JOURNAL OF INTEGRAL EQUATIONS
AND APPLICATIONS
Volume 28, Number 1, Spring 2016

PROBABILISTIC REGULARIZATION OF FREDHOLM INTEGRAL EQUATIONS OF THE FIRST KIND

ENRICO DE MICHELI AND GIOVANNI ALBERTO VIANO

Communicated by Patricia Lamm

ABSTRACT. The main purpose of this paper is to focus on various issues inherent to the regularization theory of Fredholm integral equations of the first kind. Particular attention is devoted to the probabilistic approach to regularization, and a regularizing algorithm based on statistical methods is then proposed and tested on examples. The information theory approach is studied from two different viewpoints: the first approach is the standard one based on probability theory; the second one is formulated, in analogy with communication theory, in terms of the ε-capacity in the sense elaborated by Kolmogorov and his school. The classical problem of the resolving power in optics is then used to exemplify the relation between these two approaches.

1. Introduction. Inverse problems arise in many areas of science and technology and, loosely speaking, pertain to situations where one is interested in determining the *cause* of a phenomenon from the measurements of its *effects*. Typical examples of inverse problems occur, just to mention a few, in medical imaging (e.g., computer aided tomography), image processing (deconvolution), radio-astronomy, geophysics, model fitting, and so on. Inverse problems are often ill-posed in the sense that a unique solution might not exist, or the solution does not depend continuously on the input data: slight inaccuracies in the input data may lead to a solution very *far* from the true one [**29**]. The inverse problem needs regularization [**23**].

2010 AMS *Mathematics subject classification.* Primary 45B05, 47A52, 65C20, 94A05.

Keywords and phrases. Probabilistic regularization, Fredholm integral equations, inverse problems, communication theory, ε-capacity.

The second author passed away on November 10, 2014. This work was partially supported by the Consiglio Nazionale delle Ricerche project MD.P01.004.001.

Received by the editors on May 2, 2014, and in revised form on September 25, 2015.

DOI:10.1216/JIE-2016-28-1-31

The probabilistic regularization theory of Fredholm integral equations of the first kind represents one of the fundamental mathematical tools for analyzing ill-posed inverse problems (see **[68]** for examples of practical applications), and presents subtle aspects which deserve close attention. For instance, some features and similarities between the solutions provided by deterministic (variational) methods and the solutions brought by the probabilistic regularization theory are worthy of being discussed.

As we shall see, the probabilistic approach to regularization gives the chance of developing the regularization theory in various directions: on the one hand, it opens the door to using probabilistic information theory, which, eventually, produces statistical algorithms; on the other hand, it allows for exploiting concepts and language of the communication theory, which make it possible to elaborate an information theoretical approach based on topological instead of probabilistic methods. For instance, we will see that by taking advantage of the methods of the *topological information theory* a relevant expression for the classical problem of the resolving power in optics (a problem that traces back to Lord Rayleigh) can be obtained in terms of the maximum number of messages which can be conveyed back from the image to recover the object.

Consider the Fredholm integral equation of the first kind

$$(Af)(x) = \int_a^b K(x,y)\, f(y)\, \mathrm{d}y = g(x), \quad a \leqslant x \leqslant b, \tag{1.1}$$

whose kernel $K(x,y)$ is supposed to be Hermitian and square integrable:

$$K(x,y) = \overline{K(y,x)} \tag{1.2}$$

and

$$\int_a^b \int_a^b |K(x,y)|^2 \, \mathrm{d}x \, \mathrm{d}y < \infty.$$

Then, $A : L^2(a,b) \to L^2(a,b)$ is a self-adjoint compact operator. In (1.1), the kernel K and the *data function* g are supposed to be given, and the *inverse problem* is finding the unknown function f. For simplicity, we shall suppose hereafter that the kernel K, the function g and the unknown function f are real-valued functions; in addition, we assume that the interval $[a,b]$ is a closed and bounded subset of

the real line. We make an exception to these assumptions only in an example treated in subsections 3.2 and 3.3, where the function f and, accordingly, the function g, are assumed to be complex-valued. But the general results which we obtain hold true, substantially unmodified also in that case.

The Hilbert-Schmidt theorem guarantees that the integral operator A admits a set of eigenfunctions $\{\psi_k\}_{k=1}^{\infty}$ and, accordingly, a countably infinite set of eigenvalues $\{\lambda_k\}_{k=1}^{\infty}$ such that $\lim_{k\to\infty} \lambda_k = 0$. The eigenfunctions form an orthonormal basis for the orthogonal complement of the null space of the operator A. Further, we shall suppose hereafter that the eigenvalues are ordered as follows: $\lambda_1 \geqslant \lambda_2 \geqslant \cdots$. In view of the Hilbert-Schmidt theorem we can associate with the integral equation (1.1) the following eigenfunction expansion of the sought-after solution:

$$f(x) = \sum_{k=1}^{\infty} \left(\frac{g_k}{\lambda_k} \right) \psi_k(x), \tag{1.3}$$

where $g_k = (g, \psi_k)$ ($(\cdot, \cdot)$ denotes the scalar product in $L^2(a, b)$). In this case, the series (1.3) converges in the L^2-norm.

However, in view of the fact that there always exists an inherent noise in actual data, instead of (1.1), we have to deal with the following equation:

$$Af + n = \overline{g}, \qquad \overline{g} = g + n, \tag{1.4}$$

where an additive model of deterministic noise, represented by the function n belonging to the data space $L^2(a, b)$, is assumed. Therefore, instead of expansion (1.3), we have to consider the following formal series

$$\sum_{k=1}^{\infty} \left(\frac{\overline{g}_k}{\lambda_k} \right) \psi_k(x), \tag{1.5}$$

where $\overline{g}_k = (\overline{g}, \psi_k)$. The series (1.5) can even diverge if $\overline{g}$ does not belong to the range of the operator A. This is precisely the manifestation of the ill-posed character of the Fredholm integral equation of the first kind.

Several methods of regularization have been proposed (among the extensive literature on this topic see, e.g., **[8, 28, 62, 63, 64]** and the

references quoted therein; for an introductory presentation of general issues on inverse and ill-posed problems see also **[23, 49]**). All of these aim at modifying one of the elements of the triplet $\{A, X, Y\}$, where A is the integral operator defined by (1.1), while X and Y denote the solution and the data space, respectively.

In general, the initial concern is the choice of the functional ambient space where the problem can be conveniently formulated. The Banach space setting is frequently appropriate for many applied problems and have played an important role in the recent research in regularization theory for ill-posed problems. In fact, there are various practical applications where Hilbert space models can be unrealistic and the formulation of the (linear or nonlinear) model in the more general setting of Banach spaces (typically, L^p-spaces or spaces of continuous functions) can help to overcome the limitations of Hilbert spaces (see, e.g., **[58]** for a thorough discussion on this topic). However, in view of the lack of a *spectral theorem*, variational tools (Tikhonov-type methods) **[33, 57]** and iterative regularization methods (Newton-type methods) **[34, 37]** rather than regularizing procedures based on spectral theory are in use. In the case being considered here the assumptions on the operator A and the Gaussian distributions we adopted to describe the noise lead naturally to a (separable) Hilbert space setting (in our case $X \equiv Y \equiv L^2(a, b)$). Even more importantly, the spectral methods employed in our regularization analysis allow us to make simpler and particularly transparent the comparison among variational methods, probabilistic methods and the analysis based on the *topological information theory* (see Section 3).

In order to restore the continuity of the operator A^{-1}, it could be sufficient, in principle, to provide the data space Y with the *ad hoc* norm induced by A and X such that: $\|g\|_Y = \|A^{-1}f\|_X$. However, this type of regularizing procedure turns out to be usually inadequate since the induced norm $\|\cdot\|_Y$ on the data space does not allow a feasible description of the measurement errors (as do, e.g., the topologies of C^0 and L^2). In general, the topology of the data space should instead be dictated by physical considerations **[20]**.

As alternate methods of regularization, one could suitably modify the operator A itself in order to generate, and then solve, a new problem, "close" to the old one, but now well-posed. An example of this class of methods is the *quasi-reversibility* **[46]**, developed for

the regularization of ill-posed boundary value problems for partial differential equations (see, e.g., **[13, 38]**).

Most regularizing procedures are founded on a convenient modification of the solution space X. These methodologies lend themselves also to a probabilistic analysis, which will be our concern in Section 2. Among these, the most popular ones consist in admitting only those solutions that belong to a suitable subset of the solution space X. This restriction is realized by implementing *a priori* bounds, which can be written when some prior knowledge on the solution is available. Therefore, in addition to the inequality

$$\|Af - \overline{g}\|_Y \leqslant \varepsilon, \quad \varepsilon = \text{constant}, \tag{1.6}$$

which corresponds to a bound on the noise, one also considers an a priori bound on the solution of the following form:

$$\|Cf\|_Z \leqslant E, \quad E = \text{constant}, \tag{1.7}$$

where Z denotes the *constraint space* and, accordingly, C is called the constraint operator. The operator $C : X \to Z$, which is assumed to have bounded inverse, must be chosen to express some expected properties of the solution and has to be prescribed according to the physical character of the problem. For instance, a frequent choice is assuming C to be a differential operator so that bound (1.7) represents a smoothness requirement on the solution. It is obvious that the pair of positive numbers (ε, E) must be *permissible*, that is, such that the set of functions f which satisfies bounds (1.6) and (1.7) be non-empty. This problem has been extensively treated in the literature (see, for instance, **[52, 62]**) and we will not return to it. From bounds (1.6) and (1.7), we are led to define a regularized solution as the minimum of the following functional

$$\Phi(\alpha; f) = \|Af - \overline{g}\|_Y^2 + \alpha^2 \|Cf\|_Z^2, \quad \alpha \doteq \frac{\varepsilon}{E}. \tag{1.8}$$

This minimum represents (for a given value of α) the best trade-off between data-fit and regularity of the solution itself. For the sake of simplicity and convenience, from now on we take $Z \equiv L^2(a, b)$, so that we have: $X \equiv Y \equiv Z \equiv L^2(a, b)$. It is not significantly restrictive to assume that the operator C is such that C^*C and A^*A commute (A^* denoting the adjoint operator of A) since stability estimates associated with the minimization of functional (1.8) actually depend on the

spectrum of the operator S^*S, where $S = A(C^*C)^{-1/2}$ is a compact operator if A is compact **[51]** (see also **[8, 9]** for a discussion of this issue). Then, in such a case we can write $C^*Cf = \sum_{k=1}^{\infty} c_k^2 f_k \psi_k$ where $f_k = (f, \psi_k)$ are the Fourier coefficients of the unknown solution, and the c_k^2 are the eigenvalues of C^*C. The constraint operator C has compact inverse if and only if $\lim_{k\to\infty} c_k^2 = +\infty$. By expanding f and $\overline{g}$ on the basis $\{\psi_k\}_{k=1}^{\infty}$ (induced by A), it can be easily seen that the minimum of functional (1.8) is given by:

$$\widetilde{f}(x) = \left(A^*A + \alpha^2 C^*C\right)^{-1} A^*\,\overline{g}(x) = \sum_{k=1}^{\infty} \frac{\lambda_k\ \overline{g}_k}{\lambda_k^2 + \alpha^2 c_k^2}\ \psi_k(x). \tag{1.9}$$

One can then prove that $\widetilde{f}$ is indeed a regularized solution to problem (1.1), i.e., $\lim_{\varepsilon\to 0} \|f - \widetilde{f}\| = 0$ if and only if $\lim_{k\to\infty} c_k^2 = +\infty$ (where $\|\cdot\|$ denotes the L^2-norm) **[8, 9, 52]**. Assuming this latter condition holds and supposing that the sequence $\{c_k^2\}_{k=1}^{\infty}$ is monotone, another solution to problem (1.1) can be built by truncating expansion (1.9) at the largest integer k, which will be denoted by $\kappa(\varepsilon)$, such that $\lambda_k \geqslant \alpha c_k$ (where $c_k = +\sqrt{c_k^2}$). Then, neglecting $(\alpha\, c_k)$ in comparison with λ_k when $k \leqslant \kappa(\varepsilon)$, it can be proved that the approximation

$$\mathfrak{f}(x) = \sum_{k=1}^{\kappa(\varepsilon)} \frac{\overline{g}_k}{\lambda_k}\ \psi_k(x) \tag{1.10}$$

converges to the true solution f, as $\varepsilon \to 0$, in the sense of the L^2-norm **[8, 9, 52]**.

In several cases a much milder constraint can be conveniently used: for instance, we may assume $C \equiv \mathbb{I}$ (the identity operator on $Z \equiv L^2(a, b)$) and, therefore, bound (1.7) encodes an a priori knowledge on the *energy* of the solution. Accordingly, we have $c_k = 1$ for every k. In this case, the compactness condition for C^{-1} is not satisfied, but it can however be proved that the approximation obtained by truncating expansion (1.5) at the largest integer k, denoted by $k_w(\varepsilon)$, such that $\lambda_k \geqslant \alpha$:

$$\mathfrak{f}_w(x) = \sum_{k=1}^{k_w(\varepsilon)} \frac{\overline{g}_k}{\lambda_k}\psi_k(x), \tag{1.11}$$

still converges, in the weak sense, to the true solution f as $\varepsilon \to 0$, i.e., $\lim_{\varepsilon\to 0}(f - \mathfrak{f}_w, v) = 0$ for every $v \in L^2(a,b)$ **[8, 9, 52]**.

However, all these procedures are not free from defects. In fact, these methods perform as a low-pass filter, whose action smooths out the Fourier components $\overline{g}_k$ for high values of k (see (1.9), (1.10), (1.11)). But it is very easy to exhibit examples of functions whose Fourier components are small for low values of k while the significant contributions are brought by those components at "*intermediate*" values of k, which are smoothed out by the action of the filter. In this situation the standard variational regularization methods described above fail since the bulk of the unknown function f is not actually recovered. Moreover, at low values of k the contribution of the Fourier components $\overline{g}_k$ is kept without considering whether $\overline{g}_k$ actually brings exploitable information on the solution or only noise. Explicit examples of such functions have been given in **[16]**. These defects call for the introduction and implementation of probabilistic and statistical methods, which will be considered in this paper.

The paper is organized as follows. In Section 2, we expose the main steps of the probabilistic regularization method together with some results in information theory. Its relation with the deterministic variational method of regularization will be also analyzed. In Section 3, we introduce a topological information theory which is not set up within the probabilistic framework but makes use of concepts of communication theory. The *probabilistic* and *topological* approaches are then compared in the example of the classical problem of the resolving power in optics. In Section 4, statistical tools will be used to implement the ideas developed in Section 2 in order to establish an algorithm able to provide a probabilistically regularized solution to the ill-posed problem. In Section 5, some conclusions will be drawn. A short appendix introducing the notion of weak random variable in Hilbert spaces is finally given.

