

# Inference of average Bayes accuracy

Charles Zheng and Yuval Benjamini

October 18, 2016

These are preliminary notes.

## 1 Introduction

Suppose  $X$  and  $Y$  are continuous random variables (or vectors) which have a joint distribution with density  $p(x, y)$ . Let  $p(x) = \int p(x, y)dy$  and  $p(y) = \int p(x, y)dx$  denote the respective marginal distributions, and  $p(y|x) = p(x, y)/p(x)$  denote the conditional distribution.

$\text{ABE}_k$ , or  $k$ -class Average Bayes accuracy is defined as follows. Let  $X_1, \dots, X_K$  be iid from  $p(x)$ , and draw  $Z$  uniformly from  $1, \dots, k$ . Draw  $Y \sim p(y|X_Z)$ . Then, the average Bayes accuracy is defined as

$$\text{ABA}_k[p(x, y)] = \sup_f \Pr[f(X_1, \dots, X_k, Y) = Z]$$

where the supremum is taken over all functions  $f$ . A function  $f$  which achieves the supremum is

$$f_{\text{Bayes}}(x_1, \dots, x_k, y) = \operatorname{argmax}_{z \in \{1, \dots, k\}} p(y|x_z),$$

where an arbitrary rule can be employed to break ties. Such a function  $f_{\text{Bayes}}$  is called a *Bayes classification rule*. It follows that  $\text{ABA}_k$  is given explicitly by

$$\text{ABA}_k = \frac{1}{k} \int \left[ \prod_{i=1}^k p(x_i) dx_i \right] \int dy \max_i p(y|x_i),$$

as stated in the following theorem.

**Theorem 1.1** For a joint distribution  $p(x, y)$ , define

$$ABA_k[p(x, y)] = \sup_f \Pr[f(x_1, \dots, x_k, y) = Z]$$

where  $X_1, \dots, X_K$  are iid from  $p(x)$ ,  $Z$  is uniform from  $1, \dots, k$ , and  $Y \sim p(y|X_Z)$ , and the supremum is taken over all functions  $f : \mathcal{X}^k \times \mathcal{Y} \rightarrow \{1, \dots, k\}$ . Then,

$$ABA_k = \frac{1}{k} \int \left[ \prod_{i=1}^k p(x_i) dx_i \right] \int dy \max_i p(y|x_i).$$

**Proof.** First, we claim that the supremum is attained by choosing

$$f(x_1, \dots, x_k, y) = \operatorname{argmax}_{z \in \{1, \dots, k\}} p(y|x_z).$$

To show this claim, write

$$\sup_f \Pr[f(X_1, \dots, X_k, Y) = Z] = \sup_f \frac{1}{k} \int p_X(x_1) \dots p_X(x_k) p(y|x_{f(x_1, \dots, x_k, y)}) dx_1 \dots dx_k dy$$

We see that maximizing  $\Pr[f(X_1, \dots, X_k, Y) = Z]$  over functions  $f$  additively decomposes into infinitely many subproblems, where in each subproblem we are given  $\{x_1, \dots, x_k, y\} \in \mathcal{X}^k \times \mathcal{Y}$ , and our goal is to choose  $f(x_1, \dots, x_k, y)$  from the set  $\{1, \dots, k\}$  in order to maximize the quantity  $p(y|x_{f(x_1, \dots, x_k, y)})$ . In each subproblem, the maximum is attained by setting  $f(x_1, \dots, x_k, y) = \operatorname{argmax}_z p(y|x_z)$ —and the resulting function  $f$  attains the supremum to the functional optimization problem. This proves the claim.

We therefore have

$$p(y|x_{f(x_1, \dots, x_k, y)}) = \max_{i=1}^k p(y|x_i).$$

Therefore, we can write

$$\begin{aligned} ABA_k[p(x, y)] &= \sup_f \Pr[f(X_1, \dots, X_k, Y) = Z] \\ &= \frac{1}{k} \int p_X(x_1) \dots p_X(x_k) p(y|x_{f(x_1, \dots, x_k, y)}) dx_1 \dots dx_k dy. \\ &= \frac{1}{k} \int p_X(x_1) \dots p_X(x_k) \max_{i=1}^k p(y|x_i) dx_1 \dots dx_k dy. \end{aligned}$$

## 2 Variability of Bayes Accuracy

We have

$$\text{ABA}_k = \mathbf{E}[\text{BA}(X_1, \dots, X_k)]$$

where the expectation is over the independent sampling of  $X_1, \dots, X_k$  from  $p(x)$ .

