Geophysical Journal International

Geophys. J. Int. (2024) **236,** 1309–1331
Advance Access publication 2023 December 23
GJI General Geophysical Methods

https://doi.org/10.1093/gji/ggad492

Variational Bayesian experimental design for geophysical applications: seismic source location, amplitude versus offset inversion, and estimating CO_2 saturations in a subsurface reservoir

Dominik Strutz and Andrew Curtis

School of GeoSciences, University of Edinburgh, Edinburgh EH93FE, UK. E-mail: dominik.strutz@ed.ac.uk

Accepted 2023 December 20. Received 2023 December 19; in original form 2023 July 3

SUMMARY

In geophysical surveys or experiments, recorded data are used to constrain properties of the planetary subsurface, oceans, atmosphere or cryosphere. How the experimental data are collected significantly influences which parameters can be resolved and how much confidence can be placed in the results. Bayesian experimental design methods characterize, quantify and maximize expected information post-experiment—an optimization problem. Typical design parameters that can be optimized are source and/or sensor types and locations, and the choice of modelling or data processing methods to be applied to the data. These may all be optimized subject to various physical and cost constraints. This paper introduces *variational* design methods, and discusses their benefits and limitations in the context of geophysical applications. Variational methods have recently come to prominence due to their importance in machine-learning applications. They can be used to design experiments that best resolve either all model parameters, or the answer to specific questions about the system to be interrogated. The methods are tested in three schematic geophysical applications: (i) estimating a source location given arrival times of radiating energy at sensor locations, (ii) estimating the contrast in seismic velocity across a stratal interface given measurements of the amplitudes of seismic wavefield reflections from that interface, and (iii) designing a survey to best constrain CO_2 saturation in a subsurface storage scenario. Variational methods allow the value of expected information from an experiment to be calculated and optimized simultaneously, which results in substantial savings in computational cost. In the context of designing a survey to best constrain CO_2 saturation in a subsurface storage scenario, we show that optimal designs may change substantially depending on the particular questions of interest. We also show that one method, so-called D_N design, can be effective at substantially lower computational cost than other methods. Overall, this work demonstrates that optimal design methods could be used more widely in Geophysics, as they are in other scientifically advanced fields.

Key words: Inverse theory; Machine learning; Probability distributions.

1 INTRODUCTION

Every geophysical investigation that collects data is an experiment, usually intended to estimate parameters that describe the properties of natural systems. How the experimental data are collected significantly influences which parameters can be resolved and how much confidence can be placed in the results. It is well known that the expected results can be improved by explicitly optimizing the experimental design.

The field of optimal experimental design (OED) has a long history in monitoring industrial processes (Taguchi methods, Kiefer 1959; Atkinson & Fedorov 1975) and had its first geophysical application in 1977, optimizing seismometer placement to locate seismic sources (Kijko 1977a, b). The field has developed significantly since then: in Geophysics, OED has been used to design more sophisticated source location experiments (Rabinowitz & Steinberg 1990, 2000; Steinberg *et al.* 1995; Curtis *et al.* 2004; Rawlinson *et al.* 2012; Bloem *et al*. 2020; Toledo *et al*. 2020), seismic tomography surveys (Curtis & Snieder 1997; Curtis 1999a, b; Liner *et al.* 1999; Gibson & Tzimeas 2002; Curtis *et al.* 2004; Brenders & Pratt 2007; Haber *et al.* 2008; Ajo-Franklin 2009; Coles & Morgan 2009; Maurer *et al.* 2009, 2017; Khodja *et al.* 2010; Coles &

Curtis 2011a; Djikpesse *et al.* 2012; Coles *et al.* 2013; Bernauer *et al.* 2014; Nuber *et al.* 2017; Krampe *et al.* 2021), reflected wave amplitude inversions (van Den Berg *et al.* 2003, 2005; Guest & Curtis 2009, 2010, 2011), electromagnetic and electrical resistivity tomography (Maurer & Boerner 1998; Maurer *et al.* 2000, 2010; Stummer *et al.* 2002, 2004; Furman *et al.* 2004; Wilkinson *et al.* 2006, 2012; Oldenborger & Routh 2009; Qiang *et al.* 2022; Coles & Morgan 2009), electrical impedance tomography (Hyvönen *et al.* 2014), expert elicitation (Curtis & Wood 2004; Runge *et al.* 2013), contaminant transport (Alexanderian *et al.* 2014; Zhang *et al.* 2015; Alexanderian & Saibaba 2018), CO_2 monitoring (Romdhane & Eliasson 2018), local array design (Muir & Zhan 2021), and others.

There are three pre-requisites for any experimental design problem: first, a function (physical or empirical) that relates the vector of model parameters $\boldsymbol{m}$ to the vector of synthetic observations $\boldsymbol{d}$, called the *forward function*. This relationship

$$\boldsymbol{d} = F(\boldsymbol{m}) \tag{1}$$

can be linear, but in most geophysical problems, F is nonlinear. Second, we require a description of what is already known about the values of all parameters in vector $\boldsymbol{m}$, which are necessary to evaluate the forward function; this prior knowledge is usually described by a probability distribution function (pdf—a probability density if variables are continuous rather than discrete), called the prior pdf. Third, a pdf describing the probability of observing a datum $\boldsymbol{d}$if any particular set of values for a model parameter vector $\boldsymbol{m}$ were true. The latter is commonly referred to as the likelihood, and is often approximated by a Gaussian with mean $F(\boldsymbol{m})$ and variance corresponding to the measurement uncertainty; this effectively changes eq. (1) to $\boldsymbol{d} = F(\boldsymbol{m}) + \boldsymbol{\epsilon}$, where errors $\boldsymbol{\epsilon}$ are drawn from the Gaussian distribution. Using Bayesian inference, these three states of knowledge can be combined with information about $\boldsymbol{m}$ derived from any future measured data set, the resultant state of information being described by the so-called posterior pdf.

Both the content of data vector $\boldsymbol{d}$ and the form of F are significantly influenced by how an experiment is set up, in other words, by the experimental design. The appropriate approach to design an experiment depends on the nature of the forward function $F(\boldsymbol{m})$. Fig. 1(a) shows an example of a linear model-data relationship. The observed data (green and blue) and their uncertainties (here, a Gaussian distribution) on the vertical axis are mapped into the model parameter domain by backprojecting the data uncertainty through F onto the parameter axis m. The backprojected values are then scaled by the prior pdf in model space, which is here assumed to be constant. Clearly, a more certain observation (green) yields a more certain model parameter estimate (a narrower pdf on the horizontal axis) than a less certain observation (blue). However, the data uncertainty ranges are also scaled by the reciprocal slope of the forward function: high gradients lead to more certain model parameter estimates and, therefore, lower uncertainties. Since the slope is constant, this is true for any recorded datum with similar uncertainties. Accordingly, OED for linear forward functions varies the design so as to maximize the gradient (or higher dimensional equivalents) since a higher slope results in more accurate model parameter estimates given the same data uncertainties.

The effectiveness of this design criterion deteriorates as F deviates from linearity. Most properties still hold in a related sense for slightly nonlinear functions, as shown in Fig. 1(b). Each measured datum still maps to a single point in model parameter space, and provided data uncertainties are small, the slope of the function at the backprojection point is usually the most influential factor affecting how measurement uncertainties relate to model uncertainties. However, unlike for linear functions, the slope now depends on the model parameter values, so parameter uncertainties do too: for example, in Fig. 1(b) note that compared to the blue data and parameter estimates, the more accurate green data measurement produces slightly larger parameter uncertainties due to the lower gradient of F encountered during backprojection. In this case, the range of parameter values that we expect might occur (our prior information about the parameters) thus influences which design is optimal. An average slope over the range of possible model parameter values can thus be used as a design quality metric for slightly nonlinear forward functions. Since the range of parameter values that might be encountered is described by the Bayesian prior probability distribution, optimizing this quality metric is typically referred to as Bayesian OED in the statistical literature. However, this terminology is misleading, because this approach only optimizes an average of a design quality measure that only accounts for physics that is linearized around each parameter value. It would therefore be more appropriate to refer to these as pseudo-Bayesian or linearized design methods (Pronzato & Walter 1985; Chaloner & Verdinelli 1995; Fedorov & Hackl 1997; Winterfors & Curtis 2012; Ryan *et al.* 2016).

The simple averaging method described above breaks down for generally nonlinear functions, especially for those that are multimodal (have multiple distinct peaks and troughs). An example of such a function and how it affects the mapping from data to model parameter space is given in Fig. 1(c). This shows that a single datum may be consistent with different distinct regions of model parameter space far removed from each other (green). Even a datum whose mean measurement value is consistent with only a single model can also map to a range of distinctly different model parameter values due to its measurement uncertainty around the mean (blue). If we are to define a quality measure that describes the aspects of any experimental design for a fully nonlinear forward function, clearly, it must depend on all models that might explain the data. Since the set of parameter values consistent with the data are then non-unique, possibly disjoint, and may have a varying probability of being true given the data, we describe the set of values by the Bayesian posterior pdf. The most commonly used design quality metric in substantially nonlinear problems is the expected information gain (EIG, Lindley 1956), which will be introduced in Section 3.

This work aims to introduce a set of Bayesian OED methods that are novel to Geophysics. Each method calculates the EIG in the context of geophysical applications with generally linear or nonlinear forward functions. We illustrate their relative merits and computational costs in the context of three representative Geophysical experiments: (i) locating seismic sources based on *P*- and *S*-wave arrival times, (ii) assessing the contrast in seismic velocity across a stratal interface given measurements of the amplitudes of waves reflected from that interface and (iii) designing a survey to best constrain CO_2 saturation in a subsurface storage scenario. While these examples concern seismic waves, they are representative of design problems for the location of other source types (Kim & Lees 2014; Lugrin *et al.* 2014) and for reflections of other wave types (Tarantola 1984; Hunziker *et al.* 2016), since elastic waves exhibit intermediate physical complexity between acoustic and electromagnetic waves. We also demonstrate that optimal designs may change substantially depending on which question about the subsurface we wish to answer (Arnold & Curtis 2018). These results and illustrations show that optimal design methods might usefully be adopted more widely in Geophysics as they are in other scientifically advanced disciplines, and they allow practitioners to make more informed choices between the various methods for their particular applications.

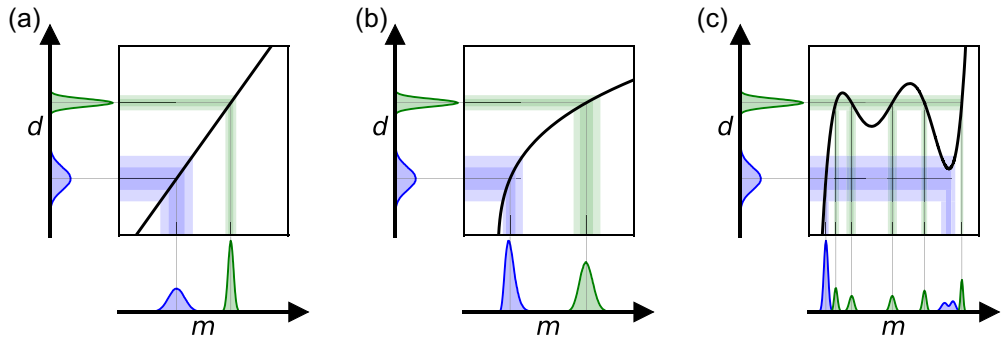

Figure 1. The green and blue Gaussian distributions on the d-axis represent two measurements with different uncertainties. Distributions on the m-axis represent the resulting backprojections of the two measurements through (a) linear, (b) slightly nonlinear and (c) more strongly nonlinear model-data relationships (black lines).

2 BAYESIAN EXPERIMENTAL DESIGN

The Bayesian posterior distribution is denoted $p(\boldsymbol{m} \mid \boldsymbol{d})$ and describes the state of knowledge post-experiment. It can be expressed using Bayes rule as

$$p(\boldsymbol{m} \mid \boldsymbol{d}) = \frac{p(\boldsymbol{m})p(\boldsymbol{d} \mid \boldsymbol{m})}{\int_{\mathbb{M}} p(\boldsymbol{m}')\, p(\boldsymbol{d} \mid \boldsymbol{m}')\, d\boldsymbol{m}'} = \frac{p(\boldsymbol{m})p(\boldsymbol{d} \mid \boldsymbol{m})}{p(\boldsymbol{d})}, \tag{2}$$

where $p(\boldsymbol{m})$is the prior pdf, $p(\boldsymbol{d} \mid \boldsymbol{m})$ is called the likelihood function and $p(\boldsymbol{d})$ is called the evidence. The prior and posterior pdf's are defined over the model parameter space $\mathbb{M}$, while the likelihood and evidence are defined over data space $\mathbb{D}$. The dimensionality of these spaces is a primary factor in the computational complexity of both inference and design problems.

An intuitive example of the model parameter space is the 3-D space for seismic source (earthquake) location, with an optional fourth dimension for the origin time. A corresponding data space might be the n-dimensional space of arrival times of first-arriving waves detected at each of n receivers. Alternatively, we might pick both P- and S-wave arrival times, in which case the data space might consist of these data at $n/2$ receivers. This example thus illustrates that the choice of data processing algorithms may change the data space substantially, and so may be a primary element of any experimental design.

Eq. (2) can be used to characterize the solution to many geophysical parameter estimation problems. Since the interest of this paper lies in experimental design, it is necessary to include a variable describing the design $\xi \in \Xi$, where Ξ denotes the space of all potential experimental setups. Provided that the experimental design does not impose a change in the parametrization of the model, the prior distribution $p(\boldsymbol{m})$ is usually not affected by a change in design. The design influences the solution to any inference problem through the likelihood $p(\boldsymbol{d} \mid \boldsymbol{m}, \xi)$, which in turn affects the evidence. The posterior of the model parameters given an observation and experimental design is therefore

$$p(\boldsymbol{m} \mid \boldsymbol{d}, \xi) = \frac{p(\boldsymbol{m})p(\boldsymbol{d} \mid \boldsymbol{m}, \xi)}{\int_{\mathbb{M}} p(\boldsymbol{m}')\, p(\boldsymbol{d} \mid \boldsymbol{m}', \xi)\, d\boldsymbol{m}'} = \frac{p(\boldsymbol{m})p(\boldsymbol{d} \mid \boldsymbol{m}, \xi)}{p(\boldsymbol{d} \mid \xi)}. \tag{3}$$

The evidence $p(\boldsymbol{d} \mid \xi)$ acts as a normalizing factor for the resulting posterior pdf and is sometimes not calculated explicitly when solving inference problems—for example, commonly used Markov chain Monte Carlo Methods (McMC) are designed to avoid its calculation (Mosegaard & Tarantola 1995). However, many nonlinear OED algorithms depend on $p(\boldsymbol{d} \mid \xi)$. For nonlinear problems, calculating either $p(\boldsymbol{m} \mid \boldsymbol{d}, \xi)$ or $p(\boldsymbol{d} \mid \xi)$generally requires many evaluations of $p(\boldsymbol{d} \mid \boldsymbol{m}, \xi)$ to estimate the integral expression in eq. (3), and each evaluation of $p(\boldsymbol{d} \mid \boldsymbol{m}, \xi)$ requires a computation of the forward function. Therefore, the tractability of design problems depends on the complexity of F and the number of evaluations required to estimate this integral, unless a way to avoid its evaluation or a sufficiently accurate approximation is found. For a general, yet detailed mathematical treatment of Bayesian inference and additional examples, we refer readers to Tarantola (2005).

Generally, we aim to optimize experiments such that they maximize information in the posterior distribution, within bounds imposed by practical constraints on the cost of performing the experiment. We therefore, need a metric that quantifies the information embodied within any probability distribution. Shannon (1948) information is an intuitive measure of information with several beneficial properties (e.g. linear additivity of information from independent sources). The Shannon information $I[\,\cdot\,]$ described by an arbitrary continuous probability density function $p(x)$ is defined as

$$\mathrm{I}[p(x)] = \mathbb{E}_{p(x)}\left[\log_b(p(x))\right] = \int_{\mathcal{X}} p(x) \log_b(p(x)) \mathrm{d}x, \tag{4}$$

where $x \in \mathcal{X}$ is a random variable distributed according to $p(x)$ and $\mathbb{E}_{p(x)}$ is the expectation with respect to $p(x)$, which is defined by the rightmost expression. Depending on the context, information is also often expressed as the negative of the entropy H, $\mathrm{I}[p(x)] = -\mathrm{H}[p(x)]$, where entropy H is defined to be the negative of either expression on the right of eq. (4). This absolute information measure can be extended to the relative information content of one pdf relative to another, also called the Kullback–Leibler (KL) divergence (Kullback & Leibler 1951)

$$\mathrm{KL}(P||Q) = \int_{\mathcal{X}} p(x) \log\left(\frac{p(x)}{q(x)}\right) \mathrm{d}x. \tag{5}$$

For further information on the properties of information, the reader is referred to Cover & Thomas (2006).

Following Lindley (1956), we now define the information gain (IG) about $\boldsymbol{m}$ obtained by recording data $\boldsymbol{d}$ using experimental design ξ to be the difference between the posterior state of information and the information about $\boldsymbol{m}$ in the prior pdf:

$$\mathrm{IG}(\xi, \boldsymbol{d}) = \mathrm{I}[p(\boldsymbol{m} \mid \boldsymbol{d}, \xi)] - \mathrm{I}[p(\boldsymbol{m})] \tag{6}$$

$$= \mathrm{KL}[p(\boldsymbol{m} \mid \boldsymbol{d}, \xi)||p(\boldsymbol{m})] \tag{7}$$

$$= \mathbb{E}_{p(\boldsymbol{m} \mid \boldsymbol{d}, \xi)}[\log(p(\boldsymbol{m} \mid \boldsymbol{d}, \xi))] - \mathbb{E}_{p(\boldsymbol{m})}[\log(p(\boldsymbol{m}))], \tag{8}$$

where eq. (8) is obtained by substituting eq. (4) into eq. (6). The IG depends on $\boldsymbol{d}$, which is not available during the design stage of

an experiment. However, the evidence $p(\boldsymbol{d}\,|\,\xi)$ provides the probability of observing any particular value for $\boldsymbol{d}$, given our prior state of information about the parameters, and the forward function and measurement uncertainties which are both included within the likelihood (eq. 3). The so-called EIG is therefore defined as the expectation of IG over $\boldsymbol{d} \sim p(\boldsymbol{d}\,|\,\xi)$, giving

$$\mathrm{EIG}(\xi) = \mathbb{E}_{p(\boldsymbol{d}\,|\,\xi)}\left[\mathrm{I}[p(\boldsymbol{m}\,|\,\boldsymbol{d},\xi)] - \mathrm{I}[p(\boldsymbol{m})]\right]. \tag{9}$$

This criterion depends only on ξ and is used in this work to assess the quality of any experimental design ξ prior to the collection of data. The EIG is equivalent to $\mathbb{E}_{p(\boldsymbol{d}\,|\,\xi)}\,\mathrm{KL}\,(p(\boldsymbol{m}\,|\,\boldsymbol{d},\xi)||\,p(\boldsymbol{m}))$, and using this definition we can show that the EIG is identical, even if the slightly different definition of the IG by Rényi (1961) is used. It also makes clear that the EIG is strictly positive, except if the models and data are completely independent so the experiment provides no information in which case the EIG is zero.

