Statistics and Probability Letters 148 (2019) 150–154

Contents lists available at ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro

Hölder's identity

W.D. Brinda [a], Jason M. Klusowski [b], Dana Yang [a,*]

[a] Department of Statistics and Data Science, Yale University, New Haven, CT, 06511, USA,
[b] Department of Statistics and Biostatistics, Rutgers University — New Brunswick, Piscataway, NJ, 08901, USA

ARTICLE INFO

Article history:
Received 1 August 2018
Received in revised form 22 December 2018
Accepted 2 January 2019
Available online 11 January 2019

Keywords:
Hölder's inequality
Measure theory
Information theory

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.

1. Introduction

Hölder's inequality is most commonly written

$$\int |f(y)g(y)|d\mu(y) \leq \|f\|_p \|g\|_q \tag{1}$$

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 $\alpha \in [0, 1]$,

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

In other words, *the integral of the point-wise geometric average of two functions is bounded by the geometric average of their integrals*.

In Section 2, we point out that (2) holds for arbitrary geometric expectations as long as μ is σ-finite. We also clarify a few points of confusion that have arisen regarding this more general inequality; a number of papers have stated the result without σ-finiteness or purported to prove it with Jensen's inequality. The section concludes with *Hölder's identity* quantifying the ratio of the two sides of Hölder's inequality. Next, Section 3 describes the *compensation identities*, two decompositions of expected relative entropy between a random probability measure and a fixed probability measure. These identities both resemble the bias–variance decomposition, and one of the variance-like terms that arises is exactly the natural logarithm of the ratio between the two sides of Hölder's inequality. Sections 2 and 3 are adapted from appendix sections of W. D. Brinda's doctoral thesis, and we will point to it for proofs.

Additional discussion is provided in a supplement to this paper. Section A recalls the context of the original paper that presented a version of Hölder's identity which arose in an analysis of the relative entropy from the Bayesian posterior

* Corresponding author.
E-mail addresses: william.brinda@yale.edu (W.D. Brinda), jason.klusowski@rutgers.edu (J.M. Klusowski), xiaoqian.yang@yale.edu (D. Yang).

https://doi.org/10.1016/j.spl.2019.01.008
0167-7152/

distribution to a particular approximation of that distribution. Section B works out the proof of the generalized Hölder's inequality that is indicated by Dunford and Schwartz (1958) to verify that it requires σ-finiteness of μ. Finally, in Section C we use Jensen's inequality to give a general version of Hölder's inequality that does not require σ-finiteness, although it does use an integrability condition that was not needed in our σ-finite version.

2. Generality of Hölder's inequality

Inequality (2) holds for *arbitrary* geometric expectations over a random element indexing functions.

Theorem 2.1 (*Hölder's Inequality*)**.** *Let $\mathcal{X}$ and $\mathcal{Y}$ be measurable spaces, and let $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^+$ be product measurable. For any σ-finite measure μ on $\mathcal{Y}$ and any $\mathcal{X}$-valued random element X,*

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

This fact is an immediate consequence of Hölder's identity (Corollary 2.5) presented later in this section.

Inequalities (1) and (2) represent the two-point distribution version of Theorem 2.1. 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 2.2. *Let $\mathcal{X}$ and $\mathcal{Y}$ be measurable spaces, and let $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^+$ be product measurable. For any σ-finite measure μ on $\mathcal{Y}$ and finite measure γ on $\mathcal{X}$,*

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

Using e^f as the function in Theorem 2.1, and taking the log of both sides gives us an equivalent inequality that is also worth stating.

Corollary 2.3. *Let $\mathcal{X}$ and $\mathcal{Y}$ be measurable spaces, and let $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ be product measurable. For any σ-finite measure μ on $\mathcal{Y}$ and any $\mathcal{X}$-valued random element X,*

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

The fact that Hölder's inequality holds in this generality is perhaps not widely known. For example, Karakostas (2008) proved an extension of Hölder's inequality to *countable* products assuming μ is σ-finite; that result was improved by Chen et al. (2016, Thm 2.11). The inequalities they present are readily subsumed by Corollary 2.2 by letting γ concentrate on a countable set.

Haussler and Opper (1997, Lemma 1) state our Corollary 2.3, 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 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 no continuous affine functional tangent to it. For a more concrete example, see Perlman (1974, Introduction).) If adequate care is taken, the logic of Jensen's inequality can be applied to this problem as we show in Section C; there, we prove a variant of Hölder's identity that does not require σ-finiteness.

