A Laplace Method for Under-Determined Bayesian Optimal Experimental Designs

Quan Long [a,b,*], Marco Scavino[a,1], Raúl Tempone[a], Suojin Wang[d]

[a] *CEMSE, King Abdullah University of Science and Technology, Thuwal, 23955-6900, KSA*
[b] *ICES, University of Texas at Austin, Austin, 78712-1299, USA*
[c] *Instituto de Estadística (IESTA), Universidad de la República, Montevideo, Uruguay*
[d] *Department of Statistics, Texas A & M University, College Station, TX, 77843, USA*

Abstract

In [1], a new method based on the Laplace approximation was developed to accelerate the estimation of the post-experimental expected information gains (Kullback-Leibler divergence) in model parameters and predictive quantities of interest in the Bayesian framework. A closed-form asymptotic approximation of the inner integral and the order of the corresponding dominant error term were obtained in the cases where the parameters are determined by the experiment. In this work, we extend that method to the general case where the model parameters cannot be determined completely by the data from the proposed experiments. We carry out the Laplace approximations in the directions orthogonal to the null space of the Jacobian matrix of the data model with respect to the parameters, so that the information gain can be reduced to an integration against the marginal density of the transformed parameters that are not determined by the experiments. Furthermore, the expected information gain can be approximated by an integration over the prior, where the integrand is a function of the posterior covariance matrix projected over the aforementioned orthogonal directions. To deal with the issue of dimensionality in a complex problem, we use either Monte Carlo sampling or sparse quadratures for the integration over the prior probability density function, depending on the regularity of the integrand function. We demonstrate the accuracy, efficiency and robustness of the proposed method via several nonlinear under-determined test cases. They include the designs of the scalar parameter in a one dimensional cubic polynomial function with two unidentifiable parameters forming a linear manifold, and the boundary source locations for impedance tomography in a square domain, where the unknown parameter is the conductivity, which is represented as a random field.

Keywords: Bayesian statistics; Optimal experimental design; Information gain; Laplace approximation; Monte Carlo Sampling; Sparse quadrature; Uncertainty quantification.

AMS subject classification: 62K05, 65N21, 65C60

*Corresponding author. Tel.: +966-02-808-0396. Email: quan.long@kaust.edu.sa

1. Introduction

In the Bayesian framework, the usefulness of a proposed experiment is usually measured by the expected information gain, i.e., the expected log ratio between the posterior and the prior [2]. The expected information gain is also equivalent to the mutual information between the parameters or quantity of interest and the observables, and is associated with the Bayesian D-optimality for a normal linear model in the literature[3, p.277]. The computation of the expected information gain, however, may become extremely expensive when the outcomes of the experiment are modeled as functions of the solution of Partial Differential Equations (PDEs). Using the Laplace approximation [4–7], we proposed in [1] a fast approach for the estimation of the expected information gain and analyzed the rates of different dominant error terms with respect to the amount of data in each experimental scenario, provided that the parameters can be determined completely by the experiments in the sense that a single dominant maximum a posteriori probability (MAP) estimate exists. When both the determinant of the posterior covariance matrix and the prior probability density functions (pdf) have enough regularities with respect to the random parameters, we demonstrated, with several nonlinear examples involving the solutions of PDEs, that sparse quadratures can be employed to carry out the resulting integrations with high efficiency.

In this work, we extend our methodology in [1] to the cases in which the random parameters cannot be completely determined by the experiments, i.e., under-determined parameters. Thus, we assume here that there is an imbedded manifold on which the parameters are not informed by the data. In this context, the posterior pdf will start to concentrate around this manifold as the amount of data increases. The key innovation of our novel extension is, hence, performing the normality approximation for the conditional posterior pdf given a fixed point on the manifold and a Laplace approximation of the conditional expected information gain. Both are carried out in directions that are orthogonal to the non-informative manifold. Asymptotic expansions of the expected Kullback-Leibler divergence between the posterior and prior pdfs for determined models have been derived by several authors using the likelihood ratio process. See, for instance, [8] for an interesting connection with an information-theoretic framework of the Bayesian Central Limit Theorem, [9] for an extension to non i.i.d. regular models for experimental designs, and [10] for the analysis of non-regular models when the posterior distribution is consistent. Other works were inspired by [11], whose motivation was to justify the use of certain prior distributions in Bayesian statistical analysis by maximizing the Shannon mutual information between the parameter vector and the data, also in the presence of nuisance parameters in the model. These works, for example, [12–14], making an explicit distinction between parameters of interest and nuisance parameters in the model, can be used, in principle, to obtain asymptotic expressions of the expected information gain in under-determined models. However, their applicability is confined to the case where the non-informative manifold can be explicitly parametrized in terms of the nuisance parameters. Instead, our approach does not require such an explicit representation of the underlying manifold where the posterior distribution concentrates. On the other hand, an explicit representation of the manifold is practically not possible. In Section 2, we formulate our new methodology for parameter inferences: we first introduce

the information gain and the expected information gain. We then reparameterize the prior and posterior pdfs, using two sets of local parameters, $\boldsymbol{t}$ and $\boldsymbol{s}$, separately. The $\boldsymbol{t}$ direction parameterizes the non-informative manifold and $\boldsymbol{s}$ parameterizes the directions orthogonal to it. Next, the Laplace approximation is carried out along the $\boldsymbol{s}$ direction, conditioned on a fixed $\boldsymbol{t}$ value. Finally, the information gain is expressed as an integral along the $\boldsymbol{t}$ direction. With an extra integral over the data, we obtain the asymptotic formulation of the expected information gain. Section 3 applies similar ideas to the prediction of quantities of interest. After a brief description of the adopted method for numerical integrations in Section 4, several numerical examples are presented in Section 5, including the designs of the scalar parameter in a one-dimensional function with two unidentifiable parameters, a one-dimensional function with two unidentifiable parameters but a different manifold, and the boundary source locations for impedance tomography in a square domain. We use both a sampling method and a polynomial-based sparse quadrature method to carry out the numerical integration, depending on the problem dimensionality and the available regularity.

Nomenclature

$\mathbf{1}_{\Omega_M}$ an indicator function that takes the value of 1 when $\boldsymbol{\theta} \in \Omega_M$, 0 otherwise

Δ_k the length of the k^{th} segment

$\boldsymbol{U}_h$ the nodal voltage vector

$h_p(\boldsymbol{s}, \boldsymbol{t})$ the logarithm of the prior weight function $p(\boldsymbol{s}, \boldsymbol{t})$ i.e., $log[p(\boldsymbol{s}, \boldsymbol{t})]$

$h(\boldsymbol{s}, \boldsymbol{t})$ the logarithm of posterior weight function, $p(\boldsymbol{s}, \boldsymbol{t}|\bar{\boldsymbol{y}})$ i.e., $\log[p(\boldsymbol{s}, \boldsymbol{t}|\bar{\boldsymbol{y}})]$

$\bar{\boldsymbol{y}} = \{\boldsymbol{y}\}_{i=1}^{M}$ a set of observed data points

ϵ_Q the prediction error

tr the trace of a matrix

$\nabla_{\boldsymbol{s}}$ the gradient in $\boldsymbol{s}$

$\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{s}|\boldsymbol{t}}$ the approximate conditional covariance matrix

$\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{s}|\boldsymbol{t}} = \frac{1}{M}\{\boldsymbol{U}^T[\boldsymbol{J}_g(\boldsymbol{f}(\hat{\boldsymbol{s}}, \boldsymbol{t}))^T \boldsymbol{\Sigma}_\epsilon^{-1} \boldsymbol{J}_g(\boldsymbol{f}(\hat{\boldsymbol{s}}, \boldsymbol{t}))]\boldsymbol{U}\}^{-1}$

$\mathbf{1}_{\boldsymbol{s}=\hat{\boldsymbol{s}}}$ an indicator function, which takes the value of 1 when $\boldsymbol{s} = \hat{\boldsymbol{s}}$, and takes the value 0 otherwise

$\boldsymbol{E_s}$ the sum of the data residuals, i.e., $\boldsymbol{E_s} = \sum_{i=1}^{M} \boldsymbol{r}_i$

$\boldsymbol{F}_h$ the force vector

$\boldsymbol{H_g}$ the Hessian of model $\boldsymbol{g}$ w.r.t. the parameter $\boldsymbol{\theta}$

$\boldsymbol{H_s}$ the Hessian of model $\boldsymbol{g}$ w.r.t. the parameter $\boldsymbol{s}$

$\boldsymbol{J_g}$ the Jacobian of model $\boldsymbol{g}$ w.r.t. the parameter $\boldsymbol{\theta}$

$\boldsymbol{J_s}$ the Jacobian of $\boldsymbol{g}$ w.r.t. the parameter $\boldsymbol{s}$

$\boldsymbol{K}(\boldsymbol{\theta})$ the stiffness matrix

$\boldsymbol{n}$ the normal vector to the boundary

$\boldsymbol{r}_i$ the i^{th} residual vector, i.e., $\boldsymbol{r}_i = \boldsymbol{g}(\boldsymbol{\theta}_0) + \boldsymbol{\epsilon}_i - \boldsymbol{g}(\boldsymbol{\theta})$

$\boldsymbol{U}$ the matrix whose columns are the basis corresponding to the positive eigenvalues of $\boldsymbol{H}(\boldsymbol{f}(\boldsymbol{0}, \boldsymbol{t}))$

$\boldsymbol{V}$ the matrix whose columns are the basis corresponding to the zero eigenvalues of $\boldsymbol{H}(\boldsymbol{f}(\boldsymbol{0}, \boldsymbol{t}))$

$\boldsymbol{y}_i$ the i^{th} $\boldsymbol{s} \times 1$ observable response vector

Λ a diagonal matrix containing the eigenvalues of $\boldsymbol{H}(\boldsymbol{f}(\boldsymbol{0}, \boldsymbol{t}))$

$\boldsymbol{\xi}$ the $r \times 1$ vector of design parameters, also known as the experimental setup

$\boldsymbol{\epsilon}_1, ..., \boldsymbol{\epsilon}_M$ i.i.d. $s \times 1$ error vectors, with $\boldsymbol{\epsilon}_1 \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_\epsilon)$

$\boldsymbol{\theta}_0$ the $d \times 1$ true parameter vector used to generate the synthetic data

a_j the subset of the boundary corresponding to the j^{th} electrode

d the dimension of parameter vector $\boldsymbol{\theta}_0$

$d_{\boldsymbol{g}}$ the dimension of measurement vector $\boldsymbol{g}$

$H^1(\Omega)$ the Sobolev space with a square integrable gradient

$H_p(\hat{\boldsymbol{s}}, \boldsymbol{t})$ the Hessian of $h_p(\hat{s}, \boldsymbol{t})$

l the total number of electrodes

M the total number of observations

NQ the number of quadrature points

$NS1$ the number of points in a one-dimensional mesh partitioning the domain of scalar Q

$O(\cdot)$ the big O notation

$O_P(\cdot)$ the big O in probability

$\mathbb{P}(\cdot)$ the probability measure

$p_{\boldsymbol{\Theta}}(\boldsymbol{\theta})$ the prior of the unknown random parameter $\boldsymbol{\theta}$

$p_{\boldsymbol{\Theta}}(\boldsymbol{\theta}|\bar{\boldsymbol{y}})$ the posterior pdf of the unknown random parameter $\boldsymbol{\theta}$

$p_{\mathcal{Y}}(\bar{\boldsymbol{y}}|\boldsymbol{\xi})$ the Bayesian evidence, defined as the marginalization of likelihood over all admissible parameters

$p_{\boldsymbol{s}}(\mathbf{0}) = \int_{T_t} p(\mathbf{0}, \boldsymbol{t}) d\boldsymbol{t}$ the marginal of prior pdf of parameter $\boldsymbol{s}$

Q the quantity of interest

U_j the measured voltage on the j^{th} electrode and part of the solution of the weak form of the Poisson equation

V_h the finite element subspace of V, i.e., $V_h := \{v \in V : v \text{ is piecewise linear continuous over } \Omega_h\}$

w_i the weights for the i^{th} quadrature points

$\boldsymbol{g} : \mathbb{R}^d \times \mathbb{R}^k \to \mathbb{R}^s$ a deterministic nonlinear mapping

2. Estimation of the expected information gain for an under-determined model

This section contains the main theoretical results about the global approximation of the expected information gain as a function of the experimental setup, when making Bayesian inferences for the parameter vector of an under-determined statistical model.

2.1. The model, information gain and expected information gain

Suppose that the data-generating process has the form

$$\boldsymbol{y}_i = \boldsymbol{g}(\boldsymbol{\theta}_0, \boldsymbol{\xi}) + \boldsymbol{\epsilon}_i\,, \quad i = 1, \ldots, M, \tag{1}$$

where we recall that $\boldsymbol{\epsilon}_1, \ldots, \boldsymbol{\epsilon}_M$ are i.i.d. $s \times 1$ random vectors, with $\boldsymbol{\epsilon}_1 \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_\epsilon)$, as defined into the nomenclature.

Let $p(\boldsymbol{\theta})$ be a prior density function for the parameter vector $\boldsymbol{\theta}$ in model (1). Given the sample of observed response vectors $(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_M) =: \bar{\boldsymbol{y}}$, the Kullback-Leibler (K-L) divergence (information gain) and the expected K-L divergence (expected information gain) for each experimental setup, $\boldsymbol{\xi}$, are defined as

$$D_{KL}(\bar{\boldsymbol{y}}, \boldsymbol{\xi}) := \int_{\boldsymbol{\Theta}} \log\left(\frac{p_{\boldsymbol{\Theta}}(\boldsymbol{\theta}|\bar{\boldsymbol{y}}, \boldsymbol{\xi})}{p_{\boldsymbol{\Theta}}(\boldsymbol{\theta})}\right) p_{\boldsymbol{\Theta}}(\boldsymbol{\theta}|\bar{\boldsymbol{y}}, \boldsymbol{\xi}) d\boldsymbol{\theta}\,,$$
$$I(\boldsymbol{\xi}) := \int_{\mathcal{Y}} \int_{\boldsymbol{\Theta}} \log\left(\frac{p_{\boldsymbol{\Theta}}(\boldsymbol{\theta}|\bar{\boldsymbol{y}}, \boldsymbol{\xi})}{p_{\boldsymbol{\Theta}}(\boldsymbol{\theta})}\right) p_{\boldsymbol{\Theta}}(\boldsymbol{\theta}|\bar{\boldsymbol{y}}, \boldsymbol{\xi}) p_{\mathcal{Y}}(\bar{\boldsymbol{y}}|\boldsymbol{\xi}) d\boldsymbol{\theta} d\bar{\boldsymbol{y}}\,.$$

The condition $\boldsymbol{\xi}$ will be dropped for the sake of conciseness in the rest of the paper.

2.2. The non-informative manifold of parameters

When an under-determined (or unidentifiable) model is considered, the true parameter vector, $\boldsymbol{\theta}_0$, cannot be uniquely inferred through point estimation, even if the number M of available observations is made arbitrarily large. In such a case, the set of points where the likelihood function reaches its maximum does not reduce to $\boldsymbol{\theta}_0$ but, regardless of the value of $\boldsymbol{\xi}$, concentrates around the following set in $\mathbb{R}^d$:

$$T(\boldsymbol{\theta}_0) := \{\boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \mathbb{R}^d \,:\, p(\bar{\boldsymbol{y}}|\boldsymbol{\theta}) - p(\bar{\boldsymbol{y}}|\boldsymbol{\theta}_0) = \mathbf{0}, \quad \text{as} \quad M \to \infty\}\,. \tag{2}$$

Under the assumption that the nonlinear mapping, $\boldsymbol{g}$, is twice continuously differentiable, $T(\boldsymbol{\theta}_0)$ is a compact $(d-r)$-dimensional $\mathcal{C}^2$-submanifold of $\mathbb{R}^d$, where r is the codimension of the manifold T. In the case where we consider Gaussian measurement noise, the above definition of the non-informative manifold is equivalent to

$$T(\boldsymbol{\theta}_0) := \{\boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \mathbb{R}^d \,:\, g(\boldsymbol{\theta}) - g(\boldsymbol{\theta}_0) = \mathbf{0}\}\,. \tag{3}$$

Consider the tubular neighborhoods

$$\Omega_M(\boldsymbol{\theta}_0) := \{\boldsymbol{\theta} \in \mathbb{R}^d : \|g(\boldsymbol{\theta}_0) - g(\boldsymbol{\theta})\|^2_{\boldsymbol{\Sigma}_\epsilon} \leq \ell(M)\}\,,$$

with $\|\boldsymbol{a}\|^2_{\Sigma_\epsilon} := \boldsymbol{a}^T \boldsymbol{\Sigma}_\epsilon^{-1} \boldsymbol{a}$, and $\ell(M)$ is a strictly positive non-increasing function of M.

As a preliminary result, we show that the posterior probability distribution concentrates on $T(\boldsymbol{\theta}_0)$. We adapt to our framework the technique used in [15] and in [16] reproduced here for the sake of completeness. The following lemma is established based on the data generating model (1).

Lemma 1. *Assume that the prior probability distribution assigns a positive mass to any neighborhood of $T(\boldsymbol{\theta}_0)$, and $\ell(M) = \ell_0 M^{-\beta}$ with $\ell_0 > 0$ and $\frac{1}{2} > \beta > 0$. Given any $c > 0$, there exists an $\tilde{M}(c) > 0$ such that, for any $M \geq \tilde{M}(c)$ it holds that*

$$\mathbb{P}_\epsilon(\mathbb{P}(\boldsymbol{\theta} \in \boldsymbol{\Theta}\backslash\Omega_M(\boldsymbol{\theta}_0) \mid \bar{\boldsymbol{y}}) > c) \leq \left(\frac{16c_g\sqrt{s}}{\sqrt{M}\ell(M) - 16c_g\sqrt{s}}\right)^2, \tag{4}$$

where $\mathbb{P}_\epsilon$ is the probability measure generated by the random vectors $\boldsymbol{\epsilon}_1, \ldots, \boldsymbol{\epsilon}_M$ under model (1) with $s = dim(\boldsymbol{\epsilon}_1)$, and given that there exists $c_g > 0$ such that

$$\|g(\boldsymbol{\theta}_0) - g(\boldsymbol{\theta})\|_{\Sigma_\epsilon} \leq c_g \tag{5}$$

for any $\boldsymbol{\theta} \in \boldsymbol{\Theta}$.

