Generalised Mixability, Constant Regret, and Bayesian Updating

Mark D. Reid
The Australian National University & NICTA

Rafael M. Frongillo
Microsoft Research

Robert C. Williamson
The Australian National University & NICTA

February 8th, 2014

Abstract

Mixability of a loss is known to characterise when constant regret bounds are achievable in games of prediction with expert advice through the use of the aggregating algorithm [Vovk, 2001]. We provide a new interpretation of mixability via convex analysis that highlights the role of the Kullback-Leibler divergence in its definition. This naturally generalises to what we call Φ-mixability where the Bregman divergence D_Φ replaces the KL divergence. We prove that losses that are Φ-mixable also enjoy constant regret bounds via a generalised aggregating algorithm that is similar to mirror descent.

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. Notions of "mixability" which play a central role in the setting of prediction with expert advice offer a more general way to aggregate, and which take account of the loss function used to evaluate predictions (how well they fit the data). As shown by Vovk [2001], his more general "aggregating algorithm" reduces to Bayesian updating when log loss is used. However, as we will show there is another design variable that to date has not been fully exploited. The aggregating algorithm makes use of a distance between the current distribution and a prior which serves as a regulariser. In particular the aggregating algorithm uses the KL-divergence. We consider the general setting of an arbitrary loss and an arbitrary regulariser (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 the generalised sense then there is a generalised aggregating algorithm which can be guaranteed to have constant regret.

In symbols (more formally defined later), if we use $\ell_x(p_\theta)$ to denote the loss of the prediction p_θ by expert θ on observation x and $D_\phi(\mu', \mu)$ is used to penalise the "distance" between the choice of updated distribution μ' from what it was previously μ then we can recover both Bayesian updating and the updates of the aggregating algorithm as minimisers of $\mathbb{E}_{\theta\sim\mu'}\left[\ell_x(p_\theta)\right] + D_\Phi(\mu', \mu)$ via the choices summarised in the table below.

Scheme	Loss	Regulariser
Bayesian updating	log loss	KL divergence
Aggregating algorithm	general mixable loss	KL divergence
This paper	general Φ-mixable loss	general Bregman divergence D_Φ

We show that there is a single notion of mixability that applies to all three of these cases and guarantees the corresponding updates can be used to achieve constant regret.

We stress that the idea of the more general regularisation and updates is hardly new. See for example the discussion of potential based methods in [Cesa-Bianchi, 2006] and other references later in the paper. The key novelty is the generalised notion of mixability, the name of which is justified by the key new technical result — a constant regret bound assuming the general mixability condition achieved via a generalised algorithm which can be seen as intimately related to mirror descent. Crucially, our result depends on some properties of the conjugates of potentials defined over probabilities that do not hold for potential functions defined over more general spaces.

1.1 Prediction With Expert Advice and Mixability

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 expert reveals 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 the typically the focus of the analysis of this style of game. In particluar, 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*. We consider a version of the game where predictions made by the player and the experts are all distributions over X. The set of such distributions will be denoted Δ_X and the probability (or density) $p \in \Delta_X$ assigns to $x \in X$ will be denoted $p(x)$. A *loss* $\ell : \Delta_X \to \mathbb{R}^X$ assigns the penalty $\ell_x(p)$ to predicting $p \in \Delta_X$ when $x \in X$ is

observed. The set of experts is denoted Θ and in each round $t = 1, \ldots, T$, each expert $\theta \in \Theta$ makes a prediction $p_\theta^t \in \Delta_X$. These are revealed to the player who makes a prediction $p^t \in \Delta_X$. Once observation $x^t \in X$ is revealed the experts receive loss $\ell_{x^t}(p_\theta^t)$ and the player receives loss $\ell_{x^t}(p^t)$. The aim of the player is to minimise its *regret* $\text{Regret}^T := L^T - \min_\theta L_\theta^T$ where $L^T := \sum_{t=1}^T \ell_{x^t}(q^t)$ and $L_\theta^T = \sum_{t=1}^T \ell_{x^t}(p_\theta^t)$.

