Randomized Maximum Likelihood via High-Dimensional Bayesian Optimization

Valentin Breaz
University of Nottingham

June 7, 2022

Introduction

- Bayesian inverse problems: target the posterior distribution $p(x|\mathcal{D})$ of the unknown parameters $x \in \mathbb{R}^D$ given the observed data $\mathcal{D} \in \mathbb{R}^m$
- Standard setting: Gaussian likelihood $\mathcal{D}|x \sim \mathcal{N}_m(f(x), \Sigma_{\text{obs}})$, where $f(x) : \mathbb{R}^D \to \mathbb{R}^m$ is known as the simulator
- The simulator is often the deterministic solution to a PDE modelling an underlying physical process [Stuart, 2010]
- Bayesian inverse problems are employed in a variety of applications, such as climate modelling, medical imaging and material sciences

Randomized Maximum Likelihood

Randomized Maximum Likelihood (RML) was introduced by [Oliver et al., 1996], as an approximate posterior sampling methodology

Algorithm Randomized Maximum Likelihood (RML)

```
$n_{RML}$ : number of samples required
for $n \in [n_{RML}]$ do
    1. Sample $\mathcal{D}_n \sim \mathcal{N}_m(\mathcal{D}, \Sigma_{\text{obs}})$ from the Gaussian likelihood
    2. Sample $\mu_n \sim \mathcal{N}_D(\mu, \Sigma)$ from the Gaussian prior
    3. Construct $\log p(\mathcal{D}|x)p(x)$ w.r.t. the randomizations $(\mathcal{D}_n, \mu_n)$

           $O_n(x) := \log \mathcal{N}_m(f(x)|\mathcal{D}_n, \Sigma_{\text{obs}}) + \log \mathcal{N}_D(x|\mu_n, \Sigma)$

    4. Obtain $x_n^\star$ as the maximizer $x_n^\star = \arg\max_x O_n(x)$.
end for
```

Randomized Maximum Likelihood

- The resulting solutions $\{x_n^\star\}_{n=1}^{n_{RML}}$ are regarded as approximate samples from the posterior distribution $p(x|\mathcal{D})$; the samples are only exact draws from the posterior when the simulator $f(x)$ is linear
- Additional care needs to be taken in more challenging scenarios, such as multi-modal posteriors with highly nonlinear simulators [Bardsley et al., 2014, Oliver, 2015, Ba et al., 2021]
- Nonetheless, good practical performance has been observed for nonlinear neural network parameterized simulators [Tang et al., 2020]

Randomized Maximum Likelihood and Active Subspaces

- In this work, we address solving the RML optimization problems efficiently in the case of a high-dimensional input space $\mathbb{R}^D$
- We focus on the specific scenario where the log-likelihood

$$\log p(\mathcal{D}|x) \propto L(x) := -||\mathcal{D} - f(x)||^2_{\Sigma_{\text{obs}}}$$

  has a low-dimensional active subspace [Constantine et al., 2015]
- In other words, we assume $L(x) \approx g(A^T x)$, where $g : \mathbb{R}^d \to \mathbb{R}$ with $d \ll D$, and $A \in \mathbb{R}^{D \times d}$ is a semi-orthogonal matrix ($A^T A = I_d$)

Active Subspaces

- We consider the expected outer product of the gradient

$$C := \int \nabla L(x) \nabla L(x)^T dp(x)$$

- In practice, we might only have access to the Monte-Carlo sum

$$\hat{C} := \frac{1}{n} \sum_{i=1}^{n} \nabla L(x_i) \nabla L(x_i)^T, \;\; x_i \sim p(x)$$

- The d dominant eigenvectors of $\hat{C}$ form the active subspace A, usually when the d^{th} eigenvalue is $\geq 10\times$ larger than the $(d+1)^{th}$
- The prior distribution in the integral can be replaced by an approximation of the posterior distribution [Zahm et al., 2018]

List of assumptions

- Computational constraints mean that we are limited to at most N simulator evaluations (for complex simulators, N may be small)
- We do not have access to gradients of the simulator
- Although we assume an active subspace A exists, we do not have access to A, and moreover, we do not have sufficient budget to estimate it from $\{x_i, L(x_i)\}_{i=1}^N$
- Even though the log-likelihood has a low-dimensional active subspace, the prior might not have such a structure

