Regularizing Priors for Linear Inverse Problems

JEAN-PIERRE FLORENS AND ANNA SIMONI

REGULARIZING PRIORS FOR LINEAR INVERSE PROBLEMS[*]

BY JEAN-PIERRE FLORENS AND ANNA SIMONI

Université Toulouse 1 Capitole and Università Bocconi

We consider statistical linear inverse problems in Hilbert spaces of the type $\hat{Y} = Kx + U$ where we want to estimate the function x from indirect noisy functional observations $\hat{Y}$. In several applications the operator K has an inverse that is not continuous on the whole space of reference; this phenomenon is known as ill-posedness of the inverse problem.

We use a Bayesian approach and a conjugate-Gaussian model. For a very general specification of the probability model the posterior distribution of x is known to be inconsistent in a frequentist sense. Our first contribution consists in constructing a class of Gaussian prior distributions on x that are shrinking with the measurement error U and we show that, under mild conditions, the corresponding posterior distribution is consistent in a frequentist sense and converges at the optimal rate of contraction. Then, a class Λ of posterior mean estimators for x is given. We propose an empirical Bayes procedure for selecting an estimator in this class that mimics the posterior mean that has the smallest risk on the true x.

1. Introduction. Given a functional noisy observation $\hat{Y}$, we consider the problem of estimating the solution x of the noisy functional equation

$$(1.1) \qquad \hat{Y} = Kx + U, \qquad x \in \mathcal{X}, \hat{Y} \in \mathcal{Y}$$

where $\mathcal{X}$ and $\mathcal{Y}$ are infinite dimensional separable Hilbert spaces over $\mathbb{R}$ with norm $|| \cdot ||$ induced by the inner product $< \cdot, \cdot >$. The residual U is a stochastic noise and $K : \mathcal{X} \to \mathcal{Y}$ is a known, bounded, linear operator with infinite dimensional range. The operator K^* will denote the adjoint of K, *i.e.* K^* is such that $< K\varphi, \psi > = < \varphi, K^*\psi >$, $\forall \varphi \in \mathcal{X}$ and $\psi \in \mathcal{Y}$.

This model is classical in inverse problem literature and it is encountered

in many applications. Examples are provided for instance by deconvolution or error-in-variables models, nonparametric regression, density estimation, mathematical models for tomography and signal and image processing. An overview of applications can be found in [34].

In certain applications the operator K is compact, but we admit also non-compact K. The bounded operator K is *compact* if the image $\{Kx_n\}$ of any bounded sequence $\{x_n\}$ of $\mathcal{X}$ contains a Cauchy subsequence, see [23]. A compact operator is characterized by a countable number of eigenvalues (or singular values) that accumulates only at zero. A non-compact operator may be characterized by a continuum of eigenvalues (or singular values) that accumulates at zero. In the particular case with $K = I$, where I is the identity operator, we call (1.1) a direct problem and it is well-posed.

When the eigenvalues of K accumulate at zero, the problem of recovering an estimation of the original signal x from $\hat{Y}$ is ill-posed because in general K^{-1} is not continuous on the whole $\mathcal{Y}$, see Chapter 2 in [11]. We are in principle motivated by inverse problems affected by this kind of ill-posedness, however, the methodology that we propose can also be applied to direct and well-posed inverse problems.

The unboundedness of K^{-1} is a problem because it prevents the estimated solution $K^{-1}\hat{Y}$ from being consistent, even if $\hat{Y}$ converges to Kx, since the noise in $\hat{Y}$ is deeply amplified by K^{-1}. To remove the ill-posedness, regularization methods have been proposed in numerical analysis and classical statistics like Spectral cut-off, Tikhonov regularization, or Landweber-Fridman regularization, to name only a few, see [25] Section 15.5 and references therein.

In reverse, in this paper we focus on Bayesian methods for solving inverse problems. This is the most complete statistical way to address inverse problems because we model a distribution for the noise (and not only a level for the noise as in numerical analysis) and we specify a prior distribution on x.

The elements x and U are modeled as Hilbert space-valued random variables (H-r.v.), that is, for a complete probability space $(S, \mathcal{S}, \mathbb{P})$, x (resp. U) defines a measurable map

$$x : (S, \mathcal{S}, \mathbb{P}) \to (\mathcal{X}, \mathfrak{B}(\mathcal{X})) \quad (\text{resp. } U : (S, \mathcal{S}, \mathbb{P}) \to (\mathcal{Y}, \mathfrak{B}(\mathcal{Y})))$$

where $\mathfrak{B}(\mathcal{X})$ (resp. $\mathfrak{B}(\mathcal{Y})$) is the Borel σ-field generated by the open sets of $\mathcal{X}$ (resp. of $\mathcal{Y}$), see, *e.g.* [26]. The Bayesian approach combines the prior and sampling information in order to overcome the ill-posedness of K^{-1} and proposes the posterior distribution of x as solution for (1.1). However, since the dimension of the problem is infinite, the posterior distribution suffers of another problem: it is, in general, inconsistent in the frequentist sense, that

is, as the stochastic measurement error degenerates to 0 the posterior distribution does not converge towards a degenerate distribution concentrated on the true value of x, see [10] for frequentist inconsistency in nonparametric Bayes estimation. On the contrary, when (1.1) is an inverse problem in finite dimensional spaces and K is a matrix, the posterior distribution is consistent in a frequentist sense (under mild assumptions) and the posterior mean is equal to the Tikhonov regularized solution (also known as *Ridge regularized* solution in finite dimensional problems), see for instance [22].

To solve the problem of frequency inconsistency of the posterior distribution, we have proposed in another paper, see [12], to regularize the posterior distribution with a Tikhonov scheme and we have defined a new object called *Regularized Posterior distribution* that plays the role of the posterior distribution. Another solution has been proposed by [27] and [29] and consists in regularizing through a restriction of the space of definition of $\hat{Y}$.

The purpose of this paper is twofold. First, we show that when we have a particular structure of the operator K and of the covariance operator of U in (1.1), a prior covariance operator can be chosen in such a way that the posterior distribution is consistent in the frequentist sense. In practice, the regularization can be automatically performed by the prior-to-posterior transformation and no ad-hoc regularization needs to be introduced. The prior covariance operator that we specify depends on a parameter $\alpha > 0$, on the level of measurement noise U and on the degree of penalization s chosen for measuring the variability of the solution (as, for instance, the highest order of derivatives in a Sobolev penalization). We show that, while the parameter s does not enter the rate of convergence of the posterior distribution, the parameter α does and moreover it plays the role of a regularization parameter.

We compute an upper bound for the quadratic risk associated with the posterior mean estimator $\hat{x}_\alpha$. We determine the α which minimizes this bound and the corresponding rate. Moreover, we compute the fastest rate of contraction of the posterior distribution.

Then, we consider the specific case of operators with geometric spectra, *i.e.* the mildly ill-posed case. In this framework, the rate of convergence previously found is shown to be the optimal rate of convergence, in a minimax sense, associated with the class of estimators $\hat{x}_\alpha$, $\alpha \in \Lambda$, where Λ is some set to be specified later. The corresponding optimal α is the solution of the minimization problem

$$\alpha_{opt}(\Lambda, x) = \arg \inf_{\alpha \in \Lambda} \mathbf{E}_x ||\hat{x}_\alpha - x||^2$$

if such a value exists, where $\mathbf{E}_x(\cdot)$ denotes the expectation taken with respect to the conditional distribution of $\hat{Y}$ given x. We call *oracle* the random function $x \mapsto \hat{x}_{\alpha_{opt}}$, that is, the best estimate of x, see [36]. However, as the oracle depends on the true value x and on its regularity, which are unknown, we cannot compute it.

Our second purpose is therefore to determine a data-driven method for optimally selecting the parameter α in the prior distribution. We propose an empirical Bayes (EB) approach for recovering α and we denote this value by $\hat{\alpha}$. We prove for the case of operators with geometric spectra an *oracle inequality*, that is, an inequality of the type

$$\mathbf{E}_x||\hat{x}_{\hat{\alpha}} - x||^2 \leq \inf_{\alpha \in \Lambda} \mathbf{E}_x||\hat{x}_\alpha - x||^2 (1 + \mathcal{O}(1))$$

for each x belonging to a subset $\mathcal{X}_\beta \subset \mathcal{X}$. Then, our EB estimator is *adaptive* to the oracle.

The paper develops as follows. In the next two subsections we specify the statistical model associated with (1.1) and the prior distribution of x; some example of application of model (1.1) are given. In Section 2 we compute the posterior distribution of x and we prove that it is consistent in a frequentist sense. We provide the minimax rate of convergence of the posterior mean when it is taken as estimator for x and when the operators have geometric spectra. In section 3, we develop an EB method for selecting a data-dependent regularization parameter $\hat{\alpha}$ with values in Λ. Optimality, in a minimax sense, of $\hat{\alpha}$ and $\hat{x}_{\hat{\alpha}}$ follows from the oracle inequality in Theorem 3.2 which is one of the main results of our paper. All the proof are provided in Section 4.

It is convenient to set up some notational convention used in the paper. For positive quantities M_δ and N_δ depending on a discrete or continuous index δ, we write $M_\delta \asymp N_\delta$ to mean that the ratio M_δ/N_δ is bounded away from zero and infinity. We write $M_\delta \sim N_\delta$ if there exist $\underline{m}, \bar{m} > 0$ such that $\underline{m}N_\delta \leq M_\delta \leq \bar{m}N_\delta$. We write $M_\delta = \mathcal{O}(N_\delta)$ if M_δ is at least of the same order as N_δ. We denote with $\mathcal{R}(\cdot)$ the range of an operator, with $\mathcal{D}(\cdot)$ its domain and with $tr(\cdot)$ its trace. For an operator $A : \mathcal{Y} \to \mathcal{X}$ (resp. $B : \mathcal{X} \to \mathcal{Y}$), we denote with A^* (resp. B^*) its adjoint and for a subset $\mathcal{Y}_1 \subset \mathcal{Y}$ (resp. $\mathcal{X}_1 \subset \mathcal{X}$), $A|_{\mathcal{Y}_1} : \mathcal{Y}_1 \to \mathcal{X}$ (resp. $B^*|_{\mathcal{X}_1} : \mathcal{X}_1 \to \mathcal{Y}$) denotes the restriction of A (resp. of B) to the domain $\mathcal{Y}_1$ (resp. $\mathcal{X}_1$). The operator I denotes the identity operator on both spaces $\mathcal{X}$ and $\mathcal{Y}$, *i.e.* $\forall \psi \in \mathcal{X}$, $\varphi \in \mathcal{Y}$, $I\psi = \psi$ and $I\varphi = \varphi$.

1.1. *Sampling distribution and examples.* Let the random error U in model (1.1) be a H-r.v. with gaussian distribution $\mathcal{N}(0, \delta\Sigma)$, where $\Sigma : \mathcal{Y} \to \mathcal{Y}$ is a one-to-one, positive definite, self-adjoint, trace class operator and $\delta > 0$ is a scale parameter. Then, the conditional distribution of $\hat{Y}$ given x, denoted by P^x, is gaussian:

$$(1.2) \qquad\qquad \hat{Y}|x \sim \mathcal{N}(Kx, \delta\Sigma)$$

and it is known in Bayes literature as *sampling distribution*. Hereafter, $\mathbf{E}_x(\cdot)$ will denote the expectation taken with respect to P^x. The scale parameter δ is the noise level in the observation $\hat{Y}$ and $\delta \downarrow 0$. In many statistical and econometrics models the observed function $\hat{Y}$ is a transformation of an n-sample of finite dimensional objects (like the empirical distribution function, the empirical characteristic function, the Nadaraya-Watson estimator of the regression function) or it is the mean of functional data (like many example in signal and image processing). In these models $\delta := \delta(n)$, where n denotes the sample size and $\delta \downarrow 0$ as $n \uparrow \infty$.

The covariance operator $\Sigma : \mathcal{Y} \to \mathcal{Y}$ is assumed to be fixed and known. It follows from the definition of covariance operator that it is linear, bounded and compact. The trace-class property of Σ rules out a covariance operator proportional to the identity operator. In fact, an operator is *trace class* if its trace is finite. For $\Sigma = cI$, $c > 0$, $tr(\Sigma)$ is clearly unbounded. Moreover, in many applications Σ is not naturally modeled as I. This implies that, in general, in infinite-dimensional inverse problems the posterior mean cannot be equal to the Tikhonov regularized estimator $x_\alpha^T := (\alpha I + K^*K)^{-1}K^*\hat{Y}$ as it happens in the finite dimensional case (*e.g.* for the *Ridge regression*). However, we show in this paper that, when Σ, K and the prior covariance operator are suitably linked (see Assumptions A.1, A.2 and C.1-C.3), the posterior mean equates the Tikhonov regularized solution in the *Hilbert Scale* induced by the prior covariance operator. We give later the definition of Hilbert Scale.

The assumption of gaussianity of the statistical model (1.2) is just made in order to construct the estimator and it is not restrictive. The proofs of our result of frequency consistency do not rely on the normality of U. In particular, our estimation procedure can be applied to cases where $\hat{Y}|x$ is only asymptotically gaussian, as in Examples 2 and 3 below.

We give now three examples of estimation problems that can be written in the form of (1.1) and treated with our methodology. Another example, concerning the estimation of a functional regression, is given in section 2.3.

EXAMPLE 1. *White noise model.* Consider the following model

$$(1.3) \qquad dY_\delta(t) = f(t)dt + \frac{1}{\sqrt{n}}dB(t), \quad t \in [0,1]$$

where $\{Y_\delta(t)\}$ is the noisy observation, $f(\cdot) \in L^2([0,1]) := \{f : \int_0^1 |f|^2(t)dt < \infty\}$ is the unknown signal and $B(t)$ is a standard Brownian motion. This model can be rewritten as

$$Y_\delta(t) = \int_0^t f(s)ds + \frac{1}{\sqrt{n}}B(t)$$

and is the limiting experiment for some curve estimation problems such as density (see [33]) and nonparametric regression (see [4]) estimation. The covariance operator Σ is $\varphi \in \mathcal{Y} \mapsto \Sigma\varphi = \int_0^1 (s \wedge t)\varphi(s)ds$ and has eigenvalues $\lambda_j^2 = \frac{4}{\pi^2 j^2}$ and eigenvectors $\varphi_j(t) = \sqrt{2}\sin(\frac{\pi j}{2}t)$ for $j = 1, 3, 5, \ldots$. The transformation K is $\psi \in \mathcal{X} \mapsto K\psi = \int_0^t \psi(s)ds$ and, since $KK^* = \Sigma$, the singular values of K are λ_j, $j = 1, 3, 5, \ldots$. Thus, Assumption A.1 below is satisfied and Assumption A.2 holds with $a = 0$, as it will result clear from subsection 2.2. The conjugate gaussian sequence space model associated with (1.3) is considered *e.g.* in [45] and [2]; they show in Theorem 5.1 and Theorem 2.1, respectively, that the minimax rate of convergence is attained by the posterior mean and the posterior distribution of the Fourier coefficients of f. We are able to attain the same rate for the posterior mean and distribution.

EXAMPLE 2. *Regression estimation.* Let (ξ, w) be a $\mathbb{R}^{1+q}$-valued random vector with distribution F and $\mathcal{X} := L_F^2(w)$ be the space of square integrable functions of w with respect to F. We want to recover the regression function of ξ given w, that is, the function $m(w) \in L_F^2(w)$ such that $\xi = m(w) + \varepsilon$, $\mathbf{E}(\varepsilon|w) = 0$ and $\mathbf{E}(\varepsilon^2|w) = \sigma^2$. This function can be characterized as the solution of a linear inverse problem.

Let $g(w,t) : \mathbb{R}^q \times \mathbb{R}^p \to \mathbb{R}$, $p \geq q$, be a known function that is square integrable with respect to $F \times \pi$, with π a measure on $\mathbb{R}^p$. Then, by denoting with $\mathbf{E}$ the expectation taken with respect to F, m is solution of $\mathbf{E}(g(w,t)\xi) = Km(t)$, where $\phi \in L_F^2(w) \mapsto (K\phi)(t) = \int g(w,t)\phi(t)dF(w)$. Hence, $K : \mathcal{X} \to \mathcal{Y} := L_\pi^2(\mathbb{R}^p)$ is a compact operator which is known if $F(w)$ is known. Moreover, the fact that ξ has finite second moment ensures that $\mathbf{E}(g(w,t)\xi) \in L_\pi^2(\mathbb{R}^p)$. The left hand side $\mathbf{E}(g(w,t)\xi)$ is unknown and can be estimated by its empirical mean, so that

$$\frac{1}{n}\sum_{i=1}^n g(w_i,t)\xi_i = Km + U,$$

where (ξ_i, w_i), $i = 1, \ldots, n$ is an *iid* sample from F. The error $U \in \mathcal{Y}$ is asymptotically gaussian with covariance operator $\frac{\sigma^2}{n} KK^*$ which is known and $\delta = \frac{\sigma^2}{n}$. Assumptions A.1, A.2, C.1-C.3 below are satisfied.

