$\mathcal{L}^2$ spaces (and their useful properties)

1 Orthogonal projections in $\mathcal{L}^2$

For a measure space $(\mathcal{X}, \mathcal{A}, \mu)$, the set $\mathcal{L}^2 = \mathcal{L}^2(\mathcal{X}, \mathcal{A}, \mu)$ of all square integrable, $\mathcal{A}\backslash\mathcal{B}(\mathbb{R})$-measurable real functions on $\mathcal{X}$ would be a Hilbert space if we worked with equivalence classes of functions that differ only on μ-negligible sets. The corresponding set $L^2(\mathcal{X}, \mathcal{A}, \mu)$ can be identified as a true Hilbert space.

Lazy probabilists (like me) often ignore the distinction between L^2 and $\mathcal{L}^2$, referring to $\|f\|_2 = \big(\mu(f^2)\big)^{1/2}$ as a norm on $\mathcal{L}^2$ (rather than using the more precise term 'semi-norm') and

$$\langle f, g\rangle = \mu(fg) \qquad \text{for } f, g \in \mathcal{L}^2(\mathcal{X}, \mathcal{A}, \mu)$$

as an inner product. It is true that $\langle f, g\rangle$ is linear in f for fixed g and linear in g for fixed f; and it is true that $\|f\|_2^2 = \langle f, f\rangle$; but we can only deduce that $f(x) = 0$ a.e.$[\mu]$ if $\|f\|_2 = 0$. The inner product satisfies the Cauchy-Schwarz inequality:

<1> $$|\langle f, g\rangle| \le \|f_2\|\, \|g\|_2 \qquad \text{for } f, g \in \mathcal{L}^2(\mathcal{X}, \mathcal{A}, \mu),$$

which is actually just a special case of the Hölder inequality.

Remark. For a surprising range of application of the humble C-S inequality see the surprising little book by Steele (2004).

As shown by HW3.1, the space $\mathcal{L}^2$ is also complete: for each Cauchy sequence $\{h_n : n \in \mathbb{N}\}$ in $\mathcal{L}^2$ there exists an h in $\mathcal{L}^2$ (unique only up to μ-equivalence) for which $\|h_n - h\|_2 \to 0$.

A subset $\mathcal{H}_0$ of $\mathcal{L}^2$ is said to be ***closed*** if it contains all its limit points: if f in $\mathcal{L}^2$ with $\|h_n - f\|_2 \to 0$ for a sequence $\{h_n\}$ in $\mathcal{H}_0$ then $f \in \mathcal{H}_0$. Equivalently, $[f] \subset \mathcal{H}_0$, where

$$[f] := \{g \in \mathcal{L}^2(\mathcal{X}, \mathcal{A}, \mu) : f = g \quad \text{a.e.}[\mu]\}.$$

<2> **Example.** A linear map $\tau : \mathcal{L}^2(\mathcal{X}, \mathcal{A}, \mu) \to \mathbb{R}$ (also known as a linear functional) is said to be continuous if there exists a finite constant C for which

$$|\tau(h)| \le C \|h\|_2 \qquad \text{for all } h \in \mathcal{L}^2(\mathcal{X}, \mathcal{A}, \mu).$$

The set $\mathcal{H}_0 = \{h \in \mathcal{L}^2 : \tau(h) = 0\}$ is a closed subset of $\mathcal{L}^2$: if $\|h_n - f\|_2 \to 0$ for a sequence $\{h_n\}$ in $\mathcal{H}_0$ then

$$|\tau(h_n) - \tau(f)| = |\tau(h_n - f)| \le C \|h_n - f\|_2 \to 0,$$

which implies $\tau(f) = 0$.

□

It is often enough to have just $[f] \cap \mathcal{H}_0 \neq \emptyset$. To avoid some tedious qualifications about negligible sets I will say that a subset $\mathcal{H}_0$ of $\mathcal{L}^2(\mathcal{X}, \mathcal{A}, \mu)$ is ***effectively closed*** if: for each f in $\mathcal{L}^2$ with $\|h_n - f\|_2 \to 0$ for a sequence $\{h_n\}$ in $\mathcal{H}_0$ there exists an h in $\mathcal{H}_0$ for which $h(x) = f(x)$ a.e.$[\mu]$.

Remark. The usual fix of working with $L^2(\mathcal{X}, \mathcal{A}, \mu)$ gets a bit dangerous when several different measures or sigma-fieldss are involved.

<3> **Example.** Suppose $\mathcal{G}$ is a sub-sigma-field of $\mathcal{A}$. You know that $\mathcal{H}_0 := \mathcal{L}^2(\mathcal{X}, \mathcal{G}, \mu)$ is complete. You can also think of $\mathcal{H}_0$ as a subspace of $\mathcal{L}^2 = \mathcal{L}^2(\mathcal{X}, \mathcal{A}, \mu)$. It is essentially closed.

If $\|g_n - f\|_2 \to 0$ for a sequence $\{g_n\}$ in $\mathcal{H}_0$ and $f \in \mathcal{L}^2$ then

$$\|g_n - g_m\|_2 \le \|g_n - f\|_2 + \|g_m - f\|_2 ,$$

which is less than any given $\epsilon > 0$ if $\min(m, n)$ is large enough. That is, $\{g_n\}$ is a Cauchy sequence in $\mathcal{H}_0$. By completeness of $\mathcal{L}^2(\mathcal{X}, \mathcal{G}, \mu)$ there exists an h in $\mathcal{H}_0$ for which $\|h_n - h\| \to 0$. It follows that $\|f - h\|_2 = 0$, so that $h = f$ a.e.$[\mu]$.

□

The following result underlies the existence of both Radon-Nikodym derivatives (densities) for measures and Kolmogorov conditional expectations.

<4> **Theorem.** *Suppose $\mathcal{H}_0$ is an essentially closed subspace of $\mathcal{L}^2(\mathcal{X}, \mathcal{A}, \mu)$. For each $f \in \mathcal{L}^2$ there exists an f_0 in $\mathcal{H}_0$ for which:*

(i) $\|f - f_0\|_2 = \delta := \inf\{\|f - h\|_2 : h \in \mathcal{H}_0\}$;

(ii) $\langle f - f_0, h\rangle = 0$ *for every* h *in* $\mathcal{H}_0$;

(iii) f_0 *is uniquely determined up to* μ*-equivalence;*

(iv) $\|f\|_2^2 = \|f_0\|_2^2 + \|f - f_0\|_2^2$.

