Bayesian Analysis (2016) **11**, Number 3, pp. 671–695

On Bayesian A- and D-Optimal Experimental Designs in Infinite Dimensions

Alen Alexanderian[*,†], Philip J. Gloor[‡], and Omar Ghattas[§,¶,‖]

Abstract. We consider Bayesian linear inverse problems in infinite-dimensional separable Hilbert spaces, with a Gaussian prior measure and additive Gaussian noise model, and provide an extension of the concept of Bayesian D-optimality to the infinite-dimensional case. To this end, we derive the infinite-dimensional version of the expression for the Kullback–Leibler divergence from the posterior measure to the prior measure, which is subsequently used to derive the expression for the expected information gain. We also study the notion of Bayesian A-optimality in the infinite-dimensional setting, and extend the well known (in the finite-dimensional case) equivalence of the Bayes risk of the MAP estimator with the trace of the posterior covariance, for the Gaussian linear case, to the infinite-dimensional Hilbert space case.

Keywords: Bayesian inference in Hilbert space, Gaussian measure, Kullback–Leibler divergence, Bayesian optimal experimental design, expected information gain, Bayes risk.

1 Introduction

In a Bayesian inference problem, one uses experimental (observed) data to update the prior state of knowledge about a parameter, which often specifies certain properties of a mathematical model. The ingredients of a Bayesian inference problem include the prior measure, which encodes our prior knowledge about the inference parameter, experimental data, and the data likelihood, which describes the conditional distribution of the experimental data for a given model parameter. The solution of a Bayesian inference problem is a posterior probability law for the inference parameter. The quality of this solution, which can be measured using different criteria, depends to a large extent on the experimental data used in solving the inference problem. In practice, acquisition of such experimental data is often costly, as it requires deployment of scarce resources. Hence, the problem of optimal collection of experimental data, i.e., that of optimal experimental design (OED) (Atkinson and Donev, 1992; Uciński, 2005; Pukelsheim, 2006), is an integral part of modeling and decision making under uncertainty. The basic problem of OED is to optimize a function of the experimental setup that describes, in a certain sense, which needs to be specified, the statistical quality of the solution to the Bayesian

[*]The Institute for Computational Engineering and Sciences, The University of Texas at Austin
[†]Now at: Department of Mathematics, North Carolina State University, alexanderian@ncsu.edu
[‡]Mathematics Department, United States Naval Academy, gloor@usna.edu
[§]The Institute for Computational Engineering and Sciences, The University of Texas at Austin, omar@ices.utexas.edu
[¶]Department of Geological Sciences, The University of Texas at Austin, Austin, TX, USA
[‖]Department of Mechanical Engineering, The University of Texas at Austin, Austin, TX, USA

 DOI: 10.1214/15-BA969

inference problem. Note that what constitutes an experimental design depends on the application at hand. For example, in a problem involving diffusive transport of a contaminant, one may use measurements of concentration at sensor sites in the physical domain (at a certain point in time) to *infer* where the contaminant originated, i.e., the initial state of the concentration field. In this problem, an experimental design specifies the locations of the sensors in the physical domain. Note also that the inference parameter in this example, i.e., the initial concentration field, is a random function (random field) whose realizations belong to an appropriate function space.

We consider the problem of design of experiments for inference problems whose inference parameter belongs to an infinite-dimensional separable Hilbert space. This is motivated by the recent interest in the Bayesian framework for inverse problems (Stuart, 2010). A *Bayesian inverse problem* involves inference of Hilbert space valued parameters that describe physical properties of mathematical models, which are often governed by partial differential equations. Study of such problems, which requires a synthesis of ideas from inverse problem theory, PDE-constrained optimization, functional analysis, and probability and statistics, has provided a host of interesting mathematical problems with a wide range of applications. The problem of design of experiments in this infinite-dimensional setting involves optimizing functionals of experimental designs, which are defined in terms of operators on Hilbert spaces.

The precise definition of what is meant by an *optimal* design leads to the choice of a design criterion. A popular experimental design criterion, in the finite-dimensional case, is that of D-optimality, which seeks to minimize the determinant of the posterior covariance operator. The geometric intuition behind D-optimality is that of minimizing the volume of the uncertainty ellipsoid. Minimizing this determinant, however, is not meaningful in infinite dimensions, as the posterior covariance operator is a trace-class linear operator with eigenvalues that accumulate at zero. In the present work, we provide an extension of the concept of D-optimal design to the infinite-dimensional Hilbert space setting. In particular, we focus on the case of Bayesian linear inverse problems whose parameter space is an infinite-dimensional separable Hilbert space $\mathscr{H}$, and work in a conjugate Gaussian setting by assuming a Gaussian prior measure and an additive Gaussian noise model. To study the concept of D-optimality in the infinite-dimensional setting we formulate the problem as that of maximizing the expected information gain, measured by the Kullback–Leibler (KL) divergence (Kullback and Leibler, 1951) from posterior to prior. To be precise, if μ_{pr} denotes the prior measure, $\boldsymbol{y}$ is a vector of experimental data, obtained using an experimental design that is specified by a vector of design parameters $\boldsymbol{\xi}$, and $\mu_{\text{post}}^{\boldsymbol{y},\boldsymbol{\xi}}$ denotes the resulting posterior measure, then the KL divergence from posterior to prior is given by,

$$D_{\text{kl}}\left(\mu_{\text{post}}^{\boldsymbol{y},\boldsymbol{\xi}} \| \mu_{\text{pr}}\right) := \int_{\mathscr{H}} \log\left\{\frac{d\mu_{\text{post}}^{\boldsymbol{y},\boldsymbol{\xi}}}{d\mu_{\text{pr}}}\right\} d\mu_{\text{post}}^{\boldsymbol{y},\boldsymbol{\xi}}.$$

(The argument of the logarithm in the above formula is the Radon–Nikodym derivative of the posterior measure with respect to the prior measure.) The experimental design criterion is then defined by averaging $D_{\text{kl}}(\mu_{\text{post}}^{\boldsymbol{y},\boldsymbol{\xi}} \| \mu_{\text{pr}})$ over all possible experimental data.

In a Bayesian inverse problem, this averaging over experimental data can be done as follows:

$$\text{expected information gain} := \int_{\mathscr{H}} \int_{\mathscr{Y}} D_{\text{kl}} \left(\mu_{\text{post}}^{\boldsymbol{y},\boldsymbol{\xi}} \| \mu_{\text{pr}} \right) \pi_{\text{like}}(\boldsymbol{y}|u;\boldsymbol{\xi}) d\boldsymbol{y} \, \mu_{\text{pr}}(du),$$

where $\boldsymbol{\xi}$ is a fixed design vector, $\mathscr{Y}$ denotes the space of experimental data, and $\pi_{\text{like}}(\boldsymbol{y}|u;\boldsymbol{\xi})$ is the data likelihood, which specifies the distribution of $\boldsymbol{y}$ for a given $u \in \mathscr{H}$.

It is known in the finite-dimensional Gaussian linear case (i.e., an inference problem with Gaussian prior and noise distributions) that maximizing this expected information gain is equivalent to minimizing the determinant of the posterior covariance operator, i.e., the usual D-optimal design problem. While this does not directly extend to the infinite-dimensional case, it suggests a mathematically rigorous path to an infinite-dimensional analogue of Bayesian D-optimality. In the present work, we derive analytic expressions for the KL divergence from posterior to prior in a Hilbert space. This enables deriving the expression for the expected information gain, leading to the infinite-dimensional version of the Bayesian D-optimal experimental design criterion.

We also discuss another popular experimental design criterion, that of A-optimality, in the infinite-dimensional setting. An A-optimal design is one that minimizes the trace of the posterior covariance operator i.e., if $\mathcal{C}_{\text{post}}(\boldsymbol{\xi}) : \mathscr{H} \to \mathscr{H}$ denotes the posterior covariance operator corresponding to an experimental design $\boldsymbol{\xi}$, we seek to minimize $\text{tr}(\mathcal{C}_{\text{post}}(\boldsymbol{\xi}))$. In the statistics literature it is known (see, e.g., Chaloner and Verdinelli (1995)) that for a Gaussian linear inference problem in $\mathscr{H} = \mathbb{R}^n$, minimizing the trace of the posterior covariance *matrix* is equivalent to minimizing the average mean square error of the maximum a posteriori probability (MAP) estimator for the inference parameter. We provide an extension of this result to the infinite-dimensional Hilbert space setting, where we show that the trace of the posterior covariance *operator*—a positive, self-adjoint, and trace-class operator on $\mathscr{H}$—coincides with the average mean square error of the MAP estimator.

1.1 Motivation

Let us begin with an example of a Bayesian inverse problem in finite dimensions, where we seek to infer a parameter $\boldsymbol{u} \in \mathbb{R}^n$, using noisy experimental data $\boldsymbol{y} \in \mathbb{R}^q$ and a linear model that relates the experimental data and the parameter $\boldsymbol{u}$. In a Bayesian formulation, we model our prior knowledge/beliefs about $\boldsymbol{u}$ with a prior probability distribution. We consider the case where this prior distribution is a Gaussian. Let us denote a Gaussian measure with mean $\boldsymbol{m}$ and (symmetric positive definite) covariance matrix $\mathbf{Q}$ by μ, and recall that the measure $\mu = \mathcal{N}(\boldsymbol{m}, \mathbf{Q})$ admits the Lebesgue density,

$$\frac{d\mu}{d\boldsymbol{x}}(\boldsymbol{x}) = \frac{1}{\sqrt{(2\pi)^n |\mathbf{Q}|}} \exp\left\{ -\frac{1}{2} \langle \boldsymbol{x} - \boldsymbol{m}, \mathbf{Q}^{-1}(\boldsymbol{x} - \boldsymbol{m}) \rangle \right\},$$

where by $\frac{d\mu}{d\boldsymbol{x}}$ we denote the Radon–Nikodym derivative of μ with respect to the n-dimensional Lebesgue measure. We consider a prior law $\mu_0 = \mathcal{N}(\boldsymbol{m}_0, \mathbf{C}_0)$ for $\boldsymbol{u}$ and assume that

$$\boldsymbol{y} = \mathbf{A}\boldsymbol{u} + \boldsymbol{\eta}, \tag{1}$$

where $\mathbf{A}\colon \mathbb{R}^n \to \mathbb{R}^q$ is a $q \times n$ matrix, and $\boldsymbol{\eta}$ is a random vector that accounts for the experimental noise. Furthermore, we assume a Gaussian noise model, i.e., $\boldsymbol{\eta} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Gamma}_{\text{noise}})$, where $\boldsymbol{\Gamma}_{\text{noise}}$ is the (symmetric positive definite) noise covariance matrix.

To solve the Bayesian inverse problem means to compute a posterior probability law for the parameter $\boldsymbol{u}$, given an experimental value of $\boldsymbol{y}$. By Bayes' theorem, the probability density function (pdf) of $\boldsymbol{u}$ conditioned on $\boldsymbol{y}$, i.e., the posterior pdf, is

$$\pi(\boldsymbol{u}|\boldsymbol{y}) = \frac{\pi(\boldsymbol{u}, \boldsymbol{y})}{\pi(\boldsymbol{y})} = \frac{\pi(\boldsymbol{y}|\boldsymbol{u})\,\pi(\boldsymbol{u})}{\int_{\mathbb{R}^n} \pi(\boldsymbol{y}|\boldsymbol{u})\,\pi(\boldsymbol{u})\,d\boldsymbol{u}}, \tag{2}$$

where $\pi(\boldsymbol{u}, \boldsymbol{y})$ is the joint pdf of $\boldsymbol{u}$ and $\boldsymbol{y}$. In the present Gaussian linear case, substituting the respective Gaussian pdfs for the prior $\pi(\boldsymbol{u})$ and the likelihood $\pi(\boldsymbol{y}|\boldsymbol{u})$ in (2) one can show that

$$\pi(\boldsymbol{u}|\boldsymbol{y}) = \frac{1}{\sqrt{(2\pi)^n|\mathbf{C}|}} \exp\Big\{-\frac{1}{2}\langle \boldsymbol{u} - \boldsymbol{m}, \mathbf{C}^{-1}(\boldsymbol{u} - \boldsymbol{m})\rangle\Big\},$$

where $\boldsymbol{m} = \mathbf{C}\big(\mathbf{A}^*\boldsymbol{\Gamma}_{\text{noise}}^{-1}\boldsymbol{y} + \mathbf{C}_0^{-1}\boldsymbol{m}_0\big)$ and $\mathbf{C}^{-1} = \mathbf{C}_0^{-1} + \mathbf{A}^*\boldsymbol{\Gamma}_{\text{noise}}^{-1}\mathbf{A}$; see, e.g., Tarantola (2005, Chapter 3).

In a problem of optimal design of experiments, one seeks an experimental setup that can be used to collect data $\boldsymbol{y}$, from which the parameter $\boldsymbol{u}$ is optimally inferred. An experimental design is usually identified with a vector of experimental design parameters, which, as before, we denote by $\boldsymbol{\xi}$. The vector $\boldsymbol{\xi}$ enters the Bayesian inverse problem through the data likelihood $\pi(\boldsymbol{y}|\boldsymbol{u})$. The exact nature of this dependence on $\boldsymbol{\xi}$ is not essential to our discussion so, for notational convenience, we suppress the vector $\boldsymbol{\xi}$ in our derivations. (See, e.g., Chaloner and Verdinelli (1995) for an overview of how an experimental design is incorporated in an inference problem in classical formulations.)

