Constant Regret, Generalized Mixability, and Mirror Descent

Zakaria Mhammedi ZAK.MHAMMEDI@ANU.EDU.AU
Robert C. Williamson BOB.WILLIAMSON@ANU.EDU.AU
The Australian National University & Data61
Canberra, ACT, Australia

Abstract

We consider the setting of prediction with expert advice; a learner makes predictions by aggregating those of a group of experts. Under this setting, and with the right choice of loss function and "mixing" algorithm, it is possible for the learner to achieve constant regret regardless of the number of prediction rounds. For example, constant regret can be achieved with *mixable* losses using the *Aggregating Algorithm* (AA). The *Generalized Aggregating Algorithm* (GAA) is a name for a family of algorithms parameterized by convex functions on simplices (entropies), which reduce to the AA when using the *Shannon entropy*. For a given entropy Φ, losses for which constant regret is possible using the GAA are called Φ-mixable. Which losses are Φ-mixable was previously left as an open question. We fully characterize Φ-mixability, and answer other open questions posed by Reid et al. (2015). We also elaborate on the tight link between the GAA and the *mirror descent algorithm* which minimizes the weighted loss of experts.

1. Introduction

Two fundamental problems in learning are how to aggregate information and under what circumstances can one learn fast. In this paper, we consider the problems jointly, extending the understanding and characterization of exponential mixing due to Vovk (1998), who showed that not only does the "*aggregating algorithm*" learn quickly when the loss is suitably chosen, but that it is in fact a generalization of classical Bayesian updating, to which it reduces when the loss is log-loss. We consider a general class of aggregating schemes, going beyond Vovk's exponential mixing, and provide a complete characterization of the mixing behavior for general losses and general mixing schemes parameterized by an arbitrary entropy function.

In the *game of prediction with expert advice* a *learner* predicts the outcome of a random variable (outcome of the *environment*) by aggregating the predictions of a pool of experts. At the end of each prediction round t, the outcome of the environment is announced and the learner and experts suffer losses based on their predictions. We are interested in algorithms that the learner can use to "aggregate" the experts' predictions and minimize the *regret* at the end of the game. In this case, the regret is defined as the difference between the cumulative loss of the learner and that of the best expert (in hindsight) after T rounds; if $\ell_t(\boldsymbol{a}_\theta^t)$ [resp. $\ell_t(\boldsymbol{a}_*^t)$] is the loss suffered by expert $\theta \in [k]$ [resp. the learner] who makes prediction $\boldsymbol{a}_\theta^t$ [resp. $\boldsymbol{a}_*^t$] at round t, then after T rounds the regret can be expressed as

$$R_\ell(T) = \sum_{1 \leq t \leq T} \ell_t(\boldsymbol{a}_*^t) - \inf_{\theta \in [k]} \sum_{1 \leq t \leq T} \ell_t(\boldsymbol{a}_\theta^t).$$

The *Aggregating Algorithm* (AA) (Vovk, 1998) achieves constant regret — a precise notion of "fast learning" — for *mixable* losses (formally defined later); that is, $R_\ell(T)$ is bounded from above by a constant R_ℓ independently of the number of rounds T. In this case, it is said that the algorithm achieves a *fast rate* for the corresponding loss. Reid et al. (2015) introduced the *Generalized Aggregating Algorithm* (GAA), going beyond the AA. The GAA is parameterized by the choice of a convex function Φ on the simplex (entropy) and reduces to the AA when Φ is the Shannon entropy. The GAA can achieve fast rates for losses satisfying a certain mixability condition (Φ-mixability). In particular, when a loss ℓ is Φ-mixable, the GAA achieves a constant regret R_ℓ^Φ which depends jointly on the *generalized mixability constant* η_ℓ^Φ — the largest η such that ℓ is $(\frac{1}{\eta}\Phi)$-mixable — and the divergence $D_\Phi(\boldsymbol{e}_\theta, \boldsymbol{q})$, where $\boldsymbol{q} \in \Delta_k$ is a prior distribution over experts (indexed by $\theta \in [k]$) and $\boldsymbol{e}_\theta$ is the θ'th standard basis element (Reid et al., 2015). Characterizing when losses are Φ-mixable was left as an open problem.

At each prediction round, the GAA can be divided into two steps; a *substitution step* where the learner picks a prediction from a set specified by the Φ-mixability condition; and an *update step* where a new distribution $\boldsymbol{q}$ over experts is computed depending on their performance on the new outcome of the environment and the previous distribution. The set of predictions specified by the Φ-mixability condition is non-empty when the loss is Φ-mixable. Interestingly, the update step of the GAA is exactly the *Mirror Descent Algorithm* (MDA) (Steinhardt and Liang, 2014; Orabona et al., 2015) which minimizes the weighted loss of experts. In fact, both the MDA and the GAA use a divergence "measure" D_Φ, generated by a convex function Φ, as a regularizer when updating the distribution $\boldsymbol{q}$.

Contributions. In this paper, we answer the questions presented by Reid et al. (2015) around the notion of *generalized mixability* using entropic duality. For an entropy Φ and a loss ℓ, we derive a necessary and sufficient condition (Theorems 15 and 16) for ℓ to be Φ-mixable. In particular, if ℓ and Φ satisfy some regularity conditions, then ℓ is Φ-mixable if and only if $\eta_\ell \Phi - \mathrm{S}$ is convex on the simplex, where S is the Shannon entropy and η_ℓ the largest η such that ℓ is η-mixable (Vovk, 1998; van Erven et al., 2012).

We derive an explicit expression for η_ℓ^Φ (Corollary 18), and hence, for the regret bound of the GAA. This allows us to compare the regret bound R_ℓ^Φ of any entropy Φ (such that ℓ is Φ-mixable) with that of the Shannon entropy S. In this case, we show (Theorem 19) that $R_\ell^{\mathrm{S}} \leq R_\ell^\Phi$; that is, the Shannon entropy achieves the lowest worst-case regret when using the GAA. This fact is similar to Vovk's result regarding the fundamental nature of log-loss (Vovk, 2015). Nevertheless, by leveraging the connection between the GAA and MDA, we discuss possible modifications to the GAA to enable better regret bounds in practice.

In Section 2, we introduce notations used throughout the paper and give some background on key notions in convex analysis which will be crucial in our proofs. We also present loss functions considered in the paper and state some new results (Theorem 3 and 4) along with corrected proofs for old ones (Theorem 5). In Section 3, we introduce the notions of classical and generalized mixability and derive some useful properties and characterizations. In particular, we derive a necessary and sufficient condition for Φ-mixability (Theorems 15 and 16). We also elucidate the tight relationship between the GAA and the MDA first observed by Reid et al. (2015). We conclude this paper by a general discussion and direction for future work.

2. Preliminaries

In this section, we introduce the notation used throughout the paper and present some key concepts in convex analysis. We also define entropies and describe loss functions considered in this paper. Furthermore, we present some new results regarding the latter.

2.1. Notations

For $n \in \mathbb{N}$, we define $\tilde{n} = n - 1$. We denote $[n] \coloneqq \{1, \dots, n\}$ the set of integers between 1 and n. Let $\langle \cdot, \cdot \rangle$ denote the standard inner product in $\mathbb{R}^n$ and $\|\cdot\|$ the corresponding norm. Let I_n and $\mathbf{1}_n$ denote the $n \times n$ identity matrix and the vector of all ones in $\mathbb{R}^n$. Let $\boldsymbol{e}_1, \dots, \boldsymbol{e}_n$ denote the *standard basis* for $\mathbb{R}^n$. For a set $\mathcal{I} \subsetneq \mathbb{N}$ and $\boldsymbol{r}_1, \dots, \boldsymbol{r}_n \in \mathbb{R}^k$, we denote $[\boldsymbol{r}_i]_{i \in \mathcal{I}} \coloneqq [\boldsymbol{r}_{i_1}, \dots, \boldsymbol{r}_{i_k}] \in \mathbb{R}^{n \times k}$, where $\mathcal{I} = \{i_1, \dots, i_k\}$ and $i_1 \leq \dots \leq i_k$. We denote its transpose by $[\boldsymbol{r}_i]_{i \in \mathcal{I}}^\mathsf{T} \in \mathbb{R}^{k \times n}$. For two vectors $\boldsymbol{p}, \boldsymbol{q} \in \mathbb{R}^n$, we write $\boldsymbol{p} \leq \boldsymbol{q}$ [resp. $\boldsymbol{p} < \boldsymbol{q}$], if $\forall i \in [n], p_i \leq q_i$ [resp. $p_i < q_i$]. We also denote $\boldsymbol{p} \odot \boldsymbol{q} = [p_i q_i]_{1 \leq i \leq n}^\mathsf{T} \in \mathbb{R}^n$ the *Hadamard product* of $\boldsymbol{p}$ and $\boldsymbol{q}$. If $(\boldsymbol{c}_k)$ is a sequence of vectors in $\mathcal{C} \subseteq \mathbb{R}^n$, we simply write $(\boldsymbol{c}_k) \subset \mathcal{C}$. For a sequence $(\boldsymbol{v}_m) \subset \mathbb{R}^n$, we write $\boldsymbol{v}_m \overset{m \to \infty}{\to} \boldsymbol{v}$ or $\lim_{m \to \infty} \boldsymbol{v}_m = \boldsymbol{v}$, if $\forall i \in [n], \lim_{m \to \infty} [\boldsymbol{v}_m]_i = v_i$. For a square matrix $A \in \mathbb{R}^{n \times n}$, $\lambda_{\min}(A)$ [resp. $\lambda_{\max}(A)$] denote its minimum [resp. maximum] eigenvalue. For $k \geq 1$, $\boldsymbol{u} \in [0, +\infty[^k$ and $\boldsymbol{w} \in \mathbb{R}^k$, we define $\log \boldsymbol{u} \coloneqq [\log u_i]_{1 \leq i \leq k}^\mathsf{T} \in \mathbb{R}^k$ and $\exp \boldsymbol{w} \coloneqq [\exp w_i]_{1 \leq i \leq k}^\mathsf{T} \in \mathbb{R}^k$.

Let $\Delta_n \coloneqq \{\boldsymbol{p} \in [0, +\infty[^n \colon \langle \boldsymbol{p}, \mathbf{1}_n \rangle = 1\}$ be the *probability simplex* in $\mathbb{R}^n$. We also define $\tilde{\Delta}_n \coloneqq \{\tilde{\boldsymbol{p}} \in [0, +\infty[^{\tilde{n}} \colon \langle \tilde{\boldsymbol{p}}, \mathbf{1}_{\tilde{n}} \rangle \leq 1\}$. We will use the notations $\Delta_n^k \coloneqq (\Delta_n)^k$ and $\tilde{\Delta}_n^k \coloneqq (\tilde{\Delta}_n)^k$. For $\mathcal{I} \subseteq [n]$, the set $\Delta_\mathcal{I} = \{\boldsymbol{q} \in \Delta_n : q_i = 0, \forall i \in [n] \setminus \mathcal{I}\}$ is a $|\mathcal{I}|$*-face* of Δ_n. We denote $\Pi_\mathcal{I}^n : \mathbb{R}^n \to \mathbb{R}^{|\mathcal{I}|}$ the linear projection operator satisfying $\Pi_\mathcal{I}^n \boldsymbol{u} = [u_i]_{i \in \mathcal{I}}^\mathsf{T}$. It is easy to verify that $\Pi_\mathcal{I}^n [\Pi_\mathcal{I}^n]^\mathsf{T} = I_{|\mathcal{I}|}$ and that $\boldsymbol{q} \mapsto \Pi_\mathcal{I}^n \boldsymbol{q}$ is a bijection from $\Delta_\mathcal{I}$ to $\Delta_{|\mathcal{I}|}$. In the special case where $\mathcal{I} = [\tilde{n}]$, we write $\Pi_n \coloneqq \Pi_{[\tilde{n}]}^n$ and we define the affine operator $\amalg_n : \mathbb{R}^{\tilde{n}} \to \mathbb{R}^n$ by $\amalg_n(\boldsymbol{u}) \coloneqq [u_1, \dots, u_{\tilde{n}}, 1 - \langle \boldsymbol{u}, \mathbf{1}_{\tilde{n}} \rangle]^\mathsf{T} = J_n \boldsymbol{u} + \boldsymbol{e}_n$, where $J_n \coloneqq \begin{bmatrix} I_{\tilde{n}} \\ -\mathbf{1}_{\tilde{n}}^\mathsf{T} \end{bmatrix} \in \mathbb{R}^{n \times \tilde{n}}$.

For $\boldsymbol{u} \in \mathbb{R}^n$ and $c \in \mathbb{R}$, we denote $\mathcal{H}_{\boldsymbol{u},c} \coloneqq \{\boldsymbol{y} \in \mathbb{R}^n : \langle \boldsymbol{y}, \boldsymbol{u} \rangle \leq c\}$ and $\mathcal{B}(\boldsymbol{u}, c) \coloneqq \{\boldsymbol{v} \in \mathbb{R}^n : \|\boldsymbol{u} - \boldsymbol{v}\| \leq c\}$. $\mathcal{H}_{\boldsymbol{u},c}$ is a closed half space and $\mathcal{B}(\boldsymbol{u}, c)$ is the c*-ball* in $\mathbb{R}^n$ centered at $\boldsymbol{u}$. Let $\mathcal{C} \subseteq \mathbb{R}^n$ be a non-empty set. We denote $\operatorname{int} \mathcal{C}$, $\operatorname{ri} \mathcal{C}$, $\operatorname{bd} \mathcal{C}$, and $\operatorname{rbd} \mathcal{C}$ the *interior*, *relative interior*, *boundary*, and *relative boundary* of a set $\mathcal{C} \in \mathbb{R}^n$, respectively (Hiriart-Urruty and Lemaréchal, 2001). We denote the *indicator function* of $\mathcal{C}$ by $\iota_\mathcal{C}$, where for $\boldsymbol{u} \in \mathcal{C}$, $\iota_\mathcal{C}(\boldsymbol{u}) = 0$, otherwise $\iota_\mathcal{C}(\boldsymbol{u}) = +\infty$. The *support function* of $\mathcal{C}$ is defined by

$$\sigma_\mathcal{C}(\boldsymbol{u}) \coloneqq \sup_{\boldsymbol{s} \in \mathcal{C}} \langle \boldsymbol{u}, \boldsymbol{s} \rangle, \ \boldsymbol{u} \in \mathbb{R}^n.$$

Let $f \colon \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$. We denote $\operatorname{dom} f \coloneqq \{\boldsymbol{u} \in \mathbb{R}^n : f(\boldsymbol{u}) < +\infty\}$ the *effective domain* of f. The function f is *proper* if $\operatorname{dom} f \neq \emptyset$. The function f is *convex* if $\forall (\boldsymbol{u}, \boldsymbol{v}) \in \mathbb{R}^n$ and $\lambda \in]0, 1[$, $f(\lambda \boldsymbol{u} + (1 - \lambda) \boldsymbol{v}) \leq \lambda f(\boldsymbol{u}) + (1 - \lambda) f(\boldsymbol{v})$. When the latter inequality is strict for all $\boldsymbol{u} \neq \boldsymbol{v}$, f is *strictly convex*. When f is convex, it is *closed* if it is *lower semi-continuous*; that is, for all $\boldsymbol{u} \in \mathbb{R}^n$, $\liminf_{\boldsymbol{v} \to \boldsymbol{u}} f(\boldsymbol{v}) \geq f(\boldsymbol{u})$. The function f is said to be 1*-homogeneous* if $\forall (\boldsymbol{u}, \alpha) \in \mathbb{R}^n \times]0, +\infty[$, $f(\alpha \boldsymbol{u}) = \alpha f(\boldsymbol{u})$, and it is said to be 1*-coercive* if $\frac{f(\boldsymbol{u})}{\|\boldsymbol{u}\|} \to +\infty$ as $\|\boldsymbol{u}\| \to \infty$. Let f be proper. The *sub-differential* of f is defined by

$$\partial f(\boldsymbol{u}) \coloneqq \{\boldsymbol{s}^* \in \mathbb{R}^n : f(\boldsymbol{v}) \geq f(\boldsymbol{u}) + \langle \boldsymbol{s}^*, \boldsymbol{v} - \boldsymbol{u} \rangle, \forall \boldsymbol{v} \in \mathbb{R}^n\}.$$

Any element $\boldsymbol{s} \in \partial f(\boldsymbol{u})$ is called a *sub-gradient* of f at $\boldsymbol{u}$. We say that f is *directionally differentiable* if for all $(\boldsymbol{u}, \boldsymbol{v}) \in \operatorname{dom} f \times \mathbb{R}^n$ the limit $\lim_{t \downarrow 0} \frac{f(\boldsymbol{u}+t\boldsymbol{v})-f(\boldsymbol{u})}{t}$ exists in $[-\infty, +\infty]$. In this case, we denote the limit by $f'(\boldsymbol{u}; \boldsymbol{v})$. When f is convex, it is directionally differentiable (Rockafellar, 1997). Let f be proper and directionally differentiable. The *divergence* generated by f is the map $D_f \colon \mathbb{R}^n \times \operatorname{dom} f \to [0, +\infty]$ defined by

$$D_f(\boldsymbol{v}, \boldsymbol{u}) := \begin{cases} f(\boldsymbol{v}) - f(\boldsymbol{u}) - f'(\boldsymbol{u}; \boldsymbol{v} - \boldsymbol{u}), & \text{if } \boldsymbol{v} \in \operatorname{dom} f; \\ +\infty, & \text{otherwise.} \end{cases}$$

For $\mathcal{I} \subset [n]$ and $f^{\mathcal{I}} := f \circ [\Pi_{\mathcal{I}}^n]^{\mathsf{T}}$, it is easy to verify that $(f^{\mathcal{I}})'(\Pi_{\mathcal{I}}^n \boldsymbol{p}; \Pi_{\mathcal{I}}^n \boldsymbol{q} - \Pi_{\mathcal{I}}^n \boldsymbol{p}) = f'(\boldsymbol{p}; \boldsymbol{q} - \boldsymbol{p}), \forall (\boldsymbol{p}, \boldsymbol{q}) \in \Delta_{\mathcal{I}}$. In this case, it holds that $D_f(\boldsymbol{q}, \boldsymbol{p}) = D_{f^{\mathcal{I}}}(\Pi_{\mathcal{I}}^n \boldsymbol{q}, \Pi_{\mathcal{I}}^n \boldsymbol{p})$. If f is differentiable [resp. twice differentiable] at $\boldsymbol{u} \in \operatorname{int} \operatorname{dom} f$, we denote $\nabla f(\boldsymbol{u}) \in \mathbb{R}^n$ [resp. $\mathsf{H} f(\boldsymbol{u}) \in \mathbb{R}^{n \times n}$] its *gradient* vector [resp. *Hessian* matrix] at $\boldsymbol{u}$. A vector-valued function $g \colon \mathbb{R}^n \to \mathbb{R}^m$ is differentiable at $\boldsymbol{u}$ if for all $i \in [m]$, g_i is differentiable at $\boldsymbol{u}$. In this case, the *differential* of g at $\boldsymbol{u}$ is the linear operator $\mathsf{D} g(\boldsymbol{u}) : \mathbb{R}^n \to \mathbb{R}^m$ defined by $\mathsf{D} g(\boldsymbol{u}) := [\nabla g_i(\boldsymbol{u})]_{1 \le i \le m}^{\mathsf{T}}$. If f has k continuous derivatives on a set $\Omega \subset \mathbb{R}^k$, we write $f \in C^k(\Omega)$.

We define $\tilde{f} : \mathbb{R}^{\tilde{n}} \to \mathbb{R} \cup \{+\infty\}$ by $\tilde{f} := f \circ \amalg_n + \iota_{\tilde{\Delta}_n}$. That is,

$$\tilde{f}(\tilde{\boldsymbol{u}}) := \begin{cases} f(J_n \tilde{\boldsymbol{u}} + \boldsymbol{e}_n), & \text{for } \tilde{\boldsymbol{u}} \in \tilde{\Delta}_n; \\ +\infty, & \text{for } \tilde{\boldsymbol{u}} \in \mathbb{R}^{n-1} \setminus \tilde{\Delta}_n. \end{cases} \tag{1}$$

If $\tilde{f}$ is directionally differentiable, then $f'(\boldsymbol{p}, \boldsymbol{q} - \boldsymbol{p}) = \tilde{f}'(\tilde{\boldsymbol{p}}, \tilde{\boldsymbol{q}} - \tilde{\boldsymbol{p}})$, for $\boldsymbol{p}, \boldsymbol{q} \in \Delta_n$. If $\tilde{f}$ is differentiable at $\tilde{\boldsymbol{p}} = \Pi_n(\boldsymbol{p})$, then $\tilde{f}'(\tilde{\boldsymbol{p}}, \tilde{\boldsymbol{q}} - \tilde{\boldsymbol{p}}) = \langle \nabla \tilde{f}(\tilde{\boldsymbol{p}}), \tilde{\boldsymbol{q}} - \tilde{\boldsymbol{p}} \rangle$. If, additionally, f is differentiable at $\boldsymbol{p} \in \operatorname{ri} \Delta_k$, the chain rule yields $\nabla \tilde{f}(\tilde{\boldsymbol{p}}) = J_n^{\mathsf{T}} \nabla f(\boldsymbol{p})$. Since $J_n(\tilde{\boldsymbol{p}} - \tilde{\boldsymbol{q}}) = \amalg_n(\tilde{\boldsymbol{p}} - \tilde{\boldsymbol{q}}) = \boldsymbol{p} - \boldsymbol{q}$, it also follows that $\langle \tilde{\boldsymbol{p}} - \tilde{\boldsymbol{q}}, \nabla \tilde{f}(\tilde{\boldsymbol{p}}) \rangle = \langle \boldsymbol{p} - \boldsymbol{q}, \nabla f(\boldsymbol{p}) \rangle$.

The *Fenchel dual* of a (proper) function f is defined by $f^*(\boldsymbol{v}) := \sup_{\boldsymbol{u} \in \operatorname{dom} f} \langle \boldsymbol{u}, \boldsymbol{v} \rangle - f(\boldsymbol{u})$, and it is a closed, convex function on $\mathbb{R}^n$ (Hiriart-Urruty and Lemaréchal, 2001). Some useful properties of the Fenchel dual are given in Appendix A.

2.2. Entropies on the simplex

A function $\Phi \colon \mathbb{R}^k \to \mathbb{R} \cup \{+\infty\}$ is an *entropy* if it is closed, convex, and $\Delta_k \subseteq \operatorname{dom} \Phi$. Its *entropic dual* $\Phi^\star : \mathbb{R}^k \to \mathbb{R} \cup \{+\infty\}$ is defined by $\Phi^\star(\boldsymbol{z}) := \sup_{\boldsymbol{q} \in \Delta_k} \langle \boldsymbol{q}, \boldsymbol{z} \rangle - \Phi(\boldsymbol{q}), \boldsymbol{z} \in \mathbb{R}^k$. For the remainder of this paper, we consider entropies defined on $\mathbb{R}^k$, where $k \ge 2$.

Let $\Phi \colon \mathbb{R}^k \to \mathbb{R} \cup \{+\infty\}$ be an entropy and $\Phi_\Delta := \Phi + \iota_{\Delta_k}$. In this case, $\Phi^\star = \Phi_\Delta^*$. It is clear that Φ_Δ is 1-coercive, and therefore, $\operatorname{dom} \Phi^\star = \operatorname{dom} \Phi_\Delta^* = \mathbb{R}^k$ (Hiriart-Urruty and Lemaréchal, 2001, Prop. E.1.3.8). The entropic dual of Φ can also be expressed using the Fenchel dual of $\tilde{\Phi} : \mathbb{R}^{k-1} \to \mathbb{R} \cup \{+\infty\}$ defined by (1) after substituting f by Φ and n by k. In fact,

$$\begin{aligned} \Phi^\star(\boldsymbol{z}) &= \sup_{\tilde{\boldsymbol{q}} \in \tilde{\Delta}_k} \langle J_k \tilde{\boldsymbol{q}} + \boldsymbol{e}_k, \boldsymbol{z} \rangle - \Phi(J_k \tilde{\boldsymbol{q}} + \boldsymbol{e}_k), \\ &= \langle \boldsymbol{e}_k, \boldsymbol{z} \rangle + \sup_{\tilde{\boldsymbol{q}} \in \tilde{\Delta}_k} \left\langle \tilde{\boldsymbol{q}}, J_k^{\mathsf{T}} \boldsymbol{z} \right\rangle - \tilde{\Phi}(\tilde{\boldsymbol{q}}), \\ &= \langle \boldsymbol{e}_k, \boldsymbol{z} \rangle + \tilde{\Phi}^*(J_k^{\mathsf{T}} \boldsymbol{z}), \end{aligned} \tag{2}$$

where (2) follows from the fact that $\operatorname{dom} \tilde{\Phi} = \tilde{\Delta}_k$. Note that when Φ is an entropy, $\tilde{\Phi}$ is a closed convex function on $\mathbb{R}^{k-1}$. Hence, it holds that $\tilde{\Phi}^{**} = \tilde{\Phi}$ (Rockafellar, 1997).

The *Shannon entropy* by $\mathrm{S}(\boldsymbol{q}) := \sum_{i\in[k]:q_i\neq 0} q_i \log q_i$,[1] if $\boldsymbol{q} \in [0,+\infty[^k$; and $+\infty$ otherwise. In the next proposition, we give the expressions of the entropic dual of S as well as the Fenchel dual $\tilde{\mathrm{S}}$. The proof is in Appendix C.1.

Proposition 1 *For the Shannon entropy* S*, it holds that* $\tilde{\mathrm{S}}^*(\boldsymbol{v}) = \log(\langle \exp(\boldsymbol{v}), \mathbf{1}_{\tilde{k}}\rangle + 1), \forall \boldsymbol{v} \in \mathbb{R}^{k-1}$, *and* $\mathrm{S}^\star(\boldsymbol{z}) = \log\langle \exp(\boldsymbol{z}), \mathbf{1}_k\rangle, \forall \boldsymbol{z} \in \mathbb{R}^k$.

2.3. Loss Functions

In general, a loss function is a map $\ell\colon \mathcal{X} \times \mathcal{A} \to [0,+\infty]$ where $\mathcal{X}$ is an outcome set and $\mathcal{A}$ is a prediction set. In this paper, we consider the case of n possible outcomes; $\mathcal{X} = [n]$. Overloading notation slightly, we define the mapping $\ell\colon \mathcal{A} \to [0,+\infty]^n$ by $[\ell(\boldsymbol{a})]_x = \ell(x,\boldsymbol{a}), \forall x \in [n]$ and denote $\ell_x(\cdot) := [\ell(\cdot)]_x$. We further extend the new definition of ℓ to the set $\bigcup_{k\geq 1} \mathcal{A}^k$ such that for $x \in [n]$ and $A := [\boldsymbol{a}_\theta]^\mathsf{T}_{\theta\in[k]} \in \mathcal{A}^k$, we have $\ell_x(A) := [\ell_x(\boldsymbol{a}_\theta)]^\mathsf{T}_{\theta\in[k]} \in [0,+\infty]^k$. We define the *effective domain* of ℓ by $\operatorname{dom}\ell := \{\boldsymbol{a} \in \mathcal{A} : \ell(\boldsymbol{a}) \in [0,+\infty[^n\}$, and the *loss surface* by $\mathcal{S}_\ell := \{\ell(\boldsymbol{a}) : \boldsymbol{a} \in \operatorname{dom}\ell\}$. We say that ℓ is *closed* if $\mathcal{S}_\ell$ is closed. The *superprediction* set of ℓ is defined by $\mathcal{S}_\ell^\oplus := \mathcal{S}_\ell \oplus [0,+\infty[^n$, where $\oplus$ denotes the Minkowski sum. The η*-exponentiated* superprediction set of ℓ is defined by $\exp(-\eta\mathcal{S}_\ell^\oplus) := \{\exp(-\eta\boldsymbol{s}) : \boldsymbol{s} \in \mathcal{S}_\ell^\oplus\}$. Let $\boldsymbol{a}_0, \boldsymbol{a}_1 \in \mathcal{A}$. The prediction $\boldsymbol{a}_0$ is said to be *better* than $\boldsymbol{a}_1$ if $\ell(\boldsymbol{a}_0) \leq \ell(\boldsymbol{a}_1)$ and for some $x \in [n]$, $\ell_x(\boldsymbol{a}_0) < \ell_x(\boldsymbol{a}_1)$ (Williamson et al., 2016). A loss ℓ is *admissible* if for any $\boldsymbol{a} \in \mathcal{A}$ there are no better predictions. A loss ℓ is said to be *trivial* if there exists a prediction $\boldsymbol{a} \in \mathcal{A}$ which minimizes ℓ over all outcomes $x \in [n]$. Thus, a closed loss is *non-trivial* if there exists $(x_0, x_1, \boldsymbol{a}_0, \boldsymbol{a}_1) \in [n] \times [n] \times \mathcal{A} \times \mathcal{A}$ such that

$$\boldsymbol{a}_1 \in \operatorname{argmin}\{\ell_{x_0}(\boldsymbol{a}) : \ell_{x_1}(\boldsymbol{a}) = \inf_{\hat{\boldsymbol{a}}\in\mathcal{A}} \ell_{x_1}(\hat{\boldsymbol{a}})\} \text{ and } \inf_{\boldsymbol{a}\in\mathcal{A}} \ell_{x_0}(\boldsymbol{a}) = \ell_{x_0}(\boldsymbol{a_0}) < \ell_{x_0}(\boldsymbol{a}_1). \tag{3}$$

In the rest of this paper, we will often make use of the following condition on losses.

Condition I *ℓ is a closed, admissible, and non-trivial loss such that* $\operatorname{dom}\ell \neq \emptyset$.

In this paragraph let $\mathcal{A} = \Delta_n$. We define the *conditional risk* $L_\ell : \Delta_n \times \Delta_n \to \mathbb{R}$ by $L_\ell(\boldsymbol{p},\boldsymbol{q}) = \mathbb{E}_{x\sim\boldsymbol{p}}[\ell_x(\boldsymbol{q})] = \langle \boldsymbol{p}, \ell(\boldsymbol{q})\rangle$ and the *Bayes risk* by $\underline{L}_\ell(\boldsymbol{p}) := \inf_{\boldsymbol{q}\in\Delta_n} \underline{L}_\ell(\boldsymbol{p},\boldsymbol{q})$. In this case, the loss ℓ is *proper* [resp. *strictly proper*] if $\underline{L}_\ell(\boldsymbol{p}) = \langle \boldsymbol{p}, \ell(\boldsymbol{p})\rangle \leq \langle \boldsymbol{p}, \ell(\boldsymbol{q})\rangle$ [resp. $\langle \boldsymbol{p}, \ell(\boldsymbol{p})\rangle < \langle \boldsymbol{p}, \ell(\boldsymbol{q})\rangle$] for all $\boldsymbol{p} \neq \boldsymbol{q}$ in Δ_n. For example, the *log-loss* $\ell_{\log} : \Delta_n \to [0,+\infty]^n$ is defined by $\ell_{\log}(\boldsymbol{p}) = -\log \boldsymbol{p}$, where the "log" of a vector applies coordinate-wise. One can easily check that $\ell_{\log}$ is strictly proper. We denote $\underline{L}_{\log}$ the corresponding Bayes risk.

The above definition of the Bayes risk is restricted to losses defined on the simplex. For a loss $\ell\colon \mathcal{A} \to [0,+\infty]^n$, we use the following definition of the Bayes risk.

Definition 2 (Bayes Risk (Williamson, 2014)) *Let $\ell\colon \mathcal{A} \to [0,+\infty]^n$ be a loss such that $\mathcal{S}_\ell^\oplus \neq \emptyset$. The* Bayes risk *$\underline{L}_\ell : \mathbb{R}^n \to \mathbb{R} \cup \{-\infty\}$ is defined by*

$$\forall \boldsymbol{u} \in \mathbb{R}^n, \quad \underline{L}_\ell(\boldsymbol{u}) := \inf_{\boldsymbol{z}\in\mathcal{S}_\ell^\oplus} \langle \boldsymbol{u}, \boldsymbol{z}\rangle = -\sigma_{\mathcal{S}_\ell^\oplus}(-\boldsymbol{u}).$$

1. The Shannon entropy is usually defined with a minus sign. However, it will be more convenient for us to work without it.

Note that since $\mathcal{S}_\ell^\oplus$ is a non-empty subset of $[0,+\infty[^n$, $\boldsymbol{u} \mapsto \sigma_{\mathcal{S}_\ell^\oplus}(-\boldsymbol{u})$ is convex and finite on $[0,+\infty[^n$; it is bounded from above by 0. In particular, $\partial\sigma_{\mathcal{S}_\ell^\oplus}(-\boldsymbol{p}) \neq \emptyset$ for all $\boldsymbol{p} \in \operatorname{ri}\Delta_n \subset \operatorname{int}\operatorname{dom}\sigma_{\mathcal{S}_\ell^\oplus} =]0,+\infty[^n$ (Rockafellar, 1997, Thm. 23.4).

We call $\underline{\ell} : \Delta_n \to [0,+\infty]^n$ a *support loss* of ℓ if $\forall \boldsymbol{p} \in \operatorname{ri}\Delta_n$, $\underline{\ell}(\boldsymbol{p}) \in \partial\sigma_{\mathcal{S}_\ell^\oplus}(-\boldsymbol{p})$, and $\forall \boldsymbol{p} \in \operatorname{rbd}\Delta_n$, there exists a sequence $(\boldsymbol{p}_m) \subset \operatorname{ri}\Delta_n$, such that $\underline{\ell}(\boldsymbol{p}_m) \overset{m\to\infty}{\to} \underline{\ell}(\boldsymbol{p})$.

Theorem 3 *Any loss $\ell\colon \mathcal{A} \to [0,+\infty]^n$ such that $\operatorname{dom}\ell \neq \emptyset$, has a proper support loss $\underline{\ell}$ whose Bayes risk is $\underline{L}_\ell$.*

The proof of the proposition is in Appendix C.2. Note that when the Bayes risk is differentiable on $]0,+\infty[^n$, or equivalently, when $\boldsymbol{u} \mapsto \sigma_{\mathcal{S}_\ell^\oplus}(-\boldsymbol{u})$ is differentiable on $]0,+\infty[^n$, the support loss of ℓ is uniquely defined on $\operatorname{ri}\Delta_n$. This is because when $\sigma_{\mathcal{S}_\ell^\oplus}$ is differentiable at $-\boldsymbol{p}$, $\partial\sigma_{\mathcal{S}_\ell^\oplus}(-\boldsymbol{p}) = \{\nabla\sigma_{\mathcal{S}_\ell^\oplus}(-\boldsymbol{p})\} = \{\underline{\ell}(\boldsymbol{p})\}$ (Hiriart-Urruty and Lemaréchal, 2001, Cor. D.2.1.4).

Theorem 4 *Let $\ell\colon \mathcal{A} \to [0,+\infty]^n$ be a loss satisfying Condition I and $\underline{\ell}$ a proper support loss of ℓ. If $\underline{L}_\ell$ is not differentiable at $\boldsymbol{p}$ then there exist $\boldsymbol{a}_0, \boldsymbol{a}_1 \in \operatorname{dom}\ell$, such that $\ell(\boldsymbol{a}_0) \neq \ell(\boldsymbol{a}_1)$ and $\underline{L}_\ell(\boldsymbol{p}) = \langle \boldsymbol{p}, \underline{\ell}(\boldsymbol{p})\rangle = \langle \boldsymbol{p}, \ell(\boldsymbol{a}_0)\rangle = \langle \boldsymbol{p}, \ell(\boldsymbol{a}_1)\rangle$.*

If the Bayes risk $\underline{L}_\ell$ is differentiable on $]0,+\infty[^n$, then

$$\begin{aligned} &\forall \boldsymbol{p} \in \operatorname{dom}\underline{\ell}, \quad \exists \boldsymbol{a}_* \in \operatorname{dom}\ell, \qquad \ell(\boldsymbol{a}_*) = \underline{\ell}(\boldsymbol{p}), \\ &\forall \boldsymbol{a} \in \operatorname{dom}\ell, \quad \exists (\boldsymbol{p}_m) \subset \operatorname{ri}\Delta_n, \quad \underline{\ell}(\boldsymbol{p}_m) \overset{m\to\infty}{\to} \ell(\boldsymbol{a}). \end{aligned}$$

The proof is in Appendix C.3. Note that we always have $\operatorname{ri}\Delta_n \subset \operatorname{dom}\underline{\ell}$. Consistent with the notations introduced in the previous subsection, we write $\tilde{\underline{\ell}} \coloneqq \underline{\ell} \circ \Pi_n$ and $\tilde{\underline{L}}_\ell \coloneqq \underline{L}_\ell \circ \Pi_n$.

In the literature, many theoretical results involving loss functions relied on the fact that the superprediction set of a proper loss is convex (Williamson et al., 2016; Dawid, 2007). An earlier proof of this result by Williamson et al. (2016) was incomplete[2]. In the next theorem, we restate this result and prove it in Appendix C.4.

Theorem 5 *If $\ell\colon \Delta_n \to [0,+\infty[^n$ is a continuous proper loss, then $\mathcal{S}_\ell^\oplus = \bigcap_{\boldsymbol{p}\in\Delta_n} \mathcal{H}_{-\boldsymbol{p},-\underline{L}_\ell(\boldsymbol{p})}$. In particular, $\mathcal{S}_\ell^\oplus$ is convex.*

3. Mixability in the Game of Prediction with Expert Advice

Here, we consider the setting of prediction with expert advice (Vovk, 1998); there is pool of k experts, parameterized by $\theta \in [k]$, which make predictions $\boldsymbol{a}_\theta^t \in \mathcal{A}$ at each round t. In the same round, the learner predicts $\boldsymbol{a}_* \in F(A^t) \in \mathcal{A}$, where $A^t = [\boldsymbol{a}_\theta^t]_{1\le\theta\le k} \in \mathcal{A}^k$ is the matrix of experts' predictions and F is set valued *aggregating* function on $\mathcal{A}^k$. At the end of the round, the outcome $x^t \in [n]$ is announced and each expert θ [resp. learner] suffers a loss $\ell_{x^t}(\boldsymbol{a}_\theta)$ [resp. $\ell_{x^t}(\boldsymbol{a}_*)$], where $\ell\colon \mathcal{A} \to [0,+\infty]^n$. We will refer to such a game by $\mathcal{G}_\ell(k,T)$.

