Lawvere Archive Posthumous Publication – July 14, 2025

The Category of Probabilistic Mappings

With Applications to Stochastic Processes, Statistics, and Pattern Recognition*

F. William Lawvere

Abstract (June 27, 2020). This brief article is being published now in response to numerous requests, most of which come from researchers wanting to apply functorial methods to probability and statistics. Thus there is a natural curiosity concerning the origins and further developments of this work and of the ideas involved in it. It does include several programmatic principles, such as

a. the need for an intrinsic metric on each hom set (so as to describe that a statistical decision function or powerful test is only correct within epsilon);

b. the role of this category as a base for diagram categories, Kan extensions, etc., (for example the category of Markov processes has as objects the endomorphisms in this category);

c. an attempt at a rational framework for the representations of general stochastic processes in terms of Markov processes.

It is unusually brief because it is a fragment of an introduction to an Appendix of a much larger document that was classified as SECRET. The origin of this document is closely intertwined with historical events that occurred in the mid-1960s. The first section is simply a review of standard material on sigma algebras and probability measures.
Section 2 establishes that there is an adjointness between deterministic and probabilistic mappings, hence there is, by composition, a monad, (according to now-standard terminology) whose unit is traditionally named after Dirac.
Section 3 contains reasonable proposals for treating decision problems, stochastic processes, and Markov processes and the relations between them. Now, I would emphasize enlarging from the Kleisli category to the Eilenberg–Moore category of such a monad, in order to facilitate the search for the indicated adjoints, and also to emphasize that there is an intrinsic measure of distance.
Considerable progress on these questions was made by my student Xiao-Qing Meng in her thesis, in particular, generalizing the temporal transitions in stochastic processes, and especially showing that the assignment of an intrinsic metric to convex sets is a monoidal functor, hence that the whole enriched category theory over convex sets (in particular those given by probabilistic generators and relations) is naturally re-enriched by a notion of nearness.

* Spring 1962

Author Commentary (July 2, 2020)

If I had to do this basic paper now, I could omit the Nuts and Bolts of σ-algebras (Section 1). Instead I would start with an initial postulate expressing the actual intention of an Averaging Monad: $P(A) =$ the part of $\mathrm{Hom}_R(R^A, R)$ consisting of those functionals that retract along $A \to 1$, thus preserving constants. Probabilistic mappings can be viewed as a refinement of classical logical relations (a quotient $P \to P_0$ of the relevant monad has as its Kleisli category the category whose morphisms are "possibilistic maps" or (finitistic) relations in the usual sense of logic). Such monads are typically on a Cartesian category $\mathcal{E}$, and their most obvious feature is that $P(1) = 1$ (which is a concentrated expression of the idea that the average of a variable that happens to be constant is that constant; indeed, it may be of some use to consider the larger part $P(A) \to \mathcal{E}(R^A, R)$ satisfying just that 'averaging' equation, as well as homomorphicity.)

The canonical map $P(A \times B) \to P(A) \times P(B)$ (called "marginals") expresses the contradiction between general joint distributions exhibiting dependence and the particular ones that are independent. In fact, the dimensionality of a fiber, $1 + AB - (A + B)$, indicates how many parameters are needed to measure "dependence". For example, if $A = B = 2$ (the classical paradigm case) then for linear P a single parameter measures dependence. Especially if the Fubini section of the marginals exists, there results an "independence surface" isomorphic to $P(A) \times P(B)$ inside $P(A \times B)$; the distance to this surface from any joint distribution can be measured.

For such categories based on averaging it is crucial to understand the resolutions of the contradiction "double dualization as a commutative monad" (as emphasized by Arens, Linton, Kock, Lucyshyn-Wright and others). That is, for sufficiently small submonads of P the even more useful structure of a symmetric monoidal closed category will be obtained. That means a hom/(tensor) adjointness within the category of P-algebras over $\mathcal{E}$; in particular,

$PA \otimes PB = P(A \times B)$ for the free algebras but more generally, due to the normalization, there is a canonical P-homomorphism $C \otimes D \to C \times D$ for any two algebras.

Note that $\mathrm{Hom}(C, D)$ is almost never free, even if C, D are free (for example, consider $C = P(3)$, $D = P(2)$ in the classical case and note that a cube is not a simplex (a common terminology for a free algebra in this context).) That Hom is almost never free is one of the important reasons why it is necessary to consider the whole category of P-algebras (= "convex sets"), not just the Kleisli category; the vast machinery of enriched categories [2] can then be applied to construct functor categories, Kan extensions, etc. in order to analyze, design, and construct natural stochastic processes and decision procedures of all sorts. Another reason is that not only the unusually intrinsic "generators", but also "relations" enable the objective representation of complexes of probabilistic measures, subject to some a priori convex constraints: for example, a homomorphism from the tetrahedron $P(4)$ corresponds to an arbitrary 4-tuple of probability measures on an arbitrary codomain space, whereas a homomorphism from its non-free rectangular quotient $P(2) \times P(2)$ corresponds to a 4-tuple subject to a single convex constraint.