Proof

Given the observations $\bar{\boldsymbol{y}}$ the log-likelihood function for the model (1) is given by

$$\begin{aligned}\log L(\boldsymbol{\theta}|\bar{\boldsymbol{y}}) &= -\frac{sM}{2}\log(2\pi) - \frac{M}{2}\log|\boldsymbol{\Sigma}_\epsilon| - \frac{M}{2}||\boldsymbol{g}(\boldsymbol{\theta}_0) - \boldsymbol{g}(\boldsymbol{\theta})||^2_{\Sigma_\epsilon} \\ &\quad - \frac{1}{2}\sum_{i=1}^{M}||\boldsymbol{\epsilon}_i||^2_{\boldsymbol{\Sigma}_\epsilon} - \sum_{i=1}^{M}\boldsymbol{\epsilon}_i^T\boldsymbol{\Sigma}_\epsilon^{-1}(\boldsymbol{g}(\boldsymbol{\theta}_0) - \boldsymbol{g}(\boldsymbol{\theta}))\,.\end{aligned}$$

We have

$$\begin{aligned}\mathbb{P}(\boldsymbol{\theta} \in \boldsymbol{\Theta}\backslash\Omega_M(\boldsymbol{\theta}_0) \mid \bar{\boldsymbol{\epsilon}}) = \mathbb{P}(\boldsymbol{\theta} &\in \boldsymbol{\Theta}\backslash\Omega_M(\boldsymbol{\theta}_0) \mid \bar{\boldsymbol{y}}) = \frac{\int_{\boldsymbol{\Theta}\backslash\Omega_M(\boldsymbol{\theta}_0)} L(\boldsymbol{\theta}|\bar{\boldsymbol{y}})p(\boldsymbol{\theta})d\boldsymbol{\theta}}{\int_{\boldsymbol{\Theta}} L(\boldsymbol{\theta}|\bar{\boldsymbol{y}})p(\boldsymbol{\theta})d\boldsymbol{\theta}} \\ &\leq \frac{\int_{\boldsymbol{\Theta}\backslash\Omega_M(\boldsymbol{\theta}_0)} e^{-\frac{M}{2}\|\boldsymbol{g}(\boldsymbol{\theta}_0)-\boldsymbol{g}(\boldsymbol{\theta})\|^2_{\boldsymbol{\Sigma}_\epsilon} - \sum_{i=1}^{M}\boldsymbol{\epsilon}_i^T\boldsymbol{\Sigma}_\epsilon^{-1}(\boldsymbol{g}(\boldsymbol{\theta}_0)-\boldsymbol{g}(\boldsymbol{\theta}))}p(\boldsymbol{\theta})d\boldsymbol{\theta}}{\int_{\boldsymbol{\Theta}} \mathbf{1}_{\{\|\boldsymbol{g}(\boldsymbol{\theta}_0)-\boldsymbol{g}(\boldsymbol{\theta})\|^2_{\boldsymbol{\Sigma}_\epsilon} \leq \frac{\ell(M)}{2}\}} e^{-\frac{M}{2}||\boldsymbol{g}(\boldsymbol{\theta}_0)-\boldsymbol{g}(\boldsymbol{\theta})||^2_{\boldsymbol{\Sigma}_\epsilon} - \sum_{i=1}^{M}\boldsymbol{\epsilon}_i^T\boldsymbol{\Sigma}_\epsilon^{-1}(\boldsymbol{g}(\boldsymbol{\theta}_0)-\boldsymbol{g}(\boldsymbol{\theta}))}p(\boldsymbol{\theta})d\boldsymbol{\theta}} \\ &\leq \frac{e^{-\frac{M}{4}\ell(M)}\int_{\boldsymbol{\Theta}\backslash\Omega_M(\boldsymbol{\theta}_0)} e^{-\sum_{i=1}^{M}\boldsymbol{\epsilon}_i^T\boldsymbol{\Sigma}_\epsilon^{-1}(\boldsymbol{g}(\boldsymbol{\theta}_0)-\boldsymbol{g}(\boldsymbol{\theta}))}p(\boldsymbol{\theta})d\boldsymbol{\theta}}{\int_{\boldsymbol{\Theta}} \mathbf{1}_{\{||\boldsymbol{g}(\boldsymbol{\theta}_0)-\boldsymbol{g}(\boldsymbol{\theta})||^2_{\boldsymbol{\Sigma}_\epsilon} \leq \frac{\ell(M)}{2}\}} e^{-\sum_{i=1}^{M}\boldsymbol{\epsilon}_i^T\boldsymbol{\Sigma}_\epsilon^{-1}(\boldsymbol{g}(\boldsymbol{\theta}_0)-\boldsymbol{g}(\boldsymbol{\theta}))}p(\boldsymbol{\theta})d\boldsymbol{\theta}} \\ &\leq \frac{e^{-\frac{M}{4}\ell(M)}e^{\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}|\sum_{i=1}^{M}\boldsymbol{\epsilon}_i^T\boldsymbol{\Sigma}_\epsilon^{-1}(\boldsymbol{g}(\boldsymbol{\theta}_0)-\boldsymbol{g}(\boldsymbol{\theta}))|}\int_{\boldsymbol{\Theta}\backslash\Omega_M(\boldsymbol{\theta}_0)} p(\boldsymbol{\theta})d\boldsymbol{\theta}}{e^{-\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}|\sum_{i=1}^{M}\boldsymbol{\epsilon}_i^T\boldsymbol{\Sigma}_\epsilon^{-1}(\boldsymbol{g}(\boldsymbol{\theta}_0)-\boldsymbol{g}(\boldsymbol{\theta}))|}\int_{\boldsymbol{\Theta}} \mathbf{1}_{\{||\boldsymbol{g}(\boldsymbol{\theta}_0)-\boldsymbol{g}(\boldsymbol{\theta})||^2_{\boldsymbol{\Sigma}_\epsilon} \leq \frac{\ell(M)}{2}\}} p(\boldsymbol{\theta})d\boldsymbol{\theta}} \\ &\leq \frac{e^{2\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}|\sum_{i=1}^{M}\boldsymbol{\epsilon}_i^T\boldsymbol{\Sigma}_\epsilon^{-1}(\boldsymbol{g}(\boldsymbol{\theta}_0)-\boldsymbol{g}(\boldsymbol{\theta}))|}e^{-\frac{M}{4}\ell(M)}}{\int_{\boldsymbol{\Theta}} \mathbf{1}_{\{||\boldsymbol{g}(\boldsymbol{\theta}_0)-\boldsymbol{g}(\boldsymbol{\theta})||^2_{\boldsymbol{\Sigma}_\epsilon} \leq \frac{\ell(M)}{2}\}} p(\boldsymbol{\theta})d\boldsymbol{\theta}}\,.\end{aligned} \tag{6}$$

The inequality (6) holds given the set of measurement noise $\bar{\boldsymbol{\epsilon}} = \{\boldsymbol{\epsilon}_i\}$, $i = 1, ..., M$. Now we consider the set $\Omega_{sup} := \{\omega : \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} |\sum_{i=1}^{M}\boldsymbol{\epsilon}_i^T(\omega)\boldsymbol{\Sigma}_\epsilon^{-1}(\boldsymbol{g}(\boldsymbol{\theta}_0) - \boldsymbol{g}(\boldsymbol{\theta}))| > \frac{M}{16}\ell(M)\}$ so that, on its complement $\Omega\backslash\Omega_{sup}$ the random variable $\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} |\sum_{i=1}^{M}\boldsymbol{\epsilon}_i^T(\omega)\boldsymbol{\Sigma}_\epsilon^{-1}(\boldsymbol{g}(\boldsymbol{\theta}_0) - \boldsymbol{g}(\boldsymbol{\theta}))|$ is bounded above by $\frac{M\ell(M)}{16}$. Combining this result with (6), we have

$$\mathbb{P}(\boldsymbol{\theta} \in \boldsymbol{\Theta}\backslash\Omega_M(\boldsymbol{\theta}_0) \mid \bar{\boldsymbol{\epsilon}}) \leq \frac{e^{-\frac{M}{8}\ell(M)}}{\int_{\boldsymbol{\Theta}} \mathbf{1}_{\{||\boldsymbol{g}(\boldsymbol{\theta}_0)-\boldsymbol{g}(\boldsymbol{\theta})||^2_{\boldsymbol{\Sigma}_\epsilon} \leq \frac{\ell(M)}{2}\}} p(\boldsymbol{\theta})d\boldsymbol{\theta}}\,. \tag{7}$$

Using a partition of the space Ω with respect to the set Ω_{sup} we have, for any $c > 0$,

$$\mathbb{P}_\epsilon(\mathbb{P}(\boldsymbol{\theta} \in \boldsymbol{\Theta}\backslash\Omega_M(\boldsymbol{\theta}_0) \mid \bar{\boldsymbol{\epsilon}}) > c) \leq \mathbb{P}_\epsilon(\mathbb{P}(\boldsymbol{\theta} \in \boldsymbol{\Theta}\backslash\Omega_M(\boldsymbol{\theta}_0) \mid \bar{\boldsymbol{\epsilon}}) > c\,, \Omega\backslash\Omega_{sup}) + \mathbb{P}_\epsilon(\Omega_{sup})\,.$$

Based on the assumption that $\ell(M) = \ell_0 M^{-\beta}$ with $\frac{1}{2} > \beta > 0$, for any $c > 0$, applying the inequality (7), there exists $\tilde{M}(c) > 0$ which makes the right hand side of (7) smaller than c such that for any $M \geq \tilde{M}(c)$ we have

$$\mathbb{P}_\epsilon(\mathbb{P}(\boldsymbol{\theta} \in \boldsymbol{\Theta}\backslash\Omega_M(\boldsymbol{\theta}_0) \mid \bar{\boldsymbol{\epsilon}}) > c\,, \Omega\backslash\Omega_{sup}) = 0\,.$$

Finally, due to (5), we obtain

$$\begin{aligned}
\mathbb{P}_\epsilon(\Omega_{sup}) &\leq \mathbb{P}_\epsilon\left(\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \|\sum_{i=1}^M \boldsymbol{\epsilon}_i\|_{\boldsymbol{\Sigma}_\epsilon}\, \|\boldsymbol{g}(\boldsymbol{\theta}_0) - \boldsymbol{g}(\boldsymbol{\theta})\|_{\boldsymbol{\Sigma}_\epsilon} > \frac{M}{16}\ell(M)\right) \\
&\leq \mathbb{P}_\epsilon\left(\|\sum_{i=1}^M \boldsymbol{\epsilon}_i\|_{\boldsymbol{\Sigma}_\epsilon} > \frac{M}{16c_g}\ell(M)\right) \\
&= \mathbb{P}_\epsilon\left(\|\sum_{i=1}^M \boldsymbol{\epsilon}_i\|_{\boldsymbol{\Sigma}_\epsilon} - \mathbb{E}\|\sum_{i=1}^M \boldsymbol{\epsilon}_i\|_{\boldsymbol{\Sigma}_\epsilon} > \frac{M}{16c_g}\ell(M) - \mathbb{E}\|\sum_{i=1}^M \boldsymbol{\epsilon}_i\|_{\boldsymbol{\Sigma}_\epsilon}\right) \\
&= \mathbb{P}_\epsilon\left(\|\sum_{i=1}^M \boldsymbol{\epsilon}_i\|_{\boldsymbol{\Sigma}_\epsilon} - \mathbb{E}\|\sum_{i=1}^M \boldsymbol{\epsilon}_i\|_{\boldsymbol{\Sigma}_\epsilon} > \frac{M}{16c_g}\ell(M) - \sqrt{Ms}\right) \\
&\leq \frac{Var\left(\left\|\sum_{i=1}^M \boldsymbol{\epsilon}_i\right\|_{\boldsymbol{\Sigma}_\epsilon}\right)}{\left(\frac{M}{16c_g}\ell(M) - \sqrt{Ms}\right)^2} \leq \left(\frac{16c_g\sqrt{s}}{\sqrt{M}\ell(M) - 16c_g\sqrt{s}}\right)^2,
\end{aligned}$$

where $\mathbb{E}$ denotes the expectation, and Cauchy-Schwarz and Chebyshev's inequalities are used. Therefore, the lemma is proved. □

Note that the choice of $\ell(M)$ in Lemma 1 is not unique.

The K-L divergence can then be written as the sum of two terms, namely

$$D_{KL}(\bar{\boldsymbol{y}}) = \int_{\Omega_M(\boldsymbol{\theta}_0)} \log\left(\frac{p_{\boldsymbol{\Theta}}(\boldsymbol{\theta}|\bar{\boldsymbol{y}})}{p_{\boldsymbol{\Theta}}(\boldsymbol{\theta})}\right) p_{\boldsymbol{\Theta}}(\boldsymbol{\theta}|\bar{\boldsymbol{y}}) d\boldsymbol{\theta} + \epsilon_{\Omega_M}\,. \tag{8}$$

Additionally, we have the following remark on the posterior probability mass in $\boldsymbol{\Theta}\backslash\Omega_M$. The denominator of (7) is proportional to the Lebesgue measure of Ω_M up to a constant, i.e. $\int_{\boldsymbol{\Theta}} \mathbf{1}_{\{||\boldsymbol{g}(\boldsymbol{\theta}_0)-\boldsymbol{g}(\boldsymbol{\theta})||^2_{\boldsymbol{\Sigma}_\epsilon} \leq \frac{\ell(M)}{2}\}} p(\boldsymbol{\theta}) d\boldsymbol{\theta} = C\hat{g}(\ell(M))^{r/2}$, where $\hat{g}$ is a positive increasing function which maps the $\ell(M)$ to the square distance from θ to the manifold T. Consequently, we obtain

$$\begin{aligned}
\mathbb{P}(\boldsymbol{\theta} \in \boldsymbol{\Theta}\backslash\Omega_M(\boldsymbol{\theta}_0)|\bar{\boldsymbol{y}}) &= \frac{e^{-\ell_0 \frac{M^{1-\beta}}{8}}}{C\hat{g}(\ell_0 M^{-\beta})^{r/2}} \\
&= O_P\left(\hat{g}(\ell_0 M^{-\beta})^{-r/2} e^{-\ell_0 \frac{M^{1-\beta}}{8}}\right),
\end{aligned}$$

which goes exponentially to zero as $M \to \infty$.

Lemma 2. *The error caused by approximating the integration of K-L divergence by the integral over $\Omega_M(\boldsymbol{\theta}_0)$ in Equation (8) is*

$$\epsilon_{\Omega_M} = \int_{\boldsymbol{\Theta}\backslash\Omega_M(\boldsymbol{\theta}_0)} \log\left(\frac{p_{\boldsymbol{\Theta}}(\boldsymbol{\theta}|\bar{\boldsymbol{y}})}{p_{\boldsymbol{\Theta}}(\boldsymbol{\theta})}\right) p_{\boldsymbol{\Theta}}(\boldsymbol{\theta}|\bar{\boldsymbol{y}}) d\boldsymbol{\theta} = O_P\left(M\hat{g}(\ell_0 M^{-\beta})^{-r/2} e^{-\ell_0 \frac{M^{1-\beta}}{8}}\right) .$$

Proof
The proof is straightforward. We first rewrite ϵ_{Ω_M} using Bayes theorem as follows:

$$\epsilon_{\Omega_M} = \int_{\boldsymbol{\Theta}\backslash\Omega_M(\boldsymbol{\theta}_0)} \left[\log\left(p_{\mathcal{Y}}(\bar{\boldsymbol{y}}|\boldsymbol{\theta})\right) - \log\left(p_{\mathcal{Y}}(\bar{\boldsymbol{y}})\right)\right] p_{\boldsymbol{\Theta}}(\boldsymbol{\theta}|\bar{\boldsymbol{y}}) d\boldsymbol{\theta} .$$

From the first equation of the proof of lemma 1, we have already seen that the order of the log-likelihood is $O_P(M)$. Meanwhile, $p_{\mathcal{Y}}(\bar{\boldsymbol{y}}) = \int_{\boldsymbol{\Theta}} p_{\mathcal{Y}}(\bar{\boldsymbol{y}}|\boldsymbol{\theta}) p_{\boldsymbol{\Theta}}(\boldsymbol{\theta}) d\boldsymbol{\theta}$ is also $O_P(M)$ according to the mean value theorem. Therefore, $\log\left(p_{\mathcal{Y}}(\bar{\boldsymbol{y}}|\boldsymbol{\theta})\right) - \log\left(p_{\mathcal{Y}}(\bar{\boldsymbol{y}})\right) = O_P(M)$. Combining this with $\mathbb{P}(\boldsymbol{\theta} \in \boldsymbol{\Theta}\backslash\Omega_M(\boldsymbol{\theta}_0)|\bar{\boldsymbol{y}}) = O_P\left(\hat{g}(\ell_0 M^{-\beta})^{-r/2} e^{-\ell_0 \frac{M^{1-\beta}}{8}}\right)$ leads to the completion of the proof of Lemma 2. □

2.3. Local coordinates and weight functions

For the purpose of conciseness, we use T and Ω_M instead of $T(\boldsymbol{\theta}_0)$ and $\Omega_M(\boldsymbol{\theta}_0)$, respectively, in the remainder of this work. We define a new set of parameters, $\boldsymbol{t}$ and $\boldsymbol{s}$, for our estimation of D_{KL}: $\boldsymbol{t}$ parameterizes the manifold T and $\boldsymbol{s}$ parameterizes the direction orthogonal to T. Specifically, the direction orthogonal to T is defined as the orthonormal complement of the kernel of the Jacobian of our model $\boldsymbol{g}$, i.e., $\mathbf{Ker}(\boldsymbol{J}_g)^{\perp}$. We observe that the $\mathbf{Ker}(\boldsymbol{J}_g)$ contains the directions that are tangent to the manifold at $\boldsymbol{t}$ given $\boldsymbol{\theta} \in T$.

We define the following diffeomorphism mapping:

Definition 1.

$$\boldsymbol{f} : \Omega_{Ms,t} \to \Omega_M , \tag{9}$$

where $\Omega_{M\boldsymbol{s},\boldsymbol{t}}$ is the $(\boldsymbol{s},\boldsymbol{t})$ space, which is rectangular, ignoring possible boundary effects, i.e., $\Omega_{M\boldsymbol{s},\boldsymbol{t}} = [-O(\sqrt{\ell(M)}), O(\sqrt{\ell(M)})] \times T_{\boldsymbol{t}}$.

The tubular neighborhood theorem [17, p.346] guarantees the existence of $\Omega_{M\boldsymbol{s},\boldsymbol{t}}$. Here, $[-O(\sqrt{\ell(M)}), O(\sqrt{\ell(M)})]$ is the range for the parameter $\boldsymbol{s}$, and $T_{\boldsymbol{t}}$ is the set containing all the values of the parameter $\boldsymbol{t}$. Observe that all these objects depend on $\boldsymbol{\theta}_0$. For the purpose of conciseness, we do not write this dependence explicitly. Instead of $T_{\boldsymbol{t}}(\boldsymbol{\theta}_0)$, we write $T_{\boldsymbol{t}}$ in the remainder of this work. Figure 1 illustrates such an non-informative manifold, T, the orthogonal direction, S, and the subdomain, Ω_M.

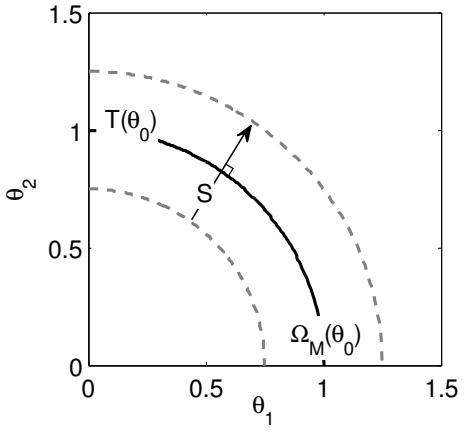


Figure 1: An illustrative non-informative manifold in two-dimensional parameter space.

Generally, we are not able to give an explicit parameterization of the manifold, T. Nevertheless, we can give the explicit form of the local coordinate, $\boldsymbol{s}$, as follows. We first define a cost function, F, given by

$$F(\boldsymbol{\theta}) := \frac{1}{2}\|(\boldsymbol{g}(\boldsymbol{\theta}) - \boldsymbol{g}(\boldsymbol{\theta}_0))\|^2_{\boldsymbol{\Sigma}_\epsilon} .$$

We subsequently perform the eigenvalue decomposition of the Hessian of $F(\boldsymbol{\theta})$ at $\boldsymbol{f}(\boldsymbol{0}, \boldsymbol{t})$ on the manifold (by the construction we have that $\boldsymbol{s} = \boldsymbol{0}$ on the manifold, T) as follows:

$$\boldsymbol{H}(\boldsymbol{f}(\boldsymbol{0}, \boldsymbol{t})) = [\boldsymbol{U}\, \boldsymbol{V}]\, \boldsymbol{\Lambda}\, [\boldsymbol{U}\, \boldsymbol{V}]^T . \tag{10}$$

Then we can locally define $\boldsymbol{s}$ at the vicinity of the point $\boldsymbol{f}(\boldsymbol{0}, \boldsymbol{t})$ as a vector of length r which equals the rank of the Hessian matrix, $\boldsymbol{H}(\boldsymbol{f}(\boldsymbol{0}, \boldsymbol{t}))$:

$$\boldsymbol{s} := \boldsymbol{U}^T(\boldsymbol{\theta} - \boldsymbol{f}(\boldsymbol{0}, \boldsymbol{t})) .$$

Meanwhile, $\boldsymbol{t}$ is a vector of length $d - r$. In this work, we assume that r does not change value w.r.t. $\boldsymbol{\theta}$.

