PROCEEDINGS A

royalsocietypublishing.org/journal/rspa

Research

Cite this article: Sapsis TP. 2020 Output-weighted optimal sampling for Bayesian regression and rare event statistics using few samples. *Proc. R. Soc. A* **476**: 20190834.
http://dx.doi.org/10.1098/rspa.2019.0834

Received: 30 November 2019
Accepted: 15 January 2020

Subject Areas:
statistics, applied mathematics

Keywords:
optimal experimental design, rare extreme events, Bayesian regression, optimal sampling, active learning

Author for correspondence:
Themistoklis P. Sapsis
e-mail: sapsis@mit.edu

Electronic supplementary material is available online at https://doi.org/10.6084/m9.figshare.c.4840293.

THE ROYAL SOCIETY PUBLISHING

Output-weighted optimal sampling for Bayesian regression and rare event statistics using few samples

Themistoklis P. Sapsis

Department of Mechanical Engineering, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139, USA

TPS, 0000-0003-0302-0691

For many important problems the quantity of interest is an unknown function of the parameters, which is a random vector with known statistics. Since the dependence of the output on this random vector is unknown, the challenge is to identify its statistics, using the minimum number of function evaluations. This problem can be seen in the context of active learning or optimal experimental design. We employ Bayesian regression to represent the derived model uncertainty due to finite and small number of input–output pairs. In this context we evaluate existing methods for optimal sample selection, such as model error minimization and mutual information maximization. We show that for the case of known output variance, the commonly employed criteria in the literature do not take into account the output values of the existing input–output pairs, while for the case of unknown output variance this dependence can be very weak. We introduce a criterion that takes into account the values of the output for the existing samples and adaptively selects inputs from regions of the parameter space which have an important contribution to the output. The new method allows for application to high-dimensional inputs, paving the way for optimal experimental design in high dimensions.

1. Introduction

For a wide range of problems in engineering and science it is essential to quantify the statistics of specific quantities of interest (or output) that depend on uncertain parameters (or input) with known statistical characteristics. The main obstacle towards this goal is

that this dependence is not known a priori and numerical or physical experiments need to be performed in order to specify it. If the problem at hand allows for the generation of many input–output pairs then one can employ standard regression methods to machine learn the input–output map over the support of the parameters and subsequently compute the statistics of the output.

However, for several problems of interest it is not possible to simulate even a moderate size of parameters. In this case it is critical to choose the input samples carefully so that they provide the best possible information for the output of interest [1–3]. A class of problems that belong in this family is the probabilistic quantification of extreme or rare events rising from high dimensional complex systems such as turbulence [4–8], networks [9], waves [10–12], and materials or structures [13,14]. Of course the considered set-up is not limited to extreme or rare events but it is also relevant for any problem where the aim is to quantify the input–output relationship with very few but carefully selected data points.

The described set-up is a typical example of an optimal experimental design or active learning problem [1]. Specifically, we will assume that we have already a sequence of input–output data and our goal will be to sequentially identify the next most informative input or experiment that one should perform in order to achieve fastest possible convergence for the output statistics. The problem has been studied extensively using criteria relying on mutual information theory or the Kullback–Leibler (KL) divergence (e.g. [15]). More recently another criterion was introduced focusing on the rapid convergence of the output statistics [16]. A common characteristic of these methods is the large computational cost associated with the resulting optimization problem that constrains applicability to low-dimensional input or parameter spaces.

The first objective of this work is to understand some fundamental limitations of popular selection criteria widely used for optimal experimental design (beyond the large computational cost). Specifically, we will examine how well these criteria distinguish and promote the parameters that have the most important influence to the quantities of interest. The second objective is the formulation of a new, output-weighted selection approach that explicitly and in a controllable manner takes into account, beyond the uncertainty of each parameter, its effect on the output variables, i.e. the quantities of interest. This is an important characteristic as it is often the case that a small number of parameters controls a specific quantity of interest. The philosophy of the developed criterion is to exploit the existing samples in order to estimate which parameters are the most influential for the input and then bias the sampling process using this information. Therefore, while traditional criteria tend to estimate the regression parameters with uniform accuracy over all input parameters (even those that do not contribute to the output), the introduced criterion adaptively detects the most influential input parameters and allocates more samples to reduce the regression error over these important input directions.

Beyond its intuitive and controllable character on selecting parameter values according to their effect to the output statistics, the new criterion has a numerically tractable form which allows for easy computation of each value and gradient. The latter property allows for the employment of gradient optimization methods and therefore the applicability of the approach even in high-dimensional input spaces. We demonstrate ideas through several examples ranging from linear to nonlinear maps with low- and high-dimensional input spaces. In particular, we show that the important dependencies of given quantities of interest can be identified and quantified using a very small number of input–output pairs, allowing also for quantification of rare event statistics with minimal computational cost.

2. Set-up

Let the input vector $\mathbf{x} \in \mathbb{R}^m$ denote the set of parameters or system variables and $\mathbf{y} \in \mathbb{R}^d$ be the output vector describing the quantities of interest. The input vector can be thought of as high-dimensional with known statistics described by the probability density function (pdf) $p(\mathbf{x})$ that corresponds to mean value μ_x and covariance $\mathbf{C}_{xx}$ (or correlation $\mathbf{R}_{xx}$). In what follows we will use p to denote pdf and an index will be used only if the random variable is not automatically implied by the argument.

A map from the input to the output variables, $\mathbf{y} = \mathbf{T}(\mathbf{x})$, exists and our aim is to approximate the statistics of the output, $p(\mathbf{y})$, using the smallest possible number of evaluations of the map $\mathbf{T}$. We will assume that we have already obtained some input–output pairs, which we employ in order to optimize the selection of the next input that one should evaluate. This problem can be seen as an optimal experimental design problem where the experimental parameters that one is optimizing coincide with the random parameters. All the results/methods presented in this work can be formulated in the standard set-up of optimal experimental design in a straightforward way.

The first step of the approach is to employ a Bayesian regression model to represent the map $\mathbf{T}$. Our choice of the Bayesian framework is dictated by our need to have a priori estimates for the model error, as those will be employed for the sample selection criteria. For simplicity we will present our results for linear regression models, although the extension for regression schemes with nonlinear basis functions or Gaussian process regression schemes is straightforward. We formulate a linear regression model with an input vector $\mathbf{x}$ that multiplies a coefficient matrix $\mathbf{A} \in \mathbb{R}^{d\times m}$ to produce an output vector $\mathbf{y}$, with Gaussian noise added:

$$\left.\begin{aligned} &\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{e}, \\ &\mathbf{e} \sim \mathcal{N}(0, \mathbf{V}) \\ \text{and} \quad &p(y \mid \mathbf{x}, \mathbf{A}, \mathbf{V}) = \mathcal{N}(\mathbf{A}\mathbf{x}, \mathbf{V}). \end{aligned}\right\} \tag{2.1}$$

We emphasize that for what follows we consider the case of a known noise variance $\mathbf{V} \in \mathbb{R}^{d\times d}$. From the perspective of applications, the motivation for known variance is that for a wide range of engineering problems in mechanics, fluids, vibrations, materials, etc. there are well-established experimental methods or high-fidelity simulation methods, which are very accurate but with very large cost. For problems like these the challenge is not to estimate the output or measurement noise (which is typically estimated or calibrated beforehand) but to identify the effect of the uncertain parameters of the problem to the quantities of interest. For this reason we focus on the case of known noise variance. The case of unknown covariance matrix $\mathbf{V}$ is discussed in the electronic supplementary material, appendix D. The basic set-up involves a given dataset of pairs $D = \{(\mathbf{y}_1, \mathbf{x}_1), (\mathbf{y}_2, \mathbf{x}_2), \ldots, (\mathbf{y}_N, \mathbf{x}_N)\}$. We set $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_N]$ and $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N]$.

For the matrix $\mathbf{A}$ we assume a Gaussian prior with mean $\mathbf{M} \in \mathbb{R}^{d\times m}$ and covariance $\mathbf{K} \in \mathbb{R}^{m\times m}$ for the columns, and $\mathbf{V}$ for the rows. This has the form:

$$p(\mathbf{A}) \sim \mathcal{N}(\mathbf{M}, \mathbf{V}, \mathbf{K}) = \frac{|\mathbf{K}|^{d/2}}{|2\pi\mathbf{V}|^{m/2}} \exp\left(-\frac{1}{2}\mathrm{tr}((\mathbf{A}-\mathbf{M})^T\mathbf{V}^{-1}(\mathbf{A}-\mathbf{M})\mathbf{K})\right). \tag{2.2}$$

Then one can obtain the posterior for the matrix $\mathbf{A}$ [17,18]

$$p(\mathbf{A} \mid D, \mathbf{V}) \sim \mathcal{N}(\mathbf{S}_{yx}\mathbf{S}_{xx}^{-1}, \mathbf{V}, \mathbf{S}_{xx}), \tag{2.3}$$

where,

$$\left.\begin{aligned} &\mathbf{S}_{xx} = \mathbf{X}\mathbf{X}^T + \mathbf{K} \\ \text{and} \quad &\mathbf{S}_{yx} = \mathbf{Y}\mathbf{X}^T + \mathbf{M}\mathbf{K}. \end{aligned}\right\} \tag{2.4}$$

Essentially, $\mathbf{X}\mathbf{X}^T$ is the data correlation of the sample input points $\mathbf{x}_i$, $i = 1, \ldots, N$. We choose $\mathbf{K} = \alpha\mathbf{I}$ ($\mathbf{I}$ is the identity matrix) and $\mathbf{M} = 0$, where α is an empirical parameter that will be optimized. Therefore, the above relations take the form

$$\mathbf{S}_{xx} = \mathbf{X}\mathbf{X}^T + \alpha\mathbf{I}$$

and

$$\mathbf{S}_{yx} = \mathbf{Y}\mathbf{X}^T.$$

Based on the above we obtain the probability distribution function for new inputs $\mathbf{x}$:

$$\left.\begin{aligned} &p(\mathbf{y} \mid \mathbf{x}, D, \mathbf{V}) = \mathcal{N}(\mathbf{S}_{yx}\mathbf{S}_{xx}^{-1}\mathbf{x}, \mathbf{V}(1+c)) \\ \text{and} \quad &c = \mathbf{x}^T\mathbf{S}_{xx}^{-1}\mathbf{x}. \end{aligned}\right\} \tag{2.5}$$

Then we can obtain an estimate for the probability density function of the output as

$$p(\mathbf{y} \mid \mathbf{D}, \mathbf{V}) = \int p(\mathbf{y} \mid \mathbf{x}, D, \mathbf{V}) p(\mathbf{x}) \, \mathrm{d}\mathbf{x}. \tag{2.6}$$

It is important to emphasize that the output $\mathbf{y}$ is random due to two sources: (i) the uncertainty of the input vector $\mathbf{x}$, and (ii) the uncertainty due to the model error expressed by the term c. The latter is directly related to the choice of data points $\mathbf{x}_i$, $i = 1, \ldots, N$ and the goal is to choose these points in such a way so that the statistics of $\mathbf{y}$ converge most rapidly.

The most notable property for this model is the fact that the model error is independent of the expected output value of the system. This fact holds true also for Gaussian Process regression (GPR) schemes or regression models that use nonlinear basis functions. This property will have very important consequences when it comes to the optimal input sample selection for the modelling of the input–output relation.

(a) Properties of the data correlation S_{xx}

One can compute the eigenvectors, $\hat{\mathbf{r}}_i$, and eigenvalues, σ_i^2, of the data correlation matrix, $\mathbf{S}_{xx}$,

$$\hat{\mathbf{R}} = [\hat{\mathbf{r}}_1 | \cdots | \hat{\mathbf{r}}_m] \in \mathbb{R}^{m \times m} \quad \text{and} \quad \sigma_i^2, \quad i = 1, \ldots, m.$$

By applying a linear transformation to the $\mathbf{S}_{xx}$ eigendirections, $\mathcal{X} = \hat{\mathbf{R}}^T \mathbf{x}$ we have

$$\mathbf{x}^T \mathbf{S}_{xx}^{-1} \mathbf{x} = \sum_{i=1}^{m} \frac{\chi_i^2}{\sigma_i^2}.$$

Thus, the eigendirections of $\mathbf{S}_{xx}$ indicate the principal directions of maximum confidence for the linear model. The eigenvalues quantify this confidence: the larger the eigenvalue the slower the uncertainty increases (quadratically) as χ_i^2 increases. For a new, arbitrary point, $\mathbf{x}_{N+1} = \mathbf{h}$, added to the family of $\mathbf{x}$ points we will have $\mathbf{X}' = [\mathbf{X} \mid \mathbf{x}_{N+1}]$. By direct computation we obtain

$$\mathbf{S}'_{xx} = \sum_{i=1}^{N} \mathbf{x}_i \mathbf{x}_i^T = \mathbf{S}_{xx} + \mathbf{h}\mathbf{h}^T. \tag{2.7}$$

If the new point belongs to the j eigendirection, $\mathbf{x}_{N+1} = \kappa \mathbf{r}_j$, where $\kappa \in \mathbb{R}$ then the new data correlation will be,

$$\mathbf{S}'_{xx} = \mathbf{S}_{xx} + \kappa^2 \mathbf{r}_j \mathbf{r}_j^T.$$

It can be easily checked that under this assumption the new matrix $\mathbf{S}'_{xx}$ will have the same eigenvectors. Moreover the j eigenvalue will be $\sigma'^2_j = \sigma_j^2 + \kappa^2$, while all other eigenvalues will remain invariant. Therefore, adding one more data point along a principal direction will increase the confidence along this direction by the magnitude of this new point.

The larger the magnitude of any point we add, the larger its impact on the covariance. One can trivially increase the magnitude of the new points but this does not offer any real benefit. Moreover, in a typical realistic scenario there will be magnitude constraints. To avoid this ambiguity, typical of linear regression problems, we will fix the magnitude of the input points, i.e. $\mathbf{x} \in \mathbb{S}^{m-1} = \{\mathbf{x} \in \mathbb{R}^m, \|\mathbf{x}\| = 1\}$, so that we can assess the direction of the new points, without being influenced by the magnitude. For nonlinear problems the input points should be chosen from a compact set, typically defined by the mechanics of the specific problem.

3. Fundamental limitations of standard optimal experimental design criteria

Here we consider two popular criteria that can be employed for the selection of the next most informative input sample $\mathbf{x}_{N+1}$. The first one is based on the minimization of the model error expressed by the parameter c (equation (2.5)), while the second one is the KL divergence or equivalently the maximization of the mutual information between input and output variables, which is the standard approach in the optimal experimental design literature [1].

