Statistical Decision Rules and Optimal Inference

N. N. Čencov
Statistical Decision Rules and Optimal Inference
Translations of

MATHEMATICAL MONOGRAPHS

Volume 53

Statistical Decision Rules and Optimal Inference

N. N. Čencov

American Mathematical Society
Providence, Rhode Island
ABSTRACT. This monograph is devoted to the general theory of statistical inference. The approaches developed here permit the author to consider, from a single point of view, the main concepts and laws of mathematical statistics, methods of constructing optimal statistical estimates, etc. The book is intended for those working in mathematical statistics, information theory, game theory, and also applications of probabilistic and statistical methods. Individual sections may be of interest to specialists in measure theory, differential geometry and nonlinear functional analysis.

Library of Congress Cataloging-in-Publication Data
Chentsov, N. N. (Nikolai Nikolaevich)
Statistical decision rules and optimal inference.
(Translations of mathematical monographs; 53)
Translation of: Statisticheskie reshaishchie pravila i optimal'nye vyvody.
Bibliography: p.
Includes index.
1. Statistical decision. 2. Distribution (Probability theory) I. Title. II. Series.
QA279.4.C4613 519.5'42 81-15039
ISBN 0-8218-4502-0 AACR2
ISSN 0065-9262

AMS softcover ISBN 978-0-8218-1347-8

Copyright © 1982 by the American Mathematical Society
Printed in the United States of America.

The paper used in this book is acid-free and falls within the guidelines established to ensure permanence and durability.
Information on copying and reprinting can be found in the back of this volume.
Visit the AMS home page at http://www.ams.org/
10 9 8 7 6 5 4 3 13 12 11 10 09 08
# TABLE OF CONTENTS

PREFACE vii

INTRODUCTION 1
1. Statistical problems and statistical decisions 1
2. Probability measures and conditional distributions 13
3. Smooth manifolds and their mappings 37
4. Category theory and geometry 50

CHAPTER I. THE FORMAL DECISION PROBLEM 65
5. The category of statistical decisions 65
6. Markov geometry of families of probability distributions 76
7. Geometry of dominated families of probability distributions 92
8. Invariant information characteristics 113

CHAPTER II. EQUIVARIANT DIFFERENTIAL GEOMETRY OF A COLLECTION OF PROBABILITY DISTRIBUTIONS 127
9. Geometry of the simplex of probability distributions on a finite algebra 127
10. Projective geometry of a collection of probability distributions 140
11. Invariant Riemannian metric on a manifold of probability distributions 156
12. Geodesic mean of probability distributions 165

CHAPTER III. SMOOTH FAMILIES OF PROBABILITY DISTRIBUTIONS AND THE INFORMATION INEQUALITY 185
13. Finite-dimensional approximation of infinite-dimensional manifolds of probability distributions 185
14. Differentiable families of probability distributions 199
15. The information inequality 213
16. Continuously differentiable families of probability distributions 228
CHAPTER IV. GEOMETRY OF EXPONENT FAMILIES OF
PROBABILITY DISTRIBUTIONS
17. Convex functions and the Legendre transformation 245
18. Exponent families of probability distributions 262
19. Natural parametrization of exponent families 279
20. Conjugate parametrizations of an exponent family of probability
distributions 290
21. Distributions of values of the directional statistic of an exponent
family and related families 303
22. Nonsymmetric Pythagorean geometry of the information devia-
tion 319
23. Charts of an exponent family of probability distributions 334

CHAPTER V. OPTIMAL DECISION RULES IN THE EQUIV-
ARIANT PROBLEM OF POINT ESTIMATION
24. Estimation of the unknown mean of a multivariate normal dis-
tribution 355
25. Estimation of the unknown density of a probability
distribution 355
26. Invariant loss functions in problems of mathematical statistics 370
27. Optimal estimators for smooth families of probability distribu-
tions 391
28. Quasi-homogeneous families of probability distributions 403

APPENDIX
29. Random probability measures 437

NOTES AND COMMENTS 451

BIBLIOGRAPHY 465

INDEX 477
PREFACE

The general concepts of a statistical decision and a statistical decision rule are basic for all of modern statistical theory. According to Wald, every particular statistical problem is a problem of decision-making: the statistician, having processed certain observational material, must draw conclusions as to the observed phenomenon. Since the outcome of each observation is random, one cannot usually expect these conclusions to be absolutely accurate. It is a job for the theory to ascertain the minimal unavoidable uncertainty of the conclusions in the problem and to indicate an optimal decision rule.

In classical problems of mathematical statistics, one is required to determine the (unknown) probability distribution of the outcomes on the basis of independent observations and certain additional information. When the number of observations used in such cases increases, one can establish various simple and quite general asymptotic relationships.

In any theory, a general law should be amenable to equivalent formulations; that is to say, the statement of the law should not vary when a situation is replaced by another one, equivalent to the former (within the framework of that theory); otherwise it would not be a general law. In classical geometry, such "changes of situation" form a group. In mathematical statistics the description of the set of equivalent situations is more complicated.

The system of all statistical decision rules for all conceivable statistical problems, together with the natural operation of composition, forms an algebraic category. This category generates a uniform geometry of families of probability laws, in which the "figures" are the families and the "motions" are the decision rules. Two families are "congruent" if and only if they possess equivalent statistical properties.

An apt name for the subject of this monograph might be "geometrical statistics". The algebra of decision rules and the natural geometry that it
generates are studied here from a statistical standpoint. The geometrical methods and language that we develop are then applied to the equivariant theory of optimal estimates.

The book is intended for specialists in mathematical statistics, information theory and game theory, and also for those interested in applications of probability-theoretical methods. The reader is expected to be familiar with probability theory and measure theory, to the extent of Kolmogorov's *Grundbegriffe der Wahrscheinlichkeitsrechnung* and Halmos' *Measure Theory*, or Neveu's *Bases Mathématiques du Calcul des Probabilités*. Formally speaking, a knowledge of statistics is not assumed; all the necessary concepts are introduced and explained in §1 of the Introduction. Also included in the Introduction is the requisite material from the theory of smooth manifolds and category theory.

My interest in the "uncertainty" of statistical estimates arose when I was engaged in the development of computer-oriented methods for the estimation of unknown densities; further encouragement came from my teacher N. V. Smirnov. Without his approval, I would probably not have risked taking my research so far from orthodox statistics.

My discussions with Ju. V. Linnik had considerable influence on my work. Many of the theorems proved here are answers to his clearly phrased questions. I am indebted to him for his unflagging interest in my research and his kind attention.

I had many useful discussions with experts concerning both the basic ideas of the theory and individual results; thanks are due in this respect to L. N. Bol'sev, B. N. Delone, I. M. Gel'fand, B. V. Gnedenko, A. M. Kagan, A. N. Kolmogorov, A. A. Ljapunov, Ju. V. Prohorov, Richard Sacksteder, Ju. M. Smirnov, Charles Stein and V. S. Vladimirov.

It is my pleasure to thank Elena Aleksandrovna Morozova for her great help and valuable advice on geometrical matters, and to Fridrih Izrailevič Karpelevič for numerous profound and important remarks.

Finally, I wish to thank the book's editor, A. V. Černavskiǐ, who read the manuscript attentively and helped to eliminate various shortcomings.

*N. N. Čencov*
NOTES AND COMMENTS

§1

A general description of equivalent statistical problems was given by Blackwell [2, 1951], [3, 1953]. At the same time, similar ideas were developed by Stein [1951, oral communication; see 3] and Sherman [1, 1951]. Earlier work had considered only equivalent reduction of statistical problems connected with the transition to sufficient statistics; see Fisher [3, 1925], and also Halmos and Savage [1, 1949], Bahadur [1, 1954], Burkholder [1], and LeCam [2, 1964]. [3, 1969]. Čencov [5, 1965] and Morse and Sacksteder [1, 1966] noticed that this equivalence relation was, roughly speaking, generated by the category of statistical decision rules. This enabled them to "algebraicize" certain concepts of statistics. Further steps in this direction were taken by Romier and coworkers in their "Introduction à la statistique mathématique" (Romier [2], [3], Littaye-Petit et al. [1], Martin and Vaugelsy [1], Laurant et al. [1], Martin et al. [1]). The idea of considering families of probability distributions not as objects on their own but as "figures" in a Kleinian geometry with a category of transformations is due to the author [5], [7].

