Mixability is Bayes Risk Curvature Relative to Log Loss

Tim van Erven
CWI Amsterdam
Tim.van.Erven@cwi.nl

Mark D. Reid
ANU and NICTA
Mark.Reid@anu.edu.au

Robert C. Williamson
ANU and NICTA
Bob.Williamson@anu.edu.au

Abstract

Mixability of a loss governs the best possible performance when aggregating expert predictions with respect to that loss. The determination of the mixability constant for binary losses is straightforward but opaque. In the binary case we make this transparent and simpler by characterising mixability in terms of the second derivative of the Bayes risk of proper losses. We then extend this result to multiclass proper losses where there are few existing results. We show that mixability is governed by the Hessian of the Bayes risk, relative to the Hessian of the Bayes risk for log loss. We conclude by comparing our result to other work that bounds prediction performance in terms of the geometry of the Bayes risk. Although all calculations are for proper losses, we also show how to carry the results across to improper losses.

1 Introduction

Mixability is an important property of a loss function that governs the performance of an aggregating forecaster in the prediction with experts setting. The notion is due to Vovk (1990, 1995). Extensions to mixability were presented by Kalnishkan and Vyugin (2002b). The motivation for studying mixability is summarised below (this summary is based on the presentation of Kalnishkan and Vyugin (2008)[1]).

Let $n \in \mathbb{N}$ and $\mathcal{Y} = \{1, \ldots, n\}$ 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 $\ell\colon \mathcal{Y} \times \mathcal{V} \to \mathbb{R}_+$ 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. 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 *merging strategy* $\mathcal{M}\colon \bigcup_{t=1}^{\infty} \left(\mathcal{Y}^{t-1} \times (\mathcal{V}^N)^t\right) \to \mathcal{V}$ 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 $\mathrm{Loss}_{\mathcal{M}}(T) = \sum_{t=1}^{T} \ell(y_t, v_t^{\mathcal{M}})$. 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) (Vovk, 1995), β-mixability (see Section 3 for the definition) implies for all $t \in \mathbb{N}$, all $i \in \{1, \ldots, N\}$,

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

Conversely, if for every $\beta \in \mathbb{R}_+$ the loss function ℓ is not β-mixable, then it is not possible to predict as well as the best expert up to an additive constant using any merging strategy.

[1] Kalnishkan and Vyugin (2008) denote mixability by $\bar{\beta} \in (0, 1)$; we use $\beta = -\ln \bar{\beta} \in (0, \infty)$.

Thus determining β_ℓ (the largest β such that ℓ is β-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 (Vovk and Zhdanov, 2009) is known. However a general characterisation of β_ℓ in terms of other key properties of the loss has been missing. The present paper shows how β_ℓ depends upon the curvature of the conditional Bayes risk for ℓ when ℓ is a strictly proper continuously differentiable multiclass loss (see Theorem 10).

We use the following notation throughout. Let $[n] := \{1, \ldots, n\}$ and denote by $\mathbb{R}_+$ the non-negative reals. The transpose of a vector x is x'. If x is a n-vector, $A = \operatorname{diag}(x)$ is the $n \times n$ matrix with entries $A_{i,i} = x_i$, $i \in [n]$ and $A_{i,j} = 0$ for $i \neq j$. We also write $\operatorname{diag}(x_i)_{i=1}^n := \operatorname{diag}(x_1, \ldots, x_n) := \operatorname{diag}((x_1, \ldots, x_n)')$. The inner product of two n-vectors x and y is denoted by matrix product $x'y$. We sometimes write $A \cdot B$ for the matrix product AB for clarity when required. If $A - B$ is positive definite (resp. semidefinite), then we write $A \succ B$ (resp. $A \succcurlyeq B$). The n-simplex $\Delta^n := \{(x_1, \ldots, x_n)' \in \mathbb{R}^n \colon x_i \geq 0,\ i \in [n],\ \sum_{i=1}^n x_i = 1\}$. Other notation (the Kronecker product $\otimes$, the derivative D, and the Hessian H) are defined in Appendix A which also includes several matrix calculus results we use.

2 Proper Multiclass Losses

We consider multiclass losses for class probability estimation. A *loss function* $\ell : \Delta^n \to \mathbb{R}_+^n$ 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 $i \in [n]$ occurs[2]. If the outcomes are distributed with probability $p \in \Delta^n$ then the *risk* for predicting q is just the expected loss

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

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 $\underline{L}_{\log}(p) = -\sum_{i=1}^n p_i \ln(p_i)$.

Proper losses are defined only on Δ^n which is a $(n-1)$-dimensional submanifold of $\mathbb{R}_+^n$. In order to define the derivatives we will need, it is necessary to project down onto $n-1$ dimensions. Let $\Pi_\Delta : \Delta^n \to \tilde{\Delta}^n$ denote the projection of the n-simplex Δ^n onto its "bottom", denoted $\tilde{\Delta}^n$. That is,

$$\Pi_\Delta(p) := (p_1, \ldots, p_{n-1}) =: \tilde{p} \in \tilde{\Delta}^n$$

is the projection of p onto its first $n-1$ coordinates. Similarly, we will project ℓ's image $\Lambda := \ell(\Delta^n)$ using $\Pi_\Lambda(\lambda) := (\lambda_1, \ldots, \lambda_{n-1})$ for $\lambda \in \Lambda$ with range denoted $\tilde{\Lambda}$. Since $p_n = p_n(\tilde{p}) := 1 - \sum_{i=1}^{n-1} \tilde{p}_i$ we see that Π_Δ is invertible. Specifically, $\Pi_\Delta^{-1}(\tilde{p}) = (\tilde{p}_1, \ldots, \tilde{p}_{n-1}, p_n(\tilde{p}))$. Thus, any function of p can be expressed as a function of $\tilde{p}$. In particular, given a loss $\ell : \Delta^n \to \mathbb{R}^n$ we can write $\ell(\tilde{p}) = \ell(\Pi_\Delta^{-1}(\tilde{p}))$ for $\tilde{p} \in \tilde{\Delta}^n$ and use $\tilde{\ell}(\tilde{p}) := \Pi_\Lambda(\ell(\tilde{p}))$ to denote its projection onto its first $n-1$ coordinates (see Figure 1).

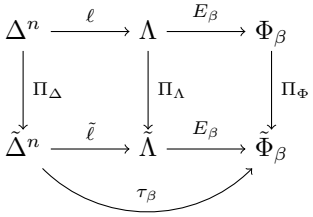

Figure 1: Mappings and spaces.

As it is central to our results, we assume all losses are proper and suitably continuously differentiable for the remainder of the paper. We will additionally assume strict properness whenever we require the Hessian of the Bayes risk to be invertible (see Lemma 5).

[2]Technically, we should allow $\ell(p) = \infty$ to allow for log loss. However, we are only concerned with the behaviour of ℓ in the relative interior of Δ^n and use Δ^n in that sense.

Lemma 1 *A continuously differentiable (strictly) proper loss ℓ has (strictly) concave Bayes risk $\underline{L}$ and a risk L that satisfies the* stationarity condition*: for each p in the relative interior of Δ^n we have*

$$p' \mathsf{D}\ell(\tilde{p}) = 0_{n-1}. \tag{2}$$

Furthermore, ℓ and Π_Λ are invertible and for all $p \in \Delta^n$, the vector p is normal to the surface $\Lambda = \ell(\Delta^n)$ at $\ell(p)$.

Proof: The Bayes risk $\underline{L}(p)$ is the infimum of a set of linear functions $p \mapsto p'\ell(q)$ and thus concave. Each linear function is tangent to $\ell(\Delta^n)$ at a single point when ℓ is strictly proper and so $\underline{L}$ is strictly concave. Properness guarantees that for all $p, q \in \Delta^n$ we have $p'\ell(p) \leq p'\ell(q)$ so the function $L_p : q \mapsto p'\ell(q)$ has a minima at $p = q$. Hence the function $\tilde{L}_p \colon \tilde{q} \mapsto p'\ell(\tilde{q})$ has a minima at $\tilde{q} = \tilde{p}$. Thus $\mathsf{D}\tilde{L}_p(\tilde{q}) = p'\mathsf{D}\ell(\tilde{q}) = 0_{n-1}$ at $\tilde{q} = \tilde{p}$ and so $p'\mathsf{D}\ell(\tilde{p}) = 0_{n-1}$. Since for every $p \in \Delta^n$, $p'\mathsf{D}\ell(\tilde{p}) = 0$ we see p is orthogonal to the tangent space of Λ at $\ell(\tilde{p})$ and thus normal to Λ at $\ell(\tilde{p}) = \ell(p)$. Now suppose there exist $p, q \in \Delta^n$ such that $\ell(p) = \ell(q)$. Since we have just shown that p and q must both be normal to Λ at $\ell(p) = \ell(q)$ and as ℓ is assumed to be continuously differentiable, it must be the case the normals are co-linear, that is, $p = \alpha q$ for some $\alpha \in \mathbb{R}$. However, since $p \in \Delta^n$, $1 = \sum_i p_i = \alpha \sum_i q_i = \alpha$ and thus $p = q$, showing ℓ is invertible.

In order to establish that Π_Λ is invertible we proceed by contradiction and assume ℓ is proper and there exist $p, q \in \Delta^n$ s.t. $\ell_i(p) = \ell_i(q)$ for $i \in [n-1]$ but $\ell_n(p) \neq \ell_n(q)$. Without loss of generality assume $\ell_n(p) < \ell_n(q)$ (otherwise just swap p and q). This means that $q'\ell(p) = \sum_{i=1}^{n} q_i \ell_i(p) = \sum_{i=1}^{n-1} q_i \ell_i(p) + q_n \ell_n(p) < q'\ell(q)$. However, this contradicts properness of ℓ and therefore the assumption that $\ell_n(p) \neq \ell_n(q)$. ■

3 Mixability

We use the following characterisation of mixability (as discussed by Vovk and Zhdanov (2009)) and motivate our main result by looking at the binary case. To define mixability we need the notions of a superprediction set and a parametrised exponential operator. The *superprediction set* S_ℓ for a loss $\ell : \Delta^n \to \mathbb{R}^n$ is the set of points in $\mathbb{R}^n$ that point-wise dominate some point on the loss surface. That is,

$$S_\ell := \{\lambda \in \mathbb{R}^n \colon \exists q \in \Delta^n,\ \forall i \in [n],\ \ell_i(q) \leq \lambda_i\}.$$

The *β-exponential operator* is defined for all $\lambda \in \mathbb{R}^n$ by

$$E_\beta(\lambda) := (e^{-\beta\lambda_1}, \ldots, e^{-\beta\lambda_n}).$$

It is clearly invertible, with inverse $E_\beta^{-1}(\phi) = -\beta^{-1}(\ln \phi_1, \ldots, \ln \phi_n)$. A loss ℓ is *β-mixable* when the set $\Phi_\beta := E_\beta(S_\ell)$ is convex. The *mixability constant* β_ℓ of a loss ℓ is the largest β such that ℓ is β-mixable:

$$\beta_\ell := \sup\{\beta > 0 \colon \ \ell \text{ is } \beta\text{-mixable}\}.$$

Now

$$\begin{aligned} E_\beta(S_\ell) &= \{E_\beta(\lambda) \colon \lambda \in \mathbb{R}^n,\ \exists q \in \Delta^n,\ \forall i \in [n],\ \ell_i(q) \leq \lambda_i\} \\ &= \{z \in \mathbb{R}^n \colon \exists q \in \Delta^n,\ \forall i \in [n],\ e^{-\beta\ell_i(q)} \geq z_i\}, \end{aligned}$$

since $x \mapsto e^{-\beta x}$ is decreasing for $\beta > 0$. Hence in order for Φ_β to be convex the function f such that $\text{graph}(f) = \{(e^{-\beta\ell_1(q)}, \ldots, e^{-\beta\ell_n(q)}) \colon q \in \Delta^n\}$ needs to be *concave*.

