JMLR: Workshop and Conference Proceedings vol 40:1–22, 2015

Generalized Mixability via Entropic Duality

Mark D. Reid MARK.REID@ANU.EDU.AU
Australian National University & NICTA

Rafael M. Frongillo RAF@CS.BERKELEY.EDU
Harvard University

Robert C. Williamson BOB.WILLIAMSON@ANU.EDU.AU
Nishant Mehta NISHANT.MEHTA@ANU.EDU.AU
Australian National University & NICTA

Abstract

Mixability is a property of a loss which characterizes when constant regret is possible in the game of prediction with expert advice. We show that a key property of mixability generalizes, and the $\exp$ and $\log$ operations present in the usual theory are not as special as one might have thought. In doing so we introduce a more general notion of Φ-mixability where Φ is a general entropy (*i.e.*, any convex function on probabilities). We show how a property shared by the convex dual of any such entropy yields a natural algorithm (the minimizer of a regret bound) which, analogous to the classical Aggregating Algorithm, is guaranteed a constant regret when used with Φ-mixable losses. We characterize which Φ have non-trivial Φ-mixable losses and relate Φ-mixability and its associated Aggregating Algorithm to potential-based methods, a Blackwell-like condition, mirror descent, and risk measures from finance. We also define a notion of "dominance" between different entropies in terms of bounds they guarantee and conjecture that classical mixability gives optimal bounds, for which we provide some supporting empirical evidence.

Keywords: online learning, prediction with expert advice, convex analysis, aggregating algorithm

1. Introduction

The combination or aggregation of predictions is central to machine learning. Traditional Bayesian updating can be viewed as a particular way of aggregating information that takes account of prior information. This is known to be special case of more general and decision theoretic "aggregating algorithms" (Vovk, 2001) which take into account loss functions when evaulating predictions. As recent work by Gravin et al. (2014) shows, there are still a number of open questions about the optimal algorithms for the aggregation of predictions from a finite number of experts. In this paper, we attempt to address these by refining and generalizing the notion of "mixability" (Vovk, 1990, 1998), which plays a central role in the theory of prediction with expert advice, characterizing the optimal learning rates in the asymptotic case of infinitely many experts.

We show there is an implicit design variable in mixability that to date has not been fully exploited. The aggregating algorithm makes use of a divergence between the current distribution and a prior which serves as a regularizer. In particular the aggregating algorithm uses the KL-divergence. We consider the general setting of an arbitrary loss and an arbitrary regularizer (in the form of a Bregman divergence) and show that we recover the core technical result of traditional mixability: if a loss is mixable in our generalized sense then there is a generalized aggregating algorithm which

can be guaranteed to have constant regret. The generalized aggregating algorithm is developed by optimizing the bound that defines our new notion of mixability.

Our approach relies heavily on the titular dual representations of convex "generalized entropy" functions defined for distributions over a fixed number of experts. By doing so we gain new insight into why the original mixability argument works and a broader understanding of when constant regret guarantees are possible. In addition, we also make a number of interesting connections between mixability, mirror descent algorithms, and notions of risk from mathematical finance.

1.1. Mixability in Prediction With Expert Advice Games

A prediction with expert advice game is defined by its loss, a collection of experts that the player must compete against, and a fixed number of rounds. Each round the experts reveal their predictions to the player and then the player makes a prediction. An observation is then revealed to the experts and the player and all receive a penalty determined by the loss. The aim of the player is to keep its total loss close to that of the best expert once all the rounds have completed. The difference between the total loss of the player and the total loss of the best expert is called the regret and is typically the focus of the analysis of this style of game. In particular, we are interested in when the regret is *constant*, that is, independent of the number of rounds played.

More formally, let X denote a set of possible *observations* and let $\mathcal{A}$ denote a set of *actions* or *predictions* the experts and player can perform. A *loss* $\ell : \mathcal{A} \to \mathbb{R}^X$ assigns the penalty $\ell_x(a)$ to predicting $a \in \mathcal{A}$ when $x \in X$ is observed. The finite set of experts is denoted[1] Θ and the set of distributions Θ is denoted Δ_Θ. In each round $t = 1, \dots, T$, each expert $\theta \in \Theta$ makes a prediction $a_\theta^t \in \mathcal{A}$. These are revealed to the player who makes a prediction $\hat{a}^t \in \mathcal{A}$. Once observation $x^t \in X$ is revealed the experts receive loss $\ell_{x^t}(a_\theta^t)$ and the player receives loss $\ell_{x^t}(\hat{a}^t)$. The aim of the player is to minimize its *regret* $\mathrm{Regret}(T) := L^T - \min_\theta L_\theta^T$ where $L^T := \sum_{t=1}^T \ell_{x^t}(\hat{a}^t)$ and $L_\theta^T = \sum_{t=1}^T \ell_{x^t}(a_\theta^t)$. We will say the game has *constant regret* if there exists a player who can always make predictions that guarantee $\mathrm{Regret}(T) \le R_{\ell,\Theta}$ for all T and all expert predictions $\{a_\theta^t\}_{t=1}^T$ where $R_{\ell,\Theta}$ is a constant that may depend on ℓ and Θ.

Vovk (1990, 1998) showed that if the loss for a game satisfies a condition called mixability then a player making predictions using the aggregating algorithm (AA) will achieve constant regret.

Definition 1 (Mixability and the Aggregating Algorithm) *Given $\eta > 0$, a loss $\ell : \mathcal{A} \to \mathbb{R}^X$ is* η-mixable *if, for all expert predictions $a_\theta \in \mathcal{A}$, $\theta \in \Theta$ and all mixture distributions $\mu \in \Delta_\Theta$ over experts there exists a prediction $\hat{a} \in \mathcal{A}$ such that for all outcomes $x \in X$,*

$$\ell_x(\hat{a}) \le -\eta^{-1} \log \sum_{\theta \in \Theta} \exp\left(-\eta \ell_x(a_\theta)\right) \mu_\theta. \tag{1}$$

The aggregating algorithm *starts with a mixture $\mu^0 \in \Delta_\Theta$ over experts. In round t, experts predict a_θ^t and the player predicts the $\hat{a}^t \in \mathcal{A}$ guaranteed by the η-mixability of ℓ so that (1) holds for $\mu = \mu^{t-1}$ and $a_\theta = a_\theta^t$. Upon observing x^t, the mixture $\mu^t \in \Delta_\Theta$ is set so that $\mu_\theta^t \propto \mu_\theta^{t-1} e^{-\eta \ell_{x^t}(a_\theta^t)}$.*

We note that our definition of mixability differs from the original in (Vovk, 1998) and instead follows the presentation of mixability in (Cesa-Bianchi and Lugosi, 2006). In particular, the original definition does not assume a fixed number of experts but instead quantifies (1) over all *simple*

1. We use this notation to emphasize two points: 1) that expert predictions are parametric models $p(x|\theta)$ in the case of Bayesian updating; and 2) many of our results generalize to infinite experts (cf. (Vovk, 1998, App. A)).

distributions (*i.e.*, with finite support) over actions, not experts. This departure from the original definition means that our definition of mixability depends on both the loss and the number of experts rather than the loss alone. Crucially, this distinction allows us to formulate our generalization using generalized entropies and focuses attention on understanding bounds in the fixed, finite expert case.

As discussed in (Cesa-Bianchi and Lugosi, 2006, §3.3), mixability can be seen as a weakening of exp-concavity that requires just enough of the loss to ensure constant regret.

Theorem 2 (Mixability implies constant regret Vovk (1998)) *If a loss ℓ is η-mixable then the aggregating algorithm will achieve* $\operatorname{Regret}(T) \leq \eta^{-1} \log |\Theta|$.

A natural question is whether there are other, similar algorithms which also enjoy constant regret guarantees or whether the specific definition in (1) is somehow special.

1.2. Contributions

The key contributions of this paper are as follows. We provide a new general definition (Definition 4) of mixability and an induced generalized aggregating algorithm (Definition 7) and show (Theorem 9) that prediction with expert advice using a Φ-mixable loss and the associated generalized aggregating algorithm is guaranteed to have constant regret. The proof illustrates that the particular form of (1) for the classical aggregating algorithm is not what guarantees constant regret, but rather it is a translation invariant property of the convex conjugate of an entropy Φ defined on a probability simplex that is the crucial property.

We characterize (Theorem 6) for which entropies Φ there exist Φ-mixable losses via the Legendre property. We show that Φ-mixability of a loss can be expressed directly in terms of the Bayes risk associated with the loss (Definition 13 and Theorem 15), reflecting the situation that holds for classical mixability (van Erven et al., 2012). As part of this analysis we show that multiclass proper losses are quasi-convex (Lemma 14) which, to the best of our knowledge appears to be a new result. We also show (Theorem 11) how entropic duals relate to the potential-based analysis of Cesa-Bianchi and Lugosi (2003).

1.3. Related Work

The starting point for mixability and the aggregating algorithm is the work of Vovk (1998, 1990). The general setting of prediction with expert advice is summarized in (Cesa-Bianchi and Lugosi, 2006, Chapters 2 and 3). There one can find a range of results that study different aggregation schemes and different assumptions on the losses (exp-concave, mixable). Variants of the aggregating algorithm have been studied for classically mixable losses, with a trade-off between tightness of the bound (in a constant factor) and the computational complexity (Kivinen and Warmuth, 1999). Weakly mixable losses are a generalization of mixable losses. They have been studied by Kalnishkan and Vyugin (2008) who show that there exists a variant of the aggregating algorithm that achieves regret $C\sqrt{T}$ for some constant C. Vovk (2001, in §2.2) makes the observation that his Aggregating Algorithm reduces to Bayesian mixtures in the case of the log loss game. See also the discussion by Cesa-Bianchi and Lugosi (2006, page 330) relating certain aggregation schemes to Bayesian updating.

The general form of updating we propose is similar to that considered by Kivinen and Warmuth (1997, 1999) who consider finding a vector w minimizing $d(w, s) + \eta L(y_t, w \cdot x_t)$ where s is

some starting vector, (x_t, y_t) is the instance/label observation at round t and L is a loss. The key difference between their formulation and ours is that our loss term is (in their notation) $w \cdot L(y_t, x_t)$ – *i.e.*, the linear combination of the losses of the x_t at y_t and not the loss of their inner product. Online methods of density estimation for exponential families are discussed by Azoury and Warmuth (2001, §3) where the authors compare the online and offline updates of the same sequence and make heavy use of the relationship between the KL divergence between members of an exponential family and an associated Bregman divergence between the parameters of those members. The analysis of mirror descent by Beck and Teboulle (2003) shows that it achieves constant regret when the entropic regularizer is used. However, there is no consideration regarding whether similar results extend to other entropies defined on the simplex.