Therefore,  $\text{BA}_k = \text{BA}(X_1, \dots, X_k)$  is already an unbiased estimator of  $\text{ABA}_k$ . However, to get confidence intervals for  $\text{ABA}_k$ , we also need to know the variability.

We have the following upper bound on the variability.

**Theorem 2.1** *Given joint density  $p(x, y)$ , for  $X_1, \dots, X_k \stackrel{iid}{\sim} p(x)$ , we have*

$$\text{Var}[\text{BA}(X_1, \dots, X_k)] \leq \frac{1}{4k}.$$

**Proof.** According to the Efron-Stein lemma,

$$\text{Var}[\text{BA}(X_1, \dots, X_k)] \leq \sum_{i=1}^k \mathbf{E}[\text{Var}[\text{BA}|X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_k]].$$

which is the same as

$$\text{Var}[\text{BA}(X_1, \dots, X_k)] \leq k \mathbf{E}[\text{Var}[\text{BA}|X_1, \dots, X_{k-1}]].$$

The term  $\text{Var}[\text{BA}|X_1, \dots, X_{k-1}]$  is the variance of  $\text{BA}(X_1, \dots, X_k)$  conditional on fixing the first  $k-1$  curves  $p(y|x_1), \dots, p(y|x_{k-1})$  and allowing the final curve  $p(y|x_k)$  to vary randomly.

Note the following trivial results

$$-p(y|x_k) + \max_{i=1}^k p(y|x_i) \leq \max_{i=1}^{k-1} p(y|x_i) \leq \max_{i=1}^k p(y|x_i).$$

This implies

$$\text{BA}(X_1, \dots, X_k) - \frac{1}{k} \leq \frac{k-1}{k} \text{BA}(X_1, \dots, X_{k-1}) \leq \text{BA}(X_1, \dots, X_k).$$

i.e. conditional on  $(X_1, \dots, X_{k-1})$ ,  $\text{BA}_k$  is supported on an interval of size  $1/k$ . Therefore,

$$\text{Var}[\text{BA}|X_1, \dots, X_{k-1}] \leq \frac{1}{4k^2}$$

since  $\frac{1}{4c^2}$  is the maximal variance for any r.v. with support of length  $c$ .  $\square$

In the next section we report on some empirical results which suggest that the constant in the bound can be greatly improved.

### 3 Empirical results

In this section we search over a class of distributions  $p(x, y)$  in an attempt to find a distribution which achieves close to the maximum variability.

The theoretical upper bound is

$$\text{sd}(\text{BA}_k) \leq \frac{1}{2\sqrt{k}}.$$

We compare the theoretical bound with the largest sd found for any distribution.

| k                     | 2     | 3     | 4     | 5     | 6     | 7     | 8     |
|-----------------------|-------|-------|-------|-------|-------|-------|-------|
| $\frac{1}{2\sqrt{k}}$ | 0.353 | 0.289 | 0.250 | 0.223 | 0.204 | 0.189 | 0.177 |
| Worst-case sd         | 0.25  | 0.194 | 0.167 | 0.150 | 0.136 | 0.126 | 0.118 |

#### 3.1 Methodology

We search over discrete distributions where  $Y \in \{1, \dots, d\}$  and  $X = (X_1, \dots, X_d)$  lies in the space of all  $d$ -permutations. The pmf  $p(x, y)$  is defined as

$$p(x, y) = \frac{1}{d!} v_{x_y}$$

where

$$0 \leq v_1 \leq v_2 \leq \dots \leq v_d$$

and

$$\sum_i v_i = 1.$$

Define the random vector  $M^{(k)}$  by

$$M_i^{(k)} = \max_{j=1}^k X_i^{(j)}$$

where  $X^{(1)}, \dots, X^{(k)}$  are independent random permutations for  $i = 1, \dots, d$ . Then define the random vector  $C^{(k)}$  by

$$C_i^{(k)} = \sum_{j=1}^d I\{M_j^{(k)} = i\}.$$

For this class of distributions, the variance is given by

$$\text{Var}(\text{BA}_k) = v' \Gamma^{(d,k)} v$$

where  $\Gamma^{(d,k)}$  is the  $d \times d$  covariance matrix with entries

$$\Gamma_{ij}^{(d,k)} = \text{Cov}(C_i^{(k)}, C_j^{(k)}).$$

Therefore, having computed  $\Gamma^{(d,k)}$ , one can find distributions which solve the constrained optimization problem:

$$\text{maximize}_{v_1, \dots, v_d} v' \Gamma^{(d,k)} v \text{ subject to } 0 \leq v_1 \leq v_2 \leq \dots \leq v_d \text{ and } \sum_{i=1}^d v_i = 1 \quad (1)$$

### 3.2 Empirical worst-case distributions

An empirical result is as follows: for all  $(d, k)$  examined, maximizers of (1) take the form

$$v = (0, \dots, 0, 1). \quad (2)$$

While the problem (1) is non-convex, to show that (2) is the global maximizer, it suffices to examine the matrix  $\Gamma^{(d,k)}$ .

