

# Upper bounds for average Bayes accuracy in terms of mutual information

Charles Zheng and Yuval Benjamini

September 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.

Mutual information is defined

$$I[p(x, y)] = \int p(x, y) \log \frac{p(x, y)}{p(x)p(y)} dx dy.$$

$ABE_k$ , 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

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

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

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

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

by

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

## 2 Problem formulation

Let  $\mathcal{P}$  denote the collection of all joint densities  $p(x, y)$  on finite-dimensional Euclidean space. For  $\iota \in [0, \infty)$  define  $C_k(\iota)$  to be the largest  $k$ -class average Bayes error attained by any distribution  $p(x, y)$  with mutual information not exceeding  $\iota$ :

$$C_k(\iota) = \sup_{p \in \mathcal{P}: I[p(x, y)] \leq \iota} \text{ABA}_k[p(x, y)].$$

A priori,  $C_k(\iota)$  exists since  $\text{ABA}_k$  is bounded between 0 and 1. Furthermore,  $C_k$  is nondecreasing since the domain of the supremum is monotonically increasing with  $\iota$ .

It follows that for any density  $p(x, y)$ , we have

$$\text{ABA}_k[p(x, y)] \leq C_k(I[p(x, y)]).$$

Hence  $C_k$  provides an upper bound for average Bayes error in terms of mutual information.

Conversely we have

$$I[p(x, y)] \geq C_k^{-1}(\text{ABA}_k[p(x, y)])$$

so that  $C_k^{-1}$  provides a lower bound for mutual information in terms of average Bayes error.

On the other hand, there is no nontrivial *lower* bound for average Bayes error in terms of mutual information, nor upper bound for mutual information in terms of average Bayes error, since

$$\inf_{p \in \mathcal{P}: I[p(x, y)] \leq \iota} \text{ABA}_k[p(x, y)] = \frac{1}{k}.$$

regardless of  $\iota$ .

The goal of this work is to attempt to compute or approximate the functions  $C_k$  and  $C_k^{-1}$ .

### 3 Special case

We work out the special case where  $p(x, y)$  lies on the unit square, and  $p(x)$  and  $p(y)$  are both the uniform distribution. Let  $\mathcal{P}^{unif}$  denote the set of such distributions, and

$$C_k^{unif}(\iota) = \sup_{p(x,y) \in \mathcal{P}^{unif}: I[p] \leq \iota} \text{ABA}_k[p].$$

We prove the following result:

**Theorem.** *For any  $\iota > 0$ , there exists  $c_\iota \geq 0$  such that defining*

$$Q_c(t) = \frac{\exp[ct^{k-1}]}{\int_0^1 \exp[ct^{k-1}]},$$

*we have*

$$\int_0^1 Q_{c_\iota}(t) \log Q_{c_\iota}(t) dt = \iota.$$

*Then,*

$$C_k^{unif} = \int_0^1 Q_c(t) t^{k-1} dt.$$

The proof depends on the following two lemmas.

**Lemma 1.** *For any measure  $G$  on  $[0, \infty]$ , let  $G^k$  denote the measure defined by*

$$G^k(A) = G(A)^k,$$

*and define*

$$E[G] = \int x dG(x).$$

$$H[G] = - \int x \log x dG(x)$$

*and*

$$\psi_k[G] = \int x d(G^k)(x).$$

*Then, defining  $Q_c$  and  $c_\iota$  as in Theorem 1, we have*

$$\sup_{G: E[G]=1, -H[G] \leq \iota} \psi_k[G] = \int_0^1 Q_{c_\iota}(t) t^{k-1} dt.$$

Furthermore, the supremum is attained by a measure  $G$  that has cdf equal to  $Q_c^{-1}$ , and thus has a density  $g$  with respect to Lebesgue measure.

**Lemma 2.** *The map*

$$\iota \rightarrow \int_0^1 Q_{c_\iota}(t) t^{k-1} dt$$

*is concave in  $\iota > 0$ .*

**Proof of Lemma 1.** (This will appear in the appendix of the paper.)

Consider the quantile function  $Q(t) = \inf_{x \in [0,1]} : G([0, x]) \geq t$ .  $Q(t)$  must be a monotonically increasing function from  $[0, 1]$  to  $[0, \infty)$ .

