Mixability is Bayes Risk Curvature Relative to Log Loss

Tim van Erven, Mark D. Reid, Robert C. Williamson

Journal of Machine Learning Research 13 (2012) 1639-1663 Submitted 8/11; Revised 3/12; Published 5/12

Mixability is Bayes Risk Curvature Relative to Log Loss

Tim van Erven TIM@TIMVANERVEN.NL
Département de Mathématiques
Université Paris-Sud
91405 Orsay Cedex, France

Mark D. Reid* MARK.REID@ANU.EDU.AU
Robert C. Williamson* BOB.WILLIAMSON@ANU.EDU.AU
Research School of Computer Science
The Australian National University
Canberra ACT 0200, Australia

Editor: Gábor Lugosi

Abstract

Mixability of a loss characterizes fast rates in the online learning setting of prediction with expert advice. 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 maximum eigenvalue of 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.

Keywords: mixability, multiclass, prediction with expert advice, proper loss, learning rates

1. Introduction

In *prediction with expert advice* (Vovk, 1990, 1995, 2001; Cesa-Bianchi and Lugosi, 2006) a learner has to predict a sequence of outcomes, which might be chosen adversarially. The setting is online, meaning that learning proceeds in rounds; and the learner is aided by a finite number of experts. At the start of each round, all experts first announce their predictions for that round, then the learner has to make a prediction, and finally the real outcome is revealed. The discrepancy between a prediction and an outcome is measured by a *loss function*, and losses add up between rounds. Finally, the goal for the learner is to minimize their *regret*, which is the difference between their cumulative loss and the cumulative loss of the best expert after T rounds.

Strategies for the learner usually come with guaranteed bounds on the regret in the worst case over all possible outcomes and expert predictions, which ensures good learning performance under all circumstances. How strong these guaranteed bounds can be depends on the loss function. Some losses are easy in the sense that the worst-case regret can be bounded by a constant, which is $O(1)$ in T. For other losses only a rate of $O(\sqrt{T})$ or worse can be guaranteed (Kalnishkan and Vyugin,

∗. Also affiliated with National ICT Australia (NICTA)

2008). Our results provide new insight and new technical tools for the class of losses for which fast, $O(1)$ rates are possible.

1.1 Fast Rates and Mixability

It is known that, under very general conditions, $O(1)$ rates are possible if and only if the loss is η-*mixable* (defined below) for some $\eta > 0$, which means that mixability characterizes fast rates. More specifically, if a loss is η-mixable and there are N experts, then using the so-called *aggregating algorithm* (Vovk, 2001) the learner is guaranteed to have regret bounded by

$$\frac{\ln N}{\eta}, \tag{1}$$

which does not grow with T. Conversely, if the loss is not η-mixable for any $\eta > 0$ and satisfies very mild regularity conditions, then it is not possible to bound the worst-case regret by an additive constant for any strategy (Kalnishkan and Vyugin, 2008; Vovk, 1995). Examples of mixable losses include the logarithmic loss, the relative entropy loss, the square loss on binary outcomes (Haussler et al., 1998) and the Brier score (Vovk and Zhdanov, 2009), which are all 1-mixable except for the square loss, which is 2-mixable.

A related condition requires the loss to be *exp-concave* (Cesa-Bianchi and Lugosi, 2006). Although exp-concavity implies mixability, the converse is not true, and therefore exp-concavity does not characterize fast rates.

Although mixability is associated with fast rates, it also appears in the analysis of losses with $O(\sqrt{T})$ rates. For example, the analysis of Kalnishkan and Vyugin (2008) may be interpreted as approximating non-mixable losses by a sequence of η-mixable losses with η going to zero (Kalnishkan and Vyugin, 2008, Remark 19). Thus mixability appears to be one of the most fundamental properties to study in the prediction with expert advice setting.

1.2 Main Results

The aggregating algorithm depends on η, and its regret bound (1) is optimized when η is as large as possible. For any loss of interest ℓ, it is thus desirable to know the largest η for which ℓ is η-mixable. We call this the *mixability constant* for ℓ.

For outcomes with two possible values, determining the mixability constant is straight-forward using a formula due to Haussler et al. (1998), but their expression has no clear interpretation. In Section 4.1 we show how, for the important class of *proper losses*, the result by Haussler et al. simplifies considerably, and may be expressed in terms of the curvature of the *Bayes risk* of the loss relative to the Bayes risk for the logarithmic loss. The relevant notions of properness and Bayes risk will first be reviewed in Section 3.

We refer to the case where outcomes have more than two possible values as the *multiclass* setting. Here no general result has previously been available, and the mixability constant has only been determined for a limited number of cases (mainly logarithmic loss and the Brier score). Our main contribution is a simple explicit formula for the mixability constant in the multiclass setting (Theorem 13 and Corollary 14), which generalises our result for binary-valued outcomes. Along the way we develop other useful characterizations of mixability in Theorem 10. We illustrate the usefulness of our results by giving a short proof for 1-mixability of the multiclass Brier score in Section 5, which is simpler than the previously known proof (Vovk and Zhdanov, 2009).

Although our results are stated for proper losses, in Section 6 we show how they carry across to losses that are not proper.

1.3 Outline

The paper is structured as follows. In the next section we introduce general notation. Then Section 3 reviews the class of proper losses and the definition of Bayes risk, along with some of their properties that are required later. It also states Condition A, which lists several continuity conditions on the loss that are required for our main results.

In Section 4 we come to the main part of the paper. There mixability is formally defined, and in Section 4.1 we state our results for binary-valued outcomes. The remainder of Section 4 is devoted to generalising this result to the multiclass setting (Theorem 13 and Corollary 14). An important intermediate result is stated in Theorem 10, and we discuss some of its direct consequences in Corollaries 11 and 12. These show that the sum of two η-mixable losses is η-mixable and that the logarithmic loss is the “most mixable” in a sense.

Section 5 contains a simplified proof for 1-mixability of the Brier score. And in Section 6 we show how our results carry across to losses that are not proper. In Section 7 we also relate our results to recent work by Abernethy et al. (2009) in a related online learning setting. Our proofs in Section 4 require some results from matrix calculus, which we review briefly in Appendix A.

2. Setting

We consider a game of *prediction with expert advice*, which goes on for rounds $t = 1,\dots,T$. At the start of each round t, N experts choose their predictions $v_t^1,\dots,v_t^N$ from a set $\mathcal{V}$; then the learner chooses their prediction $v_t \in \mathcal{V}$; and finally the true outcome $y_t \in \mathcal{Y} = \{1,\dots,n\}$ is revealed. When the outcomes are binary-valued, $n = 2$, but in the multiclass setting n can be any positive integer. Losses are measured by a function $\ell \colon \mathcal{Y}\times\mathcal{V} \to [0,\infty]$ and over the course of the game add up to $\mathrm{Loss}(T) := \sum_{t=1}^T \ell(y_t,v_t)$ for the learner and to $\mathrm{Loss}_j(T) = \sum_{t=1}^T \ell(y_t,v_t^j)$ for the j-th expert. The goal for the learner is to predict nearly as well as the best expert, as measured by the *regret*

$$R(T) = \mathrm{Loss}(T) - \min_j \mathrm{Loss}_j(T).$$

Typical strategies in the literature come with bounds on the regret that hold in the worst case, for any possible expert predictions and any possible sequence of outcomes. In particular, if the loss ℓ is η-mixable for some $\eta > 0$ and the learner predicts according to the aggregating algorithm, then the regret is bounded by

$$R(T) \le \frac{\ln N}{\eta}, \tag{2}$$

no matter what the expert predictions or the outcomes are.

2.1 Notation

We use the following notation throughout. Let $[n] := \{1,\dots,n\}$ and denote by $\mathbb{R}_+$ the non-negative reals. The transpose of a vector x is x'. If x is a n-vector, $A = \mathrm{diag}(x)$ 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 $\mathrm{diag}(x_i)_{i=1}^n := \mathrm{diag}(x_1,\dots,x_n) := \mathrm{diag}((x_1,\dots,x_n)')$. 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. semi-definite), then we write $A \succ B$ (resp. $A \succcurlyeq B$). The n-simplex $\Delta^n := \{(x_1, \dots, x_n)' \in \mathbb{R}^n \colon x_i \geq 0,\ i \in [n],\ \sum_{i=1}^n x_i = 1\}$. Other notation (the Kronecker product $\otimes$, the derivative D, and the Hessian H) is defined in Appendix A, which also includes several matrix calculus results we use.

3. Proper Multiclass Losses

We consider multiclass losses for class probability estimation, in which predictions are probability distributions: $\mathcal{V} = \Delta^n$. As we will often consider how the loss changes as a function of the predicted distribution $q \in \Delta^n$, it is convenient to define a *partial loss function* $\ell_i(q) = \ell(i, q)$ for any outcome $i \in [n]$. Together these partial loss functions make up the full *loss function* $\ell : \Delta^n \to [0, \infty]^n$, which assigns a loss vector $\ell(q) = (\ell_1(q), \dots, \ell_n(q))'$ to each distribution $q \in \Delta^n$. 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,

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

A loss is called *proper* whenever the minimal risk is always achieved by predicting the true outcome distribution, that is, $\underline{L}(p) = L(p, p)$ for all $p \in \Delta^n$. A proper loss is *strictly proper* if there exists no $q \neq p$ such that $L(p, q) = \underline{L}(p)$. For example, the *log loss* $\ell_{\log}(p) := (-\ln(p_1), \dots, -\ln(p_n))'$ is strictly proper, and its corresponding Bayes risk is the entropy $\underline{L}_{\log}(p) = -\sum_{i=1}^n p_i \ln(p_i)$.