2. Probabilistic regularization method and information theory. In this section, we shall reconsider the problem (1.4) from a probabilistic point of view. We rewrite equation (1.4) in the following form:

$$A\xi + \zeta = \eta, \tag{2.1}$$

where ξ, ζ, and η, which correspond to f, n, and $\overline{g}$, respectively, are Gaussian *weak random variables* (w.r.v.) in the Hilbert space $L^2(a,b)$ [**5**]. The notion of weak random variable allows us to extend the problem to include stochastic processes ζ with covariance operator not of trace class, notably the Gaussian *white noise* (see the Appendix for a concise introduction to the concept of Hilbert space valued weak random variable). A Gaussian w.r.v. is uniquely defined by its mean element and its covariance operator; in the present case, we denote by $R_{\xi\xi}$, $R_{\zeta\zeta}$, and $R_{\eta\eta}$ the covariance operators on $L^2(a,b)$ of ξ, ζ and η, respectively (see (A.2) in the Appendix for the definition of second moment of a w.r.v.). Next, we make the following assumptions:

(I) ξ and ζ have zero mean (see Definition (A.1) of first moment of a w.r.v.);
(II) ξ and ζ are uncorrelated, i.e., $R_{\xi\zeta} = 0$ (see Definition (A.3));
(III) $R_{\zeta\zeta}^{-1}$ exists.

The first assumption is made only for the sake of simplicity and can be easily removed. The third assumption is the mathematical formulation of the fact that all the components of the data function are affected by noise. As it has been shown by Franklin [**24,** formula (3.11)], if the signal ξ and the noise ζ satisfy assumptions (I) and (II), then

$$R_{\eta\eta} = A R_{\xi\xi} A^* + R_{\zeta\zeta}, \tag{2.2}$$

and the cross-covariance operator is given by

$$R_{\xi\eta} = R_{\xi\xi} A^*. \tag{2.3}$$

We also assume that $R_{\zeta\zeta}$ depends on a parameter ε which tends to zero when the noise vanishes:

$$R_{\zeta\zeta} = \varepsilon^2 \, \mathfrak{R}_{\zeta\zeta},$$

where $\mathfrak{R}_{\zeta\zeta}$ is a given operator (e.g., $\mathfrak{R}_{\zeta\zeta} = \mathbb{I}$, the identity operator on $L^2(a,b)$, for the white noise).

Now, our problem can be formulated as follows.

Problem 2.1. Given a value $\overline{g}$ of the w.r.v. η find an estimate of the w.r.v. ξ.

A linear estimate of ξ will be any w.r.v. $\xi_L = L\eta$, where $L : Y \to X$ is an arbitrary linear continuous operator. Then, from a value $\overline{g}$ of η, one obtains the linear estimate $L\overline{g}$ of the w.r.v. ξ. Now, a measure of the reliability of the estimator L is given by **[8, 10]**

$$\delta^2(\varepsilon, v; L) = E\left\{|(\xi - L\eta, v)|^2\right\} \quad \text{for all } v \in X = L^2(a,b), \tag{2.4}$$

where $E\{\cdot\}$ denotes the expectation value. Then, we can state the following proposition.

Proposition 2.2. *If the covariance operator $R_{\zeta\zeta}$ has bounded inverse, then there exists a unique operator L_0 which minimizes $\delta^2(\varepsilon, v; L)$ for every $v \in X$, and it is given by*

$$L_0 = R_{\xi\eta} R_{\eta\eta}^{-1} = R_{\xi\xi} A^* \left(A R_{\xi\xi} A^* + R_{\zeta\zeta}\right)^{-1}.$$

Proof. See references **[8, 10]**. □

The w.r.v. $\widetilde{\xi}_{L_0} = L_0\eta$ is called the best linear estimator of ξ and, given a value $\overline{g}$ of η, the best linear estimate $\widetilde{f}_{L_0}$ for the value of ξ is given by

$$\widetilde{f}_{L_0} = R_{\xi\xi} A^* \left(A R_{\xi\xi} A^* + R_{\zeta\zeta}\right)^{-1} \overline{g} \qquad (A^* = A).$$

If the w.r.v. ξ has finite variance (i.e., if the covariance operator $R_{\xi\xi}$ has finite trace), then the global mean-square error for the estimator $\xi_L = L\eta$ may be defined as follows:

$$\delta^2(\varepsilon; L) = E\left\{\|\xi - L\eta\|^2\right\}. \tag{2.5}$$

When the operator L_0 which minimizes (2.4) exists, then it also minimizes the global mean-square error (2.5) if and only if $L_0\eta$ has finite variance **[8, 10]**. Now, exploiting the following identity (which holds when $R_{\xi\xi}^{-1}$ and $R_{\zeta\zeta}^{-1}$ exist):

$$\left(A^* R_{\zeta\zeta}^{-1} A + R_{\xi\xi}^{-1}\right) R_{\xi\xi} A^* = A^* R_{\zeta\zeta}^{-1} \left(A R_{\xi\xi} A^* + R_{\zeta\zeta}\right),$$

the operator L_0 can be rewritten as:

$$L_0 = \left(R_{\zeta\zeta}A^* R_{\zeta\zeta}^{-1} A + R_{\zeta\zeta}R_{\xi\xi}^{-1}\right)^{-1} \left(R_{\zeta\zeta}A^* R_{\zeta\zeta}^{-1}\right).$$

If the operators A^* and $R_{\zeta\zeta}$ commute (like, e.g., in the case of white noise), the solution $\widetilde{f}_{L_0}$ can be written as follows:

$$\widetilde{f}_{L_0} = \left(A^* A + R_{\zeta\zeta}R_{\xi\xi}^{-1}\right)^{-1} A^* \,\overline{g}. \tag{2.6}$$

The best linear estimate $\widetilde{f}_{L_0}$ is therefore modulated by the operator $R_{\zeta\zeta}R_{\xi\xi}^{-1}$, which represents the *noise-to-signal spectral density ratio*. Its role in the probabilistic solution (2.6) is very similar to the role played by the penalty operator C^*C in the variational solution $\widetilde{f}$ (see (1.9)). Only the spectral components of $\overline{g}$ with high signal-to-noise ratio contribute to the estimate $\widetilde{f}_{L_0}$ whereas those with small signal-to-noise ratio are smoothed out.

Several probabilistic regularization procedures are aimed at incorporating into the problem prior information on the solution f (see **[63]**). For instance, in the Backus-Gilbert approach **[4]** one seeks for the best linear estimator that optimally balance the bias and the variance of the inversion estimate. This task is usually implemented by using a penalty approach where the smoothness information on the function is inserted explicitly. For example, in the Tikhonov-Philips method, a penalty term on the L^2-norm of the second derivative of the solution $\|f''\|_{L^2}^2$ (assuming f to be twice differentiable with $f'' \in L^2$) is adopted in order to penalize the solutions with high local curvature. This type of method, however, suffers from the non-trivial problem of determining the optimal parameter (called the "*smoothing parameter*") from the data that controls the trade-off between *similarity between data and solution* and *roughness* of the solution. Among the proposed solutions to this problem, it is worth recalling the generalized cross-validation algorithm **[66, 67]** and the *L-curve* method **[30]**.

In this paper, we follow a different route to achieve a probabilistically regularized solution. In order to obtain more tractable formulae, it is convenient to expand ξ and ζ (which are w.r.v. in L^2) in a suitable orthonormal basis. The quite natural choice is to take as a basis the eigenfunctions $\{\psi_k\}_{k=1}^{\infty}$ of the operator A. If the solution space X and the data space Y are not equal, as instead we assume hereafter, then

the problem can be worked out similarly in terms of singular values and singular functions of the operator A [**8**], and all the results that we give in what follows can be easily reformulated.

An alternative basis expansion proposed for linear inverse problems is the so-called *wavelet-vaguelette decomposition* [**21**]. Evidently the eigenfunctions $\{\psi_k\}_{k=1}^{\infty}$ carry information only on the operator A, and it could happen they are not well adapted to the function f we want to recover. On the other hand, the orthonormal wavelet bases can provide an efficient representation for ample classes of functions. The main drawback of the wavelet-vaguelette decomposition is that the Riesz basis of *vaguelettes*, i.e., the image through the operator A of the wavelet basis, are no longer orthonormal. This generates undesired correlations among the coefficients of the wavelet expansion of the solution that make it difficult to select the coefficients which are statistically meaningful.

2.1. Spectral formulation of the inverse problem. Let us return to the probabilistic formulation (2.1) of the inverse problem (1.4), and assume, in addition to conditions (I), (II) and (III) stated earlier, that the Fourier components of the w.r.v. ζ (which describes the noise) are mutually uncorrelated. This assumption often seems to be reasonable in many practical situations (e.g., in the case of stationary white noise [**50,** page 380]). Now, we project equation (2.1) on the basis $\{\psi_k\}$ of the eigenfunctions of the operator A (see [**39**] for a similar approach based on the spectral decomposition of the operator A^*A in the case the operator A is not self-adjoint). We therefore have an infinite sequence of one-dimensional equations:

$$\lambda_k \xi_k + \zeta_k = \eta_k, \quad k = 1, 2, \dots, \tag{2.7}$$

where $\xi_k = (\xi, \psi_k)$, $\zeta_k = (\zeta, \psi_k)$ and $\eta_k = (\eta, \psi_k)$ are ordinary Gaussian random variables (see the Appendix).

Consider the Fourier components of the noisy data $\overline{g}$, that is, the set $\{\overline{g}_k\}_{k=1}^{\infty}$, where $\overline{g}$ represents a realization of the w.r.v. η, and the components $\overline{g}_k$ are understood as a realization of the ordinary random variables η_k. We should, in principle (and as also suggested by equation (2.6)), split this set into two subsets.

The first subset made up of those components $\overline{g}_k$ from which we can extract information for recovering the corresponding (unknown)

Fourier components f_k of f (the components f_k being understood as a realization of the ordinary random variables ξ_k), whereas the second subset is composed of those Fourier components $\overline{g}_k$ in which the noise prevails on the noiseless coefficients g_k.

Now, returning to the ordinary random variables η_k, which are the probabilistic counterparts of the coefficients $\overline{g}_k$, we aim at splitting the set $\{\eta_k\}_{k=1}^{\infty}$ into two subsets: the first subset is composed of those Fourier components of η from which we can extract information, while the second one is composed of those Fourier components η_k in which the noise ζ_k dominates (see (2.7)). To this end, we should evaluate the amount of information about the random variable ξ_k which is contained in the random variable $\eta_{k'}$. Now, we may introduce the *variances* of the ordinary Gaussian random variables ξ_k, ζ_k and η_k: $\sigma_k^2 = (R_{\xi\xi}\psi_k, \psi_k)$, $\varepsilon^2\nu_k^2 = (R_{\zeta\zeta}\psi_k, \psi_k)$ and $\lambda_k^2\sigma_k^2 + \varepsilon^2\nu_k^2 = (R_{\eta\eta}\psi_k, \psi_k)$ (see equation (2.2)). Next, we consider the mutual information between $\xi_{k'}$ and η_k **[14, 27, 54]**:

$$J(\xi_{k'}, \eta_k) = -\frac{1}{2}\ln\left(1 - r_{\xi\eta}^2(k', k)\right), \quad k, k' = 1, 2, \ldots,$$

where the squared correlation coefficient $r_{\xi\eta}^2(k', k)$ is given by:

$$r_{\xi\eta}^2(k', k) = \frac{|E\{\xi_{k'}, \eta_k^*\}|^2}{E\{|\xi_{k'}|^2\}\, E\{|\eta_k|^2\}} = \frac{|(R_{\xi\eta}\psi_{k'}, \psi_k)|^2}{\sigma_{k'}^2\,(\lambda_k^2\sigma_k^2 + \varepsilon^2\nu_k^2)}. \tag{2.8}$$

In view of equation (2.3), we have: $(R_{\xi\eta}\psi_{k'}, \psi_k) = \lambda_k\,\rho_{k',k}$, where $\rho_{k',k} = (R_{\xi\xi}\psi_{k'}, \psi_k)$. Then, the squared correlation coefficient can be written as follows:

$$\begin{aligned} r_{\xi\eta}^2(k', k) &= \frac{|\lambda_k|^2\,|\rho_{k',k}|^2}{\sigma_{k'}^2\,(\lambda_k^2\sigma_k^2 + \varepsilon^2\nu_k^2)} = \frac{|\rho_{k',k}|^2}{\sigma_{k'}^2\,\sigma_k^2}\left(1 + \frac{\varepsilon^2\nu_k^2}{\lambda_k^2\sigma_k^2}\right)^{-1} \\ &= \frac{|\rho_{k',k}|^2}{\sigma_{k'}^2\,\sigma_k^2}\,\frac{1}{1 + s_k^{-2}}, \end{aligned} \tag{2.9}$$

where $s_k^2 = (\lambda_k\sigma_k/\varepsilon\nu_k)^2$ represents the *signal-to-noise ratio* associated with the random variable η_k. Consequently, for $k \neq k'$, we have:

$$\begin{aligned} J(\xi_{k'}, \eta_k) &\sim \frac{1}{2}\ln\left(1 + \frac{|\rho_{k',k}|^2}{\sigma_{k'}^2\,\sigma_k^2}\,\frac{s_k^2}{1 + s_k^2}\right), \\ |r_{\xi,\eta}(k', k)| &< 1; \quad k, k' = 1, 2, \ldots; k \neq k'. \end{aligned} \tag{2.10}$$

Now, equation (2.10) informs us that the amount of information about $\xi_{k'}$ provided by η_k is significant if: (i) the two random variables $\xi_{k'}$ and ξ_k are correlated enough (i.e., $(|\rho_{k',k}/(\sigma_{k'}\sigma_k)|$ is large), and (ii) the signal-to-noise ratio s_k^2 associated with η_k (which represents the kth Fourier component of the noisy data) is sufficiently large (i.e., $|\lambda_k\sigma_k| > \varepsilon|\nu_k|$). Concerning the case $k = k'$, we have:

$$J(\xi_k, \eta_k) = \frac{1}{2}\ln\left(1 + \frac{\lambda_k^2\sigma_k^2}{\varepsilon^2\nu_k^2}\right) = \frac{1}{2}\ln\left(1 + s_k^2\right), \quad k = 1, 2, \dots . \tag{2.11}$$

