Upper bounds for average Bayes accuracy in terms of mutual information

Charles Zheng and Yuval Benjamini

September 23, 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

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

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 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*

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

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

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

To show this claim, write

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

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

We therefore have

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

Therefore, we can write

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

2 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 ι:

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

A priori, $C_k(\iota)$ exists since ABA_k is bounded between 0 and 1. Furthermore, C_k is nondecreasing since the domain of the supremum is monotonically increasing with ι.

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

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

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

Conversely we have

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

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

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

regardless of ι.

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

2.1 Notation

$|\cdot|$ denotes set cardinality.

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

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

We prove the following result:

Theorem 3.1 *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,

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

The proof depends on the following lemmas.

Lemma 3.2 *Let $f(t)$ be an increasing function from $[a, b] \to \mathbb{R}$, where $a < b$, and let $g(t)$ be a bounded continuous function from $[a, b] \to \mathbb{R}$. Define the set*

$$A = \{t : f(t) \neq g(t)\}.$$

Then, we can write A as a countable union of intervals

$$A = \bigcup_{i=1}^{\infty} A_i$$

where A_i are mutually disjoint intervals, with $\inf A_i < \sup A_i$, and for each i, either $f(t) > g(t)$ for all $t \in A_i$ or $f(t) < g(t)$ for all $t \in A_i$.

Lemma 3.3 *Let $f(t)$ be a measurable function from $[a,b] \to \mathbb{R}$, where $a < b$. Then there exists sets $\mathcal{B}_0$ and $\mathcal{B}_1$, satisfying the following properties:*

- *$\mathcal{B} = \mathcal{B}_0 \cup \mathcal{B}_1$ is countable partition of $[a,b]$,*
- *$f(t)$ is constant on all $B \in \mathcal{B}_0$, but not constant on any proper super-interval $B' \supset B$, and*
- *$B \in \mathcal{B}_1$ contains no positive-length subinterval where $f(t)$ is constant.*

Lemma 3.4 *For any measure G on $[0,\infty]$, let G^k denote the measure defined by*

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

and define

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

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

and

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

Then, defining Q_c and c_ι as in Theorem 1, we have

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

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 Lesbegue measure.

Lemma 3.5 *The map*

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

is concave in $\iota > 0$.

Proof of Lemma 3.2. (This will appear in the appendix of the paper.)

The function $h(t) = f(t) - g(t)$ is measurable, since all increasing functions are measurable. Define $A^+ = \{t : f(t) > g(t)\}$ and $A^- = \{t : f(t) < g(t)\}$. Since A^+ and A^- are measurable subsets of $\mathbb{R}$, they both admit countable partitions consisting of open, closed, or half-open intervals. Let $\mathcal{H}^+$ be the

collection of all partitions of A^+ consisting of such intervals. There exists a least refined partition $\mathcal{A}^+$ within $\mathcal{H}^+$. Define $\mathcal{A}^-$ analogously, and let

$$\mathcal{A} = \mathcal{A}^+ \cup \mathcal{A}^-$$

and enumerate the elements

$$\mathcal{A} = \{A_i\}_{i=1}^{\infty}.$$

We claim that the partitions $\mathcal{A}^+$ and $\mathcal{A}^-$ have the the property that for all $t \in A^{\pm}$, ther interval $I \in \mathcal{A}^{\pm}$ containing t has endpoints $l \leq u$ defined by

$$l = \inf_{x \in [a,b]} \{x : \text{Sign}(h([x,t])) = \{\text{Sign}(h(t))\}\}$$

and

$$u = \sup_{x \in a[,b]} \{x : \text{Sign}(h([t,x])) = \{\text{Sign}(h(t))\}\}.$$

We prove the claim for the partition $\mathcal{A}^+$. Take $t \in A^+$ and define l and u as above. It is clear that $(l,u) \in A^+$, and furthermore, there is no $l' < l$ and $u' > u$ such that $(l',x) \in A^+$ or $(x,u') \in A^+$ for any $x \in I$. Let $\mathcal{H}$ be any other partition of A^+. Some disjoint union of intervals $H_i \in \mathcal{H}$ necessarily covers I for $i = 1, ...,$ and we can further require that none of the H_i are disjoint with I. Since each H_i has nonempty intersection with I, and I is an interval, this implies that $\cup_i H_i$ is also an interval. Let $l'' \leq u''$ be the endpoints of $\cup_i H_i$ Since $I \subseteq \cup_i H_i$, we have $l'' \leq l \leq u \leq u''$. However, since also $I \in A^+$, we must have $l \leq l'' \leq u'' \leq u$. This implies that $l'' = l$ and $u'' = u$. Since $\cup_i H_i = I$, and this holds for any $I \in \mathcal{A}^+$, we conclude that $\mathcal{H}$ is a refinement of $\mathcal{A}^+$. The proof of the claim for $\mathcal{A}^-$ is similar.