We call a proper loss ℓ *strongly invertible* if for all distributions $p \neq q \in \Delta^n$ there exists at least one outcome $i \in [n]$ such that $\ell_i(p) \neq \ell_i(q)$ and $p_i > 0$. Note that without the requirement that $p_i > 0$ this would be ordinary invertibility. One might also understand strong invertibility as saying that the loss should be invertible, and if we restrict the game to a face of the simplex (effectively removing one possible outcome), then the loss function for the resulting game should again be strongly invertible.

Since it is central to our results, we will assume all losses are strictly proper for the remainder of the paper (except Section 6 where we show how the assumption may be relaxed). Lemma 2 in the next section shows that strictness is not such a strong requirement.

3.1 Projecting Down to $n-1$ Dimensions

Because probabilities sum up to one, any $p \in \Delta^n$ is fully determined by its first $n-1$ components $\tilde{p} = (p_1, \dots, p_{n-1})$. It follows that any function of p can also be expressed as a function of $\tilde{p}$, which is convenient in order to use the standard rules when taking derivatives on Δ^n. To go back and forth between p and $\tilde{p}$, we define $p_n(\tilde{p}) := 1 - \sum_{i=1}^{n-1} \tilde{p}_i$ and the projection

$$\Pi_\Delta(p) := (p_1, \dots, p_{n-1})',$$

which is a continuous and invertible function from Δ^n to $\tilde{\Delta}^n := \{(p_1, \dots, p_{n-1})' \colon p \in \Delta^n\}$, with continuous inverse $\Pi_\Delta^{-1}(\tilde{p}) = (\tilde{p}_1, \dots, \tilde{p}_{n-1}, p_n(\tilde{p}))$. For similar reasons, we sometimes project loss

$$\begin{tikzcd}
\Delta^n \arrow[r, "\ell"] \arrow[d, "\Pi_\Delta"] & \Lambda \arrow[r, "E_\eta"] \arrow[d, "\Pi_\Lambda"] & \Phi_\eta \arrow[d, "\Pi_\Phi"] \\
\tilde{\Delta}^n \arrow[r, "\tilde{\ell}"] \arrow[rr, bend right, "\tau_\eta"'] & \tilde{\Lambda} \arrow[r, "E_\eta"] & \tilde{\Phi}_\eta
\end{tikzcd}$$

Figure 1: Mappings and spaces.

vectors $\ell(p)$ onto their first $n-1$ components $(\ell_1(p),\ldots,\ell_{n-1}(p))'$, using the projection

$$\Pi_\Lambda(\lambda) := (\lambda_1,\ldots,\lambda_{n-1})'.$$

We write $\Lambda := \ell(\Delta^n)$ for the domain of Π_Λ and $\tilde{\Lambda}$ for its range.

For loss functions $\ell(p)$, we will overload notation and abbreviate $\ell(\tilde{p}) := \ell(\Pi_\Delta^{-1}(\tilde{p}))$. In addition, we write

$$\tilde{\ell}(\tilde{p}) := \Pi_\Lambda(\ell(\tilde{p})) = (\ell_1(\tilde{p}),\ldots,\ell_{n-1}(\tilde{p}))'$$

for the first $n-1$ components of the loss (see Figure 1). By contrast, for $\underline{L}(p)$ we will be more careful about its domain, and use the separate notation $\underline{\tilde{L}}(\tilde{p}) := \underline{L}(\Pi_\Delta^{-1}(\tilde{p}))$ when we consider it as a function of $\tilde{p}$.

It may well be that one can avoid the explicit projection down to $n-1$ dimensions using the intrinsic methods of differential geometry (Thorpe, 1979), but we have been unable to prove our results using that machinery. In any case, in order to do calculations, one will need some coordinate system. Our projection simply defines the natural $(n-1)$-dimensional coordinate system on Δ^n.

3.2 First Properties

Our final result requires the following conditions on the loss:

Condition A *The loss $\ell(p)$ is strictly proper, continuous on Δ^n, and continuously differentiable on the relative interior* $\operatorname{rel\,int}(\Delta^n)$ *of its domain.*

As the projection Π_Δ is a linear function, differentiability of $\ell(p)$ is equivalent to differentiability of $\ell(\tilde{p})$, which will usually be easier to verify. Note that it follows from (15) below that existence of $\mathsf{D}\tilde{\ell}$ guarantees the existence of $\mathsf{H}\underline{\tilde{L}}$.

Lemma 1 *Let $\ell(p)$ be a strictly proper loss. Then the corresponding Bayes risk $\underline{L}(p)$ is strictly concave, and if $\ell(p)$ is differentiable on the relative interior* $\operatorname{rel\,int}(\Delta^n)$ *of Δ^n then it satisfies the* stationarity condition

$$p'\mathsf{D}\ell(\tilde{p}) = 0_{n-1} \qquad \textit{for } p \in \operatorname{rel\,int}(\Delta^n). \tag{3}$$

If $\ell(p)$ is also continuous on the whole simplex Δ^n, then Π_Λ, $\ell(p)$ and $\tilde{\ell}(\tilde{p})$ are all continuous and invertible, with continuous inverses.

Proof Let $p_0, p_1 \in \Delta^n$ and let $p_\lambda = (1-\lambda)p_0 + \lambda p_1$. Then for any $\lambda \in (0,1)$

$$\underline{L}(p_\lambda) = p_\lambda'\ell(p_\lambda) = (1-\lambda)L(p_0,p_\lambda) + \lambda L(p_1,p_\lambda) > (1-\lambda)\underline{L}(p_0) + \lambda\underline{L}(p_1),$$

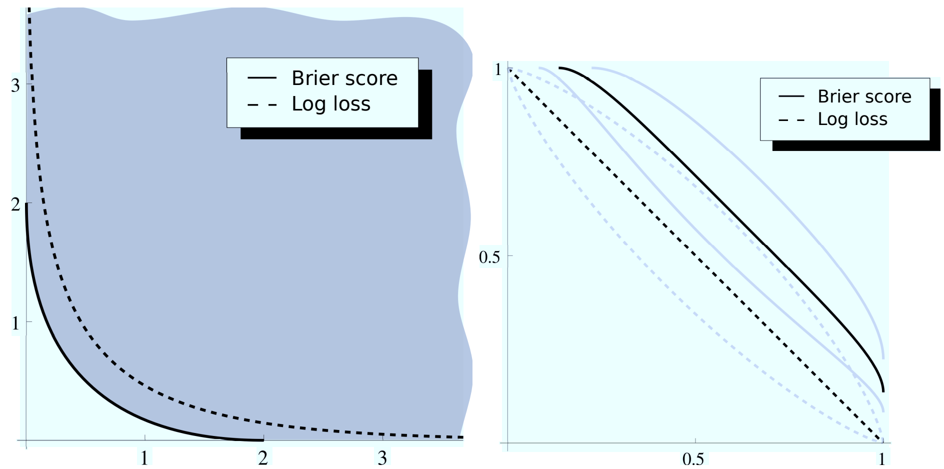

Figure 2: Left: the (boundary of the) superprediction set on two outcomes for the Brier score and the boundary of the superprediction set for log loss. Right: the same boundaries after applying the η-exponential operator for $\eta \in \{3/4, 1, 5/4\}$. The dark curves correspond to $\eta = 1$.

so $\underline{L}(p)$ is strictly concave. Properness guarantees that the function $L_p(\tilde{q}) := L(p, q(\tilde{q}))$ has a minimum at $\tilde{q} = \tilde{p}$. Hence $\mathsf{D}L_p(\tilde{q}) = p'\mathsf{D}\ell(\tilde{q}) = 0_{n-1}$ at $\tilde{q} = \tilde{p}$, giving the stationarity condition.

Now suppose ℓ is continuous on Δ^n, and observe that Π_Λ is also continuous. Then by tracing the relations in Figure 1, one sees that all remaining claims follow if we can establish invertibility of $\tilde{\ell}$ and continuity of its inverse. (Recall that Π_Δ is invertible with continuous inverse.)

To establish invertibility, suppose there exist $\tilde{p} \neq \tilde{q}$ in $\tilde{\Delta}^n$ such that $\tilde{\ell}(\tilde{p}) = \tilde{\ell}(\tilde{q})$ and assume without loss of generality that $\ell_n(p) \leq \ell_n(q)$ (otherwise, just swap them). Then $\underline{L}(q) = \tilde{q}'\tilde{\ell}(\tilde{q}) + q_n\ell_n(q) \geq \tilde{q}'\tilde{\ell}(\tilde{p}) + q_n\ell_n(p) = L(q, p)$, which contradicts strict properness. Hence $\tilde{\ell}$ must be invertible.

