Outcome-weighted sampling for Bayesian analysis

Themis Sapsis and Antoine Blanchard

Department of Mechanical Engineering
Massachusetts Institute of Technology

Funding: ONR, AFOSR, Sloan

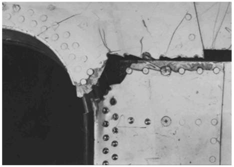
April 23, 2020

Risk Quantification

Extreme weather phenomena

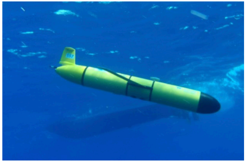
Loads/motions in FSI problems

Fatigue-crack nucleation

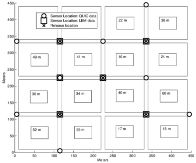


Optimization under uncertainty

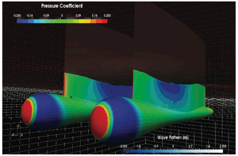
Path planning - exploration



Optimal sensor placement



Design under uncertainty



Challenges

Challenge I: High-dimensional parameter spaces

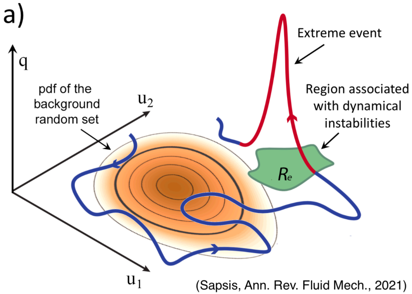
- Intrinsic instabilities
- Stochastic loads
- Random parameters

Challenge II: Need for expensive models

- Complex dynamics
- Hard to isolate dynamical mechanisms

The focus of this work

Goal: Develop sampling strategies appropriate for expensive models and high-dimensional parameter spaces



(Sapsis, Ann. Rev. Fluid Mech., 2021)

- Models in fluids: Navier-Stokes, NL Schrödinger, Euler
- Critical region of parameters is unknown
- Importance sampling based methods too expensive
- Input-space PCA focuses on subspaces, not sufficient

Risk Quantification: Problem setup

$\mathbf{x} \in \mathbb{R}^m$: Uncertain parameters; pdf: f_x
$\mathbf{y} \in \mathbb{R}^d$: Output or quantities of interest; expensive to compute

Risk Quantification Problem: Compute the statistics of y with the minimum number of experiments, i.e. input parameters $\{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N\}$

A Bayesian approach

Employ a *linear* regression model with an input vector $\mathbf{x}$ of length m that multiplies a coefficient vector $\mathbf{A}$ to produce an output vector $\mathbf{y}$ of length d, with Gaussian noise added:

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{e} \tag{1}$$

$$\mathbf{e} \sim \mathcal{N}(0, \mathbf{V}) \tag{2}$$

We are given a data set of pairs:

$D = \{(\mathbf{y}_1, \mathbf{x}_1), (\mathbf{y}_2, \mathbf{x}_2), ..., (\mathbf{y}_N, \mathbf{x}_N)\}$.

We set $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_N]$ and $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N]$.

A Bayesian approach

From Bayesian regression, we obtain the pdf for new inputs $\mathbf{x}$:

$$\begin{aligned} p(\mathbf{y}|\mathbf{x}, D, \mathbf{V}) &= \mathcal{N}(\mathbf{S}_{yx}\mathbf{S}_{xx}^{-1}\mathbf{x}, \mathbf{V}(1 + c)), \\ c &= \mathbf{x}^T\mathbf{S}_{xx}^{-1}\mathbf{x}, \\ \mathbf{S}_{xx} &= \mathbf{X}\mathbf{X}^T + \mathbf{K} \\ \mathbf{S}_{yx} &= \mathbf{Y}\mathbf{X}^T \end{aligned}$$

Question: How to choose the next input point $\mathbf{x}_{N+1} = \mathbf{h}$?

Given a hypothetical input point $\mathbf{x}_{N+1} = \mathbf{h}$, we have at $\mathbf{x}$

$$p(\mathbf{y}|\mathbf{x}, D', \mathbf{V}) = \mathcal{N}(\mathbf{S}_{yx}\mathbf{S}_{xx}^{-1}\mathbf{x}, \mathbf{V}(1+c)),$$
$$c = \mathbf{x}^T \mathbf{S}'^{-1}_{xx}\mathbf{x},$$

where $\mathbf{S}'_{yx}\mathbf{S}'^{-1}_{xx}\mathbf{x} = \mathbf{S}_{yx}\mathbf{S}_{xx}^{-1}\mathbf{x}$, assuming $\mathbf{y}_{N+1} = \mathbf{S}_{yx}\mathbf{S}_{xx}^{-1}\mathbf{h}$.

We minimize the model uncertainty by choosing $\mathbf{h}$ such that the distribution for c converges to zero (at least for the $\mathbf{x}$ we are interested):

$$\mu_c(\mathbf{h}) = \mathbb{E}[\mathbf{x}^T\mathbf{S}'^{-1}_{xx}\mathbf{x}] = tr[\mathbf{S}'^{-1}_{xx}\mathbf{C}_{xx}] + \mu_x^T\mathbf{S}'^{-1}_{xx}\mu_x = tr[\mathbf{S}'^{-1}_{xx}\mathbf{R}_{xx}]$$

(valid for any f_x)

1. The selection of the new sample does not depend on **Y**.

2. We diagonalize $\mathbf{R}_{xx}$; let $\hat{\mathbf{x}}_i, \quad i = 1, ..., m$ be the principal directions arranged according to the eigenvalues $\sigma_i^2 + \mu_{\hat{x}_i}^2$.