We hypothesize a new input point $\mathbf{x}_{N+1} = \mathbf{h}$. As the corresponding output is not a priori known we will assume that it is given by the mean regression model, $\mathbf{y}_{N+1} = \mathbf{S}_{yx}\mathbf{S}_{xx}^{-1}\mathbf{x}_{N+1}$. The new pairs of data points will be $D' = \{D, (\mathbf{x}_{N+1}, \mathbf{y}_{N+1})\}$. Under this set-up the new model error will be given by $c(\mathbf{x}; \mathbf{h}) = \mathbf{x}^T\mathbf{S}_{xx}'^{-1}\mathbf{x}$, where the new data correlation matrix is given by (2.7). In addition, the mean estimate of the new model will remain invariant since,

$$\begin{aligned}\mathbf{S}'_{yx}\mathbf{S}'^{-1}_{xx} &= [\mathbf{S}_{yx} + \mathbf{S}_{yx}\mathbf{S}_{xx}^{-1}\mathbf{h}\mathbf{h}^T][\mathbf{S}_{xx} + \mathbf{h}\mathbf{h}^T]^{-1} \\ &= \mathbf{S}_{yx}\mathbf{S}_{xx}^{-1}[\mathbf{S}_{xx} + \mathbf{h}\mathbf{h}^T][\mathbf{S}_{xx} + \mathbf{h}\mathbf{h}^T]^{-1} \\ &= \mathbf{S}_{yx}\mathbf{S}_{xx}^{-1}. \end{aligned} \tag{3.1}$$

(a) Minimization of the mean model error

The first approach we will employ is to select $\mathbf{h}$ by minimizing the mean value of the uncertainty parameter c (equation (2.5)). Using standard expressions for quadratic forms of a random variable [19] we obtain a closed expression, valid for any input distribution. More specifically, we will have:

$$\mu_c = \mathbb{E}[\mathbf{x}^T\mathbf{S}_{xx}'^{-1}\mathbf{x}] = \mathrm{tr}[\mathbf{S}_{xx}'^{-1}\mathbf{C}_{xx}] + \mu_x^T\mathbf{S}_{xx}'^{-1}\mu_x = \mathrm{tr}[\mathbf{S}_{xx}'^{-1}\mathbf{R}_{xx}]. \tag{3.2}$$

Moreover for the case of Gaussian input we also obtain [19]

$$\sigma_c^2 = \mathrm{Var}[\mathbf{x}^T\mathbf{S}_{xx}'^{-1}\mathbf{x}] = 2\mathrm{tr}[\mathbf{S}_{xx}'^{-1}\mathbf{C}_{xx}\mathbf{S}_{xx}'^{-1}\mathbf{C}_{xx}] + 4\mu_x^T\mathbf{S}_{xx}'^{-1}\mathbf{C}_{xx}\mathbf{S}_{xx}'^{-1}\mu_x. \tag{3.3}$$

We note that the model uncertainty depends only on the statistics of the input $\mathbf{x}$ (expressed through the covariance $\mathbf{C}_{xx}$) and the samples $\mathbf{X}$ (expressed through the constant (i.e. non-dependent on $\mathbf{x}$) matrix $\mathbf{S}'_{xx}$). In other words, the matrix $\mathbf{Y}$ and the output distribution play no role on the mean model uncertainty.

To understand the mechanics of selecting input samples using the mean model error we assume that $\mathbf{R}_{xx}$ is diagonal with eigenvalues $\sigma_i^2 + \mu_{x_i}^2$ $i = 1, \ldots, d$, arranged with increasing order. We also assume that samples are collected only along the principal directions of the input covariance. In this case the quantity that is minimized takes the form

$$\mu_c(\mathbf{h}) = \mathrm{tr}[\mathbf{S}_{xx}'^{-1}\mathbf{R}_{xx}] = \sum_{i=1}^{m} \frac{\sigma_i^2 + \mu_{x_i}^2}{n_i + \delta_{ik}}, \quad h_i = \delta_{ik} \in \mathbb{S}^{m-1},$$

where n_i denotes the number of samples in the ith direction. One should choose $\mathbf{h}$, or equivalently k, according to the value of the derivative of $\mu_c(\mathbf{h})$. In particular,

$$h_i = \delta_{ik}, \quad k = \arg\min_i \left(-\frac{\sigma_i^2 + \mu_{x_i}^2}{n_i^2} \right).$$

If all directions have been sampled with an equal number (e.g. each of the directions have $n_i = 1$), sampling will continue with the most uncertain direction. After sufficient sampling in this direction, the addition of a new sample will cause a smaller effect than sampling the next most important direction and this is when the scheme will change sampling direction. This behaviour guarantees that the scheme will never get 'trapped' in one direction. It will continuously evolve, as more samples in one direction lead to a very small eigenvalue of $\mathbf{S}_{xx}'^{-1}\mathbf{R}_{xx}$ along this direction, and therefore sampling along another input direction will cause a bigger contribution to the trace.

It is clear that sampling based on the uncertainty parameter c searches only in $\mathbf{x}$-directions with important uncertainty, while the impact of each input variable is completely neglected. *Therefore, even directions that have zero effect on the output variable will still be sampled as long as they are uncertain.*

(b) Maximization of the mutual information

An alternative approach for the selection of a new sample, $\mathbf{x}_{N+1} = \mathbf{h} \in \mathbb{S}^{m-1}$, is maximizing the entropy transfer or mutual information between the input and output variables, when a new sample is added [1]:

$$\mathcal{I}(\mathbf{x}, \mathbf{y} \mid D', \mathbf{V}) = \mathcal{E}_x + \mathcal{E}_{y \mid D'} - \mathcal{E}_{x,y \mid D'}. \tag{3.4}$$

where each of the entropies above are defined as,

$$\mathcal{E}_x = \int p(\mathbf{x}) \log p(\mathbf{x}) \, \mathrm{d}\mathbf{x}, \quad \mathcal{E}_{y \mid D'} = \iint p(\mathbf{y} \mid D', \mathbf{V}) \log p(\mathbf{y} \mid D', \mathbf{V}) \, \mathrm{d}\mathbf{y},$$
$$\mathcal{E}_{x,y \mid D'} = \iint p(\mathbf{y}, \mathbf{x} \mid D', \mathbf{V}) \log p(\mathbf{y}, \mathbf{x} \mid D', \mathbf{V}) \, \mathrm{d}\mathbf{x} \, \mathrm{d}\mathbf{y}.$$

This is also equivalent to maximizing the mean value of the KL divergence [2]

$$\begin{aligned}
\mathbb{E}^y[D_{KL}[p(\mathbf{x} \mid \mathbf{y}, D') || p(\mathbf{x})]] &= \int_y \int_x p(\mathbf{x} \mid \mathbf{y}, D', \mathbf{V}) \log \frac{p(\mathbf{x} | \mathbf{y}, D', \mathbf{V})}{p(\mathbf{x})} \, \mathrm{d}\mathbf{x} p(\mathbf{y} \mid D', \mathbf{V}) \, \mathrm{d}\mathbf{y} \\
&= \int_y \int_x p(\mathbf{x}, \mathbf{y} \mid D', \mathbf{V}) \log \frac{p(\mathbf{x}, \mathbf{y} \mid D', \mathbf{V})}{p(\mathbf{x}) p(\mathbf{y} \mid D', \mathbf{V})} \, \mathrm{d}\mathbf{x} \, \mathrm{d}\mathbf{y} \\
&= \mathcal{I}(\mathbf{x}, \mathbf{y} \mid D', \mathbf{V}).
\end{aligned}$$

We first compute the entropy of $p(\mathbf{x}, \mathbf{y} \mid D', \mathbf{V})$:

$$\begin{aligned}
\mathcal{E}_{x,y}(\mathbf{h}) &= \iint p(\mathbf{y}, \mathbf{x} \mid D', \mathbf{V}) \log p(\mathbf{y}, \mathbf{x} \mid D', \mathbf{V}) d\mathbf{x} d\mathbf{y} \\
&= \iint p(\mathbf{y} \mid \mathbf{x}, D', \mathbf{V}) p(\mathbf{x}) \log p(\mathbf{y} \mid \mathbf{x}, D', \mathbf{V}) \, \mathrm{d}\mathbf{x} \, \mathrm{d}\mathbf{y} + \iint p(\mathbf{y} \mid \mathbf{x}, D', \mathbf{V}) p(\mathbf{x}) \log p(\mathbf{x}) \, \mathrm{d}\mathbf{x} \, \mathrm{d}\mathbf{y} \\
&= \int \mathcal{E}_{y \mid x}(\mathbf{x}; \mathbf{h}) p(\mathbf{x}) \, \mathrm{d}\mathbf{x} + \int p(\mathbf{x}) \log p(\mathbf{x}) \, \mathrm{d}\mathbf{x} \\
&= \mathbb{E}^x[\mathcal{E}_{y \mid x}(\mathbf{x}; \mathbf{h})] + \mathcal{E}_x.
\end{aligned}$$

We focus on computing the first term on the right-hand side. For the linear regression model, the conditional error follows a Gaussian distribution. From standard expressions about the entropy of a multivariate Gaussian we have

$$\mathcal{E}_{y \mid x}(\mathbf{x}; \mathbf{h}) = \frac{1}{2} \log(1 + c)^d |2\pi e \mathbf{V}| = \frac{d}{2} \log(1 + c(\mathbf{x}; \mathbf{h})) + \frac{1}{2} \log |2\pi e \mathbf{V}|.$$

Therefore,

$$\mathbb{E}^x[\mathcal{E}_{y \mid x}(\mathbf{x}; \mathbf{h})] = \frac{d}{2} \mathbb{E}^x[\log(1 + c(\mathbf{x}; \mathbf{h}))] + \frac{1}{2} \log |2\pi e \mathbf{V}|.$$

In the general case, we cannot compute the entropy of the output, conditional on D'. To this end, the mutual information of the input and output, conditioned on D', takes the form

$$\mathcal{I}(\mathbf{x}, \mathbf{y} \mid D', \mathbf{V}) = \mathcal{E}_y(\mathbf{h}) - \frac{d}{2} \mathbb{E}^x[\log(1 + c(\mathbf{x}; \mathbf{h}))] - \frac{1}{2} \log |2\pi e \mathbf{V}|. \tag{3.5}$$

This expression is valid for any input distribution and relies only on the assumption of Bayesian linear regression. To compute the involved terms, one has to perform a Monte Carlo or importance sampling approach, even for linear regression models and Gaussian inputs. This, of course, limits the applicability of the approach to very low-dimensional input spaces. We note that the above expression is valid for the case of known noise variance, $\mathbf{V}$. The case of unknown variance $\mathbf{V}$ is considered in the electronic supplementary material, appendix D.

(i) Gaussian approximation of the output

To overcome this computational obstacle one can consider an analytical approximation of the mutual entropy, assuming Gaussian statistics for the output. This assumption is not true in general, even for Gaussian input, because of the multiplication of the (Gaussian) uncertain model parameters (matrix $\mathbf{A}$) with the Gaussian input (vector $\mathbf{x}$).

We focus on the computation of the entropy of the output $\mathbf{y}$, so that we can derive an expression for the mutual information. We will approximate the pdf for $\mathbf{y}$ through its second-order statistics. Given that the input variable is Gaussian and the exact model is linear the Gaussian approximation for the output is asymptotically accurate. Still, it will help us to obtain an understanding of how the criterion works to select new samples.

We express the covariance of the output variable using the law of total variance

$$\mathbf{C}_{yy}(D', \mathbf{V}) = \mathbb{E}^x[\mathbf{C}_{yy\,|\,x}(D', \mathbf{V})] + \mathrm{cov}[\mathbb{E}^y(\mathbf{y}\,|\,\mathbf{x}, D', \mathbf{V})]. \tag{3.6}$$

The first term is the average of the updated conditional covariance of the output variables and it is capturing the regression error. The second term expresses the covariance due to the uncertainty of the input variable $\mathbf{x}$, as measured by the estimated regression model using the input data in D'.

As we pointed out earlier the mean model using either D or D' remains invariant. Therefore, we have

$$\mathbf{C}_{yy}(D', \mathbf{V}) = \mathbf{V}(1 + \mu_c(\mathbf{h})) + \mathbf{S}_{yx}\mathbf{S}_{xx}^{-1}\mathbf{C}_{xx}\mathbf{S}_{xx}^{-1}\mathbf{S}_{yx}^{T}.$$

In this way we have the approximated entropy of the output variable using a Gaussian approximation, which is also an upper bound for any other non-Gaussian distribution with the same second-order statistics

$$\mathcal{E}_y(\mathbf{h}) = \frac{1}{2}\log|\mathbf{V}(1 + \mu_c(\mathbf{h})) + \mathbf{S}_{yx}\mathbf{S}_{xx}^{-1}\mathbf{C}_{xx}\mathbf{S}_{xx}^{-1}\mathbf{S}_{yx}^{T}| + \frac{d}{2}\log(2\pi e).$$

Therefore, we have the second-order statistics approximation of the mutual information in terms of the new sample $\mathbf{h} \in \mathbb{S}^{m-1}$, denoted as $\mathcal{I}_G$:

$$\begin{aligned}\mathcal{I}_G(\mathbf{x}, \mathbf{y}\,|\,D', \mathbf{V}) = &\frac{1}{2}\log|\mathbf{V}(1 + \mu_c(\mathbf{h})) + \mathbf{S}_{yx}\mathbf{S}_{xx}^{-1}\mathbf{C}_{xx}\mathbf{S}_{xx}^{-1}\mathbf{S}_{yx}^{T}| - \frac{1}{2}\log|\mathbf{V}| \\ &- \frac{d}{2}\mathbb{E}^x[\log(1 + c(\mathbf{x}; \mathbf{h}))].\end{aligned} \tag{3.7}$$

We observe that the second-order approximation of the mutual information criterion has minimal dependence on the output samples $\mathbf{Y}$. Specifically, (3.7) depends on the uncertainty parameter c and its statistical moments, as well as the term $\mathbf{S}_{yx}\mathbf{S}_{xx}^{-1}\mathbf{C}_{xx}\mathbf{S}_{xx}^{-1}\mathbf{S}_{yx}^{T}$. However, the latter is not coupled with the new hypothetical point $\mathbf{h}$ and to this end the minimization of this criterion does not guarantee that the output values will be taken into account in a meaningful way. Instead, the selection of the new sample depends primarily on minimizing $\mu_c = \mathrm{tr}[\mathbf{S}_{xx}'^{-1}\mathbf{R}_{xx}]$, always under the constraint $\|\mathbf{h}\| = 1$, a process that depends exclusively on the current samples $\mathbf{X}$ and the statistics of the input $\mathbf{x}$.

Therefore, regions of $\mathbf{x}$ associated with large or important values of the output $\mathbf{y}$ are not emphasized by this sampling approach and the emphasis is given in regions that minimize the mean model error μ_c. We note that these conclusions are valid for the second-order approximation of the mutual information criterion, with known output variance, σ_V^2. When one considers the mutual information criterion with unknown output variance which has to be inferred using conjugate priors, there is dependence of the criterion on the output vector $\mathbf{Y}$. However, this dependence may be very weak or even zero depending on inference parameters that are optimized based on the data. See the electronic supplementary material, appendix D, for details.

(c) Nonlinear basis regression

Similar conclusions can be made for the case where one uses nonlinear basis functions. In this case we assume that the input points 'live' within a compact set. Specifically, let the input $\mathbf{x} \in \mathcal{X} \subset \mathbb{R}^m$ be expressed as a function of another input $\mathbf{z} \in \mathcal{Z} \subset \mathbb{R}^s$ where the input value has distribution $p(\mathbf{z})$ and $\mathcal{Z}$ is a compact set. One can choose a set of basis functions

$$\mathbf{x} = \phi(\mathbf{z}). \tag{3.8}$$

In this case the distribution of the output values will be given by :

$$\left.\begin{aligned} p(\mathbf{y} \mid \mathbf{z}, D, \mathbf{V}) &= \mathcal{N}(\mathbf{S}_{y\phi}\mathbf{S}_{\phi\phi}^{-1}\phi(\mathbf{z}), \mathbf{V}(1+c)) \\ \text{and} \qquad c &= \phi(\mathbf{z})^T\mathbf{S}_{\phi\phi}^{-1}\phi(\mathbf{z}). \end{aligned}\right\} \tag{3.9}$$

The mean of the model uncertainty parameter $c = \phi(\mathbf{z})\mathbf{S}_{\phi\phi}^{-1}\phi(\mathbf{z})^T$ will become

$$\mu_c = \mathrm{tr}[\mathbf{S}_{\phi\phi}^{-1}\mathbf{C}_{\phi\phi}] + \mu_\phi^T\mathbf{S}_{\phi\phi}^{-1}\mu_\phi = \mathrm{tr}[\mathbf{S}_{\phi\phi}^{-1}\mathbf{R}_{\phi\phi}], \tag{3.10}$$

where

$$\mathbf{S}_{\phi\phi} = \sum_{i=1}^{N} \phi(\mathbf{z}_i)\phi(\mathbf{z}_i)^T \quad \text{and} \quad \mathbf{S}'_{\phi\phi} = \mathbf{S}_{\phi\phi} + \phi(\mathbf{h})\phi(\mathbf{h})^T,$$

and

$$\mu_\phi = \int \phi(\mathbf{z})p(\mathbf{z})\,\mathrm{d}\mathbf{z} \quad \text{and} \quad \mathbf{C}_{\phi\phi} = \int (\phi(\mathbf{z}) - \mu_\phi)(\phi(\mathbf{z}) - \mu_\phi)^T p(\mathbf{z})\,\mathrm{d}\mathbf{z}.$$

Following the same steps as we did for the linear model, we will have, first for the conditional entropy (assuming that the model noise in the nonlinear case is Gaussian)

$$\mathcal{E}_{y|z}(\mathbf{z}; \mathbf{h}) = \frac{1}{2}\log(1+c)^d|2\pi e\mathbf{V}| = \frac{d}{2}\log(1 + c(\phi(\mathbf{z}); \mathbf{h})) + \frac{1}{2}\log|2\pi e\mathbf{V}|.$$

Therefore,

$$\mathbb{E}^z[\mathcal{E}_{y|z}(\mathbf{z}; \mathbf{h})] = \frac{d}{2}\mathbb{E}^z[\log(1 + c(\phi(\mathbf{z}); \mathbf{h})] + \frac{1}{2}\log|2\pi e\mathbf{V}|.$$

The exact expression for the mutual information for the nonlinear case will be:

$$\mathcal{I}(\mathbf{x}, \mathbf{y} \mid D', \mathbf{V}) = \mathcal{E}_y - \frac{d}{2}\mathbb{E}^z[\log(1 + c(\phi(\mathbf{z}); \mathbf{h})] - \frac{1}{2}\log|2\pi e\mathbf{V}|. \tag{3.11}$$

To perform the second-order statistical approximation for the entropy $\mathcal{E}_y$, we follow the same steps as for the linear model case to obtain

$$\begin{aligned} \mathcal{I}_G(\mathbf{x}, \mathbf{y} \mid D', \mathbf{V}) = \frac{1}{2}\log|\mathbf{V}(1 + \mu_c(\mathbf{h})) + \mathbf{S}_{y\phi}\mathbf{S}_{\phi\phi}^{-1}\mathbf{C}_{\phi\phi}\mathbf{S}_{\phi\phi}^{-1}\mathbf{S}_{y\phi}^T| - \frac{1}{2}\log|\mathbf{V}| \\ - \frac{d}{2}\mathbb{E}^z[\log(1 + c(\phi(\mathbf{z}); \mathbf{h})]. \end{aligned} \tag{3.12}$$

The sampling strategy is more complicated in this case due to the nonlinearity of the basis elements. However, even in the present set-up the sampling depends exclusively on the statistics of the input variable $\mathbf{z}$ and the form of the basis elements ϕ. The measured output values of the modelled process do not enter explicitly into the optimization procedure for the next sample, in the same fashion with the linear model.