To establish continuity of $\tilde{\ell}^{-1}$, we need to show that $\tilde{\ell}(\tilde{p}_m) \to \tilde{\ell}(\tilde{p})$ implies $\tilde{p}_m \to \tilde{p}$ for any sequence $(\tilde{p}_m)_{m=1,2,\ldots}$ of elements from $\tilde{\Delta}^n$. To this end, let $\varepsilon > 0$ be arbitrary. Then it is sufficient to show that there exist only a finite number of elements in $(\tilde{p}_m)$ such that $\|\tilde{p}_m - \tilde{p}\| > \varepsilon$. Towards a contradiction, suppose that $(\tilde{q}_k)_{k=1,2,\ldots}$ is a subsequence of $(\tilde{p}_m)$ such that $\|\tilde{q}_k - \tilde{p}\| \geq \varepsilon$ for all $\tilde{q}_k$. Then the fact that $\tilde{\Delta}^n$ is a compact subset of $\mathbb{R}^{n-1}$ implies (by the Bolzano-Weierstrass theorem) that $(\tilde{q}_k)$ contains a converging subsequence $\tilde{r}_v \to \tilde{r}$. Since continuity of ℓ and Π_Δ^{-1} imply continuity of $\tilde{\ell}$, we have $\tilde{\ell}(\tilde{r}_v) \to \tilde{\ell}(\tilde{r})$. But since $\tilde{r}_v$ is a subsequence of $(\tilde{p}_m)$, we also have that $\tilde{\ell}(\tilde{r}_v) \to \tilde{\ell}(\tilde{p})$ and hence $\tilde{\ell}(\tilde{r}) = \tilde{\ell}(\tilde{p})$. But then strict properness implies that $\tilde{r} = \tilde{p}$, which contradicts the assumption that $\|\tilde{r}_v - \tilde{p}\| \geq \varepsilon$ for all v. ■

4. 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 $\ell\colon \Delta^n \to [0,\infty]^n$ is the set of points in $[0,\infty]^n$ that point-wise dominate some point on the loss surface. That is,

$$S_\ell := \{\lambda \in [0,\infty]^n \colon \exists q \in \Delta^n,\ \forall i \in [n],\ \ell_i(q) \le \lambda_i\}. \tag{4}$$

For any dimension m and $\eta \ge 0$, the η-*exponential operator* $E_\eta\colon [0,\infty]^m \to [0,1]^m$ is defined by

$$E_\eta(\lambda) := (e^{-\eta\lambda_1}, \dots, e^{-\eta\lambda_m}).$$

For $\eta > 0$ it is clearly invertible, with inverse $E_\eta^{-1}(\phi) = -\eta^{-1}(\ln\phi_1, \dots, \ln\phi_m)$. We will both apply it for $m = n$ and for $m = n-1$. The dimension will always be clear from the context.

A loss ℓ is η-*mixable* when the set $E_\eta(S_\ell)$ is convex. The largest η such that a loss is η-mixable is of special interest, because it determines the best possible bound in (2). We call this the *mixability constant* and denote it by η_ℓ:

$$\eta_\ell := \max\{\eta \ge 0 \colon\ \ell \text{ is } \eta\text{-mixable}\}.$$

A loss is always 0-mixable, so $\eta_\ell \ge 0$, but note that for $\eta_\ell = 0$ the bound in (2) is vacuous. A loss is therefore called *mixable* only if its mixability constant is positive, that is, $\eta_\ell > 0$.

One may rewrite the definition of $E_\eta(S_\ell)$ as follows:

$$\begin{aligned} E_\eta(S_\ell) &= \{E_\eta(\lambda) \colon \lambda \in [0,\infty]^n,\ \exists q \in \Delta^n,\ \forall i \in [n],\ \ell_i(q) \le \lambda_i\} \\ &= \{z \in [0,1]^n \colon \exists q \in \Delta^n,\ \forall i \in [n],\ e^{-\eta\ell_i(q)} \ge z_i\}, \end{aligned}$$

since $x \mapsto e^{-\eta x}$ is nonincreasing (in fact, decreasing for $\eta > 0$). Hence in order for $E_\eta(S_\ell)$ to be convex $\operatorname{graph}(f_\eta) = \Phi_\eta := \{(e^{-\eta\ell_1(q)}, \dots, e^{-\eta\ell_n(q)}) \colon q \in \Delta^n\}$ needs to be *concave*. Here f_η is the function whose graph is given by the set above. An explicit definition of f_η is given in (11) after we have introduced some more notation. Observe that Φ_η is the (upper) boundary of $E_\eta(S_\ell)$; that is why concavity of f_η corresponds to *convexity* of $E_\eta(S_\ell)$.

Lemma 2 *If a proper, strongly invertible loss ℓ is mixable, then it is strictly proper.*

An example of a mixable proper loss that is not strictly proper, is when $\ell(p)$ does not depend on p. In this case the loss is not invertible.

Proof Suppose ℓ is not strictly proper. Then there exist $p \ne q$ such that $\underline{L}(p) = L(p,q)$. In addition, mixability implies that for any $\lambda \in (0,1)$ there exists a distribution r_λ such that for all $i \in [n]$

$$\ell_i(r_\lambda) \le -\tfrac{1}{\eta_\ell}\log\left((1-\lambda)e^{-\eta_\ell\ell_i(p)} + \lambda e^{-\eta_\ell\ell_i(q)}\right) \le (1-\lambda)\ell_i(p) + \lambda\ell_i(q),$$

where the second inequality follows from (strict) convexity of $x \mapsto e^{-x}$ and is strict when $\ell_i(p) \ne \ell_i(q)$. Since $\ell_i(p) \ne \ell_i(q)$ for at least one i with $p_i > 0$, it follows that

$$L(p, r_\lambda) = p'\ell(r_\lambda) < p'\big((1-\lambda)\ell(p) + \lambda\ell(q)\big) = \underline{L}(p),$$

which contradicts the definition of $\underline{L}(p)$. Thus mixability implies that ℓ must be strictly proper. ■

4.1 The Binary Case

A loss is called binary if there are only two outcomes: $n=2$. For twice differentiable binary losses ℓ it is known (Haussler et al., 1998) that

$$\eta_\ell = \inf_{\tilde{p}\in(0,1)} \frac{\ell_1'(\tilde{p})\ell_2''(\tilde{p}) - \ell_1''(\tilde{p})\ell_2'(\tilde{p})}{\ell_1'(\tilde{p})\ell_2'(\tilde{p})(\ell_2'(\tilde{p}) - \ell_1'(\tilde{p}))}. \tag{5}$$

When a proper binary loss ℓ is differentiable, the stationarity condition (3) implies

$$\begin{aligned} & \tilde{p}\ell_1'(\tilde{p}) + (1-\tilde{p})\ell_2'(\tilde{p}) = 0 \\ \Rightarrow\quad & \tilde{p}\ell_1'(\tilde{p}) = (\tilde{p}-1)\ell_2'(\tilde{p}) \qquad (6) \\ \Rightarrow\quad & \frac{\ell_1'(\tilde{p})}{\tilde{p}-1} = \frac{\ell_2'(\tilde{p})}{\tilde{p}} =: w(\tilde{p}) =: w_\ell(\tilde{p}). \qquad (7) \end{aligned}$$

We have $\underline{\tilde{L}}(\tilde{p}) = \tilde{p}\ell_1(\tilde{p}) + (1-\tilde{p})\ell_2(\tilde{p})$. Thus by differentiating both sides of (6) and substituting into $\underline{\tilde{L}}''(\tilde{p})$ one obtains $\underline{\tilde{L}}''(\tilde{p}) = \frac{\ell_1'(\tilde{p})}{1-\tilde{p}} = -w(\tilde{p})$. (See Reid and Williamson, 2011). Equation 7 implies $\ell_1'(\tilde{p}) = (\tilde{p}-1)w(\tilde{p})$, $\ell_2'(\tilde{p}) = \tilde{p}w(\tilde{p})$ and hence $\ell_1''(\tilde{p}) = w(\tilde{p}) + (\tilde{p}-1)w'(\tilde{p})$ and $\ell_2''(\tilde{p}) = w(\tilde{p}) + \tilde{p}w'(\tilde{p})$. Substituting these expressions into (5) gives

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

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

$$\boxed{\eta_\ell = \inf_{\tilde{p}\in(0,1)} \frac{w_{\log}(\tilde{p})}{w_\ell(\tilde{p})} = \inf_{\tilde{p}\in(0,1)} \frac{\underline{\tilde{L}}''_{\log}(\tilde{p})}{\underline{\tilde{L}}''(\tilde{p})}.} \tag{8}$$

That is, the mixability constant of binary proper losses is the minimal ratio of the second derivatives of the Bayes risks for log loss and the loss in question. The rest of this paper is devoted to the generalisation of (8) 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).

By substituting $w_\ell(\tilde{p}) = \frac{\ell_1'(\tilde{p})}{\tilde{p}-1}$ and $w_{\log}(\tilde{p}) = \frac{1}{\tilde{p}(1-\tilde{p})}$ into (8), one obtains an expression to compute η_ℓ that is simpler than (5):

$$\boxed{\frac{-1}{\eta_\ell} = \inf_{\tilde{p}\in(0,1)} \tilde{p}\ell_1'(\tilde{p}).} \tag{9}$$

This result also generalizes to the multiclass case; see Corollary 14.

4.2 Mixability and the Concavity of the Function f_η

Our aim is to relate mixability of a loss to the curvature of its Bayes risk surface. Since mixability is equivalent to concavity of the function f_η, which maps the first $n-1$ coordinates of Φ_η to the n-th coordinate, we will start by giving an explicit expression for f_η. We will assume throughout that the loss ℓ is strictly proper and continuous on Δ^n.