Keeping in mind that we intend to carry out the Laplace approximation along the $\boldsymbol{s}$ direction, it is necessary to express the related pdfs, e.g., $p(\boldsymbol{\theta})$ and $p(\boldsymbol{\theta}|\bar{\boldsymbol{y}})$, in terms of the local coordinates $\boldsymbol{s}$ and $\boldsymbol{t}$. We consequently define two weight functions of $(\boldsymbol{s}, \boldsymbol{t})$ through a change of variables from the pdfs of $\boldsymbol{\theta}$:

Definition 2.

$$p(\boldsymbol{s}, \boldsymbol{t}) := p_{\boldsymbol{\Theta}}(\boldsymbol{f}(\boldsymbol{s}, \boldsymbol{t}))|\boldsymbol{J}| \tag{11}$$

$$p(\boldsymbol{s}, \boldsymbol{t}|\bar{\boldsymbol{y}}) := p_{\boldsymbol{\Theta}}(\boldsymbol{f}(\boldsymbol{s}, \boldsymbol{t})|\bar{\boldsymbol{y}})|\boldsymbol{J}| , \tag{12}$$

where $\boldsymbol{J}$ denotes the Jacobian of the diffeomorphism mapping $\boldsymbol{f}$ with respect to $(\boldsymbol{s},\boldsymbol{t})$.

Here, $p(\boldsymbol{s},\boldsymbol{t})$ and $p(\boldsymbol{s},\boldsymbol{t}|\bar{\boldsymbol{y}})$ play the role of the prior and posterior weight functions, respectively. Observe that (11) and (12) are essentially standard changes of variables for pdfs. However, the integrations of both weight functions over $\Omega_{M\boldsymbol{s},\boldsymbol{t}}$ do not equal to 1, i.e., $\mathbb{P}(\boldsymbol{\theta}\in\Omega_M)<1$ and $\mathbb{P}(\boldsymbol{\theta}\in\Omega_M|\bar{\boldsymbol{y}})<1$ for any finite M. Also note that both $p(\boldsymbol{s},\boldsymbol{t})$ and $p(\boldsymbol{s},\boldsymbol{t}|\bar{\boldsymbol{y}})$ depend on $\boldsymbol{\theta}_0$. Nevertheless, we note that $p(\boldsymbol{s},\boldsymbol{t}|\bar{\boldsymbol{y}})$ is asymptotically a pdf, since the posterior pdf $p(\boldsymbol{\theta}|\bar{\boldsymbol{y}})$ concentrates in Ω_M. In addition, since

$$p_{\boldsymbol{\Theta}}(\boldsymbol{\theta}|\bar{\boldsymbol{y}})=\frac{p_{\mathcal{Y}}(\bar{\boldsymbol{y}}|\boldsymbol{\theta})p_{\boldsymbol{\Theta}}(\boldsymbol{\theta})}{p_{\mathcal{Y}}(\bar{\boldsymbol{y}})}\quad\text{for}\quad\boldsymbol{\theta}\in\Omega_M\,,$$

we have

$$p(\boldsymbol{s},\boldsymbol{t}|\bar{\boldsymbol{y}})=\frac{p(\bar{\boldsymbol{y}}|\boldsymbol{s},\boldsymbol{t})p(\boldsymbol{s},\boldsymbol{t})}{p_{\mathcal{Y}}(\bar{\boldsymbol{y}})}\quad\text{for}\quad(\boldsymbol{s},\boldsymbol{t})\in\Omega_{M\boldsymbol{s},\boldsymbol{t}}\,. \tag{13}$$

We use these asymptotic relations in the following derivations.

Substituting (11) and (12) into (8), we obtain

$$\begin{aligned}D_{KL}(\bar{\boldsymbol{y}})&=\int_{T_{\boldsymbol{t}}}\int_{[-O(\sqrt{\ell(M)}),O(\sqrt{\ell(M)})]}\log\left(\frac{p(\boldsymbol{s},\boldsymbol{t}|\bar{\boldsymbol{y}})}{p(\boldsymbol{s},\boldsymbol{t})}\right)p(\boldsymbol{s},\boldsymbol{t}|\bar{\boldsymbol{y}})d\boldsymbol{s}d\boldsymbol{t}+\epsilon_{\Omega_M}\\&=\int_{T_{\boldsymbol{t}}}\int_{[-O(\sqrt{\ell(M)}),O(\sqrt{\ell(M)})]}\log\left(\frac{p(\boldsymbol{s},\boldsymbol{t}|\bar{\boldsymbol{y}})}{p(\boldsymbol{s},\boldsymbol{t})}\right)p(\boldsymbol{s}|\boldsymbol{t},\bar{\boldsymbol{y}})p(\boldsymbol{t}|\bar{\boldsymbol{y}})d\boldsymbol{s}d\boldsymbol{t}+\epsilon_{\Omega_M}\,.\end{aligned} \tag{14}$$

2.4. Laplace approximation for the conditional information gain

For a given $\boldsymbol{t}$, the conditional posterior weight function, $p(\boldsymbol{s}|\boldsymbol{t},\bar{\boldsymbol{y}})$, is expected to concentrate at $\hat{\boldsymbol{s}}$, which indicates the maximum likelihood estimator of the "true" parameter as M increases. We approximate $p(\boldsymbol{s}|\boldsymbol{t},\bar{\boldsymbol{y}})$, $p(\boldsymbol{s},\boldsymbol{t}|\bar{\boldsymbol{y}})$ and $p(\boldsymbol{s},\boldsymbol{t})$ by taking the exponential of the second order Taylor expansion of their corresponding logarithms at $\hat{\boldsymbol{s}}$, for a given value of $\boldsymbol{t}$. The following Gaussian posterior weight functions and the local expansion of $p(\boldsymbol{s},\boldsymbol{t})$ at $\hat{\boldsymbol{s}}$ can then be written as

$$\tilde{p}(\boldsymbol{s}|\boldsymbol{t},\bar{\boldsymbol{y}})=\frac{1}{(\sqrt{2\pi})^r|\boldsymbol{\Sigma}_{s|t}|^{1/2}}\exp\left[-\frac{(\boldsymbol{s}-\hat{\boldsymbol{s}})^T\boldsymbol{\Sigma}_{s|t}^{-1}(\boldsymbol{s}-\hat{\boldsymbol{s}})}{2}\right]\,, \tag{15}$$

$$\tilde{p}(\boldsymbol{s},\boldsymbol{t}|\bar{\boldsymbol{y}})=p(\hat{\boldsymbol{s}},\boldsymbol{t}|\bar{\boldsymbol{y}})\exp\left[-\frac{(\boldsymbol{s}-\hat{\boldsymbol{s}})^T\boldsymbol{\Sigma}_{s|t}^{-1}(\boldsymbol{s}-\hat{\boldsymbol{s}})}{2}\right]\,, \tag{16}$$

$$\tilde{p}(\boldsymbol{s},\boldsymbol{t})=p(\hat{\boldsymbol{s}},\boldsymbol{t})\exp\left[\nabla\log p(\hat{\boldsymbol{s}},\boldsymbol{t})(\boldsymbol{s}-\hat{\boldsymbol{s}})+\frac{(\boldsymbol{s}-\hat{\boldsymbol{s}})^T H_p(\hat{\boldsymbol{s}},\boldsymbol{t})(\boldsymbol{s}-\hat{\boldsymbol{s}})}{2}\right]\,. \tag{17}$$

In order to compute $\boldsymbol{\Sigma}_{s|t}$, we carry out the second order Taylor expansion of $\tilde{F}(\boldsymbol{\theta}):=-\log(p(\boldsymbol{\theta}|\bar{\boldsymbol{y}}))$ at $\boldsymbol{f}(\hat{\boldsymbol{s}},\boldsymbol{t})$ as follows:

$$\tilde{F}(\boldsymbol{f}(\hat{\boldsymbol{s}},\boldsymbol{t})+\boldsymbol{U}\boldsymbol{s})=\tilde{F}(\boldsymbol{f}(\hat{\boldsymbol{s}},\boldsymbol{t}))+\frac{(\boldsymbol{s}-\hat{\boldsymbol{s}})^T\boldsymbol{U}^T\tilde{\boldsymbol{H}}(\boldsymbol{f}(\hat{\boldsymbol{s}},\boldsymbol{t}))\boldsymbol{U}(\boldsymbol{s}-\hat{\boldsymbol{s}})}{2}+O(||\boldsymbol{s}-\hat{\boldsymbol{s}}||^3)\,,$$

where $\tilde{\boldsymbol{H}}(\boldsymbol{f}(\hat{\boldsymbol{s}},\boldsymbol{t}))$ is the Hessian matrix of $\tilde{F}(\boldsymbol{f}(\hat{\boldsymbol{s}},\boldsymbol{t}))$. Therefore, we obtain the following conditional covariance matrix after the change of variables:

$$\boldsymbol{\Sigma}_{s|t} = (\boldsymbol{U}^T(\tilde{\boldsymbol{H}}(\boldsymbol{f}(\hat{\boldsymbol{s}},\boldsymbol{t}))\boldsymbol{U})^{-1}\,. \tag{18}$$

Now, we have the following lemma regarding the approximation of the K-L divergence:

Lemma 3. *The information gain, D_{KL}, can be approximated by*

$$D_{KL} = \int_{T_{\boldsymbol{t}}} \int_{[-O(\sqrt{\ell(M)}),O(\sqrt{\ell(M)})]} \log\left(\frac{\tilde{p}(\boldsymbol{s},\boldsymbol{t}|\bar{\boldsymbol{y}})}{\tilde{p}(\boldsymbol{s},\boldsymbol{t})}\right)\tilde{p}(\boldsymbol{s}|\boldsymbol{t},\bar{\boldsymbol{y}})d\boldsymbol{s}p(\boldsymbol{t}|\bar{\boldsymbol{y}})d\boldsymbol{t} + \epsilon_{laplace} + \epsilon_{\Omega_M}\,, \tag{19}$$

where $\epsilon_{laplace} = O_P(\frac{1}{M^2})$ (its proof is given in Appendix A), and $\epsilon_{\Omega_M} = O_P\left(Me^{-\frac{M}{8}}\right)$, which decays at a rate faster than $O_P\left(\frac{1}{M^2}\right)$.

2.5. Laplace approximation for the expected information gain

We first introduce the following definition:

Definition 3.

$$D_{\boldsymbol{s}|\boldsymbol{t}} := \log\left[p(\hat{\boldsymbol{s}},\boldsymbol{t}|\bar{\boldsymbol{y}})\right] - \log\left[p(\hat{\boldsymbol{s}},\boldsymbol{t})\right] - \frac{r}{2}\,, \tag{20}$$

which is the Laplace approximation of the information gain in the $\boldsymbol{s}$ direction for a given value of $\boldsymbol{t}$. i.e.,

$$\int_{[-O(\sqrt{\ell(M)}),O(\sqrt{\ell(M)})]} \log\left(\frac{\tilde{p}(\boldsymbol{s},\boldsymbol{t}|\bar{\boldsymbol{y}})}{\tilde{p}(\boldsymbol{s},\boldsymbol{t})}\right)\tilde{p}(\boldsymbol{s}|\boldsymbol{t},\bar{\boldsymbol{y}})d\boldsymbol{s} = D_{\boldsymbol{s}|\boldsymbol{t}} + O_P\left(\frac{1}{M}\right)\,, \tag{21}$$

where the error term, $O_P\left(\frac{1}{M}\right)$, is dominated by $\frac{\mathbf{tr}(\boldsymbol{\Sigma}_{\boldsymbol{s}|\boldsymbol{t}}H_p(\hat{\boldsymbol{s}},\boldsymbol{t}))}{2}$.

The asymptotic form of the expected information gain is given below:

$$\begin{aligned} I &= \int_{\mathcal{Y}} D_{KL}\; p(\bar{\boldsymbol{y}})d\bar{\boldsymbol{y}} = \int_{\mathcal{Y}}\int_{T_{\boldsymbol{t}}} D_{\boldsymbol{s}|\boldsymbol{t}}p(\boldsymbol{t}|\bar{\boldsymbol{y}})d\boldsymbol{t}p(\bar{\boldsymbol{y}})d\bar{\boldsymbol{y}} + O\left(\frac{1}{M}\right) \\ &= \int_{\mathcal{Y}}\int_{T_t}\int_{[-O(\sqrt{\ell(M)}),O(\sqrt{\ell(M)})]} D_{\boldsymbol{s}|\boldsymbol{t}}p(\boldsymbol{t}|\bar{\boldsymbol{y}})p(\boldsymbol{s}|\boldsymbol{t},\bar{\boldsymbol{y}})d\boldsymbol{s}d\boldsymbol{t}p(\bar{\boldsymbol{y}})d\bar{\boldsymbol{y}} + O\left(\frac{1}{M}\right) \\ &= \int_{\mathcal{Y}}\int_{\Omega_{M\boldsymbol{s},\boldsymbol{t}}} D_{\boldsymbol{s}|\boldsymbol{t}}p(\boldsymbol{s},\boldsymbol{t}|\bar{\boldsymbol{y}})d\boldsymbol{s}d\boldsymbol{t}p(\bar{\boldsymbol{y}})d\bar{\boldsymbol{y}} + O\left(\frac{1}{M}\right)\,, \end{aligned}$$

where we have assumed that the error term in (21) is integrable w.r.t. the data $\bar{\boldsymbol{y}}$. Furthermore, we carry out a change of parameters such that

$$\begin{aligned} I &= \int_{\mathcal{Y}} \int_{\Omega_M} D_{\boldsymbol{s}|\boldsymbol{t}} p(\boldsymbol{\theta}_0|\bar{\boldsymbol{y}}) d\boldsymbol{\theta}_0 p(\bar{\boldsymbol{y}}) d\bar{\boldsymbol{y}} + O\left(\frac{1}{M}\right) \\ &= \int_{\mathcal{Y}} \int_{\boldsymbol{\Theta}} \mathbf{1}_{\Omega_M} D_{\boldsymbol{s}|\boldsymbol{t}} p(\boldsymbol{\theta}_0|\bar{\boldsymbol{y}}) d\boldsymbol{\theta}_0 p(\bar{\boldsymbol{y}}) d\bar{\boldsymbol{y}} + O\left(\frac{1}{M}\right) \\ &= \int_{\boldsymbol{\Theta}} \int_{\mathcal{Y}} \mathbf{1}_{\Omega_M} D_{\boldsymbol{s}|\boldsymbol{t}} p(\bar{\boldsymbol{y}}|\boldsymbol{\theta}_0) d\bar{\boldsymbol{y}} p(\boldsymbol{\theta}_0) d\boldsymbol{\theta}_0 + O\left(\frac{1}{M}\right), \end{aligned} \tag{22}$$

where the $\boldsymbol{t}$ in $D_{\boldsymbol{s}|\boldsymbol{t}}$ is implicitly given by $\boldsymbol{\theta}_0$.

Next, we rewrite the first two terms in $D_{\boldsymbol{s}|\boldsymbol{t}}$ using (13):

$$\begin{aligned} &\log\left[p(\hat{\boldsymbol{s}}, \boldsymbol{t}|\bar{\boldsymbol{y}})\right] - \log\left[p(\hat{\boldsymbol{s}}, \boldsymbol{t})\right] = \log\left[p(\bar{\boldsymbol{y}}|\hat{\boldsymbol{s}}, \boldsymbol{t})\right] - \log\left[p(\bar{\boldsymbol{y}})\right] \\ &\quad = \log\left[p(\bar{\boldsymbol{y}}|\hat{\boldsymbol{s}}, \boldsymbol{t})\right] - \log\left[\int_{T_{\boldsymbol{t}}} \int_{[-O(\sqrt{\ell(M)}), O(\sqrt{\ell(M)})]} p(\bar{\boldsymbol{y}}|\boldsymbol{s}, \boldsymbol{t}) p(\boldsymbol{s}, \boldsymbol{t}) d\boldsymbol{s} d\boldsymbol{t}\right] + O_P\left(\frac{1}{M}\right). \end{aligned} \tag{23}$$

Furthermore, the Laplace approximation for the above inner integration of $\boldsymbol{s}$ and the independence between $\boldsymbol{t}$ and $\bar{\boldsymbol{y}}$ given $\boldsymbol{s}$ (note that the tangent hyperplane to the manifold, T, is parallel to the kernel of the Jacobian of the model, i.e., $\mathbf{Ker}(\boldsymbol{J}_g)$) lead to

$$\begin{aligned} &-\log\left[\int_{T_{\boldsymbol{t}}} \int_{[-O(\sqrt{\ell(M)}), O(\sqrt{\ell(M)})]} p(\bar{\boldsymbol{y}}|\boldsymbol{s}, \boldsymbol{t}) p(\boldsymbol{s}, \boldsymbol{t}) d\boldsymbol{s} d\boldsymbol{t}\right] \\ &\quad = -\log\left[\int_{T_{\boldsymbol{t}}} p(\bar{\boldsymbol{y}}|\hat{\boldsymbol{s}}, \boldsymbol{t}) p(\hat{\boldsymbol{s}}, \boldsymbol{t}) (\sqrt{2\pi})^r |\boldsymbol{\Sigma}_{\boldsymbol{s}|\boldsymbol{t}}|^{1/2} d\boldsymbol{t}\right] + O_P\left(\frac{1}{M}\right) \\ &\quad = -\log\left[p(\bar{\boldsymbol{y}}|\hat{\boldsymbol{s}})\right] - \log\left[\int_{T_{\boldsymbol{t}}} p(\hat{\boldsymbol{s}}, \boldsymbol{t}) (\sqrt{2\pi})^r |\boldsymbol{\Sigma}_{\boldsymbol{s}|\boldsymbol{t}}|^{1/2} d\boldsymbol{t}\right] + O_P\left(\frac{1}{M}\right). \end{aligned}$$

Substituting this back into (23), we obtain

$$\log\left[p(\hat{\boldsymbol{s}}, \boldsymbol{t}|\bar{\boldsymbol{y}})\right] - \log\left[p(\hat{\boldsymbol{s}}, \boldsymbol{t})\right] = -\log\left[\int_{T_{\boldsymbol{t}}} p(\hat{\boldsymbol{s}}, \boldsymbol{t}) (\sqrt{2\pi})^r |\boldsymbol{\Sigma}_{\boldsymbol{s}|\boldsymbol{t}}|^{1/2} d\boldsymbol{t}\right] + O_p\left(\frac{1}{M}\right).$$

The new expression of $D_{\boldsymbol{s}|\boldsymbol{t}}$ is substituted back into (22) to provide the following approximation of the expected information gain.

Theorem 4. *Assume that $r < d$ and T_t is a $\mathcal{C}^2$ submanifold. Then, the expected information gain can be expressed as*

$$\begin{aligned} I = \int_{\boldsymbol{\Theta}} \int_{\mathcal{Y}} \mathbf{1}_{\Omega_M} &\left[-\log\left(\int_{T_{\boldsymbol{t}}} p(\hat{\boldsymbol{s}}, \boldsymbol{t}) |\boldsymbol{\Sigma}_{\boldsymbol{s}|\boldsymbol{t}}|^{1/2} d\boldsymbol{t}\right) - \frac{r}{2}\log(2\pi) - \frac{r}{2}\right] p(\bar{\boldsymbol{y}}|\boldsymbol{\theta}_0) p(\boldsymbol{\theta}_0) d\bar{\boldsymbol{y}} d\boldsymbol{\theta}_0 \\ &+ O\left(\frac{1}{M}\right), \end{aligned} \tag{24}$$

where the error, $O\left(\frac{1}{M}\right)$, is dominated by the standard Laplace approximation in the $\boldsymbol{s}$ direction.