4. Optimal sample selection considering the output values

We saw that selecting input samples based on either the mean model error or the mutual information does not effectively take into account the output values of the existing samples. Our goal is to develop an approach that (i) will give emphasis on the output values of the existing samples, and (ii) will be computationally tractable. In [16] a similar problem was considered

where the goal was to design a sampling method that will accelerate the convergence of the pdf in regions associated with rare events. In particular the following steps were followed in [16]:

(1) Using the existing samples the authors obtain an estimate of the map (denote it as $\mathbf{y}_0(\mathbf{x})$), as well as the map-estimation-error, $\sigma^2_{y_0}(\mathbf{x})$, at every point $\mathbf{x}$.
(2) This map-estimate and its error are then used to estimate the output pdf, denoted as $p_{y_0}(\mathbf{y})$, as well as the pdf of the perturbed map along the direction of the map-estimation-error, denoted as $p_{y_0^+}(\mathbf{y})$.
(3) A new hypothetical sample point, $\mathbf{h}$, was assumed and its impact first on the map-estimation-error, $\sigma^2_{y_0^+}(\mathbf{x};\mathbf{h})$, and then on the pdf of the perturbed map, $p_{y_0^+}(\mathbf{y};\mathbf{h})$, was quantified.
(4) Then the goal was to select the new sample that minimizes the distance of the two pdfs, i.e. between $p_{y_0^+}(\mathbf{y};\mathbf{h})$ and $p_{y_0}(\mathbf{y})$. As a distance the authors considered the L_1 difference of the logarithms of the two pdfs, instead of the KL divergence. The reason was that in the KL divergence the difference of the logarithms is multiplied with the pdf itself and therefore rare events play a less important role on the value of the criterion. By considering only the difference of the logarithms gave more emphasis in the regions associated with rare events.

The approach was very effective on computing the rare event properties (tails of the pdf) for arbitrary quantities of interest with a very few samples. However, it was limited by the large cost related to the computation of the two pdfs mentioned above, which was performed with direct Monte Carlo methods. For this reason the method could be applied in problems with relatively low-input dimensionality.

In the present work we are going to build on [16] to derive a new criterion that follows the same principles as the one just described but it is also computationally tractable and can be used beyond the context of rare events, i.e. for general optimal experimental design problems with a large number of parameters.

Specifically, we are going to apply the following steps:

(1) We will employ an asymptotic form of the criterion in [16] that will provide, analytically, the distance of the pdf logarithms, i.e. the pdf of the map-estimate and the pdf of the asymptotically perturbed map-estimate.
(2) Using standard inequalities for norms of derivatives we will bound the asymptotic form of the criterion by a more intuitive and tractable form. This new form has an interesting interpretation as it naturally weights the importance of the estimated map-error by the pdf of the input but also the inverse of the pdf of the output. In this way more emphasis is given to inputs associated with large values of the output.
(3) The final step of our analysis is to demonstrate how the derived bound can be analytically approximated in terms of second-order properties of the input pdf, as well as second-order properties of the estimated output. This last step is the most cumbersome but also crucial in order to apply the method in high-dimensional problems, as the analytical approximation of the criterion allows for the application of gradient optimization methods.

In figure 1 we provide a sketch of the main idea. A two-dimensional input space is shown where the output is a function that primarily depends on one input variable. The input variable has important variance in both dimensions. While methods relying on mutual information give emphasis primarily on the statistics of the input variable, resulting in an equally good approximation of the map over all input dimensions through a uniform coverage of all input dimensions, an output-weighted approach assesses the importance of each candidate input sample by its effect on the statistics of the output, i.e. the quantity of interest. In this way the output values for the observable interest are taken into account explicitly and in a controllable manner.

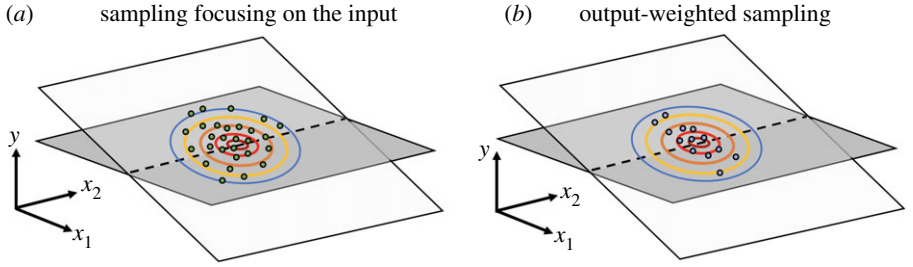

Figure 1. A schematic of an input space (grey plane), the pdf of the input (coloured contours), and the output surface (white plane). A criterion that only takes into account the characteristics of the input, $\mathbf{x}$, focuses on sampling the input directions according to their variance (*a*). However, not all input dimensions have the same effect on the output, $\mathbf{y}$. Here only x_1 contributes to the output. By using a regression trained with the existing samples we quantify the effect of each input direction to the output and select more samples from the important input dimensions (*b*). This expedites the convergence of the output statistics, p_y. (Online version in colour.)

Derivation of the output-weighted criterion

Our goal is to compute samples that accelerate the convergence of the output statistics, expressed by the probability density function, $p(\mathbf{y})$. To measure how well this convergence has occurred, we are going to rely on the distance between the probability density function of the mean model

$$\mathbf{y}_0 = \mathbf{S}_{yx}\mathbf{S}_{xx}^{-1}\mathbf{x}, \tag{4.1}$$

and the perturbed model along the most important direction of the model uncertainty (dominant eigenvector of $\mathbf{V}$), denoted as $\mathbf{r}_\mathbf{V}$:

$$\mathbf{y}_+ = \mathbf{S}_{yx}\mathbf{S}_{xx}^{-1}\mathbf{x} + \beta\mathbf{r}_V(1 + \mathbf{x}^T\mathbf{S}_{xx}'^{-1}\mathbf{x}), \tag{4.2}$$

where $\beta \ll 1$ is a small scaling factor. The corresponding probability density functions, $p_{y_0}(\mathbf{y})$ and $p_{y_+}(\mathbf{y})$ will differ only due to errors of the Bayesian regression, which vary as $\mathbf{h}$ changes. It is therefore meaningful to select the next sample that will minimize their distance. Moreover, as we are interested on capturing the probability density function equally well in regions of low and large probability we will consider the difference between the logarithms. Specifically, we define, [16],

$$D_{\mathrm{Log}^1}(\mathbf{y}_+\|\mathbf{y}_0;\mathbf{h}) = \int_{S_y} |\log p_{y_+}(\mathbf{y}) - \log p_{y_0}(\mathbf{y})|\,\mathrm{d}\mathbf{y}, \tag{4.3}$$

where S_y is a finite domain over which we are interested to define the criterion. Note that the latter has to be finite in order to have a bounded value for this distance. It can be chosen so that it contains several standard deviations of the output process. The defined criterion focuses exactly on our goal, which is the convergence of the output statistics, while the logarithm guarantees that even low probability regions have converged as well. This criterion for selecting samples was first defined in [16] and it was shown that it results in a very effective strategy for sampling processes related to low-probability extreme events. However, it is also associated with a very expensive optimization problem that has to be solved in order to minimize this distance. Apart from the cost, its complicated form does not allow for the application of gradient methods for optimization and therefore it is practical only for low-dimensional input spaces where non-gradient methods can be applied. Here one of our goals is to study its relationship with existing criteria. We are also aiming to bound it by a more trackable form that is applicable for gradient optimization methods.

To study the relationship of the criterion (4.3) with the KL divergence, we note that for bounded probability density functions the following inequality holds

$$D_{KL}(\mathbf{y}_0\|\mathbf{y}_+;S_y) = \int_{S_y}(\log p_{y_+}(\mathbf{y}) - \log p_{y_0}(\mathbf{y}))p_{y_0}(\mathbf{y})\,\mathrm{d}\mathbf{y} \le \kappa D_{\mathrm{Log}^1}(\mathbf{y}_+\|\mathbf{y}_0), \tag{4.4}$$

where κ is a constant. To this end, the criterion based on the difference of the logarithms is more conservative (i.e. harder to minimize) compared with the KL divergence (defined over the same domain).

Our next goal is to bound the D_{Log^1} criterion by one that is more tractable to optimize. We consider the criterion (4.3) for an asymptotically small value of β. This form of the criterion essentially expresses the infinitesimal difference between the mean model and the infinitesimally perturbed model by $\beta\sigma_y^2$. To compute analytically the value of the criterion for $\beta \to 0$ we employ an asymptotic result originally obtained in [16] for the study of the criterion for a large number of input samples, i.e. very small σ_y^2. For this case, or equivalently the case where β is very small that we are interested in here, we have the asymptotic form (Theorem 1 in [16])

$$D_{\mathrm{Log}^1}(\mathbf{y}_+\|\mathbf{y}_0;\mathbf{h}) \simeq \beta \int_{S_y} \frac{|(\mathrm{d}/\mathrm{d}s)\mathbb{E}[\sigma_y^2(\mathbf{x};\mathbf{h})\cdot \mathbf{1}_{y_0(x)=s}]|}{p_{y_0}(\mathbf{s})}\,\mathrm{d}\mathbf{s}, \tag{4.5}$$

where,

$$\sigma_y^2(\mathbf{x};\mathbf{h}) = \mathrm{tr}(\mathrm{cov}[\mathbf{y}\,|\,\mathbf{x}, D']) = \mathrm{tr}(\mathbf{V})(1 + c(\mathbf{x};\mathbf{h})),$$

is the conditional variance (on $\mathbf{x}$) if the output is scalar or the trace of the output conditional covariance matrix in the general case, while $\mathbf{y}_0(\mathbf{x})$ is the mean model from the input–output data collected so far.

Using standard inequalities for the derivatives of differential functions one can bound the derivative in (4.5). Specifically, if the function $\mathbb{E}[\sigma_y^2(\mathbf{x})\cdot \mathbf{1}_{y_0(x)=s}]$ has a uniformly bounded second derivative (with respect to a hypothetical new point $\mathbf{h}$), and $p_{y_0}(\mathbf{s})$ has no zeros or singular points, there exists a constant κ_0 such that ([20], Theorem 3.13, p. 109)

$$\int_{S_y} \frac{|(\mathrm{d}/\mathrm{d}s)\mathbb{E}[\sigma_y^2(\mathbf{x})\cdot \mathbf{1}_{y_0(x)=s}]|}{p_{y_0}(\mathbf{s})}\,\mathrm{d}\mathbf{s} \le \kappa_0 \int_{S_y} \frac{\mathbb{E}[\sigma_y^2(\mathbf{x})\cdot \mathbf{1}_{y_0(x)=s}]}{p_{y_0}(\mathbf{s})}\,\mathrm{d}\mathbf{s}. \tag{4.6}$$

Moreover,

$$\int_{S_y} \frac{\mathbb{E}[\sigma_y^2(\mathbf{x})\cdot \mathbf{1}_{y_0(x)=s}]}{p_{y_0}(\mathbf{s})}\,\mathrm{d}\mathbf{s} = \mathbb{E}\left[\frac{\sigma_y^2(\mathbf{x})}{p_{y_0}(\mathbf{y}_0(\mathbf{x}))}\,\middle|\, S_x\right] = \int_{S_x} \frac{p_x(\mathbf{x})}{p_{y_0}(\mathbf{y}_0(\mathbf{x}))}\sigma_y^2(\mathbf{x})\,\mathrm{d}\mathbf{x},$$

where S_x is the inverse image of the domain S_y through the map, $\mathbf{y}_0(\mathbf{x})$. Based on this we obtain the output-weighted model-error criterion, which bounds (i.e. it is more conservative) the original criterion (4.3) as well as the information-based criterion:

$$Q[\sigma_y^2] = \int_{S_x} \frac{p_x(\mathbf{x})}{p_{y_0}(\mathbf{y}_0(\mathbf{x}))}\sigma_y^2(\mathbf{x};\mathbf{h})\,\mathrm{d}\mathbf{x}. \tag{4.7}$$

In practice, S_x is chosen as $\mathbb{R}^m$ or the support of the input pdf, p_x. Because of the inequality (4.6) we can conclude that convergence of $Q[\sigma_y^2]$ also implies convergence of the metric $D_{\mathrm{Log}^1}(\mathbf{y}_+\|\mathbf{y}_0)$. However, the Q criterion is much easier to compute compared with $D_{\mathrm{Log}^1}(\mathbf{y}_+\|\mathbf{y}_0)$ and it can be employed even in high-dimensional input spaces. With the modified criterion the output data and their pdf are taken into account explicitly. In particular, the conditional variance (or uncertainty) of the model at each input point $\mathbf{x}$ is weighted by the probability of the input at this point, $p_x(\mathbf{x})$, as well as the inverse of the *estimated* probability of the output at the same input point, $p_{y_0}(\mathbf{y}_0(\mathbf{x}))$.