The idea of the more general regularization and updates is hardly new and connections between entropic duality and more general potential-based methods (Cesa-Bianchi and Lugosi, 2006, 2003) are readily made by choosing the potential to be an entropic dual, as discussed in §3.2. Interestingly, such potentials are already well studied in the mathematical finance literature where they are called convex risk measures (Föllmer and Schied, 2004), as well as in the literature on prediction markets where they are called cost functions (Abernethy et al., 2013). Thus, our work can be seen as extending existing connections between online learning and predication market mechanisms (Frongillo et al., 2012; Chen and Vaughan, 2010), as discussed in §3.3.

The key novelty is our generalized notion of mixability, the name of which is justified by the key new technical result (Theorem 9 — a constant regret bound assuming the general mixability condition achieved via a generalized algorithm that is exactly the mirror descent algorithm (*i.e.*, SANP) of Beck and Teboulle (2003) for the Bregman divergence generated by Φ. Crucially, our result depends on some properties of the conjugates of functions defined over affine spaces (*e.g.*, probabilities) that do not hold for potential functions more generally. By separating the convex geometry from the other special properties of classical entropy and mixability we hope to gain a deeper understanding of which losses admit fast rates of learning.

2. Generalized Mixability and Aggregation via Convex Duality

In this section we introduce our generalizations of mixability and the aggregating algorithm. One feature of our approach is the way the generalized aggregating algorithm falls out of the definition of generalized mixability as the minimizer of the mixability bound. Our approach relies on concepts and results from convex analysis. Terms not defined below can be found in a reference such as Hiriart-Urruty and Lemaréchal (2001).

2.1. Definitions and Notation

A function $\Phi : \Delta_\Theta \to \mathbb{R}$ is called an *entropy* (on Δ_Θ) if it is proper (*i.e.*, $-\infty < \Phi \neq +\infty$), convex[2], and lower semi-continuous. For $\eta > 0$, we write $\Phi_\eta := \eta^{-1}\Phi$. In the following example and elsewhere we use $\mathbb{1}$ to denote the vector $\mathbb{1}_\theta = 1$ for all $\theta \in \Theta$ and $|\Theta|^{-1}\mathbb{1} \in \Delta_\Theta$ to denote the uniform distribution over Θ. The distribution with unit mass on $\theta \in \Theta$ will be denoted δ_θ.

Example 1 (Entropies) *The* (negative) Shannon entropy $H(\mu) := \sum_\theta \mu_\theta \log \mu_\theta$*; the* quadratic entropy $Q(\mu) := \sum_\theta (\mu_\theta - |\Theta|^{-1}\mathbb{1})^2$*; the* Tsallis entropies $S_\alpha(\mu) := \alpha^{-1}\left(\sum_\theta \mu_\theta^{\alpha+1} - 1\right)$ *for*

2. While the information theoretic notion of Shannon entropy as a measure of uncertainty is concave, it is convenient for us to work with convex functions on the simplex which can be thought of as certainty measures.

$\alpha \in (-1, 0) \cup (0, \infty)$*; and the* Rényi entropies $R_\alpha(\mu) = \alpha^{-1}\left(\log \sum_\theta \mu_\theta^{\alpha+1}\right)$*, for* $\alpha \in (-1, 0)$*. We note that both Tsallis and Rényi entropies limit to Shannon entropy* $\alpha \to 0$*, and that Rényi entropy is convex for the given range (cf. Maszczyk and Duch (2008); van Erven and Harremoës (2014)).*

Let $\langle \mu, v \rangle$ denote a bilinear functional or *duality* between $\mu \in \Delta_\Theta$ and $v \in \Delta_\Theta^*$, where Δ_Θ and Δ_Θ^* form a dual pair[3]. The *Bregman divergence* associated with a suitably differentiable entropy Φ on Δ_Θ is given by

$$D_\Phi(\mu, \mu') = \Phi(\mu) - \Phi(\mu') - \left\langle \mu - \mu', \nabla\Phi(\mu') \right\rangle \tag{2}$$

for all $\mu \in \Delta_\Theta$ and $\mu' \in \operatorname{relint}(\Delta_\Theta)$, the relative interior of Δ_Θ. Given an entropy $\Phi : \Delta_\Theta \to \mathbb{R}$, we define its *entropic dual* to be $\Phi^*(v) := \sup_{\mu \in \Delta_\Theta} \langle \mu, v \rangle - \Phi(\mu)$ where $v \in \Delta_\Theta^*$. Note that one could also write the supremum over some larger space by setting $\Phi(\mu) = +\infty$ for $\mu \notin \Delta_\Theta$ so that Φ^* is just the usual convex dual (cf. Hiriart-Urruty and Lemaréchal (2001)). Thus, all of the standard results about convex duality also hold for entropic duals provided some care is taken with the domain of definition (Frongillo, 2013). Importantly, we note that although the unrestricted convex dual of H is $v \mapsto \sum_\theta \exp(v_\theta - 1)$ its entropic dual is $H^*(v) = \log \sum_\theta \exp(v_\theta)$.

For differentiable Φ, it is known (Hiriart-Urruty and Lemaréchal, 2001) that the supremum defining Φ^* is attained at $\mu = \nabla\Phi^*(v)$. That is,

$$\Phi^*(v) = \langle \nabla\Phi^*(v), v \rangle - \Phi(\nabla\Phi^*(v)). \tag{3}$$

A similar result holds for Φ by applying this result to Φ^* and using $\Phi = (\Phi^*)^*$. We will make repeated use of the following easily established properties of affinely restricted convex conjugation, of which entropic duality (*i.e.*, conjugation of convex functions on the simplex) is a special case. This is closely related to an observation by (Hiriart-Urruty and Lemaréchal, 2001, Prop. E.1.3.2), however we include a statement of the general result and proof of this result in Appendix A for completeness.

Lemma 3 *If* Φ *is an entropy over* Δ_Θ *then 1) for all* $\eta > 0$*,* $\Phi^*_\eta(v) = \eta^{-1}\Phi^*(\eta v)$*; and 2) the entropic dual* Φ^* *is* translation invariant *– i.e., for all* $v \in \Delta_\Theta^*$ *and* $\alpha \in \mathbb{R}$*,* $\Phi^*(v + \alpha\mathbb{1}) = \Phi^*(v) + \alpha$ *and hence for differentiable* Φ^**,* $\nabla\Phi^*(v + \alpha\mathbb{1}) = \nabla\Phi^*(v)$*.*

The translation invariance of Φ^* is central to our analysis. It is what ensures our Φ-mixability inequality (4) "telescopes" when it is summed. The proof of the original mixability result (Theorem 2) uses a similar telescoping argument that works due to the interaction of $\log$ and $\exp$ terms in Definition 1. Our results show that this telescoping property is not due to any special properties of $\log$ and $\exp$, but rather because of the translation invariance of the entropic dual of Shannon entropy, H. The remainder of our analysis generalizes that of the original work on mixability precisely because this property holds for the dual of any entropy.

Representation results from mathematical finance show that entropic duals are closely related to *convex monetary risk measures* (Föllmer and Schied, 2004), where translation invariance is called *cash invariance*. In particular, the function $\rho(v) = \Phi^*(-v)$ for an entropy Φ (a.k.a. a *penalty function*) is convex risk measure and is shown to be monotonic (*i.e.*, $v \leq v'$ pointwise implies $\rho(v) \geq \rho(v')$). The risk measure corresponding to the Shannon entropy is known as *entropic risk*.

3. In the case of finite Θ the duality is just the standard inner product. For infinite Δ_Θ, Δ_Θ^* is a space of random variables over Θ. See (Aliprantis and Border, 2007, §5.14) for details.

Furthermore, it can be shown that any measure of risk with these properties (convexity, monotonicity, and cash invariance) must be the convex conjugate of what we call an entropy (Föllmer and Schied, 2004, Theorem 4.15). We note these results hold for general spaces of probability measures, which is why our presentation does not always assume a finite set of experts Θ.

2.2. Φ-Mixability and the Generalized Aggregating Algorithm

For convenience, we will use $A \in \mathcal{A}^\Theta$ to denote the collection of expert predictions and $A_\theta \in \mathcal{A}$ to denote the prediction of expert θ. Abusing notation slightly, we will write $\ell(A) \in \mathbb{R}^{X \times \Theta}$ for the matrix of loss values $[\ell_x(A_\theta)]_{x,\theta}$, and $\ell_x(A) = [\ell_x(A_\theta)]_\theta \in \mathbb{R}^\Theta$ for the vector of losses for each expert θ on outcome x. In order to help distinguish between points, functions, distributions, etc. associated with outcomes and those associated with experts we use Roman symbols (*e.g.*, x, A, p) for the former and Greek (*e.g.*, θ, Φ, μ) for the latter.

Definition 4 (Φ-mixability) *Let Φ be an entropy on Δ_Θ. A loss $\ell : \mathcal{A} \to \mathbb{R}^X$ is Φ-mixable if for all $A \in \mathcal{A}^\Theta$, all $\mu \in \Delta_\Theta$, there exists an $\hat{a} \in \mathcal{A}$ such that for all $x \in X$*

$$\ell_x(\hat{a}) \leq \mathrm{Mix}^\Phi_{\ell,x}(A, \mu) := \inf_{\mu' \in \Delta_\Theta} \left\langle \mu', \ell_x(A) \right\rangle + D_\Phi(\mu', \mu). \tag{4}$$

The term on the right-hand side of (4) has some intuitive appeal. Since $\langle \mu', A \rangle = \mathbb{E}_{\theta \sim \mu'} [\ell_x(A_\theta)]$ (*i.e.*, the expected loss of an expert drawn at random according to μ') we can view the optimization as a trade off between finding a mixture μ' that tracks the expert with the smallest loss upon observing outcome x and keeping μ' close to μ, as measured by D_Φ. In the special case when Φ is Shannon entropy, ℓ is log loss, and expert predictions $A_\theta \in \Delta_X$ are distributions over X such an optimization is equivalent to Bayesian updating (Williams, 1980; DeSantis et al., 1988).

To see that Φ-mixability is indeed a generalization of Definition 1, we make use of an alternative form for the right-hand side of the bound in the Φ-mixability definition that "hides" the infimum inside Φ^*. As shown in Appendix A this is a straight-forward consequence of (3).

Lemma 5 *For all $A \in \mathcal{A}$ and $\mu \in \Delta_\Theta$, the mixability bound satisfies*

$$\mathrm{Mix}^\Phi_{\ell,x}(A, \mu) = \Phi^*(\nabla\Phi(\mu)) - \Phi^*(\nabla\Phi(\mu) - \ell_x(A)). \tag{5}$$

Hence, for $\Phi = \eta^{-1} H$, $\mathrm{Mix}^\Phi_{\ell,x}(A, \mu) = -\eta^{-1} \log \sum_\theta \exp(-\eta \ell_x(A_\theta)) \mu_\theta$, the bound in Definition 1.