The algorithm that witnesses the original mixability result is known as the *aggregating algorithm* (AA) [Vovk, 2001]. It works similarly to exponentiated gradient algorithms [Cesa-Bianchi, 2006] in that it updates a *mixture distribution*[1] $\mu \in \Delta_\Theta$ over experts based on their performance at the end of each round. The mixture is then used to "blend" the predictions of the experts in the next round in such a way as to achieve low regret. In the aggregating algorithm, the mixture is intially set to some "prior" $\mu^0 = \pi \in \Delta_\Theta$. After $t-1$ rounds where the observations were $x^1, \ldots, x^{t-1}$ and the expert predictions where $p_\theta^1, \ldots, p_\theta^{t-1}$ for $\theta \in \Theta$ the mixture is set to

$$\mu^{t-1}(\theta) = \frac{\exp(-\eta L_\theta^{t-1})}{\sum_{\theta \in \Theta} \exp(-\eta L_\theta^{t-1})}. \tag{1}$$

On round t, after seeing all the expert predictions p_θ^t, the AA plays a $p^t \in \Delta_X$ such that for all $x \in X$

$$\ell_x(p^t) \leq -\frac{1}{\eta} \log \sum_{\theta \in \Theta} \exp(-\eta \ell_x(p_\theta^t)) \mu^t(\theta). \tag{2}$$

Mixability is precisely the condition on the loss ℓ that guarantees that such a prediction p^t can always be found.

Definition 1. *A loss $\ell : \Delta_X \to \mathbb{R}$ is said to be* η-mixable *for $\eta > 0$ if for all mixtures $\mu^t \in \Delta_\Theta$ and all predictions $\{p_\theta^t\}_{\theta \in \Theta}$ there exists a $p^t \in \Delta_X$ such that* (2) *holds for all $x \in X$.*

The key result concerning mixability is that it characterises when constant regret is achievable.

Theorem 1 (Vovk [2001])**.** *If $\ell : \Delta_X \to \mathbb{R}^X$ is η-mixable for some $\eta > 0$ then for any game of T rounds with finitely many experts Θ the aggregating algorithm will guarantee*

$$\sum_{t=1}^T \ell_{x^t}(p^t) \leq \sum_{t=1}^T \ell_{x^t}(p_\theta^t) + \frac{\log |\Theta|}{\eta}.$$

Furthermore, Vovk [2001] also supplies a converse: that a constant regret bound is only achievable for η-mixable losses. Later work by Erven et al. [2012] has show that mixability of proper losses (see §2.1) can be characterised in terms of the curvature of the corresponding entropy for ℓ, that is, in terms of $\Phi^\ell(p) = \langle p, \ell(p) \rangle$.

[1] To keep track of the two spaces X and Θ we adopt the convention of using Roman letters for distributions in Δ_X and vectors in $\mathbb{R}^X$ and Greek letters for distributions in Δ_Θ and vectors in $\mathbb{R}^\Theta$.

1.2 Contributions

Our main contribution is a generalisation of the notion of mixability and a corresponding generalisation of Theorem 1. Specifically, for any *entropy* (*i.e.*, convex function on the simplex) $\Phi : \Delta_\Theta \to \mathbb{R}$ we define Φ-mixability for losses $\ell : \Delta_X \to \mathbb{R}^X$ (Definition 2) and provide two equivalent characterisations that lend themselves to some novel interpretations. (Lemma 4). We use these characterisations to prove the follow key result. Denote by $\delta_\theta \in \Delta_\Theta$ the unit mass on θ: $\delta_\theta(\theta) = 1$, $\delta_\theta(\theta') = 0$ for all $\theta' \neq \theta$. Let D_Φ denote the Bregman divergence induced by Φ, defined formally below in (4).

