Classification using margin pursuit

Matthew J. Holland
Osaka University

Abstract

In this work, we study a new approach to optimizing the margin distribution realized by binary classifiers, in which the learner searches the hypothesis space in such a way that a pre-set margin level ends up being a distribution-robust estimator of the margin location. This procedure is easily implemented using gradient descent, and admits finite-sample bounds on the excess risk under unbounded inputs, yielding competitive rates under mild assumptions. Empirical tests on real-world benchmark data reinforce the basic principles highlighted by the theory.

1 Introduction

Machine learning systems depend on both statistical inference procedures and efficient implementations of these procedures. These issues are reflected clearly within a risk minimization framework, in which given a known loss $L(\boldsymbol{w}; \boldsymbol{z})$ depending on data $\boldsymbol{z}$ and parameters $\boldsymbol{w}$, the ultimate objective is minimization of the risk $R(\boldsymbol{w}) := \mathbf{E}\, L(\boldsymbol{w}; \boldsymbol{z})$, where expectation is taken with respect to the data. Since R is unknown, the learner seeks to determine a candidate $\widehat{\boldsymbol{w}}$ based on a limited sample $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$ such that $R(\widehat{\boldsymbol{w}})$ is sufficiently small, with high probability over the random draw of the sample. Inference is important because R is always unknown, and the implementation is important because the only $\widehat{\boldsymbol{w}}$ we ever have in practice is one that can be computed given finite time, memory, and processing power.

Our problem of interest is binary classification, where $\boldsymbol{z} = (\boldsymbol{x}, y)$ with inputs $\boldsymbol{x} \in \mathbb{R}^d$ and labels $y \in \{-1, 1\}$. Parameter $\boldsymbol{w}$ shall determine a scoring rule $h(\cdot; \boldsymbol{w})$, where $h(\boldsymbol{x}) > 0$ implies a prediction of $y = +1$, and $h(\boldsymbol{x}) \leq 0$ implies a prediction of $y = -1$. The classification *margin* achieved by such a candidate is $y\, h(\boldsymbol{x})$, and the importance of the margin in terms of evaluating algorithm performance has been recognized for many years (Anthony and Bartlett, 1999; Langford and Shawe-Taylor, 2002). The work of Koltchinskii and Panchenko (2002) provide risk bounds that depend on the empirical mean of $I\{y\, h(\boldsymbol{x}) \leq \gamma\}$, providing useful generalization bounds for existing procedures whose on-sample margin error can be controlled. Intuitively, one might expect that having larger minimum margins on average would lead to better off-sample generalization. However, influential work by Breiman (1999) showed that the problem is not so simple, demonstrating cases in which the margins achieved are higher, but generalization is worse. In response to this, Reyzin and Schapire (2006) make the important suggestion that it is not merely the location of the margins, but properties of the entire *margin distribution* that are important to generalization.

1.1 Related work

Here we review the technical literature closely related to our work. Starting with the proposal of Garg and Roth (2003), their main theoretical results are a bound on the misclassification risk $R(h) := \mathbf{P}\{y\, h(\boldsymbol{x}) < 0\}$ of $h(\boldsymbol{x}) = \langle \boldsymbol{w}, \boldsymbol{x} \rangle + b$ for any $\boldsymbol{w} \in \mathbb{R}^d$ and $b \in \mathbb{R}$. Assuming that $\|\boldsymbol{x}\| = 1$, and given $2n$ observations, with probability no less than $1 - 4\delta$, we have

$$R(h) \leq \widehat{R}(h) + \min_d \left(\mu_d(h) + 2\sqrt{\frac{(d+2)\log(ne/(d+2)) + \log(2\delta^{-1})}{2n}} \right)$$

where $\widehat{R}(h) = n^{-1} \sum_{i=1}^n I\{y_i\, h(\boldsymbol{x}_i) < 0\}$, and the $\mu_d(h)$ term takes the form

$$\mu_d(h) := \frac{2}{\delta n} \sum_{i=1}^{2n} \min \left\{ 1, 3 \exp\left(\frac{-h(\boldsymbol{x}_i)^2 d}{2(2 + |h(\boldsymbol{x}_i)|)^2} \right), \frac{2}{h(\boldsymbol{x}_i)^2 d} \right\}.$$

Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS) 2019, Naha, Okinawa, Japan. PMLR: Volume 89.

The projection error terms are derived from the fact that

$$\mathbf{P}\{h(\boldsymbol{x})\,\widetilde{h}(\widetilde{\boldsymbol{x}}) < 0\} \leq \min\left\{1, 3\exp\left(\frac{-h(\boldsymbol{x})^2 d}{2(2+|h(\boldsymbol{x})|)^2}\right), \frac{2}{h(\boldsymbol{x})^2 d}\right\}$$

where $\widetilde{h}(\widetilde{\boldsymbol{x}}) = \langle P\boldsymbol{w}, P\boldsymbol{x}\rangle + b$, and P is a $k \times d$ random matrix of independent Gaussian random variables, $N(0, 1/d)$. Probability here is over the random draw of the matrix elements. Based on these guarantees, they construct a new loss, defined by

$$l(h;\boldsymbol{z}) = \sum_{i\in\mathcal{I}_+} \exp\left(-\alpha h(\boldsymbol{x}_i)^2\right) + \sum_{i\in\mathcal{I}_-} \exp\left(-\beta y_i\, h(\boldsymbol{x}_i)\right),$$

where $\mathcal{I}_+$ and $\mathcal{I}_-$ are respectively the indices of correctly and incorrectly classified observations. For correctly classified examples, they seek to minimize the projection error bound, whereas for incorrectly classified examples, then use a standard exponential surrogate loss. Depending on what $k \leq d$ minimizes their upper bound, the dependence on the number of parameters may be better than $O(\sqrt{d})$, but a price is paid in the form of $O(1/\delta)$ dependence on the confidence and bounded $\boldsymbol{x}$. On the computational side, proper setting of α and β is non-trivial.