From (2.11), it follows that if $|\lambda_k\sigma_k| < \varepsilon|\nu_k|$, then $J(\xi_k, \eta_k) < (1/2)\ln 2$ and the amount of information on ξ_k contained in η_k is small. Therefore, both equations (2.10) and (2.11) lead us to split the set of the Fourier components representing the noisy data into two subsets: $\{\eta_k\}_{k=1}^{\infty} = \{\eta_k\}_{k\in\mathcal{I}} \cup \{\eta_k\}_{k\in\mathcal{N}}$, where:

$$\mathcal{I} \doteq \{k : |\lambda_k\sigma_k| \geqslant \varepsilon|\nu_k|\}, \tag{2.12a}$$

$$\mathcal{N} \doteq \{k : |\lambda_k\sigma_k| < \varepsilon|\nu_k|\}. \tag{2.12b}$$

Now, we have $J(\xi_k, \eta_k) > J(\xi_{k'}, \eta_k)$ $(k' \neq k; |\rho_{k',k}/(\sigma_{k'}\sigma_k)| < 1)$. Moreover, even in the case of moderately high values of signal-to-noise ratio s_k^2 and rather correlated signal components $\xi_{k'}$ and ξ_k, the amount of information provided by η_k about ξ_k is much more than that about $\xi_{k'}$ with $k' \neq k$. For instance, for a signal-to-noise ratio as low as $s_k^2 \sim 3\,\mathrm{dB}$, the information provided by η_k on ξ_k is about twice that on $\xi_{k'}$ with $|\rho_{k',k}/(\sigma_{k'}\sigma_k)| = 0.8$, i.e.: $J(\xi_k, \eta_k) \sim 2\,J(\xi_{k'}, \eta_k)$, while in the case of poorly correlated signal components, e.g., $|\rho_{k',k}/(\sigma_{k'}\sigma_k)| = 0.5$, we have $J(\xi_k, \eta_k) \sim 6\,J(\xi_{k'}, \eta_k)$. This behavior becomes more pronounced as the signal-to-noise ratio s_k^2 increases. Therefore, since our main purpose is building upon the data set $\mathcal{I}$ (containing components with significant signal-to-noise ratio), we can adopt a (somewhat crude) model in which cross-correlations are neglected (also see **[39]**) and assume the ordinary random variables ξ_k and ζ_k to be mutually independent and normally distributed, in precisely the following way:

$$\xi_k \sim N(0, \sigma_k^2) \quad \text{and} \quad \zeta_k \sim N(0, \varepsilon^2\nu_k^2), \quad k = 1, 2, \dots .$$

Then, equations (2.7) allow us to write the conditional probability distribution of η_k given ξ_k:

$$\eta_k \mid \xi_k \sim N(\lambda_k\xi_k, \varepsilon^2\nu_k^2), \quad k = 1, 2, \dots .$$

Next, the Bayes' theorem states that the posterior distribution of ξ_k given (the occurrence of) the value $\overline{g}_k$ of the ordinary random variable η_k is normal:

$$\xi_k \mid \overline{g}_k \sim N\left(E\left\{\xi_k \mid \overline{g}_k\right\}, \mathrm{Var}\left\{\xi_k \mid \overline{g}_k\right\}\right), \tag{2.13}$$

with expectation value and variance given by:

$$E\left\{\xi_k \mid \overline{g}_k\right\} = \frac{1}{1+(1/s_k)^2}\frac{\overline{g}_k}{\lambda_k}, \tag{2.14a}$$

$$\mathrm{Var}\left\{\xi_k \mid \overline{g}_k\right\} = \frac{\sigma_k^2}{1+s_k^2}. \tag{2.14b}$$

We see from (2.14) that if $k \in \mathcal{I}$, i.e., if the signal-to-noise ratio s_k^2 is large, the distribution of ξ_k when the datum η_k takes on the value $\overline{g}_k$ is tightly peaked around the value $(\overline{g}_k/\lambda_k)$ with *small* variance $\mathrm{Var}\left\{\xi_k \mid \overline{g}_k\right\} \simeq (\varepsilon^2\nu_k^2/\lambda_k^2)$. Conversely, if $k \in \mathcal{N}$, that is, when the noise submerges the signal, we have $E\left\{\xi_k \mid \overline{g}_k\right\} \simeq 0$. Therefore, it is reasonable to consider the following approximate estimate $\langle\xi_k\rangle$ of the expected value of ξ_k (given the occurrence of the noisy datum $\overline{g}_k$):

$$\langle\xi_k\rangle = \begin{cases} \dfrac{\overline{g}_k}{\lambda_k} & \text{for} \quad k \in \mathcal{I}, \\ 0 & \text{for} \quad k \in \mathcal{N}. \end{cases}$$

Consequently, given the value $\overline{g}$ of the w.r.v. η, we are led to consider the following estimate of ξ:

$$\widehat{B}\,\overline{g} \doteq \sum_{k\in\mathcal{I}} \frac{\overline{g}_k}{\lambda_k}\,\psi_k. \tag{2.15}$$

Next, we can state the following proposition.

Proposition 2.3. *If the following conditions are satisfied*:

(a) $\lim_{k\to\infty} \nu_k^2/\sigma_k^2 = +\infty$ (*which implies the set* $\mathcal{I}$ *to be finite for any fixed positive value of* ε),
(b) *the operator* $R_{\xi\xi}$ *is of trace class,*

then the following limit holds true:

$$\lim_{\varepsilon\to 0} \delta^2(\varepsilon, \widehat{B}) = \lim_{\varepsilon\to 0} E\left\{\left\|\xi - \widehat{B}\eta\right\|^2\right\} = 0,$$

where

$$E\left\{\left\|\xi-\widehat{B}\eta\right\|^2\right\}=\operatorname{Tr}\left(R_{\xi\xi}-R_{\xi\xi}A^*\widehat{B}^*-\widehat{B}AR_{\xi\xi}+\widehat{B}R_{\eta\eta}\widehat{B}^*\right).$$

Proof. The proof follows with minor modifications the proof of Theorem 3.5 in [**16**]. □

We are now in the position of inspecting a connection between the probabilistic approximation (2.15) and the variational deterministic approximation presented in Section 1. We have seen that the variational approximation $\mathfrak{f}(x)$ given in (1.10) (which is the most frequently used) converges in the sense of the L^2-norm to the unknown function f and is obtained by truncating the expansion (1.5) at the largest integer k such that $\lambda_k \geqslant (\varepsilon/E)c_k$, where c_k^2 are the eigenvalues of the penalty operator C^*C, which encodes some a priori knowledge of the solution f. The sequence of the c_k's is also assumed to be such that $\lim_{k\to\infty} c_k^2 = +\infty$. For simplicity, but without loss of generality, let us put $E = 1$ so that the inequality above reads simply $\lambda_k \geqslant \varepsilon c_k$.

Now, let us return to the probabilistic approximation (2.15) and assume, as in Proposition 2.3, that $R_{\xi\xi}$ is an operator of trace class: $\operatorname{Tr} R_{\xi\xi} = \sum_{k=1}^{\infty} \sigma_k^2 \leqslant E < \infty$. Then $\lim_{k\to\infty} \sigma_k^2 = 0$. Further, let us suppose that $R_{\zeta\zeta}^{-1}$ is bounded. The set $\mathcal{I}$ is thus made up as follows: $\mathcal{I} = \{k : \lambda_k \geqslant \varepsilon\,|\nu_k/\sigma_k|\}$. Now, for the sake of simplicity, assume that the sequences $\{c_k^2\}_{k=1}^{\infty}$ and $\{(\nu_k/\sigma_k)^2\}_{k=1}^{\infty}$ are *monotone increasing* sequences. Therefore, the inequality which fixes the truncation index $\kappa(\varepsilon)$ in the variational approximation (i.e., $\lambda_k \geqslant \varepsilon c_k$) is strictly analogous to that which fixes the set $\mathcal{I}$ (i.e., $\lambda_k \geqslant \varepsilon|\nu_k/\sigma_k|$). The role played in the variational case by the numbers c_k^2 (i.e., the eigenvalues of the constraint operator) is played in the probabilistic case by the inverse of the k-component of the signal-to-noise ratio $(\nu_k/\sigma_k)^2$ (cf., equation (2.6) in which the operator $R_{\zeta\zeta}R_{\xi\xi}^{-1}$ acts as the penalty operator C^*C indeed). Note that, in order to be regularizing sequences, both $\{c_k^2\}_{k=1}^{\infty}$ and $\{(\nu_k/\sigma_k)^2\}_{k=1}^{\infty}$ are indeed required to be diverging (compactness condition for C^{-1} and $R_{\xi\xi}R_{\zeta\zeta}^{-1}$) for k going to infinity (see assumption (a) of Proposition 2.3).

It is, however, worth emphasizing the different rationale of the two approaches. In the (deterministic) variational paradigm the penalty operator C^*C encodes some a priori knowledge of the solution, which is

added to the initial problem. Conversely, in the probabilistic approach the operator $R_{\zeta\zeta}R_{\xi\xi}^{-1}$ describes information which is actually contained in the observed data set and that one could act to extract by means of a suitable statistical analysis of data.

The variational approximation $\mathfrak{f}(x)$ is built by summing up the contributions of all the components $\overline{g}_k$ from $k = 1$ up to $k = \kappa(\varepsilon)$, disregarding whether some $\overline{g}_k$ actually carry effective information on the solution f or, rather, are overwhelmed by the noise. Instead, the approximation $\widehat{B}\overline{g}$ in (2.15) features *frequency selectivity*, which we will aim to realize through statistical tools in Section 4. This characteristic of the probabilistic regularization avoids the drawback represented by the frequent over-smoothing of Tikhonov reconstructions using L^2-penalty terms (in this connection, see also **[58,** subsections 1.5 and 4.2] for a related discussion about regularization under sparsity constraint in the Banach space setting).

Remark 2.4. The Bayesian paradigm often is the natural framework whenever information on the statistical structure of unknown solution and noise is available (see **[36]** for an introduction to Bayesian inversion theory). The Bayesian approach has been adopted in **[24]** for linear ill-posed inverse problems and then developed in **[47]**. Examples of Bayesian formulation of nonlinear inverse problems are given in **[44, 45, 61]**. The analysis of inverse problems with non-Gaussian noise is less frequent in the literature. An excellent example of the Bayesian approach to an inverse problem with Poisson distribution of observations and non-Gaussian prior is given in **[11]**.

Recently, the focus is particularly on evaluating convergence rates under various conditions and assumptions **[32, 53]**. In **[39]**, contraction rates are obtained for mildly ill-posed problems in the case of Gaussian white noise when the operator A and the prior covariance operator are simultaneously diagonalizable. These results have then been generalized in **[1]** for non-white noise and operators not simultaneously diagonalizable. Convergence rates are given in the case of severely ill-posed problems in **[2, 40]** again in the simultaneously diagonalizable setting for the operators A and $R_{\xi\xi}$. Proposition 2.3 above guarantees that the approximation $\widehat{B}\overline{g}$ (given in (2.15)) is a probabilistic regularized solution to Problem 2.1 but does not provide any convergence rate of approximation (for this matter we refer the interested reader to

the above papers). However, our goal is different, being rather to reveal here the close connection between (deterministic) variational and Bayesian approach to ill-posed inverse problems (in this regard, see also **[2]**), and in the next Section 3 the connection with the approach based on topological information theory.

3. Topological information theory.

3.1. General theory. Standard information theory is formulated in the framework and uses language and tools of the probability theory, but the concept of *information* can be thought of as more basic and independent of probability **[41]**. For this purpose, Kolmogorov and his school introduced and developed an alternative approach to the quantitative definition of information. This theory is closely connected with concepts and language of communication theory, and the information flow is related to the concept of *direct* and *backward* channels of communication. Then, the main purpose of this theory is to give an estimate of the maximum number of *messages* which can be conveyed back from the data set to recover the unknown solution. Accordingly, solving the first kind Fredholm integral equation can be viewed as a backward channel communication problem.

Let us come back to the standard theory of regularization, and, in particular, to the a priori bound $\|Cf\|_Z \leqslant E$ (see (1.7)), where we take $C = \mathbb{I}$ (the identity on Z) as the constraint operator, $Z = L^2(a,b)$ as the constraint space and $E = 1$. We are thus led to consider the closed unit ball in the solution space (note that $X \equiv Z \equiv L^2(a,b)$), i.e., the set $\{f \in X : \|f\|_X \leqslant 1\}$. The (compact) operator A maps the closed unit ball in $L^2(a,b)$ (which, in view of the reflexivity of the Hilbert spaces, is therefore weakly compact) onto a compact ellipsoid $\mathcal{E}$ in the range of A, whose semi-axes are the positive eigenvalues λ_k of A. Let us note that $\mathcal{E} \subset Y = L^2(a,b)$, which is the data space. Further, we assume that $\overline{g} = g + n$ belongs to the range of A ($\overline{g} \in \mathrm{R}(A)$). Now, even if two data functions $\overline{g}_1$ and $\overline{g}_2$ belong to $\mathrm{R}(A)$ and their distance in Y is small, nevertheless the distance between $A^{-1}\overline{g}_1$ and $A^{-1}\overline{g}_2$ can be unlimitedly large.

We now recall some basic definitions from information theory **[41]**:

(a) in the theory of information the unit of a collection of information is the amount of information in one binary sign (that is designating whether it is 0 or 1).
(b) The entropy of a collection of possible communications undergoing transmission with a specified accuracy is defined as the number of binary signs necessary to transmit an arbitrary one of the communications with a given accuracy.
(c) The capacity of a transmitting apparatus is defined as the number of binary signs that it can transmit reliably.

Coming back to the compact ellipsoid $\mathcal{E}$, we recall some basic definitions which give a numerical estimate of its massiveness [**48**]:

(a′) a family $U_1, \dots, U_n$ of subsets of Y is an *ε-covering* of $\mathcal{E}$ if the diameter of each U_k does not exceed 2ε and if these sets cover $\mathcal{E}$: i.e., $\mathcal{E} \subset \cup_{k=1}^{n} U_k$.
(b′) Points $y_1, \dots, y_m$ of $\mathcal{E}$ are called *ε-distinguishable* if the distance between each of them exceeds ε.