Proof The argument uses completeness of $\mathcal{L}^2$ and the identity

<5> $$\|a + b\|_2^2 + \|a - b\|_2^2 = 2\,\|a\|_2^2 + \|b\|_2^2 \qquad \text{for all } a, b \in \mathcal{L}^2,$$

an equality that results from expanding $\langle a + b, a + b\rangle + \langle a - b, a - b\rangle$ then cancelling out $\langle a, b\rangle$ terms.

By definition of the infimum, for each $n \in \mathbb{N}$ there exists an $h_n \in \mathcal{H}_0$ for which

$$\|f - h_n\|_2 \le \delta_n := \delta + n^{-1}.$$

Invoke equality <5> with $a = f - h_n$ and $b = f - h_m$:

$$4\,\|f - (h_n + h_m)/2\|_2^2 + \|h_n - h_m\|_2^2 = 2\,\|f - h_n\|_2^2 + 2\,\|f - h_m\|_2^2\,.$$

The first term on the left-hand side is $\ge 4\delta^2$ because $(h_n + h_m)/2 \in \mathcal{H}$. Thus

$$\|h_n - h_m\|_2^2 \le 2\delta_n^2 + 2\delta_m^2 - 4\delta^2 \to 0 \qquad \text{as } \min(m, n) \to \infty.$$

That is $\{h_n\}$ is a Cauchy sequence, which converges in norm to an f_0 in $\mathcal{L}^2$. We may assume that $f_0 \in \mathcal{H}$ because $\mathcal{H}_0$ is essentially closed,

Equality (i) follows from

$$\delta \le \|f - f_0\|_2 \le \|f - h_n\|_2 + \|h_n - f_0\|_2 \to \delta \qquad \text{as } n \to \infty.$$

For (ii) note, for each $h \in \mathcal{H}$, that the quadratic

$$\|f - (f_0 + th)\|^2 = \|f - f_0\|^2 - 2t\langle f - f_0, h\rangle + t^2\,\|h\|_2^2$$

achieves its minimum value δ^2 at $t = 0$, which forces the coefficient of t to equal zero.

For (iii) suppose $f_0, f_1 \in \mathcal{H}$ and both $f - f_0$ and $f - f_1$ are orthogonal to each h in $\mathcal{H}$. Then the difference $f_0 - f_1$ must be orthogonal to itself, that is $\|f_0 - f_1\|_2^2 = 0$, forcing $f_0 = f_1$ a.e.$[\mu]$.

Equality (iv) follows from the fact that $\langle f - f_0, f_0 \rangle = 0$.

□

The function f_0, which is unique only up to μ-equivalence, is called an (orthogonal) projection of f onto $\mathcal{H}_0$. For each f, by arbitrarily picking one member from $[f_0] \cap \mathcal{H}_0$ as THE projection of f we get a map π_0 from $\mathcal{L}^2$ into $\mathcal{H}_0$ whose behavior involves many *almost everywhere* exceptions. For example, suppose f and g are functions in $\mathcal{L}^2$ with $\pi_0(f) = f_0$ and $\pi_0(g) = g_0$. for $i = 1, 2$. For constants c and d, part (iii) of the Theorem gives

$$c\langle f - f_0, h \rangle = 0 = d\langle g - g_0, h \rangle \qquad \text{for all } h \text{ in } \mathcal{H}_0.$$

Linearity of the inner product in its first argument gives

$$\langle c(f - f_0) + d(g - g_0), h \rangle = 0 \qquad \text{for all } h \text{ in } \mathcal{H}_0.$$

That is, $cf_0 + dg_0$ satisfies the equalities that define $\pi_0(cf + dg)$ up to a μ-equivalence, which implies

<6> $$\pi_0(cf + dg) = c\pi_0(f) + d\pi_0(g) \quad \text{a.e.}[\mu].$$

This result is as close to linearity as we can hope to get for a map that is only defined up to a μ-equivalence.

2 Continuous linear functionals

Suppose $\tau : \mathcal{L}^2(\mathcal{X}, \mathcal{A}, \mu) \to \mathbb{R}$ is a continuous linear functional, in the sense of Example <2>. From that Example you know that the linear subspace $\mathcal{H}_0 = \{h \in \mathcal{L}^2 : \tau(h) = 0\}$ is closed as a subset of $\mathcal{L}^2$.

To avoid trivialities, suppose there exists an f in $\mathcal{L}^2$ for which $\tau(f) \neq 0$. Without loss of generality we may assume $\tau(f) = 1$.

By Theorem <4>, there exists an f_0 in $\mathcal{H}_0$ for which $\langle h, f - f_0 \rangle = 0$ for all h in $\mathcal{H}_0$. Define $g_0 = f - f_0$. The construction ensures that $\tau(g_0) = \tau(f) - \tau(f_0) = 1$. By continuity, $1 \leq C \|g_0\|_2$, which implies that $\|g_0\|_2 \neq 0$.

If $h \in \mathcal{L}^2$ with $\tau(h) = d$ then $h - dg_0 \in \mathcal{H}$, because $\tau(h - dg_0) = 0$. The equality $\langle h - dg_0, g_0\rangle = 0$ rearranges to $\langle h, g_0\rangle = d\,\|g_0\|_2^2$. If we define $k := g_0/\,\|g_0\|_2^2$ then

<7> $$\tau(h) = \langle h, k\rangle \qquad \text{for all } h \text{ in } \mathcal{L}^2.$$

The $\mathcal{L}^2$ function k for which <7> holds is unique up to a μ-equivalence. For if k_1 is another $\mathcal{L}^2$ function with $\tau(h) = \langle h, k_1\rangle$ for all $h \in \mathcal{L}^2$ then

$$0 = \tau(h) - \tau(h) = \langle h, k\rangle - \langle h, k_1\rangle = \langle h, k - k_1\rangle \qquad \text{for all } h \text{ in } \mathcal{L}^2.$$

The choice $h = k - k_1$ shows that $\|k - k_1\|_2^2 = 0$, so that $k_1 = k$ a.e.$[\mu]$.

