Hölder's identity

W. D. Brinda, Jason M. Klusowski, and Dana Yang

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 Generality of Hölder's inequality

Hölder's inequality is most commonly written

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

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 fact, this relationship holds for *arbitrary* geometric expectations over a random element indexing functions.[1]

Theorem 1.1 (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 measure μ 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.

[1] 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 A, while results labeled *Corollary* are explained informally before being stated.

Corollary 1.2. *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 measure μ on $\mathcal{Y}$ and finite measure γ on $\mathcal{X}$,*

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

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

Corollary 1.3. *Let $\mathcal{X}$ and $\mathcal{Y}$ be measurable spaces, and let $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ be product measurable. For any measure μ 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, 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 1.2 by letting γ concentrate on a countable set.

[Haussler et al., 1997, 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 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].

Haussler et al. [1997] reference Symanzik [1965] where the inequality in our Theorem 1.1 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 Sec C. 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 2. For now, suffice it to say that $\widetilde{\mathbb{V}}$ represents an expected relative entropy.

Theorem 1.4. *Let $\mathcal{X}$ and $\mathcal{Y}$ be measurable spaces, and let $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ 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 1.5 (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 2.

Theorem 1.6. *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.

2 The Compensation Identities

Theorem 2.1, 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.[2]

Theorem 2.1 (The compensation identity)**.** *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* $\mathcal{Y}$*,*

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

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*.[3]

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

[3] What we call a "geometric mixture" is sometimes called a "log mixture" or "log-convex mixture," for instance by [Grünwald, 2007, Sec 19.6].

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.

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

This integral can be zero, however, in which case the geometric-mixture is not well-defined.[4]

Theorem 2.2 (The reverse compensation identity)**.** *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 $\mathcal{Y}$,*

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

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

A two-point distribution version of Theorem 2.2 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)].

Theorems 2.1 and 2.2 are perfectly analogous to the bias-variance decomposition for Hilbert-space-valued random vectors.[5] The expected divergence from the 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.[6] 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[7]

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

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

[5] 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].

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

[7] This alternative representation of $\widetilde{\mathbb{V}}$ is justified by Lemma A.4.

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 $1/2$.

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

Unnormalized Bhattacharyya divergence[8] is the $\widetilde{\mathbb{V}}$ analogue:

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

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.[9] The derivation is straightforward 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 $D_\lambda(Q\|R) := \log\frac{1}{\mu\, q^\lambda r^{1-\lambda}}$, and a random distribution that takes values Q with probability λ and R with probability $1-\lambda$ has a $\widetilde{\mathbb{V}}$ of $D_\lambda(Q\|R)$.

Acknowledgment

Our advisor Andrew Barron's explanation of Barron [1988] was instrumental in leading us to the insights of this paper.

A 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 A.1. *Let $\{q_x : x\in\mathcal{X}\}$ and $\{r_x : x\in\mathcal{X}\}$ be families of probability densities with respect to a σ-finite measure μ, 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,*

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

[8] This terminology is borrowed from [Grünwald, 2007, Eq (19.38)].

[9] 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.

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

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

□

*Proof of **Theorem 2.1**.* Lemma A.1 justifies changing the order of integration.

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

□

Lemma A.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 $\mu\, e^{\mathbb{E} \log q_X} > 0$, then for any probability measure R on $\mathcal{Y}$,*

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

Proof. Making use of the central trick from the explanations of the mean field approximation algorithm (e.g. Ormerod and Wand [2010]), 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 A.1 justifies the order interchange. □

Lemma A.3. *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 $\mu\, e^{\mathbb{E}\log q_X} > 0$, then*

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

Proof. Use $\widetilde{Q}_P$ as R in Lemma A.2. □

*Proof of **Theorem 2.2**.* Combine Lemmas A.2 and A.3. □

*Proof of **Theorem 1.4**.* We will write $f(X, \cdot)$ as f_X. The key is Lemma A.3.

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

if the geometric mixture is well-defined.

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 μ-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 that set, $\mathbb{E}[f_X - \log \mu\, e^{f_X}]$ can only be $-\infty$ if $\mathbb{E}\log \mu\, e^{f_X}$ is ∞. 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 ∞.

In the one remaining case, the geometric mixture does not exist and $\mathbb{E}f_X = -\infty$ 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.