Theorem 2. *If $\ell : \Delta_X \to \mathbb{R}^X$ is Φ-mixable then there is a family of strategies parameterised by $\pi \in \Delta_\Theta$ which, for any sequence of observations $x^1, \ldots, x^T \in X$ and sequence of expert predictions $p^1_\theta, \ldots, p^T_\theta$, plays a sequence $p^1, \ldots, p^T \in \Delta_X$ such that for all $\theta \in \Theta$*

$$\sum_{t=1}^{T} \ell_{x^t}(p^t) \leq \sum_{t=1}^{T} \ell_{x^t}(p^t_\theta) + D_\Phi(\delta_\theta, \pi). \tag{3}$$

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 1 is obtained as a corollary for π the uniform distribution over Θ. A compelling feature of our result is that it gives a natural interpretation of the constant $D_\Phi(\delta_\theta, \pi)$ in the regret bound: if π 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.

In addition, an algorithm analogous to the Aggregating Algorithm is naturally recovered to witness the above bound during the construction of the proof; see (12). Like the usual Aggregating Algorithm, our "generalised Aggregating Algorithm" updates its mixtures according to the past performances of the experts. However, our algorithm is most easily understood as doing so via updates to the *duals* of the distributions induced by Φ.

1.3 Related Work

The starting point for mixability and the aggregating algorithm is the work of Vovk [1995, 1990]. The general setting of prediction with expert advice is summarised in [Cesa-Bianchi, 2006, Chapters 2 and 3]. There one can find a range of results that study diffferent 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 tradeoff between tightness of the bound (in a constant factor) and the computational complexity [Kivinen and Warmuth, 1999]. Weakly mixable losses are a generalisation of mixable losses. They have been studied in Kalnishkan and Vyugin [2008] where it is shown there exists a variant of the aggregating algorithm that achives 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 in [Cesa-Bianchi, 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] who consider finding a vector w minimising $d(w, s)+\eta L(y_t, w \cdot x_t)$ 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 in [Azoury and Warmuth, 2001, §3] where they 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 regulariser is used. However, they do not consider whether similar results extend to other entropies defined on the simplex.

2 Generalised Mixability

This work was motivated by the observation that the original mixability definition (2) looks very closely related to the log-sum-exp function, which is known to be the simplex-restricted conjugate of Shannon entropy. We wondered whether the proof that mixability implies constant regret was due to unique properties of Shannon entropy or whether alternative notions of entropy could lead to similar results. We found that the key step of the original mixability proof (that allows the sum of bounds to telescope) holds for any convex function defined on the simplex. This is because the conjugates of such functions have a translation invariant property that allows the original telescoping series argument to go through in the general case. By re-expressing the original proof using only the tools of convex analysis we were able to naturally derive the corresponding update algorithm and express the constant term in the bound as a Bregman divergence.

2.1 Preliminaries

We begin by introducing some basic concepts and notation from convex analysis. Terms not defined here can be found in a reference such as [Hiriart-Urruty and Lemaréchal, 2001]. A convex function $\Phi : \Delta_\Theta \to \mathbb{R}$ is called an *entropy* if it is proper, convex, and lower semi-continuous. The *Bregman divergence* associated with a suitably differentiable entropy Φ is given by

$$D_\Phi(\mu, \mu') = \Phi(\mu) - \Phi(\mu') - \langle \nabla\Phi(\mu'), \mu - \mu' \rangle \tag{4}$$

for all $\mu \in \Delta_\Theta$ and $\mu' \in \mathrm{ri}(\Delta_\Theta)$, the relative interior of Δ_Θ. The *convex conjugate* of $\Phi : \Delta_\Theta \to \mathbb{R}$ is defined to be $\Phi^*(v) := \sup_{\mu \in \mathrm{dom}\,\Phi} \langle \mu, v \rangle - \Phi(\mu) = \sup_{\mu \in \Delta_\Theta} \langle \mu, v \rangle - \Phi(\mu)$ where $v \in \Delta_\Theta^*$, *i.e.*, the dual space to Δ_Θ. One could also write the supremum over $\mathbb{R}^\Theta$ by the convention of setting $\Phi(\mu) = +\infty$ for $\mu \notin \Delta_\Theta$. For differentiable Φ, it is known that the supremum defining Φ^* is attained at $\mu = \nabla\Phi^*(v)$ [Hiriart-Urruty and Lemaréchal, 2001]. That is,

$$\Phi^*(v) = \langle \nabla\Phi^*(v), v \rangle - \Phi(\nabla\Phi^*(v)). \tag{5}$$

