

# Hölder's identity

W. D. Brinda<sup>a</sup>, Jason M. Klusowski<sup>b</sup>, Dana Yang<sup>a</sup>

<sup>a</sup>*Department of Statistics and Data Science,  
Yale University, New Haven, CT, USA, 06511*

<sup>b</sup>*Department of Statistics and Biostatistics,  
Rutgers University - New Brunswick, Piscataway, NJ, USA, 08901*

---

## Abstract

We clarify that Hölder's inequality can be stated more generally than is often realized. This is an immediate consequence of an analogous information-theoretic identity which we call *Hölder's identity*. We also explain Andrew R. Barron's original use of the identity.

*Keywords:* Hölder's inequality, measure theory, information theory

---

## 1. Generality of Hölder's inequality

Hölder's inequality is most commonly written

$$\int |f(y)g(y)|d\mu(y) \leq \|f\|_p \|g\|_q \quad (1)$$

for conjugate exponents  $p$  and  $q$ . An alternative way of expressing this is to say that for any pair of non-negative functions  $f$  and  $g$  and any  $\alpha \in [0, 1]$ ,

$$\int f^\alpha(y)g^{1-\alpha}(y)d\mu(y) \leq \left(\int f(y)d\mu(y)\right)^\alpha \left(\int g(y)d\mu(y)\right)^{1-\alpha}. \quad (2)$$

In other words, *the integral of the point-wise geometric average of two functions is bounded by the geometric average of their integrals*. In fact, this relationship holds for *arbitrary* geometric expectations over a random element indexing

---

*Email addresses:* [william.brinda@yale.edu](mailto:william.brinda@yale.edu) (W. D. Brinda),  
[jason.klusowski@rutgers.edu](mailto:jason.klusowski@rutgers.edu) (Jason M. Klusowski), [xiaoqian.yang@yale.edu](mailto:xiaoqian.yang@yale.edu) (Dana Yang)

5 functions<sup>[1]</sup>

**Theorem 1.1** (Hölder’s inequality). *Let  $\mathcal{X}$  and  $\mathcal{Y}$  be measurable spaces, and let  $f : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}^+$  be product measurable. For any  $\sigma$ -finite measure  $\mu$  on  $\mathcal{Y}$  and any  $\mathcal{X}$ -valued random element  $X$ ,*

$$\int e^{\mathbb{E} \log f(X,y)} d\mu(y) \leq e^{\mathbb{E} \log \int f(X,y) d\mu(y)}.$$

Inequalities (1) and (2) represent the two-point distribution version of Theorem 1.1. The generalization for an arbitrary finite measure on  $\mathcal{X}$  is easy to derive by normalizing and then applying the result for probability measures.

**Corollary 1.2.** *Let  $\mathcal{X}$  and  $\mathcal{Y}$  be measurable spaces, and let  $f : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}^+$  be product measurable. For any  $\sigma$ -finite measure  $\mu$  on  $\mathcal{Y}$  and finite measure  $\gamma$  on  $\mathcal{X}$ ,*

$$\int e^{\int \log f(x,y) d\gamma(x)} d\mu(y) \leq e^{\frac{1}{\gamma(\mathcal{X})} \int [\log \int f(x,y)^{\gamma(\mathcal{X})} d\mu(y)] d\gamma(x)}.$$

Using  $e^f$  as the function in Theorem 1.1, and taking the log of both sides  
10 gives us an equivalent inequality that is also worth stating.

**Corollary 1.3.** *Let  $\mathcal{X}$  and  $\mathcal{Y}$  be measurable spaces, and let  $f : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$  be product measurable. For any  $\sigma$ -finite measure  $\mu$  on  $\mathcal{Y}$  and any  $\mathcal{X}$ -valued random element  $X$ ,*

$$\log \int e^{\mathbb{E} f(X,y)} d\mu(y) \leq \mathbb{E} \log \int e^{f(X,y)} d\mu(y).$$

The fact that Hölder’s inequality holds in this generality is perhaps not widely known. For example, [8] proved an extension of Hölder’s inequality to countable products assuming  $\mu$  is  $\sigma$ -finite; that result was improved by [3, Thm 2.11]. The inequalities they present are readily subsumed by Corollary 1.2 by  
15 letting  $\gamma$  concentrate on a countable set.

---

<sup>1</sup>The proof of Theorem 1.1 will come later in this paper. Every result that we label *Theorem* or *Lemma* will have a formal proof in Section 3, while results labeled *Corollary* are explained informally before being stated.

[7] Lemma 1] states our Corollary [1.3], but the justification presented there is not quite adequate. They observe, using the two-point distribution version of Hölder’s inequality, that the mapping  $f \mapsto \log \mu e^f$  is convex on the space of real-valued functions on a set. [Pettis] expectations commute with continuous affine  
 20 functionals, and Jensen’s inequality relies on the expectation commuting with a continuous affine functional tangent to the convex function. The existence of a tangent continuous affine functional is guaranteed for convex functions on finite-dimensional spaces, but not on infinite-dimensional spaces. (As a simple example, consider any discontinuous linear functional; it is convex, but it has  
 25 no continuous affine functional tangent to it. For a more concrete example, see [10, Introduction].) If adequate care is taken, the logic of Jensen’s inequality can be applied to this problem as we show in Section 3 in the supplementary material; there, we prove a variant of Hölder’s identity that does not require  $\sigma$ -finiteness.

[7] reference [14] where the inequality in our Theorem [1.1] is stated and called *generalized Hölder’s inequality*; he points to the classic text [5, VI.11 Ex 36] where it is left as an exercise. Although that exercise does not say to assume  $\sigma$ -finiteness, the proof they hint at does require it — see Section 2 in the supplementary material. For  $\sigma$ -finite measures, at least, the proof can follow a  
 35 different route from the one they hint at. We establish an identity that has an information-theoretic interpretation involving a non-negative “variance” functional  $\tilde{V}$  for random probability measures which will be defined and explained in Section [2]. For now, suffice it to say that  $\tilde{V}$  represents an expected relative entropy.