A closely related derivation in [Barron, 1988, Sec 4] was instructive; the accompanying discussion in that paper provides another interpretation of the quantities involved in Hölder's identity. □

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

For the more general statement allowing arbitrary μ and without conditions on the function, see Section D. □

*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 definition of unnormalized Renyi divergence, $\widetilde{\mathbb{V}}$ of the random distribution is equal to $D_\lambda(Q\|R)$ according to Lemma A.3. Therefore, the desired result is a direct consequence of Holder's identity, at least when μ is σ-finite. However, we deliberately omitted the σ-finiteness requirement. In fact, the reason we required

σ-finiteness in previous Lemmas and Theorems was to justify interchanges in 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 μ is σ-finite. Alternatively, the sum of any finite collection of probability measures is itself a finite dominating measure for each of their densities. □

Lemma A.4. *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 be an $\mathcal{X}$-valued random element. If $\mu\, e^{\mathbb{E}\log q_X} = 0$, then for any probability measure R, $\mathbb{E}D(R\|Q_X) = \infty$.*

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

By Lemma A.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 ∞ with R-probability 1, so $\mathbb{E}D(R\|Q_X) = \infty$. □

B Original use of Hölder's identity

Paste in Jason's tex.

C Dunford and Schwartz exercise

Theorem C.1 is a modified version of [Dunford and Schwartz, 1958, VI.11 Ex 36]. We follow the route of proof suggested by the authors of Dunford and Schwartz [1958] in order to point out that their approach relies on σ-finiteness.

Theorem C.1 (Dunford and Schwartz [1958], VI.11.36)**.** *Suppose μ and μ_1 are positive measures on spaces $\mathcal{X}$ and $\mathcal{Y}$. Assume that $\mu\mathcal{X} = 1$ and μ_1 is σ-finite. Then for any $\mu \times \mu_1-$integrable function $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^+$,*

$$\int \exp\left(\int \log f(x,y)\mu(dx)\right)\mu_1(dy) \le \exp\left(\int \log\left(\int f(x,y)\mu_1(dy)\right)\mu(dx)\right). \tag{3}$$

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 μ is the probability measure as in Theorem C.1. Throughout the remainder of this section, we will omit the dependence on the measure μ and write $|\cdot|_p$ to denote the L^p norm. The following lemma characterizes the behavior of the L^p norm as $p \to 0$ and is crucial in obtaining Theorem C.1.

Lemma C.2 (Dunford and Schwartz [1958], VI.11.32)**.** *For all $\mu-$measurable function $g : \mathcal{X} \to \mathbb{R}^+$, $\lim_{p\to 0} |g|_p$ exists and*

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

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

Lemma C.3 (Dunford and Schwartz [1958], VI.11.13, Generalized Minkowski inequality)**.** *Take σ-finite measures μ, μ_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 $r > 1$,*

$$\left(\int \left(\int f(x,y)\mu_1(dy)\right)^r \mu(dx)\right)^{1/r} \leq \int \left(\int f(x,y)^r \mu(dx)\right)^{1/r} \mu_1(dy). \quad (4)$$

Proof. Write the r'th power of the left-hand side of (4) 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 σ-finiteness of μ and μ_1 to justify the change of order of integration. Apply Hölder's inequality to further bound the above by

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

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 (4) is bounded by the RHS of (4) times the r/s'th power of the LHS. Rearrange the terms to obtain (4). □

Proof of Theorem C.1. With the two auxiliary lemmas the proof of Theorem C.1 is straightforward. By Lemma C.2, we can write the left-hand side of (3) as

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

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

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

where the inequality is obtained by Fatou's lemma. Take $r = 1/p_n$ in Lemma C.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

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

□

D Jensen's inequality in infinite-dimensional spaces

Let $\mathcal{S}$ be any set and μ be any measure on $\mathcal{S}$; the mapping[10] $f \mapsto \log \mu\, e^f$ is convex on the space of functions from $\mathcal{S}$ to $\mathbb{R}$ by Hölder's inequality (Lemma D.11). This convexity has been used to justify

$$\mathbb{E} \log \int e^{g(X,y)} d\mu(y) \geq \log \int e^{\mathbb{E} g(X,y)} d\mu(y) \tag{5}$$

for a real-valued g and random vector X, as an application of Jensen's inequality by, for instance, [Haussler et al., 1997, Lem 1].