The systematic investigation of statistical properties which are invariant (equivariant) in the category of statistical decision rules was initiated by Sacksteder and Čencov. Attention was paid previously only to invariance determined by the (group) symmetry of a specific problem; see Lehmann [1]; also Berk [1] and Brillinger [1].

(1) The history of statistics as the science of statistical inference usually begins with the amusing episode recounted by Bertrand in the preface to his course "Calcul des probabilités" [1]:

"One day in Naples the reverend Galvani saw a man from the Basilicata who, shaking three dice in a cup, wagered to throw three sixes . . . . Such luck is possible, you say. Yet the man succeeded a second time, and the bet was repeated. He put back the dice in the cup, three, four, five times, and each time he produced three sixes. "Sangue di Bacco", exclaimed the reverend, 'the dice are loaded!' And they were . . . ." [Quoted from G. Pólya, Patterns of Plausible Inference.]

(2) This approach was first put forth explicitly by Neyman and Pearson [1, 1928] in their theory of hypothesis testing.

(3) Essentially, this is a special case of the game-theoretical approach. Long ago Laplace [1, 1820] likened the derivation of an estimate to a game of chance in which the statistician suffers defeat if his estimates are inferior.

A detailed description of the concept of "statistical problem" according to Wald [1], [3] was given by Lehmann [1]. See also Birnbaum [1], De Groot and Rao [1], and Hoeffding [1].
466

NOTES AND COMMENTS

(4) Choice of action based on results of an auxiliary experiment was first systematically considered by von Neumann in game theory. See von Neumann [1, 1928], and also von Neumann and Morgenstern [1] and Blackwell and Girshick [1] (cf. footnote (3), above).

(5) The statistician deals directly with the outcomes of the experiment only in the simplest problems, involving a discrete sample space. The slightest additional complication in the problem leads to distortion of the experimental result, due to measurement errors and grouping (when of necessity the measurement is rounded off). In such situations the decision is described by a compound decision rule \( II_{\text{instrument}} \cdot II_{\text{processing}} \), where the first factor is independent of the statistician. In this book we shall consider only the ideal situation, in which the distortions due to \( II_{\text{instrument}} \) may be neglected.

§2

(1) Cap = Collection of All Probability distributions.

(2) In all topological concepts we follow Kelley [1].

(3) In §13 we shall introduce the Hellinger integrals of the Radon-Nikodým derivative, such as

\[ \int [(dQ/dP)(\omega)]^2 P(\omega) \text{d} \omega, \]

as the limits of integral sums of a special form, with specific conventions for resolution of the indeterminacy \( \infty \cdot 0 \), different from the usual.

(4) See the work of C. Ionescu Tulcea, under whose "lifting" \( P(\cdot) \rightarrow \rho(\cdot) \), to a (finite) linear combination of laws corresponds the linear combination of the densities, and to a geodesic mean (see Definition 18.2) corresponds the normalized geometric mean of the densities. The correspondence may fail for countable linear combinations (see end of §9.4).

(5) The possibility of modifying almost proper conditional distributions to obtain proper distributions depends only on the properties of the mapping \( f \). If there exists a proper conditional distribution for at least one \( Q \), it will serve to modify any \( P(\cdot|f) \).

Blackwell and Ryll-Nardzewski [1] proved that not every almost proper Borel conditional distribution can be modified by a Borel procedure to obtain a proper conditional distribution. Their arguments rely on the fact, established by Novikov [1], that it is not possible to B-uniformize an arbitrary Borel function (see subsection 14).

(6) This approach to the theory of stochastic processes was developed by Lévy [1]; see also Wiener [1]. For the connection with the theory of Monte Carlo methods, see Gel'fand, Frolov and Cencov [1], Rankin [1] and Čencov [11]. A similar approach was suggested by Blackwell [4] (see also Sazonov [1]).

(7) Provided that the initial function is not many-valued. Nevertheless, any B-function may be uniformized by an \( \mathcal{A}(B) \)-function, where \( \mathcal{A}(B) \) is the \( \sigma \)-algebra generated by the analytic sets (see Saks [1]) \( B \subset \mathcal{A}(B) \subset B^* \). See also Luzin [1], Luzin and Novikov [1], and Luzin and Sierpinski [1].

(8) The proofs of Lemma 5.11 and all other assertions of §5 rely only on general facts of measure theory and make no use of the concept of a constructive measure. Therefore the proof given here of Lemma 2.15 contains no vicious circle.

§3

(1) Applying the Taylor formula with integral remainder to the function \( f(x) = f(q_\alpha^{-1}(x)) \) of the local coordinates \( (x^{(1)}, \ldots, x^{(n)}) \), we can write (see Helgason [1])

\[ f(x) = f(x_0) + \sum_{i=1}^n (x^{(i)} - x^{(i)}_0)g_i(x), \]

where \( g_i(x) \) is a differentiable function, \( g_i(x) = X_i f(x) \), \( g_i(x_0) = (X_i f)_p \), and \( x_0 = q_\alpha(p) \). It follows at once from conditions (3.3) and (3.4) that

\[ (X f)_p = \sum_i (X_i f)_p \cdot (X x^{(i)})(p). \]

This gives (3.10).

(2) The concept of affine (linear) connection is due to Levi-Civita [1]; see also Weyl [1]. Originally the definition of linear connection referred to surfaces in euclidean space, and the
NOTES AND COMMENTS

467

correct concept was then carried over to Riemannian spaces. Any Riemannian differential metric defines a torsion-free linear connection (see, for example, Favard [1]).

§4

The invariants of a family of probability laws were first considered as invariants of an object of a category by Morse and Sacksteder [1]; see also Sacksteder [1, 2]; monotone invariants were first studied by the author in [7]. Covariants in geometry (with the group of motions) were investigated by Rozenfel’d [1].

(1) Yet another example of a category is the category of all measurable spaces \((\Omega, S)\) with all measurable mappings of one into the other. There is a closure operation \((\Omega, S) \rightarrow (\Omega, S^*)\) for the objects of this category. By Lemma 2.1 this operation is a functor.

§5

(1) A transition probability distribution describes not only a decision rule but also, for example, a communication channel with random noise, with "input alphabet" \(\Omega\) and "output alphabet" \(\mathcal{E}\). Thus, as remarked by Dobrušin, the category of Markov morphisms is at the same time a category of statistical communication channels without memory etc. (see also Csiszár [5]). Note that Theorem 9.1 (on sufficient statistics) has a very graphic interpretation in terms of a two-way communication channel. In fact, it was shown by Wrighton [1] that in a certain sense the problem of statistical inference is a degenerate problem of communication in the presence of noise.

(2) It can be shown that the integrals (5.5) are equal to the double integral

\[
\int \int_{\Omega' \times \Omega} P\{d\omega'\} \Pi(\omega' ; \, d\omega') f(\omega'),
\]

understood as a limit of Darboux-Young integral sums

\[
\sum_{(B \times A)} P\{B\} \Pi(\omega'_B, A) f(\omega'_A)
\]

with respect to the filter of finite partitions of the space \(\Omega' \times \Omega\) into measurable rectangles \(B \times A, B \in S', A \in S\). To prove the existence of the limit, one must consider the lower and upper Lebesgue integral sums for the partition with

\[
A_k = \{\omega' : k/n^2 < f(\omega') < (k + 1)/n^2\}, \quad k = 0, 1, \ldots, n;
\]

\[
B_j^k = \{\omega' : (j + 1)/n^2 < \Pi(\omega' ; A_k) < (j + 1)/n^2\}, \quad j = 0, 1, \ldots, n^2,
\]

and then let \(n \rightarrow \infty\).

(3) Incidentally, the usual \(\delta\)-function defines a functional on continuous functions, while any measurable function can be integrated with respect to a \(\delta\)-measure. Hence they define distinct functionals.

§6

(1) For example, let \(\{P_1, Q_1\} \triangleright \{P_2, Q_2\} = (P_{12})\) and, conversely, \(\{P_1, Q_1\} \preceq \{P_2, Q_2\} = (P_{21})\). It is readily checked that then \(\{P_1, Q_1\} \prec \{P_2, Q_2\} = (P_{12} \ast P_{21})\), and

\[
P_{12} \ast P_{21} = \{P_1, Q_1\} = (P_{12} \ast P_{21} \ast P_{21}).
\]

(2) All the auxiliary lemmas of §5 were proved for bounded functions. They carry over trivially to nonnegative functions which may take the value \(+\infty\).