3.1 The Binary Case

For twice differentiable binary losses ℓ it is known (Haussler et al., 1998) that

$$\beta_\ell = \min_{\tilde{p} \in [0,1]} \frac{\tilde{\ell}_1'(\tilde{p})\tilde{\ell}_2''(\tilde{p}) - \tilde{\ell}_1''(\tilde{p})\tilde{\ell}_2'(\tilde{p})}{\tilde{\ell}_1'(\tilde{p})\tilde{\ell}_2'(\tilde{p})(\tilde{\ell}_2'(\tilde{p}) - \tilde{\ell}_1'(\tilde{p}))}. \tag{3}$$

When a proper binary loss ℓ is differentiable, the stationarity condition (2) implies

$$\begin{aligned} & \tilde{p}\ell_1'(\tilde{p}) + (1-\tilde{p})\ell_2'(\tilde{p}) = 0 \\ \Rightarrow \quad & \tilde{p}\ell_1'(\tilde{p}) = (\tilde{p}-1)\ell_2'(\tilde{p}) \qquad (4) \\ \Rightarrow \quad & \frac{\ell_1'(\tilde{p})}{\tilde{p}-1} = \frac{\ell_2'(\tilde{p})}{\tilde{p}} =: w(\tilde{p}) =: w_\ell(\tilde{p}) \qquad (5) \end{aligned}$$

We have $\underline{L}(\tilde{p}) = \tilde{p}\ell_1(\tilde{p}) + (1-\tilde{p})\ell_2(\tilde{p})$. Thus by differentiating both sides of (4) and substituting into $\underline{L}''(\tilde{p})$ one obtains $\underline{L}''(\tilde{p}) = \frac{\ell_1'(\tilde{p})}{1-\tilde{p}} = -w(\tilde{p})$. (See Reid and Williamson (2011)). Equation 5 implies $\tilde{\ell}_1'(\tilde{p}) = (\tilde{p}-1)w(\tilde{p})$, $\tilde{\ell}_2'(\tilde{p}) = \tilde{p}w(\tilde{p})$ and hence $\tilde{\ell}_1''(\tilde{p}) = w(\tilde{p}) + (\tilde{p}-1)w'(\tilde{p})$ and $\tilde{\ell}_2''(\tilde{p}) = w(\tilde{p}) + \tilde{p}w'(\tilde{p})$. Substituting these expressions into (3) gives

$$\beta_\ell = \min_{\tilde{p}\in[0,1]} \frac{(\tilde{p}-1)w(\tilde{p})[w(\tilde{p}) + \tilde{p}w'(\tilde{p})] - [w(\tilde{p}) + (\tilde{p}-1)w'(\tilde{p})]\tilde{p}w(\tilde{p})}{(\tilde{p}-1)w(\tilde{p})\tilde{p}w(\tilde{p})[\tilde{p}w(\tilde{p}) - (\tilde{p}-1)w(\tilde{p})]} = \min_{\tilde{p}\in[0,1]} \frac{1}{\tilde{p}(1-\tilde{p})w(\tilde{p})}.$$

Observing that $\underline{L}_{\log}(p) = -p_1 \ln p_1 - p_2 \ln p_2$ we have $\tilde{\underline{L}}_{\log}(\tilde{p}) = -\tilde{p}\ln\tilde{p} - (1-\tilde{p})\ln(1-\tilde{p})$ and thus $\tilde{\underline{L}}''_{\log}(\tilde{p}) = \frac{-1}{\tilde{p}(1-\tilde{p})}$ and so $w_{\log}(\tilde{p}) = \frac{1}{\tilde{p}(1-\tilde{p})}$. Thus

$$\boxed{\beta_\ell = \min_{\tilde{p}\in[0,1]} \frac{w_{\log}(\tilde{p})}{w_\ell(\tilde{p})} = \min_{\tilde{p}\in[0,1]} \frac{\underline{L}''_{\log}(\tilde{p})}{\underline{L}''(\tilde{p})}.} \qquad (6)$$

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. The rest of this paper is devoted to the generalisation of (6) to the multiclass case. That there is a relationship between Bayes risk and mixability was also pointed out (in a less explicit form) by Kalnishkan et al. (2004).

3.2 Mixability and the Concavity of the function f_β

Our aim is to understand mixability in terms of other intrinsic properties of the loss function. In particular, we will relate mixability of a loss to the curvature of its Bayes risk surface. In order to do so, we need to be able to compute the curvature of the β-exponentiated superprediction set to determine when it is convex. This is done by first defining a function $f_\beta : \mathbb{R}^{n-1} \to \mathbb{R}$ with hypograph

$$\operatorname{hyp}(f_\beta) := \{(\tilde{\phi}, y) \subset \mathbb{R}^n \colon y \leq f_\beta(\tilde{\phi})\}$$

equal to $E_\beta(S_\ell)$ and then computing the curvature of f_β. Before we can define f_β we require certain properties of E_β and a mapping $\tau_\beta : \tilde{\Delta}^n \to \mathbb{R}^{n-1}$ defined by

$$\tau_\beta(\tilde{p}) := E_\beta(\tilde{\ell}(\tilde{p})) = \left(e^{-\beta\tilde{\ell}_1(\tilde{p})}, \ldots, e^{-\beta\tilde{\ell}_{n-1}(\tilde{p})}\right).$$

This takes a point $\tilde{p}$ to a point $\tilde{\phi}$ which is the projection of $\phi = E_\beta(\ell(p))$ onto its first $n-1$ coordinates. The range of τ_β is denoted $\tilde{\Phi}_\beta$ (see Figure 1).[3]

Lemma 2 *Let $\lambda \in \Lambda$ and $\phi := E_\beta(\lambda)$. Then $E_\beta^{-1}(\phi) = -\beta^{-1}(\ln\phi_1, \ldots, \ln\phi_n)$ and for all $\alpha \neq 0$, $E_{\alpha\beta}\left(E_\beta^{-1}(\phi)\right) = (\phi_1^\alpha, \ldots, \phi_n^\alpha)$. The derivatives of E_β and its inverse satisfy $\mathsf{D}E_\beta(\lambda) = -\beta \operatorname{diag}(E_\beta(\lambda))$ and $\mathsf{D}E_\beta^{-1}(\phi) = -\beta^{-1}\left[\operatorname{diag}(\phi)\right]^{-1}$. The Hessian of E_β^{-1} is*

$$\mathsf{H}E_\beta^{-1}(\phi) = \frac{1}{\beta}\begin{bmatrix} \operatorname{diag}(\phi_1^{-2}, 0, \ldots, 0) \\ \vdots \\ \operatorname{diag}(0, \ldots, 0, \phi_n^{-2}) \end{bmatrix}. \qquad (7)$$

When $\beta = 1$ and $\ell = \ell_{\log} = p \mapsto -(\ln p_1, \ldots, \ln p_n)$ the map τ_1 is the identity map—that is, $\tilde{\phi} = \tau_1(\tilde{p}) = \tilde{p}$—and $E_1^{-1}(\tilde{p}) = \tilde{\ell}_{\log}(\tilde{p})$ is the (projected) log loss.

[3] We overload E_β^{-1} using it as both a map $\Lambda \to \Phi_\beta$ and from $\tilde{\Lambda} \to \tilde{\Phi}_\beta$. This should not cause confusion because the latter is simply a codimension 1 restriction of the former. Lemma 2 holds for n and $n-1$.

Proof: The results concerning inverses and derivatives follow immediately from the definitions. By (24) the Hessian $\mathsf{H}E_\beta^{-1}(\phi) = \mathsf{D}\left(\mathsf{D}E_\beta^{-1}(\phi)\right)$ and so

$$\mathsf{H}E_\beta^{-1}(\phi) = \mathsf{D}\left(\left(-\frac{1}{\beta}\left[\operatorname{diag}(\phi)\right]^{-1}\right)'\right) = -\frac{1}{\beta}\mathsf{D}\operatorname{diag}(\phi_i^{-1})_{i=1}^n.$$

Let $h(\phi) = \operatorname{diag}(\phi_i^{-1})_{i=1}^n$. We have

$$\mathsf{D}h(\phi) := \mathsf{D}\operatorname{vec} h(\operatorname{vec}\phi) = \mathsf{D}\operatorname{vec} h(\phi) = \begin{bmatrix} \operatorname{diag}(-\phi_1^{-2}, 0, \ldots, 0) \\ \vdots \\ \operatorname{diag}(0, \ldots, 0, -\phi_n^{-2}) \end{bmatrix}.$$

The result for $\beta = 1$ and $\ell_{\log}$ follows from $\tau_1(\tilde{p}) = E_1(\tilde{\ell}(\tilde{p})) = (e^{-1.-\ln\tilde{p}_1}, \ldots, e^{-1.-\ln\tilde{p}_{n-1}})$. ■

Lemma 3 *The map $\tilde{\ell}: \tilde{\Delta}^n \to \tilde{\Lambda}$ is invertible. Also, for all $\beta > 0$, the mapping $\tau_\beta : \tilde{\Delta}^n \to \tilde{\Phi}_\beta$ is invertible with inverse $\tau_\beta^{-1} = \tilde{\ell}^{-1} \circ E_\beta^{-1}$.*

