Narrow Margins: Classification, Margins and Fat Tails

Francois Buet-Golfouse [1]

Abstract

It is well-known that, for separable data, the regularised two-class logistic regression or support vector machine re-normalised estimate converges to the maximal margin classifier as the regularisation hyper-parameter λ goes to 0. The fact that different loss functions may lead to the same solution is of theoretical and practical relevance as margin maximisation allows more straightforward considerations in terms of generalisation and geometric interpretation. We investigate the case where this convergence property is not guaranteed to hold and show that it can be fully characterised by the distribution of error terms in the latent variable interpretation of linear classifiers. In particular, if errors follow a regularly varying distribution, then the regularised and re-normalised estimate does not converge to the maximal margin classifier. This shows that classification with fat tails has a qualitatively different behaviour, which should be taken into account when considering real-life data.

Introduction

Margin maximisation, see for instance (Hastie et al., 2009; Vapnik, 1998), is an important concept that defines a property of finite sample optimality linked to separability between points from two different classes. However as the sample size grows, this property is less relevant in the sense that the dataset is unlikely to be separable.

But margin maximisation and separating hyperplanes are still appealing for (at least) two reasons: first, they are an intuitive concept and are a benchmark in any classification task, and, second, using boosting or kernel SVM in a higher dimensional space can enable separability.

(Rosset et al., 2003; 2004), building on previous work by (Bartlett et al., 2006; Freund and Schapire, 1997; Friedman et al., 2000; Schapire et al., 1997) and (Mangasarian, 1999), consider the case of a linear classifier (e.g., logistic regression or support vector machine) and investigate the convergence of a regularised estimator to a margin maximising hyperplane when data is separable. Intriguingly, they established that under an apparently mild criterion (see Eq. (3)) on the loss function, this convergence was guaranteed. This was an important result from a couple standpoints: first, it established a relationship between regularised classifiers and margin maximisation, and, second, it showed that usual loss functions shared that property, leading to the exact choice of a link function being of second order.

The key results of our work are the (partial) answer to the open question and conjecture in (Rosset et al., 2003), on the one hand, and the link between the non convergence to a margin maximising classifier and regular variation (cf. (Bingham et al., 1987)) of the loss function, on the other hand. While margin maximisation is quite specific to the linear setting, deriving analytical properties of loss functions that are also used in other settings, such as deep learning, is particularly interesting to understand choices for loss functions and their implications.

We establish a connection between this problem and heavy tails; (Taleb, 2020) offers a wide-ranging review of heavy tails in multiple applications, (Ibragimov et al., 2015) consider more specifically the role of heavy tails in finance and inference, and applications in supervised learning (mainly regression) are described in (Brownlees et al., 2015; Hsu and Sabato, 2016; Lugosi and Mendelson, 2019). Earlier approaches such as (Chatterjee and Hadi, 1986; Huber and Ronchetti, 2009; Wang et al., 2007) considered heavy tails through the lens of robust estimation. Additional research in classification tasks under a heavy-tail regime is warranted to refine the current state of understanding.

Contributions Our contributions in this paper can be articulated around three questions:

- *Is there a converse statement to (Rosset et al., 2003)'s sufficient condition?* In other words, if a normalised and regularised estimator converges to a margin-maximising hyperplane, must the loss function satisfy the same criterion? We show that, under some additional assumptions, namely the convexity and dif-

[1]Department of Mathematics, University College London, London, United Kingdom. Correspondence to: Francois Buet-Golfouse <ucahfbu@ucl.ac.uk>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

ferentiability of the loss function ℓ, this indeed holds.

- *If the ratio criterion is not verified, what can be said about the loss function?* Interestingly, we establish that such losses can be shown to be in the class of regularly varying functions under mild assumptions (see (Bingham et al., 1987) for an introduction to the theory of regularly varying functions).
- *Is there a probabilistic interpretation of these analytical results?* Using the latent interpretation of binary classification models, we show that the distribution of the latent variable must also be regularly varying, loosely characterised by heavy tails.

While the starting point of this work has to do with linear models and margin-maximising solutions, the characterisation of loss functions (and behaviour thereof) is of broad interest.

Setup and definitions We consider the case of binary classification; we thus suppose that we have n observations of a feature vector $\mathbf{x}_i \in \mathbb{R}^d$ and label $y_i \in \{-1, 1\}$, for $i = 1, ..., n$. The loss function $\ell : \mathbb{R} \to \mathbb{R}^+$ is supposed to depend only on the margin, is monotonic, non-increasing, non-negative and continuous, while the underlying model $g(\mathbf{x}) = \beta^T h(\mathbf{x}_i)$ is taken to be linear. We thus minimise the empirical risk:

$$\min_{\beta \in \mathbb{R}^{|\mathcal{H}|}} \frac{1}{n} \sum_{i=1}^{n} \ell(y_i \beta^T h(\mathbf{x}_i)), \tag{1}$$

where $\mathcal{H} = \{h_1(\mathbf{x}), \cdots\}$ is a finite dictionary of functions. The prediction at point $\mathbf{x}$ is simply sign $\left(\beta^T h(\mathbf{x})\right)$. But, as pointed out by (Rosset et al., 2003), when $|\mathcal{H}|$ is large, it is required to add some regularisation to be able to control the complexity of the classifier:

$$\min_{\beta \in \mathbb{R}^{|\mathcal{H}|}} \frac{1}{n} \sum_{i=1}^{n} \ell(y_i \beta^T h(\mathbf{x}_i)) + \lambda \|\beta\|_p^p, \tag{2}$$

for $p \geq 1$. In the following, we denote by β_λ (possible one of) the solution(s) to Problem (2).