3 Conditioning heuristics

Recall the conditioning problem. We have probability spaces $(\mathcal{X}, \mathcal{A}, \mathbb{P})$ and $(\mathcal{Y}, \mathcal{B}, \mathbb{Q})$ and an $\mathcal{A}\backslash\mathcal{B}$-measurable map T from $\mathcal{X}$ into $\mathcal{Y}$ whose distribution under $\mathbb{P}$ is $\mathbb{Q}$. That is

$$\mathbb{Q}g = \mathbb{P}g(Tx) \qquad \text{at least for } g \in \mathbb{M}^+(\mathcal{Y}, \mathcal{B}).$$

We seek a Markov kernel $\mathbb{K} = \{K_y : y \in \mathcal{Y}\}$—a family of probability measures on $\mathcal{A}$ for which $y \mapsto \mathbb{K}_y A$ is $\mathcal{B}$-measurable for each $A \in \mathcal{A}$—for which

<8> $$\mathbb{P}f(x) = \mathbb{Q}^y K_y^x f(x) \qquad \text{for each } f \in \mathcal{M}^+(\mathcal{X}, \mathcal{A})$$

and

<9> $$K_y\{x : Tx \neq y\} = 0 \qquad \text{a.e.}[\mathbb{Q}].$$

Let me show you how this conditioning problem is related to the results from Section 2.

To simplify notation, write $\langle\cdot,\cdot\rangle_{\mathbb{Q}}$ and $\|\cdot\|_{\mathbb{Q}}$ for the $\mathcal{L}^2(\mathbb{Q}) := \mathcal{L}^2(\mathcal{Y}, \mathcal{B}, \mathbb{Q})$ inner product and norm, with a similar convention for $\mathcal{L}^2(\mathbb{P}) := \mathcal{L}^2(\mathcal{X}, \mathcal{A}, \mathbb{P})$.

It is particularly informative to consider functions of the form $f(x) = g(Tx)h(x)$, with $g \in \mathcal{L}^2(\mathbb{Q})$ and $h \in \mathcal{L}^2(\mathbb{P})$. By Cauchy-Schwarz,

$$(\mathbb{P}|g(Tx)h(x)|)^2 \le \left(\mathbb{P}g(Tx)^2\right)\left(\mathbb{P}h(x)^2\right) = \|g\|_{\mathbb{Q}}^2\,\|h\|_{\mathbb{P}}^2 < \infty.$$

That is, $f \in \mathcal{L}^1(\mathbb{P})$. For each fixed h in $\mathcal{L}^2(\mathbb{P})$, the linear functional τ_h that maps g to $\mathbb{P}g(Tx)h(x)$ is well defined. Again by Cauchy-Schwarz we have

$$|\tau_h(g)| \le \|g\|_{\mathbb{Q}}\,\|h\|_{\mathbb{P}}\,,$$

which shows τ_h is continuous. The argument in Section 2 provides a function H in $\mathcal{L}^2(\mathbb{Q})$, which is unique up to $\mathbb{Q}$-equivalence, such that

<10> $$\tau_h(g) = \langle g, H\rangle_{\mathbb{Q}} \qquad \text{for all } g \in \mathcal{L}^2(\mathbb{Q}).$$

Now suppose a Markov kernel $\mathbb{K}$ satisfying equalities <8> and <9> exists. We can extend <8> to functions f in $\mathcal{L}^1(\mathcal{X}, \mathcal{A}, \mu)$ by subtraction of the equalities for f^+ and f^-.

Remark. The equality $\mathbb{P}|f| = \mathbb{Q}^y K_y^x |f(x)|$ implies that $K_y^x|f(x)| < \infty$ a.e.$[\mathbb{Q}]$. We could possibly have $K_y f^+ = K_y f^- = \infty$ for a $\mathbb{Q}$-negligible set of y's, a situation analogous to the one studied in HW7.1. A similar fix is possible.

In particular

$$\begin{aligned}
\mathbb{P}^x g(Tx)h(x) &= \mathbb{Q}^y K_y^x \left(g(Tx)h(x)\right) \qquad \text{by <8>} \\
&= \mathbb{Q}^y g(y) K_y^x h(x) \qquad \text{by <9>} \\
&= \langle g, H_1\rangle_{\mathbb{Q}} \qquad \text{where } H_1(y) = K_y^x h(x).
\end{aligned}$$

The function H_1 might take infinite values, but only on a $\mathbb{Q}$-negligible set: by Cauchy-Schwarz for K_y we have $H_1(y)^2 \le K_y^h(x)^2$ and

$$\mathbb{Q}\left(H_1(y)^2\right) \le \mathbb{Q}^y K_1^y h(x)^2 = \mathbb{P}h^2 < \infty.$$

With only $\mathbb{Q}$-negligible guilt I declare H_1 to be a candidate for the representation <10> of τ_h.

Let us ponder. You know, for each fixed h in $\mathcal{L}^2(\mathbb{P})$, that there always exists an H (unique up to $\mathbb{Q}$-equivalence) in $\mathcal{L}^2(\mathbb{Q})$ for which

<11> $$\mathbb{P}g(Tx)h(x) = \mathbb{Q}g(y)H(y) \qquad \text{for all } g \text{ in } \mathcal{L}^2(\mathbb{Q}).$$

You also know that, if conditional distributions exist, we can take H as $K_y(h)$, or perhaps $(K_y h)\,\mathbb{1}\{y \in \mathcal{Y} : \mathbb{Q}(K_y h^2 < \infty\}$ if you worry about the infinities. Put another way $\mathbb{P}(h \mid Y = y) = H(y)$ a.e.$[\mathbb{Q}]$, where $\mathbb{P}(\cdot \mid Y = y)$ is tradition shorthand for K_y.

If you are only interested in $\mathbb{P}(h \mid Y = y)$ for a single function h (or a countable set of h's) in $\mathcal{L}^2(\mathbb{P})$ it might appear unnecessary to worry about existence of $\mathbb{P}(\cdot \mid Y = y)$ as a set of probability measures on $\mathcal{A}$. Instead you could regard the conditional expected value $\mathbb{P}(h \mid Y = y)$, for a single h in $\mathcal{L}^2(\mathbb{P})$, as the value at y of function that maps $\mathcal{L}^2(\mathbb{P})$ into $\mathcal{L}^2(\mathbb{Q})$. Of course you would need to extend the map to cover functions h in $\mathcal{L}^1(\mathbb{P})$ if

you wanted a conditional expectation operator that didn't restrict its domain to square integrable functions.

The previous paragraph essentially describes the approach described by Kolmogorov (1933, Chapter 5) to define a conditional expectation map, except that he invoked the Radon-Nikodym theorem (see Section 7) to prove existence.

4 Conditioning à la Kolmogogorov

Let me formalize the ideas introduced at the end of the previous Section.