- Pettis integrals by definition commute with continuous linear functionals; not necessarily discontinuous linear functionals. point to Perlman paper for a concrete example of failure.

The topic of this section is whether or not Jensen's inequality can be applied to infer that the expectation of the log integral of exponential of a random

[10] When convenient, we use the *De Finetti* notation for integrals — see [Pollard, 2002, Sec 1.4]. We also use the *hypothetical measures* convention to avoid measurability complications; this approach is explained in CITE MY PAPER and section.

mapping is greater than the log integral of exponential of that random mapping's expectation. A first step is to clarify what we mean by the *expectation* of a random mapping. Most convenient for us is the *Pettis expectation* of the random mapping considered as a random vector taking values in the function space; when it exists, it typically coincides with the point-wise expectation. A Pettis expectation is a special case of the *Pettis integral*, which we define next.

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 μ-integral* (or simply *Pettis integral* if the measure is clear from context) of F if for every continuous linear functional $l \in \mathcal{V}'$,

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

If such a vector exists, we say that F is *Pettis integrable*.[11] We do not insist that "the" Pettis integral is unique. Uniqueness is guaranteed if $\mathcal{V}'$ separates points CITE source; for instance, the topological dual space of any locally convex Hausdorff space separates points [Aliprantis and Border, 2006, Cor 5.82]. If $l \circ F$ is measurable for all continuous linear functionals, then X is called *weakly measurable*. Rather than requiring weak measurability, we will implicitly use the *hypomeasures* idea described in CITE paper and section. That framework justifies proceeding with derivations as if every function were measurable; each "integral" is instead considered a function mapping from extensions of the measure to the ordinary integrals taken with respect to those extension.

Another notion of integral for abstract vector-valued functions is the *Bochner integral* which is constructed using limits of simple functions, analogously to the Lebesgue integral. Every Bochner integral satisfies (6) and is therefore also a Pettis integral. A real-valued function is Lebesgue integrable if and only if it is Pettis integrable, and in that case, the two integrals coincide. We will denote the Pettis integral by $\mu\, F$, the same De Finetti notation we use for Lebesgue integrals.[12] We may use the symbol $\mathbb{E}$ for a Pettis integral with respect to a probability measure, which we call a *Pettis expectation*.

D.1 Some Pettis integral properties

D.1.1 Commutation of the Pettis integral with continuous linear operators

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 show in this section. Straight-forward proofs are given along the way.

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

[11] We caution the reader that the definitions and terminology on this subject are inconsistent and do not always agree with ours.

[12] When it has been specified that the function is real-valued, one should interpret the notation to denote a Lebesgue integral, meaning that it is allowed to be infinitely-valued.

Theorem D.1. *Let $\mathcal{U}$ and $\mathcal{V}$ be rTVSs. Suppose F is a Pettis μ-integrable $\mathcal{U}$-valued function and T is a continuous linear operator from $\mathcal{U}$ to $\mathcal{V}$. Show that $T\mu\, F$ is the Pettis integral of $T \circ F$.*

Proof. This argument originally appeared in [Pettis, 1938, 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.

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

We've seen than 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 $T \circ F$. □

Lemma D.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 $\mathcal{V}$.*

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. □

Lemma D.3. *If $X : \Omega \to \mathcal{V}$ 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 D.2 to pass a scalar through the expectation.

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

□

Based on Lemmas D.2 and D.3 along with Theorem D.1, we can conclude that Pettis expectations commute with continuous *affine* operators too.

Corollary D.4. *If a is a continuous affine operator on an rTVS $\mathcal{V}$, then any Pettis integrable $\mathcal{V}$-valued random vector X has $\mathbb{E}a(X) = a(\mathbb{E}X)$.*

D.1.2 Increasing property of the Pettis integral

The purpose of this subsection is to establish Corollary D.7, a Pettis integral version of the familiar *increasing property* of the Lebesgue integral. It is a key ingredient in Theorem D.8, a generalization of Jensen's inequality. To make sense of this, we will need to bring up a few additional assumptions on the spaces in question. First, a TVS is called *locally convex* if "every neighborhood of 0 includes a convex neighborhood of 0," [Aliprantis and Border, 2006, Def 5.71] or alternatively if its topology is generated by a family of semi-norms [Aliprantis and Border, 2006, Thm 5.73]; in particular, every normed space is locally convex.