A similar result holds for Φ by applying this result to Φ^* and using $\Phi = (\Phi^*)^*$. We will make use of this result to establish the following inequality connecting a Bregman divergence D_Φ with its conjugate.

Lemma 1. *For all $\mu \in \Delta_\Theta$ and $v \in \Delta_\Theta^*$ we have*

$$\Phi^*(\nabla\Phi(\mu)) - \Phi^*(\nabla\Phi(\mu) - v) = \inf_{\mu' \in \Delta_\Theta} \langle \mu', v \rangle + D_\Phi(\mu', \mu).$$

Proof. By definition $\Phi^*(\nabla\Phi(\mu) - v) = \sup_{\mu' \in \Delta_\Theta} \langle \mu', \nabla\Phi(\mu) - v \rangle - \Phi(\mu')$ and using (5) expands $\Phi^*(\nabla\Phi(\mu))$ to $\Phi^*(\nabla\Phi(\mu)) = \langle \mu, \nabla\Phi(\mu) - \Phi(\mu) \rangle$. 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 $\inf_{\mu' \in \Delta_\Theta} \Phi(\mu') - \Phi(\mu) - \langle \nabla\Phi(\mu), \mu' - \mu \rangle + \langle \mu', v \rangle$ which then gives the result. □

We will also make use of a property of conjugates of entropies called *translation invariance* [Othman and Sandholm, 2011]. This notion is central to what are called convex and coherent risk functions in mathematical finance [Föllmer and Schied, 2004]. In the following result and throughout, we use $\mathbf{1} \in \mathbb{R}^\Theta$ for the point such that $\mathbf{1}_\theta = 1$ for all $\theta \in \Theta$.

Lemma 2. *If $\Phi : \Delta_\Theta \to \mathbb{R}$ is an entropy then its convex conjugate is* translation invariant*, that is, for all $v \in \Delta_\Theta^*$ and $\alpha \in \mathbb{R}$ we have $\Phi^*(v + \alpha\mathbf{1}) = \Phi^*(v) + \alpha$ and its gradient satisfies $\nabla\Phi^*(v + \alpha\mathbf{1}) = \nabla\Phi^*(v)$.*

Proof. By definition of the convex conjugate we have

$$\begin{aligned} \Phi^*(v + \alpha\mathbf{1}) &= \sup_{\mu \in \Delta_\Theta} \langle \mu, v + \alpha\mathbf{1} \rangle - \Phi(\mu) \\ &= \sup_{\mu \in \Delta_\Theta} \langle \mu, v \rangle - \Phi(\mu) + \alpha \\ &= \Phi^*(v) + \alpha \end{aligned}$$

since $\langle \mu, \mathbf{1} \rangle = 1$. Taking derivatives of both sides gives the second part of the lemma. □

We will also make use of the readily established fact that for any convex $\Phi : \Delta_\Theta \to \mathbb{R}$ and all $\eta > 0$ we have $(\frac{1}{\eta}\Phi)^*(v) = \frac{1}{\eta}\Phi^*(\eta v)$.

Probably the most well studied example of what we call an entropy is the negative of the Shannon entropy[2] $H(\mu) = -\sum_{\theta\in\Theta} \mu(\theta) \log \mu(\theta)$ which is known to be concave, proper, and upper semicontinuous and thus $\Phi = -H$ is an entropy. When we look at the form of the original definition of mixability, we observe that it is closely related to the conjugate of $(-H)$:

$$(-H)^*(v) = \log \sum_{\theta\in\Theta} \exp(v(\theta)) \tag{6}$$

which is sometimes called the *log-sum-exp* or *partition function*. This observation is what motivated this work and drives our generalisation to other entropies.