Again assume we have probability spaces $(\mathcal{X}, \mathcal{A}, \mathbb{P})$ and $(\mathcal{Y}, \mathcal{B}, \mathbb{Q})$ and an $\mathcal{A}\backslash\mathcal{B}$-measurable map T from $\mathcal{X}$ into $\mathcal{Y}$ whose distribution under $\mathbb{P}$ is $\mathbb{Q}$. That is

$$\mathbb{Q}g = \mathbb{P}g(Tx) \qquad \text{at least for } g \in \mathbb{M}^+(\mathcal{Y}, \mathcal{B}).$$

The following details show how to construct a map $\kappa : \mathcal{M}^+(\mathcal{X}, \mathcal{A}) \to \mathbb{M}^+(\mathcal{Y}, \mathcal{B})$ with properties analogous to those for expectation with respect to a Markov kernel. To avoid a lot of messy parentheses I'll write $\kappa(y, h)$, or $\kappa(y, h(x))$, for the value of $\kappa(h)$ at y.

<12> **Theorem.** *There exists a map $\kappa : \mathcal{M}^+(\mathcal{X}, \mathcal{A}) \to \mathbb{M}^+(\mathcal{Y}, \mathcal{B})$ for which*

<13> $$\mathbb{P}g(Tx)h(x) = \mathbb{Q}g(y)\kappa(y, h) \qquad \text{for all } h \in \mathcal{M}^+(\mathcal{X}, \mathcal{A}) \text{ and } g \in \mathcal{M}^+(\mathcal{Y}, \mathcal{B}).$$

The map has the following properties.

(i) For each fixed h, equality <13> uniquely determines $\kappa(y, h)$ up to $\mathbb{Q}$-equivalence.

(ii) $\mathcal{K}(y, 0) = 0$ and $\mathcal{K}(y, \mathbb{1}) = 1$ a.e.$[\mathbb{Q}]$.

(iii) $\mathcal{K}(y, g_1(Tx)h_1(x) + g_2(Tx)h_2(x)) = g_1(y)\mathcal{K}(y, h_1) + g_2(y)\mathcal{K}(y, h_2)$ a.e.$[\mathbb{Q}]$ *for all functions $g_i \in \mathcal{M}^+(\mathcal{Y}, \mathcal{B})$ and $h_i \in \mathcal{M}^+(\mathcal{X}, \mathcal{A})$.*

(iv) If $h_1 \le h_2$ a.e.$[\mathbb{P}]$ *then $\kappa(y, h_1) \le \kappa(y, h_2)$* a.e.$[\mathbb{Q}]$.

(v) If $h_n \in \mathcal{M}^+(\mathcal{X}, \mathcal{A})$ and $0 \le h_n(x) \le h_{n+1}(x) \uparrow h(x)$ as $n \to \infty$ a.e.$[\mathbb{P}]$ *then $\kappa(y, h_n) \uparrow \kappa(y, h)$* a.e.$[\mathbb{Q}]$.

We would get very similar properties if $\kappa(y, h) = K_y^x h(x)$ for a Markov kernel $\mathbb{K}$ satisfying <8> and <9>. Notable by its absence is an assumption corresponding to $K_y\{x \in \mathcal{X} : Tx \ne y\}$ a.e.$[\mathbb{Q}]$. In its place is property (iii),

which asserts that functions $g_i(Tx)$ behave like constants as far as κ is concerned. On the plus side, we do not need any assumptions about beyond the $\mathcal{A}\backslash\mathcal{B}$-measurability.

Remark. The Theorem could be thought of as the first step towards construction of a true Markov kernel that provides conditional distributions. The main difficulty would then be to consolidate the uncountably many a.e.$[\mathbb{Q}]$ qualifications into a single $\mathbb{Q}$-negligible set.

WARNING: Many authors would write $\mathbb{P}(h \mid T = y)$ for my function $\kappa(y, h)$, mimicking the notation for the integral with respect to a true conditional probability distribution $\mathbb{P}(\cdot \mid T = y)$. See Section 5 for one of the dangers of treating the Kolmogorov conditional expectation operator as if it had all the properties of an expectation with respect to a conditional probability distribution.

PROOF (OF THEOREM <12>) Homework 8 provides the tool needed for the proof: If $G_1, G_2 \in \mathcal{M}^+(\mathcal{Y}, \mathcal{B})$ have the property that

<14> $$\mathbb{Q}G_1(y)\mathbb{1}\{y \in B\} \le \mathbb{Q}G_1(y)\mathbb{1}\{y \in B\} \qquad \text{for all } B \text{ in } \mathcal{B}$$

then $G_1(y) \le G_2(y)$ a.e.$[\mathbb{Q}]$. In particular, if

<15> $$\mathbb{Q}G_1(y)\mathbb{1}\{y \in B\} = \mathbb{Q}G_1(y)\mathbb{1}\{y \in B\} \qquad \text{for all } B \text{ in } \mathcal{B}$$

then $G_1(y) = G_2(y)$ a.e.$[\mathbb{Q}]$. (Apply <14> with the roles of G_1 and G_2 reversed to get $G_1(y) \ge G_2(y)$ a.e.$[\mathbb{Q}]$.)

Back to the proof of the Theorem. The main challenge is to construct, for each h in $\mathcal{M}^+(\mathcal{X}, \mathcal{A})$, just one function $\kappa(y, h)$ for which <13> holds. Property (i) will then follow directly from <15>.

For property (ii) the choice is easy: if $h(x) = c \in \mathbb{R}^+$ define $\kappa(y, h) = c$, the constant function taking value c on $\mathcal{Y}$. Equality <13> then reduces to the fact that $\mathbb{Q}$ is the image of $\mathbb{P}$ under T.

To construct $\kappa(y, h)$ in general, start with the analogous $\mathcal{L}^2$ property given by equality <11>, namely, for each $h \in \mathcal{L}^2(\mathbb{P})$ we have an $H \in \mathcal{L}^2(\mathbb{Q})$ for which

$$\mathbb{P}g(Tx)h(x) = \mathbb{Q}g(y)H(y) \qquad \text{for all } g \text{ in } \mathcal{L}^2(\mathbb{Q}).$$

Write $M(y, h)$ for the function $H(y)$. That is, M is regarded as a map from $\mathcal{L}^2(\mathbb{P})$ into $\mathcal{L}^2(\mathbb{Q})$, which is only unique up to $\mathbb{Q}$-equivalence for each fixed h.