While eq. (9) is perhaps the most intuitive way to express the EIG, its value is often calculated using other expressions. Rewriting eq. (9) using eq. (2), we obtain:

$$\mathrm{EIG}(\xi) = \mathbb{E}_{p(\boldsymbol{d},\boldsymbol{m}\,|\,\xi)}\left[\log \tfrac{p(\boldsymbol{m}\,|\,\boldsymbol{d},\xi)}{p(\boldsymbol{m})}\right] \tag{10}$$

$$= \mathbb{E}_{p(\boldsymbol{d},\boldsymbol{m}\,|\,\xi)}\left[\log \tfrac{p(\boldsymbol{d},\boldsymbol{m}\,|\,\xi)}{p(\boldsymbol{m})p(\boldsymbol{d}\,|\,\xi)}\right] \tag{11}$$

$$= \mathbb{E}_{p(\boldsymbol{d},\boldsymbol{m}\,|\,\xi)}\left[\log \tfrac{p(\boldsymbol{d}\,|\,\boldsymbol{m},\xi)}{p(\boldsymbol{d}\,|\,\xi)}\right]. \tag{12}$$

In these expressions, the information is written out explicitly, and all three lines are equivalent and can be derived by repeated use of Bayes theorem. Both eqs (9) and (10) are written only in terms of pdfs in the model parameter space, meaning they assign probability densities to each vector of model parameter values (e.g. earthquake locations in the example introduced above). This means that evaluating the EIG in this form requires that we solve an inverse problem to obtain the posterior distribution $p(\boldsymbol{m}\,|\,\boldsymbol{d},\xi)$ for each data vector $\boldsymbol{d}$ (or for each of a representative subset of data samples) that is likely to be observed according to $p(\boldsymbol{d}\,|\,\xi)$. This can be impractical in several ways: first, sampling-based methods such as McMC can not be used, since they do not provide explicit probability values $p(\boldsymbol{m}_i\,|\,\boldsymbol{d},\xi)$ for given models $\boldsymbol{m}_i$, and second, solving a large number of inverse problems independently is computationally very inefficient since an expectation over a possibly high-dimensional model space needs to be evaluated for each $\boldsymbol{d}$ drawn from $p(\boldsymbol{d}\,|\,\xi)$.

Using eq. (12), on the other hand, the EIG is expressed only in terms of distributions in data space, $p(\boldsymbol{d}\,|\,\boldsymbol{m},\xi)$and $p(\boldsymbol{d}\,|\,\xi)$,

$$\mathrm{EIG}(\xi) \triangleq \mathbb{E}_{p(\boldsymbol{m})}\left[\mathrm{I}[p(\boldsymbol{d}\,|\,\boldsymbol{m},\xi)] - \mathrm{I}[p(\boldsymbol{d}\,|\,\xi)]\right] \tag{13}$$

making it possible to design experiments without explicitly solving inverse problems; this therefore, involves sampling the whole model space only once. In this formulation, the main difficulty lies in estimating $\mathrm{I}\,[p(\boldsymbol{d}\,|\,\xi)]$. Especially for high-dimensional data spaces (e.g. arrival times recorded by many receivers) this is challenging. For problems with low-dimensional model parameter and high-dimensional data spaces, it may therefore still be beneficial to calculate the EIG in the model parameter space (eq. 9). Both model parameter and data space approaches are analysed in more detail herein.

Using any expression for the EIG as our design metric, the best design is expressed mathematically as

$$\xi^* = \underset{\xi\in\Xi}{\arg\max}\ \mathrm{EIG}(\xi), \tag{14}$$

where Ξ is the set of all possible experimental designs. Ideally, this optimization is global, but in many cases a greedy (local) optimization algorithm must be used to make the optimization computationally tractable. Such algorithms usually provide a significant improvement over non-optimized experiments. Further details on the optimization process are given in Section 4.

2.1 Comments on linear design theory

This paper focuses on algorithms applicable to the complex forward functions that occur in nature, so methods designed specifically for linear or linearized models will not be covered in detail. We will however briefly show the connection between EIG and D-optimality, which is one of the so-called alphabetic design criteria (Box & Lucas 1959; Atkinson & Donev 1992), which have well-studied properties (Kiefer 1959). For a more detailed treatment of linear experimental design measures, the reader is referred to Atkinson & Donev (1992) for a general overview, and to Curtis (1999a) for an overview of most linear design measures used in geophysics.

2.1.1 Relationship between D-optimality and EIG

D-optimality is a widely used linear design criterion equal to the determinant of matrix $\boldsymbol{A}^T\boldsymbol{A}$, where $\boldsymbol{A}$ is the matrix relating $\boldsymbol{m}$ and $\boldsymbol{d}$. For Bayesian linear models it can be related to the EIG. If the prior pdf is Gaussian $\boldsymbol{m} = \mathcal{N}(m_0, \boldsymbol{\Sigma}_0)$, then for linear models the posterior pdf is proportional to $\left(\boldsymbol{A}^T\boldsymbol{A} + \boldsymbol{\Sigma}_0^{-1}\right)^{-1}$ and its information content is independent of the observed data. Making use of the analytical form of the entropy of a Gaussian, the EIG can be expressed as

$$\mathrm{EIG}(\boldsymbol{A}) = \tfrac{1}{2}\log|\boldsymbol{\Sigma}_0| - \tfrac{1}{2}\log\left|\boldsymbol{A}^T\boldsymbol{A} + \boldsymbol{\Sigma}_0^{-1}\right| + C \tag{15}$$

$$= \log\left|\boldsymbol{A}^T\boldsymbol{A}\right|\cdot C' + C'', \tag{16}$$

where C, C' and C'' are constants. Therefore, maximization of in the D-criterion results in the same design as would be obtained by maximizing the EIG.

2.1.2 Linearized models

Extensive work has been made to extend the alphabetic criteria developed for linear to nonlinear problems (Tsutakawa 1972; Chaloner & Verdinelli 1995). For this, a linearized version of eq. (1) is used

$$\boldsymbol{d} - \mathrm{F}\,(\boldsymbol{m}_0) = \boldsymbol{A}_{\boldsymbol{m}_0}(\boldsymbol{m} - \boldsymbol{m}_0) + \boldsymbol{\varepsilon}, \tag{17}$$

where $\boldsymbol{A}_{\boldsymbol{m}_0}$ is the Jacobian matrix of partial derivatives of $F(\boldsymbol{m})$ with respect to $\boldsymbol{m}$, evaluated at (linearized around) reference model $\boldsymbol{m}_0$. The physical relationships between $\boldsymbol{d}$ and $\boldsymbol{m}$are thus only approximate and are only accurate in some locality of $\boldsymbol{m}_0$. Consequently, the design that results from optimizing this relationship would no longer be independent of the reference model—which, in turn, should be free to vary according to the prior pdf. By taking the expectation over reference models, a Bayesian (linearized) version of the D-optimality criterion is given by Chaloner & Verdinelli (1995):

$$\mathbb{E}_{p(\boldsymbol{m})}\left[\log\det\tilde{\boldsymbol{L}}_{\boldsymbol{m}}\right], \tag{18}$$

where $\tilde{\boldsymbol{L}}_{\boldsymbol{m}} = \boldsymbol{A}_m^T\boldsymbol{A}_m$. The other alphabetic criteria can be extended to linearized problems using expectations in the same way.

The problems with this approach are apparent in Fig. 1: an averaging approach may work intuitively for nonlinear but approximately

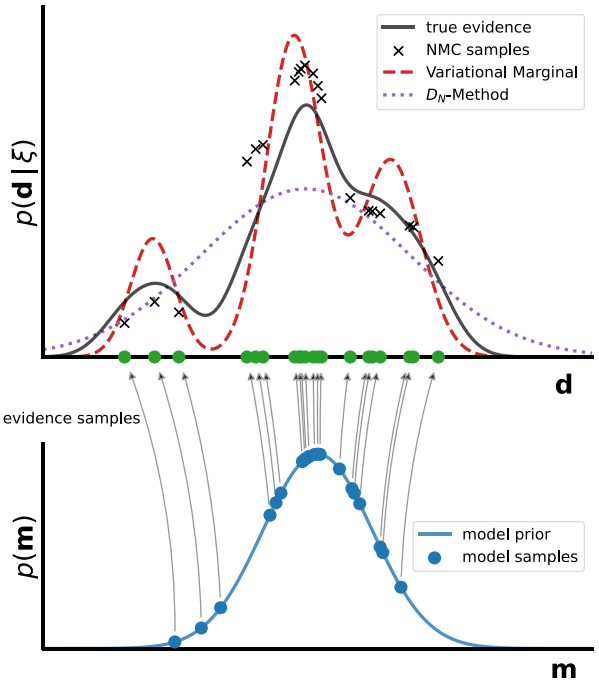

Figure 2. Illustration of EIG estimation methods in data space. Samples in model parameter space (blue dots) are generated from the prior pdf (blue curve). The likelihood (and the underlying forward model evaluation) can be used to generate samples from the evidence in data space (green dots) using the model space samples. Details of how $p(\boldsymbol{d} \mid \xi)$ is estimated using various methods (crosses, and black and red curves in the top graph) are described in Section 3. NMC denotes Nested Monte Carlo.

monotonic functions Fig. 1(b) because the resulting estimate of design quality is inversely related to the expected uncertainty of the solution approximated locally by linearized physics. However, as shown in Fig. 1(c), in the presence of multimodality, an approach based on derivatives breaks down, because the solution uncertainty approximated by the gradient of F only approximates one of the modes of the posterior pdf over parameter values that are consistent with the data (Winterfors & Curtis 2012). This is clear from eq. (18), which shows that around each prior value of the parameter space at which $\boldsymbol{L}_{\boldsymbol{m}}$ is evaluated, the posterior solution accounted for in eq. (16) consists of a single Gaussian. Therefore, even the linearized uncertainty estimate represented by the term in brackets in eq. (18) is substantially incorrect.

3 EIG ESTIMATION

The mathematical formulation of the EIG is straightforward. However, its evaluation is not, as neither $p(\boldsymbol{d} \mid \xi)$ nor $p(\boldsymbol{m} \mid \boldsymbol{d}, \xi)$ are typically known in closed form, and both $\boldsymbol{d}$ and $\boldsymbol{m}$ over which each is defined can have many dimensions. The following sections present approaches to computing an approximation of the EIG corresponding to any given design.

The basis for the following methods is to take samples $\boldsymbol{m}_i$ of the prior pdf (bottom of Fig. 2) and to project them into the data space by evaluating the forward function and taking samples $\boldsymbol{d}_i$ of the likelihoods $p(\boldsymbol{d}_i \mid \boldsymbol{m}_i, \xi)$. This procedure provides samples of the evidence (green dots in Fig. 2) but no probability values for those points in data space because the explicit evaluation of $p(\boldsymbol{d} \mid \xi)$crequires that we evaluate the normalizing integral in eq. (3) which is over the whole model space. The nested Monte Carlo (NMC), variational marginal and D_N methods below use different approaches to estimate $p(\boldsymbol{d}_i \mid \xi)$ which is needed to estimate the information described by the evidence.

3.1 Monte Carlo methods

The most straightforward and robust, but computationally expensive way of evaluating eq. (13) is to deploy a naive NMC approach as introduced by Ryan (2003) and Myung *et al.* (2013, further analysed in detail by Vincent & Rainforth 2017; Rainforth *et al.* 2017). The so-called NMC EIG estimator is defined as

$$\mathrm{EIG}_{\mathrm{NMC}}(\xi) = \frac{1}{N}\sum_{i=1}^{N} \log \frac{p\left(\boldsymbol{d}_i \mid \boldsymbol{m}_{i,0}, \xi\right)}{\frac{1}{M}\sum_{j=1}^{M} p\left(\boldsymbol{d}_i \mid \boldsymbol{m}_{i,j}, \xi\right)}, \tag{19}$$

where $\boldsymbol{m}_{i,j} \sim p(\boldsymbol{m})$ and $\boldsymbol{d}_i \sim p\left(\boldsymbol{d} \mid \left[\boldsymbol{m} = \boldsymbol{m}_{i,0}\right], \xi\right)$. The nested loop results in a slightly unusual sampling notation. An $N \times (M + 1)$ array of prior pdf samples $\boldsymbol{m}_{i,j}$ is generated from $p(\boldsymbol{m})$. The first row in this square is then referred to by $\boldsymbol{m}_{i,0}$, which means the outer loop only uses samples of this first row, while the inner loop uses a different column j (excluding the first element) of samples for each step of the outer loop.

Pointwise estimates of the evidence $p\left(\boldsymbol{d}_i \mid \xi\right)$ at N points $\boldsymbol{d}_i$ in data space are calculated using the average of the M likelihood functions $\frac{1}{M}\sum_{j=1}^{M} p\left(\boldsymbol{d}_i \mid \boldsymbol{m}_{i,j}, \xi\right)$ (inner loop of eq. 19), which estimate how likely it is that one would observe the datum $\boldsymbol{d}_i$. This results in a set of N points with an assigned probability in data space (grey crosses in Fig. 2). As these points are sampled from the prior, they can then be used to calculate an MC estimate of $\mathrm{I}[p(\boldsymbol{d} \mid \boldsymbol{\xi})]$ by $\frac{1}{N}\sum_{j=1}^{N} p\left(\boldsymbol{d}_i \mid \xi\right)$. The likelihood term $p\left(\boldsymbol{d}_i \mid \boldsymbol{m}_{i,0}, \xi\right)$ of eq. (19) takes the influence of the data noise into account which might change for different designs ξ (e.g. receiver positions close to noise sources may produce larger measurement uncertainties than those located in quiet areas).

The estimator $\mathrm{EIG}_{\mathrm{NMC}}$ has a computational cost of $T = \mathcal{O}(NM)$ with an RMSE (root-mean-square error) convergence rate of $\mathcal{O}(N^{-1/2}M^{-1})$ which means it is asymptotically optimal to set $M \propto \sqrt{N}$. The number of inner samples M controls the bias, while the number of outer samples N controls the variance of this quality estimate (Huan & Marzouk 2013). The computational cost of the NMC approach can be reduced considerably if the same samples of the inner loop (M) are reused for each iteration of the outer loop as demonstrated by Huan & Marzouk (2013), Qiang *et al.* (2022) and Zhang *et al.* (2015), resulting in a computational cost of $T = \mathcal{O}(N + M)$ total samples. This reuse of samples increases the bias in the EIG estimate, but if the bias is stationary this would not affect the subsequent design optimization. To our knowledge, no useful bounds on the size of the bias nor practically implementable conditions that guarantee stationarity for particular problems are available. The $\mathrm{EIG}_{\mathrm{NMC}}$ estimate is an upper bound and will, therefore, always be larger than the true EIG (Foster *et al.* 2019a).

3.2 Maximum entropy method

If the likelihood is independent of the design, $p(\boldsymbol{d} \mid \boldsymbol{m}, \xi) = p(\boldsymbol{d} \mid \boldsymbol{m})$, meaning that the measurement uncertainty on each datum does not vary with the design (e.g. with receiver location), eq. (13) is equivalent to

$$\mathrm{EIG}_{\mathrm{ME}}(\xi) = -\mathrm{I}[p(\boldsymbol{d} \mid \xi)] + C \tag{20}$$

and maximizing the EIG is equivalent to maximizing the entropy (negative information) of the evidence, resulting in so-called maximum entropy design (Shewry & Wynn 1987). This only slightly lowers the cost compared to the NMC method, since the main cost in evaluating eq. (13) is the estimation of $\mathrm{I}[p(\boldsymbol{d}\,|\,\xi)]$ which can be taken out of the expectation due to its independence of $\boldsymbol{m}$. The later cost is the same in both the NMC and the maximum entropy methods. $\mathrm{EIG}_{\mathrm{ME}}$ can, therefore, be estimated either by using a similar nested Monte Carlo loop as for NMC or using specific methods to calculate the entropy of a set of samples, such as *k-d* partitioning (Stowell & Plumbley 2009; Bloem *et al.* 2020). The restriction that the uncertainty on each datum does not vary is an unrealistic assumption in many cases, for example, if different receivers have different noise levels, or if noise depends on the spatial location of receivers.

Especially in cases with a large number of data dimensions (e.g. observations made using many receivers), both the NMC and the maximum entropy estimator, while consistent, will converge prohibitively slowly to be practical for most geophysical applications. Perhaps luckily then, experimental design is often particularly effective for experiments with restricted experimental resources which may have fewer observations and hence data space dimensions; MC methods are then useful in some cases. MC methods are also necessary for benchmarking algorithms that employ different approximations introduced in the following sections.

3.3 Variational methods

The bottleneck in evaluating expressions (9) and (13) occurs in the estimation of $p(\boldsymbol{m}\,|\,\boldsymbol{d},\xi)$ and $p(\boldsymbol{d}\,|\,\xi)$, respectively. The main inefficiency in the NMC method arises because the integrand in eq. (19) is estimated separately for each $\boldsymbol{d}$. Instead of evaluating this integrand directly, Foster *et al.* (2019a) proposed to find a *variational* (closed form, or otherwise analytic) approximation to either $p(\boldsymbol{m}\,|\,\boldsymbol{d},\xi)$ or $p(\boldsymbol{d}\,|\,\xi)$.

Suppose the construction of this functional approximation requires M samples. In that case, the total computational cost is on the order of $\mathcal{O}(N+M)$, which may be a substantial reduction compared to the NMC method. We therefore now introduce variational approaches to EIGestimation.

3.3.1 *Variational marginal method*

The variational marginal method operates in data space by finding a variational estimator $q_m(\boldsymbol{d}\,|\,\xi)$ that approximates the evidence $p(\boldsymbol{d}\,|\,\xi)$. The evidence is, in fact, the data space marginal posterior pdf, hence the name of this design method. Instead of evaluating the marginal density $p(\boldsymbol{d}\,|\,\xi)$ for each of the N data samples d_n, the idea is to find a variational functional emulator q_m of $p(\boldsymbol{d}\,|\,\xi)$ using M samples. Once q_m is available, using eqs (13) and (19), the EIG can be approximated as

$$\mathrm{EIG}_{\mathrm{marg}}(\xi) = \frac{1}{N}\sum_{n=1}^{N}\log\frac{p\left(\boldsymbol{d}_n\,|\,\boldsymbol{m}_n,\xi\right)}{q_m\left(\boldsymbol{d}_n\,|\,\xi\right)}, \tag{21}$$

where $\boldsymbol{m}_i \sim p(\boldsymbol{m})$ and $\boldsymbol{d}_i \sim p\left(\boldsymbol{d}|\,[\boldsymbol{m}=\boldsymbol{m}_i]\,,\xi\right)$.