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]. Specifically, if a loss $\lambda : \Delta_\Theta \to \mathbb{R}^\Theta$ is used to assign a penalty $\lambda_\theta(\mu)$ to a prediction μ upon outcome θ it is said to be *proper* if its expected value under $\theta \sim \mu$ is minimsed by predicting μ. That is, for all $\mu, \mu' \in \Delta_\Theta$ we have

$$\mathbb{E}_{\theta\sim\mu}\left[\lambda_\theta(\mu')\right] = \langle \mu, \lambda(\mu')\rangle \geq \langle \mu, \lambda(\mu)\rangle =: -\Phi^\lambda(\mu)$$

where $-\Phi^\lambda$ is the *Bayes risk* of λ and is necessarily concave [Erven et al., 2012], thus making $\Phi^\lambda : \Delta_\Theta \to \mathbb{R}$ convex and thus an entropy. The correspondence also goes the other way: given any convex function $\Phi : \Delta_\Theta \to \mathbb{R}$ we can construct a unique proper loss. The following representation can be traced back to Savage [1971] but is expressed here using conjugacy.

Lemma 3. *If $\Phi : \Delta_\Theta \to \mathbb{R}$ is a differentiable entropy then the loss $\lambda^\Phi : \Delta_\Theta \to \mathbb{R}$ defined by*

$$\lambda^\Phi(\mu) := \Phi^*(\nabla\Phi(\mu))\mathbf{1} - \nabla\Phi(\mu) \tag{7}$$

is proper.

Proof. By eq. (5) we have $\Phi^*(\nabla\Phi(\mu)) = \langle \mu, \nabla\Phi(\mu)\rangle - \Phi(\mu)$, giving us

$$\begin{aligned}
\left\langle \mu, \lambda^\Phi(\mu')\right\rangle - \left\langle \mu, \lambda^\Phi(\mu)\right\rangle &= \Big(\langle \mu', \nabla\Phi(\mu')\rangle - \Phi(\mu') - \langle \mu, \nabla\Phi(\mu')\rangle\Big) \\
&\quad - \Big(\langle \mu, \nabla\Phi(\mu)\rangle - \Phi(\mu) - \langle \mu, \nabla\Phi(\mu)\rangle\Big) \\
&= D_\Phi(\mu, \mu'),
\end{aligned}$$

from which propriety follows. □

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

[2]We write Shannon entropy here as a sum but can also consider the continuous version relative to some reference measure $\nu \in \Delta_\Theta$, that is, $H(\mu) = -\int_{\Delta_\Theta} \log(\mu(\theta))\mu(\theta)\, d\nu(\theta)$. For simplicitly, we stick to the countable case.

2.2 Φ-Mixability

For a loss $\ell : \Delta_X \to \mathbb{R}^X$ define the *assessment* $\alpha : X \to \mathbb{R}^\Theta$ to be the loss of each model/expert p_θ on observation x, *i.e.*, $\alpha_\theta(x) := \ell_x(p_\theta)$.

Definition 2. *Suppose Φ is a differentiable entropy on Δ_Θ. A loss $\ell : \Delta_X \to \mathbb{R}^X$ is* Φ-mixable *if for all $\{p_\theta\}_\theta$ and all $\mu \in \Delta_\Theta$ there is a $p \in \Delta_X$ such that for all $x \in X$,*

$$\ell_x(p) \leq -\Phi^*(-\lambda^\Phi(\mu) - \alpha(x)). \tag{8}$$

We can readily show that this definition reduces to the standard mixability definition when $\Phi = \frac{1}{\eta}(-H)$ since, in this case,