To pass from <11> to a suitable choice for κ we need first to prove an $\mathcal{L}^2$ analog of (iv): If $h_1, h_2 \in \mathcal{L}^2(\mathbb{P})$ and $0 \le h_1 \le h_2$ a.e.$[\mathbb{P}]$ then $0 \le M(y,h_1) \le M(y,h_2)$ a.e.$[\mathbb{Q}]$. To this end, invoke <11> with $g(y) = \mathbb{1}\{y \in B\}$ with $B \in \mathcal{B}$.

$$\begin{aligned}\mathbb{Q}\mathbb{1}\{y \in B\}M(y,h_1) &= \mathbb{P}\mathbb{1}\{Tx \in B\}h_1(x) \\ &\le \mathbb{P}\mathbb{1}\{Tx \in B\}h_2(x) \qquad \text{because } h_1 \le h_2 \text{ a.e.}[\mathbb{P}] \\ &= \mathbb{Q}\mathbb{1}\{y \in B\}M(y,h_2).\end{aligned}$$

All integrals in the last display are non-negative because $h_1 \ge 0$ a.e.$[\mathbb{P}]$. To get $0 \le M(y,h_1)$ a.e.$[\mathbb{Q}]$ invoke <14> with $G_1 = 0$ and $G_2(y) = M(y,h_1)$. To get $M(y,h_1) \le M(y,h_2)$ a.e.$[\mathbb{Q}]$ invoke <14> with $G_1 = (y,h_1)$ and $G_2(y) = M(y,h_2)$.

Now we have everything needed to construct κ. Suppose $g \in \mathcal{M}^+(\mathcal{Y},\mathcal{B})$ and $h \in \mathcal{M}^+(\mathcal{X},\mathcal{A})$. For all $m,n \in \mathbb{N}$ the function $g_m(y) = \min(m, g(y))$ belongs to $\mathcal{L}^2(\mathbb{Q})$ and the function $h_n(x) = \min(n, h(x))$ belongs to $\mathcal{L}^2(\mathbb{P})$. From <11>,

$$\mathbb{P}g_m(Tx)h_n(x) = \mathbb{Q}g_m(y)M(y,h_n).$$

Define $\kappa(y,h) = \sup_n M(y,h_n)$. From the previous paragraph we know that $0 \le M(y,h_n) \uparrow \kappa(y,h)$ a.e.$[\mathbb{Q}]$. Let n tend to ∞ with m held fixed, appealing twice to Monotone Convergence to get

$$\mathbb{P}g_m(Tx)h(x) = \mathbb{Q}g_m(y)\kappa(y,h).$$

Then let m tend to ∞, with two more appeals to Monotone Convergence to get <13>.

For property (iii), temporarily write $f(x)$ for $g_1(Tx)h_1(x) + g_2(Tx)h_2(x)$. By definition, $\kappa(y,f)$ is determined up to $\mathbb{Q}$-equivalence by the first in the following string of equalities. For each $G \in \mathcal{M}^+(\mathcal{Y},\mathcal{B})$,

$$\begin{aligned}&\mathbb{Q}G(y)\kappa(y,f) \\ &\quad= \mathbb{P}G(Tx)f(x) \\ &\quad= \mathbb{P}G(Tx)g_1(Tx)h_1(x) + \mathbb{P}G(Tx)g_2(Tx)h_2(x) \\ &\quad= \mathbb{Q}G(y)g_1(y)\kappa(y,h_1) + \mathbb{Q}G(y)g_2(y)\kappa(y,h_2) \\ &\quad= \mathbb{Q}G(y)H(y) \qquad \text{where } H(y) := g_1(y)\kappa(y,h_1) + g_2(y)\kappa(y,h_2).\end{aligned}$$

It follows by <15> that $\kappa(y,f) = H(y)$ a.e.$[\mathbb{Q}]$.

The argument for property (iv) is essentially the same as ints $\mathcal{L}^2$ analog. For each $G \in \mathcal{M}^+(\mathcal{Y}, \mathcal{B})$,

$$\mathbb{Q}G(y)\kappa(y, h_1) = \mathbb{P}G(Tx)h_1(x) \le \mathbb{P}G(Tx)h_2(x) = \mathbb{Q}G(y)\kappa(y, h_2).$$

By <14> we have $\kappa(y, h_1) \le \kappa(y, h_2)$ a.e.$[\mathbb{Q}]$.

For property (v) first note that, by property (iv),

$$0 \le \kappa(y, h_n) \uparrow H(y) := \sup_n \kappa(y, h_n) \qquad \text{a.e.}[\mathbb{Q}].$$

Then observe that, for each $g \in \mathcal{M}^+(\mathcal{Y}, \mathcal{B})$,

$$\begin{aligned} \mathbb{P}g(Tx)h(x) &= \lim_{n\to\infty} \mathbb{P}g(Tx)h_n(x) \qquad \text{by Monotone Convergence} \\ &= \lim_{n\to\infty} \mathbb{Q}g(y)\kappa(y, h_n) \\ &= \mathbb{Q}g(y)H(y) \qquad \text{by Monotone Convergence,} \end{aligned}$$

which identifies $H(y)$ as one possible choice for $\kappa(y, h)$.

□

I leave it to you to figure out how to extend κ to functions in $\mathcal{L}^1(\mathcal{X}, \mathcal{A}, \mathbb{P})$. The only (minor) difficulties involve a $\mathbb{Q}$-negligible set of cases where $\kappa(y, h^+)$ or $\kappa(y, h^-)$ (or both) take the value $+\infty$.

5 Dangers

As in Section 3, suppose $(\mathcal{X}, \mathcal{A}, \mathbb{P})$ and $(\mathcal{Y}, \mathcal{B}, \mathbb{Q})$ are probability spaces and T is an $\mathcal{A}\backslash\mathcal{B}$-measurable map from $\mathcal{X}$ to $\mathcal{Y}$ whose distribution under $\mathbb{P}$ equals $\mathbb{Q}$.

To begin with, suppose $\mathbb{K} = \{K_y : y \in \mathcal{Y}\}$ is a condition distribution in the sense of <8> and <9>. That is, there exists a $\mathbb{Q}$-negligible set N for which $K_y\{Tx \ne y\} = 0$ for each y in N^c and $\mathbb{P}h = \mathbb{Q}^y K_y^x h(x)$ for each h in $\mathcal{M}^+(\mathcal{X}, \mathcal{A})$.