The importance of enriched categories for the subject does not stop there. As I mentioned in my 1973 Milan paper [6] a convex set has a canonical metric. My student Xiao-Qing Meng showed in her thesis [7] that this assignment is a morphism of closed monoidal categories; that means that all the stochastic diagrams, natural transformations, etc. carry an intrinsic notion of approximation and optimization as a result of their re-enrichment via that monoidal functor into metrically enriched categories.

To facilitate understanding and use of these metrics, there are some simplifications. Since there are many variations on the basic $\mathcal{C} = \mathcal{E}^P \to \mathcal{E}$, consider relevant axioms that hold for all these $\mathcal{C}$. Already mentioned is the "normalization" implying that the functor is represented by $1 = P(1)$, but (whether we are dealing with compact convex sets or dyadically finitary

ones) the actual $P(2) = I$ plays a key role and is indeed adequate in Isbell's sense; that is, any given map $\mathcal{C}(I, C_1) \to \mathcal{C}(I, C_2)$ in $\mathcal{E}$ is induced by a map $C_1 \to C_2$ if only it commutes with the right action of every endomap in $\mathcal{C}(I, I)$ (= probabilistic maps $2 \to 2$). The submonoid of $\mathcal{C}(I, I)$ consisting of those endomaps that preserve a chosen one "0" of the two injection points is typically commutative and ordered by divisibility. Thus for each pair $x, y : 1 \to C$ we can consider the

Definition. *$M(x, y) = \{m : I \to I, m(0) = 0\}$, with $g : I \to C$, $g(0) = x$ and $(gm)(1) = y$.*

Then, for any three elements of C,

$$M(x, y)M(y, z) \subseteq M(x, z)$$

is the "triangle inequality" appropriate for a "nearness" relation (that will give information of the type distance d defined by $M = \exp(-kd)$ that will exist if M is suitably complete). In any case, analogously with the metric theory that is based on addition of non-negative reals, M is a closed category so that categories enriched over it can be identified with nearness spaces in the suggested sense.

What is being measured by this nearness can be understood as follows: the interior of such a convex set consists of distributions that are more random than the less random ones on the extreme boundary; of a given x, y the y can normally be represented as a mixture of x confounded by a more determinate point near the boundary (just extend the line...); the weaker the confounding, the nearer x is to y (it is clear that such metrics are not symmetric.)

Perhaps the simplest example of the statistical application of an intrinsic metric/nearness on the convex hom-sets is optimal decision. Suppose X is a space of parameters presumed to characterize a system of interest but not directly measurable, and suppose a morphism $X \to D$ specifies what a correct decision would be if we knew the true value of X. Suppose there is

a morphism $X \to E$ describing an experiment whose outcome can in fact be measured, thus the problem is to determine a morphism $E \to D$ that makes a decision based on the reading of the experimental outcome. Yet even if both given morphisms are deterministic, there may be no such strictly commutative diagrams, so that a statistically best solution may be sought (and usually exists), namely, a point of the convex set $\mathcal{C}(E, D)$ whose experimental transform into $\mathcal{C}(X, D)$ is as near as possible to the given correct decision. An opposite sort of triangle results if two spaces E, D are assumed to be equipped with given morphisms to (rather than from) a third space Y that is observable (rather than hidden) while D consists of (names for) "causes" via $D \to Y$. The sought-for $E \to D$ then postdicts causes at least for the part of Y parameterized by E. In the special case $E = Y$, the goal is (approximate, random) sections of given $D \to Y$, more likely to exist than deterministic sections. Both kinds of triangles occur in standard categorical "diagonal fill-in" problems, where a given map $X \to Y$ has two given factorizations through E, D respectively, resulting in a commutative square. Such a diagonal would be a single map $E \to D$ satisfying two equations so that an approximate diagonal would involve two nearness estimates. The typically statistical optimization approach specifies a tolerance α on one of these triangles and seeks $E \to D$ to optimize the nearness (to commutativity) of the other triangle, subject to that α-constraint. (Or the distance between the two composites could be minimized.)

Having outlined above some of the possible developments of the mathematics, I now return to sketch the origin of the paper.

Probabilistic Mappings in the Mid-1960s

My acceptance of the job offered by the "Think Tank" in Southern California depended on an agreement that the main topic treated would be Kennedy's Arms Control and Disarmament Agency. The preliminary interview in the Pentagon was requested by that Agency. Somewhat

more precisely, the aim of the study would be: planning for the technical support of an Arms Control Treaty between the Superpowers, for example, of a reliable verification protocol to be agreed upon.

It was envisaged that a protocol would involve three tiers of verification: Space, Stratosphere, and On-site. The passage from one tier to the next would follow probabilistically from continuing observations. What would be the mathematical framework under which this whole fantasy would function? Someone described it as a "network of probabilistic mappings". "What would that mean?" I asked myself: "It must involve diagrams in a category extending the monoid of Markov processes", and then I produced the present document, which served as an Appendix to an Appendix of a large SECRET document.

