
(No) QUADRATIC LOSS BOUNDS FOR BAYES MIXTURE PREDICTION

Jan Poland

Abstract

For Bayes mixture predictions, we study “quadratic” loss bounds of the form

$$\sum_t (\ell_t^\xi - \ell_t^\mu) = O(K_\mu + \sqrt{K_\mu \sum_t (\ell_t^\mu)^2})$$

1 Notation

This is just a short note showing a negative result. It is not very self-contained. Probably it is only comprehensible to a reader familiar with at least one of [1, 2, 3]. Looking for such a result was motivated by [4].

Let $\mathcal{C} = \{\nu_1, \nu_2, \dots\}$ be a countable class of (semi-)measures on $\{0, 1\}^*$. (We do everything on binary alphabet, although things most probably generalize to arbitrary alphabet.) Each model $\nu \in \mathcal{C}$ has a weight $w_\nu > 0$ and a complexity $K_\nu = \ln w_\nu^{-1}$. There is a *true distribution* $\mu \in \mathcal{C}$ which we require to be a *measure*. The Bayes mixture over the model class is denoted by ξ . At each time step t and possibly depending on the past $x_{<t}$, there is a *loss matrix*

$$\mathcal{L} = \mathcal{L}(t, x_{<t}) = \left(\overbrace{\begin{pmatrix} \ell^{00} & \ell^{01} \\ \ell^{10} & \ell^{11} \end{pmatrix}}^{\text{true symbol}} \right) \Big\} \text{prediction.} \quad (1)$$

We assume that the prediction schemes know the loss matrix and make according *Bayes-optimal* predictions: If ν is the predictor’s current probability that the symbol 1 will occur, then the output is 1 if

$$(1 - \nu)\ell^{10} + \nu\ell^{11} \leq (1 - \nu)\ell^{00} + \nu\ell^{01}. \quad (2)$$

In case of equality, the predictions may be chosen freely (e.g. in order to facilitate the analysis). We concentrate on the informed predictor which knows the true distribution μ , and the Bayes mixture predictor which estimates the true

distribution by ξ . We abbreviate

$$\begin{aligned}
\ell_t &= \mathbf{E}(\ell_t^\mu | x_{<t}) = \text{expected instantaneous loss of informed predictor}, \\
\delta_t &= \mathbf{E}(\ell_t^\xi - \ell_t^\mu | x_{<t}) = \text{expected regret of Bayes mixture}, \\
d_t &= D(\mu(1|x_{<t}) || \xi(1|x_{<t})) = \text{instantaneous Kullback-Leibler divergence} \\
L_T &= \sum_{t=1}^T \mathbf{E}_{x_{<t}} \ell_t = \text{cumulative expected loss}, \\
L_T^2 &= \sum_{t=1}^T \mathbf{E}_{x_{<t}} \ell_t^2 = \text{cumulative expected square loss}, \\
\Delta_T &= \sum_{t=1}^T \mathbf{E}_{x_{<t}} \delta_t = \text{cumulative expected regret}.
\end{aligned}$$

Furthermore we will abuse notation and write μ, ξ instead of $\mu(1|x_{<t}), \xi(1|x_{<t})$.

2 Reduction of the loss matrix

We construct worst-case loss matrices for each time step, depending on the actual μ and ξ .

Assume that losses are in $[0, 1]$ and consider the loss matrix (1). Define $\ell^0 = \ell^{10} - \ell^{00}$ and $\ell^1 = \ell^{01} - \ell^{11}$, then (2) is equivalent to $(1 - \nu)\ell^0 \leq \nu\ell^1$. Assume w.l.o.g. that $\ell^0, \ell^1 \geq 0$. Then, for any $\mu, \xi \in [0, 1]$ replacing ℓ^{01} by ℓ^0 , ℓ^{10} by ℓ^1 , and ℓ^{00} and ℓ^{11} in the loss matrix has the following effect: $\ell = \ell^\mu$ is reduced, while $\delta = \ell^\xi - \ell^\mu$ remains unchanged.

