Certified Dimension Reduction for Bayesian Updating with the Cross-Entropy Method*

Max Ehre†, Rafael Flock‡, Martin Fußeder§, Iason Papaioannou†, and Daniel Straub†

Abstract. In inverse problems, the parameters of a model are estimated based on observations of the model response. The Bayesian approach is powerful for solving such problems; one formulates a prior distribution for the parameter state that is updated with the observations to compute the posterior parameter distribution. Solving for the posterior distribution can be challenging when, e.g., prior and posterior significantly differ from one another and/or the parameter space is high-dimensional. We use a sequence of importance sampling measures that arise by tempering the likelihood to approach inverse problems exhibiting a significant distance between prior and posterior. Each importance sampling measure is identified by cross-entropy minimization as proposed in the context of Bayesian inverse problems in Engel et al. (2021). To efficiently address problems with high-dimensional parameter spaces we set up the minimization procedure in a low-dimensional subspace of the original parameter space. The principal idea is to analyse the spectrum of the second-moment matrix of the gradient of the log-likelihood function to identify a suitable subspace. Following Zahm et al. (2021), an upper bound on the Kullback-Leibler-divergence between full-dimensional and subspace posterior is provided, which can be utilized to determine the effective dimension of the inverse problem corresponding to a prescribed approximation error bound. We suggest heuristic criteria for optimally selecting the number of model and model gradient evaluations in each iteration of the importance sampling sequence. We investigate the performance of this approach using examples from engineering mechanics set in various parameter space dimensions.

Key words. Bayesian inverse problems, high dimensions, cross-entropy method, importance sampling, certified dimension reduction

AMS subject classifications. 62F15, 62L12, 62P30, 60G60, 65C05

1. Introduction.

We consider inverse problems in the context of a computational model f with $\boldsymbol{y} = f(\boldsymbol{\theta})$. That is, we want to characterise a cause (parameters of the computational model, $\boldsymbol{\theta}$) based on observations of the corresponding effects or consequences of said cause (output of the computational model $\boldsymbol{y}$). An example is a structural system represented with a finite element model that is parameterized by loads, geometric and material properties $\boldsymbol{\theta} \in \mathcal{X} \subseteq \mathbb{R}^d$ and that produces outputs such as stresses and deflections $\boldsymbol{y} \in \mathcal{Y} \subseteq \mathbb{R}^m$. In the majority of applications, we cannot expect the inverse problem to be well defined, i.e., there need not be a solution, the solution may not be unique or it might be very sensitive to the observations [54]. To further complicate the matter, in practice, observations are often incom-

*Submitted to the editors March 12, 2022.

Funding: We acknowledge support by the German Research Foundation (DFG) through Grants STR 1140/11-1 and PA 2901/1-1.

†Technical University of Munich, School of Engineering and Design, Engineering Risk Analysis Group (max.ehre@tum.de, iason.papaioannou@tum.de, straub@tum.de).

‡Technical University of Denmark, Department of Applied Mathematics and Computer Science, DTU Compute (raff@dtu.dk).

§Technical University of Munich, School of Engineering and Design, Chair of Structural Analysis (martin.fusseder@tum.de).

plete and/or contaminated with noise. Here, we focus on the Bayesian approach to inverse problems, which offers a consistent framework for incorporating both noisy and incomplete observations as well as addressing ill-posedness by regularizing the problem using prior information [29, 54]. [29, 54] discuss the Bayesian inverse problem (BIP) in infinite-dimensional settings, while in practice, we usually retreat to the finite-dimensional case by means of discretizing infinite-dimensional random objects such as random fields and processes. Hence, we focus on finite-dimensional BIPs in this work.

We represent $\boldsymbol{\theta}$ and $\boldsymbol{y}$ as real-valued random vectors $\boldsymbol{\Theta} : \mathcal{X}, \mathcal{B}(\mathcal{X}) \to \mathbb{R}$ and $\boldsymbol{Y} : \mathcal{Y}, \mathcal{B}(\mathcal{Y}) \to \mathbb{R}$, where $\mathcal{B}(\cdot)$ is the Borel σ-algebra, and we assume the probability measures $\mathbb{P}_{\boldsymbol{\Theta}}$, $\mathbb{P}_{\boldsymbol{Y}}$ to be absolutely continuous with respect to the respective Lebesgue measures on $\mathcal{B}(\mathcal{X})$ and $\mathcal{B}(\mathcal{Y})$. We then may use the associated probability density functions (PDF) $p(\boldsymbol{\theta})$ and $p(\boldsymbol{y})$ to characterize $\boldsymbol{\Theta}$ and $\boldsymbol{Y}$.

We start by placing a prior distribution on $\boldsymbol{\theta}$ by defining the prior PDF $p_0(\boldsymbol{\theta}) : \mathbb{R}^d \to \mathbb{R}_{\geq 0}$. As the name suggests, $p_0(\boldsymbol{\theta})$ formalizes any information one may have on $\boldsymbol{\theta}$ *prior* to considering any observations. This information may come as the outcome of an expert elicitation, selection rules [41] and/or guiding principles to construct noninformative priors such as Jeffreys's priors [26] or priors satisfying the maximum entropy principle [25]. While many of these principles rest on the idea to minimize the influence the prior exerts on the posterior distribution and thus aim at 'letting the data speak', it is flat/weak priors in particular that can lead to overly confident inference results [18]. A single layer of priors may not do justice to complex models with a large number of unobserved variables, in which case *hierarchical models* with several layers of prior distributions can be utilized [17, Section 2.8].