A D-optimal design is one that aims to minimize the volume of the uncertainty ellipsoid; this is achieved by minimizing $\log\det(\mathbf{C})$. An A-optimal design, on the other hand, minimizes $\mathrm{tr}(\mathbf{C})$, which results in minimized average posterior variance. It is known that, in the finite-dimensional case, minimizing $\log\det(\mathbf{C})$ is equivalent to maximizing the expected information gain,

$$\int_{\mathbb{R}^n}\int_{\mathbb{R}^q} D_{\text{kl}}\left(\pi(\boldsymbol{u}|\boldsymbol{y})\|\pi(\boldsymbol{u})\right)\,\pi(\boldsymbol{y}|\boldsymbol{u})d\boldsymbol{y}\,\pi(\boldsymbol{u})d\boldsymbol{u}, \tag{3}$$

where $D_{\text{kl}}\left(\pi(\boldsymbol{u}|\boldsymbol{y})\|\pi(\boldsymbol{u})\right) = \int_{\mathbb{R}^n} \log(\pi(\boldsymbol{u}|\boldsymbol{y})/\pi(\boldsymbol{u}))\,\pi(\boldsymbol{u}|\boldsymbol{y})\,d\boldsymbol{u}$. (Notice that in this finite-dimensional setting, we can write the expression for the KL divergence and the expected information gain in terms of the posterior and prior pdfs.) On the other hand, in the case of Bayesian A-optimality, it is known in the finite-dimensional case that minimization of $\mathrm{tr}(\mathbf{C})$ is equivalent to minimization of the average mean square error (see Section 5) of the estimator $\boldsymbol{m} = \mathbf{C}(\mathbf{A}^*\boldsymbol{\Gamma}_{\text{noise}}^{-1}\boldsymbol{y} + \mathbf{C}_0^{-1}\boldsymbol{m}_0)$ of $\boldsymbol{u}$.

The topic of the present paper is optimal design of experiments for Bayesian linear inverse problems, where the inference parameter belongs to an infinite-dimensional Hilbert space. In particular, we study Bayesian A- and D-optimality for such problems. It is instructive to consider an example of an infinite-dimensional Bayesian inverse problem to motivate the discussion that follows. Consider the following time-dependent partial differential equation (PDE) that describes the conduction of heat in a physical domain $\mathcal{D}$ with piecewise smooth boundary $\partial\mathcal{D}$:

$$\begin{aligned} y_t(\boldsymbol{x},t) &= -\nabla\cdot\big(\kappa(\boldsymbol{x})\nabla y(\boldsymbol{x},t)\big) && (\boldsymbol{x},t)\in\mathcal{D}\times[0,t_{\text{final}}],\\ y(\boldsymbol{x},0) &= u(\boldsymbol{x}) && \boldsymbol{x}\in\mathcal{D},\\ \kappa\nabla y(\boldsymbol{x},t)\cdot\boldsymbol{n} &= 0 && (\boldsymbol{x},t)\in\partial\mathcal{D}\times[0,t_{\text{final}}]. \end{aligned}$$

Here, $y(\boldsymbol{x},t)$ is the temperature at a given point $\boldsymbol{x}\in\mathcal{D}$ and at time t, y_t denotes the partial derivative of y with respect to t, and the function $\kappa(\boldsymbol{x})$ describes the thermal diffusivity of the medium. The function $u(\boldsymbol{x})$ is the initial state, which, in the present example, we consider to be the uncertain parameter that we wish to estimate. In particular, suppose that we have placed temperature sensors at points $\{\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_{n_s}\}$ in $\mathcal{D}$ and take measurements at times $\{t_1,t_2,\ldots,t_{n_t}\}$ in $[0,t_{\text{final}}]$. The problem of using this temperature data to infer the initial state $u(\boldsymbol{x})$, through a Bayesian formulation, is an example of a Bayesian inverse problem with a inversion parameter that lives in an infinite-dimensional function space.

Notice that the familiar form of the Bayes formula given in (2), in terms of pdfs (Lebesgue densities), is not meaningful in this case, because there is no Lebesgue measure in infinite dimensions. Thus, the formulation of the Bayesian inverse problem in a function space requires care. For Gaussian linear inverse problems in infinite dimensions, it is known (Stuart, 2010) that analogous expressions for the mean and covariance operator of the posterior measure can be derived, albeit in terms of Hilbert space operators. It is important to note that the study of Bayesian inverse problems in infinite dimensions is not only an interesting theoretical exercise, but also has important implications in practical computations. Namely, to derive appropriate formulations and consistent discretizations of such Bayesian inverse problems, suitable for computer implementations, a careful understanding of these problems in infinite dimensions is essential. We refer to the article Bui-Thanh et al. (2013), which describes consistent discretizations and numerical algorithms for solution of infinite-dimensional Bayesian linear inverse problems that are governed by PDEs.

1.2 Paper overview

Our goal is to study the Bayesian A- and D-optimal experimental design criteria in infinite dimensions. To facilitate the discussion, in Section 2 we recall the background material from analysis and probability in infinite dimensions that is needed in the rest of the article. Next, we discuss the infinite-dimensional formulation of Bayesian inverse problems in Section 3. We augment the presentation in that section with a brief discussion of the directions in the parameter space where significant uncertainty reduction occurs in the inference process. In Section 4, we study Bayesian D-optimality in infinite dimensions, where we derive expressions for the KL divergence from posterior to prior, and derive the analogue of Bayesian D-optimality in the infinite-dimensional Hilbert space setting. Finally, in Section 5, we study Bayesian A-optimality in infinite dimensions, where we extend the known result relating the trace of the posterior covariance and the Bayes risk of the MAP estimator to infinite dimensions.

2 Background concepts

In this section, we outline the background concepts that are needed in the rest of this article. In what follows, $\mathscr{H}$ denotes an infinite-dimensional separable real Hilbert space, with inner-product $\langle \cdot, \cdot \rangle_{\mathscr{H}}$ and induced norm $\|\cdot\|_{\mathscr{H}} = \langle \cdot, \cdot \rangle_{\mathscr{H}}^{1/2}$.

2.1 Trace-class operators on $\mathscr{H}$

Let $\mathscr{L}(\mathscr{H})$ denote the set of bounded linear operators on $\mathscr{H}$. For $\mathcal{A} \in \mathscr{L}(\mathscr{H})$, $|\mathcal{A}| = (\mathcal{A}^*\mathcal{A})^{1/2}$, where $\mathcal{A}^*$ denotes the adjoint of $\mathcal{A}$. We say $\mathcal{A}$ is of *trace-class* if for any orthonormal basis $\{f_j\}_{j=1}^{\infty}$ of $\mathscr{H}$,

$$\sum_{j=1}^{\infty} \langle |\mathcal{A}| f_j, f_j \rangle_{\mathscr{H}} < \infty.$$

It is straightforward to show that the value of the above summation is invariant with respect to the choice of the orthonormal basis (Reed and Simon, 1972). We denote by $\mathscr{L}_1(\mathscr{H})$ the subspace of $\mathscr{L}(\mathscr{H})$, consisting of trace-class operators. For $\mathcal{A} \in \mathscr{L}_1(\mathscr{H})$,

$$\mathrm{tr}(\mathcal{A}) = \sum_{j=1}^{\infty} \langle \mathcal{A} f_j, f_j \rangle_{\mathscr{H}},$$

where the sum is finite and its value is independent of the choice of the orthonormal basis (Conway, 2000; Reed and Simon, 1972).

We say that a linear self-adjoint operator $\mathcal{A}$ is positive if $\langle x, \mathcal{A}x \rangle_{\mathscr{H}} \geq 0$ for all $x \in \mathscr{H}$, and is strictly positive if $\langle x, \mathcal{A}x \rangle_{\mathscr{H}} > 0$ for all nonzero $x \in \mathscr{H}$. Let $\mathscr{L}_1^{\mathrm{sym}+}(\mathscr{H})$ be the subspace of positive self-adjoint operators in $\mathscr{L}_1(\mathscr{H})$, and note that for $\mathcal{A} \in \mathscr{L}_1^{\mathrm{sym}+}(\mathscr{H})$, there exists an orthonormal basis of eigenvectors, $\{e_j\}$, with corresponding (real, non-negative) eigenvalues, $\{\lambda_j\}$, and $\mathrm{tr}(A) = \sum_{j=1}^{\infty} \langle \mathcal{A} e_j, e_j \rangle_{\mathscr{H}} = \sum_{j=1}^{\infty} \lambda_j$. We also recall that for $\mathcal{A}$ in $\mathscr{L}_1^{\mathrm{sym}+}(\mathscr{H})$ (or more generally to the space of positive self-adjoint compact operators), the *square root* of $\mathcal{A}$ is defined as follows,

$$\mathcal{A}^{1/2} x = \sum_{j=1}^{\infty} \lambda_j^{1/2} \langle e_j, x \rangle_{\mathscr{H}} e_j, \quad x \in \mathscr{H}.$$

In what follows we shall make repeated use of the following result: if $\mathcal{A} \in \mathscr{L}_1(\mathscr{H})$ and $\mathcal{B} \in \mathscr{L}(\mathscr{H})$ then $\mathcal{AB}$ and $\mathcal{BA}$ both belong to $\mathscr{L}_1(\mathscr{H})$ and $\mathrm{tr}(\mathcal{AB}) = \mathrm{tr}(\mathcal{BA})$; see, e.g., Reed and Simon (1972), Da Prato and Zabczyk (2002, Chapter 1), or Gel'fand and Vilenkin (1964). Moreover, it is straightforward to show that if $\mathcal{A}$ is a trace-class operator and $\mathcal{B} : \mathscr{H} \to \mathbb{R}^q$ is a bounded linear operator, then $\mathcal{AB}^*\mathcal{B} \in \mathscr{L}_1(\mathscr{H})$ and $\mathrm{tr}(\mathcal{AB}^*\mathcal{B}) = \mathrm{tr}(\mathcal{BAB}^*)$.

2.2 Borel probability measures on $\mathscr{H}$

We work with probability measures on the measurable space $(\mathscr{H}, \mathfrak{B}(\mathscr{H}))$, where $\mathfrak{B}(\mathscr{H})$ denotes the Borel sigma-algebra on $\mathscr{H}$; we refer to such measures as Borel probability measures. Let μ be a Borel probability measure on $\mathscr{H}$ with bounded first and second moments. The mean $m \in \mathscr{H}$ and covariance operator $\mathcal{Q} \in \mathscr{L}(\mathscr{H})$ of μ are characterized as follows:

$$\langle m, x \rangle_{\mathscr{H}} = \int_{\mathscr{H}} \langle z, x \rangle_{\mathscr{H}} \, \mu(dz), \qquad \langle \mathcal{Q}x, y \rangle_{\mathscr{H}} = \int_{\mathscr{H}} \langle x, z - m \rangle_{\mathscr{H}} \langle y, z - m \rangle_{\mathscr{H}} \, \mu(dz),$$

for all $x, y \in \mathscr{H}$. It is straightforward to show (see, e.g., Da Prato (2006)) that $\mathcal{Q}$ belongs to $\mathscr{L}_1^{\text{sym}+}(\mathscr{H})$, and that

$$\int_{\mathscr{H}} \|x\|_{\mathscr{H}}^2 \, \mu(dx) = \operatorname{tr}(\mathcal{Q}) + \|m\|_{\mathscr{H}}^2 \,. \tag{4}$$

Let us also recall the notion of the Fourier transform of a Borel measure μ on $\mathscr{H}$. The function $\hat{\mu} : \mathscr{H} \to \mathbb{R}$ given by,

$$\hat{\mu}(\xi) = \int_{\mathscr{H}} e^{i\langle x, \xi \rangle} \, \mu(dx),$$

is called the Fourier transform of μ. It is known that the Fourier transform $\hat{\mu}$ uniquely determines μ; that is, if μ and ν are two Borel probability measures on $\mathscr{H}$ such that $\hat{\mu}(\xi) = \hat{\nu}(\xi)$ for all $\xi \in \mathscr{H}$ then $\mu = \nu$; see (Da Prato, 2006, Proposition 1.7).

2.3 Gaussian measures on $\mathscr{H}$

In the present work, we shall be working with Gaussian measures on Hilbert spaces (Da Prato, 2006); μ is a Gaussian measure on $(\mathscr{H}, \mathfrak{B}(\mathscr{H}))$ if for every $x \in \mathscr{H}$ the linear functional $\langle x, \cdot \rangle_{\mathscr{H}}$, considered as a random variable from $(\mathscr{H}, \mathfrak{B}(\mathscr{H}), \mu)$ to $(\mathbb{R}, \mathfrak{B}(\mathbb{R}))$, is a (one-dimensional) Gaussian random variable. Given $m \in \mathscr{H}$ and $\mathcal{Q} \in \mathscr{L}_1^{\text{sym}+}(\mathscr{H})$, the Gaussian measure $\mu = \mathcal{N}(m, \mathcal{Q})$ is the unique probability measure with

$$\hat{\mu}(\xi) = \exp\left\{ i\langle m, \xi \rangle - \frac{1}{2}\langle Q\xi, \xi \rangle \right\}, \quad \xi \in \mathscr{H}.$$

We refer the reader to Da Prato (2006), Da Prato and Zabczyk (2002), or Da Prato and Zabczyk (2014) for the theory of Gaussian measures on Hilbert spaces. If the covariance operator $\mathcal{Q}$ satisfies $\ker(\mathcal{Q}) = \{0\}$, where $\ker(\mathcal{Q})$ denotes the null space of $\mathcal{Q}$, we say that $\mathcal{N}(m, \mathcal{Q})$ is a non-degenerate Gaussian measure.

In what follows, we shall use the following result, concerning the law of an affine transformation on $\mathscr{H}$: If $\mu = \mathcal{N}(m, \mathcal{Q})$ is a Gaussian measure on $\mathscr{H}$, $\mathcal{A} \in \mathscr{L}(\mathscr{H}, \mathscr{K})$ where $\mathscr{K}$ is also a Hilbert space, and $b \in \mathscr{K}$, then $Tx = \mathcal{A}x + b$ is a random variable on $\mathscr{H}$ whose law (a probability measure on $\mathscr{K}$) is given by $\mu_T = \mu \circ T^{-1} = \mathcal{N}(\mathcal{A}m + b, \mathcal{A}\mathcal{Q}\mathcal{A}^*)$ (Da Prato, 2006). Using this result, we note that for $z_1, \ldots, z_n \in$

$\mathscr{H}$ the mapping $\mathcal{A}x = (\langle z_1, x\rangle_{\mathscr{H}}, \ldots, \langle z_n, x\rangle_{\mathscr{H}})$, a linear transformation from $\mathscr{H}$ to $\mathscr{K} = \mathbb{R}^n$, has law $\mathcal{N}(\boldsymbol{a}, \boldsymbol{\Sigma})$ with $\boldsymbol{a} \in \mathbb{R}^n$ and $\boldsymbol{\Sigma} \in \mathbb{R}^{n\times n}$ given by

$$a_i = \langle z_i, m\rangle_{\mathscr{H}}, \quad \Sigma_{ij} = \langle \mathcal{Q}z_i, z_j\rangle_{\mathscr{H}}, \quad i,j \in \{1,\ldots,n\}.$$

See (Da Prato, 2006, Corollary 1.19) for derivation of this.

