From Exp-concavity to Mixability

Tim van Erven

December 12, 2012
Updated: December 3, 2013

Abstract

In sequential prediction (online learning) with expert advice the goal is to predict a sequence of outcomes almost as well as the best advisor from a pool of experts. The quality of predictions is measured by a *loss function*, which is determined by the application one has in mind. For most loss functions, the best performance that can be guaranteed is to be within $O(\sqrt{T})$ of the best expert on T outcomes, but for some special loss functions $O(1)$ overhead is possible.

In the 1990's, people familiar with the work of Vovk called these special loss functions *mixable losses*, but nowadays the notion of mixability appears to be mostly forgotten, and the geometric concept of *exp-concavity* has taken its place. This raises the question of how the two are related, which strangely does not appear to be answered in very much detail in the literature. As I have been studying mixability quite a bit in my recent work, I was wondering about this, so here are some thoughts. **Update:** In particular, I will construct a parameterization of the squared loss in which it is 1/2-exp-concave instead of only 1/8-exp-concave like in its usual parameterization.

1 Mixability and Exp-concavity

Suppose we predict an outcome $y \in \mathcal{Y}$ by specifying a prediction $a \in \mathcal{A}$. The better our prediction, the smaller our loss $\ell(y, a)$. (I will assume $\ell(y, a)$ is nonnegative, but that does not really matter.) For example, if y and a both take values in $\{0, 1\}$, then the *0/1-loss* $\ell(y, a) = |y - a|$ is 0 if we predict correctly and 1 otherwise. Alternatively, if y and a are both real-valued, then the *squared loss* is $\ell(y, a) = (y - a)^2$. And finally, if a specifies a probability density f_a on $\mathcal{Y}$, then our loss may be the *log loss* $\ell(y, a) = -\ln f_a(y)$.

Mixability For $\eta > 0$, a loss function is called *η-mixable* [1] if for any probability distribution π on $\mathcal{A}$ there exists a prediction $a_\pi \in \mathcal{A}$ such that

$$e^{-\eta \ell(y, a_\pi)} \geq \int e^{-\eta \ell(y,a)} \pi(\mathrm{d}a) \qquad \text{for all } y \in \mathcal{Y}. \tag{1}$$

The constant in the $O(1)$ overhead compared to the best expert is proportional to $1/\eta$, so the bigger η the better.

Exp-concavity For $\eta > 0$, a loss function is called η-exp-concave if for any distribution π on $\mathcal{A}$ the prediction $a_\pi = \int a \, \pi(\mathrm{d}a)$ satisfies (1).

So exp-concavity is just mixability with a_π fixed to be the mean. This choice is appropriate in the case of log loss. In this case, for $\eta = 1$, the numbers $e^{-\eta\ell(y,a)} = f_a(y)$ just equal probability densities and (1) holds with equality.

For squared loss, however, the appropriate choice for a_π is not the mean. Suppose that y and a both take values in $[-1, +1]$. Then, while the squared loss is 1/2-mixable for $a_\pi = \frac{h_{1/2}(-1) - h_{1/2}(1)}{4}$ with $h_\eta(y) = \frac{-1}{\eta} \ln \int e^{-\eta(y-a)^2} \pi(\mathrm{d}a)$, it is only 1/8-exp-concave when parameterized by a. (See [2, 3].) This does not rule out, however, that the squared loss might be 1/2-exp-concave in a different parameterization. As we shall see, such a parameterization indeed exists ~~if we restrict y to take only two values $\{-1, +1\}$, but I have not been able to find a suitable reparameterization in general~~.

2 Relations

Clearly, exp-concavity implies mixability: it just makes the choice for a_π explicit. What is not so obvious, is when the implication also goes the other way. It turns out that in some cases it actually does if we reparameterize our predictions in a clever (one might also say: complicated) way by the elements of a certain set $\mathcal{B}_\eta$.

Theorem 1. *Suppose a loss $\ell\colon \mathcal{Y} \times \mathcal{A} \to [0, \infty]$ satisfies Conditions 1 and 2 below for some $\eta > 0$. Then ℓ is η-mixable if and only if it can be parameterized in such a way that it is η-exp-concave.*

The technical conditions I need are the following:

1. All predictions in $\mathcal{A}$ should be admissible.
2. For any element g on the north-east boundary of the set $\mathcal{B}_\eta$, there should exist a prediction $a \in \mathcal{A}$ such that $g(y) = e^{-\eta\ell(y,a)}$ for all y.