Next, one or several observations of $\boldsymbol{y}$ that we refer to as $\tilde{\boldsymbol{y}}$, are represented by the *likelihood* $L(\boldsymbol{\theta}) := p(\tilde{\boldsymbol{y}}|\boldsymbol{\theta}) : \mathbb{R}^d \to \mathbb{R}_{\geq 0}$, which states how likely these observations are to occur under any given set of parameters $\boldsymbol{\theta}$. In Bayesian inverse problems, the likelihood will be a function of f, thereby facilitating the backpropagation of information on outputs of f, $\boldsymbol{y}$, to its parameters $\boldsymbol{\theta}$. With this, the posterior PDF of $\boldsymbol{\theta}$ conditional on observations $\tilde{\boldsymbol{y}}$ follows from Bayes' theorem as

$$p_{\boldsymbol{y}}(\boldsymbol{\theta}) := p(\boldsymbol{\theta}|\tilde{\boldsymbol{y}}) = \frac{p(\tilde{\boldsymbol{y}}|\boldsymbol{\theta})p_0(\boldsymbol{\theta})}{p(\tilde{\boldsymbol{y}})} = \frac{L(\boldsymbol{\theta})p_0(\boldsymbol{\theta})}{Z}, \tag{1.1}$$

where

$$Z = \int_{\mathcal{X}} L(\boldsymbol{\theta})p_0(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} \tag{1.2}$$

is the marginal likelihood of the data also known as the *model evidence*. We assume the evidence is finite $Z < \infty$ and the likelihood is Borel-measurable. In the Bayesian approach, solving the inverse problem amounts to computing the posterior distribution of $\boldsymbol{\theta}$ and generating samples from $p_{\boldsymbol{y}}(\boldsymbol{\theta})$. In many instances, the posterior distribution cannot be computed exactly. Instead, sampling approaches such as importance sampling (IS) [19] or Markov Chain Monte Carlo (MCMC) [22] are used to sample from the posterior and construct estimates of

posterior expectations. If prior and posterior distributions differ from one another significantly, constructing efficient *biasing* or *proposal* densities for IS or MCMC, respectively, becomes difficult. Such problems can be addressed by repeatedly applying sampling methods on an artificial sequence of distributions that gradually approach the posterior starting from the prior, namely *sequential Monte Carlo* methods (SMC) [13, 37, 11]. In *adaptive* SMC [30, 24, 32], the distribution sequence is determined during runtime based on intermediate samples.

In SMC approaches, the distributions appearing in the sequence are characterized by samples that are obtained through resample-move steps; samples from each previous distribution are moved via MCMC sampling to obtain samples from the next distribution. However, MCMC produces dependent samples. Alternatively, in *cross-entropy importance sampling* (CE-IS) [50], a sequence of parametrized distributions is defined such that each new distribution in the sequence is identified through solving an optimization (cross-entropy minimization) problem. Estimation of the target distribution is then performed with IS using the final fitted parametric density. Hence, CE-IS avoids MCMC sampling and dependent samples. CE-IS was introduced in the context of rare event estimation in [48] and was recently applied to solve the BIP in [15].

Both acceptance rate and mixing time — and hence, computational efficiency — of many MCMC algorithms deteriorate as the problem dimension d increases [46, 36]; notable exceptions include the *preconditioned Crank-Nicholson* (pCN) sampler [3, 8] and Hamiltonian MCMC [38]. Therefore, different approaches have been proposed to reduce the dimension of the inverse problem by identifying low-dimensional subspaces on which the solution to the original problem may be identified in good approximation. While their existence cannot be guaranteed independent of the inverse problem, low-dimensional subspaces frequently occur in BIPs as a result of f being a smoothing operator applied to the input vector $\boldsymbol{\theta}$, e.g., in the form of solutions to a set of partial differential equations (PDEs). In [35, 57] the problem dimension is reduced by representing the prior with a truncated Karhunen-Loéve-expansion. In the context of linear BIPs, [16, 51] construct low-rank approximations of the prior-preconditioned Hessian of the log-likelihood thereby exploiting structure in both prior and likelihood. The *likelihood-informed subspace* method of [9] extends this approach to nonlinear BIPs based on a low-rank approximation of the posterior-preconditioned Hessian of the log-likelihood. [7] propose a similar approach in which they identify an *active subspace* of the BIP, i.e., a low-rank approximation of the prior-preconditioned negative log-likelihood gradient. Building on the idea of likelihood-informed subspaces, [60] proposes *certified dimension reduction* for nonlinear BIPs and derives an upper bound on the *Kullback-Leibler-divergence* between reduced and full space posterior in function of the subspace dimension.

While the CE-IS approach to BIPs of [15] circumvents MCMC altogether, its performance deteriorates with increasing parameter dimension. This is both due to an increasing degeneracy of the IS weights that are used in the context of CE-IS [49] as well as the rapidly growing number of parameters in the employed distribution models. For example, in Gaussian models with full covariance structure, the number of parameters is $p = d(d+3)/2$, implying that the

number of f-evaluations required to obtain an accurate fit scale quadratically with d. Solving BIPs with CE-IS is therefore only suitable for low parameter dimension. [58] uses CE-IS for estimating rare event probabilities of models with large parameter dimension by applying certified dimension reduction.

In this work, we devise a scheme to efficiently solve nonlinear BIPs using CE-IS and certified dimension reduction. Our method extends the approaches of [15, 58] to address high-dimensional BIPs. Moreover, we introduce heuristic rules for adaptively selecting the number of model and model gradient evaluations during the simulation. Section 2 recapitulates CE-IS for BIPs following [15]. Section 3 details the certified dimensionality reduction approach for CE-IS and Subsection 3.4 contains a discussion on methodology, algorithmic details and a summary of the final procedure. In Section 4, we investigate the efficacy of our method on two structural engineering examples both featuring large parameter dimensions. Concluding remarks are given in Section 5.