$$\Phi^*(v) = \frac{1}{\eta} \log \sum_\theta \exp(\eta v(\theta)) \tag{9}$$

by (6) and the fact that $(\frac{1}{\eta} f)^*(x^*) = \frac{1}{\eta} f^*(\eta x^*)$ for any convex f. As mentioned above, the proper loss corresponding to this choice of Φ is easily seen to be $\lambda^\Phi_\theta(\mu) = -\frac{1}{\eta} \log(\mu(\theta))$ by substitution into (7). Thus, the mixability inequality becomes $\ell_x(p) \leq -\frac{1}{\eta} \log \sum_\theta \exp(-\eta\alpha(x) + \log \mu(\theta))$ which is equivalent to (2).

We now show that the above definition is equivalent to one involving the Bregman divergence for Φ and also the difference in the "potential" Φ^* evaluated at $\nabla\Phi(\mu)$ before and after it is updated by $\alpha(x)$.

Lemma 4. *Suppose Φ is a differentiable entropy on Δ_Θ. Then the Φ-mixability condition* (8) *is equivalent to the following:*

$$\ell_x(p) \leq \inf_{\mu' \in \Delta_\Theta} \langle \mu', \alpha(x) \rangle + D_\Phi(\mu', \mu), \tag{10}$$

$$\ell_x(p) \leq \Phi^*(\nabla\Phi(\mu)) - \Phi^*(\nabla\Phi(\mu) - \alpha(x)). \tag{11}$$

Proof. Expanding the definition of $\lambda^\Phi(\mu)$ makes the right-hand side of (8) equal to

$$-\Phi^*(-\Phi^*(\nabla\Phi(\mu))\mathbf{1} + \nabla\Phi(\mu) - \alpha(x)) = -\Phi^*(\nabla\Phi(\mu) - \alpha(x)) + \Phi^*(\nabla\Phi(\mu))$$

since Φ^* is translation invariant by Lemma 2. This gives (11). Further applying Lemma 1 with $v = \alpha(x)$ gives (10). □

3 The Generalised Aggregating Algorithm

In this section we prove our main result (Theorem 2) and examine the "generalised Aggregating Algorithm" that witnesses the bound. The updating strategy we use is the one that repeatedly returns the minimiser of the right-hand side of (10).

Definition 3. *On round t, after observing $x^t \in X$, the* generalised aggregating algorithm *(GAA) updates the mixture $\mu^{t-1} \in \Delta_\Theta$ by setting*

$$\mu^t := \operatorname*{arg\,min}_{\mu \in \Delta_\Theta} \left\langle \mu, \alpha(x^t) \right\rangle + D_\Phi(\mu, \mu^{t-1}). \tag{12}$$

The next lemma shows that this updating process simply aggregates the assessments in the dual space Δ_Θ^* with $\nabla\Phi(\pi)$ as the starting point.

Lemma 5. *The GAA updates μ^t satisfy $\nabla\Phi(\mu^t) = \nabla\Phi(\mu^{t-1}) - \alpha(x^t)$ for all t and so*

$$\nabla\Phi(\mu^T) = \nabla\Phi(\pi) - \sum_{t=1}^{T} \alpha(x^t). \tag{13}$$

Proof. By considering the Lagrangian $\mathcal{L}(\mu, a) = \langle \mu, \alpha(x^t)\rangle + D_\Phi(\mu, \mu^{t-1}) + a(\langle \mu, \mathbf{1}\rangle - 1)$ and setting its derivative to zero we see that the minimising μ^t must satisfy $\nabla\Phi(\mu^t) = \nabla\Phi(\mu^{t-1}) - \alpha(x^t) - a^t\mathbf{1}$ where the $a^t \in R$ is the dual variable at step t. For convex Φ, the functions $\nabla\Phi^*$ and $\nabla\Phi$ are inverses [Hiriart-Urruty and Lemaréchal, 2001] so $\mu^t = \nabla\Phi^*(\nabla\Phi(\mu^{t-1}) - \alpha(x^t) - a^t\mathbf{1}) = \nabla\Phi^*(\nabla\Phi(\mu^{t-1}) - \alpha(x^t))$ by the translation invariance of Φ^* (Lemma 2). This means the constants a^t are arbitrary and can be ignored. Thus, the mixture updates satisfy the relation in the lemma and summing over $t = 1, \ldots, T$ gives (13). □