The work of Zhang and Zhou (2016) considers using first- and second-order moments of the margin distribution as relevant quantities to build an objective. Writing

$$\overline{m}(h) \coloneqq \frac{1}{n}\sum_{i=1}^n y_i\, h(\boldsymbol{x}_i)$$
$$\bar{v}(h) \coloneqq \frac{1}{n}\sum_{i=1}^n \left(y_i\, h(\boldsymbol{x}_i) - \overline{m}(h)\right)^2,$$

in the case of $h(\boldsymbol{x}) = \langle \boldsymbol{w}, \boldsymbol{x}\rangle$, they construct a loss

$$l(h;\boldsymbol{z}) = \frac{\|\boldsymbol{w}\|^2}{2} + \lambda_1\bar{v}(h) - \lambda_2\overline{m}(h) + \frac{\lambda_3}{n}\sum_{i=1}^n \max\{1 - y_i\, h(\boldsymbol{x}_i), 0\},$$

where the $\lambda_1, \lambda_2, \lambda_3$ are parameters to be set manually. The authors show how the optimization can be readily cast into an n-dimensional dual program of the form

$$\min_{\boldsymbol{\alpha}\in\mathbb{R}^n} \ \frac{1}{2}\boldsymbol{\alpha}^T U\boldsymbol{\alpha} + \boldsymbol{u}^T\boldsymbol{\alpha} \quad \text{s.t.} \quad 0 \leq \alpha_i \leq a_i, \quad i = 1, \ldots, n$$

for appropriate data-dependent matrix U, vector $\boldsymbol{u}$, and weight bounds a_i, and they give some examples of practical implementations using dual coordinate descent and variance-reduced stochastic gradient descent. In all cases, parameter settings are left up to the user. Furthermore, statistical guarantees leave something to be desired; the authors prove that for any $\widehat{\boldsymbol{\alpha}}$ satisfying their dual objective, risk bounds hold as

$$\mathbf{E}\, R(\widehat{\boldsymbol{\alpha}}) \leq \frac{1}{n}\,\mathbf{E}\left(\sum_{i\in\mathcal{I}_1}\widehat{\alpha}_i U_{i,i} + |\mathcal{I}_2|\right),$$

where expectation is taken with respect to the sample, $U_{i,i}$ are the diagonal elements of U, and the index sets are defined

$$\mathcal{I}_1 = \{i : 0 < \widehat{\alpha}_i < \lambda_3/n\}, \ \ \mathcal{I}_2 = \{i : \widehat{\alpha}_i = \lambda_3/n\}.$$

These bounds provide limited insight into how and when the algorithm performs well, and in practice the algorithm requires substantial effort for model selection.

Finally, we consider the analysis of Brownlees et al. (2015), which greatly extends foundational work done by Catoni (2012). Let $\varphi(u) = \max\{1-u, 0\}$ denote the hinge loss. Throughout this paper, we use $\widehat{\gamma}$ as a generic notation for Catoni estimators. In the case of the hinge-based loss, this is defined by

$$\text{any } \widehat{\gamma} \geq 0 \quad \text{s.t.} \ \sum_{i=1}^n \psi\left(\frac{\widehat{\gamma} - \varphi(y_i\, h(\boldsymbol{x}_i))}{s}\right) = 0 \tag{1}$$

where $s > 0$ is a scaling parameter, and ψ is a soft truncation function (see Figure 1) defined by

$$\psi(u) \coloneqq \begin{cases} u - u^3/6, & -\sqrt{2} \leq u \leq \sqrt{2} \\ 2\sqrt{2}/3, & u > \sqrt{2} \\ -2\sqrt{2}/3, & u < -\sqrt{2}. \end{cases} \tag{2}$$

This estimator depends on the choice of h, and Brownlees et al. (2015) provide tools for obtaining risk bounds for any procedure that minimizes $\widehat{\gamma}(h)$ as a function of $h \in \mathcal{H}$, where $\mathcal{H}$ denotes the hypothesis space our candidate lives in. Note that the 1-Lipschitz continuity of the hinge loss gives us that for any candidates g and h,

$$\begin{aligned}|\varphi(y\, g(\boldsymbol{x})) - \varphi(y\, h(\boldsymbol{x}))| &\leq |y||g(\boldsymbol{x}) - h(\boldsymbol{x})| \\ &= |g(\boldsymbol{x}) - h(\boldsymbol{x})|,\end{aligned}$$

which means we can bound distances defined on the space $\{f(\boldsymbol{x}) = \varphi(y\, h(\boldsymbol{x})) : h \in \mathcal{H}\}$ by distances on the space $\mathcal{H}$. Going back to the linear model case of $h(\boldsymbol{x}) = \langle \boldsymbol{w}, \boldsymbol{x}\rangle$, bounds in the $\mathcal{L}_2$ distance d_2 can be constructed using

$$\begin{aligned}\mathbf{E}\,|\varphi(y\, g(\boldsymbol{x})) - \varphi(y\, h(\boldsymbol{x}))|^2 &\leq \mathbf{E}\,|g(\boldsymbol{x}) - h(\boldsymbol{x})|^2 \\ &\leq \|\boldsymbol{w}_g - \boldsymbol{w}_h\|^2\,\mathbf{E}\,\|\boldsymbol{x}\|^2,\end{aligned}$$

and bounds in the $\mathcal{L}_\infty$ distance take the form

$$\sup_{\boldsymbol{x}} |\varphi(y\, g(\boldsymbol{x})) - \varphi(y\, h(\boldsymbol{x}))| \leq \sup_{\boldsymbol{x}} |g(\boldsymbol{x}) - h(\boldsymbol{x})| \leq \|\boldsymbol{w}_g - \boldsymbol{w}_h\| \sup_{\boldsymbol{x}} \|\boldsymbol{x}\|.$$