Since $\mathcal{E}$ is compact, for each $\varepsilon > 0$ there exists a finite ε-covering of $\mathcal{E}$. Hence, $\mathcal{E}$ can contain only finitely many ε-distinguishable points. For a given $\varepsilon > 0$, the number of sets U_k in a covering family depends on the family, but the minimal value of n, $N_\varepsilon(\mathcal{E}) \doteq \min n$, is an invariant (in the sense that it does not depend on the type of covering family) of the set $\mathcal{E}$, which depends on ε. Its logarithm, that is, the function $H_\varepsilon(\mathcal{E}) \doteq \log N_\varepsilon(\mathcal{E})$ (in this section, $\log x$ will denote the logarithm of x to the base 2) is the *ε-entropy* of the set $\mathcal{E}$ and gives the number of distinct binary signals in $\mathcal{E}$ that can be reconstructed up to ε-accuracy. Analogously, the number m in the definition (b′) depends on the choice of the points, but its maximum $M_\varepsilon(\mathcal{E}) \doteq \max m$ is an invariant of the set $\mathcal{E}$ which depends on ε. It represents the maximum number of ε-distinguishable messages that can be used for recovering the unknown solution. Its logarithm, that is, the function $C_\varepsilon(\mathcal{E}) \doteq \log M_\varepsilon(\mathcal{E})$ is the *ε-capacity* of the set $\mathcal{E}$ and provides the number of binary signals that can be actually transmitted for recovering the solution [**48**].

The starting point of our analysis are the following inequalities [**17**]:

$$H_\varepsilon(\mathcal{E}) \leqslant C_\varepsilon(\mathcal{E}) \leqslant H_{\varepsilon/2}(\mathcal{E}). \tag{3.1}$$

Therefore, in order to obtain estimates of the ε-capacity $C_\varepsilon(\mathcal{E})$ one can look for a lower bound for $H_\varepsilon(\mathcal{E})$ and an upper bound for $H_{\varepsilon/2}(\mathcal{E})$. We

can state the following proposition.

Proposition 3.1. *The following bounds hold true*:

$$H_\varepsilon(\mathcal{E}) \geqslant \sum_{k=1}^{k_w(\varepsilon)} \log \frac{\lambda_k}{\varepsilon}, \tag{3.2}$$

where $k_w(\varepsilon)$ *is the largest integer* k *such that* $\lambda_k \geqslant \varepsilon$;

$$H_{\varepsilon/2}(\mathcal{E}) \leqslant k_w\left(\frac{\varepsilon}{4}\right)\left[\log\left(\frac{1}{\varepsilon}\right) + \log 6 + \frac{1}{2}\log k_w\left(\frac{\varepsilon}{4}\right)\right], \tag{3.3}$$

where $k_w(\varepsilon/4)$ *represents the number of terms in the sequence* $\{\lambda_k\}_{k=1}^{\infty}$ *that are larger than or equal to* $\varepsilon/4$.

Proof. In view of its relevance, we briefly sketch the derivation of bound (3.2). For the details and the proof of bound (3.3) see **[18, 19, 26, 55]**.

The operator A maps the unit ball in $L^2(a,b)$ onto a compact ellipsoid $\mathcal{E}$ whose semi-axes are the positive eigenvalues λ_k of the operator A. Since we have to deal with an expansion in terms of the eigenfunctions $\{\psi_k\}_{k=1}^{\infty}$ of the operator A, we are led to consider the intersection of the ellipsoid $\mathcal{E}$ with the finite k-dimensional subspace Y_k of Y, i.e., the ellipsoid $\mathcal{E}_k = \mathcal{E} \cap Y_k$. The semi-axes of $\mathcal{E}_k$ are the first k eigenvalues λ_k of the operator A. The volume of $\mathcal{E}_k$ is just $\prod_{n=1}^{k} \lambda_n$ times the volume Ω_k of the unit ball in Y_k.

Now, we want to estimate how many balls of radius ε are necessary for covering the ellipsoid $\mathcal{E}_k$. Since the volume of such a ball is $\varepsilon^k \Omega_k$, then we are forced to stop the integer k at a value such that the semi-axes of the ellipsoid $\mathcal{E}_k$ are not smaller than the radius ε of the ball. In view of the fact that the values λ_k (which coincide with the semi-axes of $\mathcal{E}$) are a non-increasing sequence, we must then take a finite subspace Y_k, whose dimension equals the largest integer k (denoted by k_w) such that $\lambda_k \geqslant \varepsilon$. Now, since the volume of an ε-ball in Y_k is given by $\varepsilon^k \Omega_k$, it follows that, in order to cover the ellipsoid $\mathcal{E}$ by ε-balls, we need at least $\prod_{k=1}^{k_w(\varepsilon)} (\lambda_k/\varepsilon)$ such balls. In conclusion, it follows that an estimate of the minimal number of sets in an ε-covering of $\mathcal{E}$ is given by $N_\varepsilon(\mathcal{E}) \geqslant \prod_{k=1}^{k_w(\varepsilon)} (\lambda_k/\varepsilon)$ and, accordingly, bound (3.2) follows. □

Notice that the definition of the upper summation limit $k_w(\varepsilon)$ given here in Proposition 3.1 coincides with the one we have given in Section 2 with reference to the approximation $\mathfrak{f}_w(x)$ (see (1.11)), which was proved to converge in the weak sense to the true solution.

Now, from inequalities (3.1), (3.2) and (3.3), one can derive lower and upper bounds for $C_\varepsilon(\mathcal{E})$ which depend on the way the eigenvalues λ_k decrease as k increases. In general (that is, even if we do not specify how the eigenvalues λ_k decrease as k increases), we can still evaluate the leading term of the upper bound of $C_\varepsilon(\mathcal{E})$, as ε tends to zero. Indeed, the leading term, as $\varepsilon \to 0$, of the right hand side of formula (3.3) is given by $k_w(\varepsilon/4)\log(1/\varepsilon)$, provided that $[k_w(\varepsilon/4)]^{1/2} = o(1/\varepsilon)$ as $\varepsilon \to 0$. Strictly speaking, we cannot derive an upper bound for the ε-capacity taking into account only the leading term of $H_{\varepsilon/2}(\mathcal{E})$. But, for ε sufficiently small, we can neglect all the terms which increase more slowly than the leading term and, accordingly, write:

$$C_\varepsilon(\mathcal{E}) \lesssim k_w\left(\frac{\varepsilon}{4}\right)\log\left(\frac{1}{\varepsilon}\right). \tag{3.4}$$

Next, in view of the leftmost inequality in (3.1) and of bound (3.2), we have:

$$M_\varepsilon(\mathcal{E}) = 2^{C_\varepsilon(\mathcal{E})} \geqslant 2^{[\sum_{k=1}^{k_w(\varepsilon)} \log \lambda_k/\varepsilon]}. \tag{3.5}$$

Let us note that the above estimate of the minimal ε-covering has been obtained considering a set of closed balls of radius ε whose centers are the images, under the operator A, of the functions f belonging to the subset of the solution space satisfying the a priori bound (1.7) (with $C \equiv \mathbb{I}$ and $E = 1$). In other words, inequality (3.5) gives us an estimate of the maximum number of distinguishable messages which can be extracted from the ellipsoid $\mathcal{E}$ and sent back to recover the functions belonging to the subset of the solution space constrained by the condition (1.7), where the operator C is the identity and the constant E is equal to 1.

But, on the other hand, we are obliged to use the data $\overline{g}$, which are perturbed by the noise and do not necessarily belong to the ellipsoid $\mathcal{E}$ even if they are supposed to belong to the range of operator A. Accordingly, if we keep the whole data set $\overline{g}$ (without any suitable truncation) we cannot pretend to reconstruct the functions belonging to the subset of the solution space constrained by a priori bounds.

Therefore, on the one hand, we are forced to extract the set of the distinguishable messages from the data set $\{\overline{g}_k\}_{k=1}^{\infty}$, but, on the other hand, we know that in general there can be points that do not belong to the ellipsoid $\mathcal{E}$.

At this point it is worth recalling that the minimal covering $N_\varepsilon(\mathcal{E})$ of a compact set $\mathcal{E}$ is an invariant of the set $\mathcal{E}$ that depends only upon ε and not on the family of subsets used to cover $\mathcal{E}$ [**48,** pages 150, 151]. Now, as we said above, we have derived an estimate of the minimal covering of the ellipsoid $\mathcal{E}_k$ by keeping the family generated by the closed balls whose centers are the images of the elements f belonging to the unit ball in $L^2(a,b)$ and ε is the bound, in the norm of L^2, according to condition (1.6).

Then, we keep a covering family given by the set of closed balls of radius ε whose centers are an approximation of $\overline{g}$ such that the minimal covering of the intersection of Y_k with a set $\mathcal{E}_*$ (for the moment not specified) is given by the largest integer k such that $\lambda_k \geqslant \varepsilon$. This latter inequality implies $\mathcal{E}_* = \mathcal{E}$. Accordingly, we are led to an approximation, denoted by $\overline{g}_{\mathcal{E}}$, of the data given by the set $\{\overline{g}_k\}_{k=1}^{k_w(\varepsilon)}$, whose elements $\overline{g}_k$ satisfy the following equality:

$$(\overline{g}_{\mathcal{E}}, \psi_k) = \begin{cases} \overline{g}_k & \text{for } 1 \leqslant k \leqslant k_w(\varepsilon), \\ 0 & \text{for } k > k_w(\varepsilon), \end{cases} \tag{3.6}$$

where $k_w(\varepsilon)$ is the largest integer such that $\lambda_k \geqslant \varepsilon$. In view of these considerations we can conclude that the maximum number of distinguishable messages which can be sent back to reconstruct the subset of the solution space constrained by formulae (1.6) and (1.7) is at least given by $2^{\sum_{k=1}^{k_w(\varepsilon)} \log_2(\lambda_k/\varepsilon)}$ (i.e., $M_\varepsilon(\mathcal{E}) \geqslant 2^{\sum_{k=1}^{k_w(\varepsilon)} \log_2(\lambda_k/\varepsilon)}$), which tends to $+\infty$ as ε tends to zero. Accordingly, the truncated approximation is given by

$$\mathfrak{f}_{\mathcal{E}}(x) \doteq \sum_{k=1}^{\infty} (\overline{g}_{\mathcal{E}}, \psi_k)\, \psi_k(x) = \sum_{k=1}^{k_w(\varepsilon)} (\overline{g}_{\mathcal{E}}, \psi_k)\, \psi_k(x),$$

which coincides with the deterministic approximation (1.11), the constant E being taken equal to 1. Finally, the distinguishable messages are the ones such that $\|\overline{g}_{\mathcal{E}}^i - \overline{g}_{\mathcal{E}}^j\| > \varepsilon$, $i \neq j$, where $\overline{g}_{\mathcal{E}}$ satisfies equality (3.6) and belongs to the ellipsoid $\mathcal{E}$.

Hille and Tamarkin [**31**] have systematically explored the relationship between the regularity properties of the kernel of the Fredholm first kind integral equations and the asymptotic behavior of the eigenvalues. They found that, as the regularity of the kernel increases, passing from the class C^0 to C^∞ and then to the class of the analytic functions, the eigenvalues decrease more and more rapidly as $k \to +\infty$. We can thus say that the relationship (3.4) gives a rough estimate of the maximum number of ε-distinguishable messages which can be conveyed back from the data for recovering the solution. Recalling the results of Hille and Tamarkin, we can say that, in general, *the maximum number of messages decreases as the smoothness of the kernel increases.* Let us, however, remark that this criterion must be taken with caution in view of the fact that the results of Hille and Tamarkin refer only to the asymptotic behavior of the eigenvalues.

3.2. A remarkable example. As an illustrative example, consider the inverse optical imaging problem. Following the simplest scheme for the formulation of this problem, we consider a one-dimensional object illuminated by coherent light, and we denote by $f(x)$ the complex amplitude distribution (for a more detailed specification of the model see [**65**]).

The Fourier transform of $f(x)$ is given by:

$$F(\omega) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} f(x)\, e^{-\mathrm{i}\omega x}\, \mathrm{d}x, \quad f \in L^2(-\infty, +\infty), \tag{3.7}$$

but the pupil-stop blocks all the contributions with $|\omega|$ larger than a positive constant Ω. As a consequence, from the image, we will not recover exactly $f(x)$ but its band-limited version

$$g(y) = \frac{1}{\sqrt{2\pi}} \int_{-\Omega}^{\Omega} F(\omega)\, e^{\mathrm{i}\omega y}\, \mathrm{d}\omega. \tag{3.8}$$

Plugging the expression (3.7) of $F(\omega)$ into (3.8) and assuming that the complex amplitude distribution of the object vanishes outside the interval $-X_0/2 \leqslant x \leqslant X_0/2$, we have:

$$g(y) = \frac{1}{2\pi} \int_{-\Omega}^{\Omega} \mathrm{d}\omega\, e^{\mathrm{i}\omega y} \int_{-X_0/2}^{X_0/2} f(x)\, e^{-\mathrm{i}\omega x}\, \mathrm{d}x \tag{3.9}$$

$$= \int_{-X_0/2}^{X_0/2} \frac{\sin \Omega(y-x)}{\pi(y-x)} f(x)\,\mathrm{d}x, \quad -\frac{X_0}{2} \leqslant y \leqslant \frac{X_0}{2}.$$

The image function $g(y)$ is an entire band-limited function, and the *Sampling theorem* guarantees that it can be reconstructed when its values are known at a set of sampling points chosen in arithmetic progression with difference π/Ω. In particular, the image $g(y)$ can be reconstructed in the interval $(-X_0/2, X_0/2)$ from the knowledge of the function on a set of S points, where $S \doteq X_0/(\pi/\Omega) = \Omega X_0/\pi$ is the Shannon number of the image **[65]**.

Equation (3.9) can be rewritten in operator form as

$$(Af)(y) = \int_{-X_0/2}^{X_0/2} \frac{\sin \Omega(x-y)}{\pi(x-y)} f(x)\,\mathrm{d}x = g(y), \quad -\frac{X_0}{2} \leqslant y \leqslant \frac{X_0}{2}.$$

Then, the problem of object restoration is equivalent to solving the Fredholm integral equation of the first kind $Af = g$, where A is a self-adjoint, non-negative and compact operator; g represents the noiseless data and f is the unknown object distribution. As usual, one has to take into account the perturbation of the noise and, assuming again an additive model of noise, one is finally led to the standard form (1.4) of the equation, i.e., $Af + n = g + n = \overline{g}$. The operator A acts as follows: $A : X \to Y$, where X and Y are the solution and the data space, respectively. We take, for simplicity, $X = Y = L^2(-X_0/2, X_0/2)$. Furthermore, the unique solution of the equation $Af = 0$ is $f = 0$, and then the integral operator A admits a complete set of eigenfunctions $\{\psi_k\}_{k=1}^{\infty}$ corresponding to a countable infinite set of real positive eigenvalues $\lambda_1 > \lambda_2 > \cdots > 0$, which satisfy the limit: $\lim_{k\to\infty} \lambda_k = 0$. The properties of this integral operator have been already studied by several authors (see, e.g., **[25, 43, 59]**).