The term in the denominator comes as a result of considering the distance between the logarithms in (4.3). If we had started with $D_{KL}(\mathbf{y}_0\|\mathbf{y}_+;S_y)$, we would have cancellation of this important term. Note that a relevant approach, based on the heuristic superposition of the outcome and the mutual information criterion, was presented in [21]. However, there is no clear way how the two terms should be weighted or in what sense the outcome can be superimposed to the information content. We emphasize that the presented framework is not restricted to linear

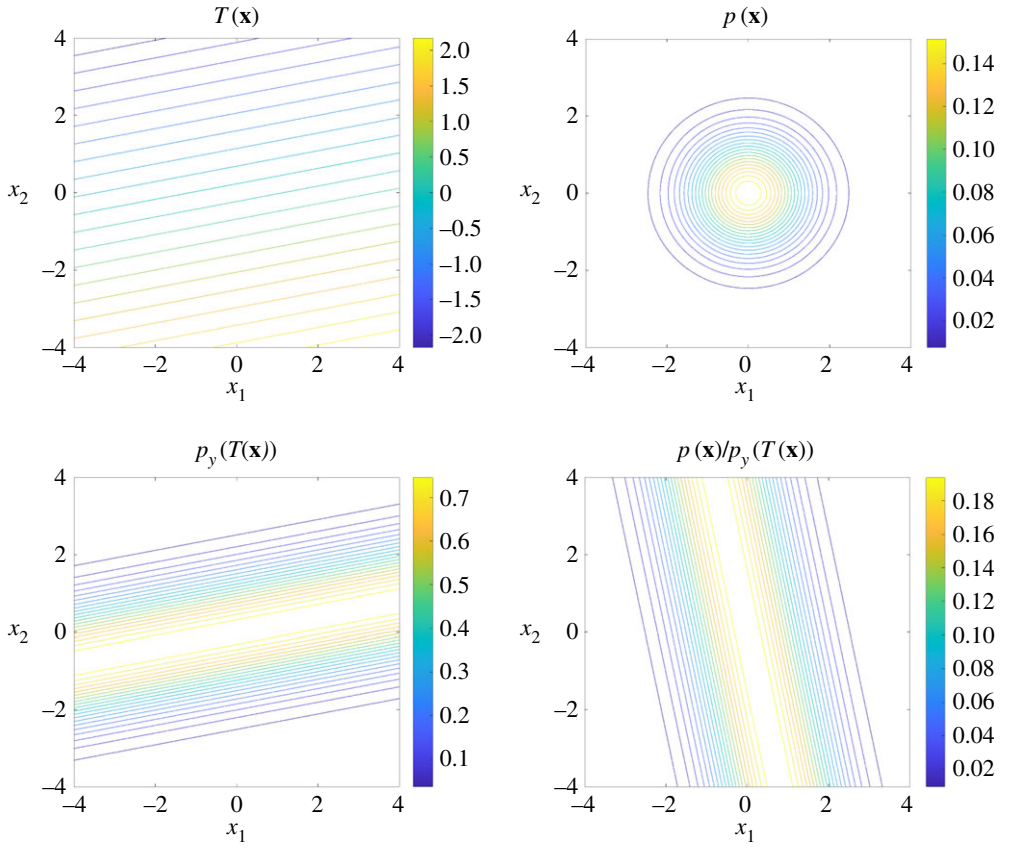

Figure 2. Illustration of the criterion for sample selection. The map, $T(\mathbf{x})$, as well as the input pdf, $p(\mathbf{x})$, are shown in the top row, while the conditional pdf of the output, $p_y(T(\mathbf{x}))$, and the weight used for sampling, $p(\mathbf{x})/p_y(T(\mathbf{x}))$, are shown in the bottom row. (Online version in colour.)

regression problems and it can also be applied to Bayesian deep learning problems (a task that will not be considered in this work). In addition, we have not made any assumption for the distribution of the input $\mathbf{x}$.

A simple demonstration

To illustrate the properties of the new criterion we consider the map

$$T(\mathbf{x}) = 0.1x_1 - 0.5x_2, \quad \text{where } \mathbf{x} \sim \mathcal{N}(0, \mathbf{I}).$$

Note that the x_2 variable is more important than x_1 in determining the value of the output, given that the two input variables have the same variance. It is therefore intuitive to require more accuracy for the second direction. However, the information distance or entropy-based criteria take into account only the input variable statistics to select the next sample, in which case, both directions will have equal importance. This is illustrated in figure 2, where we present contours of the exact map, $T(\mathbf{x})$, as well as of the input pdf, $p(\mathbf{x})$. We also present the contours of the output pdf conditional on the input, $p_y(T(\mathbf{x}))$ (bottom left), and the weight that is used in the criterion Q. Clearly, relying on the sampling criterion that uses only information about the input will not be able to approximate the map in the most important directions. On the other hand, we observe that the weight used in the Q criterion takes into account explicitly the importance of both the input variable statistics but also the information that one has estimated so far from the input–output samples. Here we used the exact map $T(\mathbf{x})$ to demonstrate the weight function but in a realistic scenario the estimated mean model $y_0(\mathbf{x})$ will be used to approximate the output pdf.

(a) Approximation of the criterion for the symmetric output pdf

Our efforts will now focus on the efficient approximation of the criterion Q. To simplify the presentation we will focus on the scalar output case, $d = 1$. The first step of the approximation focuses on the denominator. This term introduces the dependence to the output data and acts as a weight to put more emphasis on regions associated with large deviations of $y(\mathbf{x})$ from its mean. We will approximate the weight, $p_{y_0}^{-1}(y)$, by a quadratic function that optimally represents it over the region of interest, S_y. Therefore, for the scalar case we will have

$$p_{y_0}^{-1}(y) \simeq p_1 + p_2(y - \mu_y)^2, \tag{4.8}$$

where p_1, p_2 are constants chosen so that the above expression approximates the inverse of the output pdf optimally over the region of interest. We use this expression into the definition of Q (equation (4.7)) and obtain the approximation

$$Q[\sigma_y^2] \simeq p_1 \int p(\mathbf{x})\sigma_y^2(\mathbf{x})\,\mathrm{d}\mathbf{x} + p_2 \int (y_0(\mathbf{x}) - \mu_{y_0})^2 p(\mathbf{x})\sigma_y^2(\mathbf{x})\,\mathrm{d}\mathbf{x}. \tag{4.9}$$

Note that the first term does not depend on the output values but only on the input process. It is essentially the same term that appears in the entropy-based criteria. The second term however depends explicitly on the deviation of the output process from its mean and therefore on the output data. Specifically, it has large values in regions of $\mathbf{x}$ where the output process has important deviations from its mean, essentially promoting the sampling of these regions. The two constants p_1, p_2 provide the relative weight between the two contributions. They are computed for a Gaussian approximation of the output pdf in the electronic supplementary material, appendix B. For the case where the pdf p_y is expected to have important skewness, i.e. asymmetry around its mean, a linear term can be included in the expansion of $p_{y_0}^{-1}(y)$, so that this asymmetry is reflected in the sampling process.

(b) Linear regression with Gaussian input

For the case of linear regression the first term in the criterion (4.9) will take the form

$$\int p(\mathbf{x})\sigma_y^2(\mathbf{x};\mathbf{h})\,\mathrm{d}\mathbf{x} = \sigma_V^2(1 + \mu_c(\mathbf{h})) = \sigma_V^2(1 + \mathrm{tr}[\mathbf{S}_{xx}'^{-1}\mathbf{R}_{xx}]), \tag{4.10}$$

where we have considered the case of a scalar output with $\mathbf{V} = \sigma_V^2$. The second term of the criterion (4.9) will take the form

$$\begin{aligned}\frac{1}{\sigma_V^2}\int (y_0(\mathbf{x}) - \mu_{y_0})^2 p(\mathbf{x})\sigma_y^2(\mathbf{x};\mathbf{h})\,\mathrm{d}\mathbf{x} &= \int [(\mathbf{S}_{yx}\mathbf{S}_{xx}^{-1}(\mathbf{x} - \mu_x)]^2(1 + \mathbf{x}^T\mathbf{S}_{xx}'^{-1}\mathbf{x})p(\mathbf{x})\,\mathrm{d}\mathbf{x} \\ &= c_0 + \int (\mathbf{x} - \mu_x)^T\mathbf{S}_{xx}^{-1}\mathbf{S}_{yx}^T\mathbf{S}_{yx}\mathbf{S}_{xx}^{-1}(\mathbf{x} - \mu_x)\mathbf{x}^T\mathbf{S}_{xx}'^{-1}\mathbf{x}p(\mathbf{x})\,\mathrm{d}\mathbf{x},\end{aligned}$$

where c_0 is a constant that does not depend on $\mathbf{h}$

$$c_0 = \int [(\mathbf{S}_{yx}\mathbf{S}_{xx}^{-1}(\mathbf{x} - \mu_x)]^2 p(\mathbf{x})\,\mathrm{d}\mathbf{x} = \mathrm{tr}[\mathbf{S}_{xx}^{-1}\mathbf{S}_{yx}^T\mathbf{S}_{yx}\mathbf{S}_{xx}^{-1}\mathbf{C}_{xx}].$$

We observe that the second term depends on fourth-order moments of the input process $\mathbf{x}$ but also on the output values of the samples Y. This term can be computed in a closed form for the case of Gaussian input. Specifically,

$$\begin{aligned}&\int (\mathbf{x} - \mu_x)^T\mathbf{S}_{xx}^{-1}\mathbf{S}_{yx}^T\mathbf{S}_{yx}\mathbf{S}_{xx}^{-1}(\mathbf{x} - \mu_x)\mathbf{x}^T\mathbf{S}_{xx}'^{-1}\mathbf{x}p(\mathbf{x})\,\mathrm{d}\mathbf{x} \\ &\quad= \int \mathbf{x}'^T\mathbf{S}_{xx}^{-1}\mathbf{S}_{yx}^T\mathbf{S}_{yx}\mathbf{S}_{xx}^{-1}\mathbf{x}'\mathbf{x}'^T\mathbf{S}_{xx}'^{-1}\mathbf{x}'p(\mathbf{x}')\,\mathrm{d}\mathbf{x}' \\ &\qquad+ \int \mathbf{x}'^T\mathbf{S}_{xx}^{-1}\mathbf{S}_{yx}^T\mathbf{S}_{yx}\mathbf{S}_{xx}^{-1}\mathbf{x}'\mathbf{x}'^T\mathbf{S}_{xx}'^{-1}\mu_x p(\mathbf{x}')\,\mathrm{d}\mathbf{x}'\end{aligned}$$

$$+\int \mathbf{x}'^T\mathbf{S}_{xx}^{-1}\mathbf{S}_{yx}^T\mathbf{S}_{yx}\mathbf{S}_{xx}^{-1}\mathbf{x}'\mu_x^T\mathbf{S}_{xx}'^{-1}\mathbf{x}'p(\mathbf{x}')\,\mathrm{d}\mathbf{x}'$$

$$+\int \mathbf{x}'^T\mathbf{S}_{xx}^{-1}\mathbf{S}_{yx}^T\mathbf{S}_{yx}\mathbf{S}_{xx}^{-1}\mathbf{x}'\mu_x^T\mathbf{S}_{xx}'^{-1}\mu_x p(\mathbf{x}')\,\mathrm{d}\mathbf{x}',$$

where $\mathbf{x}' = \mathbf{x} - \mu_x$ and $p(\mathbf{x}')$ is the zero-mean translation pdf of the original one. The second and third term on the right-hand side vanish as they consist of third-order central moments of a Gaussian random variable. For the first term we employ a theorem for the covariance of quadratic forms, which gives for two symmetric matrices, A and B [19]:

$$\mathrm{cov}(\mathbf{x}^TA\mathbf{x}, \mathbf{x}^TB\mathbf{x}) = 2\mathrm{tr}(AC_{xx}BC_{xx}) + 4\mu_xAC_{xx}B\mu_x.$$

Therefore,

$$\mathbb{E}[\mathbf{x}^TA\mathbf{x}\mathbf{x}^TB\mathbf{x}] = 2\mathrm{tr}(AC_{xx}BC_{xx}) + 4\mu_xAC_{xx}B\mu_x - \mathrm{tr}(AC_{xx})\mathrm{tr}(BC_{xx}).$$

From this equation, it follows,

$$\int \mathbf{x}'^T\mathbf{S}_{xx}^{-1}\mathbf{S}_{yx}^T\mathbf{S}_{yx}\mathbf{S}_{xx}^{-1}\mathbf{x}'\mathbf{x}'^T\mathbf{S}_{xx}'^{-1}\mathbf{x}'p(\mathbf{x}')\,\mathrm{d}\mathbf{x}'$$

$$= 2\mathrm{tr}[\mathbf{S}_{xx}^{-1}\mathbf{S}_{yx}^T\mathbf{S}_{yx}\mathbf{S}_{xx}^{-1}\mathbf{C}_{xx}\mathbf{S}_{xx}'^{-1}\mathbf{C}_{xx}] - c_0\,\mathrm{tr}[\mathbf{S}_{xx}'^{-1}\mathbf{C}_{xx}].$$

In addition, the last term becomes

$$\int \mathbf{x}'^T\mathbf{S}_{xx}^{-1}\mathbf{S}_{yx}^T\mathbf{S}_{yx}\mathbf{S}_{xx}^{-1}\mathbf{x}'\mu_x^T\mathbf{S}_{xx}'^{-1}\mu_x p(\mathbf{x}')\,\mathrm{d}\mathbf{x}' = \mu_x^T\mathbf{S}_{xx}'^{-1}\mu_x\,\mathrm{tr}[\mathbf{S}_{xx}^{-1}\mathbf{S}_{yx}^T\mathbf{S}_{yx}\mathbf{S}_{xx}^{-1}\mathbf{C}_{xx}] = c_0\mu_x^T\mathbf{S}_{xx}'^{-1}\mu_x.$$

We collect all the computed terms and obtain

$$Q(\mathbf{h})\frac{1}{\sigma_V^2} = p_1(1 + \mathrm{tr}[\mathbf{S}_{xx}'^{-1}\mathbf{C}_{xx}] + \mu_x^T\mathbf{S}_{xx}'^{-1}\mu_x) + p_2c_0(1 + \mu_x^T\mathbf{S}_{xx}'^{-1}\mu_x - \mathrm{tr}[\mathbf{S}_{xx}'^{-1}\mathbf{C}_{xx}])$$

$$+ 2p_2\mathrm{tr}[\mathbf{S}_{xx}^{-1}\mathbf{S}_{yx}^T\mathbf{S}_{yx}\mathbf{S}_{xx}^{-1}\mathbf{C}_{xx}\mathbf{S}_{xx}'^{-1}\mathbf{C}_{xx}]. \tag{4.11}$$

This is the form of the Q criterion under the assumption of Gaussian input for the case of linear regression. For the case of zero mean input it becomes

$$Q(\mathbf{h})\frac{1}{\sigma_V^2} = (p_1 - p_2c_0)\,\mathrm{tr}[\mathbf{S}_{xx}'^{-1}\mathbf{C}_{xx}] + 2p_2\,\mathrm{tr}[\mathbf{S}_{xx}^{-1}\mathbf{S}_{yx}^T\mathbf{S}_{yx}\mathbf{S}_{xx}^{-1}\mathbf{C}_{xx}\mathbf{S}_{xx}'^{-1}\mathbf{C}_{xx}] + \mathrm{const.} \tag{4.12}$$

The coefficients p_1, p_2 are determined using the output pdf of the estimated model through the samples D (equation (4.8)), i.e. the pdf of $y_0(\mathbf{x})$. Note that the exact form of the output pdf, used in the criterion, is not important at this stage as it only defines the weights of the criterion $Q(\mathbf{h})$. For a Gaussian approximation of the output process the coefficients are given in the electronic supplementary material, appendix B.