We also note that, for $Tx = \mathcal{A}x + b$ with $\mathcal{A} \in \mathscr{L}(\mathscr{H})$ and $b \in \mathscr{H}$,

$$\int_{\mathscr{H}} \|Tx\|^2_{\mathscr{H}}\ \mu(dx) = \int_{\mathscr{H}} \|\xi\|^2_{\mathscr{H}}\ \mu_T(d\xi) = \mathrm{tr}(\mathcal{A}\mathcal{Q}\mathcal{A}^*) + \|\mathcal{A}m + b\|^2_{\mathscr{H}},$$

where the last equality uses (4). It follows that if $\mathcal{A} \in \mathscr{L}(\mathscr{H})$ is a positive, self-adjoint compact operator, and $\mu = \mathcal{N}(m, \mathcal{Q})$ is a Gaussian measure, then

$$\begin{aligned}\int_{\mathscr{H}} \langle \mathcal{A}x, x\rangle_{\mathscr{H}}\ \mu(dx) &= \int_{\mathscr{H}} \left\|\mathcal{A}^{1/2}x\right\|^2_{\mathscr{H}}\ \mu(dx) \\ &= \mathrm{tr}(\mathcal{A}^{1/2}\mathcal{Q}\mathcal{A}^{1/2}) + \left\langle \mathcal{A}^{1/2}m, \mathcal{A}^{1/2}m\right\rangle_{\mathscr{H}} \\ &= \mathrm{tr}(\mathcal{A}\mathcal{Q}) + \langle \mathcal{A}m, m\rangle_{\mathscr{H}}. \qquad (5)\end{aligned}$$

This shows that the well-known expression for the expectation of a quadratic form on $\mathbb{R}^n$ extends to the infinite-dimensional Hilbert space setting. It can be shown that, as in the finite-dimensional case, this result holds not just for Gaussian measures, but also for any Borel probability measure with mean m and covariance operator $\mathcal{Q}$; moreover, the only requirement on the operator $\mathcal{A}$ is boundedness. That is, we have the following result:

Lemma 1. *Let μ be a Borel probability measure on $\mathscr{H}$ with mean $m \in \mathscr{H}$ and covariance operator $\mathcal{Q} \in \mathscr{L}_1^{sym+}(\mathscr{H})$, and let $\mathcal{A} \in \mathscr{L}(\mathscr{H})$. Then,*

$$\int_{\mathscr{H}} \langle \mathcal{A}x, x\rangle_{\mathscr{H}}\ \mu(dx) = \mathrm{tr}(\mathcal{A}\mathcal{Q}) + \langle \mathcal{A}m, m\rangle_{\mathscr{H}}.$$

Proof. See Appendix A. □

2.4 Kullback–Leibler divergence

In probability theory the Kullback–Leibler (KL) divergence, also referred to as the relative entropy, is a measure of "distance" between two probability measures. This notion was defined in Kullback and Leibler (1951). While the KL divergence is not a metric—it is non-symmetric and does not satisfy the triangle inequality—it is used commonly in probability theory and mathematical statistics to describe the distance of a measure μ from a reference measure μ_0. In particular, in Bayesian statistics, it is common to use the KL divergence from the posterior measure to the prior measure to quantify the information gain in the inference process. Moreover, the KL divergence does satisfy some of the intuitive notions of distance, i.e., the KL divergence from μ to μ_0 is non-negative and is zero if and only if the two measures are the same.

Let μ and μ_0 be two Borel probability measures and suppose μ is absolutely continuous with respect to μ_0. The KL divergence from μ to μ_0, denoted by $D_{\text{kl}}(\mu\|\mu_0)$, is defined as

$$D_{\text{kl}}(\mu\|\mu_0) = \int_{\mathscr{H}} \log\left\{\frac{d\mu}{d\mu_0}\right\} d\mu.$$

Here $\frac{d\mu}{d\mu_0}$ denotes the Radon–Nikodym derivative of μ with respect to μ_0. In the case μ is not absolutely continuous with respect to μ_0, the KL divergence is $+\infty$. Notice that for Borel probability measures on $\mathbb{R}^n$ that admit densities with respect to the Lebesgue measure, we may rewrite the definition of the KL divergence in terms the densities; that is, if p and p_0 are Lebesgue densities, i.e., probability density functions (pdfs), of μ and μ_0, respectively, one has $D_{\text{kl}}(\mu\|\mu_0) = \int_{\mathbb{R}^n} \log(p(\boldsymbol{x})/p_0(\boldsymbol{x}))\, p(\boldsymbol{x})\, d\boldsymbol{x}$. However, in an infinite-dimensional Hilbert space, where there is no Lebesgue measure, we are forced to work with the abstract definition of the KL divergence presented above.

In this paper, we shall be dealing with Gaussian measures on infinite-dimensional Hilbert spaces. For Gaussian measures on $\mathbb{R}^n$, one can use the expression for the (multivariate) Gaussian pdfs to derive the well-known analytic expression for the KL divergence between Gaussians. In the infinite-dimensional Hilbert space setting, not only do we not have access to pdfs, but also any two given Gaussian measures may not be equivalent.[1] In fact, if we consider a centered (zero-mean) Gaussian measure $\mu = \mathcal{N}(0, \mathcal{C})$, and its shifted version $\mu' = \mathcal{N}(m, \mathcal{C})$, it is known that μ and μ' are either singular or equivalent, and that μ and μ' are equivalent if and only if the shift m belongs to the space $\mathscr{H}_\mu = \mathsf{range}(\mathcal{C}^{1/2})$. For a Gaussian measure, the space $\mathscr{H}_\mu$ so defined is called the Cameron–Martin space associated to μ. It is a known result (see, e.g., Da Prato (2006)) that if the Hilbert space $\mathscr{H}$ is infinite-dimensional, then $\mu(\mathsf{range}(\mathcal{C}^{1/2})) = 0$. Thus we see that for a Gaussian measure μ on $\mathscr{H}$, shifting the mean gives, μ-almost surely, a Gaussian measure that is singular with respect to μ; see, e.g., Da Prato (2006, Chapter 2). More generally, the conditions for the equivalence of two Gaussian measures is specified by the Feldman–Hajek Theorem (Da Prato and Zabczyk, 2002, 2014).

In the present study, we work with a special case, namely that of a Bayesian linear inverse problem on $\mathscr{H}$ with a Gaussian prior, an additive Gaussian noise model, and finite-dimensional observations; in this case the posterior measure is also Gaussian and is equivalent to the prior (Stuart, 2010), and thus, $D_{\text{kl}}(\mu_{\text{post}}^{\boldsymbol{y}}\|\mu_{\text{pr}})$ is well-defined. Later in the paper, we shall derive the expression for the KL divergence from posterior to prior in an infinite-dimensional Hilbert space, which we shall use to derive the expression for the expected information gain.

3 Bayesian linear inverse problems in a Hilbert space

We consider the problem of inference of a parameter u, which belongs to an infinite-dimensional Hilbert space $\mathscr{H}$. All our prior knowledge regarding the parameter u is encoded in a Borel probably measure on $\mathscr{H}$, which we refer to as the prior measure

[1] Recall that two measures are called equivalent if they are mutually absolutely continuous with respect to each other.

and denote by μ_{pr}; here we assume that μ_{pr} is a Gaussian measure $\mu_{\text{pr}} = \mathcal{N}(u_{\text{pr}}, \mathcal{C}_{\text{pr}})$.[2] Moreover, in what follows, we assume that $\ker(\mathcal{C}_{\text{pr}}) = \{0\}$, i.e., μ_{pr} is non-degenerate. The inference problem uses experimental data $\boldsymbol{y} \in \mathscr{Y}$ to update the prior state of knowledge of the law of the parameter u. Here $\mathscr{Y}$ is the space of the experimental data, which in the present work is $\mathscr{Y} = \mathbb{R}^q$ (endowed with the Euclidean inner product). We assume that u is a model parameter, which is related to experimental data $\boldsymbol{y} \in \mathscr{Y}$, according to the following noise model,

$$\boldsymbol{y} = \mathcal{G}u + \boldsymbol{\eta}. \tag{6}$$

The operator $\mathcal{G} : \mathscr{H} \to \mathscr{Y}$ is the *parameter-to-observable map* and is assumed to be a continuous linear mapping. In practice, for a given u, computing $\mathcal{G}u$ would involve the evaluation of a mathematical model with the parameter value u followed by the application of a restriction operator to extract data at prespecified locations in space and/or time. The discrepancy between the model output $\mathcal{G}u$ and experimental data $\boldsymbol{y}$ is modeled by $\boldsymbol{\eta}$, which is a random vector that accounts for experimental noise, i.e., noise associated with the process of collecting experimental data. We assume $\boldsymbol{\eta} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Gamma}_{\text{noise}})$, and thus, the distribution of $\boldsymbol{y}|u$ is Gaussian, $\boldsymbol{y}|u \sim \mathcal{N}(\mathcal{G}u, \boldsymbol{\Gamma}_{\text{noise}})$ with pdf

$$\pi_{\text{like}}(\boldsymbol{y}|u) = \frac{1}{\mathcal{Z}_{\text{like}}} \exp\Big\{ -\frac{1}{2}(\mathcal{G}u - \boldsymbol{y})^T \boldsymbol{\Gamma}_{\text{noise}}^{-1}(\mathcal{G}u - \boldsymbol{y})\Big\},$$

where $\mathcal{Z}_{\text{like}} = (2\pi)^{q/2} \det(\boldsymbol{\Gamma}_{\text{noise}})^{1/2}$. Note that in general $\boldsymbol{\Gamma}_{\text{noise}}$ is a symmetric positive definite matrix. In the present work, we focus on the case that the noise covariance matrix is known, based on knowledge of the experimental noise levels and the correlation lengths. This is a common assumption especially in the context of Bayesian inverse problems that are governed by PDE models, where in addition, one often assumes uncorrelated observations. An estimate of the noise levels in such cases is often available based on the error tolerances of the measurement devices that are used to collect data.

3.1 Bayes' formula and the posterior measure

The solution of the Bayesian inverse problem is the posterior measure, describing the law of the parameter u, conditioned on the experimental data $\boldsymbol{y}$, and is linked to the

[2]To specify a Gaussian prior measure, we need to specify its mean and covariance. A computationally tractable method of specifying prior covariance operators in Bayesian inverse problems in infinite dimensions, with $\mathscr{H} = L^2(\mathcal{D})$ where $\mathcal{D}$ is a bounded open set in $\mathbb{R}^d$, is to define them as inverses of differential operators. The information about the size of the variance and correlation lengths can be built into the covariance operator by considering inverses of appropriately chosen Laplacian-like operators; we refer to Stuart (2010) for theory and analytic examples and Bui-Thanh et al. (2013) for computational aspects of specifying such covariance operators. Gaussian priors so constructed are sometimes referred to as smoothing (or regularizing) priors, due to the smoothing properties of their covariance operators. The prior mean can be specified by a sufficiently regular element of $\mathscr{H}$ that reflects our prior belief about the inversion parameter. (More precisely, we require that the prior mean belong to the Cameron–Martin space $\mathsf{range}(\mathcal{C}_{\text{pr}}^{1/2})$.) See also Lindgren et al. (2011), where the authors develop theory and computational methods for constructing Matérn-type Gaussian processes.

prior measure μ_{pr} through the infinite-dimensional version of Bayes' theorem (Stuart, 2010):

$$\frac{d\mu_{\text{post}}^{\boldsymbol{y}}}{d\mu_{\text{pr}}} = \frac{1}{\mathcal{Z}(\boldsymbol{y})}\pi_{\text{like}}(\boldsymbol{y}|u), \tag{7}$$

where $\mathcal{Z}(\boldsymbol{y})$ is the normalization constant. It is convenient to use the notation

$$\Phi(u;\boldsymbol{y}) = \frac{1}{2}(\mathcal{G}u - \boldsymbol{y})^T\boldsymbol{\Gamma}_{\text{noise}}^{-1}(\mathcal{G}u - \boldsymbol{y}), \tag{8}$$

so that $\pi_{\text{like}}(\boldsymbol{y}|u) \propto \exp\{-\Phi(u;\boldsymbol{y})\}$. This enables rewriting Bayes' theorem as follows:

$$\frac{d\mu_{\text{post}}^{\boldsymbol{y}}}{d\mu_{\text{pr}}} = \frac{1}{\mathcal{Z}_0(\boldsymbol{y})}\exp\{-\Phi(u;\boldsymbol{y})\}, \tag{9}$$

with $\mathcal{Z}_0(\boldsymbol{y}) = \int_{\mathscr{H}} \exp\{-\Phi(u;\boldsymbol{y})\}\,\mu_{\text{pr}}(du)$. In the Gaussian linear case, it is possible to evaluate $\mathcal{Z}_0(\boldsymbol{y})$ analytically; see Lemma 2 below.

As discussed above, we consider Bayesian linear inverse problems, i.e., Bayesian inverse problems involving a linear parameter-to-observable map $\mathcal{G}$. It is known (Stuart, 2010) that for a Gaussian linear inverse problem, as specified above, the solution is a Gaussian posterior measure $\mu_{\text{post}}^{\boldsymbol{y}} = \mathcal{N}(u_{\text{post}}^{\boldsymbol{y}}, \mathcal{C}_{\text{post}})$ with

$$\mathcal{C}_{\text{post}} = (\mathcal{G}^*\boldsymbol{\Gamma}_{\text{noise}}^{-1}\mathcal{G} + \mathcal{C}_{\text{pr}}^{-1})^{-1}, \qquad u_{\text{post}}^{\boldsymbol{y}} = \mathcal{C}_{\text{post}}(\mathcal{G}^*\boldsymbol{\Gamma}_{\text{noise}}^{-1}\boldsymbol{y} + \mathcal{C}_{\text{pr}}^{-1}u_{\text{pr}}).$$

Notice that these expressions resemble the usual formulas for the mean and covariance of the posterior measure in a finite-dimensional Gaussian linear inverse problem. We refer to (Stuart, 2010, Example 6.32) for justification of these formulas for the infinite-dimensional Bayesian linear inverse problems considered in the present work.