The proposal was to study a projected system of verification and inspection for a possible Arms Control Agreement between the Superpowers. The system would be organized into the three levels: satellite surveillance, which could trigger the request for over-flight inspection, that in turn could trigger an on-site inspection. Of course, the trigger thresholds would be a matter of diplomacy, but the system as a whole would involve an elaborate network of "probabilistic mappings".

The whole thing had to be scrutinized by the Pentagon before the Arms Control Agency could do anything. Probably, passing through so many hands increased its exposure to espionage. The leader of the group within the Think Tank stated that an important calculation to be done by the study would be the determination of the probability of the discovery of missiles concealed on the ocean floor as part of a planned circumvention of any treaty. That was also the year of the Cuban missile crisis.

A few years later I came across a Russian document containing several of the results of my unpublished thesis, including the mistakes, (as well as the missing two lines that we later

discovered had been missed by the typist). But there was no attribution. And in Moscow the lectures were beginning on a very similar category, called the Markov category (not without justification, of course, although I don't believe Markov himself used categories).

I was surprised a couple of years later by being offered a job with French military intelligence. The one who transmitted that offer was a collaborator of M. Giry, which may explain why she knew about the "secret" developments in the US.

Apparently regarding the contact with the Arms Control Agency as dormant, the leaders of the Think Tank had a further proposal, disregarding their initial agreement with me: First, I should study books by Mao Tsetung and Che Guevara as a preparation for evaluating a large system designed to eliminate the guerilla threat in Vietnam. My last paychecks were for studying that proposal. Of course, I advised against it, after having verified mathematically that the proposed system was unfeasible. The last time I saw the director of the "Vietnam Proposal" was at the old Waltham Watch Factory, which had been taken over as a subsidiary of the California Think Tank. Naturally, my report met with utter disapproval. I took a bus from Waltham to NYC in order to defend my thesis at Hamilton Hall, in front of Eilenberg, Kadison, and Morgenbesser. Now I could complete my application for a teaching job at Reed College.

A few years later the New York Times reported on the failure of a large system that differed only in detail from the one I had analyzed. The supporters of the proposal had taken the plan and defected to another Think Tank.

The Category of Probabilistic Mappings

With Applications to Stochastic Processes, Statistics, and Pattern Recognition

1962

F. William Lawvere

I. Objects and Maps in the Category of Probabilistic Mappings

I.1 Measurable Spaces

1.1. The objects which we consider are *measurable spaces* Ω. That is, $\Omega = \langle S, \mathbb{B} \rangle$ will be an ordered pair in which S is any set and $\mathbb{B}$ is any σ-*algebra* of subsets of S. This means that:

(0) Every member of $\mathbb{B}$ is a subset of S.

(1) The empty set $\emptyset$ and the "whole space" S are members of $\mathbb{B}$.

(2) If $B \in \mathbb{B}$ (i.e., if B is a member of $\mathbb{B}$) then the complement $(S \setminus B) \in \mathbb{B}$.

(3) If $B_i, i = 0, 1, 2, \ldots$ is any countable family of members of $\mathbb{B}$, then the union $\bigcup_{i=0}^{\infty} B_i$ is also a member of $\mathbb{B}$.

We also say that $\mathbb{B}$ is the *class of measurable sets* of Ω.

1.2. If $\Omega = \langle S, \mathbb{B} \rangle$ is any measurable space and if f is a function defined on S with values *in a partially ordered set* Λ, then f is said to be Ω-Λ-*measurable* if for each $\lambda \in \Lambda$ we have

$\{\omega \mid f(\omega) \leq \lambda\} \in \mathbb{B}$; that is, if the set of all $\omega \in \Omega$ whose value under f precedes a given λ is measurable for each λ. For example, we will use this notion when $\Lambda = \mathbb{R}$, the real numbers.

1.3. More generally, if $\Omega = \langle S, \mathbb{B} \rangle$ and $\Omega' = \langle S', \mathbb{B}' \rangle$ are any measurable spaces, and if f is any function defined on S with values in S', then f is said to be a *measurable mapping* if and only if $f^{-1}(B') \in \mathbb{B}$ for every $B' \in \mathbb{B}'$, where $f^{-1}(B')$ denotes the set of all $x \in S$ for which $f(x) \in B'$. The foregoing paragraph is seen to be a special case of this by considering $\Omega' = \langle \Lambda, \mathbb{B}(\Lambda) \rangle$ where $\mathbb{B}(\Lambda)$ is the smallest σ-algebra containing all sets of the form $\{\lambda' \mid \lambda' \leq \lambda\}$ for all $\lambda \in \Lambda$.