1. The sufficient condition

Let us start by recalling the main result from (Rosset et al., 2003):

Theorem 1. *(Theorem 2.1 in (Rosset et al., 2003)) Assume that the data $\{\mathbf{x}_i, y_i\}_{i=1}^n$ is separable (i.e., there exists $\beta \in \mathbb{R}^{\mathcal{H}}$ such that $\min_i y_i \beta^T h(\mathbf{x}_i) > 0$. Let ℓ be a monotone non-increasing, non-negative loss function depending on the margin only. If $\exists T > 0$ (possible $T = +\infty$) such that*

$$\lim_{t \to T} \frac{\ell(t(1-\varepsilon))}{\ell(t)} = +\infty, \tag{3}$$

for all $\varepsilon \in (0, 1)$, then ℓ is margin maximising loss function in the sense that any convergence point of the normalised solutions $\frac{\beta_\lambda}{\|\beta_\lambda\|_p}$ to the regularised problem (Eq. (2)) as $\lambda \to 0$ is an L^p margin maximising separating hyperplane. Consequently, if the margin maximising hyperplane is unique, then the solutions converge to it

$$\lim_{\lambda \to 0} \frac{\beta_\lambda}{\|\beta_\lambda\|_p} = \arg\max_{\beta, \|\beta_\lambda\|_p = 1} \min_i y_i \beta^T h(\mathbf{x}_i). \tag{4}$$

1.1. Interpretation

The condition $\lim_{t \to T} \frac{\ell(t(1-\varepsilon))}{\ell(t)} = +\infty$ has a very natural explanation to it. For the ratio condition in Eq. (3) to hold, it must be that $\lim_{t \to T} \ell(t) = 0$ (otherwise the ratio would be finite; the case $\lim_{t \to +T} \ell(t) = +\infty$ implies that $\ell(t) = +\infty$ for all t given the non-increasingness of ℓ). Now, if we suppose that ℓ is differentiable and that ℓ' is non-zero in a neigbourhood of T, we obtain by L'Hospital's rule, that

$$\lim_{t \to T} \frac{\ell(t(1-\varepsilon))}{\ell(t)} = (1-\varepsilon) \lim_{t \to T} \frac{\ell'(t(1-\varepsilon))}{\ell'(t)}, \tag{5}$$

so that $\lim_{t \to T} \frac{-\ell'(t)}{-\ell'(t(1-\varepsilon))} = 0$. In other words, the *marginal utility* of having a margin of size t versus a margin of size $t(1-\varepsilon)$ goes to 0. Roughly speaking, this means that datapoints with a smaller margin with contribute a lot more to the empirical loss.

1.2. Usual loss functions

It is straightforward to verify that the usual loss functions verify the criterion Eq. (3), such as the exponential loss function $\ell_{\text{Exponential}} : t \mapsto e^{-t}$ (used implicity in AdaBoost, cf. (Freund and Schapire, 1997; Friedman et al., 2000)), the log-likelihood $\ell_{\text{Logistic}} : t \mapsto \log(1 + e^{-t})$ used in logistic regression, or the hinge loss $\ell_{\text{SVM}} : t \mapsto \max(0, 1-t)$, which is central to support vector machines (see (Hastie et al., 2009; Vapnik, 1998)). The case $T < +\infty$ is only of interest for hinge-type losses where a cut-off is applied.

1.3. The case of Probit regression

We can show that another well-known loss function, namely the one used in Probit regression, not considered in (Rosset et al., 2003), verifies the criterion Eq. (3). In the case of Probit regression, the associated margin loss function is defined as $\ell_{\text{Probit}} : t \mapsto -\log(\Phi(t))$, where Φ is the standard Gaussian cumulative distribution function. Since $\lim_{t \to +\infty} \Phi(t) = 1$ and $\Phi'(t) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2}$, it comes that $\lim_{t \to +\infty} \frac{\ell_{\text{Probit}}(t(1-\varepsilon))}{\ell_{\text{Probit}} t)} = (1-\varepsilon) \lim_{t \to +\infty} e^{\frac{t^2}{2}(1-(1-\varepsilon)^2)} = +\infty$, again by L'Hospital's rule.

1.4. A seemingly universal result

Theorem 1 combined with the fact that the most frequently used loss functions verify the criterion in Eq. (3) means that if the data is separable and the margin maximising hyperplane is unique, then the exact choice of loss function does not matter as all usual loss functions lead to the same end result. Our overall results and considerations in Section 5.3 somewhat qualify that statement.

2. The necessary condition

In this Section, we aim at answering an open question in (Rosset et al., 2003) around the existence of a converse to Theorem 1. In other words, if $\lim_{\lambda\to 0} \frac{\beta_\lambda}{\|\beta_\lambda\|_p} \to \beta_*$, where β_* is a margin maximising hyperplane with unit norm, is it true that $\lim_{t\to T} \frac{\ell(t(1-\varepsilon))}{\ell(t)} = +\infty$ for all $\varepsilon > 0$?

We bring a partial positive answer to the question, and focus here on the case where $T = +\infty$ and $p = 2$, and make the additional assumptions that ℓ is decreasing with $\lim_{t\to+\infty} \ell(t) = 0$, convex and differentiable with continuous derivative ℓ'. We thus consider the loss function to be minimised:

$$L(\beta;\lambda) = \frac{1}{n}\sum_{i=1}^{n} \ell\left(y_i \beta^T h(\mathbf{x}_i)\right) + \lambda \beta^T \beta. \qquad (6)$$

This section goes through a number of steps that were taken to reach the result, and start from the assumption that the normalised "ridged" solution $\beta_\lambda/\|\beta_\lambda\|_2$ converges to a margin maximising hyperplane β_* with unit norm.

First, we notice that we can give an expression for the normalised regularised solution vector as a linear combination of the feature vectors.