Later, we will use $\mathrm{Mix}^\Phi_\ell(A, \mu)$ to denote the vector in $\mathbb{R}^X$ with components $\mathrm{Mix}^\Phi_{\ell,x}(A, \mu), x \in X$.

2.3. On the existence of Φ-mixable losses

A natural question at this point is do Φ-mixable losses exist for entropies other than Shannon entropy? If so, which Φ admit Φ-mixable losses? The next theorem answers both these questions, showing that the existence of "non-trivial" Φ-mixable losses is intimately related to the behaviour of an entropy's gradient at the simplex's boundary. Specifically, we will say an entropy Φ is *Legendre* [4] if: a) Φ is differentiable and strictly convex in $\mathrm{relint}(\Delta_\Theta)$; and b) $\|\nabla\Phi(\mu)\| \to \infty$ as $\mu \to \mu_b$

4. We note that our definition is slightly different (but similar in spirit) to the standard definition of Legendre function (cf. (Rockafellar, 1997)) since it requires a function f be strictly convex on the *interior* of $\mathrm{dom}(f)$ and that the interior of $\mathrm{dom}(f)$ be non-empty, but the interior is empty in the case of $\mathrm{dom}(f) = \Delta_\Theta$.

for any μ_b on the boundary of Δ_Θ. We call a loss ℓ *nontrivial* if there exist x^*, x' and a^*, a' such that

$$a' \in \arg\min\{\ell_{x^*}(a) : \ell_{x'}(a) = \inf_{a\in\mathcal{A}} \ell_{x'}(a)\} \text{ and } \inf_{a\in\mathcal{A}} \ell_{x^*}(a) = \ell_{x^*}(a^*) < \ell_{x^*}(a') \,. \tag{6}$$

Intuitively, this means that there exist distinct actions which are optimal for different outcomes x^*, x'. In particular, among all optimum actions for x', a' has the lowest loss on x^*. This rules out constant losses – *i.e.*, $\ell(a) = k \in \mathbb{R}^X$ for all $a \in \mathcal{A}$ – which are easily[5] seen to be Φ-mixable for any Φ. For technical reasons we will further restrict our attention to *curved* losses by which we mean those losses with strictly concave Bayes risks — *i.e.*, the function $L(p) := \inf_{a\in\mathcal{A}} \mathbb{E}_{x\sim p}\left[\ell_x(a)\right]$ is strictly concave — though we conjecture that the following also holds for non-curved losses.

Theorem 6 *Non-trivial, curved Φ-mixable losses exist if and only if the entropy Φ is Legendre.*

The proof is in Appendix A.4. We can apply this result to Example 1 and deduce that there are no Q-mixable losses. Also, since it is easy to show the derivatives ∇S_α and ∇R_α are unbounded for $\alpha \in (-1, 0)$, the entropies S_α and R_α are Legendre. Thus there exist S_α- and R_α-mixable losses when $\alpha \in (-1, 0)$. Due to this result we will henceforth *restrict our attention to Legendre entropies*.

We now define a generalization of the Aggregating Algorithm of Definition 1 that very naturally relates to our definition of Φ-mixability: starting with some initial distribution over experts, the algorithm repeatedly incorporates information about the experts' performances by finding the minimizer μ' in (4).

Definition 7 (Generalized Aggregating Algorithm) *The algorithm begins with a mixture distribution $\mu^0 \in \Delta_\Theta$ over experts. On round t, after receiving expert predictions $A^t \in \mathcal{A}^\Theta$, the* generalized aggregating algorithm *(GAA) predicts any $\hat{a} \in \mathcal{A}$ such that $\ell_x(\hat{a}) \le \mathrm{Mix}^\Phi_{\ell,x}(A^t, \mu^{t-1})$ for all x which is guaranteed to exist by the Φ-mixability of ℓ. After observing $x^t \in X$, the GAA updates the mixture $\mu^{t-1} \in \Delta_\Theta$ by setting*

$$\mu^t := \operatorname*{arg\,min}_{\mu'\in\Delta_\Theta} \left\langle \mu', \ell_{x^t}(A^t)\right\rangle + D_\Phi(\mu', \mu^{t-1}). \tag{7}$$

We now show that this updating process simply aggregates the per-expert losses $\ell_x(A)$ in the dual space Δ^*_Θ with $\nabla\Phi(\mu^0)$ as the starting point. The GAA is therefore exactly the *Subgradient algorithm with nonlinear projections* (SANP) for the Bregman divergence D_Φ (and a fixed step size of 1) which is known to be equivalent to mirror descent (Beck and Teboulle, 2003) using updates based on $\nabla\Phi^*$.

Lemma 8 *The GAA updates μ^t in (7) satisfy $\nabla\Phi(\mu^t) = \nabla\Phi(\mu^{t-1}) - \ell_{x^t}(A^t)$ for all t and so*

$$\nabla\Phi(\mu^T) = \nabla\Phi(\mu^0) - \sum_{t=1}^{T} \ell_{x^t}(A^t). \tag{8}$$

The proof is given in Appendix A. Finally, to see that the above is indeed a generalization of the Aggregating Algorithm from Definition 1 we need only apply Lemma 8 and observe that for $\Phi = \eta^{-1}H$, $\nabla\Phi(\mu) = \eta^{-1}(\log(\mu) + \mathbb{1})$ and so $\log\mu^t = \log\mu^{t-1} - \eta\ell_{x^t}(A^t)$. Exponentiating this vector equality element-wise gives $\mu^t_\theta \propto \mu^{t-1}_\theta \exp(-\eta\ell_{x^t}(A^t_\theta))$.

5. The inequality in (4) reduces to $0 \le \inf_{\mu'} D_\Phi(\mu', \mu)$ which is true for all Bregman divergences.

3. Properties of Φ-mixability

In this section we establish a number of key properties for Φ-mixability, the most important of these being that Φ-mixability implies constant regret.

3.1. Φ-mixability Implies Constant Regret

Theorem 9 *If $\ell : \mathcal{A} \to \mathbb{R}^X$ is Φ-mixable then there is a family of strategies parameterized by $\mu \in \Delta_\Theta$ which, for any sequence of observations $x^1, \dots, x^T \in X$ and sequence of expert predictions $A^1, \dots, A^T \in \mathcal{A}^\Theta$, plays a sequence $\hat{a}^1, \dots, \hat{a}^T \in \mathcal{A}$ such that for all $\theta \in \Theta$*

$$\sum_{t=1}^{T} \ell_{x^t}(\hat{a}^t) \le \sum_{t=1}^{T} \ell_{x^t}(A^t_\theta) + D_\Phi(\delta_\theta, \mu). \tag{9}$$

The proof is in Appendix A.1 and is a straight-forward consequence of Lemma 5 and the translation invariance of Φ^*. The standard notion of mixability is recovered when $\Phi = \frac{1}{\eta} H$ for $\eta > 0$ and H the Shannon entropy on Δ_Θ. In this case, Theorem 2 is obtained as a corollary for $\mu = |\Theta|^{-1}\mathbb{1}$, the uniform distribution over Θ. A compelling feature of our result is that it gives a natural interpretation of the constant $D_\Phi(\delta_\theta, \mu)$ in the regret bound: if μ is the initial guess as to which expert is best before the game starts, the "price" that is paid by the player is exactly how far (as measured by D_Φ) the initial guess was from the distribution that places all its mass on the best expert. Kivinen and Warmuth (1999) give a similar interpretation to the regret bound for the special case of Φ being Shannon entropy in their Theorem 3.

The following example computes mixability bounds for the alternative entropies introduced in §2.1. They will be discussed again in §4.2 below.

Example 2 *Consider games with $K = |\Theta|$ experts and the uniform distribution $\mu = K^{-1}\mathbb{1} \in \Delta_\Theta$. For the (negative) Shannon entropy, the regret bound from Theorem 9 is $D_H(\delta_\theta, \mu) = \log K$. For the family of Tsallis entropies the regret bound given by $D_{S_\alpha}(\delta_\theta, K^{-1}\mathbb{1}) = \alpha^{-1}(1 - K^{-\alpha})$. For the family of Rényi entropies the regret bound becomes $D_{R_\alpha}(\delta_\theta, K^{-1}\mathbb{1}) = \log K$.*

A second, easily established result concerns the mixability of scaled entropies. The proof follow from the observation that in (4) the only term in the definition of $\mathrm{Mix}_{\ell,x}^{\Phi_\eta}$ involving η is $D_{\Phi_\eta} = \frac{1}{\eta} D_\Phi$. The quantification over $A, \mu, \hat{a}$ and x in the original definition has been translated into infima and suprema.

Lemma 10 *The function $M(\eta) := \inf_{A,\mu} \sup_{\hat{a}} \inf_{\mu',x} \ \mathrm{Mix}_{\ell,x}^{\Phi_\eta}(A, \mu) - \ell_x(\hat{a})$ is non-increasing.*

This implies that there is a well-defined maximal $\eta > 0$ for which a given loss ℓ is Φ_η-mixable since Φ_η-mixability is equivalent to $M(\eta) \ge 0$. We call this maximal η the *Φ-mixability constant* for ℓ and denote it $\eta(\ell, \Phi) := \sup\{\eta > 0 : M(\eta) \ge 0\}$. This constant is central to the discussion in Section 4.3 below.

3.2. Relationship to Potential-Based Methods and the Blackwell Condition

Much of the analysis of online learning in Chapters 2, 3, and 11 of the book by Cesa-Bianchi and Lugosi (2006) is based on *potential functions* (Cesa-Bianchi and Lugosi, 2003) and their associated

Bregman divergences. As shown below, when their potentials are entropic duals, we obtain some connections with their analysis. In particular, we will see that if a loss satisfies Φ-mixability then the *Blackwell condition* for the potential $\Psi = \Phi^*$ is satisfied.

A loss function ℓ satisfies the *Blackwell condition* (cf. (Cesa-Bianchi and Lugosi, 2003)) for a convex potential $\Psi : \mathbb{R}^X \to \mathbb{R}$ if for all $R \in \mathbb{R}^X$ and $A \in \mathcal{A}^\Theta$ there exists some $\hat{a} \in \mathcal{A}$ such that $\sup_{x \in X} \langle \nabla \Psi(R), r_x \rangle \leq 0$, where $r_x = \ell_x(\hat{a})\mathbb{1} - \ell_x(A)$. We now have the following result. The proof is in Section A.2.