We have

$$E[G] = \int_0^1 Q(t) dt$$

$$\psi_k[G] = \int_0^1 Q(t) x^{k-1} dt.$$

and

$$-H[G] = \int_0^1 Q(t) \log Q(t) dt.$$

For any given  $\iota$ , let  $P_\iota$  denote the class of probability distributions  $G$  on  $[0, \infty]$  such that  $E[G] = 1$  and  $-H[G] \leq \iota$ . From Markov's inequality, for any  $G \in P_\iota$  we have

$$G([x, \infty]) \leq x^{-1}$$

for any  $x \geq 0$ , hence  $P_\iota$  is tight. From tightness, we conclude that  $P_\iota$  is closed under limits with respect to weak convergence. Hence, since  $\psi_k$  is a continuous function, there exists a distribution  $G^* \in P_\iota$  which attains the supremum

$$\sup_{G \in P_\iota} \psi_k[G].$$

Let  $\mathcal{Q}_\iota$  denote the collection of quantile functions of distributions in  $P_\iota$ . Then,  $\mathcal{Q}_\iota$  consists of monotonic functions  $Q : [0, 1] \rightarrow [0, \infty]$  which satisfy

$$\int_0^1 Q(t) dt = 1,$$

and

$$\int_0^1 Q(t) \log Q(t) dt \leq \iota.$$

Let  $\mathcal{Q}$  denote the collection of *all* quantile functions from measures on  $[0, \infty]$ . And letting  $Q^*$  be the quantile function for  $G^*$ , we have that  $Q^*$  attains the supremum

$$\sup_{Q \in \mathcal{Q}} \int_0^1 Q(t) t^{k-1} dt.$$

Therefore, there exist Lagrange multipliers  $\lambda \geq 0$  and  $\nu \leq 0$  such that defining

$$\mathcal{L}[Q] = \int_0^1 Q(t) (1 + \lambda \log Q(t) + \nu t^{k-1}) dt,$$

$Q^*$  attains the infimum of  $\mathcal{L}[Q]$  over *all* quantile functions,

$$\mathcal{L}[Q^*] = \inf_{Q \in \mathcal{Q}} \mathcal{L}[Q].$$

We now claim that for such  $\lambda$  and  $\nu$ , we have

$$1 + \lambda + \lambda \log Q(t) + \nu t^{k-1} = 0.$$

Consider a perturbation function  $\xi : [0, 1] \rightarrow \mathbb{R}$ . We have

$$\mathcal{L}[Q + \xi] \approx \mathcal{L}[Q] + \int_0^1 \xi(t) (1 + \lambda + \lambda \log Q(t) + \nu t^{k-1}) dt$$

for small  $\xi$ . Define

$$\nabla Q^*(t) = (1 + \lambda + \lambda \log Q^*(t) + \nu t^{k-1}).$$

The function  $\nabla Q^*(t)$  is a *functional derivative* of the Lagrangian. Note that if we were able to show that  $\nabla Q^*(t) = 0$ , as we might naively expect, this immediately yields

$$Q^*(t) = \exp[-1 - \lambda - \log \lambda - \nu t^{k-1}]. \quad (1)$$

However, the reason why we cannot simply assume  $\nabla Q^*(t) = 0$  is because the optimization occurs on a constrained space. We will ultimately show that this is the case (up to sets of negligible measure), but some delicacy is needed.

Let  $B$  denote the set of points  $t$  such that  $\nabla Q^*(t) \neq 0$ . We would like to show that  $B$  is of measure zero, which would yield (1) up to negligible sets. What needs to be done is to show that  $\nabla Q^*(t) = 0$  on a set of non-zero measure results in a contradiction. One can verify that for any  $t$  such that  $\nabla Q^*(t) \neq 0$ , one of the following four cases must apply.