The variational approximator $q_m(\boldsymbol{d}\,|\,\xi)$ is found by first introducing a variational family $q_m(\boldsymbol{d}\,|\,\xi,\boldsymbol{\phi})$ (e.g. the family of multivariate Gaussians) with parameters $\boldsymbol{\phi}$ (e.g. describing mean and covariance) that parametrize possible forms of q_m. The key is to find parameters $\boldsymbol{\phi}$ that provide the best possible approximation q_m to $p(\boldsymbol{d}\,|\,\xi)$ on average for any $\boldsymbol{d}$. It can be shown (see the appendix of Foster *et al.* 2019a) that this formulation yields an upper bound to the true EIG, and $\mathrm{EIG}_{\mathrm{marg}} = \mathrm{EIG}$ if and only if $q_m(\boldsymbol{d}\,|\,\xi) = p(\boldsymbol{d}\,|\,\xi)$. The optimal parameters $\boldsymbol{\phi}^*$ can therefore be found efficiently using stochastic gradient descent (SGD)[1] (Robbins & Monro 1951) to solve the following optimization problem:

$$\boldsymbol{\phi}^* = \arg\min_{\boldsymbol{\phi}}\left\{-\mathbb{E}_{p(\boldsymbol{d},\boldsymbol{m}\,|\,\xi)}\left\{\log q_m(\boldsymbol{d}\,|\,\xi,\boldsymbol{\phi})\right\} + \underbrace{\mathrm{I}\left[p(\boldsymbol{d}\,|\,\boldsymbol{m},\xi)\right]}_{\text{constant}}\right\} \tag{22}$$

This means that we want to find the variational family for which the samples of evidence have the highest expected probability of being observed. The lower the value of the term in eq. (22), the closer is the variational marginal estimator is to the true EIG, since it is an upper bound on the EIG. The tightness of the bound is determined by how well $q_m(\boldsymbol{d}\,|\,\xi,\boldsymbol{\phi}^*)$ represents the true evidence.

The variational family for the variational marginal method used in this study is a Gaussian mixture model (GMM)—a sum of Gaussians (Bishop 2006). The red curve in Fig. 2 illustrates an example of a variational marginal pdf. The density of the evidence samples (green) is used to find a function, here a sum of Gaussians, which can be evaluated straightforwardly to approximate the probability of observing points in data space. These probabilities are then used to calculate the EIG. We have also successfully applied normalizing flows (Tabak & Turner 2013; Dinh *et al.* 2014; Rezende & Mohamed 2015; Durkan *et al.* 2019; Zhao *et al.* 2020) to construct the above function, but they are omitted in this paper for brevity.

3.3.2 *Variational posterior method*

Finding an approximator to $p(\boldsymbol{d}\,|\,\xi)$as above might be disadvantageous if the number of data dimensions is high and the number of model dimensions is low (e.g. when designing a seismic source location experiment with many receivers recording traveltime measurements). Eq. (9) allows us instead to calculate the EIGin the model space by finding an approximator to $p(\boldsymbol{m}\,|\,\boldsymbol{d},\xi)$. This is achieved by finding a variational function $q_p(\boldsymbol{m}\,|\,\boldsymbol{d},\xi)$ to represent the posterior pdf given an observed datum $\boldsymbol{d}$ measured under design ξ. The EIG can then be approximated by the variational posterior method as

$$\mathrm{EIG}_{\mathrm{post}}(\xi) = \frac{1}{N}\sum_{n=1}^{N}\log\frac{q_p\left(\boldsymbol{m}_n\,|\,\boldsymbol{d}_n,\xi\right)}{p\left(\boldsymbol{m}_n\right)}, \tag{23}$$

where $\boldsymbol{m}_i \sim p(\boldsymbol{m})$ and $\boldsymbol{d}_i \sim p\left(\boldsymbol{d}|\boldsymbol{m}=\boldsymbol{m}_i,\xi\right)$. To evaluate this MC estimator, it is necessary to find the function $q_p(\boldsymbol{m}\,|\,\boldsymbol{d},\xi)$. For this a family of variational distributions $q_p(\boldsymbol{m}\,|\,\boldsymbol{d},\xi,\boldsymbol{\phi})$ parametrized by $\boldsymbol{\phi}$ is introduced. To find $\boldsymbol{\phi}$ and therefore a function that is close to $p(\boldsymbol{m}\,|\,\boldsymbol{d},\xi)$, we can make use of the fact that $\mathrm{EIG}_{\mathrm{post}}(\xi)$ is a lower bound on the true EIG, which is tight strictly if and only if $q_p(\boldsymbol{m}\,|\,\boldsymbol{d},\xi,\boldsymbol{\phi}) = q(\boldsymbol{m}\,|\,\boldsymbol{d},\xi)$. Maximizing this lower bound to find the optimal choice $\boldsymbol{\phi}^*$ is equivalent to evaluating

$$\boldsymbol{\phi}^* = \arg\max_{\boldsymbol{\phi}}\left\{\mathbb{E}_{p(\boldsymbol{d},\boldsymbol{m}\,|\,\xi)}\left\{\log q_p(\boldsymbol{m}\,|\,\boldsymbol{d},\xi,\boldsymbol{\phi})\right\} - \underbrace{\mathrm{I}\left[p(\boldsymbol{m})\right]}_{\text{constant}}\right\} \tag{24}$$

which maximizes the expected probability of observing the prior sample used to generate a data sample. The parameters $\boldsymbol{\phi}$ can be optimized using stochastic gradient ascent (SGA). Then, $q_p(\boldsymbol{m}\,|\,\boldsymbol{d},\xi,\boldsymbol{\phi})$

[1]Here referred to as either SGD for minimization or SGA for maximization. However, both are essentially the same algorithm where the sign of the optimization metric is flipped.

results in a different pdf for each datum $\boldsymbol{d}$(orange pdfs on the right of Fig. 3). The maximization in eq. (24) is equivalent to minimizing the KL divergence between $q_m(\boldsymbol{m} \mid \boldsymbol{d}, \xi, \boldsymbol{\phi})$ and $p(\boldsymbol{m} \mid \boldsymbol{d}, \xi)$. A considerable advantage of this approach is that it maximizes the EIG in the model parameter space, making it suited to design experiments with a high-dimensional data space but a lower dimensional parameter space.

In this study, the variational family for the approximation of the variational posterior pdf is a Gaussian mixture density network (MDN) (Bishop 1994; Meier *et al.* 2009). We have also successfully applied conditional normalizing flows (Tabak & Turner 2013; Dinh *et al.* 2014; Rezende & Mohamed 2015; Durkan *et al.* 2019; Zhao *et al.* 2020) but these are omitted in this paper for brevity.

For both variational estimators EIG_{marg} and EIG_{post}, the quality of the final result depends on how accurately q_m or q_p represents the true evidence or posterior. This depends on the flexibility (expressiveness) of the parametrization of the two functional distributions, and on how well the gradient descent optimization has converged.

3.4 D_N method

While variational methods can substantially reduce the computational cost of calculating the EIG, their optimization using SGD is still a significant computational expense. If the evidence can be approximated adequately by a multivariate Gaussian distribution $\mathcal{N}(\mu, \boldsymbol{\Sigma}_{\text{evidence}})$ with mean μ and covariance matrix $\boldsymbol{\Sigma}_{\text{evidence}}$, its information content can be calculated as a function of its covariance matrix by

$$\text{I}\left[\mathcal{N}(\mu, \boldsymbol{\Sigma}_{\text{evidence}})\right] = -\frac{1}{k}(1 + \ln(2\pi)) - \frac{1}{2}\ln\left(|\boldsymbol{\Sigma}_{\text{evidence}}|\right), \tag{25}$$

where k is the dimensionality of the data space. Under the assumption of Gaussian data noise with covariance $\boldsymbol{\Sigma}_{\text{data}}$, the EIG can then be expressed as

$$\text{EIG}_{D_N} = \ln|\boldsymbol{\Sigma}_{\text{evidence}}| - \ln|\boldsymbol{\Sigma}_{\text{data}}| + C \tag{26}$$

which is the EIG of eq. (13) in which both terms have been approximated using a Gaussian distribution, and where the constant terms have been summarized in the constant C. This defines the so-called D_N method, equivalent to the variational marginal method with the multivariate Gaussian variational family. It follows that it can be extended to non-Gaussian noise, taking only the numerator in eq. (21) and using it to replace $\ln|(\boldsymbol{\Sigma}_{\text{data}})|$.

This estimator was first introduced by Coles & Curtis (2011a) and applied by Rawlinson *et al.* (2012), Coles *et al.* (2013) and Bloem *et al.* (2020). While eq. (26) is only valid for a Gaussian distribution of the evidence, the D_N criterion remains useful in many applications because maximizing the covariance increases the spread in data space of data points corresponding to different models. Intuitively, the farther data from different models are spread apart, the easier it is to distinguish between models in the presence of data noise.

Covariance-based measures can fail, most notably in the case of multimodality in the evidence, where the distance between modes does not necessarily influence the information but where a larger distance would lead to a higher covariance. Despite this limitation, the D_N method appears to be essential in practice due to the efficiency of its evaluation and the small number of samples necessary to obtain a stable estimate of the EIG. Fig. 2 shows how the D_N method compares to the NMC and the variational marginal method by showing the Gaussian (green) with the same mean and covariance as the evidence samples (green). It is not necessary to use this pdf to obtain sample probabilities as the information of a Gaussian is known in closed form (eq. 25).

3.5 Other methods

This section has mainly focused on methods that have been used previously in geophysical applications, with the addition of with variational methods which have been developed in reasonable generality only recently. There are, of course, many other methods. A summary of methods up to 2016 is available in the review of Ryan *et al.* (2016), since when, the field has progressed substantially, most notably through bounds on mutual information and methods that use these (e.g. Kleinegesse & Gutmann 2021; Guo *et al.* 2021). There are other more recent approaches which may, in the future, also advance EIG estimation for geophysical applications (e.g. Beck *et al.* 2018; Wu *et al.* 2020; Goda *et al.* 2020; Carlon *et al.* 2020; Alexanderian 2021; Long 2022; Englezou *et al.* 2022). All of these are beyond our scope for this paper.

4 EIG OPTIMIZATION

The maximization in eq. (14) requires an optimization algorithm to be chosen independently of the method of EIG estimation, and the choice determines how often the EIG needs to be evaluated. Each evaluation can take a substantial amount of time to compute, so the choice of algorithm can dramatically change the overall computation required for experimental design. It is also possible to extend the algorithms below to consider the logistical and financial costs of deployment. A significant difficulty with implementing that approach is the need to to define how much time or money one unit of EIG is worth.

4.1 Global optimization

Ideally, the optimization algorithm used to solve eq. (14) should consider all possible designs in Σ and choose the one that results in the highest EIG. This could be achieved most straightforwardly for a continuous design space by sampling Σ in a sufficiently dense regular grid and calculating the EIG for every gridpoint. However, this is only possible for low-dimensional design spaces, because the number of EIGevaluations scales as $(n_{\text{grid}})^D$, where n_{grid} is the number of gridpoints per dimension and D is the number of dimensions. To avoid this exponential scaling, several global optimization algorithms such as the genetic algorithm (Holland 1992) and simulated annealing (Bohachevsky *et al.* 1986) have been used to design geophysical problems (e.g. Barth & Wunsch 1990; Barth 1992; Curtis & Snieder 1997; Maurer & Boerner 1998). Their guarantee to always converge towards the global optimum given sufficient iterations comes at the cost of many EIG evaluations, albeit fewer than a grid search approach requires. This, again, makes these algorithms infeasible for all but small-scale design problems. Other popular global approaches for design optimization in general are Bayesian optimization (Jones *et al.* 1998; Kleinegesse & Gutmann 2018; Foster *et al.* 2019a) and McMC methods (Amzal *et al.* 2006; Jones *et al.* 1998).

4.2 Greedy optimization

The computational infeasibility of global optimization has led to the use of greedy algorithms that make locally optimal choices but still lead to designs that substantially improve the EIG compared to

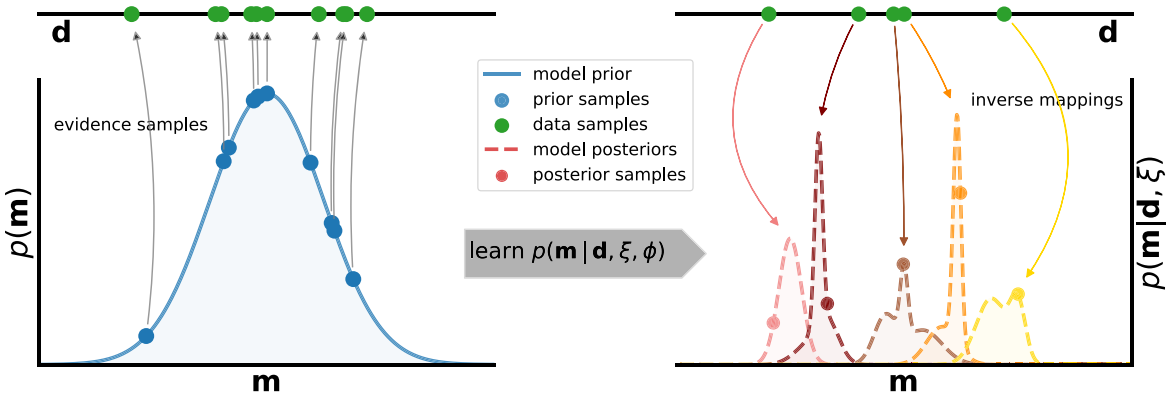

Figure 3. Illustrates the variational posterior method to estimate the EIG in model parameter space (left). The prior (blue curve) is used to generate samples in model parameter space (blue dots). The likelihood (and the underlying forward model evaluation) can be used to generate evidence samples in data space (green dots) using each model space sample. These pairs of model parameters and data samples can be used to find a variational approximating functional $q_p(\boldsymbol{m} \mid \boldsymbol{d}, \xi, \boldsymbol{\phi})$ to $p(\boldsymbol{m} \mid \boldsymbol{d}, \xi)$. Thereafter, a different set of data samples is generated in the same way (right). The corresponding posterior pdf for every such sample can be estimated cheaply using mapping q_p.

random designs. The most popular of those algorithms are sequential design optimization algorithms, in which all but one dimension in the design space is fixed, and a 1-D optimization is solved in the remaining dimension. Iterating this process through all dimensions causes the design to converge to a (locally) optimal design.

Sequential optimization can be done in different ways but the most popular and computationally cheapest is sequential construction (Curtis *et al.* 2004; Stummer *et al.* 2004) where the optimal design in a 1-D design space is selected first (e.g. a first receiver is placed at an optimal location). This locally optimal choice is then fixed, and the best choice in a second 1-D design space (a second receiver location) is selected. This process can be iterated until the desired number of design dimensions (number of receivers) is reached. Alternatives to this approach are sequential destruction (Curtis 2004b) and the sequential exchange algorithm of Mitchell (1974). Sequential design optimization algorithms are substantially cheaper than global optimization. However, in practice, they still usually result in designs that deliver an EIG close to the global optimum in practice (Coles & Curtis 2011a; Guest & Curtis 2009). A detailed mathematical treatment of sequential design optimization algorithms is given in Jagalur-Mohan & Marzouk (2021).

5 GEOPHYSICAL APPLICATIONS OF BAYESIAN OPTIMAL DESIGN METHODS

While linear and linearized OED has been used and studied extensively for geophysical applications, OED for fully nonlinear forward models is still not widely used. The first step in this direction was to use the number of modes in the misfit of linearized OED as a design criterion (Curtis & Spencer 1999; Curtis 2004a). This method was designed to alleviate some problems of linearized designs methods, but it has yet to be applied in a practical problem.

The EIG was first used as a criterion by van Den Berg *et al.* (2003, 2005). By adopting the maximum entropy sampling method they used the entropy of the evidence as a proxy for the EIG. This approach has subsequently been refined to a sequential optimization (Guest & Curtis 2009) and applied to design reflection seismic amplitude-versus-offset experiments with complex subsurface prior information (Guest & Curtis 2010). The NMC formulation was first used in geophysics by Coles & Prange (2012) and has recently been applied by Qiang *et al.* (2022) and combined with physics-informed neural networks by Wu *et al.* (2022).

The Laplace method (Tierney & Kadane 1986) allows the EIG to be calculated in the model space using the Hessian matrix of the forward model (second-order derivatives with respect to the model parameters), under the assumption that the posterior pdf is multivariate Gaussian (Long *et al.* 2013, with extensions allowing for multimodality, Long 2022). It was used by Long *et al.* (2015) for the optimal design of a full-waveform seismic source moment tensor inversion. However, the use of the determinants of Hessian matrices makes this method similar to linearized Bayesian experimental design methods.

The computationally efficient D_n method first used by Coles & Curtis (2011b) has subsequently been applied and derived using alternative approaches (Rawlinson *et al.* 2012; Coles *et al.* 2013; Bloem *et al.* 2020). In addition to the EIG, other studies on nonlinear design have used bifocal measures (Winterfors & Curtis 2008, 2012), or problem-specific measures which are not a function of the posterior pdf (López-Comino *et al.* 2017; De Landro *et al.* 2020; Ferrolino *et al.* 2020; Fichtner & Hofstede 2022; Dasgupta *et al.* 2021).

More detailed reviews of design methods and applications in geophysics are given in Curtis (2004b, a) and Maurer *et al.* (2010). For a general review of design algorithms, the reader is referred to Ryan *et al.* (2016).

6 APPLICATIONS

We now demonstrate the algorithms introduced above, and explore their relative merits in two common geophysical problems. Our aim here is to be educational rather than to provide a comprehensive study of these optimal design problems.

6.1 Seismic source location

In a seismic source location problem, the aim is to determine the location of a seismic event, such as an earthquake, using the first arrival times of *P* and *S* waves or other seismic phases. To achieve this, the subsurface structure needs to be known at least approximately.