Consider h's of the form $h(x) = f(x, Tx)$ with $f \in \mathcal{M}^+(\mathcal{X} \times \mathcal{Y}, \mathcal{A} \otimes \mathcal{B})$. By the concentration property

$$K_y^x f(x, Tx) = K_y^x f(x, y) \qquad \text{for each } y \text{ in } N^c.$$

This assertion holds simultaneously for every such f. If we use the suggestive notation $\mathbb{P}(\cdot \mid T = y\}$ for the probability measure K_y then the result becomes

$$\mathbb{P}(f(x, Tx) \mid T = y) = \mathbb{P}(f(x, y) \mid T = y) \qquad \text{for all } y \in N^c, \text{ all } f \in \mathcal{M}^+(\mathcal{X} \times \mathcal{Y}, \mathcal{A} \otimes \mathcal{B}).$$

A similar assertion need not be true if conditional distributions do not exist. In that situation we need to use the Kolmogorov conditional expectation. That is, we need to rely on existence of a map $\kappa : \mathcal{M}^+(\mathcal{X}, \mathcal{A}) \to \mathcal{M}^+(\mathcal{Y}, \mathcal{B})$ that has the properties described in Theorem <12>. Implicitly that means we have made some choice of $\mathcal{K}h$ from the $\mathbb{Q}$-equivalence class of possibilities for each h in $\mathcal{M}^+(\mathcal{X}, \mathcal{A})$. I claim it need not be true that

<16> $$\kappa(y, f(x, Tx)) \stackrel{?}{=} \kappa(y, f(x, y)) \qquad \text{a.e.}[\mathbb{Q}]??.$$

The question marks are to make sure you do not regard the assertion as true if you happen to refer back to it at some stage.

Remark. I am interpreting <16> (without all the question marks) as an assertion about every map κ, from $\mathcal{M}^+(\mathcal{X}, \mathcal{A})$ into $\mathcal{M}^+(\mathcal{Y}, \mathcal{B})$, that has property <13>.

Indeed, the map κ would interpret the $f(x, y)$ in <16> as a function of x that depends on some constant y. It helps to separate the two roles played by y in <16> by writing $f(x, t)$ or $f_t(x)$ instead of $f(x, y)$. If we define $H_t = \kappa(f_t)$ for each $t \in \mathcal{Y}$ then the equality <16> could be interpreted to assert

$$\kappa(y, f(x, Tx)) \stackrel{?}{=} H_y(y)$$

for a suitably large collection of y's.

Right away you can see that the right-hand side could cause measurability trouble because we have no reason to believe $H_t(y)$ depends on t in a nice way. Moreover, each H_t is just one function chosen somewhat arbitrarily from an equivalence class of possibilities. If $\mathbb{Q}\{t\} = 0$ for each t in $\mathcal{Y}$, we would not violate the assumptions of Theorem<12> by redefining $H_t(t) = 0$ for every t. That would cause obvious problems for <16>.

A more concrete example might clarify the issues in the previous paragraph.

<17> **Example.** Suppose $\mathbb{P} = N(0, I_2)$, the standard bivariate distribution on $\mathcal{A} := \mathcal{B}(\mathbb{R}^2)$. It has density $p(x) = (2\pi)^{-1} \exp(-|x|^2/2)$ with respect to Lebesgue measure on $\mathcal{A}$. Define $T(x) = |x|^2/2$, a suitably measurable map from $\mathcal{X} := \mathbb{R}^2$ into $\mathcal{Y} := \mathbb{R}^+$.

It is easy to show—for example, you could use the same Tonelli trick that gave the $\sqrt{2\pi}$ for the normal density—that the distribution, $\mathbb{Q}$, of T under $\mathbb{P}$ has density $q(y) = e^{-y}$ with respect to Lebesgue measure on $\mathcal{B} := \mathcal{B}(\mathcal{Y})$.

It seems natural to define $\kappa(h)$ to be the zero function on $\mathcal{Y}$ for each h in $\mathcal{M}^+(\mathcal{X},\mathcal{A})$ with $h(x)=0$ a.e.$[\mathbb{P}]$, because then we have

$$\mathbb{Q}^y g(y)\kappa(y,h) = 0 = \mathbb{P}g(Tx)h(x) \qquad \text{for each } g \text{ in } \mathcal{M}^+(\mathcal{Y},\mathcal{B}).$$

Such a choice does not violate any of the assumptions of Theorem <12>.

For each $t \in \mathcal{Y}$ define $f(x,t) = f_t(x) = \mathbb{1}\{Tx = t\}$. Note that the set $\{x : Tx = t\}$ is the circle of radius $\sqrt{2t}$ around the origin, a set that has zero $\mathbb{P}$ measure. Consequently, by the natural choice from the previous paragraph, we have $H_t := \kappa(f_t)$ equal to the zero function. We also have $f(x,Tx) = 1$ for every x, which forces $\kappa(y, f(x,Tx)) = 1$ a.e.$[\mathbb{Q}]$. We have an extreme violation of the assertion in <16>.

Remark. You might want to dismiss Example <17> as a trifle caused by too cavalier a treatment of uncountably many negligible sets. It is certainly possible to make a more careful choice of $\kappa(f_t)$ to rescue <16>. Indeed, it possible to get a true conditional distribution: take K_y to be the uniform distribution on the circle $\{x : Tx = y\}$. That is actually the whole point. Kolmogorov conditional expectation maps allow us to ignore $\mathbb{P}$-negligible changes in each $h(x)$ function; conditional distributions require us to handle uncountably many negligible sets in a clever way.

6 Bringing it all back to $\mathcal{X}$

Theorem <12> treated the case of probability spaces $(\mathcal{X},\mathcal{A},\mathbb{P})$ and $(\mathcal{Y},\mathcal{B},\mathbb{Q})$ and an $\mathcal{A}\backslash\mathcal{B}$-measurable map T whose distribution under $\mathbb{P}$ was equal to $\mathbb{Q}$. It gave existence of a map $\kappa : \mathcal{M}^+(\mathcal{X},\mathcal{A}) \to \mathbb{M}^+(\mathcal{Y},\mathcal{B})$ for which

$$\mathbb{P}g(Tx)h(x) = \mathbb{Q}g(y)\kappa(y,h) \qquad \text{for all } h \in \mathcal{M}^+(\mathcal{X},\mathcal{A}) \text{ and } g \in \mathcal{M}^+(\mathcal{Y},\mathcal{B}).$$