Bayesian Optimization

- If we ignore the prior and the multi-objective nature of our problem for now, the task of maximizing an objective $O(x) \approx g(A^T x)$ (in this case, the log-likelihood) under the assumptions mentioned above is common in high-dimensional Bayesian Optimization (HD-BO)
- Bayesian Optimization (BO, i.e. finding $\arg\max_x O(x)$ using a Gaussian process approximation $g_{\mathsf{GP}}(x) \approx O(x)$) is based on a standard exploration-exploitation principle
- Namely, an acquisition function based on $g_{\mathsf{GP}}(x)$ is used such that in the exploration phase, the target function $O(x)$ is explored globally, whereas in the exploitation phase, points $\tilde{x}$ that are likely to satisfy $\tilde{x} = \arg\max_x O(x)$ are sampled until the maximum is found

High-Dimensional Bayesian Optimization (HD-BO)

- GPs are known to deal well with small training budgets, but may struggle with high input dimensionality D [Liu and Guillas, 2016]
- As a result, the HD-BO literature mostly deals with the prevalent case where the approximation $g_{\mathsf{GP}}(x) \approx O(x)$ is unsatisfactory
- In our setting where an active subspace exists but is unknown, the most common solution is the use of random embeddings, R, instead of the true low-dimensional embedding A [Wang et al., 2013]

High-Dimensional Bayesian Optimization (HD-BO)

Algorithm Generic HD-BO with random embeddings

```
$M$ : number of evaluations of $O(\cdot)$ possible given the computational budget;
$d_e$ : chosen dimensionality of the embedding $R$;
$R \in \mathbb{R}^{D \times d_e}$ : random embedding;
$m_0$ : initial training points $\{y_i, O(Ry_i)\}_{i=1}^{m_0}$, with $y_i \in \mathbb{R}^{d_e}$
for $m \in \{m_0 + 1, \dots, M\}$ do
    1. Construct a GP approximation $O(Ry) \sim$ GP using the available objective function evaluations $\{y_i, O(Ry_i)\}_{i=1}^{m-1}$
    2. Select $y_m = \arg\max_y a_m(y)$ as the maximizer of a BO acquisition function for $O(Ry) \sim$ GP
    3. Update the training data to $\{y_i, O(Ry_i)\}_{i=1}^{m}$.
end for
Obtain $x_\star = Ry_{m_\star}$ as the maximizer
```

$$m_\star = \arg\max_m O(Ry_m), \;\; m \leq M.$$

High-Dimensional Bayesian Optimization (HD-BO)

- The random embedding, R, transforms the original high-dimensional BO problem $O(x) \sim$ GP for $x \in \mathbb{R}^D$ into a low-dimensional BO problem $O(Ry) \sim$ GP for $y \in \mathbb{R}^{d_e}$
- In other words, instead of trying to maximize $O(x)$, we try to maximize $O(Ry)$, which is the objective function on the subspace R
- In practice, we can use multiple random projections $R_1, \ldots, R_K$ giving maximizers $x_\star^1, \ldots, x_\star^K$, and select $x_\star := \arg\max_{x_\star^k} O(x_\star^k)$

Gaussian Processes and posterior sampling

- Although there is an extensive literature for HD-BO with random embeddings, there is no methodology designed for posterior sampling
- The use of GPs for posterior sampling in low-dimensional experiments can be found in the active learning literature, where the aim is to generate training points from high-posterior density regions
- Regarding high-dimensional experiments, GPs have been mostly used in cases where the prior distribution had a low-dimensional structure which was amenable to dimension reduction

RML via HD-BO (uniform prior)

- Firstly, we consider the simpler setting of a uniform prior, $x \sim U[a_i, b_i]_{i=1}^D$, where $[a_i, b_i]_{i=1}^D := [a_1, b_1] \times \cdots \times [a_D, b_D]$
- In this case, the posterior is proportional to the likelihood, and using our active subspace assumption the RML objectives become

$$O_n(x) = L_n(x) := \log \mathcal{N}_m(f(x)|\mathcal{D}_n, \Sigma_{\text{obs}}) \approx g_n(A_n^T x)$$

- Since all the objective functions $O_n(x)$ have similar structure and are based on the same underlying simulator $f(x)$, we expect that the BO exploration stage can be performed at once for all objectives

RML via HD-BO (uniform prior)

- For example, we could run HD-BO for $O_1(x)$ (T_1 iterations say), and then reuse the training data $\{(y_t^1, f(Ry_t^1))\}$ to warm-start/speed-up convergence for $O_2(x)$ ($T_2 \ll T_1$)
- Whilst this is an attractive strategy, it poses the difficulty of having to choose a stopping time T_n for every objective
- As a result, we choose the simpler strategy of performing a cyclic pass through all the objective functions $O_n(x)$