It remains to show that there are not isolated points in $\mathcal{A}$, i.e. that for all $I \in \mathcal{A}$ with endpoints $l \leq u$, we have $l < u$. Take $I \in \mathcal{A}$ with endpoints $l \leq u$ and let $t = \frac{l+u}{2}$. By definition, we have $h(t) \neq 0$. Consider the two cases $h(t) > 0$ and $h(t) < 0$.

If $h(t) > 0$, then $t' = g^{-1}(h(t)) > t$, and for all $x \in [t,t']$ we have $h(x) > 0$. Therefore, it follows from definition that $[t,t'] \in I$, and since $l \leq t < t' \leq u$, this implies that $l < u$. The case $h(t) < 0$ is handled similarly. $\square$

Proof of Lemma 3.3. (This will appear in the appendix of the paper.) To construct the interval, define

$$l(t) = \inf\{x \in [0,1] : f([x,t]) = \{f(t)\}\}$$

$$u(t) = \sup\{x \in [0,1] : f([t,x]) = \{f(t)\}\},$$

Let B_0 be the set of all t such that $l(t) < u(t)$, and let B_1 be the set of all t such that $l(t) = t = u(t)$. For all $t \in B_0$, define

$$I(t) = (l(t), u(t)) \cup \{x \in \{l(t), u(t)\} : f(x) = f(t)\}.$$

Then we claim

$$\mathcal{B}_0 = \{I(t) : t \in B_0\}$$

is a countable partition of B_0. The claim follows since the members of $\mathcal{B}_0$ are disjoint intervals of nonzero length, and B_0 has finite length. It follows from definition that for any $B \in B_0$, that f is not constant on any proper superinterval $B' \supset B$.

Meanwhile, let $\mathcal{B}_1$ be a countable partition of B_1 into intervals.

Next, we show that for all $I \in \mathcal{B}_1$, I does not contain a subinterval I' of nonzero length such that f is constant on I'. Suppose to the contrary, we could find such an interval I and subinterval I'. Then for any $t \in I'$, we have $t \in B_0$. However, this implies that $t \notin B_1$, a contradiction.

Since $t \in [a,b]$ belongs to either B_0 or B_1, letting $\mathcal{B} = \mathcal{B}_0 \cup \mathcal{B}_1$ yields the desired partition of $[a,b]$. $\square$.

Proof of Lemma 3.4. (This will appear in the appendix of the paper.)

Consider the quantile function $Q(t) = \inf_{x \in [0,1]} : G((-\infty, x]) \geq t$. $Q(t)$ must be a monotonically increasing function from $[0,1]$ to $[0,\infty)$. Let $\mathcal{Q}$ denote the collection of all such quantile functions.

We have

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

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

and

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

For any given ι, let P_ι denote the class of probability distributions G on $[0,\infty]$ such that $E[G] = 1$ and $I[G] \leq \iota$. 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_ι is tight. From tightness, we conclude that P_ι is closed under limits with respect to weak convergence. Hence, since ψ_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_ι. Then, $\mathcal{Q}_\iota$ consists of monotonic functions $Q : [0,1] \to [0,\infty]$ which satisfy

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

and

$$I[Q] = \int_0^1 Q(t) \log Q(t) dt \leq \iota.$$

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

$$\sup_{Q \in \mathcal{Q}_\iota} \phi_k[Q] = \sup_{Q \in \mathcal{Q}_\iota} \int_0^1 Q(t) t^{k-1} dt.$$

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

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

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 λ and ν, we have

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

Consider a perturbation function $\xi : [0,1] \to \mathbb{R}$. 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 ξ. 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

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

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 neglible measure), but some delicacy is needed.

The rest of the proof proceeds as follows. We will use Lemmas 3.2 and 3.3 to define a decomposition $A = D_0 \cup D_1 \cup D_2$, where D_2 is of measure zero. First, we show that for all $t \in D_0$, we have $\nabla Q^*(t) = 0$. Second, we show that for all $t \in D_1$, we have $\nabla Q^*(t) = 0$. Finally, since D_2 is a set of zero measure, this allows us to conclude that the $Q^*(t) = 0$ on all but a set of zero measure. Since sets of zero measure don't affect the integral, we conclude there exists a global optimal solution with $\nabla Q^*(t) = 0$.

We will now apply the Lemmas to obtain the necessary ingredients for constructing the sets D_i. Since $\nabla Q^*(t)$ is a difference between an increasing function and a continuous stricly increasing function, we can apply Lemma 3.2 to conclude that there exists a countable partition $\mathcal{A}$ of the set $A : \{t \in [0,1] : \nabla Q^*(t) \neq 0\}$ into intervals such that for all $J \in \mathcal{A}$, $|\text{Sign}(\nabla Q^*(J))| = 1$ and $\inf J < \sup I$. Applying Lemma 3.3 we get a countable partition $\mathcal{B} = \mathcal{B}_0 \cup \mathcal{B}_1$ of $[0,1]$ so that each element $J \in \mathcal{B}_0$ is an interval such that $\nabla Q^*(t)$ is constant on J, and furthermore is not properly contained in any interval with the same property, and each element $J \in \mathcal{B}_1$ is an interval, such that J contains no positive-length subinterval where $\nabla Q^*(t)$ is constant. Also define B_i as the union of the sets in $\mathcal{B}_i$ for $i = 0, 1$.

