

Upper bounds for average Bayes accuracy in terms of mutual information

Charles Zheng and Yuval Benjamini

October 20, 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

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

as stated in the following theorem.

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

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

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,

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

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

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

2 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 :

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

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

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

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 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 Theory

In this section we determine the value of $C_k(\iota)$, leading to 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,

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

We obtain this result by first reducing the problem to the case of densities with uniform marginals, then doing the optimization over the reduced space.

3.1 Reduction

Let $p(x, y)$ be a density supported on $\mathcal{X} \times \mathcal{Y}$, where \mathcal{X} is a subset of \mathbb{R}^{d_1} and \mathcal{Y} is a subset of \mathbb{R}^{d_2} , and such that $p(x)$ is uniform on \mathcal{X} and $p(y)$ is uniform on \mathcal{Y} .

Now let \mathcal{P}^{unif} denote the set of such distributions: in other words, \mathcal{P}^{unif} is the space of joint densities in Euclidean space with uniform marginals over the marginal supports. In this section, we prove that

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

thus reducing the problem of optimizing over the space of all densities to the problem of optimizing over densities with uniform marginals.

Also define $\mathcal{P}^{bounded}$ to be the space of all densities $p(x, y)$ with finite-volume support. Since uniform distributions can only be defined over sets of finite volume, we have

$$\mathcal{P}^{unif} \subset \mathcal{P}^{bounded} \subset \mathcal{P}.$$

Therefore, it is necessary to first show that

$$\inf_{p \in \mathcal{P}: I[p(x, y)] \leq \epsilon} \text{ABA}_k[p(x, y)] = \inf_{p \in \mathcal{P}^{\text{bounded}}: I[p(x, y)] \leq \epsilon} \text{ABA}_k[p(x, y)].$$

This is accomplished via the following lemma.

Lemma 3.2 (*Truncation*). *Let $p(x, y)$ be a density on $\mathbb{R}^{d_x} \times \mathbb{R}^{d_y}$. For all $\epsilon > 0$, there exists a subset $\mathcal{X} \subset \mathbb{R}^{d_x}$ with finite volume with respect to d_x -dimensional Lebesgue measure, and a subset $\mathcal{Y} \subset \mathbb{R}^{d_y}$ with finite volume with respect to d_y -dimensional Lebesgue measure, such that defining*

$$\tilde{p}(x, y) = \frac{I\{(x, y) \in \mathcal{X} \times \mathcal{Y}\}}{\int_{\mathcal{X} \times \mathcal{Y}} p(x, y) dx dy} p(x, y),$$

we have

$$|I[p] - I[\tilde{p}]| < \epsilon$$

and

$$|\text{ABA}_k[p] - \text{ABA}_k[\tilde{p}]| < \epsilon.$$

Proof. Recall the definition of the Shannon entropy H :

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

It is a well-known in information theory that

$$I[p(x, y)] = H[p(x)] + H[p(y)] - H[p(x, y)].$$

There exists a sequence $(\mathcal{X}_i, \mathcal{Y}_i)_{i=1}^{\infty}$ where $(\mathcal{X}_i)_{i=1}^{\infty}$ is an increasing sequence of finite-volume subsets of \mathbb{R}^{d_x} and $(\mathcal{Y}_i)_{i=1}^{\infty}$ is an increasing sequence of finite-volume subsets of \mathbb{R}^{d_y} , and $\lim_{i \rightarrow \infty} \mathcal{X}_i = \mathbb{R}^{d_x}$, $\lim_{i \rightarrow \infty} \mathcal{Y}_i = \mathbb{R}^{d_y}$. Define

$$\tilde{p}_i(x, y) = \frac{I\{(x, y) \in \mathcal{X}_i \times \mathcal{Y}_i\}}{\int_{\mathcal{X}_i \times \mathcal{Y}_i} p(x, y) dx dy} p(x, y)$$