It remains to explain what these conditions mean, and discuss their severity. I will argue that Condition 1 is very mild. Condition 2 also appears to be generally satisfied if the dimensionality of the set of predictions equals the number of possible predictions minus one, i.e. $\dim(\mathcal{A}) = |\mathcal{Y}| - 1$, but not in general. For example, for the squared loss we predict by a single number a, so $\dim(\mathcal{A}) = 1$ and hence we have $\dim(\mathcal{A}) = |\mathcal{Y}| - 1$ if y only takes two different values, but not if $\mathcal{Y}$ is the whole range $[-1, +1]$. **Update:** We can work around this, though. See below.

3 The Technical Conditions

Admissibility Condition 1 is the easiest of the two. I will call a prediction $a \in \mathcal{A}$ *admissible* if there exists no other prediction $b \in \mathcal{A}$ that is always at least as good in the sense that $\ell(y, b) \leq \ell(y, a)$ for all $y \in \mathcal{Y}$. If a is inadmissible, then we could just remove it from the set of available predictions $\mathcal{A}$, because predicting b is always at least as good anyway. So admissibility seems more of an administrative requirement (get rid of all predictions that make no sense) than a real restriction.

Condition 2 To explain the second condition, we define the new parameterization $\mathcal{B}_\eta$ as the set of functions

$$\mathcal{B}_\eta = \{g\colon \mathcal{Y} \to [0,1] \mid \text{for some distribution } \pi\colon\ g(y) = \int e^{-\eta\ell(y,a)}\pi(\mathrm{d}a)\ \forall y\}.$$

Note that the set $\mathcal{B}_\eta$ is convex by construction.

Let $\mathbb{1}(y) = 1$ be the constant function that is 1 on all $y \in \mathcal{Y}$, and for any $g \in \mathcal{B}_\eta$ let $c(g) = \sup\{c \geq 0 \mid (g + c \cdot \mathbb{1}) \in \mathcal{B}_\eta\}$. By the *north-east boundary* of $\mathcal{B}_\eta$, I mean the set of points $\{g + c(g) \mid g \in \mathcal{B}_\eta\}$. That is, if we move 'south-east' from any point in this set (in the direction of $-\mathbb{1}$), we are inside $\mathcal{B}_\eta$, but if we move further 'north-east' (in the direction of $\mathbb{1}$) we are outside.

Condition 2 implies that the north-east boundary of $\mathcal{B}_\eta$ should be equal to the set $\{e^{-\eta\ell(\cdot,a)} \mid a \in \mathcal{A}\}$, which appears to be quite typical if $\dim(A) = |\mathcal{Y}| - 1$, but not in general.

4 Construction of the Parameterization and Proof

As we have already seen that η-exp-concavity trivially implies η-mixability, it remains to construct the parameterization in which ℓ is η-exp-concave given that it is η-mixable.

The parameterization we choose is indexed by the elements of $\mathcal{B}_\eta$, which we map onto $\mathcal{A}$, with multiple elements in $\mathcal{B}_\eta$ mapping to the same element of $\mathcal{A}$. So let g be an arbitrary element of $\mathcal{B}_\eta$. How do we map it to a prediction $a \in \mathcal{A}$? We do this by choosing the prediction a such that $g(y) + c(g) = e^{-\eta\ell(y,a)}$ for all y. As $g + c(g) \cdot \mathbb{1}$ lies on the north-east boundary of $\mathcal{B}_\eta$, such a prediction exists by Condition 2.

Our construction ensures there exists a $g \in \mathcal{B}_\eta$ that maps to a for any $a \in \mathcal{A}$. To see this, suppose there was an a for which this was not the case, and let $g_a = e^{-\eta\ell(\cdot,a)}$. Then we must have $c(g_a) > 0$, because otherwise we would have $c(g) = 0$ and g_a would map to a. But then the prediction $b \in \mathcal{A}$ such that $e^{-\eta\ell(\cdot,b)} = g + c(g) \cdot \mathbb{1}$ would satisfy $e^{-\eta\ell(y,b)} > e^{-\eta\ell(y,a)}$ for all y, and hence $\ell(y,b) < \ell(y,a)$ for all y, so that a would be inadmissible, which we have ruled out by assumption.

We are now ready to prove that the loss is η-exp-concave in our parameterization. To show this, let π be an arbitrary probability distribution on $\mathcal{B}_\eta$. Then we need to show that

$$e^{-\eta\ell(y,g_\pi)} \geq \int e^{-\eta\ell(y,g)}\pi(\mathrm{d}g) \qquad \text{for all } y \in \mathcal{Y},$$

where $g_\pi = \int g\ \pi(\mathrm{d}g)$. To this end, observe that

$$\int e^{-\eta\ell(\cdot,g)}\pi(\mathrm{d}g) = \int (g + c(g) \cdot \mathbb{1})\ \pi(\mathrm{d}g) = g_\pi + c_\pi \cdot \mathbb{1},$$

where $c_\pi = \int c(g)\ \pi(\mathrm{d}g)$. Now convexity of $\mathcal{B}_\eta$ ensures that $\int e^{-\eta\ell(\cdot,g)}\pi(\mathrm{d}g) \in \mathcal{B}_\eta$, so that we must have $c_\pi \leq c(g_\pi)$. But then

$$e^{-\eta\ell(y,g_\pi)} = g_\pi(y) + c(g_\pi) \geq g_\pi(y) + c_\pi = \int e^{-\eta\ell(y,g)}\pi(\mathrm{d}g)$$

for all y, which was to be shown.