Vovk (1998) introduced the *Aggregating Algorithm* (AA) which specifies the aggregating function F of the learner. When a loss ℓ is η-mixable (see below) the AA achieves *constant regret* in

2. It was claimed that if $\mathcal{S}_\ell^\oplus$ is non-convex, there exists a point $\boldsymbol{s}_0$ on the loss surface $\mathcal{S}_\ell$ such that no hyperplane supports $\mathcal{S}_\ell^\oplus$ at $\boldsymbol{s}_0$. The non-convexity of a set by itself is not sufficient to make such a claim; the continuity of the loss ℓ is required.

the $\mathcal{G}_\ell(k,T)$ game; the difference between the cumulative loss of the learner and the best expert is upper bounded by a constant independent of the number of rounds T.

The *Generalized Aggregating Algorithm* (GAA) (Reid et al., 2015) uses an entropy function $\Phi\colon \mathbb{R}^k \to \mathbb{R}\cup\{+\infty\}$ to specify the learner's aggregating function F. The GAA also achieves a constant regret when a certain condition is satisfied (Φ-mixability). The GAA reduces to the AA when Φ is a scaled *Shannon entropy* S.

Definition 6 (η-mixability) *For $\eta > 0$, a loss $\ell\colon \mathcal{A} \to [0,+\infty]^n$ is said to be η-mixable, if $\forall \boldsymbol{q} \in \Delta_k$, $\forall A := [\boldsymbol{a}_\theta]_{\theta\in[k]} \in \mathcal{A}^k$, $\exists \boldsymbol{a}_* \in \mathcal{A}$, such that*

$$\forall x \in [n], \quad \ell_x(\boldsymbol{a}_*) \le -\eta^{-1}\log\langle \exp(-\eta\ell_x(A)), \boldsymbol{q}\rangle\,. \tag{4}$$

Chernov et al. (2010) showed that the η-mixability condition (4) is equivalent to the convexity of the η-exponentiated superprediction set $\exp(-\eta\mathcal{S}_\ell^{\oplus})$. The largest η such that ℓ is η-mixable is denoted η_ℓ. If $\eta_\ell > 0$, we say that ℓ is *classically mixable*.

For a strictly proper loss $\ell\colon \Delta_n \to [0,+\infty[^n$ whose Bayes risk satisfies $\underline{L}_\ell \in C^2(]0,+\infty[^n)$, van Erven et al. (2012) showed that the mixability constant η_ℓ is equal to

$$\underline{\eta_\ell} := \inf_{\tilde{\boldsymbol{p}}\in \operatorname{int}\tilde{\Delta}_n} (\lambda_{\max}([\mathsf{H}\underline{\tilde{L}}_{\log}(\tilde{\boldsymbol{p}})]^{-1}\mathsf{H}\underline{\tilde{L}}_\ell(\tilde{\boldsymbol{p}})))^{-1}, \tag{5}$$

The next theorem extends this result by showing that the mixability constant η_ℓ of any loss $\ell\colon \mathcal{A} \to [0,+\infty[^n$ satisfying condition I and such that $\underline{L}_\ell$ is twice differentiable is lower bounded by $\underline{\eta_\ell}$.

Theorem 7 *Let $\eta > 0$, $\ell\colon \mathcal{A} \to [0,+\infty]^n$ a loss satisfying Condition I, and $\underline{\ell}$ a proper support loss of ℓ. Suppose that $\operatorname{dom}\ell = \mathcal{A}$ and that $\underline{L}_\ell$ is twice differentiable on $]0,+\infty[^n$. Then for $\underline{\eta_\ell}$ as in (5), we have $\underline{\eta_\ell} > 0 \implies \ell$ is $\underline{\eta_\ell}$-mixable. In particular, $\eta_\ell \ge \underline{\eta_\ell}$.*

The proof of the theorem is in Appendix C.5. We will show later (Corollary 17) that, under the same conditions of Theorem 7, we actually have $\eta_\ell = \underline{\eta_\ell}$.

Definition 8 (Φ-mixability) [3] *Let $\Phi\colon \mathbb{R}^k \to \mathbb{R}\cup\{+\infty\}$ be an entropy. A loss $\ell\colon \mathcal{A} \to [0,+\infty]^n$ is Φ-mixable if $\forall \boldsymbol{q} \in \Delta_k$, $\forall A := [\boldsymbol{a}_\theta]_{1\le\theta\le k} \in \mathcal{A}^k$, $\exists \boldsymbol{a}_* \in \mathcal{A}$, such that*

$$\forall x \in [n],\ \ell_x(\boldsymbol{a}_*) \le \mathsf{M}_\Phi(\ell_x(A), \boldsymbol{q}) := \inf_{\hat{\boldsymbol{q}}\in\Delta_k} \langle \hat{\boldsymbol{q}}, \ell_x(A)\rangle + D_\Phi(\hat{\boldsymbol{q}}, \boldsymbol{q}). \tag{6}$$

As stated earlier, when ℓ is $(\frac{1}{\eta}\Phi)$-mixable, the GAA (see Figure 1) can achieve a constant regret in the $\mathcal{G}_\ell(k,T)$ game, where the constant does not depend on T. In fact, Reid et al. (2015) showed that for all $T \ge 1$ and $\theta \in [k]$,

$$\sum_{1\le i\le T} \ell_{x^t}(\boldsymbol{a}_*^t) \le \sum_{1\le i\le T} \ell_{x^t}(\boldsymbol{a}_\theta^t) + R_\ell^\Phi,$$

where $R_\ell^\Phi := \inf_{\boldsymbol{q}\in\Delta_k} \max_{\theta\in[k]} D_\Phi(\boldsymbol{e}_\theta, \boldsymbol{q})/\eta_\ell^\Phi$ is the worst-case regret, η_ℓ^Φ is the generalized mixability constant (defined formally in Corollary 18), and $(\boldsymbol{a}_*^t)$ are the outputs of the GAA with initial distribution over experts $\boldsymbol{q}^0 = \operatorname{argmin}_{\boldsymbol{q}\in\Delta_k} \max_{\theta\in[k]} D_\Phi(\boldsymbol{e}_\theta, \boldsymbol{q})$.

3. Our definition of Φ-mixability is slightly different than that introduced by Reid et al. (2015); we use the directional derivative to define the divergence D_Φ.

Aggregating Algorithm	**Generalized Aggregating Algorithm**
Input: $\boldsymbol{q}^0 \in \Delta_k$, $\eta \leq \eta_\ell$, $T \geq 1$.	**Input**: $\boldsymbol{q}^0 \in \Delta_k$, $\eta \leq \eta_\ell^\Phi$, $T \geq 1$.
for $t = 1$ to T	**for** $t = 1$ to T
Observe $A^t = [\boldsymbol{a}^t_\theta]_{1\leq\theta\leq k} \in \mathcal{A}^k$.	Observe $A^t = [\boldsymbol{a}^t_\theta]_{1\leq\theta\leq k} \in \mathcal{A}^k$.
Find $\boldsymbol{a}^t_* \in \mathcal{A}$ such that $\forall x \in [n]$,	Find $\boldsymbol{a}^t_* \in \mathcal{A}$ such that $\forall x \in [n]$,
$\ell_x(\boldsymbol{a}^t_*) \leq -\eta^{-1} \log\langle \exp[-\eta \ell_x(A)], \boldsymbol{q}^{t-1}\rangle$.	$\ell_x(\boldsymbol{a}^t_*) \leq \mathsf{M}_{\Phi_\eta}(\ell_x(A^t), \boldsymbol{q}^{t-1})$.
Observe outcome $x^t \in [n]$.	Observe outcome $x^t \in [n]$.
Set $q^t_\theta = \frac{q^{t-1}_\theta \exp(-\eta \ell_{x^t}(\boldsymbol{a}^t_\theta))}{\sum_{\theta\in[k]} q^{t-1}_\theta \exp(-\eta \ell_{x^t}(\boldsymbol{a}^t_\theta))}$, $\forall \theta \in [k]$.	Set $\boldsymbol{q}^t = \operatorname{argmin}_{\boldsymbol{\mu}\in\Delta_k} \langle \boldsymbol{\mu}, \ell_{x^t}(A^t)\rangle + D_\Phi(\boldsymbol{\mu}, \boldsymbol{q}^{t-1})$.
end for	**end for**

Figure 1: The AA versus the GAA in the $\mathcal{G}_\ell(k,T)$ game. A^t and x^t are the experts' predictions and environment outcome at round t, respectively. $\boldsymbol{q}^0$ is a prior distribution over experts and η is the learning rate. We used the notation $\Phi_\eta := \eta^{-1}\Phi$. When ℓ is η-mixable [resp. Φ-mixable] the AA [resp. GAA] achieves constant regret using the learner's predictions $(\boldsymbol{a}^t_*)_{1\leq t\leq T}$, which are guaranteed to exist. The GAA reduces to the AA when $\Phi = \eta^{-1}\,\mathrm{S}$ and S is the Shannon entropy.

Remark 9 *In order for the update distribution $\boldsymbol{q}^t$ of the GAA (Figure 1) to be well defined, the infimum of the map $\boldsymbol{\mu} \mapsto \langle \boldsymbol{\mu}, \boldsymbol{d}\rangle + D_\Phi(\boldsymbol{\mu}, \boldsymbol{q})$ must be attained at some $\boldsymbol{q}_* \in \Delta_k$ for any given $(\boldsymbol{d}, \boldsymbol{q}) \in [0, +\infty[^n \times \Delta_k$. We verify this for $\boldsymbol{q} \in \operatorname{ri}\Delta_k$; since $\tilde{\boldsymbol{q}} \in \operatorname{int}\operatorname{dom}\tilde{\Phi} = \operatorname{int}\tilde{\Delta}_k$, the function $\tilde{\boldsymbol{\mu}} \mapsto -\tilde{\Phi}'(\tilde{\boldsymbol{q}}; \tilde{\boldsymbol{\mu}} - \tilde{\boldsymbol{q}})$ is lower semicontinuous (Rockafellar, 1997, Cor. 24.5.1). Given that $\tilde{\boldsymbol{\mu}} \mapsto \langle \Pi_k(\tilde{\boldsymbol{\mu}}), \boldsymbol{d}\rangle + \tilde{\Phi}(\tilde{\boldsymbol{\mu}}) - \tilde{\Phi}(\tilde{\boldsymbol{q}})$ is a closed convex function, it is also lower semi-continuous. Therefore, the function $\tilde{\boldsymbol{\mu}} \mapsto \langle \Pi_k(\tilde{\boldsymbol{\mu}}), \boldsymbol{d}\rangle + \tilde{\Phi}(\tilde{\boldsymbol{\mu}}) - \tilde{\Phi}(\tilde{\boldsymbol{q}}) - \tilde{\Phi}'(\tilde{\boldsymbol{q}}; \tilde{\boldsymbol{\mu}} - \tilde{\boldsymbol{q}})$ is lower semicontinuous, and thus attains its minimum on the compact set $\tilde{\Delta}_k$ at some point $\tilde{\boldsymbol{q}}_*$ (Holder, 2005, Thm. 1.13). Using the fact that $D_\Phi(\boldsymbol{\mu}, \boldsymbol{q}) = D_{\tilde{\Phi}}(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{q}})$, we get that $\boldsymbol{q}_* := \Pi_k(\tilde{\boldsymbol{q}}_*) = \operatorname{argmin}_{\boldsymbol{\mu}\in\Delta_k}\langle \boldsymbol{\mu}, \boldsymbol{d}\rangle + D_\Phi(\boldsymbol{\mu}, \boldsymbol{q})$.*

When $\boldsymbol{q}$ in Remark 9 is in $\operatorname{rbd}\Delta_k$, then either $\boldsymbol{q}$ is a vertex or there exists $\mathcal{I} \subset [k]$, with $|\mathcal{I}| > 1$, such that $\boldsymbol{q} \in \operatorname{ri}\Delta_{\mathcal{I}}$. In either case, if we require the existence of a Φ-mixable loss one can use Proposition 11 below and $\Phi^{\mathcal{I}} := \Phi \circ [\Pi^k_{\mathcal{I}}]^\mathsf{T}$ to show that $\boldsymbol{\mu} \mapsto \langle \boldsymbol{\mu}, \boldsymbol{d}\rangle + D_\Phi(\boldsymbol{\mu}, \boldsymbol{q})$ attains a minimum in $\Delta_{\mathcal{I}}$ by following the steps of Remark 9. In this case, the GAA's update step is well defined.

From Figure 1, it is clear that the GAA is divided into two steps; 1) a *substitution step* which consists of finding the prediction $\boldsymbol{a}_* \in \mathcal{A}$ satisfying the mixability condition (5); and 2) an *update step* where a new distribution over experts is computed. Note that the substitution step is not well defined in the sense that there is not a unique choice of $\boldsymbol{a}_*$. One systematic way of choosing $\boldsymbol{a}_*$ is through inverse losses serving as *substitution functions* (Williamson, 2014). Kamalaruban et al. (2015) discuss other alternatives depending on the curvature of the Bayes risk. Note, however, that the choice of the substitution function does not affect the regret bound of the GAA when the loss ℓ is Φ-mixable. On the other hand, the update step is well defined and we show next that this is exactly the MDA for a certain sequence of losses.

Example 10 (GAA versus MDA) *Let $\ell\colon \mathcal{A} \to [0, +\infty[^n$ be a loss and $\Phi\colon \mathbb{R}^k \to \mathbb{R} \cup \{+\infty\}$ an entropy such that $\tilde{\Phi}$ is differentiable on $\operatorname{int}\tilde{\Delta}_k$. Let $\boldsymbol{q}^t$ be the update distribution of the GAA at round t and $\tilde{\boldsymbol{q}}^t = \Pi_k(\boldsymbol{q}^t)$. From Remark 9, it holds that*

$$
\begin{aligned}
\tilde{\boldsymbol{q}}^t &= \operatorname*{argmin}_{\tilde{\boldsymbol{q}} \in \tilde{\Delta}_k} \left\langle \Pi_k(\tilde{\boldsymbol{q}}), \ell_{x^t}(A^t) \right\rangle + \eta^{-1} D_{\tilde{\Phi}}(\tilde{\boldsymbol{q}}, \tilde{\boldsymbol{q}}^{t-1}), \\
&= \operatorname*{argmin}_{\tilde{\boldsymbol{q}} \in \tilde{\Delta}_k} \left\langle \tilde{\boldsymbol{q}}, J_k^\mathsf{T} \ell_{x^t}(A^t) \right\rangle + \eta^{-1} D_{\tilde{\Phi}}(\tilde{\boldsymbol{q}}, \tilde{\boldsymbol{q}}^{t-1}), \\
&= \operatorname*{argmin}_{\tilde{\boldsymbol{q}} \in \tilde{\Delta}_k} \left\langle \tilde{\boldsymbol{q}}, \nabla l_t(\tilde{\boldsymbol{q}}^{t-1}) \right\rangle + \eta^{-1} D_{\tilde{\Phi}}(\tilde{\boldsymbol{q}}, \tilde{\boldsymbol{q}}^{t-1}),
\end{aligned} \tag{7}
$$

where $l_t(\tilde{\boldsymbol{\mu}}) := \langle \Pi_k(\tilde{\boldsymbol{\mu}}), \ell_{x^t}(A^t) \rangle = \langle \boldsymbol{\mu}, \ell_{x^t}(A^t) \rangle$*. Update* (7) *is, by definition (Beck and Teboulle, 2003), the MDA with the sequence of losses* l_t *on* $\operatorname{int} \tilde{\Delta}_k$*, "*distance*" function* $D_{\tilde{\Phi}}(\cdot, \cdot)$*, and learning rate* η*. Therefore, the MDA is exactly the update step of the GAA.*

3.1. Useful Properties

Given a differentiable entropy on Δ_k, Reid et al. (2015) showed that for a non-trivial loss to be Φ-mixable, Φ must have gradients whose norms diverge to infinity near the relative boundary of the simplex. However, the proof of this result assumed differentiability of Φ on the boundary of Δ_k, which is imprecise given that in their definition, Φ is a function on Δ_k — a set with empty interior. In the next proposition, we show that a variant of their result holds when working with directional derivatives. This will be crucial for subsequent results. Before stating the next proposition, note that given an entropy $\Phi \colon \mathbb{R}^k \to \mathbb{R} \cup \{+\infty\}$ and $\mathcal{I} \subseteq [k]$ with $|\mathcal{I}| > 1$, the function $\Phi^{\mathcal{I}} := \Phi \circ [\Pi_{\mathcal{I}}^k]^\mathsf{T} : \mathbb{R}^{|\mathcal{I}|} \to \mathbb{R} \cup \{+\infty\}$ is also an entropy.

Proposition 11 *Let* $\Phi \colon \mathbb{R}^k \to \mathbb{R} \cup \{+\infty\}$ *be an entropy and* $\ell \colon \mathcal{A} \to [0, +\infty]^n$ *a closed non-trivial loss. If* ℓ *is* Φ*-mixable, then* $\forall \mathcal{I} \subseteq [k]$ *with* $|\mathcal{I}| > 1$*,* ℓ *is* $\Phi^{\mathcal{I}}$*-mixable and*

$$
\forall \boldsymbol{q} \in \operatorname{rbd} \Delta_{\mathcal{I}}, \forall \hat{\boldsymbol{q}} \in \operatorname{ri} \Delta_{\mathcal{I}}, \ \Phi'(\boldsymbol{q}; \hat{\boldsymbol{q}} - \boldsymbol{q}) = -\infty. \tag{8}
$$

The proof of Proposition 11 is in Appendix C.6. Contrary to what was claimed previously (Reid et al., 2015), condition (8) together with the strict convexity and differentiability of Φ on $\operatorname{ri} \Delta_k$ is not sufficient for the existence of a Φ-mixable loss (see Appendix D for a counter-example).

In the next proposition, we show that the Bayes risk and the Fenchel dual of $\tilde{\Phi}$ need to be differentiable in the interior of their respective domains for the existence of a Φ mixable loss. As we will argue later (Lemma 35), the differentiability of $\tilde{\Phi}^*$ implies the strict convexity of Φ on $\operatorname{ri} \Delta_k$. The proof of the next proposition is in Appendix C.7.

Proposition 12 *Let* $\Phi \colon \mathbb{R}^k \to \mathbb{R} \cup \{+\infty\}$ *be an entropy and* $\ell \colon \mathcal{A} \to [0, +\infty]^n$ *a loss satisfying Condition I. If* ℓ *is* Φ*-mixable, then the Bayes risk satisfies* $\underline{L}_\ell \in C^1(]0, +\infty[^n)$*. If, additionally,* $\underline{L}_\ell$ *is twice differentiable on* $]0, +\infty[^n$*, then* $\tilde{\Phi}^* \in C^1(\mathbb{R}^{k-1})$*.*

In the next lemma (proved in Appendix C.8), we provide a useful expression of $\mathsf{M}_\Phi(\boldsymbol{d}, \boldsymbol{q})$, which will be used to prove a sufficient condition for Φ-mixability (Theorem 15).

Proposition 13 *Let* $\Phi \colon \mathbb{R}^k \to \mathbb{R} \cup \{+\infty\}$ *be an entropy. Let* $x \in [n], \boldsymbol{d} \in \mathbb{R}^k, \boldsymbol{q} \in \operatorname{ri} \Delta_k$*, and* $\boldsymbol{q}_* = \operatorname{argmin}_{\boldsymbol{\mu} \in \Delta_k} \langle \boldsymbol{\mu}, \boldsymbol{d} \rangle + D_\Phi(\boldsymbol{\mu}, \boldsymbol{q})$*. Then any* $\boldsymbol{s}_{\boldsymbol{q}}^* \in \operatorname{argmax}\{\langle \boldsymbol{s}, \tilde{\boldsymbol{q}}_* - \tilde{\boldsymbol{q}} \rangle : \boldsymbol{s} \in \partial \tilde{\Phi}(\tilde{\boldsymbol{q}})\}$ *satisfies*

$$
\tilde{\boldsymbol{q}}_* \in \partial \tilde{\Phi}^*(\boldsymbol{s}_{\boldsymbol{q}}^* - J_k^\mathsf{T} \boldsymbol{d}), \tag{9}
$$

$$
\mathsf{M}_\Phi(\boldsymbol{d}, \boldsymbol{q}) = d_k + \tilde{\Phi}^*(\boldsymbol{s}_{\boldsymbol{q}}^*) - \tilde{\Phi}^*(\boldsymbol{s}_{\boldsymbol{q}}^* - J_k^\mathsf{T} \boldsymbol{d}). \tag{10}
$$

3.2. The generalized aggregating algorithm using the Shannon entropy S

The purpose of this subsection is to show that the GAA reduces to the AA when the former uses the Shannon entropy. In this case, generalized and classical mixability are equivalent.

Let $\ell\colon \mathcal{A} \to [0,+\infty]^n$ be a loss and Φ be as in Proposition 13 and suppose that Φ and $\tilde{\Phi}^*$ are differentiable on $\operatorname{ri}\Delta_k$ and $\mathbb{R}^{k-1}$, respectively. If we substitute $\boldsymbol{d}$ for $\ell_x(A)$ in (9) and (10), where $A \in (\operatorname{dom}\ell)^k$, it is possible to show that

$$\nabla\Phi^\star(\nabla\Phi(\boldsymbol{q}) - \ell_x(A)) = \operatorname*{argmin}_{\boldsymbol{\mu}\in\Delta_k}\langle\boldsymbol{\mu}, \ell_x(A)\rangle + D_\Phi(\boldsymbol{\mu},\boldsymbol{q}), \tag{11}$$

$$\mathsf{M}_\Phi(\ell_x(A),\boldsymbol{q}) = \Phi^\star(\nabla\Phi(\boldsymbol{q})) - \Phi^\star(\nabla\Phi(\boldsymbol{q}) - \ell_x(A)). \tag{12}$$

This result was already established by Reid et al. (2015).

Let $\boldsymbol{q} \in \operatorname{ri}\Delta_k$. By definition of S, $\nabla\,\mathrm{S}(\boldsymbol{q}) = \log\boldsymbol{q} + \mathbf{1}_k$, and due to Proposition 1, $\mathrm{S}^\star(\boldsymbol{z}) = \log\langle\exp\boldsymbol{z}, \mathbf{1}_k\rangle, \boldsymbol{z} \in \mathbb{R}^k$. Therefore, $\nabla\,\mathrm{S}(\boldsymbol{q}) - \eta\ell_x(A) = \log(\exp(-\eta\ell_x(A)) \odot \boldsymbol{q}) + \mathbf{1}_k$ and $\nabla\,\mathrm{S}^\star(\boldsymbol{z}) = \frac{\exp\boldsymbol{z}}{\langle\exp\boldsymbol{z},\mathbf{1}_k\rangle}, \forall (x,A) \in [n]\times(\operatorname{dom}\ell)^k$. Thus,

$$\nabla\mathrm{S}^\star(\nabla\,\mathrm{S}(\boldsymbol{q}) - \eta\ell_x(A)) = \frac{\exp(-\eta\ell_x(A))\odot\boldsymbol{q}}{\langle\exp(-\eta\ell_x(A)),\boldsymbol{q}\rangle}. \tag{13}$$

Let $\mathrm{S}_\eta := \eta^{-1}\,\mathrm{S}$. Then $\nabla\,\mathrm{S} = \eta\nabla\,\mathrm{S}_\eta$ and $\forall \boldsymbol{z} \in \mathbb{R}^k, \nabla\,\mathrm{S}_\eta^\star(\boldsymbol{z}) = \nabla\,\mathrm{S}^\star(\eta\boldsymbol{z})$ (Reid et al., 2015).[4] Then the left hand side of (13) can be written as $\nabla\,\mathrm{S}_\eta^\star(\nabla\,\mathrm{S}_\eta(\boldsymbol{q}) - \ell_x(A))$. Using this fact, (11) and (13) show that the update distribution $\boldsymbol{q}^t$ of the GAA (Figure 1) coincides with that of the AA after substituting $\boldsymbol{q}, x$, and A by $\boldsymbol{q}^{t-1}, x^t$, and $A^t := [\boldsymbol{a}_\theta]_{\theta\in[k]}$, respectively.

Now using the fact that $\mathsf{M}_{\mathrm{S}_\eta}(\ell_x(A),\boldsymbol{q}) = \eta^{-1}\mathsf{M}_{\mathrm{S}}(\eta\ell_x(A),\boldsymbol{q})$ (Reid et al., 2015) and (12), we get $\mathsf{M}_{\mathrm{S}_\eta}(\ell_x(A),\boldsymbol{q}) = \eta^{-1}[\mathrm{S}^\star(\nabla\,\mathrm{S}(\boldsymbol{q})) - \mathrm{S}^\star(\nabla\,\mathrm{S}(\boldsymbol{q}) - \eta\ell_x(A))] = -\eta^{-1}\log\langle\exp(-\eta\ell_x(A)),\boldsymbol{q}\rangle$. Suppose now that $\boldsymbol{q}$ belongs to relative interior of a face $\Delta_{\mathcal{I}} \subsetneq \Delta_k$ with $\mathcal{I} \subsetneq [k]$. We can repeat the argument above for $\mathrm{S}^{\mathcal{I}} := \mathrm{S}\circ[\Pi^k_{\mathcal{I}}]^\mathsf{T}$ to show that

$$\begin{aligned}\mathsf{M}_{\mathrm{S}^{\mathcal{I}}_\eta}(\Pi^k_{\mathcal{I}}\ell_x(A), \Pi^k_{\mathcal{I}}\boldsymbol{q}) &= -\eta^{-1}\log\langle\exp(-\eta\Pi^k_{\mathcal{I}}\ell_x(A)), \Pi^k_{\mathcal{I}}\boldsymbol{q}\rangle,\\ &= -\eta^{-1}\log\langle\exp(-\eta\ell_x(A)),\boldsymbol{q}\rangle.\end{aligned} \tag{14}$$

Equation 14 will now allow us to prove the equivalence between classical and generalized mixability when using the Shannon entropy. The proof of Theorem 14 is in Appendix C.8.

Theorem 14 *Let $\eta > 0$ and $\mathrm{S}_\eta := \eta^{-1}\,\mathrm{S}$, where S is the Shannon entropy. For a loss $\ell\colon\mathcal{A}\to[0,+\infty]^n$, ℓ is η-mixable if and only if ℓ is S_η-mixable.*

3.3. Necessary and Sufficient Conditions for Φ-Mixability

In this subsection, we show that given an entropy $\Phi\colon\mathbb{R}^k\to\mathbb{R}\cup\{+\infty\}$ and a loss $\ell\colon\mathcal{A}\to[0,+\infty[^n$ satisfying certain regularity conditions, ℓ is Φ-mixable if and only if

$$\boxed{\underline{\eta_\ell}\Phi - \mathrm{S} \text{ is convex on } \Delta_k.} \tag{15}$$

4. Reid et al. (2015) showed the equality $\nabla\Phi^\star_\eta(\boldsymbol{u}) = \nabla\Phi^\star(\eta\boldsymbol{u}), \forall\boldsymbol{u}\in\operatorname{dom}\Phi^\star$, for any entropy differentiable on Δ_k - not just for the Shannon Entropy.

Theorem 15 *Let $\Phi\colon \mathbb{R}^k \to \mathbb{R} \cup \{+\infty\}$ be an entropy and $\ell\colon \mathcal{A} \to [0, +\infty]^n$ be a loss. If $\eta_\ell \Phi - \mathrm{S}$ is convex on Δ_k, then ℓ is Φ-mixable.*

As consequence of Theorem 15 (whose proof is in Appendix C.10), if a loss ℓ is not classically mixable (that is, $\eta_\ell = 0$) it cannot be Φ-mixable for any entropy Φ. This is because $\eta_\ell \Phi - \mathrm{S} = -\mathrm{S}$ is not convex. The converse of Theorem 15 also holds under additional smoothness conditions on Φ and ℓ as stated in the next theorem. The proof is in Appendix C.11.

Theorem 16 *Let $\Phi\colon \mathbb{R}^k \to \mathbb{R} \cup \{+\infty\}$ be an entropy such that $\tilde{\Phi} := \Phi \circ \Pi_k$ is twice differentiable on $\operatorname{int} \tilde{\Delta}_k$, and $\ell\colon \mathcal{A} \to [0, +\infty]^n$ a loss satisfying Condition I and such that $\underline{L}_\ell$ is twice differentiable on $]0, +\infty[^n$. Then ℓ is Φ-mixable only if $\underline{\eta_\ell} \Phi - \mathrm{S}$ is convex on Δ_k.*

The twice differentiability condition of $\underline{L}_\ell$ in Theorem 16 is not as strong as it may first appear; we showed in Proposition 12 that ℓ is Φ-mixable only if $\underline{L}_\ell \in C^1(]0, +\infty[)^n$. Furthermore, since $\underline{L}_\ell$ is concave, Alexandrov's Theorem (see e.g. (Borwein et al., 2010, Thm. 6.7)) guarantees that $\underline{L}_\ell$ is twice differentiable almost everywhere in $]0, +\infty[^n$. A version of Theorem 16 which does not assume the twice differentiability of the Bayes risk is given in Appendix D (Theorem 38).

Corollary 17 *Let ℓ and Φ be as in Theorem 16. If $\operatorname{dom} \ell = \mathcal{A}$, then $\eta_\ell = \underline{\eta_\ell}$. Furthermore, ℓ is Φ-mixable if and only if $\underline{\eta_\ell} \Phi - \mathrm{S}$ is convex on Δ_k.*

Proof We already know from Theorem 7 that $\underline{\eta_\ell} \le \eta_\ell$. Suppose now that ℓ is classically mixable. By definition of η_ℓ, ℓ is η_ℓ-mixable, and thus from Theorem 14, ℓ is $(\eta_\ell^{-1}\,\mathrm{S})$-mixable. Substituting Φ for $\eta_\ell^{-1}\,\mathrm{S}$ in Theorem 16 implies that $(\underline{\eta_\ell}/\eta_\ell - 1)\,\mathrm{S}$ is convex on $\operatorname{ri} \Delta_k$. Consequently, $\eta_\ell \le \underline{\eta_\ell}$, and hence $\eta_\ell = \underline{\eta_\ell}$. From this fact and Theorems 15 and 16, it follows that ℓ is Φ-mixable if and only if $\underline{\eta_\ell} \Phi - \mathrm{S}$ is convex on Δ_k. ∎

Corollary 17 suggests that when the Bayes risk is twice differentiable on the interior of its domain it contains all necessary information for the characterization of classical mixability.

Corollary 18 (The Generalized Mixability Constant) *Let ℓ and Φ be as in Theorem 16. If $\operatorname{dom} \ell = \mathcal{A}$, then the largest $\eta \ge 0$ such that ℓ is Φ_η-mixable is given by*

$$\boxed{\eta_\ell^\Phi = \underline{\eta_\ell} \inf_{\tilde{\boldsymbol{q}} \in \operatorname{int} \tilde{\Delta}_k} \lambda_{\min}(\mathsf{H}\tilde{\Phi}(\tilde{\boldsymbol{q}})(\mathsf{H}\tilde{\mathrm{S}}(\tilde{\boldsymbol{q}}))^{-1}).} \tag{16}$$

The proof is in Appendix C.12. Observe that if we substitute Φ by S in (14), we get $\eta_\ell^{\mathrm{S}} = \underline{\eta_\ell}$.

4. Discussion and Future Work

So far, we showed that the Φ-mixability of losses satisfying Condition I are characterized by the convexity of $\eta^* \Phi - \mathrm{S}$, where $\eta^* \in \{\eta_\ell, \underline{\eta_\ell}\}$ (see Theorems 15 and 16). As a consequence of this, and contrary to what was conjectured previously (Reid et al., 2015), the generalized mixability condition does not induce a correspondence between losses and entropies; for a given loss ℓ, there is no particular entropy Φ^ℓ — specific to the choice of ℓ — which minimizes the regret of the GAA. Rather, the Shannon entropy S minimizes the regret regardless of the choice of ℓ (see Theorem 19 for the precise statement). This is consistent with Vovk's result regarding the fundamental nature of the log-loss (Vovk, 2015).

Nevertheless, given a loss ℓ and entropy Φ, the curvature of the loss surface $\mathcal{S}_\ell$ determines the maximum learning rate η_ℓ^Φ of the GAA; in other words, it determines for which scaled entropies $\Phi_\eta := \eta^{-1}\Phi$, ℓ is Φ_η-mixable. In fact, the curvature of $\mathcal{S}_\ell$ is linked to $\underline{\eta_\ell}$ through the Hessian of the Bayes risk (see Theorem 41 in Appendix E.1 for the precise statement). The magnitude of $\underline{\eta_\ell}$ determines the learning rate η_ℓ^Φ as per Corollary 18.

Given a loss ℓ satisfying Condition I, we will now use the expression of the generalized mixability constant η_ℓ^Φ to explicitly compare the regret bounds achieved using the GAA with different entropies. For an entropy $\Phi\colon \mathbb{R}^k \to \mathbb{R}\cup\{+\infty\}$, we define $R_\ell^\Phi(\boldsymbol{q}) := \max_\theta D_\Phi(\boldsymbol{e}_\theta, \boldsymbol{q})/\eta_\ell^\Phi$, $\boldsymbol{q} \in \operatorname{ri}\Delta_k$. The optimal regret bound achieved using the GAA is given by $R_\ell^\Phi := \inf_{\boldsymbol{q}\in\Delta_k} R_\ell^\Phi(\boldsymbol{q})$ (Reid et al., 2015). The next theorem shows that $R_\ell^{\mathrm{S}} \leq R_\ell^\Phi$. The proof is in Appendix C.13.

Theorem 19 *Let* $\mathrm{S}, \Phi\colon \mathbb{R}^k \to \mathbb{R}\cup\{+\infty\}$, *where* S *is the Shannon entropy and* Φ *is an entropy such that* $\tilde{\Phi}$ *is twice differentiable on* $\operatorname{int}\tilde{\Delta}_k$. *A loss* $\ell : \mathcal{A} \to [0,+\infty[^n$, *satisfying Condition I and with* $\underline{L}_\ell$ *twice differentiable on* $]0,+\infty[^n$, *is* Φ*-mixable only if* $R_\ell^{\mathrm{S}} \leq R_\ell^\Phi$.

We conjecture that the result of Theorem 19 still holds even if one relaxes the smoothness requirement on $\tilde{\Phi}$ and $\underline{L}_\ell$.

The result of Theorem 19 only concerns the worst case bound of the GAA. Since the update step of the GAA is exactly the MDA one can perhaps benefit from varying the entropy and the learning rate at each round to improve the regret bounds (Steinhardt and Liang, 2014; Joulani et al., 2017). In this scenario, at each round t the (adaptive) GAA would use a different entropy Φ^t computed according to the past performance of experts. Corollary 18 would still give an upper bound on the learning rate η^t at each round t as a function of the Hessian of Φ^t. Furthermore, this upper bound could be made larger by only requiring the Φ-mixability condition (5) to be satisfied locally; that is, ensuring (5) for $\boldsymbol{q} = \boldsymbol{q}^{t-1}$ and $A = A^t \in \mathcal{A}^k$, where $\boldsymbol{q}^{t-1}$ is the previous distribution over experts and A^t is their prediction matrix. This would allow higher learning rates and potentially faster convergence.

Acknowledgments

This work was supported by the Australian Research Council of Data61.

References

Ravi P. Agarwal, Maria Meehan, and Donal O'Regan. *Fixed point theory and applications*. Cambridge university press, 2001.

Amir Beck and Marc Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003.

Dennis S. Bernstein. *Matrix Mathematics: Theory, Facts, and Formulas*. Princeton University Press, second edition, 2011.

Jonathan M. Borwein, Jon D. Vanderwerff, et al. *Convex functions: constructions, characterizations and counterexamples*. Cambridge University Press, 2010.

Alexey Chernov, Yuri Kalnishkan, Fedor Zhdanov, and Vladimir Vovk. Supermartingales in prediction with expert advice. *Theoretical Computer Science*, 411(29-30):2647–2669, 2010.

A. Philip Dawid. The geometry of proper scoring rules. *Annals of the Institute of Statistical Mathematics*, 59(1):77–93, 2007.

Godfrey Harold Hardy. *A course of pure mathematics*. Cambridge University Press, 2008.

J-B. Hiriart-Urruty and C. Lemaréchal. Fundamentals of convex analysis. Springer, 2001.

Allen Holder. An introduction to mathematical optimization, 2005.

Pooria Joulani, András György, and Csaba Szepesvári. A modular analysis of adaptive (non-)convex optimization: Optimism, composite objectives, and variational bounds. In *International Conference on Algorithmic Learning Theory*, pages 681–720, 2017.

Parameswaran Kamalaruban, Robert Williamson, and Xinhua Zhang. Exp-concavity of proper composite losses. In *Conference on Learning Theory*, pages 1035–1065, 2015.

Nelson Merentes and Kazimierz Nikodem. Remarks on strongly convex functions. *Aequationes mathematicae*, 80(1):193–199, 2010.

Francesco Orabona, Koby Crammer, and Nicolò Cesa-Bianchi. A generalized online mirror descent with applications to classification and regression. *Machine Learning*, 99(3):411–435, 2015.

Mark D. Reid, Rafael M. Frongillo, Robert C. Williamson, and Nishant Mehta. Generalized mixability via entropic duality. In *Conference on Learning Theory*, pages 1501–1522, 2015.

Joel W. Robbin and Dietmar A. Salamon. Introduction to differential geometry. *ETH, Lecture Notes, preliminary version, January*, 2011.

R. Tyrrell Rockafellar. Convex analysis. Princeton University Press, Princeton, NJ, 1997.

Walter Rudin. *Principles of mathematical analysis*. McGraw-Hill, 1964.

Jacob Steinhardt and Percy Liang. Adaptivity and optimism: An improved exponentiated gradient algorithm. In *International Conference on Machine Learning*, pages 1593–1601, 2014.

John A. Thorpe. *Elementary Topics in Differential Geometry*. Springer Science & Business Media, 1994.

Tim van Erven, Mark D. Reid, and Robert C. Williamson. Mixability is Bayes risk curvature relative to log loss. *Journal of Machine Learning Research*, 13:1639–1663, 2012.

Vladimir Vovk. A game of prediction with expert advice. *Journal of Computer and System Sciences*, 56(2):153–173, 1998.

Vladimir Vovk. The fundamental nature of the log loss function. In *Fields of Logic and Computation II*, pages 307–318. Springer, 2015.

Robert C. Williamson. The geometry of losses. In *Conference on Learning Theory*, pages 1078–1108, 2014.

Robert C. Williamson, Elodie Vernet, and Mark D. Reid. Composite multiclass losses. *Journal of Machine Learning Research*, 17:223:1–223:52, 2016.

Appendix A. Additional Background Results

The following proposition gives some useful properties of the Fenchel dual which will be used in several proofs.