Note that \tilde{p}_i gives the conditional distribution of (X, Y) conditional on $(X, Y) \in \mathcal{X}_i \times \mathcal{Y}_i$. Furthermore, it is convenient to define $\tilde{p}_{\infty} = p$. We can find some i_1 , such that for all $i \geq i_1$, we have

$$\left| \int_{x \notin \mathcal{X}_i} p(x) \log p(x) dx \right| < \frac{\epsilon}{6}$$

$$\left| \int_{y \notin \mathcal{Y}_i} p(y) \log p(y) dy \right| < \frac{\epsilon}{6}$$

$$\left| \int_{(x,y) \notin \mathcal{X}_i \times \mathcal{Y}_i} p(x,y) \log p(x,y) dx dy \right| < \frac{\epsilon}{6}$$

and also such that

$$-\log \left[\int_{x,y \in \mathcal{X}_i \times \mathcal{Y}_i} p(x,y) dx dy \right] < \frac{\epsilon}{2}$$

Then, it follows that

$$|I[p] - I[\tilde{p}_i]| < \epsilon$$

for all $i \geq i_1$.

Now we turn to the analysis of average Bayes error. Let f_i denote the Bayes k -class classifier for $\tilde{p}_i(x,y)$ and f_∞ the Bayes k -class classifier for $p(x,y)$: recall that by definition,

$$\text{ABA}_k[\tilde{p}_i] = \Pr_{\tilde{p}_i}[f_i(X_1, \dots, X_k, Y) = Z]$$

Define

$$\epsilon_i = \Pr_p[(X_1, \dots, X_k, Y) \notin \mathcal{X}_i^k \times \mathcal{Y}_i];$$

by continuity of probability we have $\lim_i \epsilon_i \rightarrow 0$. We claim that

$$|\text{ABA}_k[\tilde{p}_i] - \text{ABA}_k[p]| \leq \epsilon_i.$$

Given the claim, the proof is completed by finding $i > i_1$ such that $\epsilon_i < \epsilon$, and defining $\mathcal{X} = \mathcal{X}_i$, $\mathcal{Y} = \mathcal{Y}_i$.

Consider using f_i to obtain a classification rule for $p(x,y)$: define

$$\tilde{f}_i = \begin{cases} f_i(x_1, \dots, x_k, y) & \text{when } (x_1, \dots, x_k, y) \in \mathcal{X}_i^k \times \mathcal{Y} \\ 0 & \text{otherwise.} \end{cases}$$

We have

$$\begin{aligned}
\text{ABA}_k[p] &= \sup_f \Pr_p[f(X_1, \dots, X_k, Y) = Z] \\
&\geq \\
&= (1 - \epsilon_i) \Pr_p[f_i(X_1, \dots, X_k, Y) = Z | (X_1, \dots, X_k, Y) \in \mathcal{X}_i^k \times \mathcal{Y}_i] \\
&\quad + \epsilon_i \Pr_p[f_i(X_1, \dots, X_k, Y) = Z | (X_1, \dots, X_k, Y) \notin \mathcal{X}_i^k \times \mathcal{Y}_i] \\
&= (1 - \epsilon_i) \Pr_{\tilde{p}}[f_i(X_1, \dots, X_k, Y) = Z] + \epsilon_i 0 \\
&= (1 - \epsilon_i) \text{ABA}_k[\tilde{p}_i] \geq \text{ABA}_k[\tilde{p}_i] - \epsilon_i.
\end{aligned}$$

In other words, when \tilde{p}_i is close to p , the Bayes classification rule for \tilde{p}_i obtains close to the Bayes rate when the data is generated under p .

Now consider the reverse scenario of using f_p to perform classification under \tilde{p}_i . This is equivalent to generating data under $p(x, y)$, performing classification using f , then only evaluating classification accuracy conditional on $(X_1, \dots, X_k, Y) \in \mathcal{X}_i^k \times \mathcal{Y}_i$. Therefore,

$$\begin{aligned}
\text{ABA}_k[\tilde{p}_i] &= \sup_f \Pr_{\tilde{p}_i}[f(X_1, \dots, X_k, Y) = Z] \\
&\geq \Pr_{\tilde{p}_i}[f_p(X_1, \dots, X_k, Y) = Z] \\
&= \Pr_p[f_p(X_1, \dots, X_k, Y) = Z | (X_1, \dots, X_k, Y) \in \mathcal{X}_i^k \times \mathcal{Y}_i] \\
&= \frac{1}{1 - \epsilon_i} \Pr_p[I\{(X_1, \dots, X_k, Y) \in \mathcal{X}_i^k \times \mathcal{Y}_i\} \text{ and } f_p(X_1, \dots, X_k, Y) = Z] \\
&\geq \frac{1}{1 - \epsilon_i} \left(1 - \Pr_p[I\{(X_1, \dots, X_k, Y) \notin \mathcal{X}_i^k \times \mathcal{Y}_i\}] - \Pr_p[f_p(X_1, \dots, X_k, Y) \neq Z] \right) \\
&= \frac{\text{ABA}_k[p] - \epsilon_i}{1 - \epsilon_i} \geq \text{ABA}_k[p] - \epsilon_i.
\end{aligned}$$