Theorem 11 *Let $\Phi : \Delta_X \to \mathbb{R}$ be an entropy with entropic dual $\Psi = \Phi^*$. If ℓ is a loss function that is Φ-mixable, then ℓ satisfies the Blackwell condition for the convex potential function Ψ.*

3.3. Φ-Mixability of Proper Losses and Their Bayes Risks

Entropies are known to be closely related to the Bayes risk of what are called proper losses or proper scoring rules (Dawid, 2007; Gneiting and Raftery, 2007). Here, the predictions are distributions over outcomes, *i.e.*, points in Δ_X. To highlight this we will use p, $\hat{p}$ and P instead of a, $\hat{a}$ and A to denote actions. If a loss $\ell : \Delta_X \to \mathbb{R}^X$ is used to assign a penalty $\ell_x(\hat{p})$ to a prediction $\hat{p}$ upon outcome x it is said to be *proper* if its expected value under $x \sim p$ is minimized by predicting $\hat{p} = p$. That is, for all $p, \hat{p} \in \Delta_X$,

$$\mathbb{E}_{x \sim p} \left[\ell_x(\hat{p}) \right] = \langle p, \ell(\hat{p}) \rangle \geq \langle p, \ell(p) \rangle =: -F^\ell(p),$$

where $-F^\ell$ is the *Bayes risk* of ℓ and is necessarily concave (van Erven et al., 2012), thus making $F^\ell : \Delta_X \to \mathbb{R}$ convex and thus an entropy. The correspondence also goes the other way: given any convex function $F : \Delta_X \to \mathbb{R}$ we can construct a unique proper loss (Vernet et al., 2011). The following representation can be traced back to Savage (1971) but is expressed here using convex duality.

Lemma 12 *If $F : \Delta_X \to \mathbb{R}$ is a differentiable entropy then the loss $\ell^F : \Delta_X \to \mathbb{R}$ defined by*

$$\ell^F(p) := -\nabla F(p) + F^*(\nabla F(p))\mathbb{1} = -\nabla F(p) + \left(\langle \nabla F(p), p \rangle - F(p) \right) \mathbb{1} \tag{10}$$

is proper.

By way of example, it is straight-forward to show that the proper loss associated with the negative Shannon entropy $\Phi = H$ is the log loss, that is, $\ell^H(\mu) := -\left(\log \mu(\theta)\right)_{\theta \in \Theta}$.

This connection between losses and entropies lets us define the Φ-mixability of a proper loss strictly in terms of its associated entropy. This is similar in spirit to the result by van Erven et al. (2012) which shows that the original mixability (for $\Phi = H$) can be expressed in terms of the relative curvature of Shannon entropy and the loss's Bayes risk. We use the following definition to explore the optimality of Shannon mixability in Section 4.3 below.

Definition 13 *An entropy $F : \Delta_X \to \mathbb{R}$ is* Φ-mixable *if*

$$\sup_{P, \mu} F^* \left(-\operatorname{Mix}^\Phi_{\ell^F}(P, \mu) \right) = \sup_{P, \mu} F^* \left(\left\{ \Phi^*(\nabla \Phi(\mu) - \ell^F_x(P)) \right\}_x - \Phi^*(\nabla \Phi(\mu))\mathbb{1} \right) \leq 0 \tag{11}$$

where ℓ^F is as in Lemma 12 and the supremum is over expert predictions $P \in \Delta_X^\Theta$ and mixtures over experts $\mu \in \Delta_\Theta$.

Although this definition appears complicated due to the handling of vectors in $\mathbb{R}^X$ and $\mathbb{R}^\Theta$, it has a natural interpretation in terms of *risk measures* from mathematical finance (Föllmer and Schied, 2004). Given some convex function $\alpha : \Delta_X \to \mathbb{R}$, its associated risk measure is its dual $\rho(v) := \sup_{p \in \Delta_X} \langle p, -v \rangle - \alpha(p) = \alpha^*(-v)$ where v is a *position* meaning v_x is some monetary value associated with outcome x occurring. Due to its translation invariance, the quantity $\rho(v)$ is often interpreted as the amount of "cash" (*i.e.*, outcome independent value) an agent would ask for to take on the uncertain position v. Observe that the risk ρ^F for when $\alpha = F$ satisfies $\rho^F \circ \ell^F = 0$ so that $\ell^F(p)$ is always a ρ^F-risk free position. If we now interpret $\mu^* = \nabla\Phi(\mu)$ as a position over outcomes in Θ and Φ^* as a risk for $\alpha = \Phi$ the term $\left\{\Phi^*(\mu^* - \ell^F_x(P))\right\}_x - \Phi^*(\mu^*)\mathbb{1}$ can be seen as the change in ρ^Φ risk when shifting position μ^* to $\mu^* - \ell^F_x(P)$ for each possible outcome x. Thus, the mixability condition in (11) can be viewed as a requirement that a ρ^F-risk free change in positions over Θ always be ρ^Φ-risk free.

The following theorem shows that the entropic version of Φ-mixability Definition 13 is equivalent to the loss version in Definition 4 in the case of proper losses. Its proof can be found in Appendix A.3 and relies on Sion's theorem and the facts that proper losses are *quasi-convex* (*i.e.*, $\forall \lambda \in [0,1]$, $f(\lambda x + (1-\lambda)y) \leq \max\{f(x), f(y)\}$). This latter fact appears to be new so we state it here as a separate lemma and prove it in Appendix A.

Lemma 14 *If $\ell : \Delta_X \to \mathbb{R}^X$ is proper then $p' \mapsto \langle p, \ell(p') \rangle$ is quasi-convex for all $p \in \Delta_X$.*

Theorem 15 *If $\ell : \Delta_X \to \mathbb{R}^X$ is proper and has Bayes risk $-F$ then F is an entropy and ℓ is Φ-mixable if and only if F is Φ-mixable.*

The entropic form of mixability in (11) shares some similarities with expressions for the classical mixability constants given by Haussler et al. (1998) for binary outcome games and by van Erven et al. (2012) for general games. Our expression for the mixability is more general than the previous two being both for binary and non-binary outcomes and for general entropies. Computing the optimizing argument is also more efficient than in (van Erven et al., 2012) since, for non-binary outcomes, their approach requires inverting a Hessian matrix at each point in the optimization.

4. Conclusions and Open Questions

The main purpose of this work was to shed new light on mixability by casting it within the broader notion of Φ-mixability. We showed that the constant regret bounds enjoyed by mixable losses are due to the translation invariance of entropic duals, and so are also enjoyed by any Φ-mixable loss. Our definitions and results allow us to now ask precise questions about alternative entropies and the optimality of their associated aggregating algorithms.

4.1. Are All Entropies "Equivalent"?

Since Theorem 6 shows the existence of Φ-mixable losses for Legendre Φ, we can ask about the relationship between the sets of losses that are mixable for different choices of Φ. For example, are there losses that are H-mixable but not S_α-mixable, or vice-versa? We conjecture that essentially all entropies Φ have the same Φ-mixable losses up to a scaling factor.

Conjecture 16 *Let Φ be an entropy on Δ_Θ and ℓ be a Φ-mixable loss. If Ψ is a Legendre entropy on Δ_Θ then there exists an $\eta > 0$ such that ℓ is Ψ_η-mixable.*

Some intuition for this conjecture is derived from observing that $\mathrm{Mix}^{\Psi_\eta}_{\ell,x} = \eta^{-1}\,\mathrm{Mix}^{\Psi}_{\eta\ell,x}$ and that as $\eta \to 0$ the function $\eta\ell$ behaves like a constant loss and will therefore be mixable. This means that scaling up $\mathrm{Mix}^{\Psi}_{\eta\ell,x}$ by η^{-1} should make it larger than $\mathrm{Mix}^{\Phi}_{\ell,x}$. However, some subtlety arises in ensuring that this dominance occurs uniformly. The discussion in Appendix B gives an example of two entropies where this scaling trick does not work.

4.2. Asymptotic Behaviour

There is a lower bound due to Vovk (1998) for general losses ℓ which shows that if one is allowed to vary the number of rounds T and the number of experts $K = |\Theta|$, then no regret bound can be better than the optimal regret bound obtained by Shannon mixability. Specifically, for a fixed loss ℓ with optimal Shannon mixability constant η_ℓ, suppose that for some $\eta' > \eta_\ell$ we have a regret bound of the form $(\log K)/\eta'$ as well as some strategy L for the learner that supposedly satisfies this regret bound. Vovk's lower bound shows, for this η' and L, that there exists an instantiation of the prediction with expert advice game with T large enough and K roughly exponential in T (and both are still finite) for which the alleged regret bound will fail to hold at the end of the game with non-zero probability. The regime in which Vovk's lower bound holds suggests that the best achievable regret with respect to the number of experts grows as $\log K$. Indeed, there is a lower bound for general losses ℓ that shows the regret of the best possible algorithm on games using ℓ must grow like $\Omega(\log K)$ (Haussler et al., 1998).

The above lower bound arguments apply when the number of experts is large (*i.e.*, exponential in the number of rounds) or if we consider the dynamics of the regret bound as K grows. This leaves open the question of the best possible regret bound for moderate and possibly fixed K which we formally state in the next section. This question serves as a strong motivation for the study of generalized mixability considered here. Note also that the above lower bounds are consistent with the fact that there cannot be non-trivial, Φ-mixable losses for non-Legendre Φ (*e.g.*, the quadratic entropy Q) since the growth of the regret bound as a function of K (cf. Example 2) is less than $\log K$ and hence violates the above lower bounds.

4.3. Is There An "Optimal" Entropy?

Since we believe that Φ-mixability for Legendre Φ yield the same set of losses, we can ask whether, for a fixed loss ℓ, some Φ give better regret bounds than others. These bounds depend jointly on the largest η such that ℓ is Φ_η-mixable and the value of $D_\Phi(\delta_\theta, \mu)$. We can define the optimal regret bound one can achieve for a particular loss ℓ using the generalized aggregating algorithm with Φ_η for some $\eta > 0$. This allows us to compare entropies on particular losses, and we can say that an entropy *dominates* another if its optimal regret bound is better for all losses ℓ. Recalling the definition of the maximal Φ-mixability constant from Lemma 10, we can determine a quantity of more direct interest: the best regret bound one can obtain using a scaled copy of Φ. Recall that if ℓ is Φ-mixable, then the best regret bound one can achieve from the generalized aggregating algorithm is $\inf_\mu \sup_\theta D_\Phi(\delta_\theta, \mu)$. We can therefore define the best regret bound for ℓ on a scaled version of Φ to be $R_{\ell,\Phi} := \eta(\ell, \Phi)^{-1} \inf_\mu \sup_\theta D_\Phi(\delta_\theta, \mu)$ where $\eta(\ell, \Phi)$ denotes the Φ-mixability constant for ℓ. Then $R_{\ell,\Phi}$ simply corresponds to the regret bound for the entropy $\Phi_{\eta(\ell,\Phi)}$. Note a crucial property of $R_{\ell,\Phi}$, which will be very useful in comparing entropies: $R_{\ell,\Phi} = R_{\ell,\alpha\Phi}$ for all $\alpha > 0$. (This follows from the observation that $\eta(\ell, \alpha\Phi) = \eta(\ell, \Phi)/\alpha$.) That is, $R_{\ell,\Phi}$ is independent of the particular scaling we choose for Φ.