Proposition 1. *For a given $\lambda > 0$, the normalised solution vector $\beta_{\lambda,1} = \frac{\beta_\lambda}{\|\beta_\lambda\|_2}$ can be expressed as*

$$\beta_{\lambda,1} = K_\lambda \sum_{i=1}^{n} \alpha_{i,\lambda} y_i h(\mathbf{x}_i),$$

where $\alpha_{i,\lambda} = \frac{\ell'(m_{i,\lambda})}{\sum_{j=1}^n \ell'(m_{j,\lambda})} \geq 0$ for all i, and $K_\lambda > 0$ is a normalising constant such that $\|\beta_{\lambda,1}\|_2 = 1$. In addition, it holds $0 < \frac{1}{\sqrt{n}} \min_{i=1,\cdots,n} \|h(\mathbf{x}_i)\|_2 \leq K_\lambda^{-1} \leq \max_{i=1,\cdots,n} \|h(\mathbf{x}_i)\|_2$, i.e., K_λ is always bounded by upper- and lower-bounds that are independent of λ.

Proof. The first-order condition of the problem reads $\frac{\partial L}{\partial \beta} = \frac{1}{n}\sum_{i=1}^n \ell'(m_{i,\lambda}) y_i h(\mathbf{x}_i) + 2\lambda\beta$, leading to $\beta_\lambda = -\frac{1}{2\lambda n}\sum_{i=1}^n \ell'(m_{i,\lambda}) y_i h(\mathbf{x}_i)$. Since the loss function ℓ is decreasing, $\ell'(t) < 0$ for all $t \in \mathbb{R}$, so that $\alpha_{i,\lambda} = \frac{\ell'(m_{i,\lambda})}{\sum_{j=1}^n \ell'(m_{j,\lambda})}$ is positive for all i. Now, $K_\lambda^2 = \frac{1}{\|\sum_{i=1}^n \alpha_{i,\lambda} h(\mathbf{x}_i)\|_2^2}$; since $\sum_{i=1}^n \alpha_{i,\lambda} = 1$, it is well-known that $1/n \leq \sum_{i=1}^n \alpha_{i,\lambda}^2 \leq 1$. □

From now one, we can thus focus on the behaviour of the weights $\alpha_{i\lambda}$ specifically.

Proposition 2. *Suppose that $h(\mathbf{x}_j)$ is not a support vector of the limiting margin maximising hyperplane β_*, then $\alpha_{j,\lambda} \to 0$ as $\lambda \to 0$. On the other hand, if $h(\mathbf{x}_i)$ is a support vector, then $\alpha_{i,\lambda}$ is bounded by below.*

Proof. Since $\beta_{\lambda,1}$ converges to a margin-maximising hyperplane and by continuity of the minimal margin in β, this entails that there exists $\overline{\lambda} > 0$ such that for any $\lambda < \overline{\lambda}$ and for all $i = 1, \cdots, n$, $m_{i,\lambda} \geq 0$. Similarly, given that β_* corresponds to a margin-maximising hyperplane, it holds that $\beta_* = K_* \sum_{i=1}^n \alpha_{i,*} y_i h(\mathbf{x}_i)$, where $\alpha_{i,*} > 0$ if $h(\mathbf{x}_i)$ is a support vector (in other words, on the boundary of the slab) and $\alpha_{i,*} = 0$ otherwise (this can be obtained via the dual approach to the margin maximisation problem, see (Vapnik, 1998) or Section 4.5.2. in (Hastie et al., 2009)).

By assumption and given the loss minimising property, we have the convergence of the normalised solution vector to the margin-maximising weight vector β_* as $\lambda \to 0$: $\beta_{\lambda,1} \to \beta_*$. But this is equivalent to $K_\lambda \alpha_{i,\lambda} \to K_* \alpha_{i,*}$. In particular, for non support vectors, this means $\alpha_{j,\lambda} \to 0$.

Let us now show that for any support vector $h(\mathbf{x}_i)$, its associated coefficient $\alpha_{i,\lambda}$ is bounded for λ small enough. Indeed, since $K_\lambda \alpha_{i,\lambda} \to K_* \alpha_i > 0$, then, for any $\delta > 0$, there exists λ' such that, for any $\lambda \leq \lambda'$, $\|K_\lambda \alpha_{i,\lambda} - K_* \alpha_{i,*}\|_2^2 \leq \delta$. □

This distinction between support and non-support vectors will now help us characterise the behaviour of the loss function.

Proposition 3. *For any non-support vector $h(\mathbf{x}_j)$ and any support vector $h(\mathbf{x}_i)$, it holds*

$$\frac{\ell'(m_{j,\lambda})}{\ell'(m_{i,\lambda})} \to 0, \qquad (7)$$

as $\lambda \to 0$.

Proof. Let us pick i such that $h(\mathbf{x}_i)$ is a support vector and j such that $h(\mathbf{x}_j)$ is *not* a support vector. Then

$$\begin{aligned} \alpha_{j,\lambda} &= \frac{\ell'(m_{j,\lambda})}{\sum_{k=1}^n \ell'(m_{k,\lambda})} \\ &= \frac{\ell'(m_{j,\lambda})}{\ell'(m_{i,\lambda})} \cdot \frac{\ell'(m_{i,\lambda})}{\sum_{k=1}^n \ell'(m_{k,\lambda})} \\ &= \frac{\ell'(m_{j,\lambda})}{\ell'(m_{i,\lambda})} \alpha_{i,\lambda}. \end{aligned}$$

Since $\alpha_{i,\lambda}$ is bounded by below, $\alpha_{j,\lambda} \to 0$ implies that $\frac{\ell'(m_{j,\lambda})}{\ell'(m_{i,\lambda})} \to 0$. □

We are now in a position to characterise the limiting property and tail behaviour of the ratio of the *derivative* of the loss function.

Proposition 4. *Consider a non-support vector $h(\mathbf{x}_j)$ and a support vector $h(\mathbf{x}_i)$, with respective margins $m_{j,*}, m_{i,*}$ and let $\epsilon \in (0, \frac{m_{j,*}-m_{i,*}}{2})$. Then*

$$\lim_{t\to+\infty} \frac{\ell'(t(1-\mu))}{\ell'(t)} = +\infty, \tag{8}$$

where $\mu = \frac{m_{j,*}-m_{i,*}-2\epsilon}{m_{j,*}-\epsilon} \in (0,1)$.