(c) Nonlinear regression with Gaussian input

For the case of regression with nonlinear basis the first term in the criterion (4.9) will take the form

$$\int p(\mathbf{z})\sigma_y^2(\mathbf{z};\mathbf{h})\,\mathrm{d}\mathbf{z} = \sigma_V^2(1 + \mu_c(\mathbf{h})) = \sigma_V^2(1 + \mathrm{tr}[\mathbf{S}_{\phi\phi}'^{-1}\mathbf{C}_{\phi\phi}] + \mu_\phi^T\mathbf{S}_{\phi\phi}'^{-1}\mu_\phi), \tag{4.13}$$

where we have considered the case of a scalar output with $\mathbf{V} = \sigma_V^2$. The second term of the criterion (4.9) will take the form

$$\frac{1}{\sigma_V^2}\int (y_0(\mathbf{z}) - \mu_{y_0})^2p(\mathbf{z})\sigma_y^2(\mathbf{z};\mathbf{h})\,\mathrm{d}\mathbf{x} = \int [(\mathbf{S}_{y\phi}\mathbf{S}_{\phi\phi}^{-1}(\phi(\mathbf{z}) - \mu_\phi)]^2(1 + \phi(\mathbf{z})^T\mathbf{S}_{\phi\phi}'^{-1}\phi(\mathbf{z}))p(\mathbf{z})\,\mathrm{d}\mathbf{z}$$

$$= \int [(\mathbf{S}_{y\phi}\mathbf{S}_{\phi\phi}^{-1}(\phi - \mu_\phi)]^2(1 + \phi^T\mathbf{S}_{\phi\phi}'^{-1}\phi)p(\phi)\,\mathrm{d}\phi,$$

where we expressed the integral using the pdf for the basis elements ϕ. In this way the integral is now expressed exactly as in the linear regression case. To obtain a closed approximation

we approximate the pdf for ϕ through its second-order statistics (i.e. approximate $p(\phi)$ with a Gaussian pdf). The analysis shown for the linear case with Gaussian input is then valid leading to the following expression for the Q criterion:

$$Q(\mathbf{h})\frac{1}{\sigma_V^2} = p_1(1 + \mathrm{tr}[\mathbf{S}'^{-1}_{\phi\phi}\mathbf{C}_{\phi\phi}] + \mu_\phi^T\mathbf{S}'^{-1}_{\phi\phi}\mu_\phi) + p_2c_0(1 + \mu_\phi^T\mathbf{S}'^{-1}_{\phi\phi}\mu_\phi - \mathrm{tr}[\mathbf{S}'^{-1}_{\phi\phi}\mathbf{C}_{\phi\phi}]) + 2p_2\mathrm{tr}[\mathbf{S}^{-1}_{\phi\phi}\mathbf{S}^T_{y\phi}\mathbf{S}^{-1}_{\phi\phi}\mathbf{C}_{\phi\phi}\mathbf{S}'^{-1}_{\phi\phi}\mathbf{C}_{\phi\phi}]. \tag{4.14}$$

So, for a given basis $\phi(\mathbf{z})$ one needs first to obtain the mean vector μ_ϕ and covariance $C_{\phi\phi}$ using the expressions in §3c and then follow the same steps as in the linear case. The expression for the gradient of the Q criterion, under general choice of $\phi(\mathbf{z})$ is given in the electronic supplementary material, appendix A.

5. Examples

(a) Linear map with a 2D input space

To demonstrate the properties of the new criterion we first consider the two-dimensional problem

$$T(\mathbf{x}) = \hat{a}_1x_1 + \hat{a}_2x_2 + \epsilon, \quad \text{where } \mathbf{x} \sim \mathcal{N}\left(0, \begin{bmatrix}\sigma_1^2 & 0\\ 0 & \sigma_2^2\end{bmatrix}\right) \text{ and } \sigma_V^2 = 0.05. \tag{5.1}$$

We consider two cases of parameters

— case I: $\hat{a}_1 = 0.8$, $\hat{a}_2 = 1.3$, and $\sigma_1^2 = 1.4$, $\sigma_2^2 = 0.6$.
— case II: $\hat{a}_1 = 0.01$, $\hat{a}_2 = 2.0$, and $\sigma_1^2 = 2.0$, $\sigma_2^2 = 0.2$.

The two cases are presented in figure 3 in polar coordinates. The black arrows indicate the principal directions of the input covariance, scaled according to the eigenvalues of the covariance matrix, while the green arrow indicates the direction of the gradient of the map $T(\mathbf{x})$. While for the first case the contributions of both input variables to the output are comparable and thus sampling is important for both of them, for case II the contribution of the first input variable is negligible. However, this input variable is the one with the highest uncertainty.

For each case we assess four adaptive sampling strategies according to the criteria:

(1) The directly computed mutual information, $\mathcal{I}(\mathbf{x}, \mathbf{y} \mid D')$ is maximized in $\mathbb{S}^1$,
(2) The second-order statistical approximation of the mutual information, $\mathcal{I}_G(\mathbf{x}, \mathbf{y} \mid D')$ is maximized in $\mathbb{S}^1$,
(3) The uncertainty parameter, $\mu_c(\mathbf{h})$ is minimized in $\mathbb{S}^1$,
(4) The output-weighted model error criterion $Q(\mathbf{h})$ is minimized in $\mathbb{S}^1$.

For the Q criterion we choose $p_1 = 0$ and $p_2 = 1$ to emphasize the role of the second, new term that takes into account the output samples. This case of parameters corresponds to the case where we optimally approximate $p_{y_0}^{-1}$ over the full real axis, i.e. $\beta \to \infty$ using the notation of the electronic supplementary material, appendix B. We denote this criterion as Q_∞. We also compare with a Monte Carlo approach where samples are randomly generated from the input distribution of $\mathbf{x}$ and then normalized so they belong in $\mathbb{S}^1$.

For the adaptive strategies based on μ_c and Q_∞ we use the analytical expressions (3.2) and (4.12), respectively, together with their gradient computed in the electronic supplementary material, appendix A. This allowed us to use gradient-based optimization methods. For the adaptive strategies based on the mutual information, $\mathcal{I}(\mathbf{x}, \mathbf{y} \mid D, \mathbf{h})$ and its second-order approximation, $\mathcal{I}_G(\mathbf{x}, \mathbf{y} \mid D, \mathbf{h})$, we used a random sampling approach and equations (3.5) and (3.7), respectively. Specifically, we generated 10^5 samples from the input distribution $\mathbf{x}$ and used the

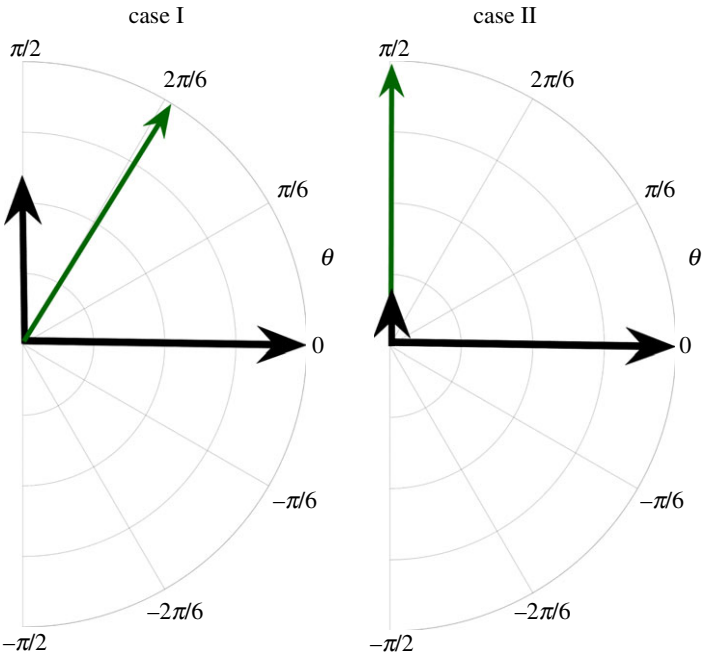

Figure 3. Black arrows indicate direction and magnitude of the principal directions (and corresponding eigenvalues) of the input covariance $\mathbf{C}_{xx}$ for each case of parameters. The green arrow indicates the gradient of the map $T(\mathbf{x})$. (Online version in colour.)

exact expression:

$$\mathbb{E}^x[\mathcal{E}_{y\,|\,x}(\mathbf{x};\mathbf{h})] = \mathbb{E}^x[\log(1+\mathbf{x}^T\mathbf{S}'^{-1}_{xx}\mathbf{x})],$$

which was numerically computed as an ensemble average. For the computation of $\mathcal{I}$ we also generated 10^5 realizations of the vector $\mathbf{a}=(a_1,a_2)$, which according to the linear regression method follows the normal distribution (2.3):

$$p(\mathbf{a}\,|\,D',\sigma_V^2)\sim\mathcal{N}(\mathbf{S}_{yx}\mathbf{S}_{xx}^{-1},\sigma_V^2\mathbf{S}'^{-1}_{xx}).$$

Next, we computed a pdf approximation for y using the generated samples and the kernel smoothing functions method [22], and we approximated the entropy of the resulted distribution by direct numerical integration. Note that this additional step, required for $\mathcal{I}$, has a vast computational cost. Most importantly, because of the absence of an analytical expression for the gradient of $\mathcal{I}$, its application to high dimensional inputs is impossible. For this example the next sample vector was parametrized as $\mathbf{h}=[\cos(\theta),\sin(\theta)]$ and the criterion was optimized by direct selection of the maximum value over a one-dimensional grid for $\theta\in[-(\pi/2),(\pi/2)]$.

All four adaptive strategies are initiated with four random samples drawn from the $\mathbf{x}$ distribution. For each case we present the average error curve over 400 experiments, i.e. experiments with different sets of initial samples and different realizations of the observation noise (figure 4: left panels). In particular, for each of these 400 experiments and for every number of N samples, we compute the conditional output variance, $\sigma^2(y\,|\,D_N)$ (obtained using the estimated map from the N samples), and we compute its difference from the exact output variance, $\sigma^2(y)$ (obtained using the exact map). Then we average the absolute difference $|\sigma^2(y\,|\,D_N)-\sigma^2(y)|$ over all 400 experiments. The standard deviation for each error curve is also presented in the shaded region. For the four adaptive sampling strategies we also present the pdf of the orientation of the samples $\mathbf{h}=(h_1,h_2)$, i.e. $\theta=\arctan(h_2/h_1)$.

In both cases of parameters we observe that the strategy based on the Q_∞ criterion outperforms the other three adaptive strategies. This difference in performance is even more pronounced for

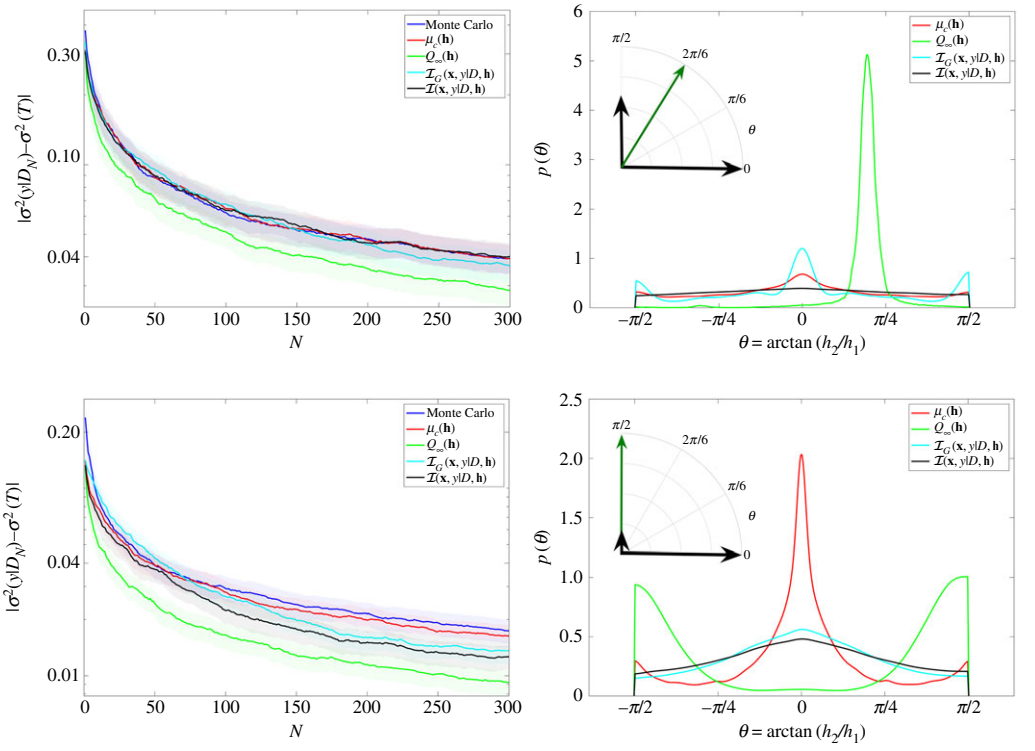

Figure 4. Comparison of the four adaptive strategies based on different criteria and the Monte Carlo method. On the left plots the average error of the output variance is shown with respect to the number of samples used over 400 experiments for each criterion. The shaded regions indicate 0.2σ based on the 400 numerical experiments. The pdf of these samples is shown for each adaptive strategy in the plots on the right. The black vectors indicate the eigenvectors of the input covariance $\mathbf{C}_{xx}$ and the green vector denotes the gradient of the exact map: $(\hat{a}_1, \hat{a}_2)$. (Online version in colour.)

case II, where one of the input variables has negligible contribution but large uncertainty. An interesting observation is that the Monte Carlo strategy performs as good as the μ_c criterion and the mutual information criteria, $\mathcal{I}$ and $\mathcal{I}_G$. This is not a surprise given that $\mathcal{I}_G$ depends primarily on μ_c and the latter is designed to give more emphasis on input directions with large uncertainty, without taking into account their expected contributions to the output, similarly with Monte Carlo. The same conclusions hold even for the full mutual information criterion, an indication that although $\mathcal{I}$ partially incorporates the output samples, it does not do it in a useful way.

Similar observations can be made if we examine the pdf for the input samples obtained from the four adaptive strategies. We can see that for each case of parameters the strategies based on μ_c, $\mathcal{I}$, and $\mathcal{I}_G$ behave very similarly and tend to place input samples in the direction of larger uncertainty. On the other hand, the Q_∞-based strategy is placing more samples in directions that compromise between large expected impact to the output variable but also with important uncertainty. We note that all the cases presented here correspond to a known output variance, σ_V^2. Results for this example corresponding to unknown variance σ_V^2 are presented in the electronic supplementary material, appendix D.

(b) A high-dimensional linear problem

The next problem to demonstrate the optimal sampling approach is a 20-dimensional linear function. Note that for this case optimization of the mutual information is an impossible task given the fact that the full expression for the mutual information is hard to optimize in the absence of expressions for its gradient.

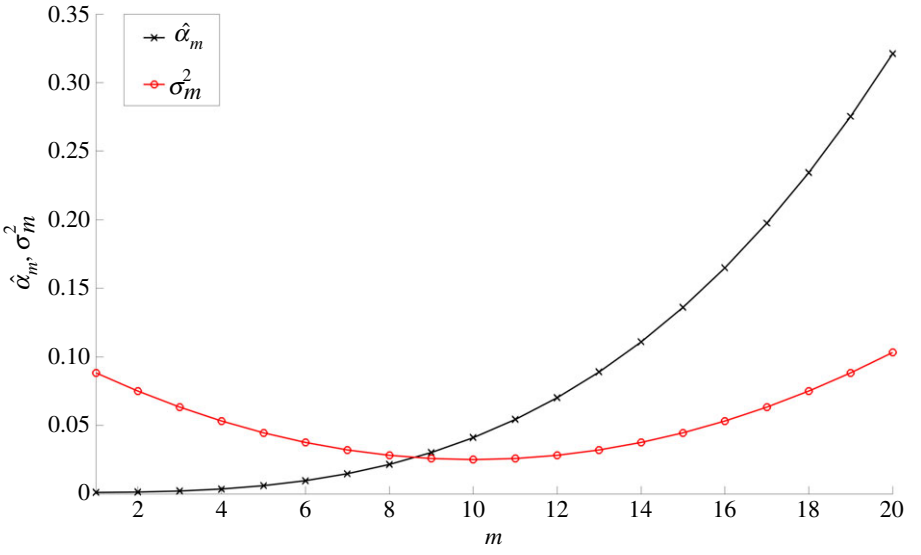

Figure 5. Coefficients, $\hat{\alpha}_m$, of the map $T(\mathbf{x})$ (black curve) plotted together with the variance of each input direction σ_m^2 (red curve) for the high dimensional problem (5.2). (Online version in colour.)

Specifically, we consider the system

$$T(\mathbf{x}) = \sum_{m=1}^{20} \hat{a}_m x_m + \epsilon, \quad \text{where } x_m \sim \mathcal{N}(0, \sigma_m^2), \quad m = 1, \ldots, 20, \tag{5.2}$$

where the coefficients and input variances are chosen as

$$\hat{a}_m = \left(1 + 40\left(\frac{m}{10}\right)^3\right) 10^{-3}, \quad m = 1, \ldots, 20$$

and

$$\sigma_m^2 = \left(\frac{1}{4} + \frac{1}{128}(m-10)^3\right) 10^{-1}, \quad m = 1, \ldots, 20.$$

This system represents a typical high-dimensional case, where we have some very influential degrees of freedom and some that have negligible impact to the output variable. The energy of these coefficients is typically not related to their influence to the output variable. In figure 5 we present the coefficients and input variances.