RML via HD-BO (uniform prior)

- We write $n := n_0$ for the initial points, $M := n_{RML}$ for the number of RML objectives and $O_m(y) := O_m(Ry)$ for $m \in [M]$

$$\begin{array}{cc} & \begin{array}{cccc} 1 & \dots & n & n+1 \end{array} \\ \begin{array}{c} O_1(y) \\ O_2(y) \\ \vdots \\ O_M(y) \end{array} & \begin{pmatrix} y_1, O_1(y_1) & \vdots & y_n, O_1(y_n) & y_{n+1} \\ y_1, O_2(y_1) & \vdots & y_n, O_2(y_n) & \\ \vdots & \vdots & \vdots & \\ y_1, O_M(y_1) & \vdots & y_n, O_M(y_n) & \end{pmatrix} \end{array} \rightarrow \begin{array}{c} \begin{array}{cccc} 1 & \dots & n & n+1 \end{array} \\ \begin{pmatrix} y_1, O_1(y_1) & \vdots & y_n, O_1(y_n) & y_{n+1}, O_1(y_{n+1}) \\ y_1, O_2(y_1) & \vdots & y_n, O_2(y_n) & y_{n+1}, O_2(y_{n+1}) \\ \vdots & \vdots & \vdots & \vdots \\ y_1, O_M(y_1) & \vdots & y_n, O_M(y_n) & y_{n+1}, O_M(y_{n+1}) \end{pmatrix} \end{array}$$

$$\begin{array}{c} \begin{array}{cccc} 1 & \dots & n+1 & n+2 \end{array} \\ \begin{pmatrix} y_1, O_1(y_1) & \vdots & y_{n+1}, O_1(y_{n+1}) & \\ y_1, O_2(y_1) & \vdots & y_{n+1}, O_2(y_{n+1}) & y_{n+2} \\ \vdots & \vdots & \vdots & \\ y_1, O_M(y_1) & \vdots & y_{n+1}, O_M(y_{n+1}) & \end{pmatrix} \end{array} \rightarrow \begin{array}{c} \begin{array}{cccc} 1 & \dots & n+1 & n+2 \end{array} \\ \begin{pmatrix} y_1, O_1(y_1) & \vdots & y_{n+1}, O_1(y_{n+1}) & y_{n+2}, O_1(y_{n+2}) \\ y_1, O_2(y_1) & \vdots & y_{n+1}, O_2(y_{n+1}) & y_{n+2}, O_2(y_{n+2}) \\ \vdots & \vdots & \vdots & \vdots \\ y_1, O_M(y_1) & \vdots & y_{n+1}, O_M(y_{n+1}) & y_{n+2}, O_M(y_{n+2}) \end{pmatrix} \end{array}$$

RML via HD-BO (uniform prior)

Algorithm HD-BO-RML with uniform priors

```
$N$ : max possible number of evaluations of $f(\cdot)$;
$d_e$ : choice of embedding dimensionality;
$R_1, \dots, R_K \in \mathbb{R}^{D \times d_e}$ : collection of random embeddings;
$n_0 \times K$ initial points: $\{y_i^k, f(R_k y_i^k)\}_{i=1}^{n_0}$, with $y_i^k \in \mathbb{R}^{d_e}$, $k \in [K]$;
for $k \in \{1, \dots, K\}$ do
    for $n \in \{n_0 + 1, \dots, \lfloor N/K \rfloor\}$ do
        1. Set $n' := n \mod n_{RML}$
        2. Construct a GP approximation to $O_{n'}(R_k y)$ using simulations $\{y_i^k, f(R_k y_i^k)\}_{i=1}^{n-1}$
        3. Select $y_n^k = \arg\max_y a_n^k(y)$ as the maximizer of a BO acquisition function using the
        GP approximation
        4. Perform simulation $f(R_k y_n^k)$ and update the shared simulation ensemble to
        $\{y_i^k, f(R_k y_i^k)\}_{i=1}^{n}$.
    end for
end for
for $n \in \{1, \dots, n_{RML}\}$ do
    1. Obtain $x_n^\star = R_{k_\star} y_{m_\star}^{k_\star}$ as the maximizer

        $$k_\star, m_\star = \arg\max_{k,m} O_n(R_k y_m^k),\ k \in [K], m \leq \lfloor N/K \rfloor$$

end for
```