Proof. Observe that we can rewrite the margin as $m_{k,\lambda} = \|\beta_\lambda\|_2 y_k \beta_{\lambda,1}^T h(\mathbf{x}_k) := \|\beta_\lambda\|_2 \cdot m_{k,\lambda,1}$ for $k = 1, \cdots, n$. In other words, $m_{k,\lambda,1}$ is the "normalised" margin. By convergence of the normalised weight vector and by continuity of the margin, we have that $m_{k,\lambda,1} \to m_{k,*} := y_k \beta_*^T h(\mathbf{x}_k)$. Since i is a support vector but j isn't, it comes $m_{i,*} < m_{j,*}$. Hence, for any $0 < \epsilon < \frac{m_{j,*}-m_{i,*}}{2}$, there exists $\lambda'' > 0$ such that for all $\lambda \leq \lambda''$,

$$\begin{aligned} 0 < m_{i,*} - \epsilon \leq m_{i,\lambda,1} \leq m_{i,*} + \epsilon \\ < m_{j,*} - \epsilon \leq m_{j,\lambda,1} \leq m_{j,*} + \epsilon. \end{aligned}$$

Since ℓ is convex, it comes that ℓ' is non-decreasing, hence the key inequality:

$$0 \leq \frac{\ell'(\|\beta_\lambda\|_2 \cdot (m_{j,*} - \epsilon))}{\ell'(\|\beta_\lambda\|_2 \cdot (m_{i,*} + \epsilon))} \leq \frac{\ell'(m_{j,\lambda})}{\ell'(m_{i,\lambda})}. \tag{9}$$

But, as $\lambda \to 0$, $\|\beta_\lambda\|_2 \to +\infty$ (since for $\lambda < \overline{\lambda}$, all margins are positive but $\ell(t) > 0$, $\|\beta_\lambda\|_2$ must diverge as $\lambda \to 0$) and $\frac{\ell'(m_{j,\lambda})}{\ell'(m_{i,\lambda})} \to 0$ thanks to Proposition 3. This now implies that

$$\lim_{t\to+\infty} \frac{\ell'(t(m_{j,*} - \epsilon))}{\ell'(t(m_{i,*} + \epsilon))} = 0.$$

By continuity of ℓ', this is equivalent to

$$\lim_{t\to+\infty} \frac{\ell'(t(1-\mu))}{\ell'(t)} = +\infty, \tag{10}$$

where $\mu = \frac{m_{j,*}-m_{i,*}-2\epsilon}{m_{j,*}-\epsilon} \in (0,1)$. □

It now remains to derive a similar result for all $\mu \in (0,1)$ and for ℓ rather than ℓ'.

Proposition 5. *Under the assumptions of this section, it holds*

$$\lim_{t\to+\infty} \frac{\ell(t(1-\mu))}{\ell(t)} = +\infty, \tag{11}$$

for any $\mu \in (0,1)$.

Proof. This result of Proposition 4 holds for any positive margins $m_{i,*}, m_{j,*}$ such that $0 < m_{i,*} < m_{j,*}$ and any $0 < \epsilon < \frac{m_{j,*}-m_{i,*}}{2}$, hence, for any $\mu \in (0,1)$, it must hold that $\lim_{t\to+\infty} \frac{\ell'(t(1-\mu))}{\ell'(t)} = +\infty$. This is not quite the desired result, but, since $\lim_{t\to+\infty} \ell(t) = 0$ and ℓ is differentiable (and such that $\ell'(t) \neq 0$ for $t > 0$ given that ℓ is decreasing), we can apply L'Hospital's rule to get $\lim_{t\to+\infty} \frac{\ell(t(1-\mu))}{\ell(t)} = (1-\mu) \lim_{t\to+\infty} \frac{\ell'(t(1-\mu))}{\ell'(t)} = +\infty$, for any $\mu \in (0,1)$. □

Remark 1. Let us point out that the same analysis can be conducted, albeit coordinate by coordinate, for any $p > 1$ (i.e., as long as the p-norm is differentiable).

3. Functional characterisation of the loss ℓ

We have shown that the condition on the convergence to infinity of the ratio $\ell((1-\varepsilon)t)/\ell(t)$ was a necessary (under strict assumptions) and sufficient (under mild assumptions) condition for the convergence of the normalised regularised estimator $\beta_{\lambda,1}$ to a margin-maximising solution.

In this Section, we are however interested in understanding the case where this ratio criterion in Eq. (3) does not hold and the consequences in terms of loss function. From now on, we thus suppose that there exists $\varepsilon \in (0,1)$, such that

$$\lim_{t\to+\infty} \frac{\ell((1-\varepsilon)t)}{\ell(t)} = \gamma \neq +\infty \tag{12}$$

We make the assumption, as in the previous section, that for any $a > 0$, $\lim_{t\to+\infty} \frac{\ell(at)}{\ell(t)} \in \overline{\mathbb{R}}_+$, i.e., the limit exist but can be $+\infty$ or a non-negative real. This assumption is made for the sake of simplicity but can be very modified, see Section 3.3.

For the sake of clarity, let us now denote $\eta := 1 - \varepsilon$.