Proposition 20 (Hiriart-Urruty and Lemaréchal (2001)) *Let $f, h : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$. If f and h are proper and there are affine functions minorizing them on $\mathbb{R}^n$, then for $\boldsymbol{v} \in \mathbb{R}^n$*

$$
\begin{array}{llll}
(i) & g(\boldsymbol{u}) = f(\boldsymbol{u}) + r & \implies & g^*(\boldsymbol{v}) = f^*(\boldsymbol{v}) - r, \\
(ii) & g(\boldsymbol{u}) = f(\boldsymbol{u}) + \langle \boldsymbol{v}_0, \boldsymbol{u} \rangle & \implies & g^*(\boldsymbol{v}) = f^*(\boldsymbol{v} - \boldsymbol{v}_0), \\
(iii) & f \leq h & \implies & f^* \geq h^*, \\
(iv) & \boldsymbol{s} \in \partial f^*(\boldsymbol{v}) & \implies & f^*(\boldsymbol{v}) = \langle \boldsymbol{v}, \boldsymbol{s} \rangle - f(\boldsymbol{s}), \\
(v) & g(\boldsymbol{u}) = f(t\boldsymbol{u}), t > 0 & \implies & g^*(\boldsymbol{v}) = f^*(\boldsymbol{v}/t),
\end{array}
$$

The following result due to Chernov et al. (2010) will be crucial to prove the convexity of the superprediction set (Theorem 5).

Lemma 21 (Chernov et al. (2010)) *Let $\Delta(\Omega)$ be the set of distributions over some set $\Omega \subseteq \mathbb{R}$. Let a function $Q : \Delta(\Omega) \times \Omega \to \mathbb{R}$ be such that $Q(\cdot, \omega)$ is continuous for all $\omega \in \Omega$. If for all $\boldsymbol{\pi} \in \Delta(\Omega)$ it holds that $\mathbb{E}_{\omega \sim \boldsymbol{\pi}} Q(\boldsymbol{\pi}, \omega) \leq r$, where $r \in \mathbb{R}$ is some constant, then*

$$\exists \boldsymbol{\pi} \in \Delta(\Omega), \forall \omega \in \Omega,\ Q(\boldsymbol{\pi}, \omega) \leq r.$$

Note that when Ω in the lemma above is $[n]$, $\Delta([n]) \equiv \Delta_n$.

We make use of the following lemma due to Bernstein (2011) in proving a necessary condition for Φ-mixability (Theorem 16).

Lemma 22 *$\forall m \geq 1, \forall A, B \in \mathbb{R}^{m \times m}$, $\lambda_{\max}(AB) = \lambda_{\max}(BA)$ and $\lambda_{\min}(AB) = \lambda_{\min}(BA)$.*

Appendix B. Technical Lemmas

This appendix presents technical lemmas which will be needed in various proofs within this paper.

For an open convex set Ω in $\mathbb{R}^n$, a function $\phi \colon \Omega \to \mathbb{R}$ is said to be α-*strongly convex* if $\boldsymbol{z} \mapsto \phi(\boldsymbol{u}) - \alpha \|\boldsymbol{u}\|^2$ is convex on Ω (Merentes and Nikodem, 2010). The next lemma is a generalization of α-strong convexity, where $\boldsymbol{u} \mapsto \|\boldsymbol{u}\|^2$ is replaced by any strictly convex function.

Lemma 23 *Let $\Omega \subseteq \mathbb{R}^n$ be an open convex set. Let $\phi, \psi \colon \Omega \to \mathbb{R}$ be twice differentiable.*

If ψ is strictly convex, then $\forall \boldsymbol{u} \in \Omega$, $\mathsf{H}\psi(\boldsymbol{u})$ is invertible, and for any $\alpha > 0$

$$\forall \boldsymbol{u} \in \Omega,\ \lambda_{\min}(\mathsf{H}\phi(\boldsymbol{u})(\mathsf{H}\psi(\boldsymbol{u}))^{-1}) \geq \alpha \iff \phi - \alpha\psi \text{ is convex}, \tag{17}$$

Furthermore, if $\alpha > 1$, then the left hand side of (17) implies that $\phi - \psi$ is strictly convex.

Proof Suppose that $\inf_{\boldsymbol{u} \in \Omega} \lambda_{\min}(\mathsf{H}\phi(\boldsymbol{u})(\mathsf{H}\psi(\boldsymbol{u}))^{-1}) \geq \alpha$. Since g is strictly convex and twice differentiable on Ω, $\mathsf{H}\psi(\boldsymbol{u})$ is symmetric positive definite, and thus invertible. Therefore, there exists a symmetric positive definite matrix $G \in \mathbb{R}^{n\times n}$ such that $GG = \mathsf{H}\psi(\boldsymbol{u})$. Lemma 22 implies

$$\begin{aligned}
& \inf_{\boldsymbol{u}\in\Omega} \lambda_{\min}(\mathsf{H}\phi(\boldsymbol{u})(\mathsf{H}\psi(\boldsymbol{u}))^{-1}) && \geq \alpha, \\
\iff \quad & \inf_{\boldsymbol{u}\in\Omega} \lambda_{\min}(G^{-1}\mathsf{H}\phi(\boldsymbol{u})G^{-1}) && \geq \alpha, \\
\iff \quad & \forall \boldsymbol{u} \in \Omega, \forall \boldsymbol{v} \in \mathbb{R}^n \setminus \{\boldsymbol{0}\}, \tfrac{\boldsymbol{v}^\mathsf{T} G^{-1}(\mathsf{H}\phi(\boldsymbol{u}))G^{-1}\boldsymbol{v}}{\boldsymbol{v}^\mathsf{T}\boldsymbol{v}} && \geq \alpha, \\
\iff \quad & \forall \boldsymbol{u} \in \Omega, \forall \boldsymbol{w} \in \mathbb{R}^n \setminus \{\boldsymbol{0}\}, \boldsymbol{w}^\mathsf{T}(\mathsf{H}\phi(\boldsymbol{u}))\boldsymbol{w} && \geq \alpha \boldsymbol{w}^\mathsf{T} GG\boldsymbol{w} = \boldsymbol{w}^\mathsf{T}(\alpha\mathsf{H}\psi(\boldsymbol{u}))\boldsymbol{w}, \\
\iff \quad & \forall \boldsymbol{u} \in \Omega, \mathsf{H}\phi(\boldsymbol{u}) && \succeq \alpha\mathsf{H}\psi(\boldsymbol{u}), \\
\iff \quad & \forall \boldsymbol{u} \in \Omega, \mathsf{H}(\phi - \alpha\psi)(\boldsymbol{u}) && \succeq 0,
\end{aligned}$$

where in the third and fifth lines we used the definition of minimum eigenvalue and performed the change of variable $\boldsymbol{w} = G^{-1}\boldsymbol{v}$, respectively. To conclude the proof of (17), note that the positive semi-definiteness of $\mathsf{H}(\phi - \alpha\psi)$ is equivalent to the convexity of $\phi - \alpha\psi$ (Hiriart-Urruty and Lemaréchal, 2001, Thm B.4.3.1).

Finally, note that the equivalences established above still hold if we replace α, "$\geq$", and "$\succeq$ " by 1, "$>$", and "$\succ$" , respectively. The strict convexity of $\phi - \psi$ would then follow from the positive definiteness of $\mathsf{H}(\phi - \psi)$ (ibid.). ■

The next crucial lemma is a slight modification of a result due to Chernov et al. (2010).

Lemma 24 *Let* $f\colon \operatorname{ri}\Delta_n \times [n] \to \mathbb{R}$ *be a continuous function in the first argument and such that* $\forall(\boldsymbol{q}, x) \in \operatorname{ri}\Delta_n \times [n], -\infty < f(\boldsymbol{q}, x)$. *Suppose that* $\forall \boldsymbol{p} \in \operatorname{ri}\Delta_n, \mathbb{E}_{x\sim\boldsymbol{p}}[f(\boldsymbol{p}, x)] \leq 0$, *then*

$$\forall \epsilon > 0, \exists \boldsymbol{p}_\epsilon \in \operatorname{ri}\Delta_n, \forall x \in [n], f(\boldsymbol{p}_\epsilon, x) \leq \epsilon.$$

Proof Pick any $\delta > 0$ such that $\delta(n-1) < 1$, and $c_0 < 0$ such that $\forall(\boldsymbol{q}, x) \in \operatorname{ri}\Delta_n \times [n], c_0 \leq f(\boldsymbol{q}, x)$. We define $\Delta_n^\delta := \{\boldsymbol{p} \in \Delta_n : \forall x \in [n], p_x \geq \delta\}$ and $g(\boldsymbol{q}, \boldsymbol{p}) := \mathbb{E}_{x\sim\boldsymbol{q}}[f(\boldsymbol{p}, x)]$. For a fixed $\boldsymbol{q}$, $\boldsymbol{p} \mapsto g(\boldsymbol{q}, \boldsymbol{p})$ is continuous, since f is continuous in the first argument. For a fixed $\boldsymbol{p}$, $\boldsymbol{q} \mapsto g(\boldsymbol{q}, \boldsymbol{p})$ is linear, and thus concave. Since Δ_n^δ is convex and compact, g satisfies Ky Fan's minimax Theorem (Agarwal et al., 2001, Thm. 11.4), and therefore, there exists $\boldsymbol{p}^\delta \in \Delta_n^\delta$ such that

$$\forall \boldsymbol{q} \in \Delta_n^\delta, \quad \mathbb{E}_{x\sim\boldsymbol{q}}[f(\boldsymbol{p}^\delta, x)] = g(\boldsymbol{q}, \boldsymbol{p}^\delta) \leq \sup_{\boldsymbol{\mu}\in\Delta_n^\delta} g(\boldsymbol{\mu}, \boldsymbol{\mu}) = \sup_{\boldsymbol{\mu}\in\Delta_n^\delta} \mathbb{E}_{x\sim\boldsymbol{\mu}}[f(\boldsymbol{\mu}, x)] \leq 0. \tag{18}$$

For $x_0 \in [n]$, let $\hat{\boldsymbol{q}} \in \Delta_n^\delta$ be such that $\hat{q}_{x_0} = 1 - \delta(n-1)$ and $\hat{q}_x = \delta$ for $x \neq x_0$ (this is a legitimate distribution since $\delta(n-1) < 1$ by construction). Substituting $\hat{\boldsymbol{q}}$ for $\boldsymbol{q}$ in (18) gives

$$\begin{aligned}
& (1 - \delta(n-1))f(\boldsymbol{p}^\delta, x_0) + \delta \textstyle\sum_{x\neq x_0} f(\boldsymbol{p}^\delta, x) && \leq 0, \\
\implies \quad & (1 - \delta(n-1))f(\boldsymbol{p}^\delta, x_0) && \leq -c_0\delta(n-1), \\
\implies \quad & f(\boldsymbol{p}^\delta, x_0) && \leq [-c_0\delta(n-1)]/[1 - \delta(n-1)].
\end{aligned}$$

Choosing $\delta^* := \epsilon/[(-c_0 + \epsilon)(n-1)]$, and $\boldsymbol{p}_\epsilon := \boldsymbol{p}^{\delta^*}$ gives the desired result.

■

Lemma 25 *Let* $f, g\colon I \to \mathbb{R}^n$, *where* $I \subseteq \mathbb{R}$ *is an open interval containing* 0. *Suppose* $g(t)$ *[resp.* f*] is continuous [resp. differentiable] at* 0. *Then* $t \mapsto \langle f(t), g(t)\rangle$ *is differentiable at* 0 *if and only if* $t \mapsto \langle f(0), g(t)\rangle$ *is differentiable at* 0, *and we have*

$$\left.\frac{d}{dt}\langle f(t), g(t)\rangle\right|_{t=0} = \left\langle \left.\frac{d}{dt} f(t)\right|_{t=0}, g(0) \right\rangle + \left.\frac{d}{dt}\langle f(0), g(t)\rangle\right|_{t=0}.$$

Proof We have

$$\frac{\langle f(t), g(t)\rangle - \langle f(0), g(0)\rangle}{t} = \frac{\langle f(t), g(t)\rangle - \langle f(0), g(t)\rangle}{t} + \frac{\langle f(0), g(t)\rangle - \langle f(0), g(0)\rangle}{t},$$
$$= \left\langle \frac{f(t) - f(0)}{t}, g(t) \right\rangle + \frac{\langle f(0), g(t)\rangle - \langle f(0), g(0)\rangle}{t}.$$

But since g [resp. f] is continuous [resp. differentiable] at 0, the first term on the right hand side of the above equation converges to $\langle \frac{d}{dt} f(t)\big|_{t=0}, g(0)\rangle$ as $t \to 0$. Therefore, $\frac{1}{t}(\langle f(0), g(t)\rangle - \langle f(0), g(0)\rangle)$ admits a limit when $t \to 0$ if and only if $\frac{1}{t}(\langle f(t), g(t)\rangle - \langle f(0), g(0)\rangle)$ admits a limit when $t \to 0$. This shows that $t \mapsto \langle f(0), g(t)\rangle$ is differentiable at 0 if an only if $t \mapsto \langle f(t), g(t)\rangle$ is differentiable at 0, and in this case the above equation yields

$$\frac{d}{dt} \langle f(t), g(t)\rangle \Big|_{t=0} = \lim_{t\to 0} \frac{\langle f(t), g(t)\rangle - \langle f(0), g(0)\rangle}{t},$$
$$= \lim_{t\to 0} \left(\left\langle \frac{f(t) - f(0)}{t}, g(t) \right\rangle + \frac{\langle f(0), g(t)\rangle - \langle f(0), g(0)\rangle}{t} \right),$$
$$= \left\langle \frac{d}{dt} f(t)\Big|_{t=0}, g(0) \right\rangle + \frac{d}{dt} \langle f(0), g(t)\rangle \Big|_{t=0}.$$

∎

Note that the differentiability of $t \mapsto \langle f(0), g(t)\rangle$ at 0 does not necessarily imply the differentiability of g at 0. Take for example $n = 3$, $f(t) = \mathbf{1}/3$ for $t \in]-1, 1[$, and

$$g(t) = \begin{cases} -t\boldsymbol{e}_1 + t\frac{\mathbf{1}}{3}, & \text{if } t \in]-1, 0[; \\ -t\frac{\mathbf{1}}{3} + t\boldsymbol{e}_2, & \text{if } t \in [0, 1[. \end{cases}$$

Then the function $t \mapsto \langle f(0), g(t)\rangle = 0$ is differentiable at 0 but g is not. The preceding Lemma will be particularly useful in settings where it is desired to compute the derivative $\frac{d}{dt}\langle f(0), g(t)\rangle|_{t=0}$ without any explicit assumptions on the differentiability of $g(t)$ at 0. For example, this will come up when computing $\frac{d}{dt}\langle \boldsymbol{p}, D\tilde{\ell}(\tilde{\boldsymbol{\alpha}}^t)\boldsymbol{v}\rangle|_{t=0}$, where $\boldsymbol{v} \in \mathbb{R}^{n-1}$ and $t \mapsto \tilde{\boldsymbol{\alpha}}^t$ is smooth curve on $\operatorname{int} \tilde{\Delta}_n$, with the only assumption that $\tilde{L}_\ell$ is twice differentiable at $\tilde{\boldsymbol{\alpha}}^0 \in \operatorname{int} \tilde{\Delta}_n$.

Lemma 26 *Let $\ell \colon \Delta_n \to [0, +\infty]^n$ be a proper loss. For any $\boldsymbol{p} \in \operatorname{ri} \Delta_n$, it holds that*

$$\ell \text{ is continuous at } \boldsymbol{p} \overset{(i)}{\iff} \underline{L}_\ell \text{ is differentiable at } \boldsymbol{p} \overset{(ii)}{\iff} \partial[-\underline{L}_\ell](\boldsymbol{p}) = \{\nabla \underline{L}_\ell(\boldsymbol{p})\} = \{\ell(\boldsymbol{p})\}.$$

Proof [$\overset{(i)}{\iff}$] This equivalence has been shown before by Williamson et al. (2016).

[$\overset{(ii)}{\iff}$] Since $\underline{L}_\ell(\boldsymbol{p}) = -\sigma_{\mathcal{S}_\ell^\oplus}(-\boldsymbol{p})$, for all $\boldsymbol{p} \in \operatorname{ri} \Delta_n$, it follows that $\underline{L}_\ell$ is differentiable at $\boldsymbol{p}$ if and only if $\partial[-\underline{L}_\ell](\boldsymbol{p}) = \partial\sigma_{\mathcal{S}_\ell^\oplus}(-\boldsymbol{p}) = \{-\nabla\sigma_{\mathcal{S}_\ell^\oplus}(-\boldsymbol{p})\} = \{\nabla \underline{L}_\ell(\boldsymbol{p})\}$ (Hiriart-Urruty and Lemaréchal, 2001, Cor. D.2.1.4). It remains to show that $\nabla \underline{L}_\ell(\boldsymbol{r}) = \ell(\boldsymbol{r})$ when $\underline{L}_\ell$ is differentiable at $\boldsymbol{r} \in \operatorname{ri} \Delta_n$. Let $\boldsymbol{\alpha}_x^t = \boldsymbol{r} + t\boldsymbol{e}_x$ and $\tilde{\boldsymbol{\alpha}}_x^t = \Pi_n(\boldsymbol{\alpha}_x^t)$, where $(\boldsymbol{e}_x)_{x\in[n]}$ is the standard basis of $\mathbb{R}^n$. For $x \in [n]$, the functions $f_x(t) \coloneqq \boldsymbol{\alpha}_x^t$ and $g_x(t) \coloneqq \tilde{\ell}(\tilde{\boldsymbol{\alpha}}_x^t)$ satisfy the conditions of Lemma 25.

Therefore, $h_x(t) := \langle f_x(0), g_x(t)\rangle = \langle \boldsymbol{r}, \tilde{\ell}(\tilde{\boldsymbol{\alpha}}_x^t)\rangle$ is differentiable at 0 and

$$\begin{aligned}\nabla \underline{\tilde{L}}(\boldsymbol{r})\boldsymbol{e}_x = \frac{d}{dt}\underline{\tilde{L}}(\boldsymbol{\alpha}_x^t)\Big|_{t=0} &= \frac{d}{dt}\left\langle f_x(t), g_x(t)\right\rangle\Big|_{t=0}, \\ &= \left\langle \boldsymbol{e}_x, \tilde{\ell}(\tilde{\boldsymbol{r}})\right\rangle + \frac{d}{dt}h_x(t)\Big|_{t=0}, \\ &= \tilde{\ell}_x(\tilde{\boldsymbol{r}}),\end{aligned}$$

where the last equality holds because h_x attains a minimum at 0 due to the properness of ℓ. The result being true for all $x \in [n]$ implies that $\nabla \underline{\tilde{L}}(\tilde{\boldsymbol{r}}) = \tilde{\ell}(\tilde{\boldsymbol{r}}) = \ell(\boldsymbol{r})$. ■

The next Lemma is a restatement of earlier results due to van Erven et al. (2012). Our proof is more concise due to our definition of the Bayes risk in terms of the support function of the superprediction set.

Lemma 27 (van Erven et al. (2012)) *Let $\ell\colon \Delta_n \to [0,+\infty]^n$ be a proper loss whose Bayes risk is twice differentiable on $]0,+\infty[^n$ and let $X_{\boldsymbol{p}} = I_{\tilde{n}} - \mathbf{1}_{\tilde{n}}\tilde{\boldsymbol{p}}^\mathsf{T}$. The following holds*

(i) $\forall \boldsymbol{p} \in \operatorname{ri}\Delta_n$, $\langle \boldsymbol{p}, \mathsf{D}\tilde{\ell}(\tilde{\boldsymbol{p}})\rangle = \mathbf{0}_{\tilde{n}}^\mathsf{T}$.

(ii) $\forall \tilde{\boldsymbol{p}} \in \operatorname{int}\tilde{\Delta}_n$, $\mathsf{D}\tilde{\ell}(\tilde{\boldsymbol{p}}) = \begin{bmatrix} X_{\boldsymbol{p}} \\ -\tilde{\boldsymbol{p}}^\mathsf{T}\end{bmatrix}\mathsf{H}\underline{\tilde{L}}_\ell(\tilde{\boldsymbol{p}})$.

(iii) $\forall \tilde{\boldsymbol{p}} \in \operatorname{int}\tilde{\Delta}_n$, $\mathsf{H}\underline{\tilde{L}}_{\log}(\tilde{\boldsymbol{p}}) = -(X_{\boldsymbol{p}})^{-1}(\operatorname{diag}(\tilde{\boldsymbol{p}}))^{-1}$.

Proof [We show (i) and (ii)] Let $\boldsymbol{p} \in \operatorname{ri}\Delta_n$ and $f(\tilde{\boldsymbol{q}}) := \langle \boldsymbol{p}, \tilde{\ell}(\tilde{\boldsymbol{q}})\rangle = \langle \boldsymbol{p}, \nabla \underline{L}_\ell(\boldsymbol{q})\rangle$, where the equality is due to Lemma 26. Since $\underline{L}_\ell$ is twice differentiable $]0,+\infty[^n$, f is differentiable on $\operatorname{int}\tilde{\Delta}_n$ and we have $\mathsf{D}f(\tilde{\boldsymbol{q}}) = \langle \boldsymbol{p}, \mathsf{D}\tilde{\ell}(\tilde{\boldsymbol{p}})\rangle$. Since ℓ is proper, f reaches a minimum at $\tilde{\boldsymbol{p}} \in \operatorname{int}\Delta_n$, and thus $\langle \boldsymbol{p}, \mathsf{D}\tilde{\ell}(\tilde{\boldsymbol{p}})\rangle = \mathbf{0}_{\tilde{n}}^\mathsf{T}$ (this shows (i)). On the other hand, we have $\nabla \underline{\tilde{L}}_\ell(\tilde{\boldsymbol{p}}) = J_n^\mathsf{T}\nabla \underline{L}_\ell(\boldsymbol{p}) = J_n^\mathsf{T}\tilde{\ell}(\tilde{\boldsymbol{p}})$. By differentiating and using the chain the rule, we get $\mathsf{H}\underline{\tilde{L}}_\ell(\tilde{\boldsymbol{p}}) = [\mathsf{D}\tilde{\ell}(\tilde{\boldsymbol{p}})]^\mathsf{T}J_n$. This means that $\forall i \in [\tilde{n}]$, $[\mathsf{H}\underline{\tilde{L}}_\ell(\tilde{\boldsymbol{p}})]_{\cdot,i} = \nabla\tilde{\ell}_i(\tilde{\boldsymbol{p}}) - \nabla\tilde{\ell}_n(\tilde{\boldsymbol{p}})$, and thus $\sum_{i=1}^{\tilde{n}} p_i[\mathsf{H}\underline{\tilde{L}}_\ell(\tilde{\boldsymbol{p}})]_{\cdot,i} = \sum_{i=1}^{\tilde{n}} p_i\nabla\tilde{\ell}_i(\tilde{\boldsymbol{p}}) - (1-p_n)\nabla\tilde{\ell}_n(\tilde{\boldsymbol{p}})$. On the other hand, it follows from point (i) of the lemma that $\sum_{i=1}^{n} p_i\nabla\tilde{\ell}_i(\tilde{\boldsymbol{p}}) = \mathbf{0}_{\tilde{n}}$. Therefore, $[\mathsf{H}\underline{\tilde{L}}_\ell(\tilde{\boldsymbol{p}})]\tilde{\boldsymbol{p}} = -\nabla\tilde{\ell}_n(\tilde{\boldsymbol{p}})$ and, as a result, $\forall i \in [\tilde{n}]$, $[\mathsf{H}\underline{\tilde{L}}_\ell(\tilde{\boldsymbol{p}})]_{\cdot,i} - [\mathsf{H}\underline{\tilde{L}}_\ell(\tilde{\boldsymbol{p}})]\tilde{\boldsymbol{p}} = \nabla\tilde{\ell}_i(\tilde{\boldsymbol{p}})$. The last two equations can be combined as $\mathsf{D}\tilde{\ell}(\tilde{\boldsymbol{p}}) = \begin{bmatrix} X_{\boldsymbol{p}} \\ -\tilde{\boldsymbol{p}}^\mathsf{T}\end{bmatrix}\mathsf{H}\underline{\tilde{L}}_\ell(\tilde{\boldsymbol{p}})$.

[We show (iii)] It follows from (ii), since $\forall i \in [\tilde{n}]$, $\nabla[\tilde{\ell}_{\log}]_i(\tilde{\boldsymbol{p}}) = \frac{1}{p_i}\boldsymbol{e}_i$, for $\tilde{\boldsymbol{p}} \in \operatorname{int}\tilde{\Delta}_n$. ■

In the next lemma we state a new result for proper losses which will be crucial to prove a necessary condition for Φ-mixability (Theorem 16) — one of the main result of the paper.

Lemma 28 *Let $\ell\colon \Delta_n \to [0,+\infty]^n$ be a proper loss whose Bayes risk is twice differentiable on $]0,+\infty[^n$. For $\boldsymbol{v} \in \mathbb{R}^{n-1}$ and $\tilde{\boldsymbol{p}} \in \operatorname{int}\tilde{\Delta}_n$,*

$$\left\langle \boldsymbol{p}, (\mathsf{D}\tilde{\ell}(\tilde{\boldsymbol{p}})\boldsymbol{v}) \odot (\mathsf{D}\tilde{\ell}(\tilde{\boldsymbol{p}})\boldsymbol{v})\right\rangle = -\boldsymbol{v}^\mathsf{T}\mathsf{H}\underline{\tilde{L}}_\ell(\tilde{\boldsymbol{p}})[\mathsf{H}\underline{\tilde{L}}_{\log}(\tilde{\boldsymbol{p}})]^{-1}\mathsf{H}\underline{\tilde{L}}_\ell(\tilde{\boldsymbol{p}})\boldsymbol{v}, \tag{19}$$

where $\boldsymbol{p} = \Pi_n(\tilde{\boldsymbol{p}})$ and $\underline{L}_{\log}$ is the Bayes risk of the log *loss.*

Furthermore, if $t \mapsto \tilde{\boldsymbol{\alpha}}^t$ *is a smooth curve in* $\operatorname{int} \tilde{\Delta}_n$ *and satisfies* $\tilde{\boldsymbol{\alpha}}^0 = \tilde{\boldsymbol{p}}$ *and* $\frac{d}{dt}\tilde{\boldsymbol{\alpha}}^t\big|_{t=0} = \boldsymbol{v}$*, then* $t \mapsto \langle \boldsymbol{p}, \mathsf{D}\tilde{\ell}(\tilde{\boldsymbol{\alpha}}^t)\boldsymbol{v}\rangle$ *is differentiable at* 0 *and we have*

$$\frac{d}{dt}\left\langle \boldsymbol{p}, \mathsf{D}\tilde{\ell}(\tilde{\boldsymbol{\alpha}}^t)\boldsymbol{v}\right\rangle\bigg|_{t=0} = -\boldsymbol{v}^{\mathsf{T}}\mathsf{H}\underline{\tilde{L}}_\ell(\tilde{\boldsymbol{p}})\boldsymbol{v}. \tag{20}$$

Proof We know from Lemma 27 that for $\tilde{\boldsymbol{p}} \in \operatorname{int} \tilde{\Delta}_n$, we have $\mathsf{D}\tilde{\ell}(\tilde{\boldsymbol{p}}) = \begin{bmatrix} X_{\boldsymbol{p}} \\ -\tilde{\boldsymbol{p}}^{\mathsf{T}} \end{bmatrix}\mathsf{H}\underline{\tilde{L}}_\ell(\tilde{\boldsymbol{p}})$, where $X_{\boldsymbol{p}} = I_{n-1} - \boldsymbol{1}_{n-1}\tilde{\boldsymbol{p}}^{\mathsf{T}}$. Thus, we can write

$$\begin{aligned}\left\langle \boldsymbol{p}, \mathsf{D}\tilde{\ell}(\tilde{\boldsymbol{p}})\boldsymbol{v} \odot \mathsf{D}\tilde{\ell}(\tilde{\boldsymbol{p}})\boldsymbol{v}\right\rangle &= \boldsymbol{v}^{\mathsf{T}}(\mathsf{D}\tilde{\ell}(\tilde{\boldsymbol{p}}))^{\mathsf{T}} \operatorname{diag}(\boldsymbol{p})\, \mathsf{D}\tilde{\ell}(\tilde{\boldsymbol{p}})\boldsymbol{v}, \\ &= \boldsymbol{v}^{\mathsf{T}}(\mathsf{H}\underline{\tilde{L}}_\ell(\tilde{\boldsymbol{p}}))^{\mathsf{T}}[X_{\boldsymbol{p}}^{\mathsf{T}},\ -\tilde{\boldsymbol{p}}] \operatorname{diag}(\boldsymbol{p}) \begin{bmatrix} X_{\boldsymbol{p}} \\ -\tilde{\boldsymbol{p}}^{\mathsf{T}} \end{bmatrix} \mathsf{H}\underline{\tilde{L}}_\ell(\tilde{\boldsymbol{p}})\boldsymbol{v}. \end{aligned} \tag{21}$$

Observe that $[X_{\boldsymbol{p}}^{\mathsf{T}},\ -\tilde{\boldsymbol{p}}] \operatorname{diag}(\boldsymbol{p}) = [I_{n-1} - \tilde{\boldsymbol{p}}\boldsymbol{1}_{n-1}^{\mathsf{T}},\ -\tilde{\boldsymbol{p}}] \operatorname{diag}(\boldsymbol{p}) = [\operatorname{diag}(\tilde{\boldsymbol{p}}) - \tilde{\boldsymbol{p}}\tilde{\boldsymbol{p}}^{\mathsf{T}},\ -\tilde{\boldsymbol{p}}p_n]$. Therefore,

$$\begin{aligned}\left[X_{\boldsymbol{p}}^{\mathsf{T}},\quad -\tilde{\boldsymbol{p}}\right] \operatorname{diag}(\boldsymbol{p}) \begin{bmatrix} X_{\boldsymbol{p}} \\ -\tilde{\boldsymbol{p}}^{\mathsf{T}} \end{bmatrix} &= \left[\operatorname{diag}(\tilde{\boldsymbol{p}}) - \tilde{\boldsymbol{p}}\tilde{\boldsymbol{p}}^{\mathsf{T}},\quad -\tilde{\boldsymbol{p}}p_n\right] \begin{bmatrix} I_{n-1} - \boldsymbol{1}_{n-1}\tilde{\boldsymbol{p}}^{\mathsf{T}} \\ -\tilde{\boldsymbol{p}}^{\mathsf{T}} \end{bmatrix}, \\ &= \operatorname{diag}(\tilde{\boldsymbol{p}}) - \tilde{\boldsymbol{p}}\tilde{\boldsymbol{p}}^{\mathsf{T}} - \tilde{\boldsymbol{p}}\tilde{\boldsymbol{p}}^{\mathsf{T}} + \tilde{\boldsymbol{p}}\tilde{\boldsymbol{p}}^{\mathsf{T}}(1 - p_n) + p_n\tilde{\boldsymbol{p}}\tilde{\boldsymbol{p}}^{\mathsf{T}}, \\ &= \operatorname{diag}(\tilde{\boldsymbol{p}}) - \tilde{\boldsymbol{p}}\tilde{\boldsymbol{p}}^{\mathsf{T}}, \\ &= \operatorname{diag}(\tilde{\boldsymbol{p}})\, X_{\boldsymbol{p}}, \\ &= -(\mathsf{H}\underline{\tilde{L}}_{\log}(\tilde{\boldsymbol{p}}))^{-1}, \end{aligned} \tag{22}$$

where the last equality is due to Lemma 27. The desired result follows by combining (21) and (22).

[**We show** (20)] Let $\tilde{\boldsymbol{p}} \in \operatorname{int} \tilde{\Delta}_n$, we define $\tilde{\boldsymbol{\alpha}}^t := \tilde{\boldsymbol{p}} + t\boldsymbol{v}$, $\boldsymbol{\alpha}^t := \amalg_n(\tilde{\boldsymbol{\alpha}}^t) = \boldsymbol{p} + tJ_n\boldsymbol{v}$, and $r(t) := \boldsymbol{\alpha}^t / \left\|\boldsymbol{\alpha}^t\right\|$, where $t \in \{s : \tilde{\boldsymbol{p}} + s\boldsymbol{v} \in \operatorname{int} \tilde{\Delta}_n\}$. Since $t \mapsto r(t)$ is differentiable at 0 and $t \mapsto \mathsf{D}\tilde{\ell}(\tilde{\boldsymbol{\alpha}}^t)\boldsymbol{v}$ is continuous at 0, it follows from Lemma 23 that

$$\begin{aligned}\frac{d}{dt}\left\langle r(0), \mathsf{D}\tilde{\ell}(\tilde{\boldsymbol{\alpha}}^t)\boldsymbol{v}\right\rangle\bigg|_{t=0} &= \frac{d}{dt}\left\langle r(t), \mathsf{D}\tilde{\ell}(\tilde{\boldsymbol{\alpha}}^t)\boldsymbol{v}\right\rangle\bigg|_{t=0} - \left\langle \frac{d}{dt}r(t)\bigg|_{t=0}, \mathsf{D}\tilde{\ell}(\tilde{\boldsymbol{p}})\boldsymbol{v}\right\rangle, \\ &= -\left\langle \frac{d}{dt}r(t)\bigg|_{t=0}, \mathsf{D}\tilde{\ell}(\tilde{\boldsymbol{p}})\boldsymbol{v}\right\rangle, \end{aligned}$$

where the second equality holds since, according to Lemma 27, we have $\langle \boldsymbol{\alpha}^t, \mathsf{D}\tilde{\ell}(\tilde{\boldsymbol{\alpha}}^t)\boldsymbol{v}\rangle = 0$. Since $r(0) = \boldsymbol{p}/\left\|\boldsymbol{p}\right\|$, $\frac{d}{dt}r(t)\big|_{t=0} = \left\|\boldsymbol{p}\right\|^{-1}(I_n - r(0)[r(0)]^{\mathsf{T}})J_n\boldsymbol{v}$, and $J_n = \begin{bmatrix} I_{n-1} \\ -\boldsymbol{1}_{n-1}^{\mathsf{T}} \end{bmatrix}$, we get

$$\begin{aligned}\left\|\tilde{\boldsymbol{p}}\right\| \frac{d}{dt}\left\langle r(0), \mathsf{D}\tilde{\ell}(\tilde{\boldsymbol{\alpha}}^t)\boldsymbol{v}\right\rangle\bigg|_{t=0} &= -\left\langle \left(I_n - r(0)[r(0)]^{\mathsf{T}}\right) J_n\boldsymbol{v}, \mathsf{D}\tilde{\ell}(\tilde{\boldsymbol{p}})\boldsymbol{v}\right\rangle, \\ &= -\left\langle J_n\boldsymbol{v}, \mathsf{D}\tilde{\ell}(\tilde{\boldsymbol{p}})\boldsymbol{v}\right\rangle, \\ &= -\left\langle J_n\boldsymbol{v}, \begin{bmatrix} X_{\boldsymbol{p}} \\ -\tilde{\boldsymbol{p}}^{\mathsf{T}} \end{bmatrix} \mathsf{H}\underline{\tilde{L}}_\ell(\tilde{\boldsymbol{p}})\boldsymbol{v}\right\rangle, \\ &= -\boldsymbol{v}^{\mathsf{T}}\mathsf{H}\underline{\tilde{L}}_\ell(\tilde{\boldsymbol{p}})\boldsymbol{v}, \end{aligned} \tag{23}$$

where the passage to (23) is due to $r(0) = \boldsymbol{p}/\left\|\boldsymbol{p}\right\| \perp \mathsf{D}\tilde{\ell}(\tilde{\boldsymbol{p}})$. In the last equality we used the fact that $J_n^{\mathsf{T}}\begin{bmatrix} X_{\boldsymbol{p}} \\ -\tilde{\boldsymbol{p}}^{\mathsf{T}} \end{bmatrix} = [I_{n-1},\ -\mathbf{1}_{n-1}]\begin{bmatrix} I_{n-1} - \mathbf{1}_{n-1}\tilde{\boldsymbol{p}} \\ -\tilde{\boldsymbol{p}}^{\mathsf{T}} \end{bmatrix} = I_{n-1}$. ■

Lemma 29 *For $\eta > 0$, $\mathrm{S}_\eta := \eta^{-1}\,\mathrm{S}$ satisfies (8), where S is the Shannon entropy.*

Proof Let $(\hat{\boldsymbol{q}}, \boldsymbol{q}) \in (\operatorname{rbd}\Delta_{\mathcal{I}}) \times (\operatorname{ri}\Delta_{\mathcal{I}})$ and $\boldsymbol{q}^\lambda := \hat{\boldsymbol{q}} + \lambda(\boldsymbol{q} - \hat{\boldsymbol{q}})$, for $\lambda \in]0,1[$. Let $\mathcal{J} := \{j \in \mathcal{I} : \hat{q}_j \neq 0\}$ and $\mathcal{H} := \mathcal{I} \setminus \mathcal{J}$. We have

$$\mathrm{S}(\hat{\boldsymbol{q}}; \boldsymbol{q} - \hat{\boldsymbol{q}}) = \lim_{\lambda\downarrow 0} \lambda^{-1}\left[\sum\nolimits_{\theta\in\mathcal{I}} q_\theta^\lambda \log q_\theta^\lambda - \sum\nolimits_{\theta'\in\mathcal{J}} \hat{q}_{\theta'} \log \hat{q}_{\theta'}\right],$$
$$= \lim_{\lambda\downarrow 0} \lambda^{-1}\left[\sum\nolimits_{\theta\in\mathcal{J}} (q_\theta^\lambda \log q_\theta^\lambda - \hat{q}_\theta \log \hat{q}_\theta) + \sum\nolimits_{\theta'\in\mathcal{H}} q_{\theta'}^\lambda \log q_{\theta'}^\lambda\right]. \tag{24}$$

Observe that the limit of either summation term inside the bracket in (24) is equal to zero. Thus, using l'Hopital's rule we get

$$\mathrm{S}(\hat{\boldsymbol{q}}; \boldsymbol{q} - \hat{\boldsymbol{q}}) = \lim_{\lambda\downarrow 0}\left[\sum\nolimits_{\theta\in\mathcal{J}}[(q_\theta - \hat{q}_\theta)\log q_\theta^\lambda + (q_\theta - \hat{q}_\theta)] + \sum\nolimits_{\theta'\in\mathcal{H}}[q_{\theta'}\log q_{\theta'}^\lambda + q_{\theta'}]\right],$$
$$= \sum\nolimits_{\theta\in\mathcal{J}} (q_\theta - \hat{q}_\theta)\log \hat{q}_\theta + \sum\nolimits_{\theta'\in\mathcal{H}} q_{\theta'}\left[\lim_{\lambda\downarrow 0} \log q_{\theta'}^\lambda\right], \tag{25}$$

where in (25) we used the fact that $\sum_{\theta\in\mathcal{J}}(q_\theta - \hat{q}_\theta) + \sum_{\theta'\in\mathcal{H}} q_{\theta'} = 0$. Since for all $\theta' \in \mathcal{H}$, $\lim_{\lambda\downarrow 0} q_{\theta'}^\lambda = 0$, the right hand side of (25) is equal to $-\infty$. Therefore S satisfies (8). Since $\mathrm{S}_\eta = \eta^{-1}\,\mathrm{S}$, it is clear that S_η also satisfies (8).