Now, using their results, for large enough s and n, one can show that with probability no less than $1-\delta$, it holds that

$$\mathbf{E}\,\varphi\left(y\,\widehat{h}(\boldsymbol{x})\right) - \inf_{h\in\mathcal{H}} \mathbf{E}\,\varphi(y\,h(\boldsymbol{x})) \leq O\left(\sqrt{\frac{\log(3\delta^{-1})}{n}} + \log(2\delta^{-1})\left(\frac{\eta_2(\mathcal{H})}{\sqrt{n}} + \frac{\eta_\infty(\mathcal{H})}{n}\right)\right),$$

where $\eta_2(\mathcal{H})$ and $\eta_\infty(\mathcal{H})$ are complexity terms. When these terms can be bounded, we can use the fact that the hinge loss is "classification calibrated," and using standard results from Bartlett et al. (2006), can obtain bounds on the excess misclassification risk based on the above inequality. The problem naturally is how to control these complexity terms. Skipping over some technical details, these terms can be bounded using covering number integrals dependent on $\mathcal{H}$. As a concrete example, we have

$$\eta_\infty(\mathcal{H}) \leq c_\infty \int_0^{\Delta(\mathcal{H};d_\infty)} \log N(\epsilon, \mathcal{H}, d_\infty)\, d\epsilon,$$

where $d_\infty(g,h) = \sup_{\boldsymbol{x}} |g(\boldsymbol{x}) - h(\boldsymbol{x})|$ is the $\mathcal{L}_\infty$ metric on $\mathcal{H}$, the covering number $N(\epsilon, \mathcal{H}, d_\infty)$ is the number of ϵ-balls in the d_∞ metric needed to cover $\mathcal{H}$, and $\Delta(\mathcal{H}; d_\infty) = \sup\{d_\infty(g,h) : g, h \in \mathcal{H}\}$. In the case of $h(\boldsymbol{x}) = \langle \boldsymbol{w}, \boldsymbol{x}\rangle$, this means $\|\boldsymbol{x}\|$ must be almost surely bounded in order for the $\mathcal{L}_\infty$ distance to be finite and the upper bounds to be meaningful. Under such assumptions, say $\boldsymbol{w}$ comes from the unit ball and $\|\boldsymbol{x}\| \leq B_X$ almost surely. Then ignoring non-dominant terms, the high-probability upper bound takes the form

$$O\left(\sqrt{\frac{\log(3\delta^{-1})}{n}} + \frac{\log(2\delta^{-1}) d B_X}{\sqrt{n}}\right).$$

While extremely flexible and applicable to a wide variety of learning tasks and algorithms, for the classification task, getting around the bound on $\boldsymbol{x}$ is impossible using the machinery of Brownlees et al. (2015). Even more serious complications are introduced by the difficulty of computation: while simple fixed-point procedures can be used to accurately approximate the robust objective $\widehat{\gamma}(h)$, it cannot be expressed explicitly, and indeed need not be convex as a function defined on $\mathcal{H}$, even in the linear model case. Approximation error is unavoidable due to early stopping, and in addition to this computational overhead, using non-linear solvers to minimize the function $\widehat{\gamma}(h)$ can be costly and unstable in high-dimensional tasks (Holland and Ikeda, 2017). A recent pre-print from Lecué et al. (2018) considers replacing the M-estimator of Brownlees et al. (2015) with a median-of-means risk estimate, which does not require bounded inputs to get strong guarantees, but which requires an expensive iterative subroutine for every loss evaluation, leading to substantial overhead for even relatively small learning tasks.

Our contributions To deal with the limitations of existing procedures highlighted above, the key idea here is to introduce a new convex loss that encourages the distribution of the margin to be tightly concentrated near a certain prescribed level. The procedure is easily implemented using gradient descent, admits formal performance guarantees reflecting both computational cost and optimization error, and aside from the usual cost of gradient computation there is virtually no computational overhead. Two key highlights are:

- The proposed algorithm enjoys high-probability risk bounds under moment bounds on $\boldsymbol{x}$, and does not require $\|\boldsymbol{x}\|$ to be bounded.
- Numerical experiments show how a simple data-dependent re-scaling procedure can reduce the need for trial-and-error tuning of regularization.

2 Proposed algorithm

We would like to utilize the strong elements of the existing procedures cited, while addressing their chief weaknesses. To do so, we begin by integrating the Catoni influence function ψ defined in (2), which results in a new function of the form

$$\rho(u) := \begin{cases} \frac{u^2}{2} - \frac{u^4}{24} & |u| \leq \sqrt{2}, \\ |u|\frac{2\sqrt{2}}{3} - \frac{1}{2} & |u| > \sqrt{2}. \end{cases} \tag{3}$$

Note that $\rho'(u) = \psi(u)$ for all $u \in \mathbb{R}$. This function satisfies $\rho(u) \geq 0$, is symmetric about zero so $\rho(u) = \rho(-u)$, and since the absolute value of the slope is bounded by $|\rho'(u)| \leq 2\sqrt{2}/3$, we have that ρ is Lipschitz continuous, namely that for any $u, v \in \mathbb{R}$, we have $|\rho(u) - \rho(v)| \leq (2\sqrt{2}/3)|u - v|$.

Recalling the Catoni estimator (1) used by Brownlees et al. (2015), we define a new objective which is closely related:

$$Q(h;\gamma) := \frac{s^2}{n}\sum_{i=1}^n \rho\left(\frac{\gamma - y_i\, h(\boldsymbol{x}_i)}{s}\right). \tag{4}$$

Here $\gamma \in \mathbb{R}$ is the desired margin level, and once again $s > 0$ is a re-scaling parameter. Note that this loss

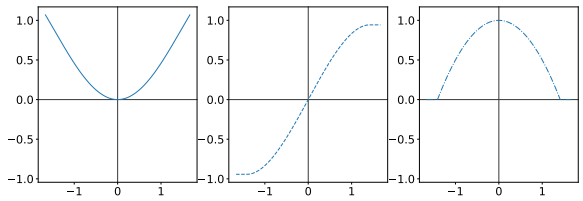

Figure 1: Graphs of ρ, $\rho' = \psi$ and ρ'' near the origin.

penalizes not only incorrectly classified examples, but also examples which are correctly classified, but *overconfident*. The intuition here is that by also penalizing overconfident correct examples to some degree, we seek to constrain the variance of the margin distribution. The nature of this penalization is controlled by γ: a larger value leads to less correct examples being penalized.