For the observation noise we consider two cases:

— case I: $\sigma_\epsilon^2 = 0.05$ (accurate observations)
— case II: $\sigma_\epsilon^2 = 0.5$ (noisy observations)

Given that $\sum_{m=1}^{20} \hat{a}_m^2 \sigma_m^2 = 0.0272$ the first case corresponds to relatively accurate observations while the second is a highly noisy observations case. We expect the adaptive sampling approach to be more valuable for the second case, given that for the first case we need very few samples anyway.

We apply the adaptive criteria after we have obtained one sample per input direction, to guarantee that the matrix $\mathbf{S}_{xx}$ is invertible. Then we run each numerical experiment $L = 400$ times to make sure that the randomness due to the observation noise does not favour any method.

In figure 6 we present the performance of the sampling approach based on μ_c and Q_∞, as well as a direct Monte Carlo approach. For the first case (accurate observations), shown in the left plot, we note a clear advantage of the Q_∞ sampling approach that takes into account the output samples. This advantage is more pronounced in the second case of noisy observations, where the

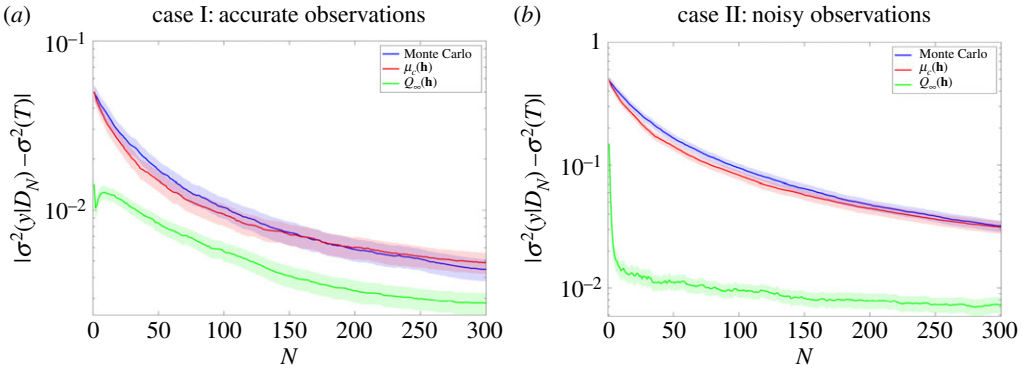

Figure 6. Performance of the two adaptive approaches based on μ_c and Q_∞ for the high-dimensional problem (5.2). The left plot (*a*) corresponds to observation noise, $\sigma_\epsilon^2 = 0.05$ (accurate observations) and the right (*b*) to $\sigma_\epsilon^2 = 0.5$ (noisy observations). (Online version in colour.)

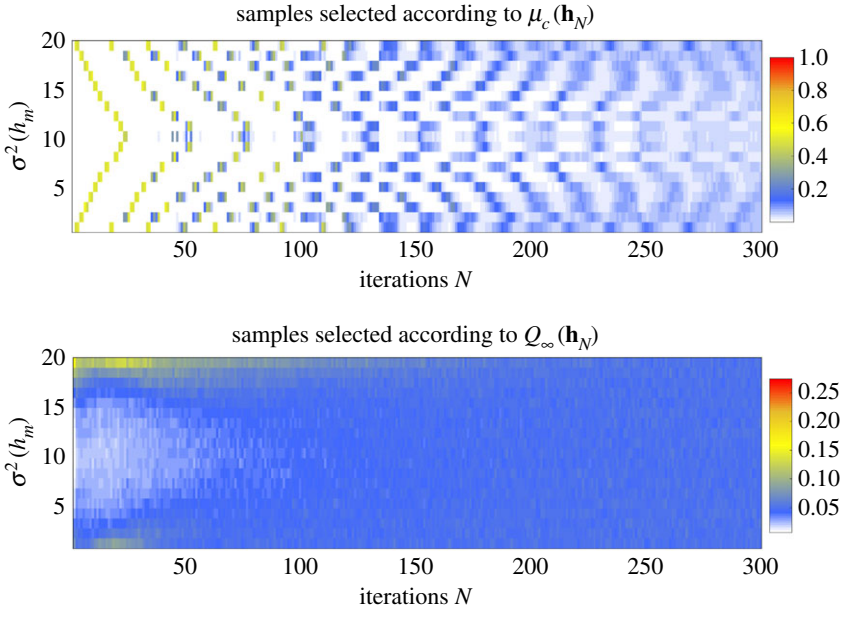

Figure 7. Energy of the different components of **h** with respect to the number of iterations *N* for case I of the high-dimensional problem. (Online version in colour.)

approach using the Q_∞ criterion obtains an order of magnitude higher accuracy from the very first samples. Note that the μ_c sampling strategy is comparable with the Monte Carlo approach, since it does not take into account the output samples.

The same conclusions can be obtained if we observe the variance of the $\mathbf{h}_N$ components, over different runs of the numerical experiments, $l = 1, \ldots, L$ (here $L = 400$), i.e. over different realizations of the observation noise:

$$\sigma^2(h_m) = \frac{1}{L}\sum_{l=1}^{L}(h_{N,m,l} - \bar{h}_{N,m})^2, \quad \text{where } \bar{h}_{N,m} = \frac{1}{L}\sum_{l=1}^{L} h_{N,m,l}, \quad m = 1, \ldots, 20. \tag{5.3}$$

Results are shown in figure 7. Sampling according to μ_c results in a distribution that is following the shape of the variance σ_m. Specifically, the scheme iteratively changes input directions based on their variance, starting from the most energetic ($m = 1$ and $m = 20$) and

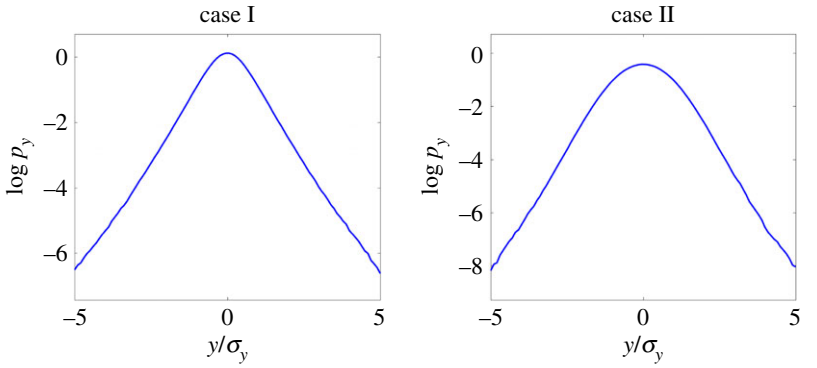

Figure 8. Exact pdf for the two cases of the nonlinear map using MC with 10^5 samples. (Online version in colour.)

moving towards the less energetic ones ($m = 10$). Then the loop begins again, until all the input directions are equally well sampled, after which point the sampling is random.

Sampling according to Q_∞, on the other hand, is performed in one loop starting from the most energetic directions, but giving more emphasis in the input directions close to $m = 20$ that have both high energy and large contribution to the output y. This 'asymmetry' in the sampling results in significantly faster convergence compared with the Monte Carlo method or the μ_c criterion.

The effect of the domain selection S_y is discussed in detail in the electronic supplementary material, appendix C, where a parametric study is also shown. Specifically, if we use a finite number of standard deviations to optimally approximate $p_{y_0}^{-1}$ (equation (4.8)), by employing Q_β (with β finite), the term μ_c in the criterion improves the behaviour of sampling for large N (electronic supplementary material, appendix C).

(c) A 2D nonlinear problem with nonlinear basis functions

The next application involves a nonlinear map with a 2D input space. Specifically, we consider the two-dimensional nonlinear problem

$$T(\mathbf{z}) = \hat{a}_1 z_1 + \hat{a}_2 z_2 + \hat{a}_3 z_1^3 + \hat{a}_4 z_2^3 + \epsilon, \quad \text{where } \mathbf{x} \sim \mathcal{N}\left(0, \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}\right) \text{ and } \sigma_V^2 = 10^{-4}. \tag{5.4}$$

We consider two cases of parameters

— case I : $\hat{a}_1 = 10^{-2}$, $\hat{a}_2 = 5$, $\hat{a}_3 = 0$, $\hat{a}_4 = 10^2$, and $\sigma_1^2 = 2.10^{-1}$, $\sigma_2^2 = 5.10^{-3}$.
— case II: $\hat{a}_1 = 10$, $\hat{a}_2 = 5$, $\hat{a}_3 = 0$, $\hat{a}_4 = 10^2$, and $\sigma_1^2 = 2.10^{-3}$, $\sigma_2^2 = 5.10^{-3}$.

In the first case the output has very weak dependence on the first variable although the latter has very large variance. Moreover, the second variable has significantly smaller variance but plays the dominant role for the output. On the other hand, for the second case, both input variables play an important role and their variance is also comparable. The exact pdf computed with an expensive Monte Carlo simulation is shown in figure 8. Both distributions are characterized by heavy tails.

We set up a nonlinear Bayesian regression scheme with the following odd basis functions:

$$\phi(\mathbf{z}) = z_1^i z_2^j, \quad (i, j) \in \{(0,1), (1,0), (1,1), (0,3), (3,0)\}. \tag{5.5}$$

This set of basis functions contains all the odd monomials with an order less or equal to 3. We observe that for the nonlinear case although the input space is two-dimensional, the regression is performed in a five-dimensional space. To avoid an ill-conditioned matrix $\mathbf{S}_{\phi\phi}$ we assume a prior with covariance $\mathbf{K} = \alpha\mathbf{I}$ (see equation (2.4)). For the cases considered we set $\alpha = 10^{-1}$. For each case we first choose randomly (using the distribution of $\mathbf{z}$) two samples. Then we use the criteria

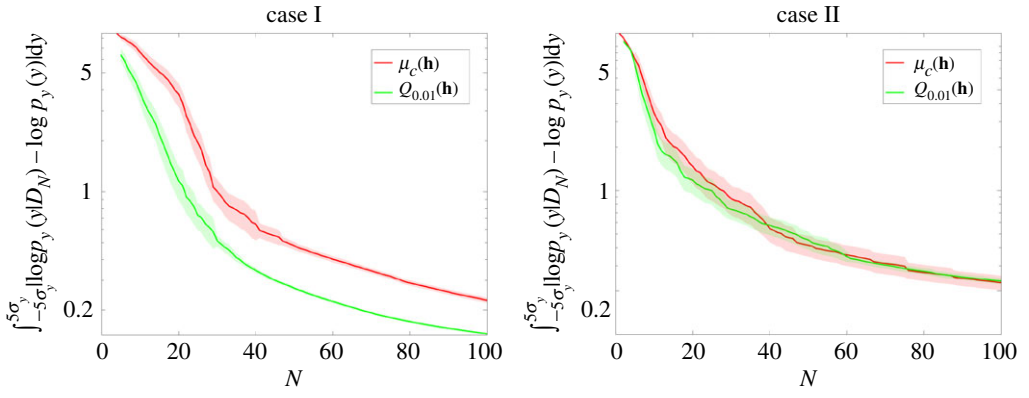

Figure 9. Performance of the two adaptive approaches based on μ_c and Q_∞ for the nonlinear, two-dimensional problem. In the second case both input directions have important roles to the output and the two methods are comparable, as expected. (Online version in colour.)

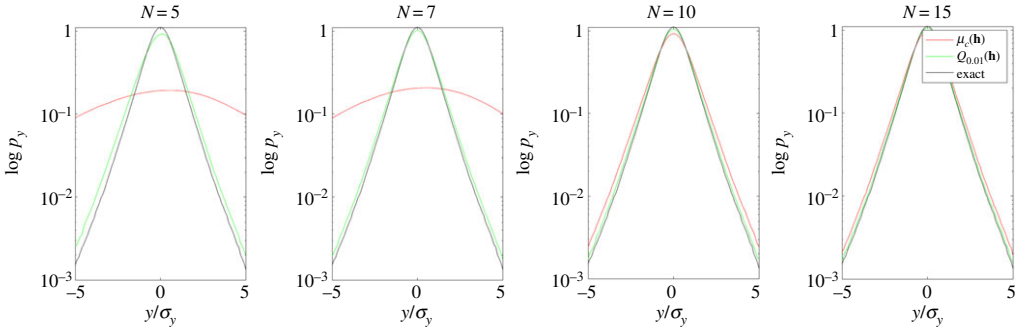

Figure 10. Performance of the two adaptive approaches based on μ_c and Q_∞ for the nonlinear, two-dimensional problem and case I parameters. The resulted pdfs for samples selected according to the two criteria are compared with the exact pdf. (Online version in colour.)

based on μ_c (equation (3.10)) and $Q_{0.01}$ (equation (4.14)) to optimize 100 samples. For each step we employed a gradient-based optimization using the expressions presented in the electronic supplementary material, appendix A and we restricted the samples in the disk to: $|\mathbf{z}| \leq 2$. For each criterion we performed 200 optimization cycles, i.e. we computed for each criterion the full sequence of 100 samples 200 times and we computed the statistics of the error (mean and standard deviation), so that the results are not sensitive on the randomness due to observational noise or the initial samples.

The convergence analysis for each criterion is presented in figure 9. The left plot shows the convergence of the two methods for the parameters of the first numerical experiment (case I). Each curve is the mean error computed from the 200 optimization cycles, while the shaded area indicates the spread across different runs. Note that for case I there is only input variable z_2 that plays a dominant role on the output, while the other input variable has negligible effect but important variance. In agreement with the results of the linear problems, the samples based on the $Q_{0.01}$ criterion achieve better performance as they rapidly align with the direction that has the most important influence on the output.

This is not the case for the samples based on the μ_c criterion that align primarily with the directions of importance variance, resulting in a slower convergence. For the case II parameters both input directions have comparable variance and comparable effect to the output. In this case,

as expected, the two criteria have comparable performance. This is clearly demonstrated by the right plot in figure 9. Finally, in figure 10 we demonstrate the convergence of the pdfs for the first case of parameters.

6. Conclusion

We have analysed fundamental limitations of popular criteria for samples selection, employed in the optimal experimental design community. These criteria are based on maximization of entropy-based quantities, typically having the form of mutual information between input and output variables. Specifically, we have shown that beyond the large computational cost associated with these criteria that restricts their applicability to very low-dimensional problems, there is weak dependence of the induced sampling process to the output values of the existing samples when the variance of the output noise is assumed to be known. Even for the case of unknown variance, although the dependence on the output values is not controllable and can become very weak. In this way, directions of the parameter space that contribute the most to the output may not be emphasized. This is not a failure of the existing criteria but rather an intrinsic property following the fact that they are designed to converge uniformly over all parameters, independently of their influence to the output.

Motivated by these limitations, we have presented a new criterion for optimally selecting training samples that significantly accelerates the convergence of Bayesian regression schemes with respect to the state of the art. The criterion explicitly takes into account the fact that different parameter values have different impacts to the output of interest, with some of them being much more influential than others. In this way, it places more samples towards the influential parameters, which are also characterized by important uncertainty. In addition, the introduced criterion is more practical to compute, compared with mutual information criteria, as its gradient can be analytically derived, allowing for the employment of gradient optimization methods. Therefore, the new method allows for the optimization of samples, even for a large number of parameters, paving the way for optimal experimental design and active learning in high dimensions.

Future work will focus on the formulation of the presented framework on the training of deep neural networks. The presented approach is expected to have an important impact in application areas such as optimal experimental design for systems where very few experiments are available (e.g. biology), adaptive sampling in complex environments with multiple objectives, uncertainty quantification and extreme event statistics in challenging problems such as fatigue-crack, coastal flooding, critical network events, and others.

