Tue Tjur

Conditional Probability Distributions

Institute of Mathematical Statistics
University of Copenhagen
1974

Lecture Notes

2

Tue Tjur

Conditional Probability Distributions

Institute of Mathematical Statistics
University of Copenhagen
1974

Lecture Notes **2**

Institute of Mathematical Statistics
University of Copenhagen
5 Universitetsparken
2100 Copenhagen Ø

CONTENTS.

PREFACE.

One of the most insufficient tools of classical probability theory is the definition of conditional distributions. It is well known, that the "almost everywhere-definition" reflects only half of the truth, as the specific choice of conditional distributions in concrete situations does not follow from the definition, but always from our common sense.

Within the theory of Radon measures, a more concrete definition, based on a wellknown idea of "differentiation", can be given. This definition, its consequences, and some applications, will be discussed in the following. I think the exposition is sufficiently complete as concerns existence theorems, special properties and connections to the classical theory, to constitute a realistic alternative to the classical approach.

The reader is assumed to have a general mathematical background. In addition, some knowledge of the theory of Radon measures is necessary; however, knowledge of abstract measure theory together with a short look into the appendix (page 333) will suffice, at least to begin with.

I am most grateful to students and teachers of the Institute of Mathematical Statistics, University of Copenhagen, who made this work possible. In particular, thanks are due to those students who in 1971-72 had the doubtful pleasure of being presented to the earlier versions of the definition. Finally, I should like to thank my former teacher and present colleague Hans Brøns; without his inspiring influence I would hardly have come to wonder about these problems.

København 1973 Tue Tjur

CHAPTER I : INTRODUCTION

1. FOUNDATIONS OF PROBABILITY.

It is well known, that there is a close analogy between concepts like

Number of elements	(sum)
interval length	(integral)
area	(double integral)
volume	(triple integral)
mass	
heat	
electric charge	
probability	(statistical expectation).

In common, they have the property of attaching numbers to sets (functions), in such a way that disjoint union of the sets (addition of the functions) corresponds to addition of the numbers.

Some of these analogies have been known for quite a long time. For example, the analogy between sums and integrals was obviously one of the main ideas behind the definition of the integral, thus going about 300 years back. The analogy between number of elements and probability is even a little older, being

a necessary part of the first definitions of probability.

The later development of measure theory, initiated by Henri Lebesgue in the beginning of this century, made it possible to state in a more precise manner what the common feature of these concepts is.

The idea of counting the elements in a set (number of elements) is, in a sense, the prototype, and the other concepts of the list are generalizations.

Among the generalizations, only the concept of probability demands a measure theory, which is essentially more abstract than that of counting measure, Lebesgue measure, and measures given by densities with respect to counting measure or Lebesgue measure. Probability occurs on sets of many different structures. Certain operations, like the construction of product measures, construction of projective limits (by the consistency theorem), construction of transformed measures and conditioning (or decomposition), which have a fundamental meaning in probability theory, require a measure theoretic framework, in which counting measure, Lebesgue measure etc. are special cases of the same thing.

Lebesgue's theory was soon generalized (Maurice Fréchet, (1915)) to a theory of measures on arbitrary sets, equipped with σ-algebras of subsets. This theory made it possible to define

the concepts of probability on solid, mathematical grounds. Kolmogorov (1933) gave a description of this branch of mathematics; his exposition has constituted the foundations of probability since then.

Unfortunately, some later discovered difficulties in probability theory seem to trace directly back to Fréchet's choice of the σ- continuity as continuity axiom (or localization axiom, see Tue Tjur(1972)). One of the more serious problems is the following:

In the theory of stochastic processes, probability measures on spaces of the form X^T are studied. Here, X denotes the state space (usually $\mathbb{R}$ or $\mathbb{R}^n$), and T denotes the time scale (usually an interval on $\mathbb{R}$ or $\mathbb{Z}$). From intuitive considerations, it seems plausible that a probability measure on such a space should be determined from its finite dimensional marginal distributions (i.e. the distributions describing the joint stochastical variation of the states at a finite number of time points). This is actually the case for $X = \mathbb{R}$, if the prescribed finite dimensional distributions satisfy the natural consistency conditions (Kolmogorov's consistency theorem). The probability measures on X^T ($= \mathbb{R}^T$) are defined on the cylinder σ-algebra, the σ- algebra spanned by the σ - algebras corresponding to the finite dimensional distributions. Two difficulties arise in connection with this theorem:

Firstly, the theorem is not valid in its natural "abstract"

form, i.e. when X is an arbitrary set with a σ-algebra. As proved by Sparre Andersen and Jessen (1948) there exists a consistent family which does not correspond to a probability measure on X^T. The validity of Kolmogorov's consistency theorem depends on assumptions, which can not be described properly in terms of sets with σ-algebras and measurability. Many special versions of the theorem exist, but they have in common the introduction of assumptions, which are essentially of topological character (this was noticed by Halmos (1950)). In all cases, the proof contains a typical compactness argument.

Secondly, the probability measures constructed by the theorem (when it is valid, for $X = \mathbb{R}$, say) are not immediately applicable as models for probabilistic phenomena (except under the additional assumption, that T be denumerable). Any set in the cylinder σ-algebra can be described in terms of a denumerable number of "coordinates" x_t. Thus, events like $\{x_t \leq 2 \text{ for all } t\}$ or {the sample function is continuous} are automatically excluded. Extension to a bigger σ-algebra can usually not be done in any natural way, because many of the interesting events have inner measure 0 and outer measure 1. This problem is usually solved in the following way: Let T_o be a dense, denumerable subset of T. The consistency theorem produces a probability measure on X^{T_o}. This probability measure is transformed by a retraction

$$j: X^{T_o} \rightarrow X^T \text{ (defined almost everywhere)}$$

into a new probability measure, defined on the σ-algebra

$$\{ A \mid j^{-1}A \text{ in the cylinder } \sigma\text{-alg. on } X^{T_o}\} .$$

The choice of j (and, thereby, the choice of the new σ-algebra) must be carried out in a more or less arbitrary manner, from suitable continuity criterions. If the T_o-samplefunction is extendable to a continuous function on T with probability 1, it seems natural to define j as the "extension-by-continuity-imbedding". If the T_o-sample function (as is often the case) has right and left limits at each point of the timescale, then it seems natural to define j as the extension by right or left continuity. General criterions may be introduced in order to make the sample functions of the new process as continuous as possible (Doob's separability concept). But a unique, reasonable choice of j is usually not possible. Consequently, one has to specify the models by introducing assumptions, which have no empirical meaning (like right continuity of the sample functions).

Also the basic definition of a probability field, as a set with a σ-algebra and a normalized, σ-additive set function, has its weak points: The concept of a σ-algebra has no relation to the situations, we pretend to describe in probability theory. The assumption of σ-continuity has (as explicitely noticed by Kolmogorov) no empirical meaning. It is no more than an ad hoc assumption, introduced in order to ensure the validity of Lebesgue's results in the abstract case.

The difficulties in the theory of stochastic processes indicate, that the choice of σ-continuity as continuity axiom may be too superficial. For a more detailed discussion of these matters, see Tue Tjur (1972).

A detailed study of the proof of the complete additivity (or σ-additivity) of the Lebesgue measure shows, that the crucial step is a compactness argument (though this point seems to have played a minor role in Lebesgue's mind; I have not been able to find the compactness argument in the thesis from 1902. In the extended exposition from 1904, the argument is hidden in a lemma (ascribed to Borel) on page 104-105). One might hope, then, that simple assumptions of topological character would ensure the validity of Lebesgues results, thus removing the need of σ-continuity as an assumption. This idea leads immediately to the theory of Radon measures. In the definition of a Radon measure, the σ-continuity condition is replaced by a regularity condition, stating that any integrable set can be approached (in measure) from the inside by compact sets. As proved by Radon (1913) (Riesz (1909)), these measures can be described as linear functionals on a space of continuous functions.

A Radon measure on a locally compact space induces an "abstract" measure (i.e. regularity implies σ-continuity). For locally compact spaces with a denumerable base for the topology, the converse is also true (σ-continuity implies regularity). This means, that the difference between the two measure concepts is a purely formal matter, as long as we are only interested in

spaces like $\hat{Z}$, $\mathbb{R}$ and $\mathbb{R}^n$; which we are most of the time, in fact . For this and for other reasons, the theory of Radon measures has never been widely accepted as an alternative to abstract measure theory. In functional analysis, Radon measures are often considered, because of their beautiful linear representation, which is not shared by the abstract measures. But in elementary mathematical analysis, abstract measures are used. A wellknown (and outstanding) exception is the integration theory of the Bourbaki Series, based exclusively on Radon measures.

In the theory of probability, Radon measures are seldom used. The concept is known among probabilists, but Radon measures are usually regarded as a special sort of very nice abstract measures on special spaces, and as a tool for proving Kolmogorov's consistency theorem in special cases. Only few exceptions are known to me:

Laurent Schwartz (1958) , in connection with a discussion of the Wiener process, has given a very short introduction to the most basic concepts in probability theory, based on the theory of Radon measures.

Edward Nelson (1959) has noticed, that the complicated and unnatural constructions of stochastic processes can be avoided, if Radon measures are applied. The state space must be compact, i.e. one has to compactify $\mathbb{R}$, $\mathbb{R}^n$ etc., but this is a minor difficulty in view of the simplicity obtained: Kolmogorov's

consistency theorem is valid, and the models constructed are applicable without any sort of modification. This result is still widely unknown (or, at least: widely unused), probably because nobody has ever written a book about stochastic processes, based exclusively on the theory of Radon measures. In fact, nobody has bothered much about the evolution of a unified theory of stochastic processes, since Doob wrote his book in 1953.

Hans Brøns (1967) introduced, in a course given at University of Copenhagen 1967-68, the Radon measures in the foundations of mathematical statistics. One reason for this was the need of a more natural formulation of the basic statistical models, in particular those given by invariance properties. The global version of the definition of conditional distributions, as discussed here, was given (the definition by the adjointness equation (see theorem 23.1) and the definition by the decomposition criterion (theorem 23.3)).

Tue Tjur (1972). In the monograph Tue Tjur (1972), I have given a short introduction to probability, based on the theory of Radon measures on compact spaces. The assumption about compactness is not restrictive -though rather unsatisfactory- because the spaces considered may always be compactified. Almost all the definitions and results of the paper are known beforehand, and the work has mainly consisted in the translation of known results from probability theory to the language of Radon measure theory. The aim has not been to prove new

results, but to show, that the theory of Radon measures, without being any more complicated than the abstract measure theory, provides a theory of probability, which is more natural and more efficient than the classical theory. Let me list some of the advantages, one obtains:

(1) The definition of a measure is based on the topology; σ-algebras come in as a secondary tool (it is a matter of taste, whether or not this should be called an advantage. I think it should, because I find a topology a much more fundamental structure than a σ-algebra).

(2) The concept of weak convergence arises in a natural way.

(3) Kolmogorov's consistency theorem is valid, and the models so constructed are immediately applicable.

(4) The relevant concept of measurability (see the appendix) solves many (if not all) measurability problems in the theory of stochastic processes.

(5) Under certain (in practice, very unrestrictive) regularity conditions, the concept of a conditional distribution can be given a concrete, pointwise meaning.

Not all these postulates are demonstrated in Tue Tjur (1972). Only the postulates demonstrated elsewhere are so, in fact. (1) and (2) are certainly true, also in the locally compact

case. (3) is true, but it remains to give a careful description of what happens, if the state space is locally compact, not compact. How does the choice of compactification affect the model ? The compactification-technique is not immediately acceptable, because it introduces this arbitrary factor in the model (though seemingly "less arbitrary" than the choice of version in the classical theory). It may be convenient to introduce measures on completely regular spaces, in order to solve this problem. The compactification-technique is no more than a technique. It may be acceptable as a tool in the theoretical part of the framework, but as soon as special spaces (like $\mathbb{R}$) are considered, the theory must be extended to the locally compact case, at least.

The postulate (4) is no more than a good guess, based on examples. Results like the measurability of "stopping rules" (see Tue Tjur (1972), page 149), indicate the power of Bourbaki's measurability concept. See also theorem 20, 21 and 22 in the appendix (here). It is a characteristic property of the abstract theory, that seemingly basic results like theorem 20 turn out to be wrong, unless new and surprisingly complicated conditions are imposed.

The postulate (5) will, of course, be discussed here. The definition of conditional distributions is the only new method in Tue Tjur (1972). Though the idea of defining conditional distributions by a differentiation procedure is certainly not new, it has always been regarded as a purely heuristic method,

incompatible with the abstract theory (see e.g. Feller, vol II, ed.1 page 157 or ed.2 page 160, or Breiman, page 68). In section 6, I shall comment on a single attempt (the only attempt, as far as I know) to define conditional distributions by "differentiation" (Søren Johansen (1967)).

Obviously, the method is incompatible with the classical theory, since it involves topological considerations. But in the framework of Radon measures, the method works quite well, as we shall see.

2. NOTATION.

All spaces X,Y, ... are assumed to be locally compact and σ-compact, if nothing else is stated (for the definition of σ-compactness, see the appendix). By a measure we always mean a Radon measure, and integration, measurability etc. are as defined in the appendix.

Measurability. The relevant concept of measurability is defined in the appendix. Notice, that the definition is very different from the usual definitions known from abstract measure theory. It is interesting, however, that for "small" spaces (like $\mathbb{R}$, $\mathbb{R}^n$ etc.) the definitions almost coincide (see Tue Tjur (1972), page 71).

Brackets [ ] , indexed by a variable, indicate that the expression in the brackets should be considered a function of this variable. For example, we write

$$[\, g(x,y) \,]_x$$

for the mapping, which, for fixed y, takes x into g(x,y). Formally, the use of brackets is defined by the equation

$$f = [\, f(x) \,]_x \quad .$$

The integral sign $\int$ is only applied in case of Lebesgue measure. The notation

$$\int f(x)\, \mu\, (dx)$$

is inexpedient, the symbol dx having no meaning in the abstract case. The only good reason for keeping this notation in use is the "variable specifier mechanism", built into it. The use of brackets, as explained above, removes this motivation. We write

$$\mu f \qquad \text{or} \qquad \mu[f(x)]_x$$

for the integral of f with respect to μ.

Example: $(\mu \otimes \nu) f = (\mu \otimes \nu)[f(x,y)]_{(x,y)}$

$$= \mu[\nu[f(x,y)]_y]_x .$$

Stochastic variables. This concept will play the role of a pure notational convention, similar to the classical concept of a variable. The formal definition of a stochastic variable as a mapping is not convenient, the theory being formulated in terms of transformations between probability fields, rather than in terms of a (classical) "background" probability field. The use of a symbol x as a stochastic variable is declared by specifications like

$$x \in (X, \mu) ,$$

read: Let x in X be chosen at random with respect to the probability measure μ , or: Let x be a stochastic variable with distribution μ .

Also <u>derived</u> stochastic variables

$$y = t(x) , \quad x \in (X, \mu) ,$$

are considered (similar to the classical <u>dependent variables</u>. But for obvious reasons, the term "dependent" can not be used in that connection here). The definition of y above requires, that

$$t: X \to Y$$

is a μ-measurable transformation. We think of y as a stochastic variable

$$y \in (Y, t(\mu)) ,$$

but the functional dependence between x and y is -of course- subsumed in the following. As for "classical" variables, y replaces -more or less- the expression t(x).

In some connections, the notation

$$\mathcal{L}(x) = \mu$$

for the distribution of x is convenient. For example,

$$\mathcal{L}(y) = \mathcal{L}(t(x)) = t(\mu) .$$

Also the wellknown notation

$$\mathcal{L}(x|\ t(x) = y_o)$$

for the conditional distribution of x, given $t(x) = y_o$, may be convenient.

Some special symbols.

$:=$	is used for definitions. For example, $y := f(x)$ means: Let y be defined as $y = f(x)$.
1_A	denotes either the indicatorfunction of the set A (i.e. $1_A(x) = 1$ for $x \in A$, 0 for $x \notin A$), or the identity $1_A : A \to A$.

See also the more complete list of symbols introduced in the text (page 363).

CHAPTER II : LOCAL DEFINITION OF A CONDITIONAL DISTRIBUTION

3. THE DEFINITION.

Let μ denote a probability measure on X, and let

$$t: X \to Y$$

be a μ-measurable transformation. By

$$\nu := t(\mu)$$

we denote the transformed measure. To specify this situation as a whole, we shall write

$$t: (X, \mu) \to (Y, \nu) ,$$

referring to t as a homomorphism between probability fields.

For a point $y \in Y$, we want to define the conditional distribution of $x \in (X, \mu)$, given $t(x) = y$.

For a measurable set $A \subseteq X$ with $\mu A > 0$, let

$$\mu^A := \frac{1}{\mu A} \cdot (1_A \cdot \mu)$$

denote the conditional distribution of x, given $x \in A$ (here, $1_A \cdot \mu$ denotes the measure given by the density 1_A with respect to μ ; see the appendix). It seems natural to define the conditional distribution of x, given $t(x) = y$, as the limit of the distributions $\mu^{t^{-1}B}$ ($\nu B > 0$), when B tends to y, in some sense.

For simplicity, we write μ^B instead of $\mu^{t^{-1}B}$ in the following.

We assume, that the point y belongs to the support of ν . If this is not the case, we can certainly not hope to be able to "approximate" y by a set of positive measure.

For the precise definition of the limit $\lim \mu^B$, we shall need the concept of a net.

Nets. By an upwards directed set, we mean a set D, equipped with a reflexive and transitive relation $\leq$ (a preordering), such that $(D, \leq)$ is upwards directed:

$$\forall\ d_1 , d_2 \in D \quad \exists\ d \in D : \ d_1 \leq d \ \text{ and } \ d_2 \leq d \ .$$

A D-net, or a generalized sequence, indexed by D, on an arbitrary topological space Z, is a mapping

$$d \rightarrow z_d$$

$$D \rightarrow Z$$

(or, equivalently, a <u>family</u> $(z_d | d \in D)$ of points in Z). We write

$$(z_d | d \in D) \quad \text{or} \quad (z_d | d \in (D, \leq))$$

for the net.

Taking as $(D, \leq)$ the set $\mathbb{N}$ of natural numbers, equipped with its usual ordering, we simply get a sequence $(z_n | n \in \mathbb{N})$; this explains the name "generalized sequence".

<u>Convergence</u> of a net is defined exactly as for sequences: The net $(z_d | d \in D)$ is said to be <u>convergent towards</u> $z \in Z$, if, for any neighbourhood W of z, there exists a $d_o \in D$, such that

$$d \geq d_o \implies z_d \in W .$$

We write

$$z = \lim_{d \uparrow \infty} z_d$$

(the point z is uniquely determined, as soon as the space Z is assumed to be Hausdorff).

For the definition of the conditional distribution, it seems natural to construct the net as follows: Let $(D, \leq)$ denote the set of (say) compact neighbourhoods of y, ordered by inverse inclusion $\supseteq$, and consider the net $(\mu^B | B \in (D, \supseteq))$.

We might then define the conditional distribution as the limit (in the weak topology) of this net, in case it exists.

There is, however, a slightly more restrictive definition, which is easier to work with, because it ensures continuity of the conditional distribution as a function of y. Define the net as follows:

Let D denote the set of pairs (V,B), consisting of an open neighbourhood V of y and an open set $B \subseteq V$ with $\nu B > 0$. The relation $\leq$ is defined as inverse inclusion with respect to V, i.e. we write

$$(V_1,B_1) \leq (V_2,B_2)$$

if and only if

$$V_1 \supseteq V_2 .$$

Obviously, $\leq$ becomes a preordering. In order to prove that $(D,\leq)$ is upwards directed, let (V_1,B_1) and (V_2,B_2) be elements of D. Then, $(V_1 \cap V_2, V_1 \cap V_2)$ is also an element of D, dominating the two given elements (notice, that we have used the assumption $y \in \operatorname{supp} \nu$; the definition of D requires $\nu(V_1 \cap V_2) > 0$). Hence,

$$(\mu^B \mid (V,B) \in (D,\leq))$$

is a net on $\mathcal{M}(X)$.

Definition: If the limit

$$\mu^y = \lim_{(V,B)\uparrow\infty} \mu^B$$

exists and belongs to $\mathcal{P}(X)$, it is called the conditional distribution of $x \in (X, \mu)$, given $t(x) = y$.

The idea behind the (rather complicated) construction of the net is not hard to understand: We want to let B be a small set, close to y, but possibly not containing y. Unfortunately, the sets B can not be ordered directly with respect to this closeness-property. But the neighbourhoods of y can, in a sense, be ordered with respect to "closeness to y" (namely, by inverse inclusion). Letting V tend to y in this sense, and letting B be a freely varying subset of V, we obtain a kind of "convergence of B towards y".

Nets of the type $(z_B | B \to y)$. The directed set $(D, \leq)$, as defined above, is an important tool in the following. The nets considered will always be of the form $(z_B | (V,B) \in (D, \leq))$, i.e. the point $z_B = z_{(V,B)}$ depends on B only (and not on V). For simplicity, we write "$B \to y$" instead of "$(V,B)\uparrow\infty$". In order to indicate, that the measure ν is the measure behind the definition of the net, we may write "$B \to y,\ \nu B > 0$" . The net will be written simply as $(z_B | B \to y)$.

Defective conditional distribution. The requirement, that the limit measure $\lim_{B \to y} \mu^B$ in the definition should be a probability measure, is not empty. It is wellknown, that probability mass may disappear during a passage to the limit (unless X is compact). If the net $(\mu^B \mid B \to y)$ converges, but not towards a probability measure, we talk about a defective conditional distribution. A defective conditional distribution has total mass < 1 (most often, $\| \mu^y \| = 0$).

4. JUSTIFICATION OF THE DEFINITION.

Most of the situations, considered in probability theory, fall within one of the following main classifications:

The discrete case: Probability distributions, given by their densities with respect to counting measure on some set (usually $\grave{Z}$ or $\grave{Z}^n$).

The continuous case: Probability distributions, given by their densities with respect to Lebesgue measure on $\mathbb{R}$ or $\mathbb{R}^n$, or -possibly- with respect to some area measure on a differentiable manifold.

Stochastic processes: Probability distributions, describing an infinite system of consistently connected probability fields of discrete or continuous type.

It is obvious, that the definition of conditional distribution given here (page 20) is equivalent to the usual, elementary definition in the discrete case. As we shall see, the definition is also applicable to the two other cases, in the sense that the conditional distributions are, in most cases, everywhere (or, at least: almost everywhere) defined, and they have the properties one would expect conditional distributions to have.

Existence of the conditional distributions in most cases of

interest does not follow from any results of a general, mathematical character. Only strong regularity conditions, like differentiability of the transformation t, continuity of the densities defining μ and ν etc., ensure the existence. But such regularity assumption are usually satisfied. It is not hard to construct examples, where the conditional distribution is nowhere defined, but these examples seem to be of no interest to us (see section 30, page 251).

The advantage of the present definition, as compared to the definition given in the classical theory, is its local character, making it possible to talk about the conditional distribution of x, given a certain, fixed value y of t(x). In the classical theory, no topological assumptions are included, and therefore the local definition is meaningless. Instead, one has to define the conditional distribution in terms of its global properties. This involves, among other things, that the conditional distributions are only determined up to equivalence (i.e., they can be changed on a null set), which makes conditioning upon a fixed value of y meaningless.

Conditional expectation. Discussing various dependence/independence-structures (conditional independence, the Markov property, asymptotic independence etc.), the concept of conditioning plays a fundamental role, as a tool for description of the (unconditioned) stochastic variation. In such cases, it would be unnatural to make assumptions about existence of the conditional distributions, which we happen to mention in

the discussion. But in such cases, conditional expectations (including conditional probabilities, defined as conditional expectations of the indicatorfunctions) can be used. Conditional expectations are always defined (up to equivalence) as L^2-functions, and they have very simple, algebraic properties. They constitute the natural framework for discussion of dependence/independence-structures, because of the close analogy between independence and orthogonality.

The conditional distributions should only be considered, when they are actually of interest as distributions; by this, I mean, that their own probability fields are considered. The typical application of conditional distributions is the "realization" of the stochastic variable $x \in (X, \mu)$ in the following way: Choose (at random) $y \in (Y, \nu)$; then, choose (at random) $x \in (X, \mu^y)$. A "correct" application of conditional distributions is conditioning upon an ancillary statistic in a statistical model. In this application, the "conditional experiment" is regarded as if it was actually carried out. In such examples, a definition of the conditional distribution, given a value of y, is certainly convenient (if not necessary).

An analogy. The definition of conditional distribution and conditional expectation has an obvious, classical analogue: Let F be a real function of a real variable. We can define the derivative f of F by the equation

$$f(x) := \lim_{h \to 0} \frac{1}{h}(F(x+h)-F(x)) .$$

We can also define f as the function, satisfying the equations

$$\int_0^{x_0} f(x)dx = F(x_0) - F(0) , \qquad x_0 \in \mathbb{R} .$$

The last definition is the more general, and it has, from a mathematical point of view, many nice properties, not shared by the first definition. But according to the second definition, f is only defined up to equivalence.

For the definition of the concept velocity, we prefer the first definition of the derivative. It requires stronger regularity assumptions than the second definition, but it makes the velocity at time t a welldefined and locally determined number. The second definition would reduce the velocity to an equivalence class of functions, the only concrete property being , that we may integrate the velocity function in order to compute certain distances.

5. THE CONDITIONAL DISTRIBUTION OF A DERIVED STOCHASTIC VARIABLE.

Consider a diagram

$$\begin{array}{ccc} (X, \mu) & \xrightarrow{\ t\ } & (Y, \nu) \\ \Big\downarrow s & & \\ (Z, \xi) & & \end{array}$$

of probability fields and homomorphisms. By the conditional distribution ξ^y of $s(x)$, given $t(x) = y$, one would immediately mean the transformed conditional distribution

$$\xi^y = s(\mu^y) .$$

But under certain conditions, we can define ξ^y, without assuming existence of μ^y (for example, in the trivial case, where Z consists of one point). Therefore, we shall give another definition:

Definition. If, for $y \in \operatorname{supp} \nu$, the limit measure

$$\xi^y = \lim_{B \to y} s(\mu^B)$$

exists and belongs to $\mathcal{P}(Z)$, it is called the conditional

distribution of $s(x)$, given $t(x) = y$.

The distribution ξ^y , defined in this way, is called a conditional distribution of a derived stochastic variable (namely, the derived stochastic variable $z = s(x)$), or, simply, a derived conditional distribution.

In the following, we write ξ^B for $s(\mu^B)$.

In case s is continuous, the above definition is consistent with the first proposal $\xi^y = s(\mu^y)$:

5.1 Theorem. Suppose that s is continuous, and let y be a point in supp ν , such that the conditional distribution μ^y is defined. Then, also the derived conditional distribution ξ^y is defined, and

$$\xi^y = s(\mu^y).$$

Proof: It follows immediately from corollary A 16 (page 346) that the mapping

$$\begin{aligned} \mu' &\to s(\mu') \\ \mathcal{P}(X) &\to \mathcal{P}(Z) \end{aligned}$$

is continuous. The conditional distribution μ^y being defined, we have (for $B \to y$)

$$\lim \xi^B = \lim s(\mu^B) = s(\lim \mu^B) = s(\mu^y) \in P(Z) ,$$

i.e. ξ^y is defined and equal to $s(\mu^y)$.

The existence of a derived conditional distribution can, in a unique way, be associated with the existence of a "proper" conditional distribution; consider the following diagram, constructed from the diagram in the beginning of this section (page 26):

$$(X, \mu) \xrightarrow{(t,s)} (Y \times Z, \gamma) \xrightarrow{p} (Y, \nu)$$

The homomorphism (t,s) is (of course) defined by $(t,s)(x) = (t(x),s(x))$, and γ denotes the transformed measure $\gamma = (t,s)(\mu)$. The transformation p is the projection on the first component Y of the product $Y \times Z$. The relation $p(\gamma) = \nu$ follows immediately from the equation $p \circ (t,s) = t$.

The transformation (t,s) reduces the original probability field to the smallest probability field, describing the joint stochastic variation of y and z. Hence, it is not surprising, that we have

5.2 Theorem. The conditional distribution ξ^{y_o} of

$z = s(x)$, given $t(x) = y_o$, is defined if and only if the conditional distribution γ^{y_o} of $(y,z) \in (Y \times Z, \gamma)$, given $y = y_o$ (i.e., given $p(y,z) = y_o$) is defined. In case of existence, we have

$$\gamma^{y_o} = \varepsilon_{y_o} \otimes \xi^{y_o} .$$

Proof: Let

$$q: Y \times Z \to Z$$

denote the projection on Z. For $B \subseteq Y$, $\nu B > 0$, we have

$$\xi^B = q(\gamma^B)$$
$$\text{and} \qquad \nu^B = p(\gamma^B) .$$

If the limit $\gamma^{y_o} = \lim_{B \to y_o} \gamma^B$ exists, we have for $B \to y_o$

$$q(\gamma^{y_o}) = q(\lim \gamma^B) = \lim q(\gamma^B) = \lim \xi^B .$$

Thus, ξ^{y_o} is defined and equal to $q(\gamma^{y_o})$. In any case, we have

$$p(\gamma^{y_o}) = p(\lim \gamma^B) = \lim p(\gamma^B) = \lim \nu^B = \varepsilon_{y_o}$$

From the two equations

$$\xi^{y_o} = q(\gamma^{y_o})$$
$$\varepsilon_{y_o} = p(\gamma^{y_o})$$

it is not hard to show that

$$\gamma^{y_o} = \varepsilon_{y_o} \otimes \xi^{y_o} .$$

Conversely, suppose that the limit measure $\xi^{y_o} = \lim \xi^B$ exists and belongs to $\mathcal{P}(Z)$. Then, an arbitrary contact point γ' for the net $(\gamma^B \mid B \to y_o)$ satisfies the relations

$$\xi^{y_o} = q(\gamma')$$
$$\varepsilon_{y_o} = p(\gamma') .$$

Hence, the only possible contact point of the net is $\varepsilon_{y_o} \otimes \xi^{y_o}$. The set of measures on $Y \times Z$ of total mass ≤ 1 being <u>compact</u>, this implies, that the net $(\gamma^B \mid B \to y_o)$ is convergent towards $\varepsilon_{y_o} \otimes \xi^{y_o}$. Thus, γ^{y_o} is defined (and equal to $\varepsilon_{y_o} \otimes \xi^{y_o}$).

The theorem shows, that the conditional distribution of z, given $y = y_o$, depends only on the simultaneous distribution γ of y and z, but not on the choice of "background probability field" (X, μ).

Moreover, it follows from the theorem, that the concept of

a derived conditional distribution represents a convenient notation, rather than a new concept. We need not (and shall not) discuss results concerning existence and properties of derived conditional distributions, such results being easily deduced from the the results concerning "proper" conditional distributions, by means of theorem 5.2.

If we regard a stochastic variable $x \in (X, \mu)$ as being derived <u>from itself</u>, theorem 5.2 yields the following result:

<u>5.3 Corollary</u>. Let

$$t: (X, \mu) \to (Y, \nu)$$

be given, and suppose for $y_o \in Y$ that the conditional distribution μ^{y_o} is defined. Put

$$\gamma := (t, 1_X)(\mu) \in \mathcal{P}(Y \times X)$$

(here, $1_X : X \to X$ denotes the identity).

Then, the conditional distribution of $(y,x) \in (Y \times X, \gamma)$, given $y = y_o$, is defined, and given by

$$\gamma^{y_o} = \varepsilon_{y_o} \otimes \mu^{y_o} .$$

6. REMARKS.

Differentiation of set functions. The technique of letting a set B tend, in some sense, towards a point y, is wellknown from the theory of differentiation of set functions, see Lebesgue (1910), Hahn and Rosenthal (1948), Saks (1937), Dunford and Schwartz (1957). This analogy becomes even more clear in section 20, where the definition and properties of essential values are modifications (or even special cases) of definitions and results from the theory of differentiation of set functions.

Let μ and λ be two measures (or just, additive set functions) on some space, and define the Radon-Nikodym-derivative

$$\frac{d\mu}{d\lambda}(x) := \lim_{(A \to x)} \frac{\mu A}{\lambda A} ,$$

where "$(A \to x)$" means, that A tends to x in the sense of section 3, but under certain regularity assumptions about the set A during the passage to the limit. In order to obtain almost-everywhere-results, strong regularity conditions may be imposed on A. For example, in case λ is Lebesgue measure on $\mathbb{R}^n$, the following result is typical: If A denotes a cube, containing x, then the limit $\frac{d\mu}{d\lambda}$ is defined λ-almost everywhere, and $\frac{d\mu}{d\lambda}$ is the density of the absolutely continuous part of μ. Thus, in case μ is absolutely continuous, the Radon-Nikodym-derivative is simply the density of μ

with respect to λ , the existence of which is known from the Radon-Nikodym theorem. The theorem cited here can be found in Dunford and Schwartz, page 214.

In our definition of the conditional distribution, no restrictions were imposed on A (or B, as we called it there), except for the -obviously necessary- condition $\lambda A > 0$ (in section 3: $\nu B > 0$). It might be of interest to see what happens, if the definition of the Radon-Nikodym-derivative is based on this "unrestricted" net.

Density. Instead of "Radon-Nikodym-derivative" , we shall use the term density, to keep closer to the terminology applied here (see page 350) . Let μ and λ be measures on a (locally compact and σ-compact) space X, and define for $x \in X$

$$d(x) := \lim_{\substack{A \to x \\ \lambda A > 0}} \frac{\mu A}{\lambda A} \quad .$$

The number $d(x) \in \mathbb{R}$ (if defined) is called the density, or the local density of μ with respect to λ at the point x.

6.1 Theorem. Suppose, that $d(x)$ is defined for all x. Then, d is continuous, and

$$\mu = d \cdot \lambda \quad .$$

This theorem is not hard to prove. But we shall not prove it here, since we are not going to use it. The theorem is very likely to be found among the results proved in the extensive litterature about differentiation of set functions.

Continuity of limits of the type $\lim_{A \to x} z_A$. The most interesting statement in theorem 6.1 is certainly the continuity of d. This continuity turns out to be a general property of limits of the type introduced in section 3:

6.2 Theorem. Let λ be a measure on X, and let Z be a completely regular topological space (i.e. Z is a Hausdorff space, and any point of Z has a neighbourhood base of closed sets). Let

$$A \to z_A$$

be a mapping, from the set of open sets $A \subseteq X$ with compact closure into the space Z. Let C denote the set of points x in X, such that the limit

$$z_x := \lim_{\substack{A \to x \\ \lambda A > 0}} z_A$$

exists. Then, the mapping

$$x \to z_x \quad , \quad C \to Z$$

is continuous.

This theorem is also easy to prove; and we do not do so as we have no use for it. The continuity of conditional distributions (section 22) , which is certainly a special case of theorem 6.2 above, comes out in a different manner, as a consequence of the properties of essential values.

Definition of a conditional distribution by differentiation was proposed by Søren Johansen (1967). The definition given there is easily seen to be equivalent to the one given here, the only difference being, that the set B is replaced by a continuous function $g \geq 0$ with compact support (the support tending to y , in the sense defined here). It was proved, that the measure μ can be represented as a mixture of measures μ_y , where μ_y denotes a contact point of the net $(\mu^g | g \to y)$, defined as indicated above. From this result it follows immediately, that μ can be represented as a mixture of the conditional distributions μ^y , in case they are all defined. The result is obtained by means of Choquet's theorem on representation of points in a convex set as barycenters of probability measures, concentrated on the extreme points. Any extreme point of the closed, convex hull of the set $\{\mu^B | \nu B > 0\} \subseteq \mathcal{P}(X)$ turns out to be a contact point for one and only one of the nets $(\mu^g | g \to y)$.

CHAPTER III: EXISTENCE OF THE CONDITIONAL DISTRIBUTIONS

7. CONDITIONING IN A DECOMPOSED MEASURE SPACE.

Let λ denote an arbitrary measure on X. Suppose we have a continuous transformation

$$t: X \to Y .$$

By a decomposition of λ with respect to t, we mean a representation of λ as a mixture of measures, concentrated on the level surfaces for t. An immediate example is the decomposition of a probability measure into its conditional distributions with respect to t (see theorem 23.3, page 196); but we can also think of the decomposition of Lebesgue measure in $\mathbb{R}^2$ as a mixture of Lebesgue measures on vertical lines (lines parallel to the y-axis).

Definition. A decomposition of λ with respect to t is a pair

$$(\lambda' , (\lambda_y | \ y \in Y)) ,$$

consisting of a measure λ' on Y and a continuous mapping

$$y \to \lambda_y \quad , \quad Y \to \mathcal{M}(X)$$

such that the following two conditions are satisfied:

(1) For $y \in Y$, $\operatorname{supp} \lambda_y \subseteq t^{-1}(y)$

(2) For $k \in \mathcal{K}(X)$, $\lambda k = \lambda'[\lambda_y k]_y$.

Condition (1) implies, together with the continuity assumption, that $[\lambda_y k]_y$ is a $\mathcal{K}(Y)$-function for $k \in \mathcal{K}(X)$ ($\lambda_y k$ is 0 for $\operatorname{supp} \lambda_y \cap \operatorname{supp} k = \emptyset$, i.e. for y in the complement of the compact set $t(\operatorname{supp} k)$). Hence, the right side in (2) is welldefined, and condition (2) expresses, that λ is the mixture of the measures λ_y with respect to λ' (see the appendix, page 351).

Now, let μ be a probability measure, given by the density f with respect to λ (page 350). We shall try to construct the conditional distribution of $x \in (X, \mu)$, given $t(x) = y_o$, for some $y_o \in Y$.

It follows from theorem A 22 (page 352) that f is λ_y-integrable for λ'-almost all y, and the λ'-almost everywhere defined function

$$g(y) = \lambda_y f$$

is λ'-integrable with

$$\lambda' g = \lambda' [\lambda_y f]_y = \lambda f = 1 .$$

Thus, the function g is a probability density with respect to λ', and it is not difficult to prove, that the corresponding probability measure is exactly the transformed measure $\nu := t(\mu)$:

For $h \in \mathcal{K}(Y)$, we have

$$\nu h = \mu(h\circ t) = (f\cdot\lambda)(h\circ t) = \lambda(f\cdot(h\circ t))$$

$$= \lambda'[\lambda_y(f\cdot(h\circ t))]_y \overset{*}{=} \lambda'[h(y)\,\lambda_y f]_y = \lambda'(h\cdot g)$$

$$= (g\cdot\lambda')h .$$

The identity $\overset{*}{=}$ follows from the fact that $h\circ t$ is constant on $\operatorname{supp}\lambda_y$, according to condition (1) in the definition of a decomposition.

7.1 Theorem. Let $y_o \in \operatorname{supp}\lambda'$ be given. Suppose, that there exists a neighbourhood V_o of y_o such that

(1) f is λ_y-integrable for all y in $V_o \cap \operatorname{supp}\lambda'$ (i.e. g is welldefined on $V_o \cap \operatorname{supp}\lambda'$).

(2) $g(y_o) > 0$.

(3) The two mappings

$$y \to g(y) \quad , \quad V_o \cap \operatorname{supp} \lambda' \to \mathbb{R}$$

$$y \to f \cdot \lambda_y \quad , \quad V_o \cap \operatorname{supp} \lambda' \to \mathcal{M}(X)$$

are continuous at the point y_o .

Then, the conditional distribution of x , given $t(x)=y_o$, is defined, and given by

$$\mu^{y_o} = \frac{1}{g(y_o)} f \cdot \lambda_{y_o} .$$

Notice, that the condition (3) is satisfied, if the densities f and g are continuous. Moreover, g need only be continuous at the point y_o .

In the proof, we shall need the following lemma:

7.2 Lemma. Let λ be a measure on X , and let f be an integrable (or just locally integrable, see page 350) function. Let x_o be a point in $\operatorname{supp} \lambda$, and suppose that the restriction of f to $\operatorname{supp} \lambda$ is continuous at the point x_o . Then

$$\lim_{A \to x_o} \frac{(1_A \cdot f)}{\lambda A} = f(x_o) .$$

Proof: For $\varepsilon > 0$, choose an open neighbourhood U of x_o with compact closure such that

$$|f(x) - f(x_o)| \leq \varepsilon \quad \text{for} \quad x \in U \cap \operatorname{supp} \lambda .$$

For an open set $A \subseteq U$ with $\lambda A > 0$, put

$$A' := A \cap \operatorname{supp} \lambda .$$

Then,

$$\left| \frac{\lambda(f \cdot 1_A)}{\lambda A} - f(x_o) \right| = \left| \frac{\lambda(f \cdot 1_{A'})}{\lambda A'} - f(x_o) \right|$$

$$= \left| \frac{\lambda(f \cdot 1_{A'} - f(x_o) 1_{A'})}{\lambda A'} \right|$$

$$\leq \frac{1}{\lambda A'} \lambda \left| (f - f(x_o)) \cdot 1_{A'} \right| \leq \frac{1}{\lambda A'} \lambda \left| \varepsilon \cdot 1_{A'} \right| = \varepsilon .$$

Proof of the theorem: It follows from the assumptions, that y_o belongs to $\operatorname{supp} \nu$.

For a function $k \in \mathcal{K}(X)$, we have for $B \to y_o$ ($\nu B > 0$)

$$\mu^B k = \frac{1}{\nu B} \mu(1_{t-1_B} \cdot k) = \frac{1}{\nu B} \lambda(f \cdot 1_{t-1_B} \cdot k)$$

$$= \frac{1}{\nu B} \lambda' \left[\lambda_y (f \cdot 1_{t-1_B} \cdot k) \right]_y = \frac{\lambda'(1_B \cdot [\lambda_y (f \cdot k)]_y)}{\lambda'(1_B \cdot g)}$$

$$= \frac{\lambda'(1_B[(f\cdot\lambda_y)k]_y)/\lambda' B}{\lambda'(1_B\cdot g)/\lambda' B}$$

(we have $\lambda' B > 0$, since $\nu B > 0$).

According to the assumption (3) , the functions

$$[(f\cdot\lambda_y)k]_y \quad \text{and} \quad g$$

are both continuous at the point y_o, relatively to supp λ' (i.e. as functions supp $\lambda' \to \mathbb{R}$). Hence, by the lemma, we have

$$\mu^{B}k \to \frac{(f\cdot\lambda_{y_o})k}{g(y_o)} \quad \text{for} \quad B \to y_o ,$$

and this proves the theorem.

Notice, that the continuity assumption about g is, in reality, an assumption about the decomposition and the density f . It does <u>not</u> suffice to assume, that g is <u>equivalent</u> to some continuous function. The argument in the proof is valid in this case also, but the conditional distribution is defective, if the normalizing factor $g(y_o)$ must be changed in order to obtain continuity.

8. CONDITIONING IN A PRODUCT SPACE.

Let

$$X := Y \times Z$$

be a product space, and correspondingly,

$$\lambda := \lambda' \otimes \lambda''$$

a product measure. Let μ be a probability measure on X, given by the density f with respect to λ. We shall try to construct a conditional distribution with respect to the projection

$$t: X \to Y$$

given by

$$t(x) = t(y,z) := y .$$

In order to apply the results from section 7 we define

$$\lambda_y := \varepsilon_y \otimes \lambda'' .$$

Obviously then, $(\lambda', (\lambda_y | y \in Y))$ is a decomposition of λ with respect to t (see the definition of the product measure, page 338). As in section 7, the function

$$g(y) := \lambda_y f = \lambda'' [f(y,z)]_z$$

is the density of $\nu = t(\mu)$ with respect to λ'. Theorem 7.1 takes the following form:

8.1 Theorem. Let $y_o \in \text{supp } \lambda'$ be given. Suppose there exists a neighbourhood V_o of y_o such that the following conditions are satisfied:

(1) For $y \in V_o \cap \text{supp } \lambda'$, the function $[f(y,z)]_z$ is λ''-integrable (i.e. g is welldefined on $V_o \cap \text{supp } \lambda'$).

(2) $g(y_o) > 0$.

(3) The mappings

$$y \to g(y) = \lambda''[f(y,z)]_z , \qquad V_o \cap \text{supp } \lambda' \to \mathbb{R}$$

$$y \to [f(y,z)]_z \cdot \lambda'' , \qquad V_o \cap \text{supp } \lambda' \to \mathcal{M}(Z)$$

are continuous at the point y_o .

Then, the conditional distribution of $(y,z) \in (Y \times Z, \mu)$, given $y = y_o$, is defined, and given by

$$\mu^{y_o} = \frac{1}{g(y_o)} \cdot f \cdot (\varepsilon_{y_o} \otimes \lambda'') .$$

The theorem is a trivial translation of theorem 7.1, except for this: Continuity at y_o of the mapping $y \to f \cdot \lambda_y$ (page 39) is, in the present case, equivalent to continuity of the mapping

$$y \to f \cdot (\varepsilon_y \otimes \lambda'') \quad , \qquad V_o \cap \operatorname{supp} \lambda' \to \mathcal{M}(Y \times Z)$$

at the point y_o . But it follows easily from theorem A 5 (page 337) that this is equivalent to continuity at y_o of the mapping

$$y \to [f(y,z)]_z \cdot \lambda'' \quad , \quad V_o \cap \operatorname{supp} \lambda' \to \mathcal{M}(Z) \, ,$$

as assumed in theorem 8.1 above.

The conclusion of theorem 8.1 looks somewhat complicated, the conditional distribution of (y,z) being concentrated on the fibre $\{y_o\} \times Z$ in the product space. The conclusion may, however, be rewritten as follows by theorem 5.2 (page 28-29):

The conditional distribution of z, given $y = y_o$ (for $(y,z) \in (Y \times Z, \mu)$), is defined and given by the density

$$\frac{1}{g(y_o)} [f(y_o,z)]_z \qquad (= \text{const} \cdot f(y_o,\cdot))$$

with respect to λ'' .

Just as in theorem 7.1, the condition (3) in theorem 8.1 is satisfied, if f and g are continuous. But continuity of f is far from necessary, as the following theorem shows:

8.2 Theorem. Suppose, that the conditions (1) and (2) of theorem 8.1 (page 43) are satisfied. Furthermore, assume that

$$(3)' \qquad \lambda''\,|\,[f(y,z) - f(y_o,z)]_z\,| \to 0 \quad \text{for} \quad y \in V_o \cap \operatorname{supp}\lambda' \;,\; y \to y_o \;.$$

Then, condition (3) of theorem 8.1 (and so the conclusion of that theorem) is also satisfied.

Proof: The condition (3)' expresses that the mapping

$$\begin{array}{ccc} y & \to & [f(y,z)]_z \\ V_o \cap \operatorname{supp}\lambda' & \to & L(\lambda'') \end{array}$$

is continuous (the space $L(\lambda'')$ is defined on page 342). Condition (3) follows immediately from the fact that the mappings

$$h \to \lambda'' h \;, \qquad L(\lambda'') \to \mathbb{R}$$

and

$$h \to h\cdot\lambda'' \;, \qquad L(\lambda'') \to \mathcal{M}(Z)$$

are continuous (which is easily proved).

Example: Put

$$Y := \mathbb{R}^m$$
$$Z := \mathbb{R}^{n-m}$$
$$X := Y \times Z = \mathbb{R}^n ,$$

and let

$$\lambda^n = \lambda^m \otimes \lambda^{n-m}$$

denote the Lebesgue measure on $\mathbb{R}^n$, expressed as the product of the Lebesgue measures on $\mathbb{R}^m$ and $\mathbb{R}^{n-m}$. Let f denote a probability density on $\mathbb{R}^n$, and define

$$g(y) := \int \cdots \int f(y_1,\ldots,y_m;z_1,\ldots,z_{n-m})dz_1\ldots dz_{n-m} .$$

The function g is defined almost everywhere (Fubini's theorem). Let

$$t: \mathbb{R}^n \to \mathbb{R}^m$$

be the projection

$$t(x) = t(y,z) := y .$$

In case f is continuous it follows from theorem 8.1, that the conditional distribution of $x = (y,z) \in (\mathbb{R}^n, f\cdot\lambda^n)$, given $y = y_o$, is defined at all points y_o where g is continuous and (strictly) positive.

In case f is not continuous, theorem 8.2 can be applied. In all cases of practical interest, the condition (3)' is satisfied, except possibly for y_o in a closed null set of singularity points (usually the discontinuity points for g). We shall not here go into a detailed discussion of the regularity conditions one may impose on f in order to obtain pointwise or almost everywhere existence of the conditional distributions. The considerations in section 17 will show that the conditional distributions are defined, at least almost everywhere, in all reasonable situations of this type. For counterexamples, see section 30.