(3) In Morse and Sacksteder [1] the integral invariants were not introduced, and their arguments were therefore a little more complicated.

(4) Or weakly continuous Markov chains with compact state space (see Bebutov [1]).
NOTES AND COMMENTS

§8

(1) In contrast to the term "divergence" (see Kullback [3]), "deviation" reflects the asymmetric way in which the probability measures enter.

(2) A similar assertion was made (without proof) by Rosenblatt-Roth in the text of [1]. A more general statement, using slightly different terminology, was however proved earlier by Csiszár [5].

(3) Of course, for arbitrary pairs the equality \( I(Q|P) = I(Q'|P') \) does not imply equivalence. We published (8.6) in [7].

(4) The functional \( \chi^2(Q, P) \) was studied by Kagan [1] as a measure of the difference between \( Q \) and \( P \) (and called the divergence of \( P \) and \( Q \)). The functional (8.10) was studied by Perez [1] and Onicescu [1], under the name of information energy.

(5) In the same way as Shannon's entropy \( i_{\text{sh}} \) describes the exponential growth \( \exp\{N_{\text{sh}}\} \) of the number of highly probable messages when the length \( N \) of the message is increased. Recall that for a discrete space \( \Omega = (\omega_1, \ldots, \omega_m) \), uniform distribution \( Q \leftrightarrow (1/m, \ldots, 1/m) \) of outcomes and distribution \( P \leftrightarrow (p_1, \ldots, p_m) \) we have

\[
i_{\text{sh}}[P] = -\sum_{k=1}^{m} p_k \ln p_k = \ln m - I(Q|P).
\]

(6) The proof of (8.21) goes back to an unpublished paper of Stein [3] and the dissertation of Joshi [1] (in this connection Chernoff [2] and Kullback [3]). The relation (8.22) is due to Chernoff [1], whose proof relies on a rather delicate limit theorem of Cramér [3]. Formula (8.23) is due to N. P. Salihov. Salihov has recently proved that in testing several hypotheses \( P_1, \ldots, P_m \) the rate of exponential decrease of the maximum probability of error is given by

\[
I = \min_{j, k} \inf_{R \in \mathcal{C}_{\text{ph}}} \max \{ I(P_j|R), I(P_k|R) \}.
\]

(7) If the quantile falls on an atom of the distribution of \( f_N(e) \), \( P_N^N(e: f_N(e) = k_N) > 0 \), then, as usual, randomization is carried out (see Lehmann [1], Chapter 3), so that the hypothesis \( P_1 \) is rejected invariably if \( f_N(e) < k_N \) and with a certain probability \( q \) if \( f_N(e) = k_N \), where

\[
P_1^N(e: f_N(e) < k_N) + qP_N^N(e: f_N(e) = k_N) = b.
\]

(8) The convergence \( k_N \to -I[P_0|P_1] \) also occurs if \( -I[P_0|P_1] = -\infty \), since the difference \( \ln p_1(\omega) - \ln p_0(\omega) \) is \( P_1 \)-quasi-integrable. A rigorous proof of this version of the law of large numbers for the likelihood function is contained in the proof of Lemma 27.18.

(9) If \( I = +\infty \), the estimation from below is trivial.

(10) The asymptotic behavior described by Mourier and Sakaguchi (see Kullback [3]) is apparently erroneous.

(11) Thus, Rényi [1] considered the invariants \( \ln J_u \) and \( d \ln J_u/du \), related to the information deviations by (8.25) and (8.26). The integral \( J_{1/2}(P, Q) \) was, however, investigated by Bhattacharyya [1, 1943].

A concept closely related to information deviation is that of information (contained in one random variable relative to another); see Kolmogorov [5] and Gel'fand, Kolmogorov and Jaglom [1]. If \( F_{12} \) is the joint distribution law of two random variables \( \xi = x(\omega) \) and \( \eta = y(\omega) \), and \( F_1 \) and \( F_2 \) are the marginal laws, then \( I = I[F_1 \times F_2]/F_{12} \).

It is worth mentioning that 2 arcoss \( J_{1/2}(P, Q) \) is the distance between \( P \) and \( Q \) in the natural Riemannian metric defined by the Fisher information tensor (see §11 and §12.9).

§9

(1) Here lies the essential difference between the semigroup of Markov transformations and the semigroup of all linear transformations; in the latter almost all transformations are invertible (within the semigroup).

(2) Lemma 9.4 has been known for a long time, and not only to specialists in Markov chains. See, for example, Krein and Rutman [1] or Bebutov [1]. The sets \( \varepsilon \) of outcomes \( \omega \) described in the
NOTES AND COMMENTS

469

condition are (except for \( e_0 \)) sets of communicating recurrent states of the Markov chain \( \Pi \). We have omitted the statement and proof of the fact that every such set has its own stationary distribution.

§10

(1) Let \( C_j^+ = \{ \omega: f(\omega) > y \} \cap A_j, C_j^- = A_j - C_j^+ \). Then, since \( A_j \) is an atom, either \( C_j^+ \) or \( C_j^- \) is a \( \mathbb{Z} \)-set. By induction, it follows that for any finite sequence \(-\infty = y_0 < y_1 < \cdots < y_N = +\infty\) only one of the sets \( \{ \omega: y_{k-1} < f(\omega) < y_k \} \cap A_j \) is not a \( \mathbb{Z} \)-set. Consider a countable sequence of numbers, dense in the real line, with \( x_0 = -\infty \) and \( x_1 = +\infty \). Subject to reordering, every finite subsequence \( x_0, x_1, \ldots, x_N \) defines a partition of the real line into half-closed intervals. Applying the previous reasoning, we see that only one of the intervals does not correspond to a \( \mathbb{Z} \)-set, and with increasing \( N \) this interval becomes smaller. As \( N \to \infty \) these half-closed intervals \((i_k', i_k''[ \) shrink to a point, which we denote by \( \tau \). Denoting \( B_N = \{ \omega: i_k'' < f(\omega) < i_k' \} \cap A_j \), we have \( A_j - B_N = C_N \subseteq \mathbb{Z} \). Since \( \mathbb{Z} \) is a \( \sigma \)-ideal, it follows that \( \bigcup_{k=1}^{\infty} C_k \subseteq A_j \) and \( C \subseteq \mathbb{Z} \), whence \( A_j \sim C \). Thus \( \bigcap_{k=1}^{\infty} B_N = B \subseteq A_j - C \) is not empty. The function \( f(\omega) \) takes the value \( \tau \) on the set \( B \), and \( A_j - B = C \subseteq \mathbb{Z} \).

(2) For any \( f^1, \ldots, f^m \) there is a corresponding class of equivalent \( S \)-measurable functions, a suitable representative of which is the function \( f(\omega) = f^j \) for \( \omega \in A_j \), where \( A_1, \ldots, A_m \) is a canonical partition of \( \Omega \) into atoms. The function \( r(\cdot) \) defined on the \( \sigma \)-algebra \( S \) by

\[
r(H) = \int_{H} f(\omega) \mu(\omega),
\]

is a \( \mu \)-dominated measure, independent of the specific choice of \( f \) from its equivalence class. Obviously,

\[
r_j = r(A_j) = \int_{A_j} f(\omega) \mu(\omega) = f^j \times \mu_j.
\]

Thus any \( S \)-measurable function \( f(\cdot) \) determines a linear transformation of the cone of measures, \( \mathbb{Z} \)-equivalent functions determining the same transformation.

(3) In a finite-dimensional space all norms are equivalent (see (14.3)). Here, therefore, the expression \( o(\tau) \) has an absolute meaning. The fact that this is not so in infinite-dimensional spaces makes it difficult to define smooth families (see §§13 and 14).

(4) A “function-measure” in the terminology of Bogoljubov [1].

(5) Computational formulas of this type are used, for example, in Gel’fand, Frolov and Čencov [1].

(6) The category \( \tau \) CAP of collections Capf with the system of positive centrally-projective homomorphisms was mentioned in Čencov [5] in connection with the study of natural equivalences of families. The connected group of invertible positive centrally-projective transformations is the translation group.

§11

Various authors have repeatedly expressed the opinion that the Fisher information tensor defines a natural Riemannian metric in the manifold of mutually absolutely continuous probability distributions (see, for example, Kullback [2], [3]). However, serious results in this area were obtained only by Kozlov [1], whose work stimulated our own investigations.