RML via HD-BO (Gaussian prior)

- Using again the active subspace assumption:

$$O_n(x) = L_n(x) + \log p_n(x) \approx g_n(A_n^T x) + \log \mathcal{N}_D(x|\mu_n, \Sigma),$$

  where $L_n(x) := \log \mathcal{N}_m(f(x)|\mathcal{D}_n, \Sigma_{\text{obs}})$

- Due to the potential lack of low-dimensional structure in the prior, running HD-BO by modelling $O_n(Ry) \sim$ GP can be unsatisfactory, and hence we cannot re-use the algorithm from the uniform prior case
- Instead, while we keep performing HD-BO with respect to the log-likelihood, we try to increase the prior value for the selected points of potentially high-likelihood $x_0 := R_k y_n^k$

RML via HD-BO (Gaussian prior)

Algorithm HD-BO-RML with Gaussian priors

```
$N$ : max possible number of evaluations of $f(\cdot)$;
$d_e$ : choice of embedding dimensionality;
$R_1, \dots, R_K \in \mathbb{R}^{D \times d_e}$ : collection of random embeddings;
$n_0 \times K$ initial points: $\{y_i^k, f(R_k y_i^k)\}_{i=1}^{n_0}$, with $y_i^k \in \mathbb{R}^{d_e}$, $k \in [K]$;
for $k \in \{1, \dots, K\}$ do
    for $n \in \{n_0 + 1, \dots, \lfloor N/2K \rfloor\}$ do
        1. Let $n' := n \mod n_{RML}$
        2. Construct a GP approximation to $L_{n'}(R_k y)$ using simulations $\{y_i^k, f(R_k y_i^k)\}_{i=1}^{n-1}$
        3. Select $y_n^k = \arg\max_y a_n^k(y)$ as the maximizer of a BO acquisition function using the GP approximation
        4. Perform $f(R_k y_n^k)$ and update the shared simulation ensemble to $\{y_i^k, f(R_k y_i^k)\}_{i=1}^{n}$
        5. Perform local optimization with respect to the prior and select $z_n^k = \arg\max_{x \in B} p_{n'}(x)$, where $B \in \mathbb{R}^D$ is a box centered at $x_0 = R_k y_n^k$.
    end for
end for
for $n \in \{1, \dots, n_{RML}\}$ do
    1. Obtain $x_n^\star = z_{m_\star}^{k_\star}$ as the maximizer

       $$k_\star, m_\star = \arg\max_{k,m} O_n(z_m^k),\ k \in [K], m \leq \lfloor N/2K \rfloor$$

end for
```

Experiments

- Elliptic-PDE simulator $f : \mathbb{R}^{100} \to \mathbb{R}^7$ with standard Gaussian prior $x \sim \mathcal{N}_{100}(0, I)$ [Constantine et al., 2015]
- This setup is common in many Bayesian inverse problems, including when non-Gaussian features $T(x)$ (e.g. channelized permeability fields in geology) generate $\mathcal{D} = f(T(x)) + \epsilon$ [Iglesias et al., 2015]
- Ebola: $f : \mathbb{R}^8 \to \mathbb{R}$, an 8-parameter dynamical system model for the geographic spread of Ebola in Liberia [Diaz et al., 2016]
- MHD: $f : \mathbb{R}^5 \to \mathbb{R}$, a 5-parameter magnetohydrodynamics power generation model [Glaws et al., 2016]
- HIV long-term model: $f : \mathbb{R}^{27} \to \mathbb{R}$ [Loudon and Pankavich, 2016]

Competing methods

- BOBYQA (trust-region method) [Powell, 2007]
- CMA-ES (evolution strategy method) [Hansen et al., 2019]
- Random design
- We are interested in the method that finds points with the highest mean return, i.e., largest

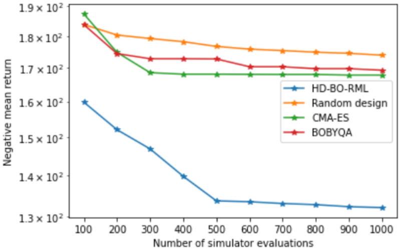
$$\frac{1}{n_{RML}} \sum_{n=1}^{n_{RML}} O_n(x_n^\star),$$

where $x_n^\star$ is the approximate maximizer of $O_n(x)$ as selected by the different methods considered

