MIXABILITY IS BAYES RISK CURVATURE RELATIVE TO LOG LOSS

Tim van Erven[1]*, Mark D. Reid*[2] *and Robert C. Williamson*[2]

[1]Centrum Wiskunde & Informatica (CWI)
P.O. Box 94079, 1090 GB Amsterdam, The Netherlands, E-mail: tim@timvanerven.nl
[2]ANU and NICTA
Canberra ACT 0200, Australia, E-mail: {Mark.Reid, Bob.Williamson}@anu.edu.au

ABSTRACT

Given K codes, a standard result from source coding tells us how to design a single universal code with codelengths within $\log(K)$ bits of the best code, on any data sequence. Translated to the online learning setting of prediction with expert advice, this result implies that for logarithmic loss one can guarantee constant regret, which does not grow with the number of outcomes that need to be predicted. In this setting, it is known for which other losses the same guarantee can be given: these are the losses that are *mixable*.

We show that among the mixable losses, log loss is special: in fact, one may understand the class of mixable losses as those that behave like log loss in an essential way. More specifically, a loss is mixable if and only if the curvature of its Bayes risk is at least as large as the curvature of the Bayes risk for log loss (for which the Bayes risk equals the entropy).

1. INTRODUCTION

For $n \in \mathbb{N}$, let $\mathcal{Y} = \{1, \ldots, n\}$ be the outcome space. We will consider a prediction game where the loss of the learner making predictions $v_1, v_2, \ldots \in \mathcal{V}$ is measured by a loss function $\ell\colon \mathcal{Y} \times \mathcal{V} \to [0, \infty]$ cumulatively: for $T \in \mathbb{N}$,

$$\mathrm{Loss}(T) := \sum_{t=1}^{T} \ell(y_t, v_t),$$

where $y_1, y_2, \ldots \in \mathcal{Y}$ are outcomes. A loss ℓ is called *η-mixable* if for every distribution P on actions $\mathcal{V}$ there exists a single action v_P such that

$$\ell(v_P, y) \leq \tfrac{-1}{\eta} \log \mathop{\mathbf{E}}_{v \sim P} \left[e^{-\eta \ell(v,y)} \right] \quad \text{for all } y \in \mathcal{Y}.$$

A loss is called *mixable* if there exists any $\eta > 0$ such that it is η-mixable.

The learner has access to predictions v_t^i, $t = 1, 2, \ldots$, $i \in \{1, \ldots, N\}$ generated by N experts $\mathcal{E}_1, \ldots, \mathcal{E}_N$ that attempt to predict the same sequence. The goal of the learner is to predict nearly as well as the best expert. A strategy for the learner, called a *merging strategy*, is a function

$$\mathcal{M}\colon \bigcup_{t=1}^{\infty} (\mathcal{Y}^{t-1} \times (\mathcal{V}^N)^t) \to \mathcal{V},$$

which takes the outcomes $y_1, \ldots, y_{t-1}$ and predictions v_s^i, $i = 1, \ldots, N$ for times $s = 1, \ldots, t$ and outputs an aggregated prediction $v_t^{\mathcal{M}}$, incurring loss $\ell(y_t, v_t^{\mathcal{M}})$ when y_t is revealed. After T rounds, the loss of $\mathcal{M}$ is $\mathrm{Loss}_{\mathcal{M}}(T) = \sum_{t=1}^{T} \ell(y_t, v_t^{\mathcal{M}})$ and the loss of expert $\mathcal{E}_i$ is $\mathrm{Loss}_{\mathcal{E}_i}(T) = \sum_{t=1}^{T} \ell(y_t, v_t^i)$. When $\mathcal{M}$ is the aggregating algorithm (which can be used for all losses considered in this paper) [2], η-mixability implies for all $t \in \mathbb{N}$, all $i \in \{1, \ldots, N\}$,

$$\mathrm{Loss}_{\mathcal{M}}(t) \leq \mathrm{Loss}_{\mathcal{E}_i}(t) + \frac{\ln N}{\eta}. \tag{1}$$

Conversely, if the loss function ℓ is not mixable, then it is not possible to predict as well as the best expert up to an additive constant using any merging strategy.

Thus determining η_ℓ (the largest η such that ℓ is η-mixable) is equivalent to precisely bounding the prediction error of the aggregating algorithm. The mixability of several binary losses and the Brier score in the multiclass case [3] is known. However a general characterisation of η_ℓ in terms of other key properties of the loss has been missing. We show how η_ℓ depends upon the curvature of the conditional Bayes risk for ℓ when ℓ is a *strictly proper* multiclass loss.

2. PROPER MULTICLASS LOSSES

Let $\Delta^n := \{(x_1, \ldots, x_n)' \in \mathbb{R}^n\colon x_i \geq 0, \sum_{i=1}^n x_i = 1\}$ denote the n-simplex, which is the set of all probability vectors on n outcomes. We consider multiclass losses for class probability estimation, where $\mathcal{Y} = \{1, \ldots, n\}$ is the set of possible classes. A *loss function* $\ell : \Delta^n \to [0, \infty]^n$ assigns a loss vector $\ell(q) = (\ell_1(q), \ldots, \ell_n(q))$ to each distribution $q \in \Delta^n$ where $\ell_i(q)$ ($= \ell(i, q)$ traditionally) is the penalty for predicting q when outcome $i \in \mathcal{Y}$ occurs. If the outcomes are distributed with probability $p \in \Delta^n$ then the *risk* for predicting q is just the expected loss

$$L(p, q) := \sum_{i=1}^{n} p_i \ell_i(q).$$

These results have previously appeared in the COLT 2011 proceedings [1]. More details can be found there.