Proof: By "diagram chasing" in Figure 1 we see that $\tilde{\ell}^{-1} = \Pi_\Lambda \circ \ell \circ \Pi_\Delta^{-1}$ and $\tau_\beta^{-1} = \Pi_\Delta \circ \ell^{-1} \circ E_\beta^{-1} \circ \Pi_\Phi^{-1}$ provided all the functions on the right hand sides exist. Π_Δ and ℓ exist by definition, Π_Δ^{-1} exists since $p_i(\tilde{p}) = \tilde{p}_i$ for $i \in [n-1]$ and $p_n(\tilde{p}) = 1 - \sum_i \tilde{p}_i$. The inverse ℓ^{-1} exists by Lemma 1 and E_β^{-1} by Lemma 2. Lastly, Π_Φ is invertible since we see $\Pi_\Phi = E_\beta \circ \Pi_\Lambda^{-1} \circ \tilde{E}_\beta^{-1}$ and $\tilde{E}_\beta^{-1}$ clearly exists due to its form and Π_Λ^{-1} because of Lemma 1. ■

We can now define

$$f_\beta \colon \tilde{\Phi}_\beta \ni \tilde{\phi} \mapsto e^{-\beta\tilde{\ell}_n\left(\tau_\beta^{-1}(\tilde{\phi})\right)} \in [0, \infty). \tag{8}$$

This can be thought of as the inverse of the projection of the β-exponentiated superprediction set Φ_β onto its first $n-1$ coordinates. That is, if $\phi \in \Phi_\beta$ and $\tilde{\phi} = \Pi_{\Phi_\beta}(\phi)$ then $\phi = (\tilde{\phi}_1, \ldots, \tilde{\phi}_{n-1}, f_\beta(\tilde{\phi}))$. This function plays a central role in the remainder of this paper because it coincides with the boundary of the β-exponentiated superprediction set.

Lemma 4 *Let $\beta > 0$ and f_β be defined as in (8). Then* $\operatorname{hyp} f_\beta = \Phi_\beta$.

Proof: We have $\phi = (\phi_1, \ldots, \phi_n)' = (e^{-\beta\ell_1(\tilde{p})}, \ldots, e^{-\beta\ell_n(\tilde{p})})'$. We express ϕ_n as a function of $\tilde{\phi} = (\phi_1, \ldots, \phi_{n-1})' = \tau_\beta(\tilde{p})$ using $\phi_n = e^{-\beta\ell_n(\tilde{p})} = e^{-\beta\ell_n(\tau_\beta^{-1}(\tilde{\phi}))}$. Hence $\operatorname{graph}(f_\beta) = \{(e^{-\beta\ell_1(p)}, \ldots, e^{-\beta\ell_n(p)})' \colon p \in \Delta^n\}$. Since for $\beta > 0$, E_β is monotone decreasing in each argument, $\lambda_i \geq \ell_i(p)$ for all $i \in [n]$ implies $E_\beta(\lambda) \leq E_\beta(\ell(p))$ (coordinatewise). ■

3.3 Relating Concavity of f_β to the Hessian of $\underline{L}$

The aim of this subsection is to express the Hessian of f_β in terms of the Bayes risk of the loss function defining f_β. We first note that a twice differentiable function $f : X \to \mathbb{R}$ defined on $X \subseteq \mathbb{R}^n$ is concave if and only if its Hessian at x, $\mathsf{H}f(x)$, is negative semi-definite for all $x \in X$ (Hiriart-Urruty and Lemaréchal, 1993). The argument that follows consists of repeated applications of the chain and inverse rules for Hessians to compute $\mathsf{H}f_\beta$.

We rely on some consequences of the strict properness of ℓ that allow us to derive simple expressions for the Jacobian and Hessian of the projected Bayes risk $\tilde{\underline{L}} := \underline{L} \circ \Pi_\Delta^{-1} : \tilde{\Delta}^n \to \mathbb{R}_+$.

Lemma 5 *Let $y(\tilde{p}) := -\left[p_n(\tilde{p})\right]^{-1}\tilde{p}$. Then $Y(\tilde{p}) := -p_n(\tilde{p})\mathsf{D}y(\tilde{p}) = \left(I_{n-1} + \frac{1}{p_n}\tilde{p}\mathbb{1}'_{n-1}\right)$ is invertible for all $\tilde{p}$, and*

$$\mathsf{D}\tilde{\ell}_n(\tilde{p}) = y(\tilde{p})' \cdot \mathsf{D}\tilde{\ell}(\tilde{p}). \tag{9}$$

The projected Bayes risk function defined by $\tilde{\underline{L}}(\tilde{p}) := \underline{L}(\Pi_\Delta^{-1}(\tilde{p}))$ *satisfies*

$$\mathsf{D}\tilde{\underline{L}}(\tilde{p}) = \tilde{\ell}(\tilde{p})' - \tilde{\ell}_n(\tilde{p})\mathbb{1}'_{n-1} \tag{10}$$

$$\mathsf{H}\tilde{\underline{L}}(\tilde{p}) = Y(\tilde{p})' \cdot \mathsf{D}\tilde{\ell}(\tilde{p}). \tag{11}$$

Furthermore, for strictly proper ℓ *the matrix* $\mathsf{H}\tilde{\underline{L}}(\tilde{p})$ *is negative definite and invertible for all* $\tilde{p}$ *and when* $\beta = 1$ *and* $\ell = \ell_{\log}$ *is the log loss,*

$$\mathsf{H}\tilde{\underline{L}}_{\log}(\tilde{p}) = -Y(\tilde{p})' \left[\operatorname{diag}(\tilde{p})\right]^{-1}. \tag{12}$$

Proof: The stationarity condition (Lemma 1) guarantees that $p'\mathsf{D}\ell(\tilde{p}) = 0_{n-1}$ for all $p \in \Delta^n$. This is equivalent to $\tilde{p}'\mathsf{D}\tilde{\ell}(\tilde{p}) + p_n(\tilde{p})\mathsf{D}\ell_n(\tilde{p}) = 0_{n-1}$, which can be rearranged to obtain (9).

By the product rule,

$$\begin{aligned}
\mathsf{D}y(\tilde{p}) &= -\tilde{p}\mathsf{D}[p_n(\tilde{p})^{-1}] - [p_n(\tilde{p})^{-1}]\mathsf{D}\tilde{p} \\
&= \tilde{p}[p_n(\tilde{p})^{-2}]\mathsf{D}p_n(\tilde{p}) - [p_n(\tilde{p})^{-1}]I_{n-1} \\
&= -\tilde{p}[p_n(\tilde{p})^{-2}]\mathbb{1}'_{n-1} - [p_n(\tilde{p})^{-1}]I_{n-1} \\
&= -\frac{1}{p_n(\tilde{p})}\left[I_{n-1} + \frac{1}{p_n(\tilde{p})}\tilde{p}\mathbb{1}'_{n-1}\right]
\end{aligned}$$

since $p_n(\tilde{p}) = 1 - \sum_{i\in[n-1]}\tilde{p}_i$ implies $\mathsf{D}p_n(\tilde{p}) = -\mathbb{1}'_{n-1}$. This establishes that $Y(\tilde{p}) = I_{n-1} + \frac{1}{p_n(\tilde{p})}\tilde{p}\mathbb{1}'_{n-1}$. That this matrix is invertible can be easily checked since $(I_{n-1} - \tilde{p}\mathbb{1}'_{n-1})(I_{n-1} + \frac{1}{p_n(\tilde{p})}\tilde{p}\mathbb{1}'_{n-1}) = I_{n-1}$ by expanding and noting $\tilde{p}\mathbb{1}'_{n-1}\tilde{p}\mathbb{1}'_{n-1} = (1-p_n)\tilde{p}\mathbb{1}'_{n-1}$.

The Bayes risk $\tilde{\underline{L}}(\tilde{p}) = \tilde{p}'\tilde{\ell}(\tilde{p}) + p_n(\tilde{p})\tilde{\ell}_n(\tilde{p})$. Taking the derivative and using the product rule ($\mathsf{D}a'b = (\mathsf{D}a')b + a'(\mathsf{D}b)$) gives

$$\begin{aligned}
\mathsf{D}\tilde{\underline{L}}(\tilde{p}) &= \mathsf{D}\left[\tilde{p}'\tilde{\ell}(\tilde{p})\right] + \mathsf{D}\left[p_n(\tilde{p})\tilde{\ell}_n(\tilde{p})\right] \\
&= \tilde{\ell}(\tilde{p}) + \tilde{p}'\mathsf{D}\tilde{\ell}(\tilde{p}) + \left[\mathsf{D}p_n(\tilde{p})\right]\tilde{\ell}_n(\tilde{p}) + p_n(\tilde{p})\mathsf{D}\tilde{\ell}_n(\tilde{p}) \\
&= \tilde{\ell}(\tilde{p}) - p_n(\tilde{p})\mathsf{D}\tilde{\ell}_n(\tilde{p}) - \tilde{\ell}_n(\tilde{p})\mathbb{1}'_{n-1} + p_n(\tilde{p})\mathsf{D}\tilde{\ell}_n(\tilde{p})
\end{aligned}$$

by (9). Thus, $\mathsf{D}\tilde{\underline{L}}(\tilde{p}) = \tilde{\ell}(\tilde{p})' - \tilde{\ell}_n(\tilde{p})\mathbb{1}'_{n-1}$, establishing (10).

Equation 11 is obtained by taking derivatives once more and using (9) again, yielding

$$\mathsf{H}\tilde{\underline{L}}(\tilde{p}) = \mathsf{D}\left(\left(\mathsf{D}\tilde{\underline{L}}(\tilde{p})\right)'\right) = \mathsf{D}\tilde{\ell}(\tilde{p}) - \mathbb{1}_{n-1}\cdot\mathsf{D}\tilde{\ell}_n(\tilde{p}) = \left(I_{n-1} + \frac{1}{p_n}\mathbb{1}_{n-1}\tilde{p}'\right)\mathsf{D}\tilde{\ell}(\tilde{p})$$

as required. Now $\tilde{\underline{L}}(\tilde{p}) = \underline{L}(p_1,\ldots,p_{n-1},p_n(\tilde{p})) = \underline{L}(p_1,\ldots,p_{n-1},1-\sum_{i=1}^{n-1}p_i) = \underline{L}(C(\tilde{p}))$ where C is affine. Since $p \mapsto \underline{L}(p)$ is strictly concave (Lemma 1) it follows (Hiriart-Urruty and Lemaréchal, 1993) that $\tilde{\underline{L}}$ is also strictly concave and thus $\mathsf{H}\tilde{\underline{L}}(\tilde{p})$ is negative definite. It is invertible since we have shown $Y(\tilde{p})$ is invertible and $\mathsf{D}\tilde{\ell}$ is invertible by the inverse function theorem and the invertibility of $\tilde{\ell}$ (Lemma 3).