EXAMPLE 3. *Instrumental regression estimation.* Let (Y, Z, W) be a random vector with values in $\mathbb{R} \times \mathbb{R}^p \times \mathbb{R}^q$ and distribution F admitting a density f. Let $L_F^2(Z)$ denote the space of square integrable functions of Z with respect to F. The instrumental regression $\varphi(Z) \in L_F^2(Z)$ is the parameter we want to estimate and is defined by

$$Y = \varphi(Z) + \varepsilon, \qquad \mathbf{E}(\varepsilon|W) = 0, \qquad Var(\varepsilon|W) = \sigma^2$$

as the solution of an integral equation of first kind: $\mathbf{E}(Y|W) = \mathbf{E}(\varphi(Z)|W)$, see [9] and [17]. We suppose to know $f(Z, W)$ while $f(Y, W)$ is unknown; then, we use a kernel estimator for $\mathbf{E}(Y|W)$ and we take into account the estimation error: $\hat{\mathbf{E}}(Y|W) = \mathbf{E}(\varphi(Z)|W) + U$. The sampling distribution can be approximated by using the asymptotic distribution of U. As the kernel estimator $\hat{\mathbf{E}}(Y|W)$ does not weakly converge toward a well-suited process (like a gaussian process) we smooth the model by re-projecting it on $L_F^2(Z)$. The instrumental regression is now characterized as the solution of $\mathbf{E}(\hat{\mathbf{E}}(Y|W)|Z) = K\varphi + V$, with $K = \mathbf{E}(\mathbf{E}(\cdot|W)|Z)$ and V a new error term. The (approximated) distribution of V is gaussian with zero mean and covariance operator $\frac{1}{n}\sigma^2 K$, where K is a known operator.

1.2. *Prior measure and main assumptions.* In the following, we introduce the prior distribution for x and three assumptions. Some of these assumptions, like Assumptions A.1 and C.1 below, are not standard in statistical inverse problems literature. This is because, in the most of the literature, U is modeled as an Hilbert space process (see *e.g.* [3], Section 2.5) with the consequence that $\Sigma = I$ and Assumptions A.1 and C.1 are automatically satisfied. Furthermore, Assumption C.1 is relevant only for Bayes inversion theory. In our framework these assumptions are necessary in order to have a consistent posterior distribution. When these assumptions are not verified we are in the general case of posterior inconsistency for which a regularized posterior distribution has been proposed in [12].

ASSUMPTION A.1. $\mathcal{R}(K) \subset \mathcal{D}(\Sigma^{-\frac{1}{2}})$.

Assumption A.1 demands that K is at least as smooth as $\Sigma^{\frac{1}{2}}$ and ensures that operator $\Sigma^{-\frac{1}{2}} K$, used in Assumption A.2 below, is well-defined. In other

words, we demand a compatibility between K and the covariance operator Σ in the sampling mechanism. This is very common in practical examples where the covariance operator is often of the form $\Sigma = (KK^*)^r$, for some $r \leq 1$, see subsection 2.3.

ASSUMPTION A.2. *There exists an unbounded densely defined operator L in the Hilbert space $\mathcal{X}$ that is self-adjoint, strictly positive and that satisfies*

$$(1.4) \qquad \underline{m}||L^{-a}x|| \leq ||\Sigma^{-\frac{1}{2}}Kx|| \leq \bar{m}||L^{-a}x||$$

on $\mathcal{X}$ for some $a > 0$ and $0 < \underline{m} \leq \bar{m} < \infty$. Moreover, L^{-2s} is trace-class for some $s > 0$.

Assumption (A.2) means that $\Sigma^{-\frac{1}{2}}K$ regularizes at least as much as L^{-a}. The operator L^{-2s} is used to construct the prior covariance operator. The H-r.v. x has a Gaussian distribution:

$$(1.5) \qquad x|g, s \sim \mathcal{N}\left(x_0, \frac{1}{g}L^{-2s}\right),$$

with $x_0 \in \mathcal{X}$ and $g = g(\delta)$ a function of δ such that $g \uparrow \infty$ as $\delta \downarrow 0$. This entails that the prior distribution shrinks towards the prior mean which is necessary for posterior consistency. The parameter g describes a class of prior distributions and it may be viewed as an hyperparameter. We provide in Section 3 an Empirical Bayes approach for selecting it. Gaussian processes for functional estimation and frequentist properties of the posterior distribution have been extensively studied in [37], [38], [40].

In the following, we use the notation $\Omega_0 = L^{-2s}$. Hence, $\Omega_0 : \mathcal{X} \to \mathcal{X}$ is a linear, bounded, positive-definite, self-adjoint, compact and trace-class operator. This choice of the prior covariance is aimed to link the prior distribution with the sampling model as it results evident from Assumption A.2. A similar idea was proposed by Zellner [43] for linear regression models for which he constructed a class of prior called g-prior. Our prior is an extension of the Zellner's g-prior and we call (1.5) *extended g-prior*.

Roughly speaking, Assumption A.2 quantifies the regularity of $\Sigma^{-\frac{1}{2}}K$ while Assumption B below quantifies the regularity of the true value of x. In order to understand Assumption A.2 we need to: *(i)* introduce the definition of *Hilbert scale*, *(ii)* explain the meaning of the parameter a, *(iii)* discuss the regularity conditions of $\Sigma^{-\frac{1}{2}}K$ and of the true value of x. *(i)*For all $s \in \mathbb{R}$, operator L in Assumption A.2 induces the Hilbert scale $(\mathcal{X}_s)_{s\in\mathbb{R}}$, where $\mathcal{X}_s$ is an Hilbert space defined as the completion of $\bigcap_{s\in\mathbb{R}} \mathcal{D}(L^s)$ with respect to the norm $||x||_s := ||L^sx||$, see [24], [11], [31]. Let notice that as

$\Sigma^{-\frac{1}{2}}K$ must satisfy (1.4) it is necessarily an injective operator.

(ii) Parameter a is the degree of ill-posedness in the statistical experiment. It is usually smaller than the degree of ill-posedness in the classical problem $\hat{Y} = Kx$ since it is determined by the rate of decreasing of the spectrum of operator $\Sigma^{-\frac{1}{2}}K$ and not by that one of K. In other words, to have consistency of the posterior mean we have to control for less ill-posedness than to have consistency of the classical solution. This is because the operator that matters is $\Sigma^{-\frac{1}{2}}K$ and it has a spectrum decreasing slower than that one of K, which is the operator that matters for the classical estimator of x.

(iii) From a frequentist point of view, there exists a true value of the parameter of interest x having generated the data $\hat{Y}$. We denote this value with x_* and it will be used in the asymptotic analysis since we care for the weak convergence of the posterior distribution of x towards a point mass in x_* as $\delta \downarrow 0$. It is a convergence with respect to the sampling probability and it is known as *posterior consistency*, see [16].

In nonparametric estimation it is customary to assume that the functional parameter x_* that we wish to estimate belongs to some known space of regularity. We find a similar idea in inverse problem theory with the difference that in this framework it is natural to impose conditions on the regularity of x_* by taking into account the behavior of the operator that characterizes the inverse problem (that is, the operator $\Sigma^{-\frac{1}{2}}K$ in our case). A way for implementing this consists in introducing a Hilbert Scale and in formulating conditions on x_* and K with respect to this common Hilbert Scale. In this paper we use the space $\mathcal{X}_s$, introduced above, as common Hilbert Scale. Therefore, the meaning of Assumption A.2 is to quantify the regularity of $\Sigma^{-\frac{1}{2}}K$, we refer to [20] Section 2 for a complete explanation of the relationship between Hilbert Scale and regularity conditions. A very similar assumption can be found in [8], see Assumptions 2.2 and 4.2 in this paper. Next assumption expresses the regularity of x_* according to $\mathcal{X}_s$.

ASSUMPTION B. *For some $\beta \geq s$, $(x_* - x_0) \in \mathcal{X}_\beta$, i.e. there exists a* $\rho_* \in \mathcal{X}$ *such that* $(x_* - x_0) = L^{-\beta}\rho_*$ $(\equiv \Omega_0^{\frac{\beta}{2s}}\rho_*)$.

The parameter β characterizes the regularity of the true function x_* and it is generally unknown. For instance, if the Hilbert scale $\mathcal{X}_\beta$ is given by the Sobolev space, then, Assumption B is equivalent to assume that $(x_* - x_0)$ has at least β square integrable derivatives, see [20].

Because $\beta \geq s$, it follows that $\mathcal{R}(\Omega_0^{\frac{\beta}{2s}}) \subset \mathcal{R}(\Omega_0^{\frac{1}{2}})$ and Assumption B implies that there exists a ξ_* such that $(x_* - x_0) = \Omega_0^{\frac{1}{2}}\xi_*$ and $\xi_* = \Omega_0^{\frac{\beta-s}{2s}}\rho_*$. Moreover,

by Proposition 3.6 in [6], we can write $\mathcal{R}(\Omega_0^{\frac{1}{2}}) = \mathcal{H}(\Omega_0)$, where $\mathcal{H}(\Omega_0)$ denotes the *Reproducing Kernel Hilbert Space* ($\mathcal{R}.\mathcal{K}.\mathcal{H}.\mathcal{S}$. in the following) associated with Ω_0 and embedded in $\mathcal{X}$, *i.e.*

$$\mathcal{H}(\Omega_0) = \Big\{ \varphi : \varphi \in \mathcal{X} \quad \text{and} \quad ||\varphi||^2_{\Omega_0} := \sum_{j=1}^{\infty} \frac{|<\varphi, \varphi_j^{\Omega_0}>|^2}{\lambda_j^{\Omega_0}} < \infty \Big\}.$$

where $|| \cdot ||_{\Omega_0}$ denotes the norm in $\mathcal{H}(\Omega_0)$ and $\{\lambda_j^{\Omega_0}, \varphi_j^{\Omega_0}\}$ is the *eigensystem* associated to Ω_0. Hence, Assumption B implies that $(x_* - x_0) \in \mathcal{H}(\Omega_0)$. The $\mathcal{R}.\mathcal{K}.\mathcal{H}.\mathcal{S}$. is a subset of $\mathcal{X}$ that gives the geometry of the distribution of x. The support of a centered Gaussian process, taking its values in an Hilbert space $\mathcal{X}$, is the closure in $\mathcal{X}$ of the $\mathcal{R}.\mathcal{K}.\mathcal{H}.\mathcal{S}$. associated with the covariance operator of this process (denoted with $\overline{\mathcal{H}(\Omega_0)}$ in our case). Then, for our prior distribution, $(x - x_0) \in \overline{\mathcal{H}(\Omega_0)}$ with probability 1, but with probability 1, $(x - x_0)$ is not in $\mathcal{H}(\Omega_0)$, see [38]. More properties about the $\mathcal{R}.\mathcal{K}.\mathcal{H}.\mathcal{S}$. associated with a gaussian measure can be found in [39].

REMARK 1.1. Assumption B is classical in Inverse Problem literature, see *e.g.* [8] and [30], and it is closely related to the so-called *source condition* which expresses the smoothness (regularity) of the function x_* according to the spectral representation of the operator K^*K defining the inverse problem, that is, with respect to the canonical Hilbert Scale, see [11] and [6]. In our case, the smoothness of $(x_* - x_0)$ is expressed according to the spectral representation of L. In the particular case considered in subsection 2.2 it will result clear how our assumption relates to similar assumptions in inverse problems literature.

REMARK 1.2. Assumption A.2 covers both the mildly ill-posed and the severely ill-posed case (under some smoothness assumptions on x_*). In the *mildly ill-posed* case the singular values of K decay slowly to zero (typically at a geometric rate) which means that the kernel of $\Sigma^{-\frac{1}{2}}K$ is finitely smooth. In this case the operator L is generally some differential operator such that L^{-1} is finitely smooth. In the *severely ill-posed* case the singular values of K decay very rapidly (typically at an exponential rate). Assumption A.2 cover also this case if the function x_* to be estimated is very smooth. This is because when the singular values of $\Sigma^{-\frac{1}{2}}K$ decay exponentially, Assumption A.2 is satisfied if also L^{-1} has an exponentially decreasing spectrum. On the other hand, L^{-1} is used to describe the smoothness of x_*; therefore Assumption B can be satisfied only if x_* is infinitely smooth.

Hereafter, we use the notation: $\alpha = \delta g$, $B = \Sigma^{-\frac{1}{2}}K\Omega_0^{\frac{1}{2}}$, $\tilde{B} = \Sigma^{-\frac{1}{2}}K$ and we can rewrite the prior covariance operator as $\frac{1}{g}\Omega_0 = \frac{\delta}{\alpha}\Omega_0$. Operator $\tilde{B}$ is

well defined under Assumption A.1 while operator B is well-defined under Assumption C.1 below. The parameter α is assumed to belong to a set Λ and plays the role of regularization parameter. For that, it must decrease to zero; the rate at which $\alpha \downarrow 0$ is crucial in order to get posterior consistency.

ASSUMPTION C.1. $\mathcal{R}(K\Omega_0^{\frac{1}{2}}) \subset \mathcal{D}(\Sigma^{-1})$.

ASSUMPTION C.2.   *The real parameters a, b and s satisfy the inequalities* $0 < a \leq s \leq \beta \leq 2s + a$.

ASSUMPTION C.3.   *There exists $\gamma \in]0,1]$ such that the operator $(B^*B)^\gamma$ is trace class, i.e. if $\{\lambda_j^2\}$ denotes the eigenvalues of B^*B, then, $\sum_j \lambda_j^{2\gamma} < \infty$ must be verified.*

Assumption C.1 concerns the degree of regularity (*e.g.* the differentiability) of the prior covariance operator with respect to the sampling covariance operator. Under Assumption C.1 and since $\mathcal{D}(\Sigma^{-1}) \subset \mathcal{D}(\Sigma^{-\frac{1}{2}})$, operator B is well-defined.

Assumption C.2 is classical in inverse problem theory, see *e.g.* [11] Section 8.5. The restriction $s \leq \beta$ means that $(x_* - x_0)$ has to be at least an element of $\mathcal{X}_s$ and it guarantees that the norm $||L^s x||$ exists $\forall x \in \mathcal{X}_\beta$. The upper bound $(2s + a)$ of β is the qualification of the regularization scheme: it says that we can at most exploit a regularity of x_* equal to $(2s + a)$.

Assumption C.3 is used in order to get a better rate of convergence of the posterior distribution. When $\gamma = 1$, Assumption C.3 is the classical Hilbert-Schmidt assumption of operator $\Sigma^{-\frac{1}{2}}K\Omega_0^{\frac{1}{2}}$. For $\gamma < 1$ this assumption is more demanding and the posterior rate is faster.

2. Main results.   The posterior distribution of x, denoted with μ_δ^Y, is the Bayesian solution of the inverse problem (1.1), see [14]. The existence of a regular version of the posterior distribution μ_δ^Y, that is, of a transition probability characterizing it, is guaranteed by the fact that $\mathcal{X}$ and $\mathcal{Y}$ are Polish spaces. In fact, any separable Hilbert space is Polish. In many applications $\mathcal{X}$ and $\mathcal{Y}$ are L^2 spaces; L^2 spaces are Polish if they are defined on a separable metric space.

The characterization of the posterior distribution μ_δ^Y of x, together with their joint distribution, is given in the following theorem. The notation $\mathfrak{B}(\mathcal{X}) \otimes \mathfrak{B}(\mathcal{Y})$ means the Borel σ-field generated by the product topology.

THEOREM 2.1. *Let x and $\hat{Y}|x$ be two gaussian H-r.v. with means x_0 and Kx and covariance operators $\frac{1}{g}\Omega_0 := \frac{1}{g}L^{-2s}$ and $\delta\Sigma$, respectively, as in (1.5) and (1.2). Then,*

   (i) *$(x, \hat{Y})$ is a measurable map from $(S, \mathcal{S}, \mathbb{P})$ to $(\mathcal{X} \times \mathcal{Y}, \mathfrak{B}(\mathcal{X}) \otimes \mathfrak{B}(\mathcal{Y}))$ and it has a Gaussian distribution: $(x, \hat{Y})|g, s \sim \mathcal{N}((x_0, Kx_0)', \Upsilon)$, where Υ is a covariance operator defined as $\Upsilon(\varphi, \psi)' = (\frac{1}{g}\Omega_0\varphi + \frac{1}{g}\Omega_0 K^*\psi, \frac{1}{g}K\Omega_0\varphi + (\delta\Sigma + \frac{1}{g}K\Omega_0 K^*)\psi)'$ for all $(\varphi, \psi) \in \mathcal{X} \times \mathcal{Y}$. Then, the marginal sampling distribution is $\hat{Y}|g, s \sim \mathcal{N}(Kx_0, (\delta\Sigma + \frac{1}{g}K\Omega_0 K^*))$.*

   (ii) *The conditional distribution μ_δ^Y of x given $\hat{Y}$ is gaussian with mean $\mathbf{E}(x|\hat{Y}, g, s) = A(\hat{Y} - Kx_0) + x_0$ and covariance operator $\frac{1}{g}[\Omega_0 - AK\Omega_0]$, where the operator $A : \mathcal{Y} \rightarrow \mathcal{X}$ is $\varphi \in \mathcal{Y} \mapsto A\varphi := \Omega_0 K^*(\alpha\Sigma + K\Omega_0 K^*)^{-1}\varphi$ and $\alpha = \delta g$.*

   (iii) *Under Assumption C.1, the linear operator $A : \mathcal{Y} \rightarrow \mathcal{X}$ satisfies*

$$||A\psi|| \leq C||\psi||, \quad \forall\psi \in \mathcal{Y}$$

   for a positive constant C. That is, A is continuous on the whole space $\mathcal{Y}$ as it can be written $A = \Omega_0^{\frac{1}{2}}(\alpha I + B^*B)^{-1}(\Sigma^{-\frac{1}{2}}B)^*$ with $B = \Sigma^{-\frac{1}{2}}K\Omega_0^{\frac{1}{2}}$.

Points *(i)* and *(ii)* in the theorem are slight modifications of Theorem 1 in [12] and Corollary 2 in [29], respectively. We rewrite them in order to have a self-containing theorem. The operator A, written as in point *(ii)*, is in general continuous only on $\mathcal{R}(\alpha\Sigma + K\Omega_0 K^*)$. Hence, it is not continuous in $\hat{Y}$ because a gaussian H-r.v. with covariance operator $\Sigma_y := \delta\Sigma + \frac{1}{g}K\Omega_0 K^*$ does not belong to $\mathcal{R}(\Sigma_y)^{\frac{1}{2}}$ with probability 1 and then $\hat{Y} \notin \mathcal{R}(\Sigma_y)$, since $\mathcal{R}(\Sigma_y) \subset \mathcal{R}(\Sigma_y)^{\frac{1}{2}}$. One of our important contributions is given in point *(iii)* of Theorem 2.1: when a compatibility assumption among K, Σ and Ω_0 holds as stated in Assumption C.1, the operator A is a continuous and bounded operator on the whole $\mathcal{Y}$.

