ASYMMETRY AND HEAVY TAILS: BUILT-IN ROBUSTNESS IN CLASSIFICATION

Francois Buet-Golfouse
Department of Mathematics
University College London
London, United Kingdom
ucahfbu@ucl.ac.uk

ABSTRACT

Real-world machine learning applications often require making seemingly innocuous choices, such as picking a classification loss function. Our objective is to show that *robustness* should already be a consideration at that stage. This paper proposes a new analysis of robustness in this set-up via a *probabilistic* analysis of margin maximisation. In particular, asymmetry and tail behaviour are important components that determine the robustness of usual functions such as exponential loss, binomial deviance, or hinge loss. We show that each loss function can be linked to an underlying distribution with different qualitative properties.

1 INTRODUCTION

Robustness in machine learning and statistics is a polysemic concept that applies to the entirety of the ML pipeline and stack. In this paper, we choose to focus on the well-known and well-trodden problem of binary classification and look at it from the perspective of built-in robustness.

By deriving connections between margin loss functions and distributions, we can characterise some of the differences amongst loss functions by probabilistic properties such as symmetry and tail behaviour. These have important connections with robustness in the ML and statistics literature (see Huber & Ronchetti (2009), Ibragimov et al. (2015) and Hsu & Sabato (2016) for some examples).

Contributions First, we establish a 1-to-1 relationship between each loss function and an underlying distribution. We characterise the latent distribution underpinning each of these losses and emphasise the fact that different properties adapting usual loss functions by choosing distributions with different properties may lead to different results.

2 LATENT MODEL OF MARGIN MAXIMISATION

Throughout this paper, we limit ourselves to the case of binary classification. In that framework, we thus have n observations of a feature vector $\mathbf{x}_i \in \mathbb{R}^p$ 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 dictionary of functions. In what follows, we denote the margin by $y_i \beta^T h(\mathbf{x}_i) := m_i$ (the i^{th} margin is thus a function of β). The prediction at point $\mathbf{x}_i$ is then simply sign $\left(\beta^T h(\mathbf{x}_i)\right)$, so that a negative margin indicates a misclassified datapoint.

As is usual (see Hastie et al. (2009); Rosset et al. (2003); Boucheron et al. (2005) for instance), we suppose that ℓ is continuous, non-negative, non-increasing and such that $\lim_{t \to -\infty} = +\infty$ and $\lim_{t \to +\infty} = 0$. In layman's terms, a negative margin will matter more than a positive margin. Such

functions are the exponential loss function $\ell_{\text{Exponential}} : t \mapsto e^{-t}$ (used implicity in AdaBoost, cf. Freund & 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 Vapnik (1998); Hastie et al. (2009)).

2.1 Latent interpretation of margin maximisation

Under the assumptions made on ℓ, it is straightforward to notice that $F(t) := \exp(-\ell(t))$ is a valid cumulative distribution function ("c.d.f."), so that Eq. (1) is equivalent to

$$\max_{\beta \in \mathbb{R}^{|\mathcal{H}|}} \prod_{i=1}^{n} F(m_i). \tag{2}$$

This lends itself to a latent variable explanation whereby there exist n latent random variables ε_i, which are independent and identically follow the distribution induced by F. The thought experiment is as follows: we suppose that for each $i = 1, \cdots, n$, we observe $m_i = y_i \beta^T h(\mathbf{x}_i)$ *and* the success of the random variable $\mathbf{1}_{\{\varepsilon_i \leq m_i\}}$. But, since ε_i is not observable, all we know is that $\mathbf{1}_{\{\varepsilon_i \leq m_i\}}$ is a Bernoulli variable with success probability $F(m_i)$. Maximising the likelihood of these n "successes" then boils down to Eq. (2).