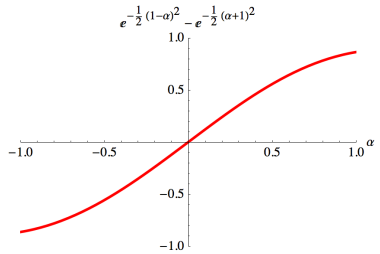

Figure 1: Plot of $f(\alpha)$ on $[-1,+1]$

5 Squared Loss

So how do things play out for the squared loss? We know that it is $1/2$-mixable, so we would like to find a parameterization in which it is also $1/2$-exp-concave. Suppose first that a takes values in $[-1,+1]$ and y takes only two values $\{-1,+1\}$. Then Condition 1 is clearly satisfied. The set $\mathcal{B}_{1/2}$ consists of all the functions $g\colon \{-1,+1\} \to [e^{-2},1]$ such that

$$g(y) = \int e^{-\frac{1}{2}(y-a)^2} \pi(\mathrm{d}a) \qquad \text{for } y \in \{-1,+1\} \tag{2}$$

for some distribution π on $\mathcal{A}$. So to verify Condition 2, we need to check that for any $g \in \mathcal{B}_{1/2}$ there exists a prediction $a_g \in \mathcal{A}$ that satisfies

$$g(y) + c(g) = e^{-\frac{1}{2}(y-a_g)^2} \qquad \text{for } y \in \{-1,+1\}. \tag{3}$$

Solving this we find that a_g indeed exists and equals

$$a_g = f^{-1}\big(g(1) - g(-1)\big), \tag{4}$$

where f^{-1} is the inverse of $f(\alpha) = e^{-\frac{1}{2}(1-\alpha)^2} - e^{-\frac{1}{2}(\alpha+1)^2}$ (see Figure 1). The existence of a_g for all g implies that Condition 2 is satisfied, and by Theorem 1 we have found a parameterization in which the squared loss is $1/2$-exp-concave, provided that y only takes the values $\{-1,+1\}$.

So what happens if we allow y to vary over the whole range $[-1,+1]$? In this case I believe that no choice of a_g will satisfy (3) for all y, and consequently Condition 2 does not hold. **Update:** However, it turns out that any parametrization that is η-exp-concave for $y \in \{-1,+1\}$ is also η-exp-concave for the whole range $y \in [-1,+1]$. This is a special property, proved by Haussler, Kivinen and Warmuth [2, Lemma 4.1], [3, Lemma 3], that only holds for certain loss functions, including the squared loss. Thus we have found a parameterization of the squared loss with $y \in [-1,+1]$ in which it is $1/2$-exp-concave (instead of only $1/8$-exp-concave like in the standard parameterization): parameterize by the functions g defined in (2), and map them to original parameters via the mapping a_g defined in (4).

6 Discussion

We have seen that exp-concavity trivially implies mixability. Conversely, mixability also implies exp-concavity roughly when the dimensionality of the set of

predictions $\dim(\mathcal{A})$ equals the number of outcomes $|\mathcal{Y}|$ minus one. In general, however, it remains unknown whether any η-mixable loss can be reparameterized to make it η-exp-concave with the same η.

As exp-concavity is a stronger requirement than mixability and introduces these complicated reparameterization problems, one might ask: why bother with it at all? One answer to this is that taking a_π to be the mean reduces the requirement (1) to ordinary concavity, which has a nice geometrical interpretation. Nevertheless, the extra flexibility offered by mixability can make it easier to satisfy (for example, for the squared loss), so in general mixability would appear to be the most convenient of the two properties.

6.1 Afterthought

It seems the proof of Theorem 1 would still work if we replaced $\mathbb{1}$ by any other positive function. I wonder whether this extra flexibility might make Condition 2 easier to satisfy.

Acknowledgements

Update: I would like to thank Sébastien Gerchinovitz for pointing me to Lemma 4.1 of Haussler, Kivinen and Warmuth.

References

[1] N. Cesa-Bianchi and G. Lugosi. *Prediction, learning, and games*. Cambridge University Press, 2006.

[2] D. Haussler, J. Kivinen, and M. K. Warmuth. Sequential prediction of individual sequences under general loss functions. *IEEE Transactions on Information Theory*, 44(5):1906–1925, 1998.

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