**Theorem 1.4.** *Let  $\mathcal{X}$  and  $\mathcal{Y}$  be measurable spaces, and let  $f : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$  be product measurable. Let  $\mu$  be a  $\sigma$ -finite measure on  $\mathcal{Y}$ , and let  $X \sim P$  be an  $\mathcal{X}$ -valued random element. If  $\int e^{f(x,y)} d\mu(y)$  is in  $(0, \infty)$   $P$ -almost surely and  $\mathbb{E} \log \int e^{f(X,y)} d\mu(y) > -\infty$ , then*

$$\mathbb{E} \log \int e^{f(X,y)} d\mu(y) - \log \int e^{\mathbb{E}f(X,y)} d\mu(y) = \tilde{V}Q_X$$

40 where  $Q_x$  has density  $q_x(y) := \frac{e^{f(x,y)}}{\int e^{f(x,y)} d\mu(y)}$  with respect to  $\mu$ .

**Corollary 1.5** (Hölder’s identity). *Let  $\mathcal{X}$  and  $\mathcal{Y}$  be measurable spaces, and let  $f : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}^+$  be product measurable. Let  $\mu$  be a  $\sigma$ -finite measure on  $\mathcal{Y}$ , and let  $X \sim P$  be an  $\mathcal{X}$ -valued random element. If  $\int f(x, y)d\mu(y)$  is in  $(0, \infty)$   $P$ -almost surely and  $\mathbb{E} \log \int f(X, y)d\mu(y) > -\infty$ , then*

$$\frac{e^{\mathbb{E} \log \int f(X, y)d\mu(y)}}{\int e^{\mathbb{E} \log f(X, y)d\mu(y)}} = e^{\tilde{\mathbb{V}}Q_X}$$

where  $Q_x$  has density  $q_x(y) := \frac{f(x, y)}{\int f(x, y)d\mu(y)}$  with respect to  $\mu$ .

An interpretation of  $\tilde{\mathbb{V}}Q_X$  will be informed by the “reverse compensation identity” which we describe in the coming section.

In the special case that  $X$  only takes two possible values,  $\tilde{\mathbb{V}}Q_X$  is an *unnorm-*  
 45 *malized Rényi divergence*  $D_\lambda$  between the two possible distributions, as defined in Section 2.

**Theorem 1.6.** *Let  $\mathcal{Y}$  be a measurable space, and let  $f : \mathcal{Y} \rightarrow \mathbb{R}^+$  and  $g : \mathcal{Y} \rightarrow \mathbb{R}^+$  have finite positive  $\mu$ -integrals. Then*

$$\frac{\int f^\lambda(y)g^{1-\lambda}(y)d\mu(y)}{\int f^\lambda(y)d\mu(y) \int g^{1-\lambda}(y)d\mu(y)} = e^{D_\lambda(Q\|R)}$$

where  $Q$  has density  $\frac{f(y)}{\int f(y)d\mu(y)}$  and  $R$  has density  $\frac{g(y)}{\int g(y)d\mu(y)}$  with respect to  $\mu$ .

## 2. The compensation identities

Theorem 2.1, called the *compensation identity* by [16, Thm 9.1], conveniently  
 50 decomposes the expected relative entropy from a random probability measure to a fixed probability measure.<sup>2</sup>

**Theorem 2.1** (The compensation identity). *Let  $\{q_x : x \in \mathcal{X}\}$  be a family of probability densities with respect to a  $\sigma$ -finite measure  $\mu$ , and suppose that  $(x, y) \mapsto q_x(y)$  is product measurable. Let  $X \sim P$  be an  $\mathcal{X}$ -valued random*

---

<sup>2</sup>In Theorem 2.1 and throughout the remainder of this paper, lower-case and upper-case letters implicitly pair probability measures with their densities.

element. For any probability measure  $R$  on  $\mathcal{Y}$ ,

$$\mathbb{E}D(Q_X \| R) = D(\bar{Q}_P \| R) + \mathbb{E}D(Q_X \| \bar{Q}_P)$$

where  $\bar{Q}_P$  represents the  $P$ -mixture over  $\{q_x\}$ .

A less familiar decomposition, which we will call the *reverse compensation identity*, holds when the expected relative entropy's *second* argument is random rather than its first. Instead of a mixture, it involves a *geometric-mixture*.<sup>3</sup> We define the  $P$ -geometric mixture of  $\{q_x\}$  to be the probability measure with density

$$\tilde{q}_P(y) := \frac{e^{\mathbb{E}_{X \sim P} \log q_X(y)}}{\int e^{\mathbb{E}_{X \sim P} \log q_X(y)} d\mu(y)}.$$

Jensen's inequality and Tonelli's theorem together provide an upper bound for the denominator.

$$\begin{aligned} \int e^{\mathbb{E} \log q_X(y)} d\mu(y) &\leq \mathbb{E} \int e^{\log q_X(y)} d\mu(y) \\ &= 1 \end{aligned}$$

This integral can be zero, however, in which case the geometric-mixture is not well-defined.<sup>4</sup>

**Theorem 2.2** (The reverse compensation identity). *Let  $\{q_x : x \in \mathcal{X}\}$  be a family of probability densities with respect to a  $\sigma$ -finite measure  $\mu$ , and suppose that  $(x, y) \mapsto q_x(y)$  is product measurable. Let  $X \sim P$  be an  $\mathcal{X}$ -valued random element. If  $\int e^{\mathbb{E} \log q_X(y)} d\mu(y) > 0$ , then for any probability measure  $R$  on  $\mathcal{Y}$ ,*

$$\mathbb{E}D(R \| Q_X) = D(R \| \tilde{Q}_P) + \mathbb{E}D(\tilde{Q}_P \| Q_X)$$

<sup>55</sup> where  $\tilde{Q}_P$  represents the  $P$ -geometric mixture over  $\{q_x\}$ .

---

<sup>3</sup>What we call a "geometric mixture" is sometimes called a "log mixture" or "log-convex mixture," for instance by [6, Sec 19.6].

<sup>4</sup>An example of such a pathological case is when  $q_X$  has positive probabilities on two densities that are mutually singular.

A two-point distribution version of Theorem 2.2 is implied by [4, Eq (3) with (4)] and similarly for any finite set of discrete distributions by [17, Equation (9)].