More clearly, the operator A written as in *(iii)* looks like a Tikhonov regularization of the unbounded inverse of B^*B. The inverse of operator B^*B is unbounded and not defined everywhere in $\mathcal{Y}$ because, as B^*B is compact, its spectrum accumulates at zero. The operator αI, introduced by the prior-to-posterior transformation, translates the eigenvalues of B^*B sufficiently far from zero or, equivalently, extends the range of B^*B to the whole space $\mathcal{Y}$ which is the same result of a Tikhonov regularization. In other words, when Assumption C.1 holds, the prior-to-posterior transformation is equivalent to apply a Tikhonov regularization scheme to the inverse of B^*B, *i.e.* to regularize the solution of the equation $B\varphi = r$, $\varphi \in \mathcal{Y}$ and $r \in \mathcal{X}$. Therefore, the

regularization effect of the prior-to-posterior transformation is not typical of all the Gaussian prior distributions in infinite dimensional spaces: a compatibility between the prior covariance operator and the sampling model must exist in order for the prior distribution to be able to regularize.

REMARK 2.1. As stressed in 1.1, the posterior mean can be interpreted as the Tikhonov regularized solution in the Hilbert scale $\mathcal{X}_s$ induced by L^s. Take for simplicity $x_0 = 0$, then

$$
\begin{aligned}
\mathbf{E}(x|\hat{Y}, g, s) &= A\hat{Y} = L^{-s}(\alpha I + L^{-s}K^*\Sigma^{-1}KL^{-s})^{-1}L^{-s}K^*\Sigma^{-\frac{1}{2}}\Sigma^{-\frac{1}{2}}\hat{Y} \\
&= (\alpha L^{2s} + \tilde{B}^*\tilde{B})^{-1}\tilde{B}^*\Sigma^{-\frac{1}{2}}\hat{Y}
\end{aligned}
$$

which is the regularized solution of the model

$$
\Sigma^{-\frac{1}{2}}\hat{Y} = \tilde{B}x + \Sigma^{-\frac{1}{2}}U.
$$

This model is the transformation of (1.1) through operator $\Sigma^{-\frac{1}{2}}$. We remark that there is no reason why the quantities $Z := \Sigma^{-\frac{1}{2}}\hat{Y}$ and $\Sigma^{-\frac{1}{2}}U$ exist, so that this specification is incorrect if we interpret Z as a Hilbert-space valued random variable. However, this transformed model makes sense if we interpret Z as a Hilbert space process in $\mathcal{Y}$, that is, we assume that $Z_\varphi := <Z, \varphi>: \mathcal{Y} \to L^2(S, \mathcal{S}, \mathbb{P})$ is a random variable with $\mathbf{E}(Z_\varphi) = 0$ and $Cov_Z = I$, where $Cov_Z : \mathcal{Y} \to \mathcal{Y}$ is the covariance operator characterized by $< Cov_Z\varphi, \psi > = \mathbf{E}(Z_\varphi Z_\varphi)$, $\forall \varphi, \psi \in \mathcal{Y}$. This assumption is quite usual in inverse problem literature, see for instance [3] and [1], and it requires that only the inner product of the noise be well-defined, which allows the covariance operator to be the identity on $\mathcal{Y}$.

REMARK 2.2. We could generalize the specification of the prior covariance operator to the form $\frac{1}{g}\Omega_0 = \frac{1}{g}QL^{-2s}Q^*$, for some bounded operator Q not necessarily compact. Then, the previous case is a particular case of this one for $Q = I$. In this setting: operator A takes the form

$$
A = QL^{-s}(\alpha I + B_Q^*B_Q)^{-1}(\Sigma^{-\frac{1}{2}}B_Q)^*,
$$

for $B_Q = \Sigma^{-\frac{1}{2}}KQL^{-s}$ and L^s is the Hilbert Scale for $\Sigma^{-\frac{1}{2}}KQ$. Assumptions A.1 and C.1 are replaced by the weaker assumptions $\mathcal{R}(KQ) \subset \mathcal{D}(\Sigma^{-\frac{1}{2}})$ and $\mathcal{R}(KQL^{-s}) \subset \mathcal{D}(\Sigma^{-1})$, respectively. Assumption B is replaced by the assumption that there exists $\tilde{\rho}_* \in \mathcal{X}$ such that $(x_* - x_0) = QL^{-\beta}\tilde{\rho}_*$.

2.1. *Asymptotic analysis.*   We analyze in this subsection frequentist asymptotic properties of the posterior distribution μ_δ^Y of x characterized in Theorem 2.1. The asymptotic analysis is for $\delta \downarrow 0$. Let P^{x_*} denote the sampling

distribution (1.2) with $x = x_*$, we remind the definition of *posterior consistency* in the case the parameter space is a separable metric space:

DEFINITION 1.   *The posterior distribution is* consistent *at x_* with respect to P^{x_*} if it weakly converges towards the Dirac measure δ_{x_*} on x_*, i.e. if for every bounded and continuous functional $a : \mathcal{X} \to \mathcal{X}$*

$$|| \int_{\mathcal{X}} a(x) \mu_\delta^Y (dx) - \int_{\mathcal{X}} a(x) \delta_{x_*}(dx)|| \to 0, \quad a.s. P^{x_*}$$

as $\delta \to 0$, see [10], [16].

If posterior consistency is verified, then the estimators constructed from μ_δ^Y can be used from both bayesian and classical statisticians.

Our asymptotic analysis is organized as follows. First, we propose the posterior mean $\mathbf{E}(x|\hat{Y}, g, s)$ as an estimator for the solution of (1.1) and we analyze its frequentist consistency and give the rate of convergence of the associated risk. Second, we study the asymptotic behavior of the posterior variance. Last, we state posterior consistency and recover the fastest rate of contraction of the posterior distribution.

To short the notation we denote the posterior mean as

(2.1) $$\hat{x}_\alpha := \mathbf{E}(x|\hat{Y}, \alpha, s) \equiv \mathbf{E}(x|\hat{Y}, g, s), \quad \alpha := \delta g$$

and we index it to $\alpha \in \Lambda$ instead of to g since the two are equivalent. The set Λ is of the type $(0, \bar{\alpha}]$, for a constant $\bar{\alpha} > 0$ not too large. As α moves in the set Λ, $\hat{x}_\alpha$ describes a class of estimators. Hereafter, $R(\alpha, x_*) := \mathbf{E}_{x_*}||\hat{x}_\alpha - x_*||^2$ denote the risk associated to our estimator and $\mathbf{E}_{x_*}$ denotes the expectation taken with respect to P^{x_*}. Furthermore, we denote with $\mathcal{X}_\beta$ the ellipsoid of the type

(2.2) $$\mathcal{X}_\beta(R) := \{\varphi : \varphi \in \mathcal{X} \text{ and } ||\varphi||_\beta^2 \le R\}.$$

Our asymptotic results will be valid uniformly on $\mathcal{X}_\beta(R)$. The following theorem gives the asymptotic behavior of $\hat{x}_\alpha$.

THEOREM 2.2.   *Let consider the observational model (1.1) with $U \sim (0, \delta\Sigma)$; let x_* be the true value of x having generated the data and $\alpha = \delta g$. Under Assumptions A.1, A.2, B and C.1-C.3 the MISE associated to $\hat{x}_\alpha$ is of order*

$$\mathbf{E}_{x_*}||\hat{x}_\alpha - x_*||^2 = \mathcal{O}\Big(\alpha^{\frac{\beta}{a+s}} + \delta\alpha^{-\frac{\gamma(a+s)+a}{a+s}}\Big).$$

Then, if α is such that $\delta\alpha^{-\frac{\gamma(a+s)+a}{a+s}} \downarrow 0$, $\hat{x}_\alpha$ is consistent in the sense that $||\hat{x}_\alpha - x_*||^2 \to 0$ in P^{x_*}-probability.

Moreover, if $(x_* - x_0) \in \mathcal{X}_\beta(R)$ and $\alpha = k_1\delta^{\frac{a+s}{\beta+a+\gamma(a+s)}}$, for some constant $k_1 > 0$, then

$$\lim_{\delta\to 0} \sup_{(x_*-x_0)\in\mathcal{X}_\beta(R)} \delta^{-\frac{\beta}{\beta+a+\gamma(a+s)}} \mathbf{E}_{x_*}||\hat{x}_\alpha - x_*||^2 < \infty.$$

The value $\alpha = k_1\delta^{\frac{a+s}{\beta+a+\gamma(a+s)}}$, given in the last part of the theorem, is the optimal one in the sense that it minimizes the upper bound of the risk associated to $\hat{x}_\alpha$. We denote this value with α^{up}, i.e. $\alpha^{up} = k_1\delta^{\frac{a+s}{\beta+a+\gamma(a+s)}}$. The corresponding optimal value g^{up} for g is obtained through the relationship $\alpha = \delta g$:

$$g^{up} = \alpha^{up}\delta^{-1} = k_1\delta^{-\frac{\beta-s+\gamma(a+s)}{\beta+a+\gamma(a+s)}},$$

with k_1 some constant. By construction, $g^{up} \uparrow \infty$ slower than δ^{-1} and faster than $\delta^{-\frac{d-1}{d}}$, with $d = \frac{\gamma(a+s)+a}{a+s}$.

We stress an important fact: the optimal rate $\delta^{\frac{\beta}{\beta+a+\gamma(a+s)}}$ of the MISE only depends on the degree of ill-posedness a and the smoothness parameter β. It does not depend on the parameter s of the prior covariance because the parameter γ, as it is determined by the spectrum of B, is a function s. Hence, the s in $\gamma(a+s)$ simplify. This will result clear in Corollary 2 where the value of γ will be explicitly computed. As s does not affect the asymptotic properties of our estimator it does not need to be estimated in an adaptive way; it simply has to be specified on the base of the prior guess of the statistician.

Assumption B is particularly suitable because it allows to express the regularity of $(x_* - x_0)$ according to the spectral representation of the prior covariance operator. However, $\hat{x}_\alpha$ is consistent at a certain rate even when Assumption B is violated and $(x_* - x_0) \notin \mathcal{H}(\Omega_0)$ are violated, that is, when $(x_* - x_0)$ is less regular. In this case, the rate can be obtained by expressing the smoothness (regularity) of $(x_* - x_0)$ according to the spectral representation of the operator B. This is the result of the following Corollary.

COROLLARY 1.   *Under Assumptions A.1, A.2, C.1-C.3, if $\Omega_0^{\frac{1}{2}}$ and $B^*\Sigma^{-\frac{1}{2}}K$ commute and if $(x_* - x_0) \in \mathcal{R}[(B^*B)^{\frac{\eta}{2}}]$, for some $\eta > 0$, then*

$$\mathbf{E}_{x_*}||\hat{x}_\alpha - x_*||^2 = \mathcal{O}(\alpha^\eta + \delta\alpha^{-\frac{\gamma(a+s)+a}{a+s}}).$$

Then, if α is such that $\delta\alpha^{-\frac{\gamma(a+s)+a}{a+s}} \downarrow 0$, $\hat{x}_\alpha$ is consistent in the sense that $||\hat{x}_\alpha - x_*||^2 \to 0$ in P^{x_*}-probability.

Because of Assumption A.2 and the result of Corollary 3 in Section 4, which is due to [11], $\mathcal{R}[(B^*B)^{\frac{\eta}{2}}] \subset \mathcal{H}(\Omega_0)$ if and only if $\eta > \frac{s}{a+s}$ and $\mathcal{R}[(B^*B)^{\frac{\eta}{2}}] = \mathcal{R}(\Omega_0^{\frac{\beta}{2s}})$ if and only if $\eta = \frac{\beta}{2(a+s)}$. Thus, the regularity assumption $(x_* - x_0) \in \mathcal{R}[(B^*B)^{\frac{\eta}{2}}]$ implies $(x_* - x_0) \in \mathcal{H}(\Omega_0)$ if and only if $\eta > \frac{s}{a+s}$; if $0 < \eta < \frac{s}{a+s}$, then Corollary 1 shows that we have a certain rate even when the true mean is not in the $\mathcal{R.K.H.S.}(\Omega_0)$.

REMARK 2.3 Corollary 1 is related to a fixed x_*. However, it may be immediately deduced from the proof that the result is true uniformly on $\Phi_\eta(M)$:

$$\sup_{(x_*-x_0)\in\Phi_\eta(M)} \mathbf{E}_{x_*}||\hat{x}_\alpha - x_*||^2 = \mathcal{O}(\alpha^\eta + \delta\alpha^{-\frac{\gamma(a+s)+a}{a+s}}).$$

where $\Phi_\eta(M) := \{\varphi \in \mathcal{X}; \sum_{j=1}^\infty \lambda_j^{-2\eta} < \varphi, \varphi_j >^2 \le M\}$ and $\{\lambda_j, \varphi_j\}$ is the eigensystem of B^*B.

The asymptotic behavior of the posterior variance is given in the following theorem. The rate given is the rate of the MISE of the posterior variance.

THEOREM 2.3. *Let consider the observational model (1.1) with $U \sim (0, \delta\Sigma)$ and $\alpha = \delta g$. Under Assumptions A.1, A.2 and C.1-C.3 the posterior variance of x converges to zero in $\mathcal{X}$-norm: $||Var(x|\hat{Y}, g, s)\phi|| \to 0, \forall\phi \in \mathcal{X}$. Furthermore, for $Var(x|\hat{Y}, g, s)|_{\mathcal{X}_\beta} : \mathcal{X}_\beta \to \mathcal{X}$, we have*

$$||Var(x|\hat{Y}, g, s)|_{\mathcal{X}_\beta}\phi||^2 = \mathcal{O}\Big(\frac{1}{g^2}\alpha^{\frac{\beta+2s}{a+s}}\Big).$$

When the optimal α^{up} and g^{up} are used, the posterior variance converges at the rate $\delta^{\frac{3\beta+2\gamma(a+s)}{\beta+a+\gamma(a+s)}}$ which is faster than the optimal rate $\delta^{\frac{\beta}{\beta+a+\gamma(a+s)}}$ of the posterior mean.

The rate of convergence depends on the subset to which we restrict the domain of the posterior covariance operator. If its domain is restricted to a bigger space than $\mathcal{X}_\beta$, then the rate will be slower.

A simple Chebishev's inequality allows to conclude that convergence of the posterior distribution towards the point mass δ_{x_*} is implied by the convergence of the posterior mean towards x_* and of the trace of the posterior variance towards 0. We state consistency of the posterior distribution in the theorem below.

THEOREM 2.4. *Let the assumptions of theorem 2.2 be satisfied. For any sequence $M_\delta \to \infty$ and $\varepsilon_\delta > 0$ small enough, the posterior probability, with $\alpha = \alpha^{up}$, satisfies*

$$\mu_\delta^Y \{x \in \mathcal{X} : \delta^{-\frac{\beta}{2(\beta+a+\gamma(a+s))}} ||x - x_*|| > M_\delta\} \to 0$$

in P^{x_*}-probability as $\delta \to 0$.

The rate of contraction $\varepsilon_\delta = \delta^{\frac{\beta}{2(\beta+a+\gamma(a+s))}}$ is the fastest one. This theorem shows that we are able to avoid the posterior inconsistency which is stressed in [10] as a typical aspect of Bayesian nonparametric estimation. This is possible because the prior distribution introduces a kind of regularization of the posterior mean due to the fact that it is shrinking. The parameter of regularization α is linked to the parameter g in the prior distribution of x.

2.2. *Operators with geometric spectra.* Of particular interest is the case where the operators K, Σ and L have geometric spectra (see EXAMPLE 1). In this case, we show that the rate of convergence given in Theorem 2.2 is optimal in a minimax sense and $\alpha_{opt} \propto \alpha^{up}$. Our estimator $\hat{x}_\alpha$ with $\alpha = \alpha_{opt}$ is optimal in the sense that it attains the optimal rate of convergence. Moreover, in this framework we show that $\hat{x}_\alpha$, with α estimated with the method proposed in Section 3, is sharp adaptive in a minimax sense on the classes of ellipsoids.
We denote with λ_j^K the singular values of K and with λ_j^Σ and λ_j^L the eigenvalues of Σ and L, respectively. We assume that they decrease as a power of j:

ASSUMPTION D. *There exist $a_0, c_0 \geq 0$, $\underline{a}, \bar{a}, \underline{c}, \bar{c}, \underline{l}, \bar{l} > 0$ such that $\underline{a}j^{-a_0} \leq \lambda_j^K \leq \bar{a}j^{-a_0}$, $\underline{c}j^{-c_0} \leq \lambda_j^\Sigma \leq \bar{c}j^{-c_0}$ and $\underline{l}j \leq \lambda_j^L \leq \bar{l}j$, $j = 1, 2, \ldots$*

Under Assumption D we rewrite Assumption B and the ellipsoid $\mathcal{X}_\beta(R)$ respectively as

$$(x_* - x_0) \in \mathcal{X}_\beta := \{\varphi \in \mathcal{X} : \sum_{j=1}^{\infty} j^{2\beta} < \varphi, \psi_j >^2 < \infty\}$$

$$\mathcal{X}_\beta(R) := \{\varphi \in \mathcal{X} : \sum_{j=1}^{\infty} j^{2\beta} < \varphi, \psi_j >^2 \leq R\}.$$

$\mathcal{X}_\beta(R)$ is the usual ellipsoid encountered in inverse problem literature, see *e.g.* [2], [7] or [45]. Assumption D is standard in literature. Under this assumption, Assumption B implies that the Fourier coefficients $< (x_* -$

$x_0, \psi_j) >$ decrease at the geometric rate j^{-b_0} for $b_0 > \frac{1}{2} + \beta$, that is, there exist $\underline{x}, \bar{x} > 0$ and $b_0 > \frac{1}{2} + \beta$ such that $\underline{x} j^{-b_0} \leq | < (x_* - x_0), \psi_j > | \leq \bar{x} j^{-b_0}$, for $\{\psi_j\}$ the eigenfunctions of B^*B.