2.6. Simplification of the integration over the manifold $T_{\boldsymbol{t}}$

In (24), there still exists an inner integral over the manifold, $T_{\boldsymbol{t}}$, and the outer integral is a non-trivial function of the inner one. Specifically, the outer one is over the space of $\boldsymbol{\theta}_0$, while the inner one is on the manifold, $T_{\boldsymbol{t}}$: $\int_{T_{\boldsymbol{t}}} p(\hat{\boldsymbol{s}}, \boldsymbol{t})|\boldsymbol{\Sigma}_{\boldsymbol{s}|\boldsymbol{t}}|^{1/2} d\boldsymbol{t}$. Therefore, as a whole, this double-integral cannot be treated computationally like a single loop in higher dimension. We now make some further simplifications of the manifold integral in (24). We first state the following lemma:

Lemma 5. *Assume that $r < d$ and T_t is a $\mathcal{C}^2$ submanifold. Then, the conditional variance matrix, $\boldsymbol{\Sigma}_{\boldsymbol{s}|\boldsymbol{t}}$, can be approximated asymptotically by a matrix, $\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{s}|\boldsymbol{t}}$, which is independent from $\boldsymbol{t}$, and the following expression holds*

$$\int_{T_{\boldsymbol{t}}} p(\hat{\boldsymbol{s}}, \boldsymbol{t})|\boldsymbol{\Sigma}_{\boldsymbol{s}|\boldsymbol{t}}|^{1/2} d\boldsymbol{t} = |\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{s}|\boldsymbol{t}}|^{1/2} \int_{T_{\boldsymbol{t}}} p(\hat{\boldsymbol{s}}, \boldsymbol{t}) d\boldsymbol{t} + O_P\left(\frac{1}{M\sqrt{M}}\right). \tag{25}$$

Proof

We know, from [1], that the Hessian of the negative log posterior $p(\boldsymbol{\theta}|\bar{\boldsymbol{y}})$ at $f(\hat{\boldsymbol{s}}, \boldsymbol{t})$ can be expressed as

$$\tilde{\boldsymbol{H}} = \boldsymbol{H}_g(\boldsymbol{f}(\hat{\boldsymbol{s}}, \boldsymbol{t}))\boldsymbol{\Sigma}_\epsilon^{-1}\boldsymbol{E}_s + M\boldsymbol{J}_g(\boldsymbol{f}(\hat{\boldsymbol{s}}, \boldsymbol{t}))^T\boldsymbol{\Sigma}_\epsilon^{-1}\boldsymbol{J}_g(\boldsymbol{f}(\hat{\boldsymbol{s}}, \boldsymbol{t})) - \boldsymbol{H}_p(\boldsymbol{f}(\hat{\boldsymbol{s}}, \boldsymbol{t}))\,, \tag{26}$$

where $\boldsymbol{H}_g(\boldsymbol{f}(\hat{\boldsymbol{s}}, \boldsymbol{t}))$ is a three-dimensional tensor, whose components can be denoted using Einstein notation by $H_{g_{i,k,l}} = \frac{\partial^2 g_i}{\partial\theta_k \partial\theta_l}$, with $i = 1, ..., d_{\boldsymbol{g}}$, and $k, l = 1, ..., d$. The (k, l) component of $\boldsymbol{H}_g\boldsymbol{\Sigma}_\epsilon^{-1}\boldsymbol{E}_s$ is defined by $H_{g_{i,k,l}}{\Sigma_\epsilon^{-1}}_{i,j}E_{sj}$, with $j = 1, ..., d_{\boldsymbol{g}}$.

Substituting (26) into (18) leads to

$$\begin{aligned}\boldsymbol{\Sigma}_{s|t} =& \frac{1}{M}\left\{\boldsymbol{U}^T\left[\boldsymbol{J}_g(\boldsymbol{f}(\hat{\boldsymbol{s}}, \boldsymbol{t}))^T\boldsymbol{\Sigma}_\epsilon^{-1}\boldsymbol{J}_g(\boldsymbol{f}(\hat{\boldsymbol{s}}, \boldsymbol{t}))\right]\boldsymbol{U}\right\}^{-1} + O_P\left(\frac{1}{M\sqrt{M}}\right) \\ =& \tilde{\boldsymbol{\Sigma}}_{s|t} + O_P\left(\frac{1}{M\sqrt{M}}\right).\end{aligned} \tag{27}$$

By construction, the model, $\boldsymbol{g}$, is independent of $\boldsymbol{t}$ for a given $\boldsymbol{s}$; therefore, $\boldsymbol{J}_g$ is independent from $\boldsymbol{t}$ given $\hat{\boldsymbol{s}}$. The approximate conditional covariance, $\tilde{\boldsymbol{\Sigma}}_{s|t}$, does not depend on $\boldsymbol{t}$. We establish the magnitude of the error in (27) using the Woodbury matrix identity [18].

Let the eigenvalue decomposition of $\boldsymbol{U}^T\left[\boldsymbol{H}_g(\boldsymbol{f}(\hat{\boldsymbol{s}}, \boldsymbol{t}))^T\boldsymbol{\Sigma}_\epsilon^{-1}\boldsymbol{E}_s - \nabla\nabla h_p(\boldsymbol{f}(\hat{\boldsymbol{s}}, \boldsymbol{t}))\right]\boldsymbol{U}$ be $\boldsymbol{RCL}$, where $\boldsymbol{R}$ is the column eigenvector matrix, $\boldsymbol{L}$ is the row eigenvector matrix, and $\boldsymbol{C}$ is the diagonal matrix containing eigenvalues. Let

$$\boldsymbol{A} = M\boldsymbol{U}^T\left[\boldsymbol{J}_g(\boldsymbol{f}(\hat{\boldsymbol{s}}, \boldsymbol{t}))^T\boldsymbol{\Sigma}_\epsilon^{-1}\boldsymbol{J}_g(\boldsymbol{f}(\hat{\boldsymbol{s}}, \boldsymbol{t}))\right]\boldsymbol{U}\,.$$

Thus, the conditional covariance matrix, $\boldsymbol{\Sigma}_{s|t}$, can be rewritten using the Woodbury matrix identity as follows

$$\boldsymbol{\Sigma}_{s|t} = (\boldsymbol{A} + \boldsymbol{RCL})^{-1} = \boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}\boldsymbol{R}(\boldsymbol{C}^{-1} + \boldsymbol{L}\boldsymbol{A}^{-1}\boldsymbol{R})^{-1}\boldsymbol{L}\boldsymbol{A}^{-1}\,. \tag{28}$$

Since the order of $\boldsymbol{C}$ is $O_P(\sqrt{M})$ and the order of $\boldsymbol{A}$ is $O_P(M)$, the error term $\boldsymbol{A}^{-1}\boldsymbol{R}(\boldsymbol{C}^{-1} + \boldsymbol{L}\boldsymbol{A}^{-1}\boldsymbol{R})^{-1}\boldsymbol{L}\boldsymbol{A}^{-1}$ is $O_P\left(\frac{1}{M\sqrt{M}}\right)$.

Now we can substitute (27) into the left hand side of (25) and pull $\tilde{\boldsymbol{\Sigma}}_{s|t}$ out of the integral. We consequently obtain the right hand side of (25). We complete the proof of Lemma 5. □

Combining Lemma 5 with Theorem 4 leads to the following theorem.

Theorem 6. *Assume that $r < d$ and T_t is a $\mathcal{C}^2$ submanifold. Then, the expected information gain can be expressed as*

$$I = \int_{\boldsymbol{\Theta}}\int_{\mathcal{Y}} \mathbf{1}_{\Omega_M}\left[-\log\left(\int_{T_{\boldsymbol{t}}} p(\hat{\boldsymbol{s}},\boldsymbol{t})d\boldsymbol{t}\right) - \frac{1}{2}\log|\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{s}|\boldsymbol{t}}| - \frac{r}{2}\log(2\pi) - \frac{r}{2}\right] p(\bar{\boldsymbol{y}}|\boldsymbol{\theta}_0)p(\boldsymbol{\theta}_0)d\bar{\boldsymbol{y}}d\boldsymbol{\theta}_0 + O\left(\frac{1}{M}\right)\,. \tag{29}$$

We can furthermore approximate the maximum posterior solution of $\boldsymbol{s}$ for a given value of $\boldsymbol{t}$, i.e., $\hat{\boldsymbol{s}}$, by $\mathbf{0}$ (remember that by construction $\boldsymbol{s} = \mathbf{0}$ on the manifold T). Theorem 6 can be simplified to the following Theorem 7.

Theorem 7. *Assume that $r < d$ and T_t is a $\mathcal{C}^2$ submanifold. Then, the expected information gain can be approximated by*

$$I = \int_{\boldsymbol{\Theta}}\left[-\log\left(\int_{T_{\boldsymbol{t}}} p(\mathbf{0},\boldsymbol{t})d\boldsymbol{t}\right) - \frac{1}{2}\log|\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{s}|\boldsymbol{t}}|\right] p(\boldsymbol{\theta}_0)d\boldsymbol{\theta}_0 - \frac{r}{2}\log(2\pi) - \frac{r}{2} + O\left(\frac{1}{M}\right)\,. \tag{30}$$

The error introduced in this approximation can be analyzed using the Taylor expansion around $\boldsymbol{s} = \mathbf{0}$ (see Appendix B for the proof).

2.7. Simplification of the marginal prior

In addition, it is in general difficult to compute the marginal pdf, $p_{\boldsymbol{s}}(\mathbf{0}) := \int_{T_{\boldsymbol{t}}} p(\mathbf{0},\boldsymbol{t})d\boldsymbol{t}$. Without losing generality, we carry out linearization of the manifold only to compute this marginal prior. In order to include most of the probability mass in this approximation, we first linearize the manifold at $\boldsymbol{\theta}^*(\boldsymbol{\theta}_0) := \arg\max_{\boldsymbol{\theta}\in T}\{p(\boldsymbol{\theta})\}$ in the case of a single modal prior or at the modes of the set $\{\boldsymbol{\theta}^*(\boldsymbol{\theta}_0)\}$, which contains all the local optimal points on T in the case of a multimodal prior.

After a linear transformation of variables, we can obtain the approximated marginal $\tilde{p}_{\boldsymbol{s}}(\mathbf{0})$ for many priors. For instance, we can obtain $\tilde{p}_{\boldsymbol{s}}(\mathbf{0})$ easily for a Gaussian or Gaussian mixture prior.

By simplifying the marginal prior, we introduce an error of order $O(1)$, which remains admissible since the expected information gain, I, is $O\left(log(M)\right)$. Furthermore, our numerical examples indicate that the error term is very small, more so than the order of the error might suggest. Up to now, we have a further simplified estimation of the expected information gain:

Theorem 8. *Assuming that $p(\boldsymbol{\theta})$ is bounded from above, we can use an approximated marginal prior such that*

$$I = \int_{\boldsymbol{\Theta}} \left[-\log\left[\tilde{p}_{\boldsymbol{s}}(\mathbf{0})\right] - \frac{1}{2}\log|\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{s}|\boldsymbol{t}}| - \frac{r}{2}\log(2\pi) - \frac{r}{2} \right] p(\boldsymbol{\theta}_0) d\boldsymbol{\theta}_0 + O\left(1\right) , \tag{31}$$

with $\tilde{p}_{\boldsymbol{s}}(\mathbf{0})$ computed by linearizing the manifold.

We emphasize that this error occurs only if the prior marginal is approximated via linearization, which is optional but brings substantial computational advantages at the cost of some small level of approximation error.

3. Estimation of the expected information gain for a quantity of interest

In [1], we approximated the posterior pdf of a scalar quantity of interest by a Gaussian distribution of a model that is completely determined by the data. We now focus on the prediction of a physical quantity of interest for an under-determined model. The scalar quantity of interest is commonly defined as a function of $\boldsymbol{\theta}$ plus some independent error, i.e.,

$$Q = \tau(\boldsymbol{\theta}) + \epsilon_Q \,,$$

where ϵ_Q is assumed to be independent of $\boldsymbol{\theta}$. We can reparameterize τ using $(\boldsymbol{s},\, \boldsymbol{t})$ defined in Section 2.3. i.e.,

$$Q = \tau(\ f(\boldsymbol{s},\, \boldsymbol{t})) + \epsilon_Q = \hat{\tau}(\boldsymbol{s},\, \boldsymbol{t}) + \epsilon_Q \quad \text{for } (\boldsymbol{s}, \boldsymbol{t}) \in \Omega_{M\boldsymbol{s},\boldsymbol{t}} \,.$$

Given a fixed value of $\boldsymbol{t}$, a Taylor expansion of τ at $(\hat{\boldsymbol{s}},\, \boldsymbol{t})$ along $\boldsymbol{s}$ leads to

$$\hat{\tau}(\boldsymbol{s},\, \boldsymbol{t}) = \hat{\tau}(\hat{\boldsymbol{s}},\, \boldsymbol{t}) + (\nabla_{\boldsymbol{s}}\hat{\tau}(\hat{\boldsymbol{s}},\, \boldsymbol{t}))(\boldsymbol{s} - \hat{\boldsymbol{s}}) + O_P(||\boldsymbol{s} - \hat{\boldsymbol{s}}||^2) \,,$$

where $\hat{\boldsymbol{s}}$ was defined in the previous section as the maximum posterior solution of $\boldsymbol{s}$ given $\boldsymbol{t}$. Since the conditional posterior pdf $p(\boldsymbol{s}|\boldsymbol{t}, \bar{\boldsymbol{y}})$ can be approximated by a Gaussian concentrated around $(\hat{\boldsymbol{s}},\, \boldsymbol{t})$ as discussed in the previous section, we can apply a small noise approximation to propagate randomness from $\boldsymbol{s}$ to the quantity of interest, Q. The resulting approximated distribution of Q, given $\boldsymbol{t}$ and $\bar{\boldsymbol{y}}$, is also Gaussian:

$$p(Q|\boldsymbol{t},\, \bar{\boldsymbol{y}}) = \frac{1}{\sqrt{2\pi}\sigma_{Q|\boldsymbol{t},\,\bar{\boldsymbol{y}}}} \exp\left[-\frac{(Q - \hat{\tau}(\hat{\boldsymbol{s}},\, \boldsymbol{t}))^2}{2\sigma^2_{Q|\boldsymbol{t},\,\bar{\boldsymbol{y}}}} \right] + O_P\left(\frac{1}{M^2}\right) = \hat{p}(Q|\boldsymbol{t},\, \bar{\boldsymbol{y}}) + O_P\left(\frac{1}{M^2}\right) , \tag{32}$$

where

$$\sigma^2_{Q|\boldsymbol{t},\,\bar{\boldsymbol{y}}} = (\nabla_{\boldsymbol{s}}\hat{\tau})^T \boldsymbol{\Sigma}_{\boldsymbol{s}|\boldsymbol{t}} \nabla_{\boldsymbol{s}}\hat{\tau} + \sigma^2_{\epsilon_Q}\,.$$

Here, $\sigma^2_{\epsilon_Q}$ denotes the variance of ϵ_Q, which is assumed to be a known constant. The derivation of the error rate in (32) is presented in Appendix C.

We now state a lemma:

Lemma 9. *Assume that $r < d$ and T_t is a $\mathcal{C}^2$ submanifold. Then, the posterior distribution of Q can be asymptotically approximated by*

$$p(Q|\bar{\boldsymbol{y}}) = \hat{p}(Q|\bar{\boldsymbol{y}}) + O_P\left(\frac{1}{M}\right)\,, \tag{33}$$

with

$$\hat{p}(Q|\bar{\boldsymbol{y}}) \propto \int_{T_{\boldsymbol{t}}} \int_{[-O(\sqrt{\ell(M)}),O(\sqrt{\ell(M)})]} \mathbf{1}_{\{\boldsymbol{s}=\hat{\boldsymbol{s}}\}} \hat{p}(Q|\boldsymbol{t},\,\bar{\boldsymbol{y}}) p(\boldsymbol{s},\boldsymbol{t}) d\boldsymbol{s} d\boldsymbol{t} + O_P\left(\frac{1}{M\sqrt{M}}\right)\,. \tag{34}$$

Proof

We first write the posterior pdf of Q as a marginalization over $\boldsymbol{t}$:

$$p(Q|\bar{\boldsymbol{y}}) = \int_{T_{\boldsymbol{t}}} \hat{p}(Q|\boldsymbol{t},\,\bar{\boldsymbol{y}}) p(\boldsymbol{t}|\bar{\boldsymbol{y}}) d\boldsymbol{t} + O_P\left(\frac{1}{M^2}\right) = \hat{p}(Q|\bar{\boldsymbol{y}}) + O_P\left(\frac{1}{M^2}\right)\,. \tag{35}$$

We can furthermore write $p(\boldsymbol{t}|\bar{\boldsymbol{y}})$ as a marginal over $\boldsymbol{s}$:

$$p(\boldsymbol{t}|\bar{\boldsymbol{y}}) = \int_{[-O(\sqrt{\ell(M)}),O(\sqrt{\ell(M)})]} p(\boldsymbol{s},\boldsymbol{t}|\bar{\boldsymbol{y}}) d\boldsymbol{s} + O_P\left(e^{-\frac{M}{8}}\right)\,.$$

Note that $p(\boldsymbol{s},\boldsymbol{t}|\bar{\boldsymbol{y}})$ should concentrate at $\boldsymbol{s} = \hat{\boldsymbol{s}}$ in $\Omega_{M\,\boldsymbol{s},\boldsymbol{t}}$ as the amount of data increases. This leads to the following Laplace approximation of $p(\boldsymbol{t}|\bar{\boldsymbol{y}})$:

$$p(\boldsymbol{t}|\bar{\boldsymbol{y}}) = p(\hat{\boldsymbol{s}},\boldsymbol{t}|\bar{\boldsymbol{y}})(\sqrt{2\pi})^r |\boldsymbol{\Sigma}_{\boldsymbol{s}|\boldsymbol{t}}|^{1/2} + O_P\left(\frac{1}{M}\right)\,. \tag{36}$$

Substituting (36) back into (35), we obtain

$$\begin{aligned} p(Q|\bar{\boldsymbol{y}}) =& \int_{T_{\boldsymbol{t}}} \hat{p}(Q|\boldsymbol{t},\,\bar{\boldsymbol{y}}) p(\hat{\boldsymbol{s}},\boldsymbol{t}|\bar{\boldsymbol{y}})(\sqrt{2\pi})^r |\boldsymbol{\Sigma}_{\boldsymbol{s}|\boldsymbol{t}}|^{1/2} d\boldsymbol{t} + O_P\left(\frac{1}{M}\right) \\ =& \hat{p}(Q|\bar{\boldsymbol{y}}) + O_P\left(\frac{1}{M}\right)\,. \end{aligned} \tag{37}$$

Using Equation (13), we rewrite $\hat{p}(Q|\bar{\boldsymbol{y}})$ up to a scaling factor (the evidence term):

$$\hat{p}(Q|\bar{\boldsymbol{y}}) \propto \int_{T_{\boldsymbol{t}}} \hat{p}(Q|\boldsymbol{t},\,\bar{\boldsymbol{y}})(\sqrt{2\pi})^r |\boldsymbol{\Sigma}_{\boldsymbol{s}|\boldsymbol{t}}|^{1/2} p(\bar{\boldsymbol{y}}|\hat{\boldsymbol{s}},\boldsymbol{t}) p(\hat{\boldsymbol{s}},\boldsymbol{t}) d\boldsymbol{t}$$

By construction, the data are independent of $\boldsymbol{t}$ given $\hat{\boldsymbol{s}}$. Therefore, we have $p(\bar{\boldsymbol{y}}|\hat{\boldsymbol{s}}, \boldsymbol{t}) = p(\bar{\boldsymbol{y}}|\hat{\boldsymbol{s}})$ and it is a constant with respect to the quantity of interest, Q. The above proportionality can be simplified to

$$\begin{aligned}\hat{p}(Q|\bar{\boldsymbol{y}}) &\propto \int_{T_{\boldsymbol{t}}} \hat{p}(Q|\boldsymbol{t},\, \bar{\boldsymbol{y}})|\boldsymbol{\Sigma}_{\boldsymbol{s}|\boldsymbol{t}}|^{1/2} p(\hat{\boldsymbol{s}}, \boldsymbol{t}) d\boldsymbol{t} = \int_{T_{\boldsymbol{t}}} \hat{p}(Q|\boldsymbol{t},\, \bar{\boldsymbol{y}}) p(\hat{\boldsymbol{s}}, \boldsymbol{t}) d\boldsymbol{t} + O_P\left(\frac{1}{M\sqrt{M}}\right) \\ &= \int_{T_{\boldsymbol{t}}} \int_{[-O(\sqrt{\ell(M)}), O(\sqrt{\ell(M)})]} \mathbf{1}_{\{\boldsymbol{s}=\hat{\boldsymbol{s}}\}} \hat{p}(Q|\boldsymbol{t},\, \bar{\boldsymbol{y}}) p(\boldsymbol{s}, \boldsymbol{t}) d\boldsymbol{s} d\boldsymbol{t} + O_P\left(\frac{1}{M\sqrt{M}}\right), \end{aligned} \tag{38}$$

where the equality in the first line is in the same spirit as Lemma 5. The proof of Lemma 9 is now completed. □