Conditioning in the continuous case. In chapter IV, conditioning problems in the continuous case (as defined on page 22) will be handled by tools from differential geometry. For the sake of completeness we shall here outline the wellknown , more elementary methods, usually applied to probability theory.

The idea is, in short, that conditioning problems of suitably regular type can always be transformed into problems of the "product type" , as discussed in this section:

Let

$$t: \mathbb{R}^n \to \mathbb{R}^m \qquad (m \leq n)$$

be a continuously differentiable transformation with differential $Dt(x) = ((\partial y_i / \partial x_j))$ of rank m . Suppose, that

we can choose a "supplementary transformation"

$$s: \mathbb{R}^n \to \mathbb{R}^{n-m}$$

such that the transformation

$$(t,s) : \mathbb{R}^n \to \mathbb{R}^m \times \mathbb{R}^{n-m}$$

maps $\mathbb{R}^n$ <u>diffeomorphic</u> into an open set $X' \subseteq \mathbb{R}^m \times \mathbb{R}^{n-m}$ (by a diffeomorphic transformation we mean a forwards and backwards continuously differentiable 1-1 mapping). Let μ be a probability measure on $\mathbb{R}^n$, given by the density f with respect to Lebesgue measure. By the integral transformation theorem, $(t,s)(\mu)$ has the density

$$f'(x') = 1_{X'}(x') \frac{f((t,s)^{-1}(x'))}{|\det D(t,s)((t,s)^{-1}(x'))|}$$

with respect to Lebesgue measure on $\mathbb{R}^m \times \mathbb{R}^{n-m}$. By theorem 8.1 (possibly via theorem 8.2), the conditional distribution of $x' \in (\mathbb{R}^m \times \mathbb{R}^{n-m}, (t,s)(\mu))$, given a fixed value of $y = t(x) = (x'_1, \ldots, x'_m)$, can now be computed, and for most purposes this conditional distribution is quite as good as the conditional distribution of x itself. A representation of μ^{y_o} as a transformed distribution is actually the best we can hope for, as long as we have no tools for a description of μ^{y_o} by its density with respect to some "area measure"

on the surface $t^{-1}(y_o)$.

The method is, in fact, more powerfull than indicated above; some of the seemingly restrictive assumptions (like the maximal rank of the differential, and the existence of a "supplementary transformation" s) may be satisfied after a simple adjustment of the problem. If the rank is not maximal, a "projection" of $\mathbb{R}^m$ on a space of lower dimension may help. The transformation s can always be chosen <u>locally</u>, if the rank condition is satisfied; hence, a <u>division</u> of $\mathbb{R}^n$ into a number of open sets and a closed null set, followed by an application of a result about "piecewise conditioning" (see section 28), may solve this problem. A closed μ-null set (for example, a set of singularity points for t) can always be removed from $\mathbb{R}^n$ without changing the problem essentially, since the methods indicated can be applied to open subsets of Euclidean spaces, as well as to Euclidean spaces.

9. CONDITIONING IN A STOCHASTIC PROCESS.

In the theory of stochastic processes, many results must be deduced from corresponding finite dimensional results, by "approximation" of a process by its finite dimensional marginal distributions.

In relation to conditioning, two different kinds of approximation seem to be of interest:

(1) Conditioning in a stochastic process: Approximation of a conditional distribution of the whole process by the conditional distribution of finitely many states $x_{t_1}, \dots, x_{t_n}$.

(2) Conditioning on a stochastic process: Approximation of a conditional distribution, given the sample function of a stochastic process, by the conditional distribution (of the same thing), given the values of the sample function at finitely many time points.

Approximations of the type (1) turn out to be legal in exactly the sense one should expect. Approximations of the type (2) is a more delicate matter; no local results seem to be valid, so we will have to return to the problem later (section 29).

Consider a compact product space

$$X_I = \prod_{i \in I} X_i \ .$$

For $I_1 \subseteq I$, we write

$$X_{I_1} := \prod_{i \in I_1} X_i \ ,$$

and for

$$I_1 \subseteq I_2 \subseteq I$$

we let

$$p_{I_2 I_1} : X_{I_2} \to X_{I_1}$$

denote the projection, i.e.

$$p_{I_2 I_1}(x_i | i \in I_2) = (x_i | \ i \in I_1) \ .$$

Points in X_{I_1} are denoted x_{I_1} , y_{I_1} etc. The finite dimensional marginal distributions corresponding to a probability measure μ_I on X_I (see the appendix, page 354) are denoted

$$\mu_M := p_{IM}(\mu_I) \in \mathcal{P}(X_M) \ , \quad M \in \mathcal{S}_o \ ,$$

where $\mathcal{S}_o$ denotes the set of finite subsets of I .

9.1 Theorem. For a homomorphism

$$t: (X_I, \mu_I) \to (Y, \nu)$$

and a point $y_o \in \operatorname{supp} \nu$, the following two conditions are equivalent:

(1) The conditional distribution $\mu_I^{y_o}$ of the process $x_I \in (X_I, \mu_I)$, given $t(x_I) = y_o$, is defined.

(2) For $M \in \mathcal{P}_o$, the derived conditional distribution $\mu_M^{y_o}$ of $x_M := p_{IM}(x_I)$, given $t(x_I) = y_o$, is defined.

In case of existence, the conditional distributions $\mu_M^{y_o}$ constitute the consistent family for the "conditioned process" $\mu_I^{y_o}$.

Proof: The statement (1) $\Rightarrow$ (2) follows immediately from theorem 5.1, and, in addition, this theorem proves that the measures $\mu_M^{y_o}$ are the finite dimensional marginal distributions for $\mu_I^{y_o}$. It remains to prove that (2) $\Rightarrow$ (1): Suppose, that the conditional distributions $\mu_M^{y_o}$ are defined. Let M and N be finite subsets of I, such that $N \subseteq M$. We may then regard the stochastic variables

$$y = t(x_I)$$

and $$x_N = p_{IN}(x_I)$$

as derived from the stochastic variable

$$(y, x_M) = (t(x_I) , p_{IM}(x_I))$$

by $(y, x_N) = (y, p_{MN}(x_M)) = (1_Y \times p_{MN})(y, x_M)$.

By theorem 5.2, the conditional distribution of (y, x_M), given $y = y_o$, is defined and equal to

$$\varepsilon_{y_o} \otimes \mu_M^{y_o} .$$

According to theorem 5.1 , the derived conditional distribution $\mu_N^{y_o}$ can be computed as a transformation of $\varepsilon_{y_o} \otimes \mu_M^{y_o}$, namely (of course) as

$$\mu_N^{y_o} = p_{MN}(\mu_M^{y_o}) .$$

This proves the consistency of the family $(\mu_M^{y_o} | M \in \mathcal{F}_o)$. By Kolmogorov's consistency theorem (page 354) a probability measure on X_I is determined. Let us (being optimistic) denote this measure by $\mu_I^{y_o}$.

For any finite subset M of I , we have

$$p_{IM}(\mu_I^{y_o}) = \mu_M^{y_o} = \lim_{B \to y_o} p_{IM}\mu_I^B \ .$$

It follows immediately from this equation, that any contact point of the net $(\mu_I^B | B \to y_o)$ has $(\mu_M^{y_o} | M \in \mathcal{P}_o)$ as its consistent family. By the consistency theorem, we conclude that $\mu_I^{y_o}$ is the only possible contact point. Since $P(X_I)$ is compact, we must have

$$\lim_{B \to y_o} \mu_I^B = \mu_I^{y_o} \ ,$$

i.e. $\mu_I^{y_o}$ is the conditional distribution of $x_I \in (X_I, \mu_I)$, given $t(x_I) = y_o$.

CHAPTER IV : CONDITIONING IN THE CONTINUOUS CASE

10. SOME REMARKS ON MATHEMATICAL PREREQUISITES.

In the following six sections (10-15) some tools from algebra and differential geometry will be introduced.

It has been somewhat difficult for me to decide how much of the mathematical framework to include here, and how much to assume to be known.

On the one hand, a complete introduction to the branches of mathematics which we are going to use would become much too extensive; furthermore it would be useless, because I might as well refer to other expositions of a higher quality. None of the results are new.

On the other hand, many of the special results which we shall need can be understood with less knowledge than normally presumed in the litterature. This can be illustrated by the following two examples:

The geometric measure on a Riemann manifold is, from a geometric and measure theoretic point of view, a very simple concept. Now, for obvious reasons the theory of this measure has been

developed by differential geometers, and within differential geometry it is easy to construct the geometric measure as a special case of the measures one can obtain by integration of differential (tensor-) forms. In our exposition, a measure theoretic approach seems more natural, and does not require any use of tensor products and forms.

For determinants on Euclidean vectorspaces we shall need a few results which are obviously special cases of much more general results from the extensive theory of multilinear algebra. The most elegant theory is obtained by means of the exterior algebra of a vectorspace (see MacLane and Birkhoff(1967), chapter XVI). A proper construction of the exterior algebra requires some knowledge of categories and functors. The results needed here can rather easily be derived from the more elementary theory of determinants for matrices.

For this reason I have decided to introduce the necessary tools here, without assuming prior knowledge of too special subjects. A further motivation for this is my belief in the relevance of the local definition of a conditional distribution. It is wellknown that the applicability of a theory depends strongly on the number of books one has to take a look in (or even read) in order to understand it, cfr. the remarks on stochastic processes on top of page 8.

Below, I shall comment on some subjects that may be more

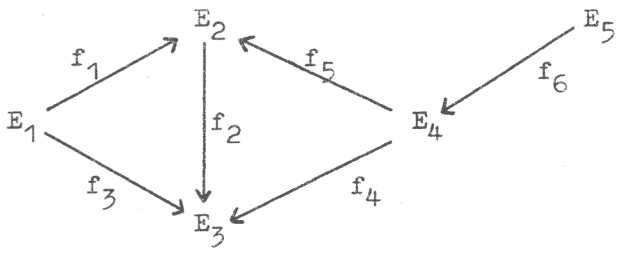
or less new to the reader.

Category theory is not really used as a tool, but the language of category theory will sometimes be used, just to clarify things. Terms from category theory will be explained to the necessary extend. Formulation in terms of categories and functors may be of help to readers knowing this theory, and to other readers it will represent no further difficulties than those of a somewhat peculiar (but convenient) terminology.

Litterature: MacLane and Birkhoff (1967), Mitchell (1965).

Diagrams can be defined as a special sort of functors, but they can also be introduced as a more elementary concept: By a diagram we mean a (usually finite) set of objects E_1 , E_2 , ... , connected by homomorphisms f_1 , f_2 , ... We draw a digram as a collection of points (representing the objects) connected by arrows (representing the homomorphisms).

Example:



As

objects	and	homomorphisms

we may take

sets	and	functions ,
topological spaces	and	continuous functions ,
vectorspaces	and	linear mappings ,
probability fields	and	homomorphisms, as defined on page 16

etc.etc.

The character of the objects and the homomorphisms is determined by the category, we are studying at present (the category of sets and functions, the category of topological spaces and continuous mappings, etc.). The idea is, in most cases, that the objects are sets with some kind of structure, and the homomorphisms are mappings, preserving the structure in some sense. Hence, for most purposes, we may regard a category as a concept of structure.

A diagram is said to commute (or to be commutative) if it is consistent with respect to compositions; for example, the diagram on page 57 commutes if and only if

$$f_3 = f_2 \circ f_1$$

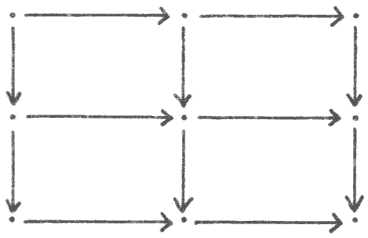
and

$$f_4 = f_2 \circ f_5 \ .$$

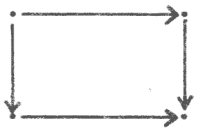
Thus, commutativity means that an element in one of the sets (objects) can be manoeuvred into at most one element in each of the other sets, by means of the homomorphisms in the diagram. Different routes between two objects in the diagram lead to the same result.

On page 26 and page 28 very simple diagrams in the category of probability fields and homomorphisms were studied. Diagrams like those do, of course, always commute.

In section 11 we shall meet diagrams like



i.e., diagrams pieced together of "squares" of the form



Such diagrams are easily seen to commute if and only if any of the squares commutes (unless the diagrams have "holes"; which they never have in our applications).

Exactness. A diagram of the form

$$\longrightarrow E_1 \xrightarrow{f_2} E_2 \xrightarrow{f_3} E_3 \xrightarrow{f_4} E_4 \longrightarrow$$

(finite or infinite) is said to be exact, if, place for place, the image of the preceding homomorphism equals the kernel of the next, i.e. if

$$\mathrm{Im}(f_i) = \mathrm{Ker}(f_{i+1})$$

$$\text{i.e.} \quad f_i(E_{i-1}) = f_{i+1}^{-1}(0) .$$

This definition requires, of course, that the concept of kernel is defined, as for example is the case in the category of vectorspaces.

Example: In the category of vectorspaces, the diagram

$$0 \longrightarrow E \xrightarrow{f} F$$

is exact if and only if f is injective. Dually, the diagram

$$E \xrightarrow{f} F \longrightarrow 0$$

is exact if and only if f is surjective (for simplicity we write 0 instead of $\{0\}$).

Product spaces and matrices. A linear mapping

$$a: \mathbb{R}^n \to \mathbb{R}^m$$

will always be identified with its $m \times n$-matrix in the usual manner. For mappings between products of vectorspaces others than $\mathbb{R}$, a similar notation is used. For example, a linear mapping

$$a: U_1 \times U_2 \to V_1 \times V_2$$

is written as a matrix

$$a = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} ,$$

the elements of which are linear mappings

$$\begin{aligned} a_{11} &: U_1 \to V_1 \\ a_{12} &: U_2 \to V_1 \\ a_{21} &: U_1 \to V_2 \\ a_{22} &: U_2 \to V_2 \end{aligned}$$

determined by

$$v = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = au = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}$$

$$= \begin{bmatrix} a_{11}u_1 + a_{12}u_2 \\ a_{21}u_1 + a_{22}u_2 \end{bmatrix} .$$

Notice, that the notation requires that elements of the product spaces are written as columns, which we shall usually not do, except when matrix notation is actually applied (but most of the time we write (u_1,u_2) etc.).

In case the vectorspaces U_1, U_2, V_1 and V_2 are spaces of the type $\mathbb{R}^n$ (i.e. Euclidean vectorspaces with selected orthonormal bases), the above notation is consistent with usual matrix notation, in the sense that the matrix a can be pieced together of the "blocks" a_{ij}.

For mappings of the form

$$a = \begin{bmatrix} a_1 & 0 \\ 0 & a_2 \end{bmatrix} : U_1 \times U_2 \to V_1 \times V_2$$

(i.e. $a(u_1,u_2) = (a_1u_1 , a_2u_2)$) we write

$$a_1 \times a_2 := a .$$

Differential geometry will be developped to the extend needed. Readers with a background in differential geometry may skip section 12 and 13, and possibly section 14. It should be

noticed that our definition of a submanifold (page 107) is different from the usual one .

Litterature: Hicks (1965) , Helgason (1962), Dieudonné (1970).

The results of section 12, 13 and 14 can, with few exceptions, be found in Hicks (1965).

Decomposition of the geometric measure. The results of section 15 are, probably, special cases of more general results, to be found in expositions like

Whitney (1957) , Federer (1969).

But the formulation given here, based on determinants on Euclidean vectorspaces, is different from anything I have seen in the litterature, so I am not able to give any explicit references.

11. DETERMINANTS ON EUCLIDEAN VECTORSPACES.

By a Euclidean vectorspace we mean a finite dimensional vectorspace E over the reals, equipped with an inner product

$$\begin{aligned} (u,v) &\to (u|v)_E \quad \text{(or just } (u|v)\text{)} \\ E \times E &\to \mathbb{R} . \end{aligned}$$

By selecting an orthonormal base, we can identify E with the space $\mathbb{R}^n$ ($n = \dim E$), equipped with the usual inner product

$$(u|v) = \sum_{i=1}^{n} u_i v_i .$$

It is essential, however, that definitions and results in the following are independent of a possible choice of base.

Adjoint linear mapping. Let E and F be Euclidean vectorspaces, and let

$$a: E \to F$$

be a linear mapping. The adjoint linear mapping

$$a^*: F \to E$$

is the mapping taking $v \in F$ into the element $a^*(v) \in E$ determined by the equations

$$(a(u)|v)_F = (u|a^*(v))_E \quad , \qquad u \in E \ .$$

The following properties of adjoint mappings are easily verified (see MacLane and Birkhoff (1967), or any introduction to linear operators on Hilbert spaces):

$$(ab)^* = b^*a^* \quad \text{for} \quad a: F \to G \quad , \quad b: E \to F \ ;$$

$$(a^*)^* = a \ .$$

$$\begin{aligned} &a \text{ injective} \iff a^* \text{ surjective} \iff a^*a \text{ bijective} \\ &a \text{ surjective} \iff a^* \text{ injective} \iff aa^* \text{ bijective} \ . \end{aligned}$$

$$(a^{-1})^* = (a^*)^{-1} \qquad \text{for } a \text{ bijective.}$$

If $a: E \to F$ has the matrix $A = ((a_{ij}))$ with respect to given, orthonormal bases in E and F, then $a^* : F \to E$ has the <u>transposed</u> matrix $A' = ((a_{ji}))$ with respect to the same bases.

<u>Subspace and quotient space</u>. Let E' be a linear subspace of a Euclidean vectorspace E. Define an inner product $(\ |\ )_{E'}$ on E', simply as the restriction of $(\ |\ )_E$. Equipped with this inner product, E' is called a <u>Euclidean subspace</u> of E.

Dually, let the vectorspace E'' be a <u>quotient space</u> in E,

i.e., let there be given a surjective linear mapping (the "projection" on the quotient space)

$$p: E \to E'' .$$

The inner product on E induces an inner product on the quotient space, as follows: Let $E' \subseteq E$ denote the kernel for p. The orthogonal complement $E'^{\perp}$ is mapped bijectively on E'' by p. The Euclidean structure (the inner product) on E'' is defined by this identification

$$E'' \approx E'^{\perp}$$

of E'' with a Euclidean subspace of E. We call E'' (under this structure) a Euclidean quotient space in E. We shall sometimes write

$$E'' = E/E' .$$

The mapping p will be called the coimbedding (dually to imbedding) on the quotient space (the more common term projection has another meaning here, as an orthogonal projection, i.e. an endomorphism $p:E \to E$ satisfying $p^* = p$ and $p^2 = p$). Our definition of the inner product on E'' seems artificial, but it will soon be obvious that the concept of Euclidean quotient space is quite analogous (dual, to be precise) to the concept of Euclidean subspace.

Isometries and coisometries. Let E and F be Euclidean vectorspaces. A linear mapping

$$a: E \to F$$

is said to be isometric, if it preserves inner products:

$$(a(u)|a(v))_F = (u|v)_E \ .$$

An isometric mapping (or an isometry) is injective, since it preserves distances. Regarded as a mapping $E \to a(E)$, an isometry is an isomorphism according to vector space structure and Euclidean structure. This means, that an isometry can, in a sense, be regarded as an imbedding of a Euclidean subspace.

It is easy to prove, that a is isometric if and only if

$$a^*a = 1_E \ .$$

Dually, a linear mapping is said to be coisometric if

$$aa^* = 1_F \ .$$

Obviously, a is coisometric if and only if a^* is isometric.

Let a be coisometric. Then a is surjective (a^* being injective). Put

$$K := \mathrm{Ker}(a) = a^{-1}(0) \ .$$

The linear mapping

$$a^*a : E \to E$$

then equals the orthogonal projection onto $K^\perp$:

The equations

$$(a^*a)(a^*a) = a^*(aa^*)a = a^*a$$

and

$$(a^*a)^* = a^*a^{**} = a^*a$$

show that a^*a is an orthogonal projection, and the kernel of a^*a equals the kernel of a , a^* being injective.

For two vectors u , $v \in K^\perp$ we have

$$(a(u)|a(v))_F = (aa^*a(u)|a(v))_F = (a^*a(u)|a^*a(v))_E$$

$$= (u|v)_E ,$$

i.e. a maps $K^\perp$ isomorphic (according to inner products) on F . This means that F can be identified with the Euclidean quotient space E/K . Hence, a coisometry can, in a sense, be regarded as the coimbedding on a quotient space.

The concepts of isometry and coisometry thus remove the need of the concepts Euclidean subspace and Euclidean quotient space. According to Euclidean structure, it makes no difference whether we think of a subspace as a concrete subset, or just

assume that an isometric "imbedding" is given.

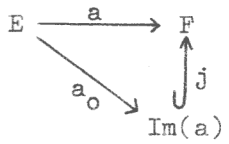
Image and coimage. Let E and F be Euclidean vectorspaces. The image of a linear mapping $a: E \to F$ is the Euclidean subspace

$$Im(a) := a(E)$$

of F. We have a unique factorization

$$a = j \circ a_o$$

of a through its image; here j denotes the imbedding of $a(E)$ in F, and a_o is uniquely defined by the equation above. We may also define a_o as the (unique) linear mapping such that the diagram



commutes. Thus a_o is simply a, equipped with a new co-domain. Arrows of the form $\hookrightarrow$ are used for isometries in the following (i.e. for imbeddings, the sign $\hookrightarrow$ being, of course, derived from $\subset$).

Our definition of the image is not quite exact: The image is, of course, more than a Euclidean space; it is a Euclidean subspace of F, regarded as a subspace in a certain manner.

This is certainly subsumed when we say that the factorization of a through its image is unique.

The dual concept can be defined as follows:

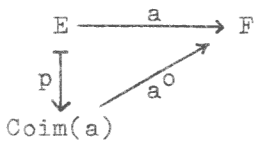
Again, let $a: E \to F$ be given. The <u>coimage</u> of a is the quotient space

$$\mathrm{Coim}(a) := E/\mathrm{Ker}(a) .$$

We have a unique factorization

$$a = a^{o} \circ p$$

of a through its coimage, illustrated by the commuting diagram



Here, p denotes the coimbedding on the quotient space. The mapping a^{o} is simply a, equipped with a new domain constructed by identification of elements with the same image under a. Arrows of the form $\longmapsto$ denote coimbeddings.

Standard symbols like $\hookrightarrow$ and $\longmapsto$ (and later we shall introduce $\xrightarrow{\sim}$ for isomorphisms) are convenient, because they enable us to build a lot of information into the diagrams. For example, in many connections we need not introduce particular

names for imbeddings and coimbeddings if it is obvious from the context what the arrows stand for.

It should be noticed that the definitions given here do not in any trivial manner come out as special cases of definitions in abstract category theory. The category of Euclidean spaces and linear mappings (i.e. the category we are studying at present) does not reflect the structures we are interested in, since the homomorphisms do not preserve the Euclidean structure. Euclidean subspaces, quotient spaces, images and coimages are constructed within the category of finite dimensional vectorspaces, and then equipped with a (forwards or backwards) induced Euclidean structure. Notice, for example, that image and coimage for a linear mapping are not Euclidean isomorphic, though they are, in a canonical sense, isomorphic as vectorspaces. The Euclidean structure of the image comes from the codomain, while the structure of the coimage is induced by the structure of the domain.

Determinants. Let E and F be Euclidean spaces of the same dimension n. Let $a: E \to F$ be a linear mapping, and let A denote the matrix for a with respect to given, orthonormal bases. By the determinant $|a|$ of a we mean the absolute value of the determinant of A, i.e.

$$|a| := |\det A|.$$

This number $|a| \geq 0$ is independent of the choice of orthonormal bases, since the matrix A_1 of a with respect to some other orthonormal bases appears from A by

$$A_1 = UAV' ,$$

where the coordinate shift matrices U and V are orthonormal (or unitary). Orthonormal matrices having determinants ± 1 , the absolute value of the determinant remains unchanged.

The use of the name determinant, instead of something more correct like "absolute determinant" or "positive determinant" will not lead to misunderstandings, since the signed determinant can be given no meaning under our assumptions. It requires, at least, that the spaces E and F are equipped with orientations, or that $E = F$. In the case $E = F$ the (signed) determinant is (and should be defined as) independent of the Euclidean structure, but such "endomorphism determinants" should not be confused with our "Euclidean determinants" .

From wellknown properties of (matrix-) determinants, the following rules are easily proved:

$$|1_E| = 1 \qquad \text{(also for } E = 0\text{)}.$$

For $a: E \to F$

$b: F \to G$,

where E, F and G are Euclidean vectorspaces of dimension n ,

$$|ba| = |b||a| .$$

$$|a^*| = |a| .$$

$$|a| = \sqrt{|a^*a|} = \sqrt{|aa^*|} .$$

The last equations suggest a generalization of the concept of a determinant: The mappings

$$a^*a : E \to E$$

and

$$aa^* : F \to F$$

having welldefined determinants, whatever the dimensions of E and F might be, we can define

$$|a|_o := \sqrt{|a^*a|}$$

and

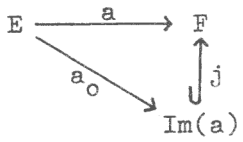
$$|a|^o := \sqrt{|aa^*|} .$$

These "generalized determinants" are, in fact,"proper" determinants, in the following sense:

11.1 Theorem. Let E and F be Euclidean vectorspaces, and let

$$a: E \to F$$

be a linear mapping. Consider the factorization



of a through its image. Then, $|a|_o$ is different from 0 if and only if dim E = dim (Im(a)) (i.e. if and only if a is injective), and in that case

$$|a|_o = |a_o| .$$

Proof: We have

$$\begin{aligned} & \dim E = \dim (\mathrm{Im}(a)) \\ \iff & a \text{ is injective} \\ \iff & a^*a \text{ is bijective} \\ \iff & |a^*a| \neq 0 \\ \iff & |a|_o \neq 0 ; \end{aligned}$$

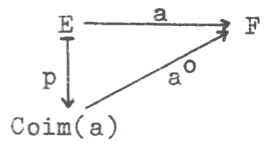
this proves the first statement of the theorem. In order to prove the identity $|a|_o = |a_o|$ in case E and Im(a) have equal dimensions, we just notice that (j being isometric)

$$(|a|_o)^2 = |a^*a| = |(ja_o)^*ja_o| = |a_o^*j^*ja_o|$$

$$= |a_o^*a_o| = |a_o|^2 .$$

The dual theorem looks like this:

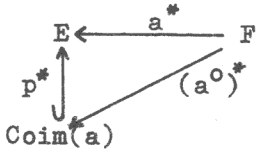
11.2 Theorem. Let



be the factorization of a through its coimage. Then, $|a|^o$ is different from 0 if and only if dim F = dim Coim(a) (i.e. if and only if a is surjective), and in that case

$$|a|^o = |a^o| .$$

Proof: The theorem can be proved quite analogously to theorem 11.1 above. But we can also make a more direct application of that theorem, by a typical "dualization" of the problem: Taking adjoints in the diagram above, we obtain a commutative diagram



Here, p^* is isometric and $(a^o)^*$ is injective; thus Coim(a) simply plays the role of the image for a^*. By the relations

$$|a|^o = |a^*|_o$$

and

$$a \text{ surjective} \iff a^* \text{ injective} ,$$

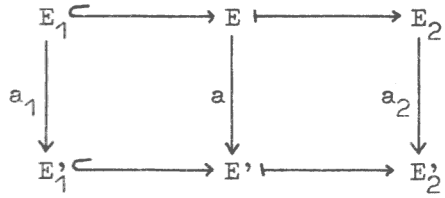
the theorem follows immediately from the preceding theorem. In addition, we have derived a canonical identification

$$\text{Coim}(a) \approx \text{Im}(a^*) .$$

We are now in the position to prove certain relations between determinants of mappings in commuting diagrams, satisfying certain exactness conditions. The results will be applied in section 15 and section 18; it may be a good idea to postpone the reading of the rest of this section until the necessity of the results becomes more obvious.

The following lemma reflects a basic property of determinants. The two theorems 11.4 and 11.5 are just more convenient formulations of -essentially- the same result.

11.3 Lemma. Consider a commutative diagram of the form



of Euclidean spaces and linear mappings (arrows $\hookrightarrow$ and $\mapsto$ denoting isometries and coisometries, respectively). Suppose that the rows of the diagram are exact, i.e. E_1 and E_1' are, as subspaces of E and E', the kernels of the two coimbeddings. Further assume that

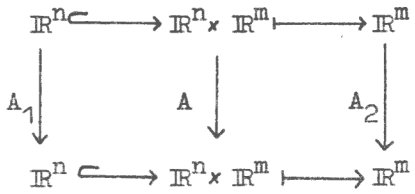
$$\dim E_1 = \dim E_1' = n$$
$$\dim E_2 = \dim E_2' = m$$
$$\text{(and thus } \quad \dim E = \dim E' = n + m \text{).}$$

Then,

$$|a| = |a_1||a_2| \ .$$

Proof: An orthonormal base for E can be chosen in such a way that the first n base vectors constitute a base for the subspace E_1. The remaining m unitvectors then constitute an orthonormal base for the subspace $E_1^{\perp} \approx E_2$, i.e. they are mapped into an orthonormal base for E_2 by the coimbedding. A similar choice of bases can be made for the spaces E_1', E' and E_2'. By means of these bases, the Euclidean vector-spaces of the diagram are identified with spaces $\mathbb{R}^n$, $\mathbb{R}^m$ etc.,

so we get a commuting diagram



where A_1 , A and A_2 denote the matrices of a_1 , a and a_2 with respect to the selected bases, and the imbeddings and coimbeddings are simply the imbedding

$$y \rightarrow \begin{bmatrix} y \\ 0 \end{bmatrix}$$

of the first component and the coimbedding (projection)

$$\begin{bmatrix} y \\ z \end{bmatrix} \rightarrow z$$

on the second component in the product space $\mathbb{R}^n \times \mathbb{R}^m$. Let the matrix A be written as a "block matrix" (see page 62) on the form

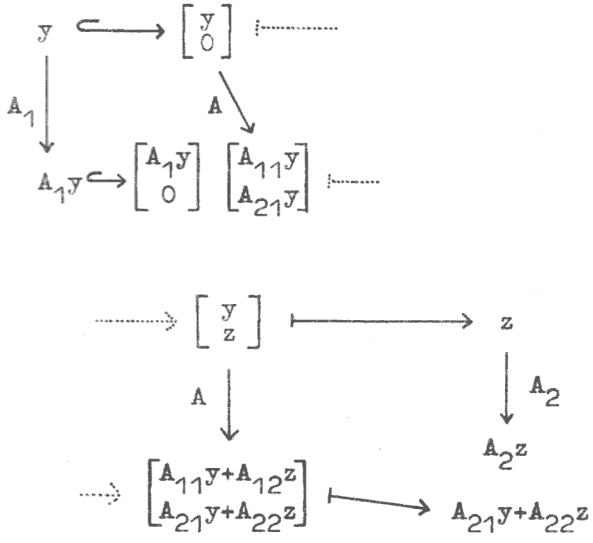
$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$$

corresponding to the product structure of its domain and

codomain. Thus

$$A \begin{bmatrix} y \\ z \end{bmatrix} = \begin{bmatrix} A_{11}y + A_{12}z \\ A_{21}y + A_{22}z \end{bmatrix} .$$

Now, let us follow an element y of $\mathbb{R}^n$ and an element $\begin{bmatrix} y \\ z \end{bmatrix}$ of $\mathbb{R}^n \times \mathbb{R}^m$ through the diagram:



The commutativity of the diagram implies that

$$\begin{bmatrix} A_{11}y \\ A_{21}y \end{bmatrix} = \begin{bmatrix} A_1y \\ 0 \end{bmatrix}$$

and $$A_{21}y + A_{22}z = A_2z .$$

These equations being valid for all $y \in \mathbb{R}^n$ and all $z \in \mathbb{R}^m$ we conclude that

$$A_{11} = A_1$$
$$A_{21} = 0$$
$$\text{and} \qquad A_{22} = A_2 \ .$$

Hence, A has the form

$$A = \begin{bmatrix} A_1 & A_{12} \\ 0 & A_2 \end{bmatrix} \ .$$

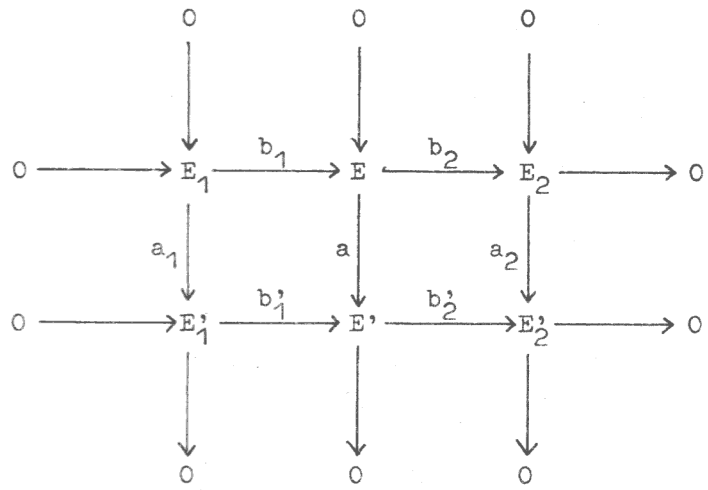
From wellknown properties of determinants we conclude that

$$\det A = \det \begin{bmatrix} A_1 & A_{12} \\ 0 & A_2 \end{bmatrix} = \det \begin{bmatrix} A_1 & 0 \\ 0 & A_2 \end{bmatrix}$$

$$= \det A_1 \cdot \det A_2$$

(if A_1 is regular, we can cancel the columns of A_{12} by subtraction of linear combinations of columns from A_1. If A_1 is <u>not</u> regular, we obviously have $\det A = 0$, the first n columns of A being linearly dependent). From the above equation, the lemma follows immediately.

<u>11.4 Theorem</u>. Let there be given a commutative diagram

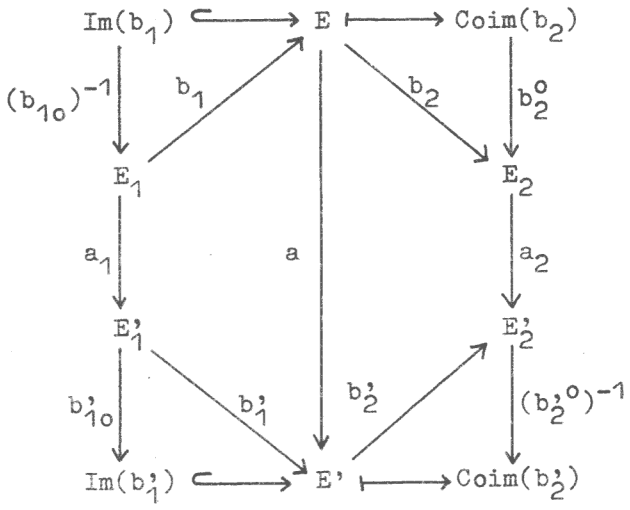


of Euclidean spaces and linear mappings. Suppose that rows and columns are exact. Then

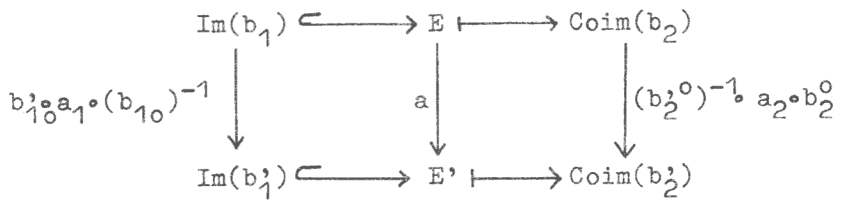
$$|b_1|_o|a||b'_2|^o = |a_1||b'_1|_o|b_2|^o|a_2| .$$

Remark: The determinants $|a|$, $|a_1|$ and $|a_2|$ are well-defined, since the exactness of the columns implies that a , a_1 and a_2 are bijections. Moreover, the exactness of the rows implies that b_1 and b'_1 are injections, while b_2 and b'_2 are surjections.

Proof: Factorizing b_1 and b'_1 through their images, and b_2 and b'_2 through their coimages, we obtain the diagram



The diagram is easily seen to commute (though the bijective mappings b_{1o} and $b_2'^{o}$ from the factorizations of b_1 and b_2' have been reversed). Contracting the columns of this diagram, we get a third commutative diagram



This diagram obviously satisfies the conditions of lemma 11.3. Thus

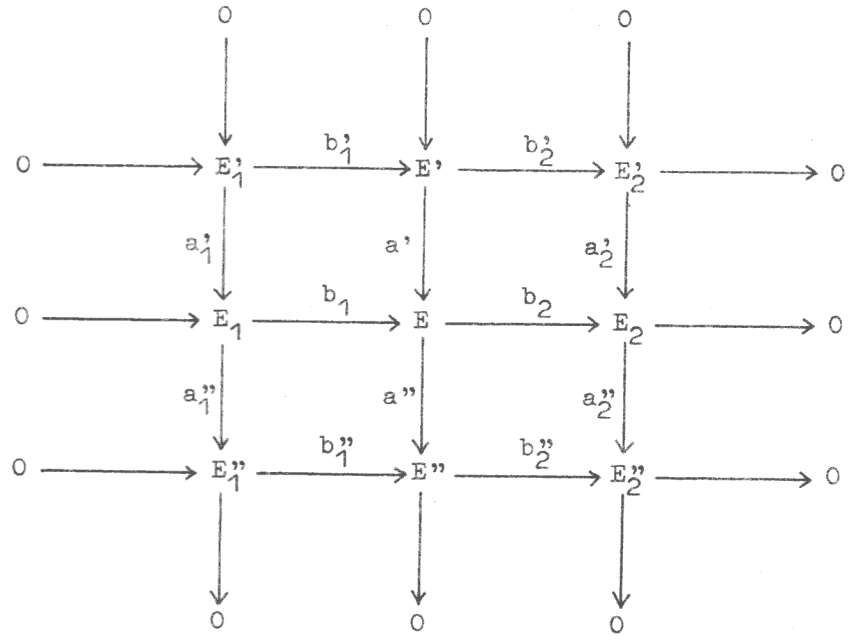
$$|a| = |b'_{1o}\circ a_1\circ (b_{1o})^{-1}||(b_2'^{o})^{-1}\circ a_2\circ b_2^{o}| = \frac{|b'_{1o}||a_1||a_2||b_2^{o}|}{|b_{1o}||b_2'^{o}|}\,.$$

Finally, applying the theorems 11.1 and 11.2 , we get

$$|a||b_1|_o|b_2'|^o = |b_1'|_o|a_1||a_2||b_2|^o \ .$$

A more general result is the following:

11.5 Theorem. Let there be given a commutative diagram

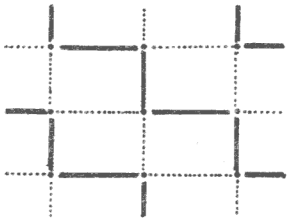


of Euclidean vectorspaces and linear mappings. Suppose that rows and columns are exact. Then

$$|b_1'|_o|a'|_o|b_2|^o|a_2''|^o|a_1''|^o|b_1''|_o$$

$$= |b_2'|^o|a_2'|_o|a_1'|_o|b_1|_o|a''|^o|b_2''|^o$$

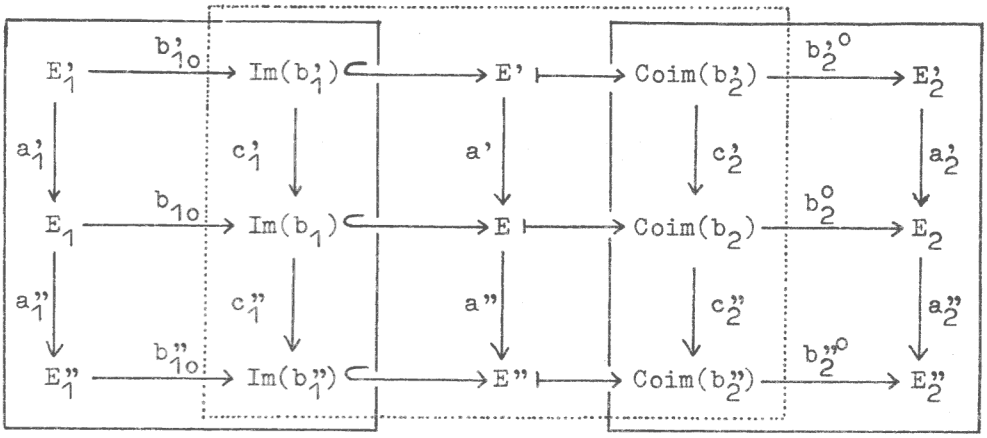
Remarks: Any of the twelve homomorphisms in the diagram occur exactly once in the formula, injections by their determinant of the type $|\ |_o$ and surjections by their determinant of type $|\ |^o$. As in theorem 11.4, it seems as if the factors are arranged at random in the formula, but if we draw the homomorphisms on the right side of the formula as dotted lines, a kind of structure becomes visible:



Thus, the factors on either side of the formula correspond to the arrows on every second zig zag line, if we imagine that the diagram is part of an infinite lattice.

Notice, that the theorem remains unchanged (except that the two sides of the formula interchange) when the diagram is reflected in its diagonal.

Proof: We extend the diagram along the middle column, by factorizing horizontal arrows through images and coimages (for convenience, the zeros are now cancelled. The frames should be ignored at present, they will be used later):



The four new homomorphisms are defined by the condition that the diagram commutes, i.e.

$$c'_1 := b_{1o} \circ a'_1 \circ (b'_{1o})^{-1}$$
$$c''_1 := b''_{1o} \circ a''_1 \circ (b_{1o})^{-1}$$
$$c'_2 := (b^o_2)^{-1} \circ a'_2 \circ b'^o_2$$
$$\text{and} \qquad c''_2 := (b''^o_2)^{-1} \circ a''_2 \circ b^o_2 \ .$$

The two upright diagrams in the full-drawn frames satisfy the conditions of theorem 11.4 (if they are overturned and equipped with a reasonable number of zeros). Hence,

$$|a'_1|_o|b_1|_o|c''_1|^o = |b'_1|_o|c'_1|_o|a''_1|^o|b''_1|_o$$

$$\text{and} \qquad |c'_2|_o|b_2|^o|a''_2|^o = |b'_2|^o|a'_2|_o|c''_2|^o|b''_2|^o \ .$$

Dividing the first formula by the second, we get (after a rearrangement of some of the factors):

$$(*)\qquad \frac{|a_1'|_o|b_1|_o|b_2'|^o|a_2'|_o|b_2''|^o}{|b_1'|_o|a_1''|^o|b_1''|_o|b_2|^o|a_2''|^o} = \frac{|c_1'|_o|c_2'|_o}{|c_1''|^o|c_2''|^o} \quad .$$

Now suppose, that the theorem we are trying to prove is valid for the diagram in the dotted frame; the conditions of the theorem are certainly satisfied. Under this assumption we have

$$1\cdot|a'|_o\cdot 1\cdot|c_2''|^o|c_1''|^o\cdot 1 = 1\cdot|c_2'|_o|c_1'|_o\cdot 1\cdot|a''|^o\cdot 1 \quad ,$$

or

$$\frac{|a'|_o}{|a''|^o} = \frac{|c_1'|_o|c_2'|_o}{|c_1''|^o|c_2''|^o} \quad .$$

This formula, together with the formula (*) above, yields the formula in the theorem.

By this argument we have reduced the problem , such that we need only prove the theorem in the special case, where all horizontal arrows are imbeddings and coimbeddings, as in the diagram in the dotted frame.

Now, in order to prove the theorem in this special case, we can reflect the diagram in its diagonal, and do the trick once more (extend along the middle column etc.). This time, the middle diagram consists of isometries and coisometries exclusively (it is easy to prove that the vertical isometries

and coisometries are not "destroyed" by the extension; actually, if a' and a" in the diagram on page 85 are isometric and coisometric, respectively, then so are c_1', c_1'', c_2' and c_2''). But for such a diagram, the theorem is trivial, all determinants being 1. This proves the theorem.

12. DIFFERENTIABLE MANIFOLDS.

By a differentiable manifold (or just a manifold) of dimension n, we mean (here) a locally compact space with a denumerable base for its topology, equipped with an n-dimensional atlas.

By an atlas $(\varphi_{U_i} \mid i \in I)$, we mean a family of (n-dimensional) charts φ_U.

By a chart, we mean a homeomorphism

$$\varphi_U : U \to U'$$

from an open subset U of X to an open subset U' of $\mathbb{R}^n$.

It is assumed that the atlas covers all localities of X, i.e. that

$$X = \bigcup_{i \in I} U_i .$$

Moreover it is assumed that overlap between charts can only give rise to differentiable deformations of the charts. In the example from which the terminology has been borrowed, we may think of the obvious requirement that longitudes and latitudes should be marked by smooth curves on the charts. The precise formulation of this consistency condition is the following:

For any two charts

$$\varphi_{U_i} : U_i \to U_i'$$

$$\varphi_{U_j} : U_j \to U_j'$$

the mapping

$$\varphi_{U_i} \circ \varphi_{U_j}^{-1} : \varphi_{U_j}(U_i \cap U_j) \to \varphi_{U_i}(U_i \cap U_j)$$

should be <u>differentiable</u>.

By <u>differentiable</u>, we mean (here and in the following) <u>infinitely often differentiable</u>.

Notice that we are a little careless about specifications of domains and codomains: We have composed the two mappings φ_{U_i} and $\varphi_{U_j}^{-1}$, though the domain of φ_{U_i} is certainly not equal to codomain of $\varphi_{U_j}^{-1}$. Such minor inaccuracies will frequently be permitted in the following.

<u>Examples</u>. An open subset of $\mathbb{R}^n$ has a canonical structure as a differentiable manifold, defined by one chart (the identity).

An m-dimensional surface X in $\mathbb{R}^n$ can be equipped as an m-dimensional manifold, for exampel by a local charting by orthogonal projection onto tangent spaces (or onto some other affine subspace of $\mathbb{R}^n$, identified with $\mathbb{R}^m$ by a choice of

origin and some base). If this is to be possible, the surface must obviously satisfy some regularity conditions. First of all it must be sufficiently smooth; it must be "open in its own dimension" (i.e. no point of the "boundary" should be included), it must not "intersect itself", nor oscillate too close to itself. The precise regularity conditions will be given under more general circumstances (see the definition of a submanifold, page 107).

Differentiability. Let X and Y be manifolds of dimensions n and m. A mapping

$$t: X \to Y$$

is said to be differentiable, if the induced mappings between charts of localities in X and Y are differentiable: For any two charts

$$\varphi_U : U \to U' \qquad (U \subseteq X , U' \subseteq \mathbb{R}^n)$$

$$\varphi_V : V \to V' \qquad (V \subseteq Y , V' \subseteq \mathbb{R}^m)$$

from the two atlases, the mapping

$$\varphi_V \circ t \circ \varphi_U^{-1} : \varphi_U^{-1}(t^{-1}V) \to V'$$

should be differentiable (i.e. , infinitely often differentiable).

A bijective mapping which, together with its inverse, is differentiable, is called a diffeomorphism. The diffeomorphisms are regarded as the isomorphisms in the category of differentiable manifolds and differentiable mappings. This means, that two differentiable structures on a set X are identified, if the identity mapping $1_X: X \to X$, regarded as a mapping from the first manifold to the second, is a diffeomorphism. The specific choice of atlas (the number of charts etc.) is not part of what we call the differentiable structure, and by a chart we mean from now on a member of some atlas inducing the given differentiable structure. Thus a chart need not be a "page" in the atlas we happened to apply in the definition, but it should be possible to include it without destroying the consistency.

Open submanifold. Let Y be an open subset of the n-dimensional differentiable manifold X. We can equip Y with a structure as differentiable manifold, taking for its atlas the family of charts of open subsets of Y. Obviously, the localities thus charted cover Y, since for any chart $\varphi_U: U \to U'$ on X, the mapping $\varphi_{Y \cap U} : Y \cap U \to \varphi_U(Y \cap U)$ is a chart.

With this differentiable structure, Y is called an open submanifold of X.

The algebras $\mathcal{C}^\infty(X)$ and $\mathcal{C}^\infty(X,x_o)$. Let $\mathcal{C}^\infty(X)$ denote the vectorspace of differentiable functions

$$f: X \to \mathbb{R} .$$

Under the usual (pointwise) multiplication, $\mathcal{C}^\infty(X)$ has the structure of a commutative $\mathbb{R}$-algebra.