1.4. If $\Omega = \langle S, \mathbb{B} \rangle$ is a measurable space, then by a *probability measure* on Ω is meant a function P which assigns to every measurable set $B \in \mathbb{B}$ a real number $P(B)$, in such a way that:

1. $0 \leq P(B) \leq 1$ for every $B \in \mathbb{B}$
2. $P(S) = 1$
3. If $B_i \in \mathbb{B}$ for $i = 1, 2, \ldots$ and if $B_i \cap B_j = \emptyset$ for $i \neq j$ (i.e., B_i are pair-wise disjoint measurable sets) then
$$P(\bigcup_{i=1}^{\infty} B_i) = \sum_{i=1}^{\infty} P(B_i).$$

1.5. In case S is a *countable* set and $\mathbb{B}$ consists of *all* subsets of S, then for any probability measure P on $\langle S, \mathbb{B} \rangle$ and any $B \in \mathbb{B}$, we have

$$P(B) = \sum_{x \in B} P(\{x\})$$

where $\{x\}$ is the "singleton" subset of S whose only member is x, for each $x \in B$. Thus, in this case, a probability measure is already determined by a function $p(x) = P(\{x\})$ of *members* of S; this function is arbitrary, save for the two conditions $0 \leq p(x) \leq 1$, $\sum_{x \in S} p(x) = 1$.

If S is not countable, then probability measures on Ω are not determined by their values at singletons. For example, if $S = \{x \mid 0 \leq x \leq 1\} =$ the "unit interval", and if $\mathbb{B} =$ the smallest σ-algebra containing all closed subintervals = the class of "Borel sets", then there are a great many probability measures P on $\Omega = \langle S, \mathbb{B} \rangle$ for which $P(\{x\}) = 0$ for all x. For example, in this case $P =$ Lebesque measure = (generalized) length is a probability measure but every singleton has zero probability.

1.6. If $\Omega = \langle S, \mathbb{B} \rangle$ and $\Omega' = \langle S', \mathbb{B}' \rangle$ are measurable spaces, if f is a measurable mapping (1.3) from Ω to Ω', and if P is a probability measure on Ω, then the *probability measure* Pf induced on Ω' *by* P *via* f is defined by

$$(Pf)(B') = P(f^{-1}(B'))$$

for every $B' \in \mathbb{B}'$.

To verify that Pf is a probability measure (i.e., satisfies the conditions 0, 1, 2 of 1.4) note that the mapping f^{-1} from $\mathbb{B}'$ to $\mathbb{B}$ is a σ-homomorphism; i.e., that

$$\begin{aligned} f^{-1}(S' \setminus B') &= S \setminus f^{-1}(B') \quad \text{for } B' \in \mathbb{B}' \\ f^{-1}(\textstyle\bigcup_{i=1}^{\infty} B_i') &= \textstyle\bigcup_{i=1}^{\infty} f^{-1}(B_i') \quad \text{for } B_i' \in \mathbb{B}' \\ f^{-1}(S') &= S \end{aligned}$$

From this it is obvious that Pf is a probability measure if P is, in fact, any mapping from

$\mathbb{B}'$ to $\mathbb{B}$ which satisfies the above conditions (whether induced by a mapping from S to S' or not) will induce a mapping from probability measures on Ω to those on Ω'.

1.7. If $\Omega = \langle S, \mathbb{B} \rangle$ is a measurable space and, if $x \in S$, then P_x defined by

$$P_x(B) = \begin{cases} 1 & \text{if } x \in B \\ 0 & \text{if } x \notin B \end{cases}$$

for any $B \in \mathbb{B}$, is a probability measure on Ω, known as a "one-point" or "Dirac" measure.

1.8. Let $\Omega = \langle S, \mathbb{B} \rangle$ be a measurable space, P a probability measure on Ω, f a bounded measurable mapping from Ω to $\mathbb{R}$ = the space of real numbers with Borel sets as the measurable sets. ("Bounded" means that for some positive real number M, $|f(x)| \leq M$ for all $x \in S$.) Such an f is often called a *bounded random variable*. We now wish to define the *P-expectation* of f, also called the *integral of f with respect to P*, denoted either by

$$\int_\Omega f \, \mathrm{d}P$$

or by

$$\int_\Omega f(x) P(\mathrm{d}x).$$

This can be done by considering approximations to the integral based on doubly infinite increasing sequences

$$\cdots \leq a_{-2} \leq a_{-1} \leq a_0 \leq a_1 \leq a_2 \leq \cdots$$

of real numbers. Given any such sequence a, define the upper approximation

$$\overline{J}(f, P, a) = \sum_{-\infty < n < \infty} f(a_n) P f(a_{n-1}, a_n]$$

and the lower approximation

$$\underline{J}(f, P, a) = \sum_{-\infty<n<\infty} f(a_n) P f(a_n, a_{n+1}].$$

Here $Pf(a, b] = P\{x \mid a < f(x) \leq b\}$ as defined in 1.6. The upper integral is defined by

$$\overline{I}(f, P) = \inf \overline{J}(f, P, a)$$

and the lower integral by

$$\underline{I}(f, P) = \sup \underline{J}(f, P, a)$$

where the infimum and supremum are taken over all doubly infinite increasing sequences a. If $\underline{I}(f, P) = \overline{I}(f, P)$, then the function f is said to be *integrable* with respect to P, and the integral is defined to be the common value

$$\underline{I}(f, P) = \int_\Omega f \, \mathrm{d}P = \overline{I}(f, P).$$

It can be shown that *every* bounded measurable function (on Ω) is integrable with respect to every probability measure (on Ω). For each individual P, there will ordinarily be many unbounded functions which are integrable with respect to P.