This is an interesting remark as this sheds some (probabilistic) light on the margin maximisation paradigm. This amounts to saying that the latent noise cannot "overtake" the margin, and thus that we do not expect large unobservable moves. Indeed, recall that classifying datapoint i correctly is equivalent to $\mathbf{1}_{\{0 \leq m_i\}} = 1$; margin maximisation is thus simply exchanging the zero bound with a randomised threshold ε_i.

2.2 The particular case of symmetric distributions

This interpretation is even more intuitive in the particular case of symmetric c.d.f's and rejoins the usual latent interpretation of logistic or Probit regressions.

It can be useful to define 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

$$\mathbf{1}_{\{y_i = -1\}} = \mathbf{1}_{\{\beta^T h(\mathbf{x}_i) + \varepsilon_i < 0\}}. \tag{3}$$

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 $\mathbb{P}(y_i = -1 | \mathbf{x}_i, \beta) = F(-\beta^T h(\mathbf{x}_i))$ and $\mathbb{P}(y_i = +1 | \mathbf{x}_i, \beta) = 1 - F(-\beta^T h(\mathbf{x}_i))$.

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

$$\mathbb{P}(y | \mathbf{x}_i, \beta) = F(y \beta^T h(\mathbf{x}_i)), \tag{4}$$

for $y \in \{-1, +1\}$. Maximising the joint likelihood is then nothing but Eq. (2).

In particular, if the noise ε is distributed according to a logistic distribution, then one recovers the binomial deviance; similarly, in the case of a standard Gaussian noise, one has $F(t) = \Phi(t)$ (where Φ is the standard Gaussian c.d.f.), and $\ell_{\text{Probit}}(t) = -\log[\Phi(t)]$.

3 Probabilistic properties of loss functions

We now turn our attention to characterising the usual loss functions and investigating their robustness from a qualitative standpoint, thanks to the properties of the implied distributions $F = e^{-\ell}$. To do so, we focus on symmetry and tail behaviour. Recall that such functions are the exponential loss function $\ell_{\text{Exponential}} : t \mapsto$ (used implicity in AdaBoost, cf. Freund & 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 $\ell_{\text{SVM}} : t \mapsto \max(0, 1 - t)$, used in SVMs (see Vapnik (1998); Hastie et al. (2009)).

In particular, we notice that two of the most popular loss functions exhibit some asymmetry, as the Gumbel distributions exhibits a *right* skew, whereas the displaced exponential distribution has a *left* skew. We return to this concept by considering below *symmetry gaps*.

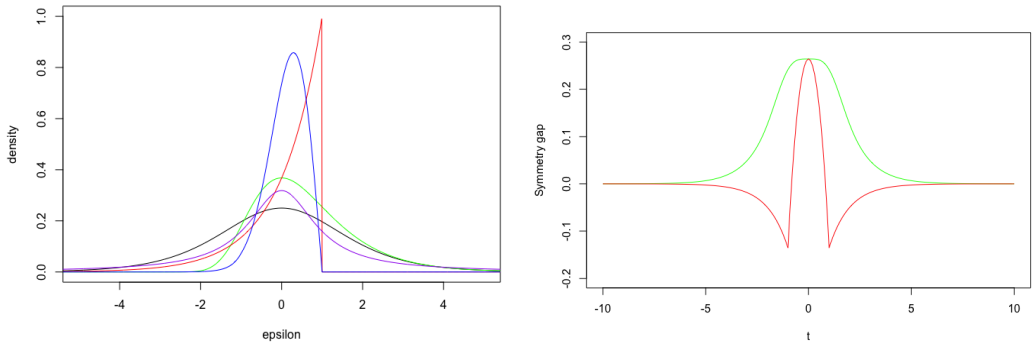

Figure 1: **(a)** Probability density functions of the underlying latent distributions: Gumbel (*green*), logistic (*black*), Gaussian (*red*), exponential-Gaussian mixture (*blue*) and Cauchy (*purple*). **(b)** Symmetry gap associated with the exponential loss/Gumbel distribution (*green*) and the hinge loss/displaced exponential distribution (*red*).