The *Bayes risk* for p is the minimal achievable risk for that outcome distribution,

$$\underline{L}(p) := \inf_{q \in \Delta^n} L(p, q).$$

We say that a loss is *proper* whenever the minimal risk is always achieved by predicting the true outcome distribution, that is, $\underline{L}(p) = L(p, p)$ for all $p \in \Delta^n$. We say a proper loss is *strictly proper* if there exists no $q \neq p$ such that $L(p, q) = \underline{L}(p)$. The log loss $\ell_{\log}(p) := (-\ln(p_1), \ldots, -\ln(p_n))'$ is strictly proper. Its corresponding Bayes risk is $\underline{L}_{\log}(p) = -\sum_{i=1}^{n} p_i \ln(p_i)$, which is the entropy of p.

3. THE BINARY CASE

Consider first the binary case, where $n = 2$. Then for continuous, twice differentiable losses ℓ it is known [4] that

$$\eta_\ell = \min_{p \in [0,1]} \frac{\ell_1'(p)\ell_2''(p) - \ell_1''(p)\ell_2'(p)}{\ell_1'(p)\ell_2'(p)(\ell_2'(p) - \ell_1'(p))}. \tag{2}$$

When a binary loss ℓ is differentiable, properness implies the *stationarity condition* [5]

$$p\ell_1'(p) + (1-p)\ell_2'(p) = 0,$$

from which it follows that

$$\frac{\ell_1'(p)}{p-1} = \frac{\ell_2'(p)}{p} =: w(p) =: w_\ell(p),$$

where w or w_ℓ is called the *weight function* [5]. By differentiating twice, one also finds that $\underline{L}''(p) = -w(p)$. Substituting these expressions into (2) and simplifying, one finds that many factors cancel, leading to

$$\eta_\ell = \min_{p \in (0,1)} \frac{1}{p(1-p)w(p)}.$$

Observing further that $\underline{L}''_{\log}(p) = \frac{-1}{p(1-p)}$ and so $w_{\log}(p) = \frac{1}{p(1-p)}$, we obtain the simple expression

$$\boxed{\eta_\ell = \min_{p \in (0,1)} \frac{w_{\log}(p)}{w_\ell(p)} = \min_{p \in (0,1)} \frac{\underline{L}''_{\log}(p)}{\underline{L}''(p)}.} \tag{3}$$

That is, the mixability constant of binary proper losses is the minimal ratio of the weight functions for log loss and the loss in question. In the next section we will show how (3) generalises to the multiclass case ($n > 2$). That there is a relationship between Bayes risk and mixability was also pointed out (in a less explicit form) by Kalnishkan, Vovk and Vyugin [6].

4. THE MULTICLASS CASE

Because probabilities sum up to one, any $p \in \Delta^n$ is fully determined by its first $n-1$ components $\tilde{p} = (p_1, \ldots, p_{n-1})$. Let $\tilde{\Delta}^n = \{\tilde{p} \colon \tilde{p} \in \Delta^n\}$ be the set of such $(n-1)$-dimensional vectors. We have been implicit about this in the previous section, but for the derivatives of $\underline{L}$ to make sense in the multiclass case, we need to define it as a function of $\tilde{p}$ rather than the full vector p:

$$\underline{L}(\tilde{p}) = \sum_{i=1}^{n-1} p_i \ell_i(\tilde{p}) + \Big(1 - \sum_{i=1}^{n-1} p_i\Big)\ell_n(\tilde{p}).$$

Let $\mathsf{H}\underline{L}(\tilde{p})$ denote the Hessian of $\underline{L}(\tilde{p})$ and for any matrix A let $\lambda_{\max} A$ denotes its maximum eigenvalue. Then in the multiclass case we obtain the following generalisation of (3):

Theorem 1. *Suppose a loss ℓ satisfies Condition 1. Then its mixability constant is*

$$\boxed{\eta_\ell = \inf_{\tilde{p} \in \operatorname{int}(\tilde{\Delta}^n)} \lambda_{\max}\left((\mathsf{H}\underline{L}(\tilde{p}))^{-1} \cdot \mathsf{H}\underline{L}_{\log}(\tilde{p})\right).} \tag{4}$$

The condition we require is as follows:

Condition 1. *The loss ℓ is strictly proper, continuous on Δ^n, and continuously differentiable on the relative interior* $\operatorname{relint}(\Delta^n)$ *of its domain.*

5. CONCLUSION

Under Condition 1, we have shown that mixability of a loss is determined by whether the curvature of its Bayes risk is as least as large as the curvature of the Bayes risk for log loss.

6. REFERENCES

[1] Tim van Erven, Mark D. Reid, and Robert C. Williamson, "Mixability is Bayes risk curvature relative to log loss," in *Proceedings of the 24th Annual Conference on Learning Theory (COLT)*, 2011.

[2] Volodya Vovk, "A game of prediction with expert advice," in *Proceedings of the Eighth Annual Conference on Computational Learning Theory*. ACM, 1995, pp. 51–60.

[3] Volodya Vovk and Fedor Zhdanov, "Prediction with expert advice for the Brier game," *Journal of Machine Learning Research*, vol. 10, pp. 2445–2471, 2009.

[4] David Haussler, Jyrki Kivinen, and Manfred K. Warmuth, "Sequential prediction of individual sequences under general loss functions," *IEEE Transactions on Information Theory*, vol. 44, no. 5, pp. 1906–1925, 1998.

[5] Mark D. Reid and Robert C. Williamson, "Information, divergence and risk for binary experiments," *Journal of Machine Learning Research*, vol. 12, pp. 731–817, March 2011.

[6] Yuri Kalnishkan, Volodya Vovk, and Michael V. Vyugin, "Loss functions, complexities, and the Legendre transformation," *Theoretical Computer Science*, vol. 313, pp. 195–207, 2004.