The following properties have been seen to hold for small  $(d, k)$ , but there may be counterexamples. (In the following, we omit the superscript in  $\Gamma^{(d,k)}$ .)

P1. For all  $j$ ,  $\max_{i>j} \Gamma_{ij} = \Gamma_{dj}$ .

P2. For all  $i$ ,

$$\alpha^2 \Gamma_{ii} + 2\alpha(1 - \alpha) \Gamma_{di} + (1 - \alpha)^2 \Gamma_{dd}$$

where  $\alpha \in [0, \frac{1}{d-i}]$  is maximized by  $\alpha = 0$ . Equivalently,

$$\max \left\{ \frac{\Gamma_{ii} + 2(i-1)\Gamma_{di}}{2i-1}, \Gamma_{di} \right\} \leq \Gamma_{dd}.$$

If P1 and P2 hold, then it follows that (2) maximizes (1), as stated in the following theorem.

**Theorem 3.1** *Consider the optimization problem*

$$\text{maximize}_{v_1, \dots, v_d} v' \Gamma^{(d,k)} v \text{ subject to } 0 \leq v_1 \leq v_2 \leq \dots \leq v_d \text{ and } \sum_{i=1}^d v_i = 1$$

*If properties P1 and P2 hold for  $\Gamma$ , then the solution is*

$$v = (0, \dots, 0, 1).$$

**Proof.**

Let  $S_i$  be the set of all feasible vectors  $v$  such that  $v_1 = \dots = v_{d-i} = 0$  for  $i = 1, \dots, d$ . Let  $v^{(i)}$  be the element of  $S_i$  which maximizes  $v'\Gamma v$  over  $S_i$ .

It is clear that  $v^{(1)} = (0, \dots, 0, 1)$ . Now we proceed to show by induction that if  $v^{(i)} = (0, \dots, 0, 1)$ , then also  $v^{(i+1)} = (0, \dots, 0, 1)$ .

Suppose  $v^{(i)} = (0, \dots, 0, 1)$ . Now, any element  $v \in S^{(i+1)}$  can be written

$$v = (0, \dots, 0, \alpha, (1 - \alpha)w)$$

where  $w \in S_i$ , for  $\alpha \in [0, 1/i]$ . Therefore, we have

$$\begin{aligned} v'\Gamma v &= \alpha^2 \Gamma_{ii} + 2\alpha(1 - \alpha)(\Gamma_{i+1,i}, \dots, \Gamma_{di})'w + (1 - \alpha)^2 w'\Gamma_{(i+1)\dots d, (i+1)\dots d} \\ &\leq \alpha^2 \Gamma_{ii} + 2\alpha(1 - \alpha)(\Gamma_{i+1,i}, \dots, \Gamma_{di})'w + (1 - \alpha)^2 \Gamma_{dd} \\ &\leq \alpha^2 \Gamma_{ii} + 2\alpha(1 - \alpha)\Gamma_{di} + (1 - \alpha)^2 \Gamma_{dd}. \\ &\leq \Gamma_{dd} \text{ by property P2.} \end{aligned}$$

But

$$v'\Gamma v \leq \Gamma_{dd} = (0, \dots, 0, 1)'\Gamma(0, \dots, 0, 1)$$

implies that

$$v^{(i+1)} = (0, \dots, 0, 1).$$

□.

*Consequences.*

Suppose that (2) does maximize all the optimization problems (1). The resulting worst-case variance for given  $(d, k)$  is given by

$$\text{Var}^{(d,k)} \approx \frac{1}{d} e^{-d/k} (1 - e^{-d/k})$$

and therefore we could obtain a variance bound of the form

$$\sup_{p(x,y)} \text{Var}(\text{BA}_k) \approx \max_{d \in 1, \dots, d} \frac{1}{d} e^{-d/k} (1 - e^{-d/k}).$$