To minimize

$$\mu_c(\mathbf{h}) = tr[\mathbf{S}_{xx}'^{-1}\mathbf{R}_{xx}] = \sum_{i=1}^{d}(\sigma_i^2 + \mu_{\hat{x}_i}^2)[\mathbf{S}_{\hat{x}\hat{x}}'^{-1}]_{ii}, \quad \mathbf{h} \in \mathbb{S}^{m-1},$$

we need to sample in directions with the largest $\sigma_i^2 + \mu_{\hat{x}_i}^2$.

3. After sufficient sampling in this direction, the scheme switches to the next most important direction and so on.

4. Emphasis on input directions with large uncertainty, <u>even those that have zero effect to the output</u>.

2. Maximizing the x,y mutual information

Maximizing the entropy transfer or mutual information between the input and output variables, when a new sample is added:

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

We have:

$$\begin{aligned}
\mathcal{E}_{x,y}(\mathbf{h}) &= \int_y \int_x f_{xy}(\mathbf{y}, \mathbf{x}|D') \log f_{xy}(\mathbf{y}, \mathbf{x}|D') \\
&= \int_x \mathcal{E}_{y|x}(\mathbf{x}|D')\; f_x(\mathbf{x}) + \int_x f_x(\mathbf{x}) \log f_x(\mathbf{x}) \\
&= \mathbb{E}^x[\mathcal{E}_{y|x}(D')] + \mathcal{E}_x.
\end{aligned}$$

2. Maximizing the x,y mutual information

Given a new input point $\mathbf{x}_{N+1} = \mathbf{h}$, we have at any input $\mathbf{x}$

$$p(\mathbf{y}|\mathbf{x}, D', \mathbf{V}) = \mathcal{N}(\mathbf{S}_{yx}\mathbf{S}_{xx}^{-1}\mathbf{x}, \mathbf{V}(1+c)),$$
$$c = \mathbf{x}^T \mathbf{S'}_{xx}^{-1} \mathbf{x},$$

Therefore,

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

Note 1: Valid for any distribution f_x
Note 2: Hard to compute for high dimensions

The Gaussian approximation of the entropy criterion:

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

Note 1: The effect of **Y** appears only through a single scalar/vector and with no coupling on the new point **h.**

Note 2: Asymptotically (i.e. for small σ_c^2) the criterion becomes

$$\begin{aligned}\mathcal{I}_G(\mathbf{x},\mathbf{y}|D') = \frac{1}{2}\log|\mathbf{I} + \mathbf{V}^{-1}\mathbf{S}_{yx}\mathbf{S}_{xx}^{-1}\mathbf{C}_{xx}\mathbf{S}_{xx}^{-1}\mathbf{S}_{yx}^T| - \\ \left(d - tr[[\mathbf{V} + \mathbf{S}_{yx}\mathbf{S}_{xx}^{-1}\mathbf{C}_{xx}\mathbf{S}_{xx}^{-1}\mathbf{S}_{yx}^T]^{-1}\mathbf{V}]\right)\frac{\mu_c(\mathbf{h})}{2} + \mathcal{O}(\mu_c^2)\end{aligned}$$

3. Output-weighted optimal sampling

Let $\mathbf{y}_0$ be the rv defined as the mean model:

$$\mathbf{y}_0 \triangleq \mathbf{S}_{yx}\mathbf{S}_{xx}^{-1}\mathbf{x}$$

We define the perturbed model:

$$\mathbf{y}_+ \triangleq \mathbf{S}_{yx}\mathbf{S}_{xx}^{-1}\mathbf{x} + \beta\mathbf{r}_V(1 + \mathbf{x}^T\mathbf{S}_{xx}^{\prime -1}\mathbf{x}),$$

where β is a scaling factor to be chosen later and $\mathbf{r}_V$ the most dominant eigenvector of $\mathbf{V}$.

We define the distance (Mohamad & Sapsis, PNAS, 2018)

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

where S_y is a finite sub-domain of $\mathbf{y}$.

We can show that for bounded pdfs:

$$D_{KL}(\mathbf{y}_+ \| \mathbf{y}_0; \mathbf{h}) \leqslant \kappa D_{Log^1}(\mathbf{y}_+ \| \mathbf{y}_0; \mathbf{h}),$$

where κ is a constant. D_{Log^1} is more conservative compared with the KL divergence.

- Significantly improved performance in terms of convergence for f_y.
- Criterion $D_{Log^1}(\mathbf{y}_+ \| \mathbf{y}_0)$ is hard to compute/optimize.

3. Output-weighted optimal sampling

Under appropriate smoothness conditions standard inequalities for derivatives of smooth functions give (Sapsis, Proc Roy Soc A, 2020):

$$lim_{\beta\to 0} D_{Log^1}(\mathbf{y}_+ \| \mathbf{y}_0; \mathbf{h}) \leq \kappa_0 \int \frac{f_x(\mathbf{x})}{f_{y_0}(\mathbf{y}_0(\mathbf{x}))} \sigma_y^2(\mathbf{x}; \mathbf{h}) d\mathbf{x}.$$

3. Output-weighted optimal sampling

We define the output-weighted model error criterion

$$Q[\mathbf{h}] \triangleq \int \frac{f_x(\mathbf{x})}{f_{y_0}(\mathbf{y}_0(\mathbf{x}))} \sigma_y^2(\mathbf{x}; \mathbf{h}) d\mathbf{x}.$$