Theorems 2.1 and 2.2 are perfectly analogous to the bias-variance decomposition for Hilbert-space-valued random vectors.<sup>5</sup> The expected divergence from a random element to a fixed element decomposes into the divergence from a “centroid” of the random element to that fixed element plus the internal variation of the random element from that centroid.<sup>6</sup> We suggest a notation that makes use of this intuition:

$$\begin{aligned}\bar{\mathbb{V}}Q_X &:= \inf_R \mathbb{E}D(Q_X \| R) \\ &= \mathbb{E}D(Q_X \| \bar{Q}_P)\end{aligned}$$

and<sup>7</sup>

$$\begin{aligned}\tilde{\mathbb{V}}Q_X &:= \inf_R \mathbb{E}D(R \| Q_X) \\ &= \begin{cases} \mathbb{E}D(\tilde{Q}_P \| Q_X), & \text{if } \int e^{\mathbb{E} \log q_X(y)} d\mu(y) > 0 \\ \infty, & \text{otherwise.} \end{cases}\end{aligned}$$

Roughly speaking,  $\bar{\mathbb{V}}Q_X$  represents the smallest possible expected code-length redundancy one can achieve when the *coding* distribution is the random  $Q_X$ ; to achieve it, one sets the decoding distribution to be  $\bar{Q}_P$ . On the other hand,  $\tilde{\mathbb{V}}Q_X$  represents the smallest possible expected code-length redundancy when the *decoding* distribution is the random  $Q_X$ ; to achieve it, one sets the coding distribution to be  $\tilde{Q}_P$ .

It is interesting to note that two-point distribution versions of these variance-like quantities are often used as divergences. The *Jensen-Shannon divergence* between probability measures  $Q$  and  $R$  is  $\bar{\mathbb{V}}$  of the random probability measure

---

<sup>5</sup>In fact, the compensation identity and bias-variance decomposition are both instances of this decomposition for Bregman divergences — see [15, Lem 3.5] and [12].

<sup>6</sup>It follows that the centroid is the choice of fixed element that has the smallest possible expected divergence from the random element.

<sup>7</sup>This alternative representation of  $\tilde{\mathbb{V}}$  is justified by Lemma 3.4

that takes values  $Q$  and  $R$  each with probability  $1/2$ .

$$D_{\text{JS}}(Q, R) := \frac{1}{2}D\left(Q\|\frac{Q+R}{2}\right) + \frac{1}{2}D\left(R\|\frac{Q+R}{2}\right)$$

*Unnormalized Bhattacharyya divergence*<sup>[8]</sup> is the  $\tilde{\mathbb{V}}$  analogue:

$$D_{\text{UB}}(Q, R) = \frac{1}{2}D\left(\frac{\sqrt{qr}}{\mu\sqrt{qr}}\|q\right) + \frac{1}{2}D\left(\frac{\sqrt{qr}}{\mu\sqrt{qr}}\|r\right)$$

where  $q$  and  $r$  are densities of  $Q$  and  $R$  with respect to  $\mu$ , and  $\mu\sqrt{qr}$  is short-hand  
65 for  $\int \sqrt{q(y)r(y)}d\mu(y)$  using de Finetti notation.<sup>[9]</sup> The derivation is straight-  
forward using the definition  $D_{\text{UB}}(Q, R) := \log \frac{1}{\mu\sqrt{qr}}$ , but it is more easily seen  
via Lemma 3.3. *Unnormalized Rényi divergence* is a generalization  $D_\lambda(Q\|R) :=$   
 $\log \frac{1}{\mu q^\lambda r^{1-\lambda}}$ , and a random distribution that takes values  $Q$  with probability  $\lambda$   
and  $R$  with probability  $1 - \lambda$  has a  $\tilde{\mathbb{V}}$  of  $D_\lambda(Q\|R)$ .

### 70 3. Proofs

It is known that relative entropy can be expressed in terms of a non-negative integrand. This fact enables us to use Tonelli's theorem to justify interchanges in the order of integration.

**Lemma 3.1.** *Let  $\{q_x : x \in \mathcal{X}\}$  and  $\{r_x : x \in \mathcal{X}\}$  be families of probability densities with respect to a  $\sigma$ -finite measure  $\mu$ , and suppose that both  $(x, y) \mapsto q_x(y)$  and  $(x, y) \mapsto r_x(y)$  are product measurable. For any  $\mathcal{X}$ -valued random element  $X$ ,*

$$\mathbb{E}\mu q_X \log \frac{q_X}{r_X} = \mu \mathbb{E}q_X \log \frac{q_X}{r_X}.$$

<sup>8</sup>This terminology is borrowed from [6] Eq (19.38)].

<sup>9</sup>The de Finetti notation writes measures like ordinary functionals that can be applied to measurable functions; it is summarized and advocated in [13] Sec 1.4]. We will use this notation extensively in the coming proofs.

*Proof.* We use the fact that  $\log z \leq z - 1$ , then invoke Tonelli's theorem.

$$\begin{aligned}
\mathbb{E} \mu q_X \log \frac{q_X}{r_X} &= \mathbb{E} \mu q_X \left[ \frac{r_X}{q_X} - 1 - \log \frac{r_X}{q_X} \right] \\
&= \mu \mathbb{E} q_X \left[ \frac{r_X}{q_X} - 1 - \log \frac{r_X}{q_X} \right] \\
&= \mu \mathbb{E} r_X - \mu \mathbb{E} q_X - \mu \mathbb{E} q_X \log \frac{r_X}{q_X} \\
&= \underbrace{\mathbb{E} \mu r_X}_1 - \underbrace{\mathbb{E} \mu q_X}_1 + \mu \mathbb{E} q_X \log \frac{q_X}{r_X}
\end{aligned}$$

□

*Proof of Theorem 2.1.* Lemma 3.1 justifies changing the order of integration.

$$\begin{aligned}
\mathbb{E} D(Q_X \| R) &= \mathbb{E} \mu q_X \log \frac{q_X}{r} \\
&= \mathbb{E} Q_X \log \frac{\bar{q}_P}{r} + \mathbb{E} Q_X \log \frac{q_X}{\bar{q}_P} \\
&= \mu \underbrace{\mathbb{E} q_X}_{\bar{q}_P} \log \frac{\bar{q}_P}{r} + \mathbb{E} D(Q_X \| \bar{Q}_P)
\end{aligned}$$