The setup used in this example (depicted in Fig. 4) employs the elastic extension of the 2-D Marmousi model (Martin *et al.* 2006),

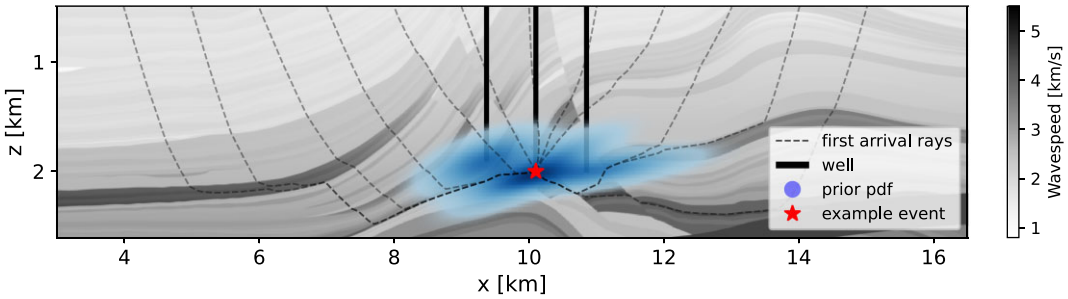

Figure 4. Seismic source location problem with three wells (thick black lines), prior pdf comprising a sum of three spatially correlated Gaussians (blue shading, darker being more probable) and seismic first arrival rays (thin black lines) originating from one of the prior samples (red star) and terminating at regularly spaced points on the ground surface.

a suitably complex subsurface structure. Three wells with depths of around 2 km are placed near the middle of the structure, and are assumed to be involved in some intervention in the subsurface which induces seismicity near their terminations. As the locations of the wells are known, the model parameter prior pdf is constrained in space, and we represent our assumed prior information as a sum of three spatially correlated Gaussians (Fig. 4).

The strong prior information combined with the 2-D nature of this synthetic example, allows a low number of receivers or even a single one to achieve a good posterior estimate of the true location. Under the assumption of a constant v_p/v_s ratio and using arrival time differences t_{diff} between P and S waves, the source time can be excluded from the source location problem similarly to Bloem *et al.* (2020), and hence can be excluded from the design process. Therefore, the model parameter space consists of the set of 2-D vectors of horizontal and vertical locations.

The fast marching method (Sethian 1996) using the openly available implementation of White *et al.* (2020) was used to calculate the seismic wave arrival times. Bloem *et al.* (2020) used the same traveltime inversion to test different linear and nonlinear experimental design algorithms, albeit with other slowness models and prior pdfs. The likelihood is modelled using a Gaussian distribution with a mean corresponding to the calculated traveltime and a standard deviation of 0.02 s as a baseline for the measurement uncertainty.

The optimal design problem is to find receivers placed on the seabed ($z = 0.5$ km) at horizontal offsets from 3.1 to 16.4 km, which give the highest EIG. The EIG of 200 possible single-receiver locations was calculated along the seabed using each of the methods described in Section 3. The sequential construction method was used to optimize multireceiver designs because of its computational efficiency and the possibility of visualizing the design process steps as receivers are added.

In most geophysical applications, the generation of data corresponding to model parameter samples is the main computational cost. We therefore compare the performance of the different methods for a two-receiver network as a function of the number of samples used (see Fig. 5). The cost of evaluating the different estimators using those samples is compared thereafter.

The variational marginal method has been implemented using a GMM with 10 Gaussians, each with a full covariance matrix. For the variational posterior method, the variational family is the output of a Gaussian MDN with a three-layer neural network consisting of 60 nodes in each layer, defining 20 Gaussians with full covariance matrices as output.

The NMC method has been implemented using new samples for each step of the outer loop, or reusing the M samples of the inner loop of eq. (19, the latter is referred to as NMC_{re}). This results in a cost of $N \times M$ or $N + M$ samples, respectively.

To put the results into context, we calculated the EIG for 1000 random designs, and for a design with receivers at 6.4 and 13.1 km, a heuristically designed network, using the NMC_{re} method (using 1×10^5 samples). All optimal design methods converge to an EIG value that outperforms almost all of the random designs as well as the heuristic design. Due to the restrictive Gaussian assumption, the D_N method results in an EIG value that is slightly lower than the optimum achieved by the other design methods. While not converging to the optimal possible design, the D_N method design substantially outperforms the median and equispaced design with as few as only 10 samples. All other methods are practically inapplicable with so few samples. Both variational methods perform similarly to the NMC_{re} method, converging towards the maximum EIG when using more than around 1×10^4 samples.

Tests using a four-receiver design lead to similar conclusions, with the exception that the variational posterior method performs better and is comparable to the NMC_{re} method, demonstrating the beneficial scaling of this method with data dimensionality. For three and four receivers, all methods converge to designs with slightly lower EIG compared to the best random design, an effect of using the sequential construction method which can converge to local maxima.

The final designs for a four-receiver network are shown in Fig. 6. For all but the D_N method, the resulting network is nearly identical, and the deviation of the D_N-derived network leads to the slight negative offset in EIG discussed above.

For now, we have assumed a constant noise level across all receiver locations, or in other words, we have assumed the noise to be independent of the design. This is rarely the case in real-world scenarios. Fig. 7 shows the effect of design-dependent noise. Here, the experimental design process uses an artificial noise function that mimics the effects of geometrical spreading and of anthropogenic noise around the wells. The best designs found change substantially, moving receivers towards regions of low noise, leading to a higher agreement in the optimal design results from different methods. The design dependence of the likelihood also stabilized the design optimization process, and indeed the result in this case might have been designed approximately using intuition alone. This demonstrates that where intuition can be applied, the design methods herein conform to expectations. And of course, a quantitative approach is necessary for more complex noise models and realistic 3-D environments in which intuition fails.

While the number of forward evaluations is often the bottleneck for the feasibility of experimental design algorithms, the cost of the EIG estimator itself is not insignificant. This is especially important

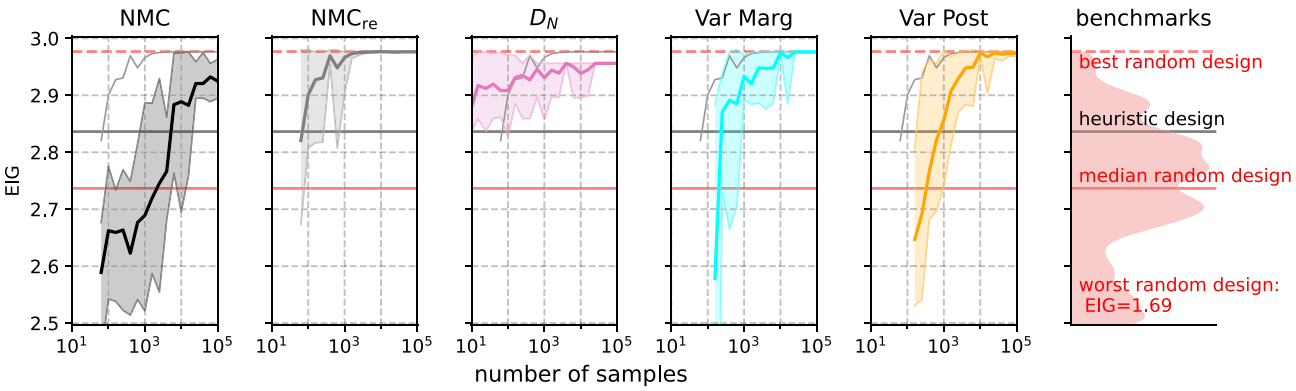

Figure 5. Benchmark results from different EIG estimation methods using sequential construction, showing the EIG for a two-receiver network for seismic source location as a function of the number of samples for which the forward function is evaluated. The solid line indicates the mean of 10 independent runs, while the shaded area indicates the respective minimum and maximum values of those runs. The mean curve of the NMC$_{re}$ results are shown in every panel to serve as a benchmark for comparison. On the right, a smoothed histogram of the EIG for 1000 randomly selected designs and for a heuristic design with receivers at 6.4 and 13.1 km is shown to put the results into context. Details on the setup and methods are in the main text.

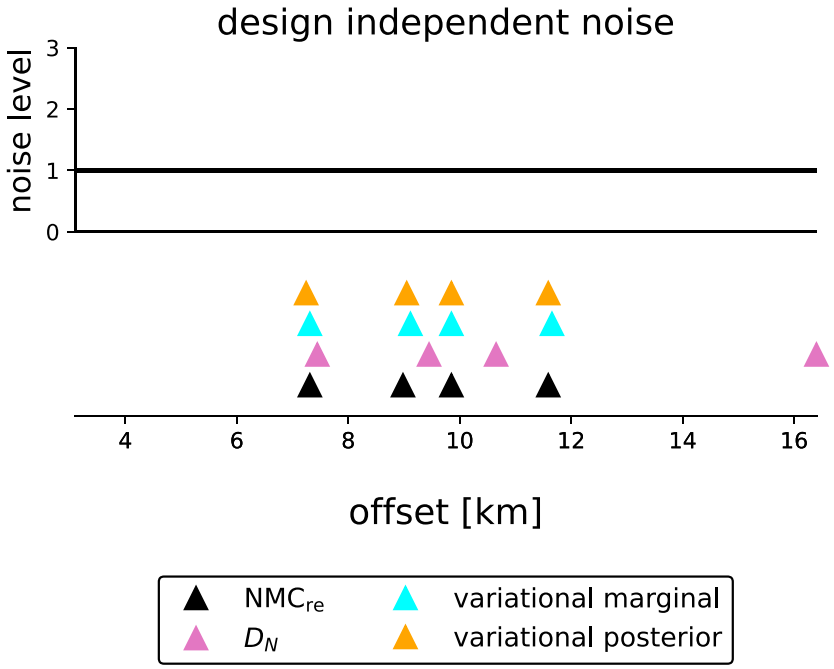

Figure 6. Optimized four-receiver networks using different EIG estimation methods and the sequential construction method for a constant noise level. The upper part shows the noise level as a function of receiver position in multiples of the base (constant) noise level, with a standard deviation of 0.02 s.

in problems where physical insights can be used to generate a large number of data samples highly efficiently. In the case of the source location example, this is possible by making use of the reciprocity of the eikonal equation and the full traveltime field generated by the fast marching method that is used to solve the eikonal equation: by treating receiver locations as eikonal source fields, many data and model parameter pairs can be generated cheaply once the traveltime field for a given receiver location has been calculated. After the pre-computed traveltimes are stored for each possible receiver position, the main bottleneck in design optimization is the cost of the EIG estimator.

The cost of the EIG calculation, excluding the cost of the forward evaluations, is shown in Fig. 8. All six methods become increasingly more expensive as more samples are used. The D_N and NMC methods are substantially cheaper than the other methods. The variational methods have a relatively large overhead since they require neural networks to be trained to represent the variational approximators. Since the training of the variational estimators is the main cost in using them for EIG estimation, they scale very well for a high number of samples if the maximum number of SGD steps is capped (here 10 000 for the variational methods). The point where this threshold is reached can be seen as the sharp change in slope for those methods in Fig. 8. The difference in computation time between the NMC with and without reused samples is twofold. First, if N_T is the total number of samples used, the standard NMC method requires N_T likelihood evaluations, while the NMC with reused samples requires $0.25 \times N_T^2$. This quadratic scaling is evident in Fig. 8. A second difference is that a numerically slightly less efficient method of computing the NMC with reused samples was used since otherwise the memory usage for the computation of the EIG for 1×10^5 samples would be more than 20 GB for storing the necessary likelihood values alone.

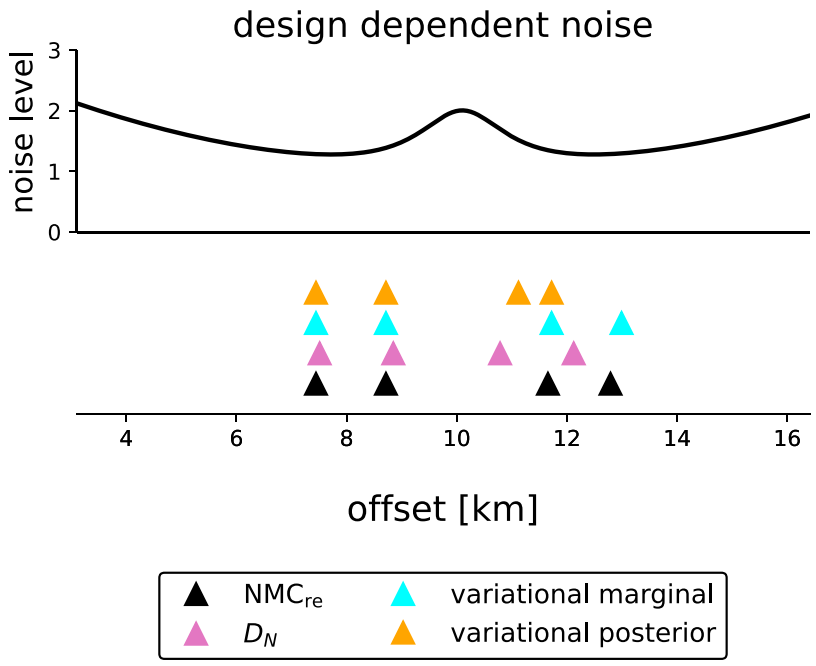

Figure 7. Optimized four-receiver networks using different EIG estimation methods and the sequential construction method for a design-dependent noise level. The upper part shows the noise level as a function of receiver position in multiples of the base (constant) noise level, with a standard deviation of 0.02 s.

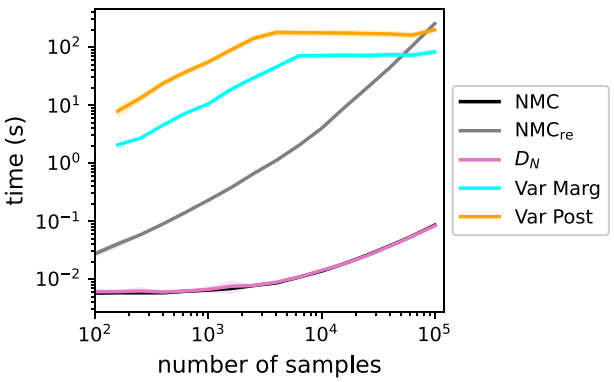

Figure 8. Benchmark results from different EIG estimation methods using sequential construction. Shows the time to calculate the EIG for a two-receiver network for seismic source location, excluding the time to generate data samples (the time spent evaluating the forward function). The solid line indicates the mean of 10 independent benchmark runs, while the shaded area indicates the respective minimum and maximum values. Details on the setup and methods are in the main text.

6.1.1 *Drawbacks of the D_N method*

The above results established the D_N method as a cheap and robust method that can produce near-optimal designs. This is also confirmed in the later Section 6.2 for an amplitude versus offset (AVO) design problem. In both cases, the assumption of a Gaussian form for the evidence only leads to a slight negative offset. Yet, we investigate conditions under which this assumption nevertheless leads the D_N method to produce a far from optimal design. We set up a scenario (see bottom of Fig. 9) with two separate areas of possible seismicity, with an inclined low-velocity layer above. For example, the cause of the seismicity might be related to drilling or injection for geothermal power production (e.g. Maurer *et al.* 2020).

In this scenario, the two-receiver network designed using the D_N method is quite different (Fig. 9) and performs substantially worse (Fig. 10) than one designed using either the NMC (reused M) or the variational marginal method (using a mixture of Gaussians). While the D_N design still performs better than the median of 1000 random designs, around a third of the randomly selected designs perform better, and it performs worse than a standard design in which the receivers are equispaced at 5000 and 10 000 m. The distribution of traveltimes as a function of horizontal receiver position illustrates why the D_N method performs so poorly in this scenario (top of Fig. 9). As long as the two bands of traveltimes due to the two areas of seismicity overlap with extremely low probability, their information content is independent of the distance between them. However, if the two bands are further apart, the standard deviation of the evidence increases, and therefore, the information content of the Gaussian used to approximate the evidence in the D_N method decreases. Due to the nature of this scenario, the distance between the two bands varies substantially across the different receiver placements. Therefore, the D_N method is, to first order, influenced by the distance between the traveltime bands and not their respective spread, whereas the latter property governs the information content of the evidence. In applications where substantial multimodality may occur *a priori* in data space, this deficiency trades off with the computational efficiency of the method.

6.1.2 *Source location interrogation*

As mentioned in Section 3.3.2, solving the EIG estimation in model parameter space allows one to design interrogation problems (Arnold & Curtis 2018, for more information see Appendix A2). Instead of maximizing the information of $p(\boldsymbol{m} \mid \boldsymbol{d}, \xi)$, the goal in such problems is to maximise the information in a target space $\mathbb{T}$, which is used to answer a specific question or set of questions Q.

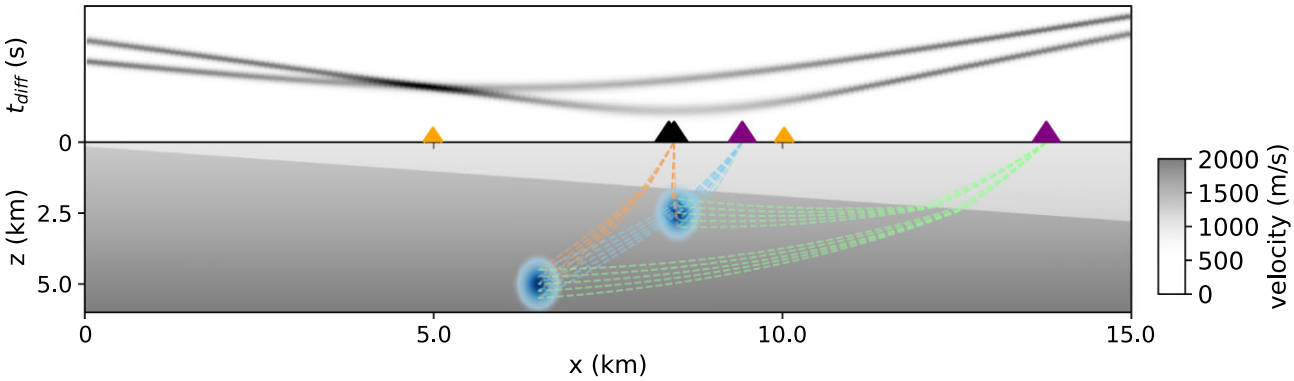

Figure 9. Setup for demonstrating shortcomings of the D_N method. The lower part shows the subsurface P-wave velocity model, including the prior pdf for event locations (blue contours). The dotted lines give examples of first arrival rays originating from 10 representative prior locations to three receiver locations. The optimal design calculated using sequential construction with the NMC method (black triangles) and the D_N method (purple triangles) to estimate the EIG are shown, as well as a heuristic equispaced design (orange triangles). The top part shows a histogram of traveltimes generated by forward modelling samples from the prior pdf for each possible receiver location. Darker grey tones indicate that more samples produce the same traveltime.