2. Cross-entropy-based importance sampling for Bayesian updating.

2.1. Importance sampling and the cross-entropy method. In this chapter, we briefly describe the *CE-based IS method for Bayesian updating* (CEBU) proposed in [15]. Importance sampling is a variance reduction method for estimating expectations of a function $G(\boldsymbol{\theta})$, $\mathbb{E}_p[G(\boldsymbol{\Theta})]$ [50, Chapter 5]. Note that we use lowercase letters for deterministic variables. We use uppercase letters for matrices and random variables/vectors with the exception of random samples, which are denoted with lowercase letters yet treated as random variables. Throughout this work we assume all random vectors to be real-valued, i.e., $\boldsymbol{X} : (\mathcal{X}, \mathcal{B}(\mathcal{X})) \to \mathbb{R}$, where $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ is a measurable space consisting of the outcome space $\mathcal{X}$ and its associated Borel σ-algebra $\mathcal{B}(\cdot)$. Further, we assume probability measures $\mathbb{P}_{\boldsymbol{X}}$ to be absolutely continuous with respect to the respective Lebesgue measures on $\mathcal{B}(\mathcal{X})$ so that we may use the associated PDFs $p(\boldsymbol{x})$ to describe $\boldsymbol{X}$. Let $q(\boldsymbol{\theta})$ be a PDF on $\mathcal{X}$ such that $q(\boldsymbol{\theta}) > 0$ whenever $p(\boldsymbol{\theta}) > 0$ and suppose we only know $\psi(\boldsymbol{\theta}) = cp(\boldsymbol{\theta})$ pointwise with unknown normalizing constant c. Then we can write

$$\mu := \mathbb{E}_p[G(\boldsymbol{\Theta})] = \frac{1}{c}\mathbb{E}_q\left[\frac{G(\boldsymbol{\Theta})\psi(\boldsymbol{\Theta})}{q(\boldsymbol{\Theta})}\right] = \frac{\mathbb{E}_q\left[G(\boldsymbol{\Theta})w(\boldsymbol{\Theta})\right]}{\mathbb{E}_q\left[w(\boldsymbol{\Theta})\right]}, \quad \text{[40, Chapter 9]} \tag{2.1}$$

where q is termed the *importance*, *auxiliary*, *instrumental* or *biasing* density and $w(\boldsymbol{\theta}) = \psi(\boldsymbol{\theta})/q(\boldsymbol{\theta})$ is referred to as the *likelihood ratio* or *IS weight*. **??** leads to the *self-normalized IS estimate*

$$\widehat{\mu}_{\mathrm{IS},q} = \frac{1}{n\widehat{c}}\sum_{k=1}^{n} G(\boldsymbol{\theta}_k)w(\boldsymbol{\theta}_k), \quad \boldsymbol{\theta}_k \overset{i.i.d.}{\sim} q(\boldsymbol{\theta}), \tag{2.2}$$

where an estimate of the normalizing constant is given as $\widehat{c} = n^{-1}\sum_{k=1}^{n} w(\boldsymbol{\theta}_k)$. For many problems q can be chosen such that (2.2) has lower variance $\widehat{\mathbb{V}}[\widehat{\mu}_{\mathrm{IS},q}]$ than the crude Monte Carlo estimate [40, Chapter 9].

In the context of BIPs p is a posterior distribution $p_{\boldsymbol{y}}$ and the normalizing constant c in

(2.1) is the model evidence Z. $p_{\boldsymbol{y}}$ is the optimal IS density to estimate the model evidence as $\mathbb{V}[\widehat{Z}] = 0$ if $q = p_{\boldsymbol{y}}$. Since sampling from the posterior is usually difficult, we continue with a discussion of how to get a parametric q close to $p_{\boldsymbol{y}}$.

[48] proposed finding a parametric IS density $q(\boldsymbol{\theta}, \boldsymbol{v})$ with parameters $\boldsymbol{v} \in \mathcal{V}$ by minimizing the *Kullback-Leibler divergence* (KLD) between $q(\boldsymbol{\theta}, \boldsymbol{v})$ and an optimal IS density in the context of rare event probability estimation. [15] builds on this principle to estimate a parametric distribution that is close to the posterior $p_{\boldsymbol{y}}$ as follows. The KLD between the posterior and the parametric density $D_{\mathrm{KL}}(p_{\boldsymbol{y}}(\boldsymbol{\theta})||q(\boldsymbol{\theta}, \boldsymbol{v}))$ is defined as [50]

$$\begin{aligned} D_{\mathrm{KL}}(p_{\boldsymbol{y}}(\boldsymbol{\theta})||q(\boldsymbol{\theta}, \boldsymbol{v})) &= \mathbb{E}_{p_{\boldsymbol{y}}}\left[\ln\left(\frac{p_{\boldsymbol{y}}(\boldsymbol{\Theta})}{q(\boldsymbol{\Theta}, \boldsymbol{v})}\right)\right] \\ &= \frac{1}{Z}\mathbb{E}_{p_0}[L(\boldsymbol{\Theta})\ln(p_{\boldsymbol{y}}(\boldsymbol{\Theta}))] \underbrace{-\frac{1}{Z}\mathbb{E}_{p_0}[L(\boldsymbol{\Theta})\ln(q(\boldsymbol{\Theta}, \boldsymbol{v}))]}_{\text{cross entropy } H(p_{\boldsymbol{y}}, q(\cdot, \boldsymbol{v}))}. \end{aligned} \tag{2.3}$$