It remains to set the scale s. To do so, first note that for any candidate h, the Catoni estimator $\widehat{\gamma}(h)$ of $\mathbf{E}\, y\, h(\boldsymbol{x})$ minimizes $Q(h;\gamma)$ as a function of γ, and enjoys a pointwise error bound dependent on s, which says

$$|\widehat{\gamma}(h) - \mathbf{E}\, y\, h(\boldsymbol{x})| \leq \frac{\operatorname{var} y\, h(\boldsymbol{x})}{s} + \frac{2s\log(2\delta^{-1})}{n}, \quad (5)$$

with probability no less than $1-\delta$. Minimizing this bound in $s > 0$ naturally leads to setting $s^2 = n \operatorname{var} y\, h(\boldsymbol{x})/2\log(2\delta^{-1})$, but in our case, a certain amount of bias is assuredly tolerable; say a certain fraction $1/k$ of the desired γ setting, plus error that vanishes as $n \to \infty$. By setting $s \geq \operatorname{var} y\, h(\boldsymbol{x})\, k/\gamma$ then, we have

$$|\widehat{\gamma}(h) - \mathbf{E}\, y\, h(\boldsymbol{x})| \leq \frac{\gamma}{k} + O\left(\frac{1}{n}\right).$$

The exact setting of $s > 0$ plays an important role both in theory and in practice; we shall look at this in more detail in sections 3–4. In practice, the true variance will of course be unknown, but we can replace the true variance with any valid upper bound on the variance; rough estimates are easily constructed using moments of the empirical distribution (see section 4).

With scaling taken care of, our proposed algorithm is simply to minimize the new loss (4) using gradient descent, namely to run the iterative update

$$\widehat{h}_{(t+1)} = \widehat{h}_{(t)} - \alpha_{(t)} \nabla Q(\widehat{h}_{(t)};\gamma),$$

where $\alpha_{(t)}$ are step sizes. We summarize the key computations in Algorithm 1 for the case of a linear model $h(\boldsymbol{x}) = \langle \boldsymbol{w}, \boldsymbol{x}\rangle$ with fixed step sizes.

Remark 1 (Algorithm 1 and distribution control). Intuitively, in running Algorithm 1 (or any generalization of it), the expectation is that with enough iterations, the approximation $\widehat{\gamma}(\widehat{h}_{(t)}) \approx \gamma$ should be rather

Algorithm 1 Margin pursuit by steepest descent.

input: $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n) \in \mathbb{R}^d \times \{-1, 1\}$

parameters: $\widehat{\boldsymbol{w}}_{(0)} \in \mathbb{R}^d$, $\gamma \in \mathbb{R}$, $k > 0$, $\alpha > 0$

scaling: $s \geq \operatorname{var} y\, h(\boldsymbol{x})\, k/\gamma$

for $t = 0, 1, \ldots, T-1$ **do**

$$\widehat{\boldsymbol{g}}_{(t)} = -\frac{s}{n}\sum_{i=1}^{n} \psi\left(\frac{\gamma - y_i \langle \widehat{\boldsymbol{w}}_{(t)}, \boldsymbol{x}_i\rangle}{s}\right) y_i \boldsymbol{x}_i$$

$$\widehat{\boldsymbol{w}}_{(t+1)} = \widehat{\boldsymbol{w}}_{(t)} - \alpha\, \widehat{\boldsymbol{g}}_{(t)}$$

end for

sharp, although arbitrary precision assuredly cannot be guaranteed. If the γ level is set too high given a hypothesis class $\mathcal{H}$ with low complexity, regardless of the choice of $h \in \mathcal{H}$, we cannot expect γ to be near the location of the margin $y\, h(\boldsymbol{x})$, which is accurately approximated by $\widehat{\gamma}(h)$. This can be easily proven: there exists a set of classifiers $\mathcal{H}$ and distribution μ under which even a perfect optimizer of the new risk has a Catoni-type estimate smaller than γ (proof in supplement).

If the approximation $\widehat{\gamma}(\widehat{h}_{(t)}) \approx \gamma$ actually is sharp, how does this relate to *control* of the margin distribution? By design, the estimator $\widehat{\gamma}(\cdot)$ is resistant to errant observations and is located near the *majority* of observations (see Proposition 2), if it turns out that $\widehat{\gamma}(\widehat{h}_{(t)})$ is close to γ, then it is *not possible* for the majority of margin points be much smaller (or much larger) than γ.[1] Conceptually, the desired outcome is similar to that of the procedure of Brownlees et al. (2015) discussed in section 1.1, but with an easy implementation and more straightforward statistical analysis. In section 3, we show that risk bounds are readily available for the proposed procedure, even without a bound on the inputs $\boldsymbol{x}$. Empirical analysis in section 4 illustrates the basic mechanisms underlying the algorithm, using real-world benchmark data sets.

3 Theoretical analysis

Notation For positive integer k, write the set of all positive integers no greater than k by $[k] \coloneqq \{1, \ldots, k\}$. The underlying distribution of interest is that of $(\boldsymbol{x}, y)$, here taking values on $\mathbb{R}^d \times \{-1, 1\}$. The data sample refers to n independent and identically distributed ("iid") copies of $(\boldsymbol{x}, y)$, denoted $(\boldsymbol{x}_i, y_i)$ for $i \in [n]$. Let $\mathcal{H}$ denote a generic class of functions $h : \mathbb{R}^d \to \mathbb{R}$. The running assumption will be that all $h \in \mathcal{H}$ are mea-

[1] Note that we still cannot rule out the possibility that the margin distribution is spread out over a wide region; a simple example is the case where the margins are symmetrically distributed around γ.

surable, and at the very least satisfy $\mathbf{E}\,|h(\boldsymbol{x})|^2 < \infty$. Denote the input variance by $v_X := \mathbf{E}\,\|\boldsymbol{x}\|^2$, and any valid median of a set A by $\operatorname{med} A$.

Scaling and location estimates Our chief interest from a theoretical standpoint is in statistical properties of Algorithm 1, in particular we seek high-probability upper bounds on the excess risk of the procedure after T iterations, given n observations, that depend on T, n, and low-order moments of the underlying distribution. We begin with some statistical properties of the motivating estimator, and a look at how scale settings impact these properties.

Proposition 2 (Scaling and location estimates)**.** *For any $h \in \mathcal{H}$ and scale $s > 0$, the estimate $\widehat{\gamma}(h)$ satisfies the following:*

1. *There exists $0 < s' < \infty$ such that for all $0 < s \leq s'$, we have $\widehat{\gamma}(h) = \operatorname{med}\{y_i\, h(\boldsymbol{x}_i)\}_{i\in[n]}$.*
2. *There exists a constant $c > 0$ such that for all $s > 0$,*

$$\left|\widehat{\gamma}(h) - \frac{1}{n}\sum_{i=1}^{n} y_i\, h(\boldsymbol{x}_i)\right| \leq \frac{c}{s^2}.$$

Remark 3. The basic facts laid out in Proposition 1 illustrate how s controls the "bias" of the Catoni estimator. A larger scale factor makes the estimator increasingly sensitive to errant data, and causes it to close in on the empirical mean. A sufficiently small value on the other hand causes the estimator to effectively ignore the distribution tails, closing in on the empirical median.

