Wald Lecture I:
Counting Bits with Kolmogorov and Shannon

David L. Donoho
Stanford University

Abstract

Shannon's Rate-Distortion Theory describes the number of bits needed to approximately represent typical realizations of a stochastic process $X = (X(t) : t \in T)$, while Kolmogorov's ϵ-entropy describes the number of bits needed to approximately represent an arbitrary member $f = (f(t) : t \in T)$ of a functional class $\mathcal{F}$. For many stochastic processes a great deal is known about the behavior of the rate distortion function, while for few functional classes $\mathcal{F}$ has there been success in determining, say, the precise asymptotics of the ϵ-entropy.

Let $W^m_{2,0}(\gamma)$ denote the class of functions $f(t)$ on $T = [0, 2\pi)$ with periodic boundary conditions and $\frac{1}{2\pi}\int_0^{2\pi} f^2(t)dt + \frac{1}{2\pi}\int_0^{2\pi} f^{(m)}(t)^2 dt \le \gamma^2$. We show that for approximating functions of this class in L^2 norm we have the precise asymptotics of the Kolmogorov ϵ-entropy:

$$H_\epsilon(W^m_{2,0}(\gamma)) \sim 2m(\log_2 e)(\gamma/2\epsilon)^{1/m}, \quad \epsilon \to 0. \tag{0.1}$$

This follows from a connection between the Shannon and Kolmogorov theories, which allows us to exploit the powerful formalism of Shannon's Rate-Distortion theory to obtain information about the Kolmogorov ϵ-entropy. In fact, the Kolmogorov ϵ-entropy is asymptotically equivalent, as $\epsilon \to 0$, to the maximum Rate-Distortion $R(D, X)$ over all stochastic processes X with sample paths in $W^m_{2,0}(\gamma)$, where we make the calibration $D = \epsilon^2$.

There is a family of Gaussian processes X^*_D which asymptotically, as $D \to 0$, take realizations in $W^m_{2,0}(\gamma)$, and for which the process at index D has essentially the highest rate-distortion $R(D, X)$ of all processes X living in $W^m_{2,0}(\gamma)$. We evaluate the rate-distortion function of members of this family, giving formula (0.1).

These results strongly parallel a key result in modern statistical decision theory, Pinsker's theorem. This points to a connection between theories of statistical estimation and data compression, which will be the theme of these Lectures.

Key Words and Phrases. Rate-Distortion, Gaussian Process, ϵ-Entropy, Ellipsoids.

Dedication. *This work is dedicated to Mark S. Pinsker.*

Acknowledgements. Thanks to Toby Berger, Lucien Birgé, Ingrid Daubechies, Ron DeVore, Mikhail Ermakov, Robert M. Gray, Iain Johnstone, Alex Samarov, Stanislaw Szarek, and Martin Vetterli for helpful discussions and correspondence.

This research was partially supported by NSF DMS-95-05151, by AFOSR MURI-95-F49620-96-1-0028, and by other sponsors.

This result was discussed (without proof) during the Wald Lectures, IMS Annual Meeting, Park City Utah, July 1997.

1 Introduction

1.1 Shannon

Fifty years ago, Claude Shannon launched the subject of information theory as a mathematical discipline by formalizing lossless data compression of discrete-valued stochastic processes and lossy data compression of continuous-valued stochastic processes. In addition to proposing a wonderfully fruitful point of view, he formulated the key results of lossless and lossy compression, and gave heuristic arguments that underlay rigorous proofs. From a modern point of view we recognize, as did A. N. Kolmogorov, that Shannon's "mathematical intuition is remarkably

precise." For example, the central formulas in the discipline of lossy compression – the rate distortion function [1] – today can be understood as applications of the theory of large deviations [5], a field which didn't exist in 1948!

Shannon's work is remarkable for providing both an enchanting formalism – entropy, mutual information – and sharp answers to technical questions – asymptotic size of optimal block codes. It contains within it many general ideas, such as the rate-distortion function, which in concrete cases (Gaussian stationary processes) lead to beautifully intuitive answers (the "Water Filling" algorithm). Here is a simple consequence of the Shannon theory, important for comparison with what follows. Suppose $X(t)$ is a Gaussian zero mean stochastic process on an interval T and suppose that the eigenvalues of its covariance obey

$$\lambda_k \sim k^{-m}, \quad k \to \infty. \tag{1.1}$$

Let $N(D, X)$ denote the minimal number of codewords needed in a codebook $\mathcal{C} = \{X'\}$ so that

$$E \min_{X' \in \mathcal{C}} \|X - X'\|^2_{L^2(T)} \le D. \tag{1.2}$$

Then

$$\frac{\log N(D, X)}{R(D, X)} \to 1, \quad D \to 0, \tag{1.3}$$

where $R(D, X)$ is the rate-distortion function for X:

$$R(D, X) = \inf\{I(X, Y) : \quad E\|X - Y\|^2_{L^2(T)} \le D\}, \tag{1.4}$$

with $I(X, Y)$ the usual mutual information

$$I(X, Y) = \int p(x, y) \log \frac{p(x, y)}{p(x)q(y)} dxdy.$$

Here $R(D, X)$ can be obtained in parametric form from a formula which involves the eigenvalues (λ_k) (first published by Kolmogorov (1956)): for $\theta > 0$ we have

$$R(D_\theta) = \sum_k \log_+(\lambda_k/\theta), \tag{1.5}$$

where

$$D_\theta = \sum \min(\theta, \lambda_k). \tag{1.6}$$

It follows that

$$R(D, X) \sim \frac{m}{(m-1)^{1/(m-1)}} \quad D^{1/(m-1)}, \quad D \to 0. \tag{1.7}$$

Note how the theory gives the sharp asymptotics of $\log N(D, X)$; it does so through introducing a new concept into the picture – the rate-distortion function, in which the mutual information appears (Deus Ex Machina) – and then actually computing $R(D, X)$.

Shannon's work is remarkable for its broad influence in intellectual life. Within a few years after the publication of *A Mathematical Theory of Communication*, "information theory" was a household phrase, discussed in connection with biology, philosophy, and art! Of course much of this discussion was far-fetched speculation. The fact that Shannon inspired a great deal of intellectual activity in a broad range of disciplines remains a visible phenomenon even today.

1.2 Kolmogorov

A. N. Kolmogorov was exposed to Shannon's work in the early 1950's and appears to have been deeply influenced by it. In Kolmogorov's seminar in Moscow, many of the bright young talents of the day were exposed to Shannon's work; Kolmogorov and students Pinsker and Dobrushin made contributions to information theory by giving proper foundations and rigorous proofs of rate-distortion theory for stationary Gaussian processes.

At the same time, Kolmogorov was interested in a number of problems in analysis, such as Hilbert's 13th Problem, which brought up issues that seemed at a formal level to parallel questions of information theory [8]. In the mid-1950's, Kolmogorov introduced the notion of the ϵ-entropy of a functional class, defined as follows. Let T be a domain, and let $\mathcal{F}$ be a class of functions $(f(t) : t \in T)$ on that domain; suppose $\mathcal{F}$ is compact for the norm $\|\cdot\|$, so that there exists an ϵ-net, i.e. a system $\mathcal{N}_\epsilon = \{f'\}$ such that

$$\sup_{f \in \mathcal{F}} \min_{f' \in \mathcal{N}_\epsilon} \|f - f'\| \leq \epsilon. \tag{1.8}$$

Let $N(\epsilon, \mathcal{F}, \|\cdot\|)$ denote the minimal cardinality of all such ϵ-nets. The Kolmogorov ϵ-entropy for $(\mathcal{F}, \|\cdot\|)$ is then

$$H_\epsilon(\mathcal{F}, \|\cdot\|) = \log_2 N(\epsilon, \mathcal{F}, \|\cdot\|). \tag{1.9}$$

It is the least number of bits required to specify any arbitrary member of $\mathcal{F}$ to within accuracy ϵ.

Simultaneous with Kolmogorov's development of a "Moscow school of information theory", he cultivated as well an interest by mathematicians in ϵ-entropy, leading to a vigorous development in the calculation of ϵ-entropies of a range of functional classes on a range of domains and norms. Here is a typical result about ϵ-entropy, reported in Kolmogorov-Tikhomirov (1959) [10] Section 5. Let T be the circle $T = [0, 2\pi)$ and let $W^m_{2,0}(\gamma)$ denote the collection of all functions $f = (f(t) : t \in T)$ such that $\|f\|^2_{L^2(T)} + \|f^{(m)}\|^2_{L^2(T)} \leq \gamma^2$. Then for finite positive constants $A_{0,m}$ and $A_{1,m}$, depending on m but not γ or $\epsilon < \epsilon_0$,

$$A_{0,m}(\gamma/\epsilon)^{1/m} \leq H_\epsilon(W^m_{2,0}(\gamma)) \leq A_{1,m}(\gamma/\epsilon)^{1/m} \quad \epsilon \to 0. \tag{1.10}$$

This result displays some paradigmatic features of the ϵ-entropy approach. First, the result does tie down the precise rate involved in the growth of the net (i.e. $H_\epsilon = O(\epsilon^{-1/m})$ as $\epsilon \to 0$). Second, it does not tie down the precise constants involved in the decay (i.e. $A_0 \neq A_1$). Third, the result (and its proof) does not directly exhibit information about the properties of an optimal ϵ-net.

1.3 Comparison

There are some formal similarities between the problems addressed by Shannon's $R(D)$ and Kolmogorov's H_ϵ. To make these clear, notice that in each case, we consider a "library" – either a function class $\mathcal{F}$ or a stochastic process X, each case yielding as typical elements functions defined on a common domain T – and we measure approximation error by the same norm $\|\cdot\|$. In both the Shannon and Kolmogorov theories we represent by first constructing finite lists of representative elements – in one case the list is called a codebook; in the other case, a net. We represent an object of interest by its closest representative in the list, and we may record simply the index into our list. The length in bits of such a recording is called in the Shannon case the rate of the codebook; in the Kolmogorov case, the entropy of the net. Our goal is to minimize the number of bits while achieving sufficient fidelity of reproduction. In the Shannon theory this is measured by mean discrepancy across random realizations; in the Kolmogorov theory this is measured by the maximum discrepancy across arbitrary members of $\mathcal{F}$. These comparisons may be summarized in tabular form:

	Shannon Theory	Kolmogorov Theory
Library	X Stochastic	$f \in \mathcal{F}$
Representers	Codebook $\mathcal{C}$	Net $\mathcal{N}$
Fidelity	$E \min_{X' \in \mathcal{C}} \|X - X'\|^2$	$\max_{f \in \mathcal{F}} \min_{f' \in \mathcal{N}} \|f - f'\|^2$
Complexity	$\log \#\mathcal{C}$	$\log \#\mathcal{N}$

In short, the two theories are entirely parallel – except that one of the theories postulates a library of samples arrived at by sampling a stochastic process, while the other selects arbitrary elements of a functional class.

While there are intriguing parallels between the $R(D)$ and H_ϵ concepts, the two approaches have developed separately, in very different contexts. Work on $R(D)$ has mostly stayed in the original context of communication/storage of random process data, while work with H_ϵ has mostly stayed in the context of questions in analysis: the Kolmogorov entropy numbers control the boundedness of Gaussian processes [6] and the properties of certain operators [3, 7] and Convex Sets [14].

Because of the different contexts, the philosophy of what it means to "calculate" $R(D)$ is usually quite different from what it means to "calculate" H_ϵ. $R(D)$ calculations are attractive because they establish absolute bounds on the number of bits required for *any* compression system as an absolute standard of comparison with practical schemes. For this purpose, $R(D)$ results are in principle only interesting when they are precise, specifying the exact minimal number of bits required to within a factor $(1 + o(1))$, where the asymptotic required is either as the domain T grows, $|T| \to \infty$, or else as the fidelity requirement becomes increasingly stringent, $D \to 0$. In contrast, H_ϵ-calculations are typically used simply to establish qualitative properties, such as the finiteness of certain norms; and then only the order asymptotics matter, i.e. results like (1.9); for such applications, constants are of little interest. Applications of Kolmogorov entropy in statistics, for example, [2, 11] have not typically required the precise evaluation of constants.

The tools available for "calculating" $R(D)$ and H_ϵ also differ considerably. In $R(D)$ there is a well-understood "grand formalism" based leading to the optimization problem (1.4) which gives the formally optimum number of bits and which has been shown in a variety of settings to rigorously provide the right answer to within $(1+o(1))$ factors. Moreover $R(D)$ theory proposed a random coding argument for constructing a rate-optimal codebook. In H_ϵ work, there is no similar "grand formalism" suggesting what the value of H_ϵ ought to be; and there is no general principle suggesting how to construct an ϵ-net.

As a result of the differences in philosophies and heuristics, there appear to be many examples where $R(D)$ is known exactly or asymptotically, but relatively few examples where H_ϵ is known precisely, or where the structure of an optimal ϵ-net is known. In essence, the infrastructure of $R(D)$ is richer, and the outcome of its application is often more precise, than the corresponding infrastructure and results in H_ϵ work.