■

Lemma 30 *Let $\Phi\colon \mathbb{R}^k \to \mathbb{R}\cup\{+\infty\}$ be an entropy. If Φ satisfies (8), then $\partial\tilde{\Phi}(\tilde{\boldsymbol{q}}) = \emptyset, \forall \tilde{\boldsymbol{q}} \in \operatorname{bd}\tilde{\Delta}_k$, and $\forall \boldsymbol{d} \in \mathbb{R}^k, \forall \boldsymbol{q} \in \operatorname{ri}\Delta_{\mathcal{I}},\ \mathsf{M}_\Phi(\boldsymbol{d}, \boldsymbol{q}) = \mathsf{M}_{\Phi^{\mathcal{I}}}(\Pi_{\mathcal{I}}^k \boldsymbol{d}, \Pi_{\mathcal{I}}^k \boldsymbol{q})$.*

Proof Let $\boldsymbol{\mu} \in \operatorname{rbd}\Delta_k$. Since Φ satisfies (8), it follows that $\forall \boldsymbol{q} \in \operatorname{ri}\Delta_k, \tilde{\Phi}(\tilde{\boldsymbol{\mu}}; \tilde{\boldsymbol{q}} - \tilde{\boldsymbol{\mu}}) = \Phi'(\boldsymbol{\mu}; \boldsymbol{q} - \boldsymbol{\mu}) = -\infty$. Therefore, $\partial\tilde{\Phi}'(\tilde{\boldsymbol{\mu}}) = \emptyset$ (Rockafellar, 1997, Thm. 23.4).

Let $\boldsymbol{d} \in \mathbb{R}^n, \mathcal{I} \subseteq [k]$, with $|\mathcal{I}| > 1$, and $\boldsymbol{q} \in \operatorname{ri}\Delta_{\mathcal{I}}$. Then

$$\mathsf{M}_{\Phi^{\mathcal{I}}}(\Pi_{\mathcal{I}}^k \boldsymbol{d}, \Pi_{\mathcal{I}}^k \boldsymbol{q}) = \inf_{\boldsymbol{\pi}\in\Delta_{|\mathcal{I}|}} \left\langle \boldsymbol{\pi}, \Pi_{\mathcal{I}}^k \boldsymbol{d}\right\rangle + D_{\Phi^{\mathcal{I}}}(\boldsymbol{\pi}, \Pi_{\mathcal{I}}^k \boldsymbol{q}),$$
$$= \inf_{\boldsymbol{\mu}\in\Delta_{\mathcal{I}}} \langle \boldsymbol{\mu}, \boldsymbol{d}\rangle + D_\Phi(\boldsymbol{\mu}, \boldsymbol{q}),$$
$$\leq \inf_{\boldsymbol{\mu}\in\Delta_k} \langle \boldsymbol{\mu}, \boldsymbol{d}\rangle + D_\Phi(\boldsymbol{\mu}, \boldsymbol{q}), \tag{26}$$
$$= \mathsf{M}_\Phi(\boldsymbol{d}, \boldsymbol{q}).$$

To complete the proof, we need to show that (26) holds with equality. For this, it suffice to prove that $\forall \boldsymbol{\mu} \in \Delta_k \setminus \Delta_{\mathcal{I}}$, $D_\Phi(\boldsymbol{\mu}, \boldsymbol{q}) = +\infty$. Let $\boldsymbol{\mu} \in \Delta_k \setminus \Delta_{\mathcal{I}}$ and $\mathcal{J} := \{\theta \in [k] : \mu_\theta \neq 0\} \cup \mathcal{I}$. In this case, we have $\boldsymbol{q} \in \operatorname{rbd}\Delta_{\mathcal{J}}$ and $\boldsymbol{q} + 2^{-1}(\boldsymbol{\mu} - \boldsymbol{q}) \in \operatorname{ri}\Delta_{\mathcal{J}}$. Thus, since Φ satisfies (8) and $\Phi'(\boldsymbol{q}; \cdot)$ is 1-homogeneous (Hiriart-Urruty and Lemaréchal, 2001, Prop. D.1.1.2), it follows that $2^{-1}\Phi'(\boldsymbol{q}; \boldsymbol{\mu} - \boldsymbol{q}) = \Phi'(\boldsymbol{q}; 2^{-1}(\boldsymbol{\mu} - \boldsymbol{q})) = -\infty$. Hence $D_\Phi(\boldsymbol{\mu}, \boldsymbol{q}) = +\infty$. ■

Lemma 31 *For $\boldsymbol{q} \in \Delta_k$, there exists a sequence $(\boldsymbol{d}_m)$ in $[0, +\infty[^k$ converging to $\boldsymbol{d} \in [0, +\infty]^k$ such that for any entropy $\Phi \colon \mathbb{R}^k \to \mathbb{R} \cup \{+\infty\}$, we have $\mathsf{M}_\Phi(\boldsymbol{d}_m, \boldsymbol{q}) \overset{m\to\infty}{\to} \mathsf{M}_\Phi(\boldsymbol{d}, \boldsymbol{q})$.*

Proof Let $\Phi \colon \mathbb{R}^k \to \mathbb{R} \cup \{+\infty\}$ be an entropy and let $\boldsymbol{d} \in [0, +\infty]^n$. Note that if $\mathcal{J} := \{\theta \in [k] : d_\theta < +\infty\} = \emptyset$, then by definition of M_Φ, it trivially holds that $\mathsf{M}_\Phi(\boldsymbol{d}_m, \boldsymbol{q}) \overset{m\to\infty}{\to} \mathsf{M}_\Phi(\boldsymbol{d}, \boldsymbol{q}) = +\infty$, for any sequence $(\boldsymbol{d}_m)$ such that $\boldsymbol{d}_m \overset{m\to\infty}{\to} \boldsymbol{d}$. Similarly, if $\mathcal{I} := \{\theta \in [k] : d_\theta \neq 0\} = \emptyset$, then the zero sequence gives the desired result. Assume now that $\mathcal{J}, \mathcal{I} \neq \emptyset$ and define the function $f_{\boldsymbol{\mu}} : \mathbb{R}^k \to \mathbb{R} \cup \{+\infty\}$ by

$$f_{\boldsymbol{\mu}}(\boldsymbol{z}) = \begin{cases} \sum_{\theta \in \mathcal{I}} \mu_i / z_i + D_\Phi(\boldsymbol{\mu}, \boldsymbol{q}), & \text{if } \boldsymbol{z} \in]0, +\infty[^k; \\ +\infty, & \text{otherwise.} \end{cases}$$

Since $\forall \boldsymbol{\mu} \in \Delta_k$, $f_{\boldsymbol{\mu}}$ is upper-semicontinuous and concave, $f_{\boldsymbol{\mu}}$ is a closed concave function. It follows that $f := \inf_{\boldsymbol{\mu} \in \Delta_k} f_{\boldsymbol{\mu}}$ is also closed and concave (Hiriart-Urruty and Lemaréchal, 2001, Prop. B.2.1.2). Let $\boldsymbol{b} \in [0, +\infty[^k$ be such that $b_\theta = \frac{1}{d_\theta}$, if $\theta \in \mathcal{I}$ (with the convention that $1/\infty = 0$); and $b_\theta = 0$ otherwise. Note that $f(\boldsymbol{b}) = \mathsf{M}_\Phi(\boldsymbol{d}, \boldsymbol{q})$. Let $\boldsymbol{z} \in \mathbb{R}^k$ be such that $z_\theta = b_\theta$, if $\theta \in \mathcal{J}$; and $z_\theta = 1$ otherwise. Note that $\boldsymbol{z} \in \operatorname{int} \operatorname{dom} f$. Thus, since f is closed and concave, we have $f(\boldsymbol{b}) = \lim_{\lambda \downarrow 0} f(\boldsymbol{b} + \lambda(\boldsymbol{z} - \boldsymbol{b}))$ (ibid., Prop. B.1.2.5). Choosing $\boldsymbol{d}_m \in [0, +\infty[^n$ such that $[\boldsymbol{d}_m]_\theta = d_\theta + 1/m(z_\theta^{-1} - d_\theta)$, if $\theta \in \mathcal{I}$; and 0 otherwise, gives the desired sequence. ■

Appendix C. Proofs of Results in the Main Body

C.1. Proof of Proposition 1

Proof Given $\boldsymbol{v} \in \mathbb{R}^{k-1}$, we first derive the expression of the Fenchel dual $\tilde{\mathrm{S}}^*(\boldsymbol{v}) := \sup_{\tilde{\boldsymbol{q}} \in \tilde{\Delta}_k} \langle \tilde{\boldsymbol{q}}, \boldsymbol{v} \rangle - \tilde{\mathrm{S}}(\tilde{\boldsymbol{q}})$. Setting the gradient of $\tilde{\boldsymbol{q}} \mapsto \langle \tilde{\boldsymbol{q}}, \boldsymbol{v} \rangle - \tilde{\mathrm{S}}(\tilde{\boldsymbol{q}})$ to $\boldsymbol{0}_{\tilde{k}}$ gives $\boldsymbol{v} = \nabla \tilde{\mathrm{S}}(\tilde{\boldsymbol{q}})$. For $\boldsymbol{q} \in]0, +\infty[^k$, we have $\nabla \mathrm{S}(\boldsymbol{q}) = \log \boldsymbol{q} + \boldsymbol{1}_k$, and from §2.1 we know that $\nabla \tilde{\mathrm{S}}(\tilde{\boldsymbol{q}}) = J_k^\mathsf{T} \nabla \mathrm{S}(\boldsymbol{q})$. Therefore,

$$\boldsymbol{v} = \nabla \tilde{\mathrm{S}}(\tilde{\boldsymbol{q}}) \implies \boldsymbol{v} = J_k^\mathsf{T} \nabla \mathrm{S}(\boldsymbol{q}) \implies \boldsymbol{v} = \log \frac{\tilde{\boldsymbol{q}}}{q_k},$$

where the right most equality is equivalent to $\tilde{\boldsymbol{q}}/q_k = \exp(\boldsymbol{v})$. Since $\langle \tilde{\boldsymbol{q}}, \boldsymbol{1}_{\tilde{k}} \rangle = 1 - q_k$, we get $q_k = (\langle \exp(\boldsymbol{v}), \boldsymbol{1}_{\tilde{k}} \rangle + 1)^{-1}$. Therefore, the supremum in the definition of $\tilde{\mathrm{S}}^*(\boldsymbol{v})$ is attained at $\tilde{\boldsymbol{q}}_* = \exp(\boldsymbol{v})(\langle \exp(\boldsymbol{v}), \boldsymbol{1}_{\tilde{k}} \rangle + 1)^{-1}$. Hence $\tilde{\mathrm{S}}^*(\boldsymbol{v}) = \langle \tilde{\boldsymbol{q}}_*, \boldsymbol{v} \rangle - \langle \tilde{\boldsymbol{q}}_*, \log \tilde{\boldsymbol{q}}_* \rangle = \log(\langle \exp(\boldsymbol{v}), \boldsymbol{1}_{\tilde{k}} \rangle + 1)$. Finally, using (2) we get $\mathrm{S}^\star(\boldsymbol{z}) = \log \langle \exp(\boldsymbol{z}), \boldsymbol{1}_k \rangle$, for $\boldsymbol{z} \in \mathbb{R}^k$. ■

C.2. Proof of Theorem 3

Proof We will construct a proper support loss $\underline{\ell}$ of ℓ.

Let $\boldsymbol{p} \in \operatorname{ri} \Delta_n$ ($-\boldsymbol{p} \in \operatorname{int} \operatorname{dom} \sigma_{\mathcal{S}_\ell^\oplus}$). Since the support function of a non-empty set is closed and convex, we have $\sigma_{\mathcal{S}_\ell^\oplus}^{**} = \sigma_{\mathcal{S}_\ell^\oplus}$ (Hiriart-Urruty and Lemaréchal, 2001, Prop. C.2.1.2). Pick any $\boldsymbol{v} \in \partial \sigma_{\mathcal{S}_\ell^\oplus}(-\boldsymbol{p}) = \partial \sigma_{\mathcal{S}_\ell^\oplus}^{**}(-\boldsymbol{p}) \neq \emptyset$. Since $\sigma_{\mathcal{S}_\ell^\oplus}^* = \iota_{\mathcal{S}_\ell^\oplus}$ (Rockafellar, 1997), we can apply Proposition 20-(iv) with f replaced by $\sigma_{\mathcal{S}_\ell^\oplus}^*$ to obtain $\langle -\boldsymbol{p}, \boldsymbol{v} \rangle = \sigma_{\mathcal{S}_\ell^\oplus}(-\boldsymbol{p}) + \iota_{\mathcal{S}_\ell^\oplus}(\boldsymbol{v})$. The fact that

$\langle -\boldsymbol{p}, \boldsymbol{v}\rangle$ and $\sigma_{\mathcal{S}_\ell^\oplus}(-\boldsymbol{p})$ are both finite implies that $\iota_{\mathcal{S}_\ell^\oplus}(\boldsymbol{v}) = 0$. Therefore, $\boldsymbol{v} \in \mathcal{S}_\ell^\oplus$ and $\langle \boldsymbol{p}, \boldsymbol{v}\rangle = -\sigma_{\mathcal{S}_\ell^\oplus}(-\boldsymbol{p}) = \underline{L}_\ell(\boldsymbol{p})$. Define $\underline{\ell}(\boldsymbol{p}) := \boldsymbol{v} \in \mathcal{S}_\ell^\oplus$.

Now let $\boldsymbol{p} \in \operatorname{rbd} \Delta_n$ and $\boldsymbol{q} := \mathbf{1}_n/n$. Since the $\underline{L}_\ell$ is a closed concave function and $\boldsymbol{q} \in \operatorname{int} \operatorname{dom} \underline{L}_\ell$, it follows that $\underline{L}_\ell(\boldsymbol{p}+m^{-1}(\boldsymbol{q}-\boldsymbol{p})) \overset{m\to\infty}{\to} \underline{L}_\ell(\boldsymbol{p})$ (Hiriart-Urruty and Lemaréchal, 2001, Prop. B.1.2.5). Note that $\boldsymbol{q}_m := \boldsymbol{p}+m^{-1}(\boldsymbol{q}-\boldsymbol{p}) \in \operatorname{ri} \Delta_n, \forall m \in \mathbb{N}$. Now let $v_{x,m} := \underline{\ell}_x(\boldsymbol{q}_m)$, where $\underline{\ell}(\boldsymbol{q}_m)$ is as constructed in the previous paragraph. If $(v_{1,m})$ is bounded [resp. unbounded], we can extract a subsequence $(v_{1,\varphi_1(m)})$ which converges [resp. diverges to $+\infty$], where $\varphi_1 : \mathbb{N} \to \mathbb{N}$ is an increasing function. By repeating this process for $(v_{2,\varphi_1(m)})$ and so on, we can construct an increasing function $\varphi := \varphi_n \circ \cdots \circ \varphi_1 : \mathbb{N} \to \mathbb{N}$, such that $\boldsymbol{v}_m := [v_{x,\varphi(m)}]^\mathsf{T}_{x\in[n]}$ has a well defined (coordinate-wise) limit in $[0, +\infty]^n$. Define $\underline{\ell}(\boldsymbol{p}) := \lim_{m\to\infty} \boldsymbol{v}_m$. By continuity of the inner product, we have

$$
\begin{aligned}
\langle \boldsymbol{p}, \underline{\ell}(\boldsymbol{p})\rangle = \lim_{m\to\infty} \langle \boldsymbol{q}_{\varphi(m)}, \boldsymbol{v}_m\rangle &= \lim_{m\to\infty} \langle \boldsymbol{q}_{\varphi(m)}, \underline{\ell}(\boldsymbol{q}_{\varphi(m)})\rangle, \\
&= \lim_{m\to\infty} \underline{L}_\ell(\boldsymbol{q}_{\varphi(m)}) = \underline{L}_\ell(\boldsymbol{p}).
\end{aligned}
$$

By construction, $\forall m \in \mathbb{N}, \boldsymbol{p}_m := \boldsymbol{q}_{\varphi(m)} \in \operatorname{ri} \Delta_n$ and $\underline{\ell}(\boldsymbol{p}_m) = \boldsymbol{v}_m \overset{m\to\infty}{\to} \underline{\ell}(\boldsymbol{p})$. Therefore, $\underline{\ell}$ is support loss of ℓ.

It remains to show that it is proper; that is $\forall \boldsymbol{p} \in \Delta_n, \forall \boldsymbol{q} \in \Delta_n$, $\langle \boldsymbol{p}, \underline{\ell}(\boldsymbol{p})\rangle \leq \langle \boldsymbol{p}, \underline{\ell}(\boldsymbol{q})\rangle$. Let $\boldsymbol{q} \in \operatorname{ri} \Delta_n$. We just showed that $\forall \boldsymbol{p} \in \Delta_n, \langle \boldsymbol{p}, \underline{\ell}(\boldsymbol{p})\rangle = \underline{L}_\ell(\boldsymbol{p})$ and that $\underline{\ell}(\boldsymbol{q}) \in \mathcal{S}_\ell^\oplus$. Using the fact that $\underline{L}_\ell(\boldsymbol{p}) = \inf_{\boldsymbol{z}\in\mathcal{S}_\ell^\oplus} \langle \boldsymbol{p}, \boldsymbol{z}\rangle$, we obtain $\langle \boldsymbol{p}, \underline{\ell}(\boldsymbol{p})\rangle \leq \langle \boldsymbol{p}, \underline{\ell}(\boldsymbol{q})\rangle$.

Now let $\boldsymbol{q} \in \operatorname{rbd} \Delta_k$. Since $\underline{\ell}$ is a support loss, we know that there exists a sequence $(\boldsymbol{q}_m) \subset \operatorname{ri} \Delta_n$ such that $\underline{\ell}(\boldsymbol{q}_m) \overset{m\to\infty}{\to} \underline{\ell}(\boldsymbol{q})$. But as we established in the previous paragraph, $\langle \boldsymbol{p}, \underline{\ell}(\boldsymbol{p})\rangle \leq \langle \boldsymbol{p}, \underline{\ell}(\boldsymbol{q}_m)\rangle$. By passing to the limit $m \to \infty$, we obtain $\langle \boldsymbol{p}, \underline{\ell}(\boldsymbol{p})\rangle \leq \langle \boldsymbol{p}, \underline{\ell}(\boldsymbol{q})\rangle$. Therefore $\underline{\ell}$ is a proper loss with Bayes risk $\underline{L}_\ell$. ■

C.3. Proof of Theorem 4

For a set $\mathcal{C}$, we denote $\operatorname{co} \mathcal{C}$ and $\overline{\operatorname{co}}\mathcal{C}$ its *convex hull* and *closed convex hull*, respectively.

Definition 32 ((Hiriart-Urruty and Lemaréchal, 2001)) *Let $\mathcal{C}$ be non-empty convex set in $\mathbb{R}^n$. We say that $\boldsymbol{u} \in \mathcal{C}$ is an extreme point of $\mathcal{C}$ if there are no two different points $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$ in $\mathcal{C}$ and $\lambda \in]0, 1[$ such that $\boldsymbol{u} = \lambda\boldsymbol{u}_1 + (1-\lambda)\boldsymbol{u}_2$.*

We denote the set of extreme points of a set $\mathcal{C}$ by $\operatorname{ext} \mathcal{C}$.

Lemma 33 *Let $\ell\colon \mathcal{A} \to [0, +\infty]^n$ be a closed loss. Then $\operatorname{ext} \overline{\operatorname{co}}\mathcal{S}_\ell^\oplus \subseteq \mathcal{S}_\ell$.*

Proof Since $\operatorname{co} \mathcal{S}_\ell^\oplus \subseteq \mathbb{R}^n$ is connected, $\operatorname{co} \mathcal{S}_\ell^\oplus = \{\boldsymbol{v} + \sum_{k=1}^n \alpha_k \ell(\boldsymbol{a}_k)\colon (\boldsymbol{a}_{k\in[n]}, \boldsymbol{\alpha}, \boldsymbol{v}) \in \mathcal{A}^n \times \Delta_n \times [0, +\infty[^n\}$ (Hiriart-Urruty and Lemaréchal, 2001, Prop. A.1.3.7).

We claim that $\overline{\operatorname{co}}\mathcal{S}_\ell^\oplus = \operatorname{co} \mathcal{S}_\ell^\oplus$. Let $(\boldsymbol{z}_m) := (\boldsymbol{v}_m + \sum_{k=1}^n \alpha_{m,k}\ell(\boldsymbol{a}_{m,k}))$ be a convergent sequence in $[0, +\infty[^n$, where $(\boldsymbol{\alpha}_m)$, $([\boldsymbol{a}_{m,k}]_{k\in[n]})$ and $(\boldsymbol{v}_m)$ are sequences in Δ_n, $\mathcal{A}^n$, and $[0, +\infty[^n$, respectively. Since Δ_n is compact, we may assume, by extraction a subsequence if necessary, that $\boldsymbol{\alpha}_m \overset{m\to\infty}{\to} \boldsymbol{\alpha}^* \in \Delta_n$. Let $\mathcal{K} := \{k \in [n] : \alpha_k^* \neq 0\}$. Since $\boldsymbol{z}_m$ converges, $([[\ell(\boldsymbol{a}_{m,k})]_{k\in\mathcal{K}}, \boldsymbol{v}_m])$

is a bounded sequence in $[0,+\infty[^{n|\mathcal{K}|+n}$. Since ℓ is closed, we may assume, by extraction a subsequence if necessary, that $\forall k \in \mathcal{K}$, $\ell(\boldsymbol{a}_{m,k}) \overset{m\to\infty}{\to} \ell(\boldsymbol{a}_k^*)$ and $\boldsymbol{v}_m \overset{m\to\infty}{\to} \boldsymbol{v}^*$, where $[\boldsymbol{a}_k^*]_{k\in\mathcal{K}} \in \mathcal{A}^{|\mathcal{K}|}$ and $\boldsymbol{v}^* \in [0,+\infty[^n$. Consequently,

$$\begin{aligned}\boldsymbol{v}^* + \sum_{k=1}^{n} \alpha_k^* \ell(\boldsymbol{a}_k^*) &= \lim_{m\to\infty}\left[\boldsymbol{v}_{m,k} + \sum_{k\in\mathcal{K}} \alpha_{m,k}\ell(\boldsymbol{a}_{m,k})\right],\\ &\leq \lim_{m\to\infty} \boldsymbol{z}_m,\end{aligned}$$

where the last inequality is coordinate-wise. Therefore, there exists $\boldsymbol{v}' \in [0,+\infty[^n$ such that $\lim_{m\to\infty}\boldsymbol{z}_m = \boldsymbol{v}' + \boldsymbol{v}^* + \sum_{k=1}^{n}\alpha_k^*\ell(\boldsymbol{a}_k^*) \in \operatorname{co}\mathcal{S}_\ell^\oplus$. This shows that $\overline{\operatorname{co}}\mathcal{S}_\ell^\oplus \subset \operatorname{co}\mathcal{S}_\ell^\oplus$, and thus $\overline{\operatorname{co}}\mathcal{S}_\ell^\oplus = \operatorname{co}\mathcal{S}_\ell^\oplus$ which proves our first claim.

By definition of an extreme point, $\operatorname{ext}\overline{\operatorname{co}}\mathcal{S}_\ell^\oplus \subseteq \overline{\operatorname{co}}\mathcal{S}_\ell^\oplus$. Let $\boldsymbol{e} \in \operatorname{ext}\overline{\operatorname{co}}\mathcal{S}_\ell^\oplus$ and $(\boldsymbol{a}_{k\in[n]}, \boldsymbol{\alpha}, \boldsymbol{v}) \in \mathcal{A}^n \times \Delta_n \times [0,+\infty[^n$ such that $\boldsymbol{e} = \sum_{k=1}^{n}\alpha_k\ell(\boldsymbol{a}_k) + \boldsymbol{v}$. If there exists $i,j\in[n]$ such that $\alpha_i\alpha_j \neq 0$ or $\alpha_i v_j \neq 0$ then $\boldsymbol{e}$ would violate the definition of an extreme point. Therefore, the only possible extreme points are of the form $\{\ell(\boldsymbol{a}) : \boldsymbol{a} \in \operatorname{dom}\ell\} = \mathcal{S}_\ell$. ■

Proof [**Theorem 4**] Suppose $\underline{L}_\ell$ is not differentiable at $\boldsymbol{p} \in \operatorname{ri}\Delta_n$. Then from the definition of the Bayes risk, $\sigma_{\mathcal{S}_\ell^\oplus}$ is not differentiable at $-\boldsymbol{p}$. This implies that $\mathcal{F}(\boldsymbol{p}) := \partial\sigma_{\mathcal{S}_\ell^\oplus}(-\boldsymbol{p})$ has more than one element (Hiriart-Urruty and Lemaréchal, 2001, Cor. D.2.1.4). Since $\sigma_{\mathcal{S}_\ell^\oplus} = \sigma_{\overline{\operatorname{co}}\mathcal{S}_\ell^\oplus}$ (ibid.. Prop. C.2.2.1), $\mathcal{F}(\boldsymbol{p}) = \partial\sigma_{\overline{\operatorname{co}}\mathcal{S}_\ell^\oplus}(-\boldsymbol{p})$ is a subset of $\overline{\operatorname{co}}\mathcal{S}_\ell^\oplus$ and every extreme point of $\mathcal{F}(\boldsymbol{p})$ is also an extreme point of $\overline{\operatorname{co}}\mathcal{S}_\ell^\oplus$ (ibid., Prop. A.2.3.7). Thus, from Lemma 33, we have $\operatorname{ext}\mathcal{F}(\boldsymbol{p}) \subset \mathcal{S}_\ell$. On the other hand, since $-\boldsymbol{p} \in \operatorname{int}\operatorname{dom}\sigma_{\mathcal{S}_\ell^\oplus}$, $\mathcal{F}(\boldsymbol{p})$ is a compact, convex set (Rockafellar, 1997, Thm. 23.4), and thus $\mathcal{F}(\boldsymbol{p}) = \operatorname{co}(\operatorname{ext}\mathcal{F}(\boldsymbol{p}))$ (Hiriart-Urruty and Lemaréchal, 2001, Thm. A.2.3.4). Hence, the fact that $\mathcal{F}(\boldsymbol{p})$ has more than one element implies that there exists $\boldsymbol{a}_0, \boldsymbol{a}_1 \in \mathcal{A}$ such that $\ell(\boldsymbol{a}_0), \ell(\boldsymbol{a}_1) \in \operatorname{ext}\mathcal{F}(\boldsymbol{p}) \subseteq \mathcal{F}(\boldsymbol{p})$ and $\ell(\boldsymbol{a}_0) \neq \ell(\boldsymbol{a}_1)$. Since $\mathcal{F}(\boldsymbol{p}) = \partial\sigma_{\mathcal{S}_\ell^\oplus}(-\boldsymbol{p})$, Proposition 20-(iv) and the fact that $\sigma^*_{\mathcal{S}_\ell^\oplus} = \iota_{\mathcal{S}_\ell^\oplus}$ imply that $\underline{L}_\ell(\boldsymbol{p}) = \langle\boldsymbol{p}, \underline{\ell}(\boldsymbol{p})\rangle = \langle\boldsymbol{p}, \ell(\boldsymbol{a}_0)\rangle = \langle\boldsymbol{p}, \ell(\boldsymbol{a}_1)\rangle$.

Let $\boldsymbol{p} \in \operatorname{ri}\Delta_n$ and suppose that $\underline{L}_\ell$ is differentiable at $\boldsymbol{p}$. In this case, $\sigma_{\mathcal{S}_\ell^\oplus}$ is differentiable at $-\boldsymbol{p}$, which implies that $\mathcal{F}(\boldsymbol{p}) = \partial\sigma_{\mathcal{S}_\ell^\oplus}(-\boldsymbol{p})$ is the singleton $\{\nabla\sigma_{\mathcal{S}_\ell^\oplus}(-\boldsymbol{p})\}$ (ibid., Cor. D.2.1.4). In this case, $\underline{\ell}(\boldsymbol{p}) = \nabla\sigma_{\mathcal{S}_\ell^\oplus}(-\boldsymbol{p})$ is the only extreme point of $\mathcal{F}(\boldsymbol{p}) \subset \overline{\operatorname{co}}\mathcal{S}_\ell^\oplus$. From Lemma 33, there exists $\boldsymbol{a}_* \in \mathcal{A}$ such that $\ell(\boldsymbol{a}_*) = \underline{\ell}(\boldsymbol{p})$. In this paragraph, we showed the following

$$\forall \boldsymbol{p} \in \operatorname{ri}\Delta_n, \exists \boldsymbol{a}_* \in \operatorname{dom}\ell,\ \ell(\boldsymbol{a}_*) = \underline{\ell}(\boldsymbol{p}). \tag{27}$$

For the rest of this proof we will assume that $\underline{L}_\ell$ is differentiable on $]0,+\infty[^n$. Let $\boldsymbol{p} \in \operatorname{rbd}\Delta_n \cap \operatorname{dom}\underline{\ell}$. Since $\underline{\ell}$ is a support loss, there exists $(\boldsymbol{p}_m)$ in $\operatorname{ri}\Delta_n$ such that $(\underline{\ell}(\boldsymbol{p}_m))_m$ converges to $\underline{\ell}(\boldsymbol{p})$. From (27) it holds that $\forall \boldsymbol{p}_m \in \operatorname{ri}\Delta_n, \exists \boldsymbol{a}_m \in \mathcal{A}, \ell(\boldsymbol{a}_m) = \underline{\ell}(\boldsymbol{p}_m)$. Since $(\ell(\boldsymbol{a}_m))_m$ converges and ℓ is closed, there exists $\boldsymbol{a}_* \in \mathcal{A}$ such that $\ell(\boldsymbol{a}_*) = \lim_{m\to\infty}\ell(\boldsymbol{a}_m) = \underline{\ell}(\boldsymbol{p})$.

Now let $\boldsymbol{a} \in \operatorname{dom}\ell$ and $f(\boldsymbol{p},x) := \underline{\ell}_x(\boldsymbol{p}) - \ell_x(\boldsymbol{a})$. Since $\ell(\boldsymbol{a}) \in \mathcal{S}_\ell^\oplus$ and $\underline{\ell}$ is proper, we have for all $\boldsymbol{p} \in \operatorname{ri}\Delta_n, \mathbb{E}_{x\sim\boldsymbol{p}}[f(\boldsymbol{p},x)] \leq 0$ and $-\infty < f(\boldsymbol{p},x), \forall x \in [n]$. Therefore, Lemma 24 implies that for all $m \in \mathbb{N}\setminus\{0\}$ there exists $\boldsymbol{p}_m \in \operatorname{ri}\Delta_n$, such that $\forall x \in [n], \underline{\ell}_x(\boldsymbol{p}_m) \leq \ell_x(\boldsymbol{a}) + 1/m$. On one hand, since $(\underline{\ell}(\boldsymbol{p}_m))$ is bounded (from the previous inequality), we may assume by extracting a subsequence if necessary, that $(\underline{\ell}(\boldsymbol{p}_m))_m$ converges. On the other hand, since $\boldsymbol{p}_m \in \operatorname{ri}\Delta_n$, (27) implies that there exists $\boldsymbol{a}_m \in \operatorname{dom}\ell$ such that $\underline{\ell}(\boldsymbol{p}_m) = \ell(\boldsymbol{a}_m)$. Since ℓ is closed and $(\ell(\boldsymbol{a}_m))_m$ converges, there exists $\boldsymbol{a}_* \in \mathcal{A}$, such that $\ell(\boldsymbol{a}_*) = \lim_{m\to\infty}\ell(\boldsymbol{a}_m) =$

$\lim_{m\to\infty} \underline{\ell}(\boldsymbol{p}_m) \leq \ell(\boldsymbol{a})$. But since ℓ is admissible, the latter component-wise inequality implies that $\ell(\boldsymbol{a}_*) = \ell(\boldsymbol{a}) = \lim_{m\to\infty} \underline{\ell}(\boldsymbol{p})$. ■

C.4. Proof of Theorem 5

Proof $[\mathcal{S}_\ell^\oplus \subseteq \bigcap_{\boldsymbol{p}\in\Delta_n} \mathcal{H}_{-\boldsymbol{p},-\underline{L}_\ell(\boldsymbol{p})}]$: Let $\boldsymbol{v} \in \mathcal{S}_\ell^\oplus$, $\boldsymbol{u} \in [0,+\infty[^n$, and $\boldsymbol{q} \in \Delta_n$ such that $\boldsymbol{v} = \ell(\boldsymbol{q})+\boldsymbol{u}$. Since ℓ is proper then $\forall \boldsymbol{p} \in \Delta_n, \underline{L}_\ell(\boldsymbol{p}) = \langle \boldsymbol{p}, \ell(\boldsymbol{p})\rangle \leq \langle \boldsymbol{p}, \ell(\boldsymbol{q})\rangle \leq \langle \boldsymbol{p}, \ell(\boldsymbol{q})+\boldsymbol{u}\rangle = \langle \boldsymbol{p}, \boldsymbol{v}\rangle$. Therefore, $\boldsymbol{v} \in \bigcap_{\boldsymbol{p}\in\Delta_n} \mathcal{H}_{-\boldsymbol{p},-\underline{L}_\ell(\boldsymbol{p})}$.

$[\bigcap_{\boldsymbol{p}\in\Delta_n} \mathcal{H}_{-\boldsymbol{p},-\underline{L}_\ell(\boldsymbol{p})} \subseteq \mathcal{S}_\ell^\oplus]$: Let $\boldsymbol{v} \in \bigcap_{\boldsymbol{p}\in\Delta_n} \mathcal{H}_{-\boldsymbol{p},-\underline{L}_\ell(\boldsymbol{p})}$. Let $\Omega = [n]$, $\Delta(\Omega) = \Delta_n$, and $Q(\boldsymbol{p}, x) = \ell_x(\boldsymbol{p}) - v_x$ for all $(\boldsymbol{p}, x) \in \Delta_n \times [n]$. Since $\boldsymbol{v} \in \bigcap_{\boldsymbol{p}\in\Delta_n} \mathcal{H}_{-\boldsymbol{p},-\underline{L}_\ell(\boldsymbol{p})}$, $\mathbb{E}_{x\sim\boldsymbol{p}} Q(\boldsymbol{p}, x) = \langle \boldsymbol{p}, \ell(\boldsymbol{p})\rangle - \langle \boldsymbol{p}, \boldsymbol{v}\rangle \leq 0$ for all $\boldsymbol{p} \in \Delta_n$. Lemma 21, implies that there exists $\boldsymbol{p}_* \in \Delta_n$ such that $Q(\boldsymbol{p}_*, x) = \ell_x(\boldsymbol{p}_*) - v_x \leq 0$, for all $x \in [n]$. This shows that $\boldsymbol{v} \in \mathcal{S}_\ell^\oplus$. ■

C.5. Proof of Theorem 7

Proof Let $\eta \coloneqq \underline{\eta_\ell}$. We will show that $\exp(-\eta\mathcal{S}_\ell^\oplus)$ is convex, which will imply that ℓ is η-mixable (Chernov et al., 2010).