- *Case 1:*  $\nabla Q^*(t) \neq 0$  on an isolated point; i.e. for all neighborhoods  $N_t$  of  $t$ ,  $B \cap N_t$  is a set of measure zero.
- *Case 2:*  $\nabla Q^*(t) \neq 0$  and there does not exist an interval such that  $[a, b] \ni t$  is  $\nabla Q^*(t)$  strictly positive or negative on the interval, but there does exist a neighborhood  $N_t$  of  $t$  such that  $B \cap N_t$  has nonzero measure.
- *Case 3:*  $\nabla Q^*(t) \neq 0$  and there exists an interval  $[a, b] \ni t$  with  $Q^*(a) < Q^*(b)$  such that  $\nabla Q^*(t)$  is either strictly positive or negative on  $[a, b]$ .
- *Case 4:*  $\nabla Q^*(t) \neq 0$  and there exists an interval  $[a, b] \ni t$  with  $Q^*(a) = Q^*(b)$  such that  $\nabla Q^*(t)$  is either strictly positive or negative on  $[a, b]$ .

The set of all points  $t$  where case 1 applies is necessarily of zero measure. Therefore if  $B$  is non-negligible, there must exist  $t$  falling in one of the three other cases must occur. But we will show that each of cases 2 through 4 result in a contradiction.

*Case 3.* Define

$$\xi^+(t) = I\{t \in [a, b]\}(Q(b) - Q(t))$$

and

$$\xi^-(t) = I\{t \in [a, b]\}(Q(a) - Q(t)).$$

, Observe that  $Q + \epsilon \xi^+ \in \mathcal{Q}$  and  $Q + \epsilon \xi^- \in \mathcal{Q}$  for any  $\epsilon \in [0, 1]$ . Now, if  $\nabla Q^*(t)$  is strictly positive on  $[a, b]$ , then for some  $\epsilon > 0$  we would have  $\mathcal{L}[Q^* + \epsilon \xi^-] < \mathcal{L}[Q^*]$ , a contradiction. A similar argument with  $\xi^+$  shows that  $\nabla Q^*(t)$  cannot be strictly negative on  $[a, b]$  either.

*Case 4.* Without loss of generality, let  $a$  and  $b$  be the endpoints of the largest interval containing  $t$  such that  $Q^*(t)$  is constant on  $(a, b)$ . Now, since  $\nabla Q^*(t) \neq 0$  on a set of nonzero measure within  $[a, b]$ , it must be the case that there exists some  $u \in [a, b]$  such that

$$\int_a^u \nabla Q^*(t) dt \neq 0.$$

If  $\int_a^u \nabla Q^*(t) dt > 0$ , then define  $\xi^+(t) = -I\{t \in (a, u)\}$  (to be contd.)

*Remark.* More specifically, the supremum is attained by a distribution with density  $p_\iota(x, y)$  where

$$p_\iota(x, y) = \begin{cases} g_\iota(y - x) & \text{for } x \geq y \\ g_\iota(1 + y - x) & \text{for } x < y \end{cases}$$

where

$$g_\iota(x) = \frac{d}{dx}G_\iota(x)$$

and  $G_\iota$  is the inverse of  $Q_c$ .

In this case, letting  $X_1, \dots, X_k \sim \text{Unif}[0, 1]$ , and  $Y \sim \text{Unif}[0, 1]$  define  $Z_i(y) = p(y|X_i)$ . We have  $\mathbf{E}(Z(y)) = 1$  and,

$$\mathbf{I}[p(x, y)] = \mathbf{E}(Z(Y) \log Z(Y))$$

while

$$\text{ABA}_k[p(x, y)] = k^{-1} \mathbf{E}(\max_i Z_i(Y)).$$

Letting  $g_y$  be the density of  $Z(y)$ , we have

$$\mathbf{I}[p(x, y)] = \mathbf{E}(-H[g_Y])$$

and

$$\text{ABA}_k[p(x, y)] = \mathbf{E}(\psi_k[g_Y])$$

where

$$H[g] = - \int g(x) x \log x dx$$

and

$$\psi_k[g] = \int xg(x)G(x)^{k-1}dx$$

for  $G(x) = \int_0^x g(t)dt$ . Additionally  $g_y$  satisfies the constraint  $\int xg(x)dx = 1$  since  $\mathbf{E}[Z(y)] = 1$ .