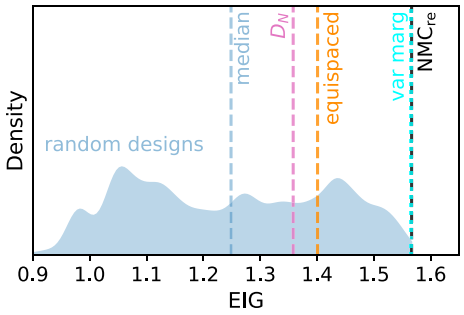

Figure 10. Comparison of the EIG values for two-receiver networks derived using sequential construction with the NMC (reused inner samples), the variational marginal, and the D_N methods to estimate the EIG, as well as an equispaced heuristic design (see Fig. 9). The blue smoothed histogram represents the EIG values of 1000 randomly selected designs.

For this, a target function $T(\boldsymbol{m} \mid Q)$, which maps samples from $\mathbb{M}$ to $\mathbb{T}$, needs to be defined.

We apply this concept to search for the optimal receiver placements to constrain only the epicentre or only the depth of the seismic source, respectively. In this case, T selects and returns one of the two coordinates, making the target space 1-D ($\mathbb{R}^1$).

A slightly simpler setup than the one above is used to demonstrate the interrogation design process; see Fig. 11. The same constant noise level as in the previous test was used. The only thing that changed is the prior, now a single multivariate Gaussian distribution, placed in the relatively shallow subsurface.

Fig. 12 compares example posterior probability distributions computed for three designs, each optimized for a different interrogation question. The EIG curves (calculated using a grid of 200 possible receiver locations) show that both the hypocentre and the epicentre design problems favour nearly identical designs with a low EIG over the source area and higher EIG at a greater distance from the expected sources. If the vertical location (source depth) is to be constrained independently of any other parameter, the resulting EIG curve and optimal receiver position are entirely different: here receiver positions over the source area are preferred, with positions at larger distances providing nearly no information. The resulting posterior probability functions also clearly show the effect of different receiver positions. The area of high probability is aligned vertically when we focus on the hypo- or epicentre, while it is aligned almost horizontally when we seek the source depth. An information tradeoff introduced by a focus on different questions is also evident since the posterior pdf for the source depth is more spread out and, therefore, is less informative than the hypocentre localization design, which corresponds to classical experimental design.

6.2 Amplitude versus offset

A well-studied problem in nonlinear geophysical optimal design is AVO inversion for seismic velocity contrasts (van Den Berg *et al.* 2003, 2005; Guest & Curtis 2009, 2010, 2011). We use this example to compare methods and discuss contrasts with results in the source location problem, and to illustrate stochastic one-step design optimization and a more realistic interrogation problem.

The objective of AVO is to determine the seismic properties of a buried layer by observing the change in seismic amplitude as a function of offset from the source. Fig. 13 depicts the setup which is identical to the one in Guest & Curtis (2009). A Gaussian prior pdf with a mean of 3750 m s^{-1} and a standard deviation of 300 m s^{-1} is assigned to the P-wave velocity α_2 in the layer of interest. Further, we assume a so-called Poisson medium in which $\beta = c\alpha$, where $c = 1/\sqrt{3}$, no density contrast between the layers, and assign a P-wave velocity α_1 of 2750 m s^{-1} and thickness of 500 m to the upper layer. The top of Fig. 13 shows the resulting reflection coefficients for a range of offsets and values for α_2, indicating the nonlinear nature of the problem, especially around the critical angle.

As for the seismic source location example, the different methods presented in Section 3 are compared as a function of the number of forward samples used (see Fig. 14). The only changes compared to the seismic source location benchmarks are that less expressive variational families are used for the variational posterior method (Gaussian MDN with three layers consisting of 30 nodes in each layer, defining 10 Gaussians). The receivers for the optimal two-receiver design for the NMC_{re}, variational marginal and variational posterior method are located at 1020 and 1160 m offset.

All variational methods perform substantially better if the NMC_{re} method is taken as a baseline. The variational posterior method especially performs well in this scenario. If inner loop samples are not reused, the NMC method performs poorly, showing no sign of convergence even for 1×10^5 total samples. This is likely due to the very low sampling density in regions with near vertical R_p curves leading to near zero probabilities. Even for 1×10^5 total samples the inner loop of the NMC method contains only 46 samples which

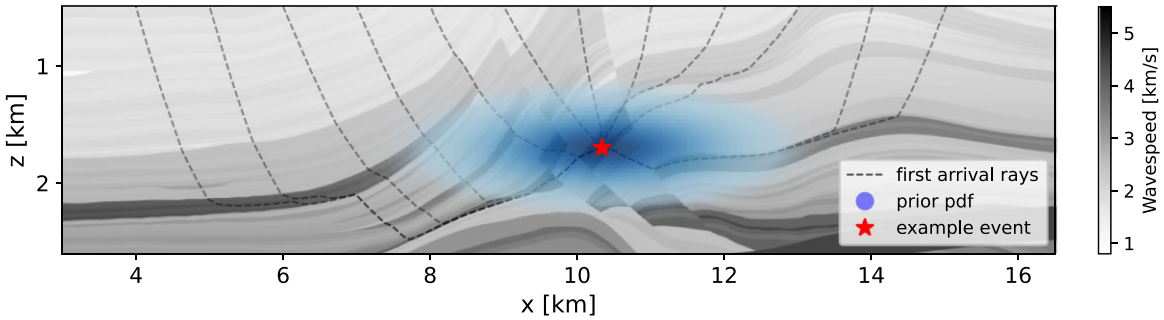

Figure 11. Seismic source location problem setup with prior samples (blue contours) and seismic rays (thin black lines) originating from one of the prior locations (red star).

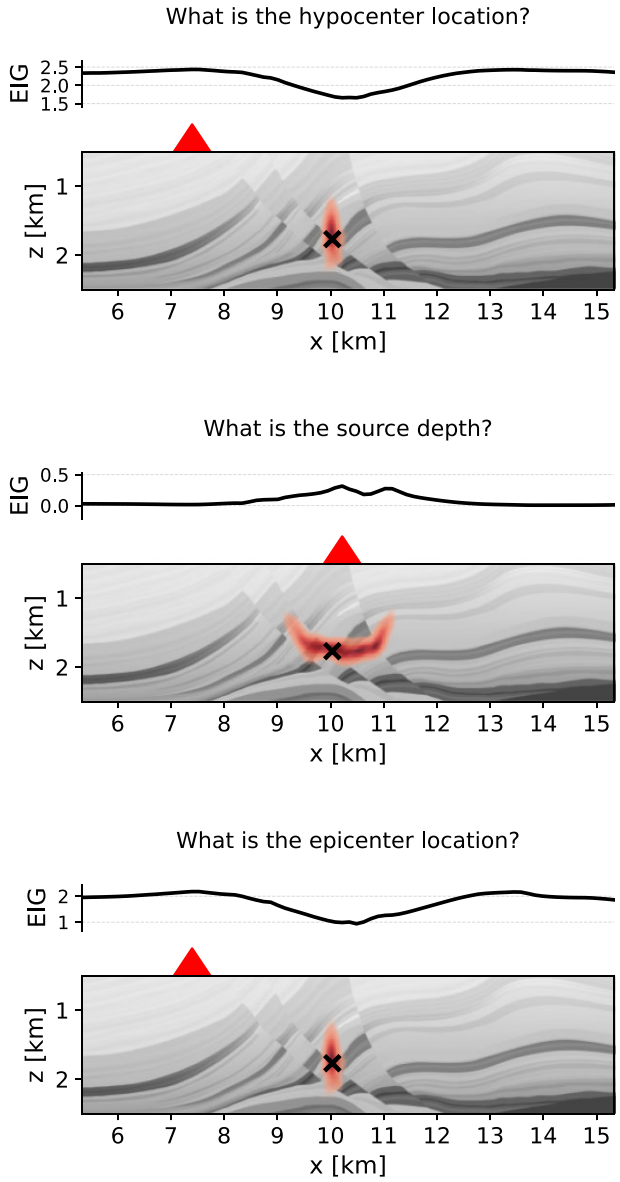

Figure 12. Comparison of optimal one-receiver networks for different interrogation goals. A high EIG in the top graph of each panel indicates good positions for answering the specific questions given in the panel title. The bottom graph of each scenario illustrates the geophysical setting and shows an example model parameter posterior pdf. The black cross indicates the true event location used to generate the datum in each example.

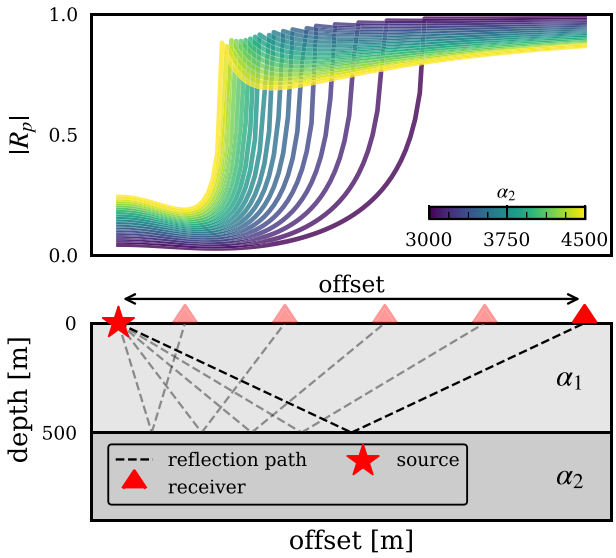

Figure 13. Schematic illustration of the AVO design problem setup. The top figure shows the change in reflection coefficient R_p (dimensionless) as a function of offset for different P-wave velocities in the lower layer, indicated by the different colours.

leads to a high variance of the EIG estimate. Around 1×10^9 total samples would be necessary for the NMC method to have the same number of inner loop samples as the NMC_{re} method at the point where it first convergences to a stable design (1×10^3 samples). The EIG value to which all methods converge is slightly lower than the best design of 1000 random trials due to the use of sequential construction for optimization. However, for a four-receiver design, the optimal designs of all but the NMC method outperform the best random design by a substantial margin (around 10 per cent). The value of experimental design is obvious, considering how much worse the average random and heuristic designs perform.

6.2.1 One-step EIG *optimization*

Using variational lower bounds, we use the AVO experimental design problem to illustrate the feasibility of SGD design optimization for geophysical problems (for more information, see Appendix A3). We use the variational posterior method and its gradients with respect to receiver positions to optimize the design ξ, while simultaneously fitting parameters $\boldsymbol{\phi}$ describing the variational approximator $q_p(\boldsymbol{m} \mid \boldsymbol{d}, \xi, \boldsymbol{\phi})$. Note that this is a slightly different use of SGD: previously we used SGD to only optimize $\boldsymbol{\phi}$ while keeping the design fixed, whereas here we use SGD to simultaneously optimize both $\boldsymbol{\phi}$

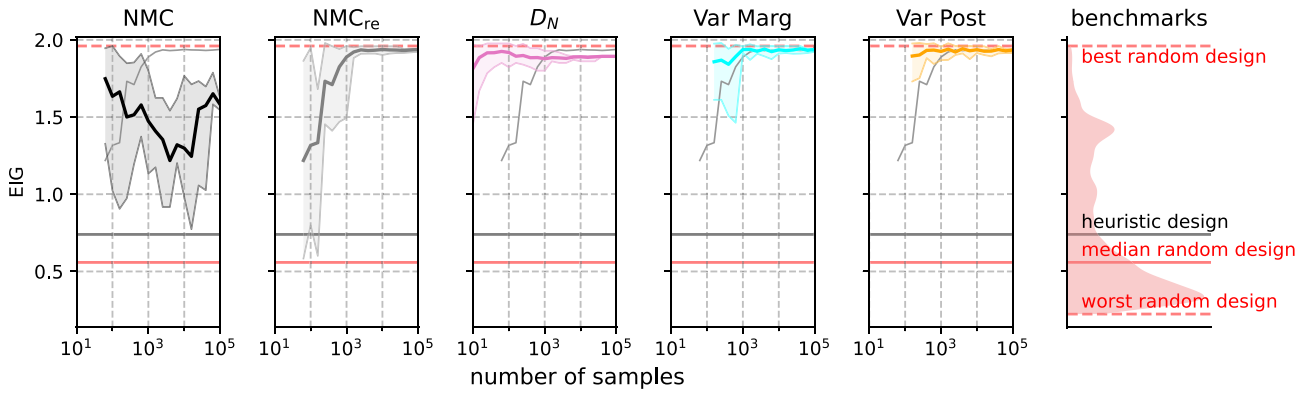

Figure 14. Benchmark results from different EIG estimation methods using sequential construction. Shows the EIG for a two-receiver network for an AVO design problem. The solid line indicates the mean of 10 independent benchmark runs, while the shaded area indicates the respective minimum and maximum values of those runs. The NMC$_{re}$ results are shown in every panel to serve as a benchmark for comparison. On the right, a smoothed histogram of the EIG for 1000 randomly selected designs and the EIG of a heuristic design is shown to put the results into context. Details on the setup and methods are in the main text. Benchmark results from different EIG estimation and design methods, showing the EIG for a two-receiver network for an AVO design problem as a function of the number of samples for which the forward function is evaluated. The solid line indicates the mean of 10 independent runs, while the shaded area indicates the respective minimum and maximum values of those runs. The mean curve of the NMC$_{re}$ results are shown in every panel to serve as a benchmark for comparison. On the right, a smoothed histogram of the EIG for 1000 randomly selected designs and for a heuristic design with receivers at 0.5 and 1.5 km is shown to put the results into context. Details on the setup and methods are in the main text.

and ξ at the same time. This approach is the BA-method introduced by Barber & Agakov (2004) and applied to experimental design first by Foster *et al.* (2019b). The variational family is the output of a Gaussian MDN with a three-layer neural network of 40 nodes in each layer defining 10 Gaussians as output. The design ξ is passed as an additional input to potentially strengthen the extrapolation capabilities of the variational mapping as proposed by Kleinegesse & Gutmann (2021) for a neural-network-only based EIGl ower bound. 1×10^5 SGA steps with a batch size of one and the optimizer Adam (Kingma & Ba 2014) were used to find values of ξ and $\boldsymbol{\phi}$ which maximize the EIG.

As a first test, we search for an optimal two-receiver design which can be benchmarked against a grid search using the NMC$_{re}$ method with 1×10^4 samples for both inner and outer loops to use as few samples as necessary while ensuring a reliable EIG estimate is obtained (see Fig. 14).

Since gradient descent algorithms are sensitive to local extrema, the initial design choice is important. We therefore ran 100 trials with random initial offsets (Fig. 15) where the EIG at the final design has been calculated using the NMC$_{re}$ method to make the results comparable. Most designs using SGA design optimization converge to the maximum EIG of the 200 $\times$ 200 grid search which is assumed to approximate the global maximum. The design calculated by sequential construction selecting locations from 200 offsets performs slightly worse than the grid search and most SGA-optimized designs. In fewer than 20 per cent of cases, one receiver gets stuck at the local minimum at zero offset. The R_p versus offset graph in Fig. 13 clearly shows this local maximum since the spread in data increases towards lower offsets from around 650 m. Due to the boundary at zero offset and the region with small gradients up to offsets of around 650 m, it is hard for the SGA algorithm to escape this local maximum in EIG. This behaviour can be seen in Fig. 15(a), which shows the change in designs during the SGA optimization process. In any case, the SGA-optimized designs outperform a heuristic design with equispaced receivers at 750 and 1250 m, since heuristically it is beneficial to place receivers between one and three times the depth of the interface (Guest & Curtis 2009). Initial designs with receivers at [50 m, 3450 m] (heuristically bad initial design) or at [1166 m, 2333 m] (heuristically good initial design) show the effects of choosing a reasonable or unreasonable initial design choice, where the unreasonable one gets stuck in a local maximum while the reasonable one converges towards the global maximum.

At the expense of potential convergence towards a local maximum, the SGA design optimization requires only 1×10^5 forward evaluations compared to the 8×10^8 for the grid search and 8×10^6 evaluations for the sequential construction using the NMC$_{re}$ method—resulting in a reduction of computational cost by a factor of 8000 and 80, respectively. If repeated computations are stored, the reduction in the number of forward evaluations drops to a factor of 20 in both cases, but savings would increase if a finer grid is used. The actual savings are hard to estimate since calculating the EIG using the NMC$_{re}$ method still involves a large number of likelihood evaluations, even if the forward model evaluations are precomputed. At the same time, the SGA algorithm introduces the overhead of calculating the gradients of the MDN for each sample using automatic differentiation. The actual reduction depends on the computational cost of the forward model, the complexity of the variational family, and the number of samples necessary for getting a stable EIG estimate.

To display the beneficial scaling properties of SGA design optimization, the AVO experimental design problem was repeated with 10 receivers and compared to the results of Guest & Curtis (2009). We used a Gaussian MDN with a three-layer neural network of 100 nodes in each layer defining 20 Gaussians as output. In this case, the SGA optimization algorithm was substantially less likely to get stuck in local minima and converged to a consistent solution in nearly all test runs. The number of SGA steps was increased to 4×10^5 to accommodate the larger number of receivers. However, as is evident from Fig. 16, the SGA-optimized design is already very close to the final design after around 5×10^4 iterations.

Two different initial conditions were tested as starting points for the SGA optimization. First, we used 10 equispaced receivers between 50 and 3450 m, which introduces little prior knowledge and is only subject to the constraint that receivers should lie between an offset of 0 and 3500 m. Second, the initial design are equispaced

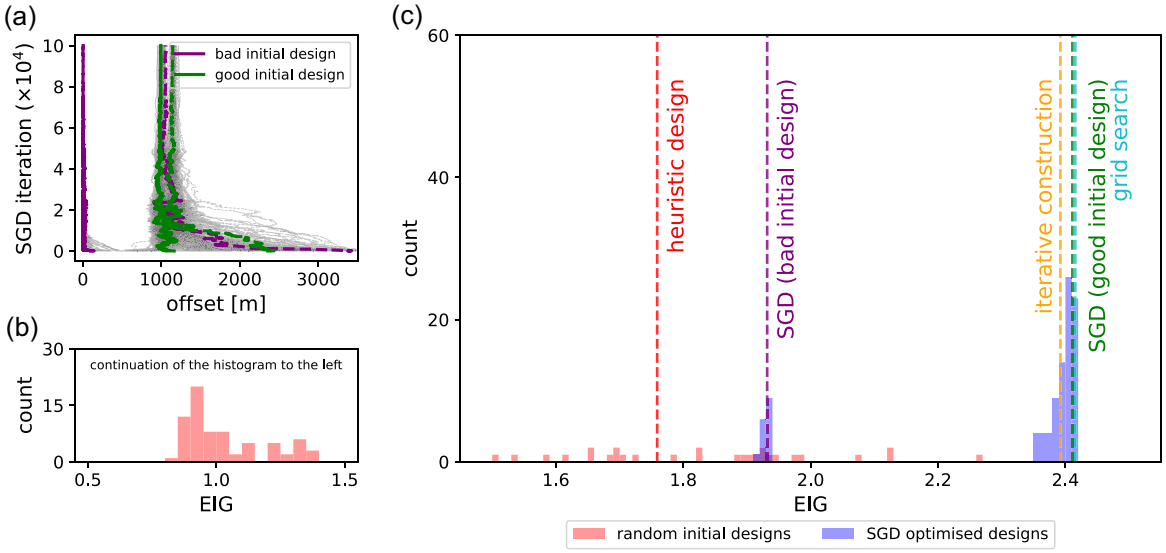

Figure 15. Summary of the results for the design of a two-receiver network for the AVO problem using SGA design optimization (c). EIGvalues of designs derived using the sequential construction method, a grid search and a heuristic design are given for comparison. All EIGvalues are calculated using the $\mathrm{NMC_{re}}$ method. (a) Offsets of the two receivers during the SGA design optimization, where solid lines indicate the receiver starting with the lower offset, and dashed lines indicate the one starting with the higher one. Red and blue histograms correspond to the EIG of starting and final designs, respectively. (b) Continuation of panel (c) towards the left with a lower resolution.

receivers between 500 and 1500 m, which is a heuristically good design, spanning between one and three times the depth of the interface. Since this starting design is closer to the final design, the SGA algorithm converges more quickly towards the final design.