Since $\underline{\eta_\ell} = \inf_{\tilde{\boldsymbol{p}}\in\operatorname{int}\tilde{\Delta}_n} (\lambda_{\max}([\mathsf{H}\underline{\tilde{L}}_{\log}(\tilde{\boldsymbol{p}})]^{-1}\mathsf{H}\underline{\tilde{L}}_\ell(\tilde{\boldsymbol{p}})))^{-1} > 0$, $\eta\underline{L}_\ell - \underline{L}_{\log}$ is convex on $\operatorname{ri}\Delta_n$ (van Erven et al., 2012, Thm. 10). Let $\boldsymbol{p} \in \operatorname{ri}\Delta_n$ and define

$$\Lambda(\boldsymbol{r}) \coloneqq \underline{L}_{\log}(\boldsymbol{r}) + \langle \boldsymbol{r}, \eta\underline{\ell}(\boldsymbol{p}) - \underline{\ell}_{\log}(\boldsymbol{p})\rangle,\ \boldsymbol{r} \in \operatorname{ri}\Delta_n.$$

Since Λ is equal to $\underline{L}_{\log}$ plus an affine function, it follows that $\eta\underline{L}_\ell - \Lambda$ is also convex on $\operatorname{ri}\Delta_n$. On one hand, since $\underline{\ell}$ and $\underline{\ell}_{\log}$ are proper losses, we have $\langle \boldsymbol{p}, \underline{\ell}(\boldsymbol{p})\rangle = \underline{L}_\ell(\boldsymbol{p})$ and $\langle \boldsymbol{p}, \underline{\ell}_{\log}(\boldsymbol{p})\rangle = \underline{L}_{\log}(\boldsymbol{p})$ which implies that

$$\eta\underline{L}_\ell(\boldsymbol{p}) - \Lambda(\boldsymbol{p}) = 0. \tag{28}$$

On the other hand, since $\underline{L}_\ell$ and $\underline{L}_{\log}$ are differentiable we have $\underline{\ell}(\boldsymbol{p}) = \nabla\underline{L}_\ell(\boldsymbol{p})$ and $\nabla\underline{L}_{\log}(\boldsymbol{p}) = \underline{\ell}_{\log}(\boldsymbol{p})$, which yields $\eta\nabla\underline{L}_\ell(\boldsymbol{p}) - \nabla\Lambda(\boldsymbol{p}) = \boldsymbol{0}_n$. This implies that $\eta\underline{L}_\ell - \Lambda$ attains a minimum at $\boldsymbol{p}$ (Hiriart-Urruty and Lemaréchal, 2001, Thm. D.2.2.1). Combining this fact and (28) gives $\eta\underline{L}_\ell(\boldsymbol{r}) \geq \Lambda(\boldsymbol{r}), \forall \boldsymbol{r} \in \operatorname{ri}\Delta_n$. By Proposition 20-(iii), $-\eta\underline{L}_\ell \leq -\Lambda$ implies

$$[-\eta\underline{L}_\ell]^* \geq [-\Lambda]^*. \tag{29}$$

Using Proposition 20-(ii), we get $[-\Lambda^*](\boldsymbol{s}) = [-\underline{L}_{\log}]^*(\boldsymbol{s} - \underline{\ell}_{\log}(\boldsymbol{p}) + \eta\underline{\ell}(\boldsymbol{p}))$ for $\boldsymbol{s} \in \mathbb{R}^n$. Since $-\eta\underline{L}_\ell(\boldsymbol{u}) = -\underline{L}_\ell(\eta\boldsymbol{u}) = \sigma_{\mathcal{S}_\ell^\oplus}(-\eta\boldsymbol{u})$ and $\sigma^*_{\mathcal{S}_\ell^\oplus} = \iota_{\mathcal{S}_\ell^\oplus}$, Proposition 20-(v) implies $[-\eta\underline{L}_\ell]^*(\boldsymbol{s}) = \iota_{\mathcal{S}_\ell^\oplus}(-\boldsymbol{s}/\eta)$. Similarly, we have $[-\underline{L}_{\log}]^*(\boldsymbol{s}) = \iota_{\mathcal{S}_{\log}^\oplus}(-\boldsymbol{s})$. Therefore, (29) implies

$$\forall \boldsymbol{s} \in \mathbb{R}^n, \quad \iota_{\mathcal{S}_\ell^\oplus}(-\boldsymbol{s}/\eta) \geq \iota_{\mathcal{S}_{\log}^\oplus}(-\boldsymbol{s} + \underline{\ell}_{\log}(\boldsymbol{p}) - \eta\underline{\ell}(\boldsymbol{p})).$$

This inequality implies that if $\boldsymbol{s} \in -\eta\mathcal{S}_\ell^\oplus$, then $\boldsymbol{s} \in -\mathcal{S}_{\log}^\oplus + \underline{\ell}_{\log}(\boldsymbol{p}) - \eta\underline{\ell}(\boldsymbol{p})$. In particular, if $\boldsymbol{u} \in e^{-\eta\mathcal{S}_\ell^\oplus}$ then

$$\boldsymbol{u} \in e^{-\mathcal{S}_{\log}^\oplus + \underline{\ell}_{\log}(\boldsymbol{p}) - \eta\underline{\ell}(\boldsymbol{p})} \subseteq \mathcal{H}_{\tau(\boldsymbol{p}),1} = \{\boldsymbol{v} \in \mathbb{R}^n : \langle \boldsymbol{v}, \boldsymbol{p} \odot e^{\eta\underline{\ell}(\boldsymbol{p})}\rangle \leq 1\}. \tag{30}$$

To see the set inclusion in (30), consider $\boldsymbol{s} \in -\mathcal{S}_{\log}^{\oplus} + \underline{\ell}_{\log}(\boldsymbol{p}) - \eta\underline{\ell}(\boldsymbol{p})$, then by definition of the superprediction set $\mathcal{S}_{\log}^{\oplus}$ there exists $\boldsymbol{r} \in \Delta_n$ and $\boldsymbol{v} \in [0, +\infty[^n$, such that $\boldsymbol{s} = \log \boldsymbol{r} - \log \boldsymbol{p} - \eta\underline{\ell}(\boldsymbol{p}) - \boldsymbol{v}$. Thus,

$$\langle e^{\boldsymbol{s}}, \boldsymbol{p} \odot e^{\eta\underline{\ell}(\boldsymbol{p})} \rangle = \langle \boldsymbol{r}, e^{-\boldsymbol{v}} \rangle \leq 1, \tag{31}$$

where the inequality is true because $\boldsymbol{r} \in \Delta_n$ and $\boldsymbol{v} \in [0, +\infty[^n$. The above argument shows that $e^{-\eta\mathcal{S}_{\ell}^{\oplus}} \subseteq \mathcal{H}_{\tau(\boldsymbol{p}),1}$. Furthermore, $e^{-\eta\mathcal{S}_{\ell}^{\oplus}} \subseteq \mathcal{H}_{\tau(\boldsymbol{p}),1} \cap [0, +\infty[^n$, since all elements of $e^{-\eta\mathcal{S}_{\ell}^{\oplus}}$ have non-negative components. The latter set inclusion still holds for $\hat{\boldsymbol{p}} \in \operatorname{rbd} \Delta_n$. In fact, from the definition of a support loss, there exists a sequence $(\boldsymbol{p}_m)$ in $\operatorname{ri} \Delta_n$ converging to $\hat{\boldsymbol{p}}$ such that $\underline{\ell}(\boldsymbol{p}_m) \overset{m\to\infty}{\to} \underline{\ell}(\hat{\boldsymbol{p}})$. Equation 31 implies that for $\boldsymbol{u} \in e^{-\eta\mathcal{S}_{\ell}^{\oplus}}$, $\langle \boldsymbol{u}, \boldsymbol{p}_m \odot e^{\eta\underline{\ell}(\boldsymbol{p}_m)} \rangle \leq 1$. Since the inner product is continuous, by passage to the limit, we obtain $\langle \boldsymbol{u}, \hat{\boldsymbol{p}} \odot e^{\eta\underline{\ell}(\hat{\boldsymbol{p}})} \rangle \leq 1$. Therefore,

$$e^{-\eta\mathcal{S}_{\ell}^{\oplus}} \subseteq \bigcap_{\boldsymbol{p}\in\Delta_n} \mathcal{H}_{\tau(\boldsymbol{p}),1} \cap [0, +\infty[^n. \tag{32}$$

Now suppose $\boldsymbol{u} \in \bigcap_{\boldsymbol{p}\in\Delta_n} \mathcal{H}_{\tau(\boldsymbol{p}),1} \cap [0, +\infty[^n$; that is, for all $\boldsymbol{p} \in \Delta_n$,

$$\begin{aligned} 1 \geq \left\langle \boldsymbol{u}, \boldsymbol{p} \odot e^{\eta\underline{\ell}(\boldsymbol{p})} \right\rangle = \left\langle \boldsymbol{p}, \boldsymbol{u} \odot e^{\eta\underline{\ell}(\boldsymbol{p})} \right\rangle &= \left\langle \boldsymbol{p}, e^{\eta\underline{\ell}(\boldsymbol{p}) + \log \boldsymbol{u}} \right\rangle, \\ &\geq e^{\langle \boldsymbol{p}, \eta\underline{\ell}(\boldsymbol{p})\rangle + \langle \boldsymbol{p}, \log \boldsymbol{u}\rangle}, \end{aligned} \tag{33}$$

where the first equality is obtained merely by expanding the expression of the inner product, and the second inequality is simply Jensen's Inequality. Since $\boldsymbol{u} \mapsto e^{\boldsymbol{u}}$ is strictly convex, the Jensen's inequality in (33) is strict unless $\exists (c, \boldsymbol{p}) \in \mathbb{R} \times \Delta_n$, such that

$$\eta\underline{\ell}(\boldsymbol{p}) + \log \boldsymbol{u} = c\mathbf{1}_n. \tag{34}$$

By substituting (34) into (33), we get $1 \geq \exp(c)$, and thus $c \leq 0$. Furthermore, (24) implies that $\boldsymbol{p} \in \operatorname{dom} \underline{\ell}$, and thus there exists $\boldsymbol{a} \in \operatorname{dom} \ell$ such that $\ell(\boldsymbol{a}) = \underline{\ell}(\boldsymbol{p})$ (Theorem 4). Using this and rearranging (34), we get $\boldsymbol{u} = \exp(-\eta\ell(\boldsymbol{a}) + c\mathbf{1})$. Since $c \leq 0$, this means that $\boldsymbol{u} \in \exp(-\eta\mathcal{S}_{\ell}^{\oplus})$. Suppose now that (34) does not hold. In this case, (33) must be a strict inequality for all $\boldsymbol{p} \in \Delta_n$. By applying the log in (33),

$$\forall \boldsymbol{p} \in \Delta_n, \underline{L}_{\ell}(\boldsymbol{p}) + \langle \boldsymbol{p}, \log \boldsymbol{u} \rangle = \langle \boldsymbol{p}, \eta\underline{\ell}(\boldsymbol{p}) \rangle + \langle \boldsymbol{p}, \log \boldsymbol{u} \rangle < 0. \tag{35}$$

Since $\boldsymbol{p} \mapsto \underline{L}_{\ell}(\boldsymbol{p}) = -\sigma_{\mathcal{S}_{\ell}^{\oplus}}(-\boldsymbol{p})$ is a closed concave function, the map $g \colon \boldsymbol{p} \mapsto \underline{L}_{\ell}(\boldsymbol{p}) + \langle \boldsymbol{p}, \log \boldsymbol{u} \rangle$ is also closed and concave, and thus upper semi-continuous. Since Δ_n is compact, the function g must attain its maximum in Δ_n (Holder, 2005, Thm. 1.13). Due to (35) this minimum is negative; there exists $c_1 > 0$ such that $\langle \boldsymbol{p}, \eta\underline{\ell}(\boldsymbol{p}) \rangle - \langle \boldsymbol{p}, -\log \boldsymbol{u} \rangle \leq -c_1$. Let $f(\boldsymbol{p}, x) := \underline{\ell}_x(\boldsymbol{p}) + \log u_x + c_1$, for $x \in [n]$. Consequently for all $\boldsymbol{p} \in \Delta_n$, $\mathbb{E}_{x\sim\boldsymbol{p}} f(\boldsymbol{p}, x) \leq 0$ and $\forall x \in [n], -\infty < f(\boldsymbol{p}, x)$. Thus, Lemma 25 applied to f with $\epsilon = c_1/2$, implies that there exists $\boldsymbol{p}_* \in \operatorname{ri} \Delta_n$, such that $\eta\underline{\ell}(\boldsymbol{p}_*) \leq -\log \boldsymbol{u} - c_1/2 \leq -\log \boldsymbol{u}$. From this inequality, $\boldsymbol{p}_* \in \operatorname{dom} \underline{\ell}$, and therefore, there exists $\boldsymbol{a}_* \in \operatorname{dom} \ell$ such that $\ell(\boldsymbol{a}_*) = \underline{\ell}(\boldsymbol{p}_*)$ (Theorem 4). This shows that $\eta\ell(\boldsymbol{a}_*) \leq -\log \boldsymbol{u}$, which implies that $\boldsymbol{u} \in \exp -\eta\mathcal{S}_{\ell}^{\oplus}$. Therefore, $\bigcap_{\boldsymbol{p}\in\Delta_n} \mathcal{H}_{\tau(\boldsymbol{p}),1} \cap [0, +\infty[^n \subseteq e^{-\eta\mathcal{S}_{\ell}^{\oplus}}$. Combining this with (32) shows that $e^{-\eta\mathcal{S}_{\ell}^{\oplus}} = \bigcap_{\boldsymbol{p}\in\Delta_n} \mathcal{H}_{\tau(\boldsymbol{p}),1} \cap [0, +\infty[^n$. Since $e^{-\eta\mathcal{S}_{\ell}^{\oplus}}$ is the intersection of convex set, it is a convex set itself. Therefore ℓ is η-mixable. ■

C.6. Proof of Proposition 11

Given an entropy $\Phi\colon \mathbb{R}^k \to \mathbb{R} \cup \{+\infty\}$ and a loss $\ell\colon \mathcal{A} \to [0, +\infty]$, we define

$$\Delta\mathsf{m}_\Phi(x, A, \boldsymbol{a}, \hat{\boldsymbol{q}}, \boldsymbol{\mu}) \coloneqq \langle \boldsymbol{\mu}, \ell_x(A)\rangle + D_\Phi(\boldsymbol{\mu}, \hat{\boldsymbol{q}}) - \ell_x(\boldsymbol{a}),$$

where $x \in [n]$, $A \in \mathcal{A}^k$, $\boldsymbol{a} \in \mathcal{A}$, and $\boldsymbol{q}, \hat{\boldsymbol{q}} \in \Delta_k$. Reid et al. (2015) showed that ℓ is Φ mixable if and only if $\Delta\mathsf{M}_\Phi \coloneqq \inf_{A\in\mathcal{A}^k, \hat{\boldsymbol{q}}\in\Delta_k} \sup_{\boldsymbol{a}_*\in\mathcal{A}} \inf_{\boldsymbol{\mu}\in\Delta_k, x\in[n]} \Delta\mathsf{m}_\Phi(x, A, \boldsymbol{a}, \hat{\boldsymbol{q}}, \boldsymbol{\mu}) \geq 0$.

Proof [**Proposition 11**]

[**We show that ℓ is $\Phi^{\mathcal{I}}$-mixable**] Let $\mathcal{I} \subseteq [k]$, with $|\mathcal{I}| > 1$, $A \in \mathcal{A}^k$, and $\boldsymbol{q} \in \Delta_{\mathcal{I}}$. Since ℓ is Φ-mixable, the following holds

$$\begin{aligned}
\exists \boldsymbol{a}_* \in \Delta_n, \forall x \in [n],\ \ell_x(\boldsymbol{a}_*) &\leq \inf_{\hat{\boldsymbol{q}}\in\Delta_k} \langle \hat{\boldsymbol{q}}, \ell_x(A)\rangle + D_\Phi(\hat{\boldsymbol{q}}, \boldsymbol{q}), && (36)\\
&\leq \inf_{\hat{\boldsymbol{q}}\in\Delta_{\mathcal{I}}} \langle \hat{\boldsymbol{q}}, \ell_x(A)\rangle + D_\Phi(\hat{\boldsymbol{q}}, \boldsymbol{q}), && (37)\\
&= \inf_{\hat{\boldsymbol{q}}\in\Delta_{\mathcal{I}}} \left\langle \Pi^k_{\mathcal{I}}\hat{\boldsymbol{q}}, \Pi^k_{\mathcal{I}}\ell_x(A)\right\rangle + D_{\Phi^{\mathcal{I}}}(\Pi^k_{\mathcal{I}}\hat{\boldsymbol{q}}, \Pi^k_{\mathcal{I}}\boldsymbol{q}),\\
&= \inf_{\hat{\boldsymbol{\mu}}\in\Delta_{|\mathcal{I}|}} \left\langle \hat{\boldsymbol{\mu}}, \ell_x(A[\Pi^k_{\mathcal{I}}]^\mathsf{T})\right\rangle + D_{\Phi^{\mathcal{I}}}(\hat{\boldsymbol{\mu}}, \Pi^k_{\mathcal{I}}\boldsymbol{q}), && (38)
\end{aligned}$$

where in (36) we used the fact that $\Phi^{\mathcal{I}}(\Pi^k_{\mathcal{I}}\boldsymbol{q}) = \Phi(\boldsymbol{q}), \forall \boldsymbol{q} \in \Delta_{\mathcal{I}}$. Given that $A \mapsto A[\Pi^k_{\mathcal{I}}]^\mathsf{T}$ [resp. $\boldsymbol{q} \mapsto \Pi^k_{\mathcal{I}}\boldsymbol{q}$] is onto from $\mathcal{A}^k$ to $\mathcal{A}^{|\mathcal{I}|}$ [resp. from $\Delta_{\mathcal{I}}$ to $\Delta_{|\mathcal{I}|}$], (38) implies that ℓ is $\Phi^{\mathcal{I}}$-mixable.

[**We show (8)**] Suppose that there exists $\hat{\boldsymbol{q}} \in \operatorname{rbd}\Delta_k$ and $\boldsymbol{q} \in \operatorname{ri}\Delta_k$ such that $|\Phi'(\hat{\boldsymbol{q}}; \boldsymbol{q} - \hat{\boldsymbol{q}})| < +\infty$. Let $f\colon [0, \epsilon] \to \mathbb{R}$ be defined by $f(\lambda) \coloneqq \Phi(\hat{\boldsymbol{q}} + \lambda(\boldsymbol{q} - \hat{\boldsymbol{q}}))$, where $\epsilon > 0$ is such that $\hat{\boldsymbol{q}} + \epsilon(\boldsymbol{q} - \hat{\boldsymbol{q}}) \in \operatorname{ri}\Delta_k$. The function f is closed and convex on $\operatorname{dom} f = [0, \epsilon]$ and $\lim_{\lambda\downarrow 0} \frac{f(\lambda) - f(0)}{\lambda} = f'(0; 1) = \Phi'(\hat{\boldsymbol{q}}; \boldsymbol{q} - \hat{\boldsymbol{q}})$ which is finite by assumption. Using this and the fact that $\lambda f'(0; 1) = f'(0; \lambda)$, we have $\lim_{\lambda\downarrow 0} \lambda^{-1}(f(\lambda) - f(0) - f'(0; \lambda)) = 0$. Substituting f by its expression in terms of Φ in the latter equality gives

$$\lim_{\lambda\downarrow 0} \lambda^{-1} D_\Phi(\hat{\boldsymbol{q}} + \lambda(\boldsymbol{q} - \hat{\boldsymbol{q}}), \hat{\boldsymbol{q}}) = 0. \tag{39}$$

Let $\eta > 0$ and $\theta^* \in [k]$ be such that $\hat{q}_{\theta^*} = 0$. Suppose that ℓ is a non-trivial Φ-mixable loss. Let x_1, x_0, $\boldsymbol{a}_1$, and $\boldsymbol{a}_0$ be as in the definition of non-trivial (see (3)). In particular, it holds that $\ell_{x_0}(\boldsymbol{a}_0) < \ell_{x_0}(\boldsymbol{a}_1)$. Fix $A \in \mathcal{A}^k$, such that $A_{\cdot,\theta^*} = \boldsymbol{a}_0$ and $A_{\cdot,\theta} = \boldsymbol{a}_1$ for $\theta \in [k] \setminus \{\theta^*\}$. Let

$$\boldsymbol{a}_* = \operatorname*{argmax}_{\boldsymbol{a}\in\Delta_n} \inf_{\boldsymbol{\mu}\in\Delta_k, x\in[n]} \Delta\mathsf{m}_\Phi(x, A, \boldsymbol{a}, \hat{\boldsymbol{q}}, \boldsymbol{\mu}),$$

with $\hat{\boldsymbol{q}} \in \operatorname{rbd}\Delta_k$ as in (39). Note that $\boldsymbol{a}_*$ exists since ℓ is closed.

If $\boldsymbol{a}_*$ is such that $\ell_{x_1}(\boldsymbol{a}_*) > \ell_{x_1}(\boldsymbol{a}_1)$, then taking $\boldsymbol{\mu} = \hat{\boldsymbol{q}}$ puts all weights on experts predicting $\boldsymbol{a}_1$, while $D_\Phi(\boldsymbol{\mu}, \hat{\boldsymbol{q}}) = 0$. Therefore,

$$\Delta\mathsf{M}_\Phi \leq \inf_{\boldsymbol{\mu}\in\Delta_k, x\in[n]} \Delta\mathsf{m}_\Phi(x, A, \boldsymbol{a}_*, \hat{\boldsymbol{q}}, \boldsymbol{\mu}) \leq \Delta\mathsf{m}_\Phi(x_1, A, \boldsymbol{a}, \hat{\boldsymbol{q}}, \hat{\boldsymbol{q}}) < 0.$$

This contradicts the Φ-mixability of ℓ. Therefore, $\ell_{x_1}(\boldsymbol{a}_*) = \ell_{x_1}(\boldsymbol{a}_1)$, which by (3) implies $\ell_{x_0}(\boldsymbol{a}_*) \geq \ell_{x_0}(\boldsymbol{a}_1)$. For $\boldsymbol{q}^\lambda = \hat{\boldsymbol{q}} + \lambda(\boldsymbol{q} - \hat{\boldsymbol{q}})$, with $\boldsymbol{q} \in \operatorname{ri} \Delta_k$ as in (36) and $\lambda \in [0, \epsilon]$,

$$
\begin{aligned}
\Delta \mathsf{M}_\Phi &\leq \inf_{\boldsymbol{\mu} \in \Delta_k, x \in [n]} \Delta \mathsf{m}_\Phi(x, A, \boldsymbol{a}_*, \hat{\boldsymbol{q}}, \boldsymbol{\mu}), \\
&\leq \Delta \mathsf{m}_\Phi(x_0, A, \boldsymbol{a}, \hat{\boldsymbol{q}}, \boldsymbol{q}^\lambda), \\
&= \langle \boldsymbol{q}^\lambda, \ell_{x_0}(A) \rangle + D_\Phi(\boldsymbol{q}^\lambda, \hat{\boldsymbol{q}}) - \ell_{x_0}(\boldsymbol{a}_*), \\
&= (1 - \lambda q_{\theta^*}) \ell_{x_0}(\boldsymbol{a}_1) + \lambda q_{\theta^*} \ell_{x_0}(\boldsymbol{a}_0) + D_\Phi(\boldsymbol{q}^\lambda, \hat{\boldsymbol{q}}) - \ell_{x_0}(\boldsymbol{a}_*), \\
&\leq (1 - \lambda q_{\theta^*}) \ell_{x_0}(\boldsymbol{a}_*) + \lambda q_{\theta^*} \ell_{x_0}(\boldsymbol{a}_0) + D_\Phi(\boldsymbol{q}^\lambda, \hat{\boldsymbol{q}}) - \ell_{x_0}(\boldsymbol{a}_*), \\
&= \lambda q_{\theta^*} (\ell_{x_0}(\boldsymbol{a}_0) - \ell_{x_0}(\boldsymbol{a}_*)) + D_\Phi(\hat{\boldsymbol{q}} + \lambda(\boldsymbol{q} - \hat{\boldsymbol{q}}), \hat{\boldsymbol{q}}).
\end{aligned}
$$

Since $q_{\theta^*} > 0$ ($\boldsymbol{q} \in \operatorname{ri} \Delta_k$) and $\ell_{x_0}(\boldsymbol{a}_0) < \ell_{x_0}(\boldsymbol{a}_1) \leq \ell_{x_0}(\boldsymbol{a}_*)$, (36) implies that there exists $\lambda_* > 0$ small enough such that $\lambda_* q_{\theta^*}(\ell_{x_0}(\boldsymbol{a}_0) - \ell_{x_0}(\boldsymbol{a}_*)) + D_\Phi(\hat{\boldsymbol{q}} + \lambda_*(\boldsymbol{q} - \hat{\boldsymbol{q}}), \hat{\boldsymbol{q}}) < 0$. But this implies that $\Delta \mathsf{M}_\Phi < 0$ which contradicts the Φ-mixability of ℓ. Therefore, $\Phi'(\hat{\boldsymbol{q}}; \boldsymbol{q} - \hat{\boldsymbol{q}})$ is either equal to $+\infty$ or $-\infty$. The former case is not possible. In fact, since Φ is convex, it must have non-decreasing slopes; In particular, it holds that $\Phi'(\hat{\boldsymbol{q}}; \boldsymbol{q} - \hat{\boldsymbol{q}}) \leq \Phi(\boldsymbol{q} - \hat{\boldsymbol{q}}) - \Phi(\hat{\boldsymbol{q}})$. Since Φ is finite on Δ_k (by definition of an entropy), we have $\Phi'(\hat{\boldsymbol{q}}; \boldsymbol{q} - \hat{\boldsymbol{q}}) < +\infty$. Therefore, we have just shown that

$$
\forall \hat{\boldsymbol{q}} \in \operatorname{rbd} \Delta_k, \forall \boldsymbol{q} \in \operatorname{ri} \Delta_k, \ \Phi'(\hat{\boldsymbol{q}}; \boldsymbol{q} - \hat{\boldsymbol{q}}) = -\infty. \tag{40}
$$

Now suppose that $(\hat{\boldsymbol{q}}, \boldsymbol{q}) \in (\operatorname{rbd} \Delta_{\mathcal{I}}) \times (\operatorname{ri} \Delta_{\mathcal{I}})$ for $\mathcal{I} \subseteq [k]$, with $|\mathcal{I}| > 1$. Note that in this case, we have $(\Phi^{\mathcal{I}})'(\Pi_{\mathcal{I}}^k \hat{\boldsymbol{q}}; \Pi_{\mathcal{I}}^k(\boldsymbol{q} - \hat{\boldsymbol{q}})) = \Phi'(\hat{\boldsymbol{q}}; \boldsymbol{q} - \hat{\boldsymbol{q}})$. We showed in the first step of this proof that under the assumptions of the proposition, ℓ must be $\Phi^{\mathcal{I}}$-mixable. Therefore, repeating the steps above that lead to (40) for Φ, $\hat{\boldsymbol{q}}$, and $\boldsymbol{q}$ substituted by $\Phi^{\mathcal{I}}$, $\Pi_{\mathcal{I}}^k \boldsymbol{q} \in \operatorname{rbd} \Delta_{|\mathcal{I}|}$, and $\Pi_{\mathcal{I}}^k \boldsymbol{q} \in \operatorname{ri} \Delta_{|\mathcal{I}|}$, we obtain $\Phi'(\hat{\boldsymbol{q}}; \boldsymbol{q} - \hat{\boldsymbol{q}}) = (\Phi^{\mathcal{I}})'(\Pi_{\mathcal{I}}^k \hat{\boldsymbol{q}}; \Pi_{\mathcal{I}}^k(\boldsymbol{q} - \hat{\boldsymbol{q}})) = -\infty$. This shows (8). ■

C.7. Proof of Proposition 12

Let "sgn" denote the *sign* function.

Proof [**Proposition 12**] Let $\mathcal{I} = \{1, 2\}$. Since ℓ is Φ-mixable, it must be $\Phi^{\mathcal{I}}$-mixable, where $\Phi^{\mathcal{I}} := \Phi^{\mathcal{I}} \circ [\Pi_{\mathcal{I}}^k]^\mathsf{T} : \mathbb{R}^2 \to \mathbb{R} \cup \{+\infty\}$ (Proposition 11). Let $\Psi := \Phi^{\mathcal{I}}$.

For $w \in]0, +\infty[$ and $z \in \operatorname{int} \operatorname{dom} \tilde{\Psi}^* = \mathbb{R}$ (see §2.2), we define $(\tilde{\Psi}^*)'_\infty(w) := \lim_{t \to +\infty} [\tilde{\Psi}^*(z + tw) - \tilde{\Psi}^*(z)]/t$. The value of $(\tilde{\Psi}^*)'_\infty(w)$ does not depend on the choice of z, and it holds that $(\tilde{\Psi}^*)'_\infty(w) = \sigma_{\operatorname{dom} \tilde{\Psi}}(w)$ and $(\tilde{\Psi}^*)'_\infty(-w) = \sigma_{\operatorname{dom} \tilde{\Psi}}(-w)$ (Hiriart-Urruty and Lemaréchal, 2001, Prop. C.1.2.2). In our case, we have $\operatorname{dom} \tilde{\Psi} = [0, 1]$ (by definition of $\tilde{\Psi}$), which implies that $\sigma_{\operatorname{dom} \tilde{\Psi}}(1) = 1$ and $\sigma_{\operatorname{dom} \tilde{\Psi}}(-1) = 0$. Therefore, $(\tilde{\Psi}^*)'_\infty(1) + (\tilde{\Psi}^*)'_\infty(-1) = 1$. As a result $\tilde{\Psi}^*$ cannot be affine; that is, it cannot have a constant slope between any two points in $\mathbb{R}$. For $\forall \delta > 0$, let $g_\delta \colon \mathbb{R} \times \{-1, 0, +1\} \to \mathbb{R}$ be defined by $g_\delta(s, u) := [\tilde{\Psi}^*(s + \delta(u+1)/2) - \tilde{\Psi}^*(s + \delta(u-1)/2)]/\delta$. Since $\tilde{\Psi}^*$ is convex it must have non-decreasing slopes (ibid., p.13). Combining this with the fact that $\tilde{\Psi}^*$ is not affine implies that

$$
\exists s_\delta^* \in \mathbb{R}, g_\delta(s_\delta^*, -1) < g_\delta(s_\delta^*, +1). \tag{41}
$$

The fact that $\tilde{\Psi}^*$ has non-decreasing slopes also implies that

$$g_\delta(s^*_\delta, +1) = [\tilde{\Psi}^*(s^*_\delta + \delta) - \tilde{\Psi}^*(s^*_\delta)]/\delta \leq \lim_{t\to\infty}[\tilde{\Psi}^*(s^*_\delta + t) - \tilde{\Psi}^*(s^*_\delta)]/t = (\tilde{\Psi}^*)'_\infty(1) = 1.$$

Similarly, we have $0 = -(\tilde{\Psi}^*)'_\infty(-1) \leq g_\delta(s^*_\delta, -1)$. Let $\tilde{\mu} \in \partial\tilde{\Psi}^*(s^*_\delta)$. Since $\tilde{\Psi}$ is a closed convex function the following equivalence holds $\tilde{\mu} \in \partial\tilde{\Psi}^*(s^*_\delta) \iff s^*_\delta \in \partial\tilde{\Psi}(\tilde{\mu})$ (ibid., Cor. D.1.4.4). Thus, if $\tilde{\mu} \in \{0, 1\} = \operatorname{bd}\tilde{\Delta}_2$, then $\partial\tilde{\Psi}(\tilde{\mu}) \neq \emptyset$, which is not possible since ℓ is Ψ-mixable (Lemma 30).

[We show $\underline{L}_\ell \in C^1(]0, +\infty[^n)$] We will now show that $\underline{L}_\ell$ is continuously differentiable on $]0, +\infty[^n$. Since $\underline{L}_\ell$ is 1-homogeneous, it suffice to check the differentiability on $\operatorname{ri}\Delta_n$. Suppose $\underline{L}_\ell$ is not differentiable at $\boldsymbol{p} \in \operatorname{ri}\Delta_n$. From Theorem 4, there exists $\boldsymbol{a}_0, \boldsymbol{a}_1 \in \mathcal{A}$ such that $\ell(\boldsymbol{a}_0), \ell(\boldsymbol{a}_1) \in \partial\sigma_{\mathcal{S}^\oplus_\ell}(-\boldsymbol{p})$ and $\ell(\boldsymbol{a}_0) \neq \ell(\boldsymbol{a}_1)$. Let $A := [\boldsymbol{a}_0, \boldsymbol{a}_1] \in \mathbb{R}^{n\times 2}$, $\delta := \min\{|\ell_x(\boldsymbol{a}_0) - \ell_x(\boldsymbol{a}_1)| : x \in [n], |\ell_x(\boldsymbol{a}_0) - \ell_x(\boldsymbol{a}_1)| > 0\}$, and $s^*_\delta \in \mathbb{R}$ as in (41). We denote $g^- := g_\delta(s^*_\delta, -1)$ and $g^+ := g_\delta(s^*_\delta, +1) \in]0, 1]$. Let $\tilde{\mu} \in \partial\tilde{\Psi}^*(s^*_\delta) \in \operatorname{int}\tilde{\Delta}_2$ and $\boldsymbol{\mu} = \Pi_2(\tilde{\mu}) \in \operatorname{ri}\Delta_2$. From the fact that ℓ is Ψ-mixable, $J_2^\top \ell_x(A) = \ell_x(\boldsymbol{a}_0) - \ell_x(\boldsymbol{a}_1)$, and (9), there must exist $\boldsymbol{a}_* \in \mathcal{A}$ such that for all $x \in [n]$,

$$\begin{aligned}\ell_x(\boldsymbol{a}_*) \leq \mathsf{M}_\Psi(\ell_x(A), \boldsymbol{\mu}) &= \ell_x(a_1) + \tilde{\Psi}^*(s^*_\delta) - \tilde{\Psi}^*(s^*_\delta - \ell_x(\boldsymbol{a}_0) + \ell_x(\boldsymbol{a}_1)),\\ &\leq \ell_x(a_1) + g_\delta(s^*_\delta, -\operatorname{sgn}[\ell_x(\boldsymbol{a}_0) - \ell_x(\boldsymbol{a}_1)])[\ell_x(\boldsymbol{a}_0) - \ell_x(\boldsymbol{a}_1)], \end{aligned} \tag{42}$$

where in (42) we used the fact that $\tilde{\Psi}^*$ has non-decreasing slopes and the definition of δ. When $\ell_x(\boldsymbol{a}_0) \leq \ell_x(\boldsymbol{a}_1)$, (42) becomes $\ell_x(\boldsymbol{a}_*) \leq (1 - g^+)\ell_x(\boldsymbol{a}_1) + g^+\ell_x(\boldsymbol{a}_0)$. Otherwise, we have $\ell_x(\boldsymbol{a}_*) \leq (1 - g^-)\ell_x(\boldsymbol{a}_1) + g^-\ell_x(\boldsymbol{a}_0) < (1 - g^+)\ell_x(\boldsymbol{a}_1) + g^+\ell_x(\boldsymbol{a}_0)$. Since ℓ is admissible, there must exist at least one $x \in [n]$ such that $\ell_x(\boldsymbol{a}_0) > \ell_x(\boldsymbol{a}_1)$. Combining this with the fact that $p_x > 0, \forall x \in [n]$ ($\boldsymbol{p} \in \operatorname{ri}\Delta_n$), implies that $\langle \boldsymbol{p}, \ell(\boldsymbol{a}_*)\rangle < \langle \boldsymbol{p}, (1 - g^+)\ell(\boldsymbol{a}_1) + g^+\ell(\boldsymbol{a}_0)\rangle = \underline{L}_\ell(\boldsymbol{p})$. This contradicts the fact that $\ell(\boldsymbol{a}_*) \in \mathcal{S}^\oplus_\ell$. Therefore, $\underline{L}_\ell$ must be differentiable at $\boldsymbol{p}$. As argued earlier, this implies that $\underline{L}_\ell$ must be differentiable on $]0, +\infty[^n$. Combining this with the fact that $\underline{L}_\ell$ is concave on $]0, +\infty[^n$, implies that $\underline{L}_\ell$ is continuously differentiable on $]0, +\infty[^n$ (ibid., Rmk. D.6.2.6).

[We show $\tilde{\Phi}^* \in C^1(\mathbb{R}^{k-1})$] Suppose that $\tilde{\Phi}^*$ is not differentiable at some $\boldsymbol{s}^* \in \mathbb{R}^{k-1}$. Then there exists $\boldsymbol{d} \in \mathbb{R}^{k-1} \setminus \{\boldsymbol{0}_{\tilde{k}}\}$ such that $(-\tilde{\Phi}^*)'(\boldsymbol{s}^*; -\boldsymbol{d}) < (\tilde{\Phi}^*)'(\boldsymbol{s}^*; \boldsymbol{d})$. Since $\boldsymbol{s}^* \in \operatorname{int}\operatorname{dom}\tilde{\Phi}^*$, $(\tilde{\Phi}^*)'(\boldsymbol{s}^*, \cdot)$ is finite and convex (Hiriart-Urruty and Lemaréchal, 2001, Prop. D.1.1.2), and thus it is continuous on $\operatorname{dom}\tilde{\Phi}^* = \mathbb{R}^{k-1}$ (ibid., Rmk. B.3.1.3). Consequently, there exists $\delta^* > 0$ such that

$$\forall \hat{\boldsymbol{d}} \in \mathbb{R}^{n-1}, \|\hat{\boldsymbol{d}} - \boldsymbol{d}\| \leq \delta^* \implies -(\tilde{\Phi}^*)'(\boldsymbol{s}^*; -\hat{\boldsymbol{d}}) < (\tilde{\Phi}^*)'(\boldsymbol{s}^*; \hat{\boldsymbol{d}}) \tag{43}$$

Let $g\colon \{-1, 1\} \to \mathbb{R}$ such that $g(u) := \sup_{\|\hat{\boldsymbol{d}}-\boldsymbol{d}\|\leq\delta^*} u(\tilde{\Phi}^*)'(\boldsymbol{s}^*; u\hat{\boldsymbol{d}})$. Note that since $\tilde{\Phi}^*$ has increasing slopes ($\tilde{\Phi}^*$ is convex), $g(1) \leq \sup_{\|\hat{\boldsymbol{d}}-\boldsymbol{d}\|\leq\delta^*}(\tilde{\Phi}^*)'_\infty(\hat{\boldsymbol{d}}) = \sup_{\|\hat{\boldsymbol{d}}-\boldsymbol{d}\|\leq\delta^*}\sigma_{\operatorname{dom}\tilde{\Phi}}(\hat{\boldsymbol{d}}) \leq 1$, where the last inequality holds because $\tilde{\Delta}_k \subset \mathcal{B}(\boldsymbol{0}_{\tilde{k}}, 1)$, and thus $\sigma_{\operatorname{dom}\tilde{\Phi}}(\hat{\boldsymbol{d}}) = \sigma_{\tilde{\Delta}_k}(\hat{\boldsymbol{d}}) \leq \sigma_{\mathcal{B}(\boldsymbol{0}_{\tilde{k}},1)}(\hat{\boldsymbol{d}}) = 1$. Let $\Delta g := g(1) - g(-1)$. From (40), it is clear that $\Delta g > 0$.