(1) Congruent embeddings of the simplexes Capf are described by linear mappings in both natural and canonical coordinates. For this reason the statement of the lemma is true for any index values which vary according to a tensor law under (linear) transformation from one canonical (or natural) coordinate system to another. This argument is applied in Lemma 11.4 to the matrix of second derivatives in canonical coordinates, and it may be applied to fields of differential operators.

Note that since the geometry is almost homogeneous, equivariant scalar fields, (contra)vector fields and tensor contravalent fields vanish identically if they vanish at some point. This is not
the case for covalent tensor fields (since they are "carried backward" by the mappings, not "taken out").

(2) The geometries of the simplex Caph in natural and canonical coordinates are very reminiscent of intrinsic geometries of the first and second kind of the simplex as hypersurface in the enveloping $m$-dimensional space Var($\Omega$, $S_m$) equipped (see Norden [1]) with a unique equivariant field of normals $n(P) = p \leftrightarrow P$.

It is highly important that the equivariant tensor $g$ coincides for "tangent" measures with the Radon-Nikodym tensor, which converts a measure on $S$ into an $S$-measurable function, i.e. into an element of the dual space.

(3) The gradients of an invariant function on different objects are represented by covectors of different dimensions. In the case of congruent embedding, however, the scalar product of any embedded tangent vector with the gradient remains unchanged, since it is equal to the derivative in the appropriate direction. This implies that the gradient is equivariant.

§12

I am indebted to E. A. Morozova for advising me to seek the natural geometry of a linear connection. Theorem 12.3 provides the answer to a question put to the author by Ju. V. Linnik. I am indebted to F. I. Karpelevič for sharpening its formulation.


(2) This follows from the uniqueness theorem for the solution of system (12.18) or of the system of second-order equations obtained by eliminating the parameters.

(3) A geodesic is a trajectory of a one-parameter subgroup of the translation group. According to the supplement to Lemma 10.2, this group is simply transitive, while the correspondence between the coordinate $s$ and the law $P$ is unique in view of Lemma 10.3.

§13

The monotonicity of the approximating sums for the information deviation was apparently first pointed out by Sanov [1] (see also the proof in Kallianpur [1]). A general approximation theory for functionals of the form (13.18) has been developed by Csiszár [2], [3] and independently by Ghurye [1] and Čencov [12], [13]. Special subclasses of such functionals have also been investigated by other authors; see Martín and Oheix [1] and their subsequent publications.

(1) Translations, as projective transformations of a simplex, are also fractional-linear. Passage to a conditional distribution may be described as a limit of translations when the canonical parameter becomes infinite (see §21.8).

(2) For the definition of projective and inductive limits, see Palamodov [1] and also Scheffer [1].

(3) Recall that in view of the conventions (13.23) the second integral is a Lebesgue integral only when $Q \gg P$. But if

$$P (\omega : (dQ/dP)(\omega) = 0) > 0,$$

then the $Q$-integral of an infinite function over this set is put equal to $+ \infty$.

§14

(1) Since the function $f$ may not be one-to-one, the surface defined by its image may be self-intersecting. The tangent plane is a local concept. If $f(x_1) = f(x_2) = y_0$, then there are two tangents at the same point $y_0$ of the space, corresponding to different points $(x_1, y_0)$ and $(x_2, y_0)$ of the surface.

It should be noted that condition (14.5) is quite restrictive. It is certainly not satisfied when $\dim X > \dim Y$ (i.e. the dimension of a smooth surface cannot exceed the dimension of the enveloping space).

(2) However, this situation is typical even in linear topological spaces (see Averbub and Smoljanov [1]). One then has to be content with functionals which are differentiable with respect to the subspace of increments of a finite "norm".

...
Note that the weakest $\ell^1(P)$-metric (which coincides with variation for dominated measures) is too weak for us (see above, subsection 3), while the strongest $\ell^\infty(P)$-metric defined by the essential supremum of the Radon-Nikodym derivative is far too strong—even the family of normal laws is not continuous in it.

(3) A definition similar to $\ell^2(P)$-differentiability, for families of one real parameter, was given by Schmetterer [1].

(4) This follows from well-known theorems of classical analysis. We shall not present the proofs, since the corresponding proposition is easier to state and to prove for continuously differentiable surfaces (see Lemma 16.3).

§15

(1) It is natural to confine the study of single-fold differentiability to differentiability in the field of $\ell^2(P)$-metrics. However, $n$-fold differentiability should be considered in the stronger $\ell^m(P)$-metrics, or, more precisely, in a hierarchy of metrics (see §26.5); otherwise such "analytic functions" as $I(Q|P)$ turn out to be nondifferentiable.

(2) That is, we obtain an estimate for the squared norm in the quotient space of $\mathbb{L}^2(P)$ by a line of constants (cf. Schmetterer [1] and Čencov [9]). The corresponding induced metric was introduced by Gel'fand (see Hille [1]).

(3) (15.12) and (15.14) are known as the (one-dimensional) information inequalities. Associated with their rigorous proofs are such names as Cramér [1], [2], C. R. Rao [1], Darmois [1], Fréchet [1], as well as Dugué [1] and many others (see van der Waerden [1]).

(4) See Cramér [1], and C. R. Rao [2].

(5) For example, $P''$ etc. (see Bhattacharyya [2], Bol'šev [1] or Seth [1]). The more the law $P$ admits unbiased estimators $f(\omega)$ for zero, $M_Pf(\omega) = 0$, the larger the lower bound for the variance of the unbiased estimator (see Kagan [4] and C. R. Rao [3]).

(6) The definition of an efficient estimator goes back to Fisher [1], [2]. Regarding estimators for which the risk coincides with the bound (15.23), see De Groot and M. Rao [1].

(7) This theorem was considered with no clearly formulated smoothness conditions by Bhattacharyya [2]. A logical error has slipped into the proof given by Kullback [3, Chapter 3]. Fraser [1] gives the proof for one parameter only. In contradistinction to the authors just listed, we have not presumed the existence of smooth densities, and we allow the appearance of formal estimators. A stronger version of this theorem is Theorem 23.6.

§16

(1) But convergence (16.2) in the $\ell^2$-metric follows from the truth of condition (16.1) in the $\ell^4$-metric (cf. §26.5 and footnote 4 to §26).

(2) Or as the limit of Riemann sums in the local $\ell^\infty$-metric.

(3) A related fact was actually used in the proof of Lemma 16.1.

(4) A fortiori, according to Corollary 1 to Lemma 16.1, differentiable with $u$th moment in the sense of Definition 14.3.

(5) Continuity of $p(\omega; t)$ is not required here. Neither is measurability of $\tau(\omega)$, since the measurability of

$$(p_{\tau}(\omega; \tau(\omega))) = [t - \theta]^{-1}[p(\omega; t) - p(\omega; \theta)]$$

follows from the definition.

(6) Integral corollaries from the information inequality for one-parameter families were first obtained by Blyth [1], Girshick and Savage [1], and Hodges and Lehmann [1]; see also Karlin [1]. The examples of Stein [1], [2] show that these corollaries do not carry over to $n > 3$ parameters. This is why Kiefer [1], in his fundamental review of the theory of optimal multivariate estimators, expressed doubt as to the applicability of the information inequality as a tool for construction of optimal estimators. The theorem stated here was proved by the present author in [9].

Integral corollaries for the variance of an asymptotically normal law were obtained by LeCam [1] and Schmetterer [1], in the case $n = 1$. 
NOTES AND COMMENTS

§17

(1) This definition was apparently first given by Minty [1]. For further research on the subject, see Rockafellar [1].

(2) Any one-to-one continuous mapping of the real line into itself, with continuous inverse, is automatically monotone and takes monotone functions to monotone functions. This is not the case for multidimensional spaces.

§18

Exponent families are perhaps the most important and well-studied class of families. Systematic investigation of their general theory began with Koopman [1, 1936], who characterized them as families having finitely many sufficient statistics (see also Dynkin [1, 1951]). They appeared independently in problems of statistical physics (see Khintchine [1]).

An account of the general theory of exponent families was given in Kullback [3, 1959] and Lehmann [1, 1960]. General analytical approaches to the theory of parameter estimation and hypothesis discrimination for these families were proposed in monographs by Linnik [2], [3]. A number of fundamental problems of the theory were also touched upon in Blackwell and Girshick [1] and Robbins [1]. Our own research has dealt mainly with the “geometrical” aspect of the theory (see Cencov [4], [6]).