Haussler and Opper (1997) reference Symanzik (1965) where the inequality in our Theorem 2.1 is stated and called *generalized Hölder's inequality*; he points to the classic text (Dunford and Schwartz, 1958, VI.11 Ex 36) where it is left as an exercise. Although that exercise does not say to assume σ-finiteness, the proof they hint at does require it — see Section B. For σ-finite measures, at least, the proof can follow a different route from the one they hint at. We establish an identity that has an information-theoretic interpretation involving a non-negative "variance" functional $\widetilde{\mathbb{V}}$ for random probability measures which will be defined and explained in Section 3. For now, suffice it to say that $\widetilde{\mathbb{V}}$ represents an expected relative entropy.

Theorem 2.4 (*See Brinda (2018, Thm B.0.8)*)**.** *Let $\mathcal{X}$ and $\mathcal{Y}$ be measurable spaces, and let $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ be product measurable. Let μ be a σ-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*

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

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

Corollary 2.5 (*Hölder's Identity*)**.** *Let $\mathcal{X}$ and $\mathcal{Y}$ be measurable spaces, and let $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^+$ be product measurable. Let μ be a σ-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*

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

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

An interpretation of $\widetilde{\mathbb{V}}Q_X$ 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, $\widetilde{\mathbb{V}}Q_X$ is an *unnormalized Rényi divergence* D_λ between the two possible distributions, as defined in Section 3.

Theorem 2.6. *Let $\mathcal{Y}$ be a measurable space, and let $f : \mathcal{Y} \to \mathbb{R}^+$ and $g : \mathcal{Y} \to \mathbb{R}^+$ have finite positive μ-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 $\frac{g(y)}{\int g(y)\mu(y)}$ with respect to μ.

3. The compensation identities

Theorem 3.1, called the *compensation identity* by Topsøe (2001, Thm 9.1), conveniently decomposes the expected relative entropy from a random probability measure to a fixed probability measure.[1]

Theorem 3.1 (*Compensation Identity (see Brinda (2018, Thm A.0.1))*)**.**
Let $\{q_x : x \in \mathcal{X}\}$ be a family of probability densities with respect to a σ-finite measure μ, and suppose that $(x, y) \mapsto q_x(y)$ is product measurable. Let $X \sim P$ be an $\mathcal{X}$-valued random element. For any probability measure R on $\mathcal{Y}$,

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

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

$$\bar{q}_P(y) = \int q_x(y)P(dx).$$

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*.[2] We define the *P-geometric mixture* of $\{q_x\}$ to be the probability measure with density

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

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

$$\int e^{\mathbb{E}\log q_X(y)}d\mu(y) \le \mathbb{E}\int e^{\log q_X(y)}d\mu(y) = 1.$$

This integral can be zero, however, in which case the geometric-mixture is not well-defined.[3]

Theorem 3.2 (*Reverse Compensation Identity (see Brinda (2018, Thm A.0.2))*)**.** *Let $\{q_x : x \in \mathcal{X}\}$ be a family of probability densities with respect to a σ-finite measure μ, 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 $\mathcal{Y}$,*

$$\mathbb{E}D(R \parallel Q_X) = D(R \parallel \widetilde{Q}_P) + \mathbb{E}D(\widetilde{Q}_P \parallel Q_X)$$

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

A two-point distribution version of Theorem 3.2 is implied by Csiszár and Matúš (2003, Eq (3) with (4)) and similarly for any finite set of discrete distributions by Veldhuis (2002, Equation (9)).

---

1 In Theorem 3.1 and throughout the remainder of this paper, lower-case and upper-case letters implicitly pair probability measures with their densities.

2 What we call a "geometric mixture" is sometimes called a "log mixture" or "log-convex mixture", for instance by Grünwald (2007, Sec 19.6).

3 An example of such a pathological case is when q_X has positive probabilities on two densities that are mutually singular.

Theorems 3.1 and 3.2 are perfectly analogous to the bias–variance decomposition for Hilbert-space-valued random vectors.[4] 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.[5] We suggest a notation that makes use of this intuition:

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

and[6]

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

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, $\widetilde{\mathbb{V}}Q_X$ 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 $\widetilde{Q}_P$.

