Stability estimates for the expected utility in Bayesian optimal experimental design

Duc-Lam Duong*, Tapio Helin* and Jose Rodrigo Rojo-Garcia*

Computational Engineering, LUT School of Engineering Science, Lappeenranta-Lahti University of Technology, 53850 Lappeenranta, Finland

E-mail: duc-lam.duong@lut.fi, tapio.helin@lut.fi and rodrigo.rojo.garcia@lut.fi

Received 25 June 2023; revised 26 September 2023
Accepted for publication 19 October 2023
Published 2 November 2023

Abstract
We study stability properties of the expected utility function in Bayesian optimal experimental design. We provide a framework for this problem in a non-parametric setting and prove a convergence rate of the expected utility with respect to a likelihood perturbation. This rate is uniform over the design space and its sharpness in the general setting is demonstrated by proving a lower bound in a special case. To make the problem more concrete we proceed by considering non-linear Bayesian inverse problems with Gaussian likelihood and prove that the assumptions set out for the general case are satisfied and regain the stability of the expected utility with respect to perturbations to the observation map. Theoretical convergence rates are demonstrated numerically in three different examples.

Keywords: Bayesian optimal experimental design, inverse problems, stability, Bayesian inference

(Some figures may appear in colour only in the online journal)

1. Introduction

Acquisition of high quality data is the crux of many challenges in science and engineering. An outstanding example is the parameter estimation problem in statistical models. Namely, data collection, whether in field experiments or in laboratory, is often restricted by limited resources. It can be difficult, expensive and time-consuming, which puts severe limits on the quality of data acquired. To maximise the value of data for inference and minimise the uncertainty of the estimated parameters, one has to design the experiment (for instance, placing the sensors) in a way that is as economical and efficient as possible. This involves choosing...
the values of the controllable variables before the experiment takes place. Carefully designed experiments can make a substantial difference in accomplishing the tasks in an appropriate manner. Optimal experimental design (OED) is a mathematical framework where a set of design variables with certain optimal criteria (based on the information matrix derived from the model) are satisfied (Steinberg and Hunter 1984, Pukelsheim 2006). The most common criteria for OED include A-optimality and D-optimality which, in finite dimensions, seeks to minimise the trace and the determinant of the Fisher information matrix, respectively.

We adopt a Bayesian approach to OED (Chaloner and Verdinelli 1995) that formulates the task as a maximisation of an expected utility. Suppose $X$ denotes our unknown parameter, $Y$ stands for the observation and $d$ the design parameter. The expected utility $U$ is given by

$$U(d) = \mathbb{E}_\mu u(X, Y; d),$$

where $u(X, Y; d)$ denotes the utility of an estimate $X$ given observation $Y$ with design $d$, and the expectation is taken w.r.t. the joint distribution $\mu$ of $X$ and $Y$. The task of optimising $U(d)$ is notoriously expensive especially if the design space is large or $X$ and $Y$ are modelled in high-dimensional space (Ryan et al 2016).

The guiding set of design problems for this paper is those emerging in modern inverse problems (Engl et al 1996), which often involve imaging a high-dimensional object by indirect observations. Bayesian approach to inverse problems has gained considerable interest during the last two decades (Kaipio and Somersalo 2006, Stuart 2010, Dashti and Stuart 2017). The underlying mathematical model in inverse problems is often governed by partial differential equations (PDEs) giving rise to complex high-dimensional likelihood models.

Exploring the complex high-dimensional posterior distributions in Bayesian inversion is computationally costly but has become standard in literature in recent years due to developments in Monte Carlo (MC) based sampling schemes and the growth of available computational resources. To accelerate this task, various approximation schemes such as surrogate modelling of the likelihood distribution can also be applied including e.g. polynomial chaos expansion (Marzouk et al 2007, Schillings and Schwab 2013) or neural networks (Herrmann et al 2020). Moreover, stability of the posterior distribution with respect to such approximations is well-understood (see Sprungk 2020, Garbuno-Inigo et al 2023 and reference therein).

In the OED framework, the computational effort compared to conventional Bayesian inversion is significantly larger due to the double expectation and optimisation task. In consequence, approximation schemes have a key role in producing optimal designs in that framework. However, questions about the stability of the expected utility have received limited attention. To the best of our knowledge, it has been only considered in terms of a fixed design, i.e. pointwise convergence of the expected utility. Tempone and others in (Long et al 2015, Beck et al 2018, 2020) have developed approximation results for nested MC methods with and without utilising an additional Laplace approximation and analyse the optimal parametrisation of the method with respect to the approximation error versus computational effort. Since MC approximation is random, any such error bound is expressed in probabilistic terms. In particular, (Beck et al 2018) provides a recipe to achieve a given probabilistic error tolerance with optimal computational effort. In another recent line of work, Foster and others (Foster et al 2019) develop error analysis for variational methods being applied in combination with MC methods and optimise the depth of variational approximation to achieve a rate of convergence $O((N + K)^{-\frac{1}{2}})$ for $N$ samples from MC and $K$ optimisation steps for the variational model.

For Bayesian OED tasks involving optimisation on a continuous design manifold (as is often the case in inverse problems), pointwise convergence does not provide the full picture of the stability in terms of the optimisation task. Instead, uniform approximation rates of given numerical schemes in a neighbourhood around the optimal design are preferable in that regard.
In this work, we study the uniform stability of the expected utility in Bayesian OED systematically where changes in likelihood function or observation map can be observed.

Non-parametric inference in Bayesian inverse problems and OED is motivated by the conviction to leave discretisation until the last possible moment (Stuart 2010), hence giving rise to opportunities to choose appropriate and robust discretisation methods. Non-parametric approach for OED in Bayesian inverse problems has been formalised by Alexanderian (see Alexanderian 2021 and references therein). Let us note that the infinite-dimensional setting arises naturally in numerous works involving Bayesian OED in inverse problems constrained by PDEs (Alexanderian et al 2014, 2016a, 2016b, Long et al 2015, Beck et al 2018, Wu et al 2020, to name a few), integral geometry (Haber et al 2008, Ruthotto et al 2018, Burger et al 2021, Helin et al 2022) or nonlinear systems (Huan and Marzouk 2013, 2014).

1.1. Our contribution

This work contributes to the rigorous study of the mathematical framework of Bayesian OED.

- We formulate the OED stability problem under a Bayesian framework in a non-parametric setting. We propose a set of assumptions (along the lines of Stuart (2010)) under which the stability problem for OED can be addressed in a systematic way (assumption 3.1). In particular, we assume that the likelihood and its approximate version are close in the Kullback–Leibler (KL) divergence.
- We establish the convergence of the expected utility for a general approximation scheme satisfying assumption 3.1 with a rate of one-half of the likelihood convergence rate (see theorem 3.4). We demonstrate by a trivial example that a faster rate is not possible without further assumptions (see example 1).
- Together with the convergence of the surrogate expected utility, we prove that their maximisers converge, up to a subsequence, to a maximiser of the true expected utility (see theorem 3.5). This ensures that the optimal design variable is also stable in the approximation.
- As an important application, we consider some Bayesian inverse problems with Gaussian noise and their observation map can be replaced by some surrogate model. We demonstrate that the assumptions we set out previously are satisfied in this case, given that the observation map and its surrogate model are close in certain norms (see proposition 4.2 and theorem 4.4).
- Finally, we carry out numerical simulations on three different design problems. We observe that the rates predicted by our main theorems are aligned with the numerical results.

1.2. Structure of the paper

The paper is organised as follows. In section 2, we give an overview of the main objects of this article, including basic notions of non-parametric Bayesian inverse problems and Bayesian experimental design. We also summarise some background in probability measure theory, commonly used metrics between measures and introduce notations that will be used throughout this paper. In section 3, we first outline the common framework including the general assumptions and main results. We proceed by establishing several lemmas and proving our main theorems. An important aspect of the main results is considered in section 4 where we study the stability of OED for some Bayesian inverse problems with Gaussian noise. In the last section, we provide three numerical examples to illustrate the results of the paper.
2. Preliminaries

2.1. Probability measures and metrics between measures

Throughout this paper, $\mathcal{X}$ will be a separable Banach space (Hilbert space) equipped with a norm $\| \cdot \|_X$ (inner product $\langle \cdot, \cdot \rangle_X$), with notice that the subscript may be ignored if no confusion arises. A bounded linear operator $C : \mathcal{X} \to \mathcal{X}$ in Hilbert space $\mathcal{X}$ is called self-adjoint if $\langle Cx, y \rangle = \langle x, Cy \rangle$ for all $x, y \in \mathcal{X}$ and positive definite (or positive) if $\langle Cx, x \rangle \geq 0$ for all $x \in \mathcal{X}$.