In other words, when \tilde{p}_i is close to p , the Bayes classification rule for p obtains close to the Bayes rate when the data is generated under \tilde{p}_i .

Combining the two directions gives $|\text{ABA}_k[\tilde{p}_i] - \text{ABA}_k[p]| \leq \epsilon_i$, as claimed.

□

One can go from bounded-volume sets to uniform distributions by adding auxillary variables. To illustrate the intuition, consider a density $p(x)$ on a

set of bounded volume, \mathcal{X} . Introduce a variable W such that conditional on $X = x$, we have w uniform on $[0, p(x)]$. It follows that the joint density $p(x, w) = 1$ and is supported on a set $\mathcal{X}' = \mathcal{X} \times [0, \infty]$. Furthermore, \mathcal{X}' is of bounded volume (in fact, of volume 1) since

$$\int_{\mathcal{X}'} dx = \int_{\mathcal{X}'} p(x, w) dx = 1.$$

Therefore, to accomplish the reduction from \mathcal{P} to \mathcal{P}^{unif} , we start with a density $p(x, y) \in \mathcal{P}$, and using Lemma 3.2, find a suitable finite-volume truncation $\tilde{p}(x, y)$. Finally, we introduce auxillary variables w and z so that the expanded joint distribution $p(x, w, y, z)$ has uniform marginals $p(x, w)$ and $p(y, z)$. However, we still need to check that the introduction of auxillary variables preserves the mutual information and average Bayes error; this is the content of the next lemma.

Lemma 3.3 *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,*

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

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

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

Also,

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

□

Combining these lemmas gives the needed reduction, given by the following theorem.

Theorem 3.4 (*Reduction.*)

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

The proof is trivial given the previous two lemmas.

3.2 Optimization

Having reduced the problem to an optimization over \mathcal{P}^{unif} , in this section we use variational calculus to find the global optimum to the optimization problem

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

The proof depends on the following lemmas.

Lemma 3.5 *Let $f(t)$ be an increasing function from $[a, b] \rightarrow \mathbb{R}$, where $a < b$, and let $g(t)$ be a bounded continuous function from $[a, b] \rightarrow \mathbb{R}$. 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 3.6 *Let $f(t)$ be a measurable function from $[a, b] \rightarrow \mathbb{R}$, 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 3.7 *Define an exponential family on $[0, 1]$ by the density function*

$$q_\beta(t) = \exp[\beta t^{k-1} - \log Z(\beta)]$$

where

$$Z(\beta) = \int_0^1 \exp[\beta t^{k-1}] dt.$$

Then, the negative entropy

$$I(\beta) = \int_0^1 q_\beta(t) \log q_\beta(t) dt$$

is decreasing in β on the interval $(-\infty, 0]$. and increasing on the interval $[0, \infty)$.

Furthermore, for any $\iota \in (0, \infty)$, there exist two solutions to $I(\beta) = \iota$: one positive and one negative.

Lemma 3.8 *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 3.9 *The map*

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

is concave in $\iota > 0$.

Proof of Lemma 3.5. (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 property that for all $t \in A^\pm$, the 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 $i = 1, \dots$, 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 $I \in \mathcal{A}^+$, 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 3.6. (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 $B \in \mathcal{B}_0$, 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 3.7.

Define $\beta(\mu)$ as the solution to

$$\mu = \int_0^1 t q_\beta(t) dt.$$

By [Wainwright and Jordan 2008], the function $\beta(\mu)$ is well-defined. Furthermore, since the sufficient statistic t^{k-1} is increasing in t , it follows that $\beta(\mu)$ is increasing.