1.9. If S is a *countable* set, $\mathbb{B}$ the family of all subsets of S, f any bounded measurable function on $\Omega = \langle S, \mathbb{B} \rangle$, and P any probability measure on Ω, then

$$\int_\Omega f(x) P(\mathrm{d}x) = \sum_{x \in S} f(x) p(x)$$

where $p(x) = P(\{x\})$ as defined in 1.5.

1.10. Let f, g be any two bounded measurable functions on a measurable space Ω, and let P be any probability measure on Ω. Then

$$\int_\Omega (f+g)\,\mathrm{d}P = \int_\Omega f\,\mathrm{d}P + \int_\Omega g\,\mathrm{d}P.$$

If a is any real number, then

$$\int_\Omega af(x)P(\mathrm{d}x) = a\int_\Omega f(x)P(\mathrm{d}x).$$

If f_n is any sequence of bounded measurable functions such that f_n is uniformly bounded ($|f_n(x)| \le M$ for all x, n) and if $\lim_{n\to\infty} f_n(x) = f(x)$ for each $x \in S$, then

$$\lim_{n\to\infty}\int_\Omega f_n\,\mathrm{d}P = \int_\Omega f\,\mathrm{d}P.$$

1.11. If $0 \le \theta \le 1$ and if P_1, P_2 are any two probability measures on the measurable space Ω, then $P = \theta P_1 + (1-\theta)P_2$ is also a probability measure, and

$$\int_\Omega f\,\mathrm{d}P = \theta\int_\Omega f\,\mathrm{d}P_1 + (1-\theta)\int_\Omega f\,\mathrm{d}P_2$$

for any bounded measurable function f on Ω.

I.2 Probabilistic Mappings

2.1. Let $\Omega = \langle S, \mathbb{B}\rangle$ and $\Omega' = \langle S', \mathbb{B}'\rangle$ be any measurable spaces. We say T *is a probabilistic mapping from* Ω *to* Ω' and write $\Omega \xrightarrow{T} \Omega'$ if and only if T assigns, to each point in Ω, a probability measure on Ω', and does so in a measurable way. More precisely, T is a function

of two variables $x \in S, B' \in \mathbb{B}'$ having the properties

(0) $0 \leq T(x, B') \leq 1$ for all $x \in S, B' \in \mathbb{B}'$

(1) $T(x, S') = 1$ for all $x \in S$

(2) $T(x, \bigcup_{i=1}^{\infty} B'_i) = \sum_{i=1}^{\infty} T(x, B'_i)$ for each $x \in S$ and for each disjoint sequence B'_i of measurable sets of Ω'.

(3) $\{x \mid T(x, B') \leq a\} \in \mathbb{B}$ for each $0 \leq a \leq 1$ and for each $B' \in \mathbb{B}'$.

We will refer to $T(x, B')$ as the (conditional) T-probability of the event B' in Ω', given the elementary event x in Ω, or as the T-probability that x is mapped into B'. In case S' is countable and $\mathbb{B}'$ consists of all subsets of S', then a probabilistic mapping $\Omega \xrightarrow{T} \Omega'$ is entirely determined by a function t of two *point* variables $x \in S, x' \in S'$. (See 1.5.)

2.2. Every measurable mapping f from Ω to Ω' (these being measurable spaces) may be regarded as a probabilistic mapping $\Omega \xrightarrow{T_f} \Omega'$ as follows:

$$T_f(x, B') = \begin{cases} 1 & f(x) \in B' \\ 0 & f(x) \notin B' \end{cases}.$$

That is, T_f assigns to x the one-point measure (on Ω') which is concentrated at $f(x)$. Probabilistic mappings of this special sort we call *deterministic*.

2.3. Let $\Omega \xrightarrow{T} \Omega' \xrightarrow{U} \Omega''$ be probabilistic mappings. We define the *composition* $\Omega \xrightarrow{UT} \Omega''$ to be the probabilistic mapping defined by

$$(UT)(x, B'') = \int_{\Omega'} U(x', B'') \cdot T(x, \mathrm{d}x')$$

That is, $(UT)(x, B'')$ is the $T(x, \_)$-expectation of $U(\_, B'')$.

This is the correct law of composition of conditional probabilities in physical and other situations.

2.4. If Ω' is a *countable* space in 2.3, then $(UT)(x, B'') = \sum_{x' \in S'} U(x', B'') \cdot T(x', x)$. If Ω'' is also countable, then

$$(UT)(x, \{x''\}) = \sum_{x' \in S'} U(x', x'') \cdot T(x, x').$$

2.5. If $\Omega \xrightarrow{f} \Omega' \xrightarrow{g} \Omega''$ are measurable mappings, then

$$T_{gf} = T_g \circ T_f$$

where gf is the usual composition of functions (thus the deterministic mappings constitute a subcategory (see 2.7) of the category of all probabilistic mappings).