The first summand on the right-hand side of (2.3) is not a function of $\boldsymbol{v}$ so that minimizing $D_{\mathrm{KL}}(p_{\boldsymbol{y}}(\boldsymbol{\theta})||q(\boldsymbol{\theta}, \boldsymbol{v}))$ is equivalent to maximizing the negative cross entropy:

$$\boldsymbol{v} = \arg\max_{\boldsymbol{v} \in \mathcal{V}} \mathbb{E}_{p_0}[L(\boldsymbol{\Theta})\ln(q(\boldsymbol{\Theta}, \boldsymbol{v}))], \tag{2.4}$$

which conveniently does not depend on the unknown Z. An approximate solution of this optimization problem based on samples from p reads

$$\widehat{\boldsymbol{v}} = \arg\max_{\boldsymbol{v} \in \mathcal{V}} \frac{1}{n}\sum_{k=1}^{n} L(\boldsymbol{\theta}_k)\ln(q(\boldsymbol{\theta}_k, \boldsymbol{v})), \quad \boldsymbol{\theta}_k \overset{i.i.d.}{\sim} p_0(\boldsymbol{\theta}). \tag{2.5}$$

The optimization problem in (2.5) is usually convex, continuous and the objective function is differentiable with respect to $\boldsymbol{v}$ such that identifying $\widehat{\boldsymbol{v}}$ is straight-forward. Closed-form solutions of (2.5) exist in various situations, e.g., if $q(\boldsymbol{\theta}, \boldsymbol{v})$ is any member of the exponential family [50, Chapter 8]. [31, 20] use a Gaussian mixture model in order to capture several disconnected failure regions. [59] and [43] use *von Mises-Fisher* and *von Mises-Fisher-Nakagami* (vMFN) mixture models, respectively, to overcome the poor performance of Gaussian models in high-dimensional rare event probability estimation problems. [15] test the performance of both Gaussian mixture and vMFN mixture models in the context of the CE method for BIPs and show that although the latter has superior performance in certain high-dimensional settings, the former possesses higher flexibility and is thus able to accurately describe complicated posteriors. In all these works, different variants of expectation maximization are used to solve for $\widehat{\boldsymbol{v}}$.

How well $q(\boldsymbol{\theta}, \widehat{\boldsymbol{v}})$ approximates $p_{\boldsymbol{y}}(\boldsymbol{\theta})$ hinges on how well samples from p can inform the objective function about $L(\boldsymbol{\theta})$. In other words, if prior and likelihood are not close to one another, we cannot expect the solution of (2.5) to yield a satisfying approximation to $p_{\boldsymbol{y}}(\boldsymbol{\theta})$ independent of the parametric model choice. This problem can be overcome by tempering the likelihood as described in the following section.

2.2. Tempering the likelihood. In order to bridge the distance between prior and likelihood one may break down the single CE problem into several smaller ones. To this end, we define a sequence of PDFs $\{q_t(\boldsymbol{\theta})\}_{j=1}^m$ with

$$q_t(\boldsymbol{\theta}) := \frac{L^{\beta_t}(\boldsymbol{\theta})p_0(\boldsymbol{\theta})}{Z_t}, \tag{2.6}$$

where $Z_t = \int_{\mathcal{X}} L^{\beta_t}(\boldsymbol{\theta})p_0(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}$ and ensuring $0 =: \beta_0 < \beta_1 < \cdots < \beta_{m-1} < \beta_m := 1$ such that $q_0(\boldsymbol{\theta}) := p_0(\boldsymbol{\theta})$ and $q_m(\boldsymbol{\theta}) := p_{\boldsymbol{y}}(\boldsymbol{\theta})$. The idea is to start with samples from $p_0(\boldsymbol{\theta})$ and select β_1 small enough to facilitate an accurate estimate $\widehat{\boldsymbol{v}}_1$. Next, upon selecting $\beta_2 \in (\beta_1, 1]$, samples from $q(\boldsymbol{\theta}, \widehat{\boldsymbol{v}}_1)$ can be used to estimate $\widehat{\boldsymbol{v}}_2$. This procedure is repeated until $\beta_m = 1$ after m steps and the CE problem is solved for the target posterior density. The t-th CE minimization problem reads

$$\widehat{\boldsymbol{v}}_t = \arg\max_{\boldsymbol{v}\in\mathcal{V}} \frac{1}{n}\sum_{k=1}^{n} \ln(q(\boldsymbol{\theta}_k, \boldsymbol{v}))w_t(\boldsymbol{\theta}_k), \quad \boldsymbol{\theta}_k \overset{i.i.d.}{\sim} q(\boldsymbol{\theta}, \widehat{\boldsymbol{v}}_{t-1}), \tag{2.7}$$

with $w_t(\boldsymbol{\theta}) = L^{\beta_t}(\boldsymbol{\theta})p_0(\boldsymbol{\theta})/q(\boldsymbol{\theta}, \widehat{\boldsymbol{v}}_{t-1})$.

In (2.7), the likelihood ratio or weight $w_t(\boldsymbol{\theta})$ accounts for the fact that the t-th PDF parameter estimate $\widehat{\boldsymbol{v}}_t$ is based on samples from the $(t-1)^{\text{th}}$ PDF $q(\boldsymbol{\theta}, \widehat{\boldsymbol{v}}_{t-1})$. The variance of $\widehat{\boldsymbol{v}}_t$ depends on the variance of the weights $w_t(\boldsymbol{\theta})$. In particular, if the numerator PDF of w has fatter tails than its denominator PDF, the weight variance blows up and the parameter estimate $\widehat{\boldsymbol{v}}_t$ deteriorates. The normalized *effective sample size* (nESS) is a common performance metric of IS that is directly related to the variance of the weights [40, Chapter 9]:

$$n_{\text{eff}} = \frac{1}{1+\delta_w^2} \quad \text{with} \quad \delta_w = \frac{\sqrt{\mathbb{V}[w(\boldsymbol{\Theta})]}}{\mathbb{E}[w(\boldsymbol{\Theta})]} \tag{2.8}$$

the coefficient of variation of the weights. Therefore in [15], β_t is computed adaptively in each step such as to achieve a target nESS n^*_{eff} by utilizing a sample-based estimate of the coefficient of variation of the weights $\widehat{\delta}_w$:

$$\beta_t = \arg\min_{\beta_{j-1}<\beta\leq 1} \left(n^*_{\text{eff}} - \frac{1}{1+\widehat{\delta}_w(\beta)^2} \right)^2 = \arg\min_{\beta_{j-1}<\beta\leq 1} \left(n^*_{\text{eff}} - \frac{(\sum_{k=1}^n w(\boldsymbol{\theta}_k;\beta))^2}{\sum_{k=1}^n w^2(\boldsymbol{\theta}_k;\beta)} \right)^2. \tag{2.9}$$