(1) The component of the identity of the group of centrally-projective transformations of $\text{Caph}(\Omega; S_m, \emptyset)$ into itself, see §10.

(2) More precisely, the statistics $q_1(\omega), \ldots, q_n(\omega)$ and $q_0(\omega) = I(\omega)$ are linearly independent.

(3) Though this density may determine a random function on smooth functions, i.e. define a generalized random variable in Gel’fand’s sense [2].

(4) Though the dimension of the family of generalized random variables is $n$.

(5) These limits form the Vorob’ev-Faddeev fine boundary of the simplex $\text{Caph}(\Omega, S_m, \emptyset)$. Under these conditions, to each strictly dominated law corresponds a whole manifold of limits. The point is that the simplex $\text{Caph}(\Omega, S_m, \emptyset)$, as a homogeneous manifold of zero curvature, has no natural boundary (see Gel’fand and Graev [1]). The boundary, obtained by Vorob’ev and Faddeev [1] from other considerations, may be obtained by the geometrical methods of Karpelevič [1].

(6) This means that the correspondence $s \rightarrow P_s$ determined by (18.1) and (18.2) defines a chart of the family.

(7) Recall that, in any semiordered space, first-order homogeneous functions are defined in an invariant manner (see Kantorovič, Vulih and Pinsker [1]).

(8) The desirability of presenting a direct proof for this theorem was pointed out to the author by Yu. V. Linik.

The theorem may also be derived from the general results of §7.

(9) This equation is transformed to a form more convenient for our purposes (cf. Bernstein [1]).

Regarding the equation $p'(x) = s g(x)p(x)$, see also Mathai [1].

§19

(1) The canonical affine parameter of a geodesic (see Favard [1]), or the canonical variable (see Khintchine [1]). This terminology is more “canonical” than that adopted by Linnik [2].

(2) The transformation from canonical to natural parameter in statistical physics is associated with the introduction of the notion of temperature (Khintchine [1]). A natural parameter was considered as an ancillary tool by Bhattacharyya [2] and Kullback [3]. The term “natural parameter” itself is due to the author [6]. The main result of the section were published in the indicated papers.

(3) See, for example, Dieudonné [1].

(4) Since for $s^{(2)} > 0$ the integral of the positive function $\exp[s^{(2)}x^2 + s^{(1)}x]$ from $-\infty$ to $+\infty$ is identically equal to $+\infty$.

(5) Unique with probability one, since both $q(\omega)$ and $p(\omega; s)$ are defined up to values on any $Z$-set.
§20

(1) Formula (20.8) is due to Huzurbazar [1]. (20.9) was proved by Kullback [3], who also provides references to the work of other authors. In the multidimensional case the Legendre conjugacy of the parameters, and, accordingly, (20.10), was first explicitly considered in Cencov [6]. The Young inequality (20.12) was proved much earlier by Sanov, by minimization of a suitable expression.

(2) Since $I[P|R] = +\infty$ when $P$ does not dominate $R$, formulas (20.15) and (20.16) permit a description of the geodesic hull spanned by mutually absolutely continuous laws $P$, as the family of probability distributions $\{P_\lambda\}$ where each $P_\lambda = R$ minimizes a sum $\sum_i \alpha_i I[P_i|R]$. Formula (20.15) recalls the expression for the moment of inertia about an arbitrary point in terms of the moment of inertia about the center of gravity. (And when the "masses" of points vary, the center of gravity runs over the entire linear span.)

(3) For nonconstructive families, all that one can assert is weak equivalence in the sense of Morse and Sacksteder (see above, Definition 6.4).

§21

(1) This simple example shows that the natural chart of an infinite-dimensional exponent family has a more intricate structure than one might expect at first glance. Example 6 (below) is due to the author [12].

(2) See also Basu [1], Doss [1] and Kale [1].

(3) This boundary is the most economical. The finer and therefore more massive boundary of Vorob’ev and Faddeev [1] converts the simplex $\text{Caph}(G, S_n, \emptyset)$ into a compact set on which the conditional probabilities of each event relative to any nonempty hypothesis are everywhere continuous.

§22

(1) The theorem may be slightly generalized by allowing the law to be a point of the boundary at infinity.

(2) The fact that the maximum likelihood method coincides with the method of minimum information deviation has been noticed by many authors, among them Kullback and Cencov; see also Kriz and Talacko [1], and Hartigan [1].

§23

(1) It is convex by Lemma 19.1. Thus, for exponent families $\gamma$ with domain $G^*$ satisfying (23.4), the domain of the natural parameter is convex.

(2) Definition 27.4 of a smooth family is a substitute, since it demands that the densities, not the measures themselves, be smooth.

(3) By Lemma 28.5, it is smooth in the sense of Definition 27.5. The definitions themselves were sought so that geodesic (exponent) families satisfy them (and for differentiable families, by analogy with finite-dimensional manifolds Caph, the information inequality be automatically valid).

(4) The inequality between the first and last members in (23.12) follows at once from (26.20). The direct proof is also quite easy.

(5) If the functional $\int \varphi[(dQ/dP)(\omega)]P(d\omega)$, where $\varphi$ is a convex function, is not continuous on Caph$(G, S, Z)$ in variation $\mu_0$, then according to Csiszar [2], [4] it does not define a uniform topology (evidently, a topology can be defined only by a hierarchy of such functionals; see §26.5). The terms used here are therefore conditional.

§24

(1) By §27.5, estimators of the form $P_\alpha$, where $\alpha$ is a parameter estimator, for an exponent family $\%$ with Gaussian loss function, form a complete class. In regard to this problem our formulation is close to that of De Groot and M. M. Rao [1] and M. M. Rao [1]; see also De Groot [1].
NOTE S AND COMMENTS

(2) Lemma 24.1 and its corollary are essentially rephrased versions of a well-known inequality of Blackwell [1], Kolmogorov [4] and C. R. Rao [1]; see also M. M. Rao [1].

(3) The proof given in our paper [9] is based not on Theorem 16.1 but on a simpler version, dealing specifically with the family $\mathcal{F}$.

(4) The author arrived at the “nonsymmetric Pythagorean geometry” of §§22 and 23 in [10] when he tried to generalize the geometry of the Gaussian method of least squares (see Gauss [1] and Linnik [1]) to arbitrary geodesic (exponent) families. The actual statement of Lemma 24.4, in some form or another, has frequently been used by statisticians to improve decision rules (see Thompson [1]).

(5) We put aside the very important and interesting (but much more complicated) Bayesian formulation of Robbins [1, 2] (see also Neyman [2]).

(6) Note that if $L(P, Q)$ is monotone and admits both expansions (24.24) and (26.10), then the constants $c_d[L]$ in both formulas are equal, but without the additional assumptions of uniform differentiability one cannot assert that the constants $c[L]$ are equal.

§25

(1) See also Frolov and Cencov [1], and the following publications: Van Ryzin [1], Schwartz [1], Kronmal and Tarter [1], Bosq [1, 2], Sesan et al. [1], and Sizova [1]. Estimators of type (25.1) have been discussed earlier by Rosenblatt [1]; see also Schuster [1].

(2) It is easy to see that $\sum \mathbb{W}(\eta) \mathbb{E}(\xi)^2$ is independent of the choice of the orthonormal basis $\eta, \ldots, \xi$ in $E$, since for each $x$ the sum $\mathbb{E}(\eta(x))^2$ is invariant under orthonormal changes of basis.

(3) The integrand in (25.31) is also independent of the choice of basis (cf. footnote 2, above).

(4) The norm of the deviation of the histogram from the graph of the density, in the metric of the space $\mathcal{C}$, has been estimated by Smirnov [1, 2] (see also Tumanjan [1]). If the grouping is optimal, it decreases almost like $N^{-1/3}$ (up to a logarithmic factor).

(5) See also Rosenblatt [1].

(6) These constraints may be relaxed somewhat by making them closer to quasi-homogeneity conditions (see Definition 28.1).

(7) We used this argument in [3]. It leads to a less precise accuracy bound than the inequality of Lemma 16.6, and is applicable only when the quality of the estimator is measured by maximum risk. On the other hand, it works under weaker constraints, and this is essential in the case of infinite-dimensional families.