We can now use $R_{\ell,\Phi}$ to define a scale-invariant relation over entropies. Define $\Phi \geq_\ell \Psi$ if $R_{\ell,\Phi} \leq R_{\ell,\Psi}$, and $\Phi \geq_* \Psi$ if $\Phi \geq_\ell \Psi$ for all losses ℓ. In the latter case we say Φ *dominates* Ψ. By construction, if one entropy dominates another its regret bound is guaranteed to be tighter and therefore its aggregating algorithm will achieve better worst-case regret. As discussed above, one natural candidate for a universally dominant entropy is the Shannon entropy.

Conjecture 17 *For all choices of Θ, the negative Shannon entropy dominates all other entropies. That is, $H \geq_* \Phi$ for all Θ and all convex Φ on Δ_Θ.*

Although we have not been able to prove this conjecture we were able to collect some positive evidence in the form of Table 1 in Appendix C. Here, we took the entropic form of Φ-mixability from Definition 13 and implemented it as an optimization problem and computed $\eta(\ell^F, \Phi)$ for F and Φ equal to the entropies introduced in Example 1 for two expert games with two outcomes. The maximal η (and hence the optimal regret bounds) for each pair was found doing a binary search for the zero-crossing of $M(\eta)$ from Lemma 10 and then applying the bounds from Example 2. Although we expected the dominant entropy for each loss ℓ^F to be its "matching" entropy (*i.e.*, $\Phi = F$), the table shows the optimal regret bound for every loss was obtained in the column for H.

One interesting feature for these matching entropy and loss cases is that the optimal η (shown in parentheses) is always equal to 1. We conjectured that ℓ^F would always be F-mixable with maximal $\eta = 1$ but found the counterexample described in Appendix B. However, we have not been able to rule out or prove the following weakenings of that conjecture. We observe that these cannot both be true due to the counterexample just described.

Conjecture 18 *Suppose $|X| = |\Theta|$ so that $\Delta_\Theta = \Delta_X$ and $\Phi : \Delta_\Theta \to \mathbb{R}$ an entropy. Then if its proper loss $\ell^\Phi : \Delta_X \to \mathbb{R}^X$ is Φ-mixable, the maximal η such that ℓ^Φ is $\eta^{-1}\Phi$-mixable is $\eta = 1$.*

Conjecture 19 *If $\Delta_\Theta = \Delta_X$ and Φ is an entropy then ℓ^Φ is $\eta^{-1}\Phi$-mixable for some $\eta > 0$.*

4.4. Future Work

Although Vovk's original mixability result has the "asymptotic" converse described in §4.2, the above conjectures highlight our lack of understanding of when fast rates of learning are achievable in the non-asymptotic regime. As well as resolving these conjectures, we hope to use this work as a basis for developing necessary conditions for constant regret for a fixed number of experts.

Finally, there have been some recent papers (Steinhardt and Liang, 2014; Orabona et al., 2015) which introduce extra time-varying updates ("hints") to the usual online mirror descent algorithms for sequential prediction to obtain a wider variety of algorithms and bounds. Given that mixability is already closely related to mirror descent, it would be interesting to see what extra structure and guarantees entropic duals provide in this setting.

Acknowledgments

We thank Matus Telgarsky for help with restricted duals, Brendan van Rooyen for noting that there are no quadratic mixable losses, Harish Guruprasad for identifying a flaw in an earlier "proof" of the quasi-convexity of proper losses, and the anonymous reviewers for valuable insights. MDR is supported by an ARC Discovery Early Career Research Award (DE130101605) and part of this work was developed while he was visiting Microsoft Research. RCW is supported by the ARC. NICTA is funded by the Australian Government through the ICT Centre of Excellence program.

References

Jacob Abernethy, Yiling Chen, and Jennifer Wortman Vaughan. Efficient market making via convex optimization, and a connection to online learning. *ACM Transactions on Economics and Computation*, 1(2):12, 2013.

Jacob D Abernethy and Rafael M Frongillo. A characterization of scoring rules for linear properties. In *Proceedings of the Conference on Learning Theory (COLT)*, volume 23 of *JMLR WCP*, pages 27–1, 2012.

Charalambos D. Aliprantis and Kim C. Border. *Infinite Dimensional Analysis: A Hitchhiker's Guide*. Springer, 2007.

Katy S Azoury and Manfred K Warmuth. Relative loss bounds for on-line density estimation with the exponential family of distributions. *Machine Learning*, 43(3):211–246, 2001.

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

Nicolo Cesa-Bianchi and Gábor Lugosi. Potential-based algorithms in on-line prediction and game theory. *Machine Learning*, 51(3):239–261, 2003.

Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge University Press, 2006.

Yiling Chen and Jennifer Wortman Vaughan. A new understanding of prediction markets via no-regret learning. In *Proceedings of the 11th ACM conference on Electronic commerce*, pages 189–198, 2010.

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

Alfredo DeSantis, George Markowsky, and Mark N Wegman. Learning probabilistic prediction functions. In *29th Annual Symposium on Foundations of Computer Science*, pages 110–119. IEEE, 1988.

Hans Föllmer and Alexander Schied. Stochastic finance, volume 27 of de gruyter studies in mathematics, 2004.

Rafael M. Frongillo. *Eliciting Private Information from Selfish Agents*. PhD thesis, University of California, Berkeley, 2013.

Rafael M Frongillo, Nicolás Della Penna, and Mark D Reid. Interpreting prediction markets: a stochastic approach. In *Proceedings of Neural Information Processing Systems*, 2012.

Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

Nick Gravin, Yuval Peres, and Balasubramanian Sivan. Towards optimal algorithms for prediction with expert advice, 2014. URL http://arxiv.org/abs/1409.3040.

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

Jean-Bapiste Hiriart-Urruty and Claude Lemaréchal. *Fundamentals of convex analysis*. Springer Verlag, 2001.

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

Jyrki Kivinen and Manfred K Warmuth. Exponentiated gradient versus gradient descent for linear predictors. *Information and Computation*, 132(1):1–63, 1997.

Jyrki Kivinen and Manfred K Warmuth. Averaging expert predictions. In *Proceedings of the 4th European Conference on Computational Learning Theory (EuroCOLT'99)*, pages 153–167. Springer, 1999.

Tomasz Maszczyk and Włodzisław Duch. Comparison of Shannon, Rényi and Tsallis entropy used in decision trees. In *Artificial Intelligence and Soft Computing–ICAISC 2008*, pages 643–651. Springer, 2008.

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

R. Tyrell Rockafellar. *Convex analysis*. Princeton University Press, 1997.

Leonard J Savage. Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, 66(336):783–801, 1971.

Jacob Steinhardt and Percy Liang. Adaptivity and optimism: An improved exponentiated gradient algorithm. In *Proceedings of the 31st International Conference on Machine Learning*, 2014.

Frederick A. Valentine. *Convex Sets*. McGraw-Hill, New York, 1964.

Tim van Erven and Peter Harremoës. Rényi divergence and Kullback-Leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820, 2014.

Tim van Erven, Mark D Reid, and Robert C Williamson. Mixability is Bayes risk curvature relative to log loss. *The Journal of Machine Learning Research*, 13(1):1639–1663, 2012.

Elodie Vernet, Robert C Williamson, and Mark D Reid. Composite multiclass losses. In *NIPS*, volume 24, pages 1224–1232, 2011.

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

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

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

Peter M Williams. Bayesian conditionalisation and the principle of minimum information. *British Journal for the Philosophy of Science*, 31(2):131–144, 1980.

Appendix A. Proofs

The following lemma is a generalization of Lemma 3.

Lemma 20 *Let W be a compact convex set for which the affine hull of $\mathrm{relint}(W)$ is an $(n-1)$-dimensional affine subspace of $\mathbb{R}^n$. Without loss of generality, assume that all $p \in W$ satisfy $\langle u, p\rangle = c$ for some $u \in \mathbb{R}^n \setminus \{0\}$ and some $c \in \mathbb{R}$. If Φ is a proper, convex, l.s.c functions over W and $\Phi_\eta := \eta^{-1}\Phi$ denotes a scaled version of Φ then 1) for all $\eta > 0$ we have $\Phi^*_\eta(v) = \eta^{-1}\Phi^*(\eta v)$; and 2) the convex conjugate Φ^* is* translation invariant *– i.e., for all $v \in W^*$ and $\alpha \in \mathbb{R}$ we have $\Phi^*(v + \alpha u) = \Phi^*(v) + c\alpha$ and hence for differentiable Φ^* we have $\nabla\Phi^*(v + \alpha u) = \nabla\Phi^*(v)$.*

Proof To show 1) we observe that $(\eta^{-1}\Phi)^*(v) = \sup_p \langle v, p\rangle - \eta^{-1}\Phi(p) = \eta^{-1}\sup_p \langle \eta v, p\rangle - \Phi(p) = \eta^{-1}\Phi^*(\eta v)$. For 2), we note that the definition of the dual implies $\Phi^*(v + \alpha u) = \sup_{\mu\in W} \langle \mu, v + \alpha u\rangle - \Phi(\mu) = \sup_{\mu\in W} \langle \mu, v\rangle - \Phi(\mu) + \alpha c = \Phi^*(v) + \alpha c$ since $\langle \mu, u\rangle = c$. Taking derivatives of both sides gives the final part of the lemma. ■

Proof [Proof of Lemma 5] By definition $\Phi^*(\nabla\Phi(\mu) - v) = \sup_{\mu'\in\Delta_\Theta} \langle \mu', \nabla\Phi(\mu) - v\rangle - \Phi(\mu')$ and using (3) gives $\Phi^*(\nabla\Phi(\mu)) = \langle \mu, \nabla\Phi(\mu)\rangle - \Phi(\mu)$. Subtracting the former from the latter gives $\langle \mu, \nabla\Phi(\mu)\rangle - \Phi(\mu) - \left[\sup_{\mu'\in\Delta_\Theta} \langle \mu', \nabla\Phi(\mu) - v\rangle - \Phi(\mu')\right]$ which, when rearranged gives $\inf_{\mu'\in\Delta_\Theta} \Phi(\mu') - \Phi(\mu) - \langle \nabla\Phi(\mu), \mu' - \mu\rangle + \langle \mu', v\rangle$ establishing the result.