75

□

**Lemma 3.2.** *Let  $\{q_x : x \in \mathcal{X}\}$  be a family of probability densities with respect to a  $\sigma$ -finite measure  $\mu$ , and suppose that  $(x, y) \mapsto q_x(y)$  is product measurable. Let  $X \sim P$  be an  $\mathcal{X}$ -valued random element. If  $\mu e^{\mathbb{E} \log q_X} > 0$ , then for any probability measure  $R$  on  $\mathcal{Y}$ ,*

$$\mathbb{E} D(R \| Q_X) = D(R \| \tilde{Q}_P) + \log \frac{1}{\mu e^{\mathbb{E} \log q_X}}.$$

*Proof.* Making use of the central trick from the explanations of the mean field



approximation algorithm (e.g. [9]), we have

$$\begin{aligned}
\mathbb{E}D(R||Q_X) &= \mathbb{E}R \log \frac{r}{q_X} \\
&= R \mathbb{E} \log \frac{r}{q_X} \\
&= R[\log r - \mathbb{E} \log q_X] \\
&= R[\log r - \log e^{\mathbb{E} \log q_X}] \\
&= R \log \frac{r}{e^{\mathbb{E} \log q_X}} \\
&= R \log \frac{r}{e^{\mathbb{E} \log q_X} / \mu e^{\mathbb{E} \log q_X}} + \log \frac{1}{\mu e^{\mathbb{E} \log q_X}}.
\end{aligned}$$

Again, Lemma [3.1] justifies the order interchange.  $\square$

**Lemma 3.3.** *Let  $\{q_x : x \in \mathcal{X}\}$  be a family of probability densities with respect to a  $\sigma$ -finite measure  $\mu$ , and suppose that  $(x, y) \mapsto q_x(y)$  is product measurable. Let  $X \sim P$  be an  $\mathcal{X}$ -valued random element. If  $\mu e^{\mathbb{E} \log q_X} > 0$ , then*

$$\mathbb{E}D(\tilde{Q}_P||Q_X) = \log \frac{1}{\mu e^{\mathbb{E} \log q_X}}.$$

*Proof.* Use  $\tilde{Q}_P$  as  $R$  in Lemma [3.2].  $\square$

*Proof of Theorem [2.2].* Combine Lemmas [3.2] and [3.3].  $\square$

*Proof of Theorem [1.4].* We will write  $f(X, \cdot)$  as  $f_X$ . The key is Lemma [3.3].

$$\begin{aligned}
\log \mu e^{\mathbb{E} f_X} &= \log \mu e^{\mathbb{E} \log [e^{f_X} / \mu \exp f_X]} + \mathbb{E} \log \mu e^{f_X} \\
&= -\mathbb{E}D(\tilde{Q}_P||Q_X) + \mathbb{E} \log \mu e^{f_X}
\end{aligned}$$

if the geometric mixture is well-defined.

80 Next, assume that the geometric mixture is not well-defined; in other words,  $\mu e^{\mathbb{E} \log (e^{f_X} / \mu e^{f_X})} = 0$ . Because the integrand is non-negative, the integral can only be zero if the integrand is zero  $\mu$ -almost everywhere. This requires the exponent, which simplifies to  $\mathbb{E}[f_X - \log \mu e^{f_X}]$ , to be  $-\infty$  almost everywhere. Assume that there exists a non-negligible set for which  $\mathbb{E} f_X > -\infty$ . Then on  
85 that set,  $\mathbb{E}[f_X - \log \mu e^{f_X}]$  can only be  $-\infty$  if  $\mathbb{E} \log \mu e^{f_X}$  is  $\infty$ . Furthermore, the contribution of that non-negligible set ensures that  $\log \mu e^{\mathbb{E} f_X}$  is also strictly

greater than  $-\infty$ , which tells us that the two sides of the proposed identity are both  $\infty$ .

In the one remaining case, the geometric mixture does not exist and  $\mathbb{E}f_X =$   
 90  $-\infty$  almost everywhere. These imply that  $\mathbb{V}Q_X = \infty$  and  $\log \mu e^{\mathbb{E}f_X} = -\infty$ ,  
 respectively. The theorem specifies that  $\mathbb{E} \log \mu e^{f_X} > -\infty$ , so again the identity  
 reduces to  $\infty = \infty$ .

An interesting observation is implicit in the above proof:  $\mathbb{E} \log \mu e^{f_X} = -\infty$   
 is only possible if  $\mathbb{E}f_X = -\infty$  almost everywhere.

95 A closely related derivation in [2, Sec 4] was instructive; the accompanying  
 discussion in that paper provides another interpretation of the quantities in-  
 volved in Hölder's identity. Barron's approach is explained in Section 1 in the  
 supplementary material.  $\square$

*Proof of Theorem 1.1.* When its conditions are met, Hölder's identity (Theo-  
 100 rem 1.5) implies the desired inequality result by non-negativity of  $\tilde{\mathbb{V}}$ .

For a variant that does not require  $\sigma$ -finiteness, see Section 3 in the supple-  
 mentary material.  $\square$

*Proof of Theorem 1.6.* Define the product measurable function  $h_x(y)$  with  $x$   
 taking values in  $\mathcal{X} = \{1, 2\}$  with  $h_1(y) = f(y)$  and  $h_2(y) = g(y)$ . By the defi-  
 105 nition of unnormalized Renyi divergence,  $\tilde{\mathbb{V}}$  of the random distribution is equal  
 to  $D_\lambda(Q||R)$  according to Lemma 3.3. Therefore, the desired result is a direct  
 consequence of Hölder's identity, at least when  $\mu$  is  $\sigma$ -finite. However, we de-  
 liberately omitted the  $\sigma$ -finiteness requirement. In fact, the reason we required  
 $\sigma$ -finiteness in previous Lemmas and Theorems was to justify interchanges in  
 110 the order of integration. When one of the integrals concentrates on a finite set  
 of atoms, then interchange is always valid by linearity of integration. Indeed,  
 when  $\mathcal{X}$  is finite, the Lemmas and Theorems of this paper are valid without  
 the condition that  $\mu$  is  $\sigma$ -finite. Alternatively, the sum of any finite collection  
 of probability measures is itself a finite dominating measure for each of their  
 115 densities.  $\square$