For a point $x_o \in X$, the vectorspace $\mathcal{C}^\infty(X,x_o)$ of differentiable functions on neighbourhoods of x_o is defined as follows:

The elements of $\mathcal{C}^\infty(X,x_o)$ are, at first, defined as functions $f \in \mathcal{C}^\infty(U)$ on open neighbourhoods U of x_o (U being regarded as an open submanifold). But functions are identified, if they coincide on some neighbourhood of x_o. It is easy to see that this equivalence relation is compatible with addition and multiplication, such that the set of equivalence classes constitutes a commutative algebra.

Vectors. We shall define the concept of a vector, or a tangent vector, at the point $x_o \in X$. To illustrate the definition, first consider the case where X is an open submanifold of $\mathbb{R}^n$. The differential geometric aspect of a vector $v \in \mathbb{R}^n$ is that we may differentiate functions in the direction of v: For $f \in \mathcal{C}^\infty(X,x_o)$, define the derivative along v at the point x_o as

$$\frac{d}{dt}f(x_o+t\cdot v)\Big|_{t=0} = \left(v_1\frac{\partial}{\partial x_1}f(x) + \ldots + v_n\frac{\partial}{\partial x_n}f(x)\right)\Big|_{x=x_o}$$

$$= Df(x_o)v \quad .$$

(this definition is obviously independent of the choice of representative f in a neighbourhood of x_o) . Since this operation, "differentiation along v" , is linear in f , we may interpret the vector v as a linear functional

$$v: \mathcal{E}^\infty(X,x_o) \to \mathbb{R} \quad ,$$

writing vf instead of $Df(x_o)v$.

The question is now: What kind of conditions can be imposed on such a linear functional to make sure that it represents differentiation along some vector ? If we can answer this question, we have an immediate proposal for a definition of vectors on arbitrary manifolds.

The question turns out to be very easy to answer. We simply demand that the rule for differentiation of products should be satisfied:

Definition. Let x_o be a point in an n-dimensional manifold X . By a *vector* v *at* x_o , we mean a linear

mapping

$$v: \mathcal{C}^\infty(X,x_o) \to \mathbb{R}$$

satisfying the "product rule"

$$v(fg) = (vf)g(x_o) + f(x_o)(vg)$$

for $f, g \in \mathcal{C}^\infty(X,x_o)$.

By

$$D(X,x_o)$$

we denote the set of vectors at the point x_o .

Obviously, $D(X,x_o)$ is a vectorspace (a subspace of the dual to $\mathcal{C}^\infty(X,x_o)$), the product rule being linear in v . The following theorem and its proof show that we have answered the question on page 93 by this definition:

12.1 Theorem. $\dim D(X,x_o) = n$.

Proof: Let $\varphi_U : U \to U'$ be a chart of a neighbourhood of x_o such that (for convenience) $\varphi_U(x_o) = 0$.

The mapping

$$f \rightarrow f \circ \varphi_U^{-1}$$

$$\mathcal{C}^\infty(U) \rightarrow \mathcal{C}^\infty(U')$$

is bijective, linear, and preserves products (i.e. it is an algebra isomorphism). Together with the similar mappings for smaller neighbourhoods of x_o, this mapping induces in an obvius manner an algebra isomorphism

$$\mathcal{C}^\infty(X,x_o) \rightarrow \mathcal{C}^\infty(\mathbb{R}^n,0)$$

(intuitively, this isomorphism is an immediate aspect of the local diffeomorphism φ_U). Hence, we need only prove the theorem in the case $X = \mathbb{R}^n$, $x_o = 0$.

We start by proving that if a function f on a neighbourhood of 0 has all its partial derivatives $\frac{\partial}{\partial x_1} f$, ... , $\frac{\partial}{\partial x_n} f$ equal to zero at the point 0, then

$$vf = 0 \qquad \text{for all} \qquad v \in D(\mathbb{R}^n,0) .$$

This proposition follows from the fact that any differentiable function satisfying this condition can, on a neighbourhood of 0, be expressed as

$$f(x) = f(0) + \sum_{i=1}^{n} x_i f_i(x) ,$$

where the functions $f_1, \ldots, f_n$ are differentiable and satisfy

$$f_1(0) = \ldots = f_n(0) = 0 .$$

We shall not prove this statement here; it follows easily from well-known results concerning differentiable functions of several variables, see Hicks (1965), page 6-7.

Let h_i denote the function $[x_i]_x$ i.e. put

$$h_i(x) := x_i .$$

Then, on a neighbourhood of 0 we have

$$f = f(0) + \sum h_i f_i .$$

It follows from the product rule that

$$vf = f(0)(v1) + \sum (h_i(0)(vf_i) + (vh_i)f_i(0))$$

$$= f(0)(v1) .$$

But from

$$v1 = v(1 \cdot 1) = 1 \cdot v1 + (v1) \cdot 1 = 2v1$$

we conclude that $v1 = 0$, and so, $vf = 0$.

Now, for an arbitrary function $f \in \mathcal{C}^\infty(\mathbb{R}^n,0)$, define a new function f_o by

$$f_o(x) := Df(0)x = \frac{\partial}{\partial x_1} f \cdot x_1 + \ldots + \frac{\partial}{\partial x_n} f \cdot x_n$$

(here and in the following it is subsumed that the partial derivatives are evaluated at the point under consideration). This function f_o is simply the differential of f at the point 0, i.e. the best linear approximation to f near 0. The difference $f-f_o$ has the differential 0, i.e. its partial derivatives vanish at 0. According to what was proved above,

$$v(f-f_o) = 0$$

$$\text{or} \qquad vf = vf_o \quad \text{for} \quad v \in D(\mathbb{R}^n,0) .$$

Writing (as above) h_i for the function taking x into the i'th coordinate, we have

$$vf = vf_o = v(\sum (\frac{\partial}{\partial x_i} f)h_i) = \sum (\frac{\partial}{\partial x_i} f)vh_i .$$

Writing

$$v_i := vh_i ,$$

we have then

$$vf = \sum v_i (\frac{\partial}{\partial x_i} f) .$$

This equation proves that the vector v is a linear combination of the functionals

$$\frac{\partial}{\partial x_1}, \ldots, \frac{\partial}{\partial x_n} .$$

These functionals are obviously vectors (the product rule being satisfied), and they are easily seen to be linearly independent (just insert $h_1, \ldots, h_n$). This proves the theorem.

Identification of $\mathbb{R}^n$ with $D(\mathbb{R}^n, x_o)$. The vector

$$v = v_1 \cdot \frac{\partial}{\partial x_1} + \ldots + v_n \cdot \frac{\partial}{\partial x_n} \in D(\mathbb{R}^n, x_o)$$

can be regarded as differentiation along the vector

$$v = (v_1, \ldots, v_n) \in \mathbb{R}^n .$$

This identification of vectors in $D(\mathbb{R}^n, x_o)$ with elements of $\mathbb{R}^n$ was the heuristic starting point in our definition of a vector. The identification of $D(\mathbb{R}^n, x_o)$ with $\mathbb{R}^n$ is defined by the formula

$$vf = Df(x_o)v$$

where $Df(x_o)$ denotes, as usually, the matrix

$$\left[\frac{\partial}{\partial x_1} f \quad \cdots \quad \frac{\partial}{\partial x_n} f \right] .$$

The usual base of coordinate vectors in $\mathbb{R}^n$ corresponds (under the identification) to the base $\frac{\partial}{\partial x_1}, \ldots, \frac{\partial}{\partial x_n}$ for $D(\mathbb{R}^n, x_o)$.

Differentials. Let X and Y be manifolds of dimensions n and m, and let

$$t: X \to Y$$

be a differentiable mapping. For a vector $v \in D(X, x_o)$, a linear functional

$$v' : \mathcal{C}^\infty(Y, t(x_o)) \to \mathbb{R}$$

is defined by

$$v'g := v(g \circ t) .$$

This functional is obviously a vector at the point $t(x_o)$. We write

$$Dt(x_o)v := v'$$

for this vector. This defines a linear mapping

$$Dt(x_o): D(X, x_o) \to D(Y, t(x_o))$$

called the differential of t at the point x_o .

Notice, that in the case of open submanifolds X of $\mathbb{R}^n$ and Y of $\mathbb{R}^m$ the definition coincides with the usual definition of the differential, under the above identification of $\mathbb{R}^n$ with $D(X,x_o)$ (and, similarly, of $\mathbb{R}^m$ with $D(Y,t(x_o))$). This is proved by means of the formula for differentiation of a composed mapping: For $g \in \mathcal{C}^\infty(\mathbb{R}^m)$ and for $j=1,\dots,n$ we have

$$(Dt(x_o)\frac{\partial}{\partial x_j})g = \frac{\partial}{\partial x_j}(g\circ t) = \sum_{i=1}^{m} \frac{\partial t_i}{\partial x_j} \frac{\partial}{\partial y_i} g ,$$

$$\text{i.e.} \qquad Dt(x_o)\frac{\partial}{\partial x_j} = \sum_{i=1}^{m} \frac{\partial t_i}{\partial x_j} \cdot \frac{\partial}{\partial y_i} \quad .$$

Thus $Dt(x_o)$ has the matrix $((\frac{\partial t_i}{\partial x_j}))$ with respect to the bases

$$\frac{\partial}{\partial x_1}, \dots, \frac{\partial}{\partial x_n} \quad \text{of} \quad D(X,x_o)$$

and $$\frac{\partial}{\partial y_1}, \dots, \frac{\partial}{\partial y_m} \quad \text{of} \quad D(Y,t(x_o)) \quad .$$

The rule for differentiation of composed mappings is also valid on "abstract" manifolds; the proof is trivial, because the concrete aspects of the rule are included in the argument above:

12.2 Theorem. Let X, Y and Z be differentiable manifolds, and let

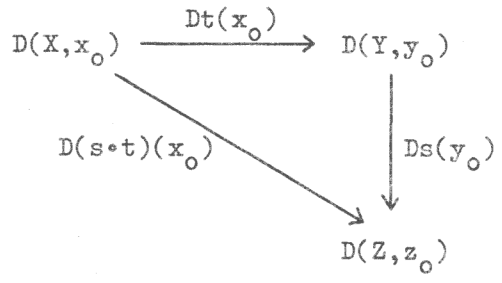
$$t: X \to Y$$

and $$s: Y \to Z$$

be differentiable mappings. Let x_o be a point in X, and define

$$y_o := t(x_o)$$
$$z_o := s(y_o) .$$

Then, the diagram



commutes, i.e.

$$D(s\circ t)(x_o) = Ds(y_o)\circ Dt(x_o) .$$

Proof: For $h \in \mathcal{E}^\infty(Z,z_o)$ and $v \in D(X,x_o)$, we have

$$(Ds(y_o)(Dt(x_o)v))h = (Dt(x_o)v)(h\circ s) = v(h\circ s\circ t)$$

$$= (D(s\circ t)(x_o)v)h .$$

Regular mappings. A differentiable mapping is said to be injectively regular at the point x_o, if the differential

$$Dt(x_o): D(X,x_o) \to D(Y,t(x_o))$$

is injective.

Correspondingly, t is said to be surjectively regular at x_o if the differential is surjective, and bijectively regular at x_o if the differential is bijective.

For short, we may say that a mapping is regular at a point, since the relevant type of regularity is determined by the dimensions of X and Y. Regularity simply means, that the differential is of maximal rank.

A mapping $t:X \to Y$ is said to be regular if it is differentiable and regular at all points of X.

Existence of local inverses.

12.3 Theorem. Let X and Y be manifolds of dimensions n and m, respectively. Let $t: X \to Y$ be regular at x_o, and put $y_o := t(x_o)$. Then

(1) For $n \geq m$ (i.e. for t surjectively regular at x_o) there exists an open neighbourhood V of y_o

and a regular mapping $s: V \to X$ such that

$$t \circ s = 1_V .$$

(2) For $n \leq m$ (i.e. for t injectively regular at x_o) there exist open neighbourhoods U of x_o and V of y_o and a regular mapping $s: V \to U$ such that

$$s \circ t = 1_U .$$

(3) For $n = m$ (i.e. for t bijectively regular at x_o) there exist open neighbourhoods U of x_o and V of y_o such that t maps U diffeomorphic on V.

Proof: In view of the local character of the theorem, we may restrict our attention to the case where X and Y are open submanifolds of $\mathbb{R}^n$ and $\mathbb{R}^m$, and $x_o = 0$ and $y_o = 0$. The general version of the theorem follows by a suitable local charting.

Under these special assumptions, the statement (3) is a wellknown result from multivariate analysis (the inverse mapping theorem). The propositions (1) and (2) are proved by means of (3) :

Proof of (1): Let L be an m-dimensional subspace of $\mathbb{R}^n$, complementary to the $(n-m)$-dimensional kernel of $Dt(0)$. Let

$$r_o: \mathbb{R}^m \to \mathbb{R}^n$$

be some linear parametrization of this subspace L, and let

$$r: W_o \to X$$

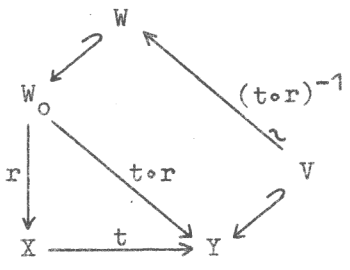
be the restriction of r_o to $W_o := r_o^{-1}(X)$. The composed mapping

$$t \circ r: W_o \to Y$$

is then regular at the point 0. According to (3), $t \circ r$ maps some neighbourhood W of 0 diffeomorphic on an open set $V \subseteq Y$. Let $(t \circ r)^{-1} : V \to W$ denote the inverse of that diffeomorphism. Then, for

$$s := r \circ (t \circ r)^{-1}$$

we have $$t \circ s = 1_V .$$



Proof of (2): Choose a linear mapping

$$r_o : \mathbb{R}^m \to \mathbb{R}^n$$

such that

$$r_o(Dt(0)(\mathbb{R}^n)) = \mathbb{R}^n$$

(i.e. the n-dimensional image of $Dt(0)$ should be mapped bijectively onto $\mathbb{R}^n$ by r_o). Then, the mapping

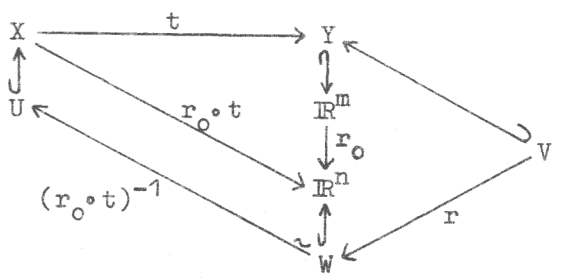
$$r_o \circ t : X \to \mathbb{R}^n$$

is regular at the point 0; according to (3), it maps a neighbourhood U of 0 in X diffeomorphic on a neighbourhood W of 0 in $\mathbb{R}^n$. Let

$$(r_o \circ t)^{-1} : W \to U$$

denote the inverse diffeomorphism. Put $V := r_o^{-1}(W) \cap Y$ and let $r: V \to W$ denote the restriction of r_o. Then, for $s := (r_o \circ t)^{-1} \circ r$, we have

$$s \circ t = 1_U .$$



Local existence of "supplementary mapping".

12.4 Theorem. Let X and Y be manifolds of dimensions n and m, and let $t: X \to Y$ be surjectively regular at x_o (thus $n \geq m$). There exists an open neighbourhood U of x_o and a surjectively regular transformation

$$s: U \to \mathbb{R}^{n-m}$$

such that

$$(t,s): U \to Y \times \mathbb{R}^{n-m}$$

maps U diffeomorphic on an open subset of $Y \times \mathbb{R}^{n-m}$.

Proof: As in the proof of the previous theorem, we need only consider the case where X and Y are open submanifolds of $\mathbb{R}^n$ and $\mathbb{R}^m$, and $x_o = 0$ and $t(x_o) = 0$. But in that case, we can obviously for s take the restriction to X of a linear mapping $s_o: \mathbb{R}^n \to \mathbb{R}^{n-m}$, with the property that

$$\begin{bmatrix} Dt(0) \\ s_o \end{bmatrix} : \mathbb{R}^n \to \mathbb{R}^m \times \mathbb{R}^{n-m}$$

is bijective (i.e. s_o maps the $n-m$-dimensional kernel of $Dt(0)$ surjectively onto $\mathbb{R}^{n-m}$).

Submanifold. Let X and Y_o be manifolds of dimensions n and m. By a parametrization, or an imbedding

$$t: Y_o \to X$$

we mean an injective and injectively regular mapping with the property that Y_o is mapped homeomorphic onto its image $Y := t(Y_o)$.

Such a mapping induces in an obvious manner a differentiable structure on the subset Y of X. Equipped with this structure, Y is called an m-dimensional submanifold of X, or an m-dimensional surface in X (for $m = 1$: a curve).

Thus a submanifold is the image of some imbedding. This terminology indicates that the differentiable structure of a submanifold is independent of its parametrization, as the following theorem shows:

12.5 Theorem. Let X, Y_1 and Y_2 be differentiable manifolds of dimensions n, m_1 and m_2, and let

$$t_1: Y_1 \to X$$
$$\text{and} \quad t_2: Y_2 \to X$$

be imbeddings. Suppose that

$$t_1(Y_1) = t_2(Y_2) .$$

Then, the bijective mapping

$$t_2^{-1} \circ t_1 : Y_1 \to Y_2$$

is a diffeomorphism (and, in particular, $m_1 = m_2$) .

Proof: It suffices to prove, that for any point $y_1 \in Y_1$ the mapping $t_2^{-1} \circ t_1$ is differentiable in a neighbourhood of y_1 . The same argument then applies to the inverse mapping

$$t_1^{-1} \circ t_2 : Y_2 \to Y_1 .$$

Put $x := t_1(y_1)$, and let $s: U \to V_2$ be a local left inverse (theorem 12.3, (2) , page 102-103) of t_2 . For $y_1' \in V_1$, we have

$$t_2^{-1} \circ t_1(y_1') = s \circ t_1(y_1') ,$$

i.e. $t_2^{-1} \circ t_1$ coincides locally with the differentiable mapping $s \circ t_1$. This proves the theorem.

The theorem shows, that the structure of a submanifold Y of X is determined by the structure on X . Hence, it has a meaning to say that a subset Y of X is a submanifold of a certain dimension. The dimension is determined by the subset, except in the trivial case $Y = \emptyset$ (the empty set

is a manifold of any dimension).

It is not hard to prove, that the submanifold-property is of local character, in the sense that $Y \subseteq X$ is an m-dimensional submanifold if and only if any point of Y has an open X-neighbourhood U such that $U \cap Y$ is an m-dimensional submanifold of U. The differentiable structure on a set Y with this "local manifold-property" is simply defined by the union of the atlases on the manifolds $Y \cap U$. It follows from theorem 12.5 that this "piecing together" is possible (the constructed atlas becomes consistent).

Notice that a submanifold of a submanifold of X can, in the obvious manner, be regarded as a submanifold of X.

Notice also, that an n-dimensional submanifold of X ($n = \dim X$) is simply an open submanifold, as defined on page 91.

Level surfaces.

12.6 Theorem. Let X and Z be differentiable manifolds of dimensions n and $n-m$, and let

$$s: X \to Z$$

be surjectively regular. Then, for $z_0 \in Z$, the level surface

$$Y := s^{-1}(z_o)$$

is an m-dimensional submanifold of X .

Proof: Let x_o be a point in Y , U a neighbourhood of x_o (relative to X), such that there exists a " supplementary transformation " (theorem 12.4, page 106) $t:U \to \mathbb{R}^m$, i.e. a surjectively regular mapping t such that

$$(s,t):U \to Z \times \mathbb{R}^m$$

maps U diffeomorphic on an open subset U_1 of $Z \times \mathbb{R}^m$.

The mapping (s,t) takes $V = U \cap Y$ into the set

$$V_1 = U_1 \cap (\{z_o\} \times \mathbb{R}^m) .$$

This set is easily seen to be an m-dimensional submanifold of U_1 . Since the concept of a submanifold is obviously invariant under diffeomorphisms, we conclude that V is a submanifold of U , and this proves the theorem (cfr. the above remarks about the local character of the submanifold-property).

Remark: Theorem 12.4 and the arguments above require, of course, that the obvious definition of a product manifold (here $Z \times \mathbb{R}^m$) is applied (definition by " product charts ").

The geometric interpretation of $D(Y,x_o)$. Let

$$j: Y \to X$$

denote the imbedding of a submanifold. For a point $x_o \in Y$, the differential

$$Dj(x_o) : D(Y,x_o) \to D(X,x_o)$$

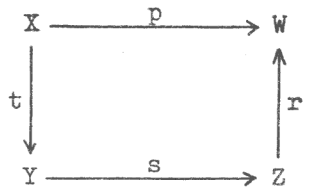
is an injective linear mapping. We shall regard the differential as an imbedding of a subspace, i.e. we identify $D(Y,x_o)$ with a subspace of $D(X,x_o)$.

In the case $X = \mathbb{R}^n$ (where $D(X,x_o) = \mathbb{R}^n$), this identification has a very simple geometric interpretation: The affine subspace $x_o + D(Y,x_o)$ is simply the tangent space for the surface Y at x_o. For this reason, vectors are sometimes called tangent vectors (also in the abstract case), and the spaces $D(X,x_o)$ are called tangent spaces.

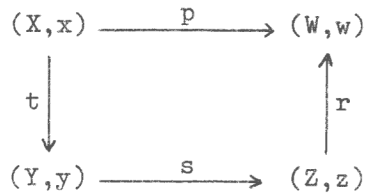
Notice that the tangent space for a level surface $Y = s^{-1}(z_o)$ is simply the kernel of $Ds(x_o)$, under the conditions of theorem 12.6: If j denotes the imbedding of Y, $s \circ j$ is constant, and so $D(s \circ j)(x_o) = Ds(x_o) \cdot Dj(x_o) = 0$; this proves that $D(Y,x_o) = \mathrm{Im}(Dj(x_o)) \subseteq \mathrm{Ker}\, Ds(x_o)$, and a consideration of the dimensions shows that this inclusion must be an identity.

The functor D. Suppose that some commutative diagram of

differentiable manifolds and differentiable mappings is given; for example



Furthermore, let there be given a point in each manifold such that the mappings take these points into each other:



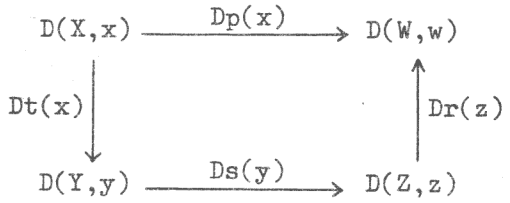
$y = t(x)$

$z = s(y)$

$w = r(z) = p(x).$

In terms of category theory, this is a diagram in the category of <u>differentiable manifolds with base point</u> (and "base point preserving" differentiable mappings as homomorphisms).

Now, replace the manifolds in the diagram by their tangent spaces (at the selected points) and the homomorphisms by their differentials (at the selected points). This yields a diagram



in the category of vectorspaces (with linear mappings as homomorphisms). It follows from theorem 12.2 (page 100-101) that this diagram commutes.

Such an operation D , taking objects and homomorphisms in a category into objects and homomorphisms of another category, in such a way that commutativity is preserved, is called a functor. Thus, D is a functor from the category of manifolds with basepoint into the category of vectorspaces.

In the following sections, this technique of transforming diagrams by D will frequently be applied. In particular, we are interested in the cases where the transformed diagrams have certain exactness properties. For this purpose it will be useful to notice that a diagram of the form

$$(\{y\},y) \hookrightarrow (Y,y) \xrightarrow{t} (X,x) \xrightarrow{s} (Z,z) \longrightarrow (\{z\},z)$$

where t is injectively regular, s surjectively regular, and t(Y) is (at least locally) a level surface for s , transforms into an exact sequence

$$0 \longrightarrow D(Y,y) \xrightarrow{Dt(y)} D(X,x) \xrightarrow{Ds(x)} D(Z,z) \longrightarrow 0 .$$

In case Y is a submanifold of X , we shall write

$$Y \hookrightarrow X$$

for the imbedding (as it was done for the 0-dimensional submanifold $\{y\}$ of Y above). The identification of $D(Y,x)$ with a subspace of $D(X,x)$ enables us to write

$$D(Y,x) \hookrightarrow D(X,x)$$

in the transformed diagram. Loosely speaking, D "preserves inclusions" $\hookrightarrow$.

For an open submanifold Y of X, we have

$$D(Y,x) = D(X,x) .$$

In diagrams, we write

$$D(Y,x) = D(X,x).$$

Vector fields. By a vector field on X we mean a family $(v_x | x \in X)$ of vectors $v_x \in D(X,x)$, such that the following "differentiability condition" is satisfied:

$$[v_x f]_x \in \mathcal{C}^\infty(X) \quad \text{for } f \in \mathcal{C}^\infty(X) .$$

In case X is an open subset of $\mathbb{R}^n$, any vector field (v_x) has a unique coordinate representation

$$v_x = v_1(x)\frac{\partial}{\partial x_1} + \ldots + v_n(x)\frac{\partial}{\partial x_n}$$

(where the partial derivatives are to be evaluted at the varying point x). The existence and uniqueness of such a representation follows immediately from the fact that $\frac{\partial}{\partial x_1}, \ldots, \frac{\partial}{\partial x_n}$ constitutes a base for each tangent space (see page 98-99). It is not hard to prove, that the above "differentiability condition" is satisfied if and only if the coordinate functions $v_1, \ldots, v_n$ belong to $\mathcal{C}^\infty(X)$.

Local existence of a base of vectorfields.

12.7 Theorem. Any point $x_o \in X$ has a neighbourhood U equipped with n vectorfields $(b_x^{(1)}), \ldots, (b_x^{(n)})$ $(x \in U)$, such that any vectorfield (v_x) on U has a unique representation

$$v_x = v_1(x)b_x^{(1)} + \ldots + v_n(x)b_x^{(n)},$$

$$v_1, \ldots, v_n \in \mathcal{C}^\infty(U).$$

Proof: Just transfer the base $\frac{\partial}{\partial x_1}, \ldots, \frac{\partial}{\partial x_n}$ from some chart.

Local extension of a vectorfield on a submanifold.

12.8 Theorem: Let Y be an m-dimensional submanifold of an n-dimensional manifold X, and let $(v_y | y \in Y)$ be a vectorfield on Y. Then, for any point $y_o \in Y$, there exists an open neighbourhood U (relative to X) and a vectorfield $(u_x | x \in U)$ on U such that $(v_y | y \in U \cap Y)$ is the restriction of (u_x) to $U \cap Y$ (i.e. $u_y = v_y$ for $y \in U \cap Y$, under the identification of $D(Y,y)$ with a subspace of $D(X,y)$).

Proof: Since the statement is of local character, we may assume that X is an open submanifold of $\mathbb{R}^n$. The vectorfield (v_y) then has a unique representation

$$v_y = v_1(y)\frac{\partial}{\partial x_1} + \ldots + v_n(y)\frac{\partial}{\partial x_n} ,$$

determined by

$$v_i(y) = v_y(h_i \circ j)$$

where j denotes the imbedding of Y into X, and $h_i = [x_i]_x$.

Locally (in a neighbourhood of y_o) we can extend the functions $v_1, \ldots, v_n$ on Y to differentiable functions $u_1, \ldots, u_n$

on X (put $u_i := v_i \circ s$, where s denotes a local left inverse of the imbedding j). Then

$$u_x := u_1(x)\frac{\partial}{\partial x_1} + \ldots + u_n(x)\frac{\partial}{\partial x_n}$$

defines a local extension of the vectorfield (v_y) on Y.

13. RIEMANN MANIFOLDS.

Loosely speaking, a differentiable manifold is a space with a local structure as a vectorspace. And a Riemann manifold is a space with a local structure as a Euclidean vectorspace. Typical examples of Riemann manifolds are Euclidean spaces (of course), and submanifolds of Euclidean spaces.

Formally, the local Euclidean structure is introduced as a Euclidean structure of the tangent spaces:

Definition: Let X be an n-dimensional differentiable manifold. A Riemann structure on X is a mapping

$$x \to (\ |\ )_x$$

taking $x \in X$ into an inner product $(\ |\ )_x$ on $D(X,x)$, in such a way that the following "differentiability condition" is satisfied:

For any two vectorfields $(v_x | x \in U)$ and $(v'_x | x \in U)$ on an open submanifold U, the function

$$[(v_x | v'_x)]_x$$

belongs to $\mathcal{C}^\infty(U)$.

A Riemann manifold is a differentiable manifold, equipped with a Riemann structure.

Roughly , the differentiability condition states that the inner product $(\mid)_x$ depends on x in a differentiable manner. In case we are given a local base $b_x^{(1)}, \ldots, b_x^{(n)}$ for the vectorfields (theorem 12.7 , page 115) the inner product $(\mid)_x$ can, in the usual manner, be described by a symmetric, positive definit matrix. Obviously then, the differentiability condition means that the elements

$$\gamma_{ij} = (b_x^{(i)} \mid b_x^{(j)})_x$$

of the matrix should be differentiable functions of x.

Remark: Actually, it suffices to assume that the differentiability condition is satisfied for vectorfields on X , since a vectorfield on an open submanifold coincides locally with a vectorfield on X . This follows from results concerning global extension of differentiable functions (see e.g. Helgason (1962), page 2-3.

Submanifolds. Let Y be an m-dimensional submanifold of an n-dimensional Riemann manifold X. The tangent space $D(Y,y)$ of Y is a subspace of the corresponding tangentspace $D(X,y)$ of X, and thus equipped with a structure as a Euclidean vectorspace (by the definition of Euclidean subspace, page 65).

This obviously suggests a Riemann structure on Y. The following argument shows that the differentiability condition is satisfied:

Let $(v_y | y \in V)$ and $(v'_y | y \in V)$ be vectorfields on an open submanifold V of Y. By theorem 12.8 (page 116) these vectorfields can be extended locally to vectorfields on an open subset U of X. Hence, the function $[(v_y | v'_y)_y]_y$ is locally the restriction of a differentiable function, and so it is differentiable.

Submanifolds of Riemann manifolds, in particular submanifolds of $\mathbb{R}^n$, are always regarded as Riemann manifolds, equipped with this Riemann structure.

The functor D. In case of Riemann manifolds, the functor D should be considered a functor into the category of Euclidean vectorspaces with linear mappings as homomorphisms.

It follows from the above definition of a sub-Riemann manifold, that imbeddings

$$Y \xhookrightarrow{j} X$$

are taken into isometries

$$D(Y,y) \xhookrightarrow{Dj(y)} D(X,y) \quad ,$$

i.e. D preserves arrows $\hookrightarrow$ as applied here and in section 11.

Orthonormal bases of vectorfields. A local base $(b_x^{(1)}, \ldots, b_x^{(n)})$ $(x \in U)$ is said to be orthonormal, if it is so pointwise, i.e. if

$$(b_x^{(i)} | b_x^{(j)})_x = \begin{cases} 1 & \text{for } i=j \\ 0 & \text{for } i \neq j \end{cases}.$$

13.1 Theorem. Let X be a Riemann manifold of dimension n. Then any point $x_o \in X$ has a neighbourhood U with an orthonormal base of vectorfields.

Proof: According to theorem 12.7, a local base exists. By the Gram-Schmidt orthonormalization procedure, this base is converted into an orthonormal base, at least pointwise. But obviously, the differentiability is not destroyed by the orthonormalization, since the coordinates of the new basis vectors with respect to the old can be expressed as nice, explicit functions of some inner products of the old basis vectors.

The geometric structure. On a Riemann manifold, many of the wellknown concepts from Euclidean geometry can be given a meaning. For example, the length of a curve can be defined, and it has a meaning to say that two submanifolds intersect orthogonally. The concept of curvature can be defined, and this gives rise to a natural definition of a line (called a geodesic) as a curve of curvature 0.

We shall restrict our attention to one aspect of the geometry on Riemann manifolds: The existence of a canonical measure, here called the geometric measure, similar to Lebesgue measure on $\mathbb{R}^n$.

14. THE GEOMETRIC MEASURE ON A RIEMANN MANIFOLD.

Let X be a Riemann manifold of dimension n and let

$$\varphi_U : U \to U'$$

be a chart of an open subset U of X. We shall construct the geometric measure on X as a piecing together of geometric measures on charted sets, defining the geometric measure on the charted set U by its density with respect to the Lebesgue measure transferred from the chart U'. This density should of course be adapted to the special case $X = \mathbb{R}^n$ such that the geometric measure on $\mathbb{R}^n$ coincides with the usual Lebesgue measure. We define

$$g: U \to \mathbb{R}$$

by

$$g(x) := \frac{1}{|D\varphi_U(x)|} ,$$

and put

$$\lambda_U := g\cdot(\varphi_U^{-1}(\lambda_U^n,))$$

where $\lambda_U^n,$ denotes the(restriction of) Lebesgue measure on the chart. It follows from the integral transformation theorem that this definition yields the restriction of Lebesgue measure to U in the special case $X = \mathbb{R}^n$. We have $|D\varphi_U(x)| \neq 0$

since

$$D\varphi_U(x): D(X,x) \to \mathbb{R}^n$$

is bijective (φ_U is a diffeomorphism). Furthermore g is continuous (and even differentiable) since the determinant can be expressed as a sum of products of elements of the matrix for $D\varphi_U(x)$ with respect to local orthonormal bases on U and U' (theorem 13.1, page 121).

In order to prove the existence of a measure λ_X on X such that λ_U is the restriction of λ_X to U for any charted set U , we must prove (according to theorem A 10 , page 340) that the measures λ_U coincide on overlapping sets. Thus for λ_{U_1} and λ_{U_2} defined as above by the charts

$$\varphi_{U_1} : U_1 \to U_1'$$

$$\varphi_{U_2} : U_2 \to U_2'$$

we must prove that

$$\lambda_{U_1}|_{U_1 \cap U_2} = \lambda_{U_2}|_{U_1 \cap U_2} .$$

Now let

$$\varphi_1 : U_1 \cap U_2 \to W_1' := \varphi_{U_1}(U_1 \cap U_2)$$

and $$\varphi_2 : U_1 \cap U_2 \to W_2' := \varphi_{U_2}(U_1 \cap U_2)$$

denote the restrictions of φ_{U_1} and φ_{U_2}, respectively. Then

$$\lambda_{U_1}|_{U_1 \cap U_2} = g_1 \cdot (\varphi_1^{-1}(\lambda^n_{W_1'}))$$

and $$\lambda_{U_2}|_{U_1 \cap U_2} = g_2 \cdot (\varphi_2^{-1}(\lambda^n_{W_2'}))$$

where $\lambda^n_{W_1'}$ and $\lambda^n_{W_2'}$ denote the restrictions of Lebesgue measure, and the two densities g_1 and g_2 are defined by

$$g_1(x) = \frac{1}{|D\varphi_1(x)|}$$

$$g_2(x) = \frac{1}{|D\varphi_2(x)|} \quad .$$

According to the integral transformation theorem we have ($\varphi_2 \circ \varphi_1^{-1}$ being a diffeomorphism $W_1' \to W_2'$)

$$(\varphi_2 \circ \varphi_1^{-1})(\lambda^n_{W_1'}) = d \cdot \lambda^n_{W_2'}$$

where

$$d(w_2') = \frac{1}{|D(\varphi_2 \circ \varphi_1^{-1})(\varphi_1(\varphi_2^{-1}(w_2')))|}$$

$$= \frac{1}{|D\varphi_2(\varphi_2^{-1}(w_2'))|} \cdot \frac{1}{|D\varphi_1^{-1}(\varphi_1(\varphi_2^{-1}(w_2')))|}$$

$$= g_2(\varphi_2^{-1}(w_2')) \cdot |D\,\varphi_1(\varphi_2^{-1}(w_2'))| = \frac{g_2(\varphi_2^{-1}(w_2'))}{g_1(\varphi_2^{-1}(w_2'))} ,$$

i.e.

$$(\varphi_2 \circ \varphi_1^{-1})(\lambda^n_{W_1'}) = ((\frac{g_2}{g_1}) \circ \varphi_2^{-1}) \cdot \lambda^n_{W_2'} .$$

This is an identity between measures on W_2'. Transforming both sides by φ_2^{-1} we get

$$\varphi_1^{-1}(\lambda^n_{W_1'}) = \frac{g_2}{g_1} \cdot (\varphi_2^{-1}(\lambda^n_{W_2'}))$$

and multiplying by g_1 on both sides yields

$$g_1 \cdot (\varphi_1^{-1}(\lambda^n_{W_1'})) = g_2 \cdot (\varphi_2^{-1}(\lambda^n_{W_2'}))$$

or

$$\lambda_{U_1}|_{U_1 \cap U_2} = \lambda_{U_2}|_{U_1 \cap U_2} .$$

Hence, the family $(\lambda_U |$ U charted subset of X) is a consistent family of restrictions (theorem A 10 , page 340), and so it determines a unique measure λ_X on X such that

$$\lambda_X|_U = \lambda_U \quad \text{for all } U .$$

This measure is called the <u>geometric measure</u> on X .

Examples: In case of an open subset X of $\mathbb{R}^n$ the geometric measure λ_X is obviously the restriction of Lebesgue measure.

The geometric measure on a two-dimensional surface in $\mathbb{R}^3$ can be interpreted as the usual area measure.

The geometric measure on an m-dimensional subspace of a Euclidean space is Lebesgue measure, normalized according to the Euclidean structure (for example such that the unit ball obtains the volume unit balls should have, namely that of the unit ball in $\mathbb{R}^m$). The geometric measure on an affine subspace of a Euclidean space has a similar interpretation.

The geometric measure on the (n-1)-dimensional unit sphere

$$S_{n-1} = \{x \in \mathbb{R}^n | \ \|x\| = 1\}$$

is the uniform distribution (i.e. the rotational invariant distribution) with a suitable normalizing factor (which we shall compute in section 32).

Similarly, the geometric measure on a Lie group (or a homogeneous space) with an invariant Riemann structure is, of course, the invariant measure (the Haar measure).

15. DECOMPOSITION OF THE GEOMETRIC MEASURE.

Let X and Y be Riemann manifolds of dimensions n and m $(n \geq m)$ and let

$$t: X \to Y$$

be a surjective and surjectively regular transformation. By

$$X_y := t^{-1}(y)$$

we denote the level surfaces for t. The geometric measures on the manifolds are denoted λ_X, λ_Y etc.

The imbedding

$$j_y: X_y \to X$$

is proper (or continuous at infinity), i.e. the inverse images of compact sets are compact. Thus, for $k \in \mathcal{K}(X)$ the function $k \circ j_y$ belongs to $\mathcal{K}(X_y)$, and so the transformed measure

$$j_y(\lambda_{X_y}) \in \mathcal{M}(X)$$

is welldefined (see the appendix, page 350). In the following we shall write λ_{X_y} instead of $j_y(\lambda_{X_y})$ most of the time, without serious danger of confusion.

According to theorem 11.2 (page 75) we have

$$|Dt(x)|^{\circ} > 0 .$$

Thus we can define a function

$$F: X \to \mathbb{R}$$

by

$$F(x) := \frac{1}{|Dt(x)|^{\circ}} .$$

This function is <u>differentiable</u>. In order to prove that, we notice that a local choice of orthonormal bases on X and Y (theorem 13.1, page 121) yields a formula

$$|Dt(x)|^{\circ} = \sqrt{|\det(M(x)M(x)')|}$$

where $M(x)$ denotes the matrix of $Dt(x)$ with respect to the selected bases. The elements of $M(x)$ are differentiable functions, and so F is obviously differentiable (the determinant is $\neq 0$). In particular F is continuous. We define a family $(\lambda_y \mid y \in Y)$ of measures on X by

$$\lambda_y := F \cdot \lambda_{X_y}$$

i.e. λ_y is defined to be the measure on X, given by the density $F = 1/|Dt|^{\circ}$ with respect to the (imbedded) geometric measure on the level surface $X_y = t^{-1}(y)$.

15.1 Theorem.

$$(\lambda_Y , (\lambda_y | y \in Y))$$

is a decomposition of λ_X with respect to t (cfr. page 36).

Proof: Obviously

$$\text{supp } \lambda_y = \text{supp } \lambda_{X_y} = X_y = t^{-1}(y) .$$

It remains to prove that the mapping

$$y \to \lambda_y$$

$$Y \to \mathcal{M}(X)$$

is continuous and that λ_X is the mixture of the measures λ_y with respect to λ_Y. In order to prove this it suffices to prove that for any $\mathcal{K}(X)$-function k the function

$$h(y) := \lambda_y k = \lambda_{X_y}(F \cdot k)$$

is a $\mathcal{K}(Y)$-function with

$$\lambda_Y h = \lambda_X k .$$

Indeed it suffices to prove this statement for $\mathcal{K}(X)$-functions with support contained in a small neighbourhood of a given point x_o. The more general statement then follows immediately by decomposition of the function k with respect to a covering by such neighbourhoods (theorem A 2, page 334).

Hence, let $x_o \in X$ be given and put $y_o := t(x_o)$. Let

$$\varphi_{V_o} : V_o \to V_o' \subseteq \mathbb{R}^m$$

be a chart of a neighbourhood of y_o and let U_o be an open neighbourhood of x_o such that $t(U_o) \subseteq V_o$. We can choose U_o such that there exists a "supplementary transformation" (theorem 12.4, page 106)

$$\psi : U_o \to \mathbb{R}^{n-m}$$

such that the mapping

$$(\varphi_{V_o} \circ t, \psi) : U_o \to \mathbb{R}^m \times \mathbb{R}^{n-m}$$

becomes diffeomorphic, i.e. it becomes a <u>chart</u>

$$\varphi_{U_o} := (\varphi_{V_o} \circ t, \psi) : U_o \to U_o' \subseteq \mathbb{R}^m \times \mathbb{R}^{n-m}.$$

Now let $V' \subseteq \mathbb{R}^m$ and $W' \subseteq \mathbb{R}^{n-m}$ be open neighbourhoods of $\varphi_{V_o}(y_o)$ and $\psi(x_o)$ such that the product

$$U' := V' \times W'$$

is contained in U'_o . Put

$$V := \varphi_{V_o}^{-1}(V')$$

and $$U := \varphi_{U_o}^{-1}(U')$$

and let

$$\varphi_V : V \to V'$$

and $$\varphi_U : U \to U'$$

denote the restrictions of φ_{V_o} and φ_{U_o} , respectively. For $y \in V$, we put

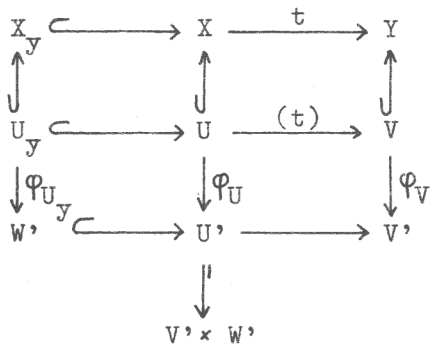
$$U_y := X_y \cap U .$$

Then U_y is an open subset of the manifold X_y , and the restriction

$$\varphi_{U_y} : U_y \to W'$$

of φ is obviously a chart.

We now have this commutative diagram of Riemann manifolds and differentiable mappings:



The arrows of the bottom row represent the imbedding

$$w' \to (\varphi_V(y), w')$$

of the fibre $W' \approx \{\varphi_V(y)\} \times W'$ and the projection

$$(v', w') \to v'$$

on the first component in the product $U' = V' \times W'$.

For a point $x \in U_y$, we apply the functor D and get the following commutative diagram of Euclidean spaces and linear mappings:

$$\begin{array}{ccccc}
D(X_y,x) & \hookrightarrow & D(X,x) & \xrightarrow{Dt(x)} & D(Y,y) \\
\uparrow & & \uparrow & & \uparrow \\
D(U_y,x) & \hookrightarrow & D(U,x) & \xrightarrow{Dt(x)} & D(V,y) \\
\downarrow{\scriptstyle D\varphi_{U_y}(x)} & & \downarrow{\scriptstyle D\varphi_U(x)} & & \downarrow{\scriptstyle D\varphi_V(y)} \\
\mathbb{R}^{n-m} & \hookrightarrow & \mathbb{R}^n & \longmapsto & \mathbb{R}^m \\
 & & \downarrow & & \\
 & & \mathbb{R}^m \times \mathbb{R}^{n-m} & &
\end{array}$$

Here the arrows of the bottom row simply stand for the imbedding of second component and the coimbedding onto first component in the product space $\mathbb{R}^n = \mathbb{R}^m \times \mathbb{R}^{n-m}$.

The lower half part of this diagram (ignoring the identity $\mathbb{R}^n \longrightarrow \mathbb{R}^m \times \mathbb{R}^{n-m}$) satisfies the conditions of theorem 11.4 (page 80). Hence

$$|D\varphi_U(x)| = |D\varphi_V(y)| \cdot |D\varphi_{U_y}(x)| \cdot |Dt(x)|^\circ .$$

For

$$u' := (v',w') := \varphi_U(x) = (\varphi_V(y), \varphi_{U_y}(x))$$

we have then

$$|D\varphi_U(\varphi_U^{-1}(u'))| = |D\varphi_V(\varphi_V^{-1}(v'))| \cdot |D\varphi_{U_y}(\varphi_{U_y}^{-1}(w'))| \cdot |Dt(\varphi_U^{-1}(u'))|^\circ .$$

Taking inverses on both sides and expressing $|Dt|^\circ$ by the function F we get

$$(*) \quad |D(\varphi_U^{-1})(u')| = |D(\varphi_V^{-1})(v')| \cdot |D(\varphi_{U_y}^{-1})(w')| \cdot F(\varphi_U^{-1}(u')) .$$

Now let k be a $\mathcal{K}(X)$-function with support contained in U. According to the definition of the geometric measure we have (with our customary carelessness about the specification of domains, restrictions etc.)

$$\lambda_X k = \lambda_U k = \lambda_U^n, (|D(\varphi_U^{-1})| \cdot (k \circ \varphi_U^{-1}))$$

where λ_U^n, denotes the restriction of Lebesgue measure to U' .

For $y \in V$ the function $F \cdot k$ is, when restricted to X_y, a $\mathcal{K}(X_y)$-function with support contained in U_y. Thus

$$h(y) = \lambda_y k = \lambda_{X_y}(F \cdot k)$$

$$= \lambda_W^{n-m}(|D(\varphi_{U_y}^{-1})| \cdot (F \cdot k) \circ \varphi_{U_y}^{-1}) .$$

By the equation (*) we get

$$h(y) = \lambda_W^{n-m}\left[\frac{|D(\varphi_U^{-1})(\varphi_V(y), w')|}{|D(\varphi_V^{-1})(\varphi_V(y))|} \cdot (k \circ \varphi_U^{-1})(\varphi_V(y), w') \right]_{w'},$$

This equation shows that h is a continuous function (theorem A 5, page 337). Obviously h has compact support $\operatorname{supp} h \subseteq V$. Thus $\lambda_Y h$ is welldefined. By the decomposition

$$\lambda_U^n, = \lambda_V^m, \otimes \lambda_W^{n-m}$$

of Lebesgue measure as a product of Lebesgue measures of lower dimensions, we get

$$\lambda_Y h = \lambda_V^m, (|D(\varphi_V^{-1})| \cdot (h \circ \varphi_V^{-1}))$$

$$= \lambda_V^m, \left[|D\varphi_V^{-1}(v')| \lambda_W^{n-m}\left[\frac{|D(\varphi_U^{-1})(v', w')|}{|D(\varphi_V^{-1})(v')|} \cdot (k \circ \varphi_U^{-1})(v', w') \right]_{w'} \right]_{v'},$$

$$= \lambda_V^m, \left[\lambda_W^{n-m} \left[|D(\varphi_U^{-1})(v',w')| \cdot (k \circ \varphi_U^{-1})(v',w') \right]_{w'} \right]_{v'},$$

$$= \lambda_U^n, [\, |D(\varphi_U^{-1})(u')| \cdot (k \circ \varphi_U^{-1})(u')]_{u'}, \qquad = \lambda_X k \, .$$

This proves the theorem.

In case the transformed measure $t(\lambda_X)$ is defined, we can easily compute its density G with respect to λ_Y, by theorem 15.1:

15.2 Theorem. Suppose that $h \circ t$ is λ_X-integrable for all $h \in \mathcal{K}(Y)$ (cfr. the appendix, page 350). Then

$$t(\lambda_X) = G \cdot \lambda_Y$$

where

$$G(y) = \| \lambda_y \| = \lambda_{X_y}(F) \, .$$

Proof: λ_X is the mixture of the measures $\lambda_y = F \cdot \lambda_{X_y}$ with respect to λ_Y. Thus for $h \in \mathcal{K}(Y)$ (see theorem A 22, page 352) we have

$$t(\lambda_X)h = \lambda_X(h \circ t) = \lambda_Y[\lambda_y(h \circ t)]_y = \lambda_Y[h(y) \, \| \lambda_y \|]_y = \lambda_Y(G \cdot h).$$

Notice that G need not be continuous (except when t is proper, and the level surfaces thus compact). But G is locally λ_Y-integrable, according to theorem A 22.