The two SGA designs are compared to their respective initial designs and two benchmarks in Fig. 16. In both cases, but especially for the equispaced starting design, the EIG(calculated using the $\mathrm{NMC_{re}}$ method for better comparison) has increased substantially. Sequential construction ($\mathrm{NMC_{re}}$ method using $200{*}2 \times 10^4$ forward evaluations) is used here as a proxy for an optimal design. Since the value added by the 10th receiver is less than one per cent (Guest & Curtis 2009), the EIG of the sequential design will be very close to the global maximum. Both SGA designs perform slightly worse than the sequential construction design, which is most likely due to a combination of bias introduced by the variational approximation, choices in the learning rate, and small gradients due to the overdetermined nature of this design problem.

The resulting designs can also be compared to an optimal design for the same setup but with a different model parameter prior pdf. Instead of a Gaussian, Guest & Curtis (2009) used a uniform distribution with upper and lower bounds of 3000 and 4500 $\mathrm{m\,s^{-1}}$, respectively. Using the maximum entropy method, this results in a design with an EIG of 2.95, which is better than the heuristic design but worse than the SGA and sequential construction designs, which shows the influence of the prior pdf in experimental design problems. Nevertheless, the design outperforms the heuristic design even for a different prior pdf.

The computational savings in this 10-D design space are substantial. Even when repeated evaluations are saved, the SGA methods require around an order of magnitude fewer forward samples. When disregarding forward evaluations, the sequential design algorithm using the $\mathrm{NMC_{re}}$ method requires 200×10 EIG evaluations, each involving $1 \times 10^{42} = 1 \times 10^8$ likelihood evaluations. In contrast, only 2×10^5–4×10^5 forward and MDN evaluations are required to calculate the SGA experimental design. The evaluation of the variational family could be sped up substantially by using GPUs, which will be especially beneficial if more complex variational families are necessary.

6.3 Interrogation for CO_2

As introduced in Section 6.1.2, we will now demonstrate the use of variational methods for the design of experiments that answer a more practically interesting scientific question using AVO data. We focus on questions relating to the CO_2 saturation in a subsurface layer. The setup of the physical parameters represents a simplified model related to the Sleipner field (Dupuy *et al.* 2017; Ghosh & Ojha 2020). The upper layer is described by its seismic properties (P-wave velocity α_1 and S-wave velocity β_1), density ρ_1 and depth d, all with ranges given in Table 1. For the lower layer, the seismic properties are modelled using Gassmann fluid substitution (Gassmann 1951; Smith *et al.* 2003), which can be used to calculate seismic parameters given properties of the drained frame (bulk modulus K_{frame}, shear modulus G_{frame}, porosity Φ), mineral grains (bulk modulus K_{grain}, density ρ_{grain}), brine occupying the pore space (bulk modulus K_{brine}, density ρ_{brine}) and liquid CO_2 that replaces it (bulk modulus K_{CO_2}, density ρ_{CO_2}). Given all of these properties, only the saturation S of the pore space by CO_2 is required in order to calculate the AVO effect, and is assumed to be uniformly distributed between 0 and 1. In contrast, all other parameters are assumed to be distributed according to a Gaussian with means and standard deviations given in Table 1.

We must first calculate the pore fluid density and bulk modulus in order to apply the Gassmann equation. For this, we use the Voigt (arithmetic) average

$$\rho_{\mathrm{fluid}} = S\rho_{\mathrm{CO_2}} + (1-S)\rho_{\mathrm{brine}}, \tag{27}$$

$$K_{\mathrm{fluid}} = SK_{\mathrm{CO_2}} + (1-S)K_{\mathrm{brine}}, \tag{28}$$

which gives a (stiff) upper bound on the bulk modulus. In real-world applications, a Reuss average or Voigt–Reuss–Hill average might be more suitable, but here the focus is on the optimal design

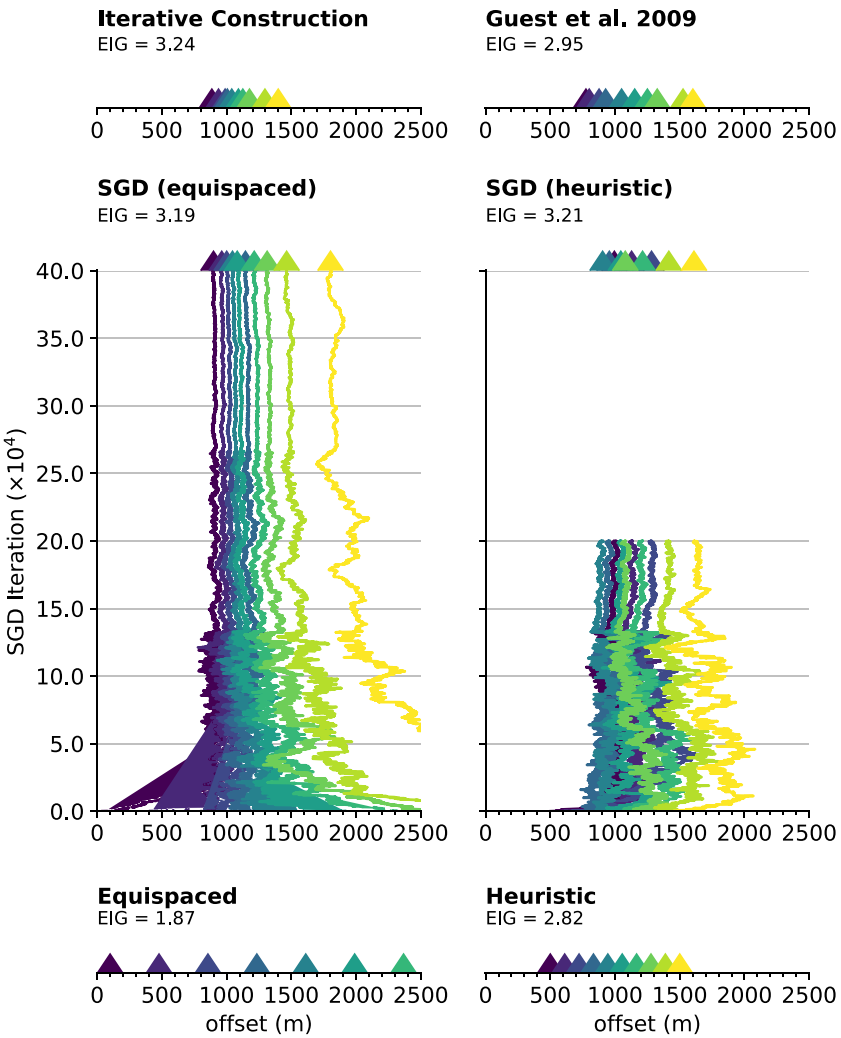

Figure 16. Designs (coloured triangles) and EIG values for 10-receiver SGA design optimization for the AVO design problem. If SGA iterations are shown on the y-axis, the design and EIG corresponds to the final (uppermost) design. SDG (equispaced) refers to an equispaced design between 0 and 3500 m which is optimized using 1×10^5 SGA steps, while SGA (heuristic) refers to a heuristically good design with equispaced receivers between 1500 and 2500 m which is optimized using 5×10^4 SGA steps. The offset axis is limited to a range from 0 to 2500 m, since the receivers for all but the equispaced design are concentrated in this area.

Table 1. Nuisance parameters in the AVO interrogation example. The quoted uncertainties correspond to the respective standard deviations.

Layer	Parameter	Value	Unit
Upper layer	α_1	2270 ± 10	$\mathrm{m\,s^{-1}}$
	β_1	854 ± 10	$\mathrm{m\,s^{-1}}$
	ρ_1	2100 ± 10	$\mathrm{kg\,m^{-3}}$
	d	1000 ± 50	m
Lower layer	K_{frame}	2.56 ± 0.77	GPa
	G_{frame}	8.5 ± 0.3	GPa
	Φ	0.37 ± 0.02	
	K_{grain}	39.3 ± 1.4	GPa
	ρ_{grain}	2664 ± 2.6	$\mathrm{kg\,m^{-3}}$
	K_{brine}	2.31 ± 0.07	GPa
	ρ_{brine}	1030 ± 20	$\mathrm{kg\,m^{-3}}$
	K_{co2}	0.08 ± 0.04	GPa
	ρ_{co2}	700 ± 77	$\mathrm{kg\,m^{-3}}$

algorithms rather than on details of the rock physical modelling. With the properties of the fluid at hand, the bulk modulus and density of the saturated rock can be modelled using the Gassmann equation

$$K_{\text{sat}} = K_{\text{frame}} + \frac{\left(1-\frac{K_{\text{frame}}}{K_{\text{grain}}}\right)^2}{\frac{\Phi}{K_{\text{fluid}}}+\frac{(1-\Phi)}{K_{\text{grain}}}-\frac{K_{\text{frame}}}{K^2_{\text{grain}}}}, \tag{29}$$

$$\rho_{\text{sat}} = \Phi\rho_{\text{grain}} + (1-\Phi)\rho_{\text{fluid}}, \tag{30}$$

which can then be used to calculate the *P*-wave velocity of the lower layer. Using Gassmann's equation, we implicitly assume that the shear modulus of the lower layer is independent of the CO_2 saturation. The seismic properties of the lower layer can now be calculated as

$$\alpha_2 = \sqrt{\frac{K_{\text{sat}}+\frac{4}{3}G_{\text{frame}}}{\rho_{\text{sat}}}}, \tag{31}$$

$$\beta_2 = \sqrt{\frac{G_{\text{frame}}}{\rho_{\text{sat}}}}, \tag{32}$$

which can then be used with the properties of the upper layer to calculate the *P*-wave reflection coefficient.

We qualitatively compare optimized two station designs for different design aims and a heuristic design with two receivers equally spaced between one and three times the interface depth (1000 m) at offsets of 1666 and 2333 m. In addition to the heuristic design, we calculated the design optimized to constrain all parameters given in Table 1 and the CO_2 saturation using the NMC_{re} method as a baseline for comparison. Sequential construction is used for this and all design optimizations in this subsection. To focus only on CO_2 saturation, we can see all parameters in Table 1 as nuisance parameters. With the variational posterior method (with the same MDN as earlier in this section), we can now derive a design that is optimal for estimating the CO_2 saturation alone, with the resulting design shown in Fig. 17. This design differs substantially from the heuristic design and the design constraining all model parameters. Since we have the mapping T^{-1} (the Gassmann equation including nuisance parameters) available in this case, it would be possible to use extended (and computationally more expensive) versions of both the NMC_{re} and variational marginal method for this interrogation design problem. This approach is not always possible if the scientific question is also a function of the CO_2 saturation.

For some applications, the exact value of the CO_2 saturation might not be of interest, but a key question is whether the saturation value is above or below a certain threshold. Changing the variational family of the variational posterior method to a neural network, taking data as input, and predicting the probability of exceeding the threshold makes it possible to design experiments optimally suited for estimating whether the CO_2 saturation is above or below a certain threshold. The last layer of the neural network is a sigmoid function, which is used to predict the probability of exceeding the threshold. The results of such an optimization for a threshold of 0.1 and 0.9 are shown in Fig. 17. The resulting designs are in a similar region as the one designed to constrain the value of the saturation, but deviate considerably to either focus on a more specific offset (threshold 0.1) or be spread further apart (threshold 0.9).

7 DISCUSSION

Recent works have shown the need for approximations in experimental design methods for fully nonlinear experimental design problems. Those approximations come in the form of linearized and Laplace methods which both assume that the forward model can be (locally) approximated by a linear model (e.g. Wilkinson *et al.* 2006; Long *et al.* 2015; Maurer *et al.* 2017; Carlon *et al.* 2020; Krampe *et al.* 2021), surrogates which approximate the forward model but put no constraints on model parameter prior or posterior pdf's (e.g. Huan & Marzouk 2013; Qiang *et al.* 2022; Wu *et al.* 2022), and functional approximations of the evidence, the posterior pdf or the mutual information between data and model parameters (e.g. Coles & Curtis 2011b; Kleinegesse & Gutmann 2018; Foster *et al.* 2019a; Kleinegesse & Gutmann 2020). In this work, we focused on variational methods which assume that the evidence or posterior pdf can be described sufficiently well by a closed-form variational approximator. They do not require any modification to either the forward problem or the prior information on the model parameters, which makes them attractive for general-purpose applications.

While functional approximations introduce additional complexity compared to straightforward double-loop Monte Carlo estimators such as the NMC method, they have significant advantages, especially since the NMC method with reused samples, while working well in the presented examples, can perform suboptimal when compared to methods using functional approximations (Englezou *et al.* 2022). Most importantly, they allow the design of experiments best suited to answer any scientific or applied question, provided a mapping from model space to the relevant target space can be defined. While this is important in its own right, the typical low dimensionality of the target space could allow experimental design methods to scale to substantially larger problems than currently possible.

Another significant advantage is the straightforward application of SGD design optimization using EIG lower bounds (variational posterior method in this study) if gradients with respect to the design parameters are available. While this is also possible using NMC or Laplace methods, they need a significant number of inner loop samples (Goda *et al.* 2020) or require an estimation of the *maximum a posteriori* estimate (Carlon *et al.* 2020), respectively, at each gradient descent step. The Laplace method can be seen as a special case of the variational posterior method in which the variational family is replaced by a multivariate Gaussian derived using the Hessian matrix of the linearized forward problem. Therefore, they should introduce a similar bias as incurred when using a well-trained MDN predicting one Gaussian with a full covariance matrix. The same goes for the consistent extensions of both methods, the VNMC (variational NMC) method of Foster *et al.* (2019a) and the Laplace-based importance sampling estimator of Carlon *et al.* (2020) and Englezou *et al.* (2022). Unlike the variational methods, the Laplace-based methods are inherently restricted to a Gaussian posterior pdf.

The D_N method performs well in the benchmarks presented in this paper. It shows the advantages of using a variational family which can be fit easily and whose information content can be evaluated analytically. Extending the method to be applicable for (some) interrogation design problems could benefit large-scale geophysical applications. It would provide a cheap and robust method that can be applied to many applications. If an inverse mapping of the target function exists or can be approximated, the variational marginal-likelihood method of Foster *et al.* (2019a) can be used to extend the D_N method to a subclass of interrogation problems. However, as has been demonstrated, the D_N method can lead to non-optimal designs if the assumption of Gaussian evidence is violated substantially. For complex high-dimensional problems, deciding whether a problem is ill-suited for the D_N method will be difficult. However, special care should be taken if the prior is multimodal or if the forward function

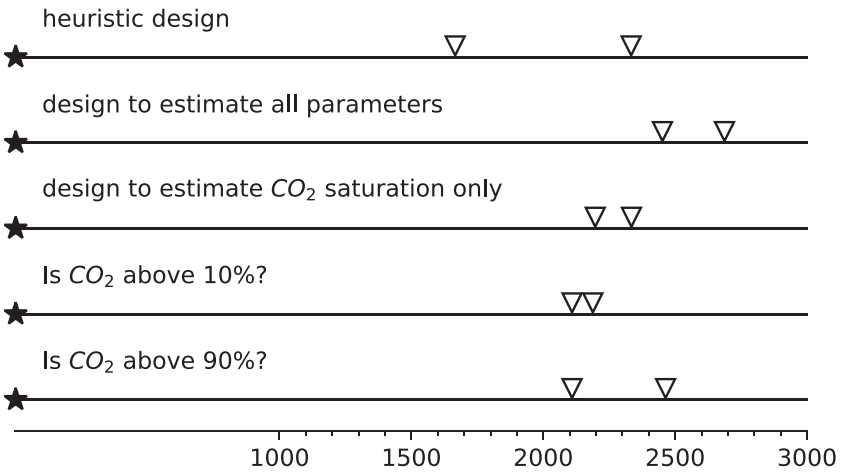

Figure 17. Offsets for two receiver designs for the AVO interrogation example. The heuristic design is chosen without optimization the other four designs are optimal designs for different scientific questions.

is known to show strong nonlinearity. Nevertheless, apart from in extreme cases, there is reason to believe that D_N design will provide at least above-average (compared to randomly selected) designs, at substantially reduced cost compared to most other design methods, and it is the only method that can produce robust designs using of the order of 10–100 forward evaluations.

Specifying variational mappings is a non-trivial task, and care needs to be taken to provide a sufficiently expressive yet computationally tractable one. Some of those problems could be alleviated using lower bounds on the EIG solely parametrized by neural networks (Kleinegesse & Gutmann 2021; Guo *et al.* 2021). Those have the same advantages as the variational posterior method but are easier to specify and can be constructed to be consistent (converge to the true EIG given enough samples) and with low variance (Guo *et al.* 2021), making them well suited for SGD design optimization. They are promising candidates for large-scale interrogation experimental design problems.

This paper aims to introduce the framework of variational experimental design methods clearly and intuitively and apply the methods to non-trivial geophysical applications. Consequently, the examples are both relatively small in scope and involve some approximations, which hinder their applications in their current form. The seismic source location example was calculated for the 2-D case, but extending it to three dimensions would be straightforward. The two-layer case of the AVO example, on the other hand, is rare to find in practical applications, and often, the interface of interest is buried under a complex overburden. In this case, a very good knowledge of the background velocities would be required to accurately model the reflectivity of the interface of interest. The reflectivity of seismic waves is nevertheless an important practical problem, and the simplified example chosen allows us to demonstrate the experimental design methods in a complex nonlinear physical setting while taking the first step towards the application of those methods to practical, real-world design problems.