Define the negative entropy as a function of μ ,

$$N(\mu) = \int_0^1 q_{\beta(\mu)}(t) \log q_{\beta(\mu)}(t) dt.$$

By Theorem 3.4 of [Wainwright and Jordan 2008], $N(\mu)$ is convex in μ . We claim that the derivative of $N(\mu) = 0$ at $\mu = \frac{1}{2}$. This implies that $N(\mu)$ is decreasing in μ for $\mu \leq \frac{1}{2}$ and increasing for $\mu \geq \frac{1}{2}$. Since $I(\beta(\mu)) = N(\mu)$, β is increasing in μ , and $\beta(\frac{1}{2}) = 0$, this implies that $I(\beta)$ is decreasing in β for $\beta \leq 0$ and increasing for $\beta \geq 0$.

We will now prove the claim. Write

$$\left. \frac{d}{d\mu} N(\mu) \right|_{\mu=1/2} = \left. \frac{d}{d\beta} I(\beta(\mu)) \right|_{\beta=0} \left. \frac{d\beta}{d\mu} \right|_{\mu=1/2}.$$

We have

$$\frac{d}{d\beta} I(\beta) = \beta \int q_\beta t^{k-1} dt - \log Z(\beta).$$

Meanwhile, $Z(0) = 1$ so $\log Z(0) = 0$. Therefore,

$$\left. \frac{d}{d\beta} I(\beta) \right|_{\beta=0} = 0.$$

This implies that $\left. \frac{d}{d\mu} N(\mu) \right|_{\mu=1/2} = 0$, as needed.

For the final statement of the lemma, note that $I(0) = 0$ since q_0 is the uniform distribution. Meanwhile, since q_β tends to a point mass as either $\beta \rightarrow \infty$ or $\beta \rightarrow -\infty$, we have

$$\lim_{\beta \rightarrow \infty} I(\beta) = \lim_{\beta \rightarrow -\infty} I(\beta) = \infty.$$

And, as we can check that $I(\beta)$ is continuous in β , this means that

$$I((-\infty, 0]) = I([0, \infty)) = [0, \infty)$$

by the mean-value theorem. Combining this fact with the monotonicity of $I(\beta)$ restricted to either the positive and negative half-line yields the fact that for any $\iota > 0$, there exists $\beta_1 < 0 < \beta_2$ such that $I(\beta_1) = I(\beta_2) = \iota$. \square .

Proof of Lemma 3.8. (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}_ι denote the collection of quantile functions of distributions in P_ι . Then, \mathcal{Q}_ι consists of monotonic functions $Q : [0, 1] \rightarrow [0, \infty]$ 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

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

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

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

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

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

The global minimizer Q^* is also necessarily a stationary point: that is, for any perturbation function $\xi : [0, 1] \rightarrow \mathbb{R}$ such that $Q^* + \xi \in \mathcal{Q}$, we have $\mathcal{L}[Q^*] \leq \mathcal{L}[Q^* + \xi]$. For sufficiently small ξ , we have

$$\mathcal{L}[Q + \xi] \approx \mathcal{L}[Q] + \int_0^1 \xi(t) (-t^{k-1} + \lambda + \nu + \nu \log Q(t)) dt. \quad (2)$$

Define

$$\nabla \mathcal{L}_{Q^*}(t) = -t^{k-1} + \lambda + \nu + \nu \log Q(t). \quad (3)$$

The function $\nabla \mathcal{L}_{Q^*}(t)$ is a *functional derivative* of the Lagrangian. Note that if we were able to show that $\nabla \mathcal{L}_{Q^*}(t) = 0$, this immediately yields

$$Q^*(t) = \exp[-1 - \lambda\nu^{-1} + \nu^{-1}t^{k-1}]. \quad (4)$$

At this point, we know that the right-hand side of (4) gives a stationary point of \mathcal{L} , but we cannot be sure that it gives the global minimizer. The reason is because the optimization occurs on a constrained space. We will show that (4) indeed gives the global minimizer Q^* , but we do so by showing that the set of points t where $\nabla \mathcal{L}_{Q^*}(t) \neq 0$ is of zero measure. Since sets of zero measure don't affect the integrals defining the optimization problem (1), we conclude there exists a global optimal solution with $\nabla \mathcal{L}_{Q^*}(t) = 0$ everywhere, which is therefore given explicitly by (4) for some $\lambda \in \mathbb{R}$, $\nu \geq 0$.