In practice, the noise covariance matrix, $\boldsymbol{\Gamma}_{\text{noise}}$ is often a multiple of the identity, $\boldsymbol{\Gamma}_{\text{noise}} = \sigma^2\mathbf{I}$, where σ is the experimental noise level. As mentioned above, we consider the case where σ is known.[3] In the derivations that follow, since there is no loss of generality, we take $\sigma = 1$. Generalizing the results to the cases where $\boldsymbol{\Gamma}_{\text{noise}}$ is an anisotropic diagonal matrix (uncorrelated observations with varying experimental noise levels) or more generally $\boldsymbol{\Gamma}_{\text{noise}}$ that is symmetric and positive definite with nonzero off diagonal entries (correlated observations) is straightforward. Moreover, for simplicity, we assume that the prior is a centered Gaussian, i.e., $u_{\text{pr}} = 0$. Again, the generalization to the case of non-centered prior measure is straightforward. With these simplifications, the mean and covariance of the posterior measure are given by

$$\mathcal{C}_{\text{post}} = (\mathcal{G}^*\mathcal{G} + \mathcal{C}_{\text{pr}}^{-1})^{-1}, \qquad u_{\text{post}}^{\boldsymbol{y}} = \mathcal{C}_{\text{post}}\mathcal{G}^*\boldsymbol{y}. \tag{10}$$

In what follows, we use the notation

$$\mathcal{H}_{\text{m}} = \mathcal{G}^*\mathcal{G}. \tag{11}$$

[3]Note that in problems where σ^2 is unknown, one often lets σ^2 be a hyper-parameter with its own (usually non-informative) prior, and then works with a marginalized (over σ^2) posterior for the inference parameter.

The motivation behind this notation is that $\mathcal{G}^*\mathcal{G}$ is the Hessian of the functional, $\Phi(u;\boldsymbol{y})$, which measures the magnitude of the *misfit* between experimental data $\boldsymbol{y}$ and model prediction $\mathcal{G}u$. Note that in statistical terms, $\mathcal{H}_\mathrm{m}$ is the Hessian of the negative log-likelihood, which is also referred to as the Fisher information operator. Another notation we shall use frequently is

$$\tilde{\mathcal{H}}_\mathrm{m} = \mathcal{C}_\mathrm{pr}^{1/2}\mathcal{H}_\mathrm{m}\mathcal{C}_\mathrm{pr}^{1/2}. \tag{12}$$

Intuitively, this *prior-preconditioned* $\mathcal{H}_\mathrm{m}$ can be thought of as the information operator that has been filtered through the prior. To further appreciate the notion of the prior-preconditioned data misfit Hessian, we note that the second moment of the parameter-to-observable map, considered as a random variable $\mathcal{G} : (\mathscr{H}, \mathfrak{B}(\mathscr{H}), \mu_\mathrm{pr}) \to (\mathbb{R}^q, \mathfrak{B}(\mathbb{R}^q))$ is given by

$$\begin{aligned}\int_{\mathscr{H}} |\mathcal{G}u|^2\,\mu_\mathrm{pr}(du) &= \int_{\mathscr{H}} \langle \mathcal{G}u, \mathcal{G}u\rangle_{\mathbb{R}^q}\;\mu_\mathrm{pr}(du)\\ &= \int_{\mathscr{H}} \langle \mathcal{H}_\mathrm{m}u, u\rangle_{\mathscr{H}}\;\mu_\mathrm{pr}(du) = \mathrm{tr}(\mathcal{C}_\mathrm{pr}\mathcal{H}_\mathrm{m}) = \mathrm{tr}(\tilde{\mathcal{H}}_\mathrm{m}).\end{aligned}$$

3.2 A spectral point of view of uncertainty reduction

Let $\tilde{\mathcal{H}}_\mathrm{m}$ be the prior-preconditioned data misfit Hessian as defined in (12) and denote

$$\mathcal{S} = (I + \tilde{\mathcal{H}}_\mathrm{m})^{-1}. \tag{13}$$

The posterior covariance operator, $\mathcal{C}_\mathrm{post}$, given in (10), can be written as $\mathcal{C}_\mathrm{post} = \mathcal{C}_\mathrm{pr}^{1/2}(I+\tilde{\mathcal{H}}_\mathrm{m})^{-1}\mathcal{C}_\mathrm{pr}^{1/2} = \mathcal{C}_\mathrm{pr}^{1/2}\mathcal{S}\mathcal{C}_\mathrm{pr}^{1/2}$. We consider the quantity

$$\delta(\mathcal{C}_\mathrm{pr}, \mathcal{C}_\mathrm{post}) := \mathrm{tr}(\mathcal{C}_\mathrm{pr}) - \mathrm{tr}(\mathcal{C}_\mathrm{post}) = \mathrm{tr}\big(\mathcal{C}_\mathrm{pr}^{1/2}(I-\mathcal{S})\mathcal{C}_\mathrm{pr}^{1/2}\big).$$

For the class of Bayesian linear inverse problems that we consider in this work, $\delta(\mathcal{C}_\mathrm{pr}, \mathcal{C}_\mathrm{post}) \geq 0$. In particular, we note that if $\{\lambda_i\}$ and $\{e_i\}$ are the eigenvalues and the respective eigenvectors of $\tilde{\mathcal{H}}_\mathrm{m}$, then

$$\langle e_i, (I-\mathcal{S})e_i\rangle_{\mathscr{H}} = 1 - \langle e_i, \mathcal{S}e_i\rangle_{\mathscr{H}} = 1 - 1/(1+\lambda_i) = \lambda_i/(1+\lambda_i) \geq 0, \quad i = 1, 2, \ldots,$$

which shows that $\delta(\mathcal{C}_\mathrm{pr}, \mathcal{C}_\mathrm{post}) = \mathrm{tr}(\mathcal{C}_\mathrm{pr}^{1/2}(I-\mathcal{S})\mathcal{C}_\mathrm{pr}^{1/2}) \geq 0$. The quantity $\delta(\mathcal{C}_\mathrm{pr}, \mathcal{C}_\mathrm{post})$ can thus be considered a measure of variance (uncertainty) reduction. More precisely, we consider for each $i \geq 1$,

$$\langle e_i, \mathcal{C}_\mathrm{post}e_i\rangle_{\mathscr{H}} = \int_{\mathscr{H}} \big\langle e_i, u - u_\mathrm{post}^{\boldsymbol{y}}\big\rangle_{\mathscr{H}}^2\,\mu_\mathrm{post}^{\boldsymbol{y}}(du),$$

which measures the posterior variance of the coordinate of u in the direction e_i.

Proposition 1. *Let $\{\lambda_i, e_i\}_1^\infty$ be eigenpairs of $\tilde{\mathcal{H}}_m$. Then, $\langle e_i, \mathcal{C}_{post}e_i\rangle_{\mathscr{H}} \leq \langle e_i, \mathcal{C}_{pr}e_i\rangle_{\mathscr{H}}$, for all $i \geq 1$.*

Proof. Note that for each $v \in \mathscr{H}$, $\mathcal{S}v = \sum_j (1+\lambda_j)^{-1} \langle e_j, v\rangle_{\mathscr{H}} e_j$. Hence,

$$\begin{aligned}\langle e_i, \mathcal{C}_{\text{post}} e_i\rangle_{\mathscr{H}} &= \left\langle e_i, \mathcal{C}_{\text{pr}}^{1/2}\mathcal{S}\mathcal{C}_{\text{pr}}^{1/2} e_i\right\rangle_{\mathscr{H}} = \left\langle \mathcal{C}_{\text{pr}}^{1/2} e_i, \mathcal{S}\mathcal{C}_{\text{pr}}^{1/2} e_i\right\rangle_{\mathscr{H}} \\ &= \sum_j (1+\lambda_j)^{-1} \left\langle e_j, \mathcal{C}_{\text{pr}}^{1/2} e_i\right\rangle_{\mathscr{H}}^2 \le \sum_j \left\langle e_j, \mathcal{C}_{\text{pr}}^{1/2} e_i\right\rangle_{\mathscr{H}}^2 = \left\|\mathcal{C}_{\text{pr}}^{1/2} e_i\right\|_{\mathscr{H}}^2 = \langle e_i, \mathcal{C}_{\text{pr}} e_i\rangle_{\mathscr{H}},\end{aligned}$$

where the penultimate equality follows from Parseval's identity. □

Also,

$$\operatorname{tr}(\mathcal{C}_{\text{post}}) = \operatorname{tr}(\mathcal{C}_{\text{pr}}) - \operatorname{tr}(\mathcal{C}_{\text{pr}}^{1/2}(I-\mathcal{S})\mathcal{C}_{\text{pr}}^{1/2}) = \sum_{j=1}^{\infty} (1-\alpha_j) \langle e_j, \mathcal{C}_{\text{pr}} e_j\rangle_{\mathscr{H}},$$

where $\alpha_j = \lambda_j/(1+\lambda_j)$. Thus, for eigenvalues λ_j that are large, we have $\alpha_j \approx 1$ which suggests that significant uncertainty reduction occurs in such directions. It is well known that for large classes of ill-posed Bayesian inverse problems, the eigenvalues λ_i of $\tilde{\mathcal{H}}_{\text{m}}$ decay rapidly to zero, with a relatively small number of dominant eigenvalues indicating the *data-informed* directions in the parameter space. This allows "focusing" the inference to low-dimensional subspaces of the parameter space $\mathscr{H}$. Such ideas have been used to develop efficient numerical algorithms for solution of infinite-dimensional Bayesian inverse problems in works such as Bui-Thanh et al. (2013); Flath et al. (2011) and for algorithms for computing A-optimal experimental designs for infinite-dimensional Bayesian linear inverse problems in Alexanderian et al. (2014).

4 KL divergence from posterior to prior and expected information gain

Let us first motivate the discussion by recalling the form of the KL divergence from the posterior to prior in the finite-dimensional case. We use boldface letters for the finite-dimensional versions of the operators appearing in the Bayesian inverse problem. To indicate that we work in $\mathbb{R}^n$, we denote by $\mu_{\text{pr,n}}$ and $\mu_{\text{post,n}}^{\boldsymbol{y}}$ the prior and posterior measures in the n-dimensional case. The following expression for $D_{\text{kl}}(\mu_{\text{post,n}}^{\boldsymbol{y}} \| \mu_{\text{pr,n}})$ is well known:

$$\begin{aligned}&D_{\text{kl}}\left(\mu_{\text{post,n}}^{\boldsymbol{y}} \| \mu_{\text{pr,n}}\right) \\ &\qquad = \frac{1}{2}\left[-\log\left(\frac{\det \mathbf{C}_{\text{post}}}{\det \mathbf{C}_{\text{pr}}}\right) - n + \operatorname{tr}(\mathbf{C}_{\text{pr}}^{-1}\mathbf{C}_{\text{post}}) + \left\langle \mathbf{C}_{\text{pr}}^{-1}\boldsymbol{u}_{\text{post}}^{\boldsymbol{y}}, \boldsymbol{u}_{\text{post}}^{\boldsymbol{y}}\right\rangle_{\mathbb{R}^n}\right]. \qquad (14)\end{aligned}$$

Note that the above expression is not meaningful in the infinite-dimensional case. For one thing, n appears explicitly in the expression. Moreover, in the infinite-dimensional case, $\mathcal{C}_{\text{pr}}$ is a trace-class operator whose eigenvalues accumulate at zero, so dividing by the determinant of the prior covariance is problematic as $n \to \infty$. Finally, in the infinite-dimensional case, $\mathcal{C}_{\text{pr}}^{-1}$ is the inverse of a compact operator and hence is unbounded; therefore, the trace term, which involves the inverse of the prior covariance, needs clarification. However, it is possible to reformulate (14) and obtain an expression that has meaning in the infinite-dimensional case as seen below.

A straightforward calculation shows that the first term on the right in (14) may be simplified:

$$\begin{aligned} -\log\left(\frac{\det \mathbf{C}_{\text{post}}}{\det \mathbf{C}_{\text{pr}}}\right) &= \log\left(\frac{\det \mathbf{C}_{\text{pr}}}{\det \mathbf{C}_{\text{post}}}\right) = \log\det\left(\mathbf{C}_{\text{pr}}\mathbf{C}_{\text{post}}^{-1}\right) \\ &= \log\det\left(\mathbf{C}_{\text{pr}}^{1/2}(\mathbf{H}_{\text{m}} + \mathbf{C}_{\text{pr}}^{-1})\mathbf{C}_{\text{pr}}^{1/2}\right) \\ &= \log\det(\tilde{\mathbf{H}}_{\text{m}} + \mathbf{I}). \end{aligned} \tag{15}$$

Recall that, in general, if $\mathbf{A}$ is Hermitian, then there exists a unitary matrix $\mathbf{U}$ such that

$$\mathbf{D} = [\lambda_i \delta_{ij}] = \mathbf{U}^* \mathbf{A} \mathbf{U}$$

is diagonal. In this case, the diagonal elements are the eigenvalues of $\mathbf{A}$, and

$$\det(\mathbf{I} + \mathbf{A}) = \det(\mathbf{U}) \det(\mathbf{I} + \mathbf{D}) \det(\mathbf{U}^*) = \prod_{i=1}^{n} (1 + \lambda_i).$$