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

$$D_{\mathrm{JS}}(Q, R) := \tfrac{1}{2}D\left(Q \parallel \tfrac{Q+R}{2}\right) + \tfrac{1}{2}D\left(R \parallel \tfrac{Q+R}{2}\right)$$

Unnormalized Bhattacharyya divergence[7] is the $\widetilde{\mathbb{V}}$ analogue:

$$D_{\mathrm{UB}}(Q, R) = \tfrac{1}{2}D\left(\tfrac{\sqrt{qr}}{\mu\sqrt{qr}} \parallel q\right) + \tfrac{1}{2}D\left(\tfrac{\sqrt{qr}}{\mu\sqrt{qr}} \parallel r\right)$$

where q and r are densities of Q and R with respect to μ, and $\mu\sqrt{qr}$ is short-hand for $\int \sqrt{q(y)r(y)}d\mu(y)$ using de Finetti notation.[8] The derivation is straight-forward using the definition $D_{\mathrm{UB}}(Q, R) := \log \frac{1}{\mu\sqrt{qr}}$, but it is more easily seen via Lemma A.3. *Unnormalized Rényi divergence* is a generalization $D_\lambda(Q \parallel R) := \log \frac{1}{\mu\, q^\lambda r^{1-\lambda}}$, and a random distribution that takes values Q with probability λ and R with probability $1-\lambda$ has a $\widetilde{\mathbb{V}}$ of $D_\lambda(Q \parallel R)$.

Information theorists have observed "Pythagorean" identities involving information projections and reverse information projections (Csiszár and Matúš, 2003, Theorem 3). Those identities are analogous to the Pythagorean identity in Euclidean space $\mathbb{R}^n$, whereas the compensation identities are analogous to the bias–variance decomposition which is itself an instance of the Pythagorean theorem applied in the $\mathcal{L}^2$-space of $\mathbb{R}^n$-valued random vectors that have finite expected squared norms. The information projection identities tell us about projecting within the space of fixed probability measures, while the compensation identities tell us how to project from the space of random probability measures onto the subset of fixed probability measures. To be more specific, the information projection identities highlight the roles of linear and geometric paths in the space of fixed probability measures, while the compensation identities reveal that the importance of linear and geometric paths extends to the space of random probability measures.

Acknowledgements

Conversations with Andrew Barron about (Barron, 1988) were instrumental in leading us to the insights of this paper. In addition, we thank the reviewer for the detailed comments and insightful suggestions for the manuscript.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.spl.2019.01.008.

4 In fact, the compensation identity and bias–variance decomposition are both instances of this decomposition for Bregman divergences — see Telgarsky and Dasgupta (2012, Lem 3.5) and Pfau (2013).

5 It follows that the centroid is the choice of fixed element that has the smallest possible expected divergence from the random element.

6 This alternative representation of $\widetilde{\mathbb{V}}$ is justified by Brinda (2018, Lem A.3.4).

7 This terminology is borrowed from Grünwald (2007, Eq (19.38)).

8 The de Finetti notation writes measures like ordinary functionals that can be applied to measurable functions; it is summarized and advocated in Pollard (2002, Sec 1.4). We will use this notation extensively in the coming proofs.

References

Barron, Andrew R., 1988. The Exponential Convergence of Posterior Probabilities with Implications for Bayes Estimators of Density Functions. Report 7. University of Illinois.
Brinda, W.D., 2018. Adaptive Estimation with Gaussian Radial Basis Mixtures (Thesis).
Chen, Wei, Jia, Longbin, Jiao, Yong, 2016. Hölders inequalities involving the infinite product and their applications in martingale spaces. Anal. Math. 42 (2), 121–141.
Csiszár, Imre, Matúš, Frantiek, 2003. Information projections revisited. IEEE Trans. Inform. Theory 49 (6), 1474–1490.
Dunford, Nelson, Schwartz, Jacob T., 1958. Linear Operators, Part 1: General Theory. Interscience Publishers, New York.
Grünwald, Peter D., 2007. The Minimum Description Length Principle. MIT Press.
Haussler, David, Opper, Manfred, 1997. Mutual information, metric entropy and cumulative relative entropy risk. Ann. Statist. 2451–2492.
Karakostas, G.L., 2008. An extension of Hölders inequality and some results on infinite products. Indian J. Math. 50, 303–307.
Perlman, Michael D., 1974. Jensen's inequality for a convex vector-valued function on an infinite-dimensional space. J. Multivariate Anal. 4 (1), 52–65.
Pfau, David, 2013. A Generalized Bias-Variance Decomposition for Bregman Divergences. Columbia University.
Pollard, David, 2002. A User's Guide to Measure Theoretic Probability. Cambridge University Press.
Symanzik, Kurt, 1965. Proof and refinements of an inequality of Feynman. J. Math. Phys. 6 (7), 1155–1156.
Telgarsky, Matus, Dasgupta, Sanjoy, 2012. Agglomerative Bregman clustering. In: International Coference on International Conference on Machine Learning, Omnipress, pp. 1011–1018.
Topsøe, Flemming, 2001. Basic concepts, identities and inequalities - the toolkit of information theory. Entropy 3 (3), 162–190.
Veldhuis, Raymond, 2002. The centroid of the symmetrical kullback-leibler distance. IEEE Signal Process. Lett. 9 (3), 96–99.