The geometric spectra case makes evident how Assumption B encompasses the assumption for instance in [42] (Theorem 4.1) and [17] (Assumption A3). Nevertheless, our Assumption B is more general because it allows the important case of exponentially declining eigenvalues which arises with an analytic kernel of K.

Under Assumption D, Assumptions A.1, A.2 and C.1-C.3 imply the following conditions on the coefficients:

(2.3) $$a = a_0 - \frac{c_0}{2} > 0, \qquad a_0 \geq c_0 - s,$$

(2.4) $$s > \frac{1}{2}, \qquad \gamma = \frac{1}{2(a_0 + s) - c_0}$$

where the last equality is up to an additive small term $\varepsilon > 0$. The following Corollary to Theorem 2.2 gives the minimax rate attained by $\hat{x}_\alpha$.

COROLLARY 2. *Let Assumptions B, D, (2.3) and (2.4) hold. Then, the Bayes estimator $\hat{x}_\alpha$, with $\alpha = k_2 \delta^{\frac{a+s}{\beta+a+\frac{1}{2}}}$ and $k_2 = (\frac{2a+1}{2\beta} \frac{c_2}{c_1})^{\frac{a+s}{\beta+a+\frac{1}{2}}}$ attains the optimal (minimax) rate $\delta^{\frac{\beta}{\beta+a+\frac{1}{2}}}$, that is,*

(2.5) $$\inf_\alpha \sup_{(x_* - x_0) \in \mathcal{X}_\beta(R)} \mathbf{E}_{x_*} ||\hat{x}_\alpha - x_*||^2 = C(R, \beta) \delta^{\frac{\beta}{\beta+a+\frac{1}{2}}},$$

where

$$C(R, \beta) = \Big[\frac{1+2a}{2\beta} \frac{c_2}{c_1} \Big]^{\frac{\beta}{\beta+a+\frac{1}{2}}} c_1 \Big(1 + \frac{2\beta}{2a+1} \Big)$$

with $c_1 := c_1(R, \beta)$ and c_2 two constants.

The value of $\alpha = k_2 \delta^{\frac{a+s}{\beta+a+\frac{1}{2}}}$ given in the previous Corollary is the optimal one in the sense that it minimizes the risk associated with $\hat{x}_\alpha$ and we denote it as $\alpha^{opt} := k_2 \delta^{\frac{a+s}{\beta+a+\frac{1}{2}}}$. Then, α_{opt} is solution of the minimization problem

(2.6) $$\alpha_{opt} = \arg\min_{\alpha \in \Lambda} R(\alpha, x_*).$$

The estimator $\hat{x}_{opt}$ computed with this value of α is called the *oracle*.

The rate $\delta^{\frac{\beta}{\beta+a+\frac{1}{2}}}$ is the same as the rate given in Theorem 2.2: Corollary 2

shows that this rate is minimax. This rate does not depend on the hyperparameter s, as previously discussed and it is the same as the rate in [7] and in [8] for the mildly ill-posed case. Furthermore, for the direct problem, *i.e.* for $a = 0$, it is the same as in [45].

The rate of contraction of the posterior distribution in the geometric spectra case is $\varepsilon_\delta = \delta^{\frac{\beta}{2(\beta+a)+1}}$. For the direct problem, it is the same rate of contraction given in Theorem 2.1 of [2].

2.3. *Covariance operators proportional to K.* Another particular case, encountered in many applications, is when the covariance operator is of the form $\Sigma = (KK^*)^r$, for some $r \in \mathbb{R}_+$ (see for instance EXAMPLE 2 and EXAMPLE 3 in Section 1.1 and EXAMPLE 4 below). In this case it is convenient to choose $L = (K^*K)^{-\frac{1}{2}}$, *i.e.* L is the canonical Hilbert Scale, and $\Omega_0 = (K^*K)^s$, for $s \in \mathbb{R}_+$.

Assumption A.1 and A.2 hold for $r \leq 1$ and $a = 1 - r$. Assumption C.1 holds for $s \geq 2r - 1$. We have to check that $(KK^*)^r$ and $(K^*K)^s$ are trace-class. We remark that K Hilbert-Schmidt does not imply that Σ and Ω_0 are trace-class and Assumption C.3 implies that Σ and Ω_0 are trace-class only if γ is such that $\gamma < \frac{r \wedge s}{s+1-r}$.

The asymptotic results of subsection 2.1 trivially apply to this particular case.

Next, we develop the example of *functional regression estimation* which is an application where the covariance operator Σ is proportional to K.

EXAMPLE 4. *Functional Regression Estimation.* The model is the following:

$$(2.7) \qquad Y = \int_0^1 h(s)X(s)ds + U, \quad \mathbf{E}(UX(s)) = 0, \quad X, h \in L^2([0,1])$$

and $U|\tau^2 \sim \mathcal{N}(0, \tau^2)$. We want to recover h. Assuming that X is a centered random variable, the most popular approach consists in multiply both sides of (2.7) by $X(s)$ and then take the expectation: $\mathbf{E}(YX(t)) = \int_0^1 h(s)Cov(X(s), X(t))ds$, for $t \in [0,1]$. This model has been studied in [18], among others. If we dispose of independent and identically distributed data $(Y_1, X_1), \ldots, (Y_n, X_n)$ we can estimate the unknown moments in the previous equation as

$$\frac{1}{n}\sum_i Y_i X_i(t) = \frac{1}{n}\sum_i < X_i, h > X_i(t) + \frac{1}{n}\sum_i U_i X_i(t),$$

where $< \cdot, \cdot >$ denotes the inner product in $L^2([0,1])$. Then, the operator K takes the form $\varphi \in L^2([0,1]) \mapsto K\varphi := \frac{1}{n}\sum_i < X_i, \varphi > X_i(t)$ and

$\delta\Sigma = \frac{\tau^2}{n}K$, so that $r = \frac{1}{2}$. Suppose that the spectrum of K has a geometric decline rate, as in subsection 2.2, and that it is the same as in [18], then $j^{-a_0} = j^{-c_0}$ and $a_0 = c_0$. If the true h satisfies Assumptions B and ??, the rate of convergence of the MISE of the posterior mean is

$$n^{-\frac{2b_0-1}{2b_0+a_0}}$$

which is the same as the minimax rate in [18].

3. An adaptive selection of α through an empirical Bayes approach. The parameter $\alpha = \delta g$ plays the role of a *smoothing* (or regularization) *parameter* and $\{\hat{x}_\alpha\}_{\alpha\in\Lambda}$ defines a class of possible estimators of x. The set Λ is of the type $(0, \bar{\alpha}]$, for a constant $\bar{\alpha} > 0$ not too large. We know that when $\alpha = \alpha_{opt}$ our estimator $\hat{x}_{opt} := \hat{x}_{\alpha_{opt}}$ is optimal in a minimax sense. However, the oracle α_{opt} cannot be computed from the data since it depends on the true value x_* and on β which are unknown. We are then satisfied if we can select a value $\hat{\alpha}$ for α depending on the data and such that $\hat{x}_{\hat{\alpha}}$ mimics the behavior of the oracle $\hat{x}_{opt}$. In order to mimic the oracle, the risk of $\hat{x}_{\hat{\alpha}}$ must be at least as small as the risk of the oracle, for δ small. This property is known as *adaptivity* of our estimator.

We propose an *empirical Bayes* procedure in order to select a data-driven $\hat{\alpha}$; we show, for the *geometric spectra* case of subsection 2.2 that the corresponding estimator $\hat{x}_{\hat{\alpha}}$ is sharp adaptive in a minimax sense. In the literature, adaptive empirical Bayes procedures have been proposed in frameworks different from the our, see for instance [21] and [44] in wavelet estimation.

3.1. *Characterization of the likelihood.* For simplicity we express g and the prior distribution of x in terms of α, so that $g = \alpha/\delta$ and $x \sim \mathcal{N}(x_0, \frac{\delta}{\alpha}\Omega_0)$. Then, α is treated as an hyperparameter. The marginal distribution of $\hat{Y}$, given s and α, is

$$(3.1) \qquad \hat{Y}|s, \alpha \sim P^\alpha, \quad P^\alpha = \mathcal{N}\left(Kx_0, \delta\Sigma + \frac{\delta}{\alpha}K\Omega_0 K^*\right).$$

When α is considered as a random variable, the marginalization of P^x with respect to the prior of x demands the implicit assumption that, conditionally on x, $\hat{Y}$ is independent on α, in symbols $\hat{Y} \perp\!\!\!\perp \alpha\,|x$.

A gaussian measure on infinite dimensional spaces does not have a density function with respect to the Lebesgue measure. Computing its likelihood is therefore not an easy task. In particular, we have to find another measure, different from the Lebesgue measure, with respect to which P^α is absolutely

continuous. The following theorem gives a characterization of such a measure.

THEOREM 3.1. *Let P^0 be a Gaussian measure with mean Kx_0 and covariance operator $\delta\Sigma$, i.e. $P^0 = \mathcal{N}(Kx_0, \delta\Sigma)$. Under Assumptions C.1 and C.3, the Gaussian measure P^α defined in (3.1) is equivalent to P^0. Moreover, the Radon-Nikodym derivative is given by*

$$(3.2) \qquad \frac{dP^\alpha}{dP^0} = \prod_{j=1}^{\infty} \sqrt{\frac{\alpha}{\lambda_j^2 + \alpha}} \, e^{\frac{\lambda_j^2}{2(\lambda_j^2 + \alpha)} z_j^2},$$

with $\frac{\lambda_j^2}{\alpha}$ the eigenvalues of $\frac{1}{\alpha}BB^*$ and z_j a standard normal random variable under P^0.

The measure P^0 is equal to P^α with $\alpha = \infty$. This theorem is an application of Theorem of Kuo (1975) [26]. The random variable z_j is defined under P^0 as $z_j = \frac{<\hat{Y} - Kx_0, \varphi_j>}{\sqrt{\delta}l_j}$, where $\{l_j^2, \varphi_j\}$ is the eigensystem associated with Σ. Thus, (3.2) gives the marginal likelihood of $\hat{Y}$ given α, with respect to P^0.

3.2. *Adaptive Empirical Bayes (EB) procedure.* We specify a non-informative prior distribution on α and select the regularization parameter that maximizes the posterior distribution of α which is proportional to the marginal likelihood of $\hat{Y}$. We define the marginal maximum likelihood estimator $\hat{\alpha}$ of α to be the maximizer of the marginal log-likelihood

$$\bar{S}(\alpha, \hat{Y}) = \frac{1}{2} \sum_{j=1}^{\infty} \left[\log\left(\frac{\alpha}{\alpha + \lambda_j^2}\right) + \frac{\lambda_j^2}{\alpha + \lambda_j^2} \frac{<\hat{Y} - Kx_0, \varphi_j>^2}{\delta l_j^2} \right].$$

In an equivalent way, $\hat{\alpha}$ is defined as the solution of the first order condition $\frac{\partial}{\partial \alpha} \bar{S}(\alpha, \hat{Y}) = 0$:

(3.3)
$$S(\alpha, \hat{Y}) := \frac{\partial}{\partial \alpha} \bar{S}(\alpha, \hat{Y}) = \frac{1}{2}\Big(\sum_{j=1}^{\infty} \frac{\lambda_j^2}{\alpha(\alpha + \lambda_j^2)} - \sum_{j=1}^{\infty} \frac{\lambda_j^2}{\delta(\alpha + \lambda_j^2)^2 l_j^2} <\hat{Y} - Kx_0, \varphi_j>^2 \Big) = 0.$$

To shorten the notation we write $S(\alpha) = S(\alpha, \hat{Y})$. Our strategy will then be to plug the value $\hat{\alpha}$ back into the prior of x and then computing the posterior mean estimator $\hat{x}_{\hat{\alpha}}$ using this value of α. The goal for estimation using EB selection of α is to obtain a risk $R(\hat{x}_{\hat{\alpha}}, x_*)$ associated with $\hat{x}_{\hat{\alpha}}$ sufficiently

small compared to the oracle $R(\hat{x}_{opt}, x_*)$.

We show in the next theorem that $\hat{\alpha}$ is of the same order as α_{opt} and we provide an oracle inequality for $R(\hat{x}_{\hat{\alpha}}, x_*)$. These results are obtained for the case where operators have geometric spectra.

THEOREM 3.2. *Let Assumptions B, D, ??, (2.3) and (2.4) hold. Then, for every $x_* \in \mathcal{X}_\beta$ which satisfies assumption ?? with $b_0 > \beta + \frac{1}{2}$, the estimator $\hat{\alpha}$ defined as the solution of (3.3) converges to 0 at the same rate as α_{opt}, i.e. $\hat{\alpha} \asymp \alpha_{opt}$. Moreover, for the estimator $\hat{x}_{\hat{\alpha}}$ with $\hat{\alpha}$ defined as the solution of (3.3), we have*

$$(3.4) \qquad R(\hat{\alpha}, x_*) \le \inf_{\alpha \in \Lambda} R(\alpha, x_*)(1 + \mathcal{O}(1))$$

uniformly in $x_* \in \mathcal{X}_\beta(R)$.

Inequality (3.4) is an *oracle inequality* on the class $(0, \bar{\alpha}]$ and it implies that our EB estimator $\hat{x}_{\hat{\alpha}}$ is *adaptive* to the oracle $\hat{x}_{opt}$ *in a minimax sense* on the family $\{\mathcal{X}_\beta(R), \beta > 0, R > 0\}$:

$$\sup_{(x_* - x_0) \in \mathcal{X}_\beta(R)} \mathbf{E}_{x_*} ||\hat{x}_{\hat{\alpha}} - x_*||^2 \le C(\beta, R)(1 + \mathcal{O}(1)) \delta^{\frac{\beta}{\beta + a + \frac{1}{2}}}$$

where $C(\beta, R)$ is a finite constant depending on $\beta > 0$ and $R > 0$. Therefore, EB procedure gives a practical rule for selecting a value for α from the data.

REMARK 3.1. Instead of doing EB estimation by plugging $\hat{\alpha}$ in the prior distribution for x we could recover the posterior distribution of $\alpha | \hat{Y}$ and then draw a sample from it through methods like *Rejection Method* or *Metropolis-Hastings*. This sample would be used for integrating out α, by Monte Carlo integration, in the posterior distribution of $x | s, \alpha, \hat{Y}$.

REMARK 3.2. The specification of a prior distribution on the hyperparameter α allowing to obtain a posterior for $\alpha | \hat{Y}$ in closed form is not an easy task. For the finite dimensional linear regression model, a prior in the gamma form has been proposed in [43]. In the Bayesian variable selection problem, different prior specifications for g have been proposed in [28]. For our model (3.2) we propose, as an alternative to the non-informative prior, a natural conjugate prior which, for some parameters $\nu_0, \mu_0 > 0$ and a sequence $(a_j)_j$ with values in $\mathbb{R}$, has kernel

$$\alpha^{-\frac{\nu_0}{2}} [det(\alpha(\alpha I + BB^*)^{-1})]^{\frac{\mu_0 + 1}{2}} \exp\left\{\frac{1}{2} \sum_j \frac{\lambda_j^2}{\alpha + \lambda_j^2} a_j^2\right\}.$$

This prior depends on operator K and we suggest that it could be used for selecting the operator itself (like model selection in finite dimensional regression models, see [28], [15]). Moreover, this prior can be think as the posterior distribution resulting from a "conceptual" sample and a non-informative prior, both the sampling and prior distributions being in the same family. A non-informative prior in the same family requires $a_j = 0$, $\forall j$.

4. Proofs. In all the proofs we use the notation $(\lambda_j, \varphi_j, \psi_j)_j$ to denote the singular value decomposition of B (or equivalently of B^*), that is, $B\psi_j = \lambda_j \varphi_j$ and $B^*\varphi_j = \lambda_j \psi_j$. In order to prove several results we make use of Corollary 8.22 in [11]. We give here a simplified version of it adapted to our framework:

COROLLARY 3. *Let $\mathcal{X}_s$, $s \in \mathbb{R}$ be a Hilbert scale induced by L and let $\Sigma^{-\frac{1}{2}}K :$ $\mathcal{X} \to \mathcal{Y}$ be a bounded operator satisfying Assumption A.2, $\forall x \in \mathcal{X}$ and for some $a > 0$. Then, for $B = \Sigma^{-\frac{1}{2}}KL^{-s}$, $s \geq 0$ and $|\nu| \leq 1$*

$$\underline{c}(\nu)||L^{-\nu(a+s)}x|| \leq ||(B^*B)^{\frac{\nu}{2}}x|| \leq \overline{c}(\nu)||L^{-\nu(a+s)}x||$$

holds on $\mathcal{D}((B^*B)^{\frac{\nu}{2}})$ with $\underline{c}(\nu) = \min(\underline{m}^\nu, \bar{m}^\nu)$ and $\overline{c}(\nu) = \max(\underline{m}^\nu, \bar{m}^\nu)$. Moreover, $\mathcal{R}((B^*B)^{\frac{\nu}{2}}) = \mathcal{X}_{\nu(a+s)} \equiv \mathcal{D}(L^{\nu(a+s)})$, where $(B^*B)^{\frac{\nu}{2}}$ has to be replaced by its extension to $\mathcal{X}$ if $\nu < 0$.