The weights on the right-hand side of (2.9) can be evaluated approximately by assuming $q(\boldsymbol{\theta}, \widehat{\boldsymbol{v}}_{t-1}) = q_{t-1}(\boldsymbol{\theta})$ in each step, such that $w(\boldsymbol{\theta}) \propto L(\boldsymbol{\theta})^{\beta_t - \beta_{t-1}}$ (the factor Z_{t-1} required here for equality cancels out in (2.9) and is immaterial to its solution).

2.3. Method. In [15], CEBU is implemented in the d-dimensional standard-normal space $(\mathcal{U}, \mathcal{B}(\mathcal{U}), \mathbb{P}_{\boldsymbol{U}})$ with $\mathcal{U} = \mathbb{R}^d$, so that the standard-normal random vector $\boldsymbol{U} \sim \varphi_d(\boldsymbol{u})$, where φ_d denotes the d-dimensional standard-normal PDF. Under a suitable isoprobabilistic transformation $T : \boldsymbol{\Theta} \to \boldsymbol{U}$, e.g., using the inverse CDF transform, the Rosenblatt transform [47] or copula models [34, 39, 55], arbitrary priors $p_0(\boldsymbol{\theta})$ are transformed to $\varphi_d(\boldsymbol{u})$ while the likelihood

in standard-normal space is given as $\tilde{L} := (L \circ T^{-1})(\boldsymbol{u})$. Similarly, we define $\tilde{p}_{\boldsymbol{y}} := \tilde{L}(u)\varphi(\boldsymbol{u})/Z$ where the evidence Z is invariant under the transformation T [15, Appendix A].

The CEBU loop terminates once $\beta_t = 1$. Then, a set of n samples is drawn from the final parametric density corresponding to $\beta_t = 1$. These samples are subsequently reweighted to generate samples from the true posterior distribution $\tilde{p}_{\boldsymbol{y}}$. To this end, a final set of weights $\mathbf{w}_{\text{final}}(\boldsymbol{u}) = \tilde{L}(\boldsymbol{u})\varphi_d(\boldsymbol{u})/q(\boldsymbol{u}, \widehat{\boldsymbol{v}}_t)$ is computed as the likelihood ratio of the unnormalized posterior in standard-normal space $\tilde{L}(\boldsymbol{u})\varphi_d(\boldsymbol{u})$ and the parametric density corresponding to $\beta_t = 1$, $q(\boldsymbol{u}, \widehat{\boldsymbol{v}}_t)$. The evidence can be written as

$$Z = \mathbb{E}_{\varphi_d}[\tilde{L}(\boldsymbol{U})] = \mathbb{E}_{q(\boldsymbol{u},\widehat{\boldsymbol{v}}_t)}[\tilde{L}(\boldsymbol{U})\varphi_d(\boldsymbol{U})/q(\boldsymbol{U}, \widehat{\boldsymbol{v}}_t)] = \mathbb{E}_{q(\boldsymbol{u},\widehat{\boldsymbol{v}}_t)}[\mathbf{w}_{\text{final}}], \tag{2.10}$$

which suggests estimating Z as

$$\widehat{Z} = \frac{1}{n}\sum_{k=1}^{n} w_{\text{final}}(\boldsymbol{u}_k), \quad \boldsymbol{u}_k \overset{i.i.d.}{\sim} q(\boldsymbol{u}, \widehat{\boldsymbol{v}}_t). \tag{2.11}$$

A desired number of N weighted posterior samples $\{\boldsymbol{u}_k\}_{k=1}^N$ may then be obtained by resampling the last set of n samples corresponding to $\beta_t = 1$ with replacement and weighted with the normalized final weights $\{\mathbf{w}_{\text{final}}(\boldsymbol{u}_k)/(n\widehat{Z})\}_{k=1}^n$. In [15], the authors use a stratified version of this resampling step based on [14]. In a final step, these samples are transformed back to $\boldsymbol{\Theta}$-space through applying the inverse transform T^{-1}. The entire procedure is summarized in Alg. 2.1.

Algorithm 2.1 CE-BU

Input Likelihood L, transform T, target nESS n^*_{eff}, # post. samples N, # samples/level n
Output posterior samples $\boldsymbol{\Theta}_{\text{post}}$, estimated evidence $\widehat{Z}$