Theorem D.5. *Let F map from a measure space into a real locally convex space. If a Pettis integral of F exists, then it must be in the closure of the convex cone generated by the range of F.*

Proof. Let C denote the closure of the convex cone. Suppose the Pettis integral v_F is outside of C. Then by [Aliprantis and Border, 2006, Cor 5.84] which invokes the Hahn-Banach Theorem, there exists a continuous linear functional l for which $l(v_F) < 0$ while l is non-negative on C. By the definition of Pettis integral, $l(v_F)$ is supposed to equal the Lebesgue integral of $l \circ F$. But $l \circ F$ is non-negative on Ω, so it cannot have a negative Lebesgue integral. □

An *ordered vector space* is a vector space with a partial order that is stable under vector addition and scalar multiplication. The set of vectors $\{v : v \geq 0\}$ is called the *positive cone*; it is a pointed convex cone. Conversely, any pointed convex cone in a vector space can be *designated as the positive cone* to imbue the vector space with a unique ordered vector space structure [Aliprantis and Border, 2006, Sec 8.1].

Any partial order that assigns each pair of elements a supremum and infimum is called a *lattice*. An ordered vector space that is also a lattice is called a *Riesz space*. In a Riesz space, each vector has a unique decomposition into a *positive part* and *negative part* $v = v_+ - v_-$ where $v_+ := v \vee 0$ and $v_- := -v \vee 0$. The *absolute value* of a vector is $|v| := v_+ + v_-$. A set in a Riesz space is called *solid* if for any x in the set, all vectors with absolute value less than or equal to that of x are also in the set.

A Riesz space that is also a TVS is called a *topological Riesz space*. A topological Riesz space is called *locally solid* if the solid neighborhoods of 0 are a local basis for 0. A locally solid Riesz space that is also locally convex is called a *locally convex-solid Riesz space*. A topological space is called Hausdorff if every pair of distinct points are separated by neighborhoods. In a locally-solid Hausdorff Riesz space, the positive cone is closed [Aliprantis and Border, 2006, Thm 8.43.1]. Any Banach lattice is a locally convex-solid Hausdorff Riesz space by [Aliprantis and Border, 2006, Thm 8.46].

Corollary D.6. *Let F map from a measure space into the positive cone of a real locally convex-solid Hausdorff Riesz space. The Pettis integral of F, if it exists, must also be in the positive cone.*

If one function is at least as large as another function point-wise, then its integral is at least as large. To see this, apply Corollary D.6 to the difference between the functions.

Corollary D.7 (Increasing property of the Pettis integral)**.** *Let $\mathcal{S}$ be a set and $\mathcal{X}$ be a real locally convex-solid Hausdorff Riesz space. Suppose $F : \mathcal{S} \to \mathcal{X}$ is point-wise greater than or equal to $G : \mathcal{S} \to \mathcal{X}$. Then for any measure μ on $\mathcal{S}$,*

$$\mu F \geq \mu G$$

if the integrals exist.

D.2 Jensen's inequality for Pettis expectations

D.2.1 Convexity and Jensen convexity

Let $\mathcal{U}$ be a real vector space and $\mathcal{V}$ be a partially ordered real vector space. An operator $f : \mathcal{U} \to \mathcal{V}$ is called *convex* if for any $u_1, u_2 \in \mathcal{X}$ and $\lambda \in [0, 1]$,

$$f(\lambda u_1 + [1 - \lambda]u_2) \leq \lambda f(u_1) + [1 - \lambda]f(u_2) \tag{7}$$

Now assume that $\mathcal{U}$ and $\mathcal{V}$ also have topologies compatible with their vector space structures. We will say that f is *Jensen convex at* u_0 if there exists a continuous affine operator $a : \mathcal{U} \to \mathcal{V}$ such that $a(u_0) = f(u_0)$ and a is *below* f, that is, $a \leq f$ point-wise on the domain of f. If a function is Jensen convex at every point in its domain, then we will simply call it *Jensen convex*.[13]

Theorem D.8 (Jensen's inequality for Pettis expectations)**.** *Let $\mathcal{V}$ be an rTVS and $\mathcal{X}$ be a real locally convex-solid Hausdorff Riesz space. For any $\mathcal{V}$-valued random vector X, if f is Jensen convex at $\mathbb{E}X$, then*

$$\mathbb{E}f(X) \geq f(\mathbb{E}X)$$

assuming the Pettis expectations exist.