Proposition 4 (Scaling and stability)**.** *Given any dataset $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$ and candidate $h \in \mathcal{H}$, construct $\widehat{\gamma}(h)$ as usual. Then consider a modified dataset $\boldsymbol{z}'_1, \ldots, \boldsymbol{z}'_n$, which is identical to the original except for one point, subject to arbitrary perturbation. Let $\widehat{\gamma}'(h)$ denote the estimator under the modified data set. Defining a sub-index as*

$$\mathcal{I} := \left\{i \in [n] : |\widehat{\gamma}(h) - y_i\, h(\boldsymbol{x}_i)| \leq s\sqrt{2}/2\right\},$$

it follows that whenever n and s are large enough that $|\mathcal{I}| \geq n/2 > 24$, we have

$$|\widehat{\gamma}(h) - \widehat{\gamma}'(h)| \leq \frac{s}{\sqrt{n}}.$$

Remark 5. The stability property highlighted in Proposition 4 is appealing because the difference $\max\{|y_i\, h(\boldsymbol{x}_i) - y'_i\, h(\boldsymbol{x}'_i)| : i \in [n]\}$ could be arbitrarily large, while the estimator $\widehat{\gamma}(h)$ in shifting to $\widehat{\gamma}'(h)$ remains close to the majority of the points, and cannot be drawn arbitrarily far away. For clarity, we have considered the case of just one modified point, but a brief glance at the proof (in the appendix) should demonstrate how analogous results can readily be obtained for the case of larger fractions of modified points.

Lemma 6 (Pointwise error bound)**.** *Fixing any $h \in \mathcal{H}$, consider the estimate $\widehat{\gamma}(h)$ defined in (1), equivalently characterized as a minimizer of $Q(h;\gamma)$ in γ, with scaling parameter s set such that $s^2 = nv/2\log(2\delta^{-1})$, where v is any upper bound $\operatorname{var} y\, h(\boldsymbol{x}) \leq v < \infty$. It follows that*

$$\mathbf{P}\left\{|\widehat{\gamma}(h) - \mathbf{E}\, y\, h(\boldsymbol{x})| > \sqrt{\frac{2v\log(2\delta^{-1})}{n}}\right\} \leq \delta.$$

Remark 7. The confidence interval in Lemma 6 is called pointwise because it holds for a pre-fixed $h \in \mathcal{H}$, in contrast with uniform bounds that hold independent of the choice of h. When considering our Algorithm 1, the candidate h will be data-dependent and thus random, meaning that pointwise bounds will have to be extended to cover all possible contingencies; see the proof of Theorem 11 for details.

Classification-calibrated loss Proceeding with our analysis, the ultimate evaluation metric of interest here is the classification risk (expectation of the zero-one loss), denoted

$$R(h) := \mathbf{P}\{\operatorname{sign}(h(\boldsymbol{x})) \neq y\}, \quad R^* := \inf_{h\in\mathcal{H}} R(h). \tag{6}$$

Using empirical estimates of the zero-one loss is not conducive to efficient learning algorithms, and our Algorithm 1 involves the minimization of a new loss $Q(\cdot;\gamma)$, defined in equation (4). To ensure that good performance in this metric implies low classification risk, the first step is to ensure that the function is *calibrated* for classification, in the sense of Bartlett et al. (2006). To start, fixing any $\gamma > 0$, define $\varphi(u) := s^2\,\rho((\gamma - u)/s)$. This furnishes the surrogate risk

$$R_\varphi(h) := \mathbf{E}\,\varphi\left(y\, h(\boldsymbol{x})\right), \quad R^*_\varphi := \inf_{h\in\mathcal{H}} R_\varphi(h). \tag{7}$$

The basic idea is that if this loss φ is calibrated, then one can show that there exists a function $\Psi_{s,\gamma}$ depending on user-specified γ and s settings, which is non-decreasing on the positive real line and satisfies

$$\Psi_{s,\gamma}(R(h) - R^*) \leq R_\varphi(h) - R^*_\varphi.$$

Our loss function ρ defined in 3 is congenial due to the fact that it is classification-calibrated, with a Ψ-transform $\Psi_{s,\gamma}(\cdot)$ that can be computed exactly, for arbitrary values of $\gamma > 0$ and $s > 0$. Details of this computation are not difficult, but are rather tedious, and thus we relegate them to the supplement. Basic facts are summarized in the following lemma.

Lemma 8. *The loss function $\varphi(u) := s^2\,\rho((\gamma-u)/s)$ is classification calibrated such that for each $\gamma > 0$, the following statements hold.*

1. *Ψ-transform: there exists a function $\Psi_{s,\gamma} : [0,1] \to \mathbb{R}_+$ for which $\Psi_{s,\gamma}(R(h) - R^*) \leq R_\varphi(h) - R_\varphi^*$, depending on ρ, s, γ, and a concave function $H_{s,\gamma}(\cdot)$ defined on $[0,1]$, specified in the proof. This Ψ-transform function takes the form*

$$\Psi_{s,\gamma}(u) = s^2\rho(\gamma/s) - H_{s,\gamma}\left(\frac{1+u}{2}\right).$$

2. *Risk convergence: given a sequence $(\widehat{h}_n)$ of sample-dependent $\{\boldsymbol{z}_1,\ldots,\boldsymbol{z}_n\} \mapsto \widehat{h}_n$, we have that convergence in our surrogate is sufficient for convergence in the zero-one risk, namely*

$$\left\{\lim_{n\to\infty} R_\varphi(\widehat{h}_n) = R_\varphi^*\right\} \subseteq \left\{\lim_{n\to\infty} R(\widehat{h}_n) = R^*\right\}.$$

3. *Invertibility: $\Psi_{s,\gamma}(u)$ is invertible on $[0,1]$, and thus for small enough excess risk, we can bound as $R(h) - R^* \leq \Psi_{s,\gamma}^{-1}(R_\varphi(h) - R_\varphi^*)$.*

Remark 9 (Generalization and γ level setting). One would naturally expect that all else equal, if a classifier achieves the same excess φ-risk for a larger value of γ, then the resulting excess classification risk should be smaller, or at least no larger. More concretely, we should expect that

$$\gamma \leq \gamma' \implies \Psi_{s,\gamma}^{-1}(a) \geq \Psi_{s,\gamma'}^{-1}(a), \quad a \in [0, s^2\,\rho(\gamma/s)].$$

This range comes from the fact that $\Psi_{s,\gamma}(0) = 0$ and $\Psi_{s,\gamma}(1) = s^2\,\rho(\gamma/s)$. This monotonicity follows from the definition of ρ and the convexity of the Ψ-transform.