3.1. The ratio $\ell(at)/\ell(t)$ has a limit for all $a > 0$

Proposition 6. *For any $n \in \mathbb{Z}$, $\lim_{t\to+\infty} \frac{\ell(\eta^n t)}{\ell(t)} = \gamma^n$.*

Proof. Given that $\lim_{t\to+\infty} \frac{\ell(\eta t)}{\ell(t)} = \gamma \geq 1$, we also have $\lim_{t\to+\infty} \frac{\ell(t)}{\ell(\eta t)} = \frac{1}{\gamma}$, and by continuity, $\lim_{t\to+\infty} \frac{\ell(t/\eta)}{\ell(t)} = \frac{1}{\gamma}$. If we consider the case $a = \eta^n$, where $n \in \mathbb{N}^*$, we can write

$$\frac{\ell(at)}{\ell(t)} = \frac{\ell(\eta^n t)}{\ell(t)} = \prod_{i=1}^{n-1} \frac{\ell(\eta^{i+1} t)}{\ell(\eta^i t)}.$$

But we observe that, by continuity, $\lim_{t\to+\infty} \frac{\ell(\eta^{i+1} t)}{\ell(\eta^i t)} = \lim_{t\to+\infty} \frac{\ell(\eta \cdot t)}{\ell(t)} = \gamma$, thus leading to $\lim_{t\to+\infty} \frac{\ell(\eta^n t)}{\ell(t)} = \gamma^n$. Bringing those two facts together, the result holds. □

Proposition 7. *There exists a function $\rho : \mathbb{R}_+^* \to \mathbb{R}_+^*$ such that*

$$\lim_{t\to+\infty} \frac{\ell(at)}{\ell(t)} = \rho(a). \tag{13}$$

In particular, $\rho(a) > 0$ for any positive a.

Proof. For any $0 < a < \eta$, there exists $n_a \in \mathbb{N}^*$ such that $\eta^{n_a+1} \leq a < \eta^{n_a}$, so that

$$\frac{\ell(\eta^{n_a}t)}{\ell(t)} \leq \frac{\ell(at)}{\ell(t)} \leq \frac{\ell(\eta^{n_a+1}t)}{\ell(t)}.$$

Based on our previous results (and the earlier assumption of the limit's existence in $\overline{\mathbb{R}}_+$), this implies that $\gamma^{n_a} \leq \lim_{t\to+\infty} \frac{\ell(at)}{\ell(t)} \leq \gamma^{n_a+1}$. The case $a \geq \eta$ is handled similarly, since there exists $n_a \in \mathbb{N}$ such that $\eta^{-n_a+1} \leq a < \eta^{-n_a}$. □

3.2. ℓ as regularly-varying function

To make use of this result, let us start by briefly recalling some fundamentals of regularly varying function theory (see (Bingham et al., 1987) for all results mentioned here related to regularly-varying functions).

Definition 1. A (measurable) function $L : \mathbb{R}_+^* \to \mathbb{R}_+^*$ is said to be *slowly varying* (at infinity) if, for all $a > 0$,

$$\lim_{t\to+\infty} \frac{L(at)}{L(t)} = 1.$$

Similarly, we can introduce *regularly varying* functions:

Definition 2. A (measurable) function $h : \mathbb{R}_+^* \to \mathbb{R}_+^*$ is said to be *regularly varying* (at infinity) if, for all $a > 0$,

$$\lim_{t\to+\infty} \frac{h(at)}{h(t)} = \rho(a),$$

where $\rho(a)$ is finite but non-zero for every $a > 0$.

This is exactly the setup that we established in the previous subsection, in particular in Proposition 7. One of the cornerstones of the theory of regularly varying functions is Karamata's *characterisation* theorem.

Theorem 2. *Every regularly varying function $h : \mathbb{R}_+^* \to \mathbb{R}_+^*$ is of the form*

$$h(t) = t^\zeta L(t),$$

where $\zeta \in \mathbb{R}$ and L is a slowly varying function.

In particular, it comes directly that $\lim_{t\to+\infty} \frac{h(at)}{h(t)} = a^\zeta$. This implies that the limit function ρ is uniquely defined as $\rho(a) = a^\zeta$ and can only be a power function. Note that a closely related result is Karamata's *representation* theorem, giving a precise representation of slowly varying functions.

In our case, we can thus conclude that if ℓ does not verify the ratio criterion, then ℓ is a regularly varying function, and it is straightforward to infer that

$$\zeta = \frac{\log(\gamma)}{\log(\eta)}. \tag{14}$$

Since we have $\eta \in (0, 1)$ and $\gamma \geq 1$, $\zeta \leq 0$. Note that $\zeta = 0$ if and only if $\gamma = 1$, in which case the loss function ℓ is slowly varying. Since ζ is non-positive, we generally consider $\xi = -\zeta$ rather than ζ directly. While the characterisation and representation of the loss function are interesting results in their own right, it is possible to make them more intuitive by adopting a probabilistic viewpoint.

3.3. Discussion

Before moving forward, let us pause briefly to discuss the assumption made to obtain this Section's results. Our assumption is that, for any $a > 0$, $\lim_{t\to+\infty} \frac{\ell(at)}{\ell(t)} \in \overline{\mathbb{R}}_+$. This is a *global* assumption which, coupled with Eq. 12, implies that the limit must then be finite everywhere. However, this assumption does not require any additional finiteness condition. The *global* aspect of the assumption can be significantly weakened if one posits that there exists at least another point such that the limit exists and is finite. Our result, in this case, is as follows:

Theorem 3. *Let $a_1, a_2 \in \mathbb{R}_+^* - \{1\}$ such that $\frac{\log a_1}{\log a_2} \notin \mathbb{Q}$ and $\lim_{t\to+\infty} \frac{\ell(a_i t)}{\ell(t)} = \rho(a_i) < +\infty$, for $i = 1, 2$, then for any $a > 0$, $\lim_{t\to+\infty} \frac{\ell(at)}{\ell(t)} = \rho(a)$, where, for any $a > 0$, $\rho(a) = a^\zeta$ for some $\zeta \in \mathbb{R}$.*

Proof. Given that ℓ is non-negative and non-increasing, we can apply "Theorem K" in (Seneta, 2002) to the function g defined as $g(u) := \log \ell(e^u)$ for $u \in \mathbb{R}$. □

The condition $\frac{\log a_1}{\log a_2} \notin \mathbb{Q}$ may, however, not be obvious to check. To summarise, the main takeaway is that Eq. 12, on its own, is –in general– not enough to guarantee that ℓ is regularly varying and an additional assumption is required.

4. Probabilistic interpretation of the characterisation result

In this section, we recall the latent interpretation of binary classification and in particular discuss the assumption of symmetry of the latent variable and its inherent limitation. This approach will then be applied to regularly varying losses and distributions in the next section.