In the infinite-dimensional setting, given a trace-class operator $\mathcal{A} \in \mathscr{L}_1^{\text{sym}+}(\mathscr{H})$,

$$\lim_{n\to\infty} \log\left(\prod_{i=1}^{n}(1 + \lambda_i(\mathcal{A}))\right) = \lim_{n\to\infty} \sum_{i=1}^{n} \log(1 + \lambda_i(\mathcal{A})) \le \lim_{n\to\infty} \sum_{i=1}^{n} \lambda_i(\mathcal{A}) < \infty,$$

so, motivated by the n-dimensional case, we may define the Fredholm determinant of $I + \mathcal{A}$ as

$$\det(I + \mathcal{A}) = \prod_{i=1}^{\infty} (1 + \lambda_i(\mathcal{A})),$$

where $\lambda_i(\mathcal{A})$ are the eigenvalues of $\mathcal{A}$ (Simon, 1977). Hence, the final expression in (15) is meaningful in infinite dimensions. Next, we consider the term $-n + \text{tr}(\mathbf{C}_{\text{pr}}^{-1}\mathbf{C}_{\text{post}})$:

$$\begin{aligned} -n + \text{tr}(\mathbf{C}_{\text{pr}}^{-1}\mathbf{C}_{\text{post}}) &= -\text{tr}(\mathbf{I}) + \text{tr}(\mathbf{C}_{\text{pr}}^{-1}\mathbf{C}_{\text{post}}) \\ &= \text{tr}(\mathbf{C}_{\text{pr}}^{-1}\mathbf{C}_{\text{post}} - \mathbf{I}) = \text{tr}\big((\mathbf{C}_{\text{pr}}^{-1} - \mathbf{C}_{\text{post}}^{-1})\mathbf{C}_{\text{post}}\big) = -\text{tr}(\mathbf{H}_{\text{m}}\mathbf{C}_{\text{post}}), \end{aligned} \tag{16}$$

where in the last step we used the fact that $\mathbf{C}_{\text{post}}^{-1} = \mathbf{H}_{\text{m}} + \mathbf{C}_{\text{pr}}^{-1}$. Notice that the argument of the trace in the final expression is in fact a trace-class operator in the infinite-dimensional case and has a well-defined trace. Combining (15) and (16) and defining the inner-product $\langle \boldsymbol{x}, \boldsymbol{y} \rangle_{\mathbf{C}_{\text{pr}}^{-1}} = \langle \mathbf{C}_{\text{pr}}^{-1/2}\boldsymbol{x}, \mathbf{C}_{\text{pr}}^{-1/2}\boldsymbol{y} \rangle_{\mathbb{R}^n}$ for $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$, we rewrite (14),

$$D_{\text{kl}}\left(\mu_{\text{post,n}}^{\boldsymbol{y}} \| \mu_{\text{pr,n}}\right) = \frac{1}{2}\Big[\log\det(\tilde{\mathbf{H}}_{\text{m}} + \mathbf{I}) - \text{tr}(\mathbf{H}_{\text{m}}\mathbf{C}_{\text{post}}) + \big\langle \boldsymbol{u}_{\text{post}}^{\boldsymbol{y}}, \boldsymbol{u}_{\text{post}}^{\boldsymbol{y}} \big\rangle_{\mathbf{C}_{\text{pr}}^{-1}}\Big]. \tag{17}$$

In Section 4.1, we derive, rigorously, alternate forms of the expression for the KL divergence from posterior to prior in the infinite-dimensional Hilbert space setting; as we shall see shortly, one of those forms is a direct extension of (17) to the infinite-dimensional case. The reason for introducing the weighted inner-product $\langle \cdot, \cdot \rangle_{\mathbf{C}_{\text{pr}}^{-1}}$ will also become clear in the discussion that follows.

4.1 The KL-divergence from posterior to prior

In what follows, we shall use the following result, which is a consequence of Proposition 1.2.8 in Da Prato and Zabczyk (2002).

Proposition 2. *Let $\mathcal{A} \in \mathscr{L}(\mathscr{H})$ be a positive self-adjoint operator, $\mu = \mathcal{N}(0, \mathcal{Q})$ a Gaussian measure on $\mathscr{H}$, and $b \in \mathscr{H}$. Then,*

$$\int_{\mathscr{H}} \exp\left\{-\frac{1}{2}\langle \mathcal{A}x, x\rangle_{\mathscr{H}} + \langle b, x\rangle_{\mathscr{H}}\right\} \mu(dx) \\ = \det(I+\tilde{\mathcal{A}})^{-1/2} \exp\left\{\frac{1}{2}\big\|(I+\tilde{\mathcal{A}})^{-1/2}\mathcal{Q}^{1/2}b\big\|^2_{\mathscr{H}}\right\},$$

where $\tilde{\mathcal{A}} = \mathcal{Q}^{1/2}\mathcal{A}\mathcal{Q}^{1/2}$.

In the following technical lemma, we calculate the expression for $\mathcal{Z}_0$, introduced in (9).

Lemma 2. *Let $\Phi(u; \boldsymbol{y}) = \frac{1}{2}(\mathcal{G}u - \boldsymbol{y})^T \boldsymbol{\Gamma}^{-1}_{noise}(\mathcal{G}u - \boldsymbol{y})$, as defined by (8). Then,*

$$\mathcal{Z}_0(\boldsymbol{y}) := \int_{\mathscr{H}} \exp\{-\Phi(u;\boldsymbol{y})\}\, \mu_{pr}(du) \\ = \exp\left\{-\frac{1}{2}|\boldsymbol{y}|^2\right\} \det(I+\tilde{\mathcal{H}}_m)^{-1/2} \exp\left\{\frac{1}{2}\langle \mathcal{C}_{post} b, b\rangle_{\mathscr{H}}\right\},$$

where $b = \mathcal{G}^*\boldsymbol{y}$ and $\mathcal{C}_{post} = (\mathcal{G}^*\mathcal{G} + \mathcal{C}_{pr}^{-1})^{-1}$, as in (10).

Proof. First note that (recall that we have assumed $\boldsymbol{\Gamma}_{\text{noise}} = \mathbf{I}$)

$$\Phi(u;\boldsymbol{y}) = \frac{1}{2}(\mathcal{G}u - \boldsymbol{y})^T(\mathcal{G}u - \boldsymbol{y}) = \frac{1}{2}\langle \mathcal{G}u, \mathcal{G}u\rangle_{\mathbb{R}^q} - \langle \mathcal{G}u, \boldsymbol{y}\rangle_{\mathbb{R}^q} + \frac{1}{2}\langle \boldsymbol{y}, \boldsymbol{y}\rangle_{\mathbb{R}^q} \\ = \frac{1}{2}\langle \mathcal{H}_{\text{m}}u, u\rangle_{\mathscr{H}} - \langle \mathcal{G}^*\boldsymbol{y}, u\rangle_{\mathscr{H}} + \frac{1}{2}|\boldsymbol{y}|^2. \tag{18}$$

Therefore,

$$\int_{\mathscr{H}} \exp\{-\Phi(u;\boldsymbol{y})\}\, \mu_{\text{pr}}(du) \\ = \exp\left\{-\frac{1}{2}|\boldsymbol{y}|^2\right\} \int_{\mathscr{H}} \exp\left\{-\frac{1}{2}\langle \mathcal{H}_{\text{m}}u, u\rangle_{\mathscr{H}} + \langle b, u\rangle_{\mathscr{H}}\right\} \mu_{\text{pr}}(du),$$

where $b = \mathcal{G}^*\boldsymbol{y}$. By Proposition 2, we have

$$\int_{\mathscr{H}} \exp\left\{-\frac{1}{2}\langle \mathcal{H}_{\text{m}}u, u\rangle_{\mathscr{H}} + \langle b, u\rangle_{\mathscr{H}}\right\} \mu_{\text{pr}}(du) \\ = \det(I + \mathcal{C}_{\text{pr}}^{1/2}\mathcal{H}_{\text{m}}\mathcal{C}_{\text{pr}}^{1/2})^{-1/2} \exp\left\{\frac{1}{2}\left\|(I + \mathcal{C}_{\text{pr}}^{1/2}\mathcal{H}_{\text{m}}\mathcal{C}_{\text{pr}}^{1/2})^{-1/2}\mathcal{C}_{\text{pr}}^{1/2}b\right\|^2_{\mathscr{H}}\right\} \\ = \det(I+\tilde{\mathcal{H}}_{\text{m}})^{-1/2} \exp\left\{\frac{1}{2}\left\|(I+\tilde{\mathcal{H}}_{\text{m}})^{-1/2}\mathcal{C}_{\text{pr}}^{1/2}b\right\|^2_{\mathscr{H}}\right\}.$$

The assertion of the lemma now follows, since $\mathcal{C}_{\text{post}} = \mathcal{C}_{\text{pr}}^{1/2}(I+\tilde{\mathcal{H}}_{\text{m}})^{-1}\mathcal{C}_{\text{pr}}^{1/2}$. □

The following result provides the expression for the KL divergence from posterior to prior:

Proposition 3. *Let μ_{pr} be a centered Gaussian measure on $\mathscr{H}$, and $\mu_{post}^{\boldsymbol{y}} = \mathcal{N}\big(u_{post}^{\boldsymbol{y}}, \mathcal{C}_{post}\big)$ be the posterior measure for a Bayesian linear inverse problem with additive Gaussian noise model as described in Section 3. Then,*

$$D_{kl}\left(\mu_{post}^{\boldsymbol{y}} \| \mu_{pr}\right) = \frac{1}{2}\left[\log\det(I + \tilde{\mathcal{H}}_m) - \mathrm{tr}(\mathcal{H}_m \mathcal{C}_{post}) - \left\langle u_{post}^{\boldsymbol{y}}, \mathcal{G}^*(\mathcal{G} u_{post}^{\boldsymbol{y}} - \boldsymbol{y})\right\rangle_{\mathscr{H}}\right]. \quad (19)$$

Proof. Consider (9), and note that

$$\begin{aligned} D_{\mathrm{kl}}\left(\mu_{\mathrm{post}}^{\boldsymbol{y}} \| \mu_{\mathrm{pr}}\right) &= \int_{\mathscr{H}} \log\left\{\frac{d\mu_{\mathrm{post}}^{\boldsymbol{y}}}{d\mu_{\mathrm{pr}}}\right\} \mu_{\mathrm{post}}^{\boldsymbol{y}}(du) \\ &= -\log \mathcal{Z}_0(\boldsymbol{y}) - \int_{\mathscr{H}} \Phi(u; \boldsymbol{y})\, \mu_{\mathrm{post}}^{\boldsymbol{y}}(du). \end{aligned} \quad (20)$$

Using (18) to expand $\Phi(u;\boldsymbol{y})$, the integral on the right becomes

$$\int_{\mathscr{H}} \Phi(u;\boldsymbol{y})\, \mu_{\mathrm{post}}^{\boldsymbol{y}}(du) = \frac{1}{2}\int_{\mathscr{H}} \langle \mathcal{H}_{\mathrm{m}} u, u\rangle_{\mathscr{H}}\, \mu_{\mathrm{post}}^{\boldsymbol{y}}(du) - \int_{\mathscr{H}} \langle \mathcal{G}^* \boldsymbol{y}, u\rangle_{\mathscr{H}}\, \mu_{\mathrm{post}}^{\boldsymbol{y}}(du) + \frac{1}{2}|\boldsymbol{y}|^2.$$

The second integral evaluates to $\left\langle \mathcal{G}^*\boldsymbol{y}, u_{\mathrm{post}}^{\boldsymbol{y}}\right\rangle_{\mathscr{H}}$, by the definition of the mean of the measure, and the first integral evaluates, via the formula for the integral of a quadratic form to

$$\int_{\mathscr{H}} \langle \mathcal{H}_{\mathrm{m}} u, u\rangle_{\mathscr{H}}\, \mu_{\mathrm{post}}^{\boldsymbol{y}}(du) = \mathrm{tr}(\mathcal{H}_{\mathrm{m}}\mathcal{C}_{\mathrm{post}}) + \left\langle u_{\mathrm{post}}^{\boldsymbol{y}}, \mathcal{H}_{\mathrm{m}} u_{\mathrm{post}}^{\boldsymbol{y}}\right\rangle_{\mathscr{H}}.$$

Using the expression for $\mathcal{Z}_0$ from Lemma 2,

$$\begin{aligned} -\log \mathcal{Z}_0(\boldsymbol{y}) &= \frac{1}{2}|\boldsymbol{y}|^2 - \log\det(I + \tilde{\mathcal{H}}_{\mathrm{m}})^{-1/2} - \frac{1}{2}\langle \mathcal{C}_{\mathrm{post}}\mathcal{G}^*\boldsymbol{y}, \mathcal{G}^*\boldsymbol{y}\rangle_{\mathscr{H}} \\ &= \frac{1}{2}|\boldsymbol{y}|^2 + \frac{1}{2}\log\det(I + \tilde{\mathcal{H}}_{\mathrm{m}}) - \frac{1}{2}\left\langle u_{\mathrm{post}}^{\boldsymbol{y}}, \mathcal{G}^*\boldsymbol{y}\right\rangle_{\mathscr{H}}, \end{aligned}$$

where we have also used the definition of $u_{\mathrm{post}}^{\boldsymbol{y}}$. Substituting into (20), we obtain

$$\begin{aligned} D_{\mathrm{kl}}\left(\mu_{\mathrm{post}}^{\boldsymbol{y}} \| \mu_{\mathrm{pr}}\right) =& \frac{1}{2}\log\det(I + \tilde{\mathcal{H}}_{\mathrm{m}}) - \frac{1}{2}\left\langle u_{\mathrm{post}}^{\boldsymbol{y}}, \mathcal{G}^*\boldsymbol{y}\right\rangle_{\mathscr{H}} \\ &- \frac{1}{2}\mathrm{tr}(\mathcal{H}_{\mathrm{m}}\mathcal{C}_{\mathrm{post}}) - \frac{1}{2}\left\langle u_{\mathrm{post}}^{\boldsymbol{y}}, \mathcal{H}_{\mathrm{m}} u_{\mathrm{post}}^{\boldsymbol{y}}\right\rangle_{\mathscr{H}} + \left\langle \mathcal{G}^*\boldsymbol{y}, u_{\mathrm{post}}^{\boldsymbol{y}}\right\rangle_{\mathscr{H}}, \end{aligned}$$

which, after some algebraic manipulation and recalling that $\mathcal{H}_{\mathrm{m}} = \mathcal{G}^*\mathcal{G}$, yields the assertion of the proposition. □

Let us note the following interpretation for the last term appearing in $D_{\mathrm{kl}}\left(\mu_{\mathrm{post}}^{\boldsymbol{y}} \| \mu_{\mathrm{pr}}\right)$ given in (19). Consider the function $\Phi(u) = \frac{1}{2}(\mathcal{G}u - \boldsymbol{y})^T(\mathcal{G}u - \boldsymbol{y})$, which is the familiar data misfit term in the deterministic interpretation of the corresponding linear inverse problem. (For notational simplicity we have suppressed the dependence of Φ on the data vector $\boldsymbol{y}$.) Note that the variational derivative of Φ at a point $u \in \mathscr{H}$ in direction $h \in \mathscr{H}$ is given by

$$\Phi'(u)h = \frac{d}{d\varepsilon}\Big|_{\varepsilon=0} \Phi(u + \varepsilon h) = \langle \mathcal{G}u - \boldsymbol{y}, \mathcal{G}h\rangle_{\mathbb{R}^q} = \langle \mathcal{G}^*(\mathcal{G}u - \boldsymbol{y}), h\rangle_{\mathscr{H}}.$$