**Lemma 3.4.** *Let  $\{q_x : x \in \mathcal{X}\}$  be a family of probability densities with respect to a  $\sigma$ -finite measure  $\mu$ , and suppose that  $(x, y) \mapsto q_x(y)$  is product measurable. Let  $X$  be an  $\mathcal{X}$ -valued random element. If  $\mu e^{\mathbb{E} \log q_X} = 0$ , then for any probability measure  $R$ ,  $\mathbb{E}D(R||Q_X) = \infty$ .*

120 *Proof.* The integrand of  $\mu e^{\mathbb{E} \log q_X}$  is non-negative, so the integral being zero implies that  $\mathbb{E} \log q_X = -\infty$   $\mu$ -almost everywhere. Since  $\mu$  dominates  $R$ , the condition also holds  $R$ -almost everywhere.

By Lemma [3.1](#),

$$\begin{aligned} \mathbb{E}D(R||Q_X) &= R \mathbb{E} \log \frac{r}{q_X} \\ &= R [\log r - \mathbb{E} \log q_X]. \end{aligned}$$

Our previous observation tells us that  $\mu e^{\mathbb{E} \log q_X} = 0$  implies that the integrand  $\log r - \mathbb{E} \log q_X$  equals  $\infty$  with  $R$ -probability 1, so  $\mathbb{E}D(R||Q_X) = \infty$ .  $\square$

## 125 Acknowledgment

Conversations with Andrew Barron about [\[2\]](#) were instrumental in leading us to the insights of this paper.

## References

- 130 [1] Charalambos D. Aliprantis and Kim C. Border. *Infinite Dimensional Analysis: A Hitchhiker's Guide*. Springer, 3rd edition, 2006.
- [2] Andrew R. Barron. The Exponential Convergence of Posterior Probabilities with Implications for Bayes Estimators of Density Functions. Report 7, University of Illinois, 1988.
- 135 [3] Wei Chen, Longbin Jia, and Yong Jiao. Hölder's inequalities involving the infinite product and their applications in martingale spaces. *Analysis Mathematica*, 42(2):121–141, 2016.

- [4] Imre Csiszár and František Matúš. Information Projections Revisited. *IEEE Transactions on Information Theory*, 49(6):1474–1490, 2003.
- [5] Nelson Dunford and Jacob T. Schwartz. *Linear Operators, Part 1: General Theory*. Interscience Publishers, New York, 1958.
- 140 [6] Peter D. Grünwald. *The Minimum Description Length Principle*. MIT Press, 2007.
- [7] David Haussler and Manfred Opper. Mutual information, metric entropy and cumulative relative entropy risk. *The Annals of Statistics*, pages 2451–  
145 2492, 1997.
- [8] G. L. Karakostas. An extension of Hölder’s inequality and some results on infinite products. *Indian Journal of Mathematics*, 50:303–307, 2008.
- [9] J. T. Ormerod and M. P. Wand. Explaining Variational Approximations. *The American Statistician*, 64(2):140–153, 2010.
- 150 [10] Michael D. Perlman. Jensen’s Inequality for a Convex Vector-Valued Function on an Infinite-Dimensional Space. *Journal of Multivariate Analysis*, 4(1):52–65, 1974.
- [11] B. J. Pettis. On Integration in Vector Spaces. *Transactions of the American Mathematical Society*, 44(2):277–304, 1938.
- 155 [12] David Pfau. A Generalized Bias-Variance Decomposition for Bregman Divergences. 2013.
- [13] David Pollard. *A User’s Guide to Measure Theoretic Probability*. Cambridge University Press, 2002.
- [14] Kurt Symanzik. Proof and Refinements of an Inequality of Feynman. *Journal of Mathematical Physics*, 6(7):1155–1156, 1965.
- 160 [15] Matus Telgarsky and Sanjoy Dasgupta. Agglomerative Bregman Clustering. In *International Conference on International Conference on Machine Learning*, pages 1011–1018. Omnipress.

- [16] Flemming Topsøe. Basic Concepts, Identities and Inequalities - the Toolkit  
of Information Theory. *Entropy*, 3(3):162–190, 2001.
- [17] Raymond Veldhuis. The Centroid of the Symmetrical Kullback-Leibler  
Distance. *IEEE Signal Processing Letters*, 9(3):96–99, 2002.

# Supplementary material

## 125 1. A Bayesian approach

Let us now establish a connection between our approach and that of [2] Sec 4]. Suppose  $q_\theta(x) = q(x|\theta)$ ,  $\theta \in \Theta$  is a family of densities dominated by a measure  $\lambda$  and  $\mu(d\theta)$  is a prior probability distribution on  $\Theta$ , reflecting our prior knowledge toward  $\Theta$ .

130 Then given  $n$  data points  $X^n$  drawn independently from the product distribution  $P^n(X^n) = \prod^n P(X_i)$  with density  $p^n(X^n) = \prod^n p(X_i)$  ( $p$  is also dominated by  $\lambda$ ), the Bayes estimator is the posterior mean density,

$$\hat{p}_n(x) = \int q(x|\theta)\mu_n(d\theta|X^n),$$

which according to Bayes' rule is the predictive density

$$\hat{p}_n(x) = \frac{m^{n+1}(X^n, x)}{m^n(X^n)},$$

where

$$m^n(x_1, \dots, x_n) = \int \left( \prod_{i=1}^n q_\theta(x_i) \right) \mu(d\theta).$$