While the dimensionalities of the presented scenarios are deliberately small, variational methods and other functional approximations offer one way of scaling Bayesian experimental design to higher dimensions. Exactly how this scaling works is hard to say currently. However, for problems where solving even one probabilistic inverse problem is challenging (e.g. full-waveform inversion), fully solving the experimental design problem will be even harder. Fortunately, it is often not required to solve the experimental design problem wholly accurately to get results better than a heuristic would provide.

While some results shown here could have been derived qualitatively using physical intuition, calculating the exact design requires a quantitative framework, as discussed in this paper. This is especially true for complex 3-D scenarios involving many receivers, complex priors, noise distributions and physical models, where heuristics and intuition break down.

8 CONCLUSIONS

In this paper, we have introduced variational experimental design methods to Geophysics, we have discussed their potential benefits and challenges, and placed them into context amongst linearized and other more established methods. We also briefly introduced the use of mutual information lower bounds for experimental design. The examples were chosen to illustrate the main concepts and encourage the use of these methods in geophysical applications.

We have compared different methods for estimating the value of an experiment and show that the naive NMC method is impractical for even small-scale geophysical problems due to the large number of samples required. In contrast, the variational methods and the NMC method with reused inner loop samples perform similarly well for both the seismic source location and AVO design problems. All three methods can fully account for the effects of nonlinearity in the physical process, but which method is preferred depends on the problem at hand.

We have also demonstrated how lower bounds on the EIG can be used to design interrogation experiments, that provide the best possible answer to specific questions of interest. We argue that this focused approach is more efficient and uses resources better than the traditional optimal design if a specific research question is of scientific interest. We also show that the optimal design can change substantially depending on the question posed.

Using AVO analysis as an example of a highly nonlinear geophysical process, we demonstrate the applicability and computational saving enabled by deploying stochastic gradient design optimization. This is especially relevant for high-dimensional designs that collect a large number of data. Even for expensive forward problems a small number of gradient descent steps can be used to refine heuristic designs to a specific problem at hand.

All methods used have been implemented in a Python package[2] to enable the use of OED for other researchers. Currently, it is in an early stage and lacks documentation. Still, it will be updated consistently over the coming years, and user-friendly documentation and tutorials will be added.

ACKNOWLEDGMENTS

The implementations of the OED algorithms in this work would not have been possible without extensive use of open-source software. Not all of them have been included in the respective sections to ease readability. All the code was written in Python (Van Rossum & Drake 2011), the libraries PyTorch (Paszke *et al.* 2019) and Zuko (Rozet 2023) were used to process probability distributions and implement the variational families, NumPy (Harris *et al.* 2020) was used for general data processing and Matplotlib (Hunter 2007) for plotting.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement no. 955515—SPIN ITN (www.spin-itn.eu)

DATA AVAILABILITY

This study only uses synthetic data. The code to generate the figures in Section 6 is available here: https://github.com/dominik-strutz/VarBEDfGP

REFERENCES

Ajo-Franklin, J. B., 2009. Optimal experiment design for time-lapse traveltime tomography, *Geophysics,* **74**(4), Q27–Q40.

Alexanderian, A., 2021. *Optimal Experimental Design for Infinite-dimensional Bayesian Inverse Problems Governed by PDEs: A Review,* https://doi.org/10.48550/arXiv.2005.12998.

Alexanderian, A. & Saibaba, A.K., 2018. *Efficient D-Optimal Design of Experiments for Infinite-dimensional Bayesian Linear Inverse Problems,* https://doi.org/10.48550/arXiv.1711.05878.

Alexanderian, A., Petra, N., Stadler, G. & Ghattas, O., 2014. *A-Optimal Design of Experiments for Infinite-dimensional Bayesian Linear Inverse Problems with Regularized ℓ_0-Sparsification,* https://doi.org/10.48550/arXiv.1308.4084.

Amzal, B., Bois, F.Y., Parent, E. & Robert, C.P., 2006. Bayesian-optimal design via interacting particle systems, *J. Am. Stat. Assoc.,* **101**(474), 773–785.

Arnold, R. & Curtis, A., 2018. Interrogation theory, *Geophys. J. Int.,* **214**(3), 1830–1846.

Atkinson, A.C. & Donev, A.N., 1992. *Optimum Experimental Designs*. Clarendon Press. doi: 10.1093/oso/9780198522546.001.0001.

Atkinson, A.C. & Fedorov, V.V., 1975. Optimal design: Experiments for discriminating between several models, *Biometrika* **62,** 289–303.

Attia, A., Alexanderian, A. & Saibaba, A.K., 2018. *Goal-oriented Optimal Design of Experiments for Large-scale Bayesian Linear Inverse Problems,* https://doi.org/10.48550/arXiv.1802.06517.

Barber, D. & Agakov, F., 2004. The IM algorithm: a variational approach to information maximization, *Adv. Neural Inf. Process. Syst.,* **16**(320), 887–905.

Barth, N. & Wunsch, C., 1990. Oceanographic experiment design by simulated annealing, *J. Phys. Oceanogr.,* **20**(9), 1249–1263.

Barth, N.H., 1992. Oceanographic experiment design II: genetic algorithms, *J. Atmos. Ocean. Technol.,* **9**(4), 434–443.

Beck, J., Dia, B.M., Espath, L. F.R., Long, Q. & Tempone, R., 2018. *Fast Bayesian Experimental Design: Laplace-based Importance Sampling for the Expected Information Gain,* https://doi.org/10.48550/arXiv.1710.03500.

Bernauer, M., Fichtner, A. & Igel, H., 2014. Optimal observables for multiparameter seismic tomography, *Geophys. J. Int.,* **198**(2), 1241–1254.

Bishop, C.M., 1994. *Mixture Density Networks,* Aston University.

Bishop, C.M., 2006. *Pattern Recognition and Machine Learning,* Springer.

Bloem, H., Curtis, A. & Maurer, H., 2020. Experimental design for fully nonlinear source location problems: which method should I choose?, *Geophys. J. Int.,* **223**(2), 944–958.

Bohachevsky, I.O., Johnson, M.E. & Stein, M.L., 1986. Generalized simulated annealing for function optimization, *Technometrics,* **28**(3), 209–217.

Box, G. E.P. & Lucas, H.L., 1959. Design of experiments in non-linear situations, *Biometrika,* **46**(1/2), 77–90.

Brenders, A.J. & Pratt, R.G., 2007. Efficient waveform tomography for lithospheric imaging: implications for realistic, two-dimensional acquisition geometries and low-frequency data, *Geophys. J. Int.,* **168**(1), 152–170.

Carlon, A.G., Dia, B.M., Espath, L., Lopez, R.H. & Tempone, R., 2020. Nesterov-aided stochastic gradient methods using laplace approximation for bayesian design optimization, *Comput. Methods Appl. Mech. Eng.,* **363,** 112909, doi: 10.1016/j.cma.2020.112909.

Chaloner, K. & Verdinelli, I., 1995. Bayesian experimental design: A review, *Stat. Sci.,* **10**(3), 273–304.

Cheng, P., Hao, W., Dai, S., Liu, J., Gan, Z. & Carin, L., 2020. *CLUB: A Contrastive Log-ratio Upper Bound of Mutual Information,* https://doi.org/10.48550/arXiv.2006.12013.

Coles, D. & Curtis, A., 2011a. A free lunch in linearized experimental design?, *Comput. Geosci.,* **37**(8), 1026–1034.

Coles, D. & Curtis, A., 2011b. Efficient nonlinear bayesian survey design using DN optimization, *Geophysics,* **76**(2), Q1–Q8.

Coles, D. & Prange, M., 2012. Toward efficient computation of the expected relative entropy for nonlinear experimental design, *Inverse Prob.,* **28**(5), http://dx.doi.org/10.1088/0266-5611/28/5/055019.

Coles, D., Yang, Y., Djikpesse, H., Prange, M. & Osypov, K., 2013. Optimal nonlinear design of marine borehole seismic surveys, *Geophysics,* **78**(3), WB17–WB29.

Coles, D.A. & Morgan, F.D., 2009. A method of fast, sequential experimental design for linearized geophysical inverse problems, *Geophys. J. Int.,* **178**(1), 145–158.

Cover, T.M. & Thomas, J.A., 2006. *Elements of Information Theory,* John Wiley & Sons, Nashville, TN , 2nd edn.

Curtis, A., 1999a. Optimal experiment design: cross-borehole tomographic examples, *Geophys. J. Int.,* **136**(3), 637–650.

Curtis, A., 1999b. Optimal design of focused experiments and surveys, *Geophys. J. Int.,* **139**(1), 205–215.

Curtis, A., 2004a. Theory of model-based geophysical survey and experimental design part B - nonlinear problems, *Leading Edge,* **23**(10), 1112–1117.

Curtis, A., 2004b. Theory of model-based geophysical survey and experimental design part a—linear problems, *Leading Edge,* **23**(10), 997–1004.

Curtis, A. & Snieder, R., 1997. Reconditioning inverse problems using the genetic algorithm and revised parameterization, *Geophysics,* **62**(5), 1524–1532.

Curtis, A. & Spencer, C., 1999. Survey design strategies for linearized nonlinear inversion, in *SEG Technical Program Expanded Abstracts 1999,* Society of Exploration Geophysicists, 1775–1778.

Curtis, A. & Wood, R., 2004. *Optimal Elicitation of Probabilistic Information from Experts,* https://www.lyellcollection.org/doi/10.1144/GSL.SP.2004.239.01.09, accessed: 2022 July 30.

Curtis, A., Michelini, A., Leslie, D. & Lomax, A., 2004. A deterministic algorithm for experimental design applied to tomographic and microseismic monitoring surveys, *Geophys. J. Int.,* **157**(2), 595–606.

Dasgupta, A., Hostache, R., Ramsankaran, R., Schumann, G. J.-P., Grimaldi, S., Pauwels, V. R.N. & Walker, J.P., 2021. On the impacts of observation location, timing, and frequency on flood extent assimilation performance, *Water Resour. Res.,* **57**(2), e2020WR028238, https://doi.org/10.1029/2020WR028238.

[2] Available under https://github.com/dominik-strutz/GeoBOED

De Landro, G., Picozzi, M., Russo, G., Adinolfi, G.M. & Zollo, A., 2020. Seismic networks layout optimization for a high-resolution monitoring of induced micro-seismicity, *J. Seismol.,* **24**(5), 953–966.

Dinh, L., Krueger, D. & Bengio, Y., 2014. *NICE: Non-linear Independent Components Estimation,* https://doi.org/10.48550/arXiv.1410.8516.

Djikpesse, H.A., Khodja, M.R., Prange, M.D., Duchenne, S. & Menkiti, H., 2012. Bayesian survey design to optimize resolution in waveform inversion, *Geophysics,* **77**(2), R81–R93.

Dupuy, B.,, Ghaderi, A., Querendez, E. & Mezyk, M., 2017. Constrained AVO for CO2 storage monitoring at sleipner, *Energy Procedia,* **114,** 3927–3936.

Durkan, C., Bekasov, A., Murray, I. & Papamakarios, G., 2019. *Neural Spline Flows,* https://doi.org/10.48550/arXiv.1906.04032.

Englezou, Y., Waite, T.W. & Woods, D.C., 2022. Approximate laplace importance sampling for the estimation of expected shannon information gain in high-dimensional bayesian design for nonlinear models, *Stat. Comput.,* **32**(5), https://doi.org/10.1007/s11222-022-10159-2.

Fedorov, V.V. & Hackl, P., 1997. *Model-Oriented Design of Experiments,* Springer Science & Business Media.

Feng, C. & Marzouk, Y.M., 2019. *A Layered Multiple Importance Sampling Scheme for Focused Optimal Bayesian Experimental Design,* https://doi.org/10.48550/arXiv.1903.11187.

Ferrolino, A.R., Lope, J. E.C. & Mendoza, R.G., 2020. Optimal location of sensors for early detection of tsunami waves, in *Computational Science – ICCS 2020,* pp. 562–575, Springer International Publishing.

Fichtner, A. & Hofstede, C., 2022. A simple algorithm for optimal design in distributed fibre-optic sensing, *Geophys. J. Int.,* **233**(1), 229–233.

Foster, A., Jankowiak, M., Bingham, E., Horsfall, P., Whye Teh, Y., Rainforth, T. & Goodman, N., 2019a. Variational bayesian optimal experimental design, in *Advances in Neural Information Processing Systems,* https://doi.org/10.48550/arXiv.1903.05480.

Foster, A., Jankowiak, M., O'Meara, M., Teh, Y.W. & Rainforth, T., 2019b. *A Unified Stochastic Gradient Approach to Designing Bayesian-Optimal Experiments,* https://doi.org/10.48550/arXiv.1911.00294.

Furman, A., Ferre, T. P.A. & Warrick, A.W., 2004. Optimization of ERT Surveys for Monitoring Transient Hydrological Events using Perturbation Sensitivity and Genetic Algorithms. *Hydrogeophysic,* **3**(4), 1230–1239.

Gassmann, F., 1951. Elastic waves through a packing of spheres, *Geophysics,* **16**(4), 673–685.

Ghosh, R. & Ojha, M., 2020. Prediction of elastic properties within CO 2 plume at sleipner field using AVS inversion modified for thin-layer reflections guided by uncertainty estimation, *J. geophys. Res. (Solid Earth),* **125**(11), e2020JB019782. https://doi.org/10.1029/2020JB019782.

Gibson, R.L. Jr & Tzimeas, C., 2002. Quantitative measures of image resolution for seismic survey design, *Geophysics,* **67**(6), 1844–1852.

Goda, T., Hironaka, T., Kitade, W. & Foster, A., 2020. *Unbiased MLMC Stochastic Sradient-based Optimization of Bayesian Experimental Designs,* https://doi.org/10.48550/arXiv.2005.08414.

Guest, T. & Curtis, A., 2009. Iteratively constructive sequential design of experiments and surveys with nonlinear parameter-data relationships, *J. Geophys. Res.[Solid Earth],* **114**(B4), https://doi.org/10.1029/2008JB005948.

Guest, T. & Curtis, A., 2010. Optimal trace selection for AVA processing of shale-sand reservoirs, *Geophysics,* **75**(4), C37–C47.

Guest, T. & Curtis, A., 2011. On standard and optimal designs of industrial-scale 2-D seismic surveys, *Geophys. J. Int.,* **186**(2), 825–836.

Guo, Q. *et al.*, 2021. *Tight Mutual Information Estimation with Contrastive Fenchel-Legendre Optimization,* https://doi.org/10.48550/arXiv.2107.01131.

Haber, E., Horesh, L. & Tenorio, L., 2008. Numerical methods for experimental design of large-scale linear ill-posed inverse problems, *Inverse Probl.,* **24**(5), https://doi.org/10.1088/0266-5611/24/5/055012.

Hainy, M., Müller, W.G. & Wagner, H., 2014. Likelihood-free simulation-based optimal design: an introduction, in *Springer Proceedings in Mathematics & Statistics,* pp. 271–278, Springer, New York, NY.

Hainy, M., Müller, W.G. & Wagner, H., 2016. Likelihood-free simulation-based optimal design with an application to spatial extremes, *Stoch. Environ. Res. Risk Assess.,* **30**(2), 481–492.

Hainy, M., Price, D.J., Restif, O. & Drovandi, C., 2018. *Optimal Bayesian Design for Model Discrimination via Classification,* https://doi.org/10.1007/s11222-022-10078-2.

Harris, C.R. *et al.*, 2020. Array programming with NumPy, *Nature,* **585**(7825), 357–362.

Holland, J.H., 1992. *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence,* MIT Press.

Huan, X. & Marzouk, Y., 2014. Gradient-based stochastic optimization methods in bayesian experimental design, *IJUQ,* **4**(6), https://doi.org/10.1615/Int.J.UncertaintyQuantification.2014006730.

Huan, X. & Marzouk, Y.M., 2013. Simulation-based optimal bayesian experimental design for nonlinear systems, *J. Comput. Phys.,* **232**(1), 288–317.

Hunter, J., 2007. Matplotlib: a 2D graphics environment, *Comput. Sci. Eng.,* **9,** 90–95.

Hunziker, J., Thorbecke, J., Brackenhoff, J. & Slob, E., 2016. Inversion of controlled-source electromagnetic reflection responses, *Geophysics,* **81**(5), F49–F57.

Hyvönen, N., Seppänen, A. & Staboulis, S., 2014. Optimizing electrode positions in electrical impedance tomography, *SIAM J. Appl. Math.,* **74**(6), 1831–1851.

Jagalur-Mohan, J. & Marzouk, Y., 2021. Batch greedy maximization of non-submodular functions: guarantees and applications to experimental design, *J. Mach. Learn. Res.,* **22**(252), 1–62.

Jones, D.R., Schonlau, M. & Welch, W.J., 1998. Efficient global optimization of expensive Black-Box functions, *J. Global Optimiz.,* **13**(4), 455–492.

Khodja, M.R., Prange, M.D. & Djikpesse, H.A., 2010. Guided bayesian optimal experimental design, *Inverse Probl.,* **26**(5), https://www.doi.org/10.1088/0266-5611/26/5/055008.

Kiefer, J., 1959. Optimum experimental designs, *J. R. Stat. Soc.,* **21**(2), 272–304.

Kijko, A., 1977a. An algorithm for the optimum distribution of a regional seismic network?i, *Pure appl. Geophys.,* **115**(4), 999–1009.

Kijko, A., 1977b. An algorithm for the optimum distribution of a regional seismic network? II. an analysis of the accuracy of location of local earthquakes depending on the number of seismic stations, *Pure appl. Geophys.,* **115**(4), 1011–1021.

Kim, K. & Lees, J.M., 2014. Local volcano infrasound and source localization investigated by 3D simulation, *Seismol. Res. Lett.,* **85**(6), 1177–1186.

Kingma, D.P. & Ba, J., 2014. *Adam: A Method for Stochastic Optimization,* https://doi.org/10.48550/arXiv.1412.6980.

Kleinegesse, S. & Gutmann, M., 2018. *Efficient Bayesian Experimental Design for Implicit Models,* https://doi.org/10.48550/arXiv.1810.09912.

Kleinegesse, S. & Gutmann, M.U., 2020. *Bayesian Experimental Design for Implicit Models by Mutual Information Neural Estimation,* https://doi.org/10.48550/arXiv.2002.08129.

Kleinegesse, S. & Gutmann, M.U., 2021. *Gradient-based Bayesian Experimental Design for Implicit Models using Mutual Information Lower Bounds,* https://doi.org/10.48550/arXiv.2105.04379.

Krampe, V., Edme, P. & Maurer, H., 2021. Optimized experimental design for seismic full waveform inversion: A computationally efficient method including a flexible implementation of acquisition costs, *Geophys. Prospect.,* **69**(1), 152–166.

Kullback, S. & Leibler, R.A., 1951. On information and sufficiency, *Ann. Math. Stat.,* **22**(1), 79–86.

Lindley, D.V., 1956. On a Measure of the Information provided by an Experiment, *Ann. Math. Statist.,* **27**(4), 986–1005.