Note that B_0 is necessarily a subset of A. That is because if $\nabla Q^*(t) = 0$ on any interval J, then that $Q^*(t)$ is necessarily not constant on the interval.

We will construct a new countable partition of A, called $\mathcal{D}$. The partition $\mathcal{D}$ is constructed by taking the union of three families of intervals,

$$\mathcal{D} = \mathcal{D}_0 \cup \mathcal{D}_1 \cup \mathcal{D}_2.$$

Define D_i to be the union of intervals in $\mathcal{D}_i$ for $i = 0, 1, 2$.

Define $\mathcal{D}_0 = \mathcal{B}_0$, Define a countable partition $\mathcal{D}_1$ by

$$\mathcal{D}_1 = \{J \cap L : J \in \mathcal{A}, L \in \mathcal{B}_1, \text{ and } |L| > 1\},$$

in order words, $\mathcal{D}_1$ consists of positive-length intervals where $\nabla Q^*(t)$ is entirely positive or negative and is not constant. Define

$$\mathcal{D}_2 = \{J \in \mathcal{B}_1 : J \subset A \text{ and } |J| = 1\},$$

i.e. $\mathcal{D}_2$ consists of isolated points in A.

One verifies that $\mathcal{D}$ is indeed a partition of A by checking that $D_0 = B_0$, $D_1 \cup D_2 = B_1 \cap A$, so that $D_0 \cup D_2 \cup D_2 = A$: it is also easy to check that elements of $\mathcal{D}$ are disjoint. Furthermore, as we mentioned earlier, the set D_2 is indeed of zero measure, since it consists of countably many isolated points.

Now we will show that for $t \in D_0$, we have $\nabla Q^*(t) = 0$. Take $t \in D$ for $D \in \mathcal{D}_0$, and let $a = \inf D$ and $b = \sup D$. Define

$$\xi^+ = I\{t \in D\}(Q^*(b) - Q^*(t))$$

and

$$\xi^- = I\{t \in D\}(Q^*(a) - Q^*(t)).$$

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 D, then for some $\epsilon > 0$ we would have $\mathcal{L}[Q^* + \epsilon\xi^-] < \mathcal{L}[Q^*]$, a contradiction. A similar argument with ξ^+ shows that $\nabla Q^*(t)$ cannot be strictly negative on D either. From this pertubation argument, we conclude that $\nabla Q^*(t) = 0$. Since this argument applies for all $t \in D_0$, we know that $\nabla Q^*(D_0) = \{0\}$.

The following observation is needed for the next stage of the proof. If we look at the function $Q^*(t)$, then up so sets of neglible measure, it is given by the expression (1) on the set D_0, and it is piecewise constant in between. But since (1) gives a strictly increasing function, and since Q^* is increasing, this implies that Q^* is discontinuous at the boundary between D_0 and D_1.

Now we are prepared to show that $\nabla Q^*(t) = 0$ for $t \in D_1$. Take $t \in D$ for $D \in \mathcal{D}_1$, and let $a = \inf D$ and $b = \sup D$. From the previous argument, there is a discontinuity at both a and b, so that $\lim_{u \to a^-} Q(u) < Q(t) < \lim_{u \to b^+} Q(u)$. Therefore, for any $\xi(t)$ which is increasing on (a, b) and zero elsewhere, there exists $\epsilon > 0$ such that $\nabla Q^* + \epsilon\xi \in \mathcal{Q}$. It remains to find such a perturbation ξ such that $\mathcal{L}[Q + \epsilon\xi] < \mathcal{L}[Q]$.

Define $F(t) = \int_a^t \nabla Q^*(u)du$ for $t \in [a, b]$. Since $\nabla Q^*(t) \neq 0$ on all but a zero-measure set within $[a, b]$, $F(t)$ cannot be constant, and falls into the following three cases:

- Case 1: $F(b) \neq 0$

- Case 2: $F(b) = 0$, and $F(t) < 0$ for some $t \in (a, b)$
- Case 3: $F(b) = 0$, and $F(t) \geq 0$ for all $t \in (a, b)$.

We will construct a suitable perturbation ξ for all three cases.

- Case 1: Construct $\xi(t) = -\text{Sign}(F(b))I\{t \in (a, b)\}$.
- Case 2: Find t_0 such that $F(t_0) < 0$. Construct

$$\xi(t) = \begin{cases} -1 & \text{for } t \in (a, t_0] \\ 0 & \text{otherwise} \end{cases}$$

- Case 3:

In all three cases, given the corresponding construction for $\xi(t)$ we get

$$\int_0^1 \xi(t) \nabla Q^*(t) dt$$

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_ι is the inverse of Q_c.

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

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

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

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

and

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

where

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

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

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 α. 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

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

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

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

Also,

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

□

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 Λ 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