Set $t \leftarrow 0$, $\beta_0 \leftarrow 0$, $\widehat{\boldsymbol{v}}_0$ (so that $q_0(\boldsymbol{u}) = \varphi_d(\boldsymbol{u})$)
while $\beta_t < 1$ **do**
  $t \leftarrow t + 1$
  Sample $\mathbf{U} \in \mathbb{R}^{n\times d} \leftarrow \{\boldsymbol{u}_k \overset{i.i.d.}{\sim} q(\boldsymbol{u}, \widehat{\boldsymbol{v}}_{t-1})\}_{k=1}^n$
  Evaluate $\boldsymbol{\ell} \in \mathbb{R}^{n\times 1} \leftarrow (L \circ T^{-1})(\mathbf{U})$
  Compute β_t with (2.9)
  Evaluate $\mathbf{w} \in \mathbb{R}^{n\times 1} \leftarrow \boldsymbol{\ell}^{\beta_t - \beta_{t-1}}$
  Compute $\widehat{\boldsymbol{v}}_t$ with (2.7)
end while
Sample $\mathbf{U} \in \mathbb{R}^{n\times d} \leftarrow \{\boldsymbol{u}_k \overset{i.i.d.}{\sim} q(\boldsymbol{u}, \widehat{\boldsymbol{v}}_t)\}_{k=1}^n$
Evaluate $\mathbf{w}_{\text{final}} \in \mathbb{R}^{n\times 1} \leftarrow \left\{\frac{(L\circ T^{-1})(\boldsymbol{u}_k)\varphi_d(\boldsymbol{u}_k)}{q(\boldsymbol{u}_k, \widehat{\boldsymbol{v}}_t)}\right\}_{k=1}^n$
Estimate evidence $\widehat{Z} \leftarrow \frac{1}{n}\sum_{k=1}^n \mathrm{w}_{\text{final},k}$
Normalize weights $\bar{\mathbf{w}}_{\text{final}} \leftarrow \mathbf{w}_{\text{final}}/(n\widehat{Z})$
$\mathbf{U}_{\text{post}} \leftarrow$ Resample (with replacement) N times from $\mathbf{U}$ with weighting $\bar{\mathbf{w}}_{\text{final}}$
$\boldsymbol{\Theta}_{\text{post}} \leftarrow T^{-1}(\mathbf{U}_{\text{post}})$
return $\boldsymbol{\Theta}_{\text{post}}$, $\widehat{Z}$

3. CEBUred: Certified dimension reduction for CEBU.

3.1. Linear subspaces of $\mathcal{U}$. [15] test both Gaussian and vMFN mixture models for q. A K-component Gaussian mixture requires fitting $Kd(d+3)/2+K-1$ parameters whereas a K-component vMFN mixture features only $K(d+3)+K-1$ parameters. In spite of the more advantageous linear scaling in d offered by vMFN mixtures, the required number of samples per CE-level can quickly exceed the computational budget if d is large.

In [15], it is shown that the CE method with the Gaussian mixture model is able to obtain accurate representations of posterior densities in various problem settings. However, the Gaussian mixture model performs poorly in high-dimensional IS [20]. This is due to the fact that IS weights with respect to Gaussian densities tend to degenerate in high dimensions. Further, the number of parameters of the GM model increases quadratically with the input dimension. The latter implies that the required number of samples per CE-level to obtain accurate parameter estimates becomes prohibitively large in high dimensions. To alleviate these problems, we draw on the ideas presented in [60] to determine a low-dimensional linear subspace of $\mathcal{X}$ in which an effective IS density can be constructed. The resulting approach can be viewed as an extension of the CE method with failure-informed dimension for rare event estimation, proposed in [58].

In each step of CEBU, for the tempered posterior distribution $\tilde{p}_{\boldsymbol{y},\beta}(\boldsymbol{u}) = Z_\beta^{-1}\tilde{L}^\beta(\boldsymbol{u})\varphi_d(\boldsymbol{u})$ with $Z_\beta = \int_{\mathbb{R}^d}\tilde{L}^\beta(\boldsymbol{u})\varphi_d(\boldsymbol{u})\mathrm{d}\boldsymbol{u}$ we seek an approximation of the form

$$\tilde{p}^{(r)}_{\boldsymbol{y},\beta} \propto (g\circ\mathbf{P}_r)(\boldsymbol{u})\varphi_d(\boldsymbol{u}), \tag{3.1}$$

where $g:\mathbb{R}^d\to\mathbb{R}_{>0}$ is a Borel-measurable function referred to as *profile function* in the following. $\mathbf{P}_r\in\mathbb{R}^{d\times d}$ is a rank-r projection matrix, i.e., $\mathbf{P}_r\circ\mathbf{P}_r=\mathbf{P}_r$. Any $\boldsymbol{u}\in\mathbb{R}^d$ can be decomposed as $\boldsymbol{u}=\mathbf{P}_r\boldsymbol{u}+\mathbf{P}_\perp\boldsymbol{u}=\boldsymbol{u}_r+\boldsymbol{u}_\perp$ with the complementary projection $\mathbf{P}_\perp:=\mathbf{I}_d-\mathbf{P}_r$ satisfying $\mathrm{Im}(\mathbf{P}_\perp)=\mathrm{Ker}(\mathbf{P}_r)$. We call $\mathcal{U}_r:=\mathrm{Im}(\mathbf{P}_r)$ the *likelihood-informed subspace* (LIS) and $\mathcal{U}_\perp:=\mathrm{Im}(\mathbf{P}_\perp)$ the *complementary subspace* (CS). The LIS and CS are at this point still subsets of the ambient space $\mathbb{R}^d$ so that no effective dimension reduction is achieved by their introduction. However, they correspond to lower-dimensional spaces we refer to as local LIS $\bar{\mathcal{U}}_r$ and local CS $\bar{\mathcal{U}}_\perp$, where in standard-normal space $\bar{\mathcal{U}}_r=\mathbb{R}^r$ and $\bar{\mathcal{U}}_\perp=\mathbb{R}^{d-r}$. We discuss the mapping to these local subspaces in more detail in Subsection 3.4. The profile function g is only a function of $\boldsymbol{u}_r\in\mathcal{U}_r$ and is constant in $\boldsymbol{u}_\perp\in\mathcal{U}_\perp$. Following [58], we first define an optimal g for the tempered posterior distributions of CEBU given a projection $\mathbf{P}_r$ in Subsection 3.2. Next, we identify the projection that minimizes the KLD between full and low-rank posterior in Subsection 3.3 and lay out the certified dimensionality reduction for CEBU in Subsection 3.4.