Finally, equation 12 holds since Lemma 2 gives us $E_1^{-1} = \tilde{\ell}_{\log}$ so (11) specialises to $\mathsf{H}\tilde{\underline{L}}_{\log}(\tilde{p}) = Y(\tilde{p})'\cdot\mathsf{D}\tilde{\ell}_{\log}(\tilde{p}) = Y(\tilde{p})'\cdot\mathsf{D}E_1^{-1}(\tilde{p}) = -Y(\tilde{p})'\cdot\left[\operatorname{diag}(\tilde{p})\right]^{-1}$, also by Lemma 2. ■

3.4 Completion of the Argument

Recall that our aim is to compute the Hessian of the boundary of the β-exponentiated super-prediction set and determine when it is negative semidefinite. The boundary is described by the function f_β which can be written as the composition $h_\beta \circ g_\beta$ where $h_\beta : \mathbb{R} \to [0,\infty)$ and $g_\beta : \tilde{\Phi}_\beta \to \mathbb{R}_+$ are defined by $h_\beta(z) := e^{-\beta z}$ and $g_\beta(\tilde{\phi}) := \tilde{\ell}_n\left(\tau_\beta^{-1}(\tilde{\phi})\right)$. The Hessian of f_β can be expanded in terms of g_β using the chain rule for the Hessian (Theorem 13) as follows.

Lemma 6 *For all $\tilde{\phi} \in \tilde{\Phi}$, the Hessian of f_β at $\tilde{\phi}$ is*

$$\mathsf{H}f_\beta(\tilde{\phi}) = \beta e^{-\beta g_\beta(\tilde{\phi})}\Gamma_\beta(\tilde{\phi}), \tag{13}$$

where $\Gamma_\beta(\tilde{\phi}) := \beta \mathsf{D}g_\beta(\tilde{\phi})' \cdot \mathsf{D}g_\beta(\tilde{\phi}) - \mathsf{H}g_\beta(\tilde{\phi})$. Furthermore, for $\beta > 0$ the negative semi-definiteness of $\mathsf{H}f_\beta(\tilde{\phi})$ (and thus the concavity of f_β) is equivalent to the negative semi-definiteness of $\Gamma_\beta(\tilde{\phi})$.

Proof: Using $f := f_\beta$ and $g := g_\beta$ temporarily and letting $z = g(\tilde{\phi})$, the chain rule for H gives

$$\begin{aligned}
\mathsf{H}f(\tilde{\phi}) &= \left(I_1 \otimes \mathsf{D}g(\tilde{\phi})'\right) \cdot (\mathsf{H}h_\beta(z)) \cdot \mathsf{D}g(\tilde{\phi}) + (\mathsf{D}h_\beta(z) \otimes I_{n-1}) \cdot \mathsf{H}g(\tilde{\phi}) \\
&= \beta^2 e^{-\beta z} \mathsf{D}g(\tilde{\phi})' \cdot \mathsf{D}g(\tilde{\phi}) - \beta e^{-\beta z} \mathsf{H}g(\tilde{\phi}) \\
&= \beta e^{-\beta g(\tilde{\phi})} \left[\beta \mathsf{D}g(\tilde{\phi})' \cdot \mathsf{D}g(\tilde{\phi}) - \mathsf{H}g(\tilde{\phi})\right]
\end{aligned}$$

since $\alpha \otimes A = \alpha A$ for scalar α and matrix A and $\mathsf{D}h_\beta(z) = \mathsf{D}[\exp(-\beta z)] = -\beta e^{-\beta z}$ so $\mathsf{H}h(z) = \beta^2 e^{-\beta z}$. Whether $\mathsf{H}f \preccurlyeq 0$ depends only on Γ_β since $\beta e^{-\beta g(\tilde{\phi})}$ is positive for all $\beta > 0$ and $\tilde{\phi}$. ■

Lemma 7 *For strictly proper ℓ and $\lambda := E_\beta^{-1}(\tilde{\phi})$ and $\tilde{p} := \tilde{\ell}^{-1}(\lambda)$,*

$$\mathsf{D}g_\beta(\tilde{\phi}) = y(\tilde{p})' A_\beta(\tilde{\phi}) \tag{14}$$

$$\mathsf{H}g_\beta(\tilde{\phi}) = -\frac{1}{p_n(\tilde{p})} A_\beta(\tilde{\phi})' \cdot \left[\beta \operatorname{diag}(\tilde{p}) + Y(\tilde{p}) \cdot \left[\mathsf{H}\underline{\tilde{L}}(\tilde{p})\right]^{-1} \cdot Y(\tilde{p})'\right] \cdot A_\beta(\tilde{\phi}), \tag{15}$$

where $A_\beta(\tilde{\phi}) := \mathsf{D}E_\beta^{-1}(\tilde{\phi})$.

Proof: By definition, $g_\beta(\tilde{\phi}) := \tilde{\ell}_n(\tau_\beta^{-1}(\tilde{\phi}))$. Since $\tau_\beta^{-1} = \tilde{\ell}^{-1} \circ E_\beta^{-1}$ we have $g_\beta = \tilde{\ell}_n \circ \tilde{\ell}^{-1} \circ E_\beta^{-1}$. Thus, by Lemma 5 equation (9), the inverse function theorem, and chain rule we have

$$\mathsf{D}g_\beta(\tilde{\phi}) = \mathsf{D}\tilde{\ell}_n(\tilde{p}) \cdot \mathsf{D}\tilde{\ell}^{-1}(\lambda) \cdot \mathsf{D}E_\beta^{-1}(\tilde{\phi}) = y(\tilde{p})' \mathsf{D}\tilde{\ell}(\tilde{p}) \cdot \left[\mathsf{D}\tilde{\ell}(\tilde{p})\right]^{-1} \cdot \left[\mathsf{D}E_\beta^{-1}(\tilde{\phi})\right] = y(\tilde{p})' A_\beta(\tilde{\phi})$$

yielding (14). Since $\tilde{p} = \tau_\beta^{-1}(\tilde{\phi})$ and $\mathsf{H}g_\beta = \mathsf{D}((\mathsf{D}g_\beta)')$ (see (24)), the chain and product rules give

$$\begin{aligned}
\mathsf{H}g_\beta(\tilde{\phi}) &= \mathsf{D}\left[\left(\mathsf{D}E_\beta^{-1}(\tilde{\phi})\right)' \cdot y\left(\tau_\beta^{-1}(\tilde{\phi})\right)\right] \\
&= \left(y(\tau_\beta^{-1}(\tilde{\phi}))' \otimes I_{n-1}\right) \cdot \mathsf{D}\left(\mathsf{D}E_\beta^{-1}(\tilde{\phi})'\right) + \left(I_1 \otimes (\mathsf{D}E_\beta^{-1}(\tilde{\phi}))'\right) \cdot \mathsf{D}\left(y\left(\tau_\beta^{-1}(\tilde{\phi})\right)\right) \\
&= (y(\tilde{p})' \otimes I_{n-1}) \cdot \mathsf{H}E_\beta^{-1}(\tilde{\phi}) + \left(\mathsf{D}E_\beta^{-1}(\tilde{\phi})\right)' \cdot \mathsf{D}y(\tilde{p}) \cdot \mathsf{D}\tau_\beta^{-1}(\tilde{\phi}) \\
&= -\frac{\beta}{p_n(\tilde{p})} A_\beta(\tilde{\phi}) \cdot \operatorname{diag}(\tilde{p}) \cdot A_\beta(\tilde{\phi}) + A_\beta(\tilde{\phi})' \cdot \mathsf{D}y(\tilde{p}) \cdot \mathsf{D}\tau_\beta^{-1}(\tilde{\phi}).
\end{aligned} \tag{16}$$

The first summand above is due to (7) and the fact that

$$\begin{aligned}
(y \otimes I_{n-1}) \cdot \mathsf{H}E_\beta^{-1}(\tilde{\phi}) &= \frac{1}{\beta}[y_1 I_{n-1}, \ldots, y_{n-1} I_{n-1}] \cdot \begin{bmatrix} \operatorname{diag}(\phi_1^{-2}, 0, \ldots, 0) \\ \vdots \\ \operatorname{diag}(0, \ldots, 0, \phi_{n-1}^{-2}) \end{bmatrix} \\
&= \frac{1}{\beta} \sum_{i=1}^{n-1} y_i \cdot I_{n-1} \cdot \operatorname{diag}(0, \ldots, 0, \phi_i^{-2}, 0, \ldots, 0) \\
&= \frac{1}{\beta} \operatorname{diag}(y_i \phi_i^{-2})_{i=1}^{n-1} \\
&= \frac{-\beta}{p_n(\tilde{p})} A_\beta(\tilde{\phi})' \cdot \operatorname{diag}(\tilde{p}) \cdot A_\beta(\tilde{\phi}).
\end{aligned}$$

The last equality holds because $A_\beta(\tilde{\phi})' \cdot A_\beta(\tilde{\phi}) = \beta^{-2}\,\mathrm{diag}(\tilde{\phi}_i^{-2})_{i=1}^{n-1}$ by Lemma 2, the definition of $y(\tilde{p}) = -[p_n(\tilde{p})]^{-1}\tilde{p}$, and because all the matrices are diagonal and thus commute.

The second summand in (16) reduces by $\mathsf{D}y(\tilde{p}) = -\frac{1}{p_n(\tilde{p})}Y(\tilde{p})$ from Lemma 5 and $\tau_\beta = E_\beta \circ \tilde{\ell}$:

$$\mathsf{D}\tau_\beta^{-1}(\tilde{\phi}) = \left[\mathsf{D}E_\beta(\lambda)\cdot \mathsf{D}\tilde{\ell}(\tilde{p})\right]^{-1} = \left[\mathsf{D}E_\beta(\lambda)\cdot (Y(\tilde{p})')^{-1}\cdot \mathsf{H}\underline{\tilde{L}}(\tilde{p})\right]^{-1} = \left[\mathsf{H}\underline{\tilde{L}}(\tilde{p})\right]^{-1}\cdot Y(\tilde{p})'\cdot \mathsf{D}E_\beta^{-1}(\lambda).$$

Substituting these into (16) gives

$$\mathsf{H}g_\beta(\tilde{\phi}) = -\frac{\beta}{p_n(\tilde{p})}A_\beta(\tilde{\phi})\cdot \mathrm{diag}(\tilde{p})\cdot A_\beta(\tilde{\phi}) - \frac{1}{p_n(\tilde{p})}A_\beta(\tilde{\phi})'\cdot Y(\tilde{p})\cdot \left[\mathsf{H}\underline{\tilde{L}}(\tilde{p})\right]^{-1}\cdot Y(\tilde{p})'\cdot A_\beta(\tilde{\phi}),$$

which can be factored into the required result. ■

We can now use the last two lemmata to express the function Γ_β in terms of the Hessian of the Bayes risk functions for the specified loss ℓ and the log loss.