We say that a self-adjoint and positive operator $C$ is of trace class if

$$\text{tr}(C) := \sum_{n=1}^{\infty} \langle Ce_n, e_n \rangle < \infty,$$

where $\{e_n\}$ is an orthonormal basis of $\mathcal{X}$.

Let $B(\mathcal{X})$ be the Borel $\sigma$-algebra on $\mathcal{X}$ and let $\mu$ be a Borel probability measure on $\mathcal{X}$, that is, $\mu$ is defined on the measurable space $(\mathcal{X}, B(\mathcal{X}))$. The mean $m \in \mathcal{X}$ and the (bounded linear operator) covariance $\sigma : \mathcal{X} \to \mathcal{X}$ of $\mu$ are defined as follows

$$\langle m, h \rangle = \int_{\mathcal{X}} \langle h, x \rangle d\mu(x), \quad \text{for all } h \in \mathcal{X},$$

$$\langle \sigma h_1, h_2 \rangle = \int_{\mathcal{X}} \langle h_1, x - m \rangle \langle h_2, x - m \rangle d\mu(x), \quad \text{for all } h_1, h_2 \in \mathcal{X}.$$

We also use the concept of weighted norm $\| \cdot \|_C = \|C^{-1/2} \cdot \|$ for any covariance operator $C$ in $\mathcal{X}$.

Let $\mu_1$ and $\mu_2$ be two Borel probability measures on $\mathcal{X}$. Let $\mu$ be a common reference measure, also defined on $(\mathcal{X}, B(\mathcal{X}))$. The following ‘distances’ between measures are utilised in the rest of the paper.

The Hellinger distance between $\mu_1$ and $\mu_2$ is defined as

$$d^2_{\text{Hell}}(\mu_1, \mu_2) = \frac{1}{2} \int_{\mathcal{X}} \left( \sqrt{\frac{d\mu_1}{d\mu}} - \sqrt{\frac{d\mu_2}{d\mu}} \right)^2 d\mu = \frac{1}{2} \int_{\mathcal{X}} \left( 1 - \sqrt{\frac{d\mu_2}{d\mu_1}} \right)^2 d\mu_1,$$

where $\mu$ is a reference measure such that $\mu_1 \ll \mu$ and $\mu_2 \ll \mu$, i.e. $\mu_1$ and $\mu_2$ are absolutely continuous with respect to $\mu$. The second identity holds if $\mu_2 \ll \mu_1$.

The KL divergence for $\mu_1$ and $\mu_2$ with $\mu_2 \ll \mu_1$ is defined as

$$D_{\text{KL}}(\mu_2 \| \mu_1) = \int_{\mathcal{X}} \log \left( \frac{d\mu_2}{d\mu_1} \right) \frac{d\mu_2}{d\mu_1} = \int_{\mathcal{X}} \log \left( \frac{d\mu_2}{d\mu_1} \right) d\mu_2.$$

Here, the second identity holds if also $\mu_1 \ll \mu_2$ (that is, $\mu_1$ and $\mu_2$ are equivalent). Notice carefully that $D_{\text{KL}}(\mu_2 \| \mu_1) \geq 0$ and that KL divergence is not symmetric.

Let us now record three well-known lemmas that will be used below. First, the Hellinger distance can be bounded by the KL divergence as follows.

**Lemma 2.1.** If $\mu_1$ and $\mu_2$ are equivalent probability measures on $\mathcal{X}$, then

$$d^2_{\text{Hell}}(\mu_1, \mu_2) \leq \frac{1}{2} D_{\text{KL}}(\mu_1 \| \mu_2).$$  \hspace{1cm} (2)

Second, the KL divergence between two Gaussian distributions has an explicit expression utilising the means and covariance matrices.
Lemma 2.2. Suppose we have two Gaussian distributions \( \mu_1 \sim \mathcal{N}(m_1, \Gamma_1) \) and \( \mu_2 \sim \mathcal{N}(m_2, \Gamma_2) \) on \( \mathbb{R}^p \). Then it holds that
\[
D_{\text{KL}}(\mu_1, \mu_2) = \frac{1}{2} \left( \text{tr} \left( \Gamma^{-1}_1 \right) - p + (m_2 - m_1)^\top \Gamma^{-1}_2 (m_2 - m_1) + \log \left( \frac{\det \Gamma_2}{\det \Gamma_1} \right) \right).
\]

Third, arbitrary moments of Gaussian distributions in Hilbert spaces are finite.

Lemma 2.3 (Da Prato and Zabczyk 2014, proposition 2.19). Let \( X \) be a separable Hilbert space. For any \( k \in \mathbb{N} \), there exists a constant \( C = C(k) \) such that
\[
\int_X \| x \|^2 \mu(dx) \leq C [\text{tr} (\Gamma)],
\]
for any Gaussian measure \( \mu = \mathcal{N}(0, \Gamma) \).

Below, we denote random variables with capital letters (\( X \) and \( Y \)), while the realisations are generally denoted by lowercase letters (\( x \) and \( y \)).

2.2. Bayesian OED

In Bayesian OED, one seeks an experimental setup providing predictive data distribution with maximal information content in terms of recovering the unknown parameter. Let us make this idea more precise: we denote the unknown latent parameter by \( x \in X \), where \( X \) is a separable Banach space. Let \( y \in Y \) be the observational data variable where \( Y \) is a finite-dimensional data space. For convenience, in this work we assume \( Y = \mathbb{R}^p \). Moreover, let \( d \in D \) be the design variable, where \( D \) is a (typically compact) metric space.

The expected utility is given by formula (1), where \( \mu \) is the joint probability distribution of \( X \) and \( Y \), and \( u \) is a utility function. In this work, we focus on the expected information gain by defining
\[
u(x, y, d) = -\log \left( \frac{\pi(y|x; d)}{\pi(y; d)} \right),
\]
where \( \pi(y|x; d) \) and \( \pi(y; d) \) are the likelihood distribution and the evidence distribution given \( d \in D \), respectively. We observe that
\[
U(d) = \int_Y \int_X \log \left( \frac{\pi(y|x; d)}{\pi(y; d)} \right) \pi(y|x; d)\mu_0(x) dy dx = \int_X D_{\text{KL}}(\pi(y|x; d), \pi(y; d)) \mu_0(dx),
\]
where \( \mu_0 \) is the prior measure on \( X \). To find the OED, the expected utility is then maximised over the design variable space \( D \). A design \( d^* \) is called optimal if it maximises \( U \), that is
\[
d^* \in \arg \max_{d \in D} U(d).
\]

We note that in general, the functional \( U: D \to \mathbb{R} \) may have several maximisers.

In inverse problems, the unknown \( x \) is connected to the data \( y \) through an observation (or parameter-to-observable) map \( G: X \times D \to Y \). The problem of inverting \( G(\cdot; d) \) with fixed \( d \in D \) is ill-posed and the likelihood distribution is governed by the observational model
\[
y = G(x; d) + \xi,
\]
where \( \xi \) represents an additive measurement noise. As a typical example, a Gaussian distribution noise distribution, \( \xi \sim \mathcal{N}(0, \Gamma) \), with the covariance matrix \( \Gamma \in \mathbb{R}^{p \times p} \) giving rise to the likelihood distribution \( y|x \sim \mathcal{N}(G(x), \Gamma) \).
2.3. Γ-convergence

We collect here the definition and some basic results of Γ-convergence that will be used later on. Standard references of this subject are Dal Maso (1993), Braides (2002).

**Definition 2.4.** Let $X$ be a metric space and assume that $F_n, F : X \to \mathbb{R}$ are functionals on $X$. We say that $F_n$ Γ-converges to $F$ if, for every $x \in X$, the following conditions hold,

(i) (liminf inequality) for every sequence $(x_n)_{n \in \mathbb{N}}$ converging to $x$,

$$F(x) \leq \liminf_{n \to \infty} F_n(x_n);$$

(ii) (limsup inequality) there exists a recovery sequence $(x_n)_{n \in \mathbb{N}}$ converging to $x$ such that

$$F(x) \geq \limsup_{n \to \infty} F_n(x_n).$$

**Theorem 2.5 (fundamental theorem of Γ-convergence).** If $F_n$ Γ-converges to $F$ and $x_n$ minimises $F_n$, then every limit point $x$ of the sequence $(x_n)_{n \in \mathbb{N}}$ is a minimiser of $F$, and

$$F(x) = \limsup_{n \to \infty} F_n(x_n).$$

3. Stability estimates

3.1. General assumptions and main results

In this section, we establish two useful results on the stability of the expected utility and of the optimal design variable. Recall now the expected information gain $U(d)$ defined in (3). Assume that we have access to a surrogate likelihood density $\pi_N(y|x; d)$, which approximates $\pi(y|x; d)$ as $N$ increases. The corresponding surrogate utility $U_N(d)$ is obtained as

$$U_N(d) = \int_{\mathcal{X}} \int_{\mathcal{Y}} \log \left( \frac{\pi_N(y|x; d)}{\pi_N(y; d)} \right) \pi_N(y|x; d) \mathrm{d}\mu_0(x) \mathrm{d}y,$$

where $\pi_N(y; d)$ is the corresponding surrogate evidence. Let us make our approximation condition precise by the following assumption.