It is convenient to introduce an auxiliary function $\tau_\eta : \tilde{\Delta}^n \to [0,1]^{n-1}$ as

$$\tau_\eta(\tilde{p}) := E_\eta(\tilde{\ell}(\tilde{p})) = \left(e^{-\eta \ell_1(\tilde{p})}, \ldots, e^{-\eta \ell_{n-1}(\tilde{p})}\right), \tag{10}$$

which maps a distribution $\tilde{p}$ to the first $n-1$ coordinates of an element in Φ_η. The range of τ_η will be denoted $\tilde{\Phi}_\eta$ (see Figure 1). In addition, let the projection $\Pi_\Phi \colon \Phi_\eta \to \tilde{\Phi}_\eta$ map any element of $\phi \in \Phi_\eta$ to its first $n-1$ coordinates $(\phi_1, \ldots, \phi_{n-1})$. Then under our assumptions, all the maps we have defined are well-behaved:

Lemma 3 *Let ℓ be a continuous, strictly proper loss. Then for $\eta > 0$ all functions in Figure 1 are continuous and invertible with continuous inverse.*

Proof Lemma 1 already covers most of the functions. Given that E_η satisfies the required properties, they can be derived for the remaining functions by writing them as a composition of functions for which the properties are known. For example, $\tau_\eta = E_\eta \circ \tilde{\ell}$ is a composition of two continuous and invertible functions, which each have a continuous inverse. ∎

It follows that, under the conditions of the lemma, the function $f_\eta \colon \tilde{\Phi}_\eta \to [0,1]$ may be defined as

$$f_\eta(\tilde{\phi}) = e^{-\eta \ell_n\left(\tau_\eta^{-1}(\tilde{\phi})\right)} \tag{11}$$

and is continuous. Moreover, as $\tilde{\Phi}_\eta$ (the domain of f_η) is the preimage under τ^{-1} of the closed set $\tilde{\Delta}^n$, continuity of τ^{-1} implies that $\tilde{\Phi}_\eta$ is closed as well. However, continuity implies that we may restrict attention to the interiors of $\tilde{\Phi}_\eta$ and of the probability simplex:

Lemma 4 *Let ℓ be a continuous, strictly proper loss. Then, for $\eta > 0$, f_η is concave if and only if it is concave on the interior $\mathrm{int}(\tilde{\Phi}_\eta)$ of its domain. Furthermore this set corresponds to a subset of the interior of the simplex: $\tau_\eta^{-1}(\mathrm{int}(\tilde{\Phi}_\eta)) \subseteq \mathrm{int}(\tilde{\Delta}^n) = \Pi_\Delta(\mathrm{rel\,int}(\Delta^n))$.*

Proof The restriction to $\mathrm{int}(\tilde{\Phi}_\eta)$ follows trivially from continuity of f_η. The set $\tau_\eta^{-1}(\mathrm{int}(\tilde{\Phi}_\eta))$ is the preimage under τ_η of the open set $\mathrm{int}(\tilde{\Phi}_\eta)$. Since τ_η is continuous, it follows that this set must also be open and hence be a subset of the interior of $\tilde{\Delta}^n$. ∎

4.3 Relating Concavity of f_η to the Hessian of $\underline{L}$

The aim of this subsection is to express the Hessian of f_η in terms of the Bayes risk of the loss function defining f_η. We first note that a twice differentiable function $f : X \to \mathbb{R}$ defined on $X \subseteq \mathbb{R}^{n-1}$ 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 $\mathsf{H} f_\eta$.

We start the analysis by considering the η-exponential operator, used in the definition of τ (10):

Lemma 5 *Suppose $\eta > 0$. Then the derivatives of E_η and E_η^{-1} are*

$$\mathsf{D}E_\eta(\lambda) = -\eta \,\mathrm{diag}\left(E_\eta(\lambda)\right) \;\; \textit{and} \;\; \mathsf{D}E_\eta^{-1}(\phi) = -\eta^{-1} \left[\mathrm{diag}(\phi)\right]^{-1}.$$

And the Hessian of E_η^{-1} is

$$\mathsf{H}E_\eta^{-1}(\phi) = \frac{1}{\eta}\begin{bmatrix} \operatorname{diag}(\phi_1^{-2}, 0, \ldots, 0) \\ \vdots \\ \operatorname{diag}(0, \ldots, 0, \phi_n^{-2}) \end{bmatrix}. \tag{12}$$

If $\eta = 1$ and $\ell = \ell_{\log} = p \mapsto -(\ln p_1, \ldots, \ln p_n)'$ is the log loss, then the map τ_1 is the identity map (i.e., $\tilde{\phi} = \tau_1(\tilde{p}) = \tilde{p}$) and $E_1^{-1}(\tilde{p}) = \tilde{\ell}_{\log}(\tilde{p})$ is the (projected) log loss.

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

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

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

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

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

Next we turn our attention to other components of f_η. Using the stationarity condition and invertibility of ℓ from Lemma 1, simple expressions can be derived for the Jacobian and Hessian of the projected Bayes risk $\underline{\tilde{L}}(\tilde{p}) := \underline{L}(\Pi_\Delta^{-1}(\tilde{p}))$:

Lemma 6 *Suppose the loss ℓ satisfied Condition A. Take $\tilde{p} \in \operatorname{int}(\tilde{\Delta}^n)$, and let $y(\tilde{p}) := -\tilde{p}/p_n(\tilde{p})$. 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

$$\mathsf{D}\ell_n(\tilde{p}) = y(\tilde{p})' \cdot \mathsf{D}\tilde{\ell}(\tilde{p}). \tag{13}$$

The projected Bayes risk function $\underline{\tilde{L}}(\tilde{p})$ satisfies

$$\mathsf{D}\underline{\tilde{L}}(\tilde{p}) = \tilde{\ell}(\tilde{p})' - \ell_n(\tilde{p})\mathbb{1}_{n-1}' \tag{14}$$

$$\text{and} \quad \mathsf{H}\underline{\tilde{L}}(\tilde{p}) = Y(\tilde{p})' \cdot \mathsf{D}\tilde{\ell}(\tilde{p}). \tag{15}$$

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

$$\mathsf{H}\underline{\tilde{L}}_{\log}(\tilde{p}) = -Y(\tilde{p})' \cdot [\operatorname{diag}(\tilde{p})]^{-1}. \tag{16}$$

Proof The stationarity condition (Lemma 1) guarantees that $p'\mathsf{D}\ell(\tilde{p}) = 0_{n-1}$ for all $p \in \operatorname{rel}\operatorname{int}(\Delta^n)$. 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 (13).

By the product rule $\mathsf{D}a'b = (\mathsf{D}a')b + a'(\mathsf{D}b)$, we obtain

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

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 is $\underline{\tilde{L}}(\tilde{p}) = \tilde{p}'\tilde{\ell}(\tilde{p}) + p_n(\tilde{p})\ell_n(\tilde{p})$. Taking the derivative and using the product rule gives

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

by (13). Thus, $\mathsf{D}\underline{\tilde{L}}(\tilde{p}) = \tilde{\ell}(\tilde{p})' - \ell_n(\tilde{p})\mathbb{1}'_{n-1}$, establishing (14).

Equation 15 is obtained by taking derivatives once more and using (13) again, yielding

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

as required. Now $\underline{\tilde{L}}(\tilde{p}) = \underline{L}(p_1,\ldots,p_{n-1},p_n(\tilde{p})) = \underline{L}(p_1,\ldots,p_{n-1},1-\sum_{i=1}^{n-1}p_i) = \underline{L}(C(\tilde{p}))$ where C is affine. Since $p \mapsto \underline{L}(p)$ is strictly concave (Lemma 1) it follows (Hiriart-Urruty and Lemaréchal, 1993) that $\tilde{p} \mapsto \underline{\tilde{L}}(\tilde{p})$ is also strictly concave and thus $\mathsf{H}\underline{\tilde{L}}(\tilde{p})$ 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 1).

Finally, Equation 16 holds since Lemma 5 gives us $E_1^{-1} = \tilde{\ell}_{\log}$ so (15) specialises to $\mathsf{H}\underline{\tilde{L}}_{\log}(\tilde{p}) = Y(\tilde{p})'\cdot\mathsf{D}\tilde{\ell}_{\log}(\tilde{p}) = Y(\tilde{p})'\cdot\mathsf{D}E_1^{-1}(\tilde{p}) = -Y(\tilde{p})'\cdot[\mathrm{diag}(\tilde{p})]^{-1}$, also by Lemma 5. ■

4.4 Completion of the Argument

Recall that our aim is to compute the Hessian of the function describing the boundary of the η-exponentiated superprediction set and determine when it is negative semi-definite. The boundary is described by the function f_η which can be written as the composition $h_\eta \circ g_\eta$ where $h_\eta(z) := e^{-\eta z}$ and $g_\eta(\tilde{\phi}) := \ell_n\left(\tau_\eta^{-1}(\tilde{\phi})\right)$. The Hessian of f_η can be expanded in terms of g_η using the chain rule for the Hessian (Theorem 21) as follows.

Lemma 7 *Suppose the loss ℓ satisfies Condition A and $\eta > 0$. Then for all $\tilde{\phi} \in \operatorname{int}(\tilde{\Phi})$, the Hessian of f_η at $\tilde{\phi}$ is*

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

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

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

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

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

We proceed to compute the derivative and Hessian of g_η:

Lemma 8 *Suppose ℓ satisfies Condition A. For $\eta > 0$ and $\tilde{\phi} \in \operatorname{int}(\tilde{\Phi}_\eta)$, let $\lambda := E_\eta^{-1}(\tilde{\phi})$ and $\tilde{p} := \tilde{\ell}^{-1}(\lambda)$. Then*

$$\mathsf{D} g_\eta(\tilde{\phi}) = y(\tilde{p})' A_\eta(\tilde{\phi}) \tag{17}$$

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

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

Proof By definition, $g_\eta(\tilde{\phi}) := \ell_n(\tau_\eta^{-1}(\tilde{\phi}))$. Since $\tau_\eta^{-1} = \tilde{\ell}^{-1} \circ E_\eta^{-1}$ we have $g_\eta = \ell_n \circ \tilde{\ell}^{-1} \circ E_\eta^{-1}$. Thus, by the chain rule, Equation 13 from Lemma 6, and the inverse function theorem, we obtain