Lemma 8 *The matrix-valued function Γ_β satisfies, for all $\tilde{\phi} \in \tilde{\Phi}$ and $\tilde{p} = \tau_\beta^{-1}(\tilde{\phi})$,*

$$\Gamma_\beta(\tilde{\phi}) = \frac{1}{p_n}A_\beta(\tilde{\phi})'\cdot Y(\tilde{p})\ \left[\left[\mathsf{H}\underline{\tilde{L}}(\tilde{p})\right]^{-1} - \beta\left[\mathsf{H}\underline{\tilde{L}}_{\log}(\tilde{p})\right]^{-1}\right]\cdot Y(\tilde{p})'\cdot A_\beta(\tilde{\phi}), \tag{17}$$

and, for each $\tilde{\phi}$, is negative semi-definite if and only if $R(\beta,\ell,\tilde{p}) := \left[\mathsf{H}\underline{\tilde{L}}(\tilde{p})\right]^{-1} - \beta\left[\mathsf{H}\underline{\tilde{L}}_{\log}(\tilde{p})\right]^{-1}$ is negative semi-definite.

Proof: Substituting the values of $\mathsf{D}g_\beta$ and $\mathsf{H}g_\beta$ from Lemma 7 into the definition of Γ_β from Lemma 6 and then using Lemma 2 and the definition of $y(\tilde{p})$, we obtain

$$\begin{aligned}\Gamma_\beta(\tilde{\phi}) &= \beta A_\beta(\tilde{\phi})'\cdot y(\tilde{p})\cdot y(\tilde{p})'\cdot A_\beta(\tilde{\phi}) + \frac{1}{p_n(\tilde{p})}A_\beta(\tilde{\phi})'\cdot\left[\beta\,\mathrm{diag}(\tilde{p}) + Y(\tilde{p})\cdot\left[\mathsf{H}\underline{\tilde{L}}(\tilde{p})\right]^{-1}\cdot Y(\tilde{p})'\right]\cdot A_\beta(\tilde{\phi})\\ &= \frac{1}{p_n}A_\beta(\tilde{\phi})'\cdot\left[\beta\frac{1}{p_n}\tilde{p}\cdot\tilde{p}' + \beta\,\mathrm{diag}(\tilde{p}) + Y(\tilde{p})\cdot\left[\mathsf{H}\underline{\tilde{L}}(\tilde{p})\right]^{-1}\cdot Y(\tilde{p})'\right]\cdot A_\beta(\tilde{\phi}).\end{aligned} \tag{18}$$

Using Lemma 5 we then see that

$$\begin{aligned}-Y(\tilde{p})\cdot\left[\mathsf{H}\underline{\tilde{L}}_{\log}(\tilde{p})\right]^{-1}\cdot Y(\tilde{p})' &= -Y(\tilde{p})\cdot\left[-Y(\tilde{p})'\,\mathrm{diag}(\tilde{p})^{-1}\right]^{-1}\cdot Y(\tilde{p})'\\ &= Y(\tilde{p})\cdot\mathrm{diag}(\tilde{p})\cdot(Y(\tilde{p})')^{-1}\cdot Y(\tilde{p})'\\ &= (I_{n-1} + \frac{1}{p_n}\mathbb{1}_{n-1}\tilde{p}')\cdot\mathrm{diag}(\tilde{p})\\ &= \mathrm{diag}(\tilde{p}) + \frac{1}{p_n}\tilde{p}\cdot\tilde{p}'.\end{aligned}$$

Substituting this for the appropriate terms in (18) gives

$$\Gamma_\beta(\tilde{\phi}) = \frac{1}{p_n}A_\beta(\tilde{\phi})'\cdot\left[Y(\tilde{p})\cdot\left[\mathsf{H}\underline{\tilde{L}}(\tilde{p})\right]^{-1}\cdot Y(\tilde{p})' - \beta Y(\tilde{p})\cdot\left[\mathsf{H}\underline{\tilde{L}}_{\log}(\tilde{p})\right]^{-1}\cdot Y(\tilde{p})'\right]\cdot A_\beta(\tilde{\phi})$$

which equals (17).

Since $\Gamma_\beta = [p_n]^{-1}BRB'$ where $B = A_\beta(\tilde{\phi})'Y(\tilde{p})$ and $R = R(\beta,\ell,\tilde{p})$ the definition of negative semi-definiteness and the positivity of p_n means we need to show that $\forall x : x'\Gamma_\beta x \le 0 \iff \forall y : y'Ry \le 0$. It suffices to show that B is invertible, since we can let $y = Bx$ to establish the equivalence. The matrix $A_\beta(\tilde{\phi})$ is invertible since, by definition, $A_\beta(\tilde{\phi}) = \mathsf{D}E_\beta^{-1}(\tilde{\phi}) = -\beta^{-1}[\mathrm{diag}(\tilde{\phi})]^{-1}$ by Lemma 2 and so has matrix inverse $-\beta\,\mathrm{diag}(\tilde{\phi})$. The matrix $Y(\tilde{p})$ is invertible by Lemma 7. Thus, B is invertible because it is the product of two invertible matrices. ■

The above arguments result in a characterisation of the concavity of the function f_β (via its Hessian)—and hence the convexity of the β-exponentiated superprediction set—in terms of the Hessian of the Bayes risk function of the loss ℓ and the log loss $\ell_{\log}$. As in the binary case (cf. (6)), this means we are now able to specify the mixability constant β_ℓ in terms of the curvature $\mathsf{H}\underline{\tilde{L}}$ of the Bayes risk for ℓ relative to the curvature $\mathsf{H}\underline{\tilde{L}}_{\log}$ of the Bayes risk for log loss.

Lemma 9 *The mixability constant β_ℓ of a twice differentiable strictly proper loss ℓ is*

$$\beta_\ell = \sup\left\{\beta > 0 : \forall \tilde{p} \in \tilde{\Delta}^n \ , \ \beta \mathsf{H}\underline{\tilde{L}}(\tilde{p}) \succcurlyeq \mathsf{H}\underline{\tilde{L}}_{\log}(\tilde{p})\right\}, \tag{19}$$

where $\underline{\tilde{L}}(\tilde{p}) := \underline{L}(p)$ is the Bayes risk of ℓ and $\underline{\tilde{L}}_{\log}$ is the Bayes risk for the log loss.

Proof: By Lemma 6 and Lemma 8 we know $\mathsf{H}f_\beta(\tilde{p}) \preccurlyeq 0 \iff R(\beta, \ell, \tilde{p}) \preccurlyeq 0$. By Lemma 5, $\mathsf{H}\underline{\tilde{L}}(\tilde{p}) \prec 0$ and $\mathsf{H}\underline{\tilde{L}}_{\log}(\tilde{p}) \prec 0$ for all $\tilde{p}$ and so we can use the fact that for positive definite matrices A and B we have $A \succcurlyeq B \iff B^{-1} \succcurlyeq A^{-1}$ (Horn and Johnson, 1985, Corollary 7.7.4). This means $R(\beta, \ell, \tilde{p}) \preccurlyeq 0 \iff \mathsf{H}\underline{\tilde{L}}(\tilde{p})^{-1} \preccurlyeq \beta \mathsf{H}\underline{\tilde{L}}_{\log}(\tilde{p})^{-1} \iff \beta^{-1}\mathsf{H}\underline{\tilde{L}}_{\log}(\tilde{p}) \preccurlyeq \mathsf{H}\underline{\tilde{L}}(\tilde{p}) \iff \beta \mathsf{H}\underline{\tilde{L}}(\tilde{p}) \succcurlyeq \mathsf{H}\underline{\tilde{L}}_{\log}(\tilde{p})$. Therefore f_β is concave at $\tilde{p}$ if and only if $\beta \mathsf{H}\underline{\tilde{L}}(\tilde{p}) \succcurlyeq \mathsf{H}\underline{\tilde{L}}_{\log}(\tilde{p})$. The mixability constant β_ℓ is defined in Section 3 to be the largest $\beta > 0$ such that the β-exponentiated superprediction set $E_\beta(S_\ell)$ is convex. This is equivalent to the function f_β being concave at all $\tilde{p}$. Thus, we have shown $\beta_\ell = \sup\{\beta > 0 \colon \forall \tilde{p} \in \tilde{\Delta}^n \ , \ \beta \underline{\tilde{L}}(\tilde{p}) \succcurlyeq \mathsf{H}\underline{\tilde{L}}_{\log}(\tilde{p})\}$ as required. ■

The mixability constant can also be expressed in terms of the maximal eigenvalue of the "ratio" of the Hessian matrices for the Bayes risk for log loss and the loss in question. In the following, $\lambda_i(A)$ will denote the ith largest (possibly repeated) eigenvalue of the $n \times n$ symmetric matrix A. That is, $\lambda_{\min}(A) := \lambda_1(A) \leq \lambda_2(A) \leq \cdots \leq \lambda_n =: \lambda_{\max}(A)$ where each $\lambda_i(A)$ satisfies $|A - \lambda_i(A)I| = 0$.

Theorem 10 *For any twice differentiable strictly proper loss ℓ, the mixability constant is*

$$\boxed{\beta_\ell = \min_{\tilde{p} \in \tilde{\Delta}^n} \lambda_{\max}\left((\mathsf{H}\underline{\tilde{L}}(\tilde{p}))^{-1} \cdot \mathsf{H}\underline{\tilde{L}}_{\log}(\tilde{p})\right).} \tag{20}$$

Equation 20 reduces to (6) when $n = 2$ since the maximum eigenvalue of a 1×1 matrix is simply its single entry.