2.6. A probabilistic mapping $1 \xrightarrow{P} \Omega$, where 1 is a one-point space, is just a probability measure on Ω. If $\Omega \xrightarrow{T} \Omega'$ is a probabilistic mapping, then TP is the induced distribution on Ω'. This is familiar in case Ω' is a Euclidean space and T a deterministic mapping (i.e., T is a "random variable"). Another special case is that where $\Omega = \langle S, \mathbb{B} \rangle$, $\Omega' = \langle S, \mathbb{B}' \rangle$, and $\mathbb{B}'$ is a sub-σ-algebra of $\mathbb{B}$, while T is the "identity" mapping; then TP is the restriction of P from $\mathbb{B}$ to $\mathbb{B}'$.

2.7. If

$$\Omega \xrightarrow{T} \Omega' \xrightarrow{U} \Omega'' \xrightarrow{V} \Omega'''$$

then

$$V(UT) = (VU)T.$$

Also, if i_Ω denotes the probabilistic mapping defined by the (deterministic) identity map on Ω, then

$$i_{\Omega'}T = T = Ti_\Omega$$

whenever $\Omega \xrightarrow{T} \Omega'$. Thus, the class $\mathcal{P}$ of all probabilistic mappings between measurable spaces, together with our notion of composition, is a *category* in the sense of Eilenberg–Mac Lane. Thus, the notions of functor, natural transformation, and adjoint functor have a well-defined meaning in connection with $\mathcal{P}$. The "objects" of $\mathcal{P}$ are arbitrary measurable spaces.

2.8. Let, for each object Ω in $\mathcal{P}$, $\mathcal{D}(\Omega) =$ the set of all probability measures on Ω, equipped with the smallest σ-algebra such that for each measurable $A \subseteq \Omega$, the evaluation $\mathcal{D}(\Omega) \to [0,1]$ at A is measurable. Thus $\mathcal{D}(\Omega)$ is also an object in $\mathcal{P}$. For any $\Omega \xrightarrow{T} \Omega'$ in $\mathcal{P}$, define the *deterministic* map $\mathcal{D}(\Omega) \xrightarrow{\mathcal{D}(T)} \mathcal{D}(\Omega')$ by

$$\mathcal{D}(T)(P)(A') = \int_\Omega P(\mathrm{d}\omega)\, T(\omega, A')$$

for every $P \in \mathcal{D}(\Omega)$ and every measurable $A' \subseteq \Omega'$. Thus, $\mathcal{D}(T)(P) = TP$ for $P \in \mathcal{D}(\Omega)$; i.e., viewed as a probabilistic mapping,

$$\mathcal{D}(T)(P, \mathcal{A}) = \begin{cases} 1 & TP \in \mathcal{A} \\ 0 & TP \notin \mathcal{A} \end{cases}$$

for every element P of $\mathcal{D}(\Omega)$, and for every measurable set $\mathcal{A}$ of probability measures on Ω'.

Define also the probabilistic mapping

$$\mathcal{D}(\Omega) \xrightarrow{\phi_\Omega} \Omega$$

for each object Ω in $\mathcal{P}$ by the formula

$$\phi_\Omega(P, A) = P(A)$$

for each element P of $\mathcal{D}(\Omega)$ and each measurable $A \subseteq \Omega$. Then for any $\Omega \xrightarrow{T} \Omega'$ in $\mathcal{P}$, the diagram

$$\begin{array}{ccc} \mathcal{D}(\Omega) & \xrightarrow{\phi_\Omega} & \Omega \\ {\scriptstyle \mathcal{D}(T)}\downarrow & & \downarrow{\scriptstyle T} \\ \mathcal{D}(\Omega') & \xrightarrow[\phi_{\Omega'}]{} & \Omega' \end{array}$$

is commutative, so that ϕ is a natural transformation of the functor $\mathcal{D}$ into the identity functor on $\mathcal{P}$.

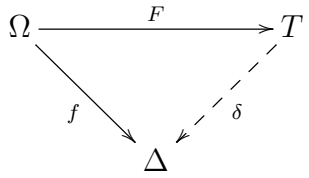
2.9. Actually $\mathcal{D}$ is *adjoint* to the inclusion of the deterministic subcategory into $\mathcal{P}$; i.e., if $\Omega \xrightarrow{T} \Omega'$ is any probabilistic mapping then there is a unique deterministic mapping f such that the diagram

$$\begin{array}{ccc} \Omega & & \\ \vdots & \searrow^{T} & \\ {\scriptstyle f}\vdots & & \\ \mathcal{D}(\Omega') & \xrightarrow[\phi_{\Omega'}]{} & \Omega' \end{array}$$

is commutative. (In particular, there is a deterministic inclusion $\Omega \to \mathcal{D}(\Omega)$ and this is actually a retract with associated retraction ϕ_Ω.) It is expected that this adjointness observation will aid in the analysis of various methodological problems such as Bohm's questions about quantum mechanics.