Since Q is a scalar, we can obtain the scaling factor for $\hat{p}(Q|\bar{\boldsymbol{y}})$ once (38) is computed using a one-dimensional grid of Q (see Section 4). We can next compute the expected conditional entropy by

$$H(Q|\bar{\boldsymbol{y}}) = \int_{\boldsymbol{\Theta}} \int_{\mathcal{Y}} \int_{Q} \log\left[\hat{p}(Q|\bar{\boldsymbol{y}})\right] \hat{p}(Q|\bar{\boldsymbol{y}}) dQ p(\bar{\boldsymbol{y}}|\boldsymbol{\theta}_0) d\bar{\boldsymbol{y}} p(\boldsymbol{\theta}_0) d\boldsymbol{\theta}_0 + O\left(\frac{1}{M}\right). \tag{39}$$

If we replace $\hat{\boldsymbol{s}}$ by $\mathbf{0}$, $\hat{p}(Q|\boldsymbol{t},\, \bar{\boldsymbol{y}})$ can be approximated using $\hat{p}(Q|\boldsymbol{t}, \boldsymbol{\theta}_0)$, which is given by

$$\hat{p}(Q|\boldsymbol{t}, \boldsymbol{\theta}_0) = \frac{1}{\sqrt{2\pi}\sigma_{Q|\boldsymbol{t},\boldsymbol{\theta}_0}} \exp\left[-\frac{(Q - \hat{\tau}(\mathbf{0}, \boldsymbol{t}))^2}{2\sigma^2_{Q|\boldsymbol{t},\boldsymbol{\theta}_0}}\right] = \hat{p}(Q|\boldsymbol{t},\, \bar{\boldsymbol{y}}) + O_P\left(\frac{1}{\sqrt{M}}\right),$$

with

$$\sigma^2_{Q|\boldsymbol{t},\boldsymbol{\theta}_0} = (\nabla_{\boldsymbol{s}} \hat{\tau})^T \boldsymbol{\Sigma}_{\boldsymbol{s}|\boldsymbol{t}}(\mathbf{0}) \nabla_{\boldsymbol{s}} \hat{\tau} + \sigma^2_{\epsilon_Q}.$$

Consequently, we have

$$\hat{p}(Q|\ \boldsymbol{\theta}_0) \propto \int_{T_{\boldsymbol{t}}} \int_{[-O(\sqrt{\ell(M)}), O(\sqrt{\ell(M)})]} \mathbf{1}_{\{\boldsymbol{s}=\mathbf{0}\}} \hat{p}(Q|\boldsymbol{t},\, \boldsymbol{\theta}_0) |\boldsymbol{\Sigma}_{\boldsymbol{s}|\boldsymbol{t}}|^{1/2} p(\boldsymbol{s}, \boldsymbol{t}) d\boldsymbol{s} d\boldsymbol{t}. \tag{40}$$

Replacing $\hat{p}(Q|\bar{\boldsymbol{y}})$ in (39) by $\hat{p}(Q|\boldsymbol{\theta}_0)$ and integrating out the $O_P(\frac{1}{\sqrt{M}})$ term over the domain $\mathcal{Y}$ (this term has mean zero; see Appendix B for details), we obtain the following asymptotic result for the expected conditional entropy:

Theorem 10.

$$H(Q|\bar{\boldsymbol{y}}) = \int_{\boldsymbol{\Theta}} \int_{Q} \log\left[\hat{p}(Q|\boldsymbol{\theta}_0)\right] \hat{p}(Q|\boldsymbol{\theta}_0) dQ p(\boldsymbol{\theta}_0) d\boldsymbol{\theta}_0 + O\left(\frac{1}{M}\right). \tag{41}$$

The rates of the dominant errors are derived in Appendix C. We observe that we have a double loop integral for the computation of $H(Q|\bar{\boldsymbol{y}})$. We show the details of the numerical

computation in the next section. Eventually the expected information gain for the quantity of interest, Q, can be approximated by

$$I = H(Q) - H(Q|\bar{\boldsymbol{y}})\,,$$

where $H(Q) = -\int_Q \log\left[p(Q)\right] p(Q) dQ$ is the prior entropy of Q. Observe that we do not need to compute $H(Q)$ in order to select the best experimental setup, since $H(Q)$ does not depend on the experimental setup, $\boldsymbol{\xi}$. A more detailed description of $H(Q)$ and its computation can be found in [1]. Thus, we will focus on the numerical computation of (41) in Section 4.

4. Numerical integration of the asymptotic forms

In most practical scenarios, we need to compute the expected information gain (31) numerically for parameter inferences and the expected conditional entropy (39) for predictions of quantities of interest.

We can approximate the integral in (31) using numerical quadratures or Monte Carlo sampling depending on the regularity of the integrand. In the case of using quadratures, the numerical integration adopts the following form:

$$I_{QU} = \sum_{i=1}^{NQ} \left[-\log\left[\tilde{p}_{\boldsymbol{s}}(\mathbf{0})\right] - \frac{1}{2}\log\left(|\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{s}|\boldsymbol{t}}|\right)\right] w_i - \frac{r}{2} - \frac{r}{2}\log\left(2\pi\right)\,, \tag{42}$$

where both $\tilde{p}_{\boldsymbol{s}}(\mathbf{0})$ and $\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{s}|\boldsymbol{t}}$ are computed using the i^{th} quadrature point, $\boldsymbol{\theta}_{0i}$, in the parameter domain. When the integrand fulfills certain regularity properties and the dimension is high, we can adopt sparse quadrature abscissas and weights for the numerical integration. A review of sparse grids can be found, for instance, in [1, 19–21]. On the other hand, if there is a lack of regularity either in the marginal prior $p_{\boldsymbol{s}}(\mathbf{0})$ or in the conditional covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{s}|\boldsymbol{t}}$, we can use Monte Carlo sampling:

$$I_{MC} = \frac{1}{NS}\sum_{j=1}^{NS} \left[-\log\left[\tilde{p}_{\boldsymbol{s}}(\mathbf{0})\right] - \frac{1}{2}\log\left(|\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{s}|\boldsymbol{t}}|\right)\right] - \frac{r}{2} - \frac{r}{2}\log(2\pi)\,, \tag{43}$$

where NS denotes the number of samples. Both $p_{\boldsymbol{s}}(\mathbf{0})$ and $\boldsymbol{\Sigma}_{\boldsymbol{s}|\boldsymbol{t}}$ are computed using the j^{th} random sample of the "true" parameter, $\boldsymbol{\theta}_{0j}$.

Due to the presence of the indicator function in the integral of $\hat{p}(Q|\boldsymbol{\theta}_0)$ in (40), we use Monte Carlo sampling for its numerical estimation. For the sake of convenience, we also adopt Monte Carlo sampling for the numerical estimation of the expected conditional entropy of the quantity of interest (41), so that we can reuse the same set of samples in both integrals. Specifically, we use the sample average w.r.t. the prior of $\boldsymbol{\theta}$ and one-dimensional binning for Q as follows:

$$H(Q|\bar{\boldsymbol{y}})_{MC} = \frac{1}{NS}\sum_{j=1}^{NS}\sum_{k=1}^{NS1} \log\left[\hat{p}(Q_k|\boldsymbol{\theta}_{0j})\right] \hat{p}(Q_k|\boldsymbol{\theta}_{0j})\Delta_k\,, \tag{44}$$

with

$$\hat{p}(Q_k|\boldsymbol{\theta}_{0j}) \propto \frac{1}{NS}\sum_{l=1}^{NS} \mathbf{1}_{\Omega}\hat{p}(Q_k|\boldsymbol{t}_l,\, \boldsymbol{\theta}_{0j})\,. \tag{45}$$

We define the domain of Q as $[\min\tau(\boldsymbol{\theta}), \max\tau(\boldsymbol{\theta})]$, where $\boldsymbol{\theta}$ takes a value from the NS samples drawn from the prior of $\boldsymbol{\theta}$. We use the same collection of samples for $\boldsymbol{\theta}$ in both (44) and (45). $\boldsymbol{t}_l$ is the local coordinate corresponding to the l^{th} sample of $\boldsymbol{\theta}$. Ideally, this sample is supposed to be on the manifold, T. In practice, we approximately consider all the samples in Ω as those on the manifold, T. Ω, which depends on $\boldsymbol{\theta}_0$, is defined by

$$\Omega(\boldsymbol{\theta}_0) = \{\boldsymbol{\theta} \in \boldsymbol{R}^d \,:\, \|g(\boldsymbol{\theta}) - g(\boldsymbol{\theta}_0)\|_{\boldsymbol{\Sigma_\epsilon}} \leq C\}\,, \tag{46}$$

where C is a small constant. We used $C = 10^{-3}$ in our computations.

5. Numerical Examples

In this section, we show the efficiency and capability of our method using two numerical examples. The first one is based on a simple nonlinear regression model with two unknown parameters and a scalar output. The second example concerns the design of the current patterns in an impedance tomography problem. The problem has 16 unknown parameters and the observations are two voltage measurements related to the solution of the Poisson equation.

5.1. Example 5.1: model with two unidentifiable parameters

We apply our Laplace method to the second example in [1]. The measurement y reads

$$y = (\alpha\theta_1 + \beta\theta_2)^3\xi^2 + (\alpha\theta_1 + \beta\theta_2)\exp[-|0.2 - \xi|] + \epsilon\,.$$

It is a single output experiment with a model having two parameters. The curve of I with respect to ξ reaches a local maximum at $\xi = 0.2$ and a global maximum at $\xi = 1$, similar to the curve of the expected information gain in the single-parameter example [1]. The model is not sensitive to a change in the parameters along the direction of $(\beta,\, -\alpha)$, where α and β are two given constants. The measurement noise is assumed to be Gaussian, i.e., $\epsilon \sim \mathcal{N}(0, \sigma_m^2)$. We firstly assume a uniform prior for the parameters $\boldsymbol{\theta} = [\theta_1,\, \theta_2]^T$, i.e.,

$$\boldsymbol{\theta} \sim \mathcal{U}(\boldsymbol{\theta}_l,\, \boldsymbol{\theta}_u)\,, \quad \text{with} \quad \boldsymbol{\theta}_l = [0, 0]^T \quad \text{and} \quad \boldsymbol{\theta}_u = [1, 1]^T\,.$$

Note that the method in [1] is not applicable here since there is no single mode in the posterior due to the non-informative prior. The Jacobian of this model with respect to $\boldsymbol{\theta}$ is

$$\boldsymbol{J} = [3\alpha(\alpha\theta_1 + \beta\theta_2)^2\xi^2 + \alpha\exp(-|0.2 - \xi|), \quad 3\beta(\alpha\theta_1 + \beta\theta_2)^2\xi^2 + \beta\exp(-|0.2 - \xi|)]\,.$$

Consider the particular case of $\alpha = 1$ and $\beta = 1$. We know that the linear manifold is defined by $\boldsymbol{V} = \left[\frac{\sqrt{2}}{2}, -\frac{\sqrt{2}}{2}\right]^T$, and its orthogonal direction is $\boldsymbol{U} = \left[\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}\right]^T$. According to our theory, the following transformation is carried out:

$$s = \frac{\sqrt{2}}{2}(\theta_1 - \theta_{10}) + \frac{\sqrt{2}}{2}(\theta_2 - \theta_{20}),$$
$$t = \frac{\sqrt{2}}{2}(\theta_1 - \theta_{10}) - \frac{\sqrt{2}}{2}(\theta_2 - \theta_{20}).$$

We can easily obtain $p_s(0)$ as

$$p_s(0) = \begin{cases} \sqrt{2}(\theta_{10} + \theta_{20}), & 0 < \theta_{10} + \theta_{20} \leq 1, \\ \sqrt{2}(2 - \theta_{10} - \theta_{20}), & 1 < \theta_{10} + \theta_{20} \leq 2. \end{cases}$$

In Figure 2, we compare the information gains computed using our Laplace approximation with sparse grid (LA + SG) numerical integration (Gauss-Legendre), our Laplace approximation with Monte Carlo (LA + MC) sampling, and double-loop Monte Carlo (DLMC) in the scenario where $M = 10$ and $\xi = 0.3$. The Laplace approximations show no bias and converge faster than the DLMC. The DLMC requires at least 10^2 times the number of samples to reach the same precision as the Laplace approximation. Note that the likelihood evaluation, in this particular case, dominates the CPU time of the DLMC. Therefore, the CPU time spent on the estimation is actually proportional to the square of number of samples, when the number of samples in the outer loop equals to the number of samples in the inner loop. In this sense, our Laplace approximation method is at least 10^4 times faster than the DLMC.

Due to the lack of smoothness of the integrand function as it is visualized in Figure 3, the convergence rate of LA + SG is similar to LA + MC. Note that there is a "kink" along the diagonal connecting $(1, 0)$ and $(0, 1)$. This makes numerical integration difficult in principle.

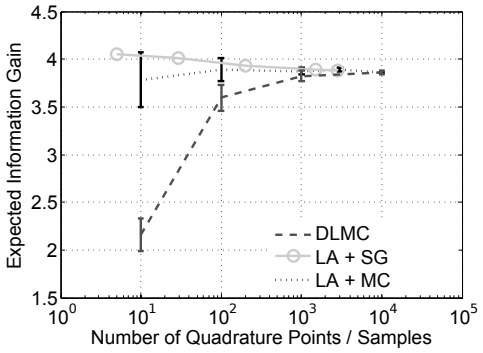


Figure 2: The convergence of the expected information gain computed using LA + MC, LA + SG, and DLMC in Example 5.1, with a uniform prior for $\boldsymbol{\theta}$ and $\xi = 0.3$. The same set of samples was used for both the inner and outer loops in the DLMC. The number of samples of DLMC is associated with this set of samples.

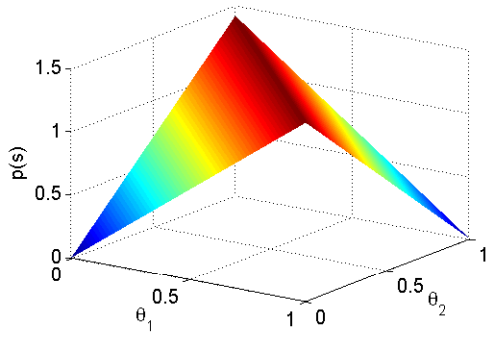


Figure 3: The tent function, $p(\boldsymbol{s})$, in Example 5.1.

5.1.1. Mixture Gaussian prior

To evaluate the robustness of (24) further, we set the prior as a mixture Gaussian, which adopts the following form:

$$p(\boldsymbol{\theta}) = 0.5 \times p_1(\boldsymbol{\theta}) + 0.5 \times p_2(\boldsymbol{\theta})\,, \tag{47}$$

where $p_1(\boldsymbol{\theta})$ and $p_2(\boldsymbol{\theta})$ are the pdfs of two multivariate Gaussians with mean vectors $[2,0]^T$ and $[0,2]^T$, respectively, and covariance matrix

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

The salient features of this prior are the two separated modes; see Figures 4(a) and 4(b) for the visualization. We compare the results obtained using LA + SG and LA + MC, and DLMC. Figure 5(a) shows the performances of the three methods in terms of the number of quadrature points (LA + SG) or sample points (LA + MC and DLMC), when $\xi = 0.3$. LA + SA and LA + MC are significantly faster than the DLMC. Similarly, Figure 5(b) shows how these methods converge to the true value of the expected information gain, $I(\xi)$, when $\xi = 1$. We used an auxiliary Gaussian pdf as a change of measure, with mean vector $[2,0]^T$ and covariance matrix

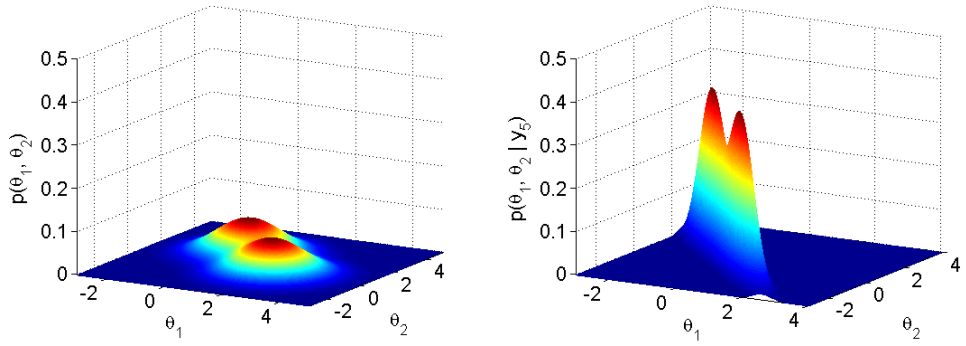
$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix},$$

in the sparse grid numerical integration.

5.1.2. Mixture Log Gaussian prior

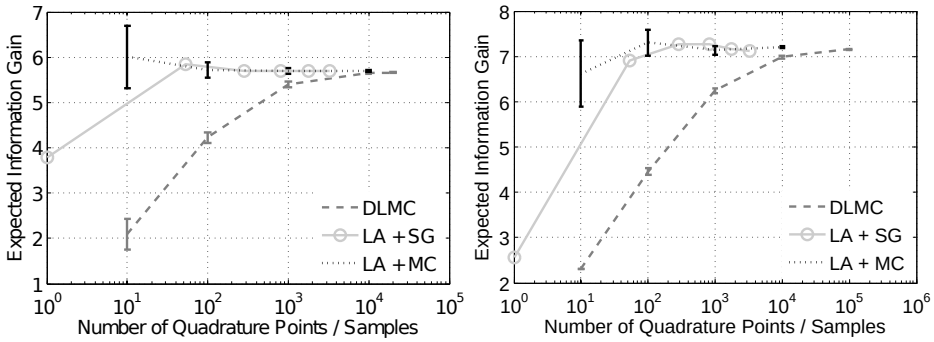
We define a new set of parameters, $\boldsymbol{\gamma} = \log \boldsymbol{\theta}$. We assume that $\boldsymbol{\gamma}$ is distributed as the mixture of two Gaussian pdfs as in Section 5.1.1. The non-informative manifold of $\boldsymbol{\gamma}$ is not a straight line anymore (see Figure 6 for the visualization of two posterior pdfs).

As seen in Figure 7, the convergence of DLMC is substantially slower than our approaches in this case (at least 10^5 times slower in terms of the number of samples) due to the change of variable. Additionally, the integral (24) against the mixture Gaussian is split into two



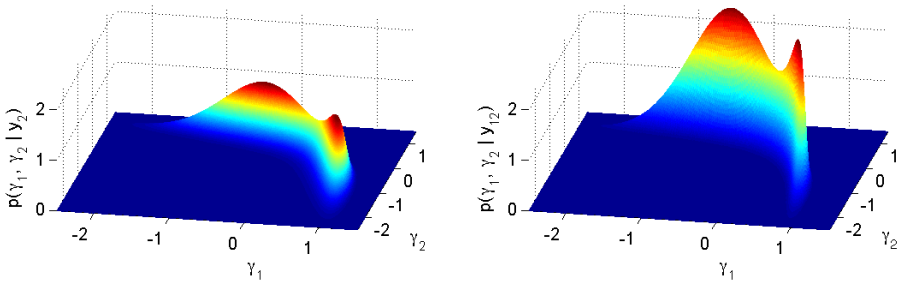
(a) Prior. (b) Posterior. $M = 5$.

Figure 4: The prior and posterior pdfs of mixture Gaussian with two separated modes in Example 5.1.



(a) $\xi = 0.3$. (b) $\xi = 1$.

Figure 5: The convergence of the expected information gain computed using the LA + MC, LA + SG, and DLMC with a mixture Gaussian prior in Example 5.1.