Proof: We define $C_\beta(\tilde{p}) := \beta \mathsf{H}\underline{\tilde{L}}(\tilde{p}) - \mathsf{H}\underline{\tilde{L}}_{\log}(\tilde{p})$ and $\rho(\tilde{p}) := \mathsf{H}\underline{\tilde{L}}(\tilde{p})^{-1} \cdot \mathsf{H}\underline{\tilde{L}}_{\log}(\tilde{p})$ and for any fixed $\tilde{p}$, we first show that zero is an eigenvalue of $C_\beta(\tilde{p})$ if and only if β is an eigenvalue of $\rho(\tilde{p})$. This can be seen since $\mathsf{H}\underline{\tilde{L}}(\tilde{p})$ is invertible (Lemma 5) so

$$\begin{aligned} &|C_\beta(\tilde{p}) - 0I| = 0 \iff |\beta \mathsf{H}\underline{\tilde{L}}(\tilde{p}) - \mathsf{H}\underline{\tilde{L}}_{\log}(\tilde{p})| = 0 \iff |\mathsf{H}\underline{\tilde{L}}(\tilde{p})^{-1}||\beta \mathsf{H}\underline{\tilde{L}}(\tilde{p}) - \mathsf{H}\underline{\tilde{L}}_{\log}(\tilde{p})| = 0 \\ &\iff \left|\mathsf{H}\underline{\tilde{L}}(\tilde{p})^{-1} \cdot \left[\beta \mathsf{H}\underline{\tilde{L}}(\tilde{p}) - \mathsf{H}\underline{\tilde{L}}_{\log}(\tilde{p})\right]\right| = 0 \iff |\beta I - \mathsf{H}\underline{\tilde{L}}(\tilde{p})^{-1} \cdot \mathsf{H}\underline{\tilde{L}}_{\log}(\tilde{p})| = 0. \end{aligned}$$

Since a symmetric matrix is p.s.d. if and only if all its eigenvalues are non-negative it must be the case that if $\lambda_{\min}(C_\beta(\tilde{p})) \geq 0$ then $C_\beta(\tilde{p}) \succcurlyeq 0$ since every other eigenvalue is bigger than the minimum one. Conversely, if $C_\beta(\tilde{p}) \not\succcurlyeq 0$ then at least one eigenvalue must be negative, thus the smallest eigenvalue must be negative. Thus, $\lambda_{\min}(C_\beta(\tilde{p})) \geq 0 \iff C_\beta(\tilde{p}) \succcurlyeq 0$. Now define $\beta(\tilde{p}) := \sup\{\beta > 0 : C_\beta(\tilde{p}) \succcurlyeq 0\} = \sup\{\beta > 0 : \lambda_{\min}(C_\beta(\tilde{p})) \geq 0\}$. We show that for each $\tilde{p}$ the function $\beta \mapsto \lambda_{\min}(C_\beta(\tilde{p}))$ is continuous and only has a single root. First, continuity is because the entries of $C_\beta(\tilde{p})$ are continuous in β for each $\tilde{p}$ and eigenvalues are continuous functions of their matrix's entries (Horn and Johnson, 1985, Appendix D). Second, as a function of its matrix arguments, the minimum eigenvalue $\lambda_{\min}$ is known to be concave (Magnus and Neudecker, 1999, §11.6). Thus, for any fixed $\tilde{p}$, its restriction to the convex set of matrices $\{C_\beta(\tilde{p}) : \beta > 0\}$ is

also concave in its entries and so in β. Since $C_0(\tilde{p}) = -\mathsf{H}\tilde{\underline{L}}_{\log}(\tilde{p})$ is positive definite for every $\tilde{p}$ (Lemma 5) we have $\lambda_{\min}(C_0(\tilde{p})) > 0$ and so, by the concavity of the map $\beta \mapsto \lambda_{\min}(C_\beta(\tilde{p}))$, there can be only one $\beta > 0$ for which $\lambda_{\min}(C_\beta(\tilde{p})) = 0$ and by continuity it must be largest non-negative one, that is, $\beta(\tilde{p})$.

Thus $\beta(\tilde{p}) = \sup\{\beta > 0 : \lambda_{\min}(C_\beta(\tilde{p})) = 0\} = \sup\{\beta > 0 : \beta \text{ is an eigenvalue of } \rho(\tilde{p})\} = \lambda_{\max}(\rho(\tilde{p}))$. Now let $\beta^* := \min_{\tilde{p}} \beta(\tilde{p}) = \min_{\tilde{p}} \lambda_{\max}(\rho(\tilde{p}))$ and let $\tilde{p}^*$ be a minimiser so that $\beta^* = \beta(\tilde{p}^*)$. We now claim that $C_{\beta^*}(\tilde{p}) \succcurlyeq 0$ for all $\tilde{p}$ since if there was some $\tilde{q} \in \tilde{\Delta}^n$ such that $C_{\beta^*}(\tilde{q}) \not\succcurlyeq 0$ we would have $\beta(\tilde{q}) < \beta^*$ since $\beta \mapsto \lambda_{\min}(C_\beta(\tilde{q}))$ only has a single root—a contradiction. Thus, since we have shown β^* is the largest β such that $C_{\beta^*}(\tilde{p}) \succcurlyeq 0$ it must be β_ℓ, by Lemma 9, as required. ∎

4 Discussion

In combination with the existing results on mixability, our result bounds the performance of certain predictors in terms of the Hessian of the Bayes risk $\mathsf{H}\underline{L}$ which depends on the choice of loss function. This implies a generalisation of the main result of Kalnishkan and Vyugin (2002a) which shows there can be no "predictive complexity" when the curvature of f_β vanishes (in the binary case). This means there can not exist a mixability constant β_ℓ of the form (1) in such a situation. This is apparent from (20) since β_ℓ is not defined when $\mathsf{H}\tilde{\underline{L}}(\tilde{p})$ is singular (which occurs when $\mathsf{H}f_\beta$ vanishes).

One can use Lemma 9 to confirm that the mixability constant for the Brier score is one, in accord with the calculation of Vovk and Zhdanov (2009). (See Appendix B for the proof.)

The main result is stated for proper losses. However it turns out that this is not really a limitation[4]. Suppose $\ell_{\text{imp}} \colon [n] \times \mathcal{V} \to [0, +\infty]$ is an *improper* loss (i.e. not proper). Let $L_{\text{imp}} \colon \Delta^n \times \mathcal{V} \to [0, +\infty]$ and $\underline{L}_{\text{imp}} \colon \Delta^n \to [0, +\infty]$ denote the corresponding conditional risk and conditional Bayes risk respectively. Let $\psi_{\text{imp}} \colon \Delta^n \to \mathcal{V}$ be a *reference link* (cf. Reid and Williamson (2010))—that is, a (possibly non-unique) function satisfying

$$L_{\text{imp}}(p, \psi_{\text{imp}}(p)) = \underline{L}_{\text{imp}}(p).$$

This function can be seen as one which "calibrates" ℓ_{imp} by returning $\psi_{\text{imp}}(p)$, the best possible prediction under labels distributed by p. Let

$$\ell(y, q) := \ell_{\text{imp}}(y, \psi_{\text{imp}}(q)), \quad y \in [n],\ q \in \Delta^n \tag{21}$$

and thus

$$L(p, q) = L_{\text{imp}}(p, \psi_{\text{imp}}(q)), \quad p, q \in \Delta^n.$$

We claim that ℓ is proper. It suffices to show that $p \in \arg\min_{q \in \Delta^n} L(p, q)$ which we demonstrate by contradiction. Thus suppose that for arbitrary $p \in \Delta^n$, there exists $p^* \neq p$ such that

$$\begin{aligned} & L(p, p^*) && < L(p, p) \\ \Leftrightarrow \quad & L_{\text{imp}}(p, \psi_{\text{imp}}(p^*)) && < L_{\text{imp}}(p, \psi_{\text{imp}}(p)) = \underline{L}_{\text{imp}}(p) = \min_{v \in \mathcal{V}} L_{\text{imp}}(p, v) \end{aligned}$$

which is indeed a contradiction. Thus ℓ defined by (21) is proper. Observe too that $\underline{L}_{\text{imp}}(p) = L_{\text{imp}}(p, \psi_{\text{imp}}(p)) = L(p, p) = \underline{L}(p)$. Thus the method of identifying the conditional Bayes risk of an improper loss with that of a proper loss (confer (Grünwald and Dawid, 2004, §3.4) and Chernov et al. (2010)) is equivalent to the above use of the reference link.

We now briefly relate our result to recent work by Abernethy et al. (2009). They formulate the problem slightly differently. They do not restrict themselves to proper losses and so the predictions are not restricted to the simplex. This means it is not necessary to go to a submanifold in order for derivatives to be well defined. (It may well be that one can avoid the explicit projection down to $\tilde{\Delta}^n$ using the intrinsic methods of differential geometry (Thorpe, 1979); we have been unable as yet to prove our result using that machinery.)

[4]We thank a referee for pointing this out by referring us to Chernov et al. (2010).

Abernethy et al. (2009) have developed their own bounds on cumulative loss in terms of the α-flatness (defined below) of $\underline{L}$. They show that α-flatness is implied by strong convexity of the loss ℓ. The duality between the loss surface and Bayes risk that they established through the use of support functions can also be seen in Lemma 5 in the relationship between the Hessian of $\tilde{\underline{L}}$ and the derivative of $\tilde{\ell}$. Although it is obscured somewhat due to our use of functions of $\tilde{p}$, this relationship is due to the properness of ℓ guaranteeing that ℓ^{-1} is the (homogeneously extended) Gauss map for the surface $\tilde{\underline{L}}$. Below we point out the relationship between α-flatness and the positive definiteness of $\mathsf{H}\underline{L}$ (we stress that in our work we used $\mathsf{H}\tilde{\underline{L}}$). The connection below suggests that the α-flatness condition is stronger than necessary.

A convex function $f\colon \mathcal{X} \to \mathbb{R}$ is said to be *α-flat* if for all $x, x_0 \in \mathcal{X}$,

$$f(x) - f(x_0) \leq \mathsf{D}f(x_0) \cdot (x - x_0) + \alpha \|x - x_0\|^2. \tag{22}$$

A concave function g is α-flat if the convex function $-g$ is α-flat.