Define a new distribution on  $\Theta$  via

$$\mu_n^*(d\theta) = \frac{e^{-nD_n(\theta)} \mu(d\theta)}{c_n},$$

where  $D_n(\theta) = (1/n)D(p^n||q^n(\cdot|\theta))$  and  $c_n = \int e^{-nD_n(\theta)} \mu(d\theta)$ . Let  $L_n^*$  be the product distribution for  $X^n$  and  $\theta$  defined by

$$L_n^*(d\theta, dx^n) = P^n(dx^n)\mu_n^*(d\theta).$$

Then  $L_n^*$  is an approximation (or surrogate distribution) to the Bayesian joint law for  $X^n$  and  $\theta$ , mainly

$$L_n^{\text{Bayes}}(dx^n, d\theta) = Q^n(dx^n | \theta)\mu(d\theta).$$

The Bayesian law  $L_n^{\text{Bayes}}$  has a joint density function  $q^n(x^n | \theta)$  with respect to the product measure  $\lambda^n \times \mu$ ; on the other hand, the approximation  $L_n^*$  has a density function  $p^n(x^n)e^{-nD_n(\theta)}/c_n$ . Thus

$$D(L_n^* || L_n^{\text{Bayes}}) = \mathbb{E} \log \frac{p^n(X^n)e^{-nD_n(\theta)}/c_n}{q^n(X^n|\theta)}.$$

[2] Sec 4, p. 21] then argues using Fubini's theorem (integrating first with respect to  $P^n$  and then with respect to  $\mu_n^*$ ) that

$$\begin{aligned} D(L_n^* || L_n^{\text{Bayes}}) &= \mathbb{E} \log \frac{p^n(X^n)e^{-nD_n(\theta)}/c_n}{q^n(X^n|\theta)} \\ &= \mathbb{E} \left[ \log \frac{p^n(X^n)}{q^n(X^n|\theta)} - nD_n(\theta) + \log(1/c_n) \right] \\ &= nD_n(\theta) - nD_n(\theta) + \log(1/c_n) \\ &= \log(1/c_n) \\ &= -\log \int e^{-nD_n(\theta)} \mu(d\theta). \end{aligned}$$

Continuing, by the chain rule for relative entropy,

$$D(L_n^* || L_n^{\text{Bayes}}) = D(p^n || m^n) + \mathbb{E}(D(\mu_n^* || \mu(\cdot | X^n))).$$

Thus, by nonnegativity of relative entropy,

$$D(p^n || m^n) \leq D(L_n^* || L_n^{\text{Bayes}}) = -\log \int e^{-nD_n(\theta)} \mu(d\theta).$$

For the case that  $n = 1$ , we have

$$D(p^n || m^n) = \int p(x) \log \frac{p(x)}{\int q_\theta(x) \mu(d\theta)} \lambda(dx),$$

and

$$-\log \int e^{-nD_n(\theta)} \mu(d\theta) = -\log \int e^{-\int p(x) \log \frac{p(x)}{q_\theta(x)} \lambda(dx)} \mu(d\theta)$$

Thus, we find that

$$\int p(x) \log \frac{p(x)}{\int q_\theta(x) \mu(d\theta)} \lambda(dx) \leq -\log \int e^{-\int p(x) \log \frac{p(x)}{q_\theta(x)} \lambda(dx)} \mu(d\theta).$$

Rearranging and then exponentiating, we find that

$$\int e^{-\int p(x) \log \frac{p(x)}{q_\theta(x)} \lambda(dx)} \mu(d\theta) \leq e^{-\int p(x) \log \frac{p(x)}{\int q_\theta(x) \mu(d\theta)} \lambda(dx)}.$$

If  $\gamma(dx) = p\lambda(dx)$ , we have

$$\int e^{\int \log \frac{q_\theta(x)}{p(x)} \gamma(dx)} \mu(d\theta) \leq e^{\int \log \left( \int \frac{q_\theta(x)}{p(x)} \mu(d\theta) \right) \gamma(dx)}. \quad (3)$$

Note that (3) is similar to (and in fact implied by) (1.1), provided  $f$  is scaled in such a way that it may be written as the ratio of two densities.

## 2. Dunford and Schwartz exercise

145 Theorem 2.1 is a modified version of [5, VI.11 Ex 36]. We follow the route of proof suggested by the authors of [5] in order to point out that their approach relies on  $\sigma$ -finiteness.

**Theorem 2.1** ([5], VI.11.36). *Suppose  $\mu$  and  $\mu_1$  are positive measures on spaces  $\mathcal{X}$  and  $\mathcal{Y}$ . Assume that  $\mu\mathcal{X} = 1$  and  $\mu_1$  is  $\sigma$ -finite. Then for any  $\mu \times \mu_1$ -integrable function  $f : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}^+$ ,*

$$\int \exp \left( \int \log f(x, y) \mu(dx) \right) \mu_1(dy) \leq \exp \left( \int \log \left( \int f(x, y) \mu_1(dy) \right) \mu(dx) \right). \quad (4)$$

To prove the theorem we need to introduce some notation. Following convention, we define the  $L^p$  norm of a measurable function on  $\mathcal{X}$  to be

$$|g|_{\mu, p} = \left( \int |g|^p \mu(dx) \right)^{1/p},$$

where  $\mu$  is the probability measure as in Theorem 2.1. Throughout the remainder of this section, we will omit the dependence on the measure  $\mu$  and write  $|\cdot|_p$  to denote the  $L^p$  norm. The following lemma characterizes the behavior of the  $L^p$  norm as  $p \rightarrow 0$  and is crucial in obtaining Theorem 2.1.

150 **Lemma 2.2** ([5], VI.11.32). *For all  $\mu$ -measurable function  $g : \mathcal{X} \rightarrow \mathbb{R}^+$ ,  $\lim_{p \rightarrow 0} |g|_p$  exists and*

$$\lim_{p \rightarrow 0} |g|_p = \exp \left( \int \log g(x) \mu(dx) \right) =: |g|_0.$$



*Proof.*

$$\log |g|_p = \frac{1}{p} \log \int g(x)^p \mu(dx) \leq \int \frac{g(x)^p - 1}{p} \mu(dx).$$

The integrand converges pointwise to  $\log g$ . To show the integral converges, notice that

$$\left| \frac{g^p - 1}{p} \right| \leq (g - 1)\mathbb{I}\{g > 1\} - \log g \mathbb{I}\{g \leq 1\}.$$

If  $\log g$  is  $\mu$ -integrable, we can apply dominated convergence to conclude that  $\lim_{p \rightarrow 0} |g|_p = |g|_0$ . If  $\int \log g d\mu = \infty$ , take a sequence of truncated  $g$  and apply monotone convergence to pass the statement to the limit.  $\square$