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

yielding (17). Since $\tilde{p} = \tau_\eta^{-1}(\tilde{\phi})$ and $\mathsf{H} g_\eta = \mathsf{D}((\mathsf{D} g_\eta)')$ (see (24)), the chain and product rules give

$$\begin{aligned}
\mathsf{H} g_\eta(\tilde{\phi}) &= \mathsf{D}\left[\left(\mathsf{D} E_\eta^{-1}(\tilde{\phi})\right)' \cdot y\left(\tau_\eta^{-1}(\tilde{\phi})\right)\right] \\
&= \left(y(\tau_\eta^{-1}(\tilde{\phi}))' \otimes I_{n-1}\right) \cdot \mathsf{D}\left(\mathsf{D} E_\eta^{-1}(\tilde{\phi})'\right) + \left(I_1 \otimes (\mathsf{D} E_\eta^{-1}(\tilde{\phi}))'\right) \cdot \mathsf{D}\left(y\left(\tau_\eta^{-1}(\tilde{\phi})\right)\right) \\
&= \left(y(\tilde{p})' \otimes I_{n-1}\right) \cdot \mathsf{H} E_\eta^{-1}(\tilde{\phi}) + \left(\mathsf{D} E_\eta^{-1}(\tilde{\phi})\right)' \cdot \mathsf{D} y(\tilde{p}) \cdot \mathsf{D} \tau_\eta^{-1}(\tilde{\phi}) \\
&= -\frac{\eta}{p_n(\tilde{p})} A_\eta(\tilde{\phi}) \cdot \operatorname{diag}(\tilde{p}) \cdot A_\eta(\tilde{\phi}) + A_\eta(\tilde{\phi})' \cdot \mathsf{D} y(\tilde{p}) \cdot \mathsf{D} \tau_\eta^{-1}(\tilde{\phi}).
\end{aligned} \tag{18}$$

The first summand in (18) is due to (12) and the fact that

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

The last equality holds because $A_\eta(\tilde{\phi})' \cdot A_\eta(\tilde{\phi}) = \eta^{-2} \operatorname{diag}(\tilde{\phi}_i^{-2})_{i=1}^{n-1}$ by Lemma 5, 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 (18) reduces by $\mathsf{D}y(\tilde{p}) = -\frac{1}{p_n(\tilde{p})} Y(\tilde{p})$ from Lemma 6 and $\tau_\eta = E_\eta \circ \tilde{\ell}$:

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

Substituting these into (18) gives

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

which can be factored into the required result. ■

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

Lemma 9 *Suppose a loss ℓ satisfies Condition A. Then for $\eta > 0$ the matrix-valued function Γ_η satisfies the following: for all $\tilde{\phi} \in \operatorname{int}(\tilde{\Phi}_\eta)$ and $\tilde{p} = \tau_\eta^{-1}(\tilde{\phi})$,*

$$\Gamma_\eta(\tilde{\phi}) = \frac{1}{p_n} A_\eta(\tilde{\phi})' \cdot Y(\tilde{p}) \cdot \left[\left[\mathsf{H}\underline{\tilde{L}}(\tilde{p})\right]^{-1} - \eta \left[\mathsf{H}\underline{\tilde{L}}_{\log}(\tilde{p})\right]^{-1}\right] \cdot Y(\tilde{p})' \cdot A_\eta(\tilde{\phi}), \tag{19}$$

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

Proof Substituting the values of $\mathsf{D}g_\eta$ and $\mathsf{H}g_\eta$ from Lemma 8 into the definition of Γ_η from Lemma 7 and then using Lemma 5 and the definition of $y(\tilde{p})$, we obtain

$$\Gamma_\eta(\tilde{\phi}) = \eta A_\eta(\tilde{\phi})' \cdot y(\tilde{p}) \cdot y(\tilde{p})' \cdot A_\eta(\tilde{\phi})$$
$$+ \frac{1}{p_n(\tilde{p})} A_\eta(\tilde{\phi})' \cdot \left[\eta \operatorname{diag}(\tilde{p}) + Y(\tilde{p}) \cdot \left[\mathsf{H}\underline{\tilde{L}}(\tilde{p})\right]^{-1} \cdot Y(\tilde{p})'\right] \cdot A_\eta(\tilde{\phi})$$
$$= \frac{1}{p_n} A_\eta(\tilde{\phi})' \cdot \left[\eta \frac{1}{p_n} \tilde{p} \cdot \tilde{p}' + \eta \operatorname{diag}(\tilde{p}) + Y(\tilde{p}) \cdot \left[\mathsf{H}\underline{\tilde{L}}(\tilde{p})\right]^{-1} \cdot Y(\tilde{p})'\right] \cdot A_\eta(\tilde{\phi}). \tag{20}$$

Using Lemma 6 we then see that

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

Substituting this for the appropriate terms in (20) gives

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

which equals (19).

Since $\Gamma_\eta=[p_n]^{-1}BRB'$ where $B=A_\eta(\tilde{\phi})'Y(\tilde{p})$ and $R=R(\eta,\ell,\tilde{p})$ the definition of negative semi-definiteness and the positivity of p_n means we need to show that $\forall x: x'\Gamma_\eta x\leq 0 \iff \forall y: y'Ry\leq 0$. It suffices to show that B is invertible, since we can let $y=Bx$ to establish the equivalence. The matrix $A_\eta(\tilde{\phi})$ is invertible since, by definition, $A_\eta(\tilde{\phi})=\mathsf{D}E_\eta^{-1}(\tilde{\phi})=-\eta^{-1}[\operatorname{diag}(\tilde{\phi})]^{-1}$ by Lemma 5 and so has matrix inverse $-\eta\operatorname{diag}(\tilde{\phi})$. The matrix $Y(\tilde{p})$ is invertible by Lemma 8. 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_η (via its Hessian)—and hence the convexity of the η-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. (8)), this means we are now able to specify the mixability constant η_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.

Theorem 10 *Suppose a loss ℓ satisfies Condition A. Let $\underline{\tilde{L}}(\tilde{p})$ be the Bayes risk for ℓ and $\underline{\tilde{L}}_{\log}(\tilde{p})$ be the Bayes risk for the log loss. Then the following statements are equivalent:*

(i.) *ℓ is η-mixable;*

(ii.) $\eta\mathsf{H}\underline{\tilde{L}}(\tilde{p})\succcurlyeq\mathsf{H}\underline{\tilde{L}}_{\log}(\tilde{p})$ *for all* $\tilde{p}\in\operatorname{int}(\tilde{\Delta}^n)$*;*

(iii.) $\eta\underline{L}(p)-\underline{L}_{\log}(p)$ *is convex on* $\operatorname{relint}(\Delta^n)$*;*

(iv.) $\eta\underline{\tilde{L}}(\tilde{p})-\underline{\tilde{L}}_{\log}(\tilde{p})$ *is convex on* $\operatorname{int}(\tilde{\Delta}^n)$.

Note that the largest η that satisfies any one of (i)–(iv) is the mixability constant for the loss. For example,

$$\eta_\ell=\max\{\eta\geq 0\colon\forall\tilde{p}\in\operatorname{int}(\tilde{\Delta}^n)\ ,\ \eta\mathsf{H}\underline{\tilde{L}}(\tilde{p})\succcurlyeq\mathsf{H}\underline{\tilde{L}}_{\log}(\tilde{p})\}.$$

Proof The case $\eta=0$ is trivial, so suppose $\eta>0$. Then by Lemmas 7 and 9 we know $\mathsf{H}f_\eta(\tilde{p})\preccurlyeq 0\iff R(\eta,\ell,\tilde{p})\preccurlyeq 0$. By Lemma 6, $\mathsf{H}\underline{\tilde{L}}(\tilde{p})\prec 0$ 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 $R(\eta,\ell,\tilde{p})\preccurlyeq 0\iff\mathsf{H}\underline{\tilde{L}}(\tilde{p})^{-1}\preccurlyeq\eta\mathsf{H}\underline{\tilde{L}}_{\log}(\tilde{p})^{-1}\iff\eta^{-1}\mathsf{H}\underline{\tilde{L}}_{\log}(\tilde{p})\preccurlyeq$

$\mathsf{H}\tilde{\underline{L}}(\tilde{p}) \iff \eta\mathsf{H}\tilde{\underline{L}}(\tilde{p}) \succcurlyeq \mathsf{H}\tilde{\underline{L}}_{\log}(\tilde{p})$. Therefore f_η is concave at $\tilde{p}$ if and only if $\eta\mathsf{H}\tilde{\underline{L}}(\tilde{p}) \succcurlyeq \mathsf{H}\tilde{\underline{L}}_{\log}(\tilde{p})$. Since concavity of f_η was equivalent to η-mixability, this establishes equivalence of (i) and (ii).

Since $\eta\mathsf{H}\tilde{\underline{L}}(\tilde{p}) \succcurlyeq \mathsf{H}\tilde{\underline{L}}_{\log}(\tilde{p}) \iff \mathsf{H}\big(\eta\tilde{\underline{L}}(\tilde{p}) - \tilde{\underline{L}}_{\log}(\tilde{p})\big) \succcurlyeq 0$, equivalence of (ii) and (iv) follows from the fact that positive semi-definiteness of the Hessian of a function on an open set is equivalent to convexity of the function (Hiriart-Urruty and Lemaréchal, 1993). Finally, equivalence of (iv) and (iii) follows by linearity of the map $p_n(\tilde{p}) = 1 - \sum_{i=1}^{n-1} \tilde{p}_i$. ■

The lemma allows one to derive η-mixability of an average of two η-mixable proper losses that satisfy its conditions:

Corollary 11 *Suppose ℓ_A and ℓ_B are two η-mixable losses that satisfy Condition A. Then, for any $\lambda \in (0,1)$, the loss $\ell = (1-\lambda)\ell_A + \lambda\ell_B$ is also η-mixable.*

Proof Clearly ℓ is continuous and continuously differentiable. And because properness of ℓ_A and ℓ_B implies that $\underline{L}_\ell(p) = (1-\lambda)\underline{L}_{\ell_A}(p) + \lambda\underline{L}_{\ell_B}(p)$, it is also strictly proper. Thus Theorem 10 applies to ℓ, and we just need to verify that $\eta\underline{L}_\ell(p) - \underline{L}_{\log}(p)$ is convex. Noting that

$$\eta\underline{L}_\ell(p) - \underline{L}_{\log}(p) = (1-\lambda)\Big(\eta\underline{L}_{\ell_A}(p) - \underline{L}_{\log}(p)\Big) + \lambda\Big(\eta\underline{L}_{\ell_B}(p) - \underline{L}_{\log}(p)\Big)$$

is a convex combination of two convex functions, the result follows. ■

One may wonder which loss is the most mixable. In the following we derive a straight-forward result that shows the (perhaps unsurprising) answer is log loss. Let $e_i \in \Delta^n$ denote the point-mass on the i-th outcome. Then we call a proper loss *fair* if $L(e_i, e_i) = \underline{L}(e_i) = 0$ for all i (Reid and Williamson, 2011). That is, if one is certain that outcome i will occur and this is correct, then it is only fair if one incurs no loss. Any loss can be made fair by subtracting the unique affine function that interpolates $\{\underline{L}(e_i) \colon i \in [n]\}$ from its Bayes risk. This does not change the curvature of $\underline{L}$ and thus by Theorem 10 it has the same mixability constant (provided the conditions of the theorem are satisfied). We will call a proper loss *normalised* if it is fair and $\max_{p\in\Delta^n} \underline{L}(p) = 1$. If a fair proper loss is not normalised, one may normalise it by dividing the loss on all outcomes by $\max_{p\in\Delta^n} \underline{L}(p)$. This scales up the mixability constant by $\max_{p\in\Delta^n} \underline{L}(p)$. For example, log loss is fair, but in order to normalise it, one needs to divide by $\max_{p\in\Delta^n} \underline{L}_{\log}(p) = \log(n)$, and the mixability constant η_ℓ for the resulting loss is $\log(n)$.