Proof. Let a be a continuous affine operator below f with $a(\mathbb{E}X) = f(\mathbb{E}X)$. By Corollaries D.7 and D.4,

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

□

[13] Some papers have used the term *Jensen convex* to mean that (7) holds with $\lambda = 1/2$. However, that concept also has the name of *midpoint convexity*, which seems much more usefully descriptive. Our definition of *Jensen convexity* ties it closely to Jensen's inequality.

A general form of the Hahn-Banach Theorem can be stated for mappings into order complete Riesz spaces (Aliprantis Thm 8.30). It follows from the Axiom of Choice; however, it is strictly weaker than Choice and can itself be taken as axiomatic instead.

Define order completeness and say which earlier examples of Riesz spaces are order complete.

ALSO point out that ... Every locally solid Riesz space can be uniquely extended to an order complete Riesz space called its Dedekind completion (use Aliprantis Thm 8.43 with Thm 8.8).

Fremlin [1974]

Aliprantis and Border [2006]

In the very general context of Theorem D.9, convexity is equivalent the existence of (not necessarily continuous) tangent[14] affine operators below it.

Theorem D.9. *Let $\mathcal{U}$ be an rTVS and $\mathcal{V}$ be an order complete real Riesz space. A function $f : \mathcal{U} \to \mathcal{V}$ is convex iff at every point in its domain there exists a tangent affine operator below it.*

Proof. First suppose f is convex, and consider any point u_0 in its domain. Define f_0 to be the translated function $f_0(u) = f(u + u_0) - f(u_0)$ which maps $0_{\mathcal{U}}$ to $0_{\mathcal{V}}$; it is also convex. Aliprantis Thm 8.30 says that any linear operator below f_0 on a subspace of $\mathcal{U}$ can be extended to a linear operator below f_0 on all of $\mathcal{U}$. A convenient linear operator to start with is the one defined only on the zero subspace $\{0_{\mathcal{U}}\}$; it maps $0_{\mathcal{U}}$ to $0_{\mathcal{V}}$ which equals $f_0(0_{\mathcal{U}})$. It has an extension to a linear operator below f_0 and tangent to it at $0_{\mathcal{U}}$. A translation of that extension is an affine functional below f and tangent to it at u_0.

Conversely, assume that f is not convex. Then there exists $u_1, u_2 \in \mathcal{U}$ and $\lambda \in (0, 1)$ such that

$$f(\lambda u_1 + [1 - \lambda]u_2) > \lambda f(u_1) + [1 - \lambda]f(u_2).$$

Let $u_0 := \lambda u_1 + [1 - \lambda]u_2$. Any line that passes through u_0 and is in the plane with u_1 and u_2 will pass strictly above either u_1 or u_2. □

Theorem D.9 helps clarify the relationship between convexity and Jensen convexity. It tells us that when the range is an order complete real Riesz space, Jensen convexity implies convexity. For any such function on a finite-dimensional domain, the two concepts become equivalent since all affine operators are continuous in that context. More generally, however, one needs additional assumptions to establish that a convex function is Jensen convex; for instance, every non-continuous linear functional is convex but not Jensen convex.

Lemma D.10. *Let $\mathcal{U}$ be an rTVS, and let $\mathcal{V}$ be an order complete real locally solid Riesz space. Suppose $f : \mathcal{U} \to \mathcal{V}$ is a convex function that is continuous at $u_0 \in \mathcal{U}$. Then f is Jensen convex at u_0.*

[14] Our use of the term "tangent" is *not* intended to mean that the slope matches the derivative of the function at the point of contact; indeed, we make no assumptions regarding differentiability.

Proof. Theorem D.9 assures us that there exists an affine operator a that is below f and tangent to it at u_0. We only need to ensure that the operator is continuous. Because every TVS has a translation-invariant topology, we can assume without loss of generality that $u_0 = 0_{\mathcal{U}}$ and that $f(u_0) = 0_{\mathcal{V}}$. In other words, we assume that a is linear.

To prove that a is continuous, we will show that for any neighborhood W of $0_{\mathcal{V}}$ one can find a neighborhood of $0_{\mathcal{U}}$ with image in W. It suffices to check any local base of $0_{\mathcal{V}}$; there is a local base defined by the continuous Riesz pseudo-norms, according to CITE Fremlin 23G. Given any continuous Riesz pseudo-norm ρ, we will identify a neighborhood of $0_{\mathcal{U}}$ with image in $\{v : \rho(v) < 1\}$.