**Lemma 2.3** ([5], VI.11.13, Generalized Minkowski inequality). *Take  $\sigma$ -finite measures  $\mu, \mu_1$  on measure spaces  $\mathcal{X}$  and  $\mathcal{Y}$ . Let  $f$  be a  $\mu \times \mu_1$ -integrable non-negative function on  $\mathcal{X} \times \mathcal{Y}$ . Then, for  $r > 1$ ,*

$$\left( \int \left( \int f(x, y) \mu_1(dy) \right)^r \mu(dx) \right)^{1/r} \leq \int \left( \int f(x, y)^r \mu(dx) \right)^{1/r} \mu_1(dy). \quad (5)$$

*Proof.* Write the  $r$ 'th power of the left-hand side of (5) as

$$\int \left( \int f d\mu_1 \right) \left( \int f d\mu_1 \right)^{r-1} d\mu = \int \left( \int f(x, \tilde{y}) \mu_1(d\tilde{y}) \right) \left( \int f(x, y) \mu_1(dy) \right)^{r-1} \mu(dx).$$

Introducing the auxiliary variable  $\tilde{y}$  allows us to move the integral over  $\tilde{y}$  before the integral over  $x$ . Note that this is also the place where we have to assume  $\sigma$ -finiteness of  $\mu$  and  $\mu_1$  to justify the change of order of integration. Apply Hölder's inequality to further bound the above by

$$\int \left( \int f(x, \tilde{y})^r \mu(dx) \right)^{1/r} \mu_1(d\tilde{y}) \left( \int \left( \int f(x, y) \mu_1(dy) \right)^{(r-1)s} \mu(dx) \right)^{1/s},$$

155 where  $s > 1$  is such that  $\frac{1}{r} + \frac{1}{s} = 1$ . Note that  $(r-1)s = r$ . We have proved that the  $r$ 'th power of the LHS of (5) is bounded by the RHS of (5) times the  $r/s$ 'th power of the LHS. Rearrange the terms to obtain (5).  $\square$

*Proof of Theorem 2.1.* With the two auxiliary lemmas the proof of Theorem 2.1 is straightforward. By Lemma 2.2, we can write the left-hand side of (4) as

$$\int |f(\cdot, y)|_0 \mu_1(dy) = \int \lim_{p \rightarrow 0} |f(\cdot, y)|_p \mu_1(dy).$$

Take a sequence  $p_n \rightarrow 0$  to rewrite the expression above as

$$\int \liminf_{n \rightarrow \infty} |f(\cdot, y)|_{p_n} \mu_1(dy) \leq \liminf_{n \rightarrow \infty} \int |f(\cdot, y)|_{p_n} \mu_1(dy),$$

where the inequality is obtained by Fatou's lemma. Take  $r = 1/p_n$  in Lemma [2.3](#), and we have

$$\int |f(\cdot, y)|_{p_n} \mu_1(dy) \leq \left| \int f(\cdot, y) \mu_1(dy) \right|_{p_n}.$$

Conclude that

$$\begin{aligned} & \int \exp \left( \int \log f(x, y) \mu(dx) \right) \mu_1(dy) \\ & \leq \liminf_{n \rightarrow \infty} \left| \int f(\cdot, y) \mu_1(dy) \right|_{p_n} \\ & = \left| \int f(\cdot, y) \mu_1(dy) \right|_0 \\ & = \exp \left( \int \log \left( \int f(x, y) \mu_1(dy) \right) \mu(dx) \right). \end{aligned}$$

□

### 3. Jensen's inequality in infinite-dimensional spaces

160 This section shows how to use Jensen's inequality to prove a variant of Corollary [1.3](#) without requiring a  $\sigma$ -finite measure. A first step is to clarify what we mean by the *expectation* of a random mapping to a topological vector space. Most convenient for us is the *Pettis expectation* of the random mapping considered as a random vector taking values in the function space. A Pettis  
165 expectation is a special case of the *Pettis integral*, which we now define.

Let  $\mathcal{V}$  be a real topological vector space (rTVS), and let  $F$  be a function from a measure space to  $\mathcal{V}$ . We will say that a vector  $v \in \mathcal{V}$  is a *Pettis  $\mu$ -integral* (or simply *Pettis integral* if the measure is clear from context) of  $F$  if for every continuous linear functional  $l \in \mathcal{V}'$ ,

$$\mu l(F) = l(v) \tag{6}$$

where  $\mu$  on the left side refers to the Lebesgue integral with respect to  $\mu$ . If such a vector exists, we say that  $F$  is *Pettis integrable*.<sup>10</sup> We will use the symbol  $\mathbb{E}$  for a Pettis integral with respect to a probability measure, which we call a *Pettis expectation*.

170 A key aspect of Jensen's inequality is the commutation of expectations with continuous affine functionals. This commutation is a simple consequence of basic facts about Pettis integrals as we review in this section. Straight-forward proofs are given along the way.

175 While the Pettis integral is *defined* by commutation with continuous linear *functionals*, it turns out that it also commutes with all continuous linear *operators*.

**Theorem 3.1.** *Let  $\mathcal{U}$  and  $\mathcal{V}$  be  $rTVS$ s. Suppose  $F$  is a Pettis  $\mu$ -integrable  $\mathcal{U}$ -valued function and  $T$  is a continuous linear operator from  $\mathcal{U}$  to  $\mathcal{V}$ . Then  $T\mu F$  is the Pettis integral of  $T \circ F$ .*

*Proof.* This argument originally appeared in [11, Thm 2.2], where Pettis integrals were introduced. Let  $l$  be a continuous linear functional on  $\mathcal{V}$ . Then  $l \circ T$  is a continuous linear functional on  $\mathcal{U}$ , so we know that it commutes with the integral.

$$\begin{aligned} l(T\mu F) &= (l \circ T)(\mu F) \\ &= \mu(l \circ T)(F) \\ &= \mu l(T \circ F) \end{aligned}$$

180 We've seen that for any  $l \in \mathcal{V}'$ , the integral of  $l(T \circ F)$  is equal to  $l(T\mu F)$ , meaning that  $T\mu F$  is by definition a Pettis integral of  $T \circ F$ .  $\square$

**Lemma 3.2.** *Let  $\mathcal{V}$  be an  $rTVS$ . For any measure space, the Pettis integral is a linear operator from the space of Pettis integrable  $\mathcal{V}$ -valued functions to  $\mathcal{V}$ .*

---

<sup>10</sup>Definitions of the Pettis integral vary somewhat; ours agrees with [10].