1. Model error weighted according to the importance (probability) of the input
2. Model error inversely weighted according to the probability of the output: *emphasis is given to outputs with low probability (rare events)*

Relevant criterion (Verdinelli & Kadane, 1992)

$$U(D') = q_1 \int \mathbf{y}_0(\mathbf{x}).\mathbf{1} d\mathbf{x} + q_2 \mathcal{E}_{xy|D'}.$$

$$Q[\sigma_y^2] \triangleq \int \frac{f_x(\mathbf{x})}{f_{y_0}(\mathbf{y}_0(\mathbf{x}))} \sigma_y^2(\mathbf{x}; \mathbf{h}) d\mathbf{x}.$$

Denominator approximation in S_y for symmetric f_y and scalar y

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

where p_1, p_2 are constants chosen so that m.s. error is min

We employ a Gaussian approximation for f_{y_0} (only for this step) and over the interval $S_y = [\mu_y, \mu_y + \beta\sigma_y]$ we obtain

$$p_1 = \sqrt{2\pi}\sigma_y \quad \text{and} \quad p_2 = \frac{5\sqrt{2\pi}}{\beta^5 \sigma_y} \left(\int_0^\beta z^2 e^{\frac{z^2}{2}} dz - \frac{\beta^3}{3} \right)$$

We collect all the computed terms and obtain (for Gaussian **x**)

$$
\begin{aligned}
Q_{\beta\sigma_y}(\mathbf{h})\frac{1}{\sigma_V^2} &= p_1(\beta)(1 + tr[\mathbf{S}_{xx}'^{-1}\mathbf{C}_{xx}] + \mu_x^T\mathbf{S}_{xx}'^{-1}\mu_x) \\
&+ p_2(\beta)c_0(1 + \mu_x^T\mathbf{S}_{xx}'^{-1}\mu_x - tr[\mathbf{S}_{xx}'^{-1}\mathbf{C}_{xx}]) \\
&+ 2p_2 tr[\mathbf{S}_{xx}^{-1}\mathbf{S}_{yx}^T\mathbf{S}_{yx}\mathbf{S}_{xx}^{-1}\mathbf{C}_{xx}\mathbf{S}_{xx}'^{-1}\mathbf{C}_{xx}].
\end{aligned}
$$

For zero mean input we have

$$
\begin{aligned}
Q_{\beta\sigma_y}(\mathbf{h})\frac{1}{\sigma_V^2} &= (p_1 - p_2c_0)tr[\mathbf{S}_{xx}'^{-1}\mathbf{C}_{xx}] \\
&+ 2p_2 tr[\mathbf{S}_{xx}'^{-1}\mathbf{C}_{xx0}\mathbf{S}_{xx}^{-1}\mathbf{S}_{yx}^T\mathbf{S}_{yx}\mathbf{S}_{xx}^{-1}\mathbf{C}_{xx}] + \text{const.}
\end{aligned}
$$

For general functions of the form

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

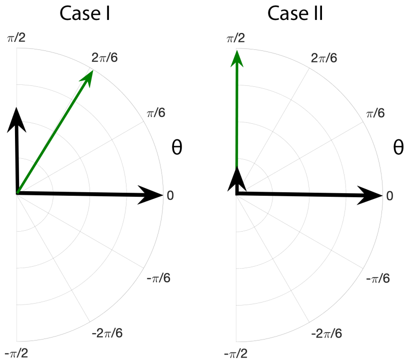
where **C** is a symmetric matrix. The gradient takes the form

$$\frac{\partial \lambda}{\partial h_k} = -2\mathbf{h}^T \mathbf{S}_{xx}'^{-1}\mathbf{C}\mathbf{S}_{xx}'^{-1}.$$

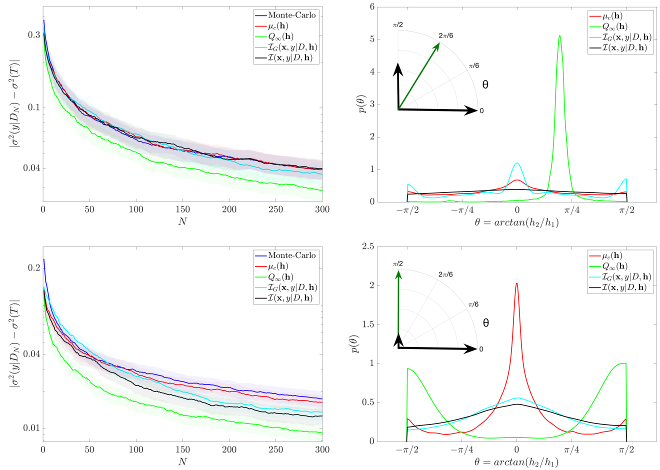
Example 1: 2-dimensional input

$$\hat{y}(\mathbf{x}) = \hat{a}_1 x_1 + \hat{a}_2 x_2 + \epsilon, \quad \text{where} \quad \mathbf{x} \sim \mathcal{N}(0, \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}) \quad \text{and} \quad \sigma_V^2 = 0.05.$$

- 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 $\sigma_1^2 = 2.0, \sigma_2^2 = 0.2$.