(a) $M = 2$ (b) $M = 12$

Figure 6: The posterior pdfs of $\boldsymbol{\gamma}$ with two separated modes in Example 5.1.

integrals with two separate Gaussian pdfs, so that an auxiliary measure is not needed. The same number of quadrature points is used in both integrations. We note that the separated integration reaches high precision when the total number of quadratures is higher than 10

(see Figure 7). The convergence of both approaches in terms of the absolute consecutive difference is shown in Figure 8.

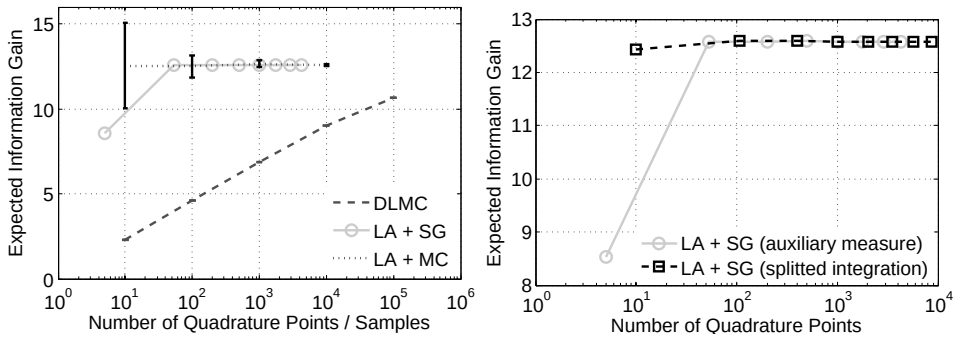


Figure 7: The convergence of the expected information gain computed using the LA + MC, LA + SG and using DLMC, with a mixture log Gaussian prior and $\xi = 1$ in Example 5.1.

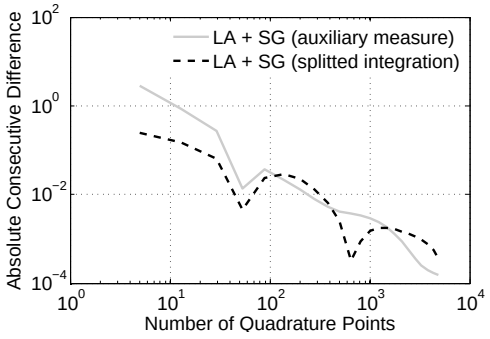


Figure 8: The absolute consecutive difference of the expected information gain computed by LA + SG in Example 5.1 with a mixture log Gaussian prior and $\xi = 1$.

5.2. Example 5.2: design of the current pattern for impedance tomography

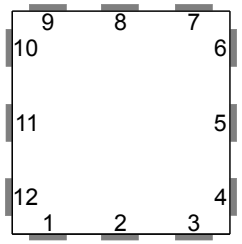
In [1], we directly applied the Laplace approximation to estimate the expected information gain for various current patterns in a square domain, to infer the material conductivities, i.e., the optimal design of so-called impedance tomography. That model was under-determined due to too few measurements.

In our current study, the Laplace approximations are carried out at the quadrature or sample points in the orthogonal directions to the low dimensional non-informative manifold. Specifically, the dimension of the manifold is two in this example, since there are only two measurements available. The physics considered is simply Ohm's law in a continuum. Electric currents are injected into the domain through electrodes attached at fixed locations on the boundary. Measurements of the voltages are taken at the electrodes to infer the resistance distribution. The admissible source locations are indexed as in Figure 9. We choose two out of 12 possible source locations in our experimental design scenario. There are in total 66 possible combinations as listed in Table 1 of [1].

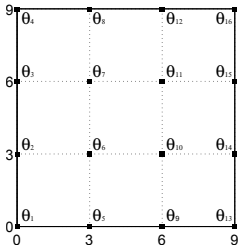
The physics is governed by a Poisson equation in a square subdomain of $\mathbb{R}^2$. Furthermore, we use the shunt model [22] of the electrodes, which assumes that the electrodes are perfect conductors. Hence, the solution of the Poisson equation (electric potential) is constant in each electrode. We refer to Appendix E for the weak form, the boundary conditions, the finite element discretization and the associated notations. We also formulate the model output (voltages at the electrodes) as a function of the solution of the discretized weak form for the Poisson equation. We continue using the same formulae for the finite element discretization and the Jacobian of the model as in [1]. We divide the subdomain Ω into nine regions (see Figure 9): $\Omega_1, \Omega_2, ..., \Omega_9$. The difference is that now we consider a piecewise bilinear continuous $\boldsymbol{\theta}(\boldsymbol{x})$ random field, i.e.,

$$\theta(\boldsymbol{x}) = \sum_{I=1}^{16} N^I(\boldsymbol{x})\theta_I \,,$$

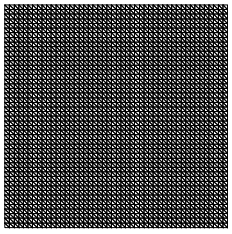
where $N^I(\boldsymbol{x})$ is the I^{th} basis function corresponding to the nodal random parameter, θ_I. Since the bilinear shape functions are zero everywhere except the neighboring elements, we normally express this interpolation in terms of the four nodal parameter values of an element. See Appendix D for the shape functions in local coordinates.



(a) The admissible locations of boundary sources and the corresponding indices in Example 5.2.



(b) The subregion and indices of the nodal parameters of the piece-wise linear random field in Example 5.2.



(c) The triangular finite element mesh.

Figure 9: The layout, parameters and finite element mesh of the impedance tomography in Example 5.2.

Now, we define the following random vector as the unknown parameter in our experimental design problem:

$$\boldsymbol{\theta} = [\theta_1,\, \theta_2,\, ..., \theta_{16}]^T \,.$$

5.2.1. Log normal prior

We assume that this random vector is a multivariate log normal, i.e., $\log(\boldsymbol{\theta}) \sim \mathcal{N}(\boldsymbol{\gamma}_0, \Sigma_p)$. The prior mean vector is $\boldsymbol{\gamma}_0 = \mathbf{0}$, and the covariance matrix is

$$\Sigma_p(4,4) = \Sigma_p(7,7) = \Sigma_p(10,10) = \Sigma_p(13,13) = 1 \,,$$

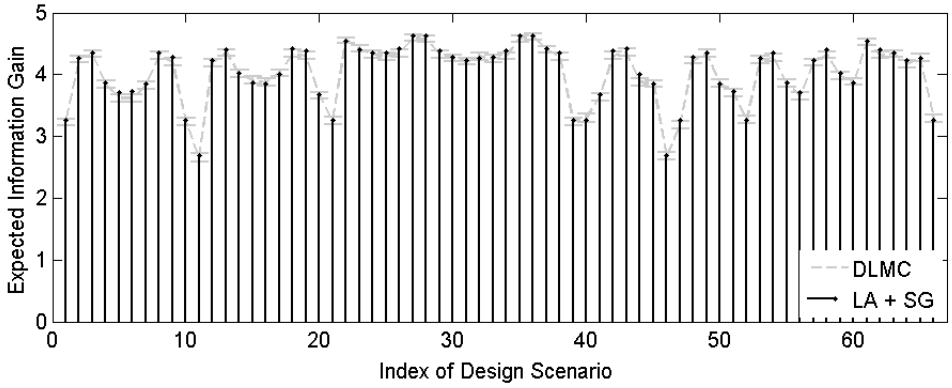


Figure 10: The expected information gains computed for all possible combinations of current sources in Example 5.2..

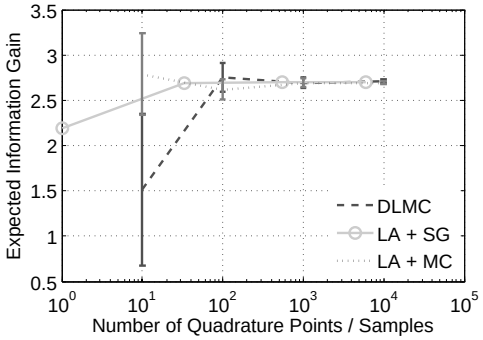


Figure 11: The convergence of the expected information gain in the 11^{th} scenario using the LA + MC, LA + SG and DLMC in Example 5.2, with a log Gaussian prior.

$$\Sigma_p(i,i) = 0.01, i \neq 4,7,10,13 \quad \text{and} \quad \Sigma_p(i,j) = 0, i \neq j\,,$$

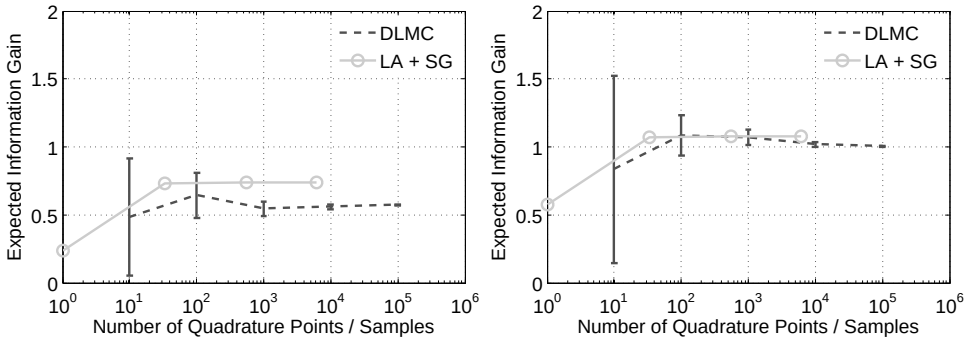
respectively. The covariance matrix for the measurement noise is assumed to be

$$\boldsymbol{\Sigma}_\epsilon = \sigma_m^2 \boldsymbol{I}_{2\times 2} \quad \text{with} \quad \sigma_m^2 = 0.001\,.$$

Similar to the results obtained in [1] where the unknown parameter field is assumed to be piecewise constant, the expected information gain reaches its maximum at the 27^{th} scenario (source locations: $3, 9$) and the 36^{th} scenario (source locations: $4, 10$). The expected information gains for the 28^{th} (source locations: $3, 10$) and the 35^{th} (source locations: $4, 9$) scenarios are in the vicinities of the maximum points. In these scenarios, most of the electric current goes through the material, lumped along the diagonal where prior variances are high. However, due to the coupling of the nodal parameters, the worst scenario is not the 52^{nd} and its symmetrical counterpart as in the case where the piecewise constant parameter field is assumed. Instead, they are 11^{th} (source locations: $1, 12$) and 46^{th} (source locations: $6, 7$) current patterns, in which a large amount of current is restricted to the bottom-left and top-right corner areas.

In Figure 11, we show the convergence of the expected information gain in the 11^{th} scenario. We have the fastest convergence of the expected information gain when using sparse grid integration (Gauss-Hermite) for (24). To reach approximately the same level of accuracy, we need to use at least 1000 times more samples in Monte Carlo sampling (both in LA + MC and DLMC) . The dashed curve in Figure 11 also shows the bias of the DLMC approach when the number of samples is smaller than 10^2.

Furthermore, we increased the noise of the measurements to $\sigma_m^2 = 0.1$. Figure 12(a) compares the convergences of the expected information gain computed using the LA + SG and DLMC, when $M = 10$. Our approach converges to a value of 0.74 when four levels of the Gauss-Hermite sparse grid with 6049 quadrature points are used. The absolute consecutive difference reduces to 10^{-4} (see Figure 13) for this number of quadrature points. In the same setting, the expected information gain computed by DLMC sampling converges to 0.574 with a 97.5% confidence interval ± 0.005, when using 10^5 samples in both the outer and inner loops. Furthermore, we increased the number of measurements to $M = 30$ and computed the expected information gain again using our method (LA + SG) and DLMC. The results are shown in Figure 12(b). Our approach converges to a value of 1.08, when four levels of the Gauss-Hermite sparse grid with 6049 quadrature points are used. The absolute consecutive difference is almost identical to the case where $M = 10$ (see Figure 13). In the same setting, the expected information gain by DLMC converges to 1.008 with a 97.5% confidence interval ± 0.005, when using 10^5 samples in both the outer and inner loops. We note that as M increases, our approximation becomes more accurate. Specifically in this case, the error is around 8% when $M = 30$, while the error is approximately 30% when $M = 10$.



(a) $M = 10$. (b) $M = 30$.

Figure 12: Convergence of the expected information gain in the 11^{th} scenario using the LA + SG and DLMC in Example 5.2 with a log Gaussian prior and $\sigma_m^2 = 0.1$.

5.2.2. Mixture log normal prior

Similar to what we did in the first example, we set the prior as a mixture log Gaussian adopting the following form:

$$p(\boldsymbol{\gamma}) = 0.5 \times p_1(\boldsymbol{\gamma}) + 0.5 \times p_2(\boldsymbol{\gamma}) \,, \tag{48}$$

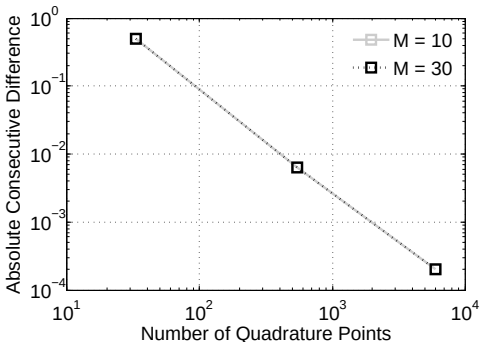


Figure 13: The absolute consecutive difference of expected information gain in the 11^{th} scenario computed using the LA + SG in Example 5.2 with a log Gaussian prior and $\sigma_m^2 = 0.1$.

where $p_1(\boldsymbol{\gamma})$ is the pdf that we used in the single Gaussian case, and $p_2(\boldsymbol{\gamma})$ is the pdf of a multivariate Gaussian with its mean vector and covariance matrix as follows:

$$\gamma_0(4) = \gamma_0(7) = \gamma_0(10) = \gamma_0(13) = 2\,, \gamma_0(i) = 0, i \neq 4, 7, 10, 13\,,$$

$$\Sigma_p(4,4) = \Sigma_p(7,7) = \Sigma_p(10,10) = \Sigma_p(13,13) = 1\,,$$
$$\Sigma_p(i,i) = 0.01, i \neq 4, 7, 10, 13\,, \quad \text{and} \quad \Sigma_p(i,j) = 0, i \neq j\ .$$

When $M = 30$ and $\sigma_m^2 = 0.1$, the difference between the results of LA + SG (0.945) and DLMC (1.007 ± 0.006) is around 6%. We show the convergence curve with respect to the number of quadrature points or samples in Figure 14(a). We reduce the magnitude of noise variance to $\sigma_m^2 = 0.001$ and keep $M = 30$ in Figure 14(b). In this case, our result is almost identical to the one computed by DLMC using 10^5 samples, i.e., 3.248 for LA + SG and 3.237 ± 0.006 for DLMC, respectively, which indicates a negligible bias of our approach in this example with mixture log Gaussian prior. In addition, we also demonstrate the fast convergence rate of the sparse quadratures in terms of the absolute consecutive difference in Figure 15.

5.2.3. Prediction of the quantity of interest

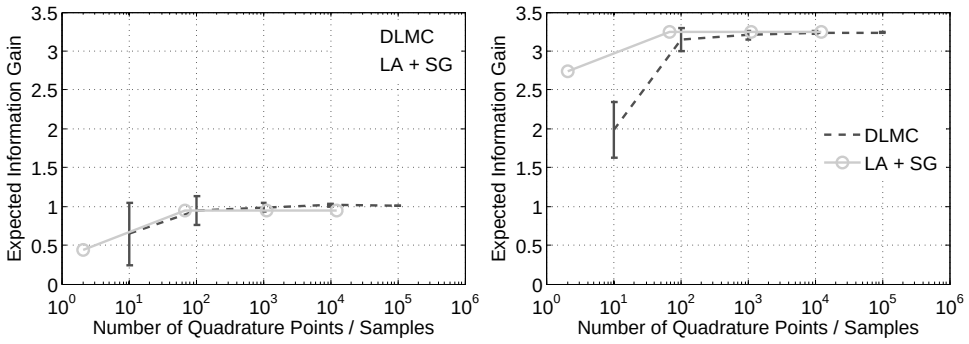
We now consider $\tau(\boldsymbol{\theta})$, the mean value of the quantity of interest, Q, in our prediction scenario as a function of the corresponding electric potential:

$$\tau(\boldsymbol{\theta}) = \boldsymbol{P}_h \boldsymbol{U}_h(\boldsymbol{\theta}) = \boldsymbol{P}_h \boldsymbol{K}_h^{-1}(\boldsymbol{\theta}) \boldsymbol{F}_h\,, \tag{49}$$

where $\boldsymbol{P}_h$ is an operator on $\boldsymbol{U}_h$. For example, we may be interested in the average voltage over a certain subdomain, $\Omega_v \in \Omega$, i.e.,

$$\tau(\boldsymbol{\theta}) = \boldsymbol{P}_{h,q} \boldsymbol{U}_{h,q}(\boldsymbol{\theta}) = \frac{1}{A_{\Omega_v}} \int_{\Omega_v} u_h(\boldsymbol{x}) d\boldsymbol{x}\,,$$

where A_{Ω_v} is the area of Ω_v. The Jacobian of $\tau(\boldsymbol{\theta})$ w.r.t. the parameters can be derived in a way similar to the derivation of $\boldsymbol{J}_g$ (see Appendix E).



(a) $\sigma_m^2 = 0.1$. (b) $\sigma_m^2 = 0.001$

Figure 14: The convergence of the expected information gain in the 11^{th} scenario using the LA + SG and DLMC in Example 5.2 with a mixture log Gaussian prior.

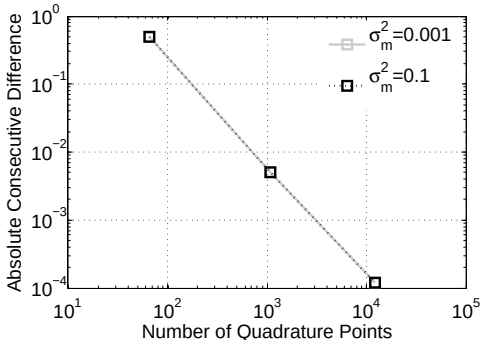


Figure 15: The absolute consecutive difference of the expected information gain in the 11^{th} scenario using the LA + SG in Example 5.2. A mixture Gaussian prior was used.

The Ω_v is a square domain with an area equal to 1.2656. We firstly consider that its four corners are $(7.5, 1.5)$, $(8.625, 1.5)$, $(8.625, 2.625)$ and $(7.5, 2.625)$. The finite element mesh complies to the boundary of this small area. We use the boundary conditions of scenario 22 in the prediction scenario. The previous log normal is used as the prior. We compute the expected conditional entropy, $H(Q|\bar{\boldsymbol{y}})$, for all the design scenarios using 1000 samples in (44). The results are shown in Figure 16. As expected, the 22^{nd} scenario is associated with the maximum information gain, and the 61^{st} scenario is associated with the minimum information gain. We plot the electric potential fields of both cases in Figures 17(a) and 17(b), where the black box indicates the Ω_v. The conductivity field is visualized in Figure 18.

Next, we consider Ω_v in the middle of the domain with corners at $(4.5, 4.5)$, $(5.625, 4.5)$, $(5.625, 5.625)$ and $(4.5, 5.625)$. While the lowest information gain remains the 61^{st} scenario and its symmetrical cases with respect to the prior parameter field, the scenarios with the highest information gains become the 41^{st} scenario and its symmetrical counterparts. Figure 20 shows the electrical potential field in the 41^{st} scenario, where the conductivity field is kept the same as in the previous computation.