(8) We are assuming here that all the moments $\mathbb{E}(\eta(x)y(x))^2$ of the initial estimator $p^*$ are measurable, which implies that $p^*$ is measurable and hence (by Lemma 17.14) that $p^n$ is measurable.

(9) Under very weak restrictions (cf. Tumanjan [1]) the square norm (25.10) of the deviation of the estimator is asymptotically normal (see Bosq [2], also [1]). Hence one can obtain sharper results for asymptotic confidence limits than we obtained in Cencov [3, 12].

(10) The question of statistical estimation of a smooth curve is very timely (see Tukey [2] and Whittle [1]).

§26

(1) Random measures have been introduced in many publications as random functionals (i.e. random generalized functions); see Prohorov [1] and Gel’fand [2]. Discrete empirical random measures (25.9) have also been considered. Both concepts admit an effective definition.

(2) At first sight it might seem natural, following Laplace [1], to adopt as loss function some invariant metric or a substitute thereof (see the review of Adhikari and Joshi [1], where more than ten such functions are listed). However, as Gauss noted [1], the theory is much simplified if one takes a quadratic loss function (see LeCam [1]). This is why our loss function is a nonsymmetric analog $2I[P|P^*]$ of the squared euclidean distance—an analytic, “approximately quadratic” functional for which the nonsymmetric Pythagorean theorem and various other geometrical theorems are valid (see subsection 5).
An interesting unsolved problem is to describe all natural matrix loss functions (see Kagan [4]).

(3) This follows from a result of Csiszár [2], [4] (see footnote 5 to §23).

(4) For example,
\[ \| R - P \|_{(p, 2)} < \| R - Q \|_{(p, 2)} + \| Q - P \|_{(p, 2)} \]
\[ < \left( \| Q \|_{(p, 2)} \right)^{3/4} \| R - Q \|_{(Q, 0)} + \| Q - P \|_{(p, 2)}. \]

(5) Moreover, \( P \{ A' \} I[P|R'] \succ I[P|R] \) as \( A' \succ \Omega. \)

§27

(1) This is a way of using an argument of Blackwell, Kolmogorov and Rao (see footnote 2 to §24).

(2) Under our assumptions, we can no longer assert that the majorant
\[ g_1(\omega) \left( h(\omega) + 3|h(\omega)|^2 + |h(\omega)|^3 \right) \]
for the third derivatives is integrable.

(3) The statement of the corollary follows from the theory of §16, since the smooth families are contained in the class of continuously differentiable families.

(4) This class contains all compact exponent families corresponding to a compact subdomain of the canonical parameter (see Lemma 28.5).

(5) If the dimension \( n \) is large, these estimators become nontrivial only for large \( N \). If one assumes that the higher moments of the majorant are bounded, the estimators become efficient at values as low as \( n \sim N^{3+\delta}. \)

(6) Another possible procedure for localizing the root is used in the proof of Theorem 27.3.

(7) Unless restrictions are imposed on the compactness of the family and of the maximization domain, the optimality and even consistency of the maximum likelihood estimator become problematic. Linnik and Mitrofanova [1] give a very delicate proof that the maximum likelihood estimator is efficient for a certain subclass of the exponent families.

§28

(1) A brief account of the contents of this section was given in our note [8] (see also [12]).

(2) Of course, only the rate of decrease of the quantity \( n \propto \Gamma(N) \) is determined; \( n \) itself may be fixed arbitrarily within certain bounds.

(3) Recall that the indices for the information matrix in canonical coordinates \( s' \) are written as subscripts: \( \gamma_{jk} \). The notation \( w^k \) is reserved for the information matrix in natural coordinates \( j \).

§29

The definition of a random probability measure (on the unit interval or on the whole real line) by its random distribution function goes back to Kolmogorov [1]. Dubins and Freedman [1] systematically considered the definition of a random distribution function \( F(t) \) as a monotone stochastic process on the real line, i.e. in terms of probabilities of quasi-intervals:

\[ \mathcal{P} \left\{ a_i < F(t_i) < a_i^*, i = 1, \ldots, n \right\}. \]

These probabilities define a Baire distribution and, according to Kakutani [1] (see Halmos [1]), a unique regular Borel distribution, which is easily shown to be concentrated on monotone functions \( F(t) \). However, as follows from a well-known result of the author [1], this distribution is generally not concentrated on distribution functions, since the set of all distribution functions is neither a Borel set nor absolutely measurable (although the set of continuous distribution functions has these properties). We therefore have to prove the existence of a non-Borel extension \( \mathcal{P} \), concentrated on the distribution functions, just as in the theory of Markov processes one has to prove the existence of a canonical modification of the process (see Itô [1], and also Doob [2]).

(1) The class of \( \mathcal{B}(S) \)-sets is fairly small. If the algebra \( S \) is uncountable, it does not even contain "singleton subfamilies". Since every family consisting of a single set function is closed in the product topology, they are all \( \mathcal{B}(S) \)-sets.

NOTES AND COMMENTS

475
(2) If $S$ has countably many generators, there is a substantial difference between the algebras $E_0(S)$ and $B_0(S)$. For example, all "singleton" families are measurable in $E_0(S)$, since every $\sigma$-additive measure is completely determined by its values on countably many generators, while an arbitrary set function is determined by its values on the whole algebra $S$.

(3) Each of conditions $1^o-3^o$ of system (2.5) (for fixed sets $H$, $H_1$ and $H_2$) describes a closed $B_0(S)$-subset of the space $X^B$. Hence axioms $1^o-3^o$, which define the collection of all normalized finitely-additive set functions, describe a closed $B(S)$-set as the intersection of the above-mentioned closed $B_0(S)$-sets. The descriptive nature of the set of all normalized countably-additive set functions (i.e. probability measures) is as yet unknown, notwithstanding several interesting studies (see Freedman [1]).

By virtue of the well-known continuity property of any finite measure relative to a dominating measure, and the $\sigma$-additivity of any finitely-additive measure which is continuous relative to a $\sigma$-additive measure (cf. the proof of Lemma 14.1), the collection $\text{Capd}(\mathcal{G}, S, Z)$ is a family of the type $F_{ad}$ in $X^B$, and hence $B(S)$-measurable, although it is not $B_0(S)$-measurable.

(4) In view of conditions $1^o-3^o$, condition $4^o$ (convergence with probability one) may be replaced by convergence in probability:

For any $\varepsilon > 0$ and every fixed sequence $H_k \searrow \emptyset$, $H_k \in S$,

$$\mathbb{P}\left\{ \Phi\{H_k\} > \varepsilon \right\} \rightarrow 0.$$  

(5) Note that Lebesgue collections $\text{Capd}(\mathcal{G}, S, Z)$ are not only $B(S)$-measurable in $X^B$ (see footnote 3, above) but also $K$-measurable, also corresponding to an $F_{ad}$-set in the sector $W$.

(6) An analogous statement is readily proved for other functionals (13.18) satisfying the conditions of Theorem 13.1.
BIBLIOGRAPHY

B. P. Adhikari and D. D. Joshi

P. S. Aleksandrov
2. Über die Urysohnschen Konstanten, Fund. Math. 20 (1933), 140–150.

V. I. Averbuh and O. G. Smoljanov

R. R. Bahadur

Stefan Banach and Casimir Kuratowski

A. P. Basu

M. Bebutoff [M. V. Bebutov]

Edwin F. Beckenbach and Richard Bellman

477
Robert H. Berk

S. N. Bernstein

J. Bertrand

A. Bhattacharyya

Herluf Bidstrup

Garrett Birkhoff

Allan Birnbaum

Richard L. Bishop and Richard J. Crittenden

David Blackwell

David Blackwell and M. A. Girshick

D. Blackwell and C. Ryll-Nardzewski
BIBLIOGRAPHY

J. R. Blum and Judah Rosenblatt

Colin R. Blyth

Salomon Bochner and William Ted Martin

N. N. Bogoljubov

L. N. Bol’sev

Denis Bosq

N. Bourbaki

David R. Brillinger

D. L. Burkholder

N. N. Čencov

Herman Chernoff

Harald Cramér

Imre Csizár

Georges Darmois
BIBLIOGRAPHY

Morris H. DeGroot

M. H. DeGroot and M. M. Rao

J. Dieudonné

J. L. Doob

S. D. A. C. Doss

Lester E. Dubins and David A. Freedman

Daniel Dugué

E. B. Dynkin

Samuel Eilenberg and Saunders Mac Lane

Samuel Eilenberg and Norman Steenrod

J. Favard

Werner Fenchel
Ronald Aylmer Fisher

D. A. S. Fraser

Maurice Fréchet

David A. Freedman

A. S. Frolov and N. N. Čencov

B. A. Fuks

C. F. Gauss
1. Abhandlungen zur Methode der kleinsten Quadrate, P. Stavkiewicz, Berlin, 1887.

I. M. Gel’fand

I. M. Gel’fand and S. V. Fomin

I. M. Gel’fand, A. S. Frolov and N. N. Čencov

I. M. Gel’fand and M. I. Graev
I. M. Gel'fand, A. N. Kolmogorov and A. M. Jaglom

S. G. Ghurye

M. A. Girshick and L. J. Savage

B. V. Gnedenko and A. N. Kolmogorov

Paul R. Halmos

Paul R. Halmos and L. J. Savage

J. A. Hartigan

Felix Hausdorff

Sigurdur Helgason

Einar Hille

J. L. Hodges, Jr., and E. L. Lehmann

Wassily Hoeffding
BIBLIOGRAPHY

V. S. Huzurbazar
1. Exact forms of some invariants for distributions admitting sufficient statistics, Biometrika 42 (1955), 533–537.