Suppose that $\underline{L}_\ell$ is twice differentiable on $]0, +\infty[^n$ and let $\underline{\ell}$ be a support loss of ℓ. By definition of a support loss, $\forall \boldsymbol{p} \in \operatorname{ri}\Delta_k, \tilde{\underline{\ell}}(\tilde{\boldsymbol{p}}) = \underline{\ell}(\boldsymbol{p}) = \nabla\underline{L}_\ell(\boldsymbol{p})$ (where $\tilde{\underline{\ell}} := \underline{\ell} \circ \Pi_n$). Thus, since $\underline{L}_\ell$ is twice differentiable on $]0, +\infty[^n$, $\tilde{\underline{\ell}}$ is differentiable on $\operatorname{int}\tilde{\Delta}_n$. Furthermore, $\underline{\ell}$ is continuous on $\operatorname{ri}\Delta_k$

given that $\underline{L}_\ell \in C^1(]0,+\infty[^n)$ as shown in the first part of this proof. Since ℓ is non-trivial, it is non-constant. Therefore, from Theorem 4, $\underline{\ell}$ is also non-constant. Consequently, the mean value theorem applied to $\underline{\ell}$ (see e.g. (Rudin, 1964, Thm. 5.10)) between any two points in $\operatorname{ri}\Delta_n$ with distinct image under $\underline{\ell}$, implies that there exists $(\tilde{\boldsymbol{p}}_*, \boldsymbol{v}_*) \in \operatorname{int}\tilde{\Delta}_n \times \mathbb{R}^{n-1}$, such that $\mathsf{D}\tilde{\underline{\ell}}(\tilde{\boldsymbol{p}}_*)\boldsymbol{v}_* \neq \mathbf{0}_{\tilde{n}}$. For the rest of the proof let $(\tilde{\boldsymbol{p}}, \boldsymbol{v}) := (\tilde{\boldsymbol{p}}_*, \boldsymbol{v}_*)$ and define $\mathcal{J} := \{x \in [n] : \mathsf{D}\tilde{\underline{\ell}}_x(\tilde{\boldsymbol{p}})\boldsymbol{v} \neq 0\}$. From Lemma 27, we have $\langle \boldsymbol{p}, \mathsf{D}\tilde{\ell}(\tilde{\boldsymbol{p}})\rangle = \mathbf{0}_{\tilde{n}}^\mathsf{T}$, which implies $\exists x \in \mathcal{J}, \mathsf{D}\tilde{\underline{\ell}}_x(\tilde{\boldsymbol{p}})\boldsymbol{v} < 0$. Let $\tilde{\boldsymbol{p}}^t := \tilde{\boldsymbol{p}} + t\boldsymbol{v}$. From Taylor's Theorem (see e.g. (Hardy, 2008, §151)) applied to the function $t \mapsto \tilde{\underline{\ell}}(\tilde{\boldsymbol{p}}^t)$, there exists $\epsilon^* > 0$ and functions $\delta_x : [-\epsilon^*, \epsilon^*] \to \mathbb{R}^n$, $x \in [n]$, such that $\lim_{t\to 0} t^{-1}\delta_x(t) = 0$ and

$$\forall |t| \leq \epsilon^*, \quad \underline{\ell}_x(\boldsymbol{p}^t) = \underline{\ell}_x(\boldsymbol{p}) + t\mathsf{D}\tilde{\underline{\ell}}_x(\boldsymbol{p})\boldsymbol{v} + \delta_x(t). \tag{44}$$

By shrinking ϵ^* if necessary, we may also assume that

$$\forall x \in \mathcal{J}, \forall \theta \in [k], \forall |t| \leq \epsilon^*,\ t^{-1}\delta_x(td_\theta) \leq \frac{\delta^* |\mathsf{D}\tilde{\underline{\ell}}_x(\tilde{\boldsymbol{p}})\boldsymbol{v}|}{\sqrt{n}\,\|\boldsymbol{d}\|}, \tag{45}$$

$$\forall x \notin \mathcal{J}, \forall \theta \in [k], \forall |t| \leq \epsilon^*,\ \delta_x(t) \leq \frac{\Delta g}{2} \min_{x\in\mathcal{J}}\{p_x\}. \tag{46}$$

Note that $\min_{x\in\mathcal{J}}\{p_x\} > 0$, since $\boldsymbol{p} \in \operatorname{ri}\Delta_n$. Let $\lambda_\theta := t^* \frac{d_\theta}{\|\boldsymbol{d}\|}$, for $\theta \in [k-1]$. From Theorem 4, there exists $[\boldsymbol{a}_\theta^*]_{\theta\in[k]} \in \mathcal{A}^k$, such that $\ell(\boldsymbol{a}_k) = \underline{\ell}(\boldsymbol{p})$ and $\ell(\boldsymbol{a}_\theta) = \underline{\ell}(\boldsymbol{p}^{\lambda_\theta}) = \underline{\ell}(\boldsymbol{p}) + t^* \frac{d_\theta}{\|\boldsymbol{d}\|} \mathsf{D}\tilde{\underline{\ell}}(\tilde{\boldsymbol{p}})\boldsymbol{v} + \delta\left(t^* \frac{d_\theta}{\|\boldsymbol{d}\|}\right)$, where $[\delta(\cdot)]_x := \delta_x(\cdot)$. Let $A := [\boldsymbol{a}_\theta]_{\theta\in[k]}$. From the fact that ℓ is Φ-mixable, it follows that there exists $\boldsymbol{a}_* \in \mathcal{A}$ such that for all $x \in [n]$,

$$\ell_x(\boldsymbol{a}_*) \leq \mathsf{M}_\Phi(\ell_x(A), \boldsymbol{\mu}) = \ell_x(\boldsymbol{a}_k) + \tilde{\Phi}^*(\boldsymbol{s}^*) - \tilde{\Phi}^*(\boldsymbol{s}^* - J_k^\mathsf{T}\ell_x(A)). \tag{47}$$

Note that for all $x \in [n]$, $J_k^\mathsf{T}\ell_x(A) = [\ell_x(\boldsymbol{a}_\theta) - \ell_x(\boldsymbol{a}_k)]_{\theta\in[\tilde{k}]}$. Thus, $\forall x \in \mathcal{J}, \forall \theta \in [\tilde{k}]$, $[J_k^\mathsf{T}\ell_x(A)]_\theta = \left[\frac{t^*}{\|\boldsymbol{d}\|}\mathsf{D}\tilde{\underline{\ell}}_x(\tilde{\boldsymbol{p}})\boldsymbol{v}\right] d_{x,\theta}$, where $d_{x,\theta} := d_\theta + \frac{\|\boldsymbol{d}\|}{t^*[\mathsf{D}\tilde{\underline{\ell}}_x(\tilde{\boldsymbol{p}})\boldsymbol{v}]}\delta_x\left(t^* \frac{d_\theta}{\|\boldsymbol{d}\|}\right)$. On the other hand, if $\theta \notin \mathcal{J}$, then $[J_k^\mathsf{T}\ell_x(A)]_\theta = \delta_x\left(t^* \frac{d_\theta}{\|\boldsymbol{d}\|}\right)$. From (45), we have $\|\boldsymbol{d}_x - \boldsymbol{d}\| \leq \delta_*, \forall x \in \mathcal{J}$, and from the monoticity of the slopes of $\tilde{\Phi}^*$

$$\begin{aligned} \forall x \in \mathcal{J}, \tilde{\Phi}^*(\boldsymbol{s}^*) - \tilde{\Phi}^*(\boldsymbol{s}^* - J_k^\mathsf{T}\ell_x(A)) &\leq (\tilde{\Phi}^*)'\left(\boldsymbol{s}^*; \frac{t^*}{\|\boldsymbol{d}\|}[\mathsf{D}\tilde{\underline{\ell}}_x(\tilde{\boldsymbol{p}})\boldsymbol{v}]\boldsymbol{d}_x\right), \\ &\leq \frac{t^*}{\|\boldsymbol{d}\|}[\mathsf{D}\tilde{\underline{\ell}}_x(\tilde{\boldsymbol{p}})\boldsymbol{v}]g(\operatorname{sgn}[\mathsf{D}\tilde{\underline{\ell}}_x(\tilde{\boldsymbol{p}})\boldsymbol{v}]), \\ &\leq \frac{t^*}{\|\boldsymbol{d}\|}[\mathsf{D}\tilde{\underline{\ell}}_x(\tilde{\boldsymbol{p}})\boldsymbol{v}]g(1) - \Delta g[\![\mathsf{D}\tilde{\underline{\ell}}_x(\tilde{\boldsymbol{p}})\boldsymbol{v} < 0]\!], \end{aligned} \tag{48}$$

where $[\![\cdot]\!]$ denotes the *Iverson* bracket. Equation 48 follows from the definition of Δg. Combining (48) with (47) yields

$$\begin{aligned} \langle \boldsymbol{p}, \ell(\boldsymbol{a}_*)\rangle &\leq \langle \boldsymbol{p}, \ell(\boldsymbol{a}_k)\rangle + \frac{t^*}{\|\boldsymbol{d}\|}\langle \boldsymbol{p}, \mathsf{D}\tilde{\underline{\ell}}(\tilde{\boldsymbol{p}})\boldsymbol{v}\rangle g(1) - \Delta g \min_{x'\in\mathcal{J}}\{p_{x'}\} + \sum_{x\notin\mathcal{J}} p_x \delta_x\left(t^* \frac{d_\theta}{\|\boldsymbol{d}\|}\right), \\ &\leq \langle \boldsymbol{p}, \ell(\boldsymbol{a}_k)\rangle - \frac{\Delta g}{2}\min_{x\in\mathcal{J}}\{p_x\}, \qquad (49) \\ &< \langle \boldsymbol{p}, \underline{\ell}(\boldsymbol{p})\rangle, \qquad (50) \end{aligned}$$

where in the first inequality we used the fact that $\exists x \in \mathcal{J}, \mathsf{D}\tilde{\underline{\ell}}_x(\tilde{\boldsymbol{p}})\boldsymbol{v} < 0$. In (49) we used (45) and the fact that $\langle \boldsymbol{p}, \mathsf{D}\tilde{\underline{\ell}}(\tilde{\boldsymbol{p}})\rangle = \mathbf{0}_{\tilde{n}}^{\mathsf{T}}$ (Lemma 27). Finally, (50) follows from the fact that $\min_{x\in\mathcal{J}}\{p_x\} > 0$ ($\boldsymbol{p} \in \operatorname{ri}\Delta_n$) and $\underline{\ell}(\boldsymbol{p}) = \ell(\boldsymbol{a}_k)$ (by definition of $\boldsymbol{a}_k$). Equation 50 shows that $\ell(\boldsymbol{a}^*) \notin \mathcal{S}_\ell^{\oplus}$, which is a contradiction. ■

C.8. Proof of Proposition 13

Proof First note that $\boldsymbol{q}_*$ is well defined; the infimum of $\boldsymbol{\mu} \mapsto \langle \boldsymbol{\mu}, \boldsymbol{q}\rangle + D_\Phi(\boldsymbol{\mu}, \boldsymbol{q})$ is attained for $\boldsymbol{q} \in \operatorname{ri}\Delta_k$ (see Remark 9).

Now since $\tilde{\Phi}$ is convex and $\tilde{\boldsymbol{q}} = \Pi_k(\boldsymbol{q}) \in \operatorname{int}\tilde{\Delta}_k = \operatorname{int}\operatorname{dom}\tilde{\Phi}$, we have $\partial\tilde{\Phi}(\tilde{\boldsymbol{q}}) \neq \emptyset$ (Rockafellar, 1997, Thm. 23.4). This means that there exists $\boldsymbol{s}_{\boldsymbol{q}}^* \in \partial\tilde{\Phi}(\tilde{\boldsymbol{q}})$ such that $\langle \boldsymbol{s}_{\boldsymbol{q}}^*, \tilde{\boldsymbol{q}}_* - \tilde{\boldsymbol{q}}\rangle = \tilde{\Phi}'(\tilde{\boldsymbol{q}}; \tilde{\boldsymbol{q}}_* - \tilde{\boldsymbol{q}})$ (Hiriart-Urruty and Lemaréchal, 2001, p.166). We will now show that $\boldsymbol{s}_{\boldsymbol{q}}^* - J_k^{\mathsf{T}}\boldsymbol{d} \in \partial\tilde{\Phi}(\tilde{\boldsymbol{q}}_*)$, which will imply that $\tilde{\boldsymbol{q}}_* \in \partial\tilde{\Phi}^*(\boldsymbol{s}_{\boldsymbol{q}}^* - J_k^{\mathsf{T}}\boldsymbol{d})$ (ibid., Cor. D.1.4.4). Let $\boldsymbol{q}_* = \operatorname{argmin}_{\boldsymbol{\mu}\in\Delta_k}\langle \boldsymbol{\mu}, \boldsymbol{d}\rangle + D_\Phi(\boldsymbol{\mu}, \boldsymbol{q})$. Thus, for all $\boldsymbol{\mu} \in \Delta_k$,

$$
\begin{aligned}
&\langle \boldsymbol{\mu}, \boldsymbol{d}\rangle + \tilde{\Phi}(\tilde{\boldsymbol{\mu}}) - \tilde{\Phi}(\tilde{\boldsymbol{q}}) - \tilde{\Phi}'(\tilde{\boldsymbol{q}}; \tilde{\boldsymbol{\mu}} - \tilde{\boldsymbol{q}}) \geq \langle \boldsymbol{q}_*, \boldsymbol{d}\rangle + \tilde{\Phi}(\tilde{\boldsymbol{q}}_*) - \tilde{\Phi}(\tilde{\boldsymbol{q}}) - \langle \boldsymbol{s}_{\boldsymbol{q}}^*, \tilde{\boldsymbol{q}}_* - \tilde{\boldsymbol{q}}\rangle, \\
\Longrightarrow \quad & \tilde{\Phi}(\tilde{\boldsymbol{\mu}}) \geq \tilde{\Phi}(\tilde{\boldsymbol{q}}_*) - \langle \tilde{\boldsymbol{\mu}} - \tilde{\boldsymbol{q}}_*, J_k^{\mathsf{T}}\boldsymbol{d}\rangle + \langle \boldsymbol{s}_{\boldsymbol{q}}^*, \tilde{\boldsymbol{q}} - \tilde{\boldsymbol{q}}_*\rangle + \Phi'(\tilde{\boldsymbol{q}}; \tilde{\boldsymbol{\mu}} - \tilde{\boldsymbol{q}}), \\
\Longrightarrow \quad & \tilde{\Phi}(\tilde{\boldsymbol{\mu}}) \geq \tilde{\Phi}(\tilde{\boldsymbol{q}}_*) - \langle \tilde{\boldsymbol{\mu}} - \tilde{\boldsymbol{q}}_*, J_k^{\mathsf{T}}\boldsymbol{d}\rangle + \langle \boldsymbol{s}_{\boldsymbol{q}}^*, \tilde{\boldsymbol{q}} - \tilde{\boldsymbol{q}}_*\rangle + \langle \boldsymbol{s}_{\boldsymbol{q}}^*, \tilde{\boldsymbol{\mu}} - \tilde{\boldsymbol{q}}\rangle, \\
\Longrightarrow \quad & \tilde{\Phi}(\tilde{\boldsymbol{\mu}}) \geq \tilde{\Phi}(\tilde{\boldsymbol{q}}_*) + \langle \tilde{\boldsymbol{\mu}} - \tilde{\boldsymbol{q}}_*, \boldsymbol{s}_{\boldsymbol{q}}^* - J_k^{\mathsf{T}}\boldsymbol{d}\rangle,
\end{aligned}
$$

where in the second line we used the fact that $\forall \boldsymbol{q} \in \Delta_k, \langle \boldsymbol{q}, \boldsymbol{d}\rangle = \langle \tilde{\boldsymbol{q}}, J_k^{\mathsf{T}}\boldsymbol{d}\rangle + d_k$, and in third line we used the fact that $\forall \boldsymbol{s} \in \partial\tilde{\Phi}(\tilde{\boldsymbol{q}}),\ \langle \boldsymbol{s}, \tilde{\boldsymbol{\mu}} - \tilde{\boldsymbol{q}}\rangle \leq \tilde{\Phi}'(\tilde{\boldsymbol{q}}; \tilde{\boldsymbol{\mu}} - \tilde{\boldsymbol{q}})$ (ibid.). This shows that $\boldsymbol{s}_{\boldsymbol{q}}^* - J_k^{\mathsf{T}}\boldsymbol{d} \in \partial\tilde{\Phi}(\tilde{\boldsymbol{q}}_*)$.

Substituting $\tilde{\Phi}'(\tilde{\boldsymbol{q}}; \tilde{\boldsymbol{q}}_* - \tilde{\boldsymbol{q}})$ by $\langle \boldsymbol{s}_{\boldsymbol{q}}^*, \boldsymbol{q}_* - \boldsymbol{q}\rangle$ in the expression of $\mathsf{M}_\Phi(\boldsymbol{d}, \boldsymbol{q})$, we get

$$
\begin{aligned}
\mathsf{M}_\Phi(\boldsymbol{d}, \boldsymbol{q}) &= d_k + \langle \tilde{\boldsymbol{q}}_*, J_k^{\mathsf{T}}\boldsymbol{d}\rangle + \tilde{\Phi}(\tilde{\boldsymbol{q}}_*) - \tilde{\Phi}(\tilde{\boldsymbol{q}}) - \langle \boldsymbol{s}_{\boldsymbol{q}}^*, \tilde{\boldsymbol{q}}_* - \tilde{\boldsymbol{q}}\rangle, \\
&= d_k + \langle \boldsymbol{s}_{\boldsymbol{q}}^*, \tilde{\boldsymbol{q}}\rangle - \tilde{\Phi}(\tilde{\boldsymbol{q}}) - [\langle \boldsymbol{s}_{\boldsymbol{q}}^* - J_k^{\mathsf{T}}\boldsymbol{d}, \tilde{\boldsymbol{q}}_*\rangle - \tilde{\Phi}(\tilde{\boldsymbol{q}}_*)], \\
&= d_k + \tilde{\Phi}^*(\boldsymbol{s}_{\boldsymbol{q}}^*) - \tilde{\Phi}^*(\boldsymbol{s}_{\boldsymbol{q}}^* - J_k^{\mathsf{T}}\boldsymbol{d}),
\end{aligned}
$$

where in the last line we used the fact that $\tilde{\Phi}$ is a closed convex function, and thus $\forall \tilde{\boldsymbol{q}} \in \tilde{\Delta}_k$, $\boldsymbol{s} \in \partial\tilde{\Phi}(\tilde{\boldsymbol{q}}) \implies \tilde{\Phi}^*(\boldsymbol{s}) = \langle \boldsymbol{s}, \tilde{\boldsymbol{q}}\rangle - \tilde{\Phi}(\tilde{\boldsymbol{q}})$ (ibid., Cor. E.1.4.4).

■

C.9. Proof of Theorem 14

Proof Let $\boldsymbol{q} \in \Delta_k$ and $A := [\boldsymbol{a}_\theta]_{1\leq\theta\leq k} \in \mathcal{A}^k$. We claim that for $x \in [n]$

$$
-\eta^{-1}\log\langle \exp(-\eta\ell_x(A)), \boldsymbol{q}\rangle = \mathsf{M}_{\mathrm{S}_\eta}(\ell_x(A), \boldsymbol{q}). \tag{51}
$$

Let $x \in [n]$. Since $\boldsymbol{q} \in \Delta_k$, $\boldsymbol{q}$ is either a vertex of Δ_k or there exists $\mathcal{I} \subseteq [n]$, with $|\mathcal{I}| > 1$, such that $\boldsymbol{q} \in \operatorname{ri}\Delta_{\mathcal{I}}$. Suppose the latter case holds. We showed (see (14)) that for $A' \in (\operatorname{dom}\ell)^k$,

$$
-\eta^{-1}\log\left\langle \exp(-\eta\ell_x(A')), \boldsymbol{q}\right\rangle = \mathsf{M}_{\mathrm{S}_\eta^{\mathcal{I}}}(\Pi_{\mathcal{I}}^k\ell_x(A'), \Pi_{\mathcal{I}}^k\boldsymbol{q}). \tag{52}
$$

Since this result was obtained without any special assumptions on ℓ — we only assumed that $\ell_x(A') \in [0, +\infty[^k$ — then (52) still holds if we replace $\ell_x(A')$ by any $\boldsymbol{d} \in [0, +\infty[^k$; that is,

$$\forall \boldsymbol{d} \in [0, +\infty[^n, \ -\eta^{-1} \log \langle \exp(-\eta \boldsymbol{d}, \boldsymbol{q} \rangle = \mathsf{M}_{\mathrm{S}_\eta^{\mathcal{I}}}(\Pi_{\mathcal{I}}^k \boldsymbol{d}, \Pi_{\mathcal{I}}^k \boldsymbol{q}),$$
$$= \mathsf{M}_{\mathrm{S}_\eta}(\boldsymbol{d}, \boldsymbol{q}), \tag{53}$$

where (53) is implied by Lemma 30 since S_η satisfies (8) (see Lemma 29).

Fix $x \in [n]$ and let $\hat{\boldsymbol{d}} := \ell_x(A) \in [0, +\infty]^k$. From Lemma 31, there exists a sequence $(\hat{\boldsymbol{d}}_m) \subset [0, +\infty[^k$ converging to $\hat{\boldsymbol{d}}$ such that $\mathsf{M}_{\mathrm{S}_\eta}(\hat{\boldsymbol{d}}_m, \boldsymbol{q}) \overset{m\to\infty}{\to} \mathsf{M}_{\mathrm{S}_\eta}(\hat{\boldsymbol{d}}, \boldsymbol{q})$. Then, from (53)

$$-\eta^{-1} \log \langle \exp(-\eta \ell_x(A)), \boldsymbol{q} \rangle = \lim_{m\to\infty} -\eta^{-1} \log \left\langle \exp(-\eta \hat{\boldsymbol{d}}_m, \boldsymbol{q} \right\rangle,$$
$$= \lim_{m\to\infty} \mathsf{M}_{\mathrm{S}_\eta}(\hat{\boldsymbol{d}}_m, \boldsymbol{q}),$$
$$= \mathsf{M}_{\mathrm{S}_\eta}(\hat{\boldsymbol{d}}, \boldsymbol{q}) = \mathsf{M}_{\mathrm{S}_\eta}(\ell_x(A), \boldsymbol{q}). \tag{54}$$

Now suppose that $\boldsymbol{q}$ is a vertex. Without loss of generality assume that $\boldsymbol{q} = \boldsymbol{e}_1$ and let $\boldsymbol{\mu} \in \Delta_k \setminus \{\boldsymbol{e}_1\}$. Then there exists $\mathcal{I}_* \subset [k]$, such that $(\boldsymbol{e}_1, \boldsymbol{\mu}) \in (\operatorname{rbd} \Delta_{\mathcal{I}_*}) \times (\operatorname{ri} \Delta_{\mathcal{I}_*})$ and by Lemma 29, $\mathrm{S}'(\boldsymbol{e}_1; \boldsymbol{\mu} - \boldsymbol{e}_1) = -\infty$. Therefore, $\forall \boldsymbol{q} \in \Delta_k \setminus \{\boldsymbol{e}_1\}$, $D_{\mathrm{S}_\eta}(\boldsymbol{q}, \boldsymbol{e}_1) = +\infty$, which implies

$$\forall x \in [n], \mathsf{M}_{\mathrm{S}_\eta}(\ell_x(A), \boldsymbol{e}_1) = \inf_{\boldsymbol{q} \in \Delta_k} \langle \boldsymbol{q}, \ell_x(A) \rangle + D_{\mathrm{S}_\eta}(\boldsymbol{q}, \boldsymbol{e}_1),$$
$$= \langle \boldsymbol{e}_1, \ell_x(A) \rangle + D_{\mathrm{S}_\eta}(\boldsymbol{e}_1, \boldsymbol{e}_1),$$
$$= \langle \boldsymbol{e}_1, \ell_x(A) \rangle,$$
$$= \ell_x(\boldsymbol{a}_1) = -\eta^{-1} \log \langle \exp(-\eta \ell_x(A)), \boldsymbol{e}_1 \rangle . \tag{55}$$

Combining (55) and (54) proves the claim in (51). The desired equivalence follows trivially from the definitions of η-mixability and S_η-mixability.

■

C.10. Proof of Theorem 15

We need the following lemma to show Theorem 15.

Lemma 34 *Let Φ be as in Theorem 15. Then $\eta_\ell \Phi - \mathrm{S}$ is convex on Δ_k only if Φ satisfies (8).*

Proof Let $\hat{\boldsymbol{q}} \in \operatorname{rbd} \Delta_k$. Suppose that there exists $\boldsymbol{q} \in \operatorname{ri} \Delta_k$ such that $\Phi'(\hat{\boldsymbol{q}}; \boldsymbol{q} - \hat{\boldsymbol{q}}) > -\infty$. Since Φ is convex, it must have non-decreasing slopes; in particular, it holds that $\Phi'(\hat{\boldsymbol{q}}; \boldsymbol{q} - \hat{\boldsymbol{q}}) \leq \Phi(\boldsymbol{q}) - \Phi(\hat{\boldsymbol{q}})$. Therefore, since Φ is finite on Δ_k (by definition of an entropy), we have $\Phi'(\hat{\boldsymbol{q}}; \boldsymbol{q} - \hat{\boldsymbol{q}}) < +\infty$. Since by assumption $\eta_\ell \Phi - \mathrm{S}$ is convex and finite on the simplex, we can use the same argument to show that $[\eta_\ell \Phi - \mathrm{S}]'(\hat{\boldsymbol{q}}; \boldsymbol{q} - \hat{\boldsymbol{q}}) = \eta_\ell \Phi'(\hat{\boldsymbol{q}}; \boldsymbol{q} - \hat{\boldsymbol{q}}) - \mathrm{S}'(\hat{\boldsymbol{q}}; \boldsymbol{q} - \hat{\boldsymbol{q}}) < +\infty$. This is a contradiction since $\mathrm{S}'(\hat{\boldsymbol{q}}; \boldsymbol{q} - \hat{\boldsymbol{q}}) = -\infty$ (Lemma 29). Therefore, it must hold that $\Phi'(\hat{\boldsymbol{q}}; \boldsymbol{q} - \hat{\boldsymbol{q}}) = -\infty$.

Suppose now that $(\hat{\boldsymbol{q}}, \boldsymbol{q}) \in (\operatorname{rbd} \Delta_{\mathcal{I}}) \times (\operatorname{ri} \Delta_{\mathcal{I}})$ for $\mathcal{I} \subseteq [k]$, with $|\mathcal{I}| > 1$. Let $\Phi^{\mathcal{I}} := \Phi \circ [\Pi_{\mathcal{I}}^k]^\mathsf{T}$ and $\mathrm{S}^{\mathcal{I}} := \mathrm{S} \circ [\Pi_{\mathcal{I}}^k]^\mathsf{T}$. Since $\eta_\ell \Phi - \mathrm{S}$ is convex on Δ_k and $\Pi_{\mathcal{I}}^k$ is a linear function, $\eta_\ell \Phi^{\mathcal{I}} - \mathrm{S}^{\mathcal{I}}$ is convex on $\Delta_{|\mathcal{I}|}$. Repeating the steps above for Φ and S substituted by $\Phi^{\mathcal{I}}$ and $\mathrm{S}^{\mathcal{I}}$, respectively, we get that $(\Phi^{\mathcal{I}})'(\Pi_{\mathcal{I}}^k \hat{\boldsymbol{q}}; \Pi_{\mathcal{I}}^k \boldsymbol{q} - \Pi_{\mathcal{I}}^k \hat{\boldsymbol{q}}) = -\infty$. Since $(\Phi^{\mathcal{I}})'(\Pi_{\mathcal{I}}^k \hat{\boldsymbol{q}}; \Pi_{\mathcal{I}}^k \boldsymbol{q} - \Pi_{\mathcal{I}}^k \hat{\boldsymbol{q}}) = \Phi'(\hat{\boldsymbol{q}}; \boldsymbol{q} - \hat{\boldsymbol{q}})$ the proof is completed. ■

Proof [**Theorem 15**] Assume $\eta_\ell \Phi - \mathrm{S}$ is convex on Δ_k. For this to hold, it is necessary that $\eta_\ell > 0$ since $-\mathrm{S}$ is strictly concave. Let $\eta := \eta_\ell$ and $\mathrm{S}_\eta := \eta^{-1}\mathrm{S}$. Then $\tilde{\mathrm{S}}_\eta = \eta^{-1}\tilde{\mathrm{S}}$ and $\tilde{\Phi} - \tilde{\mathrm{S}}_\eta = (\Phi - \mathrm{S}_\eta) \circ \Pi_k$ is convex on $\tilde{\Delta}_k$, since $\Phi - \mathrm{S}_\eta$ is convex on Δ_k and Π_k is affine.

Let $x \in [n]$, $A := [\boldsymbol{a}_\theta]_{\theta \in [k]}$, and $\boldsymbol{q} \in \Delta_k$. Suppose that $\boldsymbol{q} \in \operatorname{ri} \Delta_k$ and let $\boldsymbol{s}^*_{\boldsymbol{q}} \in \partial \tilde{\Phi}(\tilde{\boldsymbol{q}})$ be as in Proposition 13. Note that if $\ell_x(\boldsymbol{a}_\theta) = +\infty, \forall \theta \in [k]$, then the Φ-mixability condition (6) is trivially satisfied. Suppose, without loss of generality, that $\ell_x(\boldsymbol{a}_k) < +\infty$. Let $(\boldsymbol{d}_m) \subset [0, +\infty[^k$ be such that $\boldsymbol{d}_m \overset{m\to\infty}{\to} \boldsymbol{d} := \ell_x(A) \in [0, +\infty]^k$ and $\mathsf{M}_\Psi(\boldsymbol{d}_m, \boldsymbol{q}) \overset{m\to\infty}{\to} \mathsf{M}_\Psi(\boldsymbol{d}, \boldsymbol{q})$ for $\Psi \in \{\Phi, \mathrm{S}_\eta\}$. This sequence is guaranteed to exist by Lemma 31.

Let $\tilde{\Upsilon}_{\boldsymbol{q}} : \mathbb{R}^{k-1} \to \mathbb{R} \cup \{+\infty\}$ be defined by

$$\tilde{\Upsilon}_{\boldsymbol{q}}(\tilde{\boldsymbol{\mu}}) := \tilde{\mathrm{S}}_\eta(\tilde{\boldsymbol{\mu}}) + \langle \tilde{\boldsymbol{\mu}}, \boldsymbol{s}^*_{\boldsymbol{q}} - \nabla \tilde{\mathrm{S}}_\eta(\tilde{\boldsymbol{q}}) \rangle - \tilde{\Phi}^*(\boldsymbol{s}^*_{\boldsymbol{q}}) + \tilde{\mathrm{S}}^*_\eta(\nabla \tilde{\mathrm{S}}_\eta(\tilde{\boldsymbol{q}})),$$

and it's Fenchel dual follows from Proposition 20 (i+ii):

$$\tilde{\Upsilon}^*_{\boldsymbol{q}}(\boldsymbol{v}) = \tilde{\mathrm{S}}^*_\eta(\boldsymbol{v} - \boldsymbol{s}^*_{\boldsymbol{q}} + \nabla \tilde{\mathrm{S}}_\eta(\tilde{\boldsymbol{q}})) + \tilde{\Phi}^*(\boldsymbol{s}^*_{\boldsymbol{q}}) - \tilde{\mathrm{S}}^*_\eta(\nabla \tilde{\mathrm{S}}_\eta(\tilde{\boldsymbol{q}})),$$

After substituting $\boldsymbol{v}$ by $\boldsymbol{s}^*_{\boldsymbol{q}} - J_k^\mathsf{T} \boldsymbol{d}$ in the expression of $\tilde{\Upsilon}^*_{\boldsymbol{q}}$ and rearranging, we get

$$\tilde{\mathrm{S}}^*_\eta(\nabla \tilde{\mathrm{S}}_\eta(\tilde{\boldsymbol{q}})) - \tilde{\mathrm{S}}^*_\eta(\nabla \tilde{\mathrm{S}}_\eta(\tilde{\boldsymbol{q}}) - J_k^\mathsf{T} \boldsymbol{d}_m) = \tilde{\Phi}^*(\boldsymbol{s}^*_{\boldsymbol{q}}) - \tilde{\Upsilon}^*_{\boldsymbol{q}}(\boldsymbol{s}^*_{\boldsymbol{q}} - J_k^\mathsf{T} \boldsymbol{d}_m). \tag{56}$$

Since $\boldsymbol{s}^*_{\boldsymbol{q}} \in \partial \tilde{\Phi}(\tilde{\boldsymbol{q}})$ and $\tilde{\Phi}$ is a closed convex function, combining Proposition 20-(iv) and the fact that $\tilde{\Phi}^{**} = \tilde{\Phi}$ (Hiriart-Urruty and Lemaréchal, 2001, Cor. E.1.3.6) yields $\langle \tilde{\boldsymbol{q}}, \boldsymbol{s}^*_{\boldsymbol{q}} \rangle - \tilde{\Phi}^*(\boldsymbol{s}^*_{\boldsymbol{q}}) = \tilde{\Phi}(\tilde{\boldsymbol{q}})$. Thus, after substituting $\tilde{\boldsymbol{\mu}}$ by $\tilde{\boldsymbol{q}}$ in the expression of $\tilde{\Upsilon}_{\boldsymbol{q}}$, we get

$$\tilde{\Phi}(\tilde{\boldsymbol{q}}) = \tilde{\Upsilon}_{\boldsymbol{q}}(\tilde{\boldsymbol{q}}). \tag{57}$$

On the other hand, $\tilde{\Phi} - \tilde{\Upsilon}_{\boldsymbol{q}}$ is convex on $\tilde{\Delta}_k$, since $\tilde{\Upsilon}_{\boldsymbol{q}}$ is equal to $\tilde{\mathrm{S}}_\eta$ plus an affine function. Thus, $\partial[\tilde{\Phi} - \tilde{\Upsilon}_{\boldsymbol{q}}](\tilde{\boldsymbol{q}}) + \partial \tilde{\Upsilon}_{\boldsymbol{q}}(\tilde{\boldsymbol{q}}) = \partial \tilde{\Phi}(\tilde{\boldsymbol{q}})$, since $\tilde{\Phi}$ and $\tilde{\Upsilon}_{\boldsymbol{q}}$ are both convex (ibid., Thm. D.4.1.1). Since $\tilde{\Upsilon}_{\boldsymbol{q}}$ is differentiable at $\tilde{\boldsymbol{q}}$, we have $\partial \tilde{\Upsilon}_{\boldsymbol{q}}(\tilde{\boldsymbol{q}}) = \{\nabla \tilde{\Upsilon}_{\boldsymbol{q}}(\tilde{\boldsymbol{q}})\} = \{\boldsymbol{s}^*_{\boldsymbol{q}}\}$. Furthermore, since $\boldsymbol{s}^*_{\boldsymbol{q}} \in \partial \tilde{\Phi}(\tilde{\boldsymbol{q}})$, then $\boldsymbol{0}_{\tilde{k}} \in \partial \tilde{\Phi}(\boldsymbol{q}) - \partial \tilde{\Upsilon}_{\boldsymbol{q}}(\tilde{\boldsymbol{q}}) = \partial[\tilde{\Phi} - \tilde{\Upsilon}_{\boldsymbol{q}}](\tilde{\boldsymbol{q}})$. Hence, $\tilde{\Phi} - \tilde{\Upsilon}_{\boldsymbol{q}}$ attains a minimum at $\tilde{\boldsymbol{q}}$ (ibid., Thm. D.2.2.1). Due to this and (57), $\tilde{\Phi} \geq \tilde{\Upsilon}_{\boldsymbol{q}}$, which implies that $\tilde{\Phi}^* \leq \tilde{\Upsilon}^*_{\boldsymbol{q}}$ (Proposition 20-(iii)). Using this in (56) gives for all $m \in \mathbb{N}$

$$\begin{aligned} \tilde{\mathrm{S}}^*_\eta(\nabla \tilde{\mathrm{S}}_\eta(\tilde{\boldsymbol{q}})) - \tilde{\mathrm{S}}^*_\eta(\nabla \tilde{\mathrm{S}}_\eta(\tilde{\boldsymbol{q}}) - J_k^\mathsf{T} \boldsymbol{d}_m) &\leq \tilde{\Phi}^*(\boldsymbol{s}^*_{\boldsymbol{q}}) - \tilde{\Phi}^*(\boldsymbol{s}^*_{\boldsymbol{q}} - J_k^\mathsf{T} \boldsymbol{d}_m), \\ \implies \mathsf{M}_{\mathrm{S}_\eta}(\boldsymbol{d}_m, \boldsymbol{q}) &\leq \mathsf{M}_\Phi(\boldsymbol{d}_m, \boldsymbol{q}), \end{aligned}$$

where the implication is obtained by adding $[\boldsymbol{d}_m]_k$ on both sides of the first inequality and using Proposition 13.

Suppose now that $\boldsymbol{q} \in \operatorname{ri} \Delta_{\mathcal{I}}$, with $|\mathcal{I}| > 1$, and let $\Phi^{\mathcal{I}} := \Phi \circ [\Pi^k_{\mathcal{I}}]^\mathsf{T}$ and $\mathrm{S}^{\mathcal{I}} := \mathrm{S} \circ [\Pi^k_{\mathcal{I}}]^\mathsf{T}$. Note that since $\eta_\ell \Phi - \mathrm{S}$ is convex on Δ_k and $\Pi^k_{\mathcal{I}}$ is a linear function, $\eta_\ell \Phi^{\mathcal{I}} - \mathrm{S}^{\mathcal{I}}$ is convex on $\Delta_{|\mathcal{I}|}$. Repeating the steps above for Φ, S, $\boldsymbol{q}$, and A substituted by $\Phi^{\mathcal{I}}$, $\mathrm{S}^{\mathcal{I}}$, $\Pi^k_{\mathcal{I}}\boldsymbol{q}$, and $A[\Pi^k_{\mathcal{I}}]^\mathsf{T}$, respectively, yields

$$\begin{aligned} \mathsf{M}_{\mathrm{S}^{\mathcal{I}}_\eta}(\Pi^k_{\mathcal{I}} \boldsymbol{d}_m, \Pi^k_{\mathcal{I}} \boldsymbol{q}) &\leq \mathsf{M}_{\Phi^{\mathcal{I}}}(\Pi^k_{\mathcal{I}} \boldsymbol{d}_m, \Pi^k_{\mathcal{I}} \boldsymbol{q}), \\ \implies \mathsf{M}_{\mathrm{S}_\eta}(\boldsymbol{d}_m, \boldsymbol{q}) &\leq \mathsf{M}_\Phi(\boldsymbol{d}_m, \boldsymbol{q}), \\ \implies \mathsf{M}_{\mathrm{S}_\eta}(\ell_x(A), \boldsymbol{q}) &\leq \mathsf{M}_\Phi(\ell_x(A), \boldsymbol{q}), \end{aligned} \tag{58}$$

where the first implication follows from Lemma 30, since S_η and Φ both satisfy (8) (see Lemmas 29 and 34), and (58) is obtained by passage to the limit $m \to \infty$. Since $\eta = \eta_\ell > 0$, ℓ is η-mixable, which implies that ℓ is S_η-mixable (Theorem 14). Therefore, there exists $\boldsymbol{a}_* \in \mathcal{A}$, such that

$$\ell_x(\boldsymbol{a}_*) \leq \mathsf{M}_{\mathrm{S}_\eta}(\ell_x(A), \boldsymbol{q}) \leq \mathsf{M}_\Phi(\ell_x(A), \boldsymbol{q}). \tag{59}$$

To complete the proof (that is, to show that ℓ is Φ-mixable), it remains to consider the case where $\boldsymbol{q}$ is a vertex of Δ_k. Without loss of generality assume that $\boldsymbol{q} = \boldsymbol{e}_1$ and let $\boldsymbol{\mu} \in \Delta_k \setminus \{\boldsymbol{e}_1\}$. Thus, there exists $\mathcal{I}_* \subseteq [k]$, with $|\mathcal{I}_*| > 1$, such that $(\boldsymbol{e}_1, \boldsymbol{\mu}) \in (\operatorname{rbd} \Delta_{\mathcal{I}_*}) \times (\operatorname{ri} \Delta_{\mathcal{I}_*})$, and Lemma 34 implies that $\Phi'(\boldsymbol{e}_1; \boldsymbol{\mu} - \boldsymbol{e}_1) = -\infty$. Therefore, $\forall \boldsymbol{q} \in \Delta_k \setminus \{\boldsymbol{e}_1\}, D_\Phi(\boldsymbol{q}, \boldsymbol{e}_1) = +\infty$, which implies

$$\begin{aligned} \forall x \in [n], \mathsf{M}_\Phi(\ell_x(A), \boldsymbol{e}_1) &= \inf_{\boldsymbol{q} \in \Delta_k} \langle \boldsymbol{q}, \ell_x(A) \rangle + D_\Phi(\boldsymbol{q}, \boldsymbol{e}_1), \\ &= \langle \boldsymbol{e}_1, \ell_x(A) \rangle + D_\Phi(\boldsymbol{e}_1, \boldsymbol{e}_1) = \langle \boldsymbol{e}_1, \ell_x(A) \rangle, \\ &= \ell_x(\boldsymbol{a}_1). \end{aligned} \tag{60}$$

The Φ-mixability condition (6) is trivially satisfied in this case. Combining (59) and (60) shows that ℓ is Φ-mixable. ■

C.11. Proof of Theorem 16

The following Lemma gives necessary regularity conditions on the entropy Φ under the assumptions of Theorem 16.