I.3 Stochastic Processes and Decision Maps

3.1. A fairly general class of decision problems may be formulated as follows. There is a basic space Ω and a measurable partition Δ of Ω, elements of Δ being called "patterns" or "decisions". We denote the quotient mapping $\Omega \to \Delta$ by f. (Actually, for the formulation of the problem we could allow f itself to be "fuzzy"; i.e., probabilistic.) There is also a space T of "observable states" and a probabilistic mapping $\Omega \xrightarrow{F} T$ expressing the conditional probability $F(\omega, A)$ that the observed state lies in any $A \subseteq T$, given that the basic state is $\omega \in \Omega$. The problem is then to find a "best" completion δ of the diagram



One of the "virtues" of probability theory (and hence of the category $\mathcal{P}$) is that this general problem, when properly explicated, has a solution in many cases in which the corresponding deterministic problem does not; a basic reason for this is the possibility in $\mathcal{P}$ of forming convex combinations of maps, whereas there is no corresponding operation which produces deterministic maps. Of course, if there exists δ such that $\delta F = f$, we would choose such δ as the solution to our problem; unfortunately, this is not possible for many F, f of interest. One particular scheme for making definite the criterion for choosing δ is to work with a given distribution $1 \xrightarrow{P} \Omega$ on Ω, and to choose δ so as to maximize the quantity

$$\int_\Omega (\delta F)(x, \{f(x)\}) P(\mathrm{d}x)$$

which represents the average (with respect to P) of the probability of making the correct decision by first making the observation F and then following the decision rule δ. The probability measure P clearly expresses the relative importance attached to various basic

states $x \in \Omega$ when evaluating the decision rule δ. In the absence of any such P, one would choose δ so as to maximize

$$\inf_{x \in \Omega} (\delta F)(x, \{f(x)\}).$$

The existence of solutions δ to these optimization problems can be established in very great generality by topological arguments.

3.2. We consider stochastic processes with discrete time. Let N be the category with countably many objects and no non-identity maps, and let $\mathcal{P}^N$ denote the category whose objects are sequences $\Omega_0, \Omega_1, \ldots$ of objects in $\mathcal{P}$. We define a functor

$$\mathcal{P}^N \xrightarrow{\Phi} \mathcal{P}^N$$

by

$$\Phi\{\Omega_n\}_n = \left\{ \prod_{k<n} \Omega_k \right\}_n$$

for each sequence Ω of measurable spaces, where $\prod_{k<n} \Omega_k$ denotes the measurable space whose elements are all n-tuples $\langle x_0, \ldots, x_{n-1} \rangle$ with $x_i \in \Omega_i$, equipped with the smallest σ-algebra which makes each projection $\prod_{k<n} \Omega_k \to \Omega_j$ measurable. If Ω_n is thought of as the space of all possible states of a system at time n, then $\Phi(\Omega)_n$ is the space of all possible *histories* of the system up to time n. We define a general *temporally discrete stochastic process* P in Ω to be any map

$$\Phi(\Omega) \xrightarrow{P} \Omega$$

in $\mathcal{P}^N$. Given any two processes

$$\Phi(\Omega) \xrightarrow{P} \Omega, \quad \Phi(\Omega') \xrightarrow{P'} \Omega'$$

the general theory of categories indicates that a *map* $P \xrightarrow{f} P'$ of stochastic processes should be defined as a sequence

$$\Omega_n \xrightarrow{f_n} \Omega'_n$$

of maps in $\mathcal{P}$, such that for each time $n \in N$, the diagram

$$\begin{array}{ccc} \Phi(\Omega)_n & \xrightarrow{\Phi(f)_n} & \Phi(\Omega')_n \\ {\scriptstyle P_n}\downarrow & & \downarrow{\scriptstyle P'_n} \\ \Omega_n & \xrightarrow[f_n]{} & \Omega'_n \end{array}$$

is commutative. Since there is also an obvious notion of composition for such maps, all stochastic processes and all maps of such determine a category

$$(\Phi, \mathcal{P}^N)$$

which we call *the category of temporally discrete stochastic processes*. All the machinery developed in the general theory of categories, as well as that which can be developed for the particular category $\mathcal{P}$, can thus be applied to formulate, explicate, and solve many methodological problems within the category $(\Phi, \mathcal{P}^N)$.