Results for the 2D problem



Example 2: A 20-dimensional input

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

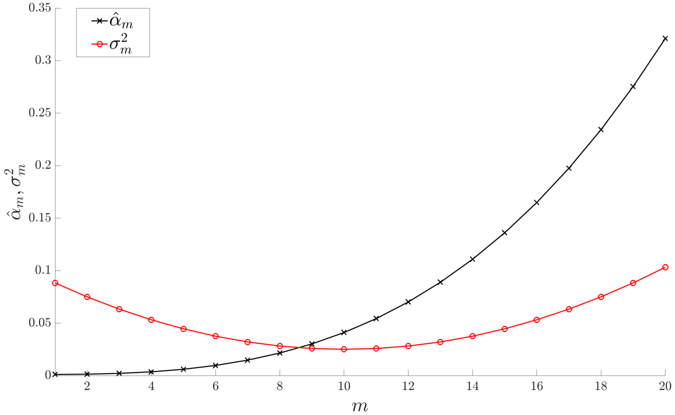
$$\hat{a}_m = \left(1 + 40\left(\frac{m}{10}\right)^3\right) 10^{-3}, \; m = 1, ..., 20,$$

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

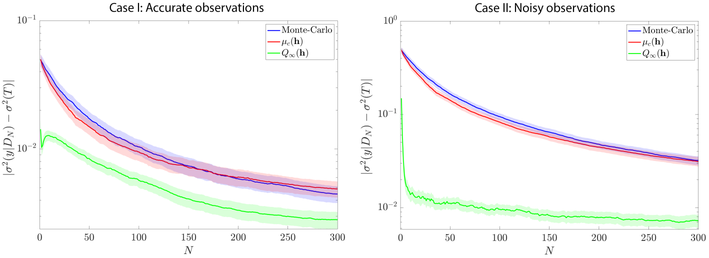
For the observation noise we consider two cases:

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

Example 2: A 20-dimensional input

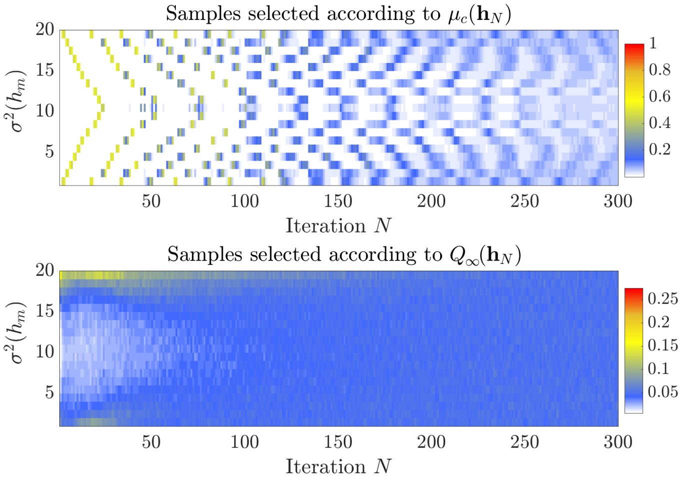


Coefficients, $\hat{\alpha}_m$, of the map $\hat{y}(\mathbf{x})$ (black curve) plotted together with the variance of each input direction σ_m^2 (red curve).



Performance of the two adaptive approaches based on μ_c and Q_∞.

Example 2: A 20-dimensional input



Energy of the different components of **h** with respect to the number of iteration N for Case I of the high dimensional problem.

Optimal sampling for nonlinear regression

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 f_z and $\mathcal{Z}$ be a compact set.

We choose a set of basis functions

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

The distribution of the output values will be

$$p(\mathbf{y}|\mathbf{z}, D, \mathbf{V}) = \mathcal{N}(\mathbf{S}_{y\phi}\mathbf{S}_{\phi\phi}^{-1}\phi(\mathbf{z}), \mathbf{V}(1 + c)),$$
$$c = \phi(\mathbf{z})^T \mathbf{S}_{\phi\phi}^{-1}\phi(\mathbf{z}),$$

$$\mathbf{S}_{\phi\phi} = \sum_{i=1}^{N} \phi(\mathbf{z}_i)\phi(\mathbf{z}_i)^T$$

Example 3: A nonlinear map

$$\hat{y}(\mathbf{z}) = \hat{a}_1 z_1 + \hat{a}_2 z_2 + \hat{a}_3 z_1^3 + \hat{a}_4 z_2^3 + \epsilon,$$

where

$$\mathbf{x} \sim \mathcal{N}(0, \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}) \text{ and } \sigma_V^2 = 10^{-4}$$

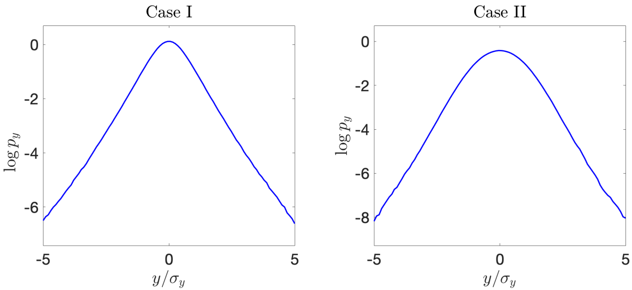
Two cases of parameters

- $\hat{a}_1 = 10^{-2}, \hat{a}_2 = 5, \hat{a}_4 = 10^2, \sigma_1^2 = 2.10^{-1}, \sigma_2^2 = 5.10^{-3}$
- $\hat{a}_1 = 10, \hat{a}_2 = 5, \hat{a}_4 = 10^2, \sigma_1^2 = 2.10^{-3}, \sigma_2^2 = 5.10^{-3}$