Define $g(u) := f(u) \vee f(-u)$; it is also continuous at $0_{\mathcal{U}}$ (CITE Fremlin 22B(b)). Let $V_\rho := g^{-1}\{v : \rho(v) < 1/2\}$; by continuity of g at $0_{\mathcal{U}}$, we know that V_ρ is a neighborhood of $0_{\mathcal{U}}$. For any $u \in V_\rho$,

$$\begin{aligned}
\rho a(u) &= \rho(a(u)_+ - a(u)_-) \\
&= \rho(a(u)_+ - a(-u)_+) \\
&\leq \rho a(u)_+ + \rho a(-u)_+ \\
&\leq \rho f(u)_+ + \rho f(-u)_+ \\
&\leq \rho f(u) + \rho f(-u) \\
&\leq 2\rho g(u) \\
&< 1.
\end{aligned}$$

The steps are justified by the definition of Riesz pseudo-norms — see CITE Fremlin 22A. □

D.2.2 The log integral of exponential function

Lemma D.11. *Let $\mathcal{F}$ be the vector space of all real-valued functions on a set S. Given any measure μ on S and constant $c \in \mathbb{R}$, the mapping $f \mapsto \log \mu\, e^{cf}$ from $\mathcal{F}$ to $\mathbb{R}$ is convex.*

Proof. Consider the real-valued functions f_1 and f_2 along with any $\lambda \in [0, 1]$. The inequality

$$\log \mu\, e^{\lambda f_1 + [1-\lambda] f_2} \leq \lambda \log \mu\, e^{f_2} + [1 - \lambda] \log \mu\, e^{f_2}$$

is equivalent to

$$\mu\,[(e^{f_1})^\lambda (e^{f_2})^{1-\lambda}] \leq (\mu\, e^{f_1})^\lambda (\mu\, e^{f_2})^{1-\lambda}$$

which is equivalent to Hölder's inequality.

The behavior of $f \mapsto \log \mu\, e^{cf}$ on the path from f_1 to f_2 matches the behavior of $f \mapsto \log \mu\, e^f$ on the path from cf_1 to cf_2. □

Convexity by itself is not sufficient to justify the use of Jensen's inequality. For discrete random vectors, however, the validity of Jensen's inequality is established in Theorem D.12.

Theorem D.12. *Let $\mathcal{F}$ be the vector space of all real-valued functions on a set S, let f_X be an $\mathcal{F}$-valued random vector, let μ be a measure on S, and let $c \in \mathbb{R}$. If either*

(i) f_X takes finitely many values or

(ii) f_X takes countably many values, the Pettis expectation $\mathbb{E}f_X$ exists, and μ is σ-finite

then

$$\log \mu\, e^{c\mathbb{E}f_X} \leq \mathbb{E} \log \mu\, e^{cf_X}. \tag{8}$$

Proof. Case (i) is equivalent to the statement of Hölder's inequality for products of finitely many functions.

For (ii), review the proof of Lemma D.11 in light of the countable product version of Hölder's inequality derived in point to Karakostas 2008. □

- USES OF Jensen's inequality WERE valid in Haussler/Opper and elsewhere, right? do their uses always satisfy criteria of my Theorem above?

More generally, if one establishes *Jensen convexity* of $f \mapsto \log \mu\, e^{cf}$ at $\mathbb{E}f_X$ (within a vector lattice of functions satisfying the conditions of Theorem D.8), then (8) can be asserted. The point-wise ordering is a natural partial order for us to use on $\mathcal{F}$ (the vector space defined in Lemma D.11 and Theorem D.12). Any real Banach lattice is a real locally convex-solid Hausdorff Riesz space (by Aliprantis Thm 8.46), but it need not be order complete. Every L^p space for $p \in [1, \infty]$ is a real Banach lattice and is also order complete — see Aliprantis Ch 13. Thus any L^p space of (equivalence classes of) functions would be convenient for us to apply Lemma D.10 and Theorem D.8.