We refer to [11] for the proof of it.

4.1. *Proof of Theorem 2.1.* (i) Because x and $\hat{Y}$ are both Borel measurable, then $(x, \hat{Y})$ is measurable in the product of the Borel σ-field $\mathfrak{B}(\mathcal{X}) \times \mathfrak{B}(\mathcal{Y})$. As $\mathcal{X}$ and $\mathcal{Y}$ are separable, it follows that the product of the Borel σ-fields is the same as the Borel σ-field of the product metric. Then, $(x, \hat{Y})$ is Borel measurable in $\mathfrak{B}(\mathcal{X}) \otimes \mathfrak{B}(\mathcal{Y})$, see Chapter 1.4 in [41].
By construction of $\hat{Y}$ in (1.1) and (1.2) we have $\hat{Y} = y_1 + y_2$, with y_1 a gaussian random variable with values in $\mathcal{R}(K)$ and y_2 a gaussian random variable with values in $\overline{\mathcal{R}.\mathcal{K}.\mathcal{H}.\mathcal{S}.}(\Sigma)$, where $\overline{\mathcal{R}.\mathcal{K}.\mathcal{H}.\mathcal{S}.}(\Sigma)$ denotes the closure in $\mathcal{Y}$ of the reproducing kernel Hilbert space associated with Σ. Remark that because Σ is one-to-one the support of y_2 is all $\mathcal{Y}$. The random elements y_1 and y_2 are independent and there exists a gaussian random variable with values in $\mathcal{D}(K) \subset \mathcal{X}$ such that $y_1 = Kx$. Let $< \cdot, \cdot >_{\mathcal{X}\mathcal{Y}}$ denote the inner product in $\mathcal{X} \times \mathcal{Y}$. For all $(\varphi, \psi) \in \mathcal{X} \times \mathcal{Y}$ we have

$$< (x, \hat{Y}), (\varphi, \psi) >_{\mathcal{X}\mathcal{Y}} = < x, \varphi > + < Kx + y_2, \psi > = < x, \varphi + K^*\psi > + < y_2, \psi >$$

that is gaussian since it is the sum of two gaussian random variables. Since this holds for all $(\varphi, \psi) \in \mathcal{X} \times \mathcal{Y}$, this proves that $(x, \hat{Y})$ is jointly gaussian. Its mean is easily computable as $< \mathbf{E}(x, \hat{Y}), (\varphi, \psi) > = < x_0, \varphi > + < Kx_0, \psi > = < (x_0, Kx_0)(\varphi, \psi) >$ and the covariance operator Υ is such that

$$< \Upsilon(\varphi, \psi), (\varphi, \psi) >_{\mathcal{X}\mathcal{Y}} = Cov(< (x, \hat{Y}), (\varphi, \psi) >_{\mathcal{X}\mathcal{Y}}, < (x, \hat{Y}), (\varphi, \psi) >_{\mathcal{X}\mathcal{Y}}).$$

Then,

$$
\begin{aligned}
<\Upsilon(\varphi,\psi),(\varphi,\psi)>_{\mathcal{X}\mathcal{Y}} &= Cov(<x,\varphi+K^*\psi>+<y_2,\psi>,<x,\varphi+K^*\psi>+<y_2,\psi>) \\
&= <\frac{1}{g}\Omega_0[\varphi+K^*\psi],\varphi+K^*\psi>+<\delta\Sigma\psi,\psi> \\
&= <\left(\begin{array}{cc}\frac{1}{g}\Omega_0, & \frac{1}{g}\Omega_0 K^* \\ \frac{1}{g}\Omega_0 K, & (\delta\Sigma+\frac{1}{g}K\Omega_0 K^*)\end{array}\right)(\varphi,\psi)',(\varphi,\psi)>_{\mathcal{X}\mathcal{Y}}
\end{aligned}
$$

for all $(\varphi,\psi)\in\mathcal{X}\times\mathcal{Y}$, where Cov denotes the covariance operator. Let denote this joint distribution with Π and the marginal distribution of $\hat{Y}|g,s$ with Q. The distribution Q is the projection of Π on $(\mathcal{Y},\mathfrak{B}(\mathcal{Y}))$. Since Π is gaussian, the projection Q must be gaussian with mean function m_Q and covariance operator R_Q. Moreover, $\forall\psi\in\mathcal{Y}$, $<m_Q,\psi>=<(x_0,Kx_0),(0,\psi)>=<Kx_0,\psi>$ and

$$
\begin{aligned}
<R_Q\psi,\psi> &= <\Upsilon(0,\psi),(0,\psi)>=<(\frac{1}{g}\Omega_0 0+\frac{1}{g}\Omega_0 K^*\psi,(\delta\Sigma+\frac{1}{g}K\Omega_0 K^*)\psi+\frac{1}{g}K\Omega_0 0),(0,\psi)> \\
&= <(\delta\Sigma+\frac{1}{g}K\Omega_0 K^*)\psi,\psi>.
\end{aligned}
$$

Hence, $m_Q=Kx_0$ and $R_Q=(\delta\Sigma+\frac{1}{g}K\Omega_0 K^*)$.

(ii) First, the posterior mean $\mathbf{E}(x|\hat{Y})$ is an affine transformation of $\hat{Y}$: $\mathbf{E}(x|\hat{Y})=A\hat{Y}+b$ for some linear operator $A:\mathcal{Y}\to\mathcal{X}$ and a function $b\in\mathcal{X}$. We refer to [29] Section 3.4, page 392 for the proof of it. The modification required by the fact that we have a non null prior mean is trivial.

To prove that μ_δ^Y is gaussian we use the characteristic function of x given $\hat{Y}$ (this is a slight modification and extension of Section 3.5, page 392 in Mandelbaum (1984)). Let $z:=x-\mathbf{E}(x|\hat{Y})=x-(A\hat{Y}+b)$. By definition of linear conditional expectation, z and $\hat{Y}$ are independent. Then, $\forall\varphi\in\mathcal{X}$

$$
\begin{aligned}
\mathbf{E}(e^{i<x,\varphi>}|\hat{Y}) &= \mathbf{E}(e^{i<z,\varphi>+i<(A\hat{Y}+b),\varphi>}|\hat{Y})=\mathbf{E}(e^{i<z,\varphi>})e^{i<(A\hat{Y}+b),\varphi>} \\
&= e^{i<(A\hat{Y}+b),\varphi>}\mathbf{E}(e^{i<x,\varphi>}\mathbf{E}_x(e^{-i<(A+b)\hat{Y},\varphi>}|x)) \\
&= e^{i<A\hat{Y},\varphi>}\mathbf{E}(e^{i<(I-AK)x,\varphi>})e^{-\frac{1}{2}<A\delta\Sigma A^*\varphi,\varphi>} \\
&= e^{i<A\hat{Y},\varphi>}e^{i<(I-AK)x_0,\varphi>}e^{-\frac{1}{2}<[\frac{1}{g}(I-AK)\Omega_0(I-K^*A^*)+A\delta\Sigma A^*]\varphi,\varphi>} \\
&= e^{i<(A\hat{Y}+(I-AK)x_0),\varphi>-\frac{1}{2}<V\varphi,\varphi>}.
\end{aligned}
$$
(4.1)

The relation (4.1) identifies the conditional distribution $\mu^{\hat{Y}}$ as being Gaussian with mean $A\hat{Y}+b$, $b=(I-AK)x_0$ and covariance $V=\frac{1}{g}(I-AK)\Omega_0(I-K^*A^*)+A\delta\Sigma A^*$. We now have to identifies the operator A. Let consider the definition of covariance operator: for all $(\varphi,\psi)\in\mathcal{X}\times\mathcal{Y}$ we have

$$
<Cov(\hat{Y},x)\varphi,\psi> = <Cov(\hat{Y},\mathbf{E}(x|\hat{Y}))\varphi,\psi>=<(\delta\Sigma+\frac{1}{g}K\Omega_0 K^*)A^*\varphi,\psi>
$$

and moreover we have that $<Cov(\hat{Y},x)\varphi,\psi>=<K\Omega_0\varphi,\psi>$. Then, A is the solution of

$$
\left(\delta\Sigma+\frac{1}{g}K\Omega_0 K^*\right)A^*\varphi=\frac{1}{g}K\Omega_0\varphi,\qquad\forall\varphi\in\mathcal{X}
$$
(4.2)

and we conclude that $A = \Omega_0 K^*(\alpha\Sigma + K\Omega_0 K^*)^{-1}$ with $\alpha = \delta g$. We substitute back this value of A in V so that we obtain $Cov(x|\hat{Y}) = \frac{1}{g}[\Omega_0 - AK\Omega_0]$.

(iii) We transform equation (4.2) in the following way

$$
\begin{aligned}
(\alpha\Sigma + K\Omega_0 K^*)A^* &= K\Omega_0 \\
\Leftrightarrow \quad \Sigma^{\frac{1}{2}}(\alpha I + \Sigma^{-\frac{1}{2}} K\Omega_0 K^* \Sigma^{-\frac{1}{2}})\Sigma^{\frac{1}{2}} A^* &= K\Omega_0 \\
\Leftrightarrow \quad (\alpha I + BB^*)\Sigma^{\frac{1}{2}} A^* &= B\Omega_0^{\frac{1}{2}} \\
\Leftrightarrow \quad \Sigma^{\frac{1}{2}} A^* &= (\alpha I + BB^*)^{-1} B\Omega_0^{\frac{1}{2}} \\
\Leftrightarrow \quad \Sigma^{\frac{1}{2}} A^* &= B(\alpha I + B^*B)^{-1}\Omega_0^{\frac{1}{2}} \\
\Leftrightarrow \quad A^* &= \Sigma^{-\frac{1}{2}} B(\alpha I + B^*B)^{-1}\Omega_0^{\frac{1}{2}}.
\end{aligned}
$$

The computations above are justified under Assumption C.1. This allows to obtain an alternative expression for A:

$$
(4.3) \qquad A = \Omega_0^{\frac{1}{2}}(\alpha I + B^*B)^{-1}(\Sigma^{-\frac{1}{2}} B)^*.
$$

The identity operator I extends the range of B^*B to the whole $\mathcal{Y}$ and then A is continuous and defined everywhere.

4.2. *Proof of Theorem 2.2.* The difference $(\hat{x}_\alpha - x_*)$ is re-written as

$$
\hat{x}_\alpha - x_* = -(I - AK)(x_* - x_0) + AU := \mathcal{C}_1 + \mathcal{C}_2,
$$

where A is defined as in Theorem 2.1. We consider the MISE associated to $\hat{x}_\alpha$: $\mathbf{E}_{x_*}||\hat{x}_\alpha - x_*||^2 = ||\mathcal{C}_1||^2 + \mathbf{E}_{x_*}||\mathcal{C}_2||^2$; by Markov inequality this is enough in order to show convergence to 0 in P^{x_*}-probability of the estimation error: $||\hat{x}_\alpha - x_*||^2 = \mathcal{O}(\mathbf{E}_{x_*}||\hat{x}_\alpha - x_*||^2)$. By Assumption B there exists a $\rho_* \in \mathcal{X}$ such that $(x_* - x_0) = L^{-\beta}\rho_*$, then

$$
\begin{aligned}
||\mathcal{C}_1||^2 &= ||[I - \Omega_0^{\frac{1}{2}}(\alpha I + B^*B)^{-1}(\Sigma^{-\frac{1}{2}} B)^* K]L^{-\beta}\rho_*||^2 \\
&= ||\Omega_0^{\frac{1}{2}}[I - (\alpha I + B^*B)^{-1}(\Sigma^{-\frac{1}{2}} B)^* K\Omega_0^{\frac{1}{2}}]L^{s-\beta}\rho_*||^2 \\
&= ||(B^*B)^{\frac{s}{2(a+s)}}[I - (\alpha I + B^*B)^{-1}B^*B](B^*B)^{\frac{\beta-s}{2(a+s)}}\tilde{v}||^2 \\
&= ||\alpha(\alpha I + B^*B)^{-1}](B^*B)^{\frac{\beta}{2(a+s)}}\tilde{v}||^2 = \mathcal{O}(\alpha^{\frac{\beta}{a+s}}).
\end{aligned}
$$

The third equality is obtained by applying Corollary 3 and $\tilde{v}$ is an element of $\mathcal{X}$ such that $L^{s-\beta}\rho_* = (B^*B)^{\frac{\beta-s}{2(a+s)}}\tilde{v}$. Next, we address the second term of the MISE: $\mathbf{E}_{x_*}||\mathcal{C}_2||^2 = tr(AVar(U)A^*)$. Corollary 3 implies that $\mathcal{R}(\Omega_0^{\frac{1}{2}}) \equiv \mathcal{D}(L^s)$ is equal to $\mathcal{R}(B^*B)^{\frac{s}{2(a+s)}}$ so that $A = (B^*B)^{\frac{s}{2(a+s)}}(\alpha I + B^*B)^{-1}(\Sigma^{-\frac{1}{2}} B)^*$ and then

$$
\begin{aligned}
\mathbf{E}_{x_*}||\mathcal{C}_2||^2 &= tr\left[(B^*B)^{\frac{s}{2(a+s)}}(\alpha I + B^*B)^{-1}(\Sigma^{-\frac{1}{2}} B)^* \delta\Sigma\Sigma^{-\frac{1}{2}} B(\alpha I + B^*B)^{-1}(B^*B)^{\frac{s}{2(a+s)}}\right] \\
&= \delta tr\left[(B^*B)^{\frac{s}{2(a+s)}}(\alpha I + B^*B)^{-1}B^*B(\alpha I + B^*B)^{-1}(B^*B)^{\frac{s}{2(a+s)}}\right]
\end{aligned}
$$

after simplification. Then,

$$\mathbf{E}_{x_*}||\mathcal{C}_2||^2 = \delta \sum_j \frac{\lambda_j^{\frac{2s}{a+s}+2}}{(\alpha+\lambda_j^2)^2} = \delta \sum_j \frac{\lambda_j^{\frac{2s}{a+s}+2-2\gamma}}{(\alpha+\lambda_j^2)^2}\lambda_j^{2\gamma}$$

$$\leq \delta \sup_j \frac{\lambda_j^{\frac{2s}{a+s}+2-2\gamma}}{(\alpha+\lambda_j^2)^2} \sum_j \lambda_j^{2\gamma} \asymp \delta\alpha^{-\frac{\gamma(a+s)+a}{a+s}},$$

where we have exploited Assumption C.3. Therefore, $\mathbf{E}_{x_*}||\mathcal{C}_2||^2 = \mathcal{O}\left(\delta\alpha^{-\frac{\gamma(a+s)+a}{a+s}}\right)$ and $\mathbf{E}_{x_*}||\hat{x}_\alpha - x_*||^2 = \mathcal{O}(\alpha^{\frac{\beta}{a+s}}+\delta\alpha^{-\frac{\gamma(a+s)+a}{a+s}})$. This proves the first part of the theorem.

To prove the second part of the Theorem it is easy to see that

$$k_1\delta^{\frac{a+s}{\beta+a+\gamma(a+s)}} = \arg\inf_{\alpha\in\Lambda}\mathbf{E}_{x_*}||\hat{x}_\alpha - x_*||^2$$

$$\sup_{(x_*-x_0)\in\mathcal{X}_\beta(R)}\mathbf{E}_{x_*}||\hat{x}_\alpha - x_*||^2 = \sup_{(x_*-x_0)\in\mathcal{X}_\beta(R)}||\mathcal{C}_1||^2 + \mathbf{E}_{x_*}||\mathcal{C}_2||^2$$

and

$$\sup_{(x_*-x_0)\in\mathcal{X}_\beta(R)}||\mathcal{C}_1||^2 = \sup_{(x_*-x_0)\in\mathcal{X}_\beta(R)}\sum_{j=1}^\infty \alpha^2 < (B^*B)^{\frac{s}{2(a+s)}}(\alpha I + B^*B)^{-1}(B^*B)^{\frac{\beta-s}{2(a+s)}}L^\beta(x_*-x_0),\psi_j >^2$$

$$= \alpha^{\frac{\beta}{a+s}}R\Big(\frac{\beta}{2(a+s)-\beta}\Big)^{\frac{\beta}{a+s}}\Big(\frac{2(a+s)-\beta}{2(a+s)}\Big)^2 \asymp ||\mathcal{C}_1||^2.$$

Therefore,

$$\inf_{\alpha\in\Lambda}\sup_{(x_*-x_0)\in\mathcal{X}_\beta(R)}\mathbf{E}_{x_*}||\hat{x}_\alpha - x_*||^2 = \mathcal{O}(\delta^{\frac{\beta}{\beta+a+\gamma(a+s)}})$$

which proves the second result of the Theorem.

4.3. *Proof of Corollary 1.* Following the proof of Theorem 2.2 we simply have to focus on term $||\mathcal{C}_1||^2$ of the risk decomposition.

$$||\mathcal{C}_1|| = ||[I - K^*\Sigma^{-\frac{1}{2}}B(\alpha I + B^*B)^{-1}\Omega_0^{\frac{1}{2}}]^*(x_*-x_0)||$$

$$= ||[I - K^*\Sigma^{-\frac{1}{2}}B(\alpha\Omega_0^{-\frac{1}{2}} + K^*\Sigma^{-1}K\Omega_0^{\frac{1}{2}})^{-1}]^*(x_*-x_0)||$$

$$= ||[\alpha\Omega_0^{-\frac{1}{2}}(\alpha\Omega_0^{-\frac{1}{2}} + K^*\Sigma^{-1}K\Omega_0^{\frac{1}{2}})^{-1}]^*(x_*-x_0)|| = ||\alpha(\alpha I + \Omega_0 K^*\Sigma^{-1}K)^{-1}(x_*-x_0)||$$

and if $\Omega_0^{\frac{1}{2}}$ commute with $B^*\Sigma^{-\frac{1}{2}}K$ we have

$$||\mathcal{C}_1||^2 = ||\alpha(\alpha I + B^*B)^{-1}(x_*-x_0)||^2 = \sum_{j=1}^\infty \frac{\alpha^2}{(\alpha+\lambda_j^2)^2} < x_*-x_0,\psi_j >^2$$

$$= \sum_{j=1}^\infty \frac{\alpha^2\lambda_j^{2\eta}}{(\alpha+\lambda_j)^2}\frac{< x_*-x_0,\psi_j >^2}{\lambda_j^{2\eta}} = \mathcal{O}(\alpha^\eta)$$

since $(x_*-x_0) \in \mathcal{R}((B^*B)^{\frac{\eta}{2}})$. By using the upper bound for $\mathbf{E}_{x_*}||\mathcal{C}_2||^2$ and the decomposition of the risk given in subsection 4.2 we get the result.