Data accessibility. A public link to the code used in the manuscript has been added at the end of the manuscript.
Competing interests. I declare I have no competing interests.
Funding. This work was supported by a Doherty Career Development Chair, a Mathworks Faculty Research Innovation Fellowship and the ARO MURI Grant W911NF-17-1-0306.
Acknowledgements. The author would like to thank Prof. Munther Dahleh, Prof. Sanjoy Mitter, Dr Mustafa Mohamad and Dr Antoine Blanchard for several stimulating discussions. This work was initiated during a sabbatical visit of the author at ETH, hosted by Prof. George Haller, which is gratefully acknowledged. The detailed comments of the referees led to several improvements and are also highly appreciated.
Code. Code for the examples shown is available at https://github.com/sapsis/Output_weighted_sampling.

References

1. Chaloner K, Verdinelli I. 1995 Bayesian experimental design: a review. *Stat. Sci.* **10**, 273–304. (doi:10.1214/ss/1177009939)
2. Huan X, Marzouk YM. 2014 Gradient-based stochastic optimization methods in Bayesian experimental design. *Int. J. Uncertain. Quantif.* **4**, 479–510. (doi:10.1615/Int.J.UncertaintyQuantification.2014006730)

3. Agrawal R, Squires C, Yang K, Shanmugam K, Uhler C. 2019 ABC-strategy: budgeted experimental design for targeted causal structure discovery. In *Proc. of Machine Learning Research 89 (AISTATS 2019), Naha, Okinawa, Japan, 16–18 April*, pp. 3400–3409.
4. Qi D, Majda AJ. 2016 Predicting fat-tailed intermittent probability distributions in passive scalar turbulence with imperfect models through empirical information theory. *Comm. Math. Sci.* **14**, 1687–1722. (doi:10.4310/CMS.2016.v14)
5. Farazmand M, Sapsis TP. 2017 A variational approach to probing extreme events in turbulent dynamical systems. *Sci. Adv.* **3**, e1701533. (doi:10.1126/sciadv.1701533)
6. Sapsis TP. 2018 New perspectives for the prediction and statistical quantification of extreme events in high-dimensional dynamical systems. *Phil. Trans. R. Soc. A* **376**, 20170133. (doi:10.1098/rsta.2017.0133)
7. Blonigan PJ, Farazmand M, Sapsis TP. 2019 Are extreme dissipation events predictable in turbulent fluid flows? *Phys. Rev. Fluids* **4**, 044606. (doi:10.1103/PhysRevFluids.4.044606)
8. Brunton S, Noack B, Koumoutsakos P. 2020 Machine learning for fluid mechanics. *Annu. Rev. Fluid Mech.* **52**, 477–508. (doi:10.1146/annurev-fluid-010719-060214)
9. Sarkar T, Roozbehani M, Dahleh MA. 2018 Robustness sensitivities in large networks. In *Emerging applications of control and systems theory* (eds R Tempo, S Yurkovich, P Misra), pp. 81–92. Cham, Switzerland: Springer.
10. Cousins W, Sapsis TP. 2016 Reduced order precursors of rare events in unidirectional nonlinear water waves. *J. Fluid Mech.* **790**, 368–388. (doi:10.1017/jfm.2016.13)
11. Mohamad MA, Cousins W, Sapsis TP. 2016 A probabilistic decomposition-synthesis method for the quantification of rare events due to internal instabilities. *J. Comput. Phys.* **322**, 288–308. (doi:10.1016/j.jcp.2016.06.047)
12. Majda AJ, Moore MNJ, Qi D. 2018 Statistical dynamical model to predict extreme events and anomalous features in shallow water waves with abrupt depth change. *Proc. Natl Acad. Sci. USA* **116**, 3982–3987. (doi:10.1073/pnas.1820467116)
13. Serebrinksy S, Ortiz M. 2005 A hysteretic cohesive-law model of fatigue-crack nucleation. *Scr. Mater.* **53**, 1193–1196. (doi:10.1016/j.scriptamat.2005.07.015)
14. Fan D, Jodin G, Consi TR, Bonfiglio L, Ma Y, Keyes LR, Karniadakis GE, Triantafyllou MS. 2019 A robotic Intelligent Towing Tank for learning complex fluid-structure dynamics. *Sci. Robot.* **4**, eaay5063. (doi:10.1126/scirobotics.aay5063)
15. Pandita P, Bilionis I, Panchal J. 2019 Bayesian optimal design of experiments for inferring the statistical expectation of a black-box function. *J. Mech. Des.* **141**, 101404.
16. Mohamad MA, Sapsis TP. 2018 Sequential sampling strategy for extreme event statistics in nonlinear dynamical systems. *Proc. Natl Acad. Sci. USA* **115**, 11 138–11 143. (doi:10.1073/pnas.1813263115)
17. Rasmussen CE, Williams CKI. 2005 *Gaussian processes in machine learning*. Cambridge, MA: The MIT Press.
18. Minka T. 2010 Bayesian linear regression.
19. Rencher AC, Schaalje BG. 2008 *Linear models in statistics*, 2nd edn, New York, NY: John Wiley & Sons.
20. Kwong MK, Zettl A. 1992 *Norm inequalities for derivatives and differences*. Berlin, Germany: Springer Verlag.
21. Verdinelli I, Kadane JB. 1992 Bayesian designs for maximizing information and outcome. *J. Am. Stat. Assoc.* **87**, 510–515. (doi:10.1080/01621459.1992.10475233)
22. Hill PD. 1985 Kernel estimation of a distribution function. *Commun. Stat. Theory Methods* **14**, 605–620. (doi:10.1080/03610928508828937)

Appendix A: Gradient of trace criteria

Several criteria in this work take the form

$$\lambda[\mathbf{h}] = \mathrm{tr}[\mathbf{S}'^{-1}_{xx}\mathbf{C}], \tag{A1}$$

where $\mathbf{C}$ is a symmetric matrix and $\mathbf{S}'_{xx} = \mathbf{S}_{xx} + \mathbf{h}\mathbf{h}^T$. The gradient of this expression can be explicitly computed. We first note that

$$\frac{\partial \mathbf{S}'^{-1}_{xx}}{\partial h_k} = -\mathbf{S}'^{-1}_{xx}\frac{\partial(\mathbf{h}\mathbf{h}^T)}{\partial h_k}\mathbf{S}'^{-1}_{xx},$$

where,

$$\frac{\partial(\mathbf{h}\mathbf{h}^T)}{\partial h_k} = \delta_{ik}h_j + \delta_{kj}h_i.$$

In this way we will have

$$\begin{aligned}
\frac{\partial \lambda}{\partial h_k} &= -\mathrm{tr}[\mathbf{S}'^{-1}_{xx}\frac{\partial(\mathbf{h}\mathbf{h}^T)}{\partial h_k}\mathbf{S}'^{-1}_{xx}\mathbf{C}] \\
&= -[\mathbf{S}'^{-1}_{xx}]_{ij}(\delta_{jk}h_m + \delta_{km}h_j)[\mathbf{S}'^{-1}_{xx}]_{mn}[\mathbf{C}]_{ni} \\
&= -h_m[\mathbf{S}'^{-1}_{xx}]_{mn}[\mathbf{C}]_{ni}[\mathbf{S}'^{-1}_{xx}]_{ik} - [\mathbf{S}'^{-1}_{xx}]_{kn}[\mathbf{C}]_{ni}[\mathbf{S}'^{-1}_{xx}]_{ij}h_j \\
&= -\mathbf{h}^T\mathbf{S}'^{-1}_{xx}\mathbf{C}\mathbf{S}'^{-1}_{xx} - (\mathbf{S}'^{-1}_{xx}\mathbf{C}\mathbf{S}'^{-1}_{xx}\mathbf{h})^T \\
&= -2\mathbf{h}^T\mathbf{S}'^{-1}_{xx}\mathbf{C}\mathbf{S}'^{-1}_{xx}.
\end{aligned}$$

For the case of nonlinear regression $\mathbf{h} = \phi(\mathbf{z})$. Then

$$\frac{\partial(\phi(\mathbf{z})\phi(\mathbf{z})^T)}{\partial z_k} = \frac{\partial \phi_i}{\partial z_k}\phi_j + \frac{\partial \phi_j}{\partial z_k}\phi_i.$$

In this way we will have

$$\begin{aligned}
\frac{\partial \lambda}{\partial z_k} &= -[\mathbf{S}'^{-1}_{\phi\phi}]_{ij}(\frac{\partial \phi_j}{\partial z_k}\phi_m + \frac{\partial \phi_m}{\partial z_k}\phi_j)[\mathbf{S}'^{-1}_{\phi\phi}]_{mn}[\mathbf{C}]_{ni} \\
&= -2[\phi^T\mathbf{S}'^{-1}_{\phi\phi}\mathbf{C}\mathbf{S}'^{-1}_{\phi\phi}]_j\frac{\partial \phi_j}{\partial z_k}.
\end{aligned} \tag{A2}$$

Appendix B: Optimal approximation of p_y^{-1}

To approximate the inverse of the output pdf, $\frac{1}{p_y}$ over S_y we are going to use a least square approach. Specifically, we will assume a symmetric output density and we will employ the approximation

$$\frac{1}{p_y(y)} \simeq \frac{1}{p_y(0)} + p_2(y-\mu_y)^2. \tag{B1}$$

The constant p_2 is chosen so that we have optimal least square approximation over the interval $[\mu_y, \mu_y + \beta\sigma_y]$ where β is a fixed parameter that defines the output region of interest, S_y. By direct minimization we obtain

$$p_2 = \frac{5}{\beta^5\sigma_y^3}\int_{\mu_y}^{\mu_y+\beta\sigma_y} \frac{y^2}{p_y(y)}dy - \frac{5}{3\beta^2 p_y(\mu_y)}.$$

For the case of Gaussian output the above expression takes the form

$$p_2 = \frac{5\sqrt{2\pi}}{\beta^5\sigma_y}\left(\int_0^\beta z^2 e^{\frac{z^2}{2}}dz - \frac{\beta^3}{3}\right).$$

In this way the least square approximation over the interval $[\mu_y, \mu_y + \beta\sigma_y]$ will be

$$\frac{1}{p_y(y)} \simeq \sqrt{2\pi}\sigma_y + \frac{5\sqrt{2\pi}}{\beta^5\sigma_y}\left(\int_0^\beta z^2 e^{\frac{z^2}{2}}dz - \frac{\beta^3}{3}\right)(y-\mu_y)^2. \tag{B2}$$

We denote the criterion with the coefficients obtained from this approximation as Q_β. Specifically,

$$\begin{aligned} Q_{\beta\sigma}(\mathbf{h})\frac{1}{\sigma_V^2} &= \sqrt{2\pi}\sigma_{y_0}(1+\mathrm{tr}[\mathbf{S}_{xx}'^{-1}\mathbf{C}_{xx}] + \mu_x^T\mathbf{S}_{xx}'^{-1}\mu_x) \\ &+ \frac{5\sqrt{2\pi}}{\beta^5\sigma_{y_0}}\left(\int_0^\beta z^2 e^{\frac{z^2}{2}}dz - \frac{\beta^3}{3}\right)\left(c_0(1+\mu_x^T\mathbf{S}_{xx}'^{-1}\mu_x) + 2\mathrm{tr}[\mathbf{S}_{xx}^{-1}\mathbf{S}_{yx}^T\mathbf{S}_{yx}\mathbf{S}_{xx}^{-1}\mathbf{C}_{xx}\mathbf{S}_{xx}'^{-1}\mathbf{C}_{xx}]\right). \end{aligned} \tag{B3}$$

For large values of β we have $\lim_{\beta\to\infty}\kappa = \infty$ and the Q criterion is essentially dominated by the output-dependent term. For the case of very small β we have $\lim_{\beta\to 0}\kappa = \frac{\sqrt{2\pi}}{\sigma_y}$.

Appendix C: Effect of the weights in the Q criterion

Here we present additional results for Case I of the high dimensional system. Specifically, in Figure 11 we present the performance of the sampling algorithm according to various choices of the β parameter. The corresponding sampling patterns are shown in (Figure 12). All cases presented are averaged over $L = 500$ numerical experiments to remove the effect of observational noise.

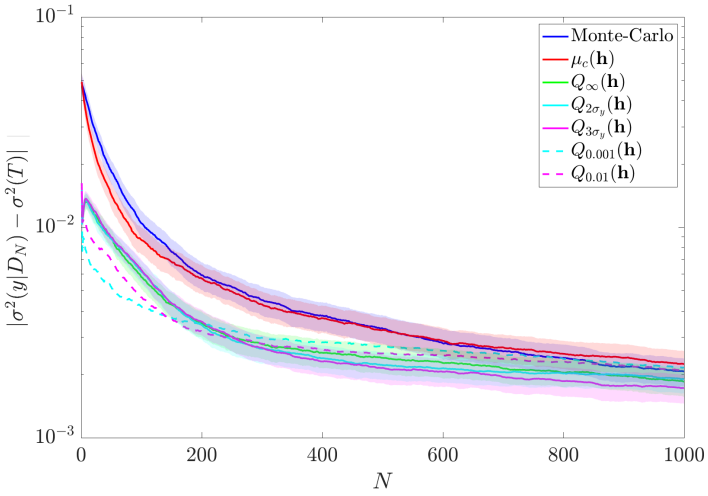

Figure 11: More detailed results for Case I of the high dimensional problem. The effect of the β parameter is shown. While for the first iterations it plays no role, asymptotically it improves the behavior of the sampling scheme.

We observe that for $\beta = 2$ or $\beta = 3$ the performance for small N is very close to the one obtained with Q_∞. For larger N, however, the performance with finite β is improved. In addition to the finite β case, we also present two cases with fixed p_1, p_2. Specifically, we have $Q_{0.01}$ representing the case $p_1 = 0.01$ and $p_2 = 1$, while $Q_{0.001}$ represents the case $p_1 = 0.001$ and $p_2 = 1$. It is interesting to observe that $Q_{0.001}$ has better performance for small N compared with all other criteria.

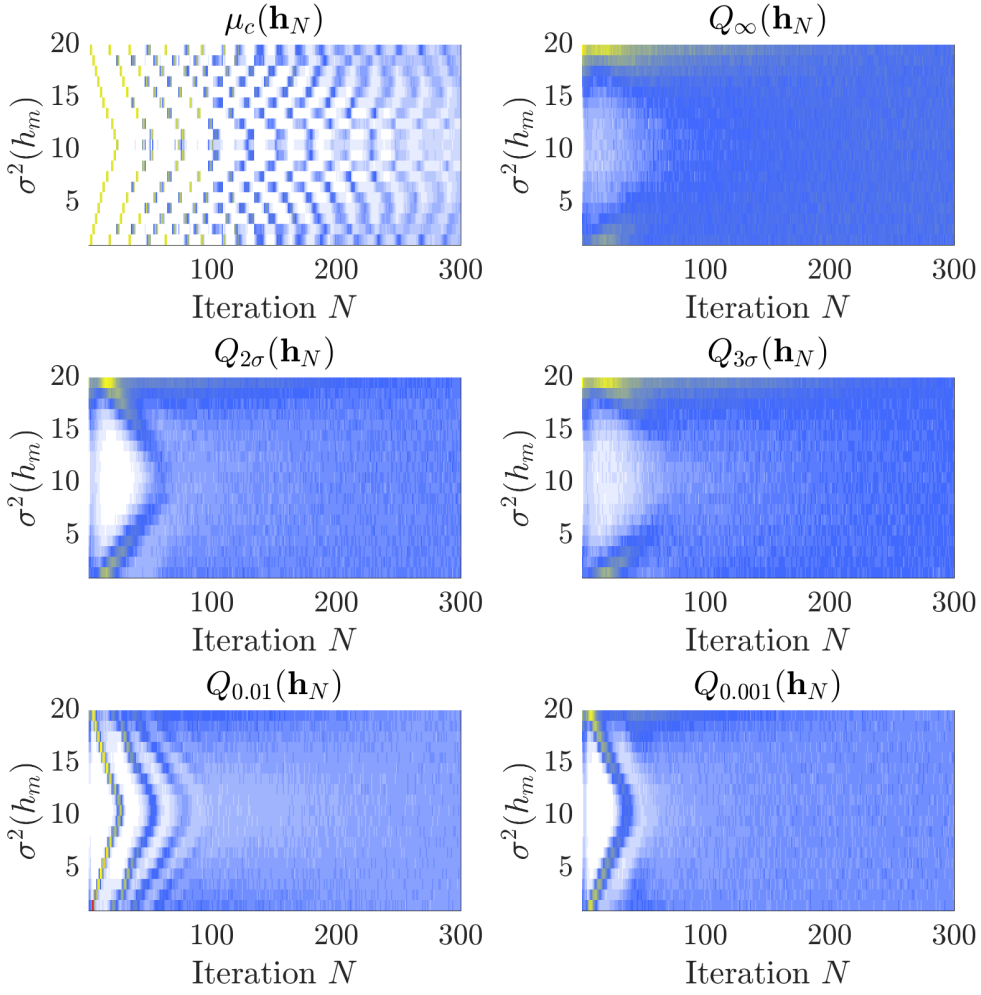

Figure 12: More detailed results for the samples of $\mathbf{h}$ with respect to the number of iteration N for Case I of the high dimensional problem. The effect of the β parameter is shown.