Next, recall that the mean of the posterior, $u^{\boldsymbol{y}}_{\text{post}}$, of the present Bayesian linear inverse problem coincides with the MAP estimator for the inference parameter u and is the global minimizer of the regularized cost functional given by

$$\mathcal{J}(u) = \Phi(u) + \frac{1}{2}\langle u, u\rangle_{\mathcal{C}_{\text{pr}}^{-1}}, \tag{21}$$

with minimization done over the Cameron–Martin space, $\mathscr{H}_{\mu_{\text{pr}}} = \mathsf{range}(\mathcal{C}_{\text{pr}}^{1/2}) \subset \mathscr{H}$; see Stuart (2010); Dashti et al. (2013). The inner-product in the regularization term is

$$\langle x, y\rangle_{\mathcal{C}_{\text{pr}}^{-1}} = \left\langle \mathcal{C}_{\text{pr}}^{-1/2}x, \mathcal{C}_{\text{pr}}^{-1/2}y\right\rangle_{\mathscr{H}}, \quad x, y \in \mathscr{H}_{\mu_{\text{pr}}}.$$

We have, by the first order optimality conditions, $\mathcal{J}'(u^{\boldsymbol{y}}_{\text{post}})h = 0$ for every $h \in \mathscr{H}_{\mu_{\text{pr}}}$, that is,

$$\left\langle \mathcal{G}^*(\mathcal{G}u^{\boldsymbol{y}}_{\text{post}} - \boldsymbol{y}), h\right\rangle_{\mathscr{H}} + \left\langle u^{\boldsymbol{y}}_{\text{post}}, h\right\rangle_{\mathcal{C}_{\text{pr}}^{-1}} = 0, \quad \text{for all } h \in \mathscr{H}_{\mu_{\text{pr}}}.$$

Thus, in particular, $-\langle \mathcal{G}^*(\mathcal{G}u^{\boldsymbol{y}}_{\text{post}} - \boldsymbol{y}), u^{\boldsymbol{y}}_{\text{post}}\rangle_{\mathscr{H}} = \langle u^{\boldsymbol{y}}_{\text{post}}, u^{\boldsymbol{y}}_{\text{post}}\rangle_{\mathcal{C}_{\text{pr}}^{-1}}$. This leads to the following alternate form of expression (19):

$$D_{\text{kl}}\left(\mu^{\boldsymbol{y}}_{\text{post}} \| \mu_{\text{pr}}\right) = \frac{1}{2}\left[\log\det(I + \tilde{\mathcal{H}}_{\text{m}}) - \operatorname{tr}(\mathcal{H}_{\text{m}}\mathcal{C}_{\text{post}}) + \left\langle u^{\boldsymbol{y}}_{\text{post}}, u^{\boldsymbol{y}}_{\text{post}}\right\rangle_{\mathcal{C}_{\text{pr}}^{-1}}\right]. \tag{22}$$

Note that this expression for the KL divergence $D_{\text{kl}}\left(\mu^{\boldsymbol{y}}_{\text{post}} \| \mu_{\text{pr}}\right)$ is the direct extension of the corresponding expression in the case of $\mathscr{H} = \mathbb{R}^n$ as given in (17) to infinite dimensions.

Remark 1. A straightforward modification of the arguments leading to (19), for the case of a prior $\mu_{\text{pr}} = \mathcal{N}(u_{\text{pr}}, \mathcal{C}_{\text{pr}})$, gives

$$D_{\text{kl}}\left(\mu^{\boldsymbol{y}}_{\text{post}} \| \mu_{\text{pr}}\right) = \frac{1}{2}\left[\log\det(I + \tilde{\mathcal{H}}_{\text{m}}) - \operatorname{tr}(\mathcal{H}_{\text{m}}\mathcal{C}_{\text{post}}) - \left\langle u^{\boldsymbol{y}}_{\text{post}} - u_{\text{pr}}, \mathcal{G}^*(\mathcal{G}u^{\boldsymbol{y}}_{\text{post}} - \boldsymbol{y})\right\rangle_{\mathscr{H}}\right].$$

Moreover, in view of the argument leading to (22), we have

$$D_{\text{kl}}\left(\mu^{\boldsymbol{y}}_{\text{post}} \| \mu_{\text{pr}}\right) = \frac{1}{2}\left[\log\det(I + \tilde{\mathcal{H}}_{\text{m}}) - \operatorname{tr}(\mathcal{H}_{\text{m}}\mathcal{C}_{\text{post}}) + \left\langle u^{\boldsymbol{y}}_{\text{post}} - u_{\text{pr}}, u^{\boldsymbol{y}}_{\text{post}} - u_{\text{pr}}\right\rangle_{\mathcal{C}_{\text{pr}}^{-1}}\right].$$

Remark 2. In the discussion above, we mentioned in passing the notion of the MAP estimator of the inference parameter. Recall that in the finite-dimensional case, a MAP estimator is defined as a point in the parameter space that maximizes the pdf (i.e., the Lebesgue density) of the posterior measure. While this definition does not extend to the infinite-dimensional setting, it is still possible to define a notion of the MAP estimator in infinite dimensions. Denote by $B_\varepsilon(z)$ the open ball of radius $\varepsilon > 0$ centered at $z \in \mathscr{H}$. A MAP estimator can be understood as a point u^* such that $B_\varepsilon(u^*)$ has maximal probability as $\varepsilon \to 0$. We refer to Dashti et al. (2013) for a rigorous treatment of the mathematical questions concerning MAP estimators of Bayesian inverse problems in separable Banach spaces. In particular, it is shown in Dashti et al. (2013) that MAP estimators are the minimizers of the regularized cost functionals of form (21), corresponding to the associated Bayesian inverse problems. In the Gaussian linear setting considered here, the functional $\mathcal{J}(u)$ has a unique global minimizer—the MAP estimator, which also coincides with the posterior mean.

4.2 Expected information gain

Here we derive the expression for the expected information gain. We first prove the following technical lemma, which is needed in the proof of the main result in this section.

Lemma 3. *The following identities hold:*

1. $\mathsf{E}_{\mu_{pr}}\left\{\mathsf{E}_{\boldsymbol{y}|u}\left\{\left\langle u^{\boldsymbol{y}}_{post}, \mathcal{G}^*\boldsymbol{y}\right\rangle_{\mathscr{H}}\right\}\right\} = \mathrm{tr}(\tilde{\mathcal{H}}_m)$,
2. $\mathsf{E}_{\mu_{pr}}\left\{\mathsf{E}_{\boldsymbol{y}|u}\left\{\left\langle u^{\boldsymbol{y}}_{post}, \mathcal{H}_m u^{\boldsymbol{y}}_{post}\right\rangle_{\mathscr{H}}\right\}\right\} = \mathrm{tr}(\mathcal{S}\tilde{\mathcal{H}}_m^2)$,

where $\tilde{\mathcal{H}}_m$ and $\mathcal{S}$ are as in (12) and (13), respectively.

Proof. We present the proof of the first statement; the second one follows from a similar argument. Let us begin from the inner expectation. Note that, by the definition of $u^{\boldsymbol{y}}_{\mathrm{post}}$ we have

$$\left\langle u^{\boldsymbol{y}}_{\mathrm{post}}, \mathcal{G}^*\boldsymbol{y}\right\rangle_{\mathscr{H}} = \langle \mathcal{C}_{\mathrm{post}}\mathcal{G}^*\boldsymbol{y}, \mathcal{G}^*\boldsymbol{y}\rangle_{\mathscr{H}} = \langle \boldsymbol{y}, \mathcal{G}\mathcal{C}_{\mathrm{post}}\mathcal{G}^*\boldsymbol{y}\rangle_{\mathbb{R}^q}.$$

For clarity, let us denote $L = \mathcal{G}\mathcal{C}_{\mathrm{post}}\mathcal{G}^*$. Recall that $\boldsymbol{y}|u$ is distributed according to $\mathcal{N}(\mathcal{G}u, \mathbf{\Gamma}_{\mathrm{noise}})$, and that we assumed $\mathbf{\Gamma}_{\mathrm{noise}} = \mathbf{I}$. Using the formula for the expectation of a quadratic form (on $\mathscr{Y} = \mathbb{R}^q$), Lemma 1, we have

$$\mathsf{E}_{\boldsymbol{y}|u}\left\{\left\langle u^{\boldsymbol{y}}_{\mathrm{post}}, \mathcal{G}^*\boldsymbol{y}\right\rangle_{\mathscr{H}}\right\} = \mathsf{E}_{\boldsymbol{y}|u}\left\{\langle \boldsymbol{y}, L\boldsymbol{y}\rangle_{\mathbb{R}^q}\right\} = \mathrm{tr}(L) + \langle \mathcal{G}u, L\mathcal{G}u\rangle_{\mathbb{R}^q} = \mathrm{tr}(L) + \langle u, \mathcal{G}^*L\mathcal{G}u\rangle_{\mathscr{H}}.$$

By the comment at the end of Section 2.1 and recalling that $\mathcal{C}_{\mathrm{post}} = \mathcal{C}_{\mathrm{pr}}^{1/2}\mathcal{S}\mathcal{C}_{\mathrm{pr}}^{1/2}$,

$$\begin{aligned}\mathrm{tr}(L) &= \mathrm{tr}(\mathcal{G}\mathcal{C}_{\mathrm{post}}\mathcal{G}^*) = \mathrm{tr}(\mathcal{C}_{\mathrm{post}}\mathcal{H}_{\mathrm{m}}) = \mathrm{tr}(\mathcal{C}_{\mathrm{pr}}^{1/2}\mathcal{S}\mathcal{C}_{\mathrm{pr}}^{1/2}\mathcal{H}_{\mathrm{m}})\\ &= \mathrm{tr}(\mathcal{S}\mathcal{C}_{\mathrm{pr}}^{1/2}\mathcal{H}_{\mathrm{m}}\mathcal{C}_{\mathrm{pr}}^{1/2}) = \mathrm{tr}(\mathcal{S}\tilde{\mathcal{H}}_{\mathrm{m}}).\end{aligned} \tag{23}$$

Therefore,

$$\mathsf{E}_{\boldsymbol{y}|u}\left\{\left\langle u^{\boldsymbol{y}}_{\mathrm{post}}, \mathcal{G}^*\boldsymbol{y}\right\rangle_{\mathscr{H}}\right\} = \mathrm{tr}(\mathcal{S}\tilde{\mathcal{H}}_{\mathrm{m}}) + \langle u, \mathcal{G}^*L\mathcal{G}u\rangle_{\mathscr{H}}. \tag{24}$$

Next, to compute the outer expectation, we proceed as follows (keep in mind that $\mu_{\mathrm{pr}} = \mathcal{N}(0, \mathcal{C}_{\mathrm{pr}})$). By Lemma 1,

$$\mathsf{E}_{\mu_{\mathrm{pr}}}\left\{\langle u, \mathcal{G}^*L\mathcal{G}u\rangle_{\mathscr{H}}\right\} = \int_{\mathscr{H}} \langle u, \mathcal{G}^*L\mathcal{G}u\rangle_{\mathscr{H}}\, \mu_{\mathrm{pr}}(du) = \mathrm{tr}(\mathcal{G}^*L\mathcal{G}\mathcal{C}_{\mathrm{pr}});$$

and

$$\begin{aligned}\operatorname{tr}(\mathcal{G}^* L \mathcal{G} \mathcal{C}_{\mathrm{pr}}) &= \operatorname{tr}(\mathcal{G}^* \mathcal{G} \mathcal{C}_{\mathrm{post}} \mathcal{G}^* \mathcal{G} \mathcal{C}_{\mathrm{pr}}) = \operatorname{tr}(\mathcal{H}_{\mathrm{m}} \mathcal{C}_{\mathrm{post}} \mathcal{H}_{\mathrm{m}} \mathcal{C}_{\mathrm{pr}}) \\ &= \operatorname{tr}(\mathcal{C}_{\mathrm{pr}}^{1/2} \mathcal{H}_{\mathrm{m}} \mathcal{C}_{\mathrm{post}} \mathcal{H}_{\mathrm{m}} \mathcal{C}_{\mathrm{pr}}^{1/2}) = \operatorname{tr}(\mathcal{C}_{\mathrm{pr}}^{1/2} \mathcal{H}_{\mathrm{m}} \mathcal{C}_{\mathrm{pr}}^{1/2} \mathcal{S} \mathcal{C}_{\mathrm{pr}}^{1/2} \mathcal{H}_{\mathrm{m}} \mathcal{C}_{\mathrm{pr}}^{1/2}) \\ &= \operatorname{tr}(\tilde{\mathcal{H}}_{\mathrm{m}} \mathcal{S} \tilde{\mathcal{H}}_{\mathrm{m}}) = \operatorname{tr}(\mathcal{S} \tilde{\mathcal{H}}_{\mathrm{m}}^2).\end{aligned}$$

Thus, combining (23), (24), and (25),

$$\begin{aligned}\mathsf{E}_{\mu_{\mathrm{pr}}}\left\{\mathsf{E}_{\boldsymbol{y}|u}\left\{\left\langle u_{\mathrm{post}}^{\boldsymbol{y}}, \mathcal{G}^* \boldsymbol{y}\right\rangle_{\mathscr{H}}\right\}\right\} &= \operatorname{tr}(\mathcal{S}\tilde{\mathcal{H}}_{\mathrm{m}}) + \operatorname{tr}(\mathcal{S}\tilde{\mathcal{H}}_{\mathrm{m}}^2) = \operatorname{tr}(\mathcal{S}\tilde{\mathcal{H}}_{\mathrm{m}}(I + \tilde{\mathcal{H}}_{\mathrm{m}})) \\ &= \operatorname{tr}(\tilde{\mathcal{H}}_{\mathrm{m}}(I + \tilde{\mathcal{H}}_{\mathrm{m}})\mathcal{S}) = \operatorname{tr}(\tilde{\mathcal{H}}_{\mathrm{m}}),\end{aligned}$$

which is the first statement of the lemma. ☐

The following theorem is the main result of this section.