**Assumption 3.1.** The following conditions hold.

(A1) There exist $C > 0$, a function $\psi : \mathbb{N} \to \mathbb{R}_+$ such that $\psi(N) \to 0$ as $N \to \infty$ and

$$\mathbb{E}^{\mu_0}[D_{KL}(\pi_N(\cdot|X; d) \parallel \pi(\cdot|X; d))] \leq C\psi(N),$$

for all $d \in \mathcal{D}$.

(A2) For any converging sequence $d_N \to d$ in $\mathcal{D}$, we have

$$\lim_{N \to \infty} \mathbb{E}^{\mu_0}[D_{KL}(\pi(\cdot|X; d_N) \parallel \pi(\cdot|X; d))] = 0.$$

(A3) There exists $C_0 > 0$ such that for every $N \in \mathbb{N}$, $d \in \mathcal{D}$ and every sequence $d_N \to d$ in $\mathcal{D}$,

$$\int_{\mathcal{X}} \int_{\mathcal{Y}} \log^2 \left( \frac{\pi(y|x; d)}{\pi(y; d)} \right) \pi(y|x; d) + \pi_N(y|x; d) + \pi(y|x; d_N) \mathrm{d}\mu_0(x) \mathrm{d}y < C_0.$$

(6)
Remark 3.2. Assumption (A1) reflects a natural condition on the approximation rate of the surrogate likelihood under the KL divergence. This is somewhat similar to the conditions commonly used in Bayesian inverse problems, albeit under different metrics (for instance, Hellinger distance, see Dashti and Stuart (2017), section 4, assumptions 2). Assumption (A2) records a continuity condition of the likelihood with respect to the design variable. (A3) is a technical assumption which, in loose terms, requires that a special divergence between likelihood and data-marginal (or equivalently, between posterior and prior) is bounded when averaged over prior. We note that the second moment of the log-ratio quantity such as $\log \frac{\pi(y|x)}{\hat{\pi}(y|x)}$ in (6) appears quite naturally in the context of Bayesian statistical inverse problems, see for instance (Nickl 2023) (proposition 1.3.1, where it is termed the V-distance).

The following proposition is the cornerstone of our main theorems. It yields in particular that the difference between the expected utility and its surrogate can be controlled by the likelihood perturbations under the KL divergence.

Proposition 3.3. Consider likelihood distributions $\pi(y|x)$ and $\hat{\pi}(y|x)$ and define

$$U = \int_Y \int_X \log \left( \frac{\pi(y|x)}{\pi(y)} \right) \pi(y|x) d\mu_0(x) dy$$

and

$$\hat{U} = \int_Y \int_X \log \left( \frac{\hat{\pi}(y|x)}{\pi(y)} \right) \hat{\pi}(y|x) d\mu_0(x) dy,$$

where $\pi(y) = \int \pi(y|x) d\mu_0(x)$ and $\hat{\pi}(y) = \int \hat{\pi}(y|x) d\mu_0(x)$. Let us denote

$$K := \int_Y \int_X \log^2 \left( \frac{\pi(y|x)}{\pi(y)} \right) \left[ \pi(y|x) + \hat{\pi}(y|x) \right] d\mu_0(x) dy.$$

It follows that

$$|U - \hat{U}| \leq \sqrt{K} \sqrt{\mathbb{E}^\mu_D KL(\pi(\cdot|x) \| \hat{\pi}(\cdot|x)) + 2\mathbb{E}^\mu_D KL(\pi(\cdot|x) \| \hat{\pi}(\cdot|x))}. \quad (7)$$

We now state the main theorems of the paper.

Theorem 3.4. Let assumption 3.1 (A1) and (A3) hold. Then there exists $C > 0$ such that for all $N$ sufficiently large,

$$\sup_{d \in \mathcal{D}} |U(d) - U_N(d)| \leq C \sqrt{\psi(N)}. \quad (8)$$

Theorem 3.5. Let assumption 3.1 hold. Suppose

$$d_N^* \in \arg\max_{d \in \mathcal{D}} U_N(d).$$

Then, the limit $d^*$ of any converging subsequence of $\{d_N^*\}_{N=1}^\infty$ is a maximiser of $U$, that is, $d^* \in \arg\max_{d \in \mathcal{D}} U(d)$. Moreover,

$$\liminf_{N \to \infty} U_N(d_N^*) = U(d^*). \quad (9)$$

In particular, if $\{d_N^*\}_{N=1}^\infty$ converges to $d^*$ in $\mathcal{D}$, then $d^*$ is a maximiser of $U$, and

$$\lim_{N \to \infty} U_N(d_N^*) = U(d^*). \quad (10)$$
Remark 3.6. Theorem 3.4 establishes the uniform convergence of the approximate expected utility. Theorem 3.5, moreover, ensures that the corresponding approximate optimal design and the maximum expected information gain also converge up to a subsequence.

3.2. Proof of the main theorems

3.2.1. Proof of proposition 3.3. Let us recall for the reader’s convenience that

\[
U = \int_{Y} \int_{X} \log \left( \frac{\pi(y|x)}{\hat{\pi}(y)} \right) \pi(y|x) \, d\mu_0(x) \, dy \quad \text{and} \quad \tilde{U} = \int_{Y} \int_{X} \log \left( \frac{\hat{\pi}(y|x)}{\pi(y)} \right) \hat{\pi}(y|x) \, d\mu_0(x) \, dy,
\]

where

\[
\pi(y) = \int \pi(y|x) \, d\mu_0(x) \quad \text{and} \quad \hat{\pi}(y) = \int \hat{\pi}(y|x) \, d\mu_0(x).
\]

The corresponding posteriors are given by

\[
\frac{d\mu}{d\mu_0}(x) = \frac{\pi(y|x)}{\pi(y)}, \quad \frac{d\hat{\mu}}{d\mu_0}(x) = \frac{\hat{\pi}(y|x)}{\hat{\pi}(y)}. \tag{11}
\]

We have

\[
U - \tilde{U} = \int_{Y} \int_{X} \left[ \log \left( \frac{\pi(y|x)}{\hat{\pi}(y)} \right) - \log \left( \frac{\hat{\pi}(y|x)}{\pi(y)} \right) \right] \pi(y|x) \, d\mu_0(x) \, dy
\]

\[
= \int_{Y} \int_{X} \log \left( \frac{\pi(y|x)}{\hat{\pi}(y)} \right) \left[ \pi(y|x) - \hat{\pi}(y|x) \right] \, d\mu_0(x) \, dy
\]

\[
+ \int_{Y} \int_{X} \log \left( \frac{\hat{\pi}(y|x)}{\pi(y|x)} \right) \hat{\pi}(y|x) \, d\mu_0(x) \, dy
\]

\[
= I + D_{KL}(\hat{\pi}(\cdot) \| \pi(\cdot)) - E^{\mu_0}D_{KL}(\hat{\pi}(\cdot|X) \| \pi(\cdot|X)),
\]

where

\[
I := \int_{Y} \int_{X} \left[ \log \left( \frac{\pi(y|x)}{\hat{\pi}(y)} \right) \right] \left[ \pi(y|x) - \hat{\pi}(y|x) \right] \, d\mu_0(x) \, dy.
\]

Since we have the identity above, we naturally have

\[
\text{d}|U - \tilde{U}| \leq |I| + |D_{KL}(\hat{\pi}(\cdot) \| \pi(\cdot))| + |E^{\mu_0}D_{KL}(\hat{\pi}(\cdot|X) \| \pi(\cdot|X))|. \tag{12}
\]

Let us first consider the term \(I\).