When $\Phi = H$ *– i.e.*, Φ is the (negative) Shannon entropy – we have that $\nabla\Phi(\mu) = \log\mu + \mathbb{1}$, that $\Phi^*(v) = \log\sum_\theta \exp(v_\theta)$, and so $\nabla\Phi^*(v) = \exp(v)/\sum_\theta \exp(v_\theta)$, where log and exp are interpreted as acting point-wise on the vector μ. By Lemma 3, $\Phi^*(\nabla\Phi(\mu)) = \Phi^*(\log\mu + \mathbb{1}) = \Phi^*(\log(\mu)) + 1 = 1$ since $\Phi^*(\log(\mu_\theta)) = \log\sum_\theta \mu_\theta = 0$. Similarly, $\Phi^*(\nabla\Phi(\mu) - \ell_x(A)) = \Phi^*(\log(\mu) - \ell_x(A)) + 1 = \log\sum_\theta \mu_\theta \exp(-\ell_x(A)) + 1$. Substituting this into Lemma 5 and applying the second part of Lemma 3 shows that $\mathrm{Mix}^{\eta^{-1}H}_{\ell,x}(A, \mu) = -\eta^{-1}\log\sum_\theta \exp(-\eta\ell_x(A_\theta))$, recovering the right-hand side of the inequality in Definition 1. ■

Proof [Proof of Lemma 12] By eq. (3) we have $F^*(\nabla F(p)) = \langle p, \nabla F(p)\rangle - F(p)$, establishing the equality in the definition of (10) and giving us

$$\begin{aligned}\langle p, \ell^F(p')\rangle - \langle p, \ell^F(p)\rangle &= \Big(\langle p', \nabla F(p')\rangle - F(p') - \langle p, \nabla F(p')\rangle\Big) \\ &\quad - \Big(\langle p, \nabla F(p)\rangle - F(p) - \langle p, \nabla F(p)\rangle\Big) \\ &= D_F(p, p'),\end{aligned}$$

from which propriety follows. ■

Proof [Proof of Lemma 8] We prove a more general result for the case of an entropy over a compact convex subset of an affine subspace W as in Lemma 3. By considering the Lagrangian

$\mathcal{L}(\mu, \alpha) = \langle \mu, \ell_{x^t}(A)\rangle + D_\Phi(\mu, \mu^{t-1}) + \alpha(\langle \mu, u\rangle - c)$ and setting its derivative to zero we see that the minimizing μ^t must satisfy $\nabla\Phi(\mu^t) = \nabla\Phi(\mu^{t-1}) - \ell_{x^t}(A^t) - \alpha^t u$ where $\alpha^t \in R$ is the dual variable at step t.

For Legendre entropies Φ, it holds that $\nabla\Phi^*(\nabla\Phi(p)) = p$, as can be seen from the fact that $\nabla\Phi^*(\nabla\Phi(p)) = \operatorname{argmax}_{q\in W}\langle q, \nabla\Phi(p)\rangle - \Phi(q)$. From the gradient-barrier property, it holds that the maximum is obtained in the interior of W, and so setting the derivative of the objective to zero we have $\nabla\Phi(p) = \nabla\Phi(q)$. Since Φ is strictly convex and differentiable, $\nabla\Phi$ is injective, and hence the optimal q is equal to p. Thus, $\nabla\Phi^*(\nabla\Phi(p)) = p$. Now, for any $p \in W$ the maps $\nabla\Phi^*$ and $\nabla\Phi$ satisfy $\nabla\Phi^*(\nabla\Phi(p)) = p$, so $\mu^t = \nabla\Phi^*(\nabla\Phi(\mu^{t-1}) - \ell_{x^t}(A^t) - \alpha^t u) = \nabla\Phi^*(\nabla\Phi(\mu^{t-1}) - \ell_{x^t}(A^t))$ by the translation invariance of Φ^* (Lemma 3). This means the constants α^t are arbitrary and can be ignored. Thus, the mixture updates satisfy the relation in the lemma and summing over $t = 1, \dots, T$ gives (8). ■

Proof [Proof of Lemma 14] Let $n = |X|$ and fix an arbitrary $p \in \Delta_X$. The function $f_p(q) = \langle p, \ell(q)\rangle$ is quasi-convex if its α sublevel sets $F_p^\alpha := \{q \in \Delta_X \colon \langle p, \ell(q)\rangle \le \alpha\}$ are convex for all $\alpha \in \mathbb{R}$. Let $g(p) := \inf_q f_p(q)$ and fix an arbitrary $\alpha > g(p)$ so that $F_p^\alpha \neq \emptyset$. Let $Q_p^\alpha := \{v \in \mathbb{R}^n \colon \langle p, v\rangle \le \alpha\}$ so $F_p^\alpha = \{q \in \Delta_X \colon \ell(q) \in Q_p^\alpha\}$. Denote by $h_q^\beta := \{v \colon \langle v, q\rangle = \beta\}$ the hyperplane in direction $q \in \Delta_X$ with offset $\beta \in \mathbb{R}$ and by $H_q^\beta := \{v \colon \langle v, q\rangle \ge \beta\}$ the corresponding half-space. Since ℓ is proper, its *superprediction set* $\mathcal{S}_\ell = \{\lambda \in \mathbb{R}^n : \exists q \in \Delta_X \forall x \in X \lambda_x \ge \ell_x(q)\}$ (see (Vernet et al., 2011, Prop. 17)) is supported at $x = \ell(q)$ by the hyperplane $h_q^{g(q)}$ and furthermore since $\mathcal{S}_\ell$ is convex, $\mathcal{S}_\ell = \bigcap_{q\in\Delta_X} H_q^{g(q)}$.

Let

$$V_p^\alpha := \bigcap_{v \in \ell(\Delta_X)\cap Q_p^\alpha} H_{\ell^{-1}(v)}^{g(\ell^{-1}(v))} = \bigcap_{q\in F_p^\alpha} H_q^{g(q)}$$

(see figure 1). Since V_p^α is the intersection of halfspaces it is convex. Note that a given half-space $H_q^{g(q)}$ is supported by exactly one hyperplane, namely $h_q^{g(q)}$. Thus the set of hyperplanes that support V_p^α is $\{h_q^{g(q)} \colon q \in F_p^\alpha\}$ If $u \in F_p^\alpha$ then there is a hyperplane in direction u that supports V_p^α and its offset is given by

$$\sigma_{V_p^\alpha}(u) := \inf_{v\in V_p^\alpha} \langle u, v\rangle = g(p) > -\infty$$

whereas if $u \notin F_p^\alpha$ then for all $\beta \in \mathbb{R}$, h_u^β does not support V_p^α and hence $\sigma_{V_p^\alpha}(u) = -\infty$. Thus we have shown

$$\left(u \notin W_p^\alpha\right) \Leftrightarrow \left(\sigma_{V_p^\alpha}(u) = -\infty\right).$$

Observe that $\sigma_{V_p^\alpha}(u) = -s_{V_p^\alpha}(-u)$ where $s_C(u) = \sup_{v\in C}\langle u, v\rangle$ is the support function of a set C. It is known (Valentine, 1964, Theorem 5.1) that the "domain of definition" of a support function $\{u \in \mathbb{R}^n \colon s_C(u) < +\infty\}$ for a convex set C is always convex. Thus $G_p^\alpha := \{u \in \Delta_X \colon \sigma_{V_p^\alpha}(u) > -\infty\} = \{u \in \mathbb{R}^n \colon \sigma_{V_p^\alpha}(u) > -\infty\} \cap \Delta_X$ is always convex because it is the intersection of convex sets. Finally by observing that

$$G_p^\alpha = \{p \in \Delta_X \colon \ell(p) \in \ell(\Delta_X) \cap Q_p^\alpha\} = F_p^\alpha$$

we have shown that F_p^α is convex. Since $p \in \Delta_X$ and $\alpha \in \mathbb{R}$ were arbitrary we have thus shown that f_p is quasi-convex for all $p \in \Delta_X$.

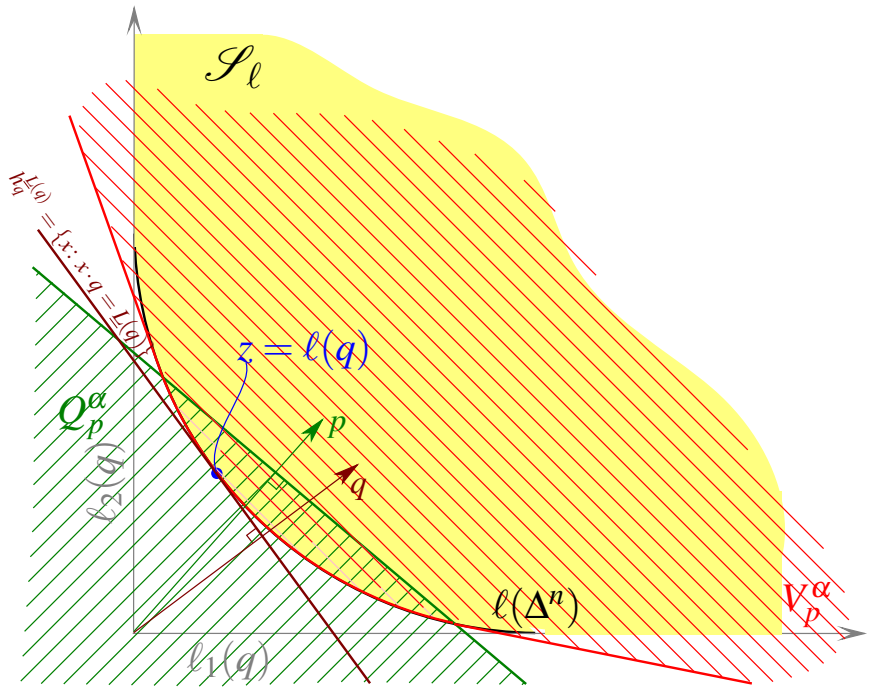

Figure 1: Visualization of construction in proof of Lemma 14.