The basis functions are chosen as

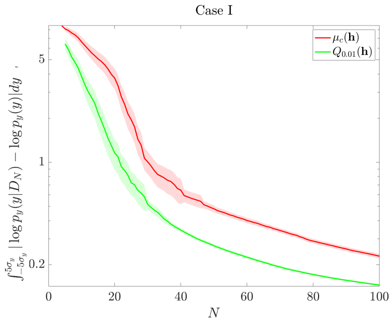
$$\phi(\mathbf{z}) = z_1^i z_2^j, \quad (i,j) \in \{(0,1),(1,0),(1,1),(0,3),(3,0)\}$$

Example 3: A nonlinear map



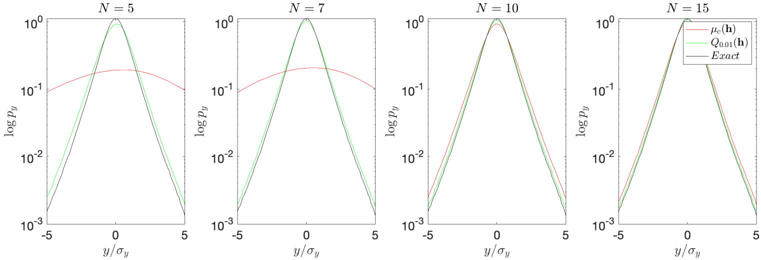
Exact pdf for the two cases of the nonlinear map using MC with 10^5 samples.

Example 3: A nonlinear map



Performance of the two adaptive approaches based on μ_c and Q_∞ for the nonlinear problem.

Example 3: A nonlinear map



Performance of the two adaptive approaches based on μ_c and Q_∞ for the nonlinear problem and Case I parameters.

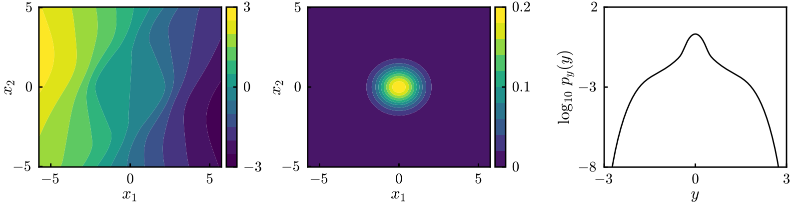
$$\ddot{u} + \delta\dot{u} + F(u) = \xi(t), \quad t \in [0, T]$$

The stochastic excitation is a parametrized by a KL expansion:

$$\xi(t) \approx \mathbf{x}\Phi(t), \quad \mathbf{x} \sim \mathcal{N}(\mathbf{0}, \Lambda)$$

The quantity of interest is the mean displacement

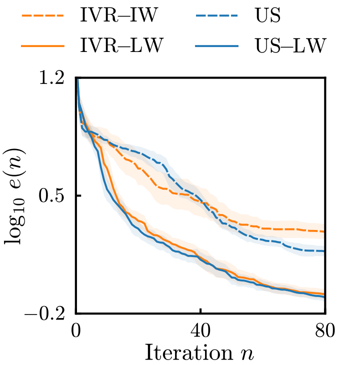
$$f(\mathbf{x}) = \frac{1}{T}\int_0^T u(t; \mathbf{x})\, \mathrm{d}t$$



Objective function | Input pdf | Output pdf

Quantifying rare events in a stochastic oscillator

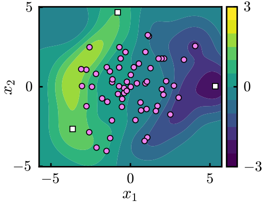
$$e(n) = \int |\log p_y(\mu) - \log p_y(f)|\, \mathrm{d}y$$



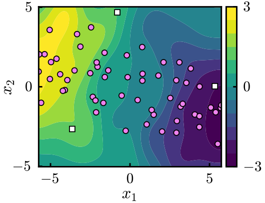
Benchmark results for the stochastic oscillator with $\sigma_\varepsilon^2 = 0$ (left) and $\sigma_\varepsilon^2 = 10^{-3}$ (right)

US: Uncertainty sampling: $min_x \sigma^2(x)$; US-LW: $min_x w(x)\sigma^2(x)$;
IVR: Integrated Variance Reduction-Input Weighted (IVR-IW): $\mu_c(x)$; IVR-LW: $Q-$criterion.

Example 4: Rare events in a stochastic oscillator



μ_c (Input-weighted variance)



$Q-$ criterion

The output-weighted criterion targets “relevant” regions more efficiently

Bayesian Optimization: Problem setup

$\mathbf{x} \in \mathcal{X} \subset \mathbb{R}^m$: Input parameters

$$\text{Minimize } y = f(\mathbf{x}) \in \mathbb{R}$$

- Starting from a set of n_{init} input-output pairs goal is to construct a surrogate of f and its global minimum
- Ingredient 1: surrogate model (here GPR)
- Ingredient 2: acquisition function

Pure exploration:

$$\text{Uncertainty Sampling} \quad a(\mathbf{x}) = -\sigma^2(\mathbf{x})$$

$$\text{Integrated Variance Reduction} \quad a(\mathbf{x}) = -\int_{\mathcal{X}} \mathrm{cov}^2(\mathbf{x}, \mathbf{x}') d\mathbf{x}' / \sigma^2(\mathbf{x})$$

Exploration–exploitation trade-off (B. Shahriari et al., IEEE 2015)

$$\text{BO-Repurposed IVR} \quad a(\mathbf{x}) = \mu(\mathbf{x}) + \kappa a_{IVR}(\mathbf{x})$$

$$\text{Lower Confidence Bound} \quad a(\mathbf{x}) = \mu(\mathbf{x}) - \kappa\sigma(x)$$