Theorem 11 *For $\alpha > 0$, f is α-flat if and only if $f - \alpha\|\cdot\|^2$ is concave.*

Proof: Hiriart-Urruty and Lemaréchal (1993, page 183) show a function h is convex if and only if

$$h(x) \geq h(x_0) + \mathsf{D}h(x_0) \cdot (x - x_0), \quad \forall x, x_0.$$

A function h is concave if and only if $-h$ is convex. Thus h is concave if and only

$$h(x) \leq h(x_0) + \mathsf{D}h(x_0) \cdot (x - x_0), \quad \forall x, x_0.$$

Let $h(x) = f(x) - \alpha\|x\|^2$. The concavity of h is equivalent to the following holding for all x, x_0:

$$\begin{aligned}
& f(x) - \alpha\|x\|^2 \leq f(x_0) - \alpha\|x_0\|^2 + (\mathsf{D}f(x_0) - 2\alpha x_0) \cdot (x - x_0) \\
\Leftrightarrow \quad & f(x) - \alpha\|x\|^2 \leq f(x_0) - \alpha\|x_0\|^2 + \mathsf{D}f(x_0) \cdot (x - x_0) - 2\alpha x_0 \cdot (x - x_0) \\
\Leftrightarrow \quad & f(x) \leq f(x_0) - \alpha\|x_0\|^2 + \mathsf{D}f(x_0) \cdot (x - x_0) + \alpha\|x\|^2 - 2\alpha x_0 \cdot x + 2\alpha\|x_0\|^2 \\
\Leftrightarrow \quad & f(x) \leq f(x_0) + \mathsf{D}f(x_0) \cdot (x - x_0) + \alpha\|x - x_0\|^2 \\
\Leftrightarrow \quad & (22).
\end{aligned}$$

■

Thus f is α-flat if and only if $\mathsf{H}(f - \alpha\|\cdot\|^2)$ is negative semidefinite, which is equivalent to $\mathsf{H}f - 2\alpha I \preccurlyeq 0 \iff \mathsf{H}f \preccurlyeq 2\alpha I$. Hence requiring $-\underline{L}$ is α-flat is a constraint on the curvature of $\underline{L}$ relative to a flat surface: $\underline{L}$ is α-flat iff $\mathsf{H}\underline{L} \succcurlyeq -2\alpha I$. However our main result shows that the mixability constant (which is the best possible constant one can have in a bound such as (1)) is governed by the curvature of $\tilde{\underline{L}}$ *normalised by the curvature of* $\tilde{\underline{L}}_{\log}$. The necessity of comparison with log loss is not that surprising in light of the observations regarding mixability by Grünwald (2007, §17.9).

5 Conclusion

We have characterised the mixability constant for strictly proper multiclass losses (and shown how the result also applies to improper losses). The result shows in a precise and intuitive way the effect of the choice of loss function on the performance of an aggregating forecaster and the special role played by Log-loss in such settings.

Acknowledgements

This work was supported by the Australian Research Council and NICTA through backing Australia's ability. Some of the work was done while all the authors were visiting Microsoft Research, Cambridge and some was done while Tim van Erven was visiting ANU and NICTA. It was also supported in part by the IST Programme of the European Community, under the PASCAL2 Network of Excellence, IST-2007-216886. This publication only reflects the authors' views.

References

Jacob Abernethy, Alekh Agarwal, Peter L. Bartlett, and Alexander Rakhlin. A stochastic view of optimal regret through minimax duality. In *Proceedings of the 22nd Annual Conference on Learning Theory*, 2009.

Nicolò Cesa-Bianchi and Gabor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

Alexey Chernov, Yuri Kalnishkan, Fedor Zhdanov, and Vladimir Vovk. Supermartingales in prediction with expert advice. *Theoretical Computer Science*, 411:2647–2669, 2010.

Paul K. Fackler. Notes on matrix calculus. North Carolina State University, 2005.

Wendell H. Fleming. *Functions of Several Variables*. Springer, 1977.

Peter D. Grünwald. *The Minimum Description Length Principle*. MIT Press, Cambridge, MA, 2007.

Peter D. Grünwald and A. Phillip Dawid. Game theory, maximum entropy, minimum discrepancy and robust bayesian decision theory. *The Annals of Statistics*, 32(4):1367–1433, 2004.

David Haussler, Jyrki Kivinen, and Manfred K. Warmuth. Sequential prediction of individual sequences under general loss functions. *IEEE Transactions on Information Theory*, 44(5): 1906–1925, 1998.

Jean-Baptiste Hiriart-Urruty and Claude Lemaréchal. *Convex Analysis and Minimization Algorithms: Part I: Fundamentals*. Springer, Berlin, 1993.

Roger A. Horn and Charles R. Johnson. *Matrix analysis*. Cambridge University Press, 1985.

Yuri Kalnishkan and Michael V. Vyugin. On the absence of predictive complexity for some games. In *Proceedings of the 13th International Conference on Algorithmic Learning Theory*, volume 2533 of *Lecture Notes in Artificial Intelligence*, pages 164–172. Springer-Verlag, 2002a.

Yuri Kalnishkan and Michael V. Vyugin. Mixability and the existence of weak complexities. In *The 15th Annual Conference on Computational Learning Theory (COLT 2002)*, volume 2375 of *Lecture Notes in Artificial Intelligence*, pages 105–120. Springer-Verlag, 2002b.

Yuri Kalnishkan and Michael V. Vyugin. The weak aggregating algorithm and weak mixability. *Journal of Computer and System Sciences*, 74:1228–1244, 2008.

Yuri Kalnishkan, Volodya Vovk, and Michael V. Vyugin. Loss functions, complexities, and the Legendre transformation. *Theoretical Computer Science*, 313:195–207, 2004.

Jan R. Magnus and Heinz Neudecker. *Matrix Differential Calculus with Applications in Statistics and Econometrics (revised edition)*. John Wiley & Sons, Ltd., 1999.

Mark D. Reid and Robert C. Williamson. Composite binary losses. *Journal of Machine Learning Research*, 11:2387–2422, 2010.

Mark D. Reid and Robert C. Williamson. Information, divergence and risk for binary experiments. *Journal of Machine Learning Research*, 12:731–817, March 2011.

John A. Thorpe. *Elementary Topics in Differential Geometry*. Springer, 1979.

Volodya Vovk. Aggregating strategies. In *Proceedings of the Third Annual Workshop on Computational Learning Theory (COLT)*, pages 371–383, 1990.

Volodya Vovk. A game of prediction with expert advice. In *Proceedings of the Eighth Annual Conference on Computational Learning Theory*, pages 51–60. ACM, 1995.

Volodya Vovk and Fedor Zhdanov. Prediction with expert advice for the Brier game. *Journal of Machine Learning Research*, 10:2445–2471, 2009.

A Matrix Calculus

We adopt notation from Magnus and Neudecker (1999): I_n is the $n \times n$ identity matrix, A' is the transpose of A, the n-vector $\mathbb{1}_n := (1, \ldots, 1)'$, and $0_{n \times m}$ denotes the zero matrix with n rows and m columns. The unit n-vector $e_i^n := (0, \ldots, 0, 1, 0, \ldots, 0)'$ has a 1 in the ith coordinate and zeroes elsewhere. If $A = [a_{ij}]$ is an $n \times m$ matrix, $\operatorname{vec} A$ is the vector of columns of A stacked on top of each other. The *Kronecker product* of an $m \times n$ matrix A with a $p \times q$ matrix B is the $mp \times nq$ matrix

$$A \otimes B := \begin{pmatrix} A_{1,1}B & \cdots & A_{1,n}B \\ \vdots & \ddots & \vdots \\ A_{m,1}B & \cdots & A_{m,n}B \end{pmatrix}.$$

We use the following properties of Kronecker products (see Chapter 2 of Magnus and Neudecker (1999)): $(A \otimes B)(C \otimes D) = (AC \otimes BD)$ for all appropriately sized A, B, C, D and $(A \otimes B)^{-1} = (A^{-1} \otimes B^{-1})$ for invertible A and B.

If $f : \mathbb{R}^n \to \mathbb{R}^m$ is differentiable at c then the *partial derivative* of f_i w.r.t. the jth coordinate at c is denoted $\mathsf{D}_j f_i(c)$ and is often[5] also written as $[\partial f_i / \partial x_j]_{x=c}$. The $m \times n$ matrix of partial derivatives of f is the *Jacobian* of f and denoted

$$(\mathsf{D}f(c))_{i,j} := \mathsf{D}_j f_i(c) \quad \text{for } i \in [m], j \in [n].$$

The *inverse function theorem* relates the Jacobians of a function and its inverse (cf. Fleming (1977, §4.5)):

Theorem 12 *Let $S \subset \mathbb{R}^n$ be an open set and $g : S \to \mathbb{R}^n$ be a C^q function with $q \geq 1$ (i.e., continuous with at least one continuous derivative). If $\mathsf{D}g(s) \neq 0$ then: there exists an open set S_0 such that $s \in S_0$ and the restriction of g to S_0 is invertible; $g(S_0)$ is open; f, the inverse of the restriction of g to S_0, is C^q; and $\mathsf{D}f(t) = [\mathsf{D}g(s)]^{-1}$ for $t = g(s)$ and $s \in S_0$.*

If F is a matrix valued function $\mathsf{D}F(X) := \mathsf{D}f(\operatorname{vec} X)$ where $f(X) = \operatorname{vec} F(X)$.

We will require the product rule for matrix valued functions (Fackler, 2005): Suppose $f \colon \mathbb{R}^n \to \mathbb{R}^{m \times p}$, $g \colon \mathbb{R}^n \to \mathbb{R}^{p \times q}$ so that $(f \times g) \colon \mathbb{R}^n \to \mathbb{R}^{m \times q}$. Then

$$\mathsf{D}(f \times g)(x) = (g(x)' \otimes I_m) \cdot \mathsf{D}f(x) + (I_q \otimes f(x)) \cdot \mathsf{D}g(x). \tag{23}$$

The *Hessian* at $x \in X \subseteq \mathbb{R}^n$ of a real-valued function $f : \mathbb{R}^n \to \mathbb{R}$ is the $n \times n$ real, symmetric matrix of second derivatives at x

$$(\mathsf{H}f(x))_{j,k} := \mathsf{D}_{k,j} f(x) = \frac{\partial^2 f}{\partial x_k \partial x_j}.$$

Note that the derivative $D_{k,j}$ is in row j, column k. It is easy to establish that the Jacobian of the transpose of the Jacobian of f is the Hessian of f. That is,

$$\mathsf{H}f(x) = \mathsf{D}\left((\mathsf{D}f(x))'\right) \tag{24}$$

[5] See Chapter 9 of Magnus and Neudecker (1999) for why the $\partial / \partial x$ notation is a poor one for multivariate differential calculus despite its popularity.

(cf. Chapter 10 of (Magnus and Neudecker, 1999)). If $f : X \to \mathbb{R}^m$ for $X \subseteq \mathbb{R}^n$ is a vector valued function then the Hessian of f at $x \in X$ is the $mn \times n$ matrix that consists of the Hessians of the functions f_i stacked vertically:

$$\mathsf{H}f(x) := \begin{pmatrix} \mathsf{H}f_1(x) \\ \vdots \\ \mathsf{H}f_m(x) \end{pmatrix}.$$

The following theorem regarding the chain rule for Hessian matrices can be found in (Magnus and Neudecker, 1999, pg. 110).