■

A.1. Proof of Theorem 9

Proof Applying Lemma 5 to the assumption that ℓ is Φ-mixable means that for μ equal to the updates μ^t from Definition 7 and A^t equal to the expert predictions at round t, there must exist an $\hat{a}^t \in \Delta_X$ such that

$$\ell_{x^t}(\hat{a}^t) \leq \Phi^*(\nabla\Phi(\mu^{t-1})) - \Phi^*(\nabla\Phi(\mu^{t-1}) - \ell_{x^t}(A^t))$$

for all $x^t \in X$. Summing these bounds over $t = 1, \ldots, T$ gives

$$\begin{aligned}
\sum_{t=1}^T \ell_{x^t}(p^t) \leq & \sum_{t=1}^T \Phi^*(\nabla\Phi(\mu^{t-1})) - \Phi^*(\nabla\Phi(\mu^{t-1}) - \ell_{x^t}(A^t)) \\
= & \Phi^*(\nabla\Phi(\mu^0)) - \Phi^*(\nabla\Phi(\mu^T)) && (12) \\
= & \inf_{\mu' \in \Delta_\Theta} \left\langle \mu', \sum_{t=1}^T \ell_{x^T}(A^t) \right\rangle + D_\Phi(\mu', \mu^0) && (13) \\
\leq & \left\langle \mu', \sum_{t=1}^T \ell_{x^t}(A^t) \right\rangle + D_\Phi(\mu', \mu^0) \qquad \text{for all } \mu' \in \Delta_\Theta && (14)
\end{aligned}$$

Line (12) above is because $\nabla\Phi(\mu^t) = \nabla\Phi(\mu^{t-1}) - \ell_{x^t}(A^t)$ by Lemma 8 and the series telescopes. Line (13) is obtained by applying (7) from Lemma 8 and matching equations (5) and (4). Setting $\mu' = \delta_\theta$ and noting $\langle \delta_\theta, \ell(A^t)\rangle = \ell_{x^t}(A^t_\theta)$ gives the required result. ■

A.2. Proof of Theorem 11

By definition, the Blackwell condition is that for all $R \in \mathbb{R}^X$, $A \in \mathcal{A}^\Theta$, there exists $\hat{a} \in \mathcal{A}$ such that for all $x \in X$

$$\langle \nabla\Phi^*(R), r_x \rangle \le 0. \tag{15}$$

Since Φ is an entropic dual with respect to the simplex, $\nabla\Phi^*(R) \in \Delta_\Theta$, and so

$$\begin{aligned}\langle \nabla\Phi^*(R), r_x \rangle &= \mathbb{E}_{\theta\sim\nabla\Phi^*(R)}\left[\ell_x(\hat{a}) - \ell_x(A_\theta)\right] \\ &= \ell_x(\hat{a}) - \mathbb{E}_{\theta\sim\nabla\Phi^*(R)}\left[\ell_x(A_\theta)\right].\end{aligned}$$

Thus, (15) is equivalent to

$$\begin{aligned}\ell_x(\hat{a}) &\le \mathbb{E}_{\theta\sim\nabla\Phi^*(R)}\left[\ell_x(A_\theta)\right] \\ &= \mathbb{E}_{\theta\sim\nabla\Phi^*(R)}\left[\ell_x(A_\theta)\right] + D_\Phi\big(\nabla\Phi^*(R), \nabla\Phi^*(R)\big).\end{aligned}$$

On the other hand, ℓ is Φ-mixable if, for all $R \in \mathbb{R}^X$, $A \in \mathcal{A}^\Theta$, there exists $\hat{a} \in \mathcal{A}$ such that for all $x \in X$:

$$\ell_x(\hat{a}) \le \inf_{\mu\in\Delta} \mathbb{E}_{\theta\sim\mu}\left[\ell_x(A_\theta)\right] + D_\Phi\big(\mu, \nabla\Phi^*(R)\big).$$

Clearly,

$$\begin{aligned}&\inf_{\mu\in\Delta} \mathbb{E}_{\theta\sim\mu}\left[\ell_x(A_\theta)\right] + D_\Phi\big(\mu, \nabla\Phi^*(R)\big) \\ &\qquad \le \mathbb{E}_{\theta\sim\nabla\Phi^*(R)}\left[\ell_x(A_\theta)\right] + D_\Phi\big(\nabla\Phi^*(R), \nabla\Phi^*(R)\big),\end{aligned}$$

and so the Φ-mixability condition implies the Blackwell condition.

A.3. Proof of Theorem 15

We first establish a general reformulation of Φ-mixability that holds for arbitrary ℓ by converting the quantifiers in the definition of Φ-mixability from Lemma 5 for ℓ into an expression involving infima and suprema. We then further refine this by assuming $\ell = \ell^F$ is proper (and thus quasi-convex) and has Bayes risk F.

$$\begin{aligned}&\inf_{A,\mu}\sup_{\hat{a}}\inf_{x}\ \Phi^*(\nabla\Phi(\mu)) - \Phi^*(\nabla\Phi(\mu) - \ell^F_x(A)) - \ell^F_x(\hat{a}) \ge 0 \\ \iff &\inf_{A,\mu}\sup_{\hat{a}}\inf_{p}\ \left\langle p, \left\{\Phi^*(\nabla\Phi(\mu)) - \Phi^*(\nabla\Phi(\mu) - \ell^F_x(P))\right\}_x\right\rangle - \left\langle p, \ell^F_x(\hat{p})\right\rangle \ge 0\end{aligned} \tag{16}$$

where the term in braces is a vector in $\mathbb{R}^X$. The infimum over x is switched to an infimum over distributions over $p \in \Delta_X$ because the optimization over p will be achieved on the vertices of the simplex as it is just an average over random variables over X.

From here on we assume that $\ell = \ell^F$ is proper and adjust our notation to emphasise that actions $\hat{a} = \hat{p}$ and $A = P$ are distributions. Note that the new expression is linear – and therefore convex in p – and, by Lemma 14, we know ℓ^F is quasi-convex and so the function being optimized in (16) is quasi-concave in $\hat{p}$. We can therefore apply Sion's theorem to swap $\inf_p$ and $\sup_{\hat{p}}$ which means ℓ^F is Φ-mixable if and only if

$$\begin{aligned}
&\inf_{P,\mu} \inf_{p} \sup_{\hat{p}} \left\langle p, \left\{\Phi^*(\nabla\Phi(\mu)) - \Phi^*(\nabla\Phi(\mu) - \ell^F_x(P))\right\}_x \right\rangle - \left\langle p, \ell^F_x(\hat{p}) \right\rangle \geq 0 \\
\iff &\inf_{P,\mu} \inf_{p} \Phi^*(\nabla\Phi(\mu)) - \left\langle p, \left\{\Phi^*(\nabla\Phi(\mu) - \ell^F_x(P))\right\}_x \right\rangle + F(p) \geq 0 \\
\iff &\inf_{P,\mu} \Phi^*(\nabla\Phi(\mu)) - F^*(\left\{\Phi^*(\nabla\Phi(\mu) - \ell^F_x(P))\right\}_x) \geq 0
\end{aligned}$$

The second line above is obtained by recalling that, by the definition of ℓ^F, its Bayes risk is F. We now note that the inner infimum over p passes through $\Phi^*(\nabla\Phi(\mu))$ so that the final two terms are just the convex dual for F evaluated at $\left\{\Phi^*(\nabla\Phi(\mu) - \ell^F_x(P))\right\}_x$. Finally, by translation invariance of F^* we can pull the $\Phi^*(\pi^*)$ term inside F^* to simplify further so that the loss ℓ^F with Bayes risk F is Φ-mixable if and only if

$$\inf_{P,\mu} -F^*\left(\left\{\Phi^*(\nabla\Phi(\mu) - \ell^F_x(P))\right\}_x - \Phi^*(\nabla\Phi(\mu))\mathbb{1}\right) \geq 0.$$

Applying Lemma 12 to write ℓ^F in terms of F and passing the sign through the infimum and converting it to a supremum gives the required result.

A.4. Proof of Theorem 6

We will make use the following formulation of mixability,

$$M(\eta) := \inf_{A\in\mathcal{A},\, \pi\in\Delta_\Theta} \sup_{\hat{a}\in\mathcal{A}} \inf_{\mu\in\Delta_\Theta,\, x\in X} \langle \mu, \ell_x(A)\rangle + \frac{1}{\eta} D_\Phi(\mu,\pi) - \ell_x(\hat{a}), \tag{17}$$

so that ℓ is Φ_η-mixable if and only if $M(\eta) \geq 0$.

Lemma 21 *Suppose ℓ has a strictly concave Bayes risk L. Then given any distinct $\mu^*, \mu' \in \Delta_\Theta$, there is some $A \in \mathcal{A}$ and $x^*, x' \in X$ such that for all $\hat{a} \in \mathcal{A}$ we have at least one of the following:*

$$\langle \mu^*, \ell_{x^*}(A)\rangle < \ell_{x^*}(\hat{a}) \,, \quad \left\langle \mu', \ell_{x'}(A)\right\rangle < \ell_{x'}(\hat{a}) \,. \tag{18}$$

Proof Let θ^* be an expert such that $\alpha := \mu^*_{\theta^*} > \mu'_{\theta^*} =: \beta$, which exists as $\mu^* \neq \mu'$. Pick arbitrary $x^*, x' \in X$ and let $p^*, p' \in \Delta_X$ with support only on $\{x^*, x'\}$ and $p^*_{x^*} = \alpha/(\alpha+\beta)$, $p'_{x^*} = (1-\alpha)/(2-\alpha-\beta)$. Now let $a^* = \arg\min_{a\in\mathcal{A}} \mathbb{E}_{x\sim p^*}[\ell_x(a)]$, $a' = \arg\min_{a\in\mathcal{A}} \mathbb{E}_{x\sim p'}[\ell_x(a)]$, and set A such that $A_{\theta^*} = a^*$ and $A_\theta = a'$ for all other $\theta \in \Theta$.

Now suppose there is some $\hat{a} \in \mathcal{A}$ violating eq. (18). Then in particular,

$$\begin{aligned}
\tfrac{1}{2}\left(\ell_{x^*}(\hat{a}) + \ell_{x'}(\hat{a})\right) &\leq \tfrac{1}{2}\left(\langle \mu^*, \ell_{x^*}(A)\rangle + \left\langle \mu', \ell_{x'}(A)\right\rangle\right) \\
&= \tfrac{1}{2}\left(\alpha\ell_{x^*}(a^*) + (1-\alpha)\ell_{x^*}(a') + \beta\ell_{x'}(a^*) + (1-\beta)\ell_{x'}(a')\right) \\
&= \tfrac{\alpha+\beta}{2}\left(\tfrac{\alpha}{\alpha+\beta}\ell_{x^*}(a^*) + \tfrac{\beta}{\alpha+\beta}\ell_{x'}(a^*)\right) + \tfrac{2-\alpha-\beta}{2}\left(\tfrac{1-\alpha}{2-\alpha-\beta}\ell_{x^*}(a') + \tfrac{1-\beta}{2-\alpha-\beta}\ell_{x'}(a')\right) \\
&= \tfrac{\alpha+\beta}{2} L(p^*) + \left(1 - \tfrac{\alpha+\beta}{2}\right) L(p') \,.
\end{aligned}$$

Letting $\bar{p} \in \Delta_X$ with $\bar{p}_{x^*} = \bar{p}_{x'} = 1/2$, observe that $\bar{p} = \frac{\alpha+\beta}{2}p^* + (1 - \frac{\alpha+\beta}{2})p'$. But by the above calculation, we have $L(\bar{p}) \leq \frac{\alpha+\beta}{2}L(p^*) + (1 - \frac{\alpha+\beta}{2})L(p')$, thus violating strict concavity of L. ■

NON-LEGENDRE $\implies$ NO NONTRIVIAL MIXABLE ℓ WITH STRICTLY CONVEX BAYES RISK:

To show that no non-constant Φ-mixable losses exist, we must exhibit a $\pi \in \Delta_\Theta$ and an $A \in \mathcal{A}$ such that for all $\hat{a} \in \mathcal{A}$ we can find a $\mu \in \Delta_\Theta$ and $x \in X$ satisfying $\langle \mu, \ell_x(A) \rangle + \frac{1}{\eta} D_\Phi(\mu, \pi) - \ell_x(\hat{a}) < 0$. Since Φ is non-Legendre it must either (1) fail strict convexity, or (2) have a point on the boundary with bounded derivative; we will consider each case separately.