Underscoring this point is the commentary of V. M. Tikhomirov in Kolmogorov's *Selected Works* [23]

> The question of finding the exact value of the ϵ-entropy ... is a very difficult one .... Besides [one specific example] ... the author of this commentary knows no meaningful examples of infinite-dimensional compact sets for which the problem of ϵ-entropy ... is solved exactly.
>
> ... A result of fundamental importance about ϵ-entropy of functions of finite smoothness was obtained by Birman and Solomyak ... [who proved, for a range of r, p, n and q]
>
> $$H_\epsilon(W_p^r, L_q(I^n)) \asymp \epsilon^{-n/r}, \quad \epsilon \to 0.$$
>
> It is easy to show that actually the following limit
>
> $$\lim_{\epsilon \to 0} \quad \epsilon^{n/r} \quad H_\epsilon(W_p^r, L_q(I^n)) = (r, p, q, n)$$
>
> exists. In [Kolmogorov and Tikhomirov] $\kappa(1, \infty, \infty, 1)$ was computed. As far as I know there is no other case where the problem of the strong asymptotics of H_ϵ is solved. [i.e. where $\kappa(r, p, q, n)$ is known].

1.4 Results

This paper will develop a link between $R(D)$ and H_ϵ ideas which will allow us to use the $R(D)$ infrastructure to determine the precise asymptotics of $H_\epsilon(W_{2,0}^m(\gamma))$, i.e. to evaluate the constant which Tikhomirov called $\kappa(r, 2, 2, 1)$. The same ideas apply easily to evaluating the precise asymptotics of $H_\epsilon(\mathcal{F})$ for many other ellipsoids, although we omit the exercise. For example,

it is rather obvious that the same method gives the value of $\kappa(r,2,2,n)$ for every dimension $n = 1, 2, 3, \ldots$ and every r.

We introduce the *worst rate distortion function of the functional class* $\mathcal{F}$ by

$$R^*(D, \mathcal{F}) = \sup\{R(D, X) : \quad P(X \in \mathcal{F}) = 1\}. \tag{1.11}$$

In words: this seeks a process living on $\mathcal{F}$ which is hardest to compress in Shannon's sense while maintaining a fidelity D.

There is a simple connection between H_ϵ and R^*: with $D = \epsilon^2$,

$$H_\epsilon(\mathcal{F}) \geq R^*(D, \mathcal{F}). \tag{1.12}$$

To see this, suppose that $\mathcal{C}$ is a D-admissible codebook for X, i.e. that

$$E \min_{X' \in \mathcal{C}} \|X - X'\|^2 \leq D;$$

then the converse source coding theorem (Theorem 3.2.2 in Berger (1971), page 70) says

$$\log \#\mathcal{C} \geq R(D, X).$$

On the other hand, if $\mathcal{N}_\epsilon$ is an ϵ-net for $\mathcal{F}$ then

$$\max_{f \in \mathcal{F}} \min_{f' \in \mathcal{N}_\epsilon} \|f - f'\|^2 \leq \epsilon^2.$$

We conclude that for any process obeying $X \in \mathcal{F}$ with probability one, the net $\mathcal{N}_\epsilon$ is a D-admissible codebook with $D = \epsilon^2$. Hence $\log \#\mathcal{N}_\epsilon \geq R(D, X)$ for any such X.

Theorem 1.1 *Let $m \in \{1, 2, \ldots\}$ and $0 < \gamma < \infty$.*

$$\lim_{\epsilon \to 0} \frac{H_\epsilon(W^m_{2,0}(\gamma))}{R^*(\epsilon^2, W^m_{2,0}(\gamma))} \quad = \quad 1.$$

In words: the leading asymptotics of $H_\epsilon(W^m_{2,0}(\gamma))$ are precisely the same as the leading asymptotics of $R^*(D, W^m_{2,0}(\gamma))$, under the calibration $D = \epsilon^2$.

Theorem 1.2 *There exists, for each D, a certain Gaussian process X^*_D, which is nearly supported in $W^m_{2,0}(\gamma)$,*

$$P(X^*_D \in W^m_{2,0}(\gamma)) \to 1, \qquad D \to 0,$$

and which is asymptotically least favorable in the sense that

$$\lim_{D \to 0} \frac{R^*(D, W^m_{2,0}(\gamma))}{R(D, X^*_D)} \quad = \quad 1.$$

Moreover, the process $\tilde{X}_D$ which is X^*_D conditioned to lie in $W^m_{2,0}(\gamma)$ obeys

$$R(D, X^*_D)/R(D, \tilde{X}_D) \to 1, \quad D \to 0.$$

In words: there is a certain Gaussian process which almost 'lives in $W^m_{2,0}(\gamma)$' and which is almost the hardest to compress to fidelity D. About this process we can say the following:

Theorem 1.3

$$R(D, X^*_D) \sim 2m(\log_2 e)(\gamma^2/2D)^{\frac{1}{2m}} \quad , \quad D \to 0.$$

In sum, H_ϵ behaves asymptotically like the rate of the least favorable stochastic process living in $W^m_{2,0}(\gamma)$; and we can essentially identify this process and calculate its rate.

1.5 Codebook Structure

The proof of Theorem 1.1 constructs an ϵ-net, $\mathcal{N}_\epsilon$. This is a finite set, which we may sample according to the uniform probability measure. An f' chosen at random from $\mathcal{N}_\epsilon$ will have a probability distribution that closely resembles the distribution of Y_D^*, where Y_D^* solves the minimum-mutual-information problem in Shannon's $R(D, X_D^*)$ (see (1.4). Indeed, in line with work of Sakrison [17, 18, 19], we know that "most" codewords in a near optimal codebook for a class of processes will be sampled from the so called "reproducing distribution" of the "worst source."

Now, a byproduct of the proof of Theorems 1.1-1.3 is information about the process Y_D^*; its Fourier coefficients take the form

$$(\hat{Y}_D^*)_{2k-1} = \beta \cdot (1 + k^{2m})^{-1/2} \cdot Z_{2k-1}, \qquad (\hat{Y}_D^*)_{2k} = \beta \cdot (1 + k^{2m})^{-1/2} \cdot Z_{2k} \quad 1 \le k \le k_0,$$

for appropriate scalars $\beta(D, \gamma)$ and $k_0(D, \gamma)$, where (Z_k) is an i.i.d. Gaussian white noise, with vanishing coefficients at $k > 2k_0$, and with $(\hat{Y}_D^*)_0 = \beta Z_0$. In short, the typical codeword is essentially a trigonometric polynomial with coefficients drawn independently from a normal distribution with variance at the k-th coefficient a simple function of k.

1.6 Wald

These results display interesting parallels with important results in a field started by Abraham Wald: statistical decision theory. In fact, it is not too much to say that a key pre-requisite for the above evaluation of $H_\epsilon(W_{2,0}^m(\gamma))$ is a strong familiarity with modern statistical decision theory.

A central result in Wald's *oeuvre* is the minimax theorem [24]. Suppose that we are given random data $Y \sim P_\theta$, where $\theta \in \Theta$. We wish to recover an estimate of θ from the data Y and we use the minimax criterion:

$$\text{Minimax Risk}(\Theta) = \inf_{\hat{\theta}} \sup_{\Theta} E_{P_\theta} \ell(\hat{\theta}(Y), \theta).$$

Here $\ell(d, \theta)$ is a loss function, and $\hat{\theta} = \hat{\theta}(Y)$ is a measurable function of the data Y. This criterion is "purely frequentist": it says that we know only that $\theta \in \Theta$ and we have no information about the "likelihood" that θ might takes values in certain subsets of Θ. The estimator is therefore determined by the geometry of the set Θ rather than our prejudices; the frequentist approach is "objective", a process of "rational estimator design". It is also challenging, since it is posing a rather difficult optimization problem, and in only a few cases has there been success in obtaining exact solutions.

In contrast, there is the Bayesian approach, where we posit that θ is in fact random, a realization of a random variable ξ distributed according to the prior distribution π. If we make a clever choice of prior π, we may simply calculate. For example, if $\ell(,)$ is a quadratic loss, then the Bayes estimate is simply $E\{\xi|Y\}$, where the expectation is over $P_\theta(dY)\pi(d\theta)$. In many cases there are priors leading to concrete algorithms for such expectations, a particularly well-known case being when π is a Gaussian prior and P_θ is Gaussian also.

Wald's minimax theorem says that

$$\text{Minimax Risk}(\Theta) = \max\{\text{Bayes Risk}(\pi) : \mathit{supp}(\pi) \subset \Theta\}. \tag{1.13}$$

To know the minimax risk for estimating a deterministic parameter $\theta \in \Theta$, one considers Bayesian estimates with prior $\theta \sim \pi$ supported in Θ and one analyzes the structure of the Bayes Risks associated with such priors. The minimax risk is precisely the largest possible Bayes Risk. The optimizing prior π^* is the "least-favorable prior"; it is the hardest prior distribution on θ which the statistician can face. The minimax procedure, although defined by a purely frequentist route, is the Bayes estimate associated with the least favorable prior.

In short, we have the ...

> **Stategy for Minimax Decision Theorists:** *To compute the minimax risk, one should think in Bayesian terms.*

That is, to understand the intrinsic difficulty of estimating an object in the worst-case sense, one should study the difficulty of specific cases in the Bayes theory, and look for the least-favorable situation in the Bayes theory.

Turn now to the main result of this lecture: the asymptotic relation of Theorem 1.1.

$$H_\epsilon(\mathcal{F}) = \sup\{R(D,X) : P(X \in \mathcal{F}) = 1\}(1 + o(1)), \quad D = \epsilon^2 \to 0. \tag{1.14}$$

This exhibits a strong parallel to Wald's theorem (1.13), and suggests the ...

Strategy for ϵ-entropy Theorists: *To compute ϵ-entropy one should think in "Shannon-ian" terms.*

That is, to understand the intrinsic difficulty of compressing an object in the worst-case theory, one should study the difficulty of specific cases in the Shannon theory, and look for the least-favorable situation in the Shannon theory.

We explain these slogans in more detail. The term on the left side of (1.14), the Kolmogorov ϵ-entropy, is based on an assumption that we only know $f \in \mathcal{F}$ and nothing more. An optimal ϵ-net for the Kolmogorov theory reflects the geometry of $\mathcal{F}$ rather than any prejudice we may have about the "likelihood" that object to be represented will occupy subsets of $\mathcal{F}$. Finding such a net poses an interesting, but challenging problem, solved in very few cases, as remarked above.

The term on the right side of (1.14) involves items from Shannon's theory, where we know the object to be compressed is a realization of a stochastic process X. Shannon's theory gives an explicit optimization problem to solve in order to determine an asymptotically optimal compression scheme for such random-process reaslizations, and for stochastic processes of nice analytic structure, one can actually solve the required optimization problem, the case of squared-L^2 norm error measure and Gaussian process X being a famous example.

We now summarize the parallels in a table:

Setting	Component	Statistical Counterpart	Data Compression Counterpart
Worst-Case			
	Icon	Wald	Kolmogorov
	Assumption	$\theta \in \Theta$	$f \in \mathcal{F}$
	Construct	Minimax Estimator	Minimal ϵ-Net
	Optimality	Minimax Risk	ϵ-entropy
Average-Case			
	Icon	Bayes	Shannon
	Assumption	θ a realization of ξ	f a realization of X
	Construct	Bayesian Estimation	Minimal D-Codebook
	Optimality	Bayes Risk	Rate-Distortion
	Easy Example	L^2-loss, Gaussian ξ	L^2-error, Gaussian X

A particularly striking parallel in this table: the "Kolmogorov-ians" in the data compression world are like the Wald-ians in the world of theoretical statistics , while "Shannon-ians" in the data compression world are like the Bayesians in the world of theoretical statistics. For example, "Shannon-ians" and Bayesians both have their optimal behavior spelled out for them, in the one case by rate-distortion theory, and in the other case by Bayes-formula; while Wald-ians and Kolmogorov-ians face in each new case difficulties of determining how to proceed.

This set of parallels suggests the "Strategy for Data Compression Theorists" above.

1.7 Pinsker

The parallel we are drawing is more than an afterthought; it is the source of Theorems 1.1-1.3 above. In fact, Theorems 1.1-1.3 have exact analogs in statistical decision theory which have been known for a long time, and which have had many interesting consequences over the years.

M. S. Pinsker solved in 1980 [13] a key problem in statistical decision theory: the problem of asymptotic minimax estimation of a function $f \in W^m_{2,0}(\gamma)$ from white noise observations

$$Y(dt) = f(t)dt + \epsilon W(dt) \quad t \in [0, 2\pi);$$

here the asymptotic is as $\epsilon \to 0$.