Corollary D.13. *Let S be a set, let μ and γ be measures on S, and let $c \in \mathbb{R}$. Suppose f_X is a Pettis integrable random vector taking values in a convex set $A \subseteq L^p(\gamma)$ and that $\mathbb{E}f_X \in A$. If the mapping from A to $\mathbb{R}$ defined by $f \mapsto \log \mu\, e^{cf}$ is continuous at $\mathbb{E}f_X$, then*

$$\log \mu\, e^{c\mathbb{E}f_X} \leq \mathbb{E} \log \mu\, e^{cf_X}.$$

Remark: If A is also closed, then it is certain that $\mathbb{E}f_X \in A$ by Theorem REF the theorem thm:pettis-closed-convex-hull.

Theorem D.14. *Let μ be a measure on $\mathbb{R}^d$, and let $c \in \mathbb{R}$. For any $\mathbb{R}^d$-valued random vector X,*

$$\log \int_S e^{c\mathbb{E}\|X-y\|^2} d\mu(y) \leq \mathbb{E} \log \int_S e^{c\|X-y\|^2} d\mu(y).$$

Proof. X is an $\mathbb{R}^d$-valued random vector. To put things into our framework, we define the random vector f_X by $y \mapsto \|X - y\|^2$. Each possible vector in the range of f_X is in the set

$$\mathcal{F} := \{y \mapsto \|x - y\|^2 : x \in \mathbb{R}^d\}$$

Need to figure out exactly what topology we want for $\mathcal{F}$ ... presumably some sort of $L^1(\gamma)$ or $L^2(\gamma)$ space that distinguishes the different functions into different equivalence classes - but we will need to make sure the point-evaluation functionals are continuous on $\mathcal{F}$ AND that the tricky Jensen function is continuous on $\mathcal{F}'$ (as defined below).

Let us first explain that the Pettis expectation of the random vector f_X is equal to the *point-wise* expectation indicated by the notation $\mathbb{E}\|X - y\|^2$. The point-evaluation functionals are continuous on $\mathcal{F}$. FIRST, PROVE THIS - at least I think it should work out ... but what's a good topology to use on $\mathcal{F}$?.... Therefore, one can commute any point-evaluation functional with the Pettis expectation; in other words, the value of the Pettis expectation at any $y \in \mathbb{R}^d$ is equal to the Pettis expectation of the value of the random function at that y.

To complete the proof, we establish that $f \mapsto \log \mu\, e^{cf}$ is continuous on the set

$$\mathcal{F}' := \{y \mapsto \mathbb{E}_{X\sim Q}\|X - y\|^2 : Q \text{ is a probability measure on } \mathbb{R}^d\}.$$

PROVE this continuity.

DO I need to require any moment conditions on X? OR do edge cases automatically take care of themselves by becoming trivial? ... IF moment conditions are needed, perhaps the existence of the Pettis integral can be guaranteed under those conditions.

□

Presumably the above argument can be be "turned around" to give a condition for which a generalization of Hölder's inequality works?

ALSO - there are other papers out there about using Jensen's inequality for infinite-dimensional vector spaces. BROWSE those papers to see how they compare to this one.

- MAYBE the countable case can be extended to arbitrary expectations more easily in the case of Bochner integrals - for σ-finite μ, try to extend from the countable Hölder result - I think we just need limits of some function to be equal to that function at the limit, so presumably we need that function to be continuous. what function are we talking about here? is it the exact same function that I'm trying to establish continuity of for the Pettis integral case? EVEN a limiting inequality might accomplish what we need if it points in the right direction.

ALIPRANTIS Lem 5.74 says that the product topology on any real function space makes it an LCHS. That might be useful.

Lemma D.15. *Let $\mathcal{X}$ be a normed space. If $l(x) \leq \|l\|t$ for all $l \in \mathcal{X}^*$, then $\|x\| \leq t$.*

Proof. We will prove the contrapositive by supposing that $\|x\| > t$. Define $l_0(x) = \|x\|$, and extend this continuous linear functional to $\mathcal{X}$ by the Hahn-

Banach theorem, while keeping its norm $\|l_0\| = 1$. Then

$$\begin{aligned} l_0(x) &= \|x\| \\ &= \|l_0\|\|x\| \\ &> \|l_0\|t. \end{aligned}$$

□

Lemma D.16. *Let f map from a set Ω to a normed space $\mathcal{X}$, and let μ be a measure on Ω such that f is Pettis integrable. Then*

$$\|\mu\, f\| \leq \mu\, \|f\|.$$

Proof. We follow the logic of David C Ulrich's answer to CITE. For any continuous linear functional $l \in \mathcal{X}^*$,

$$\begin{aligned} l(\mu\, f) &= \mu\, l(f) \\ &\leq \mu\, \|l\|\|f\| \\ &= \|l\|\mu\, \|f\|. \end{aligned}$$

The desired conclusion follows from Lemma D.15. This might hold for Dunford integrals. □

Recall that for $p < 1$, the "$\mathcal{L}^p$-norm" is not actually a semi-norm, as it fails to satisfy the triangle inequality. It has a reversed property to that of Lemma D.16.