3.1 Symmetry gap and symmetrisation

We have highlighted the role of the underlying c.d.f. and its symmetry. We thus introduce the symmetry gap, $\Delta^{\text{sym}}(t)$, as follows:

$$\Delta^{\text{sym}}(t) = 1 - F(-t) - F(t). \tag{5}$$

Let us suppose that for a given observation, the margin is worth m. Then the statistical *gain* is $F(m)$ (which, as per the maximum likelihood estimation interpretation from the previous section, we wish to maximise). On the other hand, if we flipped the margin and thus had an observation with a margin worth $-m$, then the loss we would incur is $1 - F(-m)$, in short:

$$\text{Symmetry gap} = \text{Loss of flipped margin} - \text{Gain of margin}.$$

This is thus an intrinsic measure of loss aversion for a given underlying distribution. Figure 2 shows the very different symmetry gaps for the Gumbel and displaced exponential distributions. Note in particular the change in regime around -1 and 1 for the latter. This implies different trade-offs for different loss functions.

3.2 Tail behaviour

In addition to symmetry, another important element is the tail behaviour of the latent distribution. This is a key qualitative property as *heavier* tails of a distribution. We consider here a criterion that was introduce in Rosset et al. (2003; 2004); see Sections A and B in the Appendix for an overview.

Let us introduce the survival function $\overline{F}$ defined as $\overline{F} = 1 - F(t)$ for all $t \in \mathbb{R}$. Under the assumption that ℓ is differentiable (or equivalently that F is differentiable, hence admits a probability density function f), we have

$$\lim_{t \to +\infty} \frac{\ell(t(1-\varepsilon))}{\ell(t)} = \lim_{t \to +\infty} \frac{\overline{F}(t(1-\varepsilon))}{\overline{F}(t)}, \tag{6}$$

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 will contribute a lot more to the empirical loss. For all well-known loss functions (binomial deviance, exponential, hinge, huberised hinge and Probit), this limit is worth $+\infty$.

3.3 Cauchy distribution

Here, we thus introduce a distribution that does not verify this condition. 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 the *Negative Log Arctan* loss function $\ell_{\text{Neg Log Arctan}}(t) = -\log\left(\frac{1}{2} + \frac{1}{\pi}\arctan(t)\right)$. From the fact that $\lim_{t\to+\infty} \frac{f_{\text{Cauchy}}(at)}{f_{\text{Cauchy}}(t)} = a^{-1} \neq +\infty$, we infer that the Cauchy distribution is *slowly* varying, and thus that the condition in Eq. (6) is not verified. (See Section B in the Appendix).

3.4 Comparison of loss functions

We can now summarise the probabilistic properties of margin loss functions in the following table.

Loss function	**Definition**	**Latent distribution**	**Symmetry**	**Tail criterion**
Exponential	e^{-t}	Gumbel	No	Yes
Binomial Deviance	$\log(1+e^{-t})$	Logistic	Yes	Yes
Probit	$-\log(\Phi(t))$	Standard Gaussian	Yes	Yes
Hinge	$\max(0, 1-t)$	Displaced Exponential	No	Yes
Huberised hinge	$-4t\mathbf{1}_{t<-1} + \max(0,1-t)^2\mathbf{1}_{t\geq -1}$	Exponential and Gaussian mixture	No	Yes
Negative Log Arctan	$-\log(\frac{1}{2} + \frac{1}{\pi}\arctan(t))$	Cauchy	Yes	No

Table 1: Main margin loss functions' key probabilistic properties

4 Some numerical illustrations

We ran some tests to illustrate in practice our theoretical reasoning on a real dataset and a surrogate one. Details for each dataset can be found in Section C in the appendix. Results are necessarily close since we apply different margin loss functions to the same linear model in each setting, however they do evidence the good performance of the negative log arctan loss function. Our point is not so much about this particular loss function than the fact that choosing more robust loss functions may be helpful.