Theorem 1. *Let μ_{pr} be a centered Gaussian prior measure on $\mathscr{H}$, and $\mu_{post}^{\boldsymbol{y}} = \mathcal{N}\left(u_{post}^{\boldsymbol{y}}, \mathcal{C}_{post}\right)$ be the posterior measure for a Bayesian linear inverse problem with additive Gaussian noise model as described in Section 3. Then,*

$$\mathsf{E}_{\mu_{pr}}\left\{\mathsf{E}_{\boldsymbol{y}|u}\left\{D_{kl}\left(\mu_{post}^{\boldsymbol{y}} \| \mu_{pr}\right)\right\}\right\} = \frac{1}{2}\log\det(I + \tilde{\mathcal{H}}_m).$$

Proof. By (19) we have

$$\begin{aligned}\mathsf{E}_{\mu_{\mathrm{pr}}}\left\{\mathsf{E}_{\boldsymbol{y}|u}\left\{D_{\mathrm{kl}}\left(\mu_{\mathrm{post}}^{\boldsymbol{y}} \| \mu_{\mathrm{pr}}\right)\right\}\right\} &= \frac{1}{2}\log\det(I + \tilde{\mathcal{H}}_{\mathrm{m}}) \\ &-\frac{1}{2}\operatorname{tr}(\mathcal{H}_{\mathrm{m}}\mathcal{C}_{\mathrm{post}}) - \frac{1}{2}\mathsf{E}_{\mu_{\mathrm{pr}}}\left\{\mathsf{E}_{\boldsymbol{y}|u}\left\{\left\langle u_{\mathrm{post}}^{\boldsymbol{y}}, \mathcal{G}^*(\mathcal{G}u_{\mathrm{post}}^{\boldsymbol{y}} - \boldsymbol{y})\right\rangle_{\mathscr{H}}\right\}\right\}. \quad (25)\end{aligned}$$

Using the previous lemma, we proceed as follows:

$$\begin{aligned}&\mathsf{E}_{\mu_{\mathrm{pr}}}\left\{\mathsf{E}_{\boldsymbol{y}|u}\left\{\left\langle u_{\mathrm{post}}^{\boldsymbol{y}}, \mathcal{G}^*(\mathcal{G}u_{\mathrm{post}}^{\boldsymbol{y}} - \boldsymbol{y})\right\rangle_{\mathscr{H}}\right\}\right\} \\ &\quad= \mathsf{E}_{\mu_{\mathrm{pr}}}\left\{\mathsf{E}_{\boldsymbol{y}|u}\left\{\left\langle u_{\mathrm{post}}^{\boldsymbol{y}}, \mathcal{H}_{\mathrm{m}} u_{\mathrm{post}}^{\boldsymbol{y}}\right\rangle_{\mathscr{H}}\right\}\right\} - \mathsf{E}_{\mu_{\mathrm{pr}}}\left\{\mathsf{E}_{\boldsymbol{y}|u}\left\{\left\langle u_{\mathrm{post}}^{\boldsymbol{y}}, \mathcal{G}^*\boldsymbol{y}\right\rangle_{\mathscr{H}}\right\}\right\} \\ &\quad= \operatorname{tr}(\mathcal{S}\tilde{\mathcal{H}}_{\mathrm{m}}^2) - \operatorname{tr}(\tilde{\mathcal{H}}_{\mathrm{m}}) \\ &\quad= \operatorname{tr}(\mathcal{S}(\tilde{\mathcal{H}}_{\mathrm{m}} - \mathcal{S}^{-1})\tilde{\mathcal{H}}_{\mathrm{m}}) = -\operatorname{tr}(\mathcal{S}\tilde{\mathcal{H}}_{\mathrm{m}}).\end{aligned}$$

Thus, since $\operatorname{tr}(\mathcal{H}_{\mathrm{m}}\mathcal{C}_{\mathrm{post}}) = \operatorname{tr}(\tilde{\mathcal{H}}_{\mathrm{m}}\mathcal{S}) = \operatorname{tr}(\mathcal{S}\tilde{\mathcal{H}}_{\mathrm{m}})$, the expression for the expected information gain in (25) simplifies to $\mathsf{E}_{\mu_{\mathrm{pr}}}\left\{\mathsf{E}_{\boldsymbol{y}|u}\left\{D_{\mathrm{kl}}\left(\mu_{\mathrm{post}}^{\boldsymbol{y}} \| \mu_{\mathrm{pr}}\right)\right\}\right\} = \frac{1}{2}\log\det(I + \tilde{\mathcal{H}}_{\mathrm{m}})$. ☐

The result above provides the infinite-dimensional analogue of Bayesian D-optimality. As mentioned in the introduction, an experimental design $\boldsymbol{\xi}$ enters the Bayesian inverse problem through the data likelihood. This dependence on $\boldsymbol{\xi}$, in the present Gaussian linear case, is manifested through a $\boldsymbol{\xi}$ dependent data misfit Hessian, $\mathcal{H}_{\mathrm{m}} = \mathcal{H}_{\mathrm{m}}(\boldsymbol{\xi})$.

Consequently, the D-optimal design problem in the infinite-dimensional Hilbert space setting is given by,

$$\underset{\boldsymbol{\xi}\in\Xi}{\text{maximize}}\ \log\det(I+\tilde{\mathcal{H}}_{\mathrm{m}}(\boldsymbol{\xi})),$$

where Ξ is the design space which needs to be specified in a given experimental design problem.

Remark 3. As mentioned earlier, in a large class of Bayesian inverse problems, $\tilde{\mathcal{H}}_{\mathrm{m}}$ admits a low-rank approximation,

$$\tilde{\mathcal{H}}_{\mathrm{m}}v \approx \sum_{i=1}^{r} \lambda_i \left\langle e_i, v\right\rangle_{\mathscr{H}} e_i, \quad v\in\mathscr{H},$$

where r is the numerical rank of $\tilde{\mathcal{H}}_{\mathrm{m}}$, and $\{\lambda_i\}_{i=1}^r$ are the dominant eigenvalues of $\tilde{\mathcal{H}}_{\mathrm{m}}$ with respective eigenvectors $\{e_i\}_{i=1}^r$. Thus, one can use the following approximation

$$\log\det(I+\tilde{\mathcal{H}}_{\mathrm{m}}) \approx \sum_{i=1}^{r}\log(1+\lambda_i),$$

which enables an efficient means of approximating the expected information gain.

5 Expected mean square error of the MAP estimator and Bayesian A-optimality

In this section, we consider another well-known optimal experimental design criterion, Bayesian A-optimality, which aims to minimize the trace of the posterior covariance operator. It is known in the statistics literature that for inference problems with a finite-dimensional parameter, this is equivalent to minimizing the expected mean square error of the mean posterior, which, in the case of a Bayesian linear inverse problem, coincides with the MAP estimator. In this section, we extend this result to the infinite-dimensional Hilbert space setting.

The MSE of the MAP estimator $u_{\text{post}}^{\boldsymbol{y}}$ is

$$\text{MSE}(u_{\text{post}}^{\boldsymbol{y}}; u) = \mathsf{E}_{\boldsymbol{y}|u}\left\{\left\|u - u_{\text{post}}^{\boldsymbol{y}}\right\|_{\mathscr{H}}^2\right\}.$$

The MSE is also referred to as the *risk* of the estimator $u_{\text{post}}^{\boldsymbol{y}}$, corresponding to a quadratic loss function. A straightforward calculation shows that

$$\text{MSE}(u_{\text{post}}^{\boldsymbol{y}}; u) = \left\|u - \mathsf{E}_{\boldsymbol{y}|u}\left\{u_{\text{post}}^{\boldsymbol{y}}\right\}\right\|_{\mathscr{H}}^2 + \mathsf{E}_{\boldsymbol{y}|u}\left\{\left\|u_{\text{post}}^{\boldsymbol{y}} - \mathsf{E}_{\boldsymbol{y}|u}\left\{u_{\text{post}}^{\boldsymbol{y}}\right\}\right\|_{\mathscr{H}}^2\right\}, \tag{26}$$

Note that the first term in (26) quantifies the magnitude of estimation bias, and the second term describes the variability of the estimator around its mean. The following technical lemma provides the expression for $\text{MSE}(u_{\text{post}}^{\boldsymbol{y}}; u)$ in the infinite-dimensional Hilbert space setting.

Lemma 4. *Let $u^{\boldsymbol{y}}_{post}$ be the MAP estimator for u as in (10). Then,*

$$\mathrm{MSE}(u^{\boldsymbol{y}}_{post}; u) = \|(\mathcal{C}_{post}\mathcal{H}_m - I)u\|^2_{\mathscr{H}} + \mathrm{tr}(\mathcal{C}^2_{post}\mathcal{H}_m).$$

Proof. Consider the expression for $\mathrm{MSE}(u^{\boldsymbol{y}}_{\mathrm{post}}; u)$, given in (26). For the first term in the sum, we have

$$u - \mathsf{E}_{\boldsymbol{y}|u}\left\{u^{\boldsymbol{y}}_{\mathrm{post}}\right\} = u - \mathsf{E}_{\boldsymbol{y}|u}\left\{\mathcal{C}_{\mathrm{post}}\mathcal{G}^*\boldsymbol{y}\right\} = u - \mathcal{C}_{\mathrm{post}}\mathcal{G}^*\mathcal{G}u = (I - \mathcal{C}_{\mathrm{post}}\mathcal{H}_{\mathrm{m}})u.$$

Next, note that $\xi(\boldsymbol{y}) = u^{\boldsymbol{y}}_{\mathrm{post}} - \mathsf{E}_{\boldsymbol{y}|u}\left\{u^{\boldsymbol{y}}_{\mathrm{post}}\right\}$ has law $\mu = \mathcal{N}(0, \mathcal{Q})$ with $\mathcal{Q} = (\mathcal{C}_{\mathrm{post}}\mathcal{G}^*)(\mathcal{C}_{\mathrm{post}}\mathcal{G}^*)^* = \mathcal{C}_{\mathrm{post}}\mathcal{H}_{\mathrm{m}}\mathcal{C}_{\mathrm{post}}$. Therefore,

$$\mathsf{E}_{\boldsymbol{y}|u}\left\{\left\|u^{\boldsymbol{y}}_{\mathrm{post}} - \mathsf{E}_{\boldsymbol{y}|u}\left\{u^{\boldsymbol{y}}_{\mathrm{post}}\right\}\right\|^2_{\mathscr{H}}\right\} = \int_{\mathscr{H}} \|\xi\|^2_{\mathscr{H}}\, \mu(d\xi) = \mathrm{tr}(\mathcal{C}_{\mathrm{post}}\mathcal{H}_{\mathrm{m}}\mathcal{C}_{\mathrm{post}}) = \mathrm{tr}(\mathcal{C}^2_{\mathrm{post}}\mathcal{H}_{\mathrm{m}}).$$

□

Next, we consider the average over the prior measure of the MSE,

$$\mathsf{E}_{\mu_{\mathrm{pr}}}\left\{\mathrm{MSE}(u^{\boldsymbol{y}}_{\mathrm{post}}; u)\right\} = \int_{\mathscr{H}}\int_{\mathscr{Y}} \left\|u - u^{\boldsymbol{y}}_{\mathrm{post}}\right\|^2_{\mathscr{H}} \pi_{\mathrm{like}}(\boldsymbol{y}|u)\, d\boldsymbol{y}\, \mu_{\mathrm{pr}}(du),$$

which is also known as the Bayes risk of the estimator $u^{\boldsymbol{y}}_{\mathrm{post}}$, corresponding to a quadratic loss function (Carlin and Louis, 1997; Berger, 1985). The following result extends the well-known result regarding the connection between the Bayes risk of the MAP estimator and the trace of the posterior covariance, for a Bayesian linear inverse problem, in the infinite-dimensional Hilbert space setting.

Theorem 2. *Let μ_{pr} be a centered Gaussian prior measure on $\mathscr{H}$, and $\mu^{\boldsymbol{y}}_{post} = \mathcal{N}\big(u^{\boldsymbol{y}}_{post}, \mathcal{C}_{post}\big)$ be the posterior measure for a Bayesian linear inverse problem with additive Gaussian noise model, as described in Section 3. Then, $\mathsf{E}_{\mu_{pr}}\left\{\mathrm{MSE}(u^{\boldsymbol{y}}_{post}; u)\right\} = \mathrm{tr}(\mathcal{C}_{post})$.*

Proof. By Lemma 4,

$$\mathsf{E}_{\mu_{\mathrm{pr}}}\left\{\mathrm{MSE}(u^{\boldsymbol{y}}_{\mathrm{post}}; u)\right\} = \int_{\mathscr{H}} \|(\mathcal{C}_{\mathrm{post}}\mathcal{H}_{\mathrm{m}} - I)u\|^2_{\mathscr{H}}\, \mu_{\mathrm{pr}}(du) + \mathrm{tr}(\mathcal{C}^2_{\mathrm{post}}\mathcal{H}_{\mathrm{m}}), \tag{27}$$

and since $(\mathcal{C}_{\mathrm{post}}\mathcal{H}_{\mathrm{m}} - I)u \sim \mathcal{N}(0, (\mathcal{C}_{\mathrm{post}}\mathcal{H}_{\mathrm{m}} - I)\mathcal{C}_{\mathrm{pr}}(\mathcal{C}_{\mathrm{post}}\mathcal{H}_{\mathrm{m}} - I)^*) =: \mu$,

$$\begin{aligned}\int_{\mathscr{H}} \|(\mathcal{C}_{\mathrm{post}}\mathcal{H}_{\mathrm{m}} - I)u\|^2_{\mathscr{H}}\, \mu_{\mathrm{pr}}(du) &= \int_{\mathscr{H}} \|\xi\|^2_{\mathscr{H}}\, \mu(d\xi)\\ &= \mathrm{tr}((\mathcal{C}_{\mathrm{post}}\mathcal{H}_{\mathrm{m}} - I)\mathcal{C}_{\mathrm{pr}}(\mathcal{C}_{\mathrm{post}}\mathcal{H}_{\mathrm{m}} - I)^*).\end{aligned}$$

We proceed as follows:

$$\begin{aligned}\mathrm{tr}((\mathcal{C}_{\mathrm{post}}\mathcal{H}_{\mathrm{m}} - I)\mathcal{C}_{\mathrm{pr}}(\mathcal{C}_{\mathrm{post}}\mathcal{H}_{\mathrm{m}} - I)^*) &= \mathrm{tr}((\mathcal{C}_{\mathrm{post}}\mathcal{H}_{\mathrm{m}} - I)^*(\mathcal{C}_{\mathrm{post}}\mathcal{H}_{\mathrm{m}} - I)\mathcal{C}_{\mathrm{pr}})\\ &= \mathrm{tr}(\mathcal{H}_{\mathrm{m}}\mathcal{C}^2_{\mathrm{post}}\mathcal{H}_{\mathrm{m}}\mathcal{C}_{\mathrm{pr}}) - \mathrm{tr}(\mathcal{H}_{\mathrm{m}}\mathcal{C}_{\mathrm{post}}\mathcal{C}_{\mathrm{pr}}) - \mathrm{tr}(\mathcal{C}_{\mathrm{post}}\mathcal{H}_{\mathrm{m}}\mathcal{C}_{\mathrm{pr}}) + \mathrm{tr}(\mathcal{C}_{\mathrm{pr}}).\end{aligned}$$