Now, suppose that $S = \Omega X_0/\pi$ is *sufficiently* large. Then the eigenvalues λ_k manifest the following remarkable behavior: they form a decreasing sequence with a step-like behavior, that is, they are approximately equal to 1 for $k < \lfloor S \rfloor + 1$ (the symbol $\lfloor x \rfloor$ standing for the integral part of x), whereas, successively, for $k \geqslant \lfloor S \rfloor + 1$, they fall off to zero exponentially (see **[18,** Figure 2**]**).

Next, we keep all the assumptions made in subsection 3.1 and, in particular, we assume the a priori bound $\|f\|_X \leqslant E = 1$. Then, we

have:

$$\sum_{k=1}^{k_w(\varepsilon)} \log \frac{\lambda_k}{\varepsilon} = \sum_{k=1}^{\lfloor S \rfloor} \log \frac{\lambda_k}{\varepsilon} + \sum_{k=\lfloor S \rfloor+1}^{k_w(\varepsilon)} \log \frac{\lambda_k}{\varepsilon}. \tag{3.10}$$

Since, for $k \leqslant \lfloor S \rfloor$, we have $\lambda_k \simeq 1$, the contribution of the first sum on the right hand side of (3.10) is approximately $S \log 1/\varepsilon$. For $k \geqslant \lfloor S \rfloor + 1$, we have $\lambda_k \simeq \varepsilon$, so that the second sum on the right hand side of (3.10) is nearly null. Then, from inequality (3.2) of Proposition 3.1, we obtain the following lower bound for the ε-entropy:

$$H_\varepsilon(\mathcal{E}) \sim S \log \frac{1}{\varepsilon}. \tag{3.11}$$

We can thus conclude that the maximum number of ε-distinguishable messages which can be conveyed back from the image to recover the object should be at least (see the first inequality in formula (3.1)):

$$M_\varepsilon(\mathcal{E}) \gtrsim 2^{S \log(1/\varepsilon)} \underset{\varepsilon \to 0}{\longrightarrow} \infty.$$

Let us now assume that $k > \lfloor S \rfloor + 1$; then, we can make use of the formula which gives the asymptotic behavior of the eigenvalues λ_k for $k \to +\infty$: $\lambda_k = O\{(1/k) \exp[-2k \ln(2k/(e\Omega X_0))]\}$ **[8, 43]**. Then, we can evaluate $k_w(\varepsilon/4)$. A good estimate of this latter quantity gives: $k_w(\varepsilon/4) \sim [e\pi S + \log 1/\varepsilon]/2$. We can thus obtain from the right hand side of formula (3.3):

$$H_{\varepsilon/2}(\mathcal{E}) \underset{\varepsilon \to 0}{\sim} \frac{1}{2}\left[e\pi\, S + \log\left(\frac{1}{\varepsilon}\right)\right] \log\left(\frac{1}{\varepsilon}\right),$$

which, along with (3.11), yields through (3.4):

$$S \log\left(\frac{1}{\varepsilon}\right) < C_\varepsilon(\mathcal{E}) < \frac{e\pi}{2}\, S \log\left(\frac{1}{\varepsilon}\right) + \frac{1}{2} \log^2\left(\frac{1}{\varepsilon}\right).$$

We then have estimates of the upper and lower limits for the number of messages conveyed back from the image to reconstruct the object, and, accordingly, we can write in the limit of low noise:

$$2^{S \log(1/\varepsilon)} \lesssim M_\varepsilon(\mathcal{E}) \lesssim 2^{(1/2) \log^2(1/\varepsilon)}, \quad (\text{as } \varepsilon \to 0).$$

We thus obtain a quantitative answer to the classical problem of the resolving power in optics, which traces back to Lord Rayleigh **[56]**.

3.3. Comparison between topological and probabilistic information theory. In connection with the example illustrated above we can compare topological and probabilistic information theory. Let us start with formula (2.11), which we have seen in Section 2 yields, in the case of high *signal-to-noise ratio*, the largest amount of information provided by the random variable η_k about each variable $\xi_{k'}$ for all $k' = 1, 2, \ldots$. We can then evaluate approximately the total amount of information on the w.r.v. ξ which can be extracted from the w.r.v. η by summing the contributions brought by only the components with high signal-to-noise-ratio:

$$(3.12)\quad J(\xi,\eta) \simeq \sum_{k\in\mathcal{I}} J(\xi_k,\eta_k) = \frac{1}{2}\sum_{k\in\mathcal{I}} \ln\left(1+\frac{\lambda_k^2\sigma_k^2}{\varepsilon^2\nu_k^2}\right) \simeq \sum_{k\in\mathcal{I}} \ln\left|\frac{\lambda_k\sigma_k}{\varepsilon\nu_k}\right|,$$

the rightmost approximation made in view of the fact that, for $k \in \mathcal{I}$, we have $|\lambda_k\sigma_k| \geqslant \varepsilon|\nu_k|$ (see (2.12a)). It is appropriate to stress that the expression on the right hand side of (3.12) is only an approximate estimate of the total amount of information $J(\xi,\eta)$ since we have considered here the ordinary Gaussian random variables $\xi_{k'}$ and η_k ($k \neq k'$) as mutually independent, while they actually are not (see formulae (2.8) and (2.10) in Section 2). Now, we (obviously) assume a finite energy input signal:

$$\operatorname{Tr} R_{\xi\xi} = \sum_{k=1}^{\infty} \sigma_k^2 \leqslant E.$$

For simplicity and, without loss of generality, we set $E = 1$. We have: $\sigma_k^2 \leqslant 1$ and, hence, for $k \in \mathcal{I}$,

$$\lambda_k \geqslant \varepsilon|\nu_k|/\sigma_k \geqslant \varepsilon|\nu_k| \gtrsim \varepsilon,$$

follows, where we have put $|\nu_k| \simeq 1$ ($\nu_k = 1$ in the case of white noise). We have, therefore, obtained the inequality $\lambda_k \gtrsim \varepsilon$, which is indeed the inequality that determines the approximation $\mathfrak{f}_w$ (see formula (1.11) in the introduction) in the specific case $E = 1$ (i.e., $\mathfrak{f}_w = \sum_{k=1}^{k_w(\varepsilon)} (1/\lambda_k)\overline{g}_k\psi_k$, where $k_w(\varepsilon)$ is the largest integer k such that $\lambda_k \geqslant \varepsilon$). Therefore, we write:

$$J(\xi,\eta) \simeq \sum_{k\in\mathcal{I}} J(\xi_k,\eta_k) \sim \sum_{k\in\mathcal{I}} \ln\frac{\lambda_k}{\varepsilon},$$

whose structure is actually very similar to the lower bound (3.2) for the ε-entropy $H_\varepsilon(\mathcal{E})$ we have given in Proposition 3.1. In the inverse optical imaging problem illustrated above, the eigenvalues λ_k are nearly equal to 1 for $k < \lfloor S \rfloor + 1$, and, successively, for $k \geq \lfloor S \rfloor + 1$, fall off to zero exponentially. Hence, it follows that:

$$J(\xi, \eta) \simeq \sum_{k=1}^{\lfloor S \rfloor} \ln \frac{\lambda_k}{\varepsilon} + \sum_{k=\lfloor S \rfloor + 1}^{k_w(\varepsilon)} \ln \frac{\lambda_k}{\varepsilon} \simeq S \ln \frac{1}{\varepsilon}.$$

Therefore, we can establish a connection between topological and probabilistic information theory, relating the *maximum number* $M_\varepsilon(\mathcal{E})$ *of messages conveyed back from the image to reconstruct the object* with *the total information on* ξ *which can be extracted from* η (and disregarding the difference between logarithms with different bases, which is immaterial in this context):

$$M_\varepsilon(\mathcal{E}) \sim 2^{S \log(1/\varepsilon)} \simeq 2^{S \ln(1/\varepsilon} \simeq 2^{\{\sum_{k \in \mathcal{I}} J(\xi_k, \eta_k)\}} \simeq 2^{J(\xi, \eta)}.$$

4. An algorithm based on statistical methods. It is well known that a priori bounds of the form (1.6) and (1.7) (occurring in the deterministic variational methods) are often complicated to have. In this connection, there exists a huge literature (see, for instance, **[7, 23, 66]** and the references quoted therein) where it is shown how to obtain approximate solutions with these methods even in the unfavorable cases where only one of these bounds, e.g., bound (1.6) (which seems to be in practice more accessible), is known. But, what undoubtedly results is that, in the practical cases, the necessity of some prior knowledge on the solution remains.

Within the probabilistic scheme, however, the a priori knowledge of the covariance operators $R_{\xi\xi}$ and $R_{\zeta\zeta}$ is completely unrealistic, except for some reasonable and rather general features such as, e.g., supposing the covariance operator $R_{\xi\xi}$ of the unknown solution to be an operator of trace class. Then it often turns up that the sole possibility we have to obtain information on the unknown solution is analyzing suitably the noisy data $\{\overline{g}_k\}_{k=1}^{\infty}$, so as to be able, in principle, to pick up those data from which reliable information can be extracted, and discard those where the noise prevails. That this could be a profitable strategy is supported by the analysis of Section 2 where the *noise-to-signal spectral density ratio* was shown to act as the penalty operator does

in the deterministic variational scheme (see (2.6) and the discussion right after).

The same indication is given also by formulae (2.10) and (2.11), which suggest recognizing those data components η_k with large value of s_k^2. The model elaborated in Section 2, which is based on the probabilistic information theory, aims precisely at splitting the data set into two disjoint subsets $\mathcal{I}$ and $\mathcal{N}$: one composed of those data from which information can be extracted (see (2.12a)), the other one composed of those data in which the information is submerged by the noise (see (2.12b)).

Since, in practice, *only* the noisy data set $\{\eta_k\}$ is available, a possible route for selecting those components from which effective information on ξ_k can be reliably extracted is to evaluate the amount of information about the random variable $\eta_{k'}$ provided by the random variable η_k, that is,

$$J(\eta_{k'}, \eta_k) = -\frac{1}{2}\ln(1 - r_{\eta\eta}^2(k', k)),$$

where

$$r_{\eta\eta}^2(k', k) = \frac{|E\{\eta_{k'}, \eta_k^*\}|^2}{E\{|\eta_{k'}|^2\}\, E\{|\eta_k|^2\}}. \tag{4.1}$$

Since $R_{\eta\eta} = AR_{\xi\xi}A^* + R_{\zeta\zeta}$ (see (2.2)) and recalling that the orthonormal basis $\{\psi_k\}_{k=1}^{\infty}$ diagonalizes the operator A, we have (see also (2.9)):

$$\begin{aligned} E\{\eta_{k'}, \eta_k^*\} = (R_{\eta\eta}\psi_{k'}, \psi_k) &= (AR_{\xi\xi}A^*\psi_{k'}, \psi_k) + (R_{\zeta\zeta}\psi_{k'}, \psi_k) \\ &= \lambda_{k'}\lambda_k\, \rho_{k',k} + \varepsilon^2 \nu_{k',k}, \end{aligned}$$

where $\rho_{k',k} = (R_{\xi\xi}\psi_{k'}, \psi_k)$ and $\nu_{k',k} = (R_{\zeta\zeta}\psi_{k'}, \psi_k)$. Similarly, we have: $E\{|\eta_k|^2\} = \lambda_k^2\sigma_k^2 + \varepsilon^2\nu_k^2$. Therefore, formula (4.1) yields for $k \neq k'$:

$$\begin{aligned} r_{\eta\eta}^2(k', k) &= \frac{\left|\lambda_{k'}\lambda_k\, \rho_{k',k} + \varepsilon^2\nu_{k',k}\right|^2}{(\lambda_{k'}^2\sigma_{k'}^2 + \varepsilon^2\nu_{k'}^2)(\lambda_k^2\sigma_k^2 + \varepsilon^2\nu_k^2)} \\ &\simeq \frac{s_{k'}^2}{1 + s_{k'}^2}\, \frac{s_k^2}{1 + s_k^2}\, \frac{\rho_{k',k}^2}{\sigma_{k'}^2\sigma_k^2} \\ &\quad + 2\frac{1}{s_{k'} + s_{k'}^{-1}}\, \frac{1}{s_k + s_k^{-1}}\, \frac{\rho_{k',k}}{\sigma_{k'}\sigma_k}\, \frac{\nu_{k',k}}{\nu_{k'}\nu_k}, \quad k' \neq k, \end{aligned} \tag{4.2}$$

where the $O(\varepsilon^4)$ term has been neglected. Now, the second term on the right hand side of equation (4.2) is negligible both in the cases of high and low *signal-to-noise ratio* s_k (or $s_{k'}$), and is null when the Fourier components of the noise ζ are mutually uncorrelated (e.g., in the case of white noise). Hence, large values of $r^2_{\eta\eta}(k',k)$ occur when the corresponding Fourier components $\xi_{k'}$ and ξ_k of the unknown solution are significantly correlated and, moreover, both feature a high signal-to-noise ratio. Therefore, evaluating $r^2_{\eta\eta}(k',k)$ from the noisy data might be a useful practical tool for actually constructing the sets $\mathcal{I}$ and $\mathcal{N}$ (see (2.12)).

A difficulty that is frequently encountered in practice is that, for each k, only a small number of realizations $\overline{g}_k$ of the random variable η_k is usually available. If, for each k, multiple independent realizations $\overline{g}_k$ of the random variable η_k would be available, then one could estimate the ensemble averages required by formula (4.1). But this is usually impossible from a practical point of view. In this case, we can introduce the working assumption that the process $\{\eta_k\}$ be stationary in the wide sense **[22]**, which amounts to assuming that $r^2_{\eta\eta}(k',k) = r^2_{\eta\eta}(|k'-k|)$, and then compute estimates of the autocorrelation function by means of the ergodic hypothesis connecting ensemble and *time* (i.e., the index k in our case) averages. Evidently, this assumption introduces in the process $\{\eta_k\}$ an invariance for k-translation that will have to be removed in the actual realization of the algorithm.