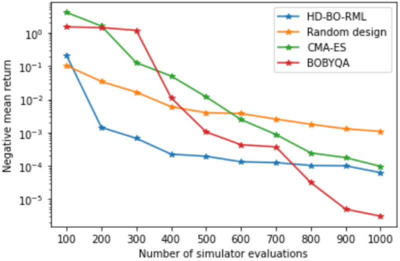
Experimental setup

- We consider 5 trials; each experiment has a fixed set of measurements $\mathcal{D}$ and thus a fixed set of RML objectives
- Each trial has a budget of at most $N = 1000$ simulations in order to find $n_{RML} = 20$ samples
- BOBYQA and CMA-ES cannot share data between different objectives $O_i(x)$ and $O_j(x)$, unlike HD-BO-RML which shares data through the common simulator via the GP training sets
- We employ BOBYQA and CMA-ES independently for each objective $O_n(x)$ for $n \in [n_{RML} = 20]$, using at most 50 simulator evaluations per objective to stay within budget

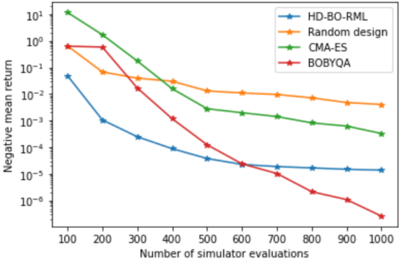
Results



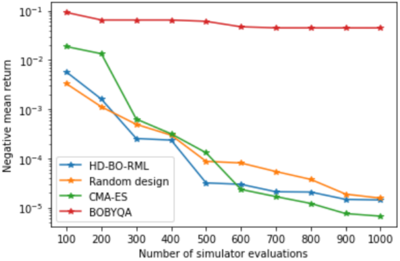
(a) PDE



(b) Ebola

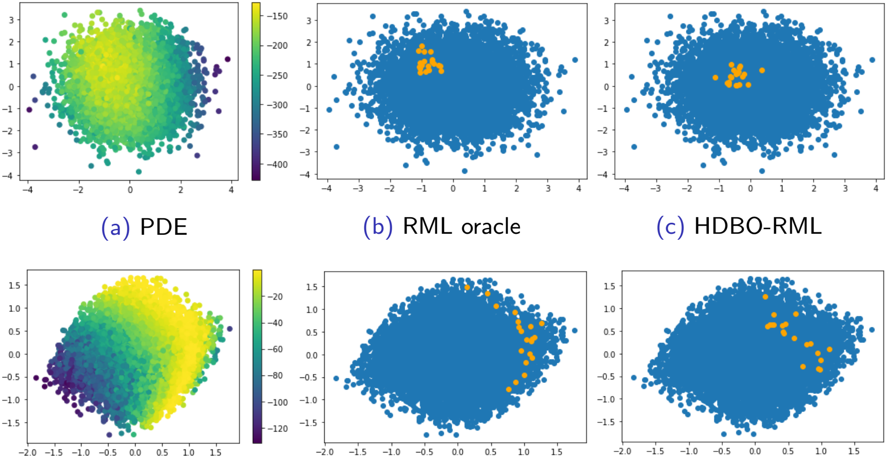


(c) MHD



(d) HIV

Visualization of samples



(a) PDE

(b) RML oracle

(c) HDBO-RML

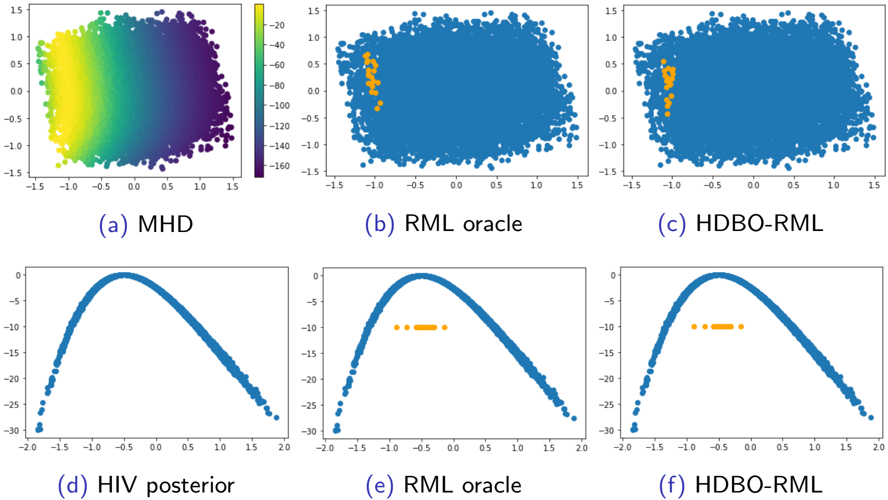
(d) Ebola

(e) RML oracle

(f) HDBO-RML

Figure: Posterior landscape in the active subspace (left), oracle RML samples with infinite budget (middle) and RML samples obtained by our procedure (right). The RML samples are shown in orange, with prior samples given in blue.