After this replacement, we further tune ℓ^0 and ℓ^1 in order to force the ξ -predictor to have maximal regret. If the μ - and the ξ -predictions coincide, then $\delta = 0$. So we must assert that the predictions differ. We choose ℓ^0 and ℓ^1 such that $(1 - \xi)\ell^0 = \xi\ell^1$ and assume that ξ outputs just the opposite prediction of μ . Together with the constraint $\ell^0, \ell^1 \in [0, 1]$, we get the following closed expressions for ℓ_t and δ_t :

$$\begin{aligned}
\delta_t &= \frac{|\xi - \mu|}{\max\{\xi, 1 - \xi\}} \\
\ell_t &= \begin{cases} (1 - \mu)\xi / (1 - \xi) & \text{if } \mu \leq \xi \leq \frac{1}{2} \\ \mu & \text{if } \xi \leq \mu \leq \frac{1}{2} \\ (1 - \xi)\mu / \xi & \text{if } \mu \leq \frac{1}{2} \leq \xi \\ \text{symmetric expressions} & \text{if } \mu \geq \frac{1}{2}. \end{cases}
\end{aligned}$$

3 Instantaneous bounds

Instantaneous bounds of a certain structure that involve only d and ℓ are very useful, since they can be generalized to cumulative bounds. The following

bounds hold:

$$\begin{aligned}\delta &\leq 2d + 2\sqrt{d\ell} \quad \text{proven by Marcus Hutter,} \\ \delta &\leq 2d + 2(\sqrt{2} - 1)\sqrt{d\ell^2} + \sqrt{2d\ell}.\end{aligned}$$

The second bound improves the first one in the constant of the leading order term $\sqrt{d\ell}$.

Unfortunately, the following bound is *not* true:

$$\delta \not\leq O(d + \sqrt{d\ell^2}).$$

The reason is that close to $(\mu, \xi) = (0, 0)$ the relative entropy d behaves like $(\mu - \xi)^2 / \max(\mu, \xi)$ and therefore the r.h.s. tends to zero at the order of $\sqrt{\mu}$ faster than the l.h.s. The “quadratic” bound does hold if μ is bounded away from 0 *or* if in the vicinity of 0, the ξ -decisions coincide with the μ -decisions. The latter condition is satisfied under additional assumptions on the losses.

4 Cumulative bounds

Here we show how to obtain cumulative bounds from the instantaneous bounds of the last section. The key property is that the bound must be *super-additive*. A function $f : [0, \infty)^2 \rightarrow [0, \infty)$ is said to be super-additive if

$$f(x_1 + x_2, y_1 + y_2) \geq f(x_1, y_1) + f(x_2, y_2).$$

The function $(d, \ell) \mapsto \sqrt{d\ell}$ satisfies this condition, as one can easily verify. So does the function $(d, \ell^2) \mapsto \sqrt{d\ell^2}$.

We use an inductive argument to prove the bounds. Assume that

$$\begin{aligned}\Delta_{t=2:T, x_1=0} &\leq 2D_{t=2:T, x_1=0} + 2\sqrt{D_{t=2:T, x_1=0}L_{t=2:T, x_1=0}} \quad \text{and} \\ \Delta_{t=2:T, x_1=1} &\leq 2D_{t=2:T, x_1=1} + 2\sqrt{D_{t=2:T, x_1=1}L_{t=2:T, x_1=1}}\end{aligned}$$

have been already proven, where the subscript $t = 2 : T, x_1 = a$ indicates that the summation is over time 2 to T after the first observed symbol is a . Using $\delta_1 \leq 2d_1 + 2\sqrt{d_1\ell_1}$, we obtain