Pinsker considered the problem

$$\min_{\hat{f}(Y)} \max_{W^m_{2,0}(\gamma)} E\|f - \hat{f}(Y)\|^2_{L^2}$$

and found that he could evaluate precisely the asymptotics of this risk, by pursuing essentially the strategy of the "Slogan for Decision Theory". He considered a Bayesian approach, viewing f as a realization of a Gaussian processes X. He restricted attention to Gaussian processes which asymptotically concentrate in $W^m_{2,0}(\gamma)$. Pinsker showed that the above worst case risk was asymptotic to the Bayes risk of the least favorable Gaussian process concentrating asymptotically in $\mathcal{F}$, and he was able to evaluate the precise asymptotics of the Bayes risk.

Conceptually, this paper is entirely parallel to Pinsker's work. Indeed if we use the above table to guide us in making the formal substitutions in Pinsker's results of terms like "Rate-Distortion" for "Bayes-Risk" and so on, we arrive at the correct form of propositions to be proved in Theorems 1.1-1.3. (The proofs, however, seem to us unrelated, and the specific evaluated constants appear to us to have no connection.)

This theme of exact parallels between data compression and statistical estimation will be developed at greater length in the sequel, Wald Lecture II.

1.8 Contents

In Section 2 we study properties of ϵ-nets for spheres and codes for Gaussian i.i.d. sequences. In section 3 we use the results on ϵ-nets for spheres to get upperbounds on H_ϵ for ellipsoids. In Section 4 we study the $R(D)$ function for Gaussian processes which lie in ellipsoids in quadratic mean. In Section 5 we show that the largest $R(D)$ is asymptotically equaivalent to $H_\epsilon(\epsilon = D^2)$. In Sections 6 and 7 we identify and evaluate the largest $R(D)$.

2 Asymptotic Properties of ℓ^2_n-balls

In this section we develop a theory in miniature for ℓ^2_n balls which exactly parallels our main results. The later results are then obtained by creatively using these results, pasting together many copies of ℓ^2_n balls.

Let $L_n(\rho)$ denote the ℓ^2_n-ball of radius $\rho\sqrt{n}$, and let

$$h_2(\rho, \epsilon; n) = n^{-1} H_{\epsilon\sqrt{n}}(L_n(\rho), \|\cdot\|_{\ell^2})$$

denote the dimension-normalized Kolmogorov entropy of the ℓ^2_n ball of radius $\rho\sqrt{n}$, at discrepancy level $\epsilon\sqrt{n}$.

Theorem 2.1 *For $\rho > \epsilon > 0$,*

$$\lim_{n\to\infty} h_2(\rho, \epsilon; n) = \log\left(\frac{\rho}{\epsilon}\right). \tag{2.1}$$

Proof. This is merely a restatement, in our terminology, of various known results. (2.1) follows immediately from the very useful chain of inequalities

$$\log_+(\rho/\epsilon) \le h_2(\rho, \epsilon; n) \le \log_+(\rho/\epsilon) + \frac{C_1 \log(n) + C_2}{n}; \tag{2.2}$$

the left-hand inequality valid for all n and the right-hand inequality valid (at least) for $n \ge 9$.

The left-hand inequality is based on very elementary 'volume arguments'. Suppose N balls of radius ϵ cover a ball of radius ρ; then

$$N \cdot \mathrm{vol}(L_n(\epsilon)) \ge \mathrm{vol}(L_n(\rho)),$$

so $N \ge \mathrm{vol}(L_n(\rho))/\mathrm{vol}(L_n(\epsilon)) = (\rho/\epsilon)^n$, and hence

$$\log N \ge n \log(\rho/\epsilon).$$

As h_2 is the minimal value of $\log(N)/n$ over all coverings, the left-hand inequality of (2.2) follows.

The right-hand imequality of (2.2) follows from Theorem 3 of C.A. Rogers (1963), which says:

> There is an absolute constant c such that if $R > 1$ and $n \geq 9$, an n-sphere of radius R can be covered by fewer than $c \cdot n \log n \cdot R^n$ spheres of radius 1, when $R > n$, and by fewer than $c \cdot n^{5/2} R^n$, when $R < n$.

We remark that the number of ϵ-spheres to cover a ρ-sphere is the same as the number of 1 spheres to cover a ρ/ϵ-sphere. Letting $N_n(\rho, \epsilon)$ denote the minimal number of covering spheres then Rogers' theorem says that, if $\rho > \epsilon$,

$$n^{-1} \cdot \log N_n(\rho, \epsilon) \leq \begin{cases} \log(\rho/\epsilon) + \log(c \cdot n \log n)/n & \rho/\epsilon > n \\ \log(\rho/\epsilon) + \log(c \cdot n^{5/2})/n & 1 \leq \rho/\epsilon \leq n \end{cases}$$

The right-hand inequality of (2.2) follows trivially. ∎

We remark that although Rogers' covering is not constructive, a result of Wyner (1967) implies that one can use random coverings.

An important feature of this result is that the form of the dependence on ρ and ϵ agrees precisely with rate-distortion theory.

To see this, let $X_1, \ldots, X_n$ be iid $N(0, \sigma^2)$. Let $\mathcal{C}_n(D)$ be a codebook of n-tuples with the property that

$$E \min_{X' \in \mathcal{C}_n} \frac{1}{n} \sum (X_i - X'_i)^2 \leq D. \tag{2.3}$$

Then a cornerstone of $R(D)$ theory is that

$$n^{-1} \log \#\mathcal{C}_n(D) \geq \frac{1}{2} \log_+ \Big(\frac{\sigma^2}{D} \Big), \tag{2.4}$$

where $\log_+(t) = (\log(t))_+$. Moreover there exists a sequence of codebooks $(\mathcal{C}^*_n)_n$ satisfying (2.3) and achieving (2.4) asymptotically:

$$\lim_{n \to \infty} \quad n^{-1} \log \#\mathcal{C}^*_n(D) = \frac{1}{2} \log_+ \Big(\frac{\sigma^2}{D} \Big). \tag{2.5}$$

For a full discussion, see Sakrison (1968). Here if we replace D by ϵ^2 and σ by ρ, we get a formal expression very similar to Theorem 2.1. We now explain this connection.

Let $R_n(D, X)$ denote the rate-distortion function for a random vector X with respect to dimension-normalized fidelity measure $\|X - Y\|^2_{2,n} = \frac{1}{n} \sum_{i=1}^n (X_i - Y_i)^2$. Define the worst such rate distortion function for the ball $L_n(\rho)$ by

$$R^*_n(D, \rho) = \sup\{n^{-1} R_n(D, X) : P(X \in L_n(\rho)) = 1\}.$$

Theorem 2.2 *For $\rho^2 > D > 0$,*

$$\lim_{n \to \infty} R^*_n(D, \rho) = \log \Big(\frac{\rho}{\sqrt{D}} \Big). \tag{2.6}$$

There is a sequence $X_{(n)}$ of Gaussian random vectors asymptotically concentrating in $L_n(\rho)$ and asymptotically achieving the indicated rate; the corresponding sequence $X'_{(n)}$ which is $X_{(n)}$ conditioned to lie in $L_n(\rho)$ achieves asymptotic equality here.

Proof.

For $n \geq 1$ let $X_{(n)} = (X_1, \ldots, X_n)$ where the components X_i are i.i.d. $N(0, \sigma^2_n)$, and $\sigma^2_n = \rho^2(1 - .1/\log(n+1))$. The choice of σ^2_n guarantees

$$P\{X_{(n)} \in L_n(\rho)\} \to 1, \qquad n \to \infty;$$

so the Gaussian vector $X_{(n)}$ is, with overwhelming probability, in $L_n(\rho)$. Define a random vector $X'_{(n)}$ as $X_{(n)}$ conditioned to lie in $L_n(\rho)$. Then $P(X'_{(n)} \in L_n(\rho)) = 1$ and so $X'_{(n)}$ is eligible for competition in the problem defining $R^*_n(D,\rho)$. We will argue in a moment that

$$\lim_{n\to\infty} R_n(D, X'_{(n)})/R_n(D, X_{(n)}) \geq 1. \tag{2.7}$$

This, combined with

$$R_n(D, X_{(n)}) = \frac{1}{2}\log_+\left(\frac{\sigma_n^2}{D}\right) \to \log\left(\frac{\rho}{\sqrt{D}}\right),$$

implies the Theorem. It remains to prove (2.7). We need a technical lemma showing the closeness of Rate-Distortion curves of nearby random variables.

Lemma 2.3 *Suppose that there is a common probability space on which random variables X and X' are defined, and that on this space we have a mapping Q yielding $X' = Q(X,Z)$ where Z is independent of X. Suppose also that $E\|X' - X\|^2 \leq \eta$, where $0 \leq \eta < D$. Then*

$$R(D, X') \geq R((\sqrt{D} + \sqrt{\eta})^2, X). \tag{2.8}$$

Proof of Lemma. Suppose that (X', Y') achieve the mutual information optimization problem associated with the Rate-Distortion curve $R(D, X')$ at distortion D. By hypothesis, we have a Markov Chain

$$X \to X' \to Y'$$

and so by the "Data Processing Inequality" of Information Theory, X' and Y' have greater mutual information than X and Y'. Compare Cover and Thomas [4] section 2.8, page 32. In addition,

$$(E\|X - Y'\|^2)^{1/2} \leq (E\|X - X'\|^2)^{1/2} + (E\|X' - Y'\|^2)^{1/2}$$

so that Y' is $(D^{1/2} + \eta^{1/2})^2$-admissible for the Rate-Distortion optimization problem associated with X. Hence $I(X, Y') \geq R((\sqrt{D} + \sqrt{\eta})^2, X)$. In short

$$R(D, X') = I(X', Y') \geq I(X, Y') \geq R((\sqrt{D} + \sqrt{\eta})^2, X).$$

and (2.8) follows. ∎

We now apply this lemma to obtain (2.7): we simply construct the appropriate mapping connecting X and X' and then calculate the distance between realizations.

We view $X_{(n)}$ and $X'_{(n)}$ as both defined on a probability space of $U = X_{(n)}/\|X_{(n)}\|_{2,n}$ and infinitely many copies S_1, S_2, ..., i.i.d. S, with $S = \|X_{(n)}\|_{2,n}$. Then $X_{(n)} = S \cdot U$ and $X'_{(n)} = S' \cdot U$, where $S = S_1$ and S' is the first among the copies S_i which is less than ρ. Then with probability close to 1, $X'_{(n)} = X_{(n)}$. The distance $E\|X'_{(n)} - X_{(n)}\|^2_{2,n} = E(S' - S)^2$. Now

$$E(S' - S)^2 = P(S > \rho) \cdot E\{(S_1 - S_2)^2 | S_1 > \rho, S_2 < \rho\}.$$

As it turns out $S^2 \cdot n/\sigma_n^2$ has a chi-squared distribution with n degrees of freedom. Standard exponential inequalities for the tails of a chi-squared distribution allow us to see that $P(S > \rho)$ is tending to zero rapidly with n, while the expectation term is $O(1)$. Hence

$$R_n(D, X'_{(n)}) \geq R_n(D + o(1), X_{(n)}) = R_n(D, X_{(n)})(1 + o(1)).$$

This completes the proof of Theorem 2.2.

3 Upper Bounds on H_ϵ by ℓ^2_n-Balls

We now construct a net which can be used to attain $H_\epsilon(1 + o(1))$. It is based on the idea of dividing Fourier coefficients into subbands, and representing the vector of coefficients in a subband as a point in a sphere.

3.1 Decomposition in Subbands

Let $(\phi_k)_{k\geq 0}$ be an orthogonal basis for $L^2(T)$ built from sinusoids, so that $\phi_0(t) = 1$ and, for $k > 0$,

$$\phi_{2k}(t) = \sqrt{2}\cos(kt); \quad \phi_{2k-1}(t) = \sqrt{2}\sin(kt).$$

Let $\theta = (\theta_k)_{k\geq 0}$ denote the vector of coefficients $\theta_k = \frac{1}{2\pi}\int_0^{2\pi} \phi_k(t)f(t)dt$. Then $f(t) = \sum_{k\geq 0}\theta_k\phi_k(t)$ and $f^{(m)}(t) = \sum_{k>0} k^m(\tilde{\theta}_{2k}\phi_{2k}(t) + \tilde{\theta}_{2k-1}\phi_{2k-1}(t))$ where for $k > 0$, $\tilde{\theta}_k = (-1)^{m/2}\theta_k$, if k is even, and $\tilde{\theta}_k = (-1)^{(m-1)/2}\theta_k$ if k is odd. Because of this, and the Parseval relation $\|f\|_{L^2}^2 = \sum\theta_k^2$, we have that the body $\Theta_2^m(\gamma)$ of all coefficient sequences arising from $W_{2,0}^m(\gamma)$, obeys

$$\Theta_2^m(\gamma) = \{\theta : \sum a_k\theta_k^2 \leq \gamma^2\},$$

where $a_0 = 1$, and $a_{2k} = a_{2k-1} = 1 + k^{2m}$, $k > 0$. In short, $\Theta_2^m(\gamma)$ is an ellipsoid.