For the practical numerical implementation of the algorithm we use the following estimator of the sample autocorrelation function, which is largely adopted by statisticians **[35]**:

(4.3)
$$\delta_{\overline{g}}(n) = \frac{\displaystyle\sum_{k=1}^{N-n} \left(\overline{g}_k - \langle \overline{g}_k \rangle\right)\left(\overline{g}_{k+n} - \langle \overline{g}_{k+n} \rangle\right)}{\left[\displaystyle\sum_{k=1}^{N-n} \left(\overline{g}_k - \langle \overline{g}_k \rangle\right)^2 \cdot \sum_{k=1}^{N-n} \left(\overline{g}_{k+n} - \langle \overline{g}_{k+n} \rangle\right)^2\right]^{1/2}}, \quad n \in [0, N-1],$$

where:

$$\langle \overline{g}_k \rangle = \frac{1}{N-n} \sum_{k=1}^{N-n} \overline{g}_k, \qquad \langle \overline{g}_{k+n} \rangle = \frac{1}{N-n} \sum_{k=1}^{N-n} \overline{g}_{k+n}.$$

Obviously, the *correlogram* (4.3) does not directly yield the Fourier

coefficients $\overline{g}_k$ that contribute to the approximate regularized solution given by formula (2.15), but provides us with estimates of the correlation coefficients between Fourier components separated n *frequencies* apart ($n = 0, 1, \ldots, N-1$). Since our goal is to identify and then reject purely random Fourier components, the matter is to implement a test of randomness for the sequence $\{\overline{g}_k\}_{k=1}^{N}$, which allows us to decide when the autocorrelation $\delta_{\bar{g}}$ at a certain lag n is (statistically) null. It can be shown that, for a set of N independent and identically distributed random variables, the variance of $\delta_{\bar{g}}(n)$ is $\mathrm{var}\,(\delta_{\bar{g}}(n)) = 1/N$, and $\delta_{\bar{g}}(n)$ is asymptotically normally distributed under the assumption of weak stationarity [**3**]. Therefore, as a first approximation, the 95 percent limit for the one-sided randomness test (notice that we test the modulus of $\delta_{\bar{g}}(n)$) can be placed at approximately $1.65/\sqrt{N}$. However, it can be shown [**3, 6**] that estimates of successive values of $\delta_{\bar{g}}(n)$ can be highly correlated, so that a certain value of $\delta_{\bar{g}}(n)$ might be large simply because the previous ones, at lower lags, are large. Obviously, this phenomenon makes it even more difficult to state at which lags the autocorrelation is significantly different from zero. To account for this interdependence of the autocorrelations at different lags, use can be made of the following Bartlett's adjusted variance of the estimates of the autocorrelation function [**6**]:

$$\mathrm{var}\,(\delta_{\bar{g}}(n)) \simeq \frac{1}{N-n}\left[1 + 2\sum_{j=1}^{n_0}\delta_{\bar{g}}^2(j)\right], \quad \text{with } n_0 < n, \tag{4.4}$$

whose square root $\varepsilon_\delta(n)$ is called the *large-lag standard error* of $\delta_{\bar{g}}(n)$. From (4.4), we see that the error of the autocorrelation coefficient at any given lag depends on the sample size N and on the estimated autocorrelation coefficients at shorter lags so that the confidence band for testing the null hypothesis appears narrower at smaller lags and gets wider at higher lags (see, for instance, Figures 1 (b) and 2 (b)). Now, the set $\mathcal{L}$ of the lags at which the autocorrelation function is significantly non-null can be defined as [**3**]:

$$\mathcal{L} = \{n : |\delta_{\bar{g}}(n)| \geqslant 1.65\,\varepsilon_\delta(n)\}, \tag{4.5}$$

and its cardinality will be denoted $N_{\mathcal{L}}$. Any integer $n_i \in \mathcal{L}$ ($i = 1, \ldots, N_{\mathcal{L}}$) just indicates a significant correlation between at least two Fourier coefficients located n_i *frequencies* apart. This means that, in principle, any couple $(\overline{g}_{k_i}, \overline{g}_{k_i+n_i})$ with $1 \leqslant k_i \leqslant N - n_i$ could have

generated such a strong correlation at the lag n_i. Thus, from the set $\mathcal{L}$, we can construct $N_{\mathcal{L}}$ families $\mathcal{F}_i$, defined as:

$$\mathcal{F}_i = \left\{(\overline{g}_{k_i}, \overline{g}_{k_i+n_i})\right\}_{k_i=1}^{N-n_i}, \quad n_i \in \mathcal{L},\ i = 1, 2, \ldots, N_{\mathcal{L}}, \tag{4.6}$$

from which, for every $i = 1, \ldots, N_{\mathcal{L}}$, the couples of Fourier coefficients $(\overline{g}_k, \overline{g}_{k+n_i})$ which are likely to be correlated may be selected. However, we remain with the problem of choosing among the couples the single Fourier coefficients that are likely to be correlated belonging to the sets $\mathcal{F}_i$. In other words, we now have to break the k-translation invariance, which was introduced initially with the assumption of wide stationarity for the process $\{\eta_k\}$. Evidently, the only satisfactory way for solving this problem would require the computation of the correlation function through *ensemble averages* of multiple realizations of $\{\eta_k\}$. Since, in practice, this is usually not possible, we limit ourselves to adopt a heuristic criterion suggested by the definition itself of autocorrelation: for any $n_i \in \mathcal{L}$, $(i = 1, 2, \ldots, N_{\mathcal{L}})$, we select the pair $(\overline{g}_{k_i^*}, \overline{g}_{k_i^*+n_i})$ which mostly contributes to the autocorrelation estimate $\delta_{\overline{g}}(n_i)$, i.e., we define k_i^* as:

$$k_i^* \doteq \arg \max_{k \in [1, N-n_i]} \left\{\left|(\overline{g}_k - \langle \overline{g}_k \rangle)\,(\overline{g}_{k+n_i} - \langle \overline{g}_{k+n_i} \rangle)\right|\right\} \tag{4.7}$$
$$n_i \in \mathcal{L}, \quad i = 1, 2, \ldots, N_{\mathcal{L}}.$$

In this way, for each family $\mathcal{F}_i$ we pick up only two components, the ones at the *frequencies* k_i^* and $k_i^* + n_i$.

However, it could happen that more than one pair of components contributes significantly to the autocorrelation at a certain lag n_i. By using the criterion given above we lose these additional pairs and, apparently, we would be unable to recover important components $\overline{g}_k$. This is only partially true, because the components which could get lost at a certain lag are likely to be selected at another lag. To illustrate this with an example, suppose that, at the lag, say $n = 2$, two couples $(\overline{g}_1, \overline{g}_3)$ and $(\overline{g}_3, \overline{g}_5)$ concur substantially to the non-null value of $|\delta_{\overline{g}}(2)|$ with, for instance, the first couple being dominant. Therefore, the set $\mathcal{I}$ would be expected to contain all three values $k = 1, 3, 5$. Instead, according to the previous criterion of selection, the coefficient $\overline{g}_5$ would be rejected. However, if the component $\overline{g}_5$ does actually emerges from the noise, then it is likely that even the couple $(\overline{g}_1, \overline{g}_5)$ will be selected from the analysis at the lag $n = 4$, leading to correctly recover the

three components $\overline{g}_1$, $\overline{g}_3$, $\overline{g}_5$.

We can thus construct the set $\mathcal{I}$ as (see (2.12a))

$$\mathcal{I} = \{k_i^*\}_{i=1}^{N_\mathcal{L}} \cup \{k_i^* + n_i\}_{i=1}^{N_\mathcal{L}}, \quad n_i \in \mathcal{L}, \tag{4.8}$$

each element of $\mathcal{I}$ being counted only once. The Fourier coefficients $\{\overline{g}_i\}_{i\in\mathcal{I}}$ selected by this procedure are expected to have a high *signal-to-noise ratio* and thus are likely to give a significant contribution to the regularized solution (2.15) since they emerge neatly from the random noise.

Conversely, it can happen that some coefficients, which had to be taken into account in order to have a good reconstruction of the solution, can get lost. The reasons for that are mainly twofold. First, referring to equation (4.2), the Fourier components ξ_k with high *signal-to-noise ratio* s_k^2 but "scarcely" correlated with other Fourier components of ξ can contribute little to the corresponding $r_{\eta\eta}^2(k', k)$ for all $k' \neq k$, and therefore cannot be detected. The second reason for failure is due to the (ever-present) sample errors arising in the statistical analysis of the finite number of true noisy data.

Finally, once the set $\mathcal{I}$ has been constructed, the regularized solution can be computed by means of (2.15). In conclusion, the main steps of the algorithm can be summarized as follows:

1. From the set $\{\overline{g}_k\}_{k=1}^N$ of Fourier coefficients of the noisy data estimate the autocorrelations $\delta_{\overline{g}}(n)$ by formula (4.3);
2. From $\{\delta_{\overline{g}}(n)\}_{n=1}^{N-1}$ construct the set $\mathcal{L}$ by selecting the lags n which correspond to *large* autocorrelations $|\delta_{\overline{g}}(n)|$ according to the criterion given in (4.5);
3. Build the $N_\mathcal{L}$ families $\mathcal{F}_i$ defined in (4.6), and, for each $\mathcal{F}_i$, pick up the pair of Fourier coefficients $(\overline{g}_{k_i^*}, \overline{g}_{k_i^*+n_i})$ which mostly contributes to $|\delta_{\overline{g}}(n_i)|$, k_i^* defined as in (4.7);
4. Construct the set $\mathcal{I}$ in (4.8) with the $N_\mathcal{L}$ pairs $(\overline{g}_{k_i^*}, \overline{g}_{k_i^*+n_i})$, $i = 1, \ldots, N_\mathcal{L}$;
5. Compute the regularized solution $(\widehat{B}\,\overline{g})(x)$ by formula (2.15).

4.1. Numerical examples. In this section, we illustrate the algorithm described above by considering two examples of the integral

equation (1.1) with kernel

$$K(x,y) = \begin{cases} (1-x)\,y & \text{if } 0 \leqslant y \leqslant x \leqslant 1, \\ x\,(1-y) & \text{if } 0 \leqslant x \leqslant y \leqslant 1. \end{cases} \tag{4.9}$$

Eigenfunctions and eigenvalues of the operator A with kernel (4.9) are, respectively:

$$\psi_k(x) = \sqrt{2}\sin(k\pi x), \qquad \lambda_k = \frac{1}{\pi^2\,k^2}, \quad k = 1, 2, \ldots .$$

In Figure 1, the main steps of the reconstructing procedure are given for the test function $f_1(x) = \exp(-3x)\sin(3\pi x)$. The noise bound is $\varepsilon = 10^{-5}$, and the "total" (that is, associated with the whole signal, not with a single Fourier component) *signal-to-noise ratio* SNR, defined as the ratio of the mean power of the noiseless data to the noise variance, is SNR $\simeq 66.5\,\mathrm{dB}$. Panel (a) of Figure 1 displays the noisy Fourier coefficients $\overline{g}_k$. It is evident that they are significantly different from zero only for the first few values of k so that we expect that both the variational solution and the statistical solution would yield a satisfactory reconstruction of the function $f_1(x)$.

In regards to the variational solutions being computed in this section, we have escaped the *ever-present* problem of choosing the regularization parameter α (see, e.g., (1.9)) **[7, 12, 15, 42, 60, 66]** by setting it by hand at the value such that the truncation index $\kappa(\varepsilon)$ (see (1.10)) coincides with the maximum value of k which is present in the set $\mathcal{I}$ (see (2.12a)). Figure 1 (c) illustrates the excellent agreement between the true function $f_1(x)$ (solid line) and the two regularized solutions (dashed line). Notice that, in this example, the variational solution (1.10) and the statistical solution (2.15) coincide since both select all first seven Fourier components $\overline{g}_k$, $k = 1, \ldots, 7$, for the reconstruction of f_1. The statistical solution $\widehat{B}\,\overline{g}$ is built on the analysis of the autocorrelation function shown in Figure 1 (b) from which the lags at which the autocorrelation function is non-null are selected (see definition (4.5) of the set $\mathcal{L}$, and the figure legend for the numerical details).

It is worth comparing the *confidence band* associated with the *large-lag variance* (4.4) (solid line) which allows us to reject many autocorrelation values abnormally inflated by sample errors, with the

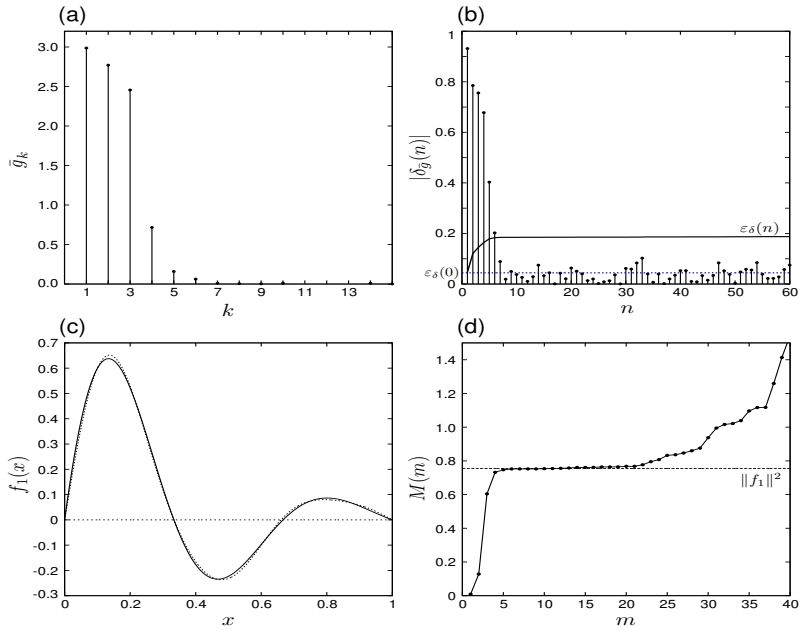

FIGURE 1. $f_1(x) = \exp(-3x)\sin(3\pi x)$, $\varepsilon = 10^{-5}$, SNR $\simeq 66.5\,\mathrm{dB}$, $N = 512$. (a) Noisy Fourier coefficients $\overline{g}_k$ ($\times 10^3$). (b) Modulus of the sample autocorrelation function vs. the lag n (the unit value of autocorrelation at $n = 0$ is not shown). The dashed horizontal line gives the 95% confidence limit for a 512-long purely random sequence. The solid curved line indicates the *large-lag* standard error $\varepsilon_\delta(n)$ of $|\delta_{\bar{g}}(n)|$. $\mathcal{L} = \{1, 2, 3, 4, 5, 6\}$, $\mathcal{I} = \{1, 2, 3, 4, 5, 6, 7\}$. (c) Regularized solutions $\widehat{B\overline{g}}$ and $\mathfrak{f}$, which in this example coincide (dotted line). $c_k = \pi k$, $\alpha = 9.0 \times 10^{-5}$, $\kappa(\varepsilon) = 7$. The solid line indicates the true solution $f_1(x)$. (d) $M(m) \doteq \sum_{k=1}^{m} (\overline{g}_k/\lambda_k)^2$ ($\times 10$) versus m. The value of $M(m)$ corresponding to the exhibited plateau, ranging approximately from $m = 7$ through $m = 20$, is nearly the squared norm of the true solution $\|f_1(x)\|^2$.