Liner, C.L., Underwood, W.D. & Gobeli, R., 1999. 3-D seismic survey design as an optimization problem, *Leading Edge,* **18**(9), 1054–1060.

Long, Q., 2022. Multimodal information gain in bayesian design of experiments, *Comput. Stat.,* **37**(2), 865–885.

Long, Q., Scavino, M., Tempone, R. & Wang, S., 2013. Fast estimation of expected information gains for bayesian experimental designs based on laplace approximations, *Comput. Methods Appl. Mech. Eng.,* **259,** 24–39.

Long, Q., Motamed, M. & Tempone, R., 2015. Fast bayesian optimal experimental design for seismic source inversion, *Comput. Methods Appl. Mech. Eng.,* **291,** 123–145.

López-Comino, J.A., Cesca, S., Kriegerowski, M., Heimann, S., Dahm, T., Mirek, J. & Lasocki, S., 2017. Monitoring performance using synthetic data for induced microseismicity by hydrofracking at the wysin site (poland), *Geophys. J. Int.,* **210**(1), 42–55.

Lugrin, G., Mora, N.M., Rachidi, F., Rubinstein, M. & Diendorfer, G., 2014. On the location of lightning discharges using time reversal of electromagnetic fields, *IEEE Trans. Electromagn. Compat.,* **56**(1), 149–158.

Martin, G.S., Wiley, R. & Marfurt, K.J., 2006. Marmousi2 an elastic upgrade for marmousi, *Leading Edge,* **25**(2), 156–166.

Maurer, H. & Boerner, D.E., 1998. Optimized and robust experimental design: a non-linear application to EM sounding, *Geophys. J. Int.,* **132**(2), 458–468.

Maurer, H., Boerner, D.E. & Curtis, A., 2000. Design strategies for electromagnetic geophysical surveys, *Inverse Probl.,* **16**(5), https://www.doi.org/10.1088/0266-5611/16/5/302.

Maurer, H., Greenhalgh, S. & Latzel, S., 2009. Frequency and spatial sampling strategies for crosshole seismic waveform spectral inversion experiments, *Geophysics,* **74**(6), WCC79–WCC89.

Maurer, H., Curtis, A. & Boerner, D.E., 2010. Recent advances in optimized geophysical survey design, *Geophysics,* **75**(5), 75A177–75A194.

Maurer, H., Nuber, A., Korta Martiartu, N., Reiser, F., Boehm, C., Manukyan, E., Schmelzbach, C. & Fichtner, A., 2017. Chapter one - optimized experimental design in the context of seismic full waveform inversion and seismic waveform imaging, in *Advances in Geophysics,* Vol. **58,** pp. 1–45, ed. ,Nielsen, L., Elsevier.

Maurer, V., Gaucher, E., Grunberg, M., Koepke, R., Pestourie, R. & Cuenot, N., 2020. Seismicity induced during the development of the rittershoffen geothermal field, france, *Geotherm. Energy,* **8**(1), 1–31.

Meier, U., Trampert, J. & Curtis, A., 2009. Global variations of temperature and water content in the mantle transition zone from higher mode surface waves, *Earth planet. Sci. Lett.,* **282**(1), 91–101.

Mitchell, T.J., 1974. An algorithm for the construction of "D-Optimal" experimental designs, *Technometrics,* **16**(2), 203–210.

Mosegaard, K. & Tarantola, A., 1995. Monte carlo sampling of solutions to inverse problems, *J. geophys. Res.,* **100**(B7), 12431–12447.

Muir, J.B. & Zhan, Z., 2021. Wavefield-based evaluation of DAS instrument response and array design, *Geophys. J. Int.,* **229**(1), 21–34.

Myung, J.I., Cavagnaro, D.R. & Pitt, M.A., 2013. A tutorial on adaptive design optimization, *J. Math. Psychol.,* **57**(3–4), 53–67.

Nuber, A., Manukyan, E. & Maurer, H., 2017. Optimizing measurement geometry for seismic near-surface full waveform inversion, *Geophys. J. Int.,* **210**(3), 1909–1921.

Oldenborger, G.A. & Routh, P.S., 2009. The point-spread function measure of resolution for the 3-D electrical resistivity experiment, *Geophys. J. Int.,* **176**(2), 405–414.

Paszke, A. *et al.*, 2019. *PyTorch: An Imperative Style, High-Performance Deep Learning Library,* https://doi.org/10.48550/arXiv.1912.01703.

Pronzato, L. & Walter, E., 1985. Robust experiment design via stochastic approximation, *Math. Biosci.,* **75**(1), 103–120.

Qiang, S., Shi, X., Kang, X. & Revil, A., 2022. Optimized arrays for electrical resistivity tomography survey using bayesian experimental design, *Geophysics,* **87**(4), E189–E203.

Rabinowitz, N. & Steinberg, D.M., 1990. Optimal configuration of a seismographic network: a statistical approach, *Bull. seism. Soc. Am.,* **80**(1), 187–196.

Rabinowitz, N. & Steinberg, D.M., 2000. A statistical outlook on the problem of seismic network configuration, in *Advances in Seismic Event Location,* pp. 51–69, eds ,Thurber, C.H.& & ,Rabinowitz, N., Springer Netherlands, Dordrecht.

Rainforth, T., Cornish, R., Yang, H., Warrington, A. & Wood, F., 2017. *On Nesting Monte Carlo Estimators,* https://doi.org/10.48550/arXiv.1709.06181.

Rawlinson, Z.J., Townend, J., Arnold, R. & Bannister, S., 2012. Derivation and implementation of a nonlinear experimental design criterion and its application to seismic network expansion at kawerau geothermal field, New Zealand, *Geophys. J. Int.,* **191**(2), 686–694.

Ren, Y., Xu, X., Yang, S., Nie, L. & Chen, Y., 2020. A physics-based neural-network way to perform seismic full waveform inversion, *IEEE Access,* **8,** https://www.doi.org/10.1109/ACCESS.2020.2997921.

Rényi, A., 1961. On measures of entropy and information, in *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability,* Vol. **1,** pp. 547–562. University of California Press.

Rezende, D.J. & Mohamed, S., 2015. *Variational Inference with Normalizing Flows,* https://doi.org/10.48550/arXiv.1505.05770.

Richardson, A., 2022. *Deepwave,* https://doi.org/10.5281/zenodo.8381177.

Robbins, H. & Monro, S., 1951. A stochastic approximation method, *Ann. Math. Stat.,* **22**(3), 400–407.

Romdhane, A. & Eliasson, P., 2018. Optimised geophysical survey design for CO2 monitoring–a synthetic study, in *14th Greenhouse Gas Control Technologies Conference Melbourne,* pp. 21–26, http://dx.doi.org/10.2139/ssrn.3366260.

Rozet, F., 2023. *Zuko 0.2.0,* https://www.doi.org/10.5281/zenodo.7625672.

Runge, A.K., Scherbaum, F., Curtis, A. & Riggelsen, C., 2013. An interactive tool for the elicitation of subjective probabilities in probabilistic Seismic–Hazard analysis, *Bull. seism. Soc. Am.,* **103**(5), 2862–2874.

Ryan, E.G., Drovandi, C.C., McGree, J.M. & Pettitt, A.N., 2016. A review of modern computational algorithms for bayesian optimal design, *Int. Stat. Rev.,* **84**(1), 128–154.

Ryan, K.J., 2003. Estimating expected information gains for experimental designs with application to the random Fatigue-Limit model, *J. Comput. Graph. Stat.,* **12**(3), 585–603.

Sethian, J.A., 1996. A fast marching level set method for monotonically advancing fronts, *Proc. Natl. Acad. Sci. U. S. A.,* **93**(4), 1591–1595.

Shannon, C.E., 1948. A mathematical theory of communication, *Bell System Tech. J.,* **27**(3), 379–423.

Shewry, M.C. & Wynn, H.P., 1987. Maximum entropy sampling, *J. Appl. Stat.,* **14**(2), 165–170.

Smith, J.D., Azizzadenesheli, K. & Ross, Z.E., 2021. EikoNet: solving the eikonal equation with deep neural networks, *IEEE Trans. Geosci. Remote Sens.,* **59**(12), 10685–10696.

Smith, T.M., Sondergeld, C.H. & Rai, C.S., 2003. Gassmann fluid substitutions: a tutorial, *Geophysics,* **68**(2), 430–440.

Steinberg, D.M., Rabinowitz, N., Shimshoni, Y. & Mizrachi, D., 1995. Configuring a seismographic network for optimal monitoring of fault lines and multiple sources, *Bull. seism. Soc. Am.,* **85**(6), 1847–1857.

Stowell, D. & Plumbley, M.D., 2009. Fast multidimensional entropy estimation by *k*-d partitioning, *IEEE Signal Process. Lett.,* **16**(6), 537–540.

Stummer, P., Maurer, H., Horstmeyer, H. & Green, A.G., 2002. Optimization of DC resistivity data acquisition: real-time experimental design and a new multielectrode system, *IEEE Trans. Geosci. Remote Sens.,* **40**(12), 2727–2735.

Stummer, P., Maurer, H. & Green, A.G., 2004. Experimental design: electrical resistivity data sets that provide optimum subsurface information, *Geophysics,* **69**(1), 120–139.

Tabak, E.G. & Turner, C.V., 2013. A family of nonparametric density estimation algorithms, *Commun. Pure appl,* **66**(2), 145–164.

Tarantola, A., 1984. Inversion of seismic reflection data in the acoustic approximation, *Geophysics,* **49**(8), 1259–1266.

Tarantola, A., 2005. *Inverse Problem Theory and Methods for Model Parameter Estimation,* SIAM.

Tierney, L. & Kadane, J.B., 1986. Accurate approximations for posterior moments and marginal densities, *J. Am. Stat. Assoc.,* **81**(393), 82–86.

Toledo, T., Jousset, P., Maurer, H. & Krawczyk, C., 2020. Optimized experimental network design for earthquake location problems: applications to geothermal and volcanic field seismic networks, *J. Volc. Geotherm. Res.,* **391,** 106433. https://doi.org/10.1016/j.jvolgeores.2018.08.011.

Tsutakawa, R.K., 1972. Design of experiment for bioassay, *J. Am. Stat. Assoc.,* **67**(339), 584–590.

van Den Berg, J., Curtis, A. & Trampert, J., 2003. Optimal nonlinear bayesian experimental design: an application to amplitude versus offset experiments, *Geophys. J. Int.,* **155**(2), 411–421.

van Den Berg, J., Curtis, A. & Trampert, J., 2005. Corrigendum, *Geophys. J. Int.,* **161**(2), 265–265.

Van Rossum, G. & Drake, F., 2011. *The Python Language Reference Manual,* Network Theory, Bristol, England.

Vincent & Rainforth, 2017. The DARC toolbox: automated, flexible, and efficient delayed and risky choice experiments using bayesian adaptive design, https://doi.org/10.31234/osf.io/yehjb.

White, M. C.A., Fang, H., Nakata, N. & Ben-Zion, Y., 2020. PyKonal: a python package for solving the eikonal equation in spherical and cartesian coordinates using the fast marching method, *Seismol. Res. Lett.,* **91**(4), 2378–2389.

Wilkinson, P.B., Chambers, J.E., Meldrum, P.I., Ogilvy, R.D. & Caunt, S., 2006. Optimization of array configurations and panel combinations for the detection and imaging of abandoned mineshafts using 3D cross-hole electrical resistivity tomography, *J. Environ. Eng. Geophys.,* **11**(3), 213–221.

Wilkinson, P.B., Loke, M.H., Meldrum, P.I., Chambers, J.E., Kuras, O., Gunn, D.A. & Ogilvy, R.D., 2012. Practical aspects of applied optimized survey design for electrical resistivity tomography, *Geophys J Int.,* **189**(1), 428–440.

Winterfors, E. & Curtis, A., 2008. Numerical detection and reduction of non-uniqueness in nonlinear inverse problems, *Inverse Probl.,* **24**(2), https://www.doi.org/10.1088/0266-5611/24/2/025016.

Winterfors, E. & Curtis, A., 2012. A bifocal measure of expected ambiguity in bayesian nonlinear parameter estimation, *Technometrics,* **54**(2), 179–190.

Wu, K., Chen, P. & Ghattas, O., 2020. *A Fast and Scalable Computational Framework for Large-scale and High-dimensional Bayesian Optimal Experimental Design,* https://doi.org/10.48550/arXiv.2010.15196.

Wu, K., Chen, P. & Ghattas, O., 2021. *An Efficient Method for Goal-oriented Linear Bayesian Optimal Experimental Design: Application to Optimal Sensor Placement,* https://doi.org/10.48550/arXiv.2102.06627.

Wu, K., O'Leary-Roseberry, T., Chen, P. & Ghattas, O., 2022. *Large-scale Bayesian optimal experimental design with derivative-informed projected neural network,* https://doi.org/10.48550/arXiv.2201.07925.

Zhang, J., Zeng, L., Chen, C., Chen, D. & Wu, L., 2015. Efficient bayesian experimental design for contaminant source identification, *Water Resour. Res.,* **51**(1), 576–598.

Zhao, X., Curtis, A. & Zhang, X., 2020. Bayesian seismic tomography using normalizing flows, *Geophys. J. Int.,* **228**(1), 213–239.

APPENDIX A: ADVANCED CONCEPTS

We now briefly describe selected implementations of more advanced design concepts, using a set of examples that hint at possible useful developments in future. They also reinforce the notion that variational and other related methods may be valuable for the design of experiments in certain types of geophysical applications.

A1 Likelihood-free experimental design

So far, we have assumed that it is possible to evaluate $p(\boldsymbol{d} \mid \boldsymbol{m}, \xi)$ pointwise (meaning the likelihood function is explicit or otherwise directly computable). This assumption is typically valid in geophysics as the likelihood is often assumed to be a Gaussian distribution around a mean predicted by the forward model. However, in certain applications, inherent randomness in the forward function exists, in which case this is not possible. An example is seismic source location in an unknown heterogeneous medium, where the randomness is due to the different realizations this medium can take. In this case, the likelihood is intractable and must be approximated. Typically, this randomness is absorbed in an explicit Gaussian data likelihood, but using a likelihood-free approach would allow us to explicitly model the effect of this randomness.

For model space methods, this is of no consequence since the likelihood is only used to generate data samples and so its value need never be evaluated explicitly. Data space methods, on the other hand, rely on explicit evaluations of $p(\boldsymbol{d} \mid \boldsymbol{m}, \xi)$. If an average over nuisance variables can be used to model the intractable likelihood, the NMC method can be used in an extended form (Feng & Marzouk 2019), but in general this will lead to high computational cost in practical problems because in this extended form the inner loop samples can not be reused. Alternative methods include the use of a variational approximation of the likelihood (e.g. Foster *et al.* 2019a; Cheng *et al.* 2020), and several works focus exclusively on experimental design algorithms for likelihood-free experimental design (Hainy *et al.* 2014, 2016, 2018; Kleinegesse & Gutmann 2018, 2020, 2021).

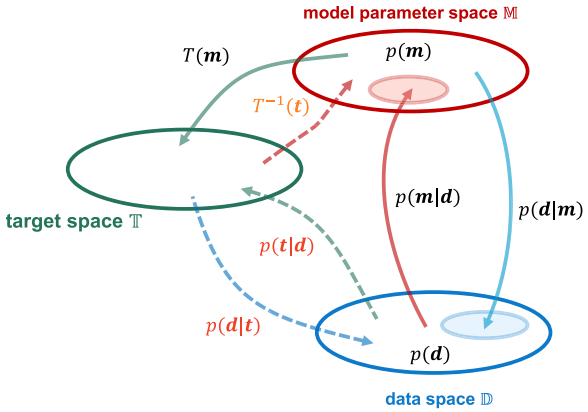

Figure A1. Schematic overview of model, data and target space and the (probabilistic) functions mapping between them.

A2 Designing experiments for interrogation problems

The objective of a scientific investigation is typically to answer a specific set of research questions. For experimental design problems, we then wish to maximize the information in a target space $\mathbb{T}$ rather than that described by the posterior distribution over model parameter space. The answers usually depend on multiple model parameter values $\boldsymbol{m}$, and as shown in Fig. A1, a target function $T(\mathbf{m}|Q)$ is defined that maps the values into a target space $\mathbb{T}$ where a question of interest Q can be answered (Arnold & Curtis 2018).

However, incorporating a target space poses challenges for data space experimental design methods since the likelihood $p(d \mid t) = \int_{\mathbb{M}} p(d \mid t, m)\, p(m)$ is typically not available directly, because it depends on model parameter values. One approach to estimate the likelihood involves conditionally sampling models from the prior distribution that map to a specific point in the target space $T^{-1} = p(m \mid t)$, and treating the model parameters as nuisance variables (Feng & Marzouk 2019). For this approach, the inverse function T^{-1} must be available, which is only the case in specific scenarios. Even if T^{-1} can be approximated, it would involve similar challenges as previously discussed for variational methods.

By contrast, model space techniques enable straightforward likelihood-free experimental design (see Appendix A1), and allow interrogation experiments to be designed without requiring T^{-1}. They can therefore be applied for any general interrogation problem, for which T or T^{-1} is computable. Only if the mapping T is linear is it possible to use linear (Bayesian) experimental design methods (Curtis 1999b; Attia *et al.* 2018; Wu *et al.* 2021)

A3 Stochastic gradient EIG optimization

Section 4 presents methods that solve eq. (14) using search algorithms where EIG calculation and optimization are carried out in separate operations. This two-step approach is a standard procedure, but it becomes increasingly difficult for higher numbers of design dimensions. Using the variational posterior method, a one-step design procedure can be constructed in which the parameters of the variational family and the design vector are optimized simultaneously (Foster *et al.* 2019b) using SGD.

SGD is a widely used optimization algorithm which allows optimizations to scale to substantially higher dimensions. The only restriction for geophysical problems in practice is the need to provide the gradients of $p(\boldsymbol{d} \mid \boldsymbol{m}, \xi)$ with respect to the design. This limitation can be challenging for some problems but is readily available for problems that are solvable analytically (AVO studies; van Den Berg *et al.* 2003, 2005) or when the forward solver can be expressed in a backwards differentiable form (Ren *et al.* 2020; Smith *et al.* 2021; Richardson 2022). Generally, any lower bound on the EIG (see Foster *et al.* 2019b; Kleinegesse & Gutmann 2021, for examples) can be straightforwardly maximized using SGD for design optimization if the gradients of $p(\boldsymbol{d} \mid \boldsymbol{m}, \xi)$ with respect to the design parameters are available or can be approximated. It is more challenging to apply upper bounds such as the NMC or variational marginal method in this one-step approach (Huan & Marzouk 2014; Foster *et al.* 2019b; Goda *et al.* 2020).