Assumptions and risk bounds, with discussion With preparatory results in place, we can now pursue an excess risk bound for Algorithm 1. To make notation more transparent, we accordingly write $R(\boldsymbol{w})$ and $R_\varphi(\boldsymbol{w})$ to denote the respective risks under $\mathcal{H} = \{h : h(\boldsymbol{x}) = \langle \boldsymbol{w}, \boldsymbol{x}\rangle, \boldsymbol{w} \in \mathcal{W}\}$, where $\mathcal{W} \subset \mathbb{R}^d$. The core technical assumptions are as follows:

A0. $\mathcal{W}$ is a compact subset of $\mathbb{R}^d$, with diameter $\Delta := \sup\{\|\boldsymbol{u} - \boldsymbol{v}\| : \boldsymbol{u}, \boldsymbol{v} \in \mathcal{W}\} < \infty$.

A1. There exists $\boldsymbol{w}^* \in \mathcal{W}$ at which $R_\varphi'(\boldsymbol{w}^*) = 0$.

A2. $R_\varphi(\boldsymbol{w})$ is κ-strongly convex on $\mathcal{W}$, with minimum[2] denoted by $\boldsymbol{w}^*$.

A3. Writing $\boldsymbol{b} := -\rho'(\gamma - y\langle \boldsymbol{w}, \boldsymbol{x}\rangle)y\,\boldsymbol{x}$ for the new loss gradient before scaling by s, and Σ for its covariance matrix, there exists some $c > 0$ such that for all $\boldsymbol{w} \in \mathcal{W}$, $a \geq 0$, and $\|\boldsymbol{u}\| = 1$, we have $\mathbf{E}\exp\left(a\langle \boldsymbol{u}, \boldsymbol{b} - \mathbf{E}\,\boldsymbol{b}\rangle\right) \leq \exp(ca^2\langle \boldsymbol{u}, \Sigma\boldsymbol{u}\rangle)$.

Remark 10 (Feasibility of assumptions). The important assumptions here are A2 and A3. The latter can be satisfied with inputs $\boldsymbol{x}$ that have sub-Gaussian tails; this does not include data with higher-order moments that are infinite, but requires no bound on $\|\boldsymbol{x}\|$ at all. As for the former assumption A2, first note that the (i,j)th element of the Hessian of the new loss function is

$$\frac{\partial^2}{\partial w_i \partial w_j} s^2\,\rho\left(\frac{\gamma - y\,\langle \boldsymbol{w}, \boldsymbol{x}\rangle}{s}\right) = \rho''\left(\frac{\gamma - y\,\langle \boldsymbol{w}, \boldsymbol{x}\rangle}{s}\right) x_i x_j$$

for $i,j \in [d]$, where $\rho''(u) = 1 - u^2/2$ for $|u| \leq \sqrt{2}$, and zero otherwise. Write $q = \boldsymbol{u}^T(\boldsymbol{x}\boldsymbol{x}^T)\boldsymbol{u}$ and $r = \rho''((\gamma - y\,\langle \boldsymbol{w}, \boldsymbol{x}\rangle)/s)$ for readability, and use $\mathbf{E}_+$ and $\mathbf{E}_-$ to denote integration over the positive and non-positive parts of q. First, observe that

$$\mathbf{E}_-\, rq = \mathbf{E}\, I\{q \leq 0\} rq \geq \mathbf{E}\, I\{q \leq 0\} q = \mathbf{E}\, q - \mathbf{E}_+\, q.$$

Using this inequality, we have

$$\begin{aligned}
\boldsymbol{u}^T R_\varphi''(\boldsymbol{w})\boldsymbol{u} &= \mathbf{E}\, rq \\
&= \mathbf{E}_+\, rq + \mathbf{E}_-\, rq \\
&\geq \mathbf{E}_+\, \rho''\left(\frac{\gamma - y\,\langle \boldsymbol{w}, \boldsymbol{x}\rangle}{s}\right) q + (\mathbf{E}\, q - \mathbf{E}_+\, q) \\
&= \mathbf{E}\, q + \mathbf{E}_+\left(\rho''\left(\frac{\gamma - y\,\langle \boldsymbol{w}, \boldsymbol{x}\rangle}{s}\right) - 1\right) q.
\end{aligned}$$

The second term on the right-hand side is a negative value that can be taken near zero for any $\boldsymbol{w} \in \mathcal{W}$ by taking $s > 0$ large enough. The first term is $\mathbf{E}\, q = \boldsymbol{u}^T\,\mathbf{E}\,\boldsymbol{x}\boldsymbol{x}^T\boldsymbol{u}$, and thus with large enough s, as long as the second moment matrix of the inputs is positive definite satisfying $\mathbf{E}\,\boldsymbol{x}\boldsymbol{x}^T \succeq cI_d$ for some $c > 0$ (a weak assumption), it follows that there exists a $\kappa > 0$ such that $R_\varphi''(\boldsymbol{w}) \succeq \kappa I_d$ holds. Since the risk is twice continuously differentiable, This implies κ-strong convexity (Nesterov, 2004, Theorem 2.1.11).

With these assumptions in place, finite-sample risk bounds can be obtained.