4.1 South African Heart Disease Dataset

Loss function	**Accuracy**	**False Positive Rate**	**False Negative Rate**
Exponential	73%	47%	16%
Binomial Deviance	73%	48%	15%
Probit	73%	48%	15%
Hinge	73%	47%	16%
Huberised hinge	74%	47%	15%
Negative Log Arctan	74%	48%	15%

4.2 Generated Dataset

Loss function	**Accuracy**	**False Positive Rate**	**False Negative Rate**
Exponential	75.4%	22.4%	27.2%
Binomial Deviance	74.8%	21.3%	29.7%
Probit	75.0%	21.6%	28.9%
Hinge	75.4%	21.6%	28.0%
Huberised hinge	75.0%	21.3%	29.3%
Negative Log Arctan	76.2%	18.3%	30.2%

5 Discussion

We have considered robustness in the narrow setting of binary classification and showed that relying on usual loss functions may actually be less neutral than thought. In particular symmetry and tail behaviour play an important role. These considerations are important because they do not only apply to linear models, but to all methods using a margin-dependent loss function, for instance deep neural networks or kernel methods.

Robustness from a statistical standpoint and heavy tails are a growing part of the recent machine learning literature Hsu & Sabato (2016); Lugosi & Mendelson (2019) and open interesting insights into real-life data where the presence of heavy tails is well-known Taleb (2020).

In summary, while robustness concerns the entire machine learning pipeline, it should be an integral part of algorithm and model design, by ensuring that choices (as basic as a loss function) are not made by default but question their robustness and introduce alternatives if necessary.

REFERENCES

Stéphane Boucheron, Olivier Bousquet, and Gábor Lugosi. Theory of classification: a survey of some recent advances. *ESAIM: PS*, 9:323–375, 2005. doi: 10.1051/ps:2005018. URL https://doi.org/10.1051/ps:2005018.

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.

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, pp. 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.

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.

A MARGIN MAXIMISATION AND TAIL BEHAVIOUR

Let us recall some of the main results in Rosset et al. (2003; 2004). First, they consider a regularised version of the empirical risk minimisation problem and thus solve for

$$\min_{\beta \in \mathbb{R}^{|\mathcal{H}|}} \frac{1}{n} \sum_{i=1}^{n} \ell(y_i \beta^T h(\mathbf{x}_i)) + \lambda \|\beta\|_p^p. \tag{7}$$

They establish a very powerful result making the link between margin maximising loss functions and margin maximising hyperplanes.

In particular, they introduce the tail criterion $\lim_{t \to T} \frac{\ell(t(1-\varepsilon))}{\ell(t)}$ and show that it is sufficient for the convergence of the regularised and normalised weight vector to a margin maximising hyperplane.

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*

$$\lim_{t \to T} \frac{\ell(t(1-\varepsilon))}{\ell(t)} = +\infty, \tag{8}$$

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. (7)) 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

$$\lim_{\lambda \to 0} \frac{\beta_\lambda}{\|\beta_\lambda\|_p} = \underset{\beta, \|\beta_\lambda\|_p = 1}{\arg\max} \min_i y_i \beta^T h(\mathbf{x}_i). \tag{9}$$

B EQUIVALENCE OF BEHAVIOUR BETWEEN LOSS FUNCTION AND SURVIVAL FUNCTION

To prove the equivalence of tail behaviour between loss and survival functions, we make repeated use of the L'Hospital rule. Under the assumption that ℓ is differentiable (so that F is differentiable and admits a probability density function f), it comes

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

In a parallel manner, since $\overline{F}'(t) = -f(t)$, we have

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

In short, we have shown that the loss function ℓ and latent distribution F have the same tail behaviour.