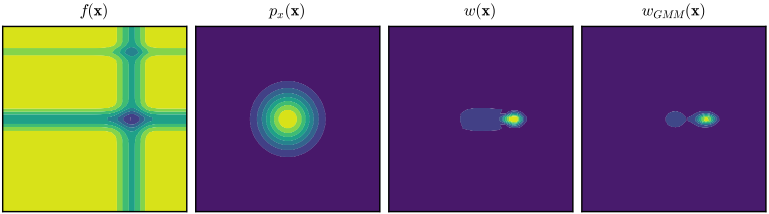
$$\text{Probability of Improvement} \quad a(\mathbf{x}) = -\Phi(\lambda(\mathbf{x}))$$

$$\text{Expected Improvement} \quad a(\mathbf{x}) = -\sigma(\mathbf{x}) \left[\lambda(\mathbf{x})\Phi(\lambda(\mathbf{x})) - \phi(\lambda(\mathbf{x}))\right]$$

where $\lambda(\mathbf{x}) = (y^* - \mu(\mathbf{x}) - \xi)/\sigma(\mathbf{x})$

The role of the likelihood ratio in BO and BED

$$\boxed{w(\mathbf{x}) = \frac{p_x(\mathbf{x})}{p_y(\mu(\mathbf{x}))}} \approx \sum_{i=1}^{n_{GMM}} \alpha_i \mathcal{N}(\mathbf{x}; \omega_i, \Sigma_i)$$



2-D Michalewicz function

The likelihood ratio

- acts as a probabilistic sampling weight
- emphasizes the most relevant regions of the input space
- can be approximated by a small number of Gaussian mixtures

Acquisition functions for BO and BED

$$w(\mathbf{x}) = \frac{p_x(\mathbf{x})}{p_y(\mu(\mathbf{x}))}$$

Pure exploration:

Uncertainty Sampling $\quad a(\mathbf{x}) = -\sigma^2(\mathbf{x})w(\mathbf{x})$

Integrated Variance Reduction $\quad a(\mathbf{x}) = -\int_{\mathcal{X}} \mathrm{cov}^2(\mathbf{x}, \mathbf{x}')w(\mathbf{x})\, \mathrm{d}\mathbf{x}'/\sigma^2(\mathbf{x})$

Exploration–exploitation trade-off :

BO-Repurposed IVR $\quad a(\mathbf{x}) = \mu(\mathbf{x}) + \kappa a_{IVR}(\mathbf{x})$

Lower Confidence Bound $\quad a(\mathbf{x}) = \mu(\mathbf{x}) - \kappa\sigma(x)w(\mathbf{x})$

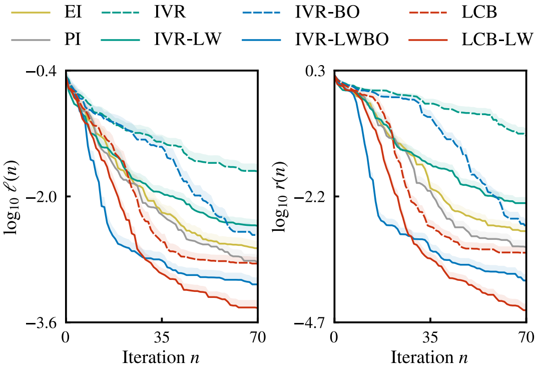
Probability of Improvement $\quad a(\mathbf{x}) = -\Phi(\lambda(\mathbf{x}))$

Expected Improvement $\quad a(\mathbf{x}) = -\sigma(\mathbf{x})\left[\lambda(\mathbf{x})\Phi(\lambda(\mathbf{x})) - \phi(\lambda(\mathbf{x}))\right]$

where $\lambda(\mathbf{x}) = (y^* - \mu(\mathbf{x}) - \xi)/\sigma(\mathbf{x})$

BO with output-weighted acquisition functions

$$\ell(n) = \min_{k\in[0,n]} \|\mathbf{x}_{true} - \mathbf{x}_k^*\|^2 \qquad r(n) = \min_{k\in[0,n]} f(\mathbf{x}_k^*) - y_{true}$$

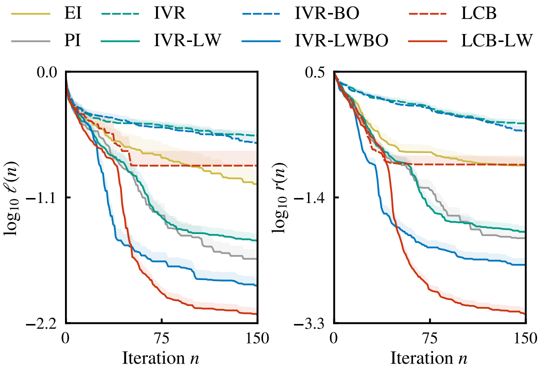


Benchmark results for 2-D Michalewicz function (distance to min and simple regret)

EI: Expected Improvement $-\sigma(\mathbf{x})\,[\lambda(\mathbf{x})\Phi(\lambda(\mathbf{x})) - \phi(\lambda(\mathbf{x}))]$, PI: Probability of Improvement $-\Phi(\lambda(\mathbf{x}))$,
IVR: integrated Variance Reduction $-\int_{\mathcal{X}} \mathrm{cov}^2(\mathbf{x},\mathbf{x}')\,\mathrm{d}\mathbf{x}'/\sigma^2(\mathbf{x})$, IVR-BO: $\mu(x) + \kappa a_{IVR}(x)$,
LCB: Lower Confidence Bound $\mu(\mathbf{x}) - \kappa\sigma(x)$, LW: Likelihood weighted: $w(\mathbf{x})$.