Define the set  $D = \{(\alpha, \beta)\}$  as the set of possible values of  $(-H[g], \psi_k[g])$  taken over all distributions  $g$  supported on  $[0, \infty)$  with  $\int xg(x)dx = 1$ . Next, let  $\mathcal{C}(D)$  denote the convex hull of  $D$ . It follows that  $(\mathbf{I}[p], \text{ABA}_k[p]) \in \mathcal{C}(D)$  since the pair is obtained via a convex average of points  $(-H[g_y], \psi_k[g_y])$ .

Define the upper envelope of  $D$  as the curve

$$d_k(\alpha) = \sup\{\beta : (\alpha, \beta) \in D\}.$$

We make the claim (to be shown in the following section) that  $d_k(\alpha)$  is convex in  $\alpha$ . As a result, the upper envelope of  $D$  is also the upper envelope of  $\mathcal{C}(D)$ . This in turn implies that  $C_k^{unif}(\iota) = d_k(\iota)$ . We establish these results, along with a open-form expression for  $C_k^{unif}$ , in the following section.

### 3.1 Variational methods

Consider the quantile function  $Q(t) = G^{-1}(t)$ .  $Q(t)$  must be a continuous function from  $[0, 1]$  to  $[0, \infty)$ . We can rewrite the moment constraint  $\mathbf{E}[g] = 1$  as

$$\int_0^1 Q(t) dt = 1.$$

Meanwhile,  $\beta = \psi_k[g]$  takes the form

$$\beta = \int_0^1 Q(t) x^{k-1} dt.$$

and  $\alpha = -H[g]$  takes the form

$$\alpha = \int_0^1 Q(t) \log Q(t) dt.$$

To find the upper envelope, it will be useful to write the Langrangian

$$\begin{aligned} \mathcal{L}[g] &= \lambda \int_0^1 Q(t) dt + \mu \int_0^1 Q(t) x^{k-1} dt + \lambda \int_0^1 Q(t) \log Q(t) dt \\ &= \int_0^1 Q(t) (\lambda + \mu x^{k-1} + \nu \log Q(t)) dt. \end{aligned}$$

In order for a quantile function  $Q(t)$  to be on the upper envelope, it must be a local maximum of  $-H$  with respect to small perturbations. Therefore, consider the functional derivative

$$D[\xi] = \lim_{\epsilon \rightarrow 0} \frac{\mathcal{L}[g + \epsilon \xi] - \mathcal{L}[g]}{\epsilon}.$$

We have

$$D[\xi] = \int_0^1 \xi(t) (\lambda + \nu + \mu x^{k-1} + \nu \log Q(t)) dt.$$

Now consider the following three cases:

- $Q(t)$  is strictly monotonic, i.e.  $Q'(t) > 0$ .
- $Q(t)$  is differentiable but not strongly monotonic:
- $Q(t)$  is not strongly monotonic: there exist intervals  $A_i = [a_i, b_i)$  such that  $Q(t)$  is constant on  $A_i$ , and isolated points  $t_i$  where  $Q'(t_i) = 0$ .



*Strictly monotonic case.* Because  $Q$  is defined on a closed interval, strict monotonicity further implies the property of *strong monotonicity* where  $\inf_{[0, 1]} Q'(t) > 0$ . Therefore, for any differentiable perturbation  $\xi(t)$  with  $\sup |\xi'(t)| < \infty$ , and further imposing that  $\xi(0) \geq 0$  in the case that  $Q(0) = 0$ , there exists some  $\epsilon > 0$  such that  $(Q + \epsilon\xi)(t)$  is still a valid quantile function. Therefore, in order for  $Q(t)$  to be a local maximum, we must have

$$0 = \lambda + \nu + \mu x^{k-1} + \nu \log Q(t)$$

for  $t \in [0, 1]$ . This implies that

$$Q(t) = c_0 e^{-c_1 x^{k-1}}$$

for some  $c_0, c_1 \geq 0$ .