*Proof.* Suppose  $F$  and  $G$  have Pettis integrals  $v_F$  and  $v_G$ . For an arbitrary coefficient  $r$ , we need to verify that the Pettis integral of  $rF + G$  is  $rv_F + v_G$ . For any continuous linear functional  $l$ ,

$$\begin{aligned}\mu l(rF + G) &= \mu [rl(F) + l(G)] \\ &= r\mu l(F) + \mu l(G) \\ &= rl(v_F) + l(v_G) \\ &= l(rv_F + v_G)\end{aligned}$$

where we used the linearity of the Lebesgue integral. □

185 **Lemma 3.3.** *If  $X : \Omega \rightarrow \mathcal{V}$  maps every  $\omega \in \Omega$  to the same vector  $v \in \mathcal{V}$ , then its Pettis expectation is  $v$ .*

*Proof.* Let  $l$  be any continuous linear functional on  $\mathcal{V}$ . We use Lemma 3.2 to pass a scalar through the expectation.

$$\begin{aligned}\mathbb{E}l(X) &= \mathbb{E}l(v) \\ &= l(v)\mathbb{E}\mathbb{1}_\Omega \\ &= l(v)\end{aligned}$$

□

Based on Lemmas 3.2 and 3.3 along with Theorem 3.1, we can conclude that Pettis expectations commute with continuous *affine* operators too.

190 **Corollary 3.4.** *If  $a$  is a continuous affine operator on an  $rTVS$   $\mathcal{V}$ , then any Pettis integrable  $\mathcal{V}$ -valued random vector  $X$  has  $\mathbb{E}a(X) = a(\mathbb{E}X)$ .*

Jensen's inequality works if the convex function has a tangent continuous affine functional (i.e. a *subdifferential*) at its Pettis expectation. Let  $f$  be a convex function and  $a$  be a satisfactory subdifferential. By the definition of

Pettis expectation and the increasing property of integrals,

$$\begin{aligned} f(\mathbb{E}X) &= a(\mathbb{E}X) \\ &= \mathbb{E}a(X) \\ &\leq \mathbb{E}f(X). \end{aligned}$$

**Theorem 3.5.** *Let  $\mathcal{X}$  and  $\mathcal{Y}$  be measurable spaces, and let each real-valued function  $\{f_x : x \in \mathcal{X}\}$  be  $\mu$ -integrable. Let  $X$  be an  $\mathcal{X}$ -valued random element, and suppose the random vector  $f_X$  has Pettis expectation  $\mathbb{E}f_X$ . Then*

$$\log \int e^{\mathbb{E}f_X(y)} d\mu(y) \leq \mathbb{E} \log \int e^{f_X(y)} d\mu(y).$$

*Proof of Theorem 1.1.* We will use De Finetti notation for the  $\mu$ -integral in this proof.

The Bronsted-Rockafellar theorem [II, Thm 7.60] states that a lower semicontinuous proper convex function on a Banach space is subdifferentiable in a dense subset of the part of its domain in which it takes finite values.  $L^1(\mu)$  is a Banach space.  $f \mapsto \log \mu e^f$  is convex by Hölder's inequality, and lower semicontinuity follows directly from Fatou's lemma

$$\liminf_{f \rightarrow f_0} \log \mu e^f \geq \log \mu e^{f_0}.$$

The term *proper* means that the convex function is finite on a non-empty domain and never equal to negative infinity. We quickly consider the non-proper cases  
195 first.  $\log \mu f_x$  can only be  $-\infty$  if  $f_x$  is  $-\infty$  almost everywhere, but that is impossible since each function is integrable. If  $\log \mu e^{f_x}$  is  $\infty$  on the entire support of  $X$ , then the right side of the desired inequality is infinite, so it is trivially satisfied.

200 Now we consider the proper case. Let  $\mathcal{G} \subseteq L^1(\mu)$  be the [dense] set where  $f \mapsto \log \mu e^f$  has a subdifferential. We define  $G$  to be the set of translated functions  $\mathcal{G} - \mathbb{E}f_X$ . Let  $\mathcal{N}$  denote the negative cone of  $L^1(\mu)$ , that is, the set of non-positive functions.  $\mathcal{N}$  is closed, so  $G$  is dense in this cone as well. Let  $(g_n)$  be a sequence of functions in  $G \cap \mathcal{N}$  that converges to zero in  $L^1(\mu)$ . Without loss

205 of generality, we can assume that  $(g_n)$  also converges to zero point-wise (because every sequence that converges in  $L^1$  has a point-wise convergent subsequence).

For every  $n \in \mathbb{N}$ , the random vector  $g_n + f_X$  has expectation  $g_n + \mathbb{E}f_X$  which is in  $\mathbb{G}$ ; Jensen's inequality can be invoked.

$$\begin{aligned} \log \mu e^{g_n + \mathbb{E}f_X} &= \log \mu e^{\mathbb{E}(g_n + f_X)} \\ &\leq \mathbb{E} \log \mu e^{g_n + f_X} \end{aligned} \tag{7}$$

Take the limit superior of each side with respect to  $n$ . On the left, note that  $e^{g_n + \mathbb{E}f_X} \leq e^{\mathbb{E}f_X}$  which is integrable, so dominated convergence applies.

$$\begin{aligned} \log \limsup \mu e^{g_n + \mathbb{E}f_X} &= \log \mu e^{\limsup g_n + \mathbb{E}f_X} \\ &= \log \mu e^{\mathbb{E}f_X} \end{aligned}$$

On the right of (7), we use two applications of the reverse Fatou lemma.

$$\begin{aligned} \limsup \mathbb{E} \log \mu e^{g_n + f_X} &\leq \mathbb{E} \log \limsup \mu e^{g_n + f_X} \\ &\leq \mathbb{E} \log \mu e^{\limsup g_n + f_X} \\ &= \mathbb{E} \log \mu e^{f_X} \end{aligned}$$

□