In case X and Y are manifolds of the same dimension the "level surfaces" are discrete subsets of X. The geometric measure on a 0-dimensional Riemann manifold is obviously the counting measure (the Riemann structure is unique, since the tangent spaces are of dimension 0). The measures λ_y thus have the form

$$\lambda_y = \sum_{x \in X_y} F(x) \cdot \varepsilon_x$$

where

$$F(x) = \frac{1}{|Dt(x)|^0} = \frac{1}{|Dt(x)|} .$$

The density G in theorem 15.2 is

$$G(y) = \sum_{x \in X_y} F(x) .$$

In particular we have, in the case where t is bijective, the integral transformation theorem for Riemann manifolds:

15.3 Theorem. Let $t: X \to Y$ be a diffeomorphism between two Riemann manifolds. Then

$$t(\lambda_X) = G \cdot \lambda_Y$$

for

$$G(y) = F(t^{-1}(y)) = |D(t^{-1})(y)| .$$

The validity of this theorem was more or less subsumed in the definition of the geometric measure, and it can easily be proved directly from the definition.

16. CONDITIONING IN A RIEMANN MANIFOLD.

The decomposition of the geometric measure (section 15) and the results on conditioning in a decomposed measure space (section 7) are immediately combined to a result about conditioning in Riemann manifolds. We shall restrict our attention to the simplest case where the densities f and g are continuous and positive. In section 17 we shall show how to trace more complicated conditioning problems back to this case. For convenience some of the formulae and results from section 15 are repeated here, such that this section summarizes the basic tools for conditioning in the continuous case. More concrete aspects of the theory are treated in section 17, 18, 31 and 32.

Let X and Y be Riemann manifolds, and let

$$t: X \to Y$$

be a surjective and surjectively regular transformation. Let us introduce and summarize some notation:

λ_X, λ_Y etc. denote the geometric measures.

$F: X \to \mathbb{R}$ is defined by $F(x) = \dfrac{1}{|Dt(x)|^\circ} = \dfrac{1}{\sqrt{|Dt(x)Dt(x)^*|}}$.

λ_y is the measure given by the density F with respect to the geometric measure on the level surface

$X_y := t^{-1}(y)$, i.e.

$$\lambda_y = F \cdot \lambda_{X_y} ;$$

λ_y and λ_{X_y} are, most of the time, regarded as measures on X.

f is a continuous and positive probability density with respect to λ_X. We put

$$\mu := f \cdot \lambda_X .$$

The transformed measure

$$\nu := t(\mu)$$

has the density

$$g(y) := \lambda_y f = \lambda_{X_y}(F \cdot f) .$$

It is assumed that g is continuous (obviously, g is positive).

The conditional distribution μ^y of $x \in (X, \mu)$, given $t(x) = y$, is then defined for all $y \in Y$ (theorem 7.1) and given by the density $\frac{1}{g(y)} f$ with respect to λ_y, i.e.

$$\mu^y = \frac{1}{g(y)}(f \cdot \lambda_y) = \frac{1}{g(y)}(F \cdot f \cdot \lambda_{X_y}) .$$

By

$$f^y(x) := \frac{1}{g(y)} F(x) f(x) = \frac{f(x)}{g(y)\,|Dt(x)|^o}$$

we denote the density of μ^y with respect to λ_{X_y}.

17. CONDITIONING IN A EUCLIDEAN SPACE.

Let

$$t_o : (\mathbb{R}^N, \mu_o) \to (\mathbb{R}^M, \nu_o)$$

be a homomorphism, i.e. a μ_o-measurable mapping, possibly undefined on a set of μ_o-measure 0 . We shall discuss the conditioning problem under reasonable regularity assumptions.

In almost all situations of relevance, the conditioning problem can be solved as follows:

Choose an n-dimensional submanifold X of $\mathbb{R}^N$ (open or of dimension $n < N$) such that the following conditions are satisfied:

(1) $Y = t_o(X)$ is an m-dimensional submanifold of $\mathbb{R}^M$.

(2) $\mu_o(X) = 1$ (and so $\nu_o(Y) = 1$).

(3) The restrictions

$$\mu \text{ of } \mu_o \text{ to } X$$
$$\nu \text{ of } \nu_o \text{ to } Y$$
$$\text{and } t:(X, \mu) \to (Y, \nu) \text{ of } t_o$$

satisfy the conditions in section 16.

Obviously then, the conditional distributions deduced in section 16 are also conditional distributions when imbedded into $\mathbb{R}^N$ (the imbedding $j: X \to \mathbb{R}^N$ is continuous, also as a mapping $j: \mathcal{P}(X) \to \mathcal{P}(\mathbb{R}^N)$).

The choice of the submanifold X can, in outline, be carried out as follows:

Suppose that μ_o is given by a density f with respect to a geometric measure on a submanifold X (it is subsumed, then, that μ_o is of "constant dimension"; we return to that point later). In case f is a usual "explicit" function (piecewise analytic, for example) we can in general reduce the domain X by a closed λ_X-null set (the closure of the discontinuity points for the restriction of f to X) such that f becomes <u>continuous</u> on the new domain X. The assumption $f > 0$ is obtained by a similar reduction (namely by removing the closed set (rel. to X) $\{x | f(x)=0\}$). We still have $\mu = f \cdot \lambda_X$, and now f is positive and continuous. The new X is certainly a submanifold of $\mathbb{R}^N$, being an open subset of the original X.

In a similar manner, we can reduce X such that the transformation t (assumed to be "explicit", pieced together of nice transformations of nice sets) becomes surjectively regular as a mapping $t: X \to Y = t(X)$. This requires, however, that Dt is (essentially) of constant rank, a point which we shall return to. A modification of Y may be neccesary in order to make Y a manifold (removal of Y's "intersections with it-

self " etc.). Finally, under the assumption that the density g of ν is a nice "explicit" function, a similar reduction of Y (and, accordingly, of X) makes g a continuous and positive function.

The assumptions about f and g are not restrictive in practice. All counterexamples known to me possess the unmistakable features of being counterexamples and nothing but that.

The assumptions about t are slightly more restrictive; not the pure regularity assumptions, but the assumption about "essential rank homogenity". Transformations, pieced together of transformations of different ranks, can not always be excluded as quite pathological. A transformation like

$$t: \mathbb{R}^2 \to \mathbb{R}^2$$

$$t(x_1,x_2) = (x_1 , x_1 \wedge x_2)$$

has an effectively varying rank (namely, 1 for $x_1 < x_2$ and 2 for $x_1 > x_2$).

We have made a similar assumption about the probability measure μ, namely that μ "is of constant dimension"; but in some cases it may be of interest to consider probability measures of "mixed dimension", like a convex combination of (say) a normal distribution and a one point measure.

Such problems of mixed dimension can be handled by a division of X or Y, followed by an application of one of the results concerning "piecewise conditioning" (section 28). The spaces X and Y should be divided into their components of different dimensions, and the components should be separated from each other (i.e. the topology should be changed) such that X and Y become disjoint unions of manifolds of different dimensions. We shall not go into details about this redefinition of the topology and its consequences, just notice, that if problems of mixed dimension are really of interest (which they seldom are) then one should be careful about the choice of topology anyway. The probabilistic structure should be reflected by the topology in such a way that points close to each other have approximately the same meaning. An atom, untimely placed in the support of some continuous distribution, has only little in common with its closest neighbours, and therefore it should rather be given its own status as an isolated point.

18. COMPUTATIONS IN RELATION TO CONDITIONING.

The considerations of section 17 showed that conditional distributions do in general exist, at least on a submanifold Y of probability 1 . Thus the conditioning problem in the continuous case can in general be "solved". But the conditional distributions are of little or no use to us, as long as their existence is the only property known to us. The somewhat more interesting algebraic problems concerning their computation (i.e. the computation of their densities with respect to the geometric measures) will be discussed in this section.

In general, we are interested in operations like

integration of a function
multiplication of a measure with a density
transformation of a measure
mixing of measures
construction of product measures
and conditioning.

We begin by a short outline of the problems involved:

Integration of a function will be discussed later, as one of the two "basic" operations.

Multiplication by a density does not in itself imply further difficulties than does the general problem of integration.

Transformation (by a dimension preserving or dimension reducing surjectively regular mapping) is carried out by the formula

$$g(y) = \lambda_{X_y}(F \cdot f)$$

for the density of the transformed measure (section 16). In order to compute g, we must compute the function

$$F(x) = \frac{1}{|Dt(x)|^{\circ}}$$

(i.e. we must compute the determinant) and then integrate the function $F \cdot f$ with respect to the geometric measure λ_{X_y}.

Mixtures of measures usually appear either as mixtures of measures given by densities with respect to the same geometric measure or as mixtures of measures concentrated on disjoint surfaces in some continuous family (or "fibre bundle") of surfaces. The first type of mixtures is a trivial matter (just "mix" the densities, i.e. integrate the "mixing variable" out of the expression for the density). The second type of mixtures can be handled as "inverse conditioning problems" where the measures μ^y and ν are known, but μ is unknown. The density of the mixture is then given by the formula $f^y(x) = \frac{1}{g(y)}F(x)f(x)$, written on the form

$$f(x) = \frac{g(y)}{F(x)} f^y(x) .$$

Thus the only technical problem in connection with this mixing procedure is computation of F (i.e. of the determinant $|Dt|^{\circ}$).

Construction of product measures does not involve new problems. Under the obvious definition of product Riemann structure, the geometric measure on the product manifold equals the product of the geometric measures on the components. Thus the density of a product probability measure is simply the ("tensor"-) product of the two densities:

$$(f\cdot\lambda_X)\otimes(g\lambda_Y) \quad = \quad (f\otimes g)\cdot(\lambda_X\otimes\lambda_Y) \quad = \quad (f\otimes g)\cdot\lambda_{X\times Y}$$

where

$$f\otimes g \quad = \quad [f(x)g(y)]_{(x,y)} \ .$$

Conditioning. Conditional distributions are given by the density

$$f^y \quad = \quad \frac{F\cdot f}{g(y)} \qquad \text{(section 16)}$$

with respect to the geometric measure on the level surface for t. The technical problems coincide with those of a transformation: Computation of F and g.

Hence, it seems as if the basic operations are integration of a function with respect to a geometric measure and computation of a determinant of the form $|Dt(x)|^{\circ}$. By means of these two operations, all those above can be carried out.

Therefore, we shall discuss these two problems in more details.

Integration of a function. Theorem 15.1 (the decomposition of the geometric measure) may, of course, be of help in facing the problem of integration of a given function. Classical tools like transformation into polar coordinates, not to mention Fubini's theorem, are typical examples. However, only little can be said about this technique of decomposition in general.

The basic, straight forward method for computation of integrals is, of course, a choice of parametrization, possibly piecewise, transforming the integral into an integral or a sum of integrals with respect to Lebesgue measure on $\mathbb{R}^n$.

The decomposition of the integral as a sum of integrals over open subsets, covering X except for a set of measure 0 (a union of closed manifolds of lower dimensions) requires no explanation, so we shall restrict our attention to the case where a parametrization

$$t: X' \to X$$

of X by an open subset X' of $\mathbb{R}^n$ exists. We have then, by the definition of the geometric measure (page 123, section 14)

$$\lambda_X f = \lambda_{X'}(|Dt|(f \circ t)),$$

where $\lambda_{X'}$ denotes the restriction of Lebesgue measure to X'.

This formula yields a new problem: Computation of $|Dt|$.

In case X is a submanifold of $\mathbb{R}^N$, we may write $|Dt|_o$ in stead of $|Dt|$, if t is regarded as a mapping with co-domain $\mathbb{R}^N$ (the tangent space for X is then identified with the image for $Dt(x')$ and theorem 11.1, page 73, is applied). From this point of view $Dt(x')$ is then a linear mapping

$$Dt(x') : \mathbb{R}^n \to \mathbb{R}^N ,$$

and the determinant is given by

$$|Dt(x')|_o = \sqrt{|Dt(x')^* Dt(x')|} = \sqrt{|\det((a_{ij}))|}$$

where

$$a_{ij} = \frac{\partial t_1(x')}{\partial x'_i} \cdot \frac{\partial t_1(x')}{\partial x'_j} + \ldots + \frac{\partial t_N(x')}{\partial x'_i} \cdot \frac{\partial t_N(x')}{\partial x'_j}$$

$$= \left(\frac{\partial t(x')}{\partial x'_i} \mid \frac{\partial t(x')}{\partial x'_j} \right) .$$

This is all we can say about computation of integrals in general. The resulting integrals with respect to Lebesgue measure may, of course, turn out to be rather complicated, and for $n > 3$ the computation of the determinant may also be a rather hopeless affair, but these problems can hardly be avoided in general.

Computation of the determinant $|Dt(x)|^{\circ}$. Let $X \subseteq \mathbb{R}^N$ and $Y \subseteq \mathbb{R}^M$ be submanifolds of dimensions n and m, respectively, and let $t: X \to Y$ be a surjective and surjectively regular transformation. We have

$$|Dt(x)|^{\circ} = \sqrt{|Dt(x)Dt(x)^*|} \quad .$$

A direct application of this formula is possible if we are able to "compute" the adjoint mapping $Dt(x)^*$, compose it with $Dt(x)$, and then compute its determinant.

In case X and Y are open submanifolds ($n=N$ and $m=M$), this is, in principle, easy. Similarly to the computation of the "injective determinant" $|Dt|_{\circ}$ on page 150, we get

$$|Dt(x)|^{\circ} = \sqrt{|\det((b_{ij}))|}$$

where

$$b_{ij} = \frac{\partial t_i(x)}{\partial x_1} \cdot \frac{\partial t_j(x)}{\partial x_1} + \ldots + \frac{\partial t_i(x)}{\partial x_n} \cdot \frac{\partial t_j(x)}{\partial x_n}$$

$$= (\operatorname{grad} t_i \,|\, \operatorname{grad} t_j) \quad .$$

For $n < N$ or $m < M$, this method can not be applied. Not even the adjoint differential $Dt(x)^*$ can be "computed" in any reasonable sense if no orthonormal bases in the tangent spaces are given. An artificial choice of orthonormal bases is in general a rather hopeless affair.

We shall illustrate a technique, based on theorem 11.5. In short, the technique is this: Construct a diagram, satisfying the conditions of theorem 11.5, containing Dt as the only mapping with unknown determinant. Thus the remaining linear mappings of the diagram should be mappings between spaces $\mathbb{R}^n$, $\mathbb{R}^m$ etc., or their determinants should be otherwise computable. This technique turns out to be very efficient, though hard to describe in general. Some ingenuity may be of help in the choice of diagrams.

Some more or less relevant cases are discussed below. It should be emphasized, that these cases are examples. It is probably true that all situations where X and Y are given as level surfaces or parameterized surfaces or combinations of such can be treated by a combination of the five cases below, but it may as well be easier to apply theorem 11.5 or some other result directly. Even in the four cases 1-4 treated here, I suggest that theorem 11.5 is applied directly. Exactness and commutativity of a concrete diagram can easily be checked. Theorem 11.5 is, in spite of its rather abstract algebraic formulation, a very simple and useful result, which I hope will be illustrated rather than obscured by the examples.

In the following, the spaces X_o, Y_o etc. may be regarded as the spaces $\mathbb{R}^N$, $\mathbb{R}^M$ etc., that is we assume that determinants of mappings between these spaces (with subscript o) are known. In all cases, X and Y are given as submanifolds of X_o and Y_o, and $t:X \to Y$ is surjective and surjectively regular. The

problem is the computation of $|Dt(x)|^{o}$.

Case 1. Let

$$t: X \to Y$$

be the restriction of a surjectively regular mapping

$$t_o: X_o \to Y_o$$

where

$$Y = Y_o \quad \text{(or } Y \text{ is an open submanifold of } Y_o\text{)}$$

and X is given as a level surface

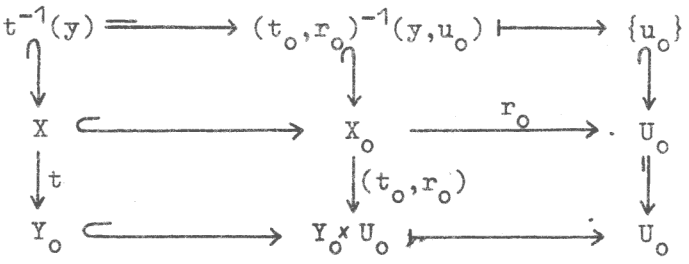
$$X := r_o^{-1}(u_o)$$

where

$$r_o: X_o \to U_o$$

is surjectively regular. It is assumed, that (t_o, r_o) is surjectively regular. Obviously then, t is also surjectively regular.

For $y \in Y$, consider the diagram



The arrows of the bottom row denote the imbedding of the fibre

$$Y_o \approx Y_o \times \{u_o\}$$

and the projection onto the second factor U_o .

Obviously this diagram commutes. The identity

$$t^{-1}(y) = (t_o, r_o)^{-1}(y, u_o)$$

follows immediately from the definition of X:

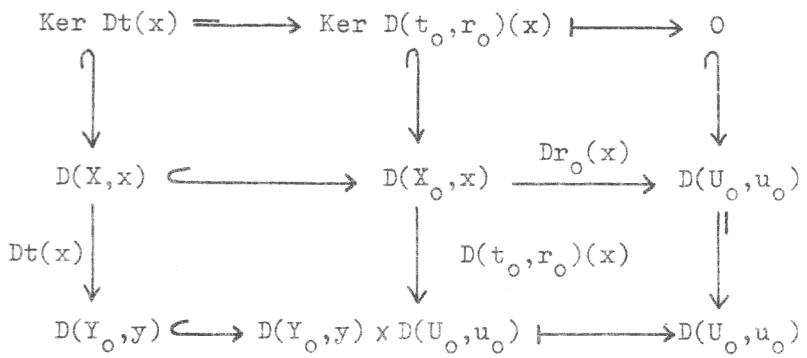
For $x \in X_o$ we have

$$\begin{aligned} & x \in t^{-1}(y) \\ \iff\ & x \in X \text{ and } t(x) = y \\ \iff\ & x \in X \text{ and } t_o(x) = y \\ \iff\ & r_o(x) = u_o \text{ and } t_o(x) = y \\ \iff\ & x \in (t_o, r_o)^{-1}(y, u_o) . \end{aligned}$$

The arrows $\longmapsto$ are used for mappings the differentials of which are coisometries (just as the imbeddings $\hookrightarrow$ have isometric differentials).

Now, for $x \in t^{-1}(y)$, the functor D is applied. The diagram so obtained obviously satisfies the conditions of theorem 11.5 (page 83), by the criterion mentioned on page 113 (the zeros are ommitted; but don't forget, that the exactness conditions of theorem 11.5 also involve, that all mappings should be

either injective or surjective).



Theorem 11.5 gives

$$1\cdot 1\cdot |Dr_o(x)|^o \cdot 1\cdot |Dt(x)|^o \cdot 1$$

$$= \quad 1\cdot 1\cdot 1\cdot 1\cdot |D(t_o,r_o)(x)|^o \cdot 1 \quad ,$$

or

$$|Dt(x)|^o \quad = \quad \frac{|D(t_o,r_o)|^o}{|Dr_o(x)|^o} \quad = \quad \frac{\left|\begin{bmatrix} Dt(x) \\ Dr_o(x) \end{bmatrix}\right|^o}{|Dr_o(x)|^o}$$

$$= \sqrt{\frac{\left|\begin{bmatrix} Dt_o(x)Dt_o(x)^* & Dt_o(x)Dr_o(x)^* \\ Dr_o(x)Dt_o(x)^* & Dr_o(x)Dr_o(x)^* \end{bmatrix}\right|}{|\,Dr_o(x)Dr_o(x)^*|}}$$

Case 2. Now, let $t: X \to Y$ be the restriction of a surjectively regular mapping $t_o: X_o \to Y_o$, where Y is given as a level surface

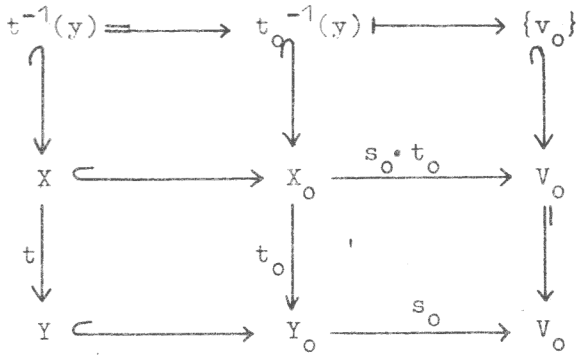
$$Y = s_o^{-1}(v_o) \quad , \qquad s_o: Y_o \to V_o \text{ surjectively regular,}$$

while

$$X = (s_o \circ t_o)^{-1}(v_o) = t_o^{-1}(Y)$$

is the corresponding surface in X. Obviously then, t is surjectively regular.

For $y \in Y$, consider the diagram



The diagram obviously commutes, and for $x \in t^{-1}(y)$ the corresponding diagram of tangent spaces and differentials satisfies the conditions of theorem 11.5 (this time we shall not draw the "linearized" diagram; just apply the remarks on page 113). We get

$$1\cdot 1\cdot 1\cdot 1\cdot |Dt_o(x)|^o \cdot |Ds_o(y)|^o = 1\cdot 1\cdot |D(s_o\circ t_o)(x)|^o \cdot 1\cdot |Dt(x)|^o \cdot 1 ,$$

or

$$|Dt(x)|^o = \frac{|Dt_o(x)|^o |Ds_o(y)|^o}{|D(s_o\circ t_o)(x)|^o} .$$

Case 3. Let $Y \subseteq Y_o$ be given, as in case 2, by

$$Y = s_o^{-1}(v_o) .$$

But now we assume that further restrictions are imposed on $x \in X$ (as in case 1) such that X is given as the level surface

$$X = (s_o\circ t_o)^{-1}(v_o) \cap r_o^{-1}(u_o)$$

$$= (s_o\circ t_o , r_o)^{-1}(v_o,u_o) .$$

As in case 1, we assume that (t_o,r_o) is surjectively regular. From this it follows easily that $(s_o\circ t_o , r_o)$ and t are surjectively regular.

For $y \in Y$, consider the diagram

$$\begin{array}{ccccc}
t^{-1}(y) & \longrightarrow & (t_o,r_o)^{-1}(y,u_o) & \longmapsto & \{(v_o,u_o)\} \\
\downarrow & & \downarrow & & \downarrow \\
X & \subset\longrightarrow & X_o & \xrightarrow{(s_o\circ t_o,r_o)} & V_o \times U_o \\
t\downarrow & & \downarrow{\scriptstyle (t_o,r_o)} & & \| \\
Y & \subset\longrightarrow & Y_o \times U_o & \xrightarrow{(s_o\times 1_{U_o})} & V_o \times U_o
\end{array}$$

The inclusion $Y \hookrightarrow Y_o \times U_o$ of the bottom row should be interpreted as the imbedding of

$$Y \approx Y \times \{u_o\} .$$

The identity $t^{-1}(y) = (t_o, r_o)^{-1}(y, u_o)$ is easily proved.

The corresponding diagram of differentials (for $x \in t^{-1}(y)$) satisfies the conditions of theorem 11.5. Thus

$$1 \cdot 1 \cdot 1 \cdot 1 \cdot |D(t_o, r_o)(x)|^o \cdot |D(s_o \times 1_{U_o})(y, u_o)|^o$$

$$= \quad 1 \cdot 1 \cdot |D(s_o \circ t_o, r_o)(x)|^o \cdot 1 \cdot |Dt(x)|^o \cdot 1 \quad ,$$

or

$$|Dt(x)|^o = \frac{|D(t_o, r_o)(x)|^o |Ds_o(y)|^o}{|D(s_o \circ t_o, r_o)(x)|^o} = \frac{\left|\begin{bmatrix} Dt_o(x) \\ Dr_o(x) \end{bmatrix}\right|^o \cdot |Ds_o(y)|^o}{\left|\begin{bmatrix} Ds_o(y)Dt_o(x) \\ Dr_o(x) \end{bmatrix}\right|^o} .$$

Notice that this formula contains the results of case 1 and case 2. Conversely, it can be derived from the two previous cases, first applying case 1 to express $|Dt|^o$ by determinants of mappings with domain $X_1 = r_o^{-1}(u_o)$ and then computing these determinants by means of the formula for case 2.

Case 4. Case 1,2 and 3 cover most of the situations where t is given as the restriction of a surjectively regular mapping t_o. In case t_o is not surjectively regular (for example, when $t_o: \mathbb{R}^N \to \mathbb{R}^M$ maps $\mathbb{R}^N$ onto some submanifold $Y \subseteq \mathbb{R}^M$ and $t: \mathbb{R}^N \to Y$ is the restriction), a formula for "reduction of codomain" may be useful.

Suppose that the image $Y = t_o(X_o)$ can be represented as a level surface

$$Y = s_o^{-1}(v_o) ,$$

$$s_o: Y_o \to V_o \quad \text{surjectively regular.}$$

Consider the diagram

$$\begin{array}{ccccc}
\operatorname{Ker} Dt(x) & \longrightarrow & \operatorname{Ker}[Dt_o(x) \quad Ds_o(y)^*] & \longmapsto & 0 \\
\downarrow & & \downarrow & & \downarrow \\
D(X_o,x) & \hookrightarrow & D(X_o,x) \times D(V_o,v_o) & \longmapsto & D(V_o,v_o) \\
\downarrow Dt(x) & & \downarrow [Dt_o(x) \quad Ds_o(y)^*] & & \downarrow Ds_o(y)Ds_o(y)^* \\
D(Y,y) & \hookrightarrow & D(Y_o,y) & \xrightarrow{Ds_o(y)} & D(V_o,v_o)
\end{array}$$

(for the definition of $[Dt_o(x) \quad Ds_o(y)^*]$, see page 61). $[Dt_o(x) \quad Ds_o(y)^*]$ is surjective since the image of $Ds_o(y)^*$ is a complement to the image of $Dt_o(x)$ (namely the orthogonal complement). From this it follows (since $Ds_o(y)^*$ is injective)

that

$$\mathrm{Ker}(Dt(x))\times\{0\} = \mathrm{Ker}[Dt_o(x) \quad Ds_o(y)^*]$$

Hence the conditions of theorem 11.5 seem to be satisfied and we get

$$1\cdot 1\cdot 1\cdot 1\cdot |[Dt_o(x) \quad Ds_o(y)^*]|^o \cdot |Ds_o(y)|^o$$

$$= \quad 1\cdot 1\cdot 1\cdot |Ds_o(y)Ds_o(y)^*| \cdot |Dt(x)|^o \cdot 1$$

or

$$|Dt(x)|^o = \frac{|[Dt_o(x) \quad Ds_o(y)^*]|^o}{|Ds_o(y)|^o} \quad .$$

In this formula, $Ds_o(y)$ may of course be replaced by any other surjective linear mapping with $D(Y,y)$ as its kernel.

Case 5. Finally, consider the case where X and Y are given by parametrizations

$$p_o: X_o' \to X_o \quad , \qquad X = p_o(X_o')$$

and

$$q_o: Y_o' \to Y_o \quad , \qquad Y = q_o(Y_o') \quad .$$

Let

$$p: X_o' \to X$$

and

$$q: Y_o' \to Y$$

denote the restrictions (thus p and q are diffeomorphisms, according to the definition of a parametrization (page 107)). The composed mapping

$$t'_o := q^{-1} \circ t \circ p$$

is assumed to be "known". This time we shall not apply theorem 11.5. The determinant $|Dt(x)|^o$ can be computed as follows:

For $y := t(x)$, $x'_o := p^{-1}(x)$, $y'_o := q^{-1}(y)$, put (for convenience)

$$T := Dt(x)$$
$$Q := Dq(y'_o)$$
$$P := Dp(x'_o)$$
$$T'_o := Dt'_o(x'_o) = Q^{-1}TP .$$

Then

$$|Dt(x)|^o = |T|^o = |QT'_oP^{-1}|^o = \sqrt{|QT'_oP^{-1}P^{*-1}T'^{*}_oQ^*|}$$

$$= \sqrt{|Q|\cdot|T'_o(P^*P)^{-1}T'^{*}_o|\cdot|Q^*|}$$

$$= |Q|\cdot\sqrt{|T'_o(P^*P)^{-1}T'^{*}_o|} .$$

By theorem 11.1 (page 73) we have

$$|Q| = |Dq(y'_o)| = |Dq_o(y'_o)|_o .$$

Further, P^*P can be computed as follows: Let $j:X \to X_o$ denote the imbedding. Then

$$p_o = j \circ p ;$$

since $Dj(x)$ is isometric we have, for $J := Dj(x)$ and $P_o := Dp_o(x'_o)$ (and thus $P_o = JP$)

$$P^*P = P^*(J^*J)P = (JP)^*JP = P_o^* P_o .$$

Hence

$$|Dt(x)|^o = |Dq_o(y'_o)|_o \cdot \sqrt{|Dt'_o(x'_o)(Dp_o(x'_o)^* Dp_o(x'_o))^{-1} Dt'_o(x'_o)^*|} .$$

The appearence of an inversed matrix in this formula is somewhat unfavorable, but there seems to be nothing to do about it. However, parametrized surfaces are of little interest to us: In case a surface has a natural parametrization, we will probably always prefer to discuss the "parameters" themselves, rather than imbedding them in a space of too high dimension. Our interest in differentiable manifolds is motivated by the occurrence of submanifolds of $\mathbb{R}^N$ without any natural parametrization, and such surfaces are in general given as level surfaces.

CHAPTER V : CONDITIONAL EXPECTATIONS

19. CONDITIONAL EXPECTATIONS.

The finite case. Let X and Y be finite sets with probability measures μ and ν, and let

$$t:(X,\mu) \to (Y,\nu)$$

be a homomorphism. The conditional distribution μ^y is given by its density with respect to counting measure:

$$\mu^y\{x\} = \begin{cases} \frac{\mu\{x\}}{\nu\{y\}} & \text{for } x \in t^{-1}(y) \\ 0 & \text{for } x \notin t^{-1}(y) \end{cases}$$

(defined for $\nu\{y\} > 0$, i.e. for $y \in \operatorname{supp} \nu$).

This aspect of the concept of conditioning was generalized by the local definition of a conditional distribution.

The elementary concept of conditioning in the finite case has, however, other aspects. One of them is the following:

Let

$$f: X \to \mathbb{R}$$

be a function, and suppose we want to determine its value $f(x)$ (as good as possible) from an observation of the derived stochastic variable

$$y = t(x), \qquad x \in (X, \mu) .$$

Our <u>estimator</u>

$$\hat{f}: Y \to \mathbb{R}$$

should be <u>optimal</u> in the sense that the expected value of the quadratic error $(f(x)-\hat{f}(y))^2$ is as small as possible.

The expected quadratic error can be expressed by inner products in the Euclidean vectorspace $L^2(\mu)$:

$$E(f(x)-\hat{f}(y))^2 = \|f - \hat{f}\circ t\|_2^2 = (f - \hat{f}\circ t | f - \hat{f}\circ t)_\mu .$$

Hence, $\hat{f}$ should be chosen such that $\hat{f}\circ t$ is as close as possible to f in this space. The set of functions of the form $g\circ t$ constitutes a subspace of $L^2(\mu)$. The point $\hat{f}\circ t$ in this subspace closest to f is of course the orthogonal projection of f onto that subspace. Hence, the difference $f - \hat{f}\circ t$ must be orthogonal to the subspace, i.e.

$$(f - \hat{f}\circ t | g\circ t)_\mu = 0 \quad \text{for all } g$$

or $$(f|g\circ t)_\mu = (\hat{f}\circ t | g\circ t)_\mu = (\hat{f}|g)_\nu \text{ for all } g .$$

In particular, inserting $g := 1_{\{y\}}$ $(y \in Y)$, we get

$$(f|1_{t^{-1}(y)})_\mu = (\hat{f}|1_{\{y\}})_\nu$$

or

$$\sum_{x \in t^{-1}(y)} f(x)\,\mu\{x\} = \hat{f}(y)\,\nu\{y\}$$

or (for $\nu\{y\} > 0$)

$$\hat{f}(y) = \frac{1}{\nu\{y\}} \sum_{x \in t^{-1}(y)} f(x)\,\mu\{x\} = \mu^y f\,.$$

Thus, the solution to our estimation problem is this: The best estimate (in the quadratic mean-sense) of $f(x)$ is the expectation of $f(x)$ when x is varying according to its <u>conditional</u> distribution μ^y, given the observed value of y.

This connection between the geometry of the L^2-spaces and the concept of conditioning can be stated in a more precise manner, as follows:

The equations

$$(f|g\circ t)_\mu = (\hat{f}|g)_\nu\,, \qquad g \in L^2(\nu)$$

express that the mapping taking f into its estimator $\hat{f}$ is the <u>adjoint</u> mapping to the operator taking g into $g\circ t$ The latter is denoted

$$L^2(t) : L^2(\nu) \to L^2(\mu)$$

$$L^2(t)g := g \circ t .$$

Hence, what we have proved is that the adjoint operator

$$L^2(t)^* : L^2(\mu) \to L^2(\nu)$$

has the pointwise representation

$$(L^2(t)^* f)(y) = \mu^y f .$$

For this reason, the operator $L^2(t)^*$ will be called the conditional expectation operator.

Remark. The notation $L^2(t)$ for the linear operator induced by t indicates that we have to do with a functor; which we have, actually: L^2 can be regarded as a functor from the category of probability fields (here: Finite probability fields) into the category of Hilbert spaces with bounded linear mappings as homomorphisms. But the functor is contravariant, i.e. it reverses arrows. Thus

$$(X,\mu) \xrightarrow{\ t\ } (Y,\nu)$$

is transformed into

$$L^2(\mu) \xleftarrow{L^2(t)} L^2(\nu) .$$

Now, the category of Hilbert spaces is contravariant isomorphic to itself by the adjointness functor *, taking objects (Hilbert spaces) into themselves and homomorphisms (bounded linear mappings) into their adjoints. Composing the adjointness functor with L^2, we obtain the covariant (i.e. arrow direction preserving) functor L^{2*}, the conditional expectation functor.

Definition of conditional expectations. While the local definition of conditional distributions required some regularity conditions, the above quadratic mean aspect of conditioning is immediately generalized:

Definition: Let

$$t: (X,\mu) \to (Y,\nu)$$

be a homomorphism between probability fields. An isometry

$$L^2(t): L^2(\nu) \to L^2(\mu)$$

is defined by

$$L^2(t)g := g \circ t .$$

The adjoint operator

$$L^2(t)^* : L^2(\mu) \to L^2(\nu)$$

is called the conditional expectation operator, and for

$f \in L^2(\mu)$, the function (or, more precisely: the equivalence class)

$$L^2(t)^* f \quad \in \quad L^2(\nu)$$

is called the conditional expectation of f , given t .

Properties of conditional expectations.

19.1 Theorem. The operator $L^2(t)^*$ has the following properties:

(1) The composed operator

$$L^2(t)L^2(t)^* : L^2(\mu) \to L^2(\mu)$$

(taking f into its conditional expectation regarded as a function on X) is the orthogonal projection onto the subspace

$$\{g \circ t \mid g \in L^2(\nu)\} .$$

(2) $L^2(t)^* 1_X = 1_Y$ (here, 1_X and 1_Y denote indicator functions, of course).

(3) $\nu(L^2(t)^* f) = \mu f$.

(4) $f \geq 0 \implies L^2(t)^* f \geq 0$.

(5) $\|L^2(t)^* f\|_1 \leq \|f\|_1$ (or $\|L^2(t)^* f\|_\nu \leq \|f\|_\mu$, see page 341).

(6) $\|L^2(t)^* f\|_2 \leq \|f\|_2$.

(7) $\|L^2(t)^* f\|_\infty \leq \|f\|_\infty$ (here, $\|\ \|_\infty$ denotes, of course, <u>essential</u> supremum norm. The inequality is also valid in case one of the sides equals $+\infty$).

<u>Proof</u>: (1) follows from the fact that $L^2(t)$ is isometric (for any isometric operator T between two Hilbert spaces, TT^* is the orthogonal projection onto the image of T).

(2): For $g \in L^2(\nu)$ we have ($L^2(t)$ being isometric)

$$(g|L^2(t)^* 1_X)_\nu = (L^2(t)g|1_X)_\mu = (L^2(t)g|L^2(t)1_Y)_\mu = (g|1_Y)_\nu .$$

This equation being valid for all g, we conclude that

$$L^2(t)^* 1_X = 1_Y .$$

(3) follows from

$$\nu(L^2(t)^*f) = (L^2(t)^*f|1_Y)_\nu = (f|L^2(t)1_Y)_\mu$$

$$= (f|1_X)_\mu = \mu f .$$

(4): Let $f \geq 0$ be given. For $g \geq 0$ ($g \in L^2(\nu)$) we have then

$$(g|L^2(t)^*f)_\nu = (L^2(t)g|f)_\mu = \mu((g\circ t)\cdot f) \geq 0 .$$

Let

$$L^2(t)^*f = \hat{f} = \hat{f}_+ - \hat{f}_- \qquad (\hat{f}_+, \hat{f}_- \geq 0)$$

be the decomposition of $\hat{f} = L^2(t)^*f$ into its positive and its negative part. The above statement for a function $g \geq 0$ is in particular valid for $g := \hat{f}_-$, and in that case we get

$$\|\hat{f}_-\|_2^2 = (\hat{f}_-|\hat{f}_-)_\nu = (\hat{f}_-|\hat{f}_-)_\nu - 0$$

$$= (\hat{f}_-|\hat{f}_-)_\nu - (\hat{f}_-|\hat{f}_+)_\nu = -(\hat{f}_-|\hat{f})_\nu = -(\hat{f}_-|L^2(t)^*f) \leq 0,$$

and so $\|\hat{f}_-\|_2^2 = 0$, i.e. $L^2(t)^*f \geq 0$.

(5): From (4) it follows that

$$L^2(t)^*(-|f|) \leq L^2(t)^*f \leq L^2(t)^*(|f|) ,$$

i.e. $$|L^2(t)^*f| \leq L^2(t)^*(|f|) .$$

Then

$$\|L^2(t)^*f\|_1 = \nu(|L^2(t)^*f|) \leq \nu(L^2(t)^*(|f|))$$

$$= \mu(|f|) = \|f\|_1 .$$

(6) is easily proved (by the formula $\|T^*\| = \|T\|$, or directly).

(7) is a consequence of (2) and (4): For $f \in L^2(\mu)$ we have

$$-\|f\|_\infty \cdot 1_X \leq f \leq \|f\|_\infty \cdot 1_X ,$$

and so

$$-\|f\|_\infty \cdot 1_Y \leq L^2(t)^* f \leq \|f\|_\infty \cdot 1_Y ,$$

i.e.

$$\|L^2(t)^* f\|_\infty \leq \|f\|_\infty .$$

This argument holds for $\|f\|_\infty < +\infty$, and for $\|f\|_\infty = +\infty$ the assertion is trivial.

Remark. It follows from (5), that the operator $L^2(t)^*$ can be extended by continuity to an L^1-operator $L(\mu) \to L(\nu)$. This means that the conditional expectation of f can be defined as soon as f is integrable. This can be done directly (i.e. without going through the L^2-case first), and that is the way it is usually done in the litterature. The self-duality of Hilbert spaces (necessary for the definition of adjoint operator)

is then replaced by the more special Radon-Nikodym theorem. The description of the L^1-operator is, however, a much more complicated affair, and the close analogy between geometric and probabilistic structure (see for example theorem 25.1,(3) page 210) is completely obscured by the L^1-definition. Though it may possibly be of interest in some cases to consider a conditional expectation of a stochastic variable with infinite variance, I think that the L^2-definition should be applied in general, in view of its canonical and unavoidable character: It arises almost immediately, when the L^2-spaces are considered. Notice that (seemingly basic) concepts like variance and correlation are closely related to the L^2-spaces: The standard deviation (the square root of the variance) of a stochastic variable $f(x)$ is simply f's distance to the one-dimensional subspace of constant functions, and the correlation coefficient for two stochastic variables $f_1(x)$ and $f_2(x)$ is the inner product of the normalized vectors $f_1/\|f_1\|_2$ and $f_2/\|f_2\|_2$ after subtraction of their constant components (i.e. after projection onto the orthogonal complement to the space of constant functions). It seems natural, that these measures of dispersion and dependence should be related to a geometry which reflects the basic ideas of stochastic independence (theorem 25.1) and conditioning.

The connection to conditional distributions. We have now given two different generalizations of the elementary concept of conditioning in the finite case: The local definition of conditional distributions, as handled in the previous

chapters, and the global definition of conditional expectations, as given in this section.

In the classical theory, only the global definition has a meaning. Conditional expections (in particular conditional probabilities defined as conditional expectations of indicator functions) were introduced by Kolmogorov (1933). The only local aspect lies, within the classical theory, in the possible existence of a pointwise representation

$$(L^2(t)^* f)(y) = \mu^y f$$

of the conditional expectation operator. This was Doob's point of view, see Doob (1953).

In our exposition, both sides of the above equation are meaningfull, and therefore we must prove it, in some sense. The left side is, however, only defined up to equivalence. Hence, the equation can not be quite true. In order to ascribe pointwise meaning to the left side, we need a definition of the value of a function at a point, in case the function is only given up to equivalence. This is the justification of the next section.

20. ESSENTIAL VALUES AND ESSENTIAL CONTINUITY.

Let λ be an arbitrary measure on X and let

$$f: X \to \mathbb{R}$$

be a locally integrable function (cfr. the appendix, page 350). Let x_o be a point in the support of λ .

Definition: If the net

$$\left(\frac{\lambda(1_A \cdot f)}{\lambda A} \mid A \to x_o \ , \ \lambda A > 0 \right)$$

is convergent towards a real number r ($\neq \pm\infty$), we say that f has the essential value r at the point x_o . we write

$$f_{ess}(x_o) := r = \lim_{A \to x_o} \frac{\lambda(1_A \cdot f)}{\lambda A}$$

for the essential value when it exists.

Notice, that the essential value is unchanged when f is changed on a set of measure 0 . The essential value (and its existence) depends on the equivalence class only.

According to lemma 7.2 (page 39), the essential value $f_{ess}(x_o)$

is defined if the restriction of f to $\operatorname{supp} \lambda$ is continuous at the point x_o .

In order to investigate further the relationship between the continuity properties of a function and the existence of its essential values, we introduce two auxiliary functions

$$\underline{f} \quad \text{and} \quad \overline{f} \; : \; X \to [-\infty, +\infty]$$

which we might call the essential hulls of f ; they are defined by

$$\underline{f}(x) := \sup\{g(x) \mid g: X \to [-\infty, +\infty] \text{ continuous}, \; g \leq f \text{ almost everywhere}\} ,$$

$$\overline{f}(x) := \inf\{g(x) \mid g: X \to [-\infty, +\infty] \text{ continuous}, \; g \geq f \text{ almost everywhere}\} .$$

20.1 Theorem. The functions $\underline{f}$ and $\overline{f}$ have the following properties:

$\underline{f}$ is lower semicontinuous.

$\overline{f}$ is upper semicontinuous.

For almost all x we have

$$\underline{f}(x) \leq f(x) \leq \overline{f}(x) ,$$

and the inequality

$$\underline{f}(x) \leq \overline{f}(x)$$

is valid for <u>all</u> $x \in \operatorname{supp} \lambda$, while for $x \notin \operatorname{supp} \lambda$ we have

$$\underline{f}(x) = +\infty \quad \text{and} \quad \overline{f}(x) = -\infty .$$

<u>Proof</u>: The semicontinuity of the functions $\underline{f}$ and $\overline{f}$ follows immediately from their definition as upper and lower bounds of sets of continuous functions.

In order to prove the inequality

$$f \leq \overline{f} \quad \text{almost sure}$$

notice first, that we need only prove it in case of a bounded function $f \geq 0$ with compact support. Then the σ-compactness of X obviously ensures that the statement holds for arbitrary bounded functions ≥ 0 , and the boundedness- and positivity-conditions are not restrictive since we may replace the interval $[-\infty, +\infty]$ by $[0,1]$ by means of some increasing homeomorphism $\varphi : [-\infty, +\infty] \to [0,1]$ (obviously, φ commutes with the constructions $\underline{\ }$ and $\overline{\ }$, in the sense that $\underline{\varphi \circ f} = \varphi \circ \underline{f}$ etc.).

For a bounded function $f \geq 0$ with compact support the function $\overline{f}$ is obviously integrable. Moreover, there exists a function $g \in \mathcal{K}(X)$ such that $g \geq f$ (theorem A 1 , page 334). It

follows from theorem A 13 (page 344) that

$$\lambda \bar{f} = \inf\{\lambda g \mid g \in \mathcal{K}(X) , g \geq f \text{ almost everywhere}\} .$$

We can choose a decreasing sequence (g_n) of $\mathcal{K}(X)$-functions such that $g_n \geq f$ almost everywhere and

$$\lim \lambda g_n = \lambda \bar{f} .$$

Then, for $g_o(x) := \lim g_n(x)$ we have

$$g_o(x) = \bar{f}(x) \text{ almost sure,}$$

and from

$$g_n(x) \geq f(x) \text{ almost sure}$$

it follows immediately (since a denumerable union of null sets is again a null set) that

$$g_o(x) \geq f(x) \text{ almost sure.}$$

Hence,

$$\bar{f} = g_o \geq f \text{ almost everywhere.}$$

The inequality $\underline{f} \leq f$ almost everywhere is proved in a similar fashion.

Now, put

$$U := \{x|\underline{f}(x) > \overline{f}(x)\} = \bigcup_{a \in [-\infty,+\infty]} \{x|\underline{f}(x) > a > \overline{f}(x)\} .$$

It follows from the semicontinuity of $\underline{f}$ and $\overline{f}$ that U is open. From

$$\underline{f}(x) \leq f(x) \leq \overline{f}(x) \quad \text{almost sure}$$

we conclude that U is a null set. Since any open null set is contained in the complement of the support, we have

$$\underline{f}(x) \leq \overline{f}(x) \quad \text{for } \underline{\text{all}} \quad x \text{ in } \operatorname{supp} \lambda .$$

Finally, for $x \notin \operatorname{supp} \lambda$ we have, according to theorem A 1 (page 334) a continuous function $g: X \to [-\infty,+\infty]$ such that g is $-\infty$ on the support and $g(x) = +\infty$. By the definition of $\underline{f}$ we have then (since $g \leq f$ almost everywhere) $g \leq \underline{f}$, and so $\underline{f}(x) = +\infty$. Similarly, $\overline{f}(x) = -\infty$.

The functions $\underline{f}$ and $\overline{f}$ are related to the concept of essential values in the following manner:

20.2 Theorem. The function f has an essential value at the point x_o if and only if

$$-\infty < \underline{f}(x_o) = \overline{f}(x_o) < +\infty .$$

In case of existence, we have

$$f_{ess}(x_o) = \underline{f}(x_o) \quad (= \overline{f}(x_o)\ ).$$

Proof: First suppose that f has the essential value r at x_o. For $\varepsilon > 0$ we can choose an open neighbourhood U of x_o with compact closure such that for any open subset A of U with $\lambda A > 0$ we have

$$(*) \qquad \left| r - \frac{\lambda(1_A \cdot f)}{\lambda A} \right| \leq \frac{\varepsilon}{2} .$$

Obviously then, this inequality holds for any measurable set $A \subseteq U$ with $\lambda A > 0$, since any measurable set can be approximated in measure from the outside by open sets (theorem A 18, page 348). In particular, the set

$$A := \{x \in U \mid f(x) < r - \varepsilon \}$$

must be a null set: If this was not the case, we should have

$$\frac{\lambda(1_A \cdot f)}{\lambda A} \leq \frac{\lambda(1_A \cdot (r - \varepsilon))}{\lambda A} = r - \varepsilon ,$$

contradicting (*) above.