Define now a partition of the "frequency domain" $k \geq 0$ into "subbands" $K_b, b = 0, 1, \ldots$ of the form $K_b = \{k_b, k_b + 1, \ldots, k_{b+1} - 1\}$ with the endpoints obeying

$$k_{b+1} - k_b \to \infty, \qquad k_{b+1}/k_b \to 1, \qquad b \to \infty. \tag{3.1}$$

We use specifically the choice that k_{b+1} is the least even integer exceeding $(1 + \frac{1}{\sqrt{b}})k_b$ for $b \geq 2$ with $k_0 = 10$ and $k_1 = 20$). The subbands K_b are getting wide asymptotically, but narrow compared to their distance from 0. For future use, put $n_b = k_{b+1} - k_b$.

Let $\alpha_b^+ = \max\{a_k : k \in K_b\}, \alpha_b^- = \min\{a_k : k \in K_b\}$. Then from (3.1)

$$\alpha_b^+/\alpha_b^- \to 1, \quad b \to \infty. \tag{3.2}$$

The sets Θ^+, Θ^- defined by

$$\Theta^\pm = \{\theta : \sum_b \alpha_b^\pm \sum_{k\in K_b} \theta_k^2 \leq \gamma^2\}$$

obey

$$\Theta^- \subset \Theta_2^m(\gamma) \subset \Theta^+;$$

in a sense all three are tail asymptotic at high frequency, which ultimately means that the number of bits required to describe any one of the sets will be asymptotically equivalent to the number of bits required to describe any of the other sets, to within accuracy ϵ, as $\epsilon \to 0$. Since Θ^+ and Θ^- treat all frequencies within a subband evenhandedly, this will show that asymptotically efficient codes can treat all frequencies in a subband evenhandedly.

3.2 Hardest Cartesian Subproblem

Θ^+ and Θ^- are in some sense simply-parametrized unions of ℓ_n^2-balls. If $(\rho_b)_{b\geq 0}$ is a sequence of radii obeying

$$\sum \alpha_b^\pm \rho_b^2 n_b \leq \gamma^2,$$

then

$$\Theta^\times((\rho_b)) \equiv \Pi_{b=1}^\infty L_{n_b}(\rho_b) \subset \Theta^\pm,$$

and in some sense the collection of all such products exhausts $\Theta^\pm$.

In the next section we will describe a simple coding scheme based on this picture of "cartesian products of ℓ^2-balls." The motivating idea is that a given vector θ which we might want to ϵ-describe can be viewed as living in one of many such cartesian products embedded in Θ^+:

$$\theta \in \Theta^\times((\rho_b)) \subset \Theta^+. \tag{3.3}$$

For each one of these cartesian products, there is a code for θ, with codelength

$$H_\epsilon(\Theta^\times((\rho_b))). \tag{3.4}$$

So we could code for θ by adaptively selecting from, among all such products, the one giving the shortest length for that θ. As it turns out, this can be achieved simply by coding subbands using codes for ℓ^2 balls. A complete code would of course also have to include a description of the selected (ρ_b). But as there is only one parameter ρ_b per subband, while the number of coefficients in a subband $n_b \to \infty$ as $b \to \infty$, in some asymptotic sense, a negligible amount of information is needed to parametrize the radii (ρ_b), as compared with specifying coefficients within subbands.

The motivation for this approach is provided by three observations.

- First, *the length in bits of an ϵ-description for Θ^+ is not essentially* **larger** *than the length in bits of an ϵ-description for the most-difficult-to-describe cartesian product inscribed in Θ^+:*
$$H_\epsilon(\Theta^+) \le \sup\{H_\epsilon(\Theta^\times((\rho_b))) : \Theta^\times((\rho_b)) \subset \Theta^+\} \cdot (1 + o(1)).$$
- Second, *the length in bits of an ϵ-description for $\Theta_2^m(\gamma)$ is never* **smaller** *than the length in bits of an ϵ-description for the most-difficult-to-describe cartesian product of balls inscribed in Θ^-:*
$$H_\epsilon(\Theta^-) \ge \sup\{H_\epsilon(\Theta^\times((\rho_b))) : \Theta^\times((\rho_b)) \subset \Theta^-\}.$$
- Third, *the upper and lower bounds are asymptotically equivalent*, which will emerge effectively from (4.6) below. Hence this approach by subband coding is asymptotically efficient.

To work with (3.4), note that if (δ_b) is any sequence of subband fidelities satisfying $\sum_b \delta_b^2 n_b \le \epsilon^2$, then
$$H_\epsilon(\Pi_{b=1}^\infty L_{n_b}(\rho_b)) \le \sum_b H_{\delta_b}(L_{n_b}(\rho_b)) = \sum_b h_2(\rho_b, \delta_b; n_b) n_b.$$
By adaptively allocating fidelities dependent on the given object θ, we can essentially produce a codelength no larger than the smallest value of the right hand side for any product obeying (3.3).

3.3 Coding Subbands

To begin with we need two parameters, which depend on ϵ, the desired distortion level, and on the parameters (m, γ).

(1) *Subband Cutoff.* Fix $B(\epsilon) = B(\epsilon; m, \gamma)$ so that k_B is the first subband border larger than $(\gamma/\epsilon^{2.1})^{1/2m}$. Then
$$\sup_{\Theta_2^m(\gamma)} \sum_{k \ge k_B} \theta_k^2 \;\le\; \epsilon^{2.1}.$$

(2) *Parameter Quantization.* Set $q_\epsilon = \epsilon^4$. Values of parameters of balls – radii and fidelities – will only be stored to accuracy q_ϵ.

We propose to code θ by partitioning into a sequence of subbands $(\theta_k : k \in K_b)$ and coding the b-th subband, for $b = 0, \ldots, B$, using the code for an ℓ_{2,n_b} ball of relative radius r_b and relative fidelity δ_b. In detail this works as follows:

A. *Quantization of Subband Radii.* Given θ, find the subband energies $\rho_b^2 = n_b^{-1} \cdot \sum_{k \in K_b} \theta_k^2$. Form quantized radii, i.e. find the smallest r_b of the form $k_b \cdot q_\epsilon$ for integer k_b so that $\rho_b \le r_b$.

B. *Allocation of Fidelity to Subbands.* Obtain a finite sequence δ_b of the form $k'_b q_\epsilon$ for integers (k'_b) which obeys
$$\sum_{b=0}^{B(\epsilon)} \delta_b^2 n_b \le \epsilon^2 - \epsilon^{2.1} \tag{3.5}$$
and, subject to this constraint, minimizes
$$\sum_{b=1}^{B(\epsilon)} h_2(r_b, \delta_b; n_b) \cdot n_b. \tag{3.6}$$

This is an optimal allocation of ϵ^2 fidelity allowance to subbands. ((3.5) will be possible for all sufficiently small $\epsilon > 0$).

C. The output code will consist of the concatenation of a binary encoding of a *prefix*

$$(k_b, b = 0, \ldots, B(\epsilon)), (k'_b, b = 0, \ldots, B(\epsilon))$$

where k_b and k'_b are as in A. and B. above, joined to the binary encoding of the *body*

$$(i_b, b = 0, \ldots, B(\epsilon))$$

where i_b is the index of the codeword closest to the subband coefficients $(\theta_k : k \in K_b)$ in an $\sqrt{n_b} \cdot \delta_b$ covering of $\ell_n^2(\sqrt{n_b} \cdot r_b)$

This coder is attempting to actively minimize the number of bits while maintaining fidelity by adaptively choosing different degrees of fidelity in different subbands. This fidelity allocation process can be viewed as a process of 'optimal bit allocation' preserving a fidelity constraint.

The length of the output code can be decomposed as

$$L(\epsilon) = L_0(\epsilon) + L_1(\epsilon),$$

where $L_0(\epsilon)$ is the length of the prefix, coding for (k_b) and (k'_b); this requires no more than

$$(B(\epsilon) + 1) \cdot 2 \log(\gamma/\epsilon^4)$$

bits, which gives $L_0(\epsilon) = O(\log^2(\epsilon^{-1}))$. (This code is fixed-length, independent of θ.) The body length $L_1(\epsilon)$ is the length for representing the different $(\theta_k : k \in K_b)$ by spherical ϵ_b-nets. The length

$$L_1(\epsilon, \theta) = \sum_b h_2(r_b, \delta_b; n_b) n_b$$

is variable; it cannot exceed

$$\max_{\theta \in \Theta_2^m(\gamma)} \min_{(\delta_b)} \sum h_2(r_b, \delta_b; n_b) n_b. \tag{3.7}$$

It will turn out that this behaves as $O(\epsilon^{-1/m})$, while the prefix length $L_0(\epsilon)$ is fixed and $O(\log^2(\epsilon^{-1}))$. Hence the leading asymptotics of the maximum code length in bits is determined by the optimization problem (3.7).

3.4 Asymptotic Optimization Problem

Consider now the optimization problem, defined on pairs $p = (\delta, \rho)$ of functions on $\mathcal{R}^+$, by

$$(\mathcal{P}_{\epsilon,\gamma}) \qquad \max_{\rho} \min_{\delta} \int_0^\infty \log_+(\frac{\rho(t)}{\delta(t)}) dt \text{ subject to } \left\{ \begin{array}{l} \int_0^\infty \rho^2(t) t^{2m} dt \le \gamma^2 \\ \int_0^\infty \delta^2(t) dt \le \epsilon^2 \end{array} \right. . \tag{3.8}$$

This represents the asymptotic form of the minimax problem defining the codelength of the previous section. Sums have become integrals, and $h_2(\rho, \delta; n)$ has become $\log_+(\rho/\delta)$.

This can be motivated as follows. Associated to a collection of subband coding parameters $(\rho_b)_b$, $(\delta_b)_b$ we can construct functions $\rho(t)$ and $\delta(t)$ of $t \in [0, \infty)$ by step-function interpolation. Let subintervals I_b, $b = 0, 1, 2, \ldots$ of the real line be defined by $I_b = [k_b, k_{b+1})$, $b \ge 0$. Then set

$$\delta(t) = \sum_b \delta_b 1_{I_b}(t), \qquad \rho(t) = \sum_b \rho_b 1_{I_b}(t). \tag{3.9}$$

The resulting pair $p = (\delta, \rho)$ obeys

$$\int_0^\infty \log_+(\frac{\rho(t)}{\delta(t)}) dt = \sum_b \log_+(\rho_b/\delta_b) n_b, \tag{3.10}$$

and also

$$\int_0^\infty \delta^2(t)dt = \sum_b \delta_b^2 n_b. \tag{3.11}$$

As for the $2m$-moment integral, we have from $a_k \sim (k/2)^{2m}$ that

$$\int_0^\infty \rho^2(t)t^{2m}dt \approx 2^{2m} \cdot \sum_b \alpha_b^{\pm} \rho_b^2 n_b, \tag{3.12}$$

in the sense that for ρ which are highly "spread out", the relative difference of the two sides will be small. By this association, the sums occurring in the optimization problem (3.7) are converted exactly or approximately into integrals.

In the appendix we prove the following technical result, justifying passage from the discrete problem (3.7) to the continuum problem (3.8):

Theorem 3.1

$$\sup_{\theta \in \Theta_2^m(\gamma)} \quad L_1(\epsilon, \theta) \le val(\mathcal{P}_{\epsilon,\gamma/2})(1 + o(1)), \qquad \epsilon \to 0.$$

Consider the form and structure of the continuum optimization problem (3.8). The crucial point is the following homogeneity:

Theorem 3.2 *For $\epsilon, \gamma > 0$,*

$$val(\mathcal{P}_{\epsilon,\gamma}) = (\gamma/\epsilon)^{1/m} val(\mathcal{P}_{1,1}).$$

In words: the behavior of $val(\mathcal{P}_{\epsilon,\gamma})$ is completely determined by the product of the constant $val(\mathcal{P}_{1,1})$, and a simple power law in γ/ϵ.