Lemma 3.7. It follows that

\[
|I| \leq K E^{\mu_0} \left[ D_{KL}(\hat{\pi}(\cdot|X) \| \pi(\cdot|X)) \right],
\]

where

\[
K = \int_{Y} \int_{X} \log^2 \left( \frac{\pi(y|x)}{\hat{\pi}(y)} \right) \left[ \pi(y|x) + \hat{\pi}(y|x) \right] \, d\mu_0(x) \, dy.
\]
Proof. By the Cauchy–Schwarz inequality,

\[ F^2 = \left( \int_X \int_Y \log \left( \frac{\pi(y|x)}{\pi(y)} \right) \left[ \sqrt{\pi(y|x)} + \sqrt{\pi(y|x)} \right] \left[ \sqrt{\pi(y|x)} - \sqrt{\pi(y|x)} \right] \, d\mu_0(x) \, dy \right)^2 \]

\[ \leq \int_X \int_Y \log^2 \left( \frac{\pi(y|x)}{\pi(y)} \right) \left[ \sqrt{\pi(y|x)} + \sqrt{\pi(y|x)} \right]^2 \, d\mu_0(x) \, dy \]

\[ \cdot \int_X \int_Y \left[ \sqrt{\pi(y|x)} - \sqrt{\pi(y|x)} \right]^2 \, d\mu_0(x) \, dy \]

\[ \leq 2 \int_X \int_Y \log^2 \left( \frac{\pi(y|x)}{\pi(y)} \right) \left[ \pi(y|x) + \tilde{\pi}(y|x) \right] \, d\mu_0(x) \, dy \cdot E^{\mu_0} \left[ d_{\text{Hell}}(\pi(\cdot|X), \tilde{\pi}(\cdot|X)) \right] \]

Now, thanks to lemma 2.1,

\[ E^{\mu_0} \left[ d_{\text{Hell}}(\tilde{\pi}(\cdot|X), \pi(\cdot|X)) \right] \leq \frac{1}{2} E^{\mu_0} \left[ D_{\text{KL}}(\tilde{\pi}(\cdot|X)) \parallel \pi(\cdot|X) \right]. \]

Therefore

\[ |F|^2 \leq K E^{\mu_0} \left[ D_{\text{KL}}(\tilde{\pi}(\cdot|X)) \parallel \pi(\cdot|X) \right], \]

as required.

Let us now consider the second term on the right-hand side of (12).

Lemma 3.8. It holds that

\[ D_{\text{KL}}(\tilde{\pi}(\cdot) \parallel \pi(\cdot)) \leq E^{\mu_0} \left[ D_{\text{KL}}(\tilde{\pi}(\cdot|X) \parallel \pi(\cdot|X)) \right]. \]

Proof. Thanks to (11) and Fubini’s theorem, we have

\[ E^{\mu_0} \left[ D_{\text{KL}}(\tilde{\pi}(\cdot|X) \parallel \pi(\cdot|X)) \right] = \int_Y \int_X \log \left( \frac{\tilde{\pi}(y|x)}{\pi(y|x)} \right) \tilde{\pi}(y|x) \, d\mu_0(x) \, dy \]

\[ = \int_Y \int_X \log \left( \frac{\tilde{\pi}(y)}{\pi(y)} \right) \tilde{\pi}(y) \, \frac{d\tilde{\mu}}{d\mu_0} \, d\mu_0(x) \, dy \]

\[ + \int_Y \int_X \log \left( \frac{d\tilde{\mu}}{d\mu_0} (x) \right) \tilde{\pi}(y|x) \, d\mu_0(x) \, dy \]

\[ = \int_X \int_Y \log \left( \frac{\tilde{\pi}(y|x)}{\pi(y|x)} \right) \tilde{\pi}(y|x) \, d\tilde{\mu}(x) \]

\[ + \int_Y \int_X \log \left( \frac{d\tilde{\mu}}{d\mu_0} (x) \right) \tilde{\pi}(y|x) \, d\mu_0(x) \, dy \]

\[ = D_{\text{KL}}(\tilde{\pi}(\cdot) \parallel \pi(\cdot)) + E^{\tilde{\pi}(\cdot)} \left[ D_{\text{KL}}(\tilde{\mu}(\cdot) \parallel \mu(\cdot)) \right]. \]

Since the KL divergence is always nonnegative, we obtain

\[ D_{\text{KL}}(\tilde{\pi}(\cdot) \parallel \pi(\cdot)) = E^{\mu_0} \left[ D_{\text{KL}}(\tilde{\pi}(\cdot|X) \parallel \pi(\cdot|X)) \right] - E^{\tilde{\pi}(\cdot)} \left[ D_{\text{KL}}(\tilde{\mu}(\cdot) \parallel \mu(\cdot)) \right] \]

\[ \leq E\mu \left[ D_{\text{KL}}(\tilde{\pi}(\cdot|X) \parallel \pi(\cdot|X)) \right], \]

which proves the claim.

Proof of proposition 3.3. Proposition 3.3 follows immediately from (12), lemmas 3.7 and 3.8.
3.2.2. Proof of theorem 3.4. Using proposition 3.3 by making the dependence of the likelihood, evidence and expected information gain on the design variable explicit, and replacing $\hat{\pi}(y|x), \hat{\pi}(y)$ and $U$ by $\pi_N(y|x;d), \pi_N(y;d)$ and $U_N(d)$, respectively, we have

$$|U(d) - U_N(d)| \leq \sqrt{K_1} \sqrt{\mathbb{E}^{d} D_{KL}(\pi_N(\cdot |X;d) \parallel \pi(\cdot |X;d))} + 2\mathbb{E}^{d} D_{KL}(\pi_N(\cdot |X;d) \parallel \pi(\cdot |X;d)),$$

where

$$K_1 := \int_{\mathcal{X}} \int_{\mathcal{Y}} \log^2 \left( \frac{\pi(y|x;d)}{\hat{\pi}(y|x;d)} \right) [\pi(y|x;d) + \pi_N(y|x;d)] \, d\mu_0(x) \, dy.$$

Thanks to assumption 3.1 (A1), \( \mathbb{E}^{d} D_{KL}(\pi_N(\cdot |X;d) \parallel \pi(\cdot |X;d)) \to 0 \) as \( N \to \infty \). Therefore, there exists \( N_0 > 0 \) large enough such that \( \mathbb{E}^{d} D_{KL}(\pi_N(\cdot |X;d) \parallel \pi(\cdot |X;d)) < 1/4 \) for \( N > N_0 \).

Note that by assumption 3.1 (A3), \( K_1 \leq C_0 \) with \( C_0 \) does not depend on \( N \). Now choose \( C = \sqrt{C_0} + 1 \), we arrive at

$$|U(d) - U_N(d)| \leq C \sqrt{\psi(N)},$$

for large \( N \). Take supremum both sides over \( d \in \mathcal{D} \) we conclude the proof of theorem 3.4.

3.2.3. Proof of theorem 3.5. Let \( d \in \mathcal{D} \). Suppose \( d_N \to d \) in \( \mathcal{D} \). We have

$$|U(d) - U_N(d_N)| \leq |U(d) - U_N(d)| + |U(d_N) - U_N(d_N)|$$

$$\leq |U(d) - U_N(d)| + \sup_{d' \in \mathcal{D}} |U(d') - U_N(d')|.$$

By proposition 3.3, where we make the dependence of the likelihood, evidence and expected information gain on the design variable explicit, and with \( \hat{\pi}(y|x), \hat{\pi}(y) \) and \( U \) being replaced by \( \pi(y|x;d_N), \pi(y;d_N) \) and \( U(d_N) \), respectively, it holds that

$$|U(d) - U_N(d_N)| \leq \sqrt{K_2} \sqrt{\mathbb{E}^{d_N} D_{KL}(\pi(\cdot |X;d_N) \parallel \pi(\cdot |X;d_N))}$$

$$+ \mathbb{E}^{d_N} D_{KL}(\pi(\cdot |X;d_N) \parallel \pi(\cdot |X;d_N)),$$

where

$$K_2 := \int_{\mathcal{X}} \int_{\mathcal{Y}} \log^2 \left( \frac{\pi(y|x;d)}{\pi(y;d_N)} \right) [\pi(y|x;d) + \pi_N(y|x;d_N)] \, d\mu_0(x) \, dy.$$

It follows from assumption 3.1 (A3) that there exists \( C_0 > 0 \), independent of \( d \), such that \( K_2 \leq C_0 \). This and (13) imply, by assumption 3.1 (A2),

$$\lim_{N \to \infty} |U(d) - U_N(d_N)| = 0.$$

Now by theorem 3.4, and note that \( \psi(N) \to 0 \) as \( N \to \infty \),

$$\lim_{N \to \infty} \sup_{d' \in \mathcal{D}} |U(d') - U_N(d')| = 0.$$

It follows that, for every \( d_N \to d \),

$$\lim_{N \to \infty} U_N(d_N) = U(d).$$

Therefore, the liminf inequality in definition 2.4 is satisfied (as an equality). The limsup inequality is trivial, since for every \( d \) we can choose the constant sequence \( d_N = d \), and it follows from (14) that \( \limsup_{N \to \infty} U_N(d_N) = \limsup_{N \to \infty} U_N(d) = U(d) \). Thus,

$$U_N \Gamma\text{-converges to } U \text{ as } N \to \infty.$$
The rest follows from the fundamental theorem of \( \Gamma \)-convergence (theorem 2.5). This completes the proof.