$$\begin{aligned}\Delta_T &= \delta_1 + (1 - \mu_1)\Delta_{t=2:T, x_1=0} + \mu_1\Delta_{t=2:T, x_1=1} \\ &\leq 2d_1 + 2\sqrt{d_1\ell_1} + 2(1 - \mu_1)D_{t=2:T, x_1=0} + 2\mu_1D_{t=2:T, x_1=1} \\ &\quad + 2(1 - \mu_1)\sqrt{D_{t=2:T, x_1=0}L_{t=2:T, x_1=0}} + 2\mu_1\sqrt{D_{t=2:T, x_1=1}L_{t=2:T, x_1=1}} \\ &\leq 2D_T + 2\sqrt{d_1\ell_1} + 2\sqrt{(1 - \mu_1)D_{t=2:T, x_1=0} + \mu_1D_{t=2:T, x_1=1}} \\ &\quad \cdot \sqrt{(1 - \mu_1)L_{t=2:T, x_1=0} + \mu_1L_{t=2:T, x_1=1}} \\ &\leq 2D_T + 2\sqrt{D_T L_t}.\end{aligned}$$

Here, the first inequality is the induction hypothesis, the second bound is Cauchy-Schwarz’s inequality, and the last estimate is the super-additivity.

In the same way the bounds

$$\Delta_T \leq 2D_T + 2(\sqrt{2} - 1)\sqrt{D_t L_T^2} + \sqrt{2D_T L_T^2}$$

(for any loss function) and $\Delta_T \leq O(D_T + \sqrt{D_t L_T^2})$ for restricted loss function or μ can be proven. The quadratic loss bound also holds if the loss function is fixed for all t , since then the decision boundary is fixed. If the decision boundary is at least at $\varepsilon > 0$ (or $1 - \varepsilon$ if it is larger than $\frac{1}{2}$), then the quadratic loss bound holds with a constant of $\varepsilon^{-\frac{1}{2}}$.

5 The gap between D and K

Although $\delta \not\leq O(d + \sqrt{d\ell^2})$, one can see that $\delta \leq 2(K + \sqrt{K\ell^2})$ does hold. So can we exploit the gap between D and K and prove a “quadratic” bound by using K instead of D ? The following counterexample suggests that the answer is negative.

Choose $\varepsilon > 0$ small and $n > 0$ arbitrary. Choose $K_1 = K = \varepsilon^{3n}$, $\mu_1 = K^{\frac{2}{3}}$, and $\xi_1 = 1 - e^{-K_1}(1 - \mu_1)$. Then $d_1 \approx K^{\frac{4}{3}}/2$, $\ell_1^2 = K^{\frac{4}{3}}$, $\delta_1 \approx K_1$. We have chosen ξ such that the complexity drops to 0 after a 0 was observed, then the distribution is perfectly learned. After a 1 is observed (which occurs with probability μ_1), the new complexity is $K_2 = K - \log \frac{\mu}{\xi} \approx K_1^{\frac{1}{3}} = \varepsilon^{3(n-1)}$. We repeat the construction, observing that $\mu_2 \approx K^{\frac{8}{9}}$, $\mu_1 \ell_2^2 \approx K^{\frac{10}{9}}$, and $\mu_1 \delta_2 \approx K$. After n steps of this construction, we get that $\Delta_n \approx nK$ and $L_n^2 \approx K(K^{\frac{1}{3}} + K^{\frac{1}{9}} + K^{\frac{1}{27}} + \dots + K^{\frac{1}{3n}}) \leq nK$, consequently $\Delta_n \gtrsim \sqrt{n}(K + \sqrt{KL_n^2})$.

References

- [1] M. Hutter. *Universal Artificial Intelligence: Sequential Decisions based on Algorithmic Probability*. Springer, Berlin, 2004.
- [2] J. Poland and M. Hutter. Convergence of discrete MDL for sequential prediction. In *17th Annual Conference on Learning Theory (COLT)*, pages 300–314, 2004.
- [3] J. Poland and M. Hutter. Asymptotics of discrete MDL for online prediction. Technical report, IDSIA, 2004. *IEEE Transactions on Information Theory*, to appear.
- [4] N. Cesa-Bianchi, Y. Mansour, and G. Stoltz. Improved second-order bounds for prediction with expert advice. In *18th Annual Conference on Learning Theory (COLT)*, Lecture Notes in Computer Science, pages 217–232. Springer, 2005.