4.1. Latent interpretation of classification

It is sometimes useful to posit a threshold model whereby a variable ε_i is unobservable but such that the observed class

label $y_i \in \{-1, +1\}$ is given by

$$\mathbf{1}_{\{y_i=-1\}} = \mathbf{1}_{\{\beta^T h(\mathbf{x}_i)+\varepsilon_i<0\}}. \tag{15}$$

The component $\beta^T h(\mathbf{x}_i)$ is observed but the ε_i's are random perturbations (usually considered to be independent and identically distributed). This leads directly to

$$\begin{aligned}\mathbb{P}(y_i = -1|x_i, \beta) &= F(-\beta^T h(\mathbf{x}_i)) \\ \mathbb{P}(y_i = +1|x_i, \beta) &= 1 - F(-\beta^T h(\mathbf{x}_i)),\end{aligned}$$

with F the cumulative distribution function ("c.d.f.") of ε. In particular, under the assumption that F is symmetric (whereby $1 - F(t) = F(-t)$ for all $t \in \mathbb{R}$), then one can succinctly rewrite the probability of observing class y as

$$\mathbb{P}(y|x_i, \beta) = F(y\beta^T h(\mathbf{x}_i)), \tag{16}$$

for $y \in \{-1, +1\}$ and the likelihood of the sample is then

$$\mathcal{L}(\{\mathbf{x}_i, y_i\}_{i=1}^n; \beta) = \prod_{i=1}^n F(y_i\beta^T h(\mathbf{x}_i)).$$

Hence, maximising the likelihood is equivalent to minimising

$$L(\{\mathbf{x}_i, y_i\}_{i=1}^n; \beta) = \frac{1}{n}\sum_{i=1}^n -\log\left(F(y_i\beta^T h(\mathbf{x}_i))\right).$$

One can thus define in a straightforward way $\ell(t) = -\log(F(t))$. Now, given a loss function ℓ, can we find a corresponding c.d.f. F?

4.2. Characterisation of losses with latent interpretation

For F to be a valid c.d.f., F must be non-negative, right continuous with left limits, non-decreasing and verify $\lim_{t\to-\infty} F(t) = 0$, $\lim_{t\to+\infty} F(t) = 1$. These conditions are guaranteed if ℓ is continuous, non-increasing and has limit $+\infty$ in $-\infty$ and 0 in $+\infty$. However, the key assumption is that of symmetry, which is difficult to obtain.

Proposition 8. *Under the assumptions of non-negativity, non-increasingness and continuity, the loss function ℓ can be expressed as a rescaled cumulative distribution function if and only if it verifies the following functional equation:*

$$2^{-\frac{\ell(t)}{\ell(0)}} + 2^{-\frac{\ell(-t)}{\ell(0)}} = 1, \tag{17}$$

for all $t \in \mathbb{R}$. In this case, $F(t) = e^{-\beta\ell(t)}$ with $\beta = \frac{\log 2}{\ell(0)}$.

The proof is very simple but this result is a negative one in the sense that not all loss functions can be written as $\ell(t) = -\log(F(t))$ for a symmetric F. A counterexample is the exponential loss function $\ell_{\text{Exponential}}$, leading to $F(t) = e^{-e^{-t}}$, which is a valid c.d.f. (namely that of a Gumbel distribution) but is not symmetric.

5. Regularly varying latent distributions

5.1. Brief overview

A concept that is closely related to that of regularly varying *functions* is that of regularly varying *distributions* (see (Cooke et al., 2014) for an introduction to the topic), which its probabilistic equivalent insofar as it characterises the tails of distributions.

Definition 3. A cumulative distribution function F is called regularly varying at infinity with tail index $\xi \in (0, +\infty)$ if

$$\lim_{t\to+\infty} \frac{\overline{F}(at)}{\overline{F}(t)} = a^{-\xi}, \tag{18}$$

for any $a > 0$, where $\overline{F} = 1 - F$ is the survival function.

It is interesting to notice that for $a > 1$, $\lim_{t\to+\infty} \frac{\overline{F}(at)}{\overline{F}(t)} = \mathbb{P}(X > at|X > t)$, where $X \sim F$.

5.2. Loss function and latent distribution tail behaviours

Under the assumption that ℓ is differentiable (or equivalently that F is differentiable, hence admits a probability density function f), we have

$$\begin{aligned}\lim_{t\to+\infty} \frac{\ell(at)}{\ell(t)} &= a \cdot \lim_{t\to+\infty} \frac{\ell'(at)}{\ell'(t)} \\ &= a \cdot \lim_{t\to+\infty} \frac{F(t)}{F(at)} \cdot \frac{f(at)}{f(t)} \\ &= a \cdot \lim_{t\to+\infty} \frac{f(at)}{f(t)},\end{aligned}$$

whence $\lim_{t\to+\infty} \frac{f(at)}{f(t)} = a^{-(\xi+1)}$. Similarly, since $\overline{F}'(t) = -f(t)$, it comes

$$\begin{aligned}\lim_{t\to+\infty} \frac{\overline{F}(at)}{\overline{F}(t)} &= a \cdot \lim_{t\to+\infty} \frac{-f(at)}{-f(t)} \\ &= a^{-\xi},\end{aligned}$$

In other words, we have shown that the loss function ℓ and its associated latent distribution F have the same tail index.

Proposition 9. *If F is regularly varying with tail index ξ and is differentiable (i.e., admits a probability density function f), then f is regularly varying with tail index $\xi + 1$ and the associated loss function $\ell := -\log F$ is regularly varying with tail index ξ.*

This is an important result linking the tail behaviour of the loss function to that of the underlying latent variable. One can understand the convergence (or not) towards a margin maximiser in terms of the distributional properties of unobservable individual noise. We have thus connected the

problem of convergence to a separating margin maximising hyperplane and heavy tails. Given Proposition 8, we can now produce loss functions with different behaviours based on different underlying tail indices.