**Remark 3.9.** It follows from (15) that \( U_N \) continuously converges to \( U \) (see Dal Maso 1993, definition 4.7), which is strictly stronger than \( \Gamma \)-convergence. In fact, this continuous convergence also implies that \(-U_N \) \( \Gamma \)-converges to \(-U\), which ensures (by symmetry) that the limit of any convergent maximising sequences of \( U_N \) is a maximiser of \( U \). (Note carefully that \( U_N \) \( \Gamma \)-converges to \( U \) does not imply \(-U_N \) \( \Gamma \)-converges to \(-U\), see Braides (2002), example 1.12.)

### 4. Gaussian likelihood in Bayesian inverse problem

In this section, we consider the inverse problem

\[
y = G(x;d) + \epsilon, \tag{16}
\]

for \( x \in \mathcal{X} \) and \( y, \epsilon \in \mathbb{R}^p \), where the noise has multivariate Gaussian distribution \( \epsilon \sim \mathcal{N}(0, \Gamma) \) with some positive definite matrix \( \Gamma \). Suppose we approximate the observation map \( G \) with a surrogate model \( G_N \). Now, thanks to lemma 2.2,

\[
\mathbb{E}^{\mu_0} D_{KL}(\pi_N(\cdot |d) \parallel \pi(\cdot |d)) = \frac{1}{2} \mathbb{E}^{\mu_0} ||G(X,d) - G_N(X,d)||^2_\Gamma,
\]

and

\[
\mathbb{E}^{\mu_0} D_{KL}(\pi(\cdot |d_N) \parallel \pi(\cdot |d)) = \frac{1}{2} \mathbb{E}^{\mu_0} ||G(X,d) - G(X,d_N)||^2_\Gamma, \tag{17}
\]

giving a natural interpretation to (A1) and (A2) in assumption 3.1 in terms of convergence in \( \Gamma \)-weighted \( L^2(\mu_0) \), i.e. \( G_N(X;d) \) should converge to \( G(X;d) \) uniformly in \( d \), while \( G(X;d) \) is required to be \( L^2(\mu_0) \)-continuous with respect to \( d \).

Let us next make the following assumption on \( G \) and \( G_N \).

**Assumption 4.1.** The observation operator and its surrogate version are bounded in \( \Gamma \)-weighted \( L^2(\mu_0) \) uniformly in \( d \), that is, there exists constant \( C_G > 0 \) and such that

\[
\sup_{d \in D} \int_{\mathcal{X}} ||G(x;d)||^2_\Gamma \, d\mu_0(x) \leq C_G, \quad \sup_{d \in D} \int_{\mathcal{X}} ||G_N(x;d)||^2_\Gamma \, d\mu_0(x) \leq C_G
\]

for all \( N \in \mathbb{N} \).

It turns out that the third condition (A3) in assumption 3.1 is implied by assumption 4.1.

**Proposition 4.2.** Let assumption 4.1 hold. It follows that, for Gaussian noise \( \epsilon \sim \mathcal{N}(0, \Gamma) \), there exists a constant \( C \) depending only on \( \text{tr}(\Gamma) \) and \( C_G \) such that

\[
K := \int_{\mathbb{R}^p} \int_{\mathcal{X}} \log^2 \left( \frac{\pi(y|x;d)}{\pi(y;d)} \right) [\pi(y|x;d) + \pi_N(y|x;d)] \, d\mu_0(y) \, dy \leq C. \tag{18}
\]

**Proof.** We rewrite \( K \) as a sum \( K = K_1 + K_2 \), where

\[
K_1 = \int_{\mathbb{R}^p} \int_{\mathcal{X}} \log^2 \left( \frac{\pi(y|x;d)}{\pi(y;d)} \right) \pi(y|x;d) \, d\mu_0(x) \, dy,
\]

and

\[
K_2 = \int_{\mathbb{R}^p} \int_{\mathcal{X}} \log^2 \left( \frac{\pi(y|x;d)}{\pi(y;d)} \right) \pi_N(y|x;d) \, d\mu_0(x) \, dy.
\]
By Cauchy inequality \((a + b)^2 \leq 2(a^2 + b^2)\),

\[
\log^2 \left( \frac{\pi (y|x; d)}{\pi (y;d)} \right) = \log^2 \left( \frac{\exp \left( -\frac{1}{2} \| y - \mathcal{G}(x; d) \|_T^2 \right)}{T \pi(\cdot|x;d)} \right)
\]

\[
= \left( -\frac{1}{2} \| y - \mathcal{G}(x; d) \|_T^2 - \log (T \pi(\cdot|x;d)) \right)^2
\]

\[
\leq \frac{1}{2} \| y - \mathcal{G}(x; d) \|_T^2 + 2 \log^2 (T \pi(\cdot|x;d)),
\]

where \(T = \int_{\mathcal{Y}} \exp \left( -\frac{1}{2} \| y - \mathcal{G}(x; d) \|_T^2 \right) \, dy = \sqrt{2 \pi^p \det(\Gamma)}\). Clearly, \(T \pi(\cdot|x;d) \in (0, 1)\) and since \(\log^2(x)\) is a convex function on \((0, 1)\), by Jensen inequality, we have

\[
\log^2 (E \pi(\cdot|x;d)) \leq E \pi(\cdot|x;d) = \frac{1}{4} E \pi(\cdot|x;d) \leq 1.
\]

Hence,

\[
K_1 \leq \frac{1}{2} \int_{\mathcal{Y}} \int_{\mathcal{X}} \| y - \mathcal{G}(x; d) \|_T^2 \pi(x,y) \, dy \, d\mu_0(x)
\]

\[
+ \frac{1}{2} \int_{\mathcal{X}} \int_{\mathcal{Y}} \int_{\mathcal{Y}} \| y - \mathcal{G}(x; d) \|_T^2 \, d\mu_0(x) \pi(y|x) \, dy \, d\mu_0(\tilde{x})
\]

\[
=: K_{1.1} + K_{1.2}. \tag{19}
\]

For the first integral in (19), we have

\[
K_{1.1} = \frac{1}{2} \int_{\mathcal{X}} \mathbb{E} \pi(\cdot|x;d) \| y - \mathcal{G}(x; d) \|_T^2 \, d\mu_0(x) \leq C, \tag{20}
\]

where the constant \(C\) depends only on \(p\). Now for the second term in (19), by applying Cauchy inequality repeatedly,

\[
K_{1.2} = \frac{1}{2} \int_{\mathcal{X}} \int_{\mathcal{Y}} \int_{\mathcal{Y}} \| y - \mathcal{G}(x; d) \|_T^2 \, d\mu_0(x) \pi(y|x) \, dy \, d\mu_0(\tilde{x})
\]

\[
\leq 4 \int_{\mathcal{X}} \int_{\mathcal{Y}} \int_{\mathcal{Y}} \left( \| y - \mathcal{G}(\tilde{x}; d) \|_T^2 + \| \mathcal{G}(\tilde{x}; d) - \mathcal{G}(x; d) \|_T^2 \right) \pi(y|x) \, dy \, d\mu_0(\tilde{x})
\]

\[
= 4 \int_{\mathcal{X}} \int_{\mathcal{Y}} \| y - \mathcal{G}(x; d) \|_T^2 \pi(y|x) \, dy \, d\mu_0(\tilde{x}) + 4 \int_{\mathcal{X}} \int_{\mathcal{Y}} \| \mathcal{G}(\tilde{x}; d) - \mathcal{G}(x; d) \|_T^2 \, d\mu_0(\tilde{x}) \, d\mu_0(x)
\]

\[
\leq 8K_{1.1} + 32 \left( \int_{\mathcal{X}} \| \mathcal{G}(\tilde{x}; d) \|_T^2 \, d\mu_0(\tilde{x}) + \int_{\mathcal{X}} \| \mathcal{G}(x; d) \|_T^2 \, d\mu_0(x) \right)
\]

\[
\leq 8C + 64C_{G}, \tag{21}
\]

thanks to (20) and assumption 4.1. Therefore, by combining (19)–(21), \(K_1\) is bounded by some universal constant \(C > 0\). Similar arguments apply to the term \(K_2\). This completes the proof. \(\Box\)
Remark 4.3. Notice that we could replace assumption 4.1 by the following conditions:

(B1) there exists \( G \in \mathbb{R}^p \) and \( C > 0 \) such that

\[
\sup_{d \in D} \| G(x, d) - G \|_T \leq C \| x - \mathbb{E}^{\mu_0} X \|_{\mathcal{X}}, \quad \sup_{d \in D} \| G_N(x, d) - G \|_T \leq C \| x - \mathbb{E}^{\mu_0} X \|_{\mathcal{X}}
\]

for all \( x \in \mathcal{X} \) and \( N \in \mathbb{N} \), and

(B2) the prior \( \mu_0 \) has a finite fourth order centred moment.