4.4. *Proof of theorem 2.3.* The asymptotic behavior of the posterior variance is similar to that one of term $\mathcal{C}_1$ considered in the proof of Theorem 2.2 except for the fact that it is scaled by the factor $\frac{1}{g}$. We have $\forall \phi \in \mathcal{X}$

$$Var(x|\hat{Y}, g, s)\phi = \frac{1}{g}[\Omega_0 - \Omega_0^{\frac{1}{2}}(\alpha I + B^*B)^{-1}(\Sigma^{-\frac{1}{2}}B)^*K\Omega_0]\phi,$$

$$||Var(x|\hat{Y}, g, s)\phi|| \leq ||Var(x|\hat{Y}, g, s)||||\phi|| = \sup_{||\psi|| \leq 1}||Var(x|\hat{Y}, g, s)\psi||||\phi||.$$

We develop the norm of the operator by using Corollary 3:

$$\sup_{||\psi|| \leq 1}||Var(x|\hat{Y}, g, s)\psi|| = \frac{1}{g}\sup_{||\psi|| \leq 1}||\Omega_0^{\frac{1}{2}}\alpha(\alpha I + B^*B)^{-1}\Omega_0^{\frac{1}{2}}\psi||$$

$$= \frac{1}{g}\sup_{||\psi|| \leq 1}||(B^*B)^{\frac{s}{2(a+s)}}\alpha(\alpha I + B^*B)^{-1}(B^*B)^{\frac{s}{2(a+s)}}\psi||$$

$$= \frac{\alpha}{g}\sup_{||\psi|| \leq 1}\Big(\sum_{j=1}^{\infty}\frac{\lambda_j^{\frac{4s}{a+s}}}{(\alpha + \lambda_j^2)^2} < \psi, \varphi_j >^2\Big)^{\frac{1}{2}}$$

$$= \frac{\alpha}{g}\Big(\sup_j \frac{\lambda_j^{\frac{2s}{a+s}}}{(\alpha + \lambda_j^2)}\Big) \asymp \frac{\alpha^{\frac{s}{a+s}}}{g}$$

which converges to zero. This proves the first statement. Consider now the restriction $Var(x|\hat{Y}, g, s)|_{\mathcal{X}_\beta}$. Since $\forall \phi \in \mathcal{X}_\beta$ there exists $v \in \mathcal{X}$ such that $\phi = \Omega_0^{\frac{\beta}{2s}}v$ and $||\phi|| = ||v||_{-\beta}$, we have

$$||Var(x|\hat{Y}, g, s)|_{\mathcal{X}_\beta}\phi|| = \frac{1}{g}||(B^*B)^{\frac{s}{2(a+s)}}\alpha(\alpha I + B^*B)^{-1}(B^*B)^{\frac{s}{2(a+s)}}\Omega_0^{\frac{\beta}{2s}}v||$$

$$= \frac{1}{g}||(B^*B)^{\frac{s}{2(a+s)}}\alpha(\alpha I + B^*B)^{-1}(B^*B)^{\frac{\beta+s}{2(a+s)}}v||$$

$$= \mathcal{O}\Big(\frac{1}{g}\alpha^{\frac{\beta+2s}{2(a+s)}}\Big).$$

By taking the square of this rate we get the result of the theorem.

4.5. *Proof of theorem 2.4.* Let $\mathbf{E}_\delta^Y$ be the expectation taken with respect to μ_δ^Y. By Chebishev's inequality and Theorem 2.2

$$\mu_\delta^Y\{x \in \mathcal{X} : ||x - x_*|| > \varepsilon_\delta M_\delta\} \leq \frac{1}{\varepsilon_\delta^2 M_\delta^2}\mathbf{E}_\delta^Y||x - x_*||^2$$

$$= \frac{1}{\varepsilon_\delta^2 M_\delta^2}\Big(||\mathbf{E}(x|\hat{Y}, g, s) - x_*||^2 + trVar(x|\hat{Y}, g, s)\Big)$$

$$\asymp \frac{1}{\varepsilon_\delta^2 M_\delta^2}(\alpha^{\frac{\beta}{a+s}} + \delta\alpha^{-\frac{\gamma(a+s)+a}{a+s}}) \quad \text{in } P^{x_*}\text{-probability}$$

since

$$
\begin{aligned}
trVar(x|\hat{Y}, g, s) &= \frac{1}{g}\sum_{j=1}^{\infty}\frac{\alpha\lambda_j^{\frac{2s}{a+s}}}{\alpha + \lambda_j^2} \leq \frac{1}{g}\Big(\sup_j \frac{\alpha\lambda_j^{\frac{2(s-\gamma(a+s))}{a+s}}}{\alpha + \lambda_j^2}\Big)\sum_{j=1}^{\infty}\lambda_j^{2\gamma} \\
&\asymp \frac{1}{g}\alpha^{\frac{s-\gamma(a+s)}{a+s}} \asymp \delta\alpha^{-\frac{a+\gamma(a+s)}{a+s}}.
\end{aligned}
$$

Hence, $\varepsilon_\delta = \inf(\alpha^{\frac{\beta}{2(a+s)}}, \delta^{\frac{1}{2}}\alpha^{-\frac{\gamma(a+s)+a}{2(a+s)}})$ and the fastest rate is obtained when $\alpha^{\frac{\beta}{a+s}} = \delta\alpha^{-\frac{\gamma(a+s)+a}{a+s}}$, so that $\varepsilon_\delta = \delta^{\frac{\beta}{2(\beta+a+\gamma(a+s))}}$.

4.6. *Proof of Corollary 2.* Let consider the eigensystem $\{\lambda_j, \varphi_j, \psi_j\}$ associated with B. Under Assumption D there exist $\underline{\lambda}, \bar{\lambda} > 0$ such that $\underline{\lambda}j^{-(a+s)} \leq \lambda_j \leq \bar{\lambda}j^{-(a+s)}$.
We rewrite the risk associated to $\hat{x}_\alpha$ as:

$$
\begin{aligned}
R(\alpha, x_*) &= \mathbf{E}_{x_*}||\hat{x}_\alpha - x_*||^2 \\
&= \sum_j \frac{\alpha^2}{(\alpha + j^{-2(a+s)})^2} < x_* - x_0, \psi_j >^2 + \delta\sum_j \frac{j^{-2s-2(a+s)}}{(\alpha + j^{-2(a+s)})^2} \\
&=: \mathcal{A}1 + \mathcal{A}2.
\end{aligned}
$$

Because of Assumption B we have that

$$
\begin{aligned}
\sup_{(x_*-x_0)\in\mathcal{X}_\beta(R)} \mathcal{A}1 &= \sup_{(x_*-x_0)\in\mathcal{X}_\beta(R)} \sum_j \frac{\alpha^2 j^{-2\beta}}{(\alpha + j^{-2(a+s)})^2}j^{2\beta} < x_* - x_0, \psi_j >^2 \\
&= R\sup_j \frac{\alpha^2 j^{-2\beta}}{(\alpha + j^{-2(a+s)})^2} = \alpha^{\frac{\beta}{a+s}}c_1, \\
c_1 &:= c_1(R, \beta) = \Big(\frac{2(a+s)-\beta}{2(a+s)}\Big)^2\Big(\frac{\beta}{2(a+s)-\beta}\Big)^{\frac{2\beta}{2(a+s)}}R.
\end{aligned}
$$

Term $\mathcal{A}2$ is developed as follows

$$
\begin{aligned}
\mathcal{A}2 &= \delta\sum_j \frac{1}{(j^{(a+2s)}(\alpha + j^{-2(a+s)}))^2} = \delta\sum_j \frac{1}{[j^{-a}(\alpha j^{2(a+s)} + 1)]^2} \\
&\sim \delta\int_0^\infty \frac{1}{[t^{-a}(\alpha t^{2(a+s)} + 1)]^2}dt = \delta\int_0^\infty \frac{\alpha^{-\frac{1}{2(a+s)}}}{[\alpha^{\frac{a}{2(a+s)}}u^{-a}(u^{2(a+s)} + 1)]^2}du \\
&= \delta\alpha^{-\frac{a}{a+s}-\frac{1}{2(a+s)}}\int_0^\infty \frac{1}{[u^{-a}(u^{2(a+s)} + 1)]^2}du := \delta\alpha^{-\frac{2a+1}{2(a+s)}}c_2.
\end{aligned}
$$

Finally,

$$
\sup_{(x_* - x_0) \in \mathcal{X}_\beta(R)} R(\alpha, x_*) \;=\; \alpha^{\frac{\beta}{(a+s)}} c_1 + \delta \alpha^{-\frac{2a+1}{2(a+s)}} c_2
$$

$$
\arg\inf_{\alpha} \sup_{(x_* - x_0) \in \mathcal{X}_\beta(R)} R(\alpha, x_*) \;=\; \left[\frac{1 + 2a}{2\beta} \frac{c_2}{c_1} \right]^{\frac{a+s}{\beta + a + \frac{1}{2}}} \delta^{\frac{a+s}{\beta + a + \frac{1}{2}}}
$$

$$
\inf_{\alpha} \sup_{(x_* - x_0) \in \mathcal{X}_\beta(R)} R(\alpha, x_*) \;=\; \delta^{\frac{\beta}{\beta + a + \frac{1}{2}}} \left[\frac{1 + 2a}{2\beta} \frac{c_2}{c_1} \right]^{\frac{\beta}{\beta + a + \frac{1}{2}}} c_1 \frac{2(a + \beta) + 1}{2a + 1}
$$

$$
\;=\; C(R, \beta) \delta^{\frac{\beta}{\beta + a + \frac{1}{2}}}.
$$

4.7. *Proof of Theorem 3.1.*   To prove theorem 3.1 we rewrite Theorem of Kuo [26] and then we just need to verify that the conditions of this theorem are verified.

THEOREM 4.1.   *Let P_2 be a Gaussian measure on $\mathcal{Y}$ with mean m and covariance operator S_2 and P_1 be another Gaussian measure on the same space with mean m and covariance operator S_1. If there exists a positive definite, bounded, invertible operator T such that $S_2 = S_1^{\frac{1}{2}} T S_1^{\frac{1}{2}}$ and $T - I$ is Hilbert-Schmidt, then P_2 is equivalent to P_1. Moreover, the Radon-Nikodym derivative is given by*

$$
(4.4) \qquad \frac{dP_2}{dP_1} = \prod_{j=1}^{\infty} \sqrt{\frac{\alpha}{\lambda_j^2 + \alpha}} e^{\frac{\lambda_j^2}{2(\lambda_j^2 + \alpha)} z_j^2},
$$

with $\frac{\lambda_j^2}{\alpha}$ the eigenvalues of $T - I$ and z_j a standard normal random variable under P_1.

We refer to Theorem in [26] for a proof of this Theorem.
In our case: $P_2 = P^\alpha$, $m = K x_0$, $S_2 = \delta\Sigma + \frac{\delta}{\alpha} K \Omega_0 K^*$, $P_1 = P^0$ and $S_1 = \delta\Sigma$. We rewrite S_2 as

$$
S_2 = \left(\delta\Sigma + \frac{\delta}{\alpha} K \Omega_0 K^* \right) \;=\; \sqrt{\delta}\Sigma^{\frac{1}{2}} \left[I + \frac{1}{\alpha} \Sigma^{-\frac{1}{2}} K \Omega_0 K^* \Sigma^{-\frac{1}{2}} \right] \Sigma^{\frac{1}{2}} \sqrt{\delta},
$$

so that $T = [I + \frac{1}{\alpha} \Sigma^{-\frac{1}{2}} K \Omega_0 K^* \Sigma^{-\frac{1}{2}}] = (I + \frac{1}{\alpha} B B^*)$ and it satisfies all the properties required in Theorem 4.1, as we show in the following.

- It is *positive definite*, that is, $(I + \frac{1}{\alpha} B B^*)$ is self-adjoint, *i.e.* $(I + \frac{1}{\alpha} B B^*)^* = (I + \frac{1}{\alpha} B B^*)$ and $\forall \varphi \in \mathcal{Y}$, $\varphi \neq 0$

$$
< (I + \frac{1}{\alpha} B B^*)\varphi, \varphi > = < \varphi, \varphi > + \frac{1}{\alpha} < B^*\varphi, B^*\varphi > = ||\varphi||^2 + \frac{1}{\alpha} ||B^*\varphi|| > 0.
$$

- It is *bounded*. The operators B and B^* are bounded if Assumption C.1 holds; the operator I is bounded by definition and a linear combination of bounded operators is bounded, see Remark 2.7 in [25].

- It is *invertible*: $(I + \frac{1}{\alpha}BB^*)$ is invertible if its inverse is bounded, *i.e.* there exists a positive number C such that $||(I + \frac{1}{\alpha}BB^*)^{-1}\varphi|| \leq C||\varphi||, \forall\varphi \in \mathcal{Y}$. We have $||(I + \frac{1}{\alpha}BB^*)^{-1}\varphi|| \leq ||(I + \frac{1}{\alpha}BB^*)^{-1}||||\varphi||$, and $||(I + \frac{1}{\alpha}BB^*)^{-1}|| \leq \sup_j \frac{\alpha}{\alpha + \lambda_j^2} = 1, \forall\varphi \in \mathcal{Y}$.

- The operator $T - I$ is Hilbert-Schmidt since $||\frac{1}{\alpha}BB^*||_{HS} = \frac{1}{\alpha}\sqrt{\sum_j \lambda_j^4} < \frac{1}{\alpha}\sqrt{\sum_j \lambda_j^2} < \infty$ by Assumption C.3, where $|| \cdot ||_{HS}$ denotes the Hilbert-Schmidt norm.

4.8. *Proof of Theorem 3.2.* We show the first part of the Theorem, *i.e.* the fact that $\hat{\alpha}$ has the same rate as α_{opt}. The oracle inequality (3.4) is proved in subsection 4.8.1.

We compute $S(\alpha)$ by using the fact that under Assumption D there exist $\underline{\lambda}, \bar{\lambda} > 0$ such that $\underline{\lambda}j^{-(a+s)} \leq \lambda_j \leq \bar{\lambda}j^{-(a+s)}$ and under Assumptions B and D there exist $\underline{x}, \bar{x} > 0$ such that $\underline{x}j^{-b_0} \leq |<x_* - x_0, \psi_j>| \leq \bar{x}j^{-b_0}$ for $b_0 > \beta + \frac{1}{2}$. Then,

$$S(\alpha) = \frac{1}{2}\sum_{j=1}^{\infty}\frac{\lambda_j^2}{\alpha(\alpha + \lambda_j^2)} - \frac{1}{2}\sum_{j=1}^{\infty}\frac{\lambda_j^2}{\delta(\alpha + \lambda_j^2)^2 l_j^2} < K(x_* - x_0), \varphi_j >^2$$

$$-\frac{1}{2}\sum_{j=1}^{\infty}\frac{\lambda_j^2}{\delta(\alpha + \lambda_j^2)^2 l_j^2} < K(x_* - x_0), \varphi_j >< U, \varphi_j > -\frac{1}{2}\sum_{j=1}^{\infty}\frac{\lambda_j^2}{\delta(\alpha + \lambda_j^2)^2 l_j^2} < U, \varphi_j >^2$$

$$:= \mathcal{S}_1 - \mathcal{S}_2 - \mathcal{S}_3 - \mathcal{S}_4.$$

Let start by computing $\mathcal{S}_2$:

$$\mathcal{S}_2 = \frac{1}{2}\sum_{j=1}^{\infty}\frac{\lambda_j^2}{\delta(\alpha + \lambda_j^2)^2 l_j^2} < K(x_* - x_0), \psi_j >^2$$

$$\sim \frac{1}{2\delta}\sum_{j=1}^{\infty}\frac{j^{-2(a+s)-2a-2b_0}}{(\alpha + j^{-2(a+s)})^2} = \frac{1}{2\delta}\sum_{j=1}^{\infty}\frac{j^{-4a-2s-2b_0}}{(\alpha + j^{-2(a+s)})^2}$$

$$= \frac{1}{2\delta}\sum_{j=1}^{\infty}\frac{j^{2s-2b_0}}{(\alpha j^{2(a+s)} + 1)^2}$$

$$\sim \frac{1}{2\delta}\alpha^{\frac{b_0-s}{a+s} - \frac{1}{2(a+s)}}\int_{\underline{u}}^{\infty}\frac{u^{2(s-b_0)}}{(u^{2(a+s)} + 1)^2}du, \quad \text{for } \underline{u} > 0$$

$$= \frac{1}{2\delta}\alpha^{\frac{b_0-s-\frac{1}{2}}{a+s}}\int_0^{\infty}\frac{u^{2(s-b_0)}}{(u^{2(a+s)} + 1)^2}du := \frac{1}{2\delta}\alpha^{\frac{\beta-s}{a+s}}c_4$$

where we have taken $b_0 = \beta + \frac{1}{2} + \varepsilon$, for a small $\varepsilon > 0$. Let ξ_j denote a $\mathcal{N}(0,1)$ random variable; we rewrite $\mathcal{S}_3$ as

$$\mathcal{S}_3 = \frac{1}{2}\sum_{j=1}^{\infty}\frac{\lambda_j^2}{\sqrt{\delta}(\alpha + \lambda_j^2)^2 l_j} < K(x_* - x_0), \psi_j > \frac{<U, \varphi_j>}{\sqrt{\delta}l_j} = \frac{1}{2\sqrt{\delta}}\sum_{j=1}^{\infty}\frac{\lambda_j^2}{(\alpha + \lambda_j^2)^2 l_j} < K(x_* - x_0), \psi_j > \xi_j.$$