Corollary 12 *Suppose a loss ℓ satisfies Condition A. Then, if ℓ is normalised and $\underline{L}(p)$ is continuous, it can only be η-mixable for $\eta \le \log(n)$. This bound is achieved if ℓ is the normalised log loss.*

Proof Since $\underline{L}(p)$ is continuous and has a compact domain, there exists a $p^* = \arg\max_{p\in\Delta^n} \underline{L}(p)$ that achieves its maximum, which is 1 by assumption. Now by Theorem 10, η-mixability implies convexity of $\eta\underline{L}(p) - \underline{L}_{\log}(p)$ on $\operatorname{int}(\Delta^n)$, which extends to convexity on Δ^n by continuity of $\underline{L}(p)$ and $\underline{L}_{\log}(p)$, and hence

$$\begin{aligned} 0 &= \mathbb{E}_{i\sim p^*}\Big[\eta\underline{L}(e_i) - \underline{L}_{\log}(e_i)\Big] \ge \eta\underline{L}(p^*) - \underline{L}_{\log}(p^*) = \eta - \underline{L}_{\log}(p^*) \\ \Rightarrow \eta &\le \underline{L}_{\log}(p^*) \le \underline{L}_{\log}\big(\tfrac{1}{n}, \ldots, \tfrac{1}{n}\big) = \log(n), \end{aligned}$$

where the first equality follows from fairness of ℓ and log loss, and the first inequality follows from Jensen's inequality. ■

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, $\lambda_{\min}(A) := \lambda_1(A) \le \lambda_2(A) \le \cdots \le \lambda_n =: \lambda_{\max}(A)$ where each $\lambda_i(A)$ satisfies $|A-\lambda_i(A)I| = 0$.

Theorem 13 *Suppose a loss ℓ satisfies Condition A. Then its mixability constant is*

$$\boxed{\eta_\ell = \inf_{\tilde{p}\in\operatorname{int}(\tilde{\Delta}^n)} \lambda_{\max}\left((\mathsf{H}\underline{\tilde{L}}(\tilde{p}))^{-1}\cdot \mathsf{H}\underline{\tilde{L}}_{\log}(\tilde{p})\right).} \tag{21}$$

Equation 21 reduces to (8) when $n=2$ since the maximum eigenvalue of a 1×1 matrix is simply its single entry. Since the maximum eigenvalue of the Hessian of a function can be thought of as the "curvature", the above result justifies the title of the paper.

Proof For $\tilde{p}\in\operatorname{int}(\tilde{\Delta}^n)$, we define $C_\eta(\tilde{p}) := \eta\mathsf{H}\underline{\tilde{L}}(\tilde{p}) - \mathsf{H}\underline{\tilde{L}}_{\log}(\tilde{p})$ and $\rho(\tilde{p}) := \mathsf{H}\underline{\tilde{L}}(\tilde{p})^{-1}\cdot\mathsf{H}\underline{\tilde{L}}_{\log}(\tilde{p})$ and first show that zero is an eigenvalue of $C_\eta(\tilde{p})$ if and only if η is an eigenvalue of $\rho(\tilde{p})$. This can be seen since $\mathsf{H}\underline{\tilde{L}}(\tilde{p})$ is invertible (Lemma 6) so

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

Since a symmetric matrix is positive semidefinite if and only if all its eigenvalues are non-negative it must be the case that if $\lambda_{\min}(C_\eta(\tilde{p})) \ge 0$ then $C_\eta(\tilde{p}) \succcurlyeq 0$ since every other eigenvalue is bigger than the minimum one. Conversely, if $C_\eta(\tilde{p}) \not\succcurlyeq 0$ then at least one eigenvalue must be negative, thus the smallest eigenvalue must be negative. Thus, $\lambda_{\min}(C_\eta(\tilde{p})) \ge 0 \iff C_\eta(\tilde{p}) \succcurlyeq 0$. Now define $\eta(\tilde{p}) := \sup\{\eta > 0 : C_\eta(\tilde{p}) \succcurlyeq 0\} = \sup\{\eta > 0 : \lambda_{\min}(C_\eta(\tilde{p})) \ge 0\}$. We show that for each $\tilde{p}$ the function $\eta \mapsto \lambda_{\min}(C_\eta(\tilde{p}))$ is continuous and only has a single root. First, continuity follows because the entries of $C_\eta(\tilde{p})$ are continuous in η 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 $\{C_\eta(\tilde{p}) : \eta > 0\}$ is also concave in its entries and so in η. Since $C_0(\tilde{p}) = -\mathsf{H}\underline{\tilde{L}}_{\log}(\tilde{p})$ is positive definite for every $\tilde{p}$ (Lemma 6) we have $\lambda_{\min}(C_0(\tilde{p})) > 0$ and so, by the concavity of the map $\eta \mapsto \lambda_{\min}(C_\eta(\tilde{p}))$, there can be only one $\eta > 0$ for which $\lambda_{\min}(C_\eta(\tilde{p})) = 0$ and by continuity it must be largest non-negative one, that is, $\eta(\tilde{p})$.

Thus

$$\eta(\tilde{p}) = \sup\{\eta > 0 : \lambda_{\min}(C_\eta(\tilde{p})) = 0\} = \sup\{\eta > 0 : \eta \text{ is an eigenvalue of } \rho(\tilde{p})\} = \lambda_{\max}(\rho(\tilde{p})).$$

Now let $\eta^* := \inf_{\tilde{p}\in\operatorname{int}(\tilde{\Delta}^n)} \eta(\tilde{p}) = \inf_{\tilde{p}\in\operatorname{int}(\tilde{\Delta}^n)} \lambda_{\max}(\rho(\tilde{p}))$. We now claim that $C_{\eta^*}(\tilde{p}) \succcurlyeq 0$ for all $\tilde{p}$ since if there was some $\tilde{q}\in\tilde{\Delta}^n$ such that $C_{\eta^*}(\tilde{q}) \not\succcurlyeq 0$ we would have $\eta(\tilde{q}) < \eta^*$ since $\eta \mapsto \lambda_{\min}(C_\eta(\tilde{q}))$ only has a single root—a contradiction. Thus, since we have shown η^* is the largest η such that $C_{\eta^*}(\tilde{p}) \succcurlyeq 0$ it must be η_ℓ, by Theorem 10, as required. ■

The following Corollary gives an expression for η_ℓ that is simpler than (21), generalising (9) from the binary case.

Corollary 14 *Suppose ℓ satisfies Condition A. Then its mixability constant satisfies*

$$\boxed{\frac{-1}{\eta_\ell} = \inf_{\tilde{p}\in\operatorname{int}(\tilde{\Delta}^n)} \lambda_{\max}\left(\operatorname{diag}(\tilde{p})\cdot \mathsf{D}\tilde{\ell}(\tilde{p})\right).} \tag{22}$$

Proof Theorem 13 combined with Lemma 6 allows us to write

$$\begin{aligned}
\eta_\ell &= \inf_{\tilde{p}\in\operatorname{int}\tilde{\Delta}^n} \lambda_{\max}\left(\left(Y(\tilde{p})'\cdot \mathsf{D}\tilde{\ell}(\tilde{p})\right)^{-1}\cdot\left(Y(\tilde{p})'\cdot \mathsf{D}\tilde{\ell}_{\log}(\tilde{p})\right)\right)\\
&= \inf_{\tilde{p}\in\operatorname{int}\tilde{\Delta}^n} \lambda_{\max}\left((\mathsf{D}\tilde{\ell}(\tilde{p}))^{-1}\cdot \mathsf{D}\tilde{\ell}_{\log}(\tilde{p})\right)\\
&= \inf_{\tilde{p}\in\operatorname{int}\tilde{\Delta}^n} \lambda_{\max}\left((\mathsf{D}\tilde{\ell}(\tilde{p}))^{-1}\cdot \operatorname{diag}(-1/p_i)_{i=1}^{n-1}\right).\\
&= -\sup_{\tilde{p}\in\operatorname{int}\tilde{\Delta}^n} \lambda_{\min}\left((\mathsf{D}\tilde{\ell}(\tilde{p}))^{-1}\cdot \operatorname{diag}(1/p_i)_{i=1}^{n-1}\right)
\end{aligned}$$

and thus (22) follows since $\lambda_{\max}(A) = 1/\lambda_{\min}(A^{-1})$. ■

5. Mixability of the Brier Score

We will now apply the results from the previous section to show that the multiclass Brier score is mixable with mixability constant 1, as first proved by Vovk and Zhdanov (2009). The n-class Brier score is[1]

$$\ell_{\text{Brier}}(y,\hat{p}) = \sum_{i=1}^{n}([\![y_i = 1]\!] - \hat{p}_i)^2,$$

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

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

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

Theorem 15 *The Brier score is mixable, with mixability constant $\eta_{\text{Brier}} = 1$.*

Proof It can be verified by basic calculus that ℓ_{Brier} is continuous and continuously differentiable on $\operatorname{int}(\tilde{\Delta}^n)$. 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

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

1. 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 $\underline{\tilde{L}}_{\text{Brier}}$ as defined above leads to $\underline{\tilde{L}}''_{\text{Brier}}(\tilde{p}) = \mathsf{H}\underline{\tilde{L}}_{\text{Brier}}(\tilde{p}) = -2(1+1) = -4$.

and the latter inequality is true because $p_i \neq \hat{p}_i$ for at least one i by assumption. Hence the conditions of Theorem 10 are satisfied.