Appendix D: The case of unknown σ_V^2

Here we consider the case of a priori unknown covariance σ_V^2. To simplify the presentation we will restrict our analysis to scalar output and vector input. We formulate a linear regression model with an input vector $\mathbf{x}$ that multiplies a coefficient vector $\mathbf{a}$ to produce an output scalar y, with Gaussian noise added that has unknown variance:

$$\begin{aligned} y &= \mathbf{a}^T\mathbf{x} + e, \\ e &\sim \mathcal{N}(0, \sigma_V^2), \\ p(y|\mathbf{x}, \mathbf{a}, \sigma^2) &= \mathcal{N}(\mathbf{a}^T\mathbf{x}, \sigma^2). \end{aligned} \tag{D1}$$

For the vector $\mathbf{a}$ we assume a Gaussian prior with mean $\mathbf{m} = \mathbf{0}$ and covariance $\mathbf{K} = \mathbf{I}\alpha$, where α is a parameter that will be optimized with respect to the evidence. This has the form:

$$p(\mathbf{a}) \sim \mathcal{N}(0, \mathbf{I}\alpha). \tag{D2}$$

A conjugate prior for σ_V^2 is the inverse Gamma (or inverse Wishart in the multi-dimensional case):

$$p(\sigma_V^2) = \frac{q(\sigma_0^2, \nu)}{(\sigma_V^2)^{1+\frac{\nu}{2}}} \exp\left(-\frac{\sigma_0^2}{2\sigma_V^2}\right), \tag{D3}$$

where q is a normalization constant: $q = \frac{\sigma_0^\nu}{2^{\frac{\nu}{2}}\Gamma(\nu/2)}$, σ_0^2 is a prior value for σ_V^2 and ν is a parameter that is optimized via empirical Bayes.

The posterior for the unknown coefficients will be given by eq. (3), while the posterior for σ_V^2 takes the form

$$p(\sigma_V^2|D) = \frac{q(\sigma_{\mathbf{Y}|\mathbf{X}}^2 + \sigma_0^2, N + \nu)}{(\sigma_V^2)^{1+\frac{N+\nu}{2}}} \exp\left(-\frac{\sigma_{\mathbf{Y}|\mathbf{X}}^2 + \sigma_0^2}{2\sigma_V^2}\right), \tag{D4}$$

where, $\sigma_{\mathbf{Y}|\mathbf{X}}^2 = \mathbf{Y}\mathbf{Y}^T - (\alpha+1)^{-1}\mathbf{S}_{yx}\mathbf{S}_{xx}^{-1}\mathbf{S}_{yx}^T$. Multiplying the predictive distribution (5) with the posterior for σ^2 and integrating over this argument we have the predictive pdf ($t-$distributed):

$$p(\mathbf{y}|\mathbf{x}, D) = \mathcal{T}(\mathbf{S}_{yx}\mathbf{S}_{xx}^{-1}\mathbf{x}, (\sigma_{\mathbf{Y}|\mathbf{X}}^2 + \sigma_0^2)(1+c)^{-1}, N + \nu + 1) \tag{D5}$$

where the parameters (α, ν) are chosen by maximizing the evidence (set the gradient of $p(\mathbf{Y}|\mathbf{X})$ equal to zero), which leads to the following fixed-point problem:

$$\begin{aligned} \hat{\sigma}^2 &= \frac{\sigma_{\mathbf{Y}|\mathbf{X}}^2 + \nu}{N + \nu}, \\ \alpha &= \frac{m}{(\hat{\sigma}^2)^{-1}\mathbf{S}_{yx}\mathbf{S}_{xx}\mathbf{S}_{yx}^T - m}, \\ \nu^{new} &= \nu \frac{\Psi(\frac{N+\nu}{2}) - \Psi(\frac{\nu}{2})}{\log\left(\frac{\sigma_{\mathbf{Y}|\mathbf{X}}^2}{\nu} + 1\right) + (\hat{\sigma}^2)^{-1} - 1}, \end{aligned} \tag{D6}$$

where $\Psi(x) = \frac{d\log\Gamma(x)}{dx}$ is the digamma function.

Selecting inputs by maximization of the mutual information

We follow the same steps as in the known variance case. We hypothesize a new sample, $\mathbf{x}_{N+1} = \mathbf{h} \in \mathbb{S}^{m-1}$ and the goal is maximizing the entropy transfer or mutual information between the input and output variables, when this new sample is added. For this case of a priori unknown variance σ_V^2 we denote the mutual information as $\hat{\mathcal{I}}$. We will have

$$\hat{\mathcal{I}}(\mathbf{x}, \mathbf{y}|D') = \mathcal{E}_x + \mathcal{E}_{y|D'} - \mathcal{E}_{x,y|D'}. \tag{D7}$$

Following the same steps with section 3.2 we have for the entropy of $p(\mathbf{x}, \mathbf{y}|D')$:

$$\mathcal{E}_{x,y}(\mathbf{h}) = \mathbb{E}^x[\mathcal{E}_{y|x}(\mathbf{x}; \mathbf{h})] + \mathcal{E}_x.$$

We focus on computing the first term on the right hand side. In this case of a priori unknown variance the conditional output follows the $t-$student distribution (eq. (D5)). Using standard expressions for its entropy we have (setting $N' = \nu + N + 1$)

$$\begin{aligned}\mathcal{E}_{y|x}(\mathbf{x}; \mathbf{h}) = \frac{N'+1}{2}\left(\Psi\left(\frac{N'+1}{2}\right) - \Psi\left(\frac{N'}{2}\right)\right) + \log\left(\sqrt{N'}B\left(\frac{N'}{2}, \frac{1}{2}\right)\right) \\ -\frac{1}{2}\log(1 + c(\mathbf{x}; \mathbf{h})) + \frac{1}{2}\log(\sigma^2_{\mathbf{Y}'|\mathbf{X}'} + \sigma_0^2),\end{aligned}$$

where,

$$\begin{aligned}\sigma^2_{\mathbf{Y}'|\mathbf{X}'} &= \mathbf{Y}'\mathbf{Y}'^T - (\alpha+1)^{-1}\mathbf{S}'_{yx}\mathbf{S}'^{-1}_{xx}\mathbf{S}'^T_{yx} \\ &= \mathbf{Y}\mathbf{Y}^T + (\mathbf{S}_{yx}\mathbf{S}_{xx}^{-1}\mathbf{h})^2 - (\alpha+1)^{-1}\mathbf{S}_{yx}\mathbf{S}_{xx}^{-1}[\mathbf{S}_{yx} + \mathbf{S}_{yx}\mathbf{S}_{xx}^{-1}\mathbf{h}\mathbf{h}^T]^T \\ &= \sigma^2_{\mathbf{Y}|\mathbf{X}} + \frac{\alpha}{1+\alpha}(\mathbf{S}_{yx}\mathbf{S}_{xx}^{-1}\mathbf{h})^2.\end{aligned} \tag{D8}$$

Note that in the second equality we used eq. (8). In general, we cannot compute analytically the entropy of the output, conditional on D'. To this end, the mutual information of the input and output, conditioned on D', takes the form

$$\hat{\mathcal{I}}(\mathbf{x}, \mathbf{y}|D') = \mathcal{E}_y(\mathbf{h}) - \frac{1}{2}\mathbb{E}^x[\log(1 + c(\mathbf{x}; \mathbf{h}))] + \frac{1}{2}\log(\sigma^2_{\mathbf{Y}'|\mathbf{X}'}(\mathbf{h}) + \sigma_0^2) + R, \tag{D9}$$

where R are terms that do not depend on the new point $\mathbf{h}$. The second and third terms are computed for each $\mathbf{h}$ with direct Monte-Carlo using 10^5 samples. It is important to emphasize that the mutual information criterion with unknown variance, (D9), depends on the output values $\mathbf{Y}$ (through the third term on the right hand side), in contrary to the known variance case (section 3.2). However, as it can be seen from eq. (D8) this dependence can be very weak or even zero depending on the value of the parameter α which is chosen based on maximization of the evidence.

In the last expression the term with the highest computational cost is the entropy of the output (first term) as one needs to estimate the histogram of y. An analytical approximation can be obtained based on a Gaussian assumption for the output, y. In this case the mutual information takes the form:

$$\hat{\mathcal{I}}_G(\mathbf{x}, \mathbf{y}|D') = \frac{1}{2}\log(2\pi e\sigma_y^2(\mathbf{h})) - \frac{1}{2}\mathbb{E}^x[\log(1 + c(\mathbf{x}; \mathbf{h}))] + \frac{1}{2}\log(\sigma^2_{\mathbf{Y}'|\mathbf{X}'}(\mathbf{h}) + \sigma_0^2) + R, \tag{D10}$$

where $\sigma_y^2(\mathbf{h})$ is the estimated variance of the output after a new candidate input $\mathbf{h}$. This is computed with Monte-Carlo simulation of y, i.e. generate 10^5 random realizations using (D4), (3) and (D1).

$Q-$criterion with unknown output variance

We emphasize that the selection approach using the Q criterion is not modified at all for the case of unknown output variance. This is because the unknown variance σ_V^2 appears as a multiplication factor in the Q criterion (eq. (31)), i.e. σ_V^2 is re-estimated each time a new data point is added using equation (31) but its value does not modify the optimal sample $\mathbf{h}$.

Numerical comparison for the 2d linear problem

Results and direct comparison with the case of known σ_V^2 are shown in Figure 13 for the linear problem of section 5.1 with a two-dimensional input. For all methods shown we have run 400 experiments (as we did for Fig. 4). In the higher dimensional problem the approach based on direct computation of mutual information is not applicable due to the vast computational cost. As we can observe the selection process based on mutual information with unknown output variance ($\hat{\mathcal{I}}_G$ or $\hat{\mathcal{I}}$), has slightly improved performance compared with the case of known σ_V^2 ($\mathcal{I}_G$ or $\mathcal{I}$) but this improvement is observed only for very small number of samples. As N increases the criterion with unknown output variance is comparable with mutual information with given output variance ($\mathcal{I}_G$ or $\mathcal{I}$). This is because the optimal value of α tends to zero as N increases and therefore the criteria with known and unknown output variance become practically identical.

It is important to emphasize that beyond the faster convergence of the $Q-$criterion, its main advantage is the low computational cost (many orders of magnitude smaller compared with the methods based on mutual information either with direct computation or through a Gaussian approximation) which allows applicability to higher dimensional problems.

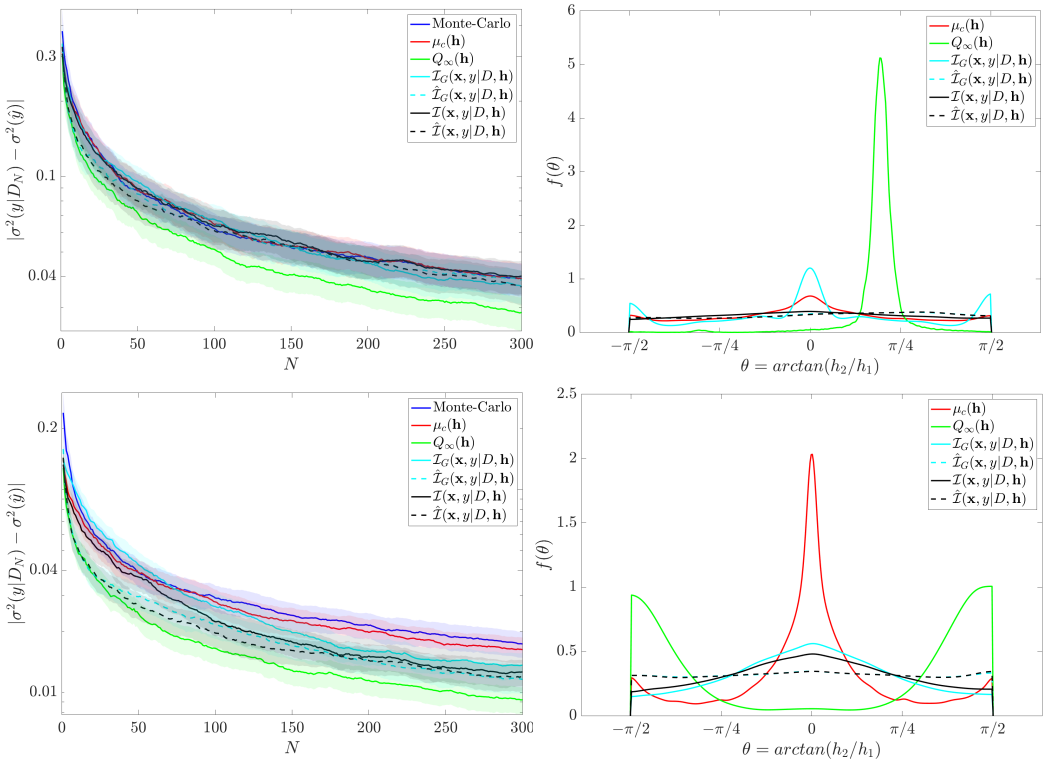

Figure 13: Comparison of selection methods based on different criteria and the Monte-Carlo method including also the case of unknown σ_V^2. Problem setup and parameters are as in section 5.1 and Figure 4. The following methods are shown: mean square error, μ_c; $Q-$criterion; Gaussian approximation of mutual information with known variance, $\mathcal{I}_G$, and with unknown variance, $\hat{\mathcal{I}}_G$; directly computed mutual information with known variance, $\mathcal{I}$, and unknown variance, $\hat{\mathcal{I}}$. The $Q-$criterion does not depend on whether the noise variance is known or estimated. Note that the pdf of samples (right plots) between $\hat{\mathcal{I}}$ and $\hat{\mathcal{I}}_G$ overlap, i.e. are indistinguishable.

Corrections to the manuscript 'Output-weighted optimal sampling for Bayesian regression and rare event statistics using few samples'

Themistoklis P. Sapsis *
Department of Mechanical Engineering,
Massachusetts Institute of Technology,
77 Massachusetts Ave., Cambridge, MA 02139

March 29, 2021

List of corrections

1. Page 11: Reference to [20] should read: 'Theorem 3.18, p. 113'.
2. After eq. (4.5), add the definition: $\frac{d}{ds} = \frac{\partial}{\partial s_1}\frac{\partial}{\partial s_2}...\frac{\partial}{\partial s_d}$.
3. Eq. (4.6) should have a square root at the right hand side:

$$\int \frac{\left|\frac{d}{ds}\mathbb{E}[\sigma^2_{\mathbf{y}}(\mathbf{x})\cdot \mathbf{1}_{\mathbf{y}_0(\mathbf{x})=\mathbf{s}}]\right|}{p_{\mathbf{y}_0}(\mathbf{s})} d\mathbf{s} \leqslant \kappa_0 \left(\int \frac{\mathbb{E}[\sigma^2_{\mathbf{y}}(\mathbf{x})\cdot \mathbf{1}_{\mathbf{y}_0(\mathbf{x})=\mathbf{s}}]}{p_{\mathbf{y}_0}(\mathbf{s})} d\mathbf{s}\right)^{1/2}.$$

*Corresponding author: sapsis@mit.edu, Tel: (617) 324-7508, Fax: (617) 253-8689