BO with output-weighted acquisition functions

$$\ell(n) = \min_{k \in [0,n]} \|\mathbf{x}_{true} - \mathbf{x}_k^*\|^2 \qquad r(n) = \min_{k \in [0,n]} f(\mathbf{x}_k^*) - y_{true}$$



Benchmark results for 6-D Hartmann function (distance to min and simple regret)

EI: Expected Improvement $-\sigma(\mathbf{x})\,[\lambda(\mathbf{x})\Phi(\lambda(\mathbf{x})) - \phi(\lambda(\mathbf{x}))]$, PI: Probability of Improvement $-\Phi(\lambda(\mathbf{x}))$,
IVR: integrated Variance Reduction $-\int_{\mathcal{X}} \mathrm{cov}^2(\mathbf{x}, \mathbf{x}')\,\mathrm{d}\mathbf{x}' / \sigma^2(\mathbf{x})$, IVR-BO: $\mu(x) + \kappa a_{IVR}(x)$,
LCB: Lower Confidence Bound $\mu(\mathbf{x}) - \kappa\sigma(x)$, LW: Likelihood weighted: $w(\mathbf{x})$.

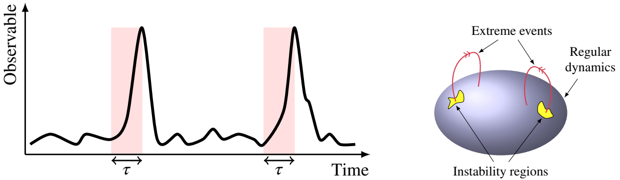
Finding extreme-event precursors by optimal sampling

For a dynamical system with flow map S_t and observable G:

- assign to each initial condition $\mathbf{x}_0$ a measure of dangerousness,

$$
\begin{aligned}
F : \mathbb{R}^d &\longrightarrow \mathbb{R} \\
\mathbf{x}_0 &\longmapsto \max_{t \in [0,\tau]} G(S_t(\mathbf{x}_0))
\end{aligned}
$$

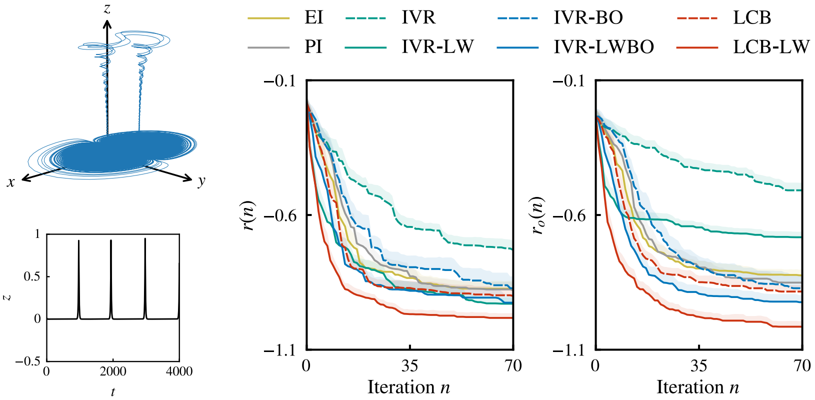
- use the sampling algorithm to probe the initial-condition space
- perform search in PCA space with Gaussian prior $p_x(\mathbf{x})$



Computation of extreme-event precursors in Gaussian PCA subspace

Finding extreme-event precursors by optimal sampling

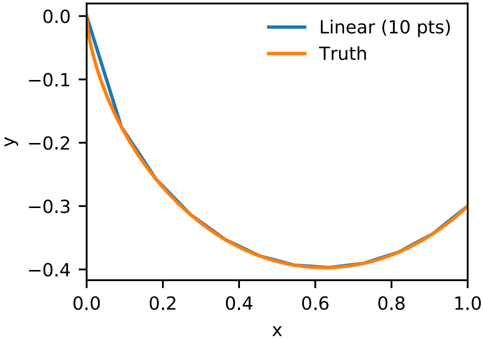
$$r(n) = \min_{k \in [0,n]} f(\mathbf{x}_k^*) \qquad r_o(n) = \min_{y_i \in \mathcal{D}_n} y_i$$



EI: Expected Improvement $-\sigma(\mathbf{x})\,[\lambda(\mathbf{x})\Phi(\lambda(\mathbf{x})) - \phi(\lambda(\mathbf{x}))]$, PI: Probability of Improvement $-\Phi(\lambda(\mathbf{x}))$,
IVR: integrated Variance Reduction $-\int_{\mathcal{X}} \mathrm{cov}^2(\mathbf{x}, \mathbf{x}')\,\mathrm{d}\mathbf{x}' / \sigma^2(\mathbf{x})$, IVR-BO: $\mu(x) + \kappa a_{IVR}(x)$,
LCB: Lower Confidence Bound $\mu(\mathbf{x}) - \kappa\sigma(x)$, LW: Likelihood weighted: $w(\mathbf{x})$.