If we write $H(y)$ for $\kappa(y,h)$ then use the fact that $\mathbb{Q}G(y) = \mathbb{P}G(Tx)$ for all $G \in \mathcal{M}^+(\mathcal{Y},\mathcal{B})$ we get

$$\mathbb{P}g(Tx) = \mathbb{Q}g(y)H(y) = \mathbb{P}g(Tx)H(Tx).$$

The functions $x \mapsto g(Tx)$ and $x \mapsto H(Tx)$ both belongs to $\mathcal{M}^+(\mathcal{X},\mathcal{B}_0)$, where

$$\mathcal{B}_0 = \{T^{-1}B : B \in \mathcal{B}\} = \sigma(T).$$

Define a map $\widetilde{\kappa} : \mathcal{M}^+(\mathcal{X},\mathcal{A}) \to \mathcal{M}^+(\mathcal{X},\mathcal{B}_0)$ by setting $\widetilde{\kappa}(x,f) := H(Tx)$ if $\kappa f = H$. Then

$$\mathbb{P}G(x)f(x) = \mathbb{P}G(x)\widetilde{\kappa}(x,f) \qquad \text{for all } G \in \mathcal{M}^+(\mathcal{X},\mathcal{B}_0).$$

In currently popular notation, the function $\widetilde{\kappa} f$ (on $\mathcal{X}$) would be written as $\mathbb{E}(f \mid T)$, which I regard as an archaic way of writing $\mathbb{P}(f \mid T)$. The value of $\mathbb{P}(f \mid T)$ at x equals $\widetilde{\kappa}(x, f)$.

Remark. Note the difference between $H(y) = \kappa(y, f) = \mathbb{P}(f \mid T = y)$ and $H(Tx) = \widetilde{\kappa}(x, f) = (\mathbb{P}(f \mid T))(x)$.

The $\mathcal{M}^+(\mathcal{X}, \mathcal{B}_0)$ function $\mathbb{P}(f \mid T)$ can also written as $\mathbb{P}(f \mid \mathcal{B}_0)$ or $\mathbb{P}_{\mathcal{B}_0}(f)$. That is, $\mathbb{P}_{\mathcal{B}_0}$ maps $\mathcal{M}^+(\mathcal{X}, \mathcal{A})$ into $\mathcal{M}^+(\mathcal{X}, \mathcal{B}_0)$.

Notice that the map T seems to have disappeared from the notation, with only the $\mathcal{B}_0$ as a reminder of the role that T played. That reminder becomes totally redundant if use a cunning notational trick.

Suppose $\mathcal{B}$ is actually a sub-sigma-field of $\mathcal{A}$, that is, $\mathcal{B}$ actually lives on $\mathcal{X}$. We could take $\mathcal{Y}$ to be $\mathcal{X}$ and regard T as the identity map on $\mathcal{X}$, that is, $Tx = x$. Then $\mathcal{B}_0$ would be exactly the same as $\mathcal{B}$, and $\mathcal{M}^+(\mathcal{Y}, \mathcal{B})$ would be the same as $\mathcal{M}^+(\mathcal{X}, \mathcal{B})$, and $\mathbb{P}_{\mathcal{B}}$ would be a map from $\mathcal{M}^+(\mathcal{X}, \mathcal{A})$ into $\mathcal{M}^+(\mathcal{X}, \mathcal{B})$ for which

<18> $$\mathbb{P}(fG) = \mathbb{P}\left[(\mathbb{P}_{\mathcal{B}} f)G\right] \qquad \text{for all } G \in \mathcal{M}^+(\mathcal{X}, \mathcal{B}).$$

The function $\mathbb{P}_{\mathcal{B}} f$ is called the Kolmogorov conditional expectation of f given the sub-sigma-field $\mathcal{B}$. It inherits from Theorem <12> the following properties.

(i) For each fixed f in $\mathcal{M}^+(\mathcal{X}, \mathcal{A})$ the equality <18> uniquely determines $\mathbb{P}_{\mathcal{B}} f$ up to $\mathbb{P}$-equivalence.

(ii) $\mathbb{P}_{\mathcal{B}} 0 = 0$ and $\mathbb{P}_{\mathcal{B}} 1 = 1$ a.e.$[\mathbb{P}]$.

(iii) $PP_{\mathcal{B}}(G_1 f_1 G_2 f_2) = G_1 \mathbb{P}_{\mathcal{B}} f_1 + G_2 \mathbb{P}_{\mathcal{B}} f_2$ a.e.$[\mathbb{P}]$ for all $G_i \in \mathcal{M}^+(\mathcal{X}, \mathcal{B})$ and $f_i \in \mathcal{M}^+(\mathcal{X}, \mathcal{A})$.

(iv) If $f_1 \le f_2$ a.e.$[\mathbb{P}]$ then $\mathbb{P}_{\mathcal{B}} f_1 \le \mathbb{P}_{\mathcal{B}} f_2$ a.e.$[\mathbb{P}]$.

(v) If $f_n \in \mathcal{M}^+(\mathcal{X}, \mathcal{A})$ and $0 \le f_1 \le f_2 \ldots \uparrow f$ as $n \to \infty$ a.e.$[\mathbb{P}]$ then $\mathbb{P}_{\mathcal{B}} f_n \uparrow \mathbb{P}_{\mathcal{B}} f$ a.e.$[\mathbb{P}]$.

Remark. Here I am using $\mathbb{P}$ to denote both a measure on $\mathcal{A}$ and its restriction to $\mathcal{B}$. Would it help to make up a new name for the restriction?

<19> **Example.** Suppose $(\mathcal{X}, \mathcal{A}, \mathbb{P})$ is a probability space and we have sub-sigma-fields $\mathcal{B}_2 \subset \mathcal{B}_1 \subset \mathcal{A}$. Show that

$$\mathbb{P}_{\mathcal{B}_2}(\mathbb{P}_{\mathcal{B}_1} f) = \mathbb{P}_{\mathcal{B}_2} f \qquad \text{a.e.}[\mathbb{P}] \text{ for each } f \text{ in } \mathcal{M}^+(\mathcal{X}, \mathcal{A}).$$

Define $h = \mathbb{P}_{\mathcal{B}_1} f$ and $k = \mathbb{P}_{\mathcal{B}_2} h$. By definition

$$\mathbb{P}G_1 f = \mathbb{P}G_1 h \qquad \text{for all } G_1 \in \mathcal{M}^+(\mathcal{X}, \mathcal{B}_1)$$
$$\mathbb{P}G_2 h = \mathbb{P}G_2 k \qquad \text{for all } G_2 \in \mathcal{M}^+(\mathcal{X}, \mathcal{B}_2)$$

Replace G_1 by G_2 in the first equality to deduce

$$\mathbb{P}G_2 f = \mathbb{P}G_2 h = \mathbb{P}G_2 k.$$

Thus k has the two properties that identify it as a possible value for $\mathbb{P}_{\mathcal{B}_2} f$. Equality a.e.$[\mathbb{P}]$ follows.