Lemma 35 *Let Φ and ℓ be as in Theorem 16. Then the following holds*

(i) $\tilde{\Phi}$ *is strictly concave on* $\operatorname{int} \tilde{\Delta}_k$.

(ii) $\tilde{\Phi}^*$ *is be continuously differentiable on* $\mathbb{R}^{k-1}$.

(iii) $\tilde{\Phi}^*$ *is twice differentiable on* $\mathbb{R}^{k-1}$ *and* $\forall \tilde{\boldsymbol{q}} \in \operatorname{int} \tilde{\Delta}_k, \mathsf{H}\tilde{\Phi}^*(\nabla \tilde{\Phi}(\tilde{\boldsymbol{q}})) = (\mathsf{H}\tilde{\Phi}(\tilde{\boldsymbol{q}}))^{-1}$.

(iv) For the Shannon entropy, we have $(\mathsf{H}\tilde{\mathrm{S}}(\tilde{\boldsymbol{q}}))^{-1} = \mathsf{H}\tilde{\mathrm{S}}^*(\nabla \tilde{\mathrm{S}}(\tilde{\boldsymbol{q}})) = \operatorname{diag} \tilde{\boldsymbol{q}} - \tilde{\boldsymbol{q}}\tilde{\boldsymbol{q}}^\mathsf{T}$.

Proof Since ℓ is Φ-mixable and $\underline{L}_\ell$ is twice differentiable on $]0, +\infty[^n$, $\tilde{\Phi}^*$ is continously differentiable on $\mathbb{R}^{n-1}$ (Proposition 12). Therefore, $\tilde{\Phi}$ is strictly convex on $\operatorname{ri} \Delta_k$ (Hiriart-Urruty and Lemaréchal, 2001, Thm. E.4.1.2).

The differentiability of $\tilde{\Phi}$ and $\tilde{\Phi}^*$ implies $\nabla \tilde{\Phi}^*(\nabla \tilde{\Phi}(\tilde{\boldsymbol{q}})) = \tilde{\boldsymbol{q}}$ (ibid.). Since $\tilde{\Phi}$ is twice differentiable on $\operatorname{int} \tilde{\Delta}_k$ (by assumption), the latter equation implies that $\tilde{\Phi}^*$ is twice differentiable on $\nabla \tilde{\Phi}(\operatorname{int} \tilde{\Delta}_k)$. Using the chain rule, we get $\mathsf{H}\tilde{\Phi}^*(\nabla \tilde{\Phi}(\boldsymbol{u}))\mathsf{H}\tilde{\Phi}(\boldsymbol{u}) = I_{\tilde{k}}$. Multiplying both sides of the equation by $(\mathsf{H}\tilde{\Phi}(\boldsymbol{u}))^{-1}$ from the right gives the expression in (iii). Note that $\mathsf{H}\tilde{\Phi}(\cdot)$ is in fact invertible on $\operatorname{int} \tilde{\Delta}_k$ since $\tilde{\Phi}$ is strictly convex on $\operatorname{int} \tilde{\Delta}_k$. It remains to show that $\nabla \tilde{\Phi}(\operatorname{int} \tilde{\Delta}_k) = \mathbb{R}^{k-1}$. This set equality follows from 1) $[\tilde{\boldsymbol{q}} \in \partial \tilde{\Phi}^*(\boldsymbol{s}) \iff \boldsymbol{s} \in \partial \tilde{\Phi}(\tilde{\boldsymbol{q}})]$ (ibid., Cor. E.1.4.4); 2) $\operatorname{dom} \tilde{\Phi}^* = \mathbb{R}^{k-1}$; and 3) $\forall \tilde{\boldsymbol{q}} \in \operatorname{bd} \tilde{\Delta}_k, \partial \tilde{\Phi}(\tilde{\boldsymbol{q}}) = \emptyset$ (Lemma 30).

For the Shannon entropy, we have $\tilde{\mathrm{S}}^*(\boldsymbol{v}) = \log(\langle \exp(\boldsymbol{v}), \boldsymbol{1}_{\tilde{k}} \rangle + 1)$ (Proposition 1) and $\nabla \tilde{\mathrm{S}}(\tilde{\boldsymbol{q}}) = \log \frac{\tilde{\boldsymbol{q}}}{q_k}$, for $(\boldsymbol{v}, \tilde{\boldsymbol{q}}) \in \mathbb{R}^{k-1} \times \tilde{\Delta}_k$. Thus $(\mathsf{H}\tilde{\mathrm{S}}(\tilde{\boldsymbol{q}}))^{-1} = \mathsf{H}\tilde{\mathrm{S}}^*(\nabla \tilde{\mathrm{S}}(\tilde{\boldsymbol{q}})) = \operatorname{diag} \tilde{\boldsymbol{q}} - \tilde{\boldsymbol{q}}\tilde{\boldsymbol{q}}^\mathsf{T}$. ■

To show Theorem 16, we analyze a particular parameterized curve defined in the next lemma.

Lemma 36 *Let $\ell \colon \Delta_n \to [0,+\infty]^n$ be a proper loss whose Bayes risk $\underline{L}_\ell$ is twice differentiable on $]0,+\infty[^n$, and let Φ be an entropy such that $\tilde{\Phi}$ and $\tilde{\Phi}^*$ are twice differentiable on $\operatorname{int}\tilde{\Delta}_k$ and $\mathbb{R}^{k-1}$, respectively. For $(\tilde{\boldsymbol{p}}, \tilde{\boldsymbol{q}}, V) \in \operatorname{int}\tilde{\Delta}_n \times \operatorname{int}\tilde{\Delta}_k \times \mathbb{R}^{\tilde{n}\times\tilde{k}}$, let $\beta : \mathbb{R} \to \mathbb{R}^n$ be the curve defined by*

$$\forall x \in [n], \quad \beta_x(t) = \tilde{\ell}_x(\tilde{\boldsymbol{p}}) + \tilde{\Phi}^*(\nabla\tilde{\Phi}(\tilde{\boldsymbol{q}})) - \tilde{\Phi}^*(\nabla\tilde{\Phi}(\tilde{\boldsymbol{q}}) - J_k^{\mathsf{T}}\tilde{\ell}_x(\tilde{P}^t)), \tag{61}$$

where $\tilde{P}^t = [\tilde{\boldsymbol{p}}\mathbf{1}_{\tilde{k}}^{\mathsf{T}} + tV,\ \tilde{\boldsymbol{p}}] \in \mathbb{R}^{\tilde{n}\times k}$ and $t \in \{s \in \mathbb{R} : \forall j \in [\tilde{k}],\ \tilde{\boldsymbol{p}} + sV_{\bullet,j} \in \operatorname{int}\tilde{\Delta}_n\}$. Then

$$\beta(0) = \tilde{\ell}(\tilde{\boldsymbol{p}}),$$

$$\dot{\beta}(0) = \mathsf{D}\tilde{\ell}(\tilde{\boldsymbol{p}})V\tilde{\boldsymbol{q}},$$

$$\frac{d}{dt}\left\langle \boldsymbol{p}, \dot{\beta}(t)\right\rangle\bigg|_{t=0} = -\sum_{j=1}^{k-1} q_j V_{\bullet,j}^{\mathsf{T}}\mathsf{H}\underline{\tilde{L}}_\ell(\tilde{\boldsymbol{p}})V_{\bullet,j} - \operatorname{tr}(\operatorname{diag}(\boldsymbol{p})\,\mathsf{D}\tilde{\ell}(\tilde{\boldsymbol{p}})V(\mathsf{H}\tilde{\Phi}(\tilde{\boldsymbol{q}}))^{-1}(\mathsf{D}\tilde{\ell}(\tilde{\boldsymbol{p}})V)^{\mathsf{T}}). \tag{62}$$

Proof Since $\tilde{P}^t = [\tilde{\boldsymbol{p}}\mathbf{1}_{\tilde{k}}^{\mathsf{T}} + tV,\ \tilde{\boldsymbol{p}}] \in \mathbb{R}^{\tilde{n}\times k}$, $\tilde{P}^0 = \tilde{\boldsymbol{p}}\mathbf{1}_k^{\mathsf{T}}$ and $\tilde{\ell}_x(\tilde{P}^0) = \tilde{\ell}_x(\tilde{\boldsymbol{p}})\mathbf{1}_k$. As a result, $J_k^{\mathsf{T}}\tilde{\ell}_x(\tilde{P}^0) = \mathbf{0}_{\tilde{k}}$, and thus $\beta_x(0) = \tilde{\ell}_x(\tilde{\boldsymbol{p}}) + \tilde{\Phi}^*(\nabla\tilde{\Phi}(\tilde{\boldsymbol{q}})) - \tilde{\Phi}^*(\nabla\tilde{\Phi}(\tilde{\boldsymbol{q}}) - \mathbf{0}_{\tilde{k}}) = \tilde{\ell}_x(\tilde{\boldsymbol{p}})$. This shows that $\beta(0) = \tilde{\ell}(\tilde{\boldsymbol{p}})$. Let $\gamma_x(t) \coloneqq \nabla\tilde{\Phi}(\tilde{\boldsymbol{q}}) - J_k^{\mathsf{T}}\tilde{\ell}_x(\tilde{P}^t)$. For $j \in [k-1]$,

$$\begin{aligned}
\frac{d}{dt}[\gamma_x(t)]_j &= \frac{d}{dt}\left([\nabla\tilde{\Phi}(\tilde{\boldsymbol{q}})]_j - [J_k^{\mathsf{T}}\tilde{\ell}_x(\tilde{P}^t)]_j\right), \\
&= -\frac{d}{dt}\left(\tilde{\ell}_x(\tilde{P}^t_{\bullet,j}) - \tilde{\ell}_x(\tilde{P}^t_{\bullet,k})\right), \\
&= -\frac{d}{dt}\left(\tilde{\ell}_x(\tilde{\boldsymbol{p}} + tV_{\bullet,j}) - \tilde{\ell}_x(\tilde{\boldsymbol{p}})\right), \qquad \left(\text{since } \frac{d}{dt}\ell_x(\tilde{P}^t_{\bullet,k}) = \frac{d}{dt}\tilde{\ell}_x(\tilde{\boldsymbol{p}}) = 0\right) \\
&= -\mathsf{D}\tilde{\ell}_x(\tilde{P}^t_{\bullet,j})V_{\bullet,j}.
\end{aligned}$$

From the definition of $\tilde{P}^t$, $\tilde{P}^0_{\bullet,j} = \tilde{\boldsymbol{p}}$, $\forall j \in [\tilde{k}]$, and therefore, $\dot{\gamma}_x(0) = -(\mathsf{D}\tilde{\ell}_x(\tilde{\boldsymbol{p}})V)^{\mathsf{T}}$. By differentiating β_x in (61) and using the chain rule, $\dot{\beta}_x(t) = -(\dot{\gamma}_x(t))^{\mathsf{T}}\nabla\tilde{\Phi}^*(\gamma_x(t))$. By setting

$t = 0$, $\dot{\beta}_x(0) = -(\dot{\gamma}_x(0))^{\mathsf{T}} \nabla \tilde{\Phi}^*(\nabla \tilde{\Phi}(\tilde{\boldsymbol{q}})) = \mathsf{D}\tilde{\ell}_x(\tilde{\boldsymbol{p}}) V \tilde{\boldsymbol{q}}$. Thus, $\dot{\beta}(0) = \mathsf{D}\tilde{\ell}(\tilde{\boldsymbol{p}}) V \tilde{\boldsymbol{q}}$. Furthermore,

$$
\begin{aligned}
\frac{d}{dt} \left\langle \boldsymbol{p}, \dot{\beta}(t) \right\rangle \Big|_{t=0} &= \frac{d}{dt} \sum_{x=1}^{n} p_x \left(\sum_{j=1}^{k-1} \mathsf{D}\tilde{\ell}_x(\tilde{P}^t_{\bullet,j}) V_{\bullet,j} [\nabla \tilde{\Phi}^*(\gamma_x(t))]_j \right) \Bigg|_{t=0}, \\
&= \sum_{j=1}^{k-1} \frac{d}{dt} \left(\sum_{x=1}^{n} p_x \mathsf{D}\tilde{\ell}_x(\tilde{P}^t_{\bullet,j}) V_{\bullet,j} [\nabla \tilde{\Phi}^*(\gamma_x(t))]_j \right) \Bigg|_{t=0}, \\
&= \sum_{j=1}^{k-1} \left(\frac{d}{dt} \left\langle \boldsymbol{p}, \mathsf{D}\tilde{\ell}(\tilde{P}^t_{\bullet,j}) V_{\bullet,j} q_j \right\rangle \Big|_{t=0} + \sum_{x=1}^{n} p_x \mathsf{D}\tilde{\ell}_x(\tilde{\boldsymbol{p}}) V_{\bullet,j} \frac{d}{dt} [\nabla \tilde{\Phi}^*(\gamma_x(t))]_j \Big|_{t=0} \right), \\
&= -\sum_{j=1}^{k-1} q_j V_{\bullet,j}^{\mathsf{T}} \mathsf{H}\underline{\tilde{L}}_\ell(\tilde{\boldsymbol{p}}) V_{\bullet,j} - \sum_{x=1}^{n} \sum_{\substack{i=1 \\ j=1}}^{k-1} p_x \mathsf{D}\tilde{\ell}_x(\tilde{\boldsymbol{p}}) V_{\bullet,j} [\mathsf{H}\tilde{\Phi}^*(\nabla \tilde{\Phi}(\tilde{\boldsymbol{q}}))]_{j,i} \mathsf{D}\tilde{\ell}_x(\tilde{\boldsymbol{p}}) V_{\bullet,i}, \\
&= -\sum_{j=1}^{k-1} q_j V_{\bullet,j}^{\mathsf{T}} \mathsf{H}\underline{\tilde{L}}_\ell(\tilde{\boldsymbol{p}}) V_{\bullet,j} - \operatorname{tr}(\operatorname{diag}(\boldsymbol{p})\, \mathsf{D}\tilde{\ell}(\tilde{\boldsymbol{p}}) V \mathsf{H}\tilde{\Phi}^*(\nabla \tilde{\Phi}(\tilde{\boldsymbol{q}})) (\mathsf{D}\tilde{\ell}(\tilde{\boldsymbol{p}}) V)^{\mathsf{T}}), \\
&= -\sum_{j=1}^{k-1} q_j V_{\bullet,j}^{\mathsf{T}} \mathsf{H}\underline{\tilde{L}}_\ell(\tilde{\boldsymbol{p}}) V_{\bullet,j} - \operatorname{tr}(\operatorname{diag}(\boldsymbol{p})\, \mathsf{D}\tilde{\ell}(\tilde{\boldsymbol{p}}) V (\mathsf{H}\tilde{\Phi}(\tilde{\boldsymbol{q}}))^{-1} (\mathsf{D}\tilde{\ell}(\tilde{\boldsymbol{p}}) V)^{\mathsf{T}}),
\end{aligned}
$$

where in the third equality we used Lemma 25, in the fourth equality we used Lemma 28, and in the sixth equality we used Lemma 35-(iii).

■

In next lemma, we state a necessary condition for Φ-mixability in terms of the parameterized curve β defined in Lemma 36.

Lemma 37 *Let ℓ, Φ, and β be as in Lemma 36. If $\exists (\tilde{\boldsymbol{p}}, \tilde{\boldsymbol{q}}, V) \in \operatorname{int} \tilde{\Delta}_n \times \operatorname{int} \tilde{\Delta}_k \times \mathbb{R}^{\tilde{n} \times \tilde{k}}$ such that the curve $\gamma(t) := \tilde{\ell}(\tilde{\boldsymbol{p}} + tV\tilde{\boldsymbol{q}})$ satisfies $\frac{d}{dt}\langle \boldsymbol{p}, \dot{\beta}(t) - \dot{\gamma}(t) \rangle \Big|_{t=0} < 0$, then ℓ is not $\Phi-$mixable. In particular, $\exists P \in \operatorname{ri} \Delta_n^k$, such that $[\mathsf{M}_\Phi(\ell_x(P), \boldsymbol{q})]^{\mathsf{T}}_{x \in [n]}$ lies outside $\mathcal{S}_\ell^{\oplus}$.*

Proof First note that for any triplet $(\tilde{\boldsymbol{p}}, \tilde{\boldsymbol{q}}, V) \in \operatorname{int} \tilde{\Delta}_n \times \operatorname{int} \tilde{\Delta}_k \times \mathbb{R}^{\tilde{n} \times \tilde{k}}$, the map $t \mapsto \left\langle \boldsymbol{p}, \dot{\beta}(t) - \dot{\gamma}(t) \right\rangle$ is differentiable at 0. This follows from Lemmas 25 and 36. Let $r(t) := \Pi_n(\tilde{\boldsymbol{p}} + tV\tilde{\boldsymbol{q}})$ and $\delta(t) := \langle r(t), \beta(t) - \gamma(t) \rangle$. Then

$$
\dot{\delta}(t) = \left\langle r(t), \dot{\beta}(t) - \dot{\gamma}(t) \right\rangle + \langle V\tilde{\boldsymbol{q}}, \beta(t) - \gamma(t) \rangle .
$$

Since $t \mapsto \langle \boldsymbol{p}, \dot{\beta}(t) - \dot{\gamma}(t)\rangle$ is differentiable at 0, it follows from Lemma 25 that $t \mapsto \dot{\delta}(t)$ is also differentiable at 0, and thus

$$\begin{aligned}
\ddot{\delta}(0) &= \frac{d}{dt}\left\langle r(t), \dot{\beta}(t) - \dot{\gamma}(t)\right\rangle\Big|_{t=0} + \left\langle J_n V\tilde{\boldsymbol{q}}, \dot{\beta}(0) - \dot{\gamma}(0)\right\rangle, \\
&= \left\langle \frac{d}{dt} r(t)\Big|_{t=0}, \dot{\beta}(0) - \dot{\gamma}(0)\right\rangle + \frac{d}{dt}\left\langle \boldsymbol{p}, \dot{\beta}(t) - \dot{\gamma}(t)\right\rangle\Big|_{t=0}, \qquad (63)\\
&= \left\langle J_n V\tilde{\boldsymbol{q}}, \dot{\beta}(0) - \dot{\gamma}(0)\right\rangle + \frac{d}{dt}\left\langle \boldsymbol{p}, \dot{\beta}(t) - \dot{\gamma}(t)\right\rangle\Big|_{t=0}, \\
&= \frac{d}{dt}\left\langle \boldsymbol{p}, \dot{\beta}(t) - \dot{\gamma}(t)\right\rangle\Big|_{t=0} < 0, \qquad (64)
\end{aligned}$$

where (63) and (64) hold because $\dot{\beta}(0) = \mathsf{D}\tilde{\ell}(\tilde{\boldsymbol{p}})V\tilde{\boldsymbol{q}} = \dot{\gamma}(0)$ (see Lemma 36). According to Taylor's theorem (see e.g. (Hardy, 2008, §151)), there exists $\epsilon > 0$ and $h : [-\epsilon, \epsilon] \to \mathbb{R}$ such that

$$\forall |t| \leq \epsilon,\ \delta(t) = \delta(0) + t\dot{\delta}(0) + \frac{t^2}{2}\ddot{\delta}(0) + h(t)t^2, \qquad (65)$$

and $\lim_{t\to 0} h(t) = 0$. From Lemma 36, $\beta(0) = \gamma(0) = 0$ and $\dot{\beta}(0) = \dot{\gamma}(0)$. Therefore, $\delta(0) = \dot{\delta}(0) = 0$ and (65) becomes $\delta(t) = \frac{t^2}{2}\ddot{\delta}(0) + h(t)t^2$. Due to (64) and the fact that $\lim_{t\to 0} h(t) = 0$, we can choose $\epsilon_* > 0$ small enough such that $\delta(\epsilon_*) = \frac{\epsilon_*^2}{2}\ddot{\delta}(0) + h(\epsilon_*)\epsilon_*^2 < 0$. This means that $\langle \Pi_n(\tilde{\boldsymbol{p}} + \epsilon_* V\tilde{\boldsymbol{q}}), \beta(\epsilon_*)\rangle < \langle \Pi_n(\tilde{\boldsymbol{p}} + \epsilon_* V\tilde{\boldsymbol{q}}), \tilde{\ell}(\tilde{\boldsymbol{p}} + \epsilon_* V\tilde{\boldsymbol{q}})\rangle = \langle \Pi_n(\tilde{\boldsymbol{p}} + \epsilon_* V\tilde{\boldsymbol{q}}), \ell(\Pi_n(\tilde{\boldsymbol{p}} + \epsilon_* V\tilde{\boldsymbol{q}})\rangle$. Therefore, $\beta(\epsilon_*)$ must lie outside the superprediction set. Thus, the mixability condition (6) does not hold for $P^{\epsilon_*} = \Pi_n[\tilde{\boldsymbol{p}}\mathbf{1}_k^{\mathsf{T}} + \epsilon_* V,\ \tilde{\boldsymbol{p}}] \in \operatorname{ri}\Delta_n^k$. This completes the proof. ■

Proof [**Theorem 16**] We will prove the contrapositive; suppose that $\underline{\eta_\ell}\Phi - \mathrm{S}$ is not convex on Δ_k and we show that ℓ cannot be Φ-mixable. Note first that from Lemma 35-(iii), $\tilde{\Phi}^*$ is twice differentiable on $\mathbb{R}^{k-1}$. Thus Lemmas 36 and 37 apply. Let $\underline{\ell}$ be a proper support loss of ℓ and suppose that $\underline{\eta_\ell}\Phi - \mathrm{S}$ is not convex on Δ_k, This implies that $\underline{\eta_\ell}\tilde{\Phi} - \tilde{\mathrm{S}}$ is not convex on $\operatorname{int}\tilde{\Delta}_k$, and by Lemma 23 there exists $\tilde{\boldsymbol{q}}_* \in \operatorname{int}\tilde{\Delta}_k$, such that $1 > \underline{\eta_\ell}\lambda_{\min}(\mathsf{H}\tilde{\Phi}(\tilde{\boldsymbol{q}}_*)(\mathsf{H}\tilde{\mathrm{S}}(\tilde{\boldsymbol{q}}_*))^{-1})$. From this and the definition of $\underline{\eta_\ell}$, there exists $\tilde{\boldsymbol{p}}_* \in \operatorname{int}\tilde{\Delta}_n$ such that

$$1 > \frac{\lambda_{\min}(\mathsf{H}\tilde{\Phi}(\tilde{\boldsymbol{q}}_*)(\mathsf{H}\tilde{\mathrm{S}}(\tilde{\boldsymbol{q}}_*))^{-1})}{\lambda_{\max}([\mathsf{H}\underline{\tilde{L}}_{\log}(\tilde{\boldsymbol{p}}_*)]^{-1}\mathsf{H}\underline{\tilde{L}}_\ell(\tilde{\boldsymbol{p}}_*))} = \frac{\lambda_{\min}(\mathsf{H}\tilde{\Phi}(\tilde{\boldsymbol{q}}_*)(\operatorname{diag}(\tilde{\boldsymbol{q}}_*) - \tilde{\boldsymbol{q}}_*\tilde{\boldsymbol{q}}_*^{\mathsf{T}}))}{\lambda_{\max}([\mathsf{H}\underline{\tilde{L}}_{\log}(\tilde{\boldsymbol{p}}_*)]^{-1}\mathsf{H}\underline{\tilde{L}}_\ell(\tilde{\boldsymbol{p}}_*))}, \qquad (66)$$

where the equality is due to Lemma 35-(iv). For the rest of this proof let $(\tilde{\boldsymbol{p}}, \tilde{\boldsymbol{q}}) = (\tilde{\boldsymbol{p}}^*, \tilde{\boldsymbol{q}}^*)$. By assumption, $\underline{\tilde{L}}_\ell$ twice differentiable and concave on $\operatorname{int}\tilde{\Delta}_n$, and thus $-\mathsf{H}\underline{\tilde{L}}_\ell(\tilde{\boldsymbol{p}})$ is symmetric positive semi-definite. Therefore, their exists a symmetric positive semi-definite matrix $\Lambda_{\boldsymbol{p}}$ such that $\Lambda_{\boldsymbol{p}}\Lambda_{\boldsymbol{p}} = -\mathsf{H}\underline{\tilde{L}}_\ell(\tilde{\boldsymbol{p}})$. From Lemma 35-(i), $\tilde{\Phi}$ is strictly convex on $\operatorname{int}\tilde{\Delta}_k$, and so there exists a symmetric positive definite matrix $K_{\boldsymbol{q}}$ such that $K_{\boldsymbol{q}}K_{\boldsymbol{q}} = \mathsf{H}\tilde{\Phi}(\tilde{\boldsymbol{q}})$. Let $\boldsymbol{w} \in \mathbb{R}^{n-1}$ be the unit norm eigenvector of $[\mathsf{H}\underline{\tilde{L}}_{\log}(\tilde{\boldsymbol{p}})]^{-1}\mathsf{H}\underline{\tilde{L}}_\ell(\tilde{\boldsymbol{p}})$ associated with $\lambda_*^\ell := \lambda_{\max}([\mathsf{H}\underline{\tilde{L}}_{\log}(\tilde{\boldsymbol{p}})]^{-1}\mathsf{H}\underline{\tilde{L}}_\ell(\tilde{\boldsymbol{p}}))$. Suppose that $c_\ell := \boldsymbol{w}^{\mathsf{T}}\mathsf{H}\underline{\tilde{L}}_\ell(\tilde{\boldsymbol{p}})\boldsymbol{w} = 0$. Since $\boldsymbol{w}^{\mathsf{T}}\Lambda_{\boldsymbol{p}}\Lambda_{\boldsymbol{p}}\boldsymbol{w} = -c_\ell = 0$, it follows from the positive semi-definiteness of $\Lambda_{\boldsymbol{p}}$ that $\Lambda_{\boldsymbol{p}}\boldsymbol{w} = \mathbf{0}_{\tilde{n}}$, and thus $\mathsf{H}\underline{\tilde{L}}_\ell(\tilde{\boldsymbol{p}})\boldsymbol{w} = -\Lambda_{\boldsymbol{p}}\Lambda_{\boldsymbol{p}}\boldsymbol{w} = \mathbf{0}_{\tilde{n}}$. This implies that $\lambda_*^\ell = 0$, which is not possible due to (66). Therefore, $\mathsf{H}\underline{\tilde{L}}_\ell(\tilde{\boldsymbol{p}})\boldsymbol{w} \neq \mathbf{0}_{\tilde{n}}$. Furthermore, the negative semi-definiteness of $\mathsf{H}\underline{\tilde{L}}_\ell(\tilde{\boldsymbol{p}})$ implies that

$$c_\ell = \boldsymbol{w}^{\mathsf{T}}\mathsf{H}\underline{\tilde{L}}_\ell(\tilde{\boldsymbol{p}})\boldsymbol{w} < 0. \qquad (67)$$

Let $\boldsymbol{v} \in \mathbb{R}^{k-1}$ be the unit norm eigenvector of $K_{\boldsymbol{q}}(\operatorname{diag}(\tilde{\boldsymbol{q}}) - \tilde{\boldsymbol{q}}\tilde{\boldsymbol{q}}^\mathsf{T})K_{\boldsymbol{q}}$ associated with $\lambda_*^\Phi := \lambda_{\min}(K_{\boldsymbol{q}}(\operatorname{diag}(\tilde{\boldsymbol{q}}) - \tilde{\boldsymbol{q}}\tilde{\boldsymbol{q}}^\mathsf{T})K_{\boldsymbol{q}}) = \lambda_{\min}(\mathsf{H}\tilde{\Phi}(\tilde{\boldsymbol{q}})(\operatorname{diag}(\tilde{\boldsymbol{q}}) - \tilde{\boldsymbol{q}}\tilde{\boldsymbol{q}}^\mathsf{T}))$, where the equality is due to Lemma 22. Let $\hat{\boldsymbol{v}} := K_{\boldsymbol{q}}\boldsymbol{v}$.

We will show that for $V = \boldsymbol{w}\hat{\boldsymbol{v}}^\mathsf{T}$, the parametrized curve β defined in Lemma 36 satisfies $\frac{d}{dt}\langle \boldsymbol{p}, \dot{\beta}(t) - \dot{\gamma}(t)\rangle\Big|_{t=0} < 0$, where $\gamma(t) = \tilde{\underline{\ell}}(\tilde{\boldsymbol{p}} + tV\tilde{\boldsymbol{q}})$. According to Lemma 37 this would imply that there exists $P \in \operatorname{ri}\Delta_n^k$, such that $[\mathsf{M}_\Phi(\underline{\ell}_x(P), \boldsymbol{q})]_{x\in[n]}^\mathsf{T}$ lies outside $\mathcal{S}_\ell^\oplus$. From Theorem 4, we know that there exists $A_* \in \mathcal{A}^k$, such that $\ell_x(A_*) = \underline{\ell}_x(P), \forall x \in [n]$. Therefore, $[\mathsf{M}_\Phi(\ell_x(A_*), \boldsymbol{q})]_{x\in[n]}^\mathsf{T} = [\mathsf{M}_\Phi(\underline{\ell}_x(P), \boldsymbol{q})]_{x\in[n]}^\mathsf{T} \notin \mathcal{S}_\ell^\oplus$, and thus ℓ is not Φ-mixable.

From Lemma 36 (Equation 62) and the fact that $V_{\cdot,j} = \hat{v}_j\boldsymbol{w}$, for $j \in [\tilde{k}]$, we can write

$$\begin{aligned}\frac{d}{dt}\left\langle \boldsymbol{p}, \dot{\beta}(t)\right\rangle\Big|_{t=0} &= -\sum_{j=1}^{k-1} q_j\hat{v}_j^2\boldsymbol{w}^\mathsf{T}\mathsf{H}\tilde{\underline{L}}_\ell(\tilde{\boldsymbol{p}})\boldsymbol{w} - \operatorname{tr}(\operatorname{diag}(\boldsymbol{p})\,\mathsf{D}\tilde{\underline{\ell}}(\tilde{\boldsymbol{p}})V(\mathsf{H}\tilde{\Phi}(\tilde{\boldsymbol{q}}))^{-1}(\mathsf{D}\tilde{\underline{\ell}}(\tilde{\boldsymbol{p}})V)^\mathsf{T}),\\ &= -\langle \tilde{\boldsymbol{q}}, \hat{\boldsymbol{v}} \odot \hat{\boldsymbol{v}}\rangle\, \boldsymbol{w}^\mathsf{T}\mathsf{H}\tilde{\underline{L}}_\ell(\tilde{\boldsymbol{p}})\boldsymbol{w} - (\hat{\boldsymbol{v}}^\mathsf{T}(\mathsf{H}\tilde{\Phi}(\boldsymbol{q}))^{-1}\hat{\boldsymbol{v}})\langle \boldsymbol{p}, [\mathsf{D}\tilde{\ell}(\tilde{\boldsymbol{p}})\boldsymbol{w}] \odot [(\mathsf{D}\tilde{\ell}(\tilde{\boldsymbol{p}})\boldsymbol{w}]\rangle,\end{aligned}$$

where the second equality is obtained by noting that 1) $(\hat{\boldsymbol{v}}^\mathsf{T}(\mathsf{H}\tilde{\Phi}(\boldsymbol{q}))^{-1}\hat{\boldsymbol{v}})$ is a scalar quantity and can be factorized out; and 2) $\operatorname{tr}(\operatorname{diag}(\boldsymbol{p})\mathsf{D}\tilde{\ell}(\tilde{\boldsymbol{p}})\boldsymbol{w}(\mathsf{D}\tilde{\ell}(\tilde{\boldsymbol{p}})\boldsymbol{w})^\mathsf{T}) = \langle \boldsymbol{p}, (\mathsf{D}\tilde{\ell}(\tilde{\boldsymbol{p}})\boldsymbol{w}) \odot (\mathsf{D}\tilde{\ell}(\tilde{\boldsymbol{p}})\boldsymbol{w})\rangle$.

On the other hand, from Lemma 28, $\frac{d}{dt}\langle \boldsymbol{p}, \dot{\gamma}(t)\rangle\big|_{t=0} = -\langle \tilde{\boldsymbol{q}}, \hat{\boldsymbol{v}}\rangle^2\boldsymbol{w}^\mathsf{T}\mathsf{H}\tilde{\underline{L}}_\ell(\tilde{\boldsymbol{q}})\boldsymbol{w}$. Using (19) and the definition of c_ℓ, we get

$$\begin{aligned}\frac{d}{dt}\left\langle \boldsymbol{p}, \dot{\beta}(t) - \dot{\gamma}(t)\right\rangle\Big|_{t=0} &= [-\langle \tilde{\boldsymbol{q}}, \hat{\boldsymbol{v}} \odot \hat{\boldsymbol{v}}\rangle + \langle \tilde{\boldsymbol{q}}, \hat{\boldsymbol{v}}\rangle^2]c_\ell +\\ &\qquad (\hat{\boldsymbol{v}}^\mathsf{T}(\mathsf{H}\tilde{\Phi}(\boldsymbol{q}))^{-1}\hat{\boldsymbol{v}})(\boldsymbol{w}^\mathsf{T}(\mathsf{H}\tilde{\underline{L}}_\ell(\tilde{\boldsymbol{p}}))(\mathsf{H}\tilde{\underline{L}}_{\log}(\tilde{\boldsymbol{p}}))^{-1}\mathsf{H}\tilde{\underline{L}}_\ell(\boldsymbol{p})\boldsymbol{w}),\\ &= -c_\ell[\langle \tilde{\boldsymbol{q}}, \hat{\boldsymbol{v}} \odot \hat{\boldsymbol{v}}\rangle - \langle \tilde{\boldsymbol{q}}, \hat{\boldsymbol{v}}\rangle^2 - \lambda_*^\ell(\hat{\boldsymbol{v}}^\mathsf{T}(\mathsf{H}\tilde{\Phi}(\boldsymbol{q}))^{-1}\hat{\boldsymbol{v}})],\\ &= -c_\ell[\hat{\boldsymbol{v}}^\mathsf{T}(\operatorname{diag}(\tilde{\boldsymbol{q}}) - \tilde{\boldsymbol{q}}\tilde{\boldsymbol{q}}^\mathsf{T})\hat{\boldsymbol{v}} - \lambda_*^\ell(\hat{\boldsymbol{v}}^\mathsf{T}(\mathsf{H}\tilde{\Phi}(\boldsymbol{q}))^{-1}\hat{\boldsymbol{v}})],\\ &= -c_\ell[\hat{\boldsymbol{v}}^\mathsf{T}(\operatorname{diag}(\tilde{\boldsymbol{q}}) - \tilde{\boldsymbol{q}}\tilde{\boldsymbol{q}}^\mathsf{T})\hat{\boldsymbol{v}} - \lambda_*^\ell(\boldsymbol{v}^\mathsf{T}K_{\boldsymbol{q}}(K_{\boldsymbol{q}}K_{\boldsymbol{q}})^{-1}K_{\boldsymbol{q}}\boldsymbol{v})],\\ &= -c_\ell[\boldsymbol{v}^\mathsf{T}K_{\boldsymbol{q}}(\operatorname{diag}(\tilde{\boldsymbol{q}}) - \tilde{\boldsymbol{q}}\tilde{\boldsymbol{q}}^\mathsf{T})K_{\boldsymbol{q}}\boldsymbol{v} - \lambda_*^\ell], \qquad (68)\\ &= -c_\ell[\lambda_*^\Phi - \lambda_*^\ell],\\ &= -c_\ell[\lambda_{\min}(\mathsf{H}\tilde{\Phi}(\boldsymbol{q})(\operatorname{diag}(\tilde{\boldsymbol{q}}) - \tilde{\boldsymbol{q}}\tilde{\boldsymbol{q}}^\mathsf{T})) - \lambda_{\max}(\mathsf{H}\tilde{\underline{L}}_\ell(\tilde{\boldsymbol{p}})(\mathsf{H}\tilde{\underline{L}}_{\log}(\tilde{\boldsymbol{p}}))^{-1})],\end{aligned}$$

where in (68) we used the fact that $\boldsymbol{v}^\mathsf{T}\boldsymbol{v} = 1$. The last equality combined with (66) and (67) shows that $\frac{d}{dt}\langle \boldsymbol{p}, \dot{\beta}(t) - \dot{\gamma}(t)\rangle\Big|_{t=0} < 0$, which completes the proof. ■

C.12. Proof of Corollary 18

Proof From Corollary 17, ℓ is Φ_η-mixable if and only if $\underline{\eta_\ell}\Phi_\eta - \mathrm{S} = \eta^{-1}\underline{\eta_\ell}\Phi - \mathrm{S}$ is convex on Δ_k. When this is the case, Lemma 23 implies that

$$1 \le \eta^{-1}\underline{\eta_\ell}(\inf_{\tilde{\boldsymbol{q}}\in\operatorname{int}\tilde{\Delta}_k} \lambda_{\min}[\mathsf{H}\tilde{\Phi}(\tilde{\boldsymbol{q}})[\mathsf{H}\tilde{\mathrm{S}}(\tilde{\boldsymbol{q}})]^{-1}]), \qquad (69)$$

where we used the facts that $\mathsf{H}(\eta^{-1}\underline{\eta_\ell}\tilde{\Phi}) = \eta^{-1}\underline{\eta_\ell}\mathsf{H}\tilde{\Phi}$, $\lambda_{\min}(\cdot)$ is linear, and $\eta^{-1}\underline{\eta_\ell}$ is independent of $\tilde{\boldsymbol{q}} \in \operatorname{int}\tilde{\Delta}_k$. Inequality 69 shows that the largest η such that ℓ is Φ_η-mixable is given by η_ℓ^Φ in (16). ■