5.3. Some examples

Let us now provide some concrete examples of latent distributions that are regularly varying. We restrict ourselves to the class of elliptical distributions for the sake of clarity (see (Anderson, 2004) for a textbook treatment), which still covers the majority of known use cases. We illustrate in Figure 1 the evolution of the ratio $\ell(at)/\ell(t)$ for different types of distribution (with different tail behaviours); as per Figure 2, this is connected to tail behaviour of the underlying loss function and distribution. The case of the Gaussian and logistic distributions has already been tackled in Sections 1.2 and 1.3.

5.3.1. Cauchy distribution

The probability density function of a (standard) Cauchy distribution is given by

$$f_{\text{Cauchy}}(t) = \frac{1}{\pi(1+t^2)}.$$

Its c.d.f. is $F_{\text{Cauchy}}(t) = \frac{1}{2} + \frac{1}{\pi}\arctan(t)$, hence

$$\ell_{\text{Cauchy}}(t) = -\log\left(\frac{1}{2} + \frac{1}{\pi}\arctan(t)\right). \qquad (19)$$

From the fact that $\lim_{t\to+\infty} \frac{f_{\text{Cauchy}}(at)}{f_{\text{Cauchy}}(t)} = a^{-1}$, we infer that $\xi_{\text{Cauchy}} = 0$, i.e., the Cauchy distribution is *slowly* varying, and so is ℓ_{Cauchy}.

5.3.2. Student-t distribution

The Cauchy distribution is actually a particular case of a Student-t distribution, whose p.d.f. reads

$$f_{\text{Student}}(t) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}},$$

where $\nu \geq 1$ –the number of degrees of freedom– is a parameter governing the tail behaviour ($\nu = 1$ corresponds to the Cauchy case and $\nu = +\infty$ to the Gaussian one). We infer that the Student-t distribution has tail index $\xi_{\text{Student}} = \frac{\nu-1}{2}$, whence it has *regularly* varying tails for $\nu > 1$. We also deduce that $F_{\text{Student}}(t) = 1 - \frac{1}{2}I_{x(t)}\left(\frac{\nu}{2}, \frac{1}{2}\right)$, and $\ell_{\text{Student}}(t) = -\log\left(1 - \frac{1}{2}I_{x(t)}\left(\frac{\nu}{2}, \frac{1}{2}\right)\right)$, where $x(t) := \frac{\nu}{t^2+\nu}$ and I is the regularised incomplete beta function. Let us point out that the heaviness of the Student-t distribution's tails has been found to be an interesting feature, for example by considered –as in (Shah et al., 2014)– Student-t processes instead of Gaussian ones.

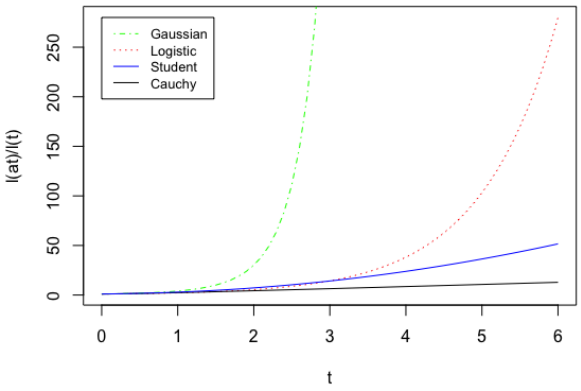

Figure 1. Evolution of the ratio $\ell(at)/\ell(t)$ as a function of t for loss functions associated respectively with the normal, logistic, Student (with $\nu = 2$ degrees of freedom) and Cauchy distributions, and $a = 0.0001$.

We can see (in Figure 1) that the ratio statistic $t \mapsto \ell(at)/\ell(t)$ explodes quickly in the Gaussian case, less quickly in the logistic case and converges in the Student and Cauchy examples. Heavy tails play a crucial role in robustness in statistics and machine learning (cf. (Hsu and Sabato, 2016; Huber and Ronchetti, 2009)) and show that loss functions may reveal different tail behaviours (cf. Figure 2) that can have an impact on an algorithm's performance.

6. Conclusion

The primary focus of the work (Freund and Schapire, 1997; Friedman et al., 2000; Rosset et al., 2003; 2004; Schapire et al., 1997) that spurred the present paper was the relationship between support vector machines and regularisation, and their respective benefits and drawbacks. To some extent, the limiting criterion in Eq. (3) has been shown to be a necessary condition too; but further research is warranted to weaken assumptions.

More importantly, we have considered the margin maximisation property of classifiers (such as support vector machines (Vapnik, 1998)) as a benchmark for classification tasks and endeavoured to determine the properties of loss functions that do not lead to the convergence of the normalised regularised estimator to a margin maximising classifier.

Surprisingly, this is the case (under mild assumptions) if and only if the loss function is regularly varying, which is equivalent to the underlying latent distribution having heavy tails. Our results, while giving a precise quantitative characterisation, are more qualitative in nature, in the sense that they highlight two possible regimes in terms of behaviour of

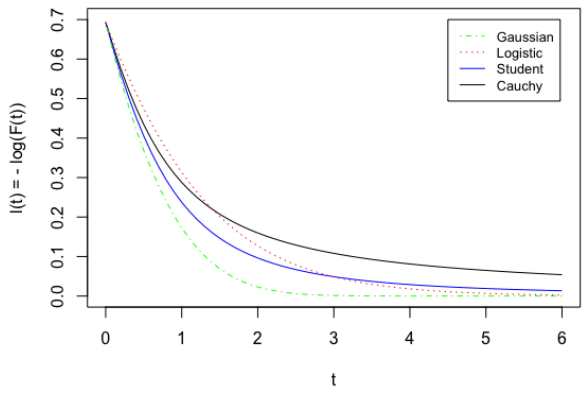

Figure 2. Tail behaviours of the respective loss functions $\ell(t) = -\log F(t)$, in the case of the normal, logistic, Student (with $\nu = 2$ degrees of freedom) and Cauchy distributions.

the normalised and regularised classifier. While usual loss functions that as the exponential, hinge, Probit or logistic loss have a similar behaviour (in terms of convergence in the separable case), heavy tailed loss functions have fundamentally different properties.