Proof. This follows from a rescaling argument. (The exact value of $val(\mathcal{P}_{1,1})$ will be obtained in section 6.) Let $\mathcal{P}_{\epsilon,\gamma}$ be the set of all pairs $p = (\delta(t), \rho(t))$ jointly feasible for $\mathcal{P}_{\epsilon,\gamma}$. For a pair p, set $(U_{a,b}p) = (a\delta(bt), a\rho(bt))$. Then obviously

$$U_{a,b}\mathcal{P}_{\epsilon,\gamma} = \mathcal{P}_{ab^{-1/2}\epsilon, ab^{-m-\frac{1}{2}}\gamma}.$$

Setting $ab^{-m-\frac{1}{2}}\gamma = 1$, $ab^{-\frac{1}{2}}\epsilon = 1$, we get

$$U_{a,b}\mathcal{P}_{\epsilon,\gamma} = \mathcal{P}_{1,1}; \qquad U_{a^{-1},b^{-1}}\mathcal{P}_{1,1} = \mathcal{P}_{\epsilon,\gamma}.$$

Now for $J(p) = \int \log_+(\rho(t)/\delta(t))dt$ we have

$$J(U_{a,b}p) = b^{-1}J(p).$$

Hence,

$$\begin{aligned} val(\mathcal{P}_{\epsilon,\gamma}) = \sup\{J(p) : p \in \mathcal{P}_{\epsilon,\gamma}\} &= \sup\{J(U_{a^{-1},b^{-1}}p) : p \in \mathcal{P}_{1,1}\} \\ &= b \sup\{J(p) : p \in \mathcal{P}_{1,1}\} \\ &= (\gamma/\epsilon)^{-1/m} val(\mathcal{P}_{1,1}). \quad \blacksquare \end{aligned}$$

The proper interpretation of the situation is that

$$\sup_{\Theta_2^m(\gamma)} \quad L(\epsilon, \theta) \le val(\mathcal{P}_{1,1})(\gamma/2\epsilon)^{1/m}(1 + o(1)),$$

and, also, that there is a limiting "shape" to the pair (δ_b^*, ρ_b^*) optimizing $L(\epsilon; \theta)$. The corresponding continuum pair

$$p_\epsilon^* = (\sum \delta_b^* 1_{I_b}, \sum \rho_b^* 1_{I_b}),$$

behaves as

$$p_\epsilon^* \sim \mathcal{U}_{a^{-1},b^{-1}} p_{1,1},$$

where $p_{1,1}$ optimizes $val(\mathcal{P}_{1,1})$, and $a = a(\epsilon, \gamma/2)$, $b = b(\epsilon, \gamma/2)$ are as in the scaling theorem. More rigorously, we have for weak convergence

$$\mathcal{U}_{a,b} p^*_\epsilon \rightarrow_w p_{1,1}$$

so that on a standard scale, the sequence of "most difficult to describe objects" and allocation of fidelities in the "most efficient description" in $W^m_{2,0}(\gamma)$ are each converging to their respective standard shapes.

4 Least-Favorable Rate Distortion

In the introduction, we ask in (1.11) for the worst rate distortion among all stochastic processes living in $W^m_{2,0}(\gamma)$. We now consider a modified concept, which will be related to (1.11) below.

Let $R^+(D, W^m_{2,0}(\gamma))$ denote the worst rate-distortion at rate D among stochastic processes X which are Gaussian zero-mean stationary processes belonging to $W^m_{2,0}(\gamma)$ "in quadratic mean" – i.e. obeying

$$E\|X\|_2^2 + \|X^{(m)}\|_2^2 \leq \gamma^2. \tag{4.1}$$

This new assumption is not exactly comparable to the previous one. It is weaker than (1.11) because we have replaced the almost-sure constraint $P(X \in W^m_{2,0}(\gamma)) = 1$ by the in-mean constraint (4.1). But it is also stronger, because we are considering now only Gaussian processes.

Geometrically, (4.1) says that the so called "concentration ellipsoid" of such an X fits inside the ellipsoid $W^m_{2,0}(\gamma)$. This geometric interpretation is crucial for what follows.

It turns out that R^+ is attained within an even smaller subclass of processes, the Gaussian zero-mean processes with independent Fourier coefficients. Such processes take the form

$$X = \sum \sqrt{\lambda_k} Z_k \phi_k(t), \tag{4.2}$$

where the $(Z_k)_{k \geq 0}$ are i.i.d. $N(0,1)$ and the λ_k obey

$$\sum a_k \lambda_k \leq \gamma^2. \tag{4.3}$$

In short, it is enough to consider those Gaussian processes whose concentration ellipsoid is aligned with the axes of the ellipsoid $W^m_{2,0}(\gamma)$.

The usual way to calculate $R(D, X)$ for such a process uses the Shannon-Kolmogorov-Pinsker "parametric form" of solution (1.5)-(1.6). This has a "water-filling" interpretation which will be useful to us. Think of the subgraph $\{(k, y) : 0 \leq y \leq \lambda_k\}$ as a vessel with profile λ. Think of 'pouring in' to this vessel a 'fluid' whose area represents a total distortion D. Think of the figure oriented with the y-axis vertical, and suppose the fluid obeys gravity. Introduce "water-level" variables $(\mu_k)_{k \geq 0}$ satisfying $0 \leq \mu_k \leq \lambda_k$, representing the amount of distortion at frequency k. We can write

$$R(D) = \min \sum_k \log_+ \left(\frac{\lambda_k}{\mu_k}\right) \quad \text{subject to} \quad \sum_k \mu_k \leq D.$$

Indeed, according to the parametric form (1.5)-(1.6), the optimal choice of μ_k, μ^*_k say, obeys

$$\mu^*_k = \begin{cases} \lambda_k & \lambda_k \leq \theta \\ \theta & \lambda_k > \theta \end{cases},$$

where $D = D_\theta$. That is, the allocated distortions behave like a fluid in the presence of gravity – assuming a configuration with a "flat top" where they don't "touch the vessel".

Consider now the problem of finding least-favorable $R(D)$ behavior:

$$R^+(D, W^m_{2,0}(\gamma)) = \max\{R(D, X) : X \text{ satisfies } (4.2) - (4.3)\}$$

or

$$\max_\lambda \min_\mu \sum_k \log_+ \left(\frac{\lambda_k}{\mu_k}\right) \quad \text{subject to} \quad \begin{cases} \sum \mu_k \leq D \\ 0 \leq \mu_k \leq \lambda_k \\ \sum a_k \lambda_k \leq \gamma^2 \end{cases}. \tag{4.4}$$

We now remark that we may relax the constraints bounding μ by λ without changing the solution of the problem, since optimization of the $\log_+$ function would only reimpose them.

This problem is closely associated with a continuum optimization problem. Let subintervals J_k, $k = 0, 1, 2, \ldots$ of the real line be defined by $J_k = [k, k+1)$, $k \geq 0$. Introduce functions λ and μ via

$$\lambda(t) = \sum_k \lambda_k 1_{J_k}(t), \quad \mu(t) = \sum_k \mu_k 1_{J_k}(t).$$

In a natural way, this correspondence converts sums into integrals:

$$\sum_k \log_+(\frac{\lambda_k}{\mu_k}) = \int_0^\infty \log_+(\frac{\lambda(t)}{\mu(t)})dt,$$

and

$$\sum_k {\mu_k}^2 = \int_0^\infty \mu(t)^2 dt,$$

while owing to the asymptotic power law structure $a_k \sim (k/2)^m$,

$$2^{2m} \cdot \sum_k a_k \lambda_k \approx \int_0^\infty t^{2m} \lambda(t) dt,$$

in the sense that for $\lambda(t)$ "spread out", the relative difference of the two sides can be small.

Consider the problem

$$(Q_{D,\gamma}) \qquad \max_{\lambda(t)} \min_{\mu(t)} \int_0^\infty \log_+ (\frac{\lambda(t)}{\mu(t)})dt \quad \text{subject to} \quad \begin{cases} \int_0^\infty \mu(t)dt \leq D \\ \int_0^\infty t^{2m}\lambda(t)dt \leq \gamma^2 \end{cases} . \qquad (4.5)$$

The following technical result connects of the discrete-index optimization problem (4.4) and the continuous problem (4.5):

Theorem 4.1

$$R^+(D, W_{2,0}^m(\gamma)) \sim val(Q_{D,\gamma/2})(1 + o(1)), \quad D \to 0.$$

Its proof is almost identical to part of the proof of Theorem 3.1 – more specifically, the argument for (9.10) below – and so we omit it.

The rescaling arguments used earlier to study $(\mathcal{P}_{\epsilon,\gamma})$ can be used here to similar effect:

Theorem 4.2 *For all $D > 0$ and $\gamma > 0$,*

$$val(Q_{D,\gamma}) = val(Q_{1,1})(\gamma^2/D)^{\frac{1}{2m}}.$$

Now compare $(Q_{D,\gamma})$ to the earlier problem $(\mathcal{P}_{\epsilon,\gamma})$. Under the correspondence $\mu \Leftrightarrow \delta^2, \lambda \Leftrightarrow \rho^2, \epsilon^2 \Leftrightarrow D$ they are the same problem.

Theorem 4.3 *Under the calibration $\epsilon^2 = D$,*

$$val(Q_{D,\gamma}) = val(\mathcal{P}_{\epsilon,\gamma}) \quad \forall \epsilon, \gamma > 0. \qquad (4.6)$$

In words: the formal asymptotic expression associated with number of bits required to code the least favorable Gaussian process living on $W_{2,0}^m(\gamma)$ "in quadratic mean" agrees exactly with the formal asymptotic expression for the worst-case performance over $W_{2,0}^m(\gamma)$ of the subband coder.

5 Lower Bounds on H_ϵ

We now explain the seeming coincidence captured in Theorem 4.3. At this point, we have established that under the calibration $\epsilon^2 = D$, $D \to 0$,

$$\begin{aligned} H_\epsilon(W^m_{2,0}(\gamma)) &\leq \mathrm{val}(\mathcal{P}_{\epsilon,\gamma/2})(1+o(1)) \\ &= \mathrm{val}(Q_{D,\gamma/2})(1+o(1)) \\ &= R^+(D, W^m_{2,0}(\gamma))(1+o(1)). \end{aligned} \tag{5.1}$$

On the other hand, it is not clear why R^+ should be connected to H_ϵ. We will construct in Section 7 below a Gaussian process X^*_D asymptotically attaining R^+ and *a process* $\tilde{X}_D$ *living in* $W^m_{2,0}(\gamma)$ *with*

$$R^+(D, W^m_{2,0}(\gamma)) \sim R(D, X^*_D) \sim R(D, \tilde{X}_D), \quad D \to 0. \tag{5.2}$$

As $\tilde{X}_D$ lives in $W^m_{2,0}(\gamma)$, it is eligible for competition in the contest posed by $R^*(D, W^m_{2,0}(\gamma))$ so we must have $R(D, \tilde{X}_D) \leq R^*(D, W^m_{2,0}(\gamma))$, and it follows that

$$R^+(D, W^m_{2,0}(\gamma)) \leq R^*(D, W^m_{2,0}(\gamma))(1+o(1)), \quad D \to 0. \tag{5.3}$$

Combining now (1.12) and (5.3) we have

$$\begin{aligned} R^+(D, W^m_{2,0}(\gamma)) &\leq R^*(D, W^m_{2,0}(\gamma))(1+o(1)), \\ &\leq H_\epsilon(W^m_{2,0}(\gamma))(1+o(1)), \quad D = \epsilon^2 \to 0. \end{aligned} \tag{5.4}$$

Comparing (5.4) with (5.1) gives Theorem 1.1, and also the formula

$$H_\epsilon(W^m_{2,0}(\gamma)) \sim R^*(D, W^m_{2,0}(\gamma)) \sim R^+(D, W^m_{2,0}(\gamma)) \sim \mathrm{val}(Q_{D,\gamma/2}), \quad D = \epsilon^2 \to 0.$$

6 Solution of $(Q_{1,1})$

By the scaling law $\mathrm{val}(Q_{D,\gamma/2}) = \mathrm{val}(Q_{1,1})(\gamma/2\epsilon)^{1/2m}$, it is now of interest to know the solution of $(Q_{1,1})$; this will also be needed to verify (5.2). This has the form

$$\max_{\lambda(t)} \min_{\mu(t)} \int_0^\infty \log_+\Big(\frac{\lambda(t)}{\mu(t)}\Big)dt \quad \text{subject to} \quad \begin{cases} \int \mu(t)dt \leq 1 \\ \int t^{2m}\lambda(t)dt \leq 1 \end{cases}.$$

We will evaluate this by a "rearrangement and truncation" argument. From the water-filling discussion we know that for a given λ, the minimizing μ has the form

$$\mu(t) = \begin{cases} \theta & \lambda(t) \geq \theta \\ \lambda & \lambda(t) < \theta \end{cases}. \tag{6.1}$$

for some parameter $\theta = \theta(\lambda)$.

Let now $S_u = \{t : \lambda(t) \geq u\}$. For a given function λ, construct its decreasing rearrangement λ^* by $\lambda^*(t) = \min\{u : \mathrm{meas}(S_u) = t\}$. Obviously $\theta(\lambda^*) = \theta(\lambda)$ and so corresponding to this λ^* is $\mu^*(t) = \min(\theta, \lambda^*(t))$. Now

$$\begin{aligned} \int \mu(t)dt &= \int \mu^*(t)dt \\ \int \log\big(\frac{\lambda(t)}{\mu(t)}\big)dt &= \int \log\big(\frac{\lambda^*(t)}{\mu^*(t)}\big)\, dt \end{aligned}$$

while

$$\int t^{2m}\lambda^*(t)dt \leq \int t^{2m}\lambda(t)dt.$$

Hence in searching for solutions to $(Q_{1,1})$, we may restrict attention to monotone decreasing λ.