We will need the following result: that for $\iota > 0$, any solution to (1) satisfies $\phi_k[Q] < 1$. This follows from the fact that

$$E[Q] - \phi_k[Q] = \int_0^1 (1 - t^{k-1})Q(t)dt,$$

where the term $(1 - t^{k-1})$ is negative, except for the one point $t = 1$. Therefore, in order for $\phi_k[Q] = 1 = E[Q]$, we must have $Q(t) = 0$ for $t < 1$. However, this yields a contradiction since $Q(t) = 0$ for $t < 1$ implies that $E[Q] = 0$, a violation of the hard constraint $E[Q] = 1$.

Let us establish that $\nu > 0$: in other words, the constraint $I[Q] = \iota$ is tight. Suppose to the contrary, that for some $\iota > 0$, the global optimum Q^* minimizes a Lagrangian with $\nu = 0$. Let $\phi^* = \phi_k[Q^*] < 1$. However, if we define $Q_\kappa(t) = I\{t \geq 1 - \frac{1}{\kappa}\}\kappa$, we have $E[Q_\kappa] = 1$, and also for some sufficiently large $\kappa > 0$, $\phi_k[Q_\kappa] > \phi^*$. But since the Lagrangian lacks a term corresponding to $I[Q]$, we conclude that $\mathcal{L}[Q_\kappa] < \mathcal{L}[Q^*]$, a contradiction.

The rest of the proof proceeds as follows. We will use Lemmas 3.5 and 3.6 to define a decomposition $A = D_0 \cup D_1 \cup D_2$, where D_2 is of measure zero. First, we show that assuming the existence of $t \in D_0$ yields a contradiction, and hence $D_0 = \emptyset$. Then, again using argument from contradiction we establish that $D_1 = \emptyset$. 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.

We will now apply the Lemmas to obtain the necessary ingredients for constructing the sets D_i . Since $\nabla \mathcal{L}_{Q^*}(t)$ is a difference between an increasing function and a continuous strictly increasing function, we can apply Lemma 3.5 to conclude that there exists a countable partition \mathcal{A} of the set $A : \{t \in [0, 1] : \nabla \mathcal{L}_{Q^*}(t) \neq 0\}$ into intervals such that for all $J \in \mathcal{A}$, $|\text{Sign}(\nabla Q^*(J))| = 1$ and $\inf J < \sup J$. Applying Lemma 3.6 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 $\nabla \mathcal{L}_{Q^*}(t)$ 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 $\nabla \mathcal{L}_{Q^*}(t)$ 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 $\nabla \mathcal{L}_{Q^*}(t) = 0$ 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 $D_0 \cup D_1 \cup D_2 = A$: 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 the existence of $t \in D_0$ implies a contradiction. 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 $\nabla \mathcal{L}_{Q^*}(t)$ 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 $\nabla \mathcal{L}_{Q^*}(t)$ cannot be strictly negative on D either. From this perturbation argument, we conclude that $\nabla \mathcal{L}_{Q^*}(t) = 0$. Since this argument applies for all $t \in D_0$, we conclude that $D_0 = \emptyset$: therefore, on the set $[0, 1] \setminus (D_1 \cup D_2)$, we have $\nabla \mathcal{L}_{Q^*}(t) = 0$.

The following observation is needed for the next stage of the proof. If we look at the function $Q^*(t)$, then up to sets of negligible measure, it is given by the expression (4) on the set $[0, 1] \setminus D_1$, and it is piecewise constant in-between. But since (4) gives a strictly increasing function, and since Q^* is increasing, this implies that Q^* is discontinuous at the boundary of D_1 .

Now we are prepared to show that $\nabla \mathcal{L}_{Q^*}(t) = 0$ 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 $\lim_{u \rightarrow a^-} Q(u) < Q(t) < \lim_{u \rightarrow b^+} Q(u)$. Therefore, for any $\xi(t)$ which is increasing on D 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 $\mathcal{L}[Q + \epsilon \xi] < \mathcal{L}[Q]$.