To see how the updates just described are indeed a generalisation of those used by the original aggregating algorithm, we can substitute $\Phi = -\frac{1}{\eta}H$ and $\pi = \frac{1}{|\Theta|}$ in (12). Because H is maximal for uniform distributions we must have $\nabla\Phi(\pi) = -\frac{1}{\eta}\nabla H(\pi) = 0$ and so $\mu^T = \nabla\Phi^*(-\sum_{t=1}^{T}\alpha(x^t))$. However, by (10) we see that

$$[\nabla\Phi^*(v)]_\theta = \frac{e^{\eta v(\theta)}}{\sum_{\theta'} e^{\eta v(\theta')}}$$

and then substituting $v(\theta) = -\sum_t \alpha(x^t)$ gives the update equation in (1).

3.1 The proof of Theorem 2

Armed with the representations of Φ-mixability in Lemma 4 and the form of the updates in Lemma 5, we now turn to the proof of our main result.

of Theorem 2. By assumption, ℓ is Φ-mixable and so, for the updates μ^t just defined we have that there exists a $p^t \in \Delta_X$ such that $\ell_{x^t}(p^t) \le -\Phi^*(-\lambda^\Phi(\mu^{t-1}) - \alpha(x^t))$ for all $x^t \in X$. Expressing these bounds using (11) from Lemma 4 and summing these over $t = 1, \ldots, T$ gives

$$\begin{aligned}
\sum_{t=1}^{T} \ell_{x^t}(p^t) &\le \sum_{t=1}^{T} \Phi^*(\nabla\Phi(\mu^{t-1})) - \Phi^*(\nabla\Phi(\mu^{t-1}) - \alpha(x^t)) \\
&= \Phi^*(\nabla\Phi(\mu^0)) - \Phi^*(\nabla\Phi(\mu^T)) && (14)\\
&= \inf_{\mu' \in \Delta_\Theta} \left\langle \mu', \sum_{t=1}^{T} \alpha(x^T) \right\rangle + D_\Phi(\mu', \pi) && (15)\\
&\le \left\langle \mu', \sum_{t=1}^{T} \alpha(x^t) \right\rangle + D_\Phi(\mu', \pi) \qquad \text{for all } \mu' \in \Delta_\Theta && (16)
\end{aligned}$$

Line (14) above is because $\nabla\Phi(\mu^t) = \nabla\Phi(\mu^{t-1}) - \alpha(x^t)$ by Lemma 5 and the series telescopes. Line (15) is obtained by applying (12) from Lemma 5 and Lemma 1. Setting $\mu' = \delta_\theta$ and noting $\langle \delta_\theta, \alpha(x^t)\rangle = \ell_{x^t}(p_\theta)$ gives the required result. □

Note that the proof above gives us something even stronger — eq. (16) states that the GAA satisfies the stronger condition that eq. (3) hold for all $\mu \in \Delta_\Theta$, in addition to all δ_θ, where the loss is an expected loss under μ. In particular, choosing $T = 1$ in eq. (15), we have

$$\ell_{x^1}(p^1) \leq \inf_{\mu\in\Delta_\Theta} \left\langle \mu, \alpha(x^1)\right\rangle + D_\Phi(\mu, \pi),$$

from which we can conclude that ℓ is actually Φ-mixable from Lemma 4. Hence, an algorithm exists which guarantees the bound in eq. (16) if and only if the loss ℓ is Φ-mixable.