constant band at $1.65/\sqrt{N}$ (dashed line). The plot of the function $M(m) = \sum_{k=1}^{m} (\overline{g}_k/\lambda_k)^2$ against m, given in Figure 1 (d), supports the correctness of the reconstructions. It exhibits a *plateau* at the value $M(m) \simeq \|f_1\|^2$, ranging from about $m = 7$ through $m = 20$, which indicates that the bulk of the function f_1 may have been actually recovered. Moreover, carefully inspecting Figures 1 (a) and 1 (d),

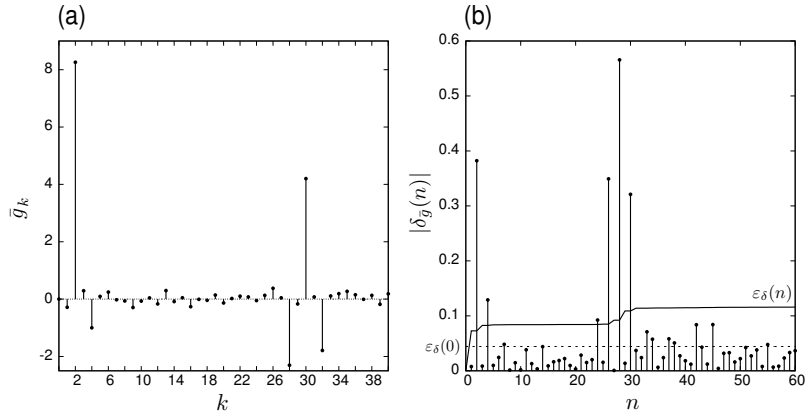

FIGURE 2. $f_2(x) = \exp[-(x-x_0)^2/(2\sigma^2)](\sin 2\pi x + 100 \sin 30\pi x)$, $x_0 = 0.5$, $\sigma = 0.2, \varepsilon = 3{\times}10^{-4}$, SNR $\simeq 80.9$ dB, $N = 512$. (a) Noisy Fourier coefficients $\overline{g}_k$ ($\times 10^3$). (b) Modulus of the sample autocorrelation function versus the lag n (the unit value of autocorrelation at $n = 0$ is not shown). The horizontal dashed line gives the 95% confidence limit for a 512-long purely random sequence. The solid curved line indicates the 95% confidence limit for a one-sided sequence, computed by using the *large-lag* standard error $\varepsilon_\delta(n)$ of $|\delta_{\bar{g}}(n)|$. From the analysis of $|\delta_{\bar{g}}(n)|$, we have: $\mathcal{L} = \{2, 4, 24, 26, 28, 30\}$ and $\mathcal{I} = \{2, 4, 28, 30, 32\}$.

one can observe (from Figure 1 (a)) that the Fourier coefficients $\overline{g}_k$ are relevant up to $k = 7$ (the coefficients $\overline{g}_k$ with $k = 8, 9, 10$ can be neglected); on the other hand, from Figure 1 (d) it can be noted the sum $M(m)$ reaches at $m = 7$ the value $\|f_1\|^2$ and then remains nearly constant up to $m \simeq 20$. This *plateau* shows precisely the transition point from the set $\mathcal{I}$ (from which information can be extracted) to the set $\mathcal{N}$ (where the noise is dominant).

In the second example, shown in Figures 2 and 3, the input test function $f_2(x) = \exp[-(x-x_0)^2/(2s^2)](\sin(2\pi x) + 100 \sin(30\pi x)$ represents a Gaussian burst signal with two main spectral bulks, centered at $k = 2$ and $k = 30$, respectively. Figure 2 (a) shows indeed that the coefficients $\overline{g}_k$ are grouped around $k = 2$ and $k = 30$, so that the variational method is expected to fail for its lack of frequency selectivity, which would be necessary for handling correctly this example. This is what actually happens as the variational reconstruction $\mathfrak{f}(x)$ (see the dashed

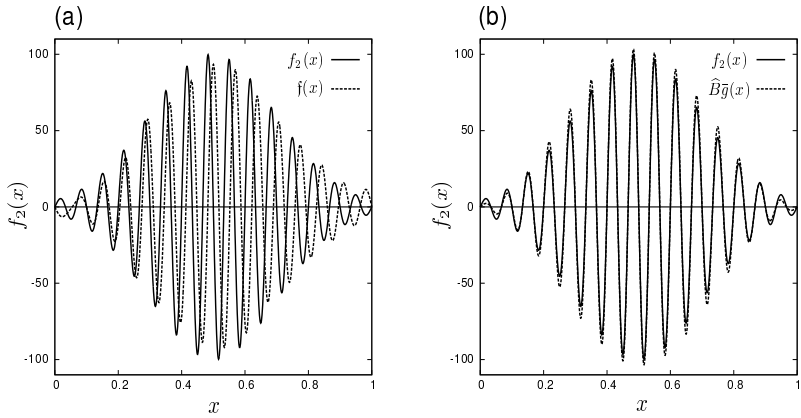

FIGURE 3. Regularized solutions. The input test function (solid line) is: $f_2(x) = \exp[-(x-x_0)^2/(2\sigma^2)](\sin 2\pi x + 100\sin 30\pi x)$, $x_0 = 0.5$, $\sigma = 0.2$, $\varepsilon = 3\times 10^{-4}$, SNR $\simeq$ 80.9 dB, $N = 512$. (a) The dashed line represents the variational solution $\mathfrak{f}(x)$ (see (1.10)). The approximation $\mathfrak{f}(x)$ has been obtained by using $c_k = \pi k$ and $\alpha = 9.5\times 10^{-7}$ (see (1.8)); $\kappa(\varepsilon) = 32$. (b) The dashed line represents the probabilistically regularized solution $\widehat{B}\overline{g}(x)$ (see (2.15)), with the set $\mathcal{I}$ obtained from the analysis of the autocorrelation function shown in Figure 2 (b).

line in Figure 3 (a)) differs evidently from the true function $f_2(x)$. The amplitude of the wavy signal is reconstructed quite correctly but it is manifestly phase-shifted as a consequence of the contribution of many (though small) undue components with $6 \lesssim k \lesssim 22$, which are added to the final solution. Conversely, the statistical reconstruction $(\widehat{B}\overline{g})(x)$ (see (2.15)) is almost indistinguishable from the true function (see the dashed line in Figure 3 (b)). The autocorrelation function shown in Figure 2 (b), from which the statistical solution is constructed, clearly exhibits two main *groups of lags*, whose correct identification leads to a set $\mathcal{L}$ which yields the most (and only) significant Fourier components of the input signal.

5. Concluding remarks. We can now draw the following conclusions.

(1) Deterministic variational methods in the Hilbert space setting could fail in those cases where the Fourier coefficients are sparse or

whenever the significant contribution of the signals are brought by those components which are smoothed by the action of the filter. Probabilistic methods are in general more flexible than the variational ones, and can be used more conveniently in those cases where the Fourier coefficients $\overline{g}_k$ which are relevant for recovering the solution are actually sparse. Sparsity constraints can effectively be incorporated even into regularization procedures developed in Banach spaces (see, e.g., **[15, 58]**).

(2) When one adopts the truncated solutions derived by the methods of variational regularization (like (1.10) and (1.11)), the point where to stop the eigenfunction expansion is obtained by comparing the eigenvalues λ_k with the ratio $(\varepsilon c_k)/E$ (or, simply, (ε/E)). In both cases, this approach appears quite unnatural from the viewpoint of the experimental or physical science, whose methodology would rather suggest smoothing out those Fourier coefficients of the noiseless data which are smaller or of the same order of magnitude of the noise. These remarks lead us to implement probabilistic and statistical methods.

(3) In the probabilistic approach of Section 2, the evaluation at each value of k of the amount of information on ξ_k contained in η_k suggests to split the data set into two subsets: the set $\mathcal{I}$ associated with the Fourier components η_k which contain information on ξ_k, and the set $\mathcal{N}$ associated with the components η_k which are significantly corrupted by the noise and from which no information can be reliably extracted. This strategy aims at extracting the most information from data, keeping at the minimum the use of a priori information on the solution (e.g., only for guiding the choice of functional spaces and norms). The approximation $\widehat{B}\,\overline{g}$, which is indeed based on the set $\mathcal{I}$, is then proved to be a regularized solution to the Fredholm integral equation of the first kind.

(4) The regularization of Fredholm integral equations of the first kind can be approached by using a topological information theory, regarding the inverse problem as a backward channel communication problem. Then one can estimate the maximum number of messages that can be conveyed back from the data for recovering the solution. We find that, in general, this number decreases as the smoothness of the kernel increases. A particularly relevant result obtained in this context concerns the inverse imaging problem. In the standard formulation of this inverse optical problem, we obtain an expression of the resolving

power in terms of the maximum number of messages that can be conveyed back for reconstructing the object. This estimate is given by: $M_\varepsilon(\mathcal{E}) \underset{\varepsilon\to 0}{\sim} 2^{S\log(1/\varepsilon)}$, where S is the Shannon number of the image.

(5) The results provided by topological information theory can be related to those obtained by probabilistic information theory. Within the (simplified) model of the inverse imaging problem of Section 3, we find that the number of binary signals $C_\varepsilon(\mathcal{E})$ that can be transmitted back from the image for recovering the object (i.e., the solution) is (approximately) equal to the amount of information $J(\xi,\eta)$ provided by the w.r.v. η (representing the data) about the w.r.v. ξ (representing the unknown solution).

APPENDIX

A. Weak random variables in Hilbert spaces. In this appendix, we shall briefly recall the notions of random variable and weak random variable in Hilbert spaces. For an extensive analysis of this topic the reader is referred to the book by Balakrishnan [**5**].

Consider the triplet $(\Omega,\mathcal{B},P)$, where Ω is an abstract point set (for instance, the set of the outcomes ω of an experiment), $\mathcal{B}$ is a σ-algebra of subsets of Ω and P is a countably additive probability measure on $\mathcal{B}$ with $P(\Omega)=1$. This triplet is a probability space. Let $\mathcal{H}$ be a separable Hilbert space. Then, a *Hilbert-space random variable* h is a measurable (with respect to $\mathcal{B}$) function of Ω on $\mathcal{H}$, $h:\Omega\to\mathcal{H}$, i.e., $h(\omega)$ is an element of $\mathcal{H}$. Now, the critical point is that not all Hilbert-space processes are Hilbert-space random variables, notably Gaussian processes with covariances not of trace class (e.g., the *white noise*, which is one of the more standard stochastic noise models considered in statistics). Therefore, in order to describe these latter processes, we have to consider a more general setting which embraces the notion of *weak random variable*.

Let us introduce cylinder sets in Hilbert spaces and probability measures thereon. Let $\mathcal{H}_m$ denote a finite-dimensional subspace of $\mathcal{H}$. By a cylinder set, we mean any set of the form $B+\mathcal{H}_m^\perp$, where B is a Borel subset of $\mathcal{H}_m$ (the *base* of the cylinder) and $\mathcal{H}_m^\perp$ is the orthogonal complement of $\mathcal{H}_m$. The class $\mathcal{C}$ of the *cylinder sets* in $\mathcal{H}$ forms a *field* of sets, not a σ-algebra, since the countable unions of sets in $\mathcal{C}$ are not necessarily in $\mathcal{C}$. However, the crucial observation is that the class $\mathcal{B}$

of the Borel sets in $\mathcal{H}$ represents the smallest σ-algebra containing all cylinder sets.

In order to have a probability space, we have to see how to induce a probability measure on $\mathcal{B}$ starting from a measure on the field $\mathcal{C}$ of cylinder sets. Let Z be a cylinder set with base in $\mathcal{H}_m$. Then the *cylinder set measure* μ (also called a *weak distribution*) is defined as

$$\mu(Z) \doteq \nu_m(B),$$

where ν_m is a countably additive probability measure on the σ-algebra of Borel subsets of $\mathcal{H}_m$. This basically means that a measure induced on cylinder sets of $\mathcal{H}$ can be thought of as an ordinary probability measure if one looks only at subspaces of $\mathcal{H}_m$ of finite dimension (i.e., $\mathcal{H}_m$ where the base B of the cylinder *lives*).

For our purposes (though primarily important in its own right) it is sufficient to consider the cylinder measure μ_R induced by a self-adjoint positive definite operator $R : \mathcal{H} \to \mathcal{H}$. The cylinder measure $\mu_R(Z)$ is defined as the Gaussian measure in $\mathcal{H}_m$ with covariance matrix $(R\psi_j, \psi_k)$, $\{\psi_k\}_{k=1}^m$ an orthonormal basis in $\mathcal{H}_m$. The key fact is that, in general, the R-induced cylinder measure $\mu_R(Z)$ cannot be extended to be countably additive on the class of Borel sets $\mathcal{B}$ on $\mathcal{H}$ (the dimension of $\mathcal{H}$ being not finite). Therefore, when dealing with Gaussian distributions we have to deal, in general, with *finitely additive measures*. Countably additivity of μ_R on $\mathcal{B}$ is guaranteed in only one important case by the following theorem.

Theorem A.1 [**5**]. *Let R be a nonnegative self-adjoint operator mapping $\mathcal{H}$ into $\mathcal{H}$. Then, in order that μ_R be countably additive, it is necessary and sufficient that R be of trace class.*

Therefore, in order to consider Gaussian processes not of trace class such as, e.g., the *white noise*, we need to allow for *finitely additive measures* μ on the field (not necessarily a σ-algebra) $\mathcal{C}$ of cylinder subsets of $\Omega \equiv \mathcal{H}$ and, correspondingly, introduce the notion of weak random variable (w.r.v.). Then, given a cylinder-probability triple $(\mathcal{H}, \mathcal{C}, \mu)$ on $\mathcal{H}$, a function $f(\omega)$, mapping $\mathcal{H}$ into $\mathcal{H}$, is a *weak random variable* if, for any finite number m:

(i) the inverse image of any cylinder set of $\mathcal{H}$ with base in $\mathcal{H}_m$ is in $\mathcal{C}$;

(ii) the measure so induced on the Borel sets is countably additive for each m.

Note that the inverse image of Borel sets of $\mathcal{H}$, i.e., $\{\omega : f(\omega) \in$ Borel set in $\mathcal{H}\}$, need not be in $\mathcal{C}$, and hence the *probability* of the corresponding *event* is in general not defined. However, condition (ii) informs us that if $\{\psi_k\}_{k=1}^{\infty}$ is an orthonormal basis in $\mathcal{H}$, then $(f(\omega), \psi_k)_{\mathcal{H}_m}$, $k = 1, \ldots, m$, defines, for each m, an ordinary random variable.

The concept of w.r.v. is, however, similar to that of an ordinary random variable. The main difference is that an ordinary random variable has the property that the inverse image of a Borel set in $\mathcal{H}$ is an event, whereas in the case of a w.r.v., it is only required that the inverse image of any cylinder set in $\mathcal{H}$ be an event.