*Other cases.* (TODO) We have to show that these cannot be local maxima.

## 4 General case

We claim that the constants  $C_k^{unif}(\iota)$  obtained for the special case also apply for the general case, i.e.

$$C_k(\iota) = C_k^{unif}(\iota).$$

We make use of the following Lemma:

**Lemma.** *Suppose  $X, Y, W, Z$  are continuous random variables, and that  $W \perp Y|Z$ ,  $Z \perp X|Y$ , and  $W \perp Z|(X, Y)$ . Then,*

$$I[p(x, y)] = I[p((x, w), (y, z))]$$

and

$$ABA_k[p(x, y)] = ABA_k[p((x, w), (y, z))].$$

**Proof.** Due to conditional independence relationships, we have

$$p((x, w), (y, z)) = p(x, y)p(w|x)p(z|y).$$

It follows that

$$\begin{aligned}
\mathbb{I}[p((x, w), (y, z))] &= \int dx dw dy dz \, p(x, y) p(w|x) p(z|w) \log \frac{p((x, w), (y, z))}{p(x, w) p(y, z)} \\
&= \int dx dw dy dz \, p(x, y) p(w|x) p(z|w) \log \frac{p(x, y) p(w|x) p(z|y)}{p(x) p(y) p(w|x) p(z|y)} \\
&= \int dx dw dy dz \, p(x, y) p(w|x) p(z|w) \log \frac{p(x, y)}{p(x) p(y)} \\
&= \int dx dy \, p(x, y) \log \frac{p(x, y)}{p(x) p(y)} = \mathbb{I}[p(x, y)].
\end{aligned}$$

Also,

$$\begin{aligned}
\text{ABA}_k[p((x, w), (y, z))] &= \int \left[ \prod_{i=1}^k p(x_i, w_i) dx_i dw_i \right] \int dy dz \, \max_i p(y, z|x_i, w_i). \\
&= \int \left[ \prod_{i=1}^k p(x_i, w_i) dx_i dw_i \right] \int dy \, \max_i p(y|x_i) \int dz \, p(z|y). \\
&= \int \left[ \prod_{i=1}^k p(x_i) dx_i \right] \left[ \prod_{i=1}^k \int dw_i p(w_i|x_i) \right] \int dy \, \max_i p(y|x_i) \\
&= \text{ABA}_k[p(x, y)].
\end{aligned}$$

□

Next, we use the fact that for any  $p(x, y)$  and  $\epsilon > 0$ , there exists a discrete distribution  $p_\epsilon(\tilde{x}, \tilde{y})$  such that

$$|\mathbb{I}[p(x, y)] - \mathbb{I}[p_\epsilon(\tilde{x}, \tilde{y})]| < \epsilon,$$

where for discrete distributions, one defines

$$\mathbb{I}[p(x, y)] = \sum_x \sum_y p(x, y) \log \frac{p(x, y)}{p(x) p(y)}.$$

We require the additional condition that the marginals of the discrete distribution are close to uniform: that is, for some  $\delta > 0$ , we have

$$\sup_{x, x': p_\epsilon(x) > 0 \text{ and } p_\epsilon(x') > 0} \frac{p_\epsilon(x)}{p_\epsilon(x')} \leq 1 + \delta.$$

and likewise

$$\sup_{y, y': p_\epsilon(y) > 0 \text{ and } p_\epsilon(y') > 0} \frac{p_\epsilon(y)}{p_\epsilon(y')} \leq 1 + \delta.$$

To construct the discretization with the required properties, choose a regular rectangular grid  $\Lambda$  over the domain of  $p(x, y)$  sufficiently fine so that partitioning  $X, Y$  into grid cells, we have

$$|I[p(x, y)] - I[\tilde{p}(\tilde{x}, \tilde{y})]| < \epsilon.$$

[NOTE: to be written more clearly] Next, define