Theorem 13 *Let S be a subset of $\mathbb{R}^n$, and $f : S \to \mathbb{R}^m$ be twice differentiable at a point c in the interior of S. Let T be a subset of $\mathbb{R}^m$ containing $f(S)$, and $g : T \to \mathbb{R}^p$ be twice differentiable at the interior point $b = f(c)$. Then the function $h(x) := g(f(x))$ is twice differentiable at c and*

$$\mathsf{H}h(c) = (I_p \otimes \mathsf{D}f(c))' \cdot (\mathsf{H}g(b)) \cdot \mathsf{D}f(c) + (\mathsf{D}g(b) \otimes I_n) \cdot \mathsf{H}f(c).$$

Applying the chain rule to functions that are inverses of each other gives the following corollary.

Corollary 14 *Suppose $f : \mathbb{R}^n \to \mathbb{R}^n$ is invertible with inverse $g := f^{-1}$. If $b = f(c)$ then*

$$\mathsf{H}f^{-1}(b) = -\left(G \otimes G'\right) \mathsf{H}f(c) G$$

where $G := [\mathsf{D}f(c)]^{-1} = \mathsf{D}g(b)$.

Proof: Since $f \circ g = \mathrm{id}$ and $\mathsf{H}[\mathrm{id}] = 0_{n^2 \times n}$ Theorem 13 implies that for c in the interior of the domain of f and $b = f(c)$

$$\mathsf{H}(g \circ f)(c) = (I_n \otimes \mathsf{D}f(c))' \cdot \mathsf{H}g(b) \cdot \mathsf{D}f(c) + (\mathsf{D}g(b) \otimes I_n) \cdot \mathsf{H}f(c) = 0_{n^2 \times n}.$$

Solving this for $\mathsf{H}g(b)$ gives

$$\mathsf{H}g(b) = -\left[(I_n \otimes \mathsf{D}f(c))'\right]^{-1} (\mathsf{D}g(b)) \otimes I_n) \cdot \mathsf{H}f(c) \cdot [\mathsf{D}f(c)]^{-1}.$$

Since $(A \otimes B)^{-1} = (A^{-1} \otimes B^{-1})$ and $(A')^{-1} = (A^{-1})'$ we have $[(I \otimes B)']^{-1} = [(I \otimes B)^{-1}]' = (I^{-1} \otimes B^{-1})' = (I \otimes B^{-1})'$ so the first term in the above product simplifies to $-\left[(I_n \otimes \mathsf{D}f(c)^{-1})\right]'$. The inverse function theorem implies $\mathsf{D}g(b) = [\mathsf{D}f(c)]^{-1} =: G$ and so

$$\begin{aligned} \mathsf{H}g(b) &= -(I_n \otimes G)' \cdot (G \otimes I_n) \cdot \mathsf{H}f(c) \cdot G \\ &= -(G \otimes G') \cdot \mathsf{H}f(c) \cdot G \end{aligned}$$

as required, since $(A \otimes B)(C \otimes D) = (AC \otimes BD)$. ■

B Mixability of the Brier Score

The n-class Brier score is[6]

$$\ell_{\mathrm{Brier}}(y, \hat{p}) = \sum_{i=1}^{n} (\llbracket y_i = 1 \rrbracket - \hat{p}_i)^2,$$

where $y \in \{0, 1\}^n$ and $\hat{p} \in \Delta^n$. Thus

$$L_{\mathrm{Brier}}(p, \hat{p}) = \sum_{i=1}^{n} \mathbb{E}_{\mathsf{Y} \sim p} (\llbracket \mathsf{Y}_i = 1 \rrbracket - \hat{p}_i)^2 = \sum_{i=1}^{n} (p_i - 2 p_i \hat{p}_i + \hat{p}_i^2).$$

[6]This is the definition used by Vovk and Zhdanov (2009). Cesa-Bianchi and Lugosi (2006) use a different definition (for the binary case) which differs by a constant. Their definition results in $\underline{\tilde{L}}(\tilde{p}) = \tilde{p}(1 - \tilde{p})$ and thus $\underline{\tilde{L}}''(\tilde{p}) = -2$. If $n = 2$, then $\underline{\tilde{L}}_{\mathrm{Brier}}$ as defined above leads to $\underline{\tilde{L}}''_{\mathrm{Brier}}(\tilde{p}) = H\underline{\tilde{L}}_{\mathrm{Brier}}(\tilde{p}) = -2(1 + 1) = -4$.

Hence $\underline{L}_{\text{Brier}}(p) = L_{\text{Brier}}(p,p) = \sum_{i=1}^n (p_i - 2p_i p_i + p_i^2) = 1 - \sum_{i=1}^n p_i^2$ since $\sum_{i=1}^n p_i = 1$, and $\tilde{\underline{L}}_{\text{Brier}}(\tilde{p}) = 1 - \sum_{i=1}^{n-1} p_i^2 - \left(1 - \sum_{i=1}^{n-1} p_i\right)^2$.

As first proved by Vovk and Zhdanov (2009), the Brier score is mixable with mixability constant 1. We will reprove this result using the following restatement of Lemma 9:

Lemma 15 *Let ℓ be a twice differentiable, strictly proper loss, with Bayes risk $\tilde{\underline{L}}(\tilde{p}) := \underline{L}(p)$. Let $\tilde{\underline{L}}_{\log}(\tilde{p}) := \underline{L}_{\log}(p)$ be the Bayes risk for the log loss. Then the following statements are equivalent:*

(i.) ℓ is β-mixable;

(ii.) $\beta\underline{L}(p) - \underline{L}_{\log}(p)$ is convex;

(iii.) $\beta\tilde{\underline{L}}(\tilde{p}) - \tilde{\underline{L}}_{\log}(\tilde{p})$ is convex.

Proof: Equivalence of (i) and (iii) follows from Lemma 9 upon observing that $\beta\tilde{\underline{L}}(\tilde{p}) - \tilde{\underline{L}}_{\log}(\tilde{p})$ is convex if and only if $\beta\mathsf{H}\tilde{\underline{L}}(\tilde{p}) \succcurlyeq \mathsf{H}\tilde{\underline{L}}_{\log}(\tilde{p})$ (Hiriart-Urruty and Lemaréchal, 1993). Equivalence of (ii) and (iii) follows by linearity of the map $p_n(\tilde{p}) = 1 - \sum_{i=1}^{n-1} \tilde{p}_i$. ■

Theorem 16 *The Brier score is mixable, with mixability constant $\beta_{\text{Brier}} = 1$.*

Proof: It can be verified by basic calculus that ℓ_{Brier} is twice differentiable. To see that it is strictly proper, note that for $\hat{p} \neq p$ the inequality $L_{\text{Brier}}(p,\hat{p}) > \underline{L}_{\text{Brier}}(p)$ is equivalent to

$$\sum_{i=1}^n (p_i^2 - 2p_i\hat{p}_i + \hat{p}_i^2) > 0 \qquad \text{or} \qquad \sum_{i=1}^n (p_i - \hat{p}_i)^2 > 0,$$

and the latter inequality is true because $p_i \neq \hat{p}_i$ for at least one i by assumption. Hence the conditions of Lemma 15 are satisfied.

We will first prove that $\beta_{\text{Brier}} \leq 1$ by showing that convexity of $\beta\tilde{\underline{L}}_{\text{Brier}}(\tilde{p}) - \tilde{\underline{L}}_{\log}(\tilde{p})$ implies $\beta \leq 1$. If $\beta\tilde{\underline{L}}_{\text{Brier}}(\tilde{p}) - \tilde{\underline{L}}_{\log}(\tilde{p})$ is convex, then it is convex as a function of p_1 when all other elements of $\tilde{p}$ are kept fixed. Consequently, the second derivative with respect to p_1 must be nonnegative:

$$0 \leq \frac{\partial^2}{\partial p_1^2}\left(\beta\tilde{\underline{L}}_{\text{Brier}}(\tilde{p}) - \tilde{\underline{L}}_{\log}(\tilde{p})\right) = \frac{1}{p_1} + \frac{1}{p_n} - 4\beta.$$

By evaluating at $p_1 = p_n = 1/2$, it follows that $\beta \leq 1$.

It remains to show that $\beta_{\text{Brier}} \geq 1$. By Lemma 15 it is sufficient to show that, for $\beta \leq 1$, $\beta\underline{L}_{\text{Brier}}(p) - \underline{L}_{\log}(p)$ is convex. We proceed by induction. For $n = 1$, the required convexity holds trivially. Suppose the lemma holds for $n-1$, and let $f_n(p_1,\ldots,p_n) = \beta\underline{L}_{\text{Brier}}(p) - \underline{L}_{\log}(p)$ for all n. Then for $n \geq 2$

$$f_n(p_1,\ldots,p_n) = f_{n-1}(p_1 + p_2, p_3, \ldots, p_n) + g(p_1,p_2),$$

where $g(p_1,p_2) = -\beta p_1^2 - \beta p_2^2 + \beta(p_1+p_2)^2 + p_1\ln p_1 + p_2 \ln p_2 - (p_1+p_2)\ln(p_1+p_2)$. As f_{n-1} is convex by inductive assumption and the sum of two convex functions is convex, it is therefore sufficient to show that $g(p_1,p_2)$ is convex or, equivalently, that its Hessian is positive semi-definite. Abbreviating $q = p_1 + p_2$, we have that

$$\mathsf{H}g(p_1,p_2) = \begin{pmatrix} 1/p_1 - 1/q & 2\beta - 1/q \\ 2\beta - 1/q & 1/p_2 - 1/q \end{pmatrix}.$$

A 2×2 matrix is positive semi-definite if its trace and determinant are both non-negative, which is easily verified in the present case: $\mathsf{Tr}(\mathsf{H}g(p_1,p_2)) = 1/p_1 + 1/p_2 - 2/q \geq 0$ and $|\mathsf{H}g(p_1,p_2)| =$

$(1/p_1 - 1/q)(1/p_2 - 1/q) - (2\beta - 1/q)^2$, which is non-negative if

$$\begin{aligned} \frac{1}{p_1 p_2} - \frac{1}{p_1 q} - \frac{1}{p_2 q} &\geq 4\beta^2 - \frac{4\beta}{q} \\ 0 &\geq 4\beta^2 q - 4\beta \\ \beta q &\leq 1. \end{aligned}$$

As $q = p_1 + p_2 \leq 1$, this inequality holds for $\beta \leq 1$, which shows that $g(p_1, p_2)$ is convex and thereby completes the proof. ■