A w.r.v. $\xi : \mathcal{H} \to \mathcal{H}$ is Gaussian if, for any element $h \in \mathcal{H}$, the *ordinary* random variable $f_w = (\xi, h)_{\mathcal{H}}$ is Gaussian. A Gaussian w.r.v. is uniquely defined by its mean element m_ξ and covariance operator $R_{\xi\xi}$. The mean element is the unique vector of $\mathcal{H}$ such that, for any $h \in \mathcal{H}$:

$$(A.1)\quad (m_\xi, h)_{\mathcal{H}} = E\{(\xi, h)_{\mathcal{H}}\} = \int_{\mathcal{H}} (\xi(\omega), h)_{\mathcal{H}} \, \mathrm{d}\mu(\omega), \quad \text{for all } h \in \mathcal{H},$$

where the cylinder measure is the Gaussian measure μ. The covariance operator $R_{\xi\xi}$ is the unique, bounded, linear, self-adjoint, non-negative operator on $\mathcal{H}$ such that, for any $h \in \mathcal{H}$:

$$\begin{aligned}(A.2)\qquad (R_{\xi\xi}h, h) &= E\{|(\xi - m_\xi, h)_{\mathcal{H}}|^2\} \\ &= \int_{\mathcal{H}} |(\xi(\omega) - m_\xi, h)_{\mathcal{H}}|^2 \, \mathrm{d}\mu(\omega), \quad \text{for all } h \in \mathcal{H}.\end{aligned}$$

Similarly, given another Gaussian w.r.v. $\eta : \mathcal{H} \to \mathcal{H}$, the cross-covariance operator $R_{\xi\eta}$ is the unique linear continuous operator on $\mathcal{H}$ defined for any $h_1, h_2 \in \mathcal{H}$ by:

$$\begin{aligned}(A.3)\qquad (R_{\xi\eta}h_1, h_2) &= E\{(\xi - m_\xi, h_2)_{\mathcal{H}}(\eta - m_\eta, h_1)_{\mathcal{H}}\} \\ &= \int_{\mathcal{H}} (\xi(\omega) - m_\xi, h_2)_{\mathcal{H}}(\eta(\omega) - m_\eta, h_1)_{\mathcal{H}} \, \mathrm{d}\mu(\omega).\end{aligned}$$

A typical example of w.r.v. is the *white noise*, a Gaussian process, which we denote by ζ, whose covariance operator $R_{\zeta\zeta}$ is given by

$R_{\zeta\zeta} = \varepsilon^2 \mathbb{I}$, where $\mathbb{I}$ is the identity operator in $\mathcal{H}$. In this case the Fourier components $\zeta_k(\omega) = (\zeta(\omega), \psi_k)_{\mathcal{H}}$ are independent ordinary Gaussian random variables with zero mean and ε^2 variance. It is worth noting that $\sum_{k=1}^{\infty} [\zeta_k(\omega)]^2 = \|\zeta\|_{\mathcal{H}}^2 < \infty$ for every ω, differently from the usual theory of random variables where the sum of squares of Gaussians with equal variances diverges with probability one. In the latter case, the sample space is the space of all sequences with the corresponding measure countably additive on the Borel sets, whereas in the former case we are dealing only with a finitely additive measure on the field $\mathcal{C}$.

Acknowledgments. We wish to thank the anonymous referees for valuable comments and insightful suggestions which helped to improve the paper.

REFERENCES

1. S. Agapiou, S. Larsson and A.M. Stuart, *Posterior contraction rates for the Bayesian approach to linear ill-posed inverse problems*, Stoch. Proc. Appl. **123** (2013), 3828–3860.

2. S. Agapiou, A.M. Stuart and Y. Zhang, *Bayesian posterior contraction rates for linear severely ill-posed inverse problems*, J. Inv. Ill-Posed Prob. **22** (2014), 297–321.

3. R.L. Anderson, *Distribution of the serial correlation coefficient*, Ann. Math. Stat. **13** (1942), 1–13.

4. G. Backus and F. Gilbert, *The resolving power of gross earth data*, J. Roy. Astronom. Soc. **266** (1968), 169–205.

5. A.V. Balakrishnan, *Applied functional analysis*, Springer-Verlag, New York, 1976.

6. M.S. Bartlett, *An introduction to stochastic processes. With special reference to methods and applications*, Cambridge University Press, Cambridge, 1978.

7. M. Bertero, P. Boccacci, G. Talenti, R. Zanella and L. Zanni, *A discrepancy principle for Poisson data*, Inv. Prob. **26** (2010), 105004.

8. M. Bertero, C. De Mol and G.A. Viano, *The stability of inverse problems*, in *Inverse scattering problems in optics*, Topics Curr. Phys. **20**, Springer, Berlin, 1980.

9. ———, *On the problems of object restoration and image extrapolation in optics*, J. Math. Phys. **20** (1979), 509–521.

10. M. Bertero and G.A. Viano, *On probabilistic methods for the solution of improperly posed problems*, Boll. Un. Mat. Ital. **15** (1978), 483–508.

11. N. Bochkina, *Consistency of the posterior distribution in generalized linear inverse problems*, Inv. Prob. **29** (2013), 095010.

12. T. Bonesky, *Morozov's discrepancy principle and Tikhonov-type functionals*, Inv. Prob. **25** (2009), 015015.

13. C. Clason and M.V. Klibanov, *The quasi-reversibility method for thermoacoustic tomography in a heterogeneous medium*, SIAM J. Sci. Comp. **30** (2008), 1–23.

14. T.M. Cover and J.A. Thomas, *Elements of information theory*, John Wiley, New York, 2005.

15. I. Daubechies, M. Defrise and C. De Mol, *An iterative thresholding algorithm for linear inverse problems with a sparsity constraint*, Comm. Pure Appl. Math. **57** (2004), 1413–1457.

16. E. De Micheli, N. Magnoli and G.A. Viano, *On the regularization of Fredholm integral equations of the first kind*, SIAM J. Math. Anal. **29** (1998), 855–877.

17. E. De Micheli and G.A. Viano, *Metric and probabilistic information associated with Fredholm integral equations of the first kind*, J. Integral Equations Appl. **14** (2002), 283–310.

18. ———, *Inverse optical imaging viewed as a backward channel communication problem*, J. Opt. Soc. Amer. **26** (2009), 1393–1402.

19. ———, *Fredholm integral equations of the first kind and topological information theory*, Int. Equat. Oper. Th. **73** (2012), 553–571.

20. C. De Mol, *Sur la Régularisation des Problèmes Inverses Linéaires*, Ph.D. thesis, Faculté de Sciences, Université Libre de Bruxelles, 1979.

21. D.L. Donoho, *Nonlinear solution of linear inverse problems by wavelet-vaguelette decomposition*, Appl. Comp. Harm. Anal. **2** (1995), 101–126.

22. J.L. Doob, *Stochastic processes*, John Wiley, New York, 1953.

23. H.W. Engl, M. Hanke and N. Neubauer, *Regularization of inverse problems*, Math. Appl. **375**, Kluwer, Dordrecht, 1996.

24. J.N. Franklin, *Well-posed stochastic extensions of ill-posed linear problems*, J. Math. Anal. Appl. **31** (1970), 682–716.

25. B.R. Frieden, *Evaluation, design, and extrapolation methods for optical signals, based on the use of the prolate functions*, in *Progress in Optics*, Vol. 9, North Holland, Amsterdam, 1971, pp. 311-407.

26. I.M. Gelf'and and N.Ya. Vilenkin, *Generalized functions* IV, *Applications of harmonic analysis*, Academic Press, New York, 1964.

27. I.M. Gel'fand and A.M. Yaglom, *Calculation of the amount of information about a random function contained in another such function*, Amer. Math. Soc. Transl. **12** (1959), 199–246.

28. C.W. Groetsch, *The theory of Tikhonov regularization for Fredholm equations of the first kind*, Res. Notes Math. **105**, Pitman, Boston, MA, 1984.

29. J. Hadamard, *Lectures on Cauchy's problem in linear differential equations*, Dover, New York, 1953.

30. P.C. Hansen, *Analysis of discrete ill-posed problems by means of the L-curve*, SIAM Rev. **34** (1992), 561–580.

31. E. Hille and J.D. Tamarkin, *On the characteristic values of linear integral equations*, Acta Math. **57** (1931), 1–76.

32. A. Hofinger and H.K. Pikkarainen, *Convergence rate for the Bayesian approach to linear inverse problems*, Inv. Prob. **23** (2007), 2469–2484.

33. B. Hofmann, B. Kaltenbacher, C. Pöschl and O. Scherzer, *A convergence rates result for Tikhonov regularization in Banach spaces with non-smooth operators*, Inv. Prob. **23** (2007), 987–1010.

34. T. Hohage and F. Werner, *Iteratively regularized Newton-type methods for general data misfit functionals and applications to Poisson data*, Numer. Math. **123** (2013), 745–779.

35. G.M. Jenkins and D.G. Watts, *Spectral analysis and its applications*, Holden-Day, San Francisco, 1968.

36. J.P. Kaipio and E. Somersalo, *Statistical and computational inverse problems*, Appl. Math. Sci. **160**, Springer Verlag, Berlin, 2005.

37. B. Kaltenbacher and B. Hofmann, *Convergence rates for the iteratively regularized Gauss-Newton method in Banach spaces*, Inv. Prob. **26** (2010), 035007.

38. M.V. Klibanov and F. Santosa, *A computational quasi-reversibility method for Cauchy problems for Laplace's equation*, SIAM J. Appl. Math. **51** (1991), 1653–1675.

39. B.T. Knapik, A.W. van der Vaart and J.H. van Zanten, *Bayesian inverse problems with Gaussian priors*, Ann. Stat. **39** (2011), 2626–2657.

40. ———, *Bayesian recovery of the initial condition for the heat equation*, Comm. Stat. Th. Meth. **42** (2013), 1294–1313.

41. A.N. Kolmogorov and V.M. Tihomirov, *ε-entropy and ε-capacity of sets in functional spaces*, Amer. Math. Soc. Trans. **17** (1961), 277–364.

42. J. Krebs, A.K. Louis and H. Wendland, *Sobolev error estimates and a priori parameter selection for semi-discrete Tikhonov regularization*, J. Inv. Ill-Posed Prob. **17** (2009), 845–869.

43. H.J. Landau, *The eigenvalue behavior of certain convolution equations*, Trans. Amer. Math. Soc. **115** (1965), 242–256.

44. S. Lasanen, *Non-Gaussian statistical inverse problems*, *Part* I: *Posterior distributions*, Inv. Prob. Imag. **6** (2012), 215–266.

45. ———, *Non-Gaussian statistical inverse problems*, *Part* II: *Posterior convergence for approximated unknowns*, Inv. Prob. Imag. **6** (2012), 267–287.

46. R. Lattès and J.L. Lions, *Méthode de Quasi-Réversibilité et Applications*, Trav. Rec. Math. **15**, Dunod, Paris, 1967.

47. M.S. Lehtinen, L. Päivärinta and E. Somersalo, *Linear inverse problems for generalized random variables*, Inv. Prob. **5** (1989), 599–612.

48. G.G. Lorentz, *Approximation of functions*, Holt, Rinehart and Winston, New York, 1966.

49. A.K. Louis, *Inverse und schlecht gestellte Probleme*, Teubner Stud. Math., B.G. Teubner, Stuttgart, 1989.

50. D. Middleton, *An introduction to statistical communication theory*, McGraw-Hill, New York, 1960.

51. K. Miller, *Least squares methods for ill-posed problems with a prescribed bound*, SIAM J. Math. Anal. **1** (1970), 52–74.

52. K. Miller and G.A. Viano, *On the necessity of nearly-best-possible methods for analytic continuation of scattering data*, J. Math. Phys. **14** (1973), 1037–1048.

53. A. Neubauer and H.K. Pikkarainen, *Convergence results for the Bayesian inversion theory*, J. Inv. Ill-Posed Prob. **16** (2008), 601–613.

54. Y. Polyanskiy and Y. Wu, *Lecture notes on information theory*, `http://people.lids.mit.edu/yp-/homepage/data/itlectures_v2.pdf`, 2014.

55. R.T. Prosser, *The ε-entropy and the ε-capacity of certain time-varying channels*, J. Math. Anal. Appl. **16** (1966), 553–573.

56. F.R.S. Lord Rayleigh, *Investigations in optics, with special reference to the spectroscope*, Phil. Mag. **8** (1879), 261–274.

57. E. Resmerita and O. Scherzer, *Error estimates for non-quadratic regularization and the relation to enhancement*, Inv. Prob. **22** (2006), 801–814.

58. T. Schuster, B. Kaltenbacher, B. Hofmann and K.S. Kazimierski, *Regularization methods in Banach spaces*, De Gruyter, Berlin, 2012.

59. D. Slepian and H.O. Pollack, *Prolate spheroidal wave functions, Fourier analysis and uncertainty*–I, Bell Syst. Tech. J. **40** (1961), 43–64.

60. D. Krawczyk-Stando and M. Rudnicki, *Regularization parameter selection in discrete ill-posed problems–The use of the U-curve*, Int. J. Appl. Math. Comp. Sci. **17** (2007), 157–164.

61. A.M. Stuart, *Inverse problems*: *A Bayesian perspective*, Acta Num. **19** (2010), 451–559.

62. G. Talenti, *Sui problemi mal posti*, Boll. Un. Mat. Ital. **15** (1978), 1–29.

63. L. Tenorio, *Statistical regularization on inverse problems*, SIAM Rev. **43** (2001), 347–366.

64. A.N. Tikhonov and Ya.V. Arsenin, *Méthodes de Resolution de Problèmes Mal Posés*, Mir, Moscow, 1976.

65. G. Toraldo di Francia, *Degrees of freedom of an image*, J. Opt. Soc. Amer. **59** (1969), 799–804.

66. G. Wahba, *Practical approximate solutions to linear operator equations when the data are noisy*, SIAM J. Num. Anal. **14** (1977), 651–667.

67. G. Wahba and Y. Wang, *When is the optimal regularization parameter insensitive to the choice of the loss function?*, Comm. Stat. Th. Meth. **19** (1990), 1685–1700.

68. G.M. Wing and J.D. Zahtr, *A primer on integral equations of the first kind. The problem of deconvolution and unfolding*, SIAM, Philadelphia, 1991.

IBF - Consiglio Nazionale delle Ricerche, Via De Marini, 6 - 16149 Genova, Italy
Email address: **enrico.demicheli@cnr.it**

Scuola di Scienze Matematiche, Fisiche e Naturali, Università di Genova, Via Dodecaneso, 33 - 16146 Genova, Italy