3.3. If $\mathbb{N}$ denotes the additive monoid of non-negative integers, considered as a category with one object 0, then the functor category

$$\mathcal{P}^{\mathbb{N}}$$

is the *category of temporally discrete Markov processes*. Explicitly, an object in $\mathcal{P}^{\mathbb{N}}$ is just a measurable space Ω together with a probabilistic mapping $\Omega \xrightarrow{T} \Omega$, and maps f in $\mathcal{P}^{\mathbb{N}}$

satisfying a commutative diagram

$$\begin{array}{ccc} \Omega & \xrightarrow{f} & \Omega' \\ {\scriptstyle T}\downarrow & & \downarrow{\scriptstyle T'} \\ \Omega & \xrightarrow[f]{} & \Omega' \end{array}$$

If we are given a Markov process $\langle \Omega, T \rangle$ together with an initial distribution $1 \xrightarrow{P_0} \Omega$, we can view our situation as a general stochastic process in which

1. $\Omega_n = \Omega$ for all $n \in \mathbb{N}$
2. $\Phi(\Omega)_0 \to \Omega_0$ is just P_0
3. $\Phi(\Omega)_n \to \Omega_n$ is just the composition

$$\prod_{k<n} \Omega_k \to \Omega_{n-1} \xrightarrow{T} \Omega$$

where the first is the projection; i.e., the dependence on the past is really only on the preceding moment and, furthermore, the law of transition from one time to the next does not change with time.

If we denote by $(1, \mathcal{P}^{\mathbb{N}})$ the category of Markov processes augmented with initial distributions, then the foregoing discussion determines a functor

$$(1, \mathcal{P}^{\mathbb{N}}) \to (\Phi, \mathcal{P}^N).$$

This assertion carries the additional information that the various mappings match up properly, and also raises the question of whether the above functor has an adjoint. That is, is it possible to extend any process to a Markov process in a fashion which is universal with respect to

maps to (or from) Markov processes?

References

[1] S. Eilenberg and G. M. Kelly. "Closed categories". In: *Proc. Conf. Categorical Algebra (La Jolla, Calif. 1965)*. Springer, New York, 1966, pp. 421–562.

[2] S. Eilenberg and J. Moore. "Adjoint functors and triples". In: *Illinois J. Math.* 9 (1965). (They used the term "triple" that was later superceded by "monad".), pp. 381–398.

[3] M. Giry. "A categorical approach to probability theory". In: *Categorical Aspects of Topology and Analysis (Ottawa, Ont., 1980)*. Vol. 915. Lecture Notes in Mathematics. Springer, Berlin-New York, 1982, pp. 68–85.

[4] P. J. Huber. "Homotopy Theory in General Categories". PhD thesis with Benno Eckmann, ETH Zurich, 1962.

[5] H. Kleisli. "Homotopy Theory in Abelian Categories". PhD thesis with Benno Eckmann, ETH Zurich, 1960.

[6] F. W. Lawvere. "Metric spaces, generalized logic, and closed categories". In: *Rendiconti Sem. Mat. Fis. Milano* 43 (1973). REPRINTS Theory Appl. Categ. **1** 2002. With Author Commentary: Enriched categories in the logic of geometry and analysis. Available at tac.mta.ca/tac/reprints/articles/1/tr1abs.html.

[7] X.-Q. Meng. "Categories of convex sets and of metric spaces, with applications to stochastic programming and related areas". Available at ncatlab. PhD thesis with F. W. Lawvere, State University of New York at Buffalo, 1988.

Acknowledgments

First, we warmly thank Tobias Fritz for his generous assistance, expertise and participation in reformatting the 1962 original, and Bill's later Abstract and Author Commentary. We also thank Jeremy Gibbons' initial push to reformat the paper and persistence in trying to have it published almost fifteen years ago. We thank Michael Barr and Bob Rosebrugh for assisting in the effort in many exchanges of emails and drafts in 2011, 2016 and 2020.

The history of the writing of the original paper is outlined by the latest Commentary, whose path to publication has been interesting. A version was given as a talk at the University of Insubria in 2012, and then Bill periodically revised many drafts through 2020. Finally, in 2025, the Lawvere Family thanks everyone involved in bringing Bill's earliest-known paper and new Commentary to publication in the Lawvere Archives.

The Archive Family

Part of Bill's email to Bob Rosebrugh (01/25/2016)

I realize that there is a group of younger researchers who would like to know more about this topic (so do I). Some are claiming that it will become a key ingredient in DARPA's[1] thrust toward "genuine" artificial intelligence.

1. It was after 1962 that Godement's notion of standard construction became developed by Kleisli, Huber, Eilenberg & Moore, and Beck, into the theory of algebras for a Monad. Once that theory is made explicit, an extremely compact description of the basic construction can be given, namely probabilistic mappings are just the morphisms in the Kleisli category of the probability monad. In fact, there are several reasons for considering instead the larger

[1] Defense Advanced Research Projects Agency

Eilenberg-Moore category of the same monad, because it is a symmetric monoidal closed category whose unit object is terminal; that permits numerous constructions involved in inference, et cetera to be expressed explicitly in terms of Kan extensions.

2. The possibility of an intrinsic metric for gauging the accuracy of statistical decisions was realized much later in the doctoral thesis of my student X.Q. Meng, based on my 1973 Milan paper concerning the closed structure of the intrinsic metric on convex sets.

Apart from the completion of the mathematical structure mentioned in the above two paragraphs, what interests me greatly is the question of why category theory is so completely unused by statisticians during the last 50 years. There are surprising conjectures.