In particular, Gaussian prior \( \mu_0 =\mathcal{N}(0, \Sigma) \) satisfies (B2) and we find that, thanks to lemma 2.3,

\[
K \leq C(1 + \text{tr}(\Sigma)),
\]

for some universal constant \( C \) depending on \( p \), if (B1) is also satisfied.

Proposition 4.2 leads to the following main result of this section.

Theorem 4.4. Consider the inverse problem (16) with an observation operator and surrogate \( G, G_N : \mathcal{X} \times D \rightarrow \mathbb{R}^p \). Suppose the noise is zero-mean Gaussian, say, \( \epsilon \sim \mathcal{N}(0, \Gamma) \). Let assumption 4.1 hold and assume that

\[
\mathbb{E}^{\mu_0} \| (G(x; d) - G_N(x; d)) \|_T^2 < C \psi(N), \quad \psi(N) \to 0 \text{ as } N \to \infty.
\]

Assume further that \( G \) is a continuous function in \( d \in D \). Then there exists \( C > 0 \) such that

\[
\sup_{d \in D} | U(d) - U_N(d) | \leq C \sqrt{\psi(N)},
\]

for all \( N \) sufficiently large. Moreover, if \( \{d_N^*\} \) is a maximising sequence of \( U_N \) then the limit of any converging subsequence of \( \{d_N^*\} \) is a maximiser of \( U \).

Remark 4.5. Consider the case of linear observation mappings \( G(x; d) = G(d)x \), \( G_N(x; d) = G_N(d)x \), and a Gaussian prior \( \mu_0 =\mathcal{N}(0, C_0) \). It follows that

\[
\mathbb{E}^{\mu_0} \mathbb{D}_{KL}(\pi_N(\cdot | X; d) \parallel \pi(\cdot | X; d)) = \frac{1}{2} \mathbb{E}^{\mu_0} \| (G(d) - G_N(d)) X \|_T^2
\]

\[
= C_0^\frac{1}{2} (G(d) - G_N(d)) \Gamma^{-\frac{1}{2}} \|_{\text{HS}}^2, \tag{22}
\]

where \( \| \cdot \|_{\text{HS}} \) stands for the Hilbert–Schmidt norm. Notice that in the case of linear observation map and Gaussian prior, the expected information gain can be explicitly solved. Consequently, the identity (22) can provide intuition regarding the sharpness of theorem 3.5 as demonstrated in the following example.

Example 1. Consider the simple inference problem with observation map \( \tilde{G}(x) = ax \), \( a > 0 \), with additive normally distributed noise \( \epsilon \). Here, we omit the dependence on \( d \). Suppose the prior distribution \( \mu_0 \) is also normal with Gaussian noise \( \epsilon \sim \mathcal{N}(0, 1) \) and prior \( \mu_0 =\mathcal{N}(0, 1) \). Moreover, suppose we approximate \( \tilde{G}(x) \) with a surrogate \( \tilde{G}_N(x) = a_Nx \) for some \( a \leq a_N \leq C \). A straightforward calculation will now yield
\[ |U_N - U| = \frac{1}{2} \log \frac{a_N^2 + 1}{a^2 + 1}. \]

Now since \( 1 - \frac{1}{2} \leq \log(x) \leq x - 1 \), for all \( x > 0 \), we deduce that
\[ \frac{|a_N - a| (a_N + a)}{2 (a_N^2 + 1)} \leq |U_N - U| \leq \frac{|a_N - a| (a + a)}{2 (a^2 + 1)}. \]

We observe from (23) that the best possible convergence rate is \(|a_N - a|\) and, indeed, by lemma 2.2 we have
\[ \mathbb{E}^\mu D_{KL} (\pi_N (\cdot | X) \| \pi (\cdot | X)) = \frac{1}{2} |a_N - a|^2. \]

In consequence, the convergence rate of theorem 3.4 is asymptotically sharp.

5. Numerical simulations

Let us numerically demonstrate the convergence rates predicted by theorems 3.4 and 3.5 with three examples. Note that these examples were also featured in Huan and Marzouk (2013, 2014) for numerical demonstrations.

5.1. Piecewise linear interpolation in one dimension

Consider a measurement model
\[ y(x, d) = \mathcal{G}(x, d) + \eta \in \mathbb{R}^2, \]
for \( x \in \mathcal{X} = [0, 1] \) and \( d \in \mathcal{D} = [0, 1]^2 \), where \( \eta \sim \mathcal{N} (0, 10^{-4} I) \) and
\[ \mathcal{G}(x, d) = \left[ \begin{array}{c} x^2 d_1^2 + x \exp (-0.2 - d_1) \\ x^2 d_2^2 + x \exp (-0.2 - d_2) \end{array} \right]. \]

As the prior distribution \( \mu_0 \) we assume a uniform distribution on the unit interval, that is \( \mu_0 = U(0, 1) \).

Here we consider a surrogate model \( \mathcal{G}_N \) obtained by piecewise linear interpolation with respect to \( x \). More precisely, the expression of the surrogate model is given by
\[ \mathcal{G}_N (x, d) = \frac{x_i - x}{h_i} \mathcal{G} (x_{i-1}, d) + \frac{x - x_{i-1}}{h_i} \mathcal{G} (x_i, d), \quad x \in [x_{i-1}, x_i], \]
where \( h_i = x_i - x_{i-1} \), and \( i = 1, \ldots, N \).

It is well known that the interpolation of \( f \in H^2 (0, 1) \) on equidistant nodes \( x_0 = 0 < x_1 < x_2 < \cdots < x_N = 1 \) satisfies \( \| f - f_N \|_{L^2(0,1)} \leq CN^{-2} \| f'' \|_{L^2(0,1)} \), see e.g. (Han and Atkinson 2009). Also, notice carefully that for \( d_1 = d_2 = 0 \) we have \( \mathcal{G} \) is linear and the approximation is accurate. In consequence, we have
\[ \sup_{d \in \mathcal{D}} \mathbb{E}^\mu \| \mathcal{G}(X; d) - \mathcal{G}_N (X; d) \|^2 = \sup_{d \in \mathcal{D}} \int_0^1 \| \mathcal{G}(x; d) - \mathcal{G}_N (x; d) \|^2 \, dx \leq CN^{-4}. \]

Moreover, it is straightforward to see that the mapping \( x \mapsto \mathcal{G}(x; d) \) is bounded on the interval \([0, 1]\) uniformly in \( d \) and, therefore, satisfies the assumption 4.1.

We have numerically evaluated both the uniform error \( E_N := \sup_{d \in \mathcal{D}} |U(d) - U_N (d)| \) and the left-hand side term in inequality (27) with varying \( N \). For evaluating \( U(d) \) and \( U_N (d) \) we
use trapezoidal rule for discretising \( X \) using an equidistant grid with 251 nodes. For the data space \( \mathbb{R}^2 \) we utilised Gauss–Hermite–Smolyak quadrature with 3843 nodes. For estimating \( E_N \) and the left-hand side of (27) we fix a \( 21 \times 21 \) grid for the design space \( D \) over which we optimise. Moreover, for any \( d \) we numerically solve \( E_{\mathbb{R}^2} \| G(X; d) - G_N(X; d) \|^2 \) using midpoint rule for the expectation with an equidistant grid in \( X \) with 1001 nodes. The supreme norm is approximated by calculating the maximum value of the evaluations in the grid of \( D \).

In figure 1 we have plotted the expected information gain \( U \) and the errors \( |U_N - U| \) with values \( N = 9, 33 \) and 257. The errors \( E_N \) and \( \sup_{d \in D} E_{\mathbb{R}^2} \| G(X; d) - G_N(X; d) \|^2 \) are plotted in figure 2 for values varying between \( N = 2 \) and \( N = 10^3 \). Moreover, we have also added the theoretical upper bound \( O(N^{-2}) \). We observe that the quantities have the same asymptotic behaviour following the theoretical bound.