A. D. Joffe and V. M. Tihomirov

Alexander Ionescu Tulcea and Cassius Ionescu Tulcea

C. Ionescu Tulcea

C. T. Irelanand S. Kullback

Kiyoshi Itô

W. James and Charles Stein

Devi Datt Joshi

A. M. Kagan
4. Estimation theory for families with shift or scale parameters and for exponent families, Dissertation, Leningrad State Univ., Leningrad, 1967. (Russian)

Shizuo Kakutani

B. K. Kale
BIBLIOGRAPHY

Gopinath Kallianpur

L. V. Kantorovič, B. Z. Vulih and A. G. Pinsker
1. *Functional analysis in partially ordered spaces*, GITTL, Moscow, 1950. (Russian)

Samuel Karlin

F. I. Karpelevič

John L. Kelley

A. Ja. Khintchine

J. Kiefer

Felix Klein

A. N. Kolmogorov
3. *On the proof of the method of least squares*, Uspehi Mat. Nauk 1 (1946), no. 1 (11), 57–70. (Russian)

B. O. Koopman
V. P. Kozlov

M. G. Krein and M. A. Rutman

T. A. Kriz and J. V. Talacko
1. Equivalence of the maximum likelihood estimator to a minimum entropy estimator, Trabajos Estadist. 19 (1968), 55–65.

Richard Kronmal and Michael Tarter

Solomon Kullback

S. Kullback and M. A. Khairat

S. Kullback and R. A. Leibler

A. G. Kuroš, A. H. Livšic and E. G. Šul’geifer

Serge Lang

Pierre Simon, Marquis de Laplace

F. Laurent, M. Oheix and J.-P. Raoult

Henri Lebesgue
BIBLIOGRAPHY

Lucien LeCam

E. L. Lehmann

Tullio Levi-Civita

Paul Lévy

Ju. V. Linnik

Ju. V. Linnik and N. M. Mitrofanova

M. Littaye-Petit, J.-L. Piednoir and B. Van Cutsem

A. A. Ljapunov
1. *On choosing from a finite number of distribution laws*, Uspehi Mat. Nauk 6 (1951), no. 1 (41), 178–186. (Russian)

Michel Loève

N. N. Lusin [Luzin]

N. Lusin and P. Novikoff [N. N. Luzin and P. S. Novikov]
N. Lusin [N. N. Luzin] and W. Sierpiński

George W. Mackey

François Martin and Madeleine Oheix

François Martin, Jean-Luc Petit and Monique Petit-Litay [Litay-Petit]

François Martin and Daniel Vaguely

A. M. Mathai

George J. Minty

Norman Morse and Richard Sacksteder

John von Neumann

John von Neumann and Oskar Morgenstern

Jacques Neveu

J. Neyman
J. Neyman and E. S. Pearson

A. P. Norden

Pierre Novikoff [P. S. Novikov]

Octav Onicescu

V. P. Palamodov

Emanuel Parzen

Albert Perez

M. M. Postnikov

Ju. V. Prohorov

Bayard Rankin

C. Radhakrishna Rao
M. M. Rao

Alfréd Rényi

Herbert E. Robbins

R. Tyrrell Rockafellar

V. A. Rohlin

Guy Romier

Murray Rosenblatt

M. Rozenblat-Rot [Millu Rosenblatt-Roth]

B. A. Rozenfel’d
1. Non-Euclidean geometries, GITTL, Moscow, 1955. (Russian)

Richard Sacksteder

Stanislaw Saks
O. N. Salaevskiǐ

I. N. Sanov

Leonard J. Savage

V. V. Sazonov

Carel L. Scheffer

Leopold Schmetterer

Eugene F. Schuster

Stuart C. Schwartz

Anton Șesan, Silvia Covali, N. Gheorghiu and Adrian Vulpe

G. R. Seth

Seymour Sherman

A. F. Sizova
A. V. Skorohod

N. V. Smirnov

Charles M. Stein
3. Comparison of experiments, Mimeographed Notes, Univ. of Chicago, Chicago, Ill., 1951.

P. L. Tchebycheff

James R. Thompson

V. M. Tihomirov

John W. T. Tukey

S. H. Tumanjan

J. R. Van Ryzin

V. G. Vinokurov
D. A. Vladimirov

V. S. Vladimirov

N. N. Vorob'ev and D. K. Faddeev

Georges Vranceanu and Georges-Georges Vranceanu

B. L. van der Waerden

Abraham Wald

Geoffrey S. Watson

Hermann Weyl

Peter Whittle

Norbert Wiener

J. Wolfowitz

R. F. Wrighton
This page intentionally left blank
INDEX

\(\Theta\)-chart of a geodesic family, 265
\(\Theta\)-local coordinate system, 415
\(\Theta\)-local distance between laws, 415
\(\Theta\)-reduced likelihood function, 426
a priori information, 4
absolute completion of an \(\omega\)-algebra, 15
absolute covariant, 59
absolute invariant, 58
absolutely equivariant linear connection, 63
accompanying law, 420
affine barycentric coordinates, 269
algebra \(E_\omega(S)\) of distribution families, 452
almost homogeneous geometry, 63
almost proper conditional probability distribution, 23
approximate maximum likelihood estimator, 448
arbitrarily convergent points, 63
arithmetic mean estimator, 356
asymptotic difficulty of a statistical estimation problem, 363
atlas, 37
atom of a ring, 127

boundary \(\delta\) of a set, 246
boundary at infinity of an exponent family, 315
boundary of an exponent family, 334
boundary point of a lower body, 314
boundary point of a convex body, 314
canonical parametrization of an exponent family, 263
canonical affine coordinate systems, 146
Cartesian product of probability distribution families, 80
category, 50
CAPD, 92
CAPF, 92
CAPHF, 139
CAP, 72
FAM, 76
FAMD, 92
FAMH, 92
FAML, 83
of mappings, 50
of statistical decision rules, 71
VAR, 73
Ceva line, 135, 148
charge, 13
Christoffel symbol, 46
closed set, 246
closure of a \(\sigma\)-algebra, 15
collection Cap of probability measures, 13
collection Capd of probability measures, 20
collection Caph of probability measures, 20
collection Conh of measures, 140
commutative diagram, 58
compact family of probability distributions, 419
compatibly constructive family of distributions, 29
complete measure, 15
complete product of families of probability laws, 79
completely finite distance of a family from a law, 324
concave function, 250
conditional probability distribution, 2
cone of positive measures (Conh), 141, 188
congruence, 52, 56
conjugate norm, 260
INDEX