3.2. Optimal profile function g. [60] show that for a given projection matrix $\mathbf{P}_r$, the optimal profile function $g^\star_\beta(\boldsymbol{u})$ that minimizes the KLD $D_{\mathrm{KL}}(\tilde{p}_{\boldsymbol{y},\beta}||\tilde{p}^{(r)}_{\boldsymbol{y},\beta})$, is the following conditional expectation

$$\mathbb{E}_p[\tilde{L}^\beta(\boldsymbol{U})|\mathbf{P}_r\boldsymbol{u}]:\boldsymbol{u}\to\int_{\mathbb{R}^{d-r}}\tilde{L}^\beta(\mathbf{P}_r\boldsymbol{u}+\boldsymbol{\Phi}_\perp\bar{\boldsymbol{u}}_\perp)p_\perp(\bar{\boldsymbol{u}}_\perp|\mathbf{P}_r\boldsymbol{u})\mathrm{d}\bar{\boldsymbol{u}}_\perp, \tag{3.2}$$

where $\boldsymbol{\Phi}_\perp \in \mathbb{R}^{d\times d-r}$ such that $\mathrm{span}(\boldsymbol{\Phi}_\perp) = \mathrm{Ker}(\mathbf{P}_r)$ and $\bar{\boldsymbol{u}}_\perp \in \mathbb{R}^{d-r}$. The conditional PDF $p_\perp(\bar{\boldsymbol{u}}_\perp|\mathbf{P}_r\boldsymbol{u})$ reads

$$p_\perp(\bar{\boldsymbol{u}}_\perp|\mathbf{P}_r\boldsymbol{u}) = \frac{\varphi_d(\mathbf{P}_r\boldsymbol{u} + \boldsymbol{\Phi}_\perp\bar{\boldsymbol{u}}_\perp)}{\int_{\mathbb{R}^{d-r}} \varphi_d(\mathbf{P}_r\boldsymbol{u} + \boldsymbol{\Phi}_\perp\bar{\boldsymbol{u}}'_\perp)\mathrm{d}\boldsymbol{u}'_\perp}, \tag{3.3}$$

which, by convention, equals zero whenever the denominator of (3.3) equals zero. Following from the optimality of (3.2), the optimal reduced posterior reconstruction in standard-normal space reads

$$\tilde{p}^{(r,\star)}_{\boldsymbol{y},\beta} \propto \mathbb{E}_p[\tilde{L}^\beta(\boldsymbol{U})|\mathbf{P}_r\boldsymbol{u}]\varphi_d(\boldsymbol{u}). \tag{3.4}$$

[60] also remarks that the conditional expectation (3.2) is not only optimal with respect to the KL divergence but also minimizes the mean-square reconstruction error of the likelihood function with respect to the prior measure $\mathbb{E}_p[(\tilde{L}^\beta(\boldsymbol{U}) - (g\circ\mathbf{P}_r)(\boldsymbol{U}))^2]$.

3.3. Optimal projection $\mathbf{P}_r$. Under assumptions on the prior distribution that hold in the standard-normal setting [60, Example 2.6], the *subspace logarithmic Sobolev inequality* in [60, Theorem 2.9] states that $\int_{\mathbb{R}^d}\|\nabla h(\boldsymbol{u})\|^2\varphi_d(\boldsymbol{u})\mathrm{d}\boldsymbol{u} \leq \infty$ for any continuously differentiable function $h:\mathbb{R}^d\to\mathbb{R}$ and for any projection $\mathbf{P}_r \in \mathbb{R}^{d\times d}$,

$$\int_{\mathbb{R}^d} h^2(\boldsymbol{u}) \ln\left(\frac{h^2(\boldsymbol{u})}{\mathbb{E}_{\varphi_d}[h(\boldsymbol{U})|\mathbf{P}_r\boldsymbol{u}]}\right)\varphi_d(\boldsymbol{u})\mathrm{d}\boldsymbol{u} \leq 2\int_{\mathbb{R}^d}\|(\mathbf{I}-\mathbf{P}_r^\mathrm{T})\nabla h(\boldsymbol{u})\|^2\varphi_d(\boldsymbol{u})\mathrm{d}\boldsymbol{u}. \tag{3.5}$$

By choosing $h^2(\boldsymbol{u}) = Z_\beta^{-1}\tilde{L}^\beta(\boldsymbol{u})$ we obtain the KLD $D_{\mathrm{KL}}(\tilde{p}_{\boldsymbol{y},\beta}||\tilde{p}^{(r,\star)}_{\boldsymbol{y},\beta})$ on the left-hand side of (3.5). With $\nabla h(\boldsymbol{u}) = \frac{1}{2}(Z_\beta^{-1}\tilde{L}^\beta(\boldsymbol{u}))^{\frac{1}{2}}\nabla\ln\tilde{L}^\beta(\boldsymbol{u})$, an upper bound on $D_{\mathrm{KL}}(\tilde{p}_{\boldsymbol{y},\beta}||\tilde{p}^{(r,\star)}_{\boldsymbol{y},\beta})$ emerges on the right-hand side of (3.5) as