Visualization of samples



(a) MHD

(b) RML oracle

(c) HDBO-RML

(d) HIV posterior

(e) RML oracle

(f) HDBO-RML

Figure: Posterior landscape in the active subspace (left), oracle RML samples with infinite budget (middle) and RML samples obtained by our procedure (right). The RML samples are shown in orange, with prior samples given in blue.

Conclusions and future work

- We have introduced an RML approach based on HD-BO that outperforms competing gradient-free optimization methods when there are tight computational budget constraints
- To demonstrate the potential of our procedure, we presented a vanilla version using default choices of embeddings, GP approximations, and acquisition function
- For future work, it would be interesting to investigate multi-output GPs for our procedure and for multi-objective HD-BO in general, complementing the findings of [Dai et al., 2020] regarding low-dimensional experiments

Ba, Y., de Wiljes, J., Oliver, D. S., and Reich, S. (2021).
Randomized maximum likelihood based posterior sampling.
arXiv e-prints, page arXiv:2101.03612.

Bardsley, J. M., Solonen, A., Haario, H., and Laine, M. (2014).
Randomize-then-optimize: A method for sampling from posterior distributions in nonlinear inverse problems.
SIAM Journal on Scientific Computing, 36(4):A1895–A1910.

Constantine, P. G., Kent, C., and Bui-Thanh, T. (2015).
Accelerating MCMC with active subspaces.
arXiv e-prints, page arXiv:1510.00024.

Dai, S., Song, J., and Yue, Y. (2020).
Multi-task bayesian optimization via gaussian process upper confidence bound.

Diaz, P., Constantine, P., Kalmbach, K., Jones, E., and Pankavich, S. (2016).
A Modified SEIR Model for the Spread of Ebola in Western Africa and Metrics for Resource Allocation.

arXiv e-prints, page arXiv:1603.04955.

Glaws, A., Constantine, P. G., Shadid, J., and Wildey, T. M. (2016).
Dimension reduction in MHD power generation models: dimensional analysis and active subspaces.
arXiv e-prints, page arXiv:1609.01255.

Hansen, N., Akimoto, Y., and Baudis, P. (2019).
CMA-ES/pycma on Github.
Zenodo, DOI:10.5281/zenodo.2559634.
BSD 3-Clause License (2014).

Iglesias, M. A., Lu, Y., and Stuart, A. M. (2015).
A Bayesian Level Set Method for Geometric Inverse Problems.
arXiv e-prints, page arXiv:1504.00313.

Liu, X. and Guillas, S. (2016).
Dimension reduction for emulation: application to the influence of bathymetry on tsunami heights.
SIAM/ASA Journal on Uncertainty Quantification, 5.

Loudon, T. and Pankavich, S. (2016).

Mathematical Analysis and Dynamic Active Subspaces for a Long term model of HIV.
arXiv e-prints, page arXiv:1604.04588.

Oliver, D. S. (2015).
Metropolized Randomized Maximum Likelihood for sampling from multimodal distributions.
arXiv e-prints, page arXiv:1507.08563.

Oliver, D. S., He, N., and Reynolds, A. C. (1996).
Conditioning permeability fields to pressure data.

Powell, M. J. D. (2007).
A view of algorithms for optimization without derivatives 1.

Stuart, A. M. (2010).
Inverse problems: A bayesian perspective.
Acta Numerica, 19:451–559.

Tang, M., Liu, Y., and Durlofsky, L. J. (2020).
A deep-learning-based surrogate model for data assimilation in dynamic subsurface flow problems.

Journal of Computational Physics, 413:109456.

Wang, Z., Hutter, F., Zoghi, M., Matheson, D., and de Freitas, N. (2013).
Bayesian Optimization in a Billion Dimensions via Random Embeddings.
arXiv e-prints, page arXiv:1301.1942.

Zahm, O., Cui, T., Law, K., Spantini, A., and Marzouk, Y. (2018).
Certified dimension reduction in nonlinear Bayesian inverse problems.
arXiv e-prints, page arXiv:1807.03712.