conjugate parametrization of an exponent family, 291
constructive probability distribution, 29
continuously differentiable family of probability distributions, 237
continuously differentiable simple family of probability distributions, 228
contravariant, 60
contravariant functor, 57
contravariant of a family of distributions, 94
covenient partition, 206
conventions for resolution of indeterminacies, 190, 195, 197, 202
convex function, 248
convex quasi-closed subfamily, 329
convex hull (conv) of a set, 247
convex quasi-closed subfamily, 329
convex set, 245
convex support (conv supp) of a measure, 304
convex support of a directional statistic, 305
convolution of a function and a charge, 202
counting measure, 288
covariant, 59
covariant functor, 57
cubic family of probability distributions, 419
cumulant of order k, 281
decrease of type of g(n), 379
derivative of a measure with respect to a measure, 20
deterministic decision rule, 5
diffeomorphism, 47
differential function, 38
differentiable manifold, 37
differentiable mapping, 200
differentiable structure, 37
differentiable surface, 200
difficulty of the problem of point estimation, 356
dimension of an exponent family, 272
direct product of probability spaces, 1
directional sufficient statistic, 263
disjoint probability distributions, 86
distance from a point to a set (body), 248, 378
distance of a family from a law, 324
domain (Dom) of a function, 250
domain of a morphism, 50
dominated measure, 19
domination of measures, 19
deviation of a set from a set, 378
dyad, 16
diff of a convex body, 314
efficient estimator of a parameter, 219
elementary category, 60
elementary category geometry, 60
elementary Klein geometry, 60
equivalent random variables, 19
equivalence relation, 52
equivalent measurable mappings, 19
equivariant, 59
differential field, 61
linear connection, 59
tensor field, 61
vector field, 61
errors of the first and second kinds, 10, 122
essentially ponderable face of a convex support, 315
exact subcategory, 53
expectation, 18
exponent family with canonical affine parameter, 263
exponential family, 264
extended gradient, 257
extreme point of a convex body, 246
face of a convex body, 314
faithful parametrization of an exponent family, 272
faithful representation of a geodesic family, 272
family of binomial laws, 288
family of normal laws, 276, 287, 355
family of Pearson distributions, 277
family of Poisson laws, 288
family of probability distributions, 4
family of probability distributions differentiable with nth moment, 209, 212
final object, 50
finite distance of a family from a law, 324
Fisher information matrix, 162, 213
formal estimator, 289
frequency polygon, 384
full subcategory, 53
function continuous on an interval up to the endpoints, 250
functor, 57
Gaussian loss function, 356
generator of a convex body, 314
genuine conditional probability distributions, 23
INDEX

geodesic, 46
geodesic family of distributions in the canonical (affine) parametrization, 189, 263
geodesic mean of probability laws, 166
geosodesically convex family of probability distributions, 274
gradient, 256

H-truncation of a probability distribution, 312
Hahn decomposition, 14
hereditarily equivariant linear connection, 63
hereditary covariant, 59
hereditary invariant, 58
histogram method, 384

I-closed family, 348
indicator (characteristic function) of a set, 70
inductive limit, 129
infinite distance of a family from a law, 324
information, 468
information deviation (relative entropy), 115, 195
information inequality, 215
initial object, 50
inner product in the space of charges, 213
inner product of functions, 213
integral convex hull of a family, 421
"integral" information inequality, 242
interior (Int), 246
invariant, 58
invariant scalar field, 61

Jordan decomposition, 14
Kolmogorov family of probability distributions, 457

Lebesgue family of probability distributions, 32
Lebesgue measurable space, 32
Lebesgue probability distribution, 32
Legendre transform, 258
likelihood function, 289
limit point of a Θ-family of charges, 342
linear connection on a manifold, 45
linear span (Lin) of a set, 246
local coordinates, 37
local mapping, 200
local mapping differentiable with nth moment, 208
localized maximum likelihood rule, 433

loss function, 7
lower μ-variation, 14
lower continuous function, 250
lower semicontinuous function, 250

Markov chain, 1
Markov geometry, 76
Markov homotopy, 179
Markov morphism, 19, 65
maximum likelihood estimator, 289, 290
mean risk, 364
measurable space, 1
measure, 13
minimax estimator, 361
monotone functional of distributions, 393
monotone invariant, 58
monotone vector field, 256
morphism, 50
μ-variation, 14
mutually absolutely continuous measures, 19

n-dimensional disk, 446
n-dimensional information width, 446
n-dimensional internal information radius, 447
n-dimensional internal radius, 379
n-dimensional width of a set, 378
natural coordinate system, 142
natural parametrization of an exponent family, 279
necessary statistic, 95
negative variation of a charge, 14
nondecreasing vector field, 256
nonincreasing vector field, 256
nonsymmetric Pythagorean geometry, 306
normalizing divisor of a canonical parametrization, 263
null-set, 19

object of a category, 50
open chart, 37
orthoprojection, 320

parallel family of vectors, 46
parametrized set, 55
partial derivative of a family of charges, 208
perfect measure, 18
point of I-contact, 348
point of weak contact for a family of charges, 342
ponderable face of a convex support, 314
positive variation of a charge, 14
probability distribution, 1, 13
probability distribution law for inferences, 7
probability space, 1
problem of projection on a geodesic family, 323
problem of projection onto a “hyperplane”, 327
projection estimator of a density, 373
projection on a convex compact family, 410
projection on a convex quasi-closed subfamily, 329
projection on a convex set, 363
projective limit, 128
pseudo-Riemannian metric, 43
“punctured” neighborhood, 250

quasi-equivalent measures, 19
quasi-homogeneous distribution family, 438
quasi-homogeneous geometry, 63
quasi-smooth category of manifolds, 60

Radon-Nikodym derivative, 20
random variables, 18
range of a morphism, 50
rational probability distribution, 137
rectangle, 1
reduced form of an exponent family, 278
reduced representation of the density of a geodesic family, 290
reduction of data, 28
regular conditional distribution, 21
regular exponent family, 306
regular functional, 190, 394
regular functional of distributions, 190, 394
regularly differentiable convex function, 255
relative entropy, 115
reparametrization of a continuously differentiable family, 233
reparametrization of an exponent family, 265
restriction of a measure to a subalgebra, 186
Riemannian metric, 43
Riemannian metric invariant in a category, 157
risk function, 356
riskiness of a decision rule, 358
separating hyperplane, 247
separating partition, 238
set of finiteness (Fin) of a function, 249
Shannon entropy, 468
simple differentiable surface without boundary, 212
simple family of probability distributions, 410
simple smooth compact family, 415
simplex of all probability distributions, 3, 128
smooth family with boundary, 237
smooth family of probability distributions, 411
space $L^p (P)$ of charges, 203
space $L^q (P)$ of functions, 202
space of bounded charges (Var), 14
space of inferences, 4
space of possible actions, 4
special chart, 40
standard parametrization of an exponent family, 290
standard representation of the density of a geodesic family, 290

statistic, 2, 17
statistical decision rule, 6, 65
statistical morphism, 81
statistically equivalent families, 12
stochastic matrix, 6
strictly concave function, 250
strictly convex function, 249
strictly decreasing vector field, 256
strictly increasing vector field, 256
subcategory CAPF, 80
subcategory FAMF, 132
submanifold, 41
sufficient statistic, 28
support (supp) of a measure, 304
supporting hyperplane, 247
symmetric tensor field, 43
system of maximum likelihood equations, 289, 426

tangent space, 40
tangent statistic, 438
tangent vector field, 42
tangent vector, 40, 438
theta-chart of a geodesic family, 265
theta-local coordinate system, 415
theta-local distance between laws, 415
theta-reduced likelihood function, 426
topology $T$ of a collection of probability measures, 201
total variation of a charge, 14
totally geodesic submanifold, 47
transition measure, 155
transition probability distribution, 2, 19
INDEX

translation in a collection of probability laws, 144
trivial ideal, 20
uncertainty of a decision rule, 358
uncertainty of the problem of point estimation, 404
\(n\)-dimensional disk, 446
uniformizable measurable wrapping, 31
upper \(\mu\)-variation, 14
variance of the likelihood function, 120
vector field, 39
vector of outcome probabilities, 3, 128
vertex of a convex body, 314
Wald risk, 7
weak limit point of a \(\Theta\)-family of charges, 342
weakly continuous mapping, 208
weakly differentiable mapping, 208
weakly smooth monotone functional of distributions, 394
weighted geodesic mean of probability distributions, 270, 418
Young's inequality, 258
Copying and reprinting. Individual readers of this publication, and nonprofit libraries acting for them, are permitted to make fair use of the material, such as to copy a chapter for use in teaching or research. Permission is granted to quote brief passages from this publication in reviews, provided the customary acknowledgment of the source is given.

Republication, systematic copying, or multiple reproduction of any material in this publication is permitted only under license from the American Mathematical Society. Requests for such permission should be addressed to the Assistant to the Publisher, American Mathematical Society, P. O. Box 6248, Providence, Rhode Island 02940-6248. Requests can also be made by e-mail to reprint-permissions@ams.org.