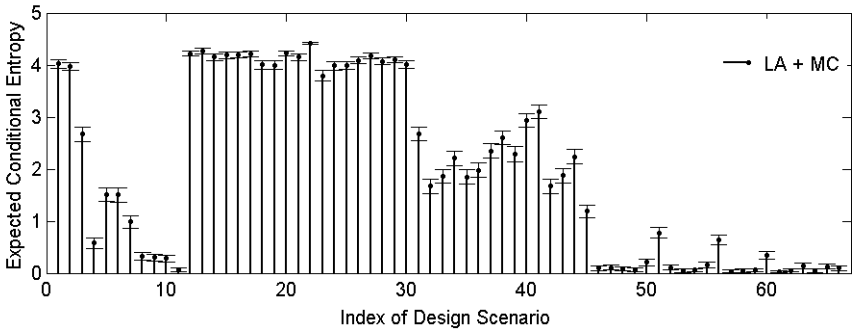
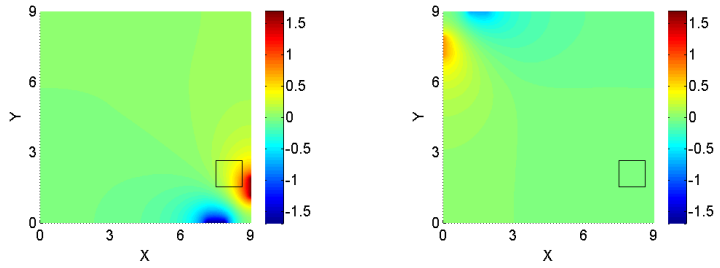


Figure 16: The expected conditional entropy of the average electric potential in a small square region computed for all possible combinations of current sources in Example 5.2.



(a) 22^{nd} Scenario (maximum information gain)

(b) 61^{st} Scenario (minimum information gain)

Figure 17: Two samples of the potential fields from the scenarios of maximum and minimum information gains, respectively, in Example 5.2. The black box indicates the area of the integrated quantity of interest.

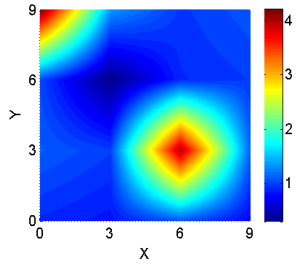


Figure 18: A sample of the conductivity field in Example 5.2.

6. Conclusion

In this work, we extended the Bayesian experimental design methodology based on the Laplace approximation from determined cases to under-determined cases. Instead of carrying out the Laplace approximation at a single well-defined posterior mode, we perform the Laplace approximation in the orthogonal directions of a submanifold on which the parameters are not identifiable after a local reparameterization. The reparameterization is driven by the low-rank Hessian of the cost function defined as a quadratic form of the difference

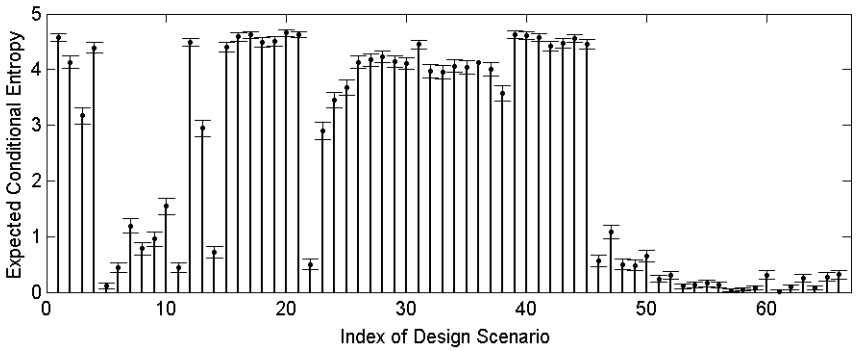


Figure 19: The expected conditional entropy of the average electric potential in a small square region computed for all possible combinations of current sources in Example 5.2.

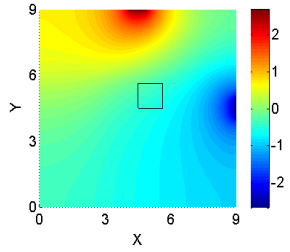


Figure 20: A sample of the potential field from the scenario of maximum information gains in Example 5.2. The black box indicates the area of the integrated quantity of interest.

between the true model and a proposed model. Eventually, the expected information gain can be approximated asymptotically as an integration over the parameter domain similar to the cases where the model parameters are determined completely by the experiments. In the final formulation shown in Theorem 8, we approximated the marginal pdf of the parameters orthogonal to the non-informative submanifold, using a linearized submanifold at the modes found by a constrained optimization. We also developed the techniques for the prediction of quantities of interest based on the same strategy. The proposed formulae (Theorems 7, 8 and 10) are able to deal with the designs based on an under-determined model and multimodal or non-informative (uniform) priors. To carry out the numerical integration to compute the approximated expected information gain, we can use sparse quadratures or Monte Carlo sampling depending on the regularity of the integrand of the proposed formulae. We have demonstrated the efficiency and accuracy of our method using several numerical examples. They include the designs of the scalar experimental setup in a one-dimensional cubic polynomial function with two unidentifiable parameters and the boundary source locations of impedance tomography in a square domain for both inferences and predictions, considering a piecewise linear continuous parameter field. Future work could be carried out on diminishing the cost of computing the Hessian matrix by using surrogate models or multi-level Monte Carlo.

Acknowledgement

We are thankful for support from the Academic Excellency Alliance UT Austin-KAUST project–Uncertainty quantification for predictive modeling of the dissolution of porous and fractured media, and the Institute of Applied Mathematics and Computational Sciences at TAMU. Part of this work was carried out while M. Scavino and S. Wang were Visiting Professors at KAUST. S. Wang's research was also partially supported by Award Number KUS-CI-016-04, made by King Abdullah University of Science and Technology (KAUST). Q. Long, M. Scavino and R. Tempone are members of the KAUST SRI Center for Uncertainty Quantification in Computational Science and Engineering.

Appendix A. Proof of the error estimate in Equation (19)

The Laplace approximation error, $\epsilon_{laplace}$, in Equation (19) can be expressed as follows:

$$\int_{\boldsymbol{T}}\int_{\boldsymbol{S}} \log\left[\frac{\tilde{p}(\boldsymbol{s},\boldsymbol{t})}{p(\boldsymbol{s},\boldsymbol{t})}\right]\tilde{p}(\boldsymbol{s}|\boldsymbol{t},\bar{\boldsymbol{y}})d\boldsymbol{s}d\boldsymbol{t} + \int_{\boldsymbol{T}}\int_{\boldsymbol{S}} \log\left[\frac{p(\boldsymbol{s},\boldsymbol{t}|\bar{\boldsymbol{y}})}{\tilde{p}(\boldsymbol{s},\boldsymbol{t}|\bar{\boldsymbol{y}})}\right]\tilde{p}(\boldsymbol{s}|\boldsymbol{t},\bar{\boldsymbol{y}})d\boldsymbol{s}d\boldsymbol{t}$$
$$+ \int_{\boldsymbol{T}}\int_{\boldsymbol{S}} \log\left[\frac{p(\boldsymbol{s},\boldsymbol{t}|\bar{\boldsymbol{y}})}{p(\boldsymbol{s},\boldsymbol{t})}\right]\left[p(\boldsymbol{s}|\boldsymbol{t},\bar{\boldsymbol{y}}) - \tilde{p}(\boldsymbol{s}|\boldsymbol{t},\bar{\boldsymbol{y}})\right]d\boldsymbol{s}d\boldsymbol{t} = E_1 + E_2 + E_3\,.$$

To prove the error order for E_1, we consider the inner integration of $\boldsymbol{s}$ for a fixed value of $\boldsymbol{t}$. We write the Taylor series of $h_p(\boldsymbol{s},\boldsymbol{t})$, defined in Section 2.4, in the vicinity of $\hat{\boldsymbol{s}}$ as follows

$$h_p(\boldsymbol{s},\boldsymbol{t}) = \sum_{|\boldsymbol{\alpha}|\leq 4}\frac{D^{\boldsymbol{\alpha}}h_p(\hat{\boldsymbol{s}},\boldsymbol{t})}{\boldsymbol{\alpha}!}(\boldsymbol{s}-\hat{\boldsymbol{s}})^{\boldsymbol{\alpha}} + O_P\left(\|\boldsymbol{s}-\hat{\boldsymbol{s}}\|^5\right)\,,$$

where we use the multi-index notation, $\boldsymbol{\alpha}$, with the following properties:

$$|\boldsymbol{\alpha}| = \alpha_1 + \cdots + \alpha_d\,,\ \alpha! = \alpha_1!\cdots\alpha_d!\,,\ (\boldsymbol{s})^{\boldsymbol{\alpha}} = s_1^{\alpha_1}\cdots s_d^{\alpha_d}\,.$$

The odd central moments of the multivariate Gaussian are zero and the parameter posterior covariance, $\boldsymbol{\Sigma}$, is of $O_P\left(\frac{1}{M}\right)$. It is straightforward to see that the fourth and sixth moments of this multivariate Gaussian are $O_P\left(\frac{1}{M^2}\right)$ and $O_P\left(\frac{1}{M^3}\right)$, respectively. Consequently, the conditional expectation of $h_p(\boldsymbol{s},\boldsymbol{t})$ is

$$\begin{aligned}\int_{\boldsymbol{S}} h_p(\boldsymbol{s},\boldsymbol{t})\tilde{p}(\boldsymbol{s}|\boldsymbol{t},\bar{\boldsymbol{y}})d\boldsymbol{s} =& h_p(\hat{\boldsymbol{s}},\boldsymbol{t}) + \frac{\boldsymbol{\Sigma}_{\boldsymbol{s}|\boldsymbol{t}} : \nabla\nabla h_p(\hat{\boldsymbol{s}},\boldsymbol{t})}{2}\\ &+ \frac{1}{4!}\sum_{i,j,k,l}(\partial_{ijkl}h_p)(\Sigma_{\boldsymbol{s}|\boldsymbol{t},ij}\Sigma_{\boldsymbol{s}|\boldsymbol{t},kl} + \Sigma_{\boldsymbol{s}|\boldsymbol{t},ik}\Sigma_{\boldsymbol{s}|\boldsymbol{t},jl} + \Sigma_{\boldsymbol{s}|\boldsymbol{t},il}\Sigma_{\boldsymbol{s}|\boldsymbol{t},jk})\\ &+ O_P\left(\frac{1}{M^3}\right)\\ =& h_p(\hat{\boldsymbol{s}},\boldsymbol{t}) + \frac{\boldsymbol{\Sigma}_{\boldsymbol{s}|\boldsymbol{t}} : \nabla\nabla h_p(\hat{\boldsymbol{s}},\boldsymbol{t})}{2} + O_P\left(\frac{1}{M^2}\right)\\ =& \int_{\boldsymbol{S}}\log(\tilde{p}(\boldsymbol{s},\boldsymbol{t}))\tilde{p}(\boldsymbol{s}|\boldsymbol{t},\bar{\boldsymbol{y}})d\boldsymbol{s} + O_P\left(\frac{1}{M^2}\right)\,,\end{aligned}$$

with i, j, k, $l = 1,...,\dim(\boldsymbol{s})$ and $\partial_{ijkl}h = \frac{\partial^4 h(\hat{\boldsymbol{s}},\boldsymbol{t})}{\partial s_i \partial s_j \partial s_k \partial s_l}$. Therefore, $E_1 = O_P(\frac{1}{M^2})$.

Next, regarding E_2,

$$\log\left[\frac{p(\boldsymbol{s},\boldsymbol{t}|\bar{\boldsymbol{y}})}{\tilde{p}(\boldsymbol{s},\boldsymbol{t}|\bar{\boldsymbol{y}})}\right]$$

consists of terms of order higher than the quadratic in the Taylor series of the posterior pdf in $\boldsymbol{s}$, i.e.,

$$\log\left[\frac{p(\boldsymbol{s},\boldsymbol{t}|\bar{\boldsymbol{y}})}{\tilde{p}(\boldsymbol{s},\boldsymbol{t}|\bar{\boldsymbol{y}})}\right] = \sum_{|\boldsymbol{\alpha}|=3,4} \frac{D^{\boldsymbol{\alpha}} h(\hat{\boldsymbol{s}},\boldsymbol{t})}{\boldsymbol{\alpha}!}(\boldsymbol{s}-\hat{\boldsymbol{s}})^{\boldsymbol{\alpha}} + O_P\left(\|\boldsymbol{s}-\hat{\boldsymbol{s}}\|^5\right).$$

Similar to the analysis of the expectation of $h_p(\boldsymbol{s},\boldsymbol{t})$ above, the expectation of this log ratio is

$$\begin{aligned}\int_{\boldsymbol{S}} \log\left[\frac{p(\boldsymbol{s},\boldsymbol{t}|\bar{\boldsymbol{y}})}{\tilde{p}(\boldsymbol{s},\boldsymbol{t}|\bar{\boldsymbol{y}})}\right]\tilde{p}(\boldsymbol{s}|\boldsymbol{t},\bar{\boldsymbol{y}})d\boldsymbol{s} =& \frac{1}{4!}\sum_{i,j,k,l}(\partial_{ijkl}h_p)\left(\Sigma_{\boldsymbol{s}|\boldsymbol{t},ij}\Sigma_{\boldsymbol{s}|\boldsymbol{t},kl} + \Sigma_{\boldsymbol{s}|\boldsymbol{t},ik}\Sigma_{\boldsymbol{s}|\boldsymbol{t},jl} + \Sigma_{\boldsymbol{s}|\boldsymbol{t},il}\Sigma_{\boldsymbol{s}|\boldsymbol{t},jk}\right) \\ &+ O_P\left(\frac{1}{M^3}\right) = O_P\left(\frac{1}{M^2}\right).\end{aligned}$$

Finally, regarding to the third term, E_3, we have

$$\begin{aligned}&\int_{\boldsymbol{T}}\int_{\boldsymbol{S}} \log\left[\frac{p(\boldsymbol{s},\boldsymbol{t}|\bar{\boldsymbol{y}})}{p(\boldsymbol{s},\boldsymbol{t})}\right](p(\boldsymbol{s}|\boldsymbol{t},\bar{\boldsymbol{y}}) - \tilde{p}(\boldsymbol{s}|\boldsymbol{t},\bar{\boldsymbol{y}}))d\boldsymbol{s}d\boldsymbol{t} \\ &\quad= \int_{\boldsymbol{T}}\int_{\boldsymbol{S}} \log\left[\frac{p(\boldsymbol{s},\boldsymbol{t}|\bar{\boldsymbol{y}})}{p(\boldsymbol{s},\boldsymbol{t})}\right]\left\{\exp\left[\sum_{|\boldsymbol{\alpha}|=3}\frac{D^{\boldsymbol{\alpha}}h_p(\hat{\boldsymbol{s}})}{\boldsymbol{\alpha}!}(\boldsymbol{s}-\hat{\boldsymbol{s}})^{\boldsymbol{\alpha}} + O_P(||\boldsymbol{s}-\hat{\boldsymbol{s}}||^4)\right] - 1\right\}\tilde{p}(\boldsymbol{s}|\boldsymbol{t},\bar{\boldsymbol{y}})d\boldsymbol{s}d\boldsymbol{t}.\end{aligned}$$

After the first order expansion of the exponential term, we obtain

$$\int_{\boldsymbol{T}}\int_{\boldsymbol{S}} \log\left[\frac{p(\boldsymbol{s},\boldsymbol{t}|\bar{\boldsymbol{y}})}{p(\boldsymbol{s},\boldsymbol{t})}\right]\left\{\sum_{|\boldsymbol{\alpha}|=3}\frac{D^{\boldsymbol{\alpha}}h_p(\hat{\boldsymbol{s}})}{\boldsymbol{\alpha}!}(\boldsymbol{s}-\hat{\boldsymbol{s}})^{\boldsymbol{\alpha}} + O_P(||\boldsymbol{s}-\hat{\boldsymbol{s}}||^4)\right\}\tilde{p}(\boldsymbol{s}|\boldsymbol{t},\bar{\boldsymbol{y}})d\boldsymbol{s}d\boldsymbol{t}.$$

Since $\log\left[\frac{p(\boldsymbol{s},\boldsymbol{t}|\bar{\boldsymbol{y}})}{p(\boldsymbol{s},\boldsymbol{t})}\right]$ is $O_P(1)$ in $\boldsymbol{s}$ and the third moment of a multivariate Gaussian is zero, the rate of this error is dominated by

$$\int_{\boldsymbol{T}}\int_{\boldsymbol{S}} O_P(||\boldsymbol{s}-\hat{\boldsymbol{s}}||^4)\tilde{p}(\boldsymbol{s}|\boldsymbol{t}\bar{\boldsymbol{y}})d\boldsymbol{s}d\boldsymbol{t},$$

which has already been shown to be inversely proportional to M^2. Now, it is straightforward to observe that the dominant term is $O_P\left(\frac{1}{M^2}\right)$. These three error terms are similar to the error terms in Appendices A, B and C in [1], respectively.

Appendix B. Proof of the error estimate of the conditional maximum posterior estimator, $\hat{\boldsymbol{s}}$

Let

$$R(\boldsymbol{s}) = \frac{1}{2}M(\boldsymbol{g}(\boldsymbol{\theta}_0)-\boldsymbol{g}(\boldsymbol{f}(\boldsymbol{s},\boldsymbol{t}))^T\boldsymbol{\Sigma}_\epsilon^{-1}(\boldsymbol{g}(\boldsymbol{\theta}_0)-\boldsymbol{g}(\boldsymbol{f}(\boldsymbol{s},\boldsymbol{t})))+\boldsymbol{E}_s{}^T\boldsymbol{\Sigma}_\epsilon^{-1}(\boldsymbol{g}(\boldsymbol{f}(\boldsymbol{s},\boldsymbol{t}))-\boldsymbol{g}(\boldsymbol{\theta}_0))-h_p(\boldsymbol{s},\boldsymbol{t})\,.$$

We then have that

$$\nabla R(\boldsymbol{s}) = M\boldsymbol{J}_s^T\boldsymbol{\Sigma}_\epsilon^{-1}(\boldsymbol{g}(\boldsymbol{f}(\boldsymbol{s},\boldsymbol{t})) - \boldsymbol{g}(\boldsymbol{\theta}_0)) + \boldsymbol{J}_s^T\boldsymbol{\Sigma}_\epsilon^{-1}\boldsymbol{E}_s - \nabla h_p(\boldsymbol{s},\boldsymbol{t})\,,$$

where $\boldsymbol{E}_s$ is the summation of residual vectors defined in [1]. If we ignore the higher order terms, the first order expansion of $\nabla R(\boldsymbol{s})$ at $(\boldsymbol{0},\boldsymbol{t})$ reads

$$\nabla R(\boldsymbol{s}) = \nabla R(\boldsymbol{0}) + \nabla\nabla R(\boldsymbol{0})\boldsymbol{s}\,.$$

Therefore, using Newton's method, $\nabla R(\hat{\boldsymbol{s}}) = \boldsymbol{0}$ implies that $\nabla R(\boldsymbol{0}) + \nabla\nabla R(\boldsymbol{0})\hat{\boldsymbol{s}} = \boldsymbol{0}$. Thus,

$$\begin{aligned}\hat{\boldsymbol{s}} =&\boldsymbol{0} - (M\boldsymbol{J}_s^T\boldsymbol{\Sigma}_\epsilon^T\boldsymbol{J}_s + \boldsymbol{H}_s^T\boldsymbol{\Sigma}_\epsilon^{-1}\boldsymbol{E}_s - \nabla\nabla h_p(\boldsymbol{0},\boldsymbol{t}))^{-1}(\boldsymbol{J}_s^T\boldsymbol{\Sigma}_\epsilon^{-1}\boldsymbol{E}_s - \nabla h_p(\boldsymbol{0},\boldsymbol{t}))\\ =&\boldsymbol{0} - (M\boldsymbol{J}_s^T\boldsymbol{\Sigma}_\epsilon^T\boldsymbol{J}_s + \boldsymbol{H}_s^T\boldsymbol{\Sigma}_\epsilon^{-1}\boldsymbol{E}_s - \nabla\nabla h_p(\boldsymbol{0},\boldsymbol{t}))^{-1}(\boldsymbol{J}_s^T\boldsymbol{\Sigma}_\epsilon^{-1}\boldsymbol{E}_s)\\ &+ (M\boldsymbol{J}_s^T\boldsymbol{\Sigma}_\epsilon^T\boldsymbol{J}_s + \boldsymbol{H}_s^T\boldsymbol{\Sigma}_\epsilon^{-1}\boldsymbol{E}_s - \nabla\nabla h_p(\boldsymbol{0},\boldsymbol{t}))^{-1}\nabla h_p(\boldsymbol{0},\boldsymbol{t})\,.\end{aligned}$$

As $M\to\infty$, $\hat{\boldsymbol{s}} = \boldsymbol{0} - (M\boldsymbol{J}_s^T\boldsymbol{\Sigma}_\epsilon^T\boldsymbol{J}_s + \boldsymbol{H}_s^T\boldsymbol{\Sigma}_\epsilon^{-1}\boldsymbol{E}_s - \nabla\nabla h_p(\boldsymbol{0},\boldsymbol{t}))^{-1}(\boldsymbol{J}_s^T\boldsymbol{\Sigma}_\epsilon^{-1}\boldsymbol{E}_s) + O_P\left(\frac{1}{M}\right) = O_P\left(\frac{1}{\sqrt{M}}\right)$.