Let us first consider the last two terms; recalling that $\mathcal{C}_{\text{post}} = \mathcal{C}_{\text{pr}}^{1/2}\mathcal{S}\mathcal{C}_{\text{pr}}^{1/2}$,

$$\begin{aligned}-\text{tr}(\mathcal{C}_{\text{post}}\mathcal{H}_{\text{m}}\mathcal{C}_{\text{pr}}) + \text{tr}(\mathcal{C}_{\text{pr}}) &= -\text{tr}(\mathcal{C}_{\text{pr}}^{1/2}\mathcal{S}\mathcal{C}_{\text{pr}}^{1/2}\mathcal{H}_{\text{m}}\mathcal{C}_{\text{pr}}) + \text{tr}(\mathcal{C}_{\text{pr}})\\ &= -\text{tr}(\mathcal{C}_{\text{pr}}\mathcal{S}\tilde{\mathcal{H}}_{\text{m}}) + \text{tr}(\mathcal{C}_{\text{pr}})\\ &= \text{tr}(\mathcal{C}_{\text{pr}}\mathcal{S}(\mathcal{S}^{-1} - \tilde{\mathcal{H}}_{\text{m}})) = \text{tr}(\mathcal{C}_{\text{pr}}\mathcal{S}) = \text{tr}(\mathcal{C}_{\text{post}}).\end{aligned}$$

Thus, by (27),

$$\mathsf{E}_{\mu_{\text{pr}}}\left\{\text{MSE}(u_{\text{post}}^{\boldsymbol{y}}; u)\right\} = \text{tr}(\mathcal{C}_{\text{post}}^2\mathcal{H}_{\text{m}}) + \text{tr}(\mathcal{H}_{\text{m}}\mathcal{C}_{\text{post}}^2\mathcal{H}_{\text{m}}\mathcal{C}_{\text{pr}}) - \text{tr}(\mathcal{H}_{\text{m}}\mathcal{C}_{\text{post}}\mathcal{C}_{\text{pr}}) + \text{tr}(\mathcal{C}_{\text{post}}).$$

Hence, showing that the first three terms sum to zero completes the proof. To this end, we note that $\text{tr}(\mathcal{H}_{\text{m}}\mathcal{C}_{\text{post}}\mathcal{C}_{\text{pr}}) = \text{tr}(\tilde{\mathcal{H}}_{\text{m}}\mathcal{S}\mathcal{C}_{\text{pr}})$ and that

$$\begin{aligned}\text{tr}(\mathcal{C}_{\text{post}}^2\mathcal{H}_{\text{m}}) + \text{tr}(\mathcal{H}_{\text{m}}\mathcal{C}_{\text{post}}^2\mathcal{H}_{\text{m}}\mathcal{C}_{\text{pr}}) &= \text{tr}(\tilde{\mathcal{H}}_{\text{m}}\mathcal{S}\mathcal{C}_{\text{pr}}\mathcal{S}) + \text{tr}(\tilde{\mathcal{H}}_{\text{m}}\mathcal{S}\mathcal{C}_{\text{pr}}\mathcal{S}\tilde{\mathcal{H}}_{\text{m}})\\ &= \text{tr}\big(\tilde{\mathcal{H}}_{\text{m}}\mathcal{S}\mathcal{C}_{\text{pr}}\mathcal{S}(I + \tilde{\mathcal{H}}_{\text{m}})\big) = \text{tr}(\tilde{\mathcal{H}}_{\text{m}}\mathcal{S}\mathcal{C}_{\text{pr}}). \quad \square\end{aligned}$$

Appendix A: Proof of Lemma 1

Let $\{e_i\}_1^\infty$ be a complete orthonormal set in $\mathscr{H}$, and denote by Π_n the orthogonal projection of $\mathscr{H}$ onto $\mathsf{Span}\{e_1, \ldots, e_n\}$; that is, for $x \in \mathscr{H}$, $\Pi_n(x) = \sum_{i=1}^n \langle e_i, x\rangle_{\mathscr{H}} e_i$. First note that

$$\begin{aligned}\int_{\mathscr{H}} \langle \mathcal{A}x, x\rangle_{\mathscr{H}} = &\int_{\mathscr{H}} \langle \mathcal{A}(x-m), x-m\rangle_{\mathscr{H}}\ \mu(dx)\\ &+ \int_{\mathscr{H}} \langle \mathcal{A}x, m\rangle_{\mathscr{H}}\ \mu(dx) + \int_{\mathscr{H}} \langle \mathcal{A}m, x\rangle_{\mathscr{H}}\ \mu(dx) - \langle \mathcal{A}m, m\rangle_{\mathscr{H}}.\end{aligned}$$

By the definition of the mean of the measure, the last three terms sum to $\langle \mathcal{A}m, m\rangle_{\mathscr{H}}$. Thus, the rest of the proof consists of showing $\int_{\mathscr{H}} \langle \mathcal{A}(x-m), x-m\rangle_{\mathscr{H}}\ \mu(dx) = \text{tr}(\mathcal{A}\mathcal{Q})$. Note that for every $x \in \mathscr{H}$, $x - m = \lim_{n\to\infty} \Pi_n(x-m)$, so

$$\lim_{n\to\infty} \langle \mathcal{A}\Pi_n(x-m), \Pi_n(x-m)\rangle_{\mathscr{H}} = \langle \mathcal{A}(x-m), x-m\rangle_{\mathscr{H}}.$$

Moreover, $|\langle \mathcal{A}\Pi_n(x-m), \Pi_n(x-m)\rangle_{\mathscr{H}}| \le \|A\|\,\|x-m\|_{\mathscr{H}}^2$, and since $\mathcal{Q}$ is trace-class, the measure μ has a bounded second moment. Hence, $\int_{\mathscr{H}} \|x-m\|_{\mathscr{H}}^2\ \mu(dx) < \infty$. Therefore, we can apply the Lebesgue Dominated Convergence Theorem to get

$$\lim_{n\to\infty} \int_{\mathscr{H}} \langle \mathcal{A}\Pi_n(x-m), \Pi_n(x-m)\rangle_{\mathscr{H}}\ \mu(dx) = \int_{\mathscr{H}} \langle \mathcal{A}(x-m), x-m\rangle_{\mathscr{H}}\ \mu(dx). \quad (28)$$

Next, let $\{e_i\}_1^\infty$ be the (complete) set of eigenvectors of $\mathcal{Q}$ with corresponding (real) eigenvalues $\{\lambda_i\}_1^\infty$. We know that $\mathcal{A}\mathcal{Q}$ is trace-class with

$$\text{tr}(\mathcal{A}\mathcal{Q}) = \sum_i \langle \mathcal{A}\mathcal{Q}e_i, e_i\rangle_{\mathscr{H}} = \sum_i \lambda_i \langle \mathcal{A}e_i, e_i\rangle_{\mathscr{H}}. \quad (29)$$

Also,

$$
\begin{aligned}
&\int_{\mathscr{H}} \langle \mathcal{A}\Pi_n(x-m), \Pi_n(x-m)\rangle_{\mathscr{H}} \; \mu(dx) \\
&\quad = \sum_{i,j=1}^{n} \int_{\mathscr{H}} \langle \mathcal{A}e_i, e_j\rangle_{\mathscr{H}} \langle x-m, e_i\rangle_{\mathscr{H}} \langle x-m, e_j\rangle_{\mathscr{H}} \, \mu(dx) \\
&\quad = \sum_{i,j=1}^{n} \langle \mathcal{A}e_i, e_j\rangle_{\mathscr{H}} \int_{\mathscr{H}} \langle x-m, e_i\rangle_{\mathscr{H}} \langle x-m, e_j\rangle_{\mathscr{H}} \, \mu(dx) \\
&\quad = \sum_{i,j=1}^{n} \langle \mathcal{A}e_i, e_j\rangle_{\mathscr{H}} \langle \mathcal{Q}e_i, e_j\rangle_{\mathscr{H}} = \sum_{i=1}^{n} \lambda_i \langle \mathcal{A}e_i, e_i\rangle_{\mathscr{H}} .
\end{aligned}
$$

Combining this last result with (28) and (29), we get

$$
\begin{aligned}
\int_{\mathscr{H}} \langle \mathcal{A}(x-m), x-m\rangle_{\mathscr{H}} \; \mu(dx) &= \lim_{n\to\infty} \int_{\mathscr{H}} \langle \mathcal{A}\Pi_n(x-m), \Pi_n(x-m)\rangle_{\mathscr{H}} \; \mu(dx) \\
&= \lim_{n\to\infty} \sum_{i=1}^{n} \lambda_i \langle \mathcal{A}e_i, e_i\rangle_{\mathscr{H}} = \mathrm{tr}(\mathcal{A}\mathcal{Q}). \; \square
\end{aligned}
$$

References

Alexanderian, A., Petra, N., Stadler, G., and Ghattas, O. (2014). "A-optimal design of experiments for infinite-dimensional Bayesian linear inverse problems with regularized ℓ_0-sparsification." *SIAM Journal on Scientific Computing*, 36(5): A2122–A2148. MR3257642. doi: http://dx.doi.org/10.1137/130933381. 683

Atkinson, A. C. and Donev, A. N. (1992). *Optimum Experimental Designs*. Oxford. 671

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*. Springer. MR0804611. doi: http://dx.doi.org/10.1007/978-1-4757-4286-2. 691

Bui-Thanh, T., Ghattas, O., Martin, J., and Stadler, G. (2013). "A computational framework for infinite-dimensional Bayesian inverse problems Part I: The linearized case, with application to global seismic inversion." *SIAM Journal on Scientific Computing*, 35(6): A2494–A2523. MR3126997. doi: http://dx.doi.org/10.1137/12089586X. 675, 680, 683

Carlin, B. P. and Louis, T. A. (1997). "Bayes and empirical Bayes methods for data analysis." *Statistics and Computing*, 7(2): 153–154. MR1427749. doi: http://dx.doi.org/10.1023/A:1018577817064. 691

Chaloner, K. and Verdinelli, I. (1995). "Bayesian experimental design: A review." *Statistical Science*, 10(3): 273–304. MR1390519. doi: http://dx.doi.org/10.1214/ss/1177009939. 673, 674

Conway, J. B. (2000). *A Course in Operator Theory*. American Mathematical Society. MR1721402. 676

Da Prato, G. (2006). *An Introduction to Infinite-Dimensional Analysis*. Springer. MR2244975. doi: http://dx.doi.org/10.1007/3-540-29021-4. 677, 678, 679

Da Prato, G. and Zabczyk, J. (2002). *Second Order Partial Differential Equations in Hilbert Spaces*, volume 293. Cambridge University Press. MR1985790. doi: http://dx.doi.org/10.1017/CBO9780511543210. 676, 677, 679, 685

— (2014). *Stochastic Equations in Infinite Dimensions*, volume 152. Cambridge University Press. MR3236753. doi: http://dx.doi.org/10.1017/CBO9781107295513. 677, 679

Dashti, M., Law, K. J., Stuart, A. M., and Voss, J. (2013). "MAP estimators and their consistency in Bayesian nonparametric inverse problems." *Inverse Problems*, 29. MR3104933. doi: http://dx.doi.org/10.1088/0266-5611/29/9/095017. 687, 688

Flath, H. P., Wilcox, L. C., Akçelik, V., Hill, J., van Bloemen Waanders, B., and Ghattas, O. (2011). "Fast algorithms for Bayesian uncertainty quantification in large-scale linear inverse problems based on low-rank partial Hessian approximations." *SIAM Journal on Scientific Computing*, 33(1). MR2783201. doi: http://dx.doi.org/10.1137/090780717. 683

Gel'fand, I. M. and Vilenkin, N. Y. (1964). *Generalized Functions. Vol. 4: Applications of Harmonic Analysis*. Translated by Amiel Feinstein. Academic Press, New York - London, 1964. MR0173945. 676

Kullback, S. and Leibler, R. A. (1951). "On Information and Sufficiency." *Annals of Mathematical Statistics*, 22(1): 79–86. MR0039968. doi: http://dx.doi.org/10.1214/aoms/1177729694. 672, 678

Lindgren, F., Rue, H., and Lindström, J. (2011). "An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(4): 423–498. MR2853727. doi: http://dx.doi.org/10.1111/j.1467-9868.2011.00777.x. 680

Pukelsheim, F. (2006). *Optimal Design of Experiments*, volume 50. SIAM. MR2224698. doi: http://dx.doi.org/10.1137/1.9780898719109. 671

Reed, M. and Simon, B. (1972). *Methods of Modern Mathematical Physics. Vol. 1.: Functional Analysis*. Academic Press. 676

Simon, B. (1977). "Notes on Infinite Determinants of Hilbert Space Operators." *Advances in Mathematics*, 24: 244–273. MR0482328. doi: http://dx.doi.org/10.1016/0001-8708(77)90057-3. 684

Stuart, A. M. (2010). "Inverse problems: a Bayesian perspective." *Acta Numerica*, 19: 451–559. MR2652785. doi: http://dx.doi.org/10.1017/S0962492910000061. 672, 675, 679, 680, 681, 687

Tarantola, A. (2005). *Inverse problem theory and methods for model parameter estimation*. siam. MR2130010. doi: http://dx.doi.org/10.1137/1.9780898717921. 674

Uciński, D. (2005). *Optimal Measurement Methods for Distributed Parameter System Identification*. Boca Raton: CRC Press. MR2465165. 671

Acknowledgments

We would like to thank the anonymous referees for their helpful comments and suggestions that led to significant improvements in the final version of the article. The work of A. Alexanderian and O. Ghattas was partially supported by NSF grant CMMI-1028889, DOE grants DE-SC0009286, DE-SC0010518, DE-SC0006656, DE-SC0002710, and AFOSR grant FA9550-12-1-0484.