Now let

$$g: X \to [-\infty, +\infty]$$

be a continuous function such that

$$g(x) = -\infty \quad \text{for} \quad x \in X\backslash U$$
$$g(x) \leq r - \varepsilon \quad \text{for} \quad x \in U$$
$$g(x_0) = r - \varepsilon \ .$$

The existence of such a function follows easily from theorem A 1 . Obviously we have

$$g \leq f \quad \text{almost everywhere}$$

and by the definition of $\underline{f}$ we conclude that

$$\underline{f}(x_0) \geq g(x_0) = r - \varepsilon \ .$$

This being true for all $\varepsilon > 0$, we have

$$\underline{f}(x_0) \geq r \ ,$$

and from the (similarly proved) inequality

$$\overline{f}(x_0) \leq r$$

and by theorem 20.1 , we conclude that

$$\underline{f}(x_0) = \overline{f}(x_0) = r = f_{ess}(x_0)$$

($x_0 \in \operatorname{supp} \lambda$ follows from the assumption that $f_{ess}(x_0)$ is defined).

Conversely, suppose that

$$\underline{f}(x_o) = \overline{f}(x_o) = r \in \mathbb{R} .$$

Then (by the last statement of theorem 20.1) $x_o \in \operatorname{supp} \lambda$. For $\varepsilon > 0$ we can choose continuous functions

$$\underline{g} , \overline{g} : X \to [-\infty, +\infty]$$

such that

$$\underline{g} \leq f \leq \overline{g} \quad \text{almost everywhere,}$$

and

$$r - \tfrac{\varepsilon}{2} \leq \underline{g}(x_o) \leq \overline{g}(x_o) \leq r + \tfrac{\varepsilon}{2} .$$

Let U be the open neighbourhood of x_o given by

$$U := \{x \mid r - \varepsilon < \underline{g}(x) , \ \overline{g}(x) < r + \varepsilon \} .$$

For any open subset $A \subseteq U$ with $0 < \lambda A < +\infty$ we have then

$$r - \varepsilon \leq \frac{\lambda(1_A \cdot f)}{\lambda A} \leq r + \varepsilon .$$

But this argument shows that the net $\left(\frac{\lambda(1_A \cdot f)}{\lambda A} \mid A \to x_o \right)$ is convergent towards r , i.e. $f_{ess}(x_o) = r$.

20.3 Theorem. Let C denote the set of points x such that the essential value $f_{ess}(x)$ is defined. Then there exists a function

$$f_o : \operatorname{supp} \lambda \to \mathbb{R}$$

with the following properties:

(1) $f_o(x) = f(x)$ for almost all x.

(2) $f_o(x) = f_{ess}(x)$ for $x \in C$.

(3) f_o is continuous at any point x in C (relatively to $\operatorname{supp} \lambda$).

Proof: For $x \in \operatorname{supp} \lambda$ we define

$$f_o(x) := (f(x) \vee \underline{f}(x)) \wedge \overline{f}(x) .$$

That is, we construct f_o from f by changing f at the points where f is not between the two hulls $\underline{f}$ and $\overline{f}$. According to theorem 20.1 f_o equals f almost everywhere. For $x_o \in C$ we have by theorem 20.2

$$\underline{f}(x_o) = \overline{f}(x_o) = f_{ess}(x_o)$$

and so from $\underline{f} \leq f_o \leq \overline{f}$ we conclude that $f_o(x_o) = f_{ess}(x_o)$.

Moreover, for $\varepsilon > 0$

$$\{x \in \operatorname{supp} \lambda \mid f_o(x) \in \,]f_o(x_o) - \varepsilon\ ,\ f_o(x_o) + \varepsilon[\,\}$$

$$\supseteq \{x \in \operatorname{supp} \lambda \mid f_o(x_o) - \varepsilon < \underline{f}(x)\} \cap \{x \in \operatorname{supp} \lambda \mid \overline{f}(x) < f_o(x_o) + \varepsilon\}.$$

It follows from the semicontinuity of $\underline{f}$ and $\overline{f}$ that this set is open relatively to the support; this proves that f_o is continuous at x_o .

20.4 Corollary. The mapping

$$\begin{aligned} x &\to f_{ess}(x) \\ C &\to \mathbb{R} \end{aligned}$$

is continuous.

In case $\operatorname{supp} \lambda = X$, we have in particular that $f_{ess}(x)$ is defined for all x if and only if f is equivalent to a continuous function (and this function is $[f_{ess}(x)]_x$) .

Essential continuity. It follows from theorem 20.3 and lemma 7.2 that existence of the essential value at a point x_o is, roughly, equivalent to continuity at that point. In order to make a precise statement out of this, we shall need the following definition:

Definition: A locally integrable function (or an equivalence class of such) f is said to be *essentially continuous* at x_o if x_o is a point of the support and f is equivalent to (or contains) a function which is continuous at x_o.

We have then

20.5 Theorem. The essential value $f_{ess}(x_o)$ is defined if and only if f is essentially continuous at x_o.

Proof: The "if" part follows immediately from lemma 7.2. The "only if" part is proved as follows: In case $f_{ess}(x_o)$ is defined, put

$$f_1(x) := \begin{cases} f_o(x) & \text{for } x \in \operatorname{supp} \lambda \\ f_o(x_o) & \text{for } x \notin \operatorname{supp} \lambda \end{cases}$$

where f_o is defined as in theorem 20.3. Then, f_1 is equivalent to f, and f_1 is continuous at x_o.

21. THE CONNECTION BETWEEN CONDITIONAL EXPECTATIONS AND CONDITIONAL DISTRIBUTIONS.

Let

$$t: (X, \mu) \to (Y, \nu)$$

be given. By means of the concepts of essential value and essential continuity we can ascribe precise meaning to the equation $(L^2(t)^* f)(y) = \mu^y f$:

21.1 Theorem. Let y_o be a point in supp ν such that the conditional distribution μ^{y_o} is defined. Then for any $\mathcal{K}(X)$-function k the conditional expectation $L^2(t)^* k$ is essentially continuous at y_o, and the essential value is given by

$$(L^2(t)^* k)_{ess}(y_o) = \mu^{y_o} k .$$

Proof: For $B \subseteq Y$, $\nu B > 0$, we have

$$\frac{1}{\nu B} \nu(1_B \cdot L^2(t)^* k) = \frac{1}{\nu B} (1_B | L^2(t)^* k)_\nu$$

$$= \frac{1}{\nu B} (L^2(t) 1_B | k)_\mu = \frac{1}{\nu B} \mu(1_{t^{-1}B} \cdot k) = \mu^B k .$$

For $B \to y_o$ we have then

$$\frac{1}{\nu B}\,\nu(1_B \cdot L^2(t)^* k) \to \mu^{y_o} k \quad ,$$

i.e.

$$(L^2(t)^* k)_{ess}(y_o) = \mu^{y_o} k \quad .$$

The converse theorem is almost valid:

21.2 Theorem. Let y_o be a point in $\operatorname{supp} \nu$. Suppose that for all $k \in \mathcal{K}(X)$, the function $L^2(t)^* k$ is essentially continuous at y_o . Then, the net $(\mu^B \mid B \to y_o)$ is convergent. Thus the conditional distribution is "defined but possibly defective".

Remark: In case X is compact, the reservation concerning defectiveness is unnecessary (and, as a matter of fact: Defective conditional distributions seldom occur in practice).

Proof: The theorem follows immediately from the relation

$$\frac{1}{\nu B}\,\nu(1_B \cdot L^2(t)^* k) = \mu^B k$$

(see the proof of the previous theorem).

21.3 Corollary: Let μ' be a probability measure on X such that for all $k \in \mathcal{K}(X)$ we have

$$(L^2(t)^* k)_{ess}(y_o) = \mu' k .$$

Then, the conditional distribution of $x \in (X, \mu)$, given $t(x) = y_o$, is defined and equal to μ'.

Summarizing these results, we can say that except for the possibility of defectivenes in theorem 21.2, existence of a conditional distribution is equivalent to essential continuity of the conditional expectations of $\mathcal{K}(X)$-functions.

Theorem 21.1 can be strengthened considerably; it is valid for $\mathcal{C}_b(X)$-functions, and even weaker continuity assumptions are sufficient:

21.4 Theorem. Suppose that the conditional distribution μ^{y_o} is defined. Let $f: X \to \mathbb{R}$ be a bounded, μ-integrable function, continuous at μ^{y_o}-almost all points. Then, $L^2(t)^* f$ is essentially continuous at y_o with

$$(L^2(t)^* f)_{ess}(y_o) = \mu^{y_o} f .$$

Proof: For μ^{y_o}-almost all x we have (lemma 7.2, page 39)

$$f_{ess}(x) = f(x)$$

and so (by theorem 20.2, page 178)

$$\underline{f}(x) = f(x) = \overline{f}(x) .$$

The functions $\underline{f}$ and $\overline{f}$ are μ^{y_o}-integrable, since they are bounded and semicontinuous. Thus f is μ^{y_o}-integrable and

$$\mu^{y_o}\underline{f} = \mu^{y_o}f = \mu^{y_o}\overline{f} .$$

Now let $\varepsilon > 0$ be given. By theorem A 15 (page 345 ; the semicontinuity of the function $\mu' \to \mu' f$ when f is semicontinuous) we can choose a neighbourhood V of y_o such that

$$\mu^{B}\underline{f} > \mu^{y_o}\underline{f} - \varepsilon$$

and

$$\mu^{B}\overline{f} < \mu^{y_o}\overline{f} + \varepsilon$$

for $B \subseteq V$. Then

$$\mu^{B}f - \varepsilon \leqslant \mu^{B}\overline{f} - \varepsilon < \mu^{y_o}\overline{f} = \mu^{y_o}f = \mu^{y_o}\underline{f} < \mu^{B}\underline{f} + \varepsilon \leqslant \mu^{B}f + \varepsilon ,$$

and so

$$|\mu^{B}f - \mu^{y_o}f| < \varepsilon .$$

This argument shows that $\mu^{B}f$ converges to $\mu^{y_o}f$ for $B \to y_o$. The theorem follows immediately from the identity

$$\mu^B f = \frac{1}{\nu B} \nu(1_B \cdot L^2(t)^* f) ,$$

proved as in the proof of theorem 21.1 (page 185; the boundedness of f implies $f \in L^2(\mu)$).

The continuity assumptions about f may as well be replaced by assumptions about essential continuity:

21.5 Corollary: Suppose that μ^{y_o} is defined. Let $f : X \to \mathbb{R}$ be a μ-integrable, μ-essentially bounded function, μ-essentially continuous at μ^{y_o}-almost all points. Then, the conditional expectation $L^2(t)^* f$ of f is ν-essentially continuous at y_o with

$$(L^2(t)^* f)_{ess}(y_o) = \mu^{y_o}[f_{ess}(x)]_x .$$

Proof: Just apply theorem 21.4 to a representative f_o, satisfying the conditions of theorem 20.3 (page 182).

The results of this section can be characterized as local versions of the equation $(L^2(t)^* f)(y) = \mu^y f$, based on the idea that the left side should be interpreted as an essential value. An immediate global interpretation of the equation is this: In case μ^y is defined for almost all y, the equation is valid for almost all y, independently of the choice of representative $L^2(t)^* f$. This result is valid for

arbitrary $L^2(\mu)$-functions f, and we shall prove it in section 24 (theorem 24.5, page 202). A special version of this result follows immediately from the results of this section:

21.6 Theorem. Suppose that μ^y is defined for almost all y. Then, for any bounded, continuous function f the almost everywhere defined function

$$[\mu^y f]_y$$

is a representative for the conditional expectation of f.

Proof: Let g denote an arbitrary representative of $L^2(t)^* f$. By theorem 20.3 (page 182) and theorem 21.4 above, we have for ν-almost all y

$$g(y) = g_{ess}(y) = \mu^y f .$$

CHAPTER VI : GLOBAL PROPERTIES OF CONDITIONAL DISTRIBUTIONS

22. CONTINUITY OF THE CONDITIONAL DISTRIBUTIONS.

By the results of section 21, continuity properties of essential values can be transferred into continuity properties of conditional distributions. Let

$$t: (X, \mu) \to (Y, \nu)$$

be given.

22.1 Theorem. Let C_o be the set of points y such that the conditional distribution μ^y is defined. Then the mapping

$$\begin{aligned} y &\to \mu^y \\ C_o &\to \mathcal{P}(X) \end{aligned}$$

is continuous.

Proof: For $k \in \mathcal{K}(X)$, the mapping

$$\begin{aligned} y &\to \mu^y k = (L^2(t)^* k)_{ess}(y) \\ C_o &\to \mathbb{R} \end{aligned}$$

is continuous (theorem 20.3, page 182).

22.2 Corollary: Suppose that μ^y is defined for all $y \in Y$. Then the mapping

$$y \to \mu^y$$
$$Y \to \mathcal{P}(X)$$

is continuous.

22.3 Corollary: Suppose that μ^y is defined for almost all y. Then the (almost everywhere defined) mapping

$$y \to \mu^y$$

is measurable (see the appendix, page 347).

23. EVERYWHERE DEFINED CONDITIONAL DISTRIBUTIONS.

As we shall see in section 24, locally defined conditional distributions have all the global properties of classical conditional distributions , as soon as they are almost everywhere defined.

However, the global results and their proofs are considerably simpler in case of everywhere defined conditional distributions. As we have seen in section 17, the class of such cases is by no means exclusive. It therefore seems reasonable to treat this simple case for itself.

We shall discuss the properties of the family $(\mu^y | y \in Y)$ of conditional distributions, in particular those properties characterizing the family. Throughout this section, let

$$t: (X, \mu) \to (Y, \nu)$$

be given, and assume that $\operatorname{supp} \nu = Y$.

The adjointness equation.

23.1 Theorem. For a continuous mapping

$$y \to \mu_y$$
$$Y \to \mathcal{P}(X)$$

the following conditions are equivalent:

(1) The conditional distribution μ^y is defined and equal to μ_y for all y.

(2) For all $f \in \mathcal{C}_b(X)$ and $g \in \mathcal{C}_b(Y)$ we have

$$\nu[g(y)\mu_y f]_y = \mu[g(t(x))f(x)]_x .$$

(3) For all $f \in \mathcal{C}_b(X)$ the function $[\mu_y f]_y$ is a representative of the conditional expectation of f.

Proof: (2) and (3) are obviously equivalent: Writing the equation (2) by inner products we get

$$(g|[\mu_y f]_y)_\nu = (L^2(t)g|f)_\mu = (g|L^2(t)^* f)_\nu .$$

For fixed f, the validity of this equation for all g in the dense subspace $\mathcal{C}_b(Y)$ of $L^2(\nu)$ implies that

$$[\mu_y f]_y = L^2(t)^* f \qquad (\text{in } L^2(\nu)).$$

Conversely, (2) is obviously valid when $[\mu_y f]_y$ is a representative of the conditional expectation.

(1) $\Longrightarrow$ (3) was proved in section 21 (theorem 21.6, page 190).

(3) $\Longrightarrow$ (1) follows from corollary 21.3 (page 186): For $k \in \mathcal{K}(X)$, the function $[\mu_y k]_y$ is continuous, and so it admits its own value as essential value, i.e.

$$(L^2(t)^* k)_{ess}(y) = \mu_y k \quad \text{for all } y .$$

By corollary 21.3 we have then $\mu^y = \mu_y$.

Remark. The above theorem is valid (and the proof is unchanged) when $\mathcal{C}_b(X)$ and $\mathcal{C}_b(Y)$ in (2) and (3) are replaced by $\mathcal{K}(X)$ and $\mathcal{K}(Y)$.

The equation in (2) will be called the adjointness equation, for obvious reasons (it states, when written on L^2-form, that $f \to [\mu_y f]_y$ is the adjoint mapping to $g \to g \circ t$).

Representation of μ as the mixture with respect to ν of the conditional distributions μ^y. It follows immediately from theorem 23.1 (inserting $g := 1_Y$ and $f := k \in \mathcal{K}(X)$) that we have

23.2 Corollary: If μ^y is defined for all y, then μ is the mixture of the conditional distributions μ^y with respect to ν :

$$\mu = \nu[\mu^y]_y \quad \text{(see the appendix, page 351).}$$

The decomposition criterion. In case t is continuous, a very simple criterion exists:

23.3 Theorem. Suppose that t is continuous (and supp $\nu = Y$). For a continuous mapping

$$y \to \mu_y$$
$$Y \to \mathcal{P}(X)$$

the following two conditions are equivalent:

(1) The conditional distribution μ^y is defined and equal to μ_y for all y.

(2) μ is the mixture of the measures μ_y with respect to ν, and for all $y \in Y$ we have

$$t(\mu_y) = \varepsilon_y .$$

Notice that (2) is equivalent to

(2)' $(\nu, (\mu_y | y \in Y))$ is a decomposition of μ with respect to t, as defined in section 7 (page 36-37).

Proof: First suppose that μ^y is defined and equal to μ_y for all y. By the corollary above, μ is then the mixture

of the measures μ_y , and the equation $t(\mu_y) = \varepsilon_y$ follows from the definition of μ^y and the continuity of t : For $h \in \mathcal{K}(Y)$ we have, by lemma 7.2 and corollary A 16 (page 346)

$$t(\mu_y)h = t(\mu^y)h = \mu^y(h \circ t) = \lim_{B \to y} \mu^B(h \circ t)$$

$$= \lim_{B \to y} \frac{1}{\nu B} \nu(1_B \cdot h) = h(y) .$$

The converse statement (2) $\Longrightarrow$ (1) follows immediately from the theorem about conditioning on a decomposed measure space (theorem 7.1, page 38; put $\lambda := \mu$, $\lambda' := \nu$, $\lambda_y := \mu_y$ and $f := 1_X$) .

24. ALMOST EVERYWHERE DEFINED CONDITIONAL DISTRIBUTIONS.

In this and the following sections we shall frequently, without explicit reference, apply the theorems A 20 and A 22 (see page 351-352) about preservation of integrability under transformations and mixings.

Throughout this section, let

$$t: (X, \mu) \to (Y, \nu)$$

be given.

The adjointness equation. First notice that if the conditional distribution μ^y is defined for almost all y, then the adjointness equation

$$\nu[g(y)\mu^y f]_y = \mu[g(t(x))f(x)]_x$$

is valid for $f \in \mathcal{L}_b(X)$, $g \in \mathcal{L}_b(Y)$. This follows immediately from theorem 21.6 (page 190), as in case of everywhere defined conditional distributions. Also the representation of μ as a mixture of the conditional distributions is easily generalized:

24.1 Theorem. If μ^y is defined for almost all y,

then

$$\mu = \nu[\mu^y]_y \ .$$

Proof: The mapping $y \to \mu^y$ is measurable (corollary 22.3, page 192), and the identity

$$\nu[\mu^y k]_y = \mu k \qquad \text{for } k \in \mathcal{K}(X)$$

follows immediately when $g := 1_Y$ and $f := k$ are inserted in the adjointness equation.

The adjointness equation can also be applied as a criterion in the "almost everywhere-case" :

24.2 Theorem. Let $C \subseteq Y$ be a set with $\nu C = 1$, and consider a continuous mapping

$$\begin{aligned} y &\to \mu_y \\ C &\to \mathcal{P}(X) \ . \end{aligned}$$

Suppose that for all $k \in \mathcal{K}(X)$, $h \in \mathcal{K}(Y)$, we have

$$\nu[h(y)\mu_y k]_y = \mu[h(t(x))k(x)]_x \ .$$

Then, the conditional distribution μ^y is defined for all $y \in C \cap \operatorname{supp} \nu$ by $\mu^y = \mu_y$.

Proof: Let $y_o \in C \cap \text{supp}\, \nu$ be given. For $k \in \mathcal{K}(X)$, the function $\nu[\mu_y k]_y$ (defined almost everywhere) is, obviously, essentially continuous at y_o with the essential value $\mu_{y_o} k$. Thus, by corollary 21.3 (page 186) the conditional distribution of x given $t(x) = y_o$ is defined and equal to μ_{y_o}.

It follows from this theorem that the mapping $y \to \mu^y$ is maximal, in the following sense:

24.3 Corollary: Let C_o denote the set of points y such that μ^y is defined, and suppose that $\nu C_o = 1$. Let C be a set such that

$$C_o \subseteq C \subseteq \text{supp}\, \nu ,$$

and let

$$y \to \mu_y$$
$$C \to \mathcal{P}(X)$$

be a continuous mapping with the property that $\mu_y = \mu^y$ for $y \in C_o$ (i.e. $y \to \mu_y$ is a continuous extension of the conditional distributions to a bigger domain, contained in the support). Then, $C = C_o$.

Hence, the points of $\operatorname{supp}\nu$ where μ^y is undefined give rise to "proper singularities". Notice, however, that this result is only valid in case of almost everywhere defined conditional distributions.

Representation of the conditional expectation operator.

In order to prove the almost everywhere-version of the equation $(L^2(t)^* f)(y) = \mu^y f$, we shall need the following lemma:

24.4 Lemma: Let $\mu = \nu[\mu_y]_y$ be a mixture (as defined in the appendix, page 351), where μ, ν and μ_y are probability measures ($\nu \in \mathcal{P}(Y)$, $\mu_y, \mu \in \mathcal{P}(X)$). Then, for any $f \in L^2(\mu)$, the function $[\mu_y f]_y$ (defined almost everywhere according to theorem A 22) belongs to $L^2(\nu)$. The so defined mapping

$$f \to [\mu_y f]_y$$
$$L^2(\mu) \to L^2(\nu)$$

is a linear operator with norm ≤ 1.

Proof: According to a wellknown convexity-inequality we have

$$(\mu_y f)^2 \leq \mu_y(f^2)$$

(both sides are defined for almost all y). Hence, the function

$[\mu_y f]_y$ belongs to $L^2(\nu)$, its square being measurable and dominated by an integrable function. Further, we have

$$\|[\mu_y f]_y\|_2^2 = \nu[(\mu_y f)^2]_y \leq \nu[\mu_y(f^2)]_y$$

$$= \mu(f^2) = \|f\|_2^2 .$$

24.5 Theorem. Suppose that the conditional distribution μ^y is defined for almost all y. Then, for all f in $L^2(\mu)$, the almost everywhere defined function

$$[\mu^y f]_y$$

is a representative for the conditional expectation of f.

Proof: The bounded linear operators

$$f \to [\mu^y f]_y$$

and

$$f \to L^2(t)^* f$$

(both mapping $L^2(\mu)$ into $L^2(\nu)$) coincide, since they coincide on the dense subspace $\mathcal{C}_b(X)$ (corollary 21.6, page 190).

24.6 Corollary: The adjointness equation

$$\nu[g(y)\mu^y f]_y = \mu[g(t(x))f(x)]_x$$

or $$(g|[\mu^y f]_y)_\nu = (g\circ t|f)_\mu$$

is valid for arbitrary L^2- functions $f \in L^2(\mu)$ and $g \in L^2(\nu)$, when the conditional distributions are defined almost everywhere.

The decomposition criterion. We have proved that the representation of μ as the mixture with respect to ν of the conditional distributions holds, as soon as the conditional distributions are defined almost everywhere. The other half part of the decomposition criterion is valid in the following sense:

24.7 Theorem. Suppose that μ^y is defined for almost all y. Then, for almost all y, t is μ^y-measurable and

$$t(\mu^y) = \varepsilon_y .$$

Remark: Notice, that $(\nu, (\mu^y | y \in Y))$ need not be a decomposition in the strict sense of section 7, even if t

is continuous.

Proof: Consider the transformation

$$(1_X , t) : X \to X \times Y$$

(where $1_X : X \to X$ denotes the identity). Put

$$\gamma := (1_X , t)\mu .$$

According to corollary 5.3 (page 31) the projection

$$p: (X \times Y , \gamma) \to (Y, \nu)$$

has the conditional distributions

$$\gamma^y = \mu^y \otimes \varepsilon_y$$

(defined whenever μ^y is defined, i.e. for almost all y).

The relation $y = t(x)$ is satisfied for γ-almost all (x,y) (it is satisfied for all (x,y) in the image of $(1_X , t)$, and the image has γ-measure 1, according to theorem A 20).

Now, γ is the mixture of the measures $\gamma^y = \mu^y \otimes \varepsilon_y$ with respect to ν. Thus (by theorem A 22) we have for ν-almost all y_o

$$(\mu^{y_o} \otimes \varepsilon_{y_o})\{(x,y) | t(x) = y\} = 1 .$$

Since we have $y = y_o$ for $(\mu^{y_o} \otimes \varepsilon_{y_o})$-almost all (x,y), we conclude that (still for ν-almost all y_o)

$$(\mu^{y_o} \otimes \varepsilon_{y_o})\{(x,y_o) | t(x) = y_o\} = 1 ,$$

$$\text{or} \quad \mu^{y_o}(t^{-1}(y_o)) = 1 .$$

Hence, for almost all y_o, t is μ^{y_o}-measurable (namely, equivalent to a constant mapping) with

$$t(\mu^{y_o}) = \varepsilon_{y_o} .$$

This is the global version of the decomposition criterion. The pointwise version (theorem 23.3, page 196) can be sharpened as follows:

24.8 Theorem. For $y_o \in Y$, suppose that μ^{y_o} is defined and that t is continuous at μ^{y_o}-almost all points. Then, t is μ^{y_o}-measurable with $t(\mu^{y_o}) = \varepsilon_{y_o}$.

Proof: Obviously, t is μ^{y_o}-measurable. For $h \in \mathcal{K}(Y)$, the function $h \circ t$ is continuous at μ^{y_o}-almost all points,

and therefore we have, by theorem 21.4 (page 187)

$$t(\mu^{y_o})h = \mu^{y_o}(h\circ t) = (L^2(t)^*(h\circ t))_{ess}(y_o)$$

$$= (L^2(t)^*L^2(t)h)_{ess}(y_o) = h_{ess}(y_o) = h(y_o) .$$

Finally, we shall prove that the decomposition criterion is also applicable as a <u>criterion</u> in the "almost everywhere-case" :

<u>24.9 Theorem</u>. Let C be a subset of Y with $\nu C = 1$ and let

$$y \to \mu_y$$
$$C \to \mathcal{P}(X)$$

be a continuous mapping. Suppose that we have

(1) For almost all $y \in C$, t is μ_y-measurable with $t(\mu_y) = \varepsilon_y$.

(2) $\mu = \nu[\mu_y]_y$.

Then, the conditional distribution μ^y is defined for <u>all</u> $y \in C \cap \operatorname{supp}\nu$ by $\mu^y = \mu_y$.

Proof: For almost all y we have

$$\mu_y(t^{-1}(y)) = 1 ,$$

and so, for $k \in \mathcal{K}(X)$ and $h \in \mathcal{K}(Y)$,

$$\mu_y((h\circ t)k) = h(y)\, \mu_y k .$$

Applying this, we get

$$\nu[h(y)\, \mu_y k]_y = \nu[\, \mu_y((h\circ t)k)]_y = \mu((h\circ t)k)$$

$$= \mu[h(t(x))k(x)]_x .$$

Thus, the adjointness equation is satisfied, and the theorem follows immediately from theorem 24.2 (page 199).

25. STOCHASTIC INDEPENDENCE.

Let (X, μ) be a probability field and let

$$t_i : X \to Y_i \quad , \qquad i \in I$$

be a family of measurable transformations. Suppose that the mapping

$$(t_i | i \in I) : X \to \prod_{i \in I} Y_i$$

is measurable (this follows from the measurability of the single transformations t_i in case I is finite or denumerable). We say that the stochastic variables

$$y_i = t_i(x) \quad , \qquad x \in (X, \mu)$$

(or the transformations t_i) are <u>stochastically independent</u>, if their joint distribution is a product measure, i.e. if

$$(t_i | i \in I)\mu = \bigotimes_{i \in I} \nu_i \quad , \qquad \nu_i \in \mathcal{P}(Y_i);$$

(in case I is infinite, it is assumed that the spaces Y_i are compact; a product measure $\otimes \nu_i$ is then defined in the obvious manner by its consistent family).

We shall formulate criterions for independence of two stochastic variables in terms of conditional distributions and conditional

expectations.

> Definition: Two subspaces of a Hilbert space are said to be geometrically orthogonal, if their orthogonal projections commute.

Loosely speaking, geometric orthogonality of two subspaces means that they are orthogonal, except that they may have a nontrivial intersection.

Examples: Two planes in $\mathbb{R}^3$ are geometrically orthogonal, if they are orthogonal in the usual sense (that is why we use the term "geometrical". An alternative proposal is "conditional orthogonality", but that may be confusing in this particular connection). Statisticians will recognize the concept of geometrical orthogonality from two way variance analysis, where the spaces corresponding to row- and column-homogenity are geometrically orthogonal.

A purely geometrical description of the concept is this: Two subspaces U and V are geometrically orthogonal if and only if the two subspaces

$$U \cap (U \cap V)^{\perp} \quad \text{and} \quad V \cap (U \cap V)^{\perp}$$

are orthogonal (in the usual sense). Thus, U and V should

be orthogonal relatively to the complement of their intersection .

25.1 Theorem. For a diagram

$$\begin{array}{ccc} (X,\mu) & \xrightarrow{\ t\ } & (Y,\nu) \\ {\scriptstyle s}\downarrow & & \\ (Z,\pi) & & \end{array}$$

of probability fields and homomorphisms, the following conditions are equivalent:

(1) The stochastic variables

$$\left.\begin{array}{rl} & y = t(x) \\ \text{and} & z = s(x) \end{array}\right\} \qquad x \in (X,\mu)$$

are stochastically independent.

(2) For all $y_o \in \operatorname{supp} \nu$ the conditional distribution π^{y_o} of $z = s(x)$, given $t(x) = y_o$, is defined and equal to π.

(3) The subspaces

$$\begin{array}{rcl} & V_t := & \{g \circ t \mid g \in L^2(\nu)\} \\ \text{and} & V_s := & \{h \circ s \mid h \in L^2(\pi)\} \end{array}$$

are geometrically orthogonal with intersection

$$V_t \cap V_s = V := \{c \cdot 1_X \mid c \in \mathbb{R}\}$$

(= the line of constant functions).

Proof:

(1) $\Rightarrow$ (3) : Let E^t, E^s and E denote the orthogonal projections onto V_t, V_s and V, respectively. Thus, by theorem 19.1 (1) (page 168)

$$E^t = L^2(t)L^2(t)^*$$

$$E^s = L^2(s)L^2(s)^*$$

and

$$Ef = \mu f \cdot 1_X .$$

For two functions f_1 and f_2 from $L^2(\mu)$ it follows immediately from the independence assumption (1) that

$$(E^s f_1 | E^t f_2)_\mu = \mu\big(((L^2(s)^* f_1) \circ s) \cdot ((L^2(t)^* f_2) \circ t)\big)$$

$$= \mu((L^2(s)^* f_1) \circ s) \cdot \mu((L^2(t)^* f_2) \circ t) = (Ef_1 | Ef_2)_\mu .$$

Thus, by the properties of orthogonal projections we have

$$(E^tE^sf_1|f_2)_\mu = (E^sf_1|E^tf_2)_\mu = (Ef_1|Ef_2)_\mu = (Ef_1|f_2)_\mu .$$

We conclude that

$$E^tE^s = E$$

and (similarly, or by taking adjoints on both sides above)

$$E^sE^t = E .$$

This proves (3) .

(3)$\Longrightarrow$(2) : Suppose that (3) is satisfied. For $B \subseteq Y$ and $h \in \mathcal{K}(Z)$ we have then

$$\pi^B h = \mu^B(h\circ s) = \frac{1}{\nu B}(1_B\circ t|h\circ s)_\mu$$

$$= \frac{1}{\nu B}(E^t(1_B\circ t)|E^s(h\circ s))_\mu = \frac{1}{\nu B}(1_B\circ t|E^tE^s(h\circ s))_\mu$$

$$= \frac{1}{\nu B}(1_B\circ t|E(h\circ s))_\mu = \frac{1}{\nu B}\nu(1_B)\cdot\pi(h) = \pi h .$$

Thus, for $B \to y_o \in \operatorname{supp}\nu$,

$$\pi^{y_o} = \lim \pi^B = \pi .$$

(2)$\Longrightarrow$(1) : Assuming (2) , by theorem 5.2 (page 28) the conditional distribution of $(y,z) = (t(x),s(x))$, given $t(x) = y_o$, is defined and equal to $\varepsilon_{y_o} \otimes \pi$ for $y_o \in \operatorname{supp}\nu$.

The (unconditioned) distribution of (y,z) equals the mixture

$$\nu[\varepsilon_y \otimes \pi]_y = \nu \otimes \pi.$$

Thus, y and z are independent.

Remark: Similar results for finitely or infinitely many stochastic variables $y_i = t_i(x)$ are easily deduced. Just apply that the y_i's are independent if and only if for any two disjoint subsets I_1 and I_2 of I the two variables $(y_i | i \in I_1)$ and $(y_i | i \in I_2)$ are independent.

26. RESULTS RELATED TO SUCCESIVE CONDITIONING.

Suppose we have a diagram

$$(X, \mu) \xrightarrow{t} (Y, \nu) \xrightarrow{s} (Z, \pi)$$

of finite probability fields and homomorphisms. Consider the following conditional distributions:

μ^y : the conditional distribution of $x \in (X, \mu)$, given $t(x) = y$.

μ^z : the conditional distribution of $x \in (X, \mu)$, given $s(t(x)) = z$.

ν^z : the conditional distribution of $y \in (Y, \nu)$, given $s(y) = z$.

$(\mu^z)^y$: the conditional distribution of the "conditioned variable" $x \in (X, \mu^z)$, given $s(x) = y$.

These conditional distributions are related to each other in the following manner:

(1) $t(\mu^z) = \nu^z$ (the conditional distribution of y, given z, equals the conditional distribution of $t(x)$, given z).

(2) $\mu^z = \nu^z[\mu^y]_y$ (the conditional distribution of x, given z, is the mixture of the conditional distributions given y, with respect to the conditional distribution of y, given z).

(3) $(\mu^z)^y = \mu^y$ for $z = s(y)$ (the conditional distribution of x, given z and <u>then</u> y, can be computed immediately by conditioning on the specification of y).

The formulae (1), (2) and (3) are valid in the sense that the right side is defined if and only if the left side is defined (the left side of (3) is undefined for $z \neq s(y)$).

We shall generalize these elementary results to the case of arbitrary probability fields. Thus, throughout this section, we study a diagram

$$(X, \mu) \xrightarrow{t} (Y, \nu) \xrightarrow{s} (Z, \pi)$$

of probability fields and homomorphisms.

We shall prove global as well as pointwise versions of the formulae. Unfortunately, various degrees of regularity conditions give rise to particular results; it is hard to decide, at the present stage of the theory, which results to regard as

interesting and which to throw out. Emphasis should be put on the corollaries 26.2, 26.7, 26.10 and 26.13 which give the main results in the case of everywhere defined conditional distributions.

The formula $t(\mu^z) = \nu^z$.

26.1 Theorem. Suppose for $z_o \in Z$ that

(1) μ^{z_o} is defined

(2) t is continuous at μ^{z_o}-almost all points.

Then, the conditional distribution ν^{z_o} is also defined and given by

$$\nu^{z_o} = t(\mu^{z_o}) .$$

Notice that this theorem yields theorem 24.8 (page 205) as a special case (namely, for $Z := Y$ and $s := 1_Y : Y \to Y$).

Proof: The transformed measure $t(\mu^{z_o})$ is obviously defined (t is μ^{z_o}-measurable, being continuous μ^{z_o}-almost everywhere). According to corollary 21.3 (page 186) it suffices to prove that for h in $\mathcal{K}(Y)$ we have

$$(L^2(s)^* h)_{ess}(z_o) = (t\mu^{z_o})h .$$

But from theorem 21.4 (page 187) it follows immediately, since $h \circ t$ is bounded and continuous at μ^{z_o}-almost all points, that

$$(L^2(s \circ t)^*(h \circ t))_{ess}(z_o) = \mu^{z_o}(h \circ t) = (t\mu^{z_o})h ,$$

and the left side of this equation equals $(L^2(s)^* h)_{ess}(z_o)$, since

$$L^2(s \circ t)^*(h \circ t) = L^2(s)^* L^2(t)^*(h \circ t)$$

$$= L^2(s)^* L^2(t)^* L^2(t) h = L^2(s)^* h .$$

26.2 Corollary. If μ^z is defined for all z and t is continuous, then ν^z is defined for all z and

$$\nu^z = t(\mu^z) .$$

Remark: The corollary can also be regarded as a special case of theorem 5.1 (page 27) .

Theorem 26.1 yields the following almost everywhere-result:

26.3 Corollary: Suppose that μ^z is defined for almost all z and that t is continuous at μ-almost all points. Then, for almost all z the conditional distribution ν^z is defined and equal to $t(\mu^z)$.

The proof is immediate.

This result is not valid without the continuity condition on t (counterexample: Put $Z := X$, $s := t^{-1}$, where t is injective and measurable but not having continuous restriction to any set of measure 1). But the result becomes valid again, if we assume in addition that ν^z is defined; this is the only non-trivial result of this section:

> 26.4 Theorem. Suppose that μ^z and ν^z are defined for almost all z. Then, for almost all z, the transformed measure $t(\mu^z)$ is defined and equal to ν^z.

Notice that theorem 24.7 (page 203) comes out as a special case of this theorem.

Proof: Let (K_n) be an increasing sequence of compact sets with $\mu(UK_n) = 1$, such that for each n the restriction of t to K_n is continuous. Then (by theorem A 22, page 352, and theorem 24.1, page 198) we have

$$\mu^z(UK_n) = 1 \quad \text{for almost all } z .$$

Hence, t is μ^z-measurable for almost all z.

Now, suppose we are able to prove that the (almost everywhere

defined) mapping $z \to t(\mu^z)$ is measurable. In that case, the theorem can be proved as follows:

For $\varepsilon > 0$, let G be a compact subset of Z with $\pi G > 1 - \varepsilon$ such that the mappings

$$z \to t(\mu^z)$$
$$\text{and} \qquad z \to \nu^z$$

are continuous, when restricted to G. For $h \in \mathcal{K}(Y)$ put

$$G_h := \{z \in G \mid \nu^z h = (t\mu^z)h\} .$$

Then, G_h is compact. For almost all z we have, by theorem 24.5 (page 202)

$$\nu^z h = (L^2(s)^* h)(z)$$

and

$$t(\mu^z)h = \mu^z(h \circ t) = L^2(s \circ t)^*(h \circ t)(z)$$

$$= (L^2(s)^* h)(z)$$

(these equations are valid π-almost sure for arbitrary representatives for $L^2(s)^* h$ and $L^2(s \circ t)^*(h \circ t)$). Thus

$$\nu^z h = (t\mu^z)h \quad \text{for almost all } z .$$

For any finite number of functions $h_1, \dots, h_n \in \mathcal{K}(Y)$ we have then

$$\pi(G_{h_1} \cap \dots \cap G_{h_n}) = \pi G ;$$

Hence, by theorem A 14 (page 345) we have

$$\pi\{z \mid \nu^z = t(\mu^z)\} \geq \pi\Big(\bigcap_{h \in \mathcal{K}(Y)} G_h\Big) = \pi G \geq 1 - \varepsilon .$$

This being true for all $\varepsilon > 0$ we conclude that $\nu^z = t(\mu^z)$ for almost all z .

It remains to prove that the mapping $z \to t(\mu^z)$ is measurable; we shall need a lemma:

26.5 Lemma: For $\mu \in \mathcal{P}(X)$, let

$$x \to \nu_x^{(n)} , \qquad n \in \mathbb{N}$$

be a sequence of measurable mappings

$$X \to \mathcal{M}_b(Y) .$$

Let

$$x \to \nu_x$$
$$X \to \mathcal{M}_b(Y)$$

be a mapping such that for almost all x

$$\| \nu_x^{(n)} - \nu_x \|_\infty \to 0$$

(i.e. $\nu_x^{(n)} h \to \nu_x h$ uniformely for $\|h\|_\infty \leq 1$).

Then, the mapping $x \to \nu_x$ is measurable.

Proof: Let $\varepsilon > 0$ be given. For each natural number i, choose n_i such that

$$\mu\{x| \ \| \nu_x^{(n_i)} - \nu_x \|_\infty > \tfrac{1}{i} \} < \varepsilon \cdot 2^{-i-1} .$$

Further, choose compact sets

$$K_i \subseteq \{x| \ \| \nu_x^{(n_i)} - \nu_x \|_\infty \leq \tfrac{1}{i} \}$$

such that

$$\mu K_i > \mu\{x| \ \| \nu_x^{(n_i)} - \nu_x \|_\infty \leq \tfrac{1}{i} \} - \varepsilon \cdot 2^{-i-1}$$

$$> 1 - \varepsilon \cdot 2^{-i}$$

with the property that the restriction of the mapping $x \to \nu_x^{(n_i)}$ to K_i is continuous. Then, $\| \nu_x^{(n_i)} - \nu_x \|_\infty$ converges uniformely to 0 on the compact set $K := \cap K_i$ with $\mu K > 1 - \varepsilon$. Obviously, the limit function $x \to \lim_i \nu_x^{(n_i)}$

is then continuous, when restricted to K. We conclude that the mapping

$$x \to \nu_x \quad (= \lim \nu_x^{(n)} \text{ for almost all } x)$$

is measurable.

Now (returning to the proof of theorem 26.4) let K be a compact set such that the restriction of t to K is continuous. We can choose a decreasing sequence (k_n) of $\mathcal{K}(X)$-functions such that

$$k_n(x) \to 1_K(x) \qquad \mu\text{-almost everywhere.}$$

For almost all z we have then

$$k_n(x) \to 1_K(x) \qquad \mu^z\text{- almost everywhere,}$$

and so

$$\| k_n \cdot \mu^z - 1_K \cdot \mu^z \|_\infty = \mu^z |k_n - 1_K| \to 0$$

for almost all z.

The mapping $z \to k_n \cdot \mu^z$ is measurable (as composed of a continuous mapping and a measurable mapping), and we conclude from the lemma that the mapping

$$z \to 1_K \cdot \mu^z$$

is measurable.

Now, put

$$P_o(X) := \{\lambda \in \mathcal{M}_b(X) \mid \|\lambda\|_\infty \leq 1\}$$

and, similarly

$$P_o(K) := \{\xi \in \mathcal{M}_b(K) \mid \|\xi\|_\infty \leq 1\} .$$

Let $j:K \to X$ denote the imbedding. The continuous mapping

$$\xi \to j(\xi)$$

$$P_o(K) \to P_o(X)$$

is injective. Since $P_o(K)$ is compact, the mapping is a homeomorphism of $P_o(K)$ onto its image. Hence, the mapping

$$z \to (1_K \cdot \mu^z)$$

is also measurable, when regarded as a mapping into $P_o(K)$.

The restriction of t to K is continuous. Thus, the mapping

$$z \to t(1_K \cdot \mu^z)$$

is also measurable. Finally, letting K run through a sequence (K_n) with $\mu(\cup K_n) = 1$, we conclude (applying the lemma once more) that the mapping

$$z \to t(\mu^z)$$

is measurable. This completes the proof of theorem 26.4 .

The formula $\mu^z = \nu^z[\mu^y]_y$.

26.6 Theorem. Suppose for $z_o \in Z$ that

(1) ν^{z_o} is defined

(2) μ^y is defined for ν^{z_o}-almost all y .

Then, μ^{z_o} is defined and given by

$$\mu^{z_o} = \nu^{z_o}[\mu^y]_y .$$

Proof: The mixture $\nu^{z_o}[\mu^y]_y$ is welldefined, the mapping $y \to \mu^y$ being continuous on its domain. By corollary 21.3 (page 186) , it suffices to prove that

$$(L^2(s\circ t)^* k)_{ess}(z_o) = \nu^{z_o}[\mu^y k]_y$$

for all $k \in \mathcal{K}(X)$. But for ν^{z_o}-almost all y (namely, for μ^y defined) we have

$$\mu^y k = (L^2(t)^* k)_{ess}(y) ,$$

and corollary 21.5 (page 189), applied to the function $L^2(t)^* k$,

gives

$$(L^2(s\circ t)^*k)_{ess}(z_o) = (L^2(s)^*(L^2(t)^*k))_{ess}(z_o)$$

$$= \nu^z\circ[(L^2(t)^*k)_{ess}(y)]_y = \nu^z\circ[\mu^y k]_y \ .$$

26.7 Corollary: Suppose that ν^z is defined for all z and μ^y is defined for all y. Then, μ^z is defined for all z and given by

$$\mu^z = \nu^z[\mu^y]_y \ .$$

The almost everywhere-version looks like this:

26.8 Corollary. Suppose that ν^z is defined for almost all z and that μ^y is defined for almost all y. Then, for almost all z, μ^z is defined and

$$\mu^z = \nu^z[\mu^y]_y \ .$$

The proof is immediate.

Notice that the results 26.6, 26.7 and 26.8 are analogous to 26.1, 26.2 and 26.3, respectively. The deeper theorem 26.4 has no analogue, because the mapping $y \to \mu^y$ is continuous,

if it is defined (as opposed to what is the mapping t).

The formula $(\mu^z)^y = \mu^y$.

26.9 Theorem. For $z_o \in Z$, assume

(1) μ^{z_o} is defined

(1)' ν^{z_o} is defined

(2) t is continuous at μ^{z_o}-almost all points.

(3) μ^y is defined for ν^{z_o}-almost all y.

Then, for all $y \in \operatorname{supp} \nu^{z_o}$ such that μ^y is defined, the conditional distribution of $x \in (X, \mu^{z_o})$, given $t(x) = y$, is defined and given by

$$(\mu^{z_o})^y = \mu^y .$$

Remark: Any of the two conditions (1) and (1)' can be removed, since it follows from the remaining three conditions (by theorem 26.1 and 26.6).

Proof: We apply the decomposition criterion: Put

$$C := \{y \mid \mu^y \text{ is defined}\}.$$

The mapping

$$\begin{aligned} y &\to \mu^y \\ C &\to \mathcal{P}(X) \end{aligned}$$

is continuous, and by theorem 26.6 we have

$$\nu^{z_o}[\mu^y]_y = \mu^{z_o} .$$

According to theorem 24.8 (page 205) we have

$$t(\mu^y) = \varepsilon_y$$

for all points y with the properties that μ^y is defined and t is continuous at μ^y-almost all points (i.e. for ν^{z_o}-almost all y). Hence, the theorem follows immediately from theorem 24.9 (page 206).

26.10 Corollary: Suppose that t is continuous and the conditional distributions μ^z, ν^z and μ^y are everywhere defined. Then, for all $z_o \in Z$, the homomorphism

$$t: (X, \mu^{z_o}) \to (Y, \nu^{z_o})$$

has conditional distributions defined for all y in the

support of ν^{z_o} by

$$(\mu^{z_o})^y = \mu^y .$$

The almost everywhere-version requires no continuity assumptions about t :

26.11 Theorem. Suppose that the conditional distributions ν^z and μ^y are defined almost everywhere. Then, for almost all z, μ^z is defined (corollary 26.8) and (by theorem 26.4) t is a homomorphism

$$t: (X,\mu^z) \to (Y,\nu^z) .$$

Moreover, for almost all such z, the following statement is true: For ν^z-almost all y, the conditional distribution $(\mu^z)^y$ is defined and equal to μ^y.

Proof: We apply the decomposition criterion. For almost all z we have (corollary 26.8)

$$\mu^z = \nu^z[\mu^y]_y .$$

Moreover, by theorem 24.7 (page 203) we have

$$t(\mu^y) = \varepsilon_y \quad \text{for almost all } y .$$

Thus, for almost all z we have $t(\mu^y) = \varepsilon_y$ for ν^z-almost all y. These arguments show, that for almost all z the mapping

$$y \to \mu^y$$

$$\{y \mid \mu^y \text{ is defined}\} \to \mathcal{P}(X)$$

satisfies the conditions of the decomposition criterion (theorem 24.9) as conditional distributions for $t:(X,\mu^z) \to (Y,\nu^z)$, and this proves the theorem.

Finally, we shall prove a result about existence of μ^y when $(\mu^z)^y$ is defined almost everywhere:

26.12 Theorem. Suppose that

(1) the conditional distributions μ^z and ν^z are defined for almost all z.

(2) For almost all z, the conditional distributions $(\mu^z)^y$ are defined for ν^z-almost all y.

Let $C \subseteq Y$ be a set with $\nu C = 1$ such that $(\mu^z)^y$ is defined for all (y,z) with $y \in C$, $z = s(y)$, and such that

(3) the mapping

$$\begin{aligned} y &\to (\mu^{s(y)})^y \\ C &\to \mathcal{P}(X) \end{aligned}$$

is continuous.