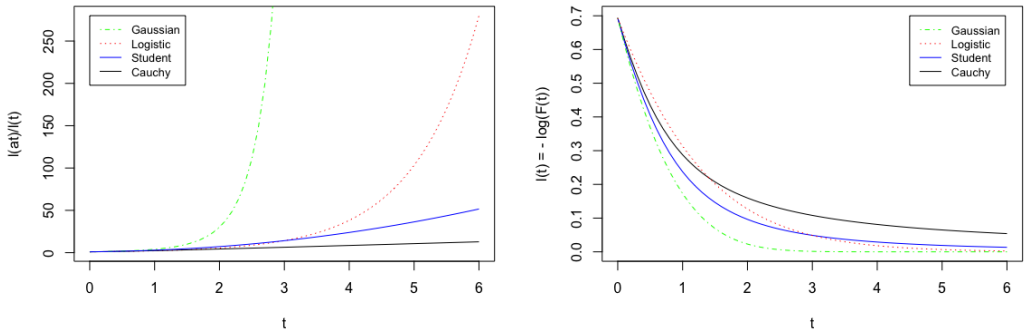

Figure 2: **(a)** Evolution of the ratio $\ell(at)/\ell(t)$ as a function of t for loss functions associated respectively with the normal, logistic, Student (with $\nu = 2$ degrees of freedom) and Cauchy distributions, and $a = 0.0001$. **(b)** Tail behaviours of the respective loss functions $\ell(t) = -\log F(t)$, in the case of the normal, logistic, Student (with $\nu = 2$ degrees of freedom) and Cauchy distributions.

C DATASETS AND IMPLEMENTATION

In this paper, we have used two different datasets with a binary response, which we present below.

C.1 SOUTH AFRICAN HEART DISEASE

The South African Heart Disease dataset [1] is a retrospective sample of males in a heart-disease high-risk region of the Western Cape, South Africa. It contains 463 observations of 8 input features and 1 binary response.

The features are

- systolic blood pressure
- cumulative tobacco (kg)
- low density lipoprotein cholesterol
- adiposity
- family history of heart disease (Present, Absent)
- obesity
- current alcohol consumption
- age at onset
- *response variable*, coronary heart disease (± 1).

C.2 GENERATED DATASET

We have also generated a dataset of 500 observations made up of 2 features and 1 binary response variable. For each observation, the features and response variable are independently generated according to the following procedure:

$$\begin{aligned} x_1 &\sim \text{Bernoulli}(1/2) \\ x_2 &\sim \text{Normal}(0, 1) \\ y &= \text{sign}(\alpha + \beta_1 x_1 + \beta_2 x_2 + \sigma\varepsilon), \end{aligned}$$

where $\varepsilon \sim \text{Student}(3)$, $\alpha = 0.1$. $\beta_1 = 1$, $\beta_2 = -2$ and $\sigma = 2$.

In addition to the classification results present in the main text, we also measure the difference between the real weights specified above and the ones derived from minimising each loss function. We renormalise the weights so that each weight vector has unit norm.

Loss function	Mean Absolute Error (MAE)	Root Mean Squared Error (RMSE)
Exponential	9.45%	7.11%
Binomial Deviance	5.76%	4.29%
Probit	7.72%	5.78%
Hinge	16.36%	12.45%
Huberised hinge	8.31%	6.23%
Negative Log Arctan	2.12%	1.60%

We see that performance is varied across estimators and that the Negative Log Arctan has the best results on this surrogate dataset.

C.3 FITTING ALGORITHMS

In this work, we have only considered underlying linear models and not sought to engineer features, for example via kernel regression, and leave this for future research. The solutions for each loss function were found via a generalised reduced gradient algorithm with multiple starts. While it is possible to solve certain problems with more adequate and bespoke methods (in particular for SVM), we did not use these so as to be consistent across all loss functions.

[1] The South African Heart Disease dataset is available at `http://www-stat.stanford.edu/~tibs/ElemStatLearn/datasets/SAheart.data`