C.13. Proof of Theorem 19

Proof Suppose ℓ is Φ-mixable. Then from Corollary 17, $\underline{\eta_\ell}\Phi - \mathrm{S}$ is convex on Δ_k, and thus $\underline{\eta_\ell} = \eta_\ell^{\mathrm{S}} > 0$ (Corollary 18). Furthermore, $\underline{\eta_\ell}\tilde{\Phi} - \tilde{\mathrm{S}} = [\underline{\eta_\ell}\Phi - \mathrm{S}] \circ \Pi_k$ is convex on $\operatorname{int} \tilde{\Delta}_k$, since Π_k is an affine function. It follows from Lemma 23 and Corollary 18 that

$$\eta_\ell^\Phi = \underline{\eta_\ell} \inf_{\tilde{\boldsymbol{q}} \in \operatorname{int} \tilde{\Delta}_k} \lambda_{\min}(\mathsf{H}\tilde{\Phi}(\tilde{\boldsymbol{q}})(\mathsf{H}\tilde{\mathrm{S}}(\tilde{\boldsymbol{q}}))^{-1}) \geq 1 > 0.$$

Let $\boldsymbol{\mu} \in \operatorname{ri} \Delta_k$ and $\theta_* := \operatorname{argmax}_\theta D_{\mathrm{S}}(\boldsymbol{e}_\theta, \boldsymbol{\mu})$. By definition of an entropy and the fact that the directional derivatives $\Phi'(\boldsymbol{\mu}; \cdot)$ and $\mathrm{S}'(\boldsymbol{\mu}; \cdot)$ are finite on Δ_k (Hiriart-Urruty and Lemaréchal, 2001, Prop. D.1.1.2), it holds that $D_\Phi(\boldsymbol{e}_{\theta_*}, \boldsymbol{\mu}), D_{\mathrm{S}}(\boldsymbol{e}_{\theta_*}, \boldsymbol{\mu}) \in]0, +\infty[$. Therefore, there exists $\alpha > 0$ such that $\alpha^{-1} D_\Phi(\boldsymbol{e}_{\theta_*}, \boldsymbol{\mu}) = D_{\mathrm{S}}(\boldsymbol{e}_{\theta_*}, \boldsymbol{\mu})$. If we let $\Psi := \alpha^{-1}\Phi$, we get

$$D_\Psi(\boldsymbol{e}_{\theta_*}, \boldsymbol{\mu}) = D_{\mathrm{S}}(\boldsymbol{e}_{\theta_*}, \boldsymbol{\mu}). \tag{70}$$

Let $d_\Psi(\tilde{\boldsymbol{q}}) := \tilde{\Psi}(\tilde{\boldsymbol{q}}) - \tilde{\Psi}(\tilde{\boldsymbol{\mu}}) - \langle \tilde{\boldsymbol{q}} - \tilde{\boldsymbol{\mu}}, \nabla \tilde{\Psi}(\tilde{\boldsymbol{\mu}}) \rangle$. Observe that $d_\Psi(\tilde{\boldsymbol{q}}) = \Psi(\boldsymbol{q}) - \Psi(\boldsymbol{\mu}) - \langle \boldsymbol{q} - \boldsymbol{\mu}, \nabla \Psi(\boldsymbol{\mu}) \rangle = D_\Psi(\boldsymbol{q}, \boldsymbol{\mu})$. We define d_{S} similarly. Suppose that $\eta_\ell^\Psi > \eta_\ell^{\mathrm{S}} = \underline{\eta_\ell}$. Then, from Corollary 18, $\forall \tilde{\boldsymbol{q}} \in \operatorname{int} \tilde{\Delta}_k$, $\lambda_{\min}(\mathsf{H}\tilde{\Psi}(\tilde{\boldsymbol{q}})(\mathsf{H}\tilde{\mathrm{S}}(\tilde{\boldsymbol{q}}))^{-1}) > 1$. This implies that $\forall \tilde{\boldsymbol{q}} \in \operatorname{int} \tilde{\Delta}_k$, $\lambda_{\min}(\mathsf{H} d_\Psi(\tilde{\boldsymbol{q}})(\mathsf{H} d_{\mathrm{S}}(\tilde{\boldsymbol{q}}))^{-1}) > 1$, and from Lemma 23, $d_\Psi - d_{\mathrm{S}}$ must be strictly convex on $\operatorname{int} \tilde{\Delta}_k$. We also have $\nabla d_\Psi(\tilde{\boldsymbol{\mu}}) - \nabla d_{\mathrm{S}}(\tilde{\boldsymbol{\mu}}) = 0$ and $d_\Psi(\tilde{\boldsymbol{\mu}}) - d_{\mathrm{S}}(\tilde{\boldsymbol{\mu}}) = 0$. Therefore, $d_\Psi - d_{\mathrm{S}}$ attains a strict minimum at $\tilde{\boldsymbol{\mu}}$ (ibid., Thm. D.2.2.1); that is, $d_\Psi(\tilde{\boldsymbol{q}}) > d_{\mathrm{S}}(\tilde{\boldsymbol{q}})$, $\forall \tilde{\boldsymbol{q}} \in \tilde{\Delta}_k \setminus \{\tilde{\boldsymbol{\mu}}\}$. In particular, for $\tilde{\boldsymbol{q}} = \Pi_k(\boldsymbol{e}_{\theta_*})$, we get $D_\Psi(\boldsymbol{e}_{\theta_*}, \boldsymbol{\mu}) = d_\Psi(\tilde{\boldsymbol{q}}) > d_{\mathrm{S}}(\tilde{\boldsymbol{q}}) = D_{\mathrm{S}}(\boldsymbol{e}_{\theta_*}, \boldsymbol{\mu})$, which contradicts (70). Therefore, $\eta_\ell^\Psi \leq \eta_\ell^{\mathrm{S}}$, and thus

$$\begin{aligned}
R_\ell^{\mathrm{S}}(\boldsymbol{\mu}) = \max\nolimits_\theta D_{\mathrm{S}}(\boldsymbol{e}_\theta, \boldsymbol{\mu})/\eta_\ell^{\mathrm{S}} &= D_{\mathrm{S}}(\boldsymbol{e}_{\theta_*}, \boldsymbol{\mu})/\eta_\ell^{\mathrm{S}}, \\
&\leq D_\Psi(\boldsymbol{e}_{\theta_*}, \boldsymbol{\mu})/\eta_\ell^\Psi, \qquad (71) \\
&\leq \max\nolimits_\theta D_\Psi(\boldsymbol{e}_\theta, \boldsymbol{\mu})/\eta_\ell^\Psi, \\
&= R_\ell^\Psi(\boldsymbol{\mu}), \qquad (72)
\end{aligned}$$

where (71) is due to $D_\Psi(\boldsymbol{e}_{\theta_*}, \boldsymbol{\mu}) = D_{\mathrm{S}}(\boldsymbol{e}_{\theta_*}, \boldsymbol{\mu})$ and $\eta_\ell^\Psi \leq \eta_\ell^{\mathrm{S}}$. Equation 72, implies that $R_\ell^{\mathrm{S}}(\boldsymbol{\mu}) \leq R_\ell^\Phi(\boldsymbol{\mu})$, since $R_\ell^\Psi(\boldsymbol{\mu}) = R_\ell^{\alpha\Phi}(\boldsymbol{\mu}) = R_\ell^\Phi(\boldsymbol{\mu})$ (Reid et al., 2015). Therefore,

$$\forall \boldsymbol{\mu} \in \operatorname{ri} \Delta_k, \; R_\ell^{\mathrm{S}}(\boldsymbol{u}) \leq R_\ell^\Phi(\boldsymbol{u}). \tag{73}$$

It remains to consider the case where $\boldsymbol{\mu}$ is in the relative boundary of Δ_k. Let $\boldsymbol{\mu} \in \operatorname{rbd} \Delta_k$. There exists $\mathcal{I}_0 \subsetneq [k]$ such that $\boldsymbol{\mu} \in \Delta_{\mathcal{I}_0}$. Let $\theta^* \in [k] \setminus \mathcal{I}_0$ and $\mathcal{I} := \mathcal{I}_0 \cup \{\theta^*\}$. It holds that $\boldsymbol{\mu} \in \operatorname{rbd} \Delta_{\mathcal{I}}$ and $\boldsymbol{\mu} + 2^{-1}(\boldsymbol{e}_{\theta^*} - \boldsymbol{\mu}) \in \operatorname{ri} \Delta_{\mathcal{I}}$. Since ℓ is Φ-mixable, it follows from Proposition 11 and the 1-homogeneity of $\Phi'(\boldsymbol{\mu}; \cdot)$ (Hiriart-Urruty and Lemaréchal, 2001, Prop. D.1.1.2) that

$$\Phi'(\boldsymbol{\mu}; \boldsymbol{e}_{\theta^*} - \boldsymbol{\mu}) = 2\Phi'(\boldsymbol{\mu}; [\boldsymbol{\mu} + 2^{-1}(\boldsymbol{e}_{\theta^*} - \boldsymbol{\mu})] - \boldsymbol{\mu}) = -\infty.$$

Hence,

$$\begin{aligned}
R_\ell^\Phi(\boldsymbol{\mu}) &= \max\nolimits_{\theta \in [k]} D_\Phi(\boldsymbol{e}_\theta, \boldsymbol{\mu}), \\
&\geq D_\Phi(\boldsymbol{e}_{\theta^*}, \boldsymbol{\mu}) = \Phi(\boldsymbol{e}_{\theta^*}) - \Phi(\boldsymbol{\mu}) - \Phi'(\boldsymbol{\mu}; \boldsymbol{e}_{\theta^*} - \boldsymbol{\mu}) = +\infty. \qquad (74)
\end{aligned}$$

Inequality 74 also applies to S, since ℓ is $(\underline{\eta_\ell}^{-1}\,\mathrm{S})$-mixable. From (74) and (73), we conclude that $\forall \boldsymbol{\mu} \in \Delta_k, \; R_\ell^{\mathrm{S}}(\boldsymbol{\mu}) \leq R_\ell^\Phi(\boldsymbol{\mu})$. ■

Appendix D. Legendre Φ, but no Φ-mixable ℓ

In this appendix, we construct a *Legendre type* entropy (Rockafellar, 1997) for which there are no Φ-mixable losses satisfying a weak condition (see below).

Let $\ell : \mathcal{A} \to [0, +\infty]^n$ be a loss satisfying condition I. According to Alexandrov's Theorem, a concave function is twice differentiable almost everywhere (see e.g. (Borwein et al., 2010, Thm. 6.7)). Now we give a version of Theorem 16 which does not assume the twice differentiability of the Bayes risk. The proof is almost identical to that of Theorem 16 with only minor modifications.

Theorem 38 *Let $\Phi\colon \mathbb{R}^k \to \mathbb{R} \cup \{+\infty\}$ be an entropy such that $\tilde{\Phi}$ is twice differentiable on* $\operatorname{int} \tilde{\Delta}_k$*, and $\ell\colon \mathcal{A} \to [0, +\infty]^n$ a loss satisfying Condition I and such that $\exists(\tilde{\boldsymbol{p}}, \boldsymbol{v}) \in \mathcal{D} \times \mathbb{R}^{\tilde{n}}, \mathsf{H}\underline{\tilde{L}}_\ell(\tilde{\boldsymbol{p}})\boldsymbol{v} \neq \mathbf{0}_{\tilde{n}}$, where $\mathcal{D} \subset \operatorname{int} \tilde{\Delta}_n$ is a set of Lebesgue measure 1 where $\underline{\tilde{L}}_\ell$ is twice differentiable, and define*

$$\underline{\eta_\ell}^* := \inf_{\tilde{\boldsymbol{p}} \in \mathcal{D}} (\lambda_{\max}([\mathsf{H}\underline{\tilde{L}}_{\log}(\tilde{\boldsymbol{p}})]^{-1} \mathsf{H}\underline{\tilde{L}}_\ell(\tilde{\boldsymbol{p}})))^{-1}. \tag{75}$$

Then ℓ is Φ-mixable only if $\underline{\eta_\ell}^* \Phi - \mathrm{S}$ is convex on Δ_k.

The new condition on the Bayes risk is much weaker than requiring $\underline{L}_\ell$ to be twice differentiability on $]0, +\infty[^n$. In the next example, we will show that there exists a Legendre type entropy for which there are no Φ-mixable losses satisfying the condition of Theorem 38.

Example 39 *Let $\Phi : \mathbb{R}^2 \to \mathbb{R} \cup \{+\infty\}$ be an entropy such that*

$$\forall q \in]0, 1[,\ \Phi(q, 1 - q) = \tilde{\Phi}(q) = \int_{1/2}^{q} \log\left(\frac{\log(1 - t)}{\log t}\right) dt.$$

$\tilde{\Phi}$ is differentiable and strictly convex on the open set $(0, 1)$. Furthermore, it satisfies (8) which makes it a function of Legendre type (Rockafellar, 1997, Lem. 26.2). In fact, (8) is satisfied due to

$$\left|\frac{d}{dq}\tilde{\Phi}(q)\right| = \left|\log\left(\frac{\log(1 - q)}{\log q}\right)\right| \overset{q \to b}{\to} +\infty,\ \text{where } b \in \{0, 1\},$$
$$\frac{d^2}{dq^2}\tilde{\Phi}(q) = \frac{-1}{q \log q} + \frac{-1}{(1 - q)\log(1 - q)} > 0,\ \forall q \in]0, 1[.$$

The Shannon entropy on Δ_2 is defined by $\mathrm{S}(q, 1 - q) = \tilde{\mathrm{S}}(q) = q \log q + (1 - q) \log(1 - q)$, for $q \in]0, 1[$. Thus, $\frac{d^2}{dq^2}\tilde{\mathrm{S}}(q) = \frac{1}{q(1-q)}$.

Suppose now that there exists a Φ-mixable loss $\ell\colon \mathcal{A} \to [0, +\infty]^n$ satisfying condition I and such that $\exists(\tilde{\boldsymbol{p}}, \boldsymbol{v}) \in \mathcal{D} \times \mathbb{R}^{\tilde{n}}, \mathsf{H}\underline{\tilde{L}}_\ell(\tilde{\boldsymbol{p}})\boldsymbol{v} \neq \mathbf{0}_{\tilde{n}}$. Let $\underline{\eta_\ell}^*$ be as in (75). By definition, we have $\underline{\eta_\ell}^* < +\infty$, and thus

$$\underline{\eta_\ell}^* \left[\frac{d^2}{dq^2}\tilde{\Phi}(q)\right]\left[\frac{d^2}{dq^2}\tilde{\mathrm{S}}(q)\right]^{-1} = \underline{\eta_\ell}^* \left(\frac{q - 1}{\log q} + \frac{-q}{\log(1 - q)}\right) \overset{q \to b}{\to} 0, \tag{76}$$

where $b \in \{0, 1\}$. From Lemma 23, (76) implies that $\underline{\eta_\ell}^* \Phi - \mathrm{S}$ is not convex on Δ_k, which is a contradiction according to Theorem 38.

Appendix E. Loss Surface and Superprediction Set

In this appendix, we derive an expression for the curvature of the image of a proper loss function. We will need the following lemma.

Lemma 40 *Let $\sigma : [0, +\infty[^n \to \mathbb{R}$ be a 1-homogeneous, twice differentiable function on $]0, +\infty[^n$. Then σ is concave on $]0, +\infty[^n$ if and only if $\tilde{\sigma} = \sigma \circ \Pi_n$ is concave on $\operatorname{int} \tilde{\Delta}_n$.*

Proof The forward implication is immediate; if σ is concave on $]0, +\infty[^n$, then $\sigma \circ \Pi_k$ is concave on $\operatorname{int} \tilde{\Delta}_k$, since Π_k is an affine function.

Now assume that $\tilde{\sigma}$ is concave on $\operatorname{int} \tilde{\Delta}_k$. Let $\lambda \in [0, 1]$ and $(\boldsymbol{p}, \boldsymbol{q}) \in [0, +\infty[^n \times [0, +\infty[^n$. We need to show that

$$\lambda\sigma(\boldsymbol{p}) + (1 - \lambda)\sigma(\boldsymbol{q}) \leq \sigma(\lambda\boldsymbol{p} + (1 - \lambda)\boldsymbol{q}). \tag{77}$$

Note that if $\boldsymbol{p} = \mathbf{0}$ or $\boldsymbol{q} = \mathbf{0}$, (77) is trivially with equality due to the 1-homogeneity of σ. Now assume that $\boldsymbol{p}$ and $\boldsymbol{q}$ are non-zero and let $c := \lambda \|\boldsymbol{p}\|_1 + (1 - \lambda) \|\boldsymbol{q}\|_1$. For convenience, we also denote $\boldsymbol{p}_1 = \boldsymbol{p}/ \|\boldsymbol{p}\|_1$ and $\boldsymbol{q}_1 = \boldsymbol{q}/ \|\boldsymbol{q}\|_1$ which are both in Δ_n. It follows that

$$\begin{aligned}
\lambda\sigma(\boldsymbol{p}) + (1 - \lambda)\sigma(\boldsymbol{q}) &= cM \left(\lambda\frac{\|\boldsymbol{p}\|_1}{c}\sigma(\boldsymbol{p}_1) + (1 - \lambda)\frac{\|\boldsymbol{q}\|_1}{c}\sigma(\boldsymbol{q}_1)\right), \\
&= c \left(\lambda\frac{\|\boldsymbol{p}\|_1}{c}\tilde{\sigma}(\tilde{\boldsymbol{p}}_1) + (1 - \lambda)\frac{\|\boldsymbol{q}\|_1}{c}\tilde{\sigma}(\tilde{\boldsymbol{q}}_1)\right), \\
&\leq c\tilde{\sigma} \left(\lambda\frac{\|\boldsymbol{p}\|_1}{c}\tilde{\boldsymbol{p}}_1 + (1 - \lambda)\frac{\|\boldsymbol{q}\|_1}{c}\tilde{\boldsymbol{q}}_1\right), \\
&= c\sigma \left(\lambda\frac{\|\boldsymbol{p}\|_1}{c}\boldsymbol{p}_1 + (1 - \lambda)\frac{\|\boldsymbol{q}\|_1}{c}\boldsymbol{q}_1\right), \\
&= \sigma(\lambda\boldsymbol{p} + (1 - \lambda)\boldsymbol{q}),
\end{aligned}$$

where the first and last equalities are due the 1-homogeneity of σ and the inequality is due to $\tilde{\sigma}$ being concave on the $\operatorname{int} \tilde{\Delta}_n$. ■

E.1. Curvature of the Loss Surface

The *normal curvature* of a $\tilde{n}$-manifold $\mathcal{S}$ (Thorpe, 1994) at a point $\boldsymbol{r} \in \mathcal{S}$ in the direction of $\boldsymbol{w} \in T_{\boldsymbol{r}}\mathcal{S}$, where $T_{\boldsymbol{r}}\mathcal{S}$ is the *tangent space* of $\mathcal{S}$ at $\boldsymbol{r} \in \mathcal{S}$, is defined by

$$\kappa(\boldsymbol{r}, \boldsymbol{w}) = \frac{\langle \boldsymbol{w}, \mathsf{DN}^{\mathcal{S}}(\boldsymbol{r})\boldsymbol{w}\rangle}{\langle \boldsymbol{w}, \boldsymbol{w}\rangle}, \tag{78}$$

where $\mathsf{N}^{\mathcal{S}}(\boldsymbol{r})$ is the normal vector to the surface at $\boldsymbol{r}$. The *minimum principal curvature* of $\mathcal{S}$ at $\boldsymbol{r}$ is expressed as $\underline{\kappa}(\boldsymbol{r}) := \inf\{\kappa(\boldsymbol{r}, \boldsymbol{w}) : \boldsymbol{w} \in T_{\boldsymbol{r}}\mathcal{S} \cap \mathcal{B}(\boldsymbol{r}, 1)\}$.

In the next theorem, we establish a direct link between the curvature of a loss surface and the Hessian of the loss' Bayes risk.

Theorem 41 *Let $\ell\colon \operatorname{ri} \Delta_n \to [0, +\infty[^n$ be a loss whose Bayes risk is twice differentiable and strictly concave on $]0, +\infty[^n$. Let $\boldsymbol{p} \in \operatorname{ri} \Delta_n$, $X_{\boldsymbol{p}} := I_{\tilde{n}} - \tilde{\boldsymbol{p}}\mathbf{1}_{\tilde{n}}^{\mathsf{T}}$, and $\boldsymbol{w} \in T_{\tilde{\ell}(\tilde{\boldsymbol{p}})}\mathcal{S}_{\ell}$. Then*

1. $\exists \boldsymbol{v} \in \mathbb{R}^{n-1}$ *such that* $\mathsf{D}\tilde{\ell}(\tilde{\boldsymbol{p}})\boldsymbol{v} = \boldsymbol{w}$.
2. $\mathcal{S}_\ell$ *is a* $\tilde{n}$*-manifold.*
3. *The normal curvature of* $\mathcal{S}_\ell$ *at* $\ell(\boldsymbol{p}) = \tilde{\ell}(\tilde{\boldsymbol{p}})$ *in the direction* $\boldsymbol{w}$ *is given by*

$$\kappa_\ell(\ell(\boldsymbol{p}), \boldsymbol{w}) = \left\| \begin{bmatrix} X_{\boldsymbol{p}} \\ -\tilde{\boldsymbol{p}}^\mathsf{T} \end{bmatrix} (-\mathsf{H}\underline{\tilde{L}}_\ell(\tilde{\boldsymbol{p}}))^{\frac{1}{2}} \boldsymbol{u} \right\|^{-1}, \tag{79}$$

where $\boldsymbol{u} = (-\mathsf{H}\underline{\tilde{L}}_\ell(\tilde{\boldsymbol{p}}))^{\frac{1}{2}}\boldsymbol{v} / \|(-\mathsf{H}\underline{\tilde{L}}_\ell(\tilde{\boldsymbol{p}}))^{\frac{1}{2}}\boldsymbol{v}\|$.

It becomes clear from (79) that smaller eigenvalues of $-\mathsf{H}\underline{\tilde{L}}_\ell(\tilde{\boldsymbol{p}})$ will tend to make the loss surface more curved at $\ell(\boldsymbol{p})$, and vice versa.

Before proving Theorem 41, we first define parameterizations on manifolds.

Definition 42 (Local and Global Parameterization) *Let* $\mathcal{S} \subseteq \mathbb{R}^n$ *be a* $\tilde{n}$*-manifold and* $\mathcal{U}$ *an open set in* $\mathbb{R}^{\tilde{n}}$*. The map* $\varphi : \mathcal{U} \to \mathcal{S}$ *is called a* local parameterization *of* $\mathcal{S}$ *if* $\mathsf{D}\varphi(\boldsymbol{u}) : \mathbb{R}^{\tilde{n}} \to T_{\varphi(\boldsymbol{u})}\mathcal{S}$ *is injective for all* $\boldsymbol{u} \in \mathcal{U}$*, where* $T_{\varphi(\boldsymbol{u})}\mathcal{S}$ *is the tangent space of* $\mathcal{S}$ *at* $\varphi(\boldsymbol{u}) \in \mathcal{S}$*.* φ *is called a* global *parameterization of* $\mathcal{S}$ *if it is, additionally, onto.*

Let φ be a global parameterization of $\mathcal{S}$ and $\mathsf{N}^\varphi := \mathsf{N}^{\mathcal{S}} \circ \varphi$. By a direct application of the chain rule, (78) can be written as

$$\kappa(\varphi(\boldsymbol{u}), \boldsymbol{w}) = \frac{\langle \boldsymbol{w}, \mathsf{D}\mathsf{N}^\varphi(\boldsymbol{u})\boldsymbol{v} \rangle}{\langle \boldsymbol{w}, \boldsymbol{w} \rangle}, \tag{80}$$

where $\boldsymbol{v}$ is such that $\mathsf{D}\varphi(\boldsymbol{u})\boldsymbol{v} = \boldsymbol{w}$. The existence of such a $\boldsymbol{v}$ is guaranteed by the fact that $\mathsf{D}\varphi$ is injective and $\dim \mathbb{R}^{\tilde{n}} = \dim T_{\varphi(\boldsymbol{u})}\mathcal{S} = \tilde{n}$.

Proof [**Theorem 41**] First we show that $\mathcal{S}_\ell$ is a $\tilde{n}$-manifold. Consider the map $\tilde{\ell} : \operatorname{int} \tilde{\Delta}_n \to \mathcal{S}_\ell$ and note that $\operatorname{int} \tilde{\Delta}_n$ is trivially a $\tilde{n}$-manifold. Due to the strict concavity of the Bayes risk, $\tilde{\ell}$ is injective (van Erven et al., 2012) and from Lemmas 27 and 40, $\mathsf{D}\tilde{\ell}(\tilde{\boldsymbol{p}}) : \mathbb{R}^{\tilde{n}} \to T_{\tilde{\ell}(\tilde{\boldsymbol{p}})}\mathcal{S}_\ell$ is also injective. Therefore, $\tilde{\ell}$ is an *immersion* (Robbin and Salamon, 2011). $\tilde{\ell}$ is also *proper* in the sense that the preimage of every compact subset of $\mathcal{S}_\ell$ is compact. Therefore, $\tilde{\ell}$ is a proper injective immersion, and thus it is an embedding from the $\tilde{n}$-manifold $\operatorname{int} \tilde{\Delta}_n$ to $\mathcal{S}_\ell$ (ibid.). Hence, $\mathcal{S}_\ell$ is a manifold.

Now we prove (79). The map $\tilde{\ell}$ is a global parameterization of $\mathcal{S}_\ell$. In fact, from Lemma 27, $\mathsf{D}\tilde{\ell}(\tilde{\boldsymbol{p}})$ has rank $\tilde{n}$, for all $\tilde{\boldsymbol{p}} \in \operatorname{int} \tilde{\Delta}_n$, which implies that $\mathsf{D}\tilde{\ell}(\tilde{\boldsymbol{p}})$ is onto from $\mathbb{R}^{\tilde{n}}$ to $T_{\tilde{\ell}(\tilde{\boldsymbol{p}})}\mathcal{S}_\ell$. Therefore, given $\boldsymbol{w} \in T_{\tilde{\ell}(\tilde{\boldsymbol{p}})}\mathcal{S}_\ell$, there exists $\boldsymbol{v} \in \mathbb{R}^{\tilde{n}}$ such that $\boldsymbol{w} = \mathsf{D}\tilde{\ell}(\tilde{\boldsymbol{p}})\boldsymbol{v}$. Furthermore, Lemma 27 implies that $\mathsf{N}^{\tilde{\ell}}(\tilde{\boldsymbol{p}}) = \boldsymbol{p}$, since $\langle \boldsymbol{p}, \mathsf{D}\tilde{\ell}(\tilde{\boldsymbol{p}}) \rangle = \mathbf{0}_{\tilde{n}}^\mathsf{T}$. Substituting $\mathsf{N}^{\tilde{\ell}}$ into (80) yields

$$\kappa_\ell(\tilde{\ell}(\tilde{\boldsymbol{p}}),\boldsymbol{w}) = \frac{\boldsymbol{v}^\mathsf{T}(\mathsf{D}\tilde{\ell}(\tilde{\boldsymbol{p}}))^\mathsf{T}\begin{bmatrix} I_{\tilde{n}}, \\ \mathbf{1}_{\tilde{n}} \end{bmatrix}\boldsymbol{v}}{\left\langle \mathsf{D}\tilde{\ell}(\tilde{\boldsymbol{p}})\boldsymbol{v}, \mathsf{D}\tilde{\ell}(\tilde{\boldsymbol{p}})\boldsymbol{v}\right\rangle},$$

$$= \frac{\boldsymbol{v}^\mathsf{T}\mathsf{H}\underline{\tilde{L}}_\ell(\tilde{\boldsymbol{p}})\left[X_{\boldsymbol{p}}^\mathsf{T}, \quad -\tilde{\boldsymbol{p}}\right]\begin{bmatrix} I_{\tilde{n}} \\ \mathbf{1}_{\tilde{n}} \end{bmatrix}\boldsymbol{v}}{\left\langle \mathsf{D}\tilde{\ell}(\tilde{\boldsymbol{p}})\boldsymbol{v}, \mathsf{D}\tilde{\ell}(\tilde{\boldsymbol{p}})\boldsymbol{v}\right\rangle},$$

$$= \frac{\boldsymbol{v}^\mathsf{T}\mathsf{H}\underline{\tilde{L}}_\ell(\tilde{\boldsymbol{p}})\boldsymbol{v}}{\boldsymbol{v}^\mathsf{T}\mathsf{H}\underline{\tilde{L}}_\ell(\tilde{\boldsymbol{p}})\left[X_{\boldsymbol{p}}^\mathsf{T}, \quad -\tilde{\boldsymbol{p}}\right]\begin{bmatrix} X_{\boldsymbol{p}} \\ -\tilde{\boldsymbol{p}}^\mathsf{T} \end{bmatrix}\mathsf{H}\underline{\tilde{L}}_\ell(\tilde{\boldsymbol{p}})\boldsymbol{v}}. \tag{81}$$

Setting $\boldsymbol{u} = (-\mathsf{H}\underline{\tilde{L}}_\ell(\tilde{\boldsymbol{p}}))^{\frac{1}{2}}\boldsymbol{v}/\|(-\mathsf{H}\underline{\tilde{L}}_\ell(\tilde{\boldsymbol{p}}))^{\frac{1}{2}}\boldsymbol{v}\|$ in (81) gives the desired result. ■

Appendix F. Classical Mixability

In this appendix, we provide a more concise proof of the necessary and sufficient conditions for the convexity of the superprediction set (van Erven et al., 2012).

Theorem 43 *Let $\ell\colon \Delta_n \to [0,+\infty[^n$ be a strictly proper loss whose Bayes risk is twice differentiable on $]0,+\infty[^n$. The following points are equivalent;*

(i) $\forall \tilde{\boldsymbol{p}} \in \operatorname{int}\tilde{\Delta}_n,\ \eta\mathsf{H}\underline{\tilde{L}}_\ell(\tilde{\boldsymbol{p}}) \succeq \mathsf{H}\underline{\tilde{L}}_{\log}(\tilde{\boldsymbol{p}})$.

(ii) $e^{-\eta\mathcal{S}_\ell^\oplus} = \bigcap_{\boldsymbol{p}\in\Delta_n}\mathcal{H}_{\tau(\boldsymbol{p}),1}\bigcap[0,+\infty[^n$*, where* $\tau(\boldsymbol{p}) := \boldsymbol{p}\odot e^{\eta\ell(\boldsymbol{p})}$.

(iii) $e^{-\eta\mathcal{S}_\ell^\oplus}$ *is convex.*

Proof We already showed (i) $\implies$ (ii) $\implies$ (iii) in the proof of Theorem 7.

We now show (iii) $\implies$ (i). Since $e^{-\eta\mathcal{S}_\ell^\oplus}$ is convex, any point $\boldsymbol{s} \in \operatorname{bd} e^{-\eta\mathcal{S}_\ell^\oplus}$ is supported by a hyperplane (Hiriart-Urruty and Lemaréchal, 2001, Lem. A.4.2.1). Since $\boldsymbol{u} \to e^{-\eta\boldsymbol{u}}$ is a homeomorphism, it maps boundaries to boundaries. From this and Lemma 33, $\operatorname{bd} e^{-\eta\mathcal{S}_\ell^\oplus} = e^{-\eta\mathcal{S}_\ell}$. Thus, for $\boldsymbol{p} \in \operatorname{ri}\Delta_n$, there exists a unit-norm vector $\boldsymbol{u}\in\mathbb{R}^n$ such that for all $\boldsymbol{s}\in\mathcal{S}_\ell^\oplus$ it either holds that $\langle \boldsymbol{u}, e^{-\eta\ell(\boldsymbol{p})}\rangle \leq \langle \boldsymbol{u}, e^{-\eta\boldsymbol{s}}\rangle$; or $\langle \boldsymbol{u}, e^{-\eta\ell(\boldsymbol{p})}\rangle \geq \langle \boldsymbol{u}, e^{-\eta\boldsymbol{s}}\rangle$. It is easy to see that it is the latter case that holds, since we can choose $\boldsymbol{s} = \ell(\boldsymbol{r}) + c\mathbf{1} \in \mathcal{S}_\ell^\oplus$, for $\boldsymbol{r}\in\Delta_n$, and make $\langle \boldsymbol{u}, e^{-\eta\boldsymbol{s}}\rangle$ arbitrarily small by making $c\in\mathbb{R}$ large. Therefore, $\forall \boldsymbol{r}\in\operatorname{ri}\Delta_n$, $\langle \boldsymbol{u}, e^{-\eta\tilde{\ell}(\tilde{\boldsymbol{p}})}\rangle = \langle \boldsymbol{u}, e^{-\eta\ell(\boldsymbol{p})}\rangle \geq \langle \boldsymbol{u}, e^{-\eta\ell(\boldsymbol{r})}\rangle = \langle \boldsymbol{u}, e^{-\eta\tilde{\ell}(\tilde{\boldsymbol{r}})}\rangle$ and $\tilde{\boldsymbol{p}}$ is a critical point of the function $f(\tilde{\boldsymbol{r}}) := \langle \boldsymbol{u}, e^{-\eta\tilde{\ell}(\tilde{\boldsymbol{r}})}\rangle$ on $\operatorname{int}\tilde{\Delta}_n$. This implies that $\nabla f(\tilde{\boldsymbol{p}}) = \mathbf{0}_{\tilde{n}}$; that is, $-\eta\langle \boldsymbol{u}, \operatorname{diag}(e^{-\eta\tilde{\ell}(\tilde{\boldsymbol{p}})})\mathsf{D}\tilde{\ell}(\tilde{\boldsymbol{p}})\rangle = -\eta\langle \operatorname{diag}(e^{-\eta\tilde{\ell}(\tilde{\boldsymbol{p}})})\boldsymbol{u}, \mathsf{D}\tilde{\ell}(\tilde{\boldsymbol{p}})\rangle = \mathbf{0}_{\tilde{n}}^\mathsf{T}$. From Lemma 27, there exists $\lambda\in\mathbb{R}$ such that $\operatorname{diag}(e^{-\eta\tilde{\ell}(\tilde{\boldsymbol{p}})})\boldsymbol{u} = \lambda\boldsymbol{p}$. Therefore, $\boldsymbol{u} = \lambda\boldsymbol{p}\odot e^{\eta\tilde{\ell}(\tilde{\boldsymbol{p}})}$, where $\lambda = \|\boldsymbol{p}\odot e^{\eta\tilde{\ell}(\tilde{\boldsymbol{p}})}\|^{-1}$, since $\|\boldsymbol{u}\| = 1$. For $\boldsymbol{v}\in\mathbb{R}^{n-1}$, let $\tilde{\boldsymbol{\alpha}}^t := \tilde{\boldsymbol{p}} + t\boldsymbol{v}$, where $t \in \{s : \tilde{\boldsymbol{p}} + s\boldsymbol{v} \in$

$\operatorname{int}\tilde{\Delta}_n\}$. Since f is twice differentiable and attains a maximum at $\tilde{\boldsymbol{p}}$,

$$
\begin{aligned}
0 \geq \frac{1}{\lambda\eta}\,\frac{d^2}{dt^2} f \circ \tilde{\boldsymbol{\alpha}}^t\Big|_{t=0} &= \frac{1}{\lambda}\,\frac{d}{dt}\left\langle \boldsymbol{u}, \operatorname{diag} e^{-\eta\tilde{\ell}(\tilde{\boldsymbol{\alpha}}^t)}\mathsf{D}\tilde{\ell}(\tilde{\boldsymbol{\alpha}}^t)\boldsymbol{v}\right\rangle\Big|_{t=0}, \\
&= \frac{d}{dt}\left\langle \boldsymbol{p}\odot e^{\eta\tilde{\ell}(\tilde{\boldsymbol{p}})}, \operatorname{diag} e^{-\eta\tilde{\ell}(\tilde{\boldsymbol{\alpha}}^t)}\mathsf{D}\tilde{\ell}(\tilde{\boldsymbol{p}})\boldsymbol{v}\right\rangle\Big|_{t=0} + \frac{d}{dt}\left\langle \boldsymbol{p}, \mathsf{D}\tilde{\ell}(\tilde{\boldsymbol{\alpha}}^t)\boldsymbol{v}\right\rangle\Big|_{t=0}, \\
&= \eta\boldsymbol{v}^\mathsf{T}\mathsf{H}\underline{\tilde{L}}_\ell(\tilde{\boldsymbol{p}})(\mathsf{H}\underline{\tilde{L}}_{\log}(\tilde{\boldsymbol{p}}))^{-1}\mathsf{H}\underline{\tilde{L}}_\ell(\tilde{\boldsymbol{p}})\boldsymbol{v} - \boldsymbol{v}^\mathsf{T}\mathsf{H}\underline{\tilde{L}}_\ell(\tilde{\boldsymbol{p}})\boldsymbol{v}, \qquad (82)
\end{aligned}
$$

where in the second equality we substituted $\boldsymbol{u}$ by $\lambda\boldsymbol{p}\odot e^{\eta\tilde{\ell}(\tilde{\boldsymbol{p}})}$ and in (82) we used (19) and (20) from Lemma 28. Note that by the assumptions on ℓ it follows that the Bayes risk $\underline{\tilde{L}}_\ell$ is strictly concave (van Erven et al., 2012, Lemma 6) and $-\mathsf{H}\underline{\tilde{L}}_\ell(\tilde{\boldsymbol{p}})$ is symmetric negative-definite. In particular, $\mathsf{H}\underline{\tilde{L}}_\ell(\tilde{\boldsymbol{p}})$ is invertible. Setting $\hat{\boldsymbol{v}} := \mathsf{H}\underline{\tilde{L}}_\ell(\tilde{\boldsymbol{p}})\boldsymbol{v}$ in (82) yields

$$0 \geq \eta\hat{\boldsymbol{v}}(\mathsf{H}\underline{\tilde{L}}_{\log}(\tilde{\boldsymbol{p}}))^{-1}\hat{\boldsymbol{v}} - \hat{\boldsymbol{v}}(\mathsf{H}\underline{\tilde{L}}_\ell(\tilde{\boldsymbol{p}}))^{-1}\hat{\boldsymbol{v}}.$$

Since $\boldsymbol{v} \in \mathbb{R}^{n-1}$ was chosen arbitrarily, $(\mathsf{H}\underline{\tilde{L}}_\ell(\tilde{\boldsymbol{p}}))^{-1} \succeq \eta(\mathsf{H}\underline{\tilde{L}}_{\log}(\tilde{\boldsymbol{p}}))^{-1}, \forall\tilde{\boldsymbol{p}} \in \operatorname{int}\tilde{\Delta}_n$. This is equivalent to the condition $\forall\tilde{\boldsymbol{p}} \in \operatorname{int}\tilde{\Delta}_n,\ \eta\mathsf{H}\underline{\tilde{L}}_\ell(\tilde{\boldsymbol{p}}) \succeq \mathsf{H}\underline{\tilde{L}}_{\log}(\tilde{\boldsymbol{p}})$. ■