Then, the conditional distribution μ^y is defined for all $y \in C$ and given by

$$\mu^y = (\mu^{s(y)})^y .$$

Proof: Again, the decomposition criterion is applied: For almost all $z \in C$ we have

$$\begin{cases} t \text{ is } \mu^z\text{-measurable with } t(\mu^z) = \nu^z & \text{(theorem 26.4, page 218)} \\ \mu^z = \nu^z[(\mu^z)^y]_y & \text{(theorem 24.1, page 198)} \\ t((\mu^z)^y) = \varepsilon_y \text{ for } \nu^z\text{-almost all } y & \text{(theorem 24.7, page 203)} \\ s(\nu^z) = \varepsilon_z & \text{(theorem 24.7, page 203).} \end{cases}$$

Let z be a point satisfying these conditions. Then, for ν^z-almost all y we have

$$s(y) = z \qquad (\text{since } s(\nu^z) = \varepsilon_z)$$

and so $(\mu^z)^y = (\mu^{s(y)})^y$ for ν^z-almost all y.

Thus, for ν^z-almost all y we have

$$t((\mu^{s(y)})^y) = t((\mu^z)^y) = \varepsilon_y$$

and $$\nu^z[(\mu^{s(y)})^y]_y = \nu^z[(\mu^z)^y]_y = \mu^z .$$

This being true for almost all z, we conclude that the mapping

$$\begin{aligned} y &\to (\mu^{s(y)})^y \\ C &\to \mathcal{P}(X) \end{aligned}$$

satisfies the conditions of theorem 24.9:

$$t((\mu^{s(y)})^y) = \varepsilon_y \text{ for } \nu\text{-almost all y}$$

and $$\nu[(\mu^{s(y)})^y]_y = \pi[\nu^z[(\mu^{s(y)})^y]_y]_z$$

$$= \pi[\mu^z]_z = \mu.$$

26.13 Corollary: Suppose that the conditional distributions μ^z and ν^z are everywhere defined. Further, suppose that the conditional distribution $(\mu^z)^y$ is defined for all z and y with $z = s(y)$, and that the mapping

$$y \rightarrow (\mu^{s(y)})^y$$

$$Y \rightarrow \mathcal{P}(X)$$

is continuous. Then, the conditional distribution μ^y is defined for all y by

$$\mu^y = (\mu^{s(y)})^y.$$

27. INTERCHANGING TWO CONDITIONING OPERATIONS.

Consider a diagram

$$\begin{array}{ccc} (X,\mu) & \xrightarrow{\ t\ } & (Y,\nu) \\ {\scriptstyle s}\downarrow & & \\ (Z,\xi) & & \end{array}$$

We shall apply the results of the previous section to the diagram

$$(X,\mu) \xrightarrow{(t,s)} (Y\times Z\ ,\gamma) \xrightarrow{\ p\ } (Y,\nu)$$

studied in section 5 (see page 28), where $\gamma := (t,s)\mu$ and p denotes the projection.

The following table "translates" the conditional distributions and the formulae studied in section 26 (see page 214-215) into the terminology of this particular case:

section 26	section 27
μ^y	$\mu^{(y,z)}$
μ^z	μ^y
ν^z	$\gamma^y = \varepsilon_y \otimes \xi^y$ (cfr. theorem 5.2, page 28)
$(\mu^z)^y$	$(\mu^y)^{(y,z)} = (\mu^y)^z$

(section 26)	(section 27)
(1) $t(\mu^z) = \nu^z$	$(t,s)\,\mu^y = \varepsilon_y \otimes \xi^y$ (or $s(\mu^y) = \xi^y$)
(2) $\mu^z = \nu^z[\mu^y]_y$	$\mu^y = \xi^y[\mu^{(y,z)}]_z$
(3) $(\mu^z)^y = \mu^y$	$(\mu^y)^z = \mu^{(y,z)}$

The formula corresponding to (1), written on the form $s(\mu^y) = \xi^y$, is known from section 5, and the results 26.1, 26.2, 26.3 and 26.4 are simply transferred into stronger versions of theorem 5.1 (page 27). For example,

theorem 26.1 gives

Suppose for $y_o \in Y$ that

(1) μ^{y_o} is defined

(2) s and t are continuous at μ^{y_o}-almost all points.

Then, the (derived) conditional distribution ξ^{y_o} is defined, and

$$(t,s)\,\mu^{y_o} = \varepsilon_{y_o} \otimes \xi^{y_o}$$

(and so $s(\mu^{y_o}) = \xi^{y_o}$).

Theorem 26.4 gives

> Suppose that μ^y and ξ^y are defined for almost all y. Then, for almost all y, s is μ^y-measurable with
>
> $$s(\mu^y) = \xi^y .$$

Thus, the two possible definitions of a "derived" conditional distribution (section 5) do agree, also in the almost everywhere-sense, when <u>both</u> are defined almost everywhere.

The formula (2) yields a formula for "cancellation of a condition". For example,

theorem 26.6 gives

> Suppose for $y_o \in Y$ that
>
> (1) ξ^{y_o} is defined
>
> (2) $\mu^{(y_o,z)}$ is defined for ξ^{y_o}-almost all z .
>
> Then, μ^{y_o} is defined and given by
>
> $$\mu^{y_o} = \xi^{y_o}[\mu^{(y_o,z)}]_z .$$

Thus, in order to cancel the conditioning upon z (changing $\mu^{(y_o,z)}$ to μ^{y_o}) we must mix according to the distribution of z under the remaining condition y_o.

The formula (3) yields a formula for succesive conditioning. For example,

theorem 26.11 gives

> Suppose that ξ^y and $\mu^{(y,z)}$ are defined almost everywhere. Then, for almost all y the conditional distribution μ^y is defined, and s is a homomorphism
>
> $$s: (X, \mu^y) \to (Z, \xi^y)$$
>
> with almost everywhere defined conditional distributions
>
> $$(\mu^y)^z = \mu^{(y,z)} .$$

Thus, in order to condition upon (y,z), we may condition first on y and then on z.

In particular, we conclude that the <u>order</u> of the two conditioning operations is irrelevant. This brings us to the point of this section:

27.1 Theorem. Suppose that the conditional distributions $\mu^{(y,z)}$ and the derived conditional distributions ξ^y and ν^z are defined almost everywhere. For almost all $(y,z) = (t(x),s(x))$, $x \in (X,\mu)$, we have then

$$(\mu^y)^z = \mu^{(y,z)} = (\mu^z)^y .$$

Proof: Just apply the above result (derived from theorem 26.11) twice, interchanging Y and Z in the second application.

Applying corollary 26.10 in stead of theorem 26.11, we get

27.2 Theorem. Suppose that s and t are continuous and that the conditional distributions

$$\mu^{(y,z)}$$
$$\mu^y \quad \text{(or, alternatively } \xi^y\text{)}$$
$$\text{and } \mu^z \quad \text{(or, alternatively } \nu^z\text{)}$$

are everywhere defined. Then, for all (y,z) such that y belongs to the support of ν^z and z belongs to the support of ξ^y, we have

$$(\mu^y)^z = \mu^{(y,z)} \quad (\mu^z)^y .$$

28. DECOMPOSITION OF A CONDITIONING PROBLEM.

In section 8 (page 49) and section 17 (page 145) we saw that particularly tedious conditioning problems may offer piecewise solutions only. For example, if we want to apply the classical methods outlined in section 8, it may be necessary to divide the domain X into smaller sets X_i for which "supplementary transformations" exist. This problem does not arise when the methods of chapter IV are applied, but still, problems of mixed dimension may force upon us a partitioning of X or Y.

It is a trivial matter to piece the solutions together, when a partitioning of Y is given. The more interesting problem of handling a partitioning of X can be solved by the methods indicated in section 27.

Partitionings of the codomain Y.

28.1 Theorem. Let $t: (X, \mu) \to (Y, \nu)$ be given. Let $(Y_i \mid i \in I)$ be a family of pairwise disjoint, open sets, such that

$$\nu Y_i > 0 \quad \text{for all } i$$

$$\text{and} \quad \sum \nu Y_i = 1$$

(then I is at most denumerable, and $Y \setminus (\cup Y_i)$ is a closed

null set). Consider the homomorphism

$$t:(X, \mu^{Y_i}) \rightarrow (Y, \nu^{Y_i})$$

(where $\mu^{Y_i} = \mu^{t^{-1}Y_i} = \frac{1}{\nu^{Y_i}} 1_{t^{-1}Y_i} \cdot \mu$).

For a point $y_o \in Y_i$ we have then:

The conditional distribution $(\mu^{Y_i})^{y_o}$ is defined if and only if the conditional distribution μ^{y_o} is defined, and in case of existence we have

$$(\mu^{Y_i})^{y_o} = \mu^{y_o} .$$

Proof: From a certain step in the passage to the limit $B \rightarrow y_o$ we have $B \subseteq Y_i$, and so

$$\mu^B = \mu^{Y_i \cap B} = (\mu^{Y_i})^B .$$

Thus, the two limits

$$\mu^{y_o} = \lim \mu^B$$

and $$(\mu^{Y_i})^{y_o} = \lim (\mu^{Y_i})^B$$

are defined at the same time and equal to each other when defined.

Partitionings of the domain X .

28.2 Theorem. Let $t:(X,\mu) \to (Y,\nu)$ be given and let $(X_i \mid i \in I)$ be a family of pairwise disjoint open sets in X such that

$$\mu X_i > 0 \quad \text{for all } i$$

and $$\sum \mu X_i = 1 .$$

For a point $y_o \in Y$, suppose

(1) $g_i(y_o) := \lim_{B \to y_o} \mu^B(X_i)$ exists for all i .

(2) $\sum g_i(y_o) = 1$.

(3) For all i with $g_i(y_o) > 0$ the conditional distribution $(\mu^{X_i})^{y_o}$ for the homomorphism

$$t: (X_i , \mu^{X_i}) \to (Y, t(\mu^{X_i}))$$

is defined.

Then, the conditional distribution μ^{y_o} is defined and given by

$$\mu^{y_o} = \sum_{\substack{i \in I \\ g_i(y_o) > 0}} g_i(y_o)(\mu^{X_i})^{y_o} .$$

Proof: We regard I as a locally compact and σ-compact (namely, denumerable) space (in its discrete topology). Let

$$s: (X, \mu) \to (I, \xi)$$

be the almost everywhere defined transformation

$$s(x) := \begin{cases} i \text{ for } x \in X_i \\ \text{undefined for } x \in X \setminus (\cup X_i). \end{cases}$$

The conditions (1) and (2) state, exactly, that the (derived) conditional distribution ξ^{y_o} is defined and has the density $[g_i(y_o)]_i$ with respect to counting measure: (1) states that the limit $\xi^{y_o} = \lim_{B \to y_o} \xi^B$ exists, since the density converges pointwise, and (2) states that this limit is not defective.

Consider the diagram

$$(X, \mu) \xrightarrow{(t,s)} (Y \times I, \gamma) \xrightarrow{p} (Y, \nu)$$

($\gamma := (t,s)(\mu)$, and p denotes the projection). By theorem 5.2 (page 28) the conditional distribution γ^{y_o} is defined and equal to $\varepsilon_{y_o} \otimes \xi^{y_o}$. Moreover, for all i with $g_i(y_o) > 0$ (i.e. for ξ-almost all i) the conditional distribution $\mu^{(y_o,i)}$ of $x \in (X, \mu)$, given $(t(x),s(x)) = (y_o,i)$, is defined and equal to $(\mu^{X_i})^{y_o}$: For $D \to (y_o,i)$, $\gamma D > 0$,

we have from a certain stage

$$D \subseteq Y \times \{i\}$$

(the fibre $Y \times \{i\}$ is open since I is discrete) and thus

$$D = B \times \{i\} , \qquad B \subseteq Y ;$$

by the condition (3) we have then

$$\lim_{D \to (y_o,i)} \mu^D = \lim_{B \to y_o} \mu^{B \times \{i\}} = \lim_{B \to y_o} \mu^{X_i \cap t^{-1}B} = (\mu^{X_i})^{y_o}.$$

We have now proved, that the diagram satisfies the conditions of theorem 26.6 (page 224). Hence, we conclude that the conditional distribution μ^{y_o} is defined and equal to

$$\gamma^{y_o}[\mu^{(y,i)}]_{(y,i)} = (\varepsilon_{y_o} \otimes \xi^{y_o})[\mu^{(y,i)}]_{(y,i)}$$

$$\xi^{y_o}[\mu^{(y_o,i)}]_i = \xi^{y_o}[(\mu^{X_i})^{y_o}]_i$$

$$= \sum_i g_i(y_o)(\mu^{X_i})^{y_o} .$$

29. CONDITIONING ON A STOCHASTIC PROCESS.

By means of the adjointness equation, applied as a criterion for a given family to be the family of conditional distributions, we can handle the problem of approximating the conditional distributions, given a stochastic process, by the conditional distributions given finitely many values of the samplefunction (cfr. section 9 , page 50).

Let $(Y_i | i \in I)$ be a family of compact spaces. We introduce notation similar to that of section 9 (see page 51):

$$Y_{I_1} := \prod_{i \in I_1} Y_i \quad \text{for } I_1 \subseteq I ,$$

$$p_{I_2 I_1} : Y_{I_2} \to Y_{I_1} \quad \text{for } I_1 \subseteq I_2 \subseteq I ,$$

$$\mathcal{P}_o := \text{the set of finite subsets of } I ,$$

$$\nu_I \in \mathcal{P}(Y_I) , \quad \nu_M := p_{IM} \nu_I \in \mathcal{P}(Y_M) .$$

29.1 Theorem. Let

$$t: (X, \mu) \to (Y_I, \nu_I)$$

be given. Suppose that

(1) for any finite subset M of I, the homomorphism

$$t_M := p_{IM} \circ t : (X, \mu) \to (Y_M, \nu_M)$$

has almost everywhere defined conditional distributions μ^{y_M}, $y_M \in Y_M$.

(2) For ν_I-almost all y_I, the limit

$$\mu_{y_I} := \lim_{M \uparrow I} \mu^{p_{IM}(y_I)}$$

exists and belongs to $\mathcal{P}(X)$ (the existence requires, of course, existence of $\mu^{p_{IM}(y_I)}$ from a certain stage $M \supseteq M_o$. By "$M \uparrow I$" we mean "$M \to \infty$ in the directed set $(\mathcal{F}_o, \subseteq)$").

(3) There exists a set $C \subseteq Y_I$ with $\nu_I C = 1$ such that the limit μ_{y_I} in (2) is defined for <u>all</u> samplefunctions y_I in C, and such that the mapping

$$y_I \to \mu_{y_I}$$
$$C \to \mathcal{P}(X)$$

is continuous.

(4) For all $k \in \mathcal{K}(X)$, the $L^2(\nu_I)$-function

$$[\mu^{p_{IM}(y_I)}k]_{y_I} = (L^2(t_M)^* k) \circ p_{IM}$$

converges to the $L^2(\nu_I)$-function

$$[\mu_{y_I}{}^k]_{y_I}$$

in quadratic mean, i.e.

$$\| [(\mu^{p_{IM}y_I} - \mu_{y_I})k]_{y_I} \|_2 \to 0 \text{ for } M \uparrow I .$$

Then, the conditional distribution of $x \in (X, \mu)$, given $t(x) = y_I$, is defined for all $y_I \in C$ by

$$\mu^{y_I} = \mu_{y_I} .$$

Remarks: Notice that the condition (4) follows from (2) by the dominated convergence principle in case I (and thereby φ_o) is denumerable. The condition (4) is just a regularity condition imposed on the convergence in (2). It is difficult to invent examples where (1),(2) and (3), but not (4), are satisfied. The necessity of (4) arises from the fact that a net of L^2-functions may converge pointwise towards one

function and in L^2-norm towards another. It is not hard to prove that $(L^2(t_M)^* k) \circ p_{IM}$ does converge in quadratic mean, namely towards $L^2(t)^* k$, so what (4) essentially says is that the net considered is not of this particular (pathological) type.

In view of these remarks, the consequences of the theorem boil down to this: If the limit distributions $\mu_{y_I} = \lim_{M \uparrow I} \mu^{p_{IM}(y_I)}$ are defined for almost all y_I and depend continuously upon y_I, then they constitute the conditional distributions for the transformation t.

In case $I = \mathbb{N}$, the convergence "$M \uparrow I$" can be replaced by "$n \to \infty$, $M = \{1, \ldots, n\}$"; this will be obvious from the proof.

Notice that the L^2-convergence in (4) is equivalent to L^1-convergence or convergence in probability, since the functions under consideration are uniformely bounded.

The continuity condition (3) is the really restrictive condition among the four. If only (1), (2) and (4) are satisfied and the mapping $y_I \to \mu_{y_I}$ is measurable, then the distributions μ_{y_I} can be regarded as "classical conditional distributions" in the sense that they give rise to a pointwise representation of the conditional expectation operator $L^2(t)^*$. In order to make "proper" conditional distributions out of the so defined distributions, the topology on Y_I can be changed, such that

μ_{y_I} becomes continuous as a function of y_I . In general, this is possible, but it may of course give rise to rather unnatural constructions.

Proof of the theorem: We apply theorem 24.2 (page 199). Obviously, we need only prove the adjointness equation

$$\nu_I[(\mu_{y_I} k)\cdot h(y_I)]_{y_I} = \mu[k(x)h(t(x))]_x$$

for h in a dense subspace of $\mathcal{K}(Y_I)$. We shall use the dense subspace of continuous function depending on a finite number of coordinates only (see page 354):

Let h be a $\mathcal{K}(Y_I)$ -function of the form

$$h = h_{M_o} \circ p_{IM_o} \qquad (M_o \in \mathcal{P}_o)$$

where $$h_{M_o} \in \mathcal{K}(Y_{M_o}) .$$

For $M \supseteq M_o$ (i.e. from a certain step in the limit procedure $M \uparrow I$) we have then

$$h = h_M \circ p_{IM}$$

where $h_M = h_{M_o} \circ p_{MM_o} \in \mathcal{K}(Y_M)$, and for $k \in \mathcal{K}(X)$ we get

$$\mu[k(x)h(t(x))]_x = \mu[k(x)h_M(t_M(x))]_x$$

$$= \nu_M[\mu^{y_M}k \cdot h_M(y_M)]_{y_M} = \nu_I[\mu^{p_{IM}(y_I)}k \cdot h_M(p_{IM}(y_I))]_{y_I}$$

$$= \nu_I[\mu^{p_{IM}(y_I)}k \cdot h(y_I)]_{y_I} .$$

For $M \uparrow I$ it follows immediately from (4) that

$$\mu[k(x)h(t(x))]_x = \nu_I[\mu_{y_I}{}^k \cdot h(y_I)]_{y_I} ,$$

and this proves the theorem (by theorem 24.2).

CHAPTER VII : EXAMPLES AND APPLICATIONS

30. CONDITIONING ON THE FIRST COORDINATE IN A TWODIMENSIONAL DISTRIBUTION.

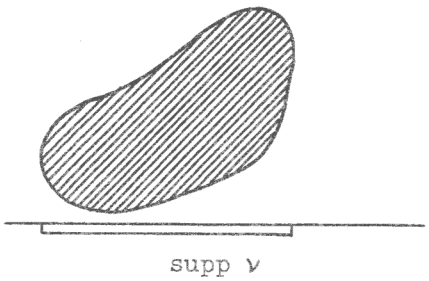
To illustrate the kind of regularity conditions necessary for existence of conditional distributions, pointwise or almost everywhere, we shall study some examples of distributions in the plane and their conditional distributions given the projection onto a line.

In most of the examples below, the distribution μ is defined in geometric terms, and the statements about existence or nonexistence of conditional distributions are correspondingly unprecise. The transformation t under consideration is the orthogonal projection onto a horizontally drawn line. Points of this line (the x-axis) are denoted x, x_0 etc. , not y, y_0 etc. as we are used to for points of the codomain Y .

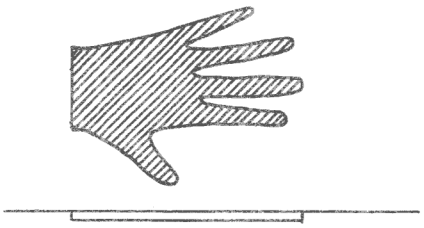
First consider some examples where μ is given by a density of the form

$$f = \frac{1}{\lambda^2 A} \cdot 1_A$$

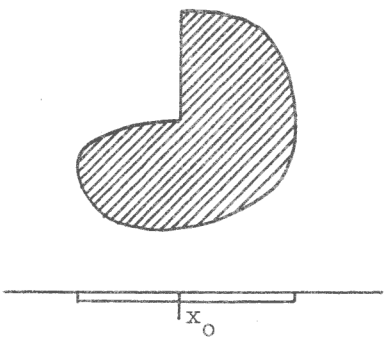
with respect to Lebesgue measure λ^2 . Thus, μ is the uniform distribution on the domain A .



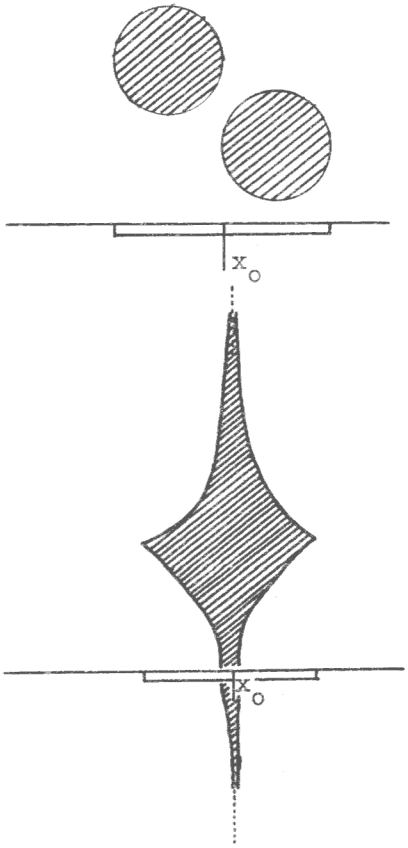
Example 1: For a plain area like this, the conditional distribution μ^x is defined for all x in the support of ν. The conditional distributions are simply uniform distributions on segments of vertical lines, except at the endpoints of the support interval, where the conditional distributions degenerate to one point measures.



Example 2: Also in this case the conditional distribution is defined everywhere on the support. At the left endpoint we get a uniform distribution on a line segment.

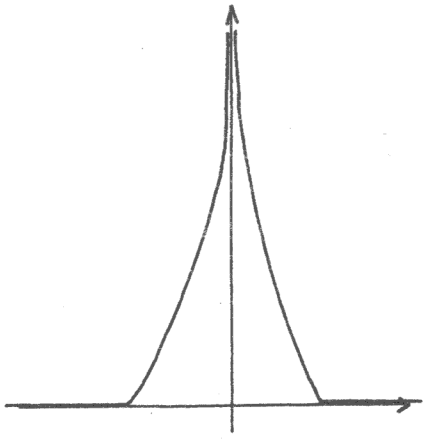


Example 3: This area yields a singularity point at x_o, where the "left conditional distribution" is different from the "right conditional distribution". Notice, however, that the conditional distribution is almost everywhere defined. The singularity point x_o is a discontinuity point for the density g of ν.



Example 4: This example shows that a singularity may occur even if the boundary of A does not contain a segment of a vertical line. Here, $g(x_o) = 0$.

Example 5: An unbounded domain of the shape indicated here gives rise to a defective conditional distribution at x_o. The density g admits the value $+\infty$ at x_o, when properly defined.



Example 6: We can construct A such that the conditional distribution is nowhere defined in the following manner:

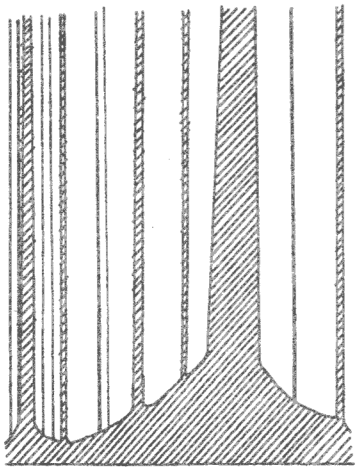
Let $h_o: \mathbb{R} \to \mathbb{R}$ be an integrable function of the shape indicated by the figure (we can take the density g from the previous example, or we can define h_o explicitely by, say,

$$h_o(x) := (-\log|x|) \vee 0 \ .$$

Next, define $h: \mathbb{R} \to \mathbb{R}$ by

$$h(x) := \sum_{n=1}^{\infty} \frac{1}{2^n} h_o(x-q_n)$$

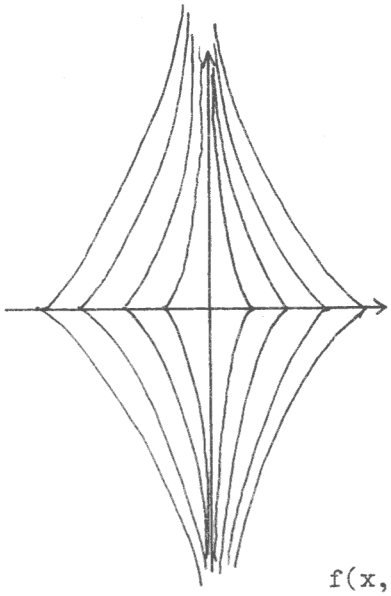
where (q_n) is a dense sequence of numbers (for example, the rational numbers, enumerated in some manner). Then h is integrable. For A we take the subgraph of h, i.e. we define



$$A := \{(x,y) \mid 0 \leq y \leq h(x)\} \ .$$

The scrawl to the left of the text is the closest we can come to a picture of this set A. The artist has been somewhat restrained by the fact that the set is dense in a half plane. The conditional distribution is nowhere defined (defective at the points q_n and completely undefined at all other points).

Example 7: Now, let us turn to the case where μ is given by some continuous density f with respect to Lebesgue measure in the plane. We saw in section 8, that the conditional distributions are then, in most cases, defined; but counter-examples do, of course, exist:



Let $f: \mathbb{R}^2 \to \mathbb{R}$ be a positive, continuous probability density which is constant on the y-axis and decrasing away from the y-axis, faster the larger is the distance to the x-axis. The level curves may look as indicated by the figure. As a concrete example, take

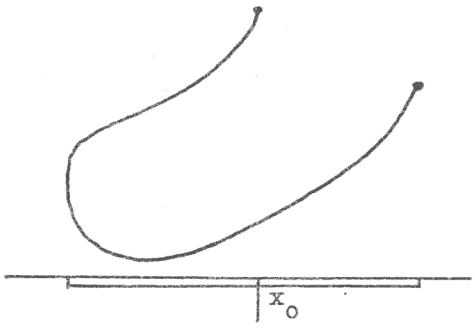
$$f(x,y) := \frac{1}{4} e^{-|x \cdot e^{|y|}|} .$$

Obviously, the conditional distribution of (x,y), given $x = 0$, is defective.

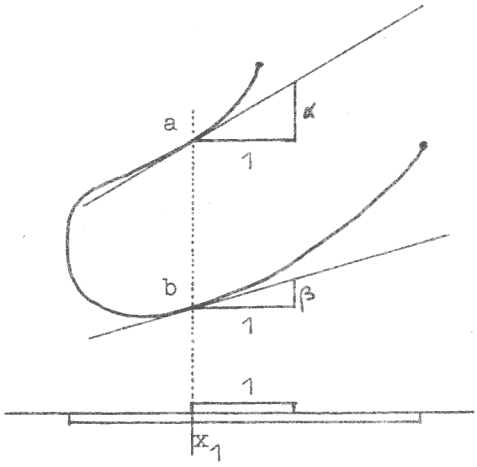
Example 8: Copying the construction of example 6, we can construct a continuous probability density f in the plane such that the conditional distribution of (x,y), given x, is nowhere defined: Put

$$f_1(x,y) := \sum_{n=1}^{\infty} \frac{1}{2^n} f(x-q_n, y)$$

where f denotes the density from example 7 above. Then, f_1 is a continuous probability density, and it is not hard to show that the conditional distributions are defective for $x = q_n$ and undefined at all other points. The drawing of the curves of constance for f_1 is left to the reader as an exercise.



Example 9: To illustrate some of the ideas in section 16 and 17, consider the case where μ is the uniform distribution (arc-length) on a smooth curve of length 1. For the curve drawn here the conditional distribution is defined at all points of the support, except at the point x_o. The conditional distributions to the right of x_o are one point measures, while the conditional distributions to the left of x_o (except at the left endpoint) are concentrated at two points.



At the point x_1 (see the second figure) the conditional distribution is concentrated at the two points a and b. The length of an infinitesimal piece of the curve at a, corresponding to an interval $[x_1, x_1+dx]$ on the x-axis is (by Pythagoras' theorem) $\sqrt{1+\alpha^2}\cdot dx$, where α denotes the slope of the tangent at a. Similarly we get the arc-length $\sqrt{1+\beta^2}\cdot dx$ at b, and so the conditional distribution of (x,y), given $x \in [x_1, x_1+dx]$, equals

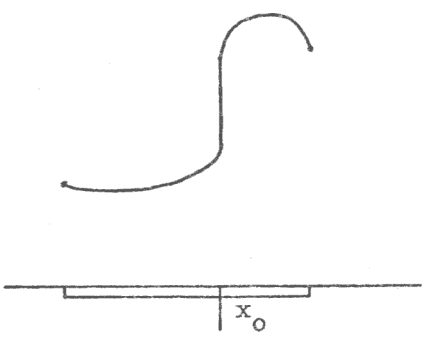
$$\text{const.}\cdot\left(\sqrt{1+\alpha^2}\cdot dx\cdot \varepsilon_a + \sqrt{1+\beta^2}\cdot dx\cdot \varepsilon_b\right)$$

$$= \frac{1}{\sqrt{1+\alpha^2}+\sqrt{1+\beta^2}}\cdot\left(\sqrt{1+\alpha^2}\cdot \varepsilon_a + \sqrt{1+\beta^2}\cdot \varepsilon_b\right).$$

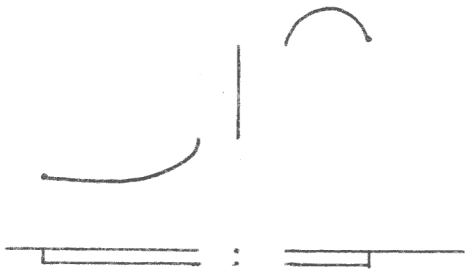
This is the conditional distribution of (x,y), given $x = x_1$. These considerations can easily be made precise : If the curve is given by an infinitesimally isometric parametrization

$$(x(s),y(s)) \quad , \quad s \in [0,1]$$

then the geometric measure on the curve is simply the transformed Lebesgue measure from $[0,1]$, and the formula for the conditional density f^y (section 16, page 141) is immediately applicable.



Example 10: This curve contains a segment of a vertical line; thus the transformed distribution ν has an atom at x_o, and the conditional distribution at x_o is not defined. Thus the conditional distribution is not defined almost everywhere. This is a typical "mixed dimension irregularity", where a seemingly



unimportant singularity in a certain dimension gives rise to a more serious singularity in a lower dimension. The solution of the problem is immediate, from the remarks in section 17 (page 145): We change the topology of the co-domain (and, possibly, on the domain) as indicated by the figure.

31. SOME APPLICATIONS OF THE METHODS IN CHAPTER IV.

Let μ be a probability measure in the plane $\mathbb{R}^2$, given by a density f with respect to Lebesgue measure, and let

$$t : \mathbb{R}^2 \to \mathbb{R}$$

be a surjectively regular transformation.

Suppose we have parametrizations

$$s_y : Z_y \to \mathbb{R}^2$$

of the level curves $X_y = t^{-1}(y)$, where Z_y is $\mathbb{R}$ or some disjoint union of open intervals on $\mathbb{R}$.

Under these assumptions, we can derive formulae for the density g of $y = t(x)$, the conditional density f^y of x with respect to the geometric measure on X_y and the density of the derived conditional distribution of, say, x_1, given $t(x_1,x_2) = y$. Throughout the section, the formulae of section 16 (page 139-141) are in constant use.

We shall not be very careful about variable specifications etc., since it will always be clear from the context which variables are to be integrated out, and partial derivatives etc. should always be evaluated at the varying point under consideration.

For example, we write

$$Dt(x) = Dt = \left[\frac{\partial y}{\partial x_1} \quad \frac{\partial y}{\partial x_2} \right] ;$$

thus

$$|Dt(x)|^{o} = \sqrt{|\det\left[(\frac{\partial y}{\partial x_1})^2+(\frac{\partial y}{\partial x_2})^2\right]|}$$

and so

$$F(x) = \frac{1}{|Dt(x)|^{o}} = \frac{1}{\sqrt{(\frac{\partial y}{\partial x_1})^2 + (\frac{\partial y}{\partial x_2})^2}} .$$

The geometric measure on the curve X_y has the density (cfr. the remarks in section 18, page 150) $|Ds_y|_o$ with respect to the Lebesgue measure "dz" , transformed by the parametrization, i.e.

$$\lambda_{X_y} = s_y(|Ds_y|_o \cdot \lambda_{Z_y}) ,$$

where

$$Ds_y(z) = \begin{bmatrix} \frac{\partial x_1}{\partial z} \\ \frac{\partial x_2}{\partial z} \end{bmatrix} ,$$

and so

$$|Ds_y|_o = \sqrt{|\det(Ds_y^* \, Ds_y)|}$$

$$= \sqrt{(\frac{\partial x_1}{\partial z})^2 + (\frac{\partial x_2}{\partial z})^2} .$$

Thus the density g of ν is given by

$$g(y) = \lambda_{X_y}(F \cdot f) =$$

$$\int_{Z_y} \sqrt{\frac{(\frac{\partial x_1}{\partial z})^2 + (\frac{\partial x_2}{\partial z})^2}{(\frac{\partial y}{\partial x_1})^2 + (\frac{\partial y}{\partial x_2})^2}} \cdot f(s_y(z))\, dz \quad .$$

The "conditional density" f^y of μ^y with respect to λ_{X_y} is given by

$$f^y(x) = \frac{1}{g(y)} \cdot F(x)f(x) \ .$$

Hence,

$$\mu^y = f^y \cdot \lambda_{X_y} = f^y \cdot s_y(|Ds_y|_o \cdot \lambda_{Z_y})$$

$$= \frac{1}{g(y)} \cdot F \cdot f \cdot s_y(|Ds_y|_o \cdot \lambda_{Z_y}) \ .$$

This formula tells us that the conditional distribution of the parameter

$$z = s_y^{-1}(x) \ ,$$

given $t(x) = y$, has the density

$$h^y(z) = \frac{1}{g(y)} \cdot F(s_y(z))f(s_y(z)) \cdot |Ds_y|_o$$

with respect to Lebesgue measure on Z_y . The conditional distribution of x_1 , given y , can then be computed by transformation of this distribution by the transformation $z \to x_1$, the first coordinate function of the parametrization s_y . In the three examples below, the curves X_y can be parametrized as graphs of functions , i.e. such that the parameter z equals the first coordinate $x_1 = (s_y)_1(z)$. This means that the density h^y above is simply the density of the conditional distribution of x_1 , given y .

Example 1: $y = t(x_1,x_2) = x_1 + x_2$;

$$Dt = [1 \quad 1] .$$

The level curves

$$X_y = \{(x_1,x_2) \mid x_1+x_2 = y\}$$

are lines, parametrizable as

$$X_y = \{(z,y-z) \mid z \in \mathbb{R}\} .$$

Then

$$Ds_y = \begin{bmatrix} 1 \\ -1 \end{bmatrix} , \text{ and}$$

$$g(y) = \int \sqrt{\frac{1^2 + (-1)^2}{1^2 + 1^2}} \cdot f(z,y-z)\, dz$$

$$= \int f(z,y-z)\, dz \quad .$$

The conditional distribution of x_1 (or z), given $x_1+x_2 = y$, has the density

$$h^y(z) = \frac{1}{g(y)} \cdot \frac{1}{\sqrt{2}} \cdot f(z,y-z) \cdot \sqrt{2} = \frac{1}{g(y)} \cdot f(z,y-z) \quad .$$

Example 2: $y = t(x_1,x_2) = x_1 \cdot x_2$;

$$Dt = \begin{bmatrix} x_2 & x_1 \end{bmatrix} .$$

The level curves X_y are hyperbolas, parametrizable by

$$(x_1,x_2) = s_y(z) = (z,\tfrac{y}{z}) \quad , \quad z \in \mathbb{R}\setminus\{0\} \quad .$$

Then

$$Ds_y = \begin{bmatrix} 1 \\ -\frac{y}{z^2} \end{bmatrix} .$$

Thus, the distribution of $t(x_1,x_2) = x_1 \cdot x_2$ has the density

$$g(y) = \int \sqrt{\frac{1^2 + \frac{y^2}{z^4}}{(\frac{y}{z})^2 + z^2}} \cdot f(z, \tfrac{y}{z})\, dz$$

$$= \int \frac{1}{|z|} f(z, \tfrac{y}{z})\, dz$$

and the distribution of x_1, given $x_1 \cdot x_2 = y$, has the density

$$h^y(z) = \frac{1}{g(y)} \cdot \frac{1}{|z|} \cdot f(z, \tfrac{y}{z}) .$$

As a special case, suppose that x_1 and x_2 are independent, uniformely distributed on $[0,1]$. Then the product $x_1 \cdot x_2$ has the density

$$g(y) = \int_y^1 \frac{1}{|z|}\, dz = -\log y \quad (y \in [0,1])$$

while the distribution of x_1, given $x_1 \cdot x_2 = y$, has the density

$$h^y(z) = \frac{1}{|z| \cdot g(y)} = \frac{1}{z \cdot (-\log y)} \quad (z \in [y,1]).$$

Example 3: $$y = t(x_1, x_2) = \frac{x_2}{x_1} ;$$

$$Dt = \left[-\frac{x_2}{x_1^2} \quad \frac{1}{x_1} \right] .$$

The level curves are lines, parametrizable by

$$(x_1,x_2) = s_y(z) = (z,\ yz)\ ,$$

$$Ds_y = \begin{bmatrix} 1 \\ y \end{bmatrix}\ .$$

Thus

$$g(y) = \int \sqrt{\frac{1^2 + y^2}{(-\frac{yz}{z^2})^2 + (\frac{1}{z})^2}}\ \cdot f(z,yz)\ dz$$

$$= \int |z| \cdot f(z,yz)\ dz\ ,$$

and the conditional distribution of x_1 is given by

$$h^y(z) = \frac{1}{g(y)} \cdot |z| \cdot f(z,yz)\ .$$

As a special case, suppose that x_1 and x_2 are independent, normally distributed $(0,1)$, i.e.

$$f(x_1,x_2) = \frac{1}{2\pi}\ e^{-(x_1^2+x_2^2)/2}\ .$$

The distribution of $y = x_2/x_1$ is then given by the density

$$g(y) = \int |z| \cdot \frac{1}{2\pi} \cdot e^{-(z^2+\ y^2z^2)/2}\ dz$$

$$= 2 \int_0^\infty z \cdot \frac{1}{2\pi} \cdot e^{-z^2(1+y^2)/2} \, dz$$

$$= \frac{1}{\pi(1+y^2)} \int_0^\infty e^{-z^2(1+y^2)/2} \, (1+y^2) \, d\frac{z^2}{2}$$

$$= \frac{1}{\pi(1+y^2)} \int_0^\infty e^{-w} \, dw$$

$$= \frac{1}{\pi(1+y^2)} \, .$$

Thus, the distribution of x_2/x_1 is a normalized Cauchy distribution. The conditional distribution of x_1, given $x_2/x_1 = y$, has the density

$$h^y(z) = \frac{1}{g(y)} \cdot |z| \frac{1}{2\pi} e^{-z^2(1+y^2)/2} \, .$$

32. THE NORMALIZED NORMAL DISTRIBUTION.

Many classical distributions, considered in the statistical applications of the normal distribution, are easily deduced by the methods of chapter IV. This approach becomes geometric, as opposed to the more analytic, classical approach. For example, normalizing factors involve areas of spheres rather than values of the Γ-function, and the formulation is purely coordinate-free (except, of course, for the solutions to classical problems, posing themselves in coordinate terms).

Notation:

E_n denotes an n-dimensional Euclidean space.

λ_n denotes the geometric measure on E_n.

$S_{n-1}(a) = \{ x \in E_n \mid \|x\| = a \}$ is the sphere of radius a in E_n. $S_{n-1}(a)$ is (for $a>0$) an (n-1)-dimensional submanifold of E_n (namely, a level surface for the surjectively regular mapping $x \to \|x\|^2$, $E_n \setminus \{0\} \to]0,+\infty[$). We put

$S_{n-1} := S_{n-1}(1)$.

For the open ball of radius a we write

$B_n(a) := \{x \in E_n \mid \|x\| < a\}$, and the open unit ball

is denoted by

$B_n := B_n(1)$.

The manifold $S_{n-1}(a)$ is compact, and so its geometric measure must be bounded. We write

$A_{n-1}(a) := \| \lambda_{S_{n-1}(a)} \|_\infty$ for its total mass, the "area" of an (n-1)-dimensional sphere of radius a. In particular, we put

$A_{n-1} := A_{n-1}(1)$.

The "volume" of the ball $B_n(a)$ is denoted by

$V_n(a) := \lambda_n(B_n(a))$, in particular

$V_n := V_n(1)$.

<u>The normalized normal distribution on</u> E_n is defined as

$\nu_n := \varphi_n \cdot \lambda_n$, where $\varphi_n(x) = c_n \cdot e^{-\|x\|^2/2}$

and c_n is a normalizing factor, for the present only known as

$c_n = 1/\lambda_n[e^{-\|x\|^2/2}]_x$.

We start by proving a wellknown theorem of fundamental importance in multivariate statistical analysis:

32.1 Theorem. Let

$$p_i : E_n \to E_{n_i} \quad , \quad i = 1,2, \ldots , k$$

be linear mappings satisfying the following three conditions:

(1) $p_i p_j^* = 0$ for $i \neq j$

(2) $p_i p_i^* = 1_{E_{n_i}}$ (i.e. p_i is coisometric)

(3) $\sum_1^k p_i^* p_i = 1_{E_n}$.

Then, for $x \in (E_n, \nu_n)$, the stochastic variables

$$\begin{aligned} y_1 &= p_1(x) \\ &\vdots \\ y_k &= p_k(x) \end{aligned}$$

are independent and normally distributed, i.e. the transformation

$$(p_1, \ldots, p_k) : E_n \to E_{n_1} \times \ldots \times E_{n_k}$$

takes ν_n into $\nu_{n_1} \otimes \ldots \otimes \nu_{n_k}$.

Proof: It follows immediately from the conditions (1), (2) and (3) that the linear mapping $(p_1, \ldots, p_k)$ is bijective and isometric, i.e. it preserves Euclidean structure forwards and backwards, when $E_{n_1} \times \ldots \times E_{n_k}$ is equipped as a Euclidean space in the usual manner by the inner product

$$((y_1,\ldots,y_k)|(y_1',\ldots,y_k')) = (y_1|y_1') + \ldots + (y_k|y_k') .$$

Consequently, $(p_1,\ldots,p_k)$ transforms the normal distribution on E_n into the normal distribution on $E_{n_1} \times \ldots \times E_{n_k}$. The latter is given by the density

$$c_n \cdot e^{-\|y\|^2/2} = c_n \cdot e^{-\frac{1}{2}\sum \|y_i\|^2}$$

$$= c_n \cdot \prod_1^k e^{-\|y_i\|^2/2}$$

with respect to the geometric measure

$$\lambda_{E_{n_1} \times \ldots \times E_{n_k}} = \lambda_{n_1} \otimes \ldots \otimes \lambda_{n_k} .$$

From this, the theorem follows immediately. In addition, we

have proved that

$$c_n = c_{n_1} \cdot \ldots \cdot c_{n_k} .$$

Example: Let $e_1, \ldots, e_n$ be an orthonormal base in E_n, and consider the coordinate variables

$$y_i := p_i(x) := (e_i | x) .$$

The so defined mappings $p_i : E_n \to \mathbb{R}$ are easily seen to satisfy the conditions of the theorem. Thus $y_1, \ldots, y_n$ are independent, normally distributed with parameters $(0,1)$. For the normalizing factors we get

$$c_n = c_1^n .$$

The χ^2-distribution. Consider the transformation

$$t : E_n \to]0, +\infty[$$

defined by

$$t(x) := \|x\|^2$$

(the transformation is not defined at $x = 0$, since we exclude 0 from the codomain. Thus E_n ought to be replaced by $E_n \setminus \{0\}$. Such minor changes (removal of a closed null set) are ignored here and in the following).

The distribution of the stochastic variable

$$y := t(x) = \|x\|^2 \quad , \quad x \in (E_n, \nu_n) \ ,$$

is called the χ^2-distribution with n degrees of freedom. In order to compute its density, we must know the differential of t:

Identifying E_n with $D(E_n, x_o)$ and $\mathbb{R}$ with $D(\,]0,+\infty[, y_o)$ in the usual manner, we get

$$\begin{aligned}(Dt(x_o))x &= \lim_{h \to 0} \frac{1}{h} (t(x_o+h\cdot x) - t(x_o)) \\ &= \lim_{h \to 0} \frac{1}{h} (2h(x_o|x) + h^2 \|x\|^2) \\ &= 2(x_o|x) \ , \text{ i.e.}\end{aligned}$$

$$Dt(x_o) = [2(x_o|x)]_x : E_n \to \mathbb{R} \ .$$

The adjoint mapping $Dt(x_o)^*: \mathbb{R} \to E_n$ is determined by

$$\begin{aligned}(Dt(x_o)^* y|x)_{E_n} &= (y|Dt(x_o)x)_{\mathbb{R}} = y\cdot 2(x_o|x)_{E_n} \\ &= (2y\cdot x_o|x)_{E_n} \ ,\end{aligned}$$

i.e.

$$Dt(x_o)^* y = 2y\cdot x_o \ .$$

The composed mapping $Dt(x_o)Dt(x_o)^* : \mathbb{R} \to \mathbb{R}$ is then

$$Dt(x_o)Dt(x_o)^* y = Dt(x_o)(2y \cdot x_o) = 2(x_o | 2y \cdot x_o)$$

$$= 4y(x_o | x_o) = 4y \|x_o\|^2 .$$

Hence

$$F(x_o) = \frac{1}{|Dt(x_o)|^o} = \frac{1}{\sqrt{4\|x_o\|^2}} = \frac{1}{2\|x_o\|} .$$

By the formula for $g(y)$ (cfr. section 16) we get

$$g(y) = \lambda_{X_y}(f \cdot F) = \lambda_{S_{n-1}(\sqrt{y})}(f \cdot F)$$

$$= \lambda_{S_{n-1}(\sqrt{y})} \left[c_n e^{-\|x\|^2/2} \cdot \frac{1}{2\|x\|} \right]_x$$

$$= c_1 {}^n e^{-y/2} \cdot \frac{1}{2\sqrt{y}} A_{n-1}(\sqrt{y})$$

(the integrand is constant on the level surface).

Now, we need the following (intuitively obvious) result:

32.2 Lemma: For $a > 0$, we have $A_{n-1}(a) = a^{n-1} A_{n-1}$.

Proof: Consider the diagram

$$\begin{array}{ccccc} S_{n-1} & \hookrightarrow & E_n & \xrightarrow{t} &]0,+\infty[\\ \downarrow a\cdot & & \downarrow a\cdot & & \downarrow a^2\cdot \\ S_{n-1}(a) & \hookrightarrow & E_n & \xrightarrow{t} &]0,+\infty[\end{array} .$$

Obviously the diagram commutes: Multiplication by a maps S_{n-1} into $S_{n-1}(a)$, and the squared norms are multiplied by a^2. Taking differentials, we get a diagram satisfying the conditions of theorem 11.4 (page 80). The determinants of the diagram are

$$\begin{array}{ccc} 1 & & 2\,\|x\| \\ \downarrow ? & & \downarrow a^n \quad \downarrow a^2 \\ 1 & & 2\,\|a\cdot x\| \end{array} .$$

Hence, the determinant in question is

$$\frac{1\cdot a^n\cdot 2\,\|a\cdot x\|}{1\cdot 2\,\|x\|\cdot a^2} = a^{n-1} .$$

From this we conclude immediately (by the integral transformation theorem) that the geometric measure on $S_{n-1}(a)$ is obtained by multiplication of the transformed geometric measure from S_{n-1} by the factor a^{n-1}, and so

$$A_{n-1}(a) = a^{n-1}A_{n-1} .$$

The lemma enables us to write the density of the χ^2-distribution as

$$g(y) = c_1{}^n e^{-y/2} \frac{1}{2\sqrt{y}} A_{n-1}(\sqrt{y})$$

$$= c_1{}^n e^{-y/2} \frac{1}{2\sqrt{y}} (\sqrt{y})^{n-1} A_{n-1} ,$$

i.e.

$$\boxed{g(y) = \frac{1}{2} c_1{}^n A_{n-1} \cdot y^{\frac{n-2}{2}} \cdot e^{-y/2}}$$

Computation of c_1. For $n = 2$, the χ^2-density equals

$$g(y) = \frac{1}{2} \cdot c_1{}^2 \cdot A_1 \cdot e^{-y/2} .$$

Since g is a probability density on $]0,+\infty[$, we must have

$$1 = \int_1^\infty g(y)dy = c_1{}^2 A_1 \int_0^\infty e^{-y/2} d\frac{y}{2} = c_1{}^2 A_1 ,$$

and so

$$c_1 = \frac{1}{\sqrt{A_1}} .$$

Introducing the more customary notation

$$\pi := \frac{1}{2} A_1 \approx 3.1416$$

we get

$$c_1 = \frac{1}{\sqrt{2\pi}} \approx 0.3989 .$$

The uniform distribution on $S_{n-1}(a)$. By the uniform distribution on the sphere $S_{n-1}(a)$, we mean the probability measure

$$\frac{1}{A_{n-1}(a)} \lambda_{S_{n-1}(a)} .$$

This distribution appears as the conditional distribution of $x \in (E_n, \nu_n)$, given $t(x) = \|x\|^2 = a^2$, since the conditional density

$$f^y(x) = \frac{1}{g(y)} F(x)f(x)$$

is constant on the sphere.