Also, since by definition $\nabla \mathcal{L}_{Q^*}(t)$ is constant on D , follows from (3) that ∇Q^* is strictly decreasing, and thus either

- Case 1: $\nabla \mathcal{L}_{Q^*}(t) \geq 0$ on D ,
- Case 2: $\nabla \mathcal{L}_{Q^*}(t) \leq 0$ on D , or
- Case 3: $\nabla \mathcal{L}_{Q^*}(t) \geq 0$ for all $t \in D \cap [a, t_0]$ and $\nabla \mathcal{L}_{Q^*}(t) \leq 0$ for all $t \in D \cap [t_0, b]$.

Depending on the case, we construct a suitable perturbation ξ :

- Case 1: Construct $\xi(t) = -I\{t \in D\}$.
- Case 2: Construct $\xi(t) = I\{t \in D\}$
- Case 2: Construct

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

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

$$\int_0^1 \xi(t) \nabla \mathcal{L}_{Q^*}(t) dt < 0.$$

Therefore, from (2), there exists some $\epsilon > 0$ such that $\mathcal{L}[Q + \epsilon \xi] < \mathcal{L}[Q]$, a contradiction. Again, since the contradiction applies for all $t \in D_1$, we conclude that $D_1 = \emptyset$.

By now we have established that a global optimum for (1) exists, and is given by (4) for some $\lambda \in \mathbb{R}$, $\nu > 0$. It remains to determine the values of λ and ν .

Reparameterize $\alpha = \exp[-1 - \lambda \nu^{-1}]$ and $\beta = \nu^{-1}$. Therefore,

$$Q^*(t) = \alpha \exp[\beta t^{k-1}]$$

for $\alpha > 0$, $\beta > 0$. There is a one-to-one mapping from $(\alpha, \beta) \in (0, \infty)^2$ to $(\lambda, \nu) \in \mathbb{R} \times (0, \infty)$.

Now, from the constraint

$$1 = E[Q^*] = \int_0^1 \alpha \exp[\beta t^{k-1}] dt.$$

we conclude that

$$\alpha = \frac{1}{\int_0^1 \exp[\beta t^{k-1}] dt}.$$

Therefore, we have reduced the set of possible solutions Q^* to a one-parameter family,

$$Q^*(t) = \frac{\exp[\beta t^{k-1}]}{Z(\beta)}.$$

where

$$Z(\beta) = \int_0^1 \exp[\beta t^{k-1}] dt.$$

Next, note that

$$I[Q^*] = \int_0^1 Q^*(t) \log Q^*(t) = \beta \mu_\beta - \log Z(\beta),$$

as a function of β , is completely characterized by Lemma 3.7. Let us define c_ι as the unique positive solution to the equation

$$c_\iota \mu_{c_\iota} - \log Z(c_\iota) = \iota$$

given by Lemma 3.7. We therefore have

$$Q^*(t) = \frac{\exp[c_\iota t^{k-1}]}{\int_0^1 \exp[c_\iota t^{k-1}]},$$

as needed. \square

Proof of Lemma 3.9. It is equivalent to show that the inverse function

$$C_k^{-1}(p) = \inf_{G: E[G]=1, \phi_k[G]=p} I[G]$$

is convex. Let $p_1, p_2 \in [0, 1]$. From lemma 3.8, we can find measures G_1, G_2 on $[0, \infty)$ which minimize $I[G_i]$ subject to $E[G_i] = 1$ and $\phi_k[G_i] = p_i$. Define the measure

$$H = \frac{G_1 + G_2}{2}.$$

Since ϕ_k is a linear functional,

$$\phi_k[H] = \frac{\phi_k[G_1] + \phi_k[G_2]}{2} = \frac{p_1 + p_2}{2}.$$

But since I is a convex functional,

$$I[H] \leq \frac{I[G_1] + I[G_2]}{2}.$$

Therefore,

$$C_k^{-1}\left(\frac{p_1 + p_2}{2}\right) \leq I[H] = \frac{I[G_1] + I[G_2]}{2} = \frac{C_k^{-1}(p_1) + C_k^{-1}(p_2)}{2}.$$

□.

Proof of theorem 3.1

Using Theorem 3.4, we have

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

Define $f(\iota) = \int_0^1 Q_{c_\iota}(t) t^{k-1} dt$: our goal is to establish that $C_k(\iota) = f(\iota)$. Note that $f(\iota)$ is the same function which appears in Lemma 3.9 and the same bound as established in Lemma 3.8.