5.2. Sparse piecewise linear interpolation in three dimensions

Consider the observation mapping \( G : [0, 1] \times D \to \mathbb{R}^2 \) for \( D = [0, 2, 1]^2 \) defined by formula (25). In this subsection, we formulate a surrogate model \( G_N \) by interpolating data in both \( x \) and \( d \) variables. We apply a piecewise linear interpolation using a sparse grid with
inverse problems 39 (2023) 125008 d-l duong et al

figure 2. the convergence rate predicted by theorem 3.4 is demonstrated for example 5.1. we plot the uniform errors of the expected utility (red curve) and uniform $l^2$-distance of the observation map and the surrogate with respect to the prior distribution (blue curve) with varying $n$. for reference, a line proportional to $n^{-2}$ is plotted.

clenshaw–curtis configuration (le maître and knio 2010). since $g$ has continuous second partial derivatives on its domain, the error of this interpolation can be bounded by

$$
\| g - g_N \|_{L^\infty(X \times d)} = O \left( n^{-2} (\log n)^{3(n-1)} \right),
$$

where $n = 3$ stands for the dimension of the domain, see e.g. (novak and ritter 1996, bungartz 1998, barthelmann et al 2000). for prior, we also assume here that the prior measure is a uniform distribution, $\mu_0 = U(0,1)$. following (27) we immediately observe

$$
\sup_{d \in d} \sqrt{\mathbb{E}^{\mu_0} \| g(X,d) - g_N(X,d) \|^2} = O \left( n^{-2} (\log n)^6 \right).
$$

similar to section 5.1 the mappings $g$ and $g_N$ are bounded and satisfy assumption 4.1.

we again evaluate the difference $e_N := \sup_{d \in d} |u(d) - u_N(d)|$ and the left-hand side term in inequality (29) with varying $n$. in the same manner as the previous example, we estimated the expected information gains $u(d)$ and $u_N(d)$ using quadrature rules. for the integral over the space $x$, we used again trapezoidal rule with 251 equidistant points, while for the space $y$ the gauss–hermite–smolyak quadrature with 1730 nodes. for estimating $e_N$ and the left-hand side of (27) we fix a $31 \times 31$ grid for the design space $d$ over which we optimise. for constructing the surrogate we utilised the algorithm and toolbox detailed in klimke and wohlmuth (2005), klimke (2008). the supreme norm and the expectation with respect to the prior measure are estimated in the same way as the first example, the left-hand side of (29) is evaluated with a grid of $1001 \times 31 \times 31$ nodes on $x \times d$.

figure 3 plots the expected information gain $u$ and the errors $|u_N - u|$ with values $n = 18, 108$ and 632. again, the errors $e_N$ and $\sup_{d \in d} \sqrt{\mathbb{E}^{\mu_0} \| g(X,d) - g_N(X,d) \|^2}$ are plotted in figure 4 for values varying between $n = 5$ and $n = 10^4$. we observe that the numerical
Figure 3. Convergence of the expected utility for example 5.2 is illustrated. The true utility $U(d)$ is plotted w.r.t. $d \in D = [0, 2, 1]^2$ on the left upper corner. The approximation error $|U(d) - U_N(d)|$ is plotted with $N = 18$ (right upper corner), 108 (left bottom corner) and 632 (right bottom corner). Smaller error towards the origin $d_1 = d_2 = 0$ is due to the linearity of $G$ at this point.

Figure 4. The convergence rate predicted by theorem 3.4 is demonstrated for example 5.2. We plot the uniform errors of the expected utility (red curve) and the uniform $L_2$-distance of the observation map $G$ and the surrogate $G_N$ with respect to the prior distribution (blue curve) with varying $N$. For reference, a curve proportional to $N^{-2} (\log N)^6$ is plotted.
convergence rates of the error terms are of the same order while the theoretical upper bound also seems to align asymptotically with these rates asymptotically. After $N = 10^5$ the convergence rates saturate around the value $10^{-7}$ due to limited numerical precision of our implementation.

5.3. Optimal sensor placement for an inverse heat equation

Consider the heat equation on the domain $\Omega \times T = [0, 1]^2 \times [0, 0.4]$ with a source term $S$ with zero Neumann boundary condition and initial value according to

$$\begin{align*}
\frac{\partial v}{\partial t} - \Delta v &= S(\cdot, x), \quad (z,t) \in (0,1)^2 \times T \\
\nabla v \cdot n &= 0, \quad (z,t) \in \partial \Omega \times T \\
v(z,0) &= 0, \quad z \in \Omega
\end{align*}$$

where $n$ is a boundary normal vector. We assume that the source term is given by

$$S(z,t,x) = \frac{s^2}{\pi h^2} e^{-\frac{\|z-x\|^2}{2h^2}}, \quad 0 \leq t < \tau, \quad 0, \quad t \geq \tau.$$ 

with parameter values $s = 2$, $h = 0.05$ and $\tau = 0.3$. Moreover, the parameter $x$ is interpreted as the position of the source. We assume that we can observe $v$ at a location $d_2 \Omega := [0, 1, 0, 0.9]^2 \Omega$ at predefined times $t_i, i = 1, \ldots, 5$. The inverse problem in this setting is to estimate the source location $x \in X := \Omega$ given the data $y = (v(d, t_i))_{i=1}^5 \in \mathbb{R}^5$, i.e. invert the mapping

$$G : \Omega \times D \to \mathbb{R}^5, \quad (x,d) \mapsto y.$$

Here, we consider a Bayesian design problem with the aim to optimise the measurement location $d$ given a uniform prior of the source location $x$ on $(0,1)^2$ and an additive Gaussian noise $\eta \sim \mathcal{N}(0,0.01I)$ in the measurement.

Numerical implementation of $G$ was carried out with a finite difference discretisation for the spatial grid, while a fourth order backward differentiation was used for the temporal discretisation. The surrogate observation mapping $G_N$ is obtained by polynomial chaos expansion with Legendre polynomials. Here, we implemented the projection on an extended domain $\Omega \times \Omega$ instead of $\Omega \times D$ to avoid any potential boundary issues. The implementation follows (Huan and Marzouk 2014) and more details can be found therein. In short, we define $G_N$ as a sum of polynomials

$$G_N(x,d) = \sum_{j=0}^{N} G_j \Psi_j(x,d),$$

where $j \in \mathbb{N}^4$ and $\{\Psi_j\}_{j \in \mathbb{N}^4}$ are a Legendre polynomial basis in $\Omega \times \Omega$ with the standard inner product, i.e. the $L^2$-inner product weighted by the uniform prior. Moreover, the coefficients satisfy

$$G_j = \frac{\int_{\Omega \times \Omega} G(x; d) \Psi_j(x, d) \, dx \, dy}{\int_{\Omega \times \Omega} \Psi_j^2(x, d) \, dx \, dy}.$$ (30)
The parameter $N$ is the truncation level of the polynomial chaos. Notice carefully that we have included the design parameter $d$ in the approximation.

We compute the coefficients $G_j$ with a Gauss–Legendre–Smolyak quadrature with Clenshaw–Curtis configuration with a high number (of the order $10^7$) of grid points and assume in the following that the truncation level $N$ is the dominating factor for the surrogate error $\sup_{d \in \Omega} \| G - G_N \|$.

The utility functions were estimated as follows: for the integral over the data space $\mathbb{R}^5$ we used the Gauss–Hermite–Smolyak quadrature with 117 nodes, while on the domain $\Omega$ we implemented a bidimensional Clenshaw–Curtis–Smolyak quadrature with 7682 nodes. For the evaluation of maximal difference of the expected utility, we fixed an equidistant grid with $n = 40$ nodes in $D$. For evaluating the expectation in the term $\sup_{d \in D} \mathbb{E}_{\mu^d} \left[ \| G(X; d) - G_N(X; d) \|^2 \right]$ we use midpoint rule and a grid of 25 nodes in each direction on $\mathcal{X}$. The supreme norm is approximated by calculating the maximum element of $25 \times 25$ nodes in a grid of $D$.

In figure 5 we show the approximation of the utility functions for every polynomial degree. In both cases we can see a visual convergence with respect to the polynomial degree. In the spirit of previous examples, figure 6 contains the errors between the surrogate models and the utility functions. Compared to the previous examples, the complexity of this computational task is substantially larger inducing larger numerical errors especially in the evaluation of $E_N$.