Projection of the uniform distribution. For $k < n$, let

$$p_o : E_n \to E_k$$

be an arbitrary coisometry (page 67). Then

$$p_o(S_{n-1}) = \overline{B_k} \quad (= \text{the closure of } B_k).$$

By

$$p: S_{n-1} \to B_k$$

we denote the (almost everywhere defined) restriction of p_o. We shall compute the density of the transformed measure

$$p\left(\frac{1}{A_{n-1}} \lambda_{S_{n-1}}\right)$$

with respect to the geometric measure on B_k.

For convenience, we think of E_k as a subspace of E_n, imbedded by p_o^*. Then, p_o is simply the orthogonal projection onto that subspace, and the level surfaces for p are

$$(S_{n-1})_z = p^{-1}(z) = \{x \in S_{n-1} \mid px = z\}$$

$$= \{x \in S_{n-1} \mid x-z \in E_k^{\perp}\}$$

$$= \{x \in E_n \mid \|x\|^2 = 1 \text{ and } x-z \in E_k^{\perp}\}$$

$$= \{x \in E_n \mid \|x-z\|^2 = 1-\|z\|^2 \text{ and } x-z \in E_k^{\perp}\}$$

$$= \{z+y \mid \|y\|^2 = 1-\|z\|^2 \text{ and } y \in E_k^{\perp}\}$$

$$= z + S_{n-k-1}(\sqrt{1-\|z\|^2}),$$

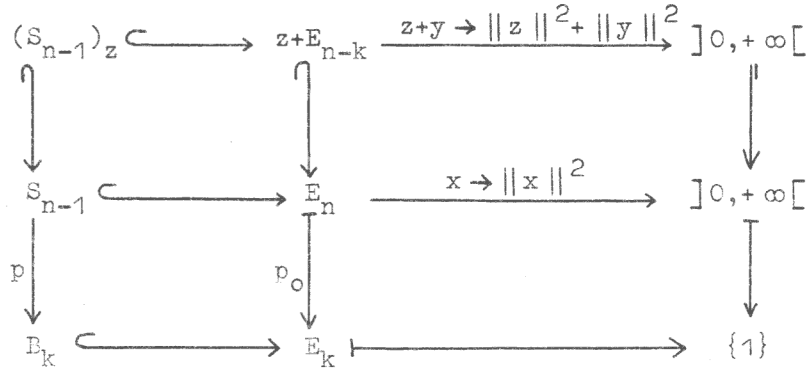
where $S_{n-k-1}(\sqrt{1-\|z\|^2})$ denotes the sphere of radius

$\sqrt{1-\|z\|^2}$ in the subspace

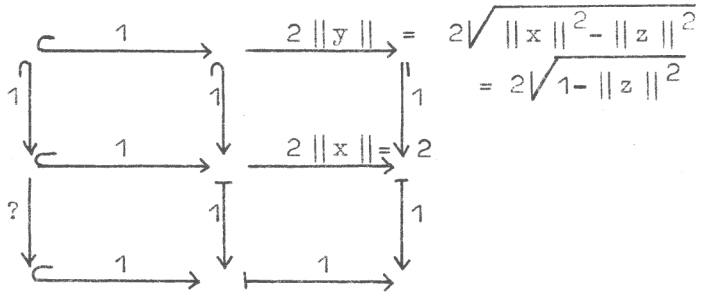
$$E_{n-k} := E_k^{\perp} .$$

Hence, the level surface $(S_{n-1})_z$ is simply an (n-k-1)-dimensional sphere.

In order to compute the function F corresponding to the transformation p, consider the diagram



The diagram commutes, and the corresponding diagram of differentials satisfies the conditions of theorem 11.5 . The determinants are



Thus the determinant in question is

$$|Dp(x)|^{o} = \frac{2\sqrt{1-\|z\|^2}}{2} = \sqrt{1-\|z\|^2}\,.$$

Now, the density of the distribution of $z = p(x)$ can be computed (by the formula $g(y) = \lambda_{X_y}(F\cdot f)$):

$$h(z) = \lambda_{(S_{n-1})_z}\left[\frac{1}{\sqrt{1-\|z\|^2}} \cdot \frac{1}{A_{n-1}}\right]_x$$

$$= \frac{1}{\sqrt{1-\|z\|^2}} \cdot \frac{A_{n-k-1}(\sqrt{1-\|z\|^2})}{A_{n-1}}$$

$$\boxed{= \frac{A_{n-k-1}}{A_{n-1}}\,(1-\|z\|^2)^{\frac{n-k-2}{2}}}$$

Example: For $k = 1$, any coisometry

$$p_o: E_n \to \mathbb{R} \quad (\approx E_1)$$

has the form

$$p_o(x) = (x_o|x)\,,$$

where x_o is a unit vector. Thus, the inner product of a fixed unit vector with a uniformely distributed unit vector has the density

$$\frac{A_{n-2}}{A_{n-1}} (1-z^2)^{\frac{n-3}{2}} \quad , \quad z \in B_1 =]-1,1[\ .$$

In particular, for $n = 3$ we get a constant density, i.e. z is uniformely distributed on $]-1,1[$. Hence

$$\frac{A_{3-2}}{A_{3-1}} V_1 = 1$$

or

$$A_2 = A_1 \cdot V_1 = 4\pi \ .$$

A more general version of this argument gives a recursive formula for A_{n-1} :

Computation of A_{n-1} . According to theorem 15.1 (the decomposition of the geometric measure) Lebesgue measure on E_n can be represented as the mixture of the measures

$$\lambda_y = \frac{1}{2\sqrt{y}} \lambda_{S_{n-1}(\sqrt{y})}$$

with respect to Lebesgue measure on $]0,+\infty[$. From this it follows immediately that

$$V_n = \lambda_{E_n}(B_n) = \int_0^\infty \frac{1}{2\sqrt{y}} \lambda_{S_{n-1}(\sqrt{y})}(B_n) \, dy$$

$$= \int_0^1 \frac{1}{2\sqrt{y}} A_{n-1}(\sqrt{y})\, dy = \int_0^1 \frac{1}{2\sqrt{y}} (\sqrt{y})^{n-1} A_{n-1}\, dy$$

$$= \frac{1}{2}A_{n-1} \int_0^1 y^{\frac{n-2}{2}}\, dy = \frac{1}{2}A_{n-1} \cdot \frac{1}{\frac{n-2}{2} + 1} = \frac{1}{n} A_{n-1} .$$

Now, suppose a coisometry

$$p_o : E_n \to E_{n-2}$$

is given. We have proved that the uniform distribution on S_{n-1} is transformed by p_o into the distribution with density

$$\frac{A_{n-(n-2)-1}}{A_{n-1}} (1 - \|z\|^2)^{\frac{n-(n-2)-2}{2}} = \frac{A_1}{A_{n-1}}$$

with respect to $\lambda_{B_{n-2}}$. Thus

$$1 = \frac{A_1}{A_{n-1}} \lambda_{B_{n-2}}(B_{n-2}) = \frac{A_1}{A_{n-1}} V_{n-2} = \frac{A_1 \cdot A_{n-3}}{A_{n-1} \cdot (n-2)}$$

and so

$$\boxed{A_{n-1} = \frac{2\pi}{n-2} A_{n-3}}$$

Succesive applications of this formula gives

For n even:

$$A_{n-1} = \frac{2\pi}{n-2}\cdot\frac{2\pi}{n-4}\cdot \ldots \cdot\frac{2\pi}{2}\cdot 2\pi \quad .$$

For n odd:

$$A_{n-1} = \frac{2\pi}{n-2}\cdot\frac{2\pi}{n-4}\cdot \ldots \cdot\frac{2\pi}{1}\cdot 2 \quad .$$

n	A_{n-1}		
1	2	=	2.0000
2	2π	≈	6.2832
3	4π	≈	12.5664
4	$2\pi^2$	≈	19.7392
5	$\frac{8}{3}\pi^2$	≈	26.319
6	π^3	≈	31.006
7	$\frac{16}{15}\pi^3$	≈	33.073
8	$\frac{1}{3}\pi^4$	≈	32.470
9	$\frac{32}{105}\pi^4$	≈	29.687
10	$\frac{1}{12}\pi^5$	≈	25.502

n	A_{n-1}
11	20.725
12	16.023
13	11.838
14	8.390
15	5.722
16	3.765
17	2.397
18	1.479
19	0.8858
20	0.5161

The distribution of the correlation coefficient. Let

$$x_1, y_1, \quad x_2, y_2, \quad x_3, y_3, \quad \ldots \quad , \quad x_n, y_n$$

be stochastically independent, normally distributed with parameters

$$E\, x_i = \xi \quad , \quad E\, y_i = \eta$$

$$E(x_i-\xi)^2 = \sigma^2 \quad , \quad E(y_i-\eta)^2 = \omega^2 \, .$$

Consider the <u>empirical correlation coefficient</u>

$$r = \frac{\sum (x_i-\bar{x})(y_i-\bar{y})}{\sqrt{(\sum (x_i-\bar{x})^2)(\sum (y_i-\bar{y})^2)}}$$

where

$$\bar{x} = \frac{1}{n}(x_1 + \ldots + x_n)$$

$$\bar{y} = \frac{1}{n}(y_1 + \ldots + y_n) \, .$$

In order to compute the distribution of r, first notice that this ditribution does not depend upon the four parameters ξ, η, σ^2 and ω^2 : Defining

$$u_i := \frac{x_i - \xi}{\sigma}$$

$$v_i := \frac{y_i - \eta}{\omega}$$

we may as well write the correlation coefficient as

$$r = \frac{\sum (u_i-\bar{u})(v_i-\bar{v})}{\sqrt{(\sum (u_i-\bar{u})^2)(\sum (v_i-\bar{v})^2)}} ,$$

and here, $u_1, v_1, \ldots, u_n, v_n$ are <u>normalized</u> normally distributed.

The vectors

$$u' := (u_1-\bar{u}, \ldots, u_n-\bar{u})$$

and

$$v' := (v_1-\bar{v}, \ldots, v_n-\bar{v})$$

are the orthogonal projections of $u = (u_1, \ldots, u_n)$ and $v = (v_1, \ldots, v_n)$ onto the $(n-1)$-dimensional subspace

$$E_{n-1} := \{(u_1, \ldots, u_n) \mid \sum u_i = 0\} .$$

By theorem 32.1, u' and v' are normalized normally distributed on E_{n-1}. Thus, the correlation coefficient has the form

$$r = \left(\frac{u'}{\|u'\|} \mid \frac{v'}{\|v'\|} \right) ,$$

where $u'/\|u'\|$ and $v'/\|v'\|$ are independent and uniformely distributed on the unit sphere in E_{n-1}. Then, <u>the conditional distribution of</u> r, <u>given that</u> $u'/\|u'\|$ <u>equals a fixed unitvector</u> e_o, is the distribution of an inner product of a fixed unit vector with a uniformely distributed

unit vector, as deduced earlier (page 277-278). This distribution being independent of the fixed unit vector, we conclude that it must also be the unconditioned distribution of r. Hence, the distribution of the empirical correlation coefficient has the density

$$\boxed{\frac{A_{n-3}}{A_{n-2}} (1-r^2)^{\frac{n-4}{2}} \quad , \quad r \in]-1,1[}$$

In addition, the argument above showed that r is stochastically independent of u (and, similarly, of v, but certainly not of (u,v)).

33. CONDITIONING ON A LINEAR FUNCTION IN A NORMAL PROCESS.

By a normal process (or a Gaussian process) we mean a process given by a consistent family of multidimensional normal distributions (i.e. affine transformations of normalized normal distributions).

In order to apply Kolmogorovs consistency theorem, we must compactify the state space. Thus, the processes considered are of the form

$$(x_t) \quad \in \quad ([-\infty,+\infty]^T \, , \, \mu) \, ,$$

and the normality simply means that for any finite subset $\{t_1, \ldots, t_k\}$ of T, the distribution of $(x_{t_1}, \ldots, x_{t_k})$ is a usual normal distribution on $\mathbb{R}^n$ (possibly concentrated on an affine subspace), imbedded in $[-\infty,+\infty]^n$.

In a finite dimensional normal distribution, conditional distributions given a linear function of the observations, are always defined on the support. In order to prove this statement, just reduce to the problem of conditioning on a coisometry in a normalized normal distribution, and apply theorem 32.1 (page 267).

It follows immediately from theorem 9.1 (page 52) that conditioning upon a linear function

$$y = t(x_{t_1}, \dots, x_{t_k}), \quad t: \mathbb{R}^k \to \mathbb{R}^n$$

of finitely many coordinates is possible.

We can extend this result a little; essentially, conditioning upon any finite dimensional linear function in a normal process is possible:

Let $L^2(x_t)$ be the closed subspace of $L^2(\mu)$, spanned by the variables x_t (regarded as L^2-functions). It can be proved (not surprisingly) that for any finite number $f_1, \dots, f_n$ of functions from $L^2(x_t)$, the n-dimensional stochastic variable

$$y = (y_1, \dots, y_n) = (f_1((x_t)), \dots, f_n((x_t)))$$

is normally distributed. By means of this result, it follows immediately from theorem 9.1, that conditioning on any such variable y is possible (the joint distribution of $y_1, \dots y_n, x_{t_1}, \dots, x_{t_k}$ is normal for any finite subset $\{t_1, \dots t_k\}$ of T).

Example: For the Wiener process

$$(x_t) \in ([-\infty, +\infty]^{[0,+\infty[}, \mu)$$

given by

$$E\, x_t = 0, \qquad E\, x_t x_s = t \wedge s,$$

it can be proved that the sample function is continuous almost surely. Thus, a variable like

$$y := \int_0^1 x_t dt$$

is welldefined. From

$$y = \lim \frac{1}{n} \sum_{i=1}^{n} x_{i/n}$$

it follows easily that y belongs to $L^2(x_t)$ (or, more precise: y as a function of (x_t) belongs to $L^2(x_t)$), and so the conditional distribution of a Wiener process, given $\int_0^1 x_t dt = y_o$, is welldefined for all $y_o \in \mathbb{R}$.

It should be emphasized, that the result is only valid for finite dimensional linear functions of a process. Conditioning upon an infinite number of linear variables is in general not possible, unless the topology of the codomain is chosen finer than the product topology (cfr. the remarks in section 29, page 246).

Example: Let $(x_t | t \in \dot{Z})$ be a stationary normal process (i.e. its distribution is invariant under translations of the timescale) and consider the autoregression, or the predictor

$$y_t := E(x_{t+1} \mid x_t , x_{t-1} , x_{t-2} , \dots) ,$$

the conditional expectation of the next observation x_{t+1}, given $x_t, x_{t-1}, \ldots$ (the observed part of the process). It can be proved that y_t (as a function of (x_t)) belongs to the space $L^2(x_t)$.

Now, one would expect the conditional distribution of x_{t+1}, given $x_t, x_{t-1}, \ldots$, to be the normal distribution with mean y_t and variance

$$\sigma^2 = E(x_{t+1}-y_t)^2 = \text{the variance of the prediction error.}$$

This is true, if the conditional distribution is defined; but it is not always defined: Let $(x_t^o, x_{t-1}^o, \ldots)$ be "an observed past" for which the conditional distribution of the next observation is defined. Then, by our definition of a conditional distribution, there exists a neighbourhood $V \subseteq [-\infty,+\infty]^{\{t, t-1, \ldots\}}$ of the observed past such that conditioning on different subsets of that neighbourhood give approximately the same conditional distribution. Now, the neighbourhood V (in the product topology) contains a neighbourhood which is a cylinder with finite dimensional base, i.e. a set determined by conditions on finitely many coordinates only. In view of the linearity of the autoregression, we conclude that it must be independent of all coordinates but these. This means, that the process is autoregressive of finite order: the autoregression depends on a finite number of past observations only. It is not extremely difficult to make a precise argument out of these loose

remarks, proving that for a stationary normal process with time $\grave{Z}$, the following two conditions are equivalent:

(1) For all (or just for almost all) "pasts" $(x_t, x_{t-1}, \dots) \in \mathbb{R}^{\{t,t-1,\dots\}}$ the conditional distribution of x_{t+1} is defined.

(2) The autoregression has the form

$$y_t = c_o x_t + c_1 x_{t-1} + \dots + c_n x_{t-n} .$$

(the "almost all"-reservation in (1) plays no role, unless the process is finitely generated, i.e. the variables x_t satisfy linear relations, such that the whole process can be generated by a finite dimensional normal distribution).

34. MARKOV PROCESSES.

Let X be a compact space with a denumerable base for the topology, and let T denote an interval on $\mathbb{R}$ or $\dot{\mathbb{Z}}$. Consider a process

$$(x_t | t \in T) \quad \in \quad (X^T, \mu) .$$

For $t_o \in T$, let

$$V_- , V_o \text{ and } V_+ \quad \subseteq L^2(\mu)$$

denote the subspaces of functions of the past, present and future, respectively; that is

V_- is the space of L^2-functions, depending on $(x_t | t \leq t_o)$ only,

V_o is the space of L^2-functions depending on x_{t_o} only,

V_+ is the space of L^2-functions depending on $(x_t | t \geq t_o)$ only.

Definition: The process is said to be a Markov process, if for all t_o the two spaces spaces V_- and V_+ are geometrically orthogonal (cfr. page 209) with intersection $V_- \cap V_+ = V_o$.

Remark: For most processes, the equation $V_- \cap V_+ = V_o$ is satisfied (exceptions are periodical processes etc.). Thus the essential contents of the definition is the geometric

orthogonality.

Conditional independence. Define two stochastic variables to be independent, given a third stochastic variable, if the two corresponding L^2-subspaces are geometrically orthogonal with intersection contained in the third. Then, the Markov property simply states that past and future are conditionally independent, given the present.

It is not hard to prove, that the above definition of a Markov process coincides with the usual one. The definition is merely included in order to illustrate, how manipulations with L^2-subspaces can replace wellknown manipulations with sub-σ-algebras.

The strong Markov property. Call a Markov process with time-scale $[0,+\infty[$ a Feller-process, if it can be constructed in the usual manner from an initial distribution π_0 of x_0 and a family of transition probability distributions

$$\pi_x^{(s,t)} \in \mathcal{P}(X) \quad , \quad s, t \in T \, , \quad s \leq t \quad , \quad x \in X$$

where $\pi_x^{(s,t)}$ depends continuously upon its three arguments (jointly) and satisfies the conditions

$$\pi_x^{(t,t)} = \varepsilon_x$$

and (the Chapman-Kolmogorov-equations)

$$\pi_x^{(s,t)}\left[\pi_y^{(t,r)}\right]_y = \pi_x^{(s,r)} \quad , s \leq t \leq r .$$

The interpretation of the transition distributions is that $\pi_x^{(s,t)}$ is the conditional distribution of x_t, given $x_s = x$, and it is easy to prove from the construction that this is in fact true, when the local definition of conditional distributions is applied (for x in the support of the distribution of x_s).

We shall summarize some results from Tue Tjur (1972). The results will not be proved here, since they involve sample function properties and a non-trivial measurability problem.

Let

$$\tau : \quad X[0,+\infty[\rightarrow [0,+\infty]$$

be a <u>stopping time</u>, i.e. a μ-measurable mapping with the following property: For any samplefunction (x_t) such that

$$\tau((x_t)) = t_o$$

and for any other sample function (y_t) <u>coinciding with</u> (x_t) <u>on an interval of the form</u> $[0, t_o+\varepsilon[$ $(\varepsilon > 0)$, we have

$$\tau((y_t)) = \tau((x_t)) \quad (=t_o).$$

Loosely speaking, the condition means that the stopping time depends on the past, and possibly on the "infinitesimal future",

but not on the future. The intuitive interpretation is, that a stopping time is a time where we stop the process (or stop observing it), and the decision of whether to stop or not is consecutively based on the observed part of the process. For example, we may stop the process the first time it runs into a certain subset of X (open, to make τ measurable), or we may stop (say, for X discrete) at the third jump (state-shift) of the process.

Now, define a new process (y_t) from the Feller process $x = (x_t)$ by

$$y_t := \begin{cases} x_t & \text{for } t \leq \tau(x) \\ x_{\tau(x)+} & \text{for } t > \tau(x) \end{cases}$$

($x_{\tau(x)+}$ stands for the right limit of the samplefunction at $\tau(x)$. It can be proved that (with probability one) the samplefunction has right and left limits at all points of the timescale).

This is the stopped process; it replaces, together with the stopping time τ , the σ-algebra induced by τ , as usually introduced in treatments of the strong Markov property. It can be proved, that the mapping taking (x_t) into (y_t) is measurable. Moreover, it can be proved that the conditional distribution of (x_t) , given a fixed samplefunction (y_t^o) of the stopped process and a fixed value t_o of the stopping time τ , is defined for $((y_t^o),t_o)$ in the support of the distribution of $((y_t), \tau(x))$. The conditional distribution

can be described as follows: For $t \leq t_o$, the "conditioned process" simply follows the sample path (y_t^o) of the given, stopped process. After time t_o, the process acts as a Feller process, given by the transition distributions of the original process and the initial state y_∞^o at time t_o, where y_∞^o denotes the constant value of the given stopped sample function after time t_o; the requirement that $((y_t^o), t_o)$ belongs to the support is easily seen to imply that y_t^o is constant for $t > t_o$. In case $t_o = +\infty$, the above description should be modified in the obvious manner (then, y_∞^o is undefined, but not needed in the description, since the description of the first part of the process gives a deterministic process, following the given sample path up to time $+\infty$).

My reason for including the above result (without proof) is that I find it a very illustrative example of the power of the local definition of conditional distributions: Not only does the definition work in this very complex case; it also forces upon us a very clear formulation of the strong Markov property: The (to my opinion, rather obscure) concept of a stopping time-σ-algebra is replaced by an intuitively simple concept of a stopped sample function. The form of the conditional distribution we want to end up with, tells us that we have to condition upon the behaviour of the samplefunction up to time τ, the "right-limit-state" at time τ and the time point τ itself.

35. A CONDITIONAL DISTRIBUTION OF A BIRTH PROCESS.

We shall discuss one more example of a conditioning problem in a stochastic process.

Let X denote the one point compactification of the natural numbers (without 0), i.e.

$$X := \{1,2, \ldots ; \infty\} ,$$

and let T denote the interval $[0,+\infty[$. We shall study a Markov process

$$(x_t) \in (X^T, \mu)$$

given by the initial condition

$$x_0 = 1$$

and the infinitesimal transition probabilities

$$P\{ x_{t+dt} = a+1 \mid x_t = a \} \approx a \cdot dt$$

$$P\{ x_{t+dt} = b \mid x_t = a \} \approx 0 \text{ for } b \neq a, a+1 ;$$

these are the transition specifications for $a \neq \infty$. For $a = \infty$, we must prescribe $P\{x_{t+dt} = \infty \mid x_t = \infty\} = 1$ (i.e. ∞ is an absorbing point) in order to end up with a Feller semigroup, but that is irrelevant for the interpretation

of the process since it never reaches the point ∞ (as follows from the theorem below).

This process is called a <u>birth process</u>, since it obviously describes the behaviour of the size x_t of a population where all individuals produce "children" according to a Poisson process, independently of each other and of their own age.

We shall prove the following theorem:

35.1 Theorem. For almost all samplefunctions

$$(x_t) \in (X^T, \mu)$$

the limit

$$w = \lim_{t \to \infty} \frac{x_t}{e^t}$$

exists. The distribution of the so defined stochastic variable w is a normalized exponential distribution (density e^{-w} on $[0,+\infty[$).

The conditional distribution μ^{w_o} of the process, given $w = w_o$, is defined for all $w_o \in [0,+\infty[$, and it can be described as follows:

The "conditioned process"

$$(x_t^{w_o}) \quad \in \quad (X^T, \mu^{w_o})$$

is a Markov process, defined by the initial condition

$$x_0^{w_o} = 1$$

and the infinitesimal transition probabilities

$$P\{x_{t+dt}^{w_o} = a+1 \mid x_t^{w_o} = a\} \approx w_o \cdot e^t \cdot dt$$

$$P\{x_{t+dt}^{w_o} = b \mid x_t^{w_o} = a\} \approx 0 \quad \text{for} \quad b \neq a, a+1$$

(and with ∞ as absorbing point).

Remarks: It is interesting and (to me, at least) surprising that the jump intensity of the conditioned process depends on the time parameter only, while the intensity of the original process depended on the state only.

The convergence of x_t/e^t specifies the extend to which the classical hypothesis of exponential increase is valid in the stochastic model: The quotient x_t/e^t converges, but not towards a constant. The limit depends on the rather uncertain early history of the population. However, for large values of t, the process becomes stable enough to allow approximation

by an exponential curve, if a correction in terms of a constant factor w or a time delay $-\log w$ is introduced, as indicated by the equation $w = \lim x_t/e^t$ when written on the forms

$$x_t \approx w \cdot e^t$$

and

$$x_t \approx e^{t-(-\log w)} .$$

The theorem was first proved by D.G.Kendall (1966) ; the mixture of the proposed conditional distributions was deduced by a direct computation of the generating function for $(x_{t_1}, \ldots, x_{t_n})$, and the convergence of x_t/e^t follows immediately from the martingale convergence theorem.

W.A.Waugh (1970) has given a proof, based on an expression of w as a function of the waiting times of the process.

The proof given here is based on the fact that the processes considered have the same backwards transition mechanism.

Proof of the theorem: For $w_o \in [0,+\infty[$, let μ_{w_o} denote the proposed conditional distribution. A Markov process of this type, increasing by single steps with an intensity depending on the time parameter only, can always be constructed from a normalized Poisson process by a transformation of the time scale. In the present situation, the measure μ_{w_o} can obviously be characterized as the distribution of

$$(y_{w_o} \cdot (e^t - 1) + 1 \mid t \in T)$$

where $(y_s \mid s \in [0, +\infty[)$ is a normalized Poisson process, that is a Markov process with states 0,1,2, .. given by the initial condition

$$y_0 = 0$$

and the transition specifications

$$P\{y_{s+ds} = a + 1 \mid y_s = a\} \approx ds$$

$$P\{y_{s+ds} = b \mid y_s = a\} \approx 0 \quad \text{for} \quad b \neq a, a+1 .$$

Now we define an auxiliary probability measure γ on

$$[0, +\infty[\times X^T$$

as follows: Let π denote the normalized exponential distribution on $[0, +\infty[$, and put

$$\gamma := \pi[\varepsilon_w \otimes \mu_w]_w .$$

Thus, γ is simply the mixture of the proposed conditional distributions with respect to the proposed distribution of w, with the "mixing variable" w included as an auxiliary stochastic variable.

The mapping

$$w \to \mu_w$$

$$[0,+\infty[\to \mathcal{P}(X^T)$$

is obviously continuous. Hence, according to the decomposition criterion (theorem 24.9, page 206), the conditional distribution of

$$(w, (x_t)) \in ([0,+\infty[\times X^T, \gamma) ,$$

given $w = w_o$, is defined for all $w_o \in [0,+\infty[$ and given by

$$\gamma^{w_o} = \varepsilon_{w_o} \otimes \mu_{w_o} .$$

Now, suppose we are able to prove the following two statements:

(1) The derived process

$$(x_t) \quad , \quad (w,(x_t)) \in ([0,+\infty[\times X^T, \gamma)$$

is a birth process, as defined in the beginning of this section.

(2) For γ-almost all $(w,(x_t))$ we have

$$w = \lim_{t \to \infty} \frac{x_t}{e^t} .$$

Then, the theorem follows immediately by an application of corollary 26.2 (page 217) to the diagram

$$([0,+\infty[\times X^T, \gamma) \longrightarrow (X^T, \mu) \longrightarrow ([0,+\infty[, \pi)$$

$$(w,(x_t)) \longrightarrow (x_t) \longrightarrow \lim \frac{x_t}{e^t} \ (=w)$$

It remains to prove the two statements (1) and (2) concerning the distribution γ:

Proof of (1): It suffices to prove that for all $t_o \in T$ the distribution of

$$(x_t | \ t \in [0,t_o]) \quad , \quad (w,(x_t)) \in ([0,+\infty[\times X^T, \gamma)$$

equals the distribution we would obtain if (x_t) was a birth process. The former is simply the mixture of the distributions of

$$(x_t^w | \ t \in [0,t_o]) \quad , \quad (x_t^w) \in (X^T, \mu_w)$$

with respect to π.

The process (x_t^w) is a Markov process, also when regarded as a process going backwards in time. This means, that the distribution of $(x_t^w | \ t \in [0,t_o])$ is specified by a "final" distribution (the distribution of $x_{t_o}^w$) and a bacwards

transition mechanism.

The distribution of $x^w_{t_o}$ is easily computed: If (y_s) denotes a normalized Poisson process, we have

$$P\{x^w_{t_o} = a\} = P\{y_{w(e^{t_o}-1)} + 1 = a\}$$

$$= \frac{1}{(a-1)!} e^{-w(e^{t_o}-1)} (w(e^{t_o}-1))^{a-1}, \quad a = 1,2, \ldots$$

The backwards jump intensity is deduced from the forwards jump intensity and the distribution of x^w_t by

$$P\{x^w_{t-dt} = a \mid x^w_t = a+1\}$$

$$= P\{x^w_t = a+1 \mid x^w_{t-dt} = a\} \cdot \frac{P\{x^w_{t-dt} = a\}}{P\{x^w_t = a+1\}}$$

$$\approx w \cdot e^{t-dt} \cdot dt \cdot \frac{\frac{1}{(a-1)!} e^{-w(e^{t-dt}-1)} (w(e^{t-dt}-1))^{a-1}}{\frac{1}{a!} e^{-w(e^t-1)} (w(e^t-1))^a}$$

$$\approx w \cdot e^t \cdot dt \cdot \frac{1}{\frac{1}{a} w(e^t-1)} = \frac{a}{1-e^{-t}} \cdot dt .$$

It is crucial, now, that the jump intensity $a/(1-e^{-t})$ <u>does not depend on the "parameter"</u> w ; this reduces the mixing

of the measures μ_w to a trivial matter: Since only the "final" distribution, but not the backwards transition mechanism, depends on the mixing parameter, the mixture comes out, simply, as the Markov process with the same transition mechanism and with the mixture of the final distributions as its final distribution (if the reader does not recognize this result, just translate it into a "forwards" (and well-known) statement).

Hence, the distribution of the derived process (x_t) is given by the bacwards transition mechanism

$$P\{x_{t-dt} = a \mid x_t = a+1\} \approx \frac{a}{1-e^{-t}} \cdot dt$$

and the final distribution given (for $t:=t_0$) by

$$P\{x_t = a\} = \pi[P\{x_t^w = a\}]_w = \int_0^\infty P\{x_t^w = a\}\, e^{-w} dw$$

$$= \int_0^\infty \frac{1}{(a-1)!} \cdot e^{-w(e^t-1)} \cdot (w(e^t-1))^{a-1} e^{-w}\, dw$$

$$= \frac{1}{(a-1)!} (e^t-1)^{a-1} \int_0^\infty w^{a-1}\, e^{-w e^t}\, dw$$

$$= \frac{1}{(a-1)!}(e^t-1)^{a-1}\frac{1}{(e^t)^{a-1}e^t}\int_0^\infty (e^t w)^{a-1}e^{-w}e^t\, d(e^t w)$$

$$= \frac{1}{(a-1)!}(e^t-1)^{a-1}\frac{1}{(e^t)^{a-1}e^t}(a-1)!$$

$$= (1-e^{-t})^{a-1}e^{-t} \quad .$$

The <u>forwards</u> jump intensity can now be deduced:

$$P\{x_{t+dt} = a+1 \mid x_t = a\} = P\{x_t=a \mid x_{t+dt}=a+1\}\frac{P\{x_{t+dt}=a+1\}}{P\{x_t=a\}}$$

$$\approx \frac{a\cdot dt}{1-e^{-t}}\,\frac{e^{-t}(1-e^{-t})^a}{e^{-t}(1-e^{-t})^{a-1}} = a\cdot dt \ .$$

Since we obviously have $P\{x_0 = 1\} = 1$, we conclude that the constructed process (x_t) is a birth process.

In addition, we happened to run into the distribution of x_t in a birth process:

$$P\{x_t = a\} = e^{-t}(1-e^{-t})^{a-1}, \quad a = 1,2, \ldots .$$

<u>Proof of</u> (2): In view of the definition of γ as a mixture, we need only prove (according to theorem A 22 , page 352)

that for all $w > 0$ we have

$$\lim \frac{x_t^w}{e^t} = w \quad \text{for } \mu_w\text{-almost all } (x_t^w) .$$

Introducing the representation of (x_t^w) by a Poisson process (y_s), we must prove that

$$\lim_{t \to \infty} \frac{y_{w(e^t-1)} + 1}{e^t} = w \quad \text{almost surely.}$$

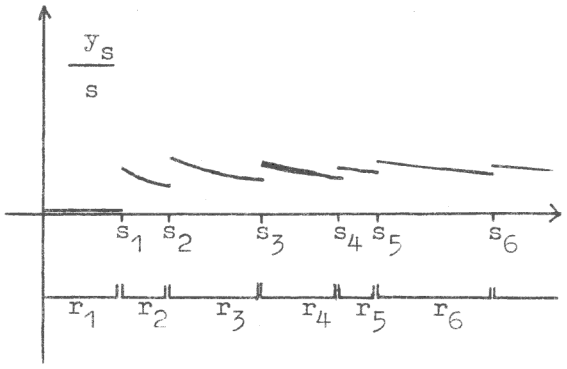
This is elementary: For $w > 0$, the "homogeneous" time parameter $s = w(e^t-1)$ diverges to ∞ for $t \to \infty$, and so it suffices to prove that for almost all Poisson-samplefunctions (y_s) we have

$$\lim_{s \to \infty} \frac{y_s+1}{\frac{s}{w}+1} = w$$

or (equivalently)

$$\lim_{s \to \infty} \frac{y_s}{s} = 1 .$$

This follows from the strong law of large numbers: As a function of s, the quotient $\frac{y_s}{s}$ looks approximately like this:



The jump times s_1 , s_2 , ... are sums of the independent, normalized exponentially distributed waiting times r_1 , r_2 , ...:

$$s_n = r_1 + \ldots + r_n .$$

The left and right limits of y_s/s at the jump times are

$$\lim_{s \uparrow s_n} \frac{y_s}{s} = \frac{n-1}{s_n} = (s_n/(n-1))^{-1}$$

and $$\lim_{s \downarrow s_n} \frac{y_s}{s} = \frac{n}{s_n} = (s_n/n)^{-1} .$$

Since these "extreme values" both converge almost surely to 1 (by the law of large numbers), we obviously have

$$\lim \frac{y_s}{s} = 1 \quad \text{almost surely.}$$

36. CONDITIONING ON A SUM OF INDEPENDENT VARIABLES.

Conditioning on the sum of two variables. Let μ_1 and μ_2 be probability distributions on $\mathbb{R}$, and consider the product measure $\mu := \mu_1 \otimes \mu_2$ on $\mathbb{R}^2$. We are interested in sufficient conditions for existence of the conditional distribution of x_1, given $x_1+x_2 = y$.

First notice, that if μ_1 and μ_2 are given by densities f_1 and f_2 with respect to Lebesgue measure, then, under weak regularity assumptions, the conditional distribution is defined and has the density

$$h^y(z) = \frac{1}{g(y)} f_1(z)f_2(y-z)$$

with respect to Lebesgue measure (cfr. section 31, page 260-261).

The following theorem shows that only one of the two measures μ_1 and μ_2 need be wellbehaved:

36.1 Theorem. Let μ_1 and μ_2 be probability measures on $\mathbb{R}$ and consider the stochastic variable

$$(x_1 , x_2) \in (\mathbb{R}^2, \mu_1 \otimes \mu_2) .$$

Suppose that μ_2 has a <u>continuous</u>, <u>bounded</u> density f_2 with respect to Lebesgue measure.

Then, for all $y_o \in \mathbb{R}$ such that

$$g(y_o) := \mu_1[f_2(y_o-z)]_z > 0 ,$$

the conditional distribution of x_1, given $x_1+x_2 = y_o$, is defined and given by the density

$$\frac{1}{g(y_o)} [f_2(y_o-z)]_z$$

with respect to μ_1.

Notice that the function g is simply the density of the distribution of x_1+x_2 (the convolution of μ_1 and μ_2). Hence, the theorem implies that the conditional distribution is almost everywhere defined.

<u>Proof</u>: Let λ denote Lebesgue measure on $\mathbb{R}$. For $B \to y_o$, we have for any $\mathcal{K}(\mathbb{R})$-function k

$$\mu_1{}^{B}k = (\mu_1 \otimes \mu_2)^{B}[k(x_1)]_{(x_1,x_2)}$$

$$= \frac{(\mu_1 \otimes \mu_2)[1_B(x_1+x_2)k(x_1)]_{(x_1,x_2)}}{(\mu_1 \otimes \mu_2)[1_B(x_1+x_2)]_{(x_1,x_2)}}$$

$$= \frac{\mu_1[k(x_1)\,\mu_2[1_B(x_1+x_2)]_{x_2}]_{x_1}}{\mu_1[\,\mu_2[1_B(x_1+x_2)]_{x_2}]_{x_1}}$$

Dividing nominator and denominator by λB we get (by the translation invariance of Lebesgue measure)

$$\mu_1{}^{B}k = \frac{\mu_1[k(x_1)\,\frac{\mu_2(B-x_1)}{\lambda(B-x_1)}\,]_{x_1}}{\mu_1[\frac{\mu_2(B-x_1)}{\lambda(B-x_1)}\,]_{x_1}}$$

(writing $B-x_1$ for the set $\{y-x_1 | y \in B\}$).

For $B \to y_o$, the integrands of nominator and denominator converge to

$$k(x_1)f_2(y_o-x_1)$$

and

$$f_2(y_o-x_1)$$

respectively. It is not hard to prove, that this convergence is uniform on compact sets (just apply the uniform continuity of f_2 on compact sets), and, since both integrands are bounded (uniformely in B) and μ_1 is a bounded measure, we get

$$\lim_{B \to y_o} \mu_1^{B_k} = \frac{\mu_1[k(x_1)f_2(y_o-x_1)]_{x_1}}{\mu_1[f_2(y_o-x_1)]_{x_1}}$$

and the theorem is proved.

Remark: The result can easily be reformulated as to concern existence of the conditional distributions of (x_1,x_2) or x_2.

Conditioning on the sum of n identically distributed variables. Let μ be a probability measure on $\mathbb{R}$ and consider the stochastic variable

$$(x_1, \dots, x_n) \in (\mathbb{R}^n, \mu \otimes \dots \otimes \mu).$$

We shall study the asymptotic properties of the conditional distribution of x_1, given

$$\frac{x_1+\dots+x_n}{n} = y_o .$$

Conditioning on the expected value. First suppose that

$$Ex_i = \mu[x]_x = y_o = 0 .$$

In this case, a heuristic argument exists: Think of $x_1,\dots,x_n$ as measurements of some property of individuals in a population, with known distribution μ. What does the information

$\bar{x} = 0$ tell us about the first individual of a very large sample? Intuitively, the answer is: Nothing. The mean $\bar{x}$ is going to be close to 0 anyway, due to the condition $Ex_i = 0$ and the law of large numbers. If, say, a very large sample mean $\bar{x}$ was observed, we would, by accident, have drawn a very biased sample, and that might indicate a very large value of x_1, but a sample mean equal to or close to the population mean tells us nothing. Hence, in this special case, we suggest that the conditional distribution of x_1 converges to the unconditioned distribution μ.

Very strong regularity assumptions must be imposed in order to make a true statement out of this. It is easy to construct examples, where the statement is completely false. Suppose, for example, that μ is concentrated at the three points $-\sqrt{2}$, 0 and $\sqrt{3}$, with probabilities $1/(2\sqrt{2})$, $1 - 1/(2\sqrt{2}) - 1/(2\sqrt{3})$ and $1/(2\sqrt{3})$. Then, from the information $\bar{x} = 0$, we conclude that all observations of the sample are 0; in particular, $x_1 = 0$. In fact, the statistic $\bar{x}$ contains information about the complete sample, except, of course, for the order of the observations.

If we assume that μ has a density f with respect to Lebesgue measure, a more refined heuristic argument can be given: Let f_n denote the density of the distribution of $x_1 + \ldots + x_n$, i.e.

$$f_n := f * \ldots * f .$$

Then, by theorem 36.1, the conditional distribution of x_1, given $x_1+(x_2+\ldots+x_n) = 0$, has the density

$$\frac{1}{f_n(0)} f_{n-1}(-x_1)f(x_1)$$

with respect to Lebesgue measure. Now, if the convolution f_{n-1} is reasonably smooth for large values of n, so smooth that we can regard it as a constant, the asymptotic density of the conditional distribution of x_1 becomes $\text{const}\cdot f(x_1)$, or $f(x_1)$. This argument indicates the sort of regularity conditions required for the exact proof:

36.2 Theorem. Let $x_1, \ldots, x_n$ be independent with distribution μ, and suppose that μ has mean 0 and variance 1:

$$E\, x_i = 0 ,$$
$$E\, x_i^2 = 1 .$$

Further, assume that μ has the density f with respect to Lebesgue measure, and let f_n denote the density of the distribution of $x_1+\ldots+x_n$. Suppose that f_n is continuous and bounded for n bigger than some n_o.

Then, from a certain stage $n \geq N_o$, the conditional distribution

$$\mathcal{L}(x_1 \mid x_1+\ldots+x_n = 0)$$

is defined, and for $n \to \infty$

$$\mathcal{L}(x_1 \mid x_1 + \ldots + x_n = 0) \to \mu \ .$$

Remark: The condition that f_n be continuous and bounded from a certain stage is satisfied if f belongs to $L^2(\mathbb{R})$ (and so, in particular, for f bounded). Then the characteristic function $\varphi(s) = \mu[e^{ixs}]_x$ is also square integrable, and so φ^2 is integrable and f_2 thus bounded and continuous (see Feller, vol. II, XV.3 Theorem 3).

More precisely, f_n is bounded for n bigger than some n_0 if and only if φ^n is integrable for n bigger than some n_1 : For φ^n integrable, f_n is bounded and continuous. Conversely, for f_n bounded and continuous, let $\hat{f}_n$ denote the function $\hat{f}_n(x) = f_n(-x)$; then $f_n * \hat{f}_n$ is bounded, continuous and symmetric with characteristic function $|\varphi|^{2n}$, and thus $|\varphi|^{2n}$ (and so φ^{2n}) is integrable (any characteristic function ≥ 0 of a distribution given by a bounded density is integrable, see Feller, vol. II, XV.3 , the corollary to theorem 3).

Proof: We apply a density version of the central limit theorem: Under the assumptions we have imposed on f , it can be proved that the density function $\sqrt{n}\, f_n(z\sqrt{n})$ (for the distribution of $(x_1+\ldots+x_n)/\sqrt{n}$) converges uniformely to the normal density:

$$\left| \sqrt{n}\, f_n(z\sqrt{n}) - \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \right| \to 0$$

uniformely for $z \in \mathbb{R}$. The theorem can be found in Feller, vol. II, XV.5, theorem 2 (1.ed. page 489, 2.ed. page 516); the proof is not difficult.

Since we are more interested in the function f_n, we write the statement on the form

$$\left| \sqrt{n}\, f_n(z) - \frac{1}{\sqrt{2\pi}} e^{-z^2/2n} \right| \to 0$$

uniformely for $z \in \mathbb{R}$. It follows immediately that for n bigger than some N_o we have

$$\begin{cases} f_{n-1} \text{ is bounded and continuous} \\ f_n(0) > 0 \\ \sqrt{n-1}\, f_{n-1}(z) \leq 1 \quad \text{for all } z. \end{cases}$$

Hence, by theorem 36.1 the conditional distribution of x_1, given $x_1+(x_2+\ldots+x_n) = 0$, is defined for $n > N_o$ and given by the density

$$\frac{f_{n-1}(-z)}{f_n(0)} f(z)$$

with respect to Lebesgue measure. For $k \in \mathcal{K}(\mathbb{R})$ we get,

by the dominated convergence principle

$$\frac{1}{f_n(0)} \int f_{n-1}(-z)f(z)k(z)\,dz \approx \frac{\sqrt{n-1}}{\sqrt{n}\, f_n(0)} \int f_{n-1}(-z)f(z)k(z)dz$$

$$\to \frac{1}{1/\sqrt{2\pi}} \int \frac{1}{\sqrt{2\pi}} f(z)k(z)\,dz = \int f(z)k(z)dz = \mu k ,$$

and this proves the theorem.

Conditioning on a biased value of the sum. Now, let us see what happens if we condition on a biased value of $\bar{x}$, i.e. a value different from the mean $E\, x_i$. The answer is fascinating: The conditional distribution of x_1, given $\bar{x}=y$, converges to a distribution given by a density of the form

$$\text{const} \cdot e^{-ax_1}$$

with respect to the unconditioned distribution μ; the parameter a is determined such that the mean of x_1 in the limit distribution equals the biased value we are conditioning on.

Obviously, this result requires very strong regularity conditions: Theorem 36.2 comes out as a special case, and the existence of the proposed asymptotic conditional distribution is also a restrictive assumption.

Firstly, a heuristic argument will be given, in order to explain the somewhat mysterious occurence of the exponential function: Suppose we can choose a such that the function e^{ax_1} is μ-integrable and the probability measure

$$\mu_a := \frac{1}{\mu[e^{ax_1}]_{x_1}} [e^{ax_1}]_{x_1} \cdot \mu$$

has y_o as its centre of gravity. Consider the n-dimensional stochastic variable

$$(x_1,\ldots,x_n) \in (\mathbb{R}^n, \mu_a \otimes \ldots \otimes \mu_a) .$$

Then, according to theorem 36.2, the conditional distribution of x_1, given $\frac{1}{n}(x_1 + \ldots + x_n) = y_o$, exists and converges to μ_a, under suitable regularity conditions. Now, the point is, that the conditional distribution, given the mean, <u>does not depend on</u> a: For $B \to y_o$, the conditional distribution of $(x_1, \ldots, x_n)$, given $\frac{1}{n}(x_1 + \ldots + x_n) \in B$, has the form

$$\text{const.} \cdot 1_{B_o} \cdot [e^{a(x_1+\ldots+x_n)}]_{(x_1,\ldots,x_n)} \cdot (\mu \otimes \ldots \otimes \mu)$$

where

$$B_o := \{(x_1,\ldots,x_n) | \tfrac{1}{n}(x_1+\ldots+x_n) \in B\} .$$

For $B \to y_o$, the density $e^{a(x_1+\ldots+x_n)}$ is approximately

constant on the set B_o, and that makes the limit measure the same as we would get for $a = 0$. Hence, multiplication of μ by the density $\text{const.}e^{ax}$ transfers the problem to the "central" case covered by theorem 36.2, without changing the conditional distributions.

The crucial point of the above argument is an application of a very simple result stating that

> Multiplication of a distribution by a density does not affect its conditional distributions, if the density is a continuous function of the statistic we are conditioning on.

Among statisticians, an immediate consequence of this result is known as Neymann's criterion for sufficiency. The proof is immediate.

The property of the exponential function that makes the argument work is expressed by the fact that multiplication of the factors of the product measure $\mu \otimes \ldots \otimes \mu$ by the exponential density amounts to multiplication of the joint product measure by a function of the sum: The product $e^{ax_1} \ldots e^{ax_n}$ is a function of $x_1 + \ldots + x_n$.

