K.R., Zien, A., Sonnenburg, S.: Efficient and accurate lp-norm multiple kernel learning. In: Advances in Neural Information Processing Systems 22, pp. 997–1005. Curran Associates, Inc. (2009)
11. Kloft, M., Rückert, U., Bartlett, P.L.: A unifying view of multiple kernel learning. In: Proceedings of the European Conference on Machine Learning, pp. 66–81. Springer (2010)
12. Koltchinskii, V.: Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems, vol. 2033. Springer (2011)
13. Koshiba, Y., Abe, S.: Comparison of l1 and l2 support vector machines. In: Proceedings of the International Joint Conference on Neural Networks. vol. 3, pp. 2054–2059. IEEE (2003)
14. Kujala, J., Aho, T., Elomaa, T.: A walk from 2-norm SVM to 1-norm SVM. In: Proceedings of the IEEE Conference on Data Mining. pp. 836–841. IEEE (2009)
15. Lichman, M.: UCI machine learning repository (2013), http://archive.ics.uci.edu/ml
16. Mohri, M., Rostamizadeh, A., Talwalkar, A.: Foundations of Machine Learning. MIT press (2012)
17. Owen, A.: A robust hybrid of lasso and ridge regression. Contemporary Mathematics 443, 59–72 (2007)
18. Robert, T.: The Lasso method for variable selection in the Cox model. Statistics in Medicine 16, 385–395 (1997)
19. Steinwart, I., Christmann, A.: Support Vector Machines. Springer Science & Business Media (2008)
20. Suykens, J.A., De Brabanter, J., Lukas, L., Vandewalle, J.: Weighted least squares support vector machines: robustness and sparse approximation. Neurocomputing 48(1), 85–105 (2002)
21. Unger, M., Pock, T., Werlberger, M., Bischof, H.: A convex approach for variational super-resolution. In: Joint Pattern Recognition Symposium. pp. 313–322 (2010)
22. Vapnik, V.: The nature of statistical learning theory. Springer (1995)
23. Wang, L., Jia, H., Li, J.: Training robust support vector machine with smooth ramp loss in the primal space. Neurocomputing 71(13), 3020–3025 (2008)
24. Yuille, A.L., Rangarajan, A.: The concave-convex procedure. Neural computation 15(4), 915–936 (2003)
25. Zhu Ji, S.R., Trevor, H., Rob, T.: 1-norm support vector machines. Advances of Neural Information Processing Systems (2004)
26. Zou, H., Hastie, T.: Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B 67(2), 301–320 (2005)