Suppose then that λ is decreasing, and that, for $\theta = \theta(\lambda)$ chosen so $\int \mu(t)dt = 1$, we have $\mathrm{supp}(\lambda) \neq S_\theta$. Then we may define $\lambda^*_\theta(t) = \lambda(t)1_{S_\theta}(t)$. Then there is a corresponding μ^*_θ and we have

$$\begin{aligned} \int \mu^*_\theta(t)dt &\leq \int \mu(t)dt, \\ \int t^{2m}\lambda^*_\theta(t)dt &\leq \int t^{2m}\lambda(t)dt, \\ \int \log\Big(\frac{\lambda^*_\theta(t)}{\mu^*_\theta(t)}\Big)dt &= \int \log\Big(\frac{\lambda(t)}{\mu(t)}\Big)dt. \end{aligned}$$

We conclude that we may restrict attention to functions λ which are monotone decreasing, supported in a finite interval, $[0,\tau]$ say, and which obey $\lambda(t) \geq 1/\tau$ throughout $[0,\tau]$. Let Λ_τ be the collection of all such λ for fixed τ.

Consider now the auxiliary optimization problem

$$\int_0^\tau \log(\lambda(t)\cdot\tau)dt \quad \text{subject to} \quad \left\{ \begin{array}{l} \int_0^\tau t^{2m}\lambda(t)dt \leq 1 \\ \lambda \in \Lambda_\tau \end{array} \right. . \tag{6.2}$$

We will find, among solutions of this problem over varying τ, a particular τ for which (6.2) gives a solution to $(Q_{1,1})$. Taking the first variation, we get that an optimum λ^*_τ for (6.2) must obey

$$\int_0^\tau \frac{v(t)}{\lambda^*(t)}dt \leq 0,$$

whenever $v(t)$ is a variation obeying

$$\int_0^\tau t^{2m}v(t)dt \leq 0, \qquad \lambda^*_\tau + \eta v \in \Lambda_\tau \quad \eta \to 0.$$

In other words, the solution λ^*_τ to (6.2) obeys:

$$\lambda^*_\tau(t) = \Big(\frac{t}{\tau}\Big)^{-2m}\frac{1}{\tau} \quad 1_{[0,\tau]}(t).$$

Now we note that $\int_0^\infty \lambda^*_\tau(t)t^{2m}dt = \tau^{-2m}$, so that λ^*_τ is feasible for problem $(Q_{1,1})$ only if $\tau \geq 1$. On the other hand, the associated $\mu^*(t)$ obeys $\int_0^\infty \mu^*(t)dt = \tau$, and so is feasible for the original $(Q_{1,1})$ only if $\tau \leq 1$. We conclude that λ^*_1 solves $(Q_{1,1})$. Now

$$\begin{aligned} \int_0^1 \log(\lambda^*_1(t))dt &= \int_0^1 \log\Big(t^{-2m}\Big)dt \\ &= 2m\int_0^1 \log(u^{-1})du \\ &= 2m\log_2 e. \end{aligned}$$

In short,

$$val(Q_{1,1}) = 2m\log_2 e.$$

7 The Least-Favorable Process

We now translate insights on the solution of $(Q_{1,1})$ back into information about the least-favorable process for R^+.

Lemma 7.1 *For scalars $\beta = \beta(D,\gamma)$, and $k_0 = k_0(D,\gamma)$ we have that the solution of (4.4) for the least-favorable Gaussian process X^+_D takes the form*

$$\lambda^+_k = \beta^2 \cdot a_k^{-1} \cdot 1_{\{0\leq k\leq k_0\}}. \tag{7.1}$$

Here we have

$$k_0 \sim (\gamma^2/D)^{1/2m}, \quad D \to 0, \tag{7.2}$$

and

$$\beta \sim (D/k_0)^{1/2}, \quad D \to 0. \tag{7.3}$$

Proof. Apply the 'rearrangement and truncation' reasoning of the previous section. The general form of (9.4) follows exactly as there. The constants k_0 and β must satisfy

$$\gamma^2 = \sum_k a_k \lambda_k^+ = (k_0 + 1) \cdot \beta^2,$$

as well as

$$D = \sum_{k=0}^{k_0} \beta^2 \cdot a_{k_0}^{-1} = \beta^2 \cdot (k_0 + 1)/a_{k_0}.$$

Solving these two equations for the two unknowns β and k_0 gives (7.2)-(7.3). ∎

In short, the least favorable process is a random trigonometric polynomial. We now show that this process can be slightly modified to be a process in $W_{2,0}^m(\gamma)$ with probability one.

Lemma 7.2 *For $\gamma > \tilde{\gamma} > 0$, let $X_{D;\tilde{\gamma}}^+$ denote the least favorable process for R^+ at distortion D and constraint $\tilde{\gamma}$. Then*

$$P\{X_{D;\tilde{\gamma}}^+ \in W_{2,0}^m(\gamma)\} \to 1, \quad D \to 0.$$

Proof. The event $\{X_{D;\tilde{\gamma}}^+ \in W_{2,0}^m(\gamma)\}$ is the same as

$$\sum_k a_k \tilde{\lambda}_k^+ Z_k^2 \le \gamma^2,$$

where the Z_k are i.i.d. $N(0,1)$ and $\tilde{\lambda}_k^+ = \lambda_k^+(D, \tilde{\gamma})$. Now, owing to the formula for λ_k^+,

$$\sum_k a_k \tilde{\lambda}_k^+ Z_k^2 = \tilde{\beta}^2 \sum_{k=0}^{\tilde{k}_0} Z_k^2,$$

where $\tilde{k}_0 = k_0(D, \tilde{\gamma})$ and $\tilde{\beta} = \beta(D, \tilde{\gamma})$. Now

$$E\ \tilde{\beta}^2 \sum_{k=0}^{\tilde{k}_0} Z_k^2 = \tilde{\beta}^2 \cdot (\tilde{k}_0 + 1) = \tilde{\gamma}^2.$$

Hence

$$\{\sum_k a_k \tilde{\lambda}_k^+ Z_k^2 > \gamma^2\} = \{\sum_{k=0}^{\tilde{k}_0} Z_k^2 > \frac{\gamma^2}{\tilde{\gamma}^2} \cdot (1 + \tilde{k}_0)\}.$$

We conclude that for an appropriate $\delta > 0$,

$$\{\tilde{X}_D \notin W_{2,0}^m(\gamma)\} \subset \{(1 + \tilde{k}_0)^{-1} \sum_{k=0}^{\tilde{k}_0} Z_k^2 > (1 + \delta)\}$$

Now

$$E \sum_{k=0}^{\tilde{k}_0} Z_k^2 = (1 + \tilde{k}_0).$$

The random variable $\sum_{k=0}^{\tilde{k}_0} Z_k^2$ has a χ^2 distribution on $\tilde{k}_0 + 1$ degrees of freedom. Hence we are asking for the event that a chi-squared random variable exceeds its mean by an amount proportional to its degrees of freedom. Owing to exponential bounds for tail probabilities of χ^2

random variables, this last event has a probability which tends to zero exponentially fast with increasing $\tilde{k}_0$. ∎

The reader should note that the previous lemma shows why we can replace the condition "$P(X \in W^m_{2,0}(\gamma)) = 1$"in R^* by the condition "the ellipsoid of concentration of X is inscribed in $W^m_{2,0}(\gamma)$" in R^+ and get similar results. In effect, there is a "Concentration of Measure" phenomenon which makes the two conditions almost the same.

We now use Lemma 2.3 to show that when the event $\{X^+ \notin W^m_{2,0}(\gamma)\}$ is very rare, conditioning on the complement of this event does not significantly change the Rate-Distortion.

Lemma 7.3 *For $\gamma > \eta > 0$, let $\overline{X}_{D;\gamma,\eta}$ be the process $X^+_{D;\gamma-\eta}$ conditioned to lie in $W^m_{2,0}(\gamma)$. Then*

$$\frac{R(D, \overline{X}_{D;\gamma,\eta})}{R(D, X^+_{D;\gamma-\eta})} \to 1, \quad D \to 0.$$

Proof. We will set things up to use Lemma 2.3.

To do so, we note that the random process $X = X^+_{D;\gamma-\eta}$ has a representation as $\sum_{k=0}^{k_0} \sqrt{\lambda_k^+} Z_k \phi_k$, where $(Z_k : 0 \le k \le k_0)$ is an i.i.d. $N(0,1)$ sequence. We can represent the vector (Z_k) as $S \cdot U$ – a vector U of euclidean length $\sqrt{1+k_0}$ times a random scale factor S, with S independent of U. Then $X = T(S \cdot U)$, where T is a linear operator.

Now $S^2 \cdot (1+k_0)$ has a χ^2 distribution on $1+k_0$ degrees of freedom. As mentioned in the previous Lemma, when the size of S^2 is larger than a constant depending on γ and η, ($(1+\delta)^2$, say) then X violates the condition $X \in W^m_{2,0}(\gamma)$.

Suppose now that we define a random variable S' on the same probability space as S, so that it is equal to S when $S < (1+\delta)$ and that when $S \ge (1+\delta)$ it is independently sampled from the conditional distribution of S given $\{S < (1+\delta)\}$. For the *same realization of U* that generated X, put $X' = T(S' \cdot U)$. Then either $X = X'$, which will happen with overwhelming probability, or we have proportionality $\frac{1}{S} X = \frac{1}{S'} X'$ on the rare event $\{S \ge (1+\delta)\}$. Now

$$\begin{aligned} E\|X - X'\|^2 &= E\|T((S-S') \cdot U)\|^2 \\ &= E(S-S')^2 \cdot E\|T(U)\|^2 \\ &\le E(S'-S)^2 \cdot \gamma^2, \end{aligned}$$

where we used independence of S and S' from U. Now note that

$$E(S'-S)^2 = P\{S \ge (1+\delta)\} \cdot E\{(S_1 - S_2)^2 | S_1 < (1+\delta), S_2 > (1+\delta)\}$$

where the S_i are i.i.d. with the same distribution as S. By the remark in Lemma 7.2, $P\{S \ge (1+\delta)\}$ decays exponentially fast with $D \to 0$. The expectation term is of size comparable to δ^2. We conclude that

$$E(S'-S)^2 = o(D).$$

Now note that X' has exactly the same distribution as $\overline{X}_{D;\gamma,\eta}$. It follows that on a common probability space, $\overline{X}_{D;\gamma,\eta}$ can be realized as the resultant of a randomized mapping applied to $X^+_{D;\gamma-\eta}$, for which $E\|X^+_{D;\gamma-\eta} - \overline{X}_{D;\gamma,\eta}\|^2 = o(D)$. We can apply Lemma 2.3, getting

$$R(\overline{X}_{D;\gamma,\eta}, D) \ge R(X^+_{D;\gamma-\eta}, (\sqrt{D} + o(\sqrt{D}))^2)$$

We can obviously also go in the other direction. Given X' we can create a randomized mapping defined on X' such that the resultant X has the same distribution as $X^+_{D;\gamma-\eta}$ and that on a common space $E\|X - X'\|^2 = o(D)$. The idea is as follows: toss a coin with probability $p = P\{S \ge (1+\delta)\}$ of heads. If it comes up heads, let X be obtained by independent sampling from the distribution of X^+, otherwise let $X = X'$.

Apply now the Lemma 2.3 to get that

$$R(D - o(D), X^+_{D;\gamma-\eta}) \ge R(D, \overline{X}_{D;\gamma,\eta});$$

then invoke

$$\frac{R(D+o(D), X^+_{D;\gamma-\eta})}{R(D-o(D), X^+_{D;\gamma-\eta})} \to 1, \quad \text{as } D \to 0,$$

which is easy to see by explicitly performing waterfilling on the known covariances of the random process $X^+_{D,\gamma-\eta}$. ∎

We are now in a position to conclude.

It follows from the Lemmata of this section that we can select $\eta(D) \to 0$ so that $X^*_D = X^+_{D;\gamma-\eta(D)}$ is the Gaussian process having the required properties for Theorem 1.2 and 1.3. The corresponding conditioned process $\tilde{X}_D = \overline{X}_{D;\gamma,\eta(D)}$ has the properties required for (5.2). The proof of Theorems 1.1, 1.2, and 1.3 is now complete.

8 Postscript

We briefly draw comparisons with some important earlier work.

8.1 V.I. Arnol'd

Our result on the ϵ-entropy of $W^m_{2,0}(\gamma)$ can be restated using the notation that Tikhomirov formulated in his commentary to Kolmogorov's work, as quoted in Section 1.3 above.