It remains to give an exact formulation of the result:

36.3 Theorem. Let a function

$$f: \mathbb{R} \to [0, +\infty[$$

be given. Then, the real numbers a, such that the function

$$[\, f(z)e^{az}]_z$$

is bounded, constitutes an interval. Let A denote the interior of that interval. For all $a \in A$, the function $[f(z)e^{az}]_z$ is then integrable with respect to Lebesgue measure; define

$$\varphi : A \to \mathbb{R}$$

by

$$\varphi(a) := \int f(z)e^{az}dz \ .$$

For $a \in A$ we denote by μ_a the probability measure given by the density

$$f_a(z) := \frac{1}{\varphi(a)} f(z)e^{az}$$

with respect to Lebesgue measure. For all $a \in A$, the mean

$$\mu_a[z]_z$$

is defined, and the mapping

$$a \to \mu_a[z]_z$$
$$A \to \mathbb{R}$$

is continuous and strictly increasing. Hence, the image of that mapping is an open interval

$$Y := \{ \mu_a[z]_z \mid a \in A \} .$$

For all $y \in Y$ and all $a_o \in A$, the following statement holds:

Let $x_1, \ldots, x_n$ be independent, stochastic variables with distribution μ_{a_o}. Then, for n bigger than some n_o, the conditional distribution

$$\mathcal{L}(x_1 \mid x_1+\ldots+x_n = ny)$$

is defined. For $n \to \infty$, this distribution converges to μ_a, where a is determined by

$$\mu_a[z]_z = y$$

(i.e. a is the "parameter" corresponding to the value y we are conditioning on).

This theorem is, essentially, a special case of a theorem due to Per Martin-Löf (1970). We shall study his approach in

section 37.

Proof: Let a_1 and a_2 , $a_1 < a_2$, be two real numbers such that the function $f(z)e^{az}$ is bounded for $a=a_1$ and $a=a_2$. Then, it is obviously bounded for a between a_1 and a_2 , as for $a \in]a_1, a_2[$ we have

$$f(z)e^{az} = (f(z)e^{a_1 z})e^{-(a_1-a)z} \leq \text{const.}\cdot e^{-(a_1-a)z}$$

and $$f(z)e^{az} = (f(z)e^{a_2 z})e^{-(a_2-a)z} \leq \text{const.}\cdot e^{-(a_2-a)z} .$$

Moreover, we see that $f(z)e^{az}$ is integrable, and even that $z\cdot f(z)e^{az}$, $z^2\cdot f(z)e^{az}$ etc. are integrable, since $f(z)e^{az}$ is dominated by a function pieced together of two exponential tails. The argument shows that the set of points a such that $f(z)e^{az}$ is bounded constitutes an interval, and that the density

$$f_a(z) = \frac{1}{\varphi(a)} f(z)e^{az} \quad , \quad \varphi(a) = \int f(z)e^{az}dz$$

defines a probability measure μ_a with mean, variance etc. for all a in the interior of the interval. It follows from the dominated convergence principle that the mapping $\varphi : A \to \mathbb{R}$ is continuous, and similarly that the mapping

$$a \to \mu_a[z]_z = \int z\, f_a(z)dz$$

is continuous. In order to prove that this mapping is strictly

increasing, just differentiate with respect to a (the derivative is obviously defined, by the dominated convergence principle):

$$\frac{d}{da}\int z\cdot f_a(z)dz = \frac{d}{da}\,\frac{\int z\cdot f(z)e^{az}dz}{\int f(z)e^{az}dz}$$

$$= \frac{(\int z^2 f(z)e^{az}dz)\cdot(\int f(z)e^{az}dz) - (\int z\cdot f(z)e^{az}dz)^2}{(\int f(z)e^{az}dz)^2}$$

$$= \mu_a[z^2]_z - (\mu_a[z]_z)^2 = \mu_a[(z' - \mu_a[z]_z)^2]_z,$$

$$= \mathrm{var}(z) \quad , \quad z \in (\mathbb{R}, \mu_a) .$$

Since the derivative is strictly positive, the mapping $z \to \int z f_a(z)dz$ is strictly increasing.

The interesting part of the theorem follows immediately from theorem 36.2: For $y \in Y$, the density $[f_a(z-y)]_z$ (where $\int z f_a(z)dz = y$) satisfies the conditions of theorem 36.2 (except that the variance need not be 1, but obviously that makes no difference). Hence the conditional distribution

$$\mathcal{L}(x_1 \mid x_1+\ldots+x_n=ny) \quad , \quad (x_1,\ldots,x_n) \in (\mathbb{R}^n, \mu_a\otimes\ldots\otimes\mu_a)$$

is defined for n bigger than some n_o, and it converges to μ_a. From the same stage the conditional distribution

$$\mathcal{L}(x_1 | x_1+\ldots+x_n = ny) \ , \ (x_1,\ldots,x_n) \in (\mathbb{R}^n, \mu_{a_o} \otimes \ldots \otimes \mu_{a_o})$$

is defined and equal to the one above (cfr. the remark on page 316), and this proves the theorem.

37. EXPONENTIAL FAMILIES AND BOLTZMANN'S LAW.

Per Martin-Löf (1970) proves a more general version of theorem 36.3 (page 317). His regularity assumptions are slightly weaker and the one-dimensional measure μ is replaced by an n-dimensional measure. Strictly speaking, our proof is not much different from Per Martin-Löf's proof; also his proof depends crucially on a density version of the central limit theorem, as did the proof of theorem 36.2.

Boltzmann's law. Martin-Löf's approach to the problem arises from his attempt towards an application of statistical-mechanical principles to statistical problems. His result can be sketched as follows:

Let λ be a measure on X, and let

$$t: X \to E$$

be a mapping into a finite dimensional vectorspace E. Suppose that for all n we have a decomposition (cfr. page 36-37)

$$(\lambda_n' , (\lambda_y^{(n)} | y \in E))$$

of $\lambda \otimes \ldots \otimes \lambda$ with respect to the transformation

$$t_n : X^n \to E$$

$$t_n(x_1,\ldots,x_n) = \frac{1}{n}(t(x_1)+\ldots+t(x_n))$$

such that the measures $\lambda_y^{(n)}$ are probability measures. Let $\mu_y^{(n)}$ denote the distribution of x_1 for

$$(x_1,\ldots,x_n) \in (X^n, \lambda_y^{(n)}) .$$

Then, under suitable regularity conditions, $\mu_y^{(n)}$ converges to a distribution μ_a, given by a density of the form

$$\text{const.}e^{a(t(x))}$$

with respect to λ, where the "parameter" a is a linear mapping

$$a: E \to \mathbb{R}$$

determined by

$$\frac{\lambda[t(x)e^{a(t(x))}]_x}{\lambda[e^{a(t(x))}]_x} = y.$$

Similarly, the distribution of $(x_1,\ldots,x_k)$ (for k fixed) converges to $\mu_a \otimes \ldots \otimes \mu_a$ for $n \to \infty$.

Theorem 36.3 comes out as a special case for $X = \mathbb{R}$, $\lambda = \mu_{a_o}$.

In statistical mechanics, a version of the above result is

known as Boltzmann's law. For a large system of particles (typically, the molecules of a gas in a closed container) the behaviour of the system can be described by the uniform distribution on the energy surface in the statespace. The behaviour of one molecule (or a small collection of molecules) is then described by a projection of that uniform distribution. But according to Boltzmann's law this distribution can be described by a density of the form $\text{const} \cdot e^{a \cdot E}$, where E denotes the energy of the particle considered.

Per Martin-Löf proves the theorem under certain regularity conditions in the two cases

$$X = \mathbb{R}^n \quad , \quad \lambda = \text{Lebesgue measure}$$

and

$$X = \mathbb{Z}^n \quad , \quad \lambda = \text{Counting measure} \; .$$

His regularity assumptions seem to be indispensable. However, it is somewhat unsatisfactory that the two cases $X = \mathbb{R}^n$ and $X = \mathbb{Z}^n$ require seperate proofs. It is possible to formulate a more general version of the theorem, which seems to contain Martin-Löf's results as special cases (though not as immediate corollaries). The regularity corditions are existence of the proposed limit distribution μ_{a_o} and an equicontinuity condition on the functions $y \to \mu_y^{(n)}$, $n = 1,2,3,\ldots$.

But first, the relation of the results to mathematical statistics will be explained.

Exponential families; sufficiency. Recall that a statistical model is a family $(\mu_a \mid a \in A)$ of probability measures on a space X ; the idea is that we want to draw inference about the (unknown) parameter a from the observation $x \in (X, \mu_a)$.

By an exponential family we mean a statistical model where A is a subset of the dual to some finite dimensional vector-space E , and μ_a is given by a density of the form

$$\frac{1}{\varphi(a)} e^{a(t(x))} \quad = \quad \text{const.} e^{a(t(x))}$$

with respect to a measure $\lambda \in \mathcal{M}(X)$, where $t: X \to E$ is a given transformation. Then, the normalizing factor is given by

$$\varphi(a) \quad = \quad \lambda \left[e^{a(t(x))} \right]_x .$$

A transformation

$$s: X \to Y$$

is said to be sufficient in the statistical model $(\mu_a \mid a \in A)$, if the conditional distribution of $x \in (X, \mu_a)$, given $s(x) = y$, does not depend on a . This means that $s(x)$ contains all information about a, the remaining "conditional experiment" $x \in (X, \mu_a^y)$ being irrelevant.

Obviously, the transformation t occuring in the exponent

in the definition of the exponential family, is sufficient (if the conditional distributions are defined; this is a consequence of Neymann's criterion, cfr. page 316). Moreover, exponential families have the following nice property: Repeated, independent sampling gives rise to the product model

$$(\mu_a | a \in A)^n := (\mu_a \otimes \ldots \otimes \mu_a | a \in A)$$

with sample space $X \times \ldots \times X$; the form of μ_a obviously ensures that the product model is again an exponential family, with

$$t(x_1)+\ldots+t(x_n)$$

as its sufficient statistic. This means, that combination of samples is easy: The statistics corresponding to the samples should simply be added.

It is a consequence of Boltzmann's law that exponential families have the following property, under certain regularity conditions:

For large samples $x_1, \ldots, x_n$, the conditional distribution of x_1, given the sufficient statistic $t(x_1)+\ldots+t(x_n)$, equals the distribution μ_a of the model with $\mu_a(t) = \frac{1}{n}(t(x_1)+\ldots+t(x_n))$, <u>independently of the true value</u> a_o <u>of the parameter</u>.

I am not in the position to explain why this should be a particularly desirable property; however, the consequences of this property are fascinating, and seemingly closely connected with the fundamental concepts of statistics: Changing the parameter a amounts to conditioning on a biased value of the sufficient statistic in a large sample.

The statistical-mechanical approach to statistics. Per Martin-Löf's approach to mathematical statistics is not explained by the above property of exponential families. He suggests that the model should be specified by a measure (Lebesgue measure or counting measure) and a statistic t, playing the role of sufficient statistic. The first description of the model is in terms of uniform distributions on level surfaces of the sufficient statistic, and the classical "parametrized" models come in as a secondary tool, via Boltzmann's law.

Example: For $X = \mathbb{R}$, $\lambda =$ Lebesgue measure, consider the statistic

$$t: X \to \mathbb{R} \quad (:= E)$$

defined by

$$t(x) := x^2 .$$

This model is determined by the choice of Lebesgue measure as "underlying measure" and the requirement that the square sum $x_1^2+\ldots+x_n^2$ should be sufficient, when a sample of n

independent observations is given. A decomposition of $\lambda \otimes \ldots \otimes \lambda$ with respect to the transformation

$$t_n : \mathbb{R}^n \to \mathbb{R}$$

$$t_n(x_1,\ldots,x_n) = \frac{1}{n}(x_1^2+\ldots+x_n^2)$$

is given by theorem 15.1 (page 130); A trivial modification changes the measures on the level surfaces to probability measures. Thus, we have a decomposition

$$(\lambda_n' , (\lambda_y^{(n)} | \ y \in]0,+\infty[ ))$$

of $\lambda \otimes \ldots \otimes \lambda$ with respect to t_n such that $\lambda_y^{(n)}$ is the uniform distribution on the sphere

$$S_{n-1}(\sqrt{ny}) = \{(x_1,\ldots,x_n) | \ \frac{1}{n}(x_1^2+\ldots+x_n^2) = y \} .$$

By the formula on page 277 (the density of a projection of a uniform distribution on a sphere) the measure $\mu_y^{(n)}$ (the distribution of x_1 for $(x_1,\ldots,x_n) \in (\mathbb{R}^n, \lambda_y^{(n)}))$ has the density

$$\frac{A_{n-2}}{A_{n-1}} \left(1-\left(\frac{x_1}{\sqrt{ny}}\right)^2\right)^{\frac{n-3}{2}} \frac{1}{\sqrt{ny}} = \text{const.}\cdot\left(1-\frac{x_1^2/y}{n}\right)^{\frac{n-3}{2}}$$

with respect to Lebesgue measure on the interval $|x_1| < \sqrt{ny}$.

In this simple case, a direct proof of Boltzmann's law can be given, as indicated by the approximate formulae

$$\text{const}\cdot(1-\frac{x_1^2/y}{n})^{\frac{n-3}{2}} \approx \text{const}\cdot(1-\frac{x_1^2/y}{n})^{n/2}$$

$$= \text{const}\cdot(1-\frac{x_1^2/2y}{n/2})^{n/2} \approx \text{const}\cdot e^{-x_1^2/2y} \quad \text{for} \quad n\to\infty .$$

Thus, the corresponding <u>parametrized</u> model is the family of normal distributions with mean 0 and variance y. The natural parameter is $a = -\frac{1}{2y}$.

Our version of Boltzmann's law looks like this:

<u>37.1 Theorem</u>. Let λ be a measure on X, $t:X\to E$ a continuous mapping of X into a finite dimensional vectorspace E, and (for all n)

$$(\lambda_n' , (\lambda_y^{(n)} | y \in \overline{t_n(X^n)}))$$

a decomposition of $\lambda\otimes\ldots\otimes\lambda$ with respect to the transformation

$$t_n(x_1 , \ldots , x_n) = \frac{1}{n}(t(x_1)+\ldots+t(x_n))$$

such that $(\lambda_y^{(n)} \mid y \in \overline{t_n(X^n)})$ is a family of probability

measures. By $\mu_y^{(n)}$ we denote the distribution of x_1 for $(x_1, \dots, x_n) \in (X^n, \lambda_y^{(n)})$.

Let

$$a_o : E \to \mathbb{R}$$

be a linear mapping such that

$$\varphi(a_o) := \lambda\left[e^{a_o(t(x))}\right]_x < +\infty$$

and define

$$\mu_{a_o} := \frac{1}{\varphi(a_o)}\left[e^{a_o(t(x))}\right]_x \cdot \lambda \quad .$$

Suppose that the expectation

$$y_o := \mu_{a_o}[t(x)]_x$$

of $t(x)$ for $x \in (X, \mu_{a_o})$ exists.

Further, suppose that the following condition is satisfied: For all $\varepsilon > 0$ and for all $k \in \mathcal{K}(X)$ there exists a neighbourhood V_o of y_o and a number n_o such that

$$|\mu_{y'}^{(n)}k - \mu_y^{(n)}k| \leq \varepsilon \quad \text{for } n \geq n_o \,,\; y' \,,\; y \in V_o \cap \overline{t_n(X^n)} \,.$$

Then, $\mu_y^{(n)}$ converges to μ_{a_o} as $n \to \infty$ and $y \to y_o$

in the following sense: For $\varepsilon > 0$, $k \in \mathcal{K}(X)$, there exists a number n_o and a neighbourhood V_o of y_o such that

$$| \mu_y^{(n)} k - \mu_{a_o} k | \leq \varepsilon \quad \text{for} \quad n \geq n_o \quad \text{and} \quad y \in V_o \cap \overline{t_n(X^n)} .$$

Thus, for any sequence (y_n) such that $y_o = \lim y_n$ and $y_n \in t_n(X^n)$, we have $\mu_{y_n}^{(n)} \to \mu_{a_o}$. In particular, if $\mu_{y_o}^{(n)}$ is defined from a certain stage (or just for infinitely many n) we have $\mu_{y_o}^{(n)} \to \mu_{a_o}$.

Proof: Consider the stochastic variable

$$(x_1, \ldots, x_n) \in (X^n, \mu_{a_o} \otimes \ldots \otimes \mu_{a_o}) .$$

By

$$\nu_{a_o}^{(n)} := t_n(\mu_{a_o} \otimes \ldots \otimes \mu_{a_o})$$

we denote the distribution of $t_n(x_1, \ldots, x_n) = \frac{1}{n}(t(x_1) + \ldots + t(x_n))$. By the decomposition criterion (or by theorem 7.1, page 38)

$$t_n : (X^n, \mu_{a_o} \otimes \ldots \otimes \mu_{a_o}) \to (E, \nu_{a_o}^{(n)})$$

has the conditional distributions $\lambda_y^{(n)}$ for $y \in \overline{t_n(X^n)}$; thus, the projected measure μ_{a_o} equals the mixture of the

projected measures $\mu_y^{(n)}$ with respect to $\nu_{a_o}^{(n)}$.

Let $\varepsilon>0$ and $k \in \mathcal{K}(X)$ be given. Choose a neighbourhood V_o of y_o and a number n_1 such that

$$|\mu_{y'}^{(n)}k - \mu_y^{(n)}k| \leq \frac{\varepsilon}{2} \quad \text{for} \quad n \geq n_1 \ , \ y' \ , \ y \in V_o \cap \overline{t_n(X^n)}.$$

By the law of large numbers, we can choose $n_o \geq n_1$ such that

$$\nu_{a_o}^{(n)}(V_o) \geq 1 - \frac{\varepsilon}{2\|k\|_\infty} \quad \text{for} \quad n \geq n_o \ ;$$

Then, for $n \geq n_o$, $y \in V_o \cap \overline{t_n(X^n)}$, we have

$$|\mu_{a_o}k - \mu_y^{(n)}k| = |\nu_{a_o}^{(n)}[\mu_{y'}^{(n)}k - \mu_y^{(n)}k]_{y'}|$$

$$\leq \nu_{a_o}^{(n)}[1_{V_o}(y')|\mu_{y'}^{(n)}k - \mu_y^{(n)}k|]_{y'},$$

$$+ \quad \nu_{a_o}^{(n)}[1_{E\setminus V_o}(y')|\mu_{y'}^{(n)}k - \mu_y^{(n)}k|]_{y'},$$

$$\leq \quad 1 \cdot \frac{\varepsilon}{2} + \frac{\varepsilon}{2\|k\|_\infty} \cdot \|k\|_\infty = \varepsilon \ .$$

This argument proves the theorem.

APPENDIX ON MEASURE THEORY.

The following pages are devoted to a very short outline of the theory of Radon measures, as applied in this exposition. The proofs can be found in (or immediately deduced from results in)

Bourbaki: Intégration, chap. 1-6 ;

or Tue Tjur (1972) (the compact case only) ;

or Tue Tjur (1971) (in Danish) .

Locally compact spaces. A topological space is said to be locally compact if it is a Hausdorff space, and there is a base of compact neighbourhoods of each point. All spaces X, Y etc. in the following are assumed to be locally compact, when nothing else is stated.

The following spaces of realvalued functions on X are considered:

$\mathcal{C}(X)$ the continuous functions

$\mathcal{C}_b(X)$ the bounded, continuous functions

$\mathcal{K}(X)$ the continuous functions with compact support

(the support of a function f is the closure of the set of points x such that $f(x) \neq 0$).

We write

$$\|f\|_\infty := \sup \{ |f(x)| \mid x \in X \}$$

for the supremum norm.

A 1 Theorem. A locally compact space is completely regular:

For

$K \subseteq X$ compact

$A \subseteq X$ closed

$K \cap A = \emptyset$

there exists a $\mathcal{K}(X)$-function k such that

$$1_K \leq k \leq 1_{X \setminus A} .$$

Decomposition of a $\mathcal{K}(X)$-function with respect to an open covering.

A 2 Theorem. Let $\mathcal{U}$ be a set of open sets, covering

the support of the $\mathcal{K}(X)$-function k. Then, there exists a representation

$$k = k_1 + \ldots + k_n$$

of k as a sum of finitely many $\mathcal{K}(X)$-functions k_i, each having support contained in some set from $\mathcal{U}$.

Measures. A (Radon) measure on X is a positive, linear functional

$$\mu : \mathcal{K}(X) \to \mathbb{R}$$

(positive means, of course: $k \geq 0 \implies \mu k \geq 0$).

The set of measures on X is denoted $\mathcal{M}(X)$.

The weak topology on $\mathcal{M}(X)$ is the topology induced by the mappings

$$\left.\begin{array}{ccc} \mu & \to & \mu k \\ \mathcal{M}(X) & \to & \mathbb{R} \end{array}\right\} \; k \in \mathcal{K}(X)$$

The set $\mathcal{M}(X)$ and its subsets $\mathcal{M}_b(X)$ and $\mathcal{P}(X)$ defined below are always regarded as topological spaces in this topology.

A measure μ is said to be bounded, if it is bounded (as a linear functional) with respect to the supremum norm. We

write

$$\|\mu\| \quad \text{or} \quad \|\mu\|_\infty := \sup \{ \mu k \mid \|k\|_\infty \leq 1 \}$$

for the (operator-) norm of μ, also called the total mass of μ.

Measures of total mass 1 are called probability measures. By $\mathcal{M}_b(X)$ and $\mathcal{P}(X)$ we denote the set of bounded measures and the set of probability measures.

A 3 Theorem. The set

$$\{ \mu \in \mathcal{M}_b(X) \mid \|\mu\| \leq 1 \}$$

is compact. If X is compact, also $\mathcal{P}(X)$ is compact.

The support $\operatorname{supp} \mu$ of a measure μ is the complement to the union of the sets $\{x \mid k(x) > 0\}$, where $k \geq 0$, $\mu k = 0$. The complement of the support can also be characterized as the greatest open null set ("null set " is defined later).

A discrete measure is a measure with finite support; or, equivalently: A measure of the form

$$\mu k = \sum_{i=1}^{n} a_i k(x_i) \quad (a_i \geq 0) .$$

The one point measure ε_x at x is defined by

$$\varepsilon_x k := k(x) ;$$

thus, the discrete measures can be described as all positive (finite) linear combinations of one point measures.

A 4 Theorem. The set of discrete measures is dense in $\mathcal{M}(X)$. The set of discrete probability measures is dense in $\mathcal{P}(X)$.

Product measures.

A 5 Theorem. Let $g \in \mathcal{K}(X \times Y)$ be given. Then, there exists a $\mathcal{K}(Y)$-function h_o, such that for all $x_o \in X$ and for all $\varepsilon > 0$, there exists a neighbourhood U of x_o with

$$|g(x_o, y) - g(x, y)| \leq \varepsilon \cdot h_o(y) \quad \text{for} \quad x \in U, \ y \in Y .$$

A 6 Corollary: The mapping

$$x \to [\, g(x,y) \,]_y$$

$$X \to \mathcal{K}(Y)$$

is continuous (for $\mathcal{K}(Y)$ equipped with the supremum-norm topology).

Now, let $\mu \in \mathcal{M}(X)$ and $\nu \in \mathcal{M}(Y)$ be given. It follows from theorem A 5 that the function

$$x \to \nu[\, g(x,y) \,]_y$$

is a $\mathcal{K}(X)$-function. Thus, we can define a measure ξ on $X \times Y$ by

$$\xi g := \mu[\, \nu[\, g(x,y) \,]_y \,]_x \ .$$

This proves the existence part of

A 7 Theorem. There exists one and only one measure ξ on $X \times Y$ such that, for all $k \in \mathcal{K}(X)$ and $h \in \mathcal{K}(Y)$,

$$\xi [\, k(x)h(y) \,]_{(x,y)} = (\mu k)(\nu h) \ .$$

This measure ξ is called the product measure, and denoted

$$\mu \otimes \nu := \xi .$$

A 8 Theorem. The mapping

$$(\mu, \nu) \to \mu \otimes \nu$$

$$\mathcal{M}(X) \times \mathcal{M}(Y) \to \mathcal{M}(X \times Y)$$

is continuous.

A 9 Theorem. For $\mu \in \mathcal{M}(X)$, $\nu \in \mathcal{M}(Y)$, $\pi \in \mathcal{M}(Z)$, we have

$$\mu \otimes (\nu \otimes \pi) = (\mu \otimes \nu) \otimes \pi$$

(under the natural identification $X \times (Y \times Z) = (X \times Y) \times Z$).

Piecing measures together. Let μ be a measure on X, U an open subset of X. The restriction

$$\mu|_U \in \mathcal{M}(U)$$

of μ to U is defined as follows: For $h \in \mathcal{K}(U)$, the extension

$$\hat{h}(x) = \begin{cases} h(x) & \text{for } x \in U \\ 0 & \text{for } x \notin U \end{cases}$$

is a $\mathcal{K}(X)$-function. The transformation

$$h \to \hat{h}$$

is a positive, linear mapping

$$\mathcal{K}(U) \to \mathcal{K}(X) ;$$

thus, we can define a measure $\mu|_U$ on U by

$$(\mu|_U)h := \mu\hat{h} .$$

A 10 Theorem. Let $(U_i | i \in I)$ be a family of open sets covering X, and let, for each $i \in I$, μ_i be a measure on U_i. Suppose that for $i, j \in I$, the restrictions of μ_i and μ_j to $U_i \cap U_j$ coincide. Then, there exists one and only one measure μ on X such that

$$\mu|_{U_i} = \mu_i \quad \text{for all } i \in I .$$

Integration. Let μ be a fixed measure on X. For an arbitrary function

$$f: X \to [-\infty, +\infty]$$

and a set M of $\mathcal{K}(X)$-functions, we write

$$M \uparrow \geq f$$

if M is <u>upwards directed</u> (for k', $k'' \in M$ there exists $k \in M$ such that $k \geq k' \vee k''$) and, for all x

$$\sup \{k(x) \mid k \in M\} \geq f(x) .$$

The μ-<u>norm</u> (or the 1-<u>norm</u>) $\|f\|_\mu$ (or $\|f\|_1$) of f is the greatest lower bound of the numbers

$$\sup \mu M \quad , \quad M \subseteq \mathcal{K}(X) \quad , \quad M \uparrow \geq |f| .$$

Notice, that we may have $\|f\|_\mu = \infty$, and also that we may have $\|f\|_\mu = 0$ for $f \neq 0$.

Now put

$$B(\mu) := \{f: X \to \mathbb{R} \mid \|f\|_\mu < +\infty\} ,$$

$$N(\mu) := \{f: X \to \mathbb{R} \mid \|f\|_\mu = 0\} .$$

It is not hard to prove that $B(\mu)$ and $N(\mu)$ are linear subspaces of $\mathbb{R}^X$.

Restricted to the space $B(\mu)$, the μ-norm becomes a semi-

norm; we can even make it a norm by considering the quotient space $B(\mu)/N(\mu)$. The functional

$$\mu : \mathcal{K}(X) \to \mathbb{R}$$

is bounded with respect to this seminorm (this follows from the formula $\|k\|_\mu = \mu|k|$, which is, in this exposition, the key to integration theory; the proof of this formula requires a compactness argument (an application of Dini's theorem)).

Now, denote by $L(\mu)$ (or $L^1(\mu)$) the closure of the subspace $\mathcal{K}(X) \subseteq B(\mu)$ in the (non-Hausdorff) seminorm topology. We have a unique, $\|\ \|_\mu$-bounded extension

$$\mu : L(\mu) \to \mathbb{R}$$

of μ to the closure of its domain. The functions in $L(\mu)$ are called integrable, and for $f \in L(\mu)$, μf is called the integral of f .

A set $A \subseteq X$ is said to be integrable if its indicator function 1_A is so, and we write

$$\mu A := \mu 1_A$$

for its integral or measure.

A null-set is a set of measure 0 . It can be proved (from the monotone convergence principle below) that a function is a null-function (i.e. a $N(\mu)$-function) if and only if it is 0 except on a null-set.

Integrable functions are, in many connections, considered equivalent if their difference is a null-function, or (equivalently) if they coincide almost everywhere (i.e. everywhere except on a null-set). Equivalent functions are identical with respect to integrals and μ-norms. We may talk about integrability of a function , as soon as we know its values almost everywhere; a similar remark applies to functions taking the values $\pm\infty$ on a set of measure 0 .

The space $L(\mu)$ becomes a Banach space, when equivalent functions are identified (i.e. $L(\mu)/N(\mu)$ is a Banach space, when equipped with the norm (induced by) $\| \ \|_\mu$).

A 11 Theorem (the monotone convergence principle). Let (f_n) be an increasing sequence of integrable functions, such that

$$\lim \mu f_n < +\infty .$$

Then, the function

$$f(x) = \lim f_n(x)$$

is integrable, and

$$\lim \| f-f_n \|_\mu = 0$$

(and, consequently, $\mu f = \lim \mu f_n$).

A 12 Theorem (the dominated convergence principle). Let (f_n) be a sequence of integrable functions and $g \geq 0$ an integrable function such that for all n

$$|f_n(x)| \leq g(x) \quad \text{for almost all } x .$$

Suppose that the limit

$$f(x) = \lim f_n(x)$$

is defined for almost all x . Then, f is integrable, and

$$\lim \| f-f_n \|_\mu = 0$$

(and, consequently, $\mu f = \lim \mu f_n$).

A 13 Theorem. Let M be an upwards directed subset of $\mathcal{K}(X)$ such that

$$\sup \mu M < +\infty .$$

Then, the function

$$f(x) = \sup_{k \in M} k(x)$$

is integrable, with (cfr. page 18)

$$\lim_{\substack{k \uparrow \infty \\ k \in (M, \leq)}} \| k - f \|_\mu = 0$$

(and, consequently, $\mu f = \sup \mu M$).

A 14 Theorem. Any compact set is integrable. For a left (downwards) directed set $\mathcal{P}$ of compact sets (ordered by inclusion) ,

$$\mu(\cap \mathcal{P}) = \inf \mu \mathcal{P}.$$

A 15 Theorem. Suppose that μ is bounded, and let f be bounded and lower semicontinuous. Then, f is integrable. For such a function f the mapping

$$\mu \rightarrow \mu f$$

$$\mathcal{P}(X) \rightarrow \mathbb{R}$$

is also lower semicontinuous.

A 16 Corollary: For $f \in \mathcal{C}_b(X)$, the mapping

$$\mu \rightarrow \mu f$$

$$\mathcal{P}(X) \rightarrow \mathbb{R}$$

is continuous.

Remark: For $f \geq 0$, theorem 15 is a trivial consequence of theorem 13, and in this case, the semicontinuity postulate holds even if $\mathcal{P}(X)$ is replaced by $\mathcal{M}_b(X)$. The more general statement in case of $\mathcal{P}(X)$ follows immediately by addition of a constant function c to f (such that c+f becomes positive). The corollary follows from this result together with the corresponding upper semicontinuity-result.

Measurability. A locally compact space X is said to be σ-compact, if there exists a sequence of compact sets covering X. In the following, the spaces X, Y etc. are assumed to be locally compact and σ-compact. This assumption is rather unrestrictive in practice, and it makes it possible to avoid a tedious distinction between local and global null-sets.

Let μ be a (fixed) measure on X and let

$$t: X \to T$$

be a mapping into an arbitrary topological space. Then, t is said to be measurable with respect to μ, if there exists an increasing sequence (K_n) of compact sets with

$$\mu(X \setminus \bigcup_n K_n) = 0$$

(i.e. the K_n's cover almost all of X), such that the restrictions

$$t|_{K_n} : K_n \to T$$

are continuous.

Remark: In Tue Tjur (1972), the term "almost continuous" was applied for what is here called "measurable", to avoid confusion with Borel-measurability etc. I have changed the terminology, because the measurability definition above almost removes the need of "measure-independent" measurability concepts. Borel sets and Borel functions, the basic elements of "abstract" measure theory, play a minor role in the theory of Radon measures (namely the role of universal measurability criterions in case T has a denumerable base).

Notice, that any continuous mapping is measurable. Also mappings which are continuous at almost all points, or

(slightly weaker) continuous when restricted to the complement of some null set, are measurable.

Measurability is a property of the equivalence class, in the sense that a change of t on a null set does not affect measurability. This is a consequence of theorem A 18 below.

Notice that if μ is bounded, the following condition is equivalent to measurability: For all $\varepsilon > 0$ there exists a compact set K with $\mu(X \setminus K) < \varepsilon$, such that the restriction of t to K is continuous.

A 17 Theorem. A function

$$f: X \to [-\infty, +\infty]$$

is integrable if and only if it is measurable with

$$\|f\|_\mu < +\infty .$$

A 18 Theorem (regularity of Radon measures). Let $A \subseteq X$ be integrable. Then, for $\varepsilon > 0$, there exists a compact set $K \subseteq A$ such that

$$\mu(A \setminus K) < \varepsilon .$$

A set $M \subseteq X$ is said to be measurable if its indicator function 1_M is measurable.

A 19 Theorem. $M \subseteq X$ is measurable if and only if for each $\varepsilon > 0$ there is a closed set F and an open set U such that

$$F \subseteq M \subseteq U$$

and

$$\mu(U \setminus F) \leq \varepsilon .$$

The Hilbert space $L^2(\mu)$. Let $L^2(\mu)$ denote the set of real (or extended real) functions f on X such that f is measurable and f^2 is integrable (i.e. $\|f^2\|_\mu < +\infty$). Equivalent functions are identified.

For $f, g \in L^2(\mu)$ put

$$(f|g)_\mu := \mu(f \cdot g) .$$

This defines an inner product $(\ |\ )_\mu$, giving $L^2(\mu)$ structure as a Hilbert space. The norm is called the 2-norm, written

$$\|f\|_2 = \sqrt{(f|f)_\mu} .$$

Transformation. Let μ be a measure on X and let

$$t: X \to Y \qquad (Y \text{ locally compact and } \sigma\text{-compact})$$

be a μ-measurable transformation such that

$$h \circ t \in L(\mu) \quad \text{for} \quad h \in \mathcal{K}(Y) .$$

The transformed measure $t(\mu) \in \mathcal{M}(Y)$ is defined by

$$t(\mu)h := \mu(h \circ t) \quad \text{for} \quad h \in \mathcal{K}(Y) .$$

Density. Let μ be a measure on X. A function f on X is said to be locally integrable if

$$f \cdot k \in L(\mu) \quad \text{for all} \quad k \in \mathcal{K}(X) .$$

A locally integrable function is measurable.

A density d is a locally integrable function $d \geq 0$. For a density d, define the measure $d \cdot \mu$ on X by

$$(d \cdot \mu)k := \mu(d \cdot k) .$$

Mixture. Let

$$x \to \nu_x$$

$$X \to \mathcal{M}(Y)$$

be a measurable mapping such that

$$[\nu_x h]_X \in L(\mu) \quad \text{for all} \quad h \in \mathcal{K}(Y) .$$

Define a measure $\mu[\nu_x]_X$ on Y by

$$(\mu[\nu_x]_X)h := \mu[\nu_x h]_X .$$

This measure is called the mixture of the measures ν_x with respect to μ.

Preservation of integrals under linear operations. The following three theorems give integrability criterions for measures constructed by transformation, multiplication by density or mixing.

A 20 Theorem. Let

$$t: X \to Y$$

be μ-measurable, and such that the transformed measure $t(\mu)$ is defined. Then, a function g on Y is $t(\mu)$-integrable if and only if $g \circ t$ is μ-integrable; if so,

$$t(\mu)g = \mu(g \circ t) .$$

A 21 Theorem. Let $d \geq 0$ be locally μ-integrable. A function f on X is $(d \cdot \mu)$-integrable if and only if $d \cdot f$ is μ-integrable; if so,

$$(d \cdot \mu) f = \mu(d \cdot f) .$$

A 22 Theorem. Let

$$x \to \nu_x$$

$$X \to \mathcal{M}(Y)$$

be a μ-measurable mapping, satisfying the condition for existence of the mixture $\mu[\nu_x]_x$. Let g be a $\mu[\nu_x]_x$-integrable function on Y. Then, for μ-almost all x, g is ν_x-integrable. The (almost everywhere defined) function

$$[\nu_x g]_x$$

is μ-integrable with

$$\mu[\nu_x g]_x = (\mu[\nu_x]_x) g .$$

A 23 Corollary (Fubini's theorem). Let g be a $\mu \otimes \nu$-integrable function on $X \times Y$. Then, for μ-almost all

x , the function $[g(x,y)]_y$ is ν-integrable, and the (almost everywhere defined) function $[\nu[g(x,y)]_y]_x$ on X is μ-integrable with

$$\mu[\nu[g(x,y)]_y]_x = (\mu \otimes \nu)g \ .$$

Probability measures on a compact product space. Let $(X_i | i \in I)$ be a family of compact spaces. Then, the product space

$$X_I := \prod_{i \in I} X_i$$

is compact (by Tychonoff's theorem). For $I_1 \subseteq I$, put

$$X_{I_1} := \prod_{i \in I_1} X_i \ .$$

For subsets I_1 and I_2 of I such that $I_1 \subseteq I_2$, we denote by

$$p_{I_2 I_1} : X_{I_2} \to X_{I_1}$$

the projection

$$p_{I_2 I_1}(x_i | i \in I_2) = (x_i | i \in I_1) \ .$$

Now, let μ_I be a probability measure on X_I and let $\mathcal{P}_o$ denote the set of finite subsets of I . For $M \in \mathcal{P}_o$, put

$$\mu_M := p_{IM}\,\mu_I \ .$$

The family $(\mu_M \mid M \in \mathcal{S}_o)$ of finite dimensional marginal distributions for μ_I satisfies the following consistency condition:

For $N \subseteq M \in \mathcal{S}_o$,

$$p_{MN}\,\mu_M = \mu_N \ .$$

Conversely, we have

A 24 Theorem (Kolmogorov's consistency theorem). Let $(\mu_M \mid M \in \mathcal{S}_o)$ be a family of probability measures $\mu_M \in \mathcal{P}(X_M)$, satisfying the above consistency condition. Then, there exists one and only one measure μ_I on X_I such that

$$p_{IM}\,\mu_I = \mu_M \qquad \text{for } M \in \mathcal{S}_o \ .$$

This theorem relies on the fact that the space of continuous functions $f: X_I \to \mathbb{R}$ depending on a finite number of coordinates only (i.e. the functions of the form $k_M \circ p_{IM}$, $k_M \in \mathcal{K}(X_M)$) is dense in $\mathcal{K}(X_I)$ $(= \mathcal{C}(X_I))$.

DANSK RESUME.

Med udgangspunkt i den formulering af sandsynlighedsregningen der baseres på Radon mål på lokalkompakte rum, indføres betingede fordelinger ved en differentiationslignende proces. Eksistens af sådanne „lokale" betingede fordelinger punktvis, overalt eller næsten overalt, diskuteres i nogle specialtilfælde, herunder i tilfælde af sædvanlige „kontinuerte" fordelinger på reelle talrum (differentialgeometrisk formulering) og i tilfælde af stokastiske processer. Lokale betingede fordelingers globale egenskaber diskuteres. Herunder vises det, at næsten overalt definerede lokale betingede fordelinger har de egenskaber der forudsættes i den sædvanligvis benyttede definition. Forskellige regneregler for betingede fordelinger (succesiv betingning, ombytning af betingningsoperationer) bevises. Endelig anvendes den lokale definition på mere konkrete problemer, hvoraf især kan fremhæves følgende:

Udledning af visse fordelinger i tilknytning til den flerdimensionale normale fordeling.

En formulering af den stærke Markovegenskab for Fellerprocesser (kun gengivet summarisk, med henvisning til Tue Tjur (1972)).

Bevis for et resultat af D.G.Kendall vedrørende betingning i en fødselsproces.

Diskussion af asymptotiske forhold ved betingning med en sum af uafhængige, identisk fordelte stokastiske variable.

En generel formulering af et resultat af Per Martin-Löf (Boltzmanns lov).

LITTERATURE.

Erik Sparre Andersen, Børge Jessen:
ON THE INTRODUCTION OF MEASURES IN INFINITE PRODUCT SETS.
Danske Videnskabernes Selskab, Matematisk-Fysiske Meddelelser 25, 1948, No. 4 .

Nicolaus Bourbaki:
ÉLÉMENTS DE MATHÉMATIQUE,

Topologie Générale	chap. 1-2	1965
Intégration	chap. 1-4	1965 (1952)
Intégration	chap. 5	1956
Intégration	chap. 6	1959
Variétés Différentielle et Analytiques. Fascicule de Résultats	§§ 1-7	1967
	§§ 8-15	1971

Paris, Hermann.

Leo Breiman:
PROBABILITY.
Addison-Wesley 1968.

Hans Brøns:
LECTURE NOTES 1967-68 : DEN NORMALE FORDELINGS TEORI.
Unpublished.

J. Dieudonné:
ÉLÉMENTS D'ANALYSE 3 .
Gauthier-Villars 1970.

J. L. Doob:
STOCHASTIC PROCESSES.
Wiley 1953 .

Nelson Dunford, Jakob T. Schwartz:
LINEAR OPERATORS, Part I .
Interscience Publishers 1957.

Herbert Federer:
GEOMETRIC MEASURE THEORY.
Springer 1969.

William Feller:
AN INTRODUCTION TO PROBABILITY THEORY AND ITS APPLICATIONS, vol. 2.
Wiley 1971 (1966).

Maurice Fréchet:
SUR L'INTÉGRALE D'UNE FONCTIONELLE ÉTENDUE A UN ENSEMBLE ABSTRAIT.
Bulletin de la Société Mathématique de France 43, 1915, 248-265.

H. Hahn, A. Rosenthal:
SET FUNCTIONS.
University of New Mexico Press, Albuquerque 1948.

P. R. Halmos:
MEASURE THEORY.
Van Nostrand, New York 1950.

Sigurdur Helgason:
DIFFERENTIAL GEOMETRY AND SYMMETRIC SPACES.
Academic Press 1962.

Noel J. Hicks:
NOTES ON DIFFERENTIAL GEOMETRY.
Van Nostrand 1965.

A. T. James:
NORMAL MULTIVARIATE ANALYSIS AND THE ORTHOGONAL GROUP.
Annals of Mathematical Statistics 25, 1954, 40-75.

Søren Johansen:
ANVENDELSE AF EKSTREMALPUNKTSMETODER I SANDSYNLIGHEDSREGNINGEN.
Københavns Universitet Institut for Matematisk Statistik 1967.

D. G. Kendall:
BRANCHING PROCESSES SINCE 1873.
Journal of the London Mathematical Society 41, 1966, 385-406.

A. N. Kolmogorov:
GRUNDBEGRIFFE DER WAHRSCHEINLICHKEITSRECHNUNG.
Ergebnisse der Mathematik und Ihrer Grenzgebiete, Springer, 1933.

A. N. Kolmogorov:
FOUNDATIONS OF THE THEORY OF PROBABILITY.
Chelsea, New York 1956 (translation of Kolmogorov (1933)).

Henri Lebesgue:
LEÇONS SUR L'INTÉGRATION ET LA RECHERCHE DES FONCTIONS PRIMITIVES.
Gauthier-Villars, Paris 1904.

Henri Lebesgue:
SUR L'INTÉGRATION DES FONCTIONS DISCONTINUES.
Annales scientifiques de l'école normale supérieure (3) 27,
1910, 361-450.

M. Loève:
PROBABILITY THEORY.
Van Nostrand 1960 (1955).

Saunders MacLane, Garrett Birkhoff:
ALGEBRA.
The Macmillan Company 1967.

Per Martin-Löf / Rolf Sundberg:
STATISTISKA MODELLER.
Institutet för försäkringsmatematik och matematisk statistik
vid Stockholms universitet 1969-70.

Barry Mitchell:
THEORY OF CATEGORIES.
Academic Press 1965.

E. Nelson:
REGULAR PROBABILITY MEASURES ON FUNCTION SPACES.
Annals of Mathematics vol. 69, 1959, 630-643.

J. Neveu:
BASES MATHEMATIQUES DU CALCUL DES PROBABILITÉS.
Masson et C^{ie} , Paris 1964.

J. Radon:
THEORIE UND ANWENDUNGEN DER ABSOLUT ADDITIVEN MENGENFUNKTIONEN.
S.-B. Akad. Wiss., Wien, 122, 1913, 1295-1438.

F. Riesz:
SUR LES OPÉRATIONS FONCTIONELLE LINÉAIRES.
C. R. Acad. Sci. Paris 149, 1909, 974-977.

S. Saks:
THEORY OF THE INTEGRAL.
Warsaw 1937. Reprinted Stechert-Hafner Pub. Co., New York.

Laurent Schwartz:
LA FONCTION ALÉATOIRE DU MOUVEMENT BROWNIEN.
Séminaire Bourbaki 1957/58, 161.

Tue Tjur:
MÅL PÅ LOKALKOMPAKTE RUM.
Københavns Universitet Institut for Matematisk Statistik 1971.

Tue Tjur:
ON THE MATHEMATICAL FOUNDATIONS OF PROBABILITY.
Institute of Mathematical Statistics, University of Copenhagen
Lecture Notes No. 1 , 1972.

W. A. Waugh:
TRANSFORMATION OF A BIRTH PROCESS INTO A POISSON PROCESS.
Journal of the Royal Statistical Society B 31, 1970, 418-431.

Hassler Whitney:
GEOMETRIC INTEGRATION THEORY.
Princeton University Press 1957.

SYMBOLS.

Standard notation.

symbol	use	meaning
[ ]	$[a,b]$, $[a,b[$	closed and half open interval.
[ ]	$[f(x)]_x$	see page 12
[ ]	$\mu[\nu_x]_x$	see page 351
:=		see page 15
1	1_A	see page 15
$\mathcal{L}$	$\mathcal{L}(x)$, $\mathcal{L}(x\|y)$	see page 14,15
$\to$	$x \to f(x)$ $X \to Y$	denotes the mapping $f: X \to Y$.
$\to$	$x_\alpha \to x_0$	$x_0 = \lim x_\alpha$.
$\|$	$f\|_A$	the restriction of f to the set A .
$\|$	$\mu\|_U$	see page 339
$\otimes$	$\mu \otimes \nu$	see page 338-339
$\|\| \ \|\|$	$\|\|\mu\|\|$	see page 336
$\|\| \ \|\|_\infty$	$\|\|f\|\|_\infty$	see page 334
$\|\| \ \|\|_\infty$	$\|\|\mu\|\|_\infty$	see page 336
$\|\| \ \|\|_1$	$\|\|f\|\|_1$	see page 341
$\|\| \ \|\|_\mu$	$\|\|f\|\|_\mu$	see page 341
$\|\| \ \|\|_2$	$\|\|f\|\|_2$	see page 349
$(\ \| \ )_\mu$	$(f\|g)_\mu$	see page 349
$-$	$\bar{A}$	the closure of A .

$*$	$\mu * \nu$, $f * g$	convolutions
$*$	a^*	adjoint mapping (see page 64)
$\vee$	$a \vee b$	$= \max\{a,b\}$
$\wedge$	$a \wedge b$	$= \min\{a,b\}$
$'$	A'	transposed matrix
$\mathcal{C}$	$\mathcal{C}(X)$	see page 333
$\mathcal{C}_b$	$\mathcal{C}_b(X)$	see page 333
E	E x	the expectation of the stochastic variable x
$\mathcal{K}$	$\mathcal{K}(X)$	see page 333
L	$L(\mu)$	see page 342
L^1	$L^1(\mu)$	see page 342
L^2	$L^2(\mu)$	see page 349
L^2	$L^2(t)$	see page 167
$\mathcal{M}$	$\mathcal{M}(X)$	see page 335
$\mathcal{M}_b$	$\mathcal{M}_b(X)$	see page 336
$\mathcal{P}$	$\mathcal{P}(X)$	see page 336
$\mathbb{R}$		the real axis
$\dot{Z}$ ($\mathbb{N}$)		the (positive) integers
ε	ε_x	see page 337
var	var(x)	the variance of the stochastic variable x
$t(\mu)$	$t(\mu)$	transformed measure, see page 350
$d \cdot \mu$	$d \cdot \mu$	see page 350
μA	μA	see page 342

Special notation, introduced in the text:

INDEX.

Tue Tjur

Conditional Probability Distributions

Published by the
Institute of Mathematical Statistics
University of Copenhagen

5 Universitetsparken, 2100 Copenhagen Ø, Denmark

Net Price D.Kr. 50.–