(1) Assume that Φ is not strictly convex; then we have some $\mu^* \neq \mu'$ such that $D_\Phi(\mu^*, \mu') = 0$. By Lemma 21 with these two distributions, we have some A and x^*, x' such that for all $\hat{a}$, either (i) $\langle \mu^*, \ell_{x^*}(A) \rangle < \ell_{x^*}(\hat{a})$ or (ii) $\langle \mu', \ell_{x'}(A) \rangle < \ell_{x'}(\hat{a})$. We set $\pi = \mu'$; in case (i) we take $\mu = \mu^*$ and $x = x^*$, and in (ii) we take $\mu = \mu'$ and $x = x'$, but as $\frac{1}{\eta} D_\Phi(\mu, \pi) = 0$ in both cases, we have $M(\eta) < 0$ for all η.

(2) Now assume instead that we have some μ' on the boundary of Δ_Θ with bounded $\|\nabla \Phi(\mu')\| = C < \infty$. Because μ' is on the boundary of Δ_Θ there is at least one expert $\theta^* \in \Theta$ for which $\mu'_{\theta^*} = 0$. Pick x^*, x', a^*, a' from the definition of nontrivial, eq. (6). In particular, note that $\ell_{x^*}(a^*) < \ell_{x^*}(a')$. Let $\pi = \mu'$ and $A \in \mathcal{A}$ such that $A_{\theta^*} = a^*$ and $A_\theta = a'$ for all other θ.

Now suppose $\hat{a} \in \mathcal{A}$ has $\ell_{x'}(\hat{a}) > \ell_{x'}(a')$. Then taking $\mu = \pi$ puts all weights on experts predicting a' while keeping $D_\Phi(\mu, \pi) = 0$, so choosing $x = x'$ gives $M(\eta) < 0$ for all η. Otherwise, $\ell_{x'}(\hat{a}) = \ell_{x'}(a')$, which by eq. (6) implies $\ell_{x^*}(\hat{a}) \geq \ell_{x^*}(a')$. Let $\mu^\alpha = \pi + \alpha(\delta_{\theta^*} - \pi)$, where δ_{θ^*} denotes the point distribution on θ^*. Calculating, we have

$$\begin{aligned}
M(\eta) &= \langle \mu^\alpha, \ell_{x^*}(A) \rangle + \tfrac{1}{\eta} D_\Phi(\mu^\alpha, \pi) - \ell_{x^*}(\hat{a}) \\
&= (1-\alpha)\ell_{x^*}(a') + \alpha \ell_{x^*}(a^*) + \tfrac{1}{\eta} D_\Phi(\mu^\alpha, \pi) - \ell_{x^*}(\hat{a}) \\
&\leq (1-\alpha)\ell_{x^*}(\hat{a}) + \alpha \ell_{x^*}(a^*) + \tfrac{1}{\eta} D_\Phi(\mu^\alpha, \pi) - \ell_{x^*}(\hat{a}) \\
&= \alpha(\ell_{x^*}(a^*) - \ell_{x^*}(\hat{a})) + \tfrac{1}{\eta} D_f(\alpha, 0),
\end{aligned}$$

where $f(\alpha) = \Phi(\mu^\alpha) = \Phi(\pi + \alpha(\delta_{\theta^*} - \pi))$. As $\nabla_\pi \Phi$ is bounded, so is $f'(0)$. Now as $\lim_{\epsilon \to 0} D_f(x+\epsilon, x)/\epsilon = 0$ for any scalar convex f with bounded $f'(x)$ (see e.g. (Rockafellar, 1997, Theorem 24.1) and Abernethy and Frongillo (2012)), we see that for any $c > 0$ we have some $\alpha > 0$ such that $D_f(\alpha, 0) < c\alpha$. Taking $c = \eta(\ell_{x^*}(\hat{a}) - \ell_{x^*}(a^*)) > 0$ then gives $M(\eta) < 0$.

LEGENDRE $\implies \exists$ MIXABLE ℓ:

Assuming Φ is Legendre, we need only show that some non-constant ℓ is Φ-mixable. As $\nabla_\pi \Phi$ is infinite on the boundary, π must be in the relative interior of Δ_Θ; otherwise $D_\Phi(\mu, \pi) = \infty$ for $\mu \neq \pi$.

Take $\mathcal{A} = \Delta_X$ and $\ell(p, x) = \|p - \delta_x\|^2$ to be the 2-norm squared loss. Now for all μ in the interior of Δ_Θ and $P \in \Delta_X^\Theta$, we have $\langle \mu, \ell_x(P) \rangle = \sum_\theta \mu_\theta \|P_\theta - \delta_x\|^2 \geq \|\bar{p} - \delta_x\|^2$ by convexity, where $\bar{p} = \sum_\theta \mu_\theta P_\theta$. In fact, as μ is in the interior, this inequality is strict, and remains so if replace μ by μ' with $\|\mu' - \mu\| < \epsilon$ for some ϵ sufficiently small. Now for all μ, P the algorithm can take $\hat{p} = \bar{p}$, and we can always choose $\eta = \inf_{x, \mu' : \|\mu' - \mu\| = \epsilon} D_\Phi(\mu', \mu)/(\epsilon \ell_{\max}) > 0$,

so either $\|\mu - \pi\| < \epsilon$ in which case we are fine by the above, or μ is far enough away that the D_Φ term dominates the algorithm's loss. (Here $\ell_{\max}$ is just $\max_{p,x} \ell_x(p)$, which is bounded, and $D_\Phi(\mu', \mu) > 0$ as Φ is strictly convex.) So if Φ is Legendre, squared loss is Φ-mixable.

Appendix B. A Loss with Bayes Risk $-B$ that is not B-Mixable

Let $b : \mathbb{R} \to \mathbb{R}$ be $b(p) := (\log p)(1 - \frac{1}{2}\log p)$ and for $p \in \Delta_X$ define the "bentropy"[6] $B(p) := \sum_x p_x f(p_x)$. For binary outcomes expert and learner predictions are of the form $(p, 1-p) \in \Delta_2$ and the loss associated with B (the "bentropic loss"), constructed using Lemma 12

$$\ell^B(p, 1-p) = \left(-f(p) + (1-p)\log\left(\tfrac{p}{1-p}\right), -f(1-p) + p\log\left(\tfrac{1-p}{p}\right)\right) \tag{19}$$

has Bayes risk $-B(p)$. One can verify that B is Legendre since $B'(p) = \frac{1}{2}\left((\log p)^2 - (\log(1-p))^2\right)$, and that $B''(p) = \frac{p\log(1-p)+(1-p)\log p}{p(1-p)}$.

Using the analysis of mixability in §4.1 of (van Erven et al., 2012), a proper, binary loss ℓ has a mixability constant η_ℓ given by the smallest ratio of curvatures between the Bayes risk for log loss and the Bayes risk for ℓ. That is, $\eta_\ell = \inf_{p\in(0,1)} \frac{H''(p)}{-F''(p)}$ where H is Shannon entropy and $H''(p) = [p(1-p)]^{-1}$. For $F = B$ we see $\eta_{\ell^B} = \inf_p \frac{1}{-p\log(1-p)-(1-p)\log p} = 0$. We have thus estabilised the following:

Lemma 22 *The binary proper loss ℓ^B defined in (19) is not classically mixable.*

However, we were able to determine numerically that ℓ^B is also not B-mixable. We did so by considering the two outcome/two expert case and looking for specific expert predictions $(p^A, 1-p^A)$ and $(p^B, 1-p^B)$ and mixture $(\mu, 1-\mu)$ so that the bound in (11) is violated. We found one in the case where $p^A = 0.4$, $p^B = 0.01$, and $\mu = 0.4$ which gives the value 0.145 on the left side of (11).

Finally, to give some intuition as to why Conjecture 16 is subtle, we note that the mixability Mix_ℓ^Φ only depends of Φ through the Bregman divergence term $D_\Phi(\mu', \mu)$. Since a Bregman divergence is the second-order and higher tail of the Taylor series expansion of Φ about μ, the ability to scale the mixability term for Φ so that it dominates Mix^Ψ depends on whether the ratio Ψ''/Φ'' can be uniformly bounded. In the case consider here, where $\Psi = B$ and $\Phi = H$ we have $\frac{-B''(p)}{H''(p)} = -p\log(1-p) - (1-p)\log p$ which is unbounded for $p \in (0, 1)$.

6. The name was chosen to highlight the new entropy similar to regular entropy but "bent" by the $1 - \frac{1}{2}\log p$ terms.

Appendix C. Table of Maximum Mixability Constants and Regrets

Table 1 is discussed in Section 4.3.

Table 1: Mixability and optimal regrets for pairs of losses and entropies in 2 outcome/2 experts games. Entries show the regret bound $\eta^{-1} D_\Phi(\delta_\theta, \frac{1}{2}\mathbb{1})$ for the maximum η (in parentheses).

	Entropy						
Loss	H	$S_{-.1}$	$S_{-.5}$	$S_{-.9}$	$R_{-.1}$	$R_{-.5}$	$R_{-.9}$
log	0.69 (1*)	0.74 (.97)	1.17 (.71)	5.15 (.19)	0.77 (0.9)	1.38 (0.5)	6.92 (0.1)
ℓ^Q	0.34 (2)	0.37 (1.9)	0.58 (1.4)	2.57 (0.4)	0.38 (1.8)	0.69 (1)	3.45 (0.2)
$\ell^{S_{-.5}}$	0.49 (1.4)	0.53 (1.4)	0.82 (1*)	3.64 (.26)	0.54 (1.3)	0.98 (.71)	4.90 (.14)
$\ell^{R_{-.5}}$	0.34 (2)	0.37 (1.9)	0.58 (1.4)	2.57 (.37)	0.38 (1.8)	0.69 (1*)	3.46 (0.2)