Finally, we briefly note some similarities between the Generalised Aggregating Algorithm and the literature on automated market makers for prediction markets. The now-standard framework of Abernethy et al. [2013] defines the cost of a purchase of some bundle of securities as the difference in a convex potential function. Formally, for some convex $C : \mathbb{R}^n \to \mathbb{R}$, a purchase of bundle $r \in \mathbb{R}^n$ given current market state q is given by $C(q + r) - C(q)$. The instantaneous prices in the market at state q are therefore $p = \nabla C(q)$. As the prices correspond to probabilities in their framework, it must be the case that $R := C^*$ satisfies $\mathrm{dom}(R) = \Delta_n$. From this we can conclude as we have done above that C is translation invariant, and thus one can restate the cost of the bundle r as $R^*(\nabla R(p) + r) - R^*(\nabla R(p))$.

We now are in a position to draw an anology with our GAA. The formulation of Φ-mixability in eq. (11) says that the loss upon observing x must be bounded above by $\Phi^*(\nabla\Phi(\mu)) - \Phi^*(\nabla\Phi(\mu) + \alpha(x))$, which is exactly the negative of the expression above, where $R = \Phi$, $p = \mu$, and $r = \alpha(x)$. Thus, Φ-mixability is saying the loss must be at least as good for the algorithm than in the market making setting, and hence it is not surprising that the loss bounds are the same in both settings; see Abernethy et al. [2013] for more details.

4 Future Work

Our exploration into a generalized notion of mixability opens more doors than it closes. In the following, we briefly outline several open directions.

Relation to original mixability result

The proof of our main result, Theorem 2, shows that in essence, an algorithm can guarantee constant regret, expressed in terms of a Φ-divergence between a starting point and the best expert, if and only if the underlying loss ℓ is Φ-mixable. The original mixability result of Vovk [2001] states that one achieves

a constant regret of $\log|\Theta|/\eta$ if and only if ℓ is, in our terminology, $(-\eta^{-1}H)$-mixable. But of course for any Φ which is bounded on Δ_Θ, the penalty $D_\Phi(\delta_\theta, \pi)$ is also bounded, and hence it would seem that for all bounded $\Phi : \Delta_\Theta \to \mathbb{R}$, a loss ℓ is mixable in the sense of Vovk if and only if it is $\eta^{-1}\Phi$-mixable for some $\eta > 0$.

Relation to curvatures of ℓ and Φ

A recent result of Erven et al. [2012] shows that the mixability constant η from the original Definition 1 can be calculated as the ratio of curvatures between the Bayes risk of the loss ℓ and Shannon entropy. It would stand to reason therefore that for any Φ, the Φ-mixability constant η for a loss ℓ, defined as the largest η such that ℓ is $\eta^{-1}\Phi$-mixable, would be similarly defined as the ratio to $-\Phi$ instead of H.

Optimal regret bound

The curvature discussion above addresses the question of finding, given a Φ, the largest η such that ℓ is $\eta^{-1}\Phi$-mixable. Note that the larger η is, the smaller the corresponding regret term $\eta^{-1}D_\Phi(\delta_\theta, \pi)$ is. Hence, for fixed Φ, this Φ-mixability constant yields the tightest bound. The question remains, however, what is the tightest bound one can acheive across *all* choices of Φ? Again in reference to Vovk, it seems that the choice of Φ may not matter, at least as long as $D_\Phi(\delta_\theta, \pi)$ is a constant independent of θ. It would be clarifying to directly assert this claim or find a counter-example.

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. URL `http://dl.acm.org/citation.cfm?id=2465777`.

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. *Prediction, learning, and games*. Cambridge University Press, 2006.

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

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*, 13:1639–1663, 2012.

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

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

J.B. Hiriart-Urruty and C. 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 *Computational Learning Theory*, pages 153–167. Springer, 1999.

Abraham Othman and Tuomas Sandholm. Liquidity-sensitive automated market makers via homogeneous risk measures. In *Internet and Network Economics*, pages 314–325. Springer, 2011.

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

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

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

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