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

Charles Zheng and Yuval Benjamini

September 7, 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

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

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

$$\mathrm{ABA}_k[p(x, y)] = \sup_f \Pr[f(x_1, ..., x_k, y) = Z]$$

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

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

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

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

## 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 $\iota$:

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

A priori, $C_k(\iota)$ exists since $\mathrm{ABA}_k$ 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

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

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

Conversely we have

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

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

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

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

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

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

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

while

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

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

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

and

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

where

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

and

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

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 $(\mathrm{I}[p], \mathrm{ABA}_k[p]) \in \mathcal{C}(D)$ since the pair is obtained via a convex average of points $(-H[g_y], \psi_k[g])$.

We trivially obtain the following theorem.

**Theorem.** *Let $d_k(\iota) = \sup\{\beta : (\iota,\beta) \in \mathcal{C}(D)\}$. Then*

$$C_k(\iota) \leq d_k(\iota).$$

$\square$

Hence, we can obtain an upper bound on $C_k$ via properties of the set $D$. It suffices to determine the upper envelope, which can be determined by solving the continuous optimization problems

$$\text{maximize} \int z_0^\infty g(z) G^{k-1}(z) dz$$

subject to the constraints

- $g : [0, \infty) \to [0, \infty)$.
- $\int_0^\infty g(z) dz = 1$.
- $\int_0^\infty z g(z) dz = 1$.
- $\int_0^\infty g(z) \log g(z) \leq \alpha$.

We can let $\mathcal{G}_\alpha$ denote the set of densities $g$ satisfying the constraints; it is evident that $\mathcal{G}_\alpha$ is convex. Meanwhile, the objective function, which can also be written

$$\int_0^\infty z g(z) \left( \int_0^z g(t) dt \right)^{k-1} dz$$

# 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

$$\begin{aligned}
\mathrm{I}[p((x,w),(y,z))] &= \int dxdwdydz\ p(x,y)p(w|x)p(z|w)\log\frac{p((x,w),(y,z))}{p(x,w)p(y,z)} \\
&= \int dxdwdydz\ p(x,y)p(w|x)p(z|w)\log\frac{p(x,y)p(w|x)p(z|y)}{p(x)p(y)p(w|x)p(z|y)} \\
&= \int dxdwdydz\ p(x,y)p(w|x)p(z|w)\log\frac{p(x,y)}{p(x)p(y)} \\
&= \int dxdy\ p(x,y)\log\frac{p(x,y)}{p(x)p(y)} = \mathrm{I}[p(x,y)].
\end{aligned}$$

Also,

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

□

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

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

where for discrete distributions, one defines

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

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

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

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