From a more practical perspective, it also shows that relying on usual loss functions may be actually less innocuous than anticipated, in the sense that it assumes that there are no heavy tails. This finding is not limited to purely linear or dictionary learning models, but extends to all methods using a margin-dependent loss function (i.e., the dictionary $\mathcal{H}$ need not be fixed). Heavy tails are a growing and exciting part of the recent machine learning literature (Hsu and Sabato, 2016; Lugosi and Mendelson, 2019) and distribution estimation (Ben-Hamou et al., 2017), and open interesting perspectives for real-life data as the presence of heavy tails is well-documented (Taleb, 2020). Some questions remain around the application of these insights to multi-class classification and the impact of regularly varying loss functions in other settings such as deep neural networks or Gaussian Processes for classification.

Acknowledgements

The author wishes to acknowledge discussions with Dr Andrea Macrina and suggestions made by the anonymous referees that have greatly contributed to improving this paper.

References

Theodore W. Anderson. *An Introduction to Multivariate Statistical Analysis*. Wiley, third edition, 2004.

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

Anna Ben-Hamou, Stéphane Boucheron, and Mesrob I. Ohannessian. Concentration inequalities in the infinite urn scheme for occupancy counts and the missing mass, with applications. *Bernoulli*, 23(1):249 – 287, 2017.

N. H. Bingham, C. M. Goldie, and J. L. Teugels. *Regular Variation*. Encyclopedia of Mathematics and its Applications. Cambridge University Press, 1987. doi: 10.1017/CBO9780511721434.

Christian Brownlees, Emilien Joly, and Gábor Lugosi. Empirical risk minimization for heavy-tailed losses. *Ann. Statist.*, 43(6):2507–2536, 12 2015. doi: 10.1214/15-AOS1350. URL https://doi.org/10.1214/15-AOS1350.

Samprit Chatterjee and Ali S. Hadi. Influential observations, high leverage points, and outliers in linear regression. *Statist. Sci.*, 1(3):379–393, 08 1986. doi: 10.1214/ss/1177013622. URL https://doi.org/10.1214/ss/1177013622.

Roger M. Cooke, Daan Nieboer, and Jolanta Misiewicz. *Regularly Varying and Subexponential Distributions*, chapter 4, pages 49–63. John Wiley Sons, Ltd, 2014. ISBN 9781119054207. doi: https://doi.org/10.1002/9781119054207.ch4. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/9781119054207.ch4.

Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.*, 55(1):119–139, August 1997. ISSN 0022-0000. doi: 10.1006/jcss.1997.1504. URL https://doi.org/10.1006/jcss.1997.1504.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Additive Logistic Regression: a Statistical View of Boosting. *The Annals of Statistics*, 38(2), 2000.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA, second edition, 2009.

Daniel Hsu and Sivan Sabato. Loss minimization and parameter estimation with heavy tails. *Journal of Machine Learning Research*, 17(18):1–40, 2016. URL http://jmlr.org/papers/v17/14-273.html.

Peter J. Huber and Elvezio M. Ronchetti. *Robust Statistics*. Wiley, second edition, 2009.

Marat Ibragimov, Rustam Ibragimov, and Johan Walden. *Heavy-tailed distributions and robustness in economics and finance*, volume 214 of *Lecture Notes in Statistics*. Springer, 2015. doi: 10.1007/978-3-319-16877-7.

Gabor Lugosi and Shahar Mendelson. Mean estimation and regression under heavy-tailed distributions: A survey. *Foundations of Computational Mathematics*, 19:1145–1190, 2019.

O. L. Mangasarian. Arbitrary-norm separating plane. *Oper. Res. Lett.*, 24(1–2):15–23, February 1999. ISSN 0167-6377. doi: 10.1016/S0167-6377(98)00049-2. URL https://doi.org/10.1016/S0167-6377(98)00049-2.

Saharon Rosset, Ji Zhu, and Trevor Hastie. Margin maximizing loss functions. In *Proceedings of the 16th International Conference on Neural Information Processing Systems*, NIPS'03, page 1237–1244, Cambridge, MA, USA, 2003. MIT Press.

Saharon Rosset, Ji Zhu, and Trevor Hastie. Boosting as a regularized path to a maximum margin classifier. *J. Mach. Learn. Res.*, 5:941–973, December 2004. ISSN 1532-4435.

Robert E. Schapire, Yoav Freund, Peter Barlett, and Wee Sun Lee. Boosting the margin: A new explanation for the effectiveness of voting methods. In *Proceedings of the Fourteenth International Conference on Machine Learning*, ICML '97, page 322–330, San Francisco, CA, USA, 1997. Morgan Kaufmann Publishers Inc. ISBN 1558604863.

Eugene Seneta. Karamata's characterization theorem, feller, and regular variation in probability theory. *Publications de l'institut mathematique*, 71(85):79–89, 2002.

Amar Shah, Andrew Wilson, and Zoubin Ghahramani. Student-t Processes as Alternatives to Gaussian Processes. In Samuel Kaski and Jukka Corander, editors, *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, volume 33 of *Proceedings of Machine Learning Research*, pages 877–885, Reykjavik, Iceland, 22–25 Apr 2014. PMLR. URL http://proceedings.mlr.press/v33/shah14.html.

Nassim Nicholas Taleb. Statistical consequences of fat tails: Real world preasymptotics, epistemology, and applications, 2020. URL https://arxiv.org/pdf/2001.10488.pdf.

Vladimir N. Vapnik. *Statistical Learning Theory*. Wiley-Interscience, 1998.

Hansheng Wang, Guodong Li, and Guohua Jiang. Robust regression shrinkage and consistent variable selection through the lad-lasso. *Journal of Business & Economic Statistics*, 25(3):347–355, 2007. doi: 10.1198/073500106000000251.