Theorem 11. *Running Algorithm 1 for T iterations, the final output produced, written $\widehat{\boldsymbol{w}}_{(T)}$, for constant $c > 0$ and $\beta := 2\kappa v_X/(\kappa + v_X)$ satisfies*

$$\Psi_{s,\gamma}\left(R(\widehat{\boldsymbol{w}}_{(T)}) - R^*\right) \leq$$
$$(1 - \alpha\beta)^T v_X \|\widehat{\boldsymbol{w}}_{(0)} - \boldsymbol{w}^*\|^2 + \frac{4v_X}{\beta^2 n}\left((1+\delta)v_X + 2s\,\varepsilon^*\right)^2$$

with probability no less than $1 - 2\delta$ over the random draw of the sample, where the dominant term ε^* is

[2] Assuming we can take the derivative under the integral, the smoothness of ρ implies differentiability of R_φ. Then using the compactness of $\mathcal{W}$, it follows that $\boldsymbol{w}^* \in \mathcal{W}$.

defined

$$\varepsilon^* := \sqrt{c\rho'(\sqrt{2})^2 \, \mathbf{E}\,\|\boldsymbol{x}\boldsymbol{x}^T\|(d\log(3\sqrt{n}(2\delta)^{-1}) + \log(\delta^{-1}))}.$$

Remark 12 (Interpretation and tradeoffs). Excess risk bounds give in Theorem 11 are composed of two key terms, one of a computational nature, and one of a statistical nature. The first term is optimization error, which decreases as T grows, and depends on the initial estimate $\widehat{\boldsymbol{w}}_{(0)}$, the step-size α, and the convexity of the surrogate risk through β. The second term is statistical error, and depends on the sample size, scale s, the number of parameters, and second-order moments of the inputs $\boldsymbol{x}$. Note that there is a clear tradeoff due to s: a sufficiently large scale factor is needed to ensure A2 holds (yielding large enough β), but setting s too large impacts the statistical error in a negative way.

Finally, we note the d factor in ε^* is due to a covering number argument used to obtain a bound on the empirical gradient error that holds uniformly over $\boldsymbol{w} \in \mathcal{W}$. Does there exist another computational procedure, with the *same optimization error*, and without this seemingly superfluous d factor in the statistical error? We pursue such analysis in future work.

4 Empirical analysis

In our numerical experiments, we aim to complement the theoretical analysis carried out in the previous section. We look at how algorithm parameter settings impact generalization guarantees, and using real-world datasets, investigate how Algorithm 1 performs, comparing its behavior with a benchmark procedure.

Benchmark data tests: experimental setup In all the experiments discussed here, we consider binary classification on real-world data sets, modified to control for unbalanced ratios of positive and negative labels. Training for each data set is done using pair $(\boldsymbol{X}, \boldsymbol{y})$, where $\boldsymbol{X}$ is $n \times d$, and $\boldsymbol{y}$ is $n \times 1$, and testing is done on a disjoint subset. The train-test sequence is repeated over 25 trials, and all numerical performance metrics displayed henceforth should be assumed to be averages taken over all trials.

We use four data sets, denoted COV, DIGIT5, PROTEIN, and SIDO, creating subsets under the following constraints: (1) Sample size n is no more than ten times the nominal dimension d, and (2) both the training and testing data sets have balanced ratios of labels (as close as possible to 50% each). Starting with COV ($n = 540$, $d = 54$, non-zero: 22%), this is the "Forest CoverType dataset" on the UC Irvine repository, converted into a binary task identifying class 1 against the rest. DIGIT5 ($n = 5000$, $d = 784$, non-zero: 19%) is the MNIST hand-written digit data, converted into a binary task for the digit 5. PROTEIN ($n = 740$, $d = 74$, non-zero: 99%) is the protein homology dataset (KDD Cup 2004). SIDO ($n = 425$, $d = 4932$, non-zero: 11%) is the molecular descriptor data set (NIPS 2008 causality challenge), with binary-valued features. In each trial, from the full original data set, we take a random sub-sample of the specified size, without replacement, for training, and for test data we use as much of the remaining data as possible, within the confines of constraint (2) above.

As a well-known benchmark algorithm against which we can compare the behaviour and performance of the proposed Algorithm 1, we implement and run the well-known Pegasos algorithm of Shalev-Shwartz et al. (2011). For both methods, the initial value $\widehat{\boldsymbol{w}}_{(0)}$ is determined randomly in each trial. We explore multiple settings of Algorithm 1 described further below, but in all cases we take the stochastic optimization approach: instead of using all n training examples at each step, we randomly select one at a time for computing the update direction, and use a step size of . For direct comparison with Pegasos, we set the margin level to $\gamma = 1$, add a squared ℓ_2-norm regularization term with coefficient λ, utilizing a step size of $\alpha = (s\sqrt{\lambda}(1+t))^{-1}$, and projecting to the $1/\sqrt{\lambda}$-radius ball. That is, we run a stochastic projected gradient descent version of Algorithm 1, and evaluate the impact of the proposed loss function.

Benchmark data tests: generalization with naive scaling We begin with the simplest setting of Algorithm 1, where $s = 1$ is fixed throughout. In Figure 2, we plot training error, test error, and location statistics of the empirical margin distribution, all as a function of cost incurred (equal to number of gradients computed). For each dataset, we experimented with $\lambda \in \{10^0, 10^{-6}, 10^{-6}, \ldots, 10^{-1}\}$ and display the results for the case of λ that resulted in the best performance, as measured by the lowest test error achieved over all iterations.

We see that our proposed procedure is as good or better than the best setting of Pegasos, and results in a margin distribution very distinct from that of the competing procedure. On the whole, we see a much more symmetrical distribution, with smaller variance, that over iterations pushes the margin location up in a monotonic fashion, in stark contrast to that of Pegasos, whose empirical distribution peaks early and slowly settles down over time. The smaller variance and higher degree of symmetry is precisely what we would expect given the definition of ρ, which assigns a penalty for correctly classified examples that are over-