This series is convergent if, for a fixed α, $\mathbf{E}||\mathcal{S}_3||^2 < \infty$. This is always verified because:

$$\mathbf{E}||\mathcal{S}_3||^2 \sim \frac{1}{2\delta}\sum_{j=1}^{\infty}\frac{\lambda_j^4}{(\alpha+\lambda_j^2)^4}j^{-2(a+b_0)}$$

$$\leq \frac{1}{2\delta}\Big(\sup_j \frac{\lambda_j^2}{(\alpha+\lambda_j^2)^2}\Big)^2\sum_{j=1}^{\infty}j^{-2(a+b_0)} = \frac{1}{32\delta\alpha}\sum_{j=1}^{\infty}j^{-2(a+b_0)}$$

and it is finite if and only if $b_0 > \frac{1}{2} - a$ which is always true. Therefore, $\mathcal{S}_3$ is a gaussian random variable with 0 mean and variance $\frac{1}{2\delta}\sum_{j=1}^{\infty}\frac{\lambda_j^4}{(\alpha+\lambda_j^2)^4}j^{-2(a+b_0)}$:

$$\mathcal{S}_3 \sim \frac{1}{2\sqrt{\delta}}\xi\Big(\sum_{j=1}^{\infty}\frac{j^{-4(a+s)-2a-2b_0}}{(\alpha+j^{-2(a+s)})^4}\Big)^{\frac{1}{2}} = \frac{1}{2\sqrt{\delta}}\xi\Big(\sum_{j=1}^{\infty}\frac{j^{2a+4s-2b_0}}{(\alpha j^{2(a+s)}+1)^4}\Big)^{\frac{1}{2}}$$

$$\sim \frac{1}{2\sqrt{\delta}}\xi\alpha^{-\frac{1}{4(a+s)}-\frac{a+2s-b_0}{2(a+s)}}\Big(\int_{\underline{u}}^{\infty}\frac{u^{2a+4s-2b_0}}{(u^{2(a+s)}+1)^4}du\Big)^{\frac{1}{2}} := \frac{1}{2\sqrt{\delta}}\xi\alpha^{-\frac{a+2s-\beta}{2(a+s)}}c_5$$

where we have taken $b_0 = \beta + \frac{1}{2} + \varepsilon$, for a small $\varepsilon > 0$. Because $\frac{<U,\varphi_j>^2}{\delta l_j^2} \sim i.i.d. \chi_1^2$, we center term $\mathcal{S}_4$ around its mean and apply the Central Limit Theorem

$$\mathcal{S}_4 = \frac{1}{2}\sum_{j=1}^{\infty}\frac{\lambda_j^2}{(\alpha+\lambda_j^2)^2}\Big(\frac{<U,\varphi_j>^2}{\delta l_j^2}-1\Big) + \frac{1}{2}\sum_{j=1}^{\infty}\frac{\lambda_j^2}{(\alpha+\lambda_j^2)^2}$$

$$= \frac{1}{2}\Big(2\sum_{j=1}^{\infty}\frac{\lambda_j^4}{(\alpha+\lambda_j^2)^4}\Big)^{\frac{1}{2}}\xi + \frac{1}{2}\sum_{j=1}^{\infty}\frac{\lambda_j^2}{(\alpha+\lambda_j^2)^2} := \mathcal{S}_{4a} + \mathcal{S}_{4b},$$

where $\xi \sim \mathcal{N}(0,1)$ and $2\sum_{j=1}^{\infty}\frac{\lambda_j^4}{(\alpha+\lambda_j^2)^4}$ is the variance of $\sum_{j=1}^{\infty}\frac{\lambda_j^2}{(\alpha+\lambda_j^2)^2}(\xi^2-1)$. We remark that the same approximations for $\mathcal{S}_3$ and $\mathcal{S}_4$ remain valid even if U is only asymptotically gaussian. In this case an application of the Central Limit Theorem would provide the same result.

Term $\mathcal{S}_{4b}$ is non random and we subtract it from $\mathcal{S}_1$:

$$\mathcal{S}_1 - \mathcal{S}_{4b} \sim \frac{1}{2}\Big[\sum_{j=1}^{\infty}\frac{j^{-2(a+s)}}{\alpha(\alpha+j^{-2(a+s)})} - \sum_{j=1}^{\infty}\frac{j^{-2(a+s)}}{(\alpha+j^{-2(a+s)})^2}\Big]$$

$$= \frac{1}{2\alpha}\sum_{j=1}^{\infty}\frac{j^{-4(a+s)}}{(\alpha+j^{-2(a+s)})^2}$$

$$= \frac{1}{2\alpha}\sum_{j=1}^{\infty}\frac{1}{(\alpha j^{2(a+s)}+1)^2} \sim \frac{1}{2\alpha}\int_0^{\infty}\frac{1}{(\alpha t^{2(a+s)}+1)^2}dt$$

$$= \frac{1}{2}\alpha^{-1-\frac{1}{2(a+s)}}\int_0^{\infty}\frac{1}{(u^{2(a+s)}+1)^2}du := \frac{1}{2}\alpha^{-\frac{2(a+s)+1}{2(a+s)}}c_3.$$

Lastly, term $\mathcal{S}_{4a}$ is

$$
\begin{aligned}
\mathcal{S}_{4a} &\sim \frac{1}{\sqrt{2}}\Big(\sum_{j=1}^{\infty}\frac{j^{-4(a+s)}}{(\alpha+j^{-2(a+s)})^4}\Big)^{\frac{1}{2}}\xi = \frac{1}{\sqrt{2}}\Big(\sum_{j=1}^{\infty}\frac{j^{4(a+s)}}{(\alpha j^{2(a+s)}+1)^4}\Big)^{\frac{1}{2}}\xi \\
&\sim \frac{1}{\sqrt{2}}\Big(\int_0^{\infty}\frac{t^{4(a+s)}}{(\alpha t^{2(a+s)}+1)^4}dt\Big)^{\frac{1}{2}}\xi \\
&= \frac{1}{\sqrt{2}}\alpha^{-1-\frac{1}{4(a+s)}}\Big(\int_0^{\infty}\frac{u^{4(a+s)}}{(u^{2(a+s)}+1)^4}du\Big)^{\frac{1}{2}}\xi = \frac{1}{\sqrt{2}}\xi\alpha^{-\frac{4(a+s)+1}{4(a+s)}}c_6.
\end{aligned}
$$

By putting together all the terms we get

$$
S(\alpha) = \frac{1}{2}\alpha^{-1-\frac{1}{2(a+s)}}c_3 - \frac{1}{2\delta}\alpha^{\frac{\beta-s}{a+s}}c_4 - \frac{1}{2\sqrt{\delta}}\xi\alpha^{-\frac{a+2s-\beta}{2(a+s)}}c_5 - \frac{1}{\sqrt{2}}\xi\alpha^{-\frac{4(a+s)+1}{4(a+s)}}c_6
$$

and $\hat{\alpha}$ is solution of $S(\alpha) = 0$ or equivalently of $\tilde{S}(\alpha) = 0$ with

$$
(4.5) \qquad \tilde{S}(\alpha) := \delta(c_3 - \sqrt{2}\xi\alpha^{\frac{1}{4(a+s)}}c_6) - \alpha^{\frac{2\beta+2a+1}{2(a+s)}}c_4 - \sqrt{\delta}\xi\alpha^{\frac{\beta+a+1}{2(a+s)}}c_5 = 0.
$$

A solution to this equation does not exist in closed form and it must be computed through numerical methods. However, we can easily compute the rate of convergence to 0 of $\hat{\alpha}$. In fact, let notice that $\sqrt{2}\delta\xi\alpha^{\frac{1}{4(a+s)}}c_6$ is negligible with respect to δc_3 as $\delta \downarrow 0$ and $\sqrt{\delta}\xi\alpha^{\frac{\beta+a+1}{2(a+s)}}c_5$ is negligible with respect to $\alpha^{\frac{2\beta+2a+1}{2(a+s)}}c_4$ if $\sqrt{\delta}\alpha^{-\frac{\beta+a}{2(a+s)}} \downarrow 0$. The $\hat{\alpha}$ which solves $\tilde{S}(\alpha) = 0$ is then such that $\delta c_3 = \alpha^{\frac{2\beta+2a+1}{2(a+s)}}c_4$, $i.e.$

$$
\hat{\alpha} = \mathcal{O}_p(\delta^{\frac{a+s}{\beta+a+\frac{1}{2}}})
$$

and $\sqrt{\delta}\hat{\alpha}^{-\frac{\beta+a}{2(a+s)}} \downarrow 0$. Furthermore, let $\hat{\alpha}_1$ be such that $\delta c_3 = \hat{\alpha}_1^{\frac{2\beta+2a+1}{2(a+s)}}c_4$. Then, $\tilde{S}(\hat{\alpha}_1) = \delta^{\frac{a+s}{\beta+a+\frac{1}{2}}}\xi(\sqrt{2}c_6\delta^{\frac{4\beta-4s+3}{\beta+a+\frac{1}{2}}} - c_5\delta^{\frac{3\beta-a-4s+2}{4\beta+4a+2}})$ and $\tilde{S}(\hat{\alpha}_1) = \delta^{\frac{a+s}{\beta+a+\frac{1}{2}}}o_p(1)$. Then, we can write

$$
(4.6) \qquad \hat{\alpha} = \delta^{\frac{a+s}{\beta+a+\frac{1}{2}}}\Big(\frac{c_3}{c_4} + o_p(1)\Big) \asymp \alpha_{opt}.
$$

4.8.1. *Proving the oracle inequality (3.4).* In this proof we use the notation $B(\hat{\alpha}) := (\hat{\alpha}I + B^*B)^{-1}$ and $B(\alpha_{opt}) := (\alpha_{opt}I + B^*B)^{-1}$. Let consider the norm defined by $\sqrt{\mathbf{E}_{x_*}||\cdot||^2}$ and the risk defined as $R(\alpha, x_*) = \mathbf{E}_{x_*}||\hat{x}_\alpha - x_*||^2$ with $\alpha = \hat{\alpha}$ or $\alpha = \alpha_{opt}$. The risk can be bounded above by

$$
\begin{aligned}
R(\hat{\alpha}, x_*) &= \mathbf{E}_{x_*}||\hat{x}_{\hat{\alpha}} - \hat{x}_{opt} + \hat{x}_{opt} - x_*||^2 \\
&\leq \mathbf{E}_{x_*}(||\hat{x}_{\hat{\alpha}} - \hat{x}_{opt}|| + ||\hat{x}_{opt} - x_*||)^2 \\
(4.7) \qquad &\leq 2(\mathbf{E}_{x_*}||\hat{x}_{\hat{\alpha}} - \hat{x}_{opt}||^2 + R(\alpha_{opt}, x_*)).
\end{aligned}
$$

Next, we develop $\mathbf{E}_{x_*}||\hat{x}_{\hat{\alpha}} - \hat{x}_{\alpha_{opt}}||^2$. From the definition of $\hat{x}_\alpha$ in (2.1) we write

$$
\begin{aligned}
\mathbf{E}_{x_*}||\hat{x}_{\hat{\alpha}} - \hat{x}_{\alpha_{opt}}||^2 &= \mathbf{E}_{x_*}||\Omega_0^{\frac{1}{2}}B(\hat{\alpha})B^*\Sigma^{-\frac{1}{2}}\hat{Y} - \Omega_0^{\frac{1}{2}}B(\alpha_{opt})B^*\Sigma^{-\frac{1}{2}}\hat{Y}||^2 \\
&= \mathbf{E}_{x_*}||\left(\Omega_0^{\frac{1}{2}}B(\hat{\alpha})B^*\Sigma^{-\frac{1}{2}}\hat{Y} - x_*\right) - \left(\Omega_0^{\frac{1}{2}}B(\alpha_{opt})B^*\Sigma^{-\frac{1}{2}}\hat{Y} - x_*\right)||^2 \\
&= \mathbf{E}_{x_*}||\left(\Omega_0^{\frac{1}{2}}B(\hat{\alpha})B^*\Sigma^{-\frac{1}{2}}K - I\right)x_* + \Omega_0^{\frac{1}{2}}B(\hat{\alpha})B^*\Sigma^{-\frac{1}{2}}U \\
&\quad - \left(\Omega_0^{\frac{1}{2}}B(\alpha_{opt})B^*\Sigma^{-\frac{1}{2}}K - I\right)x_* - \Omega_0^{\frac{1}{2}}B(\alpha_{opt})B^*\Sigma^{-\frac{1}{2}}U||^2 \\
&\leq 2\mathbf{E}_{x_*}|| - \Omega_0^{\frac{1}{2}}\hat{\alpha}B(\hat{\alpha})\xi_* + \Omega_0^{\frac{1}{2}}\alpha_{opt}B(\alpha_{opt})\xi_*||^2 \\
&\quad + 2\mathbf{E}_{x_*}||\Omega_0^{\frac{1}{2}}\left[B(\hat{\alpha}) - B(\alpha_{opt})\right]B^*\Sigma^{-\frac{1}{2}}U||^2 := 2(\mathcal{A}1 + \mathcal{A}2),
\end{aligned}
$$

where ξ_* is such that $x_* = \Omega_0^{\frac{1}{2}}\xi_*$ and it exists by Assumption B.

$$
\begin{aligned}
\mathcal{A}1 &= \mathbf{E}_{x_*}||\Omega_0^{\frac{1}{2}}B(\hat{\alpha})(-\hat{\alpha}\alpha_{opt}I - \hat{\alpha}B^*B + \hat{\alpha}\alpha_{opt}I + \alpha_{opt}B^*B)B(\alpha_{opt})\xi_*||^2 \\
&= \mathbf{E}_{x_*}||\Omega_0^{\frac{1}{2}}B(\hat{\alpha})(\alpha_{opt} - \hat{\alpha})B^*B\frac{1}{\alpha_{opt}}\alpha_{opt}B(\alpha_{opt})\xi_*||^2 \\
&\leq \mathbf{E}_{x_*}||(B^*B)^{\frac{s}{2(a+s)}}B(\hat{\alpha})\frac{\alpha_{opt} - \hat{\alpha}}{\alpha_{opt}}(B^*B)^{1-\frac{s}{2(a+s)}}||^2||(B^*B)^{\frac{s}{2(a+s)}}\alpha_{opt}B(\alpha_{opt})\xi_*||^2.
\end{aligned}
$$

The last norm is exactly term $||\mathcal{C}_1||^2$ evaluated at α_{opt} in the proof of Theorem 2.2 (and also term $\mathcal{A}_1$ in the proof of Corollary 2) and then it is part of $R(\alpha_{opt}, x_*)$. We denote it by $||\mathcal{C}_1^{opt}||^2$ and, by Lemma 2 we have

$$(4.8) \qquad \mathcal{A}_1 \leq \frac{\mathbf{E}_{x_*}(\alpha_{opt} - \alpha_*)^2}{\alpha_{opt}^2}||\mathcal{C}_1^{opt}||^2.$$

Consider now term $\mathcal{A}2$.

$$
\begin{aligned}
\mathcal{A}2 &= \mathbf{E}_{x_*}||\Omega_0^{\frac{1}{2}}B(\hat{\alpha})(\alpha_{opt} - \hat{\alpha})B(\alpha_{opt})B^*\Sigma^{-\frac{1}{2}}U||^2 \\
&\leq \left(\mathbf{E}_{x_*}||\Omega_0^{\frac{1}{2}}B(\hat{\alpha})(\alpha_{opt} - \hat{\alpha})(B^*B)^{-\frac{s}{2(a+s)}}||^2\right)\left(\mathbf{E}_{x_*}||(B^*B)^{\frac{s}{2(a+s)}}B(\alpha_{opt})B^*\Sigma^{-\frac{1}{2}}U||^2\right).
\end{aligned}
$$

The last term in brackets in $\mathcal{A}2$ is term $\mathbf{E}_{x_*}||\mathcal{C}_2||^2$ in the proof of Theorem 2.2, evaluated at α_{opt} (and also term $\mathcal{A}_2$ in the proof of Corollary 2). We denote it by $\mathbf{E}_{x_*}||\mathcal{C}_2^{opt}||^2$. Let $\bar{\mathcal{X}} = \{v \in \mathcal{X}$ such that $||v|| \leq 1\}$, then

$$
\begin{aligned}
\mathcal{A}2 &= \mathbf{E}_{x_*}||(B^*B)^{\frac{s}{2(a+s)}}B(\hat{\alpha})(\alpha_{opt} - \hat{\alpha})(B^*B)^{-\frac{s}{2(a+s)}}||^2\mathbf{E}_{x_*}||\mathcal{C}_2^{opt}||^2 \\
&= \mathbf{E}_{x_*}\left[(\alpha_{opt} - \hat{\alpha})^2 \sup_{v \in \bar{\mathcal{X}}}\sum_{j=1}^{\infty}\frac{1}{(\hat{\alpha} + \lambda_j^2)^2} <v, \varphi_j>^2\right]\mathbf{E}_{x_*}||\mathcal{C}_2^{opt}||^2 \\
&= \mathbf{E}_{x_*}\left[\frac{(\alpha_{opt} - \hat{\alpha})^2}{\alpha_{opt}^2}\sup_{v \in \bar{\mathcal{X}}}\sum_{j=1}^{\infty}\frac{\alpha_{opt}^2}{(\hat{\alpha} + \lambda_j^2)^2} <v, \varphi_j>^2\right]\mathbf{E}_{x_*}||\mathcal{C}_2^{opt}||^2
\end{aligned}
$$

$$\mathcal{A}2 \quad \leq \quad \mathbf{E}_{x_*}\Big[\frac{(\alpha_{opt} - \hat{\alpha})^2}{\alpha_{opt}^2} \sup_{v \in \bar{\mathcal{X}}} \sum_{j=1}^{\infty} <v, \varphi_j>^2\Big]\mathbf{E}_{x_*}||\mathcal{C}_2^{opt}||^2$$