We will first prove that $\eta_{\text{Brier}} \leq 1$ by showing that convexity of $\eta\underline{\tilde{L}}_{\text{Brier}}(\tilde{p}) - \underline{\tilde{L}}_{\log}(\tilde{p})$ on $\operatorname{int}(\tilde{\Delta}^n)$ implies $\eta \leq 1$. If $\eta\underline{\tilde{L}}_{\text{Brier}}(\tilde{p}) - \underline{\tilde{L}}_{\log}(\tilde{p})$ 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:

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

By letting p_1 and p_n both tend to $1/2$, it follows that $\eta \leq 1$.

It remains to show that $\eta_{\text{Brier}} \geq 1$. By Theorem 10 it is sufficient to show that, for $\eta \leq 1$, $\eta\underline{L}_{\text{Brier}}(p) - \underline{L}_{\log}(p)$ is convex on $\operatorname{relint}(\Delta^n)$. We proceed by induction. For $n = 1$, the required convexity holds trivially. Suppose the lemma holds for $n-1$, and let $f_n(p_1,\ldots,p_n) = \eta\underline{L}_{\text{Brier}}(p) - \underline{L}_{\log}(p)$ 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 $g(p_1,p_2) = -\eta p_1^2 - \eta p_2^2 + \eta(p_1+p_2)^2 + p_1\ln p_1 + p_2\ln p_2 - (p_1+p_2)\ln(p_1+p_2)$. Since 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

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

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: $\operatorname{Tr}(\mathsf{H}g(p_1,p_2)) = 1/p_1 + 1/p_2 - 2/q \geq 0$ and $|\mathsf{H}g(p_1,p_2)| = (1/p_1 - 1/q)(1/p_2 - 1/q) - (2\eta - 1/q)^2$, which is non-negative if

$$\begin{aligned} \frac{1}{p_1p_2} - \frac{1}{p_1q} - \frac{1}{p_2q} &\geq 4\eta^2 - \frac{4\eta}{q} \\ 0 &\geq 4\eta^2 q - 4\eta \\ \eta q &\leq 1. \end{aligned}$$

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

6. Extension to Improper Losses

Our results are stated for proper losses. However, they also extend to a large class of *improper* (i.e., not proper) loss functions $\ell_{\text{imp}}\colon \mathcal{V} \to [0,\infty]$, which may be related to a proper loss ℓ with the same mixability constant using the following construction.

For any distribution $p \in \Delta^n$ and action $v \in \mathcal{V}$, let $L_{\text{imp}}(p,v) = p'\ell_{\text{imp}}(v)$ denote the risk and let $\underline{L}_{\text{imp}}(p) = \inf_{v\in\mathcal{V}} L_{\text{imp}}(p,v)$ denote the Bayes risk for ℓ_{imp}. If the infimum in the definition of the Bayes risk is achieved for all p, there exists a (possibly non-unique) *reference link* $\psi_{\text{imp}}\colon \Delta^n \to \mathcal{V}$ (Reid and Williamson, 2010), which is a 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 outcomes distributed by p. The loss function defined by

$$\ell(q) := \ell_{\text{imp}}(\psi_{\text{imp}}(q)) \qquad (q \in \Delta^n)$$

is proper by definition of the reference link.

If for every action $v \in \mathcal{V}$ there exists a distribution $p \in \Delta^n$ such that $\psi_{\text{imp}}(p) = v$ (i.e., the reference link is surjective), then ℓ is just a reparametrization of ℓ_{imp} and their superprediction sets S_ℓ and $S_{\ell_{\text{imp}}}$, as defined in (4), are the same. It then follows that $E_\eta(S_\ell) = E_\eta(S_{\ell_{\text{imp}}})$ for all η, such that ℓ and ℓ_{imp} must have the same mixability constants.

It turns out that the superprediction sets of ℓ and ℓ_{imp} are often the same even if ψ_{imp} is not surjective. This follows from Theorem 20 of Chernov et al. (2010) and its proof,[2] which may be reformulated as follows.

Theorem 16 (Chernov et al., 2010) *Let $\Lambda_{\text{imp}} = \ell_{\text{imp}}(\mathcal{V})$ be the set of achievable loss vectors. Suppose ℓ_{imp} is mixable and satisfies the following conditions:*

(i.) Λ_{imp} is a compact subset of $[0,\infty]^n$ (in the extended topology);

(ii.) There exists an action $v \in \mathcal{V}$ such that all components of $\ell_{\text{imp}}(v)$ are finite;

(iii.) For every distribution $p \in \Delta^n$ such that $p_i = p_j = 0$ for some $i \neq j$, the minimum of $L_{\text{imp}}(p,\cdot)$ is unique.

Then a unique reference link ψ_{imp} exists and $S_\ell = S_{\ell_{\text{imp}}}$, so ℓ and ℓ_{imp} have the same mixability constants. Moreover, ℓ is continuous and strictly proper.

Remark 17 *To see the equivalence between our version and Theorem 20 of Chernov et al. (2010), note that mixability of ℓ_{imp} implies that $\Sigma_\Lambda^\eta = \Sigma_\Lambda$ in their notation, for any $\eta > 0$ such that ℓ_{imp} is η-mixable.*

It seems likely that the mixability constants for ℓ_{imp} and ℓ will be the same even under weaker conditions than those of Theorem 16. In particular, we suspect that mixability of ℓ_{imp} is not always necessary, and Chernov and Vovk (2010) suggest that Condition iii may be removed. See also the discussion on mixability of composite losses by Vernet et al. (2012).

In the absence of such strengthenings of Theorem 16, it may be useful to recall that exp-concavity of ℓ_{imp} implies mixability (Cesa-Bianchi and Lugosi, 2006). An easy test to determine the mixability constant for ℓ_{imp} in some cases where it is 0, is given by the following observation (Kalnishkan and Vyugin, 2008):

Lemma 18 *If $S_{\ell_{\text{imp}}}$ is not convex, then ℓ_{imp} is not mixable.*

Proof Suppose ℓ_{imp} is η-mixable for some $\eta > 0$. Then, for any $x, y \in S_{\ell_{\text{imp}}}$ and any $\lambda \in [0,1]$, the set $E_\eta(S_{\ell_{\text{imp}}})$ contains the point $z = (1-\lambda)E_\eta(x) + \lambda E_\eta(y)$. Consequently, $S_{\ell_{\text{imp}}}$ itself contains $z' = E_\eta^{-1}(z)$, and by construction each component of z' satisfies

$$z'_i = -\frac{1}{\eta}\ln\left((1-\lambda)e^{-\eta x_i} + \lambda e^{-\eta y_i}\right) \leq (1-\lambda)x_i + \lambda y_i \qquad (i = 1,\ldots,n)$$

2. We thank a COLT2011 referee for referring us to this result.

by convexity of the exponential function. It follows that the point $(1-\lambda)x+\lambda y$ dominates z' and hence is also contained in $S_{\ell_{\mathrm{imp}}}$. Thus $S_{\ell_{\mathrm{imp}}}$ is convex, and we have shown that mixability implies convexity of $S_{\ell_{\mathrm{imp}}}$, from which the result follows. ∎

7. Connection to α-Flatness and Strong Convexity

We now briefly relate our result to recent work by Abernethy et al. (2009). They formulate the learning 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 the submanifold $\tilde{\Delta}^n$ in order for derivatives to be well defined.

Abernethy et al. (2009) have developed their own bounds on cumulative loss in terms of the α-flatness (defined below) of $\underline{L}(p)$. 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 6 in the relationship between the Hessian of $\underline{\tilde{L}}$ 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 $\underline{\tilde{L}}$. Below we point out the relationship between α-flatness and the positive definiteness of $\mathsf{H}\underline{L}(p)$ (we stress that in our work we used $\mathsf{H}\underline{\tilde{L}}(\tilde{p})$). Whilst the two results are not precisely comparable, the comparison below seems to suggest that the condition of Abernethy et al. (2009) is stronger than necessary.

Suppose $\mathcal{X}$ is a Banach space with norm $\|\cdot\|$. Given a real number $\alpha>0$ and a function $\sigma:\mathbb{R}_+\to[0,\infty]$ such that $\sigma(0)=0$, a convex function $f\colon\mathcal{X}\to\mathbb{R}$ is said to be $(\alpha,\sigma,\|\cdot\|)$-*flat* (or $(\alpha,\sigma,\|\cdot\|)$-smooth)[3] if for all $x,x_0\in\mathcal{X}$,

$$f(x)-f(x_0)\leq \mathsf{D}f(x_0)\cdot(x-x_0)+\alpha\sigma(\|x-x_0\|).$$

A concave function g is flat if the convex function $-g$ is flat. When $\|\cdot\|=\|\cdot\|_2$, and $\sigma(x)=x^2$, it is known (Hiriart-Urruty and Lemaréchal, 1993) that for $\alpha>0$, f is $(\alpha,x\mapsto x^2,\|\cdot\|_2)$-flat if and only if $f-\alpha\|\cdot\|^2$ is concave. Thus f is α-flat if and only if $\mathsf{H}(f-\alpha\|\cdot\|^2)$ is negative semi-definite, which is equivalent to $\mathsf{H}f-2\alpha I\preccurlyeq 0 \iff \mathsf{H}f\preccurlyeq 2\alpha I$.