□

7 Radon-Nikodym theorem

The simplest form of the theorem concerns two finite measures μ and ν defined on some $(\mathcal{X}, \mathcal{A})$.

<20> **Theorem.** *If $\mu\mathcal{X} < \infty$ and $\mu f \geq \nu f$ for each f in $\mathcal{M}^+(\mathcal{X}, \mathcal{A})$ then there exists an $\mathcal{A}$-measurable function Δ with $0 \leq \Delta(x) \leq 1$ for all x such that*

$$\nu f = \mu(f\Delta) \qquad \text{for each } f \text{ in } \mathcal{M}^+(\mathcal{X}, \mathcal{A}).$$

That is, the measure ν has density Δ with respect to the measure μ. The function Δ is unique up to μ-equivalence.

Proof (sketch of a proof due to von Neumann, 1940, page 127)

Write $\|\cdot\|_\mu$ and $\langle \cdot, \cdot \rangle_\mu$ for the norm and inner product on $\mathcal{L}^2 := \mathcal{L}^2(\mathcal{X}, \mathcal{A}, \mu)$. Regard ν as a linear functional on $\mathcal{L}^2$. By Cauchy-Schwarz, the functional is continuous, in the sense of Example <2>:

$$|\nu f| \leq \nu |f\mathbb{1}| \leq \sqrt{(\nu f^2)(\nu \mathbb{1}^2)} \leq \|f\|_\mu \sqrt{\nu \mathcal{X}} \qquad \text{for all} \in \mathcal{L}^2.$$

Section 2 tells us that there exists a function k in $\mathcal{L}^2$ for which

$$\nu f = \langle f, k \rangle_\mu = \mu(fk) \qquad \text{for all} \in \mathcal{L}^2.$$

I'll show that, $0 \leq k \leq 1$ a.e.$[\mu]$.

For each $\epsilon > 0$,

$$0 \le \nu\{k \le -\epsilon\} = \mu k\{k \le -\epsilon\} \le -\epsilon\mu\{k \le -\epsilon\}.$$

It follows that $\mu\{k \le -\epsilon\} = 0$ for each $\epsilon > 0$ and hence $\mu\{k < 0\} = 0$. Similarly, the inequality

$$\mu\{k \ge 1+\epsilon\} \ge \nu\{k \ge 1+\epsilon\} = \mu k\{k \ge 1+\epsilon\} \ge (1+\epsilon)\mu\{k \ge 1+\epsilon\}$$

implies $\mu\{k \ge 1+\epsilon\} = 0$ for each $\epsilon > 0$ and hence $\mu\{k > 1\} = 0$. Define $\Delta = k\mathbb{1}\{0 \le k \le 1\}$. Then $\nu f = \mu(f\Delta)$ for all $f \in \mathcal{L}^2$.

For functions f in $\mathcal{M}^+(\mathcal{X}, \mathcal{A})$ we have $f \wedge n \in \mathcal{L}^2$ so that

$$\nu(f \wedge n) = \mu((f \wedge n)\Delta).$$

Let n tend to infinity, appealing twice to Monotone Convergence to deduce that $\nu f = \mu f\Delta)$.

The uniqueness a.e.$[\mu]$ of Δ follows from the μ analog of <15>.

□

Theorem <20> has an extension to sigma-finite measures μ and ν with ν ***dominated*** by μ, that is: for each $A \in \mathcal{A}$, if $\mu A = 0$ then $\nu A = 0$.

Remark. Domination is sometimes expressed as "ν is absolutely continuous with respect to μ", which is often denoted by $\nu \ll \mu$. This terminology borrows from the classical concept of absolute continuity of a function defined on the real line (Pollard, 2001, Section 3.4). The density Δ is often denoted by $d\nu/d\mu$ and is called the ***Radon-Nikodym derivative of ν with respect to μ***.

<21> **Theorem.** *If μ and ν are both sigma-finite measures with ν dominated by μ then there exists a real-valued function $\Delta_0 \in \mathcal{M}^+(\mathcal{X}, \mathcal{A})$ for which*

$$\nu f = \mu(f\Delta_0) \qquad \text{for each } f \text{ in } \mathcal{M}^+(\mathcal{X}, \mathcal{A}).$$

The function Δ is unique up to μ-equivalence.

Proof (Sketch. For details see Pollard (2001, Section 3.2).)

The idea is to first treat the case where μ and ν are finite measures, with $\nu \ll \mu$, by appealing to Theorem <20> with μ replaced by $\lambda = \mu + \nu$. One argues from the inequality

$$\nu\{\Delta \ge 1\} = \nu\Delta\{\Delta \ge 1\} + \mu\Delta\{\Delta \ge 1\}$$

that $\mu\{\Delta \geq 1\} = 0$, which implies $\nu\{\Delta \geq 1\} = 0$. For $h \in \mathcal{M}^+(\mathcal{X}, \mathcal{A})$ and $n \in \mathbb{N}$ one needs to rearrange the equality

$$\nu f = \nu(f\Delta) + \mu(f\Delta) \qquad \text{for } f = \frac{(h \wedge n)\mathbb{1}\{\Delta \leq 1 - n^{-1}\}}{1 - \Delta}$$

to get $\nu f(1 - \Delta) = \mu f\Delta$. Two appeals to Monotone Convergence then give $\nu h = \mu h\Delta_0$ with

$$\Delta_0 = \frac{\Delta}{1 - \Delta}\mathbb{1}\{0 \leq \Delta < 1\}.$$

The sigma-finite version of the Theorem is then proved by partitioning $\mathcal{X}$ into a sequence of sets $\mathcal{X}_i$ for which $\nu\mathcal{X}_i + \mu\mathcal{X}_i < \infty$ and applying the result for finite ν and μ on each $\mathcal{X}_i$.

□

References

Kolmogorov, A. N. (1933). *Grundbegriffe der Wahrscheinlichkeitsrechnung.* Berlin: Springer-Verlag. Second English Edition, *Foundations of Probability* 1950, published by Chelsea, New York.

Pollard, D. (2001). *A User's Guide to Measure Theoretic Probability.* Cambridge University Press.

Steele, J. M. (2004). *The Cauchy-Schwarz Master Class: An Introduction to the Art of Mathematical Inequalities.* Cambridge University Press.

von Neumann, J. (1940). On rings of operators. III. *Annals of Mathematics 41*(1), 94–161.