Also, as is seen from figure 5, the error close to the boundary of $\Omega$ is converging substantially slower. However, as illustrated by figure 6 the two rates of convergence are aligned with $\mathcal{D}$. 

Figure 5. Convergence of the expected utility for example 5.3 is illustrated. The true utility $U(d)$ is plotted on the left upper corner. The approximation error $|U(d) - U_N(d)|$ is plotted with $N = 2$ (right upper corner), 6 (left bottom corner) and 14 (right bottom corner). The design space $\mathcal{D}$ is highlighted with a black box.
Figure 6. The convergence rate predicted by theorem 3.4 is demonstrated for example 5.3. We plot the uniform errors of the expected utility (red curve) and the uniform $L^2$-distance of the observation map $G$ and the polynomial chaos surrogate $G_n$ with respect to the prior distribution (blue curve) with varying $N$.

6. Conclusion

We have developed a framework to study the stability of the expected information gain for optimal experimental designs in infinite-dimensional Bayesian inverse problems. We showed a uniform convergence for the expected information gain, with a sharp rate, given approximations of the likelihood. In the case of Bayesian inverse problems with Gaussian noise, this rate is proved to coincide with the $L^2$-convergence of the observation maps with respect to the prior distribution. Moreover, we also showed that the optimal design variable is also stable in the approximation. The results are illustrated by three numerical experiments.

Possible extensions of this work naturally include considering the stability of various other utilities such as the negative least square loss (Chaloner and Verdinelli 1995) or utilities related to Bayesian optimisation (Shahriari et al. 2015). Here, we only considered perturbations of the utility induced by a surrogate likelihood model. However, a further numerical (such as MC based) approximation for the expected utility is most of the time needed. A number of results considering convergence with respect to MC error for fixed design have been provided (see e.g. Ryan et al. 2016 and references therein) while the uniform case has not been addressed to the best of our knowledge. Finally, $\Gamma$-convergence does not directly provide a convergence rate for the optimal designs, which remains an interesting open problem.

Data availability statement

All data that support the findings of this study are included within the article (and any supplementary files).
Acknowledgments

The work of T H and D L D was supported by the Academy of Finland (Decision Numbers 326961 and 345720). J R R G was supported by the LUT Doctoral School.

ORCID IDs

Duc-Lam Duong https://orcid.org/0000-0002-3139-7309
Jose Rodrigo Rojo-Garcia https://orcid.org/0009-0008-2072-3480

References

Alexanderian A 2021 Optimal experimental design for infinite-dimensional Bayesian inverse problems governed by PDEs: a review Inverse Problems 37 043001
Alexanderian A, Gloor P J and Ghattas O 2016a On Bayesian A- and D-optimal experimental designs in infinite dimensions Bayesian Anal. 11 671–95
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 J. Sci. Comput. 36 A2122–48
Alexanderian A, Petra N, Stadler G and Ghattas O 2016b A fast and scalable method for A-optimal design of experiments for infinite-dimensional Bayesian nonlinear inverse problems SIAM J. Sci. Comput. 38 A243–72
Barthelmann V, Novak E and Ritter K 2000 High dimensional polynomial interpolation on sparse grids Adv. Comput. Math. 12 273–88
Beck J, Dia B M, Espath L F, Long Q and Tempone R 2018 Fast Bayesian experimental design: Laplace-based importance sampling for the expected information gain Comput. Methods Appl. Mech. Eng. 334 523–53
Beck J, Mansour Dia B, Espath L and Tempone R 2020 Multilevel double loop Monte Carlo and stochastic collocation methods with importance sampling for Bayesian optimal experimental design Int. J. Numer. Methods Eng. 121 3482–503
Braides A 2002 Gamma-Convergence for Beginners vol 22 (Clarendon)
Bungartz H J 1998 Finite Elements of Higher Order on Sparse Grids (Berichte aus der Informatik) (Shaker)
Burger M, Hauptmann A, Helin T, Hyvönen N and Puska J-P 2021 Sequentially optimized projections in x-ray imaging Inverse Problems 37 075006
Chaloner K and Verdinelli I 1995 Bayesian experimental design: a review Stat. Sci. 10 273–304
Da Prato G and Zabczyk J 2014 Stochastic Equations in Infinite Dimensions (Cambridge University Press)
Dal Maso G 1993 An Introduction to $\Gamma$-Convergence vol 8 (Springer)
Dashi M and Stuart A M 2017 The Bayesian approach to inverse problems Handbook of Uncertainty Quantification (Springer) pp 311–428
Engl H W, Hanke M and Neubauer A 1996 Regularization of Inverse Problems vol 375 (Springer)
Foster A, Jankowiak M, Bingham E, Horsfall P, Teh Y W, Rainforth T and Goodman N 2019 Variational Bayesian optimal experimental design Neural Information Processing Systems (NIPS’19) p 32
Garbuno-Inigo A, Helin T, Hoffmann F and Hosseini B 2023 Bayesian posterior perturbation analysis with integral probability metrics (arXiv:2303.01512)
Haber E, Horesh L and Tenorio I 2008 Numerical methods for experimental design of large-scale linear ill-posed inverse problems Inverse Problems 24 055012
Han W and Atkinson K E 2009 Theoretical Numerical Analysis: A Functional Analysis Framework, 3rd edn (Springer)
Helin T, Hyvönen N and Puska J-P 2022 Edge-promoting adaptive Bayesian experimental design for x-ray imaging SIAM J. Sci. Comput. 44 B506–30
Herrmann L, Schwab C and Zech J 2020 Deep neural network expression of posterior expectations in Bayesian PDE inversion Inverse Problems 36 125011
Huan X and Marzouk Y M 2013 Simulation-based optimal Bayesian experimental design for nonlinear systems J. Comput. Phys. 232 288–317
Huan X and Marzouk Y 2014 Gradient-based stochastic optimization methods in Bayesian experimental design Int. J. Uncertain. Quantification 4 479–510
Kaipio J and Somersalo E 2006 Statistical and Computational Inverse Problems vol 160 (Springer)
Klimke A 2008 Sparse grid interpolation toolbox–User’s guide (Universität Stuttgart) (available at: www.researchgate.net/publication/330956900_Sparse_Grid_Interpolation_Toolbox_ User’s_Guide)
Klimke A and Wohlmuth B 2005 Algorithm 847: spinterp: piecewise multilinear hierarchical sparse grid interpolation in MATLAB ACM Trans. Math. Softw. 31 561–79
Le Maître O and Knio O M 2010 Spectral Methods for Uncertainty Quantification: With Applications to Computational Fluid Dynamics (Springer)
Long Q, Motamed M and Tempone R 2015 Fast Bayesian optimal experimental design for seismic source inversion Comput. Methods Appl. Mech. Eng. 291 123–45
Long Q, Scavino M, Tempone R and Wang S 2015 A Laplace method for under-determined Bayesian optimal experimental designs Comput. Methods Appl. Mech. Eng. 285 849–76
Marzouk Y M, Najm H N and Rahn L A 2007 Stochastic spectral methods for efficient Bayesian solution of inverse problems J. Comput. Phys. 224 560–86
Nickl R 2023 Bayesian Non-Linear Statistical Inverse Problems (Zurich Lectures in Advanced Mathematics)(EMS Press) (https://doi.org/10.4171/ZLAM/30)
Novak E and Ritter K 1996 High dimensional integration of smooth functions over cubes Numer. Math. 75 79–97
Pukelshein F 2006 Optimal Design of Experiments (SIAM)
Ruthotto L, Chung J and Chung M 2018 Optimal experimental design for inverse problems with state constraints SIAM J. Sci. Comput. 40 B1080–100
Ryan E G, Drovandi C C, McGree J M and Pettitt A N 2016 A review of modern computational algorithms for Bayesian optimal design Int. Stat. Rev. 84 128–54
Schillings C and Schwab C 2013 Sparse, adaptive Smolyak quadratures for Bayesian inverse problems Inverse Problems 29 065011
Shahriari B, Swersky K, Wang Z, Adams R P and De Freitas N 2015 Taking the human out of the loop: a review of Bayesian optimization Proc. IEEE 104 148–75
Sprungk B 2020 On the local Lipschitz stability of Bayesian inverse problems Inverse Problems 36 055015
Steinberg D M and Hunter W G 1984 Experimental design: review and comment Technometrics 26 71–97
Stuart A M 2010 Inverse problems: a Bayesian perspective Acta Numer. 19 451–559
Wu K, Chen P and Ghattas O 2020 A fast and scalable computational framework for large-scale and high-dimensional Bayesian optimal experimental design (arXiv:2010.15196)