(4.9) $$\leq \quad \frac{\mathbf{E}_{x_*}(\alpha_{opt} - \hat{\alpha})^2}{\alpha_{opt}^2}||v||^2\mathbf{E}_{x_*}||\mathcal{C}_2^{opt}||^2 \leq \frac{\mathbf{E}_{x_*}(\alpha_{opt} - \hat{\alpha})^2}{\alpha_{opt}^2}\mathbf{E}_{x_*}||\mathcal{C}_2^{opt}||^2$$

where we have used the following fact:

$$\frac{\alpha_{opt}}{\hat{\alpha} + \lambda_j^2} \quad = \quad \frac{\alpha_{opt} - \hat{\alpha}}{\hat{\alpha} + \lambda_j^2} + \frac{\hat{\alpha}}{\hat{\alpha} + \lambda_j^2} \leq \frac{\alpha_{opt} - \hat{\alpha}}{\hat{\alpha} - \alpha_{opt}} + \frac{\alpha_{opt} - \hat{\alpha}}{\alpha_{opt} + \lambda_j^2} + \frac{\hat{\alpha}}{\hat{\alpha} + \lambda_j^2}$$

$$\leq \quad -1 + \frac{\alpha_{opt}}{\alpha_{opt} + \lambda_j^2} + \frac{\hat{\alpha}}{\hat{\alpha} + \lambda_j^2} \leq -1 + 1 + 1 = 1.$$

By putting together (4.8) and (4.9) we have

$$\mathbf{E}_{x_*}||\hat{x}_{\hat{\alpha}} - \hat{x}_{\alpha_{opt}}||^2 \leq \frac{2\mathbf{E}_{x_*}(\alpha_{opt} - \hat{\alpha})^2}{\alpha_{opt}^2}\Big(||\mathcal{C}_1^{opt}||^2 + \mathbf{E}_{x_*}||\mathcal{C}_2^{opt}||^2\Big)$$

where $||\mathcal{C}_1^{opt}||^2 + \mathbf{E}_{x_*}||\mathcal{C}_2^{opt}||^2 = R(\alpha_{opt}, x_*)$. Hence, by (4.7) and Lemma 1 we have

$$R(\hat{\alpha}, x_*) \quad \leq \quad R(\alpha_{opt}, x_*)\Big(\frac{4\mathbf{E}_{x_*}(\alpha_{opt} - \hat{\alpha})^2}{\alpha_{opt}^2} + 2\Big)$$

$$\leq \quad \inf_{\alpha \in \Lambda} R(\alpha, x_*)\Big(1 + \frac{\alpha_{opt}^2 + \alpha_{opt}^2 4\Big(1 + \frac{c_3^2}{k_3^2 \tilde{c}_3^2} + o(1)\Big)}{\alpha_{opt}^2}\Big)$$

$$\leq \quad \inf_{\alpha \in \Lambda} R(\alpha, x_*)(1 + \mathcal{O}(1)).$$

This concludes the proof of Theorem 3.2.

LEMMA 1.   *Let α_{opt} be the oracle defined in (2.6) and $\hat{\alpha}(\hat{Y})$ be the solution of $S(\alpha, \hat{Y}) = 0$ defined in 3.3. Under the assumptions of Theorem 3.2, we have*

(4.10) $$\mathbf{E}_{x_*}(\alpha_{opt} - \hat{\alpha}(\hat{Y}))^2 \leq \alpha_{opt}^2\Big(1 + \frac{c_3^2}{k_3^2 \tilde{c}_3^2} + o_p(1)\Big)$$

where $c_3 := \int_0^{\infty} \frac{1}{(u^{2(a+s)}+1)^2}du$, $k_3 := \int_0^{\infty} u^{2s-2a+1}du$ ans $\tilde{c}_3 := \int_0^{\infty} \frac{u^{4a+2s}}{(u^{2(a+s)}+1)^3}du$.

PROOF.   Under Assumptions B and D the risk $R(\alpha, x_*)$ can be rewritten as

$$R(\alpha, x_*) = \sum_j \Big(1 - \frac{j^{-2(a+s)}}{(\alpha + j^{-2(a+s)})}\Big)^2 <x_* - x_0, \psi_j>^2 + \delta\sum_j \frac{j^{-2(a+2s)}}{(\alpha + j^{-2(a+s)})^2}.$$

Then, α_{opt} is the solution of $\frac{\partial}{\partial\alpha}R(\alpha, x_*) = 0$ and $\frac{\partial}{\partial\alpha}R(\alpha, x_*)$ is approximately equal to

$$\frac{\partial}{\partial\alpha}R(\alpha, x_*) \sim 2\Big(\alpha^{-\frac{a+s+b_0-\frac{1}{2}}{a+s}}\tilde{c}_4 - \delta\alpha^{-\frac{2a+s+\frac{1}{2}}{a+s}}\tilde{c}_3\Big),$$

with $\tilde{c}_4 := \int_0^\infty \frac{u^{4(a+s)-2b_0}}{(u^{2(a+s)}+1)^3}$ and $\tilde{c}_3 := \int_0^\infty \frac{u^{4a+2s}}{(u^{2(a+s)}+1)^3} du$. This can be shown by using a similar procedure as for determine $S(\alpha)$ above. In an equivalent way, α_{opt} is solution of

$$\delta \tilde{c}_3 - \alpha^{\frac{\beta+a+\frac{1}{2}}{a+s}} \tilde{c}_4 = 0$$

since $b_0 = \beta + \frac{1}{2} + \epsilon$, $\epsilon > 0$. Hence,

(4.11) $$\alpha_{opt} = \delta^{\frac{a+s}{\beta+a+\frac{1}{2}}} \frac{\tilde{c}_3}{\tilde{c}_4}.$$

From (4.6), we have $\hat{\alpha} = \alpha_{opt}(\frac{c_3\tilde{c}_4}{c_4\tilde{c}_3} + o_p(1))$ and by the Cauchy-Schwartz inequality we have $c_4 \leq \tilde{c}_4 k_3$. Putting all together we conclude that

$$\begin{aligned} \mathbf{E}_{x_*}(\alpha_{opt} - \hat{\alpha})^2 &= \alpha_{opt}^2 \left(1 - \frac{c_3\tilde{c}_4}{c_4\tilde{c}_3} - o(1)\right) \\ &\leq \alpha_{opt}^2 \left(1 + \frac{c_3^2}{k_3^2\tilde{c}_3^2} + o(1)\right). \end{aligned}$$

$\square$

LEMMA 2. *Let* $\hat{\alpha}, \alpha_{opt} \in \Lambda := (0, \bar{\alpha}]$ *where* $\bar{\alpha} > 0$ *is a finite constant and* $\{\lambda_j^2\}$ *denote the eigenvalues of* B^*B. *Then, we have*

(4.12) $$\mathbf{E}_{x_*}\left(||(B^*B)^{\frac{s}{2(a+s)}}B(\hat{\alpha})(\alpha_{opt} - \hat{\alpha})(B^*B)^{1-\frac{s}{2(a+s)}}||^2\right) = \mathcal{O}_p\left(\delta^{\frac{2(a+s)}{\beta+a+\frac{1}{2}}}\right).$$

where $B(\hat{\alpha}) := (\hat{\alpha}I + B^*B)^{-1}$.

PROOF. Let consider the norm $|| \cdot || := \sqrt{\mathbf{E}_{x_*}| \cdot |^2}$. By definition of the norm of an operator, we have

$$||(B^*B)^{\frac{s}{2(a+s)}}B(\hat{\alpha})(B^*B)^{1-\frac{s}{2(a+s)}}||^2 := \sup_{v\in\bar{\mathcal{X}}} ||(B^*B)^{\frac{s}{2(a+s)}}B(\hat{\alpha})(\alpha_{opt}-\hat{\alpha})(B^*B)^{1-\frac{s}{2(a+s)}}v||^2,$$

where the sup is taken over the set $\bar{\mathcal{X}} = \{v \in \mathcal{X}$ such that $||v|| \leq 1\}$. We develop the right hand side of the previous equality as

$$\mathbf{E}_{x_*}\left[(\alpha_{opt} - \hat{\alpha})^2 \sup_{v\in\bar{\mathcal{X}}} \sum_{j=1}^\infty \frac{\lambda_j^4}{(\hat{\alpha} + \lambda_j^2)^2} <v, \varphi_j>^2\right]$$

$$\leq \mathbf{E}_{x_*}(\alpha_{opt} - \hat{\alpha})^2 \left(\sup_j \frac{\lambda_j^2}{(\hat{\alpha} + \lambda_j^2)}\right)^2 \sup_{v\in\bar{\mathcal{X}}} \sum_{j=1}^\infty <v, \varphi_j>^2$$

$$\leq \mathbf{E}_{x_*}(\alpha_{opt} - \hat{\alpha})^2$$

since $\sup_j \frac{\lambda_j^2}{(\hat{\alpha}+\lambda_j^2)} \leq 1$ and $\sup_{v\in\bar{\mathcal{X}}} ||v||^2 = 1$. $\square$

References.

[1] Bauer, F., Hohage, T. and A., Munk (2009), 'Iteratively regularized Gauss-Newton method for nonlinear inverse problems with random noise', *SIAM J. Numer. Anal.*, **47**, 1827-1846.

[2] Belitser, E. and Ghosal, S. (2003), 'Adaptive Bayesian inference on the mean of an infinite-dimensional normal distribution', *Annals of Statistics*, **31**, 536–559.

[3] Bissantz, N., Hohage, T., Munk,A. and F., Ruymgaart (2007), 'Convergence rates of general regularization methods for statistical inverse problems and applications', *SIAM J. Numer. Anal.*, **45**, 2610-2636.

[4] Brwon, L.D. and Low, M.G. (1996), 'Asymptotic equivalence of nonparametric regression and white noise', *Annals of Statistics*, **24**, 2384–2398.

[5] Carrasco, M. and J.P., Florens (2007), 'Spectral Method for Deconvolving a Density', preprint.

[6] Carrasco, M., Florens, J.P., and E., Renault (2005), 'Estimation based on Spectral Decomposition and Regularization', *Handbook of Econometrics*, J.J. Heckman and E. Leamer, eds., **6**, Elsevier, North Holland.

[7] Cavalier, L. and A.B., Tsybakov (2002), 'Sharp adaptation for inverse problems with random noise', *Proba. Theory and Related Fields*, **123**, 323–354.

[8] Chen, X. and M., Reiss, 2007, On rate optimality for ill-posed inverse problems in econometrics, Cowles Foundation discussion paper.

[9] Darolles, S., Florens, J.P., and E., Renault (2003), 'Nonparametric instrumental regression', Working paper.

[10] Diaconis, F., and D., Freedman (1986), 'On the Consistency of Bayes Estimates', *Annals of Statistics*, **14**, 1-26.

[11] Engl, H.W., Hanke, M. and A., Neubauer (2000), 'Regularization of Inverse Problems', *Kluwer Academic, Dordrecht*.

[12] Florens, J.P. and A., Simoni (2008), 'Regularized Posteriors in Linear ill-Posed Inverse Problems', preprint.

[13] Florens, J.P. and A., Simoni (2008), 'Regularized Posteriors in Linear ill-Posed Inverse Problems: Extensions', preprint.

[14] Franklin, J.N. (1970), 'Well-posed stochastic extension of ill-posed linear problems', *Journal of Mathematical Analysis and Applications*, **31**, 682 - 716.

[15] George, E.I. and D.P., Foster (2000), 'Calibration and Empirical Bayes Variable Selection', *Biometrika*, **87**, 4, 731 - 747.

[16] Ghosh, J.K. and R.V., Ramamoorthi (2003), 'Bayesian Nonparametrics', *Springer Series in Statistics*.

[17] Hall, P. and J., Horowitz (2005), 'Nonparametric methods for inference in the presence of instrumental variables', *Annals of Statistics*, **33**, 2904-2929.

[18] Hall,P. and J.L., Horowitz (2007), 'Methodology and Convergence Rates for Functional Linear Regression', *Annals of Statistics*, **35**, 70 - 91.

[19] Hiroshi, S. and O., Yoshiaki (1975), 'Separabilities of a Gaussian Measure', *Annales de l'I.H.P., section B*, tome 11, **3**, 287 - 298.

[20] Johannes, J., Van Bellegem, S. and A., Vanhems (2009), 'Convergence Rates for ill-posed Inverse Problems with an Unknown Operator', preprint.

[21] Johnstone, I.M. and B.W., Silverman (2004), 'Needles and straw in haystacks: empirical Bayes estimates of possibly sparse sequences', *Annals of Statistics*, **32**, No.4, 1594–1649.

[22] Kaipio, J., and E., Somersalo (2004), 'Statistical and Computational Inverse Problems', *Applied Mathematical Series, vol.160, Springer, Berlin*.

[23] Kato, T. (1995), 'Perturbation theory for linear operators', Springer.
[24] Krein, S.G. and J.I., Petunin (1966), 'Scales of Banach Spaces', *Russian Math. Surveys*, **21**, 85 - 160.
[25] Kress, R. (1999), 'Linear Integral Equation', Springer.
[26] Kuo, H.H. (1975), 'Gaussian Measures in Banach Spaces', Springer.
[27] Lehtinen, M.S., Päivärinta, L. and E., Somersalo (1989), 'Linear Inverse Problems for Generalised Random Variables', *Inverse Problems*, 5, 599-612.
[28] Liang, F., Paulo, R., Molina, G., Clyde, M. and Berger, J. (2005), 'Mixtures of g-priors for Bayesian Variable Selection', ISDS Technical Report, Duke University.
[29] Mandelbaum, A. (1984), 'Linear Estimators and Measurable Linear Transformations on a Hilbert Space', *Z. Wahrcheinlichkeitstheorie*, 3, 385-98.
[30] Nair, M.T., Pereverzev, S.V. and Tautenhahn, U. (2005), 'Regularization in Hilbert scales under general smoothing conditions', *Inverse Problems*, **21**, 1851-1969.
[31] Neubauer, A. (1988), 'When do Sobolev Spaces Form a Hilbert Scale?', *Proc. Amer. Math. Soc.*, **103**, 557 - 562.
[32] Neveu, J. (1965), 'Mathematical Fundations of the Calculus of Probability', *San Francisco: Holden-Day*.
[33] Nussbaum, M. (1996), 'Asymptotic equivalence of density estimation and gaussian white noise', *Annals of Statistics*, **24**, No.6, 2399–2430.
[34] Ramsay, J. and Silverman, B. (1997), 'Functional data analysis'. *Springer, New York*.
[35] Simoni, A. (2009), 'Bayesian Analysis of Linear Inverse Problems with Applications in Economics and Finance', *PhD Dissertation* - Université de Science Sociales, Toulouse.
[36] Tsybakov, A.B. (2004), 'Introduction à l'estimation non-paramétrique', Springer.
[37] Van der Vaart, A.W. and J.H., Van Zanten (2007), 'Bayesian Inference with Rescaled Gaussian Process Priors', *Electronic Journal of Statistics*, 1, 433 - 448.
[38] Van der Vaart, A.W. and J.H., Van Zanten (2008), 'Rates of Contraction of Posterior Distributions Based on Gaussian Process Priors', *Annals of Statistics*, Vol.36, **3**, 1435–1463.
[39] Van der Vaart, A.W. and J.H., Van Zanten (2008), 'Reproducing Kernel Hilbert Spaces of Gaussian Priors', volume 3 of *IMS Collections*, pages 220 - 222. Institute of Mathematical Statistics, 2008.
[40] Van der Vaart, A.W. and J.H., Van Zanten (2009), 'Adaptive Bayesian Estimation Using a Gaussian Random Field with Inverse Gamma Bandwidth', *Annals of Statistics*, Vol. 37, **5B**, 2655–2675.
[41] Van der Vaart, A.W. and J.A., Wellner (1996), 'Weak Convergence and Empirical Processes', Springer, New York.
[42] Van Rooij, A.C.M. and F.H., Ruymgaart (1999), 'On inverse estimation', in: Ghosh, S. (Ed) *Asymptotic, Nonparametric and Time Series*, pp. 579-613 (Dekker).
[43] Zellner, A. (1986), 'On Assessing Prior Distributions and Bayesian Regression Analysis with g-prior Distribution', in: Goel, P.K. and Zellner, A. (Eds) *Bayesian Inference and Decision Techniques: essays in honour of Bruno de Finetti*, pp. 233-243 (Amsterdam, North Holland).
[44] Zhang, C.-H. (2005), 'General empirical Bayes wavelet methods and exactly adaptive minimax estimation', *Annals of Statistics*, **33**, No.1, 54–100.
[45] Zhao, L.H. (2000), 'Bayesian aspects of some nonparametric problems', *Annals of Statistics*, **28**, 532–552.

Toulouse School of Economics
Université Toulouse 1 Capitole
21, allée de Brienne
31000 Toulouse (France)
E-mail: florens@cict.fr

Department of Decision Sciences
Università Bocconi
via Röntgen, 1
20136 Milano (Italy)
E-mail: anna.simoni@unibocconi.it