Lemma D.17. *Let $p \in (0, 1]$. Let f map from a set Ω to a normed space $\mathcal{X}$, and let μ be a measure on Ω such that f is Pettis integrable. Then*

$$\|\mu\, f\|_p \geq \mu\, \|f\|_p.$$

Proof. Apply Lemma D.16 to the function $f^{1/p}$. - MORE CONDITIONS probably: NEED some additional things to be Pettis integrable? - CITE: Idea comes from Dunford and Schwarz Ex 14. □

the following derivation is about a POINT-WISE geometric expectation:

Given a set Ω with measure γ and a normed space $\mathcal{X}$, define the $\mathcal{L}^p_{\mathcal{X}}(\gamma)$-norm of any $f : \Omega \to \mathcal{X}$ by $\|f\|_p := (\gamma\, \|f\|^p)^{1/p}$ for $p \in (-\infty, 0) \cup (0, \infty)$. Guided by limiting behavior, we define $\|f\|_{-\infty} = \inf f$, $\|f\|_\infty = \sup f$, and $\|f\|_0 = e^{\gamma \log f}$.

MAYBE strict convexity is enough to ensure tangent continuous affine functionals below the function.

MIGHT be worth including the Dunford and Schwarz result for L^p spaces also just to ensure that the point-wise integral version is out there too.

HOW DO I RECONCILE the point-wise and Pettis integral stories???

POINT-WISE version of theorem works for inner product spaces. See Ulrich's response:

References

Charalambos D Aliprantis and Kim Border. *Infinite dimensional analysis: a hitchhiker's guide*. Springer Science & Business Media, 2006.

Andrew R. Barron. The exponential convergence of posterior probabilities with implications for bayes estimators of density functions. Report 7, University of Illinois, 1988.

W. Chen, L. B. Jia, and Y. Jiao. Hölder's inequalities involving the infinite product and their applications in martingale spaces. *Analysis Mathematica*, 42(2):121–141, Jun 2016.

Imre Csiszár and František Matúš. Information projections revisited. *IEEE Transactions on Information Theory*, 49(6):1474–1490, 2003.

Nelson Dunford and Jacob T Schwartz. *Linear operators*. Interscience Publishers, 1958.

David H Fremlin. *Topological Riesz spaces and measure theory*. Cambridge Univ. Press, 1974.

Peter D. Grünwald. *The minimum description length principle*. MIT press, 2007. ISBN 0262072815.

David Haussler, Manfred Opper, et al. Mutual information, metric entropy and cumulative relative entropy risk. *The Annals of Statistics*, 25(6):2451–2492, 1997.

George L. Karakostas. An extension of hölder's inequality and some results on infinite products. *Indian Journal of Mathematics*, 50(2):303–307, 2008.

J. T. Ormerod and M. P. Wand. Explaining variational approximations. *The American Statistician*, 64(2):140–153, 2010.

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.

B. J. Pettis. On integration in vector spaces. *Transactions of the American Mathematical Society*, 44(2):277–304, 1938.

David Pfau. A generalized bias-variance decomposition for bregman divergences. Available at `davidpfau.com`, 2013.

David Pollard. *A user's guide to measure theoretic probability*, volume 8. Cambridge University Press, 2002.

Kurt Symanzik. Proof and refinements of an inequality of feynman. *Journal of Mathematical Physics*, 6(7):1155–1156, 1965.

Matus Telgarsky and Sanjoy Dasgupta. Agglomerative bregman clustering. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, pages 1011–1018. Omnipress, 2012.

Flemming Topsøe. Basic concepts, identities and inequalities — the toolkit of information theory. *Entropy*, 3(3):162–190, 2001.

Raymond Veldhuis. The centroid of the symmetrical kullback-leibler distance. *IEEE signal processing letters*, 9(3):96–99, 2002.