Appendix C. Proof of the error estimate in Equation (39)

We express the posterior distribution of the quantity of interest, Q, as

$$\begin{aligned}p(Q|\bar{\boldsymbol{y}}) &= \int_{\boldsymbol{T}} p(Q|\boldsymbol{t},\bar{\boldsymbol{y}})p(\boldsymbol{t}|\bar{\boldsymbol{y}})d\boldsymbol{t}\\ &= \int_{\boldsymbol{T}} \tilde{p}(Q|\boldsymbol{t},\bar{\boldsymbol{y}})p(\boldsymbol{t}|\bar{\boldsymbol{y}})d\boldsymbol{t} + \int_{\boldsymbol{T}}\left[p(Q|\boldsymbol{t},\bar{\boldsymbol{y}}) - \tilde{p}(Q|\boldsymbol{t},\bar{\boldsymbol{y}})\right]p(\boldsymbol{t}|\bar{\boldsymbol{y}})d\boldsymbol{t}\\ &= \int_{\boldsymbol{T}} \tilde{p}(Q|\boldsymbol{t},\bar{\boldsymbol{y}})(\int_{\boldsymbol{S}}\tilde{p}(\boldsymbol{s},\boldsymbol{t}|\bar{\boldsymbol{y}})d\boldsymbol{s})d\boldsymbol{t} + \int_{\boldsymbol{T}} \tilde{p}(Q|\boldsymbol{t},\bar{\boldsymbol{y}})(\int_{\boldsymbol{S}} p(\boldsymbol{s},\boldsymbol{t}|\bar{\boldsymbol{y}}) - \tilde{p}(\boldsymbol{s},\boldsymbol{t}|\bar{\boldsymbol{y}})d\boldsymbol{s})d\boldsymbol{t}\\ &\quad + \int_{\boldsymbol{T}}\left[p(Q|\boldsymbol{t},\bar{\boldsymbol{y}}) - \tilde{p}(Q|\boldsymbol{t},\bar{\boldsymbol{y}})\right]p(\boldsymbol{t}|\bar{\boldsymbol{y}})d\boldsymbol{t}\\ &= \tilde{p}(Q|\bar{\boldsymbol{y}}) + \int_{\boldsymbol{T}}\left[p(Q|\boldsymbol{t},\bar{\boldsymbol{y}}) - \tilde{p}(Q|\boldsymbol{t},\bar{\boldsymbol{y}})\right]p(\boldsymbol{t}|\bar{\boldsymbol{y}})dt + O_P(\frac{1}{M^2})\,, \qquad \text{(C.1)}\end{aligned}$$

where the error rate is obtained in a way similar to the derivation of E_3. Furthermore, we have

$$
\begin{aligned}
&p(Q|\boldsymbol{t},\bar{\boldsymbol{y}}) \\
&= \int_{\boldsymbol{S}} p(Q|\boldsymbol{s},\boldsymbol{t},\bar{\boldsymbol{y}}) p(\boldsymbol{s}|\boldsymbol{t},\bar{\boldsymbol{y}}) d\boldsymbol{s} \\
&= \int_{\boldsymbol{S}} p(Q|\boldsymbol{s},\boldsymbol{t},\bar{\boldsymbol{y}}) \tilde{p}(\boldsymbol{s}|\boldsymbol{t},\bar{\boldsymbol{y}}) d\boldsymbol{s} + \int_{\boldsymbol{S}} p(Q|\boldsymbol{s},\boldsymbol{t},\bar{\boldsymbol{y}}) \left[p(\boldsymbol{s}|\boldsymbol{t},\bar{\boldsymbol{y}}) - \tilde{p}(\boldsymbol{s}|\boldsymbol{t},\bar{\boldsymbol{y}})\right] d\boldsymbol{s} \\
&= \int_{\boldsymbol{S}} \exp\left\{ \frac{[Q - \tau(\hat{\boldsymbol{s}},\boldsymbol{t}) - \nabla\tau(\hat{\boldsymbol{s}},\boldsymbol{t})(\boldsymbol{s}-\hat{\boldsymbol{s}})]^2 - [Q-\tau(\boldsymbol{s},\boldsymbol{t})]^2}{2\sigma_Q^2} \right\} \tilde{p}(Q|\boldsymbol{s},\boldsymbol{t},\bar{\boldsymbol{y}}) \tilde{p}(\boldsymbol{s}|\boldsymbol{t},\bar{\boldsymbol{y}}) d\boldsymbol{s} \\
&\quad + \int_{\boldsymbol{S}} p(Q|\boldsymbol{s},\boldsymbol{t},\bar{\boldsymbol{y}}) \left[p(\boldsymbol{s}|\boldsymbol{t},\bar{\boldsymbol{y}}) - \tilde{p}(\boldsymbol{s}|\boldsymbol{t},\bar{\boldsymbol{y}})\right] d\boldsymbol{s} \\
&= \int_{\boldsymbol{S}} \exp\left\{ [(\boldsymbol{s}-\hat{\boldsymbol{s}})^T \nabla\nabla\tau(\hat{\boldsymbol{s}},\boldsymbol{t})(\boldsymbol{s}-\hat{\boldsymbol{s}}) + O_P(||\boldsymbol{s}-\hat{\boldsymbol{s}}||^3)] O_P(||\boldsymbol{s}-\hat{\boldsymbol{s}}||) \right\} \tilde{p}(Q|\boldsymbol{s},\boldsymbol{t},\bar{\boldsymbol{y}}) \tilde{p}(\boldsymbol{s}|\boldsymbol{t},\bar{\boldsymbol{y}}) d\boldsymbol{s} \\
&\quad + \int_{\boldsymbol{S}} p(Q|\boldsymbol{s},\boldsymbol{t}) \left[p(\boldsymbol{s}|\boldsymbol{t},\bar{\boldsymbol{y}}) - \tilde{p}(\boldsymbol{s}|\boldsymbol{t},\bar{\boldsymbol{y}})\right] d\boldsymbol{s} \\
&= \int_{\boldsymbol{S}} \left\{ [(\boldsymbol{s}-\hat{\boldsymbol{s}})^T \nabla\nabla\tau(\hat{\boldsymbol{s}},\boldsymbol{t})(\boldsymbol{s}-\hat{\boldsymbol{s}}) + O_P(||\boldsymbol{s}-\hat{\boldsymbol{s}}||^3)] O_P(||\boldsymbol{s}-\hat{\boldsymbol{s}}||) + 1 \right\} \tilde{p}(Q|\boldsymbol{s},\boldsymbol{t}) \tilde{p}(\boldsymbol{s}|\boldsymbol{t},\bar{\boldsymbol{y}}) d\boldsymbol{s} \\
&\quad + \int_{\boldsymbol{S}} p(Q|\boldsymbol{s},\boldsymbol{t}) \left[p(\boldsymbol{s}|\boldsymbol{t},\bar{\boldsymbol{y}}) - \tilde{p}(\boldsymbol{s}|\boldsymbol{t},\bar{\boldsymbol{y}})\right] d\boldsymbol{s} \\
&= \int_{\boldsymbol{S}} \tilde{p}(Q|\boldsymbol{s},\boldsymbol{t}) \tilde{p}(\boldsymbol{s}|\boldsymbol{t},\bar{\boldsymbol{y}}) d\boldsymbol{s} + \int_{\boldsymbol{S}} p(Q|\boldsymbol{s},\boldsymbol{t}) \left[p(\boldsymbol{s}|\boldsymbol{t},\bar{\boldsymbol{y}}) - \tilde{p}(\boldsymbol{s}|\boldsymbol{t},\bar{\boldsymbol{y}})\right] d\boldsymbol{s} + O_P\left(\frac{1}{M^2}\right).
\end{aligned}
$$

By reusing the expansion of $p(\boldsymbol{s}|\boldsymbol{t},\bar{\boldsymbol{y}}) - \tilde{p}(\boldsymbol{s}|\boldsymbol{t},\bar{\boldsymbol{y}})$ derived in E_3, we obtain

$$
p(Q|\boldsymbol{t},\bar{\boldsymbol{y}}) - \tilde{p}(Q|\boldsymbol{t},\bar{\boldsymbol{y}}) = \int_{\boldsymbol{S}} p(Q|\boldsymbol{s},\boldsymbol{t}) \left[p(\boldsymbol{s}|\boldsymbol{t},\bar{\boldsymbol{y}}) - \tilde{p}(\boldsymbol{s}|\boldsymbol{t},\bar{\boldsymbol{y}})\right] d\boldsymbol{s} + O_P\left(\frac{1}{M^2}\right) = O_P\left(\frac{1}{M^2}\right).
$$

Therefore,

$$
p(Q|\bar{\boldsymbol{y}}) = \tilde{p}(Q|\bar{\boldsymbol{y}}) + O_P(\frac{1}{M^2}). \tag{C.2}
$$

In this connection, the first error term in (39), i.e., the error term before marginalizing over data, can be obtained as

$$
\begin{aligned}
&\int_{\mathcal{Q}} \log\left[\frac{p(Q|\bar{\boldsymbol{y}})}{\tilde{p}(Q|\bar{\boldsymbol{y}})}\right] \tilde{p}(Q|\bar{\boldsymbol{y}}) dQ + \int_{\mathcal{Q}} p(Q|\bar{\boldsymbol{y}}) \left(p(Q|\bar{\boldsymbol{y}}) - \hat{p}(Q|\bar{\boldsymbol{y}})\right) dQ \\
&\quad = \int_{\mathcal{Q}} \log\left[\frac{p(Q|\bar{\boldsymbol{y}}) - \tilde{p}(Q|\bar{\boldsymbol{y}})}{\tilde{p}(Q|\bar{\boldsymbol{y}})} + 1\right] \tilde{p}(Q|\bar{\boldsymbol{y}}) dQ + O_P\left(\frac{1}{M^2}\right) \\
&\quad = \int_{\mathcal{Q}} \left[\frac{p(Q|\bar{\boldsymbol{y}}) - \tilde{p}(Q|\bar{\boldsymbol{y}})}{\tilde{p}(Q|\bar{\boldsymbol{y}})} + O_P\left(\frac{p(Q|\bar{\boldsymbol{y}}) - \tilde{p}(Q|\bar{\boldsymbol{y}})}{\tilde{p}(Q|\bar{\boldsymbol{y}})}\right)^2\right] \tilde{p}(Q|\bar{\boldsymbol{y}}) dQ + O_P\left(\frac{1}{M^2}\right) \\
&\quad = O_P\left(\frac{1}{M^2}\right).
\end{aligned}
$$

Appendix D. Bilinear shape functions

The four bilinear shape functions for one square element are

$$\begin{aligned} N_1(\xi,\eta) &= \frac{1}{4}(1-\xi)(1-\eta)\,, \\ N_2(\xi,\eta) &= \frac{1}{4}(1+\xi)(1-\eta)\,, \\ N_3(\xi,\eta) &= \frac{1}{4}(1+\xi)(1+\eta)\,, \\ N_4(\xi,\eta) &= \frac{1}{4}(1-\xi)(1+\eta)\,, \end{aligned} \tag{D.1}$$

where ξ and η are two local coordinates of a quadrature point in the square.

Appendix E. Brief summary of the impedance tomography problem

We briefly summarize the problem setup of the impedance tomography example. Readers should refer to [1] for details.

We introduce V, the space of $H^1(\Omega)$ functions that are constant over the electrodes, i.e.,

$$V = \{v \in H^1(\Omega) \,:\, v|_{a_j} = U_j,\, j = 1, ..., l\,, \text{ and } \sum_{j=1}^{l} U_j = 0\}\,,$$

where $H^1(\Omega)$ is the Sobolev space with a square integrable gradient. The electric potential field, $u \in V$, should satisfy

$$B(u,v) = L(v)\,, \quad \forall v \in V \subset H^1(\Omega)\,,$$

with

$$B(u,v) := \int_\Omega \theta\, \nabla u \cdot \nabla v \, d\boldsymbol{x} \quad \text{and} \quad L(v) := \sum_{j=1}^{l} v_j I_j\,, \quad \forall v \in V \subset H^1(\Omega)\,, \tag{E.1}$$

as well as the conditions:

$$\boldsymbol{q}\cdot\boldsymbol{n} = 0 \quad \text{on} \quad \delta\Omega / \bigcup_{j=1}^{l} a_j, \quad \sum_{j=1}^{l} U_j = 0\,, \quad \text{and} \quad \sum_{j=1}^{l} I_j = 0\,.$$

We use piecewise linear continuous finite elements on a uniform triangulation Ω_h (see Figure 9) to solve the Poisson equation numerically given the value of $\boldsymbol{\theta}$. The regular mesh is comprised of 4608 triangles. The mesh size h is chosen to be small enough so that the discretization error is negligible. Then we want to find $u_h \in V_h \subseteq V$, such that

$$B(u_h, v_h) = L(v_h),\, \forall v_h \in V_h\,,$$

with $B(\cdot,\cdot)$ and $L(\cdot)$ as in (E.1). This leads to a linear system of equations, namely

$$\boldsymbol{K}(\boldsymbol{\theta})\boldsymbol{U}_h = \boldsymbol{F}_h\,.$$

The measurements are the voltages at the electrodes, which are approximated by area boundary sources on segments of length 0.75, each of which is resolved by four elements. The 12 admissible source coordinates (centroids of the segments) are $(1.5,0)$, $(4.5,0)$, $(7.5,0)$, $(9,1.5)$, $(9,4.5)$, $(9,7.5)$, $(7.5,9)$, $(4.5,9)$, $(1.5,9)$, $(0,7.5)$, $(0,4.5)$ and $(0,1.5)$. We choose only two out of these 12 to perform the measurement. Thus, there are 66 possible combinations as listed in Table 1 in [1].

The model of the i^{th} voltage measurement after applying the finite element approximation reads (note that the discrepancy between the true model and the finite element approximation is assumed to be negligible):

$$\boldsymbol{y}_i = \boldsymbol{g}(\boldsymbol{\theta}_t) + \boldsymbol{\epsilon}_i = \boldsymbol{P}_h\boldsymbol{U}_h(\boldsymbol{\theta}_t) + \boldsymbol{\epsilon}_i\,, \quad i = 1,...,M\,,$$

where the length of the $\boldsymbol{y}_i$ vector is the number of electrodes from which we take the measurements. Each of the rows of $\boldsymbol{P}_h$ is such that we have

$$(\boldsymbol{P}_h\boldsymbol{U}_h)_j = \frac{1}{|a_j|}\int_{a_j} u_h(\boldsymbol{x})d\boldsymbol{x}\,, \quad j = 1,...,l\,,$$

for each of the l electrodes. Here, the coordinates of the electrodes comply with the nodal positions of the mesh.

The Jacobian of the measurement model, $\boldsymbol{g}$, with respect to the conductivity, $\boldsymbol{\theta}$, is

$$\boldsymbol{J}_g = \frac{\partial \boldsymbol{g}}{\partial \boldsymbol{\theta}} = -\boldsymbol{P}_h\boldsymbol{K}^{-1}\sum_{e=1}^{NE}\frac{\partial \boldsymbol{K}_e}{\partial \boldsymbol{\theta}}\boldsymbol{U}_h\,,$$

where NE denotes the number of elements in the finite element discretization, $\boldsymbol{K}_e$ is the element stiffness matrix, which sums to $\boldsymbol{K}$. In particular, the k^{th} column of matrix $\boldsymbol{J}_g$ is

$$\boldsymbol{J}_{gk} = \frac{\partial \boldsymbol{g}}{\partial \theta_k} = -\boldsymbol{P}_h\boldsymbol{K}^{-1}\sum_{e=1}^{NE}\frac{\partial \boldsymbol{K}_e}{\partial \theta_k}\boldsymbol{K}^{-1}\boldsymbol{F}_h\,, \quad k = 1,...,16\,,$$

where the summation applies to all the elements in region Ω_h.

References

[1] Q. Long, M. Scavino, R. Tempone, S. Wang, Fast estimation of expected information gains for bayesian experimental designs based on laplace approximations, Computer Methods in Applied Mechanics and Engineering 259 (2013) 24–39.

[2] J. Ginebra, On the measure of the information in a statistical experiment, Bayesian Analysis 2 (2007) 167–211.

[3] K. Chaloner, I. Verdinelli, Bayesian experimental design: a review, Statistical Science 10 (1995) 273–304.
[4] S. M. Stigler, Laplace's 1774 memoir on inverse probability, Statistical Science 1 (1986) 359–363.
[5] L. Tierney, R. E. Kass, J. B. Kadane, Fully exponential Laplace approximations to expectations and variances of nonpositive functions, Journal of the American Statistical Association 84 (1989) 710–716.
[6] R. E. Kass, L. Tierney, J. B. Kadane, The validity of posterior expansions based on Laplace's method, in: Essays in Honor of George Barnard (eds. S. Geisser, J. S. Hodges, S. J. Press, and A. Zellner), North–Holland, Amsterdam, 1990, pp. 473–488.
[7] R. E. Kass, L. Tierney, J. B. Kadane, Laplace's method in Bayesian analysis, Contemporary Mathematics 115 (1991) 89–99.
[8] B. S. Clarke, A. R. Barron, Entropy Risk and the Bayesian central limit theorem, Technical Report 91-56, Department of Statistics, Purdue University, 1991.
[9] N. G. Polson, On the expected amount of information from a non-linear model, Journal of the Royal Statistical Society, Series B 54 (1992) 889–895.
[10] S. Ghosal, T. Samanta, Expansion of Bayes risk for entropy loss and reference prior in nonregular cases, Statistics & Decisions 15 (1997) 129–140.
[11] J. M. Bernardo, Reference posterior distributions for Bayesian inference, Journal of the Royal Statistical Society, Series B 41 (1979) 113–147.
[12] N. G. Polson, Bayesian Perspectives on Statistical Modelling, Ph.D. thesis, Department of Mathematics, University of Nottingham, 1988.
[13] B. S. Clarke, L. Wasserman, Noninformative priors and nuisance parameters, Journal of the American Statistical Association 88 (1993) 1427–1432.
[14] B. S. Clarke, A. Yuan, Partial information reference priors: derivation and interpretations, Journal of Statistical Planning and Inference 123 (2004) 313–345.
[15] M. Joannides, F. Le Gland, Small noise asymptotics of the Bayesian estimator in nonidentifiable models, Statistical Inference for Stochastic Processes 5 (2002) 95–130.
[16] C.-R. Hwang, Laplace's method revisited: weak convergence of probability measures, Annals of Probability 8 (1980) 1177–1182.
[17] M. Spivak, A Comprehensive Introduction to Differential Geometry, volume I, Publish or Perish, Inc., Houston, Texas, 3rd edition, 1999.
[18] W. W. Hager, Updating the inverse of a matrix, SIAM Review 31 (1989) pp. 221–239.
[19] S. A. Smolyak, Quadrature and interpolation formulas for tensor products of certain classes of functions, Soviet Math. Dokl. 4 (1963) 240–243.
[20] V. Barthelmann, E. Novak, K. Ritter, High dimensional polynomial interpolation on sparse grids, Advances in Computational Mathematics 12 (2000) 273–288.
[21] F. Nobile, R. Tempone, C. G. Webster, A sparse grid stochastic collocation method for partial differential equations with random input data, SIAM Journal on Numerical Analysis 46 (2008) 2309–2345.
[22] E. Somersalo, M. Cheney, D. Isaacson, Existence and uniqueness for electrode models for electric current computed tomography, SIAM Journal on Applied Mathematics 52 (1992) 1023–1040.