Define the 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_ι} .

One can verify that $I[p_\iota] = \iota$, and

$$\text{ABA}_k[p] = \int_0^1 Q_{c_\iota}(t) t^{k-1} dt.$$

This establishes that

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

It remains to show that for all $p \in \mathcal{P}^{unif}$ with $I[p] \leq \iota$, that $\text{ABA}_k[p] \leq \text{ABA}_k[p_\iota]$.

Take $p \in \mathcal{P}^{unif}$ such that $I[p] \leq \iota$. 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,

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

while

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

Letting G_y be the distribution of $Z(y)$, we have

$$E[G_y] = 1$$

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

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

where the expectation is taken over $Y \sim \text{Unif}[0, 1]$ and where $E[G]$, $I[G]$, and $\psi_k[G]$ are defined as in Lemma 3.8.

Define the random variable $J = I[G_Y]$. We have

$$\begin{aligned} \text{ABA}_k[p(x, y)] &= \mathbf{E}(\psi_k[G_Y]) \\ &= \int_0^1 \psi_k[G_y] dy \\ &\leq \int_0^1 \left(\sup_{G: I[G] \leq I[G_y]} \psi_k[G] \right) dy \\ &= \int_0^1 f(I[G_y]) dy = \mathbf{E}[f(J)]. \end{aligned}$$

Now, since f is concave by Lemma 3.9, we can apply Jensen's inequality to conclude that

$$\text{ABA}_k[p(x, y)] = \mathbf{E}[f(J)] \leq f(\mathbf{E}[J]) = f(\iota),$$

which completes the proof. \square

Appendix

3.3 Computation

Noting that Q_β in theorem 3.1 defines an exponential family on $[0, 1]$, we obtain the following result:

$$C_k(\iota) = \sup_{\beta: I^{(k)}(\beta) = \iota} \text{ABA}^{(k)}(\beta)$$

where

$$I^{(k)}(\beta) = \beta \dot{\Lambda}^{(k)}(\beta) - \Lambda^{(k)}(\beta)$$

and

$$\text{ABA}^{(k)}(\beta) = \dot{\Lambda}^{(k)}(\beta)$$

for

$$\Lambda^{(k)}(\beta) = \log \int_0^1 \exp[\beta t^{k-1}] dt.$$

Therefore, computing $C_k(\iota)$ depends in turn on computing $\Lambda^{(k)}(\beta)$ and its first derivative.

The following result provides a rapidly converging power series for $\Lambda^{(k)}(\beta)$ and also a limiting result which facilitates large- k approximation.

Theorem 3.10

$$\Lambda^{(k)}(\beta) = \log \sum_{j=0}^{\infty} \frac{\beta^j}{j!(j + \frac{1}{k-1})} - \log(k-1)$$

and therefore,

$$\lim_{k \rightarrow \infty} (k-1)(\exp(\Lambda^{(k)}) - 1) = W(\beta),$$

where

$$W(\beta) = \sum_{j=1}^{\infty} \frac{\beta^j}{j!j}.$$

Proof. Write

$$\begin{aligned}
(k-1)(e^{\Lambda^{(k)}(\beta)} - 1) &= (k-1)\left(\int_0^1 \exp(\beta t^{k-1}) dt - 1\right) \\
&= (k-1)\left(\int_0^1 \sum_{j=0}^{\infty} \frac{\beta^j t^{j(k-1)}}{j!} dt - 1\right) \\
&= (k-1)\left(\sum_{j=0}^{\infty} \int_0^1 \frac{\beta^j t^{j(k-1)}}{j!} dt - 1\right) \\
&= (k-1)\left(\sum_{j=0}^{\infty} \frac{\beta^j}{j!(j(k-1)+1)} dt - 1\right) \\
&= \left(\sum_{j=0}^{\infty} \frac{\beta^j}{j!(j + \frac{1}{k-1})} dt - (k-1)\right) \\
&= \sum_{j=1}^{\infty} \frac{\beta^j}{j!(j + \frac{1}{k-1})} dt.
\end{aligned}$$

Taking the limit as $k \rightarrow \infty$ yields the stated formula for $W(\beta)$. \square