Put

$$\kappa(m,2,2,1) = lim_{\epsilon\to 0}\epsilon^{1/m} H_\epsilon(W^m_{2,0}(\gamma), L^2[0,2\pi)).$$

Then we have proven in this paper that

$$\kappa(m,2,2,1) = 2m\log_2(e)2^{-1/m}, \quad m = 1,2,3\ldots.$$

A precursor to our result was established in Kolmogorov-Tikhomirov (1959), who credit a key step to V.I. Arnol'd. Based on the behavior of $A_{0,m}$ and $A_{1,m}$ in a relation of the form (1.10), they were able to show that

$$\lim_{m\to\infty} \frac{\kappa(m,2,2,1)}{2m\log_2(e)} = 1$$

In a sense, with the insertion of an extra factor $2^{1/m}$ we have sharpened the Arnol'd-Kolmogorov-Tikhomirov asymptotic result $\kappa(m,2,2,1) \sim 2m\log_2(e)$, $m \to \infty$ into an exact evaluation, good for all m.

8.2 Tikhomirov

Tikhomirov, in a publication that has appeared only in Russian [22], attempted to find a partial generalization the Arnol'd-Kolmogorov-Tikhomirov result to the non L^2 case. He assumed that we have a set $\Theta^m_p(\gamma)$ of objects with coefficients $\sum_k a_k|\theta_k|^p \le \gamma^p$, defined by weights $a_0 = 1$, and $a_{2k} = a_{2k-1} = (1+k^{mp})$ for $k > 0$, and that we wish to code objects in this class with fidelity measured using an ℓ^p norm in a sequence space. Put

$$\kappa(m,p,p,1) = \lim_{\epsilon\to 0} \epsilon^{1/m} H_\epsilon(\Theta^m_p(\gamma), \ell^p).$$

Then he was able to show that

$$\lim_{m\to\infty} \frac{\kappa(m,p,p,1)}{2m\log_2(e)} = 1.$$

In principle the approach of this paper can be carried out in that setting, allowing to derive the exact value of $\kappa(m,p,p,1)$. This would require a study of the problem of coding ℓ^p balls and of minimax rate distortion theory for ℓ^p balls. Purely formal calculations suggest that the exact value of $\kappa(m,p,p,1) = 2m\log_2(e)2^{-1/m}$, and that the least favorable distribution has asymptotically independent coefficients θ_k with density $p_k(\theta) \sim c_p \cdot \exp\{-|\theta|^p/\lambda_k\}/\lambda_k$, for appropriate scaling factors λ_k obeying $\sum_k a_k\lambda_k^p \le \gamma^p$ and depending on ϵ. The appearance of this Generalized Gaussian density is due to its role as the maximum entropy distribution under

p-th power constraint, which arises naturally in determining the worst rate distortion curve for ℓ^p balls for ℓ^p distortion.

The generalization to a more general collection of non-L^2 cases is an interesting problem which we leave for future work.

9 Appendix: Proof of Theorem 3.1

Our proof is organized by defining a sequence of four optimization problems, the first of which is discrete, but visibly related to the continuous optimization problem, the final one of which gives the minimax codelength of the subband coder described in Section 3.3. We will establish a chain of asymptotic equivalences and inequalities among the problems which combine to show the desired asymptotic inequality of Theorem 3.1.

We will use material from sections 4-7, but there is no risk of circularity.

We begin with a discrete problem most naturally related to the continuum problem $(\mathcal{P}_{\epsilon,\gamma})$, namely

$$(\mathcal{P}^1_{\epsilon,\gamma}) \quad \max_{(\rho_b)} \min_{\delta} \sum_{b=0}^{\infty} n_b \log_+(\rho_b/\delta) \quad s.t. \quad \sum_b \min(\rho_b^2, \delta^2) n_b \le \epsilon^2 \\ \sum_b \rho_b^2 \alpha_b^- n_b \le \gamma^2/2^{2m}$$

Here we use discussion from Sections 6 and 7 and impose in advance that the subband distortion allocations have precisely the water-filling form $\delta_b^2 = \min(\rho_b^2, \delta^2)$.

Note that this discrete-index minimax problem has a similar form as the problem in Lemma 7.1 (after relabelling $\epsilon^2 \Leftrightarrow D$, etc.), and a solution can be characterized in the same way.

Lemma 9.1 *For scalars $\beta = \beta(\epsilon, \gamma)$, and $b_0 = b_0(\epsilon, \gamma)$, the solution of $(\mathcal{P}^1_{\epsilon,\gamma})$ takes the form*

$$(\rho_b^*)^2 = \beta/\alpha_b^- \cdot 1_{\{0 \le b \le b_0\}}. \tag{9.4}$$

Here b_0 obeys

$$k_{b_0} \sim (\gamma/\epsilon)^{1/m}, \quad \epsilon \to 0, \tag{9.5}$$

and

$$\beta \sim (\epsilon^2/k_{b_0}), \quad \epsilon \to 0. \tag{9.6}$$

The reasoning is the same as in the proof of Lemma 7.1, and we omit it.

We now argue that

$$val(\mathcal{P}^1_{\epsilon,\gamma}) \le val(\mathcal{P}_{\epsilon,\tilde{\gamma}}), \tag{9.7}$$

where $\tilde{\gamma}(\epsilon)$ is to be defined. Indeed, let $(\rho_b^*), \delta^*$ be an optimizing pair for $(\mathcal{P}^1_{\epsilon,\gamma})$. Define corresponding step functions $\rho_\epsilon^*(t)$ and $\delta_\epsilon^*(t)$ as in (3.9). Because of (3.11) and (3.12) these step functions are evidently almost feasible for $(\mathcal{P}_{\epsilon,\gamma})$ and in fact are feasible for $(\mathcal{P}_{\epsilon,\tilde{\gamma}})$, where

$$\tilde{\gamma}(\epsilon)^2 \equiv \int_0^\infty (\rho_\epsilon^*(t))^2 t^{2m} dt.$$

We also check that $\delta_\epsilon^*(t)$ solves the inner minimization over functions $\delta(t)$ which is contemplated in $(P_{\epsilon,\tilde{\gamma}})$. Indeed, by the water-filling discussion, the optimum such function will take the form $\delta(t) = \min(\delta, \rho_\epsilon^*(t))$ for a constant δ determined by

$$\int \min(\delta^2, (\rho_\epsilon^*)^2(t)) dt = \epsilon^2;$$

using (3.10) we can check that this is simply an alternate description of the step function δ_ϵ^*. This establishes (9.7).

Now we show that $\tilde{\gamma}(\epsilon) = \gamma + o(1)$ as $\epsilon \to 0$. Owing to the explicit form of ρ_b^*,

$$\begin{aligned}\int_0^\infty (\rho_\epsilon^*(t))^2 t^{2m} dt &= \sum_{b=0}^{b_0} \rho_b^2 \int_{I_b} t^{2m} dt \\ &= \sum_{b=0}^{b_0} (\beta/\alpha_b^-) \cdot (\bar{\alpha}_b \cdot n_b) \\ &= \beta \sum_{b=0}^{b_0} w_b n_b\end{aligned}$$

where $w_b = \bar{\alpha}_b/\alpha_b^-$, with

$$\bar{\alpha}_b = n_b^{-1} \int_{I_b} t^{2m} dt = \mathrm{Ave}\{t^2 m : t \in I_b\},$$

and where we used the fact that $n_b = |I_b|$. Define now

$$w_b^+ = \sup\{t^{2m} : t \in I_b\}/\alpha_b^-,$$

$$w_b^- = \inf\{t^{2m} : t \in I_b\}/\alpha_b^-,$$

so that we have the bracketing

$$w_b^- \le w_b \le w_b^+, \quad b = 0, 1, 2, \ldots.$$

Our construction of subbands, (3.1), gives

$$w_b^+/w_b^- \to 1, \quad b \to \infty.$$

Now using the fact that subband boundaries occur at even k, $\alpha_b^- = 1 + (k_b/2)^{2m}$, so

$$\sum n_b w_b^- / \sum n_b = \sum \left\{ n_b k_b^{2m} / (1 + (k_b/2)^{2m}) \right\} / \sum n_b \sim 2^{2m} \quad \epsilon \to 0.$$

It follows that the weighted average

$$\sum n_b w_b / \sum n_b \to 2^{2m}, \qquad \text{as } \epsilon \to 0,$$

simply because our subband construction has monotone subband widths: $n_0 \le n_1 \le ... \le n_b \le n_{b+1} \le ....$ Now in the omitted proof of Lemma 9.1 (compare the very analogous Lemma 7.1), β is defined so that

$$\beta \sum_{b=0}^{b_0} n_b = \gamma^2/2^{2m}.$$

Hence

$$\begin{aligned}\int_0^\infty (\rho_\epsilon^*(t))^2 t^{2m} dt &= \beta \sum_{b=0}^{b_0} w_b n_b \\ &= \left(\beta \sum_{b=0}^{b_0} n_b \right) \cdot \left(\sum_{b=0}^{b_0} w_b n_b / \sum_{b=0}^{b_0} n_b \right) \\ &\to (\gamma^2/2^{2m}) \cdot 2^{2m} = \gamma^2.\end{aligned}$$

In short, $\tilde{\gamma}(\epsilon) \to \gamma$ as $\epsilon \to 0$.

We now consider the other direction:

$$val(\mathcal{P}^1_{\epsilon,\gamma}) \ge val(\mathcal{P}_{\epsilon,\gamma}). \tag{9.8}$$

Let $\rho_\epsilon^*(t)$ and $\delta_\epsilon^*(t)$ furnish a solution of $(\mathcal{P}_{\epsilon,\gamma})$. This has the explicit form given in Section 6. Define sequences

$$\rho_b^* = \inf\{\rho_\epsilon^*(t) : t \in I_b\}, \quad \delta_b^* = \inf\{\delta_\epsilon^*(t) : t \in I_b\}, \quad b = 0, \ldots, B(\epsilon).$$

Then the step-functions $\rho_\epsilon^+(t)$ and $\delta_\epsilon^+(t)$ constructed using these sequences in (3.9) obey

$$\begin{aligned} \sum_b n_b \min((\rho_b^+)^2, \delta_b^2) &= \int \min((\rho_\epsilon^+)^2, (\delta_\epsilon^+)^2) dt \\ &\leq \int \min((\rho_\epsilon^*)^2, (\delta_\epsilon^*)^2) dt = \epsilon^2. \end{aligned}$$

Similarly,

$$\begin{aligned} \sum_b n_b \alpha_b^- \rho_b^2 &= \int \tau_{2m}(t)(\rho_\epsilon^+)^2(t) dt \\ &\leq \int t^{2m}(\rho_\epsilon^+)^2(t) dt \leq \int t^{2m}(\rho_\epsilon^*)^2(t) dt, \end{aligned}$$

where $\tau_{2m}(t)$ is the step function $\sum_b (t_b^*)^{2m} 1_{I_b}(t)$, with $t_b^* = \inf\{t : t \in I_b\}$. Then (9.8) follows.

Now we remark that, from the explicit form of the solution for $(\mathcal{P}_{\epsilon,\gamma})$ given in Section 7, we know that for any pair of functions $\tilde{\epsilon}(\epsilon) = \epsilon(1 + o(1))$ and $\tilde{\gamma}(\epsilon) = \gamma(1 + o(1))$, then

$$val(\mathcal{P}_{\epsilon,\gamma}) \sim val(\mathcal{P}_{\tilde{\epsilon},\tilde{\gamma}}), \quad \epsilon \to 0. \tag{9.9}$$

It follows that asymptotically negligible recalibrations of noise level and smoothness constraint cause asymptotically negligible changes in the value of the problem $(\mathcal{P}_{\epsilon,\gamma})$. We conclude from the above that

$$val(\mathcal{P}_{\epsilon,\gamma}^1) \sim val(\mathcal{P}_{\epsilon,\gamma}), \quad \epsilon \to 0. \tag{9.10}$$

We now consider our second optimization problem in our chain, $(\mathcal{P}_{\epsilon,\gamma}^2)$, which has the same constraints as $(\mathcal{P}_{\epsilon,\gamma}^1)$, but which has for objective the finite sum $\sum_{b=0}^{B(\epsilon)}$ rather than $\sum_{b=0}^{\infty}$. We now remark that Lemma 9.1 gave an explicit solution to $(\mathcal{P}_{\epsilon,\gamma}^1)$; our definition of $B(\epsilon)$ in Section 3.3 was chosen expressly so that $b_0(\epsilon) < B(\epsilon)$ for small ϵ, i.e. so the solution would be supported entirely in the subbands occuring before $B(\epsilon)$. As a result, the two problems $(\mathcal{P}_{\epsilon,\gamma}^1)$ and $(\mathcal{P}_{\epsilon,\gamma}^2)$ have the same optimal solution and the same value:

$$val(\mathcal{P}_{\epsilon,\gamma}^1) = val(\mathcal{P}_{\epsilon,\gamma}^2), \quad \epsilon \to 0. \tag{9.11}$$

Rather than continuing to the third problem in the chain, it is convenient to define the fourth and final one, and to work backwards in the chain. In the fourth problem our ultimate goal (3.6) is spelled out in detail:

$$(\mathcal{P}_{\epsilon,\gamma}^4) \qquad \max_{(\rho_b)} \min_{\nu} \sum_{b=0}^{B(\epsilon)} n_b h_2(r_b, \nu; n_b), \tag{9.12}$$

where the new optimization variables (r_b) and ν are quantized with quantum $q \equiv q_\epsilon = \epsilon^4$:

$$r_b = \lceil \rho_b / q \rceil \cdot q, \tag{9.13}$$

$$\nu = k \cdot q, \quad k \in \mathbf{Z}, \tag{9.14}$$

and the optimization variables (ρ_b) and ν obey the constraints

$$\sum_b \min(\nu^2, r_b^2) n_b \leq \epsilon^2 - \epsilon^{2.1}, \tag{9.15}$$

$$\sum_b \rho_b^2 \alpha_b^- n_b \leq \gamma^2 / 2^{2m}. \tag{9.16}$$

We will relate this to the last of our intermediate optimization problems:

$$(\mathcal{P}^3_{\epsilon,\gamma}) \qquad \max_{(\rho_b)} \min_{\nu} \sum_{b=0}^{B(\epsilon)} n_b \log_+(r_b/\nu), \tag{9.17}$$

subject to the same constraints (9.13)-(9.14) and (9.15)-(9.16) as $(\mathcal{P}^4_{\epsilon,\gamma})$. In this problem, we have replaced the dimension-normalized entropy h_2 in the objective of $(\mathcal{P}^4_{\epsilon,\gamma})$ by its asymptotically equivalent partner $\log_+(\cdot)$, yielding a problem more obviously connected to, say, $(\mathcal{P}^2_{\epsilon,\gamma})$.