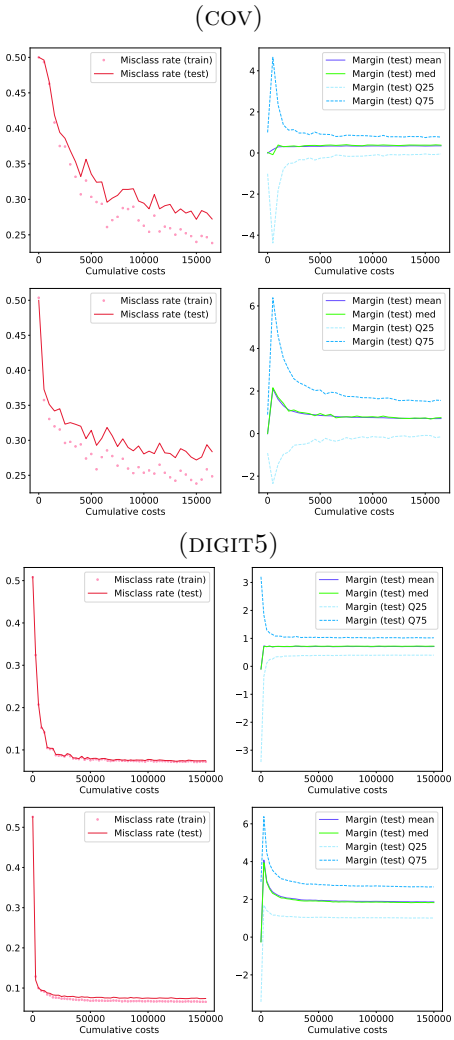

Figure 2: For each dataset, the top row is Algorithm 1, bottom row is Pegasos.

confidently classified, as discussed in section 2.

Benchmark data tests: scaling and regularization Next, we look at the impact of a fixed scale, determined by observed data, as follows. Each run of Algorithm 1 starts with $s = 1$ fixed just as in the previous tests, but after a pre-fixed number of steps, updates the scale just once, to take a value of $s \geq \sqrt{n v_X / (2\lambda \log(\delta^{-1}))}$ (see Lemma 6), where v_X is approximated using the 75th quantile of the empirical distribution induced by $\{|y_i \langle \widehat{\boldsymbol{w}}_{(t)}, \boldsymbol{x}_i \rangle| : i \in [n]\}$. This time, we intentionally under-regularize, setting λ at less than 1/100th of the best setting found in the previous tests. Representative results are given in Figure 3.

When highly under-regularized, *and* without scaling, the learning algorithm just wanders about, overwhelmed by the variance of the per-iteration sub-

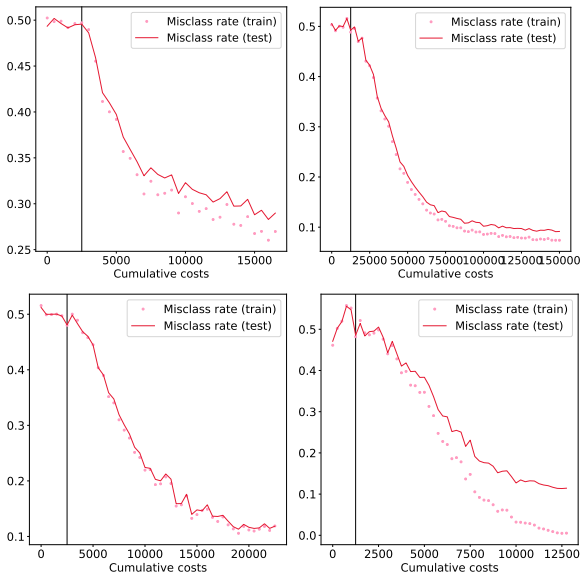

Figure 3: Algorithm 1 with data-based s setting starting from the point marked by a black vertical line. From first row, reading left to right, COV, DIGIT5, PROTEIN (all $\lambda = 10^{-5}$), and SIDO ($\lambda = 10^{-3}$).

sampling; when the procedure is left to run like this, a good solution can rarely be found before the step size grows small, highly inefficient. On the other hand, using the simple data-driven scaling procedure just described to fix a "safe" value of s, we find that the learning algorithm is almost immediately accelerated, and in less time essentially catches up with the performance achieved under the best regularization possible. This is extremely encouraging, as it suggests that a safe, inexpensive, automated scaling procedure can make up for our lack of knowledge about the ideal regularization parameter, allowing for potentially significant savings in hyper-parameter exploration.

5 Concluding remarks

In this paper, we introduced and analyzed a learning algorithm based on a new convex loss and re-scaled margin deviations. Statistical guarantees are available for a routine which can be easily implemented as-is, and practical utility in experiments using real-world datasets was confirmed in our empirical analysis. As a natural future line of work, consideration of a procedure which iteratively minimizes our loss, checks $\widehat{\gamma}$, and updates the desired γ threshold would be an interesting next step to take this work in.

Acknowledgements

This work was supported in part by JSPS *KAKENHI* Grant Number 18H06477.

References

Anthony, M. and Bartlett, P. L. (1999). *Neural Network Learning: Theoretical Foundations*. Cambridge University Press.

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

Breiman, L. (1999). Prediction games and arcing algorithms. *Neural Computation*, 11(7):1493–1517.

Brownlees, C., Joly, E., and Lugosi, G. (2015). Empirical risk minimization for heavy-tailed losses. *Annals of Statistics*, 43(6):2507–2536.

Catoni, O. (2012). Challenging the empirical mean and empirical variance: a deviation study. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 48(4):1148–1185.

Garg, A. and Roth, D. (2003). Margin distribution and learning. In *Proceedings of the 20th International Conference on Machine Learning*, pages 210–217.

Holland, M. J. and Ikeda, K. (2017). Efficient learning with robust gradient descent. *arXiv preprint arXiv:1706.00182*.

Koltchinskii, V. and Panchenko, D. (2002). Empirical margin distributions and bounding the generalization error of combined classifiers. *Annals of Statistics*, 30(1):1–50.

Langford, J. and Shawe-Taylor, J. (2002). PAC-Bayes & Margins. In *Advances in Neural Information Processing Systems 15*, pages 439–446.

Lecué, G., Lerasle, M., and Mathieu, T. (2018). Robust classification via mom minimization. *arXiv preprint arXiv:1808.03106*.

Nesterov, Y. (2004). *Introductory Lectures on Convex Optimization: A Basic Course*. Springer.

Reyzin, L. and Schapire, R. E. (2006). How boosting the margin can also boost classifier complexity. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 753–760.

Shalev-Shwartz, S., Singer, Y., Srebro, N., and Cotter, A. (2011). Pegasos: Primal Estimated sub-GrAdient SOlver for SVM. *Mathematical Programming*, 127(1):3–30.

Zhang, T. and Zhou, Z.-H. (2016). Optimal margin distribution machine. *arXiv preprint arXiv:1604.03348*.