Abernethy et al. (2009) show that if $\underline{L}$ is $(\alpha,x\mapsto x^2,\|\cdot\|_1)$-flat, then the minimax regret for a prediction game with T rounds is bounded above by $4\alpha\log T$. It is thus of interest to relate their assumption on $\underline{L}$ to the mixability condition (which guarantees constant regret, in the prediction with experts setting).

In contrast to the above quoted result for $\|\cdot\|_2$, we only get a one-way implication for $\|\cdot\|_1$.

Lemma 19 *If $f-\alpha\|\cdot\|_1^2$ is concave on $\mathbb{R}_+^n$ then f is $(\alpha,x\mapsto x^2,\|\cdot\|_1)$-flat.*

Proof It is known (Hiriart-Urruty and Lemaréchal, 1993, page 183) that a function h is concave if and only if $h(x)\leq h(x_0)+\mathsf{D}h(x_0)\cdot(x-x_0)$ for all x,x_0. Hence $f-\alpha\|\cdot\|_1^2$ is concave on $\mathbb{R}_+^n$ if and

3. This definition is redundantly parametrised: $(\alpha,\sigma,\|\cdot\|)$-flatness is equivalent to $(1,\alpha\sigma,\|\cdot\|)$-flatness. We have defined the notion as above in order to relate to existing definitions and because in fact one sometimes fixes σ and then is interested in the effect of varying α. When $\sigma(x)=x^2$, Abernethy et al. (2009) and Kakade et al. (2010) call this α-flat with respect to $\|\cdot\|$. Azé and Penot (1995) and Zălinescu (1983) would say f is σ-flat with respect to an implicitly given norm if f is (in our definition) $(\alpha,\sigma,\|\cdot\|)$-flat for some $\alpha>0$ (which in their setup is effectively bundled into σ). These differences do not matter (unless one wishes to use results from the earlier literature, which we do not).

only if for all $x, x_0 \in \mathbb{R}^n_+$,

$$\begin{aligned} & f(x) - \alpha\|x\|_1^2 \;\leq\; f(x_0) - \alpha\|x_0\|_1^2 + \mathsf{D}(f(x_0) - \alpha\|x_0\|_1^2)\cdot(x - x_0) \\ \Leftrightarrow\;\; & f(x) - f(x_0) \;\leq\; \alpha\|x\|_1^2 - \alpha\|x_0\|_1^2 + \mathsf{D}f(x_0)\cdot(x-x_0) - \alpha\mathsf{D}(\|x_0\|_1^2)\cdot(x-x_0). \end{aligned} \tag{23}$$

Since $\mathsf{D}(\|x_0\|_1^2) = 2\|x_0\|_1\mathbb{1}$ and $2\|x_0\|_1(\mathbb{1}\cdot(x-x_0)) = 2\|x_0\|_1(\|x\|_1 - \|x_0\|_1) = 2\|x_0\|_1\|x\|_1 - 2\|x_0\|_1^2$,

$$\begin{aligned} (23) \Leftrightarrow\;\; & f(x) - f(x_0) \;\leq\; \alpha\left(\|x\|_1^2 + \|x_0\|^2 - 2\|x_0\|_1\|x\|_1\right) + \mathsf{D}f(x_0)\cdot(x-x_0) \\ \Leftrightarrow\;\; & f(x) - f(x_0) \;\leq\; Df(x_0)\cdot(x-x_0) + \alpha\left(\|x\|_1 - \|x_0\|_1\right)^2. \end{aligned}$$

By the reverse triangle inequality $\|x - x_0\|_1 \geq |\|x\|_1 - \|x_0\|_1| \geq \|x\|_1 - \|x_0\|_1$ and thus $\|x-x_0\|_1^2 \geq (\|x\|_1 - \|x_0\|_1)^2$, which gives

$$\Rightarrow\;\; f(x) - f(x_0) \;\leq\; \mathsf{D}f(x_0)\cdot(x-x_0) + \alpha\|x-x_0\|_1^2.$$

$\blacksquare$

Now $f - \alpha\|\cdot\|_1^2$ is concave if and only if $\mathsf{H}(f - \alpha\|\cdot\|_1^2) \preccurlyeq 0$. We have (again for $x \in \mathbb{R}^n_+$) $\mathsf{H}(f - \alpha\|\cdot\|_1^2) = \mathsf{H}f - \alpha\mathsf{H}(\|\cdot\|_1^2)$. Let $\phi(x) = \|x\|_1^2$. Then $\mathsf{D}\phi(x) = 2\|x\|_1\mathsf{D}(\|x\|_1) = 2\|x\|_1\mathbb{1}$. Hence $\mathsf{H}\phi(x) = \mathsf{D}\left(\mathsf{D}\phi(x)\right)') = \mathsf{D}(2\|x\|_1\mathbb{1}') = 2\mathbb{1}\cdot\mathbb{1}'$. Thus $(\alpha, x \mapsto x^2, \|\cdot\|_1)$-flatness of $\underline{L}$ is implied by negative semi-definiteness of the Hessian of $\underline{L}$ relative to $2\alpha\mathbb{1}\cdot\mathbb{1}'$, instead of $\underline{L}_{\log}$ (see Theorem 10, part ii). The comparison with log loss is not that surprising in light of the observations regarding mixability by Grünwald (2007, §17.9).

The above analysis is not entirely satisfactory for three reasons: 1) Lemma 19 does not characterise the flatness condition (it is only a sufficient condition); 2) we have glossed over the fact that in order to compute derivatives one needs to work in $\tilde{\Delta}^n$; and 3) the learning protocols for the two situations are not identical. These last two points can be potentially addressed in future work. However the first seems impossible since there can not exist a characterisation of $(\alpha, x \mapsto x^2, \|\cdot\|_1)$-flatness in terms of concavity of some function. To see this, consider the one dimensional case and suppose there was some function g such that f was flat if g was concave. Then we would require $\mathsf{D}g(x)\cdot(x-x_0) = \alpha\|x-x_0\|_1^2 \Rightarrow \mathsf{D}g(x)(x-x_0) = \alpha|x-x_0|^2 = \alpha(x-x_0)^2 \Rightarrow \mathsf{D}g(x) = \alpha(x-x_0)$ which is impossible because the left hand side $\mathsf{D}g(x)$ does not depend upon x_0. On the other hand, perhaps it is not worth further investigation since the result due to Abernethy et al. (2009) is only a *sufficient* condition for logarithmic regret.

8. Conclusion

Mixability characterizes fast rates in the prediction with expert advice setting in terms of the mixability constant. An explicit formula to determine the mixability constant was previously available only for binary-valued outcomes, and the formula did not have a clear interpretation.

For strictly proper losses, Theorem 13 simplifies this formula and generalises it to outcomes with any finite number of possible values. The new formula has a clear interpretation as the minimal curvature of the Bayes risk for the loss relative to log loss. This shows in a precise and intuitive way the effect of the choice of loss function on the worst-case regret of the learner, and the special role played by log loss in such settings. Closely related characterizations of mixability are given in Theorem 10 and Corollary 14.

Although our main results are stated only for proper losses, Section 6 shows that many losses that are not proper can be related to a proper loss with the same mixability constant, which implies that our results cover these improper losses as well.

Acknowledgments

We thank the JMLR referees for helpful comments which improved the presentation, and Elodie Vernet for useful technical discussions. This work was supported by the Australian Research Council and NICTA which is funded by the Australian Government through the ICT Centre of Excellence program. It was done while Tim van Erven was affiliated with the Centrum Wiskunde & Informatica, Amsterdam, the Netherlands. 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. An earlier and shorter version of this paper appeared in the proceedings of COLT2011, and the present version has benefited from comments from the COLT referees.

Appendix A. Matrix Calculus

We adopt the notation of 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 Magnus and Neudecker, 1999, Chapter 2): $(A \otimes B)(C \otimes D) = (AC \otimes BD)$ 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 with respect to the jth coordinate at c is denoted $\mathsf{D}_j f_i(c)$ and is often[4] 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):

4. 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.

Theorem 20 *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

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

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,

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

(Magnus and Neudecker, 1999, Chapter 10). 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 the book of Magnus and Neudecker (1999, pg. 110).

Theorem 21 *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 22 *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 21 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

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

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 $(A \otimes B)(C \otimes D) = (AC \otimes BD)$. ■

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.

Dominique Azé and Jean-Paul Penot. Uniformly convex and uniformly smooth convex functions. *Annales de la faculté des sciences de Toulouse*, 4(4):705–730, 1995.

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

Alexey Chernov and Vladimir Vovk. Prediction with advice of unknown number of experts. In *Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence (UAI 2010)*, 2010.

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.

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.

Sham M. Kakade, Shai Shalev-Shwartz, and Ambuj Tewari. Regularization techniques for learning with matrices. arXiv:0910.0610v2, October 2010.

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.

Elodie Vernet, Robert C. Williamson, and Mark D. Reid. Composite multiclass losses. *Journal of Machine Learning Research*, March 2012. To be submitted.

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. Competitive on-line statistics. *International Statistical Review*, 69(2):213–248, 2001.

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

Constantin Zălinescu. On uniformly convex functions. *Journal of Mathematical Analysis and Applications*, 95:344–374, 1983.