To make a link between $(\mathcal{P}^4_{\epsilon,\gamma})$ and $(\mathcal{P}^3_{\epsilon,\gamma})$, we compare the objective functions in each and recall the crucial inequality from Section 2, (2.2), which gives

$$h_2(r_b, \nu; n_b) \le \log_+(r_b/\nu) + \frac{C_1 \log(n_b) + C_2}{n_b}.$$

Now, in the construction of our subbands (see Section 3.3), we may assume $B(\epsilon) \le \epsilon^{1/2m}$; it then results that

$$\sum_{b=0}^{B(\epsilon)} (C_1 \log(n_b) + C_2) = o(\epsilon^{1/m}),$$

so the two objectives always differ by a term which is $o(H_\epsilon)$, and so

$$val(\mathcal{P}^3_{\epsilon,\gamma}) = val(\mathcal{P}^4_{\epsilon,\gamma})(1 + o(1)), \quad \epsilon \to 0. \tag{9.18}$$

Our final step is to make a link between $(\mathcal{P}^3_{\epsilon,\gamma})$ and $(\mathcal{P}^2_{\epsilon,\gamma})$. In comparing the two problems, we note the quantization of ρ_b to r_b in these expressions, and of δ to ν. We will wish to show that the small perturbations involved in these quantizations have asymptotically negligible effects.

We need in two places below to control the *smallest* value that the water level δ can be, *uniformly* over all feasible candidates. So consider the problem

$$(\mathcal{D}_{\epsilon,\gamma}) \quad \inf \delta \quad : \quad \sum_b \min(\rho_b^2, \delta^2) n_b \ge \epsilon^2$$
$$\sum_b \rho_b^2 \alpha_b^- n_b \le \gamma^2$$

Lemma 9.2

$$val(\mathcal{D}_{\epsilon,\gamma}) > c \cdot \epsilon^2, \tag{9.19}$$

for all sufficiently small $\epsilon > 0$, where $c = c(\gamma)$.

In other words, it generally does not pay to use a water level δ which is substantially smaller than ϵ^2.

Proof. The problem can be converted into a linear programming problem in variables $d = \delta^2$ and $v_b = \rho_b^2$, of the form

$$\inf d \quad : \quad 0 \le v_b \le d,$$
$$\sum_b v_b n_b \ge \epsilon^2,$$
$$\sum_b v_b \alpha_b^- n_b \le \gamma^2.$$

The solution of this problem is at an extreme point of the feasible polytope, of the form $v_0 = v_1 = \ldots = v_k = d$, where k is chosen so that $kd \ge \epsilon^2$ and also $d \sum_b \alpha_b^- n_b \le \gamma^2$. Working with this solution, we obtain (9.19) ∎

We now return to study the differences that quantization of variables ρ_b and δ can cause between the feasible sets, and later in the objective functions, of the two problems $(\mathcal{P}^2_{\epsilon,\gamma})$ and $(\mathcal{P}_{\epsilon,\gamma}3)$.

As far as feasibility goes, note that when a sequence (ρ_b) considered in $(\mathcal{P}^2_{\epsilon,\gamma})$ is associated to a certain optimum water-level δ, the coder contemplated in $(\mathcal{P}^4_{\epsilon,\gamma})$, (9.14) has available to it a quantized water level ν which lies in the range $\delta \le \nu \le \delta + q$, with q the quantum. With this choice of ν, the coder has a distortion greater than that produced in $(\mathcal{P}^2_{\epsilon,\gamma})$. We can control this extra distortion:

$$\begin{aligned}\sum_b \min(r_b^2, \nu^2) n_b &- \sum_b \min(\rho_b^2, \delta^2) n_b \\ &\le \sum (r_b^2 - \rho_b^2) n_b + (\nu^2 - \delta^2) \cdot \sum_b n_b \\ &= \sum (r_b - \rho_b)(r_b + \rho_b) n_b + (\nu - \delta)(\nu + \delta) \cdot \sum_b n_b \\ &\le 3q \cdot \sum_b n_b r_b + 3\delta q \sum n_b\end{aligned}$$

where we used (9.19), which implies that $\delta > q_\epsilon = \epsilon^4$. Now from the subband definitions in Section 3.3,

$$\sum_{b=0}^{B(\epsilon)} n_b \le 2(\gamma^2/\epsilon^{2.1})^{1/2m} \tag{9.20}$$

so, as $m \ge 1$, $q \sum_{b=0}^{B(\epsilon)} n_b = O(\epsilon^4 \epsilon^{-1.05/m}) = o(\epsilon^2)$. We conclude that if

$$\sum_{b=0}^{B(\epsilon)} \min(\rho_b^2, \delta^2) n_b \le \epsilon^2$$

then a corresponding choice of of ν will obey

$$\sum_{b=0}^{B(\epsilon)} \min(r_b^2, \nu^2) n_b \le \epsilon^2 + o(\epsilon^2),$$

with $o(\cdot)$ uniform over admissible choices of (ρ_b). It follows that there is a function $\tilde{\epsilon}(\epsilon) = \epsilon(1 + o(1))$ which is such that to every $((\rho_b), \delta)$ feasible for $(\mathcal{P}^2_{\epsilon,\gamma})$ corresponds a $((r_b), \nu)$ feasible for $(\mathcal{P}^3_{\tilde{\epsilon},\gamma})$. In the other direction, we have from $\rho_b \le r_b$ and $\delta \le \nu$ that every $((r_b), \nu)$ feasible for $(\mathcal{P}^3_{\epsilon,\gamma})$ yields a $((\rho_b), \delta)$ feasible for $(\mathcal{P}^2_{\epsilon,\gamma})$, i.e. without any difference in ϵ.

We now compare the values of objective functions in the two problems. Evidently from $\rho_b \le r_b$ and $\delta \le \nu$, and the fact that $\log_+(\cdot)$ is Lipschitz,

$$\begin{aligned}\log_+(r_b/\nu) &\le \log_+(\rho_b/\nu) + (r_b - \rho_b)/\nu \\ &\le \log(\rho_b/\delta) + q/\delta.\end{aligned}$$

It follows that

$$\sum_b \log_+(r_b/\nu) n_b \le \sum_b \log_+(\rho_b/\delta) n_b + q/\delta \sum_b n_b. \tag{9.21}$$

Now by (9.19), $q/\delta = O(\epsilon^2)$, and combined with (9.20) we have

$$q/\delta \sum_b n_b = O(\epsilon^{2-1.05/m}) = o(\epsilon^{-1/m}).$$

We conclude that whenever $((\rho_b), \delta)$ are feasible for $(\mathcal{P}^2_{\epsilon,\gamma})$ then

$$\sum_b \log_+(r_b/\nu) \le \sum_b \log_+(\rho_b/\delta) + o(\epsilon^{-1/m})$$

with the $o(\cdot)$ term uniform over such feasible choices.

Combining the above remarks on feasibility and objective values,

$$val(\mathcal{P}^3_{\tilde{\epsilon},\gamma}) \le val(\mathcal{P}^2_{\epsilon,\gamma}) + o(\epsilon^{-1/m}).$$

Combining now all links in our chain of problems, and the remark (9.9),

$$val(\mathcal{P}^4_{\epsilon,\gamma}) \le val(\mathcal{P}_{\epsilon,\gamma})(1 + o(1)), \qquad \epsilon \to 0,$$

establishing Theorem 3.1. ∎

References

[1] Berger. T. (1971) *Rate Distortion Theory: A mathematical basis for data compression.* Prentice Hall: Englewood Cliffs, N.J.

[2] Birgé, L. (1983) Approximation dans les espaces métriques et théorie de l'estimation. (French) [Approximation in metric spaces and the theory of estimation] *Z. Wahrsch. Verw. Gebiete* **65** (1983), no. 2, 181–237.

[3] Carl, B. and Stephani, I. (1990) *Entropy, Compactness, and the Approximation of Operators.* Cambridge.

[4] Cover T. M. and Thomas, J.A. (1991) *Elements of Information Theory*, Wiley, NY 1991.

[5] Dembo, A. and Zeitouni, O. (1993) *Large deviations techniques and applications.* Jones & Bartlett: Boston.

[6] Dudley, R. M. (1967) The sizes of compact subsets of Hilbert spaces and continuity of Gaussian Processes, *J. Funct. Anal.* **1**, 290-330.

[7] Edmunds, D.E. and H. Triebel (1996) *Function spaces, entropy numbers, and differential operators.* Cambridge Univ. Press.

[8] Kolmogorov, A. N. (1956) On the Shannon theory of information transmission in the case of continuous signals, *Trans IRE*, **IT-2** 102-108.

[9] Kolmogorov, A.N. (1956) Some fundamental problems in the approximate and exact representation of functions of one or several variables. *Proc III Math Congress USSR Vol. 2* pp. 28-29. MCU Press, Moscow.

[10] Kolmogorov, A.N. and Tikhomirov, V.M. (1959) ϵ-entropy and ϵ-capacity. *Uspekhi Mat.* Nauk 14 3-86. (Engl Transl. *Amer. Math. Soc. Transl.* Ser 2 Vol 17 277-364.

[11] Le Cam, L. (1973) Convergence of Estimates under Dimensionality Restrictions. *Ann., Statist.* **1**, 38-53.

[12] Mitjagin, B. (1961) *Uspekhi Mat. Nauk* **16** 63-132. [In English: *Russian Math Surveys* **16**, 59-128.]

[13] Pinsker, M.S. (1980) Optimal Filtering of Square-Integrable Signals in Gaussian white noise. *Problems Information Transmission*, 120-133.

[14] Pisier, G. (1989) *The Volume of Convex Bodies and Banach Space Geometry*, Cambridge University Press.

[15] Rogers, C.A. (1963) Covering a sphere with spheres. *Mathematika*, **10**, 157-164.

[16] Sakrison, D.J. (1968) A geometric treatment of the problems of source encoding a Gaussian random variable, *IEEE Trans. IT*, **14**, 481-486.

[17] Sakrison, D.J. (1969) The rate distortion function of a class of sources. *Information and Control* , **15**, 165-195.

[18] Sakrison, D.J. (1970) The rate of a class of random processes. *IEEE Trans. IT*, **16**, 10-16.

[19] Sakrison, D.J. (1975) Worst sources and robust codes for difference distortion measures. *IEEE Trans. IT*, **21**, 301-309.

[20] Shannon, C.E. (1948) A mathematical theory of communication. *Bell System Technical Journal*, **27**, 379-423, 623-656.

[21] Shannon, C.E. (1959) Coding theorems for a discrete source with a fidelity constraint. *1959 IRE Conv. Record, Part 4*, pp. 142-163.

[22] Tikhomirov, V.M. (1976) *Certain questions of Approximation Theory.* Moscow, Moscow University Press.

[23] Tikhomirov, V.M. (1992) Commentary: ϵ-Entropy and ϵ-Capacity. in *A.N. Kolmogorov: Selected Works. III. Information Theory and Theory of Algorithms*. A.N. Shiryaev, Ed. Kluwer Academic Publishers.

[24] Wald, A. (1950) *Statistical Decision Functions*. Wiley N.Y.

[25] Wyner, A.D. (1967) Random packings and coverings of the unit N-sphere. *Bell System Technical Journal*, 2111-2118.