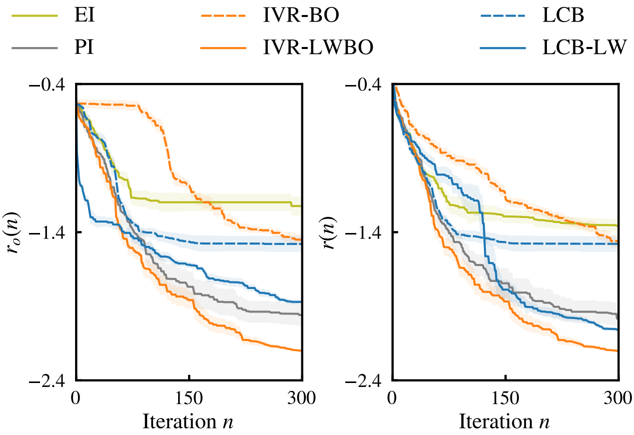
$$f(\mathbf{x}) = \log(T(\mathbf{x}) - t_c)$$

$T(\mathbf{x})$	Travel time for given parametrization $\mathbf{x}$
t_c	Best travel time possible (cycloid)



The Brachistochrone problem

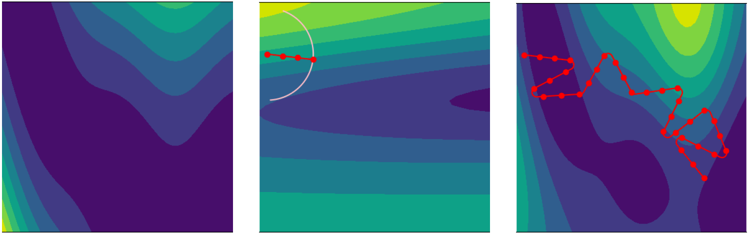
$$r_o(n) = \min_{y_i \in \mathcal{D}_n} y_i \qquad r(n) = \min_{k \in [0,n]} f(\mathbf{x}_k^*) - y_{true}$$



EI: Expected Improvement $-\sigma(\mathbf{x})\,[\lambda(\mathbf{x})\Phi(\lambda(\mathbf{x})) - \phi(\lambda(\mathbf{x}))]$, PI: Probability of Improvement $-\Phi(\lambda(\mathbf{x}))$,
IVR: integrated Variance Reduction $-\int_{\mathcal{X}} \mathrm{cov}^2(\mathbf{x}, \mathbf{x}')\,\mathrm{d}\mathbf{x}' / \sigma^2(\mathbf{x})$, IVR-BO: $\mu(x) + \kappa a_{IVR}(x)$,
LCB: Lower Confidence Bound $\mu(\mathbf{x}) - \kappa\sigma(x)$, LW: Likelihood weighted: $w(\mathbf{x})$.

Informative path planning for terrain exploration

A UAV is tasked with reconstructing a terrain elevation map $f(\mathbf{x})$



The unknown terrain | First (random) iteration | After 11 iterations

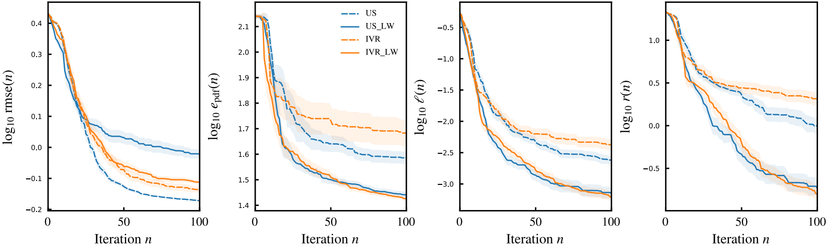
Next best destination:

$$\mathbf{x}_f^* = \operatorname*{argmin}_{\mathbf{x}_f} \int_{S(\mathbf{x}_c, \mathbf{x}_f)} a(\mathbf{x}(s))\, \mathrm{d}s$$

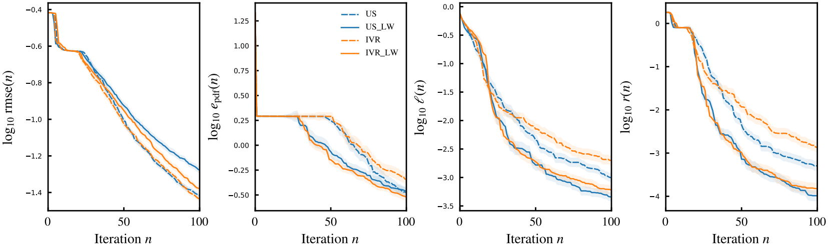
where $S(\mathbf{x}_c, \mathbf{x}_f)$ is the shortest Dubins curve from $\mathbf{x}_c$ to candidate $\mathbf{x}_f$

Reconstruction of strongly anomalous terrain

The Ackley function



The Michalewicz function



US: Uncertainty sampling: $min_x \sigma^2(x)$; US-LW: $min_x w(x)\sigma^2(x)$;
IVR: Integrated Variance Reduction-Input Weighted (IVR-IW): $\mu_c(x)$; IVR-LW: $Q-$criterion.

Conclusions

- Samples based on maximum mutual information or minimum model error do not effectively take into account the contribution to the output.
- A new criterion allows for sampling of points in regions that have important influence to the output.
- The criterion can be approximated analytically so that we can apply it to high dimensional parameter spaces.
- Application to risk quantification and optimization

Sapsis, Output-weighted optimal sampling for Bayesian regression and rare event statistics using few samples, **Proceedings of the Royal Society A**, (2020).

Blanchard & Sapsis, Bayesian optimization with output-weighted importance sampling, **arXiv**, (2020).