$$\begin{aligned} D_{\mathrm{KL}}(\tilde{p}_{\boldsymbol{y},\beta}||\tilde{p}^{(r,\star)}_{\boldsymbol{y},\beta}) &\leq \frac{1}{2}\int_{\mathbb{R}^d}\|(\mathbf{I}-\mathbf{P}_r^\mathrm{T})\nabla\ln\tilde{L}^\beta(\boldsymbol{u})\|^2\tilde{p}_{\boldsymbol{y},\beta}(\boldsymbol{u})\mathrm{d}\boldsymbol{u} \\ &= \frac{1}{2}\int_{\mathbb{R}^d}\mathrm{tr}\left[(\mathbf{I}-\mathbf{P}_r^\mathrm{T})\beta^2\nabla\ln\tilde{L}(\boldsymbol{u})(\nabla\ln\tilde{L}(\boldsymbol{u}))^\mathrm{T}(\mathbf{I}-\mathbf{P}_r)\right]\tilde{p}_{\boldsymbol{y},\beta}(\boldsymbol{u})\mathrm{d}\boldsymbol{u} \\ &= \frac{1}{2}\mathrm{tr}\left[(\mathbf{I}-\mathbf{P}_r^\mathrm{T})\mathbf{H}(\mathbf{I}-\mathbf{P}_r)\right] =: \frac{1}{2}\mathcal{R}(\mathbf{P}_r,\mathbf{H}), \end{aligned}$$

where $\mathcal{R}(\mathbf{P}_r,\mathbf{H})$ is the mean-squared error incurred by approximating $\nabla\ln\tilde{L}(\boldsymbol{U})$ with $\mathbf{P}_r^\mathrm{T}\nabla\ln\tilde{L}(\boldsymbol{U})$ when $\boldsymbol{U}\sim\tilde{p}_{\boldsymbol{y},\beta}(\boldsymbol{u})$ and we define

$$\mathbf{H} := \beta^2\mathbb{E}_{\tilde{p}_{\boldsymbol{y},\beta}}\left[\nabla\ln\tilde{L}(\boldsymbol{U})(\nabla\ln\tilde{L}(\boldsymbol{U}))^\mathrm{T}\right]. \tag{3.6}$$

Our goal is to find the rank-r projection that minimizes $\mathcal{R}(\mathbf{P}_r,\mathbf{H})$. [60, Proposition 2.11] states that a minimizer of $\mathcal{R}(\mathbf{P}_r,\mathbf{H})$ over all viable projections of rank r is given by the r eigenvectors of $\mathbf{H}$ corresponding to its r leading eigenvalues. Let the solutions of the eigenproblem $\mathbf{H}\boldsymbol{\phi}_i = \boldsymbol{\phi}_i\lambda_i$, $\{\boldsymbol{\phi}_i,\lambda_i\}_{i=1}^d$, be ordered so that $\lambda_1\geq\lambda_2\geq\cdots\geq\lambda_d$, then collecting $\boldsymbol{\Phi}_r := [\boldsymbol{\phi}_1,\boldsymbol{\phi}_2,\ldots,\boldsymbol{\phi}_r] \in \mathbb{R}^{d\times r}$, the optimal projector is given as $\mathbf{P}_r = \boldsymbol{\Phi}_r\boldsymbol{\Phi}_r^\mathrm{T}$. With this

definition of the projection and since the standard-normal prior satisfies inequality (3.5), the accuracy of the reduced posterior can be controlled using a tolerance ϵ as

$$D_{\mathrm{KL}}(\tilde{p}_{\boldsymbol{y},\beta} || \tilde{p}^{(r,\star)}_{\boldsymbol{y},\beta}) \leq \frac{\beta^2}{2} \sum_{i=r+1}^{d} \lambda_i \leq \epsilon. \tag{3.7}$$

Upon selecting ϵ and computing the eigenpairs of $\mathbf{H}$, we choose r as small as possible so that (3.7) holds. Efficient dimension reduction is therefore contingent on a sharp decay of the $\mathbf{H}$-spectrum, which is a property of the computational model f and the observation model $p(\boldsymbol{u}, \boldsymbol{y})$ (i.e., prior and likelihood).

3.4. Method. $\boldsymbol{\Phi}_r$ is the eigenspace of the symmetric matrix $\mathbf{H}$ and thus is an orthogonal basis of $\mathcal{U}_r$. $\boldsymbol{\Phi}_r$ maps the ambient LIS coordinate $\boldsymbol{u}_r \in \mathcal{U}_r$ to its local counterpart $\bar{\boldsymbol{u}}_r \in \bar{\mathcal{U}}_r = \mathbb{R}^r$ as $\bar{\boldsymbol{u}}_r = \boldsymbol{\Phi}_r^{\mathrm{T}} \boldsymbol{u}_r$. In the same way, we define $\boldsymbol{\Phi}_\perp := [\boldsymbol{\phi}_{r+1}, \boldsymbol{\phi}_{r+2}, \ldots, \boldsymbol{\phi}_d] \in \mathbb{R}^{d \times d-r}$, so that the ambient CS coordinate $\boldsymbol{u}_\perp$ is mapped to its local counterpart $\bar{\boldsymbol{u}}_\perp \in \bar{\mathcal{U}}_\perp = \mathbb{R}^{d-r}$ as $\bar{\boldsymbol{u}}_\perp = \boldsymbol{\Phi}_\perp^{\mathrm{T}} \boldsymbol{u}_\perp$. Thus, we can write any $\boldsymbol{u} \in \mathbb{R}^d$ as $\boldsymbol{u} = \boldsymbol{\Phi}_r \bar{\boldsymbol{u}}_r + \boldsymbol{\Phi}_\perp \bar{\boldsymbol{u}}_\perp$ and collect the ambient coordinate of $\mathbb{R}^d$ with respect to the basis defined by $\boldsymbol{\Phi} = [\boldsymbol{\Phi}_r, \boldsymbol{\Phi}_\perp]$ as

$$\bar{\boldsymbol{u}} = \begin{bmatrix} \bar{\boldsymbol{u}}_r \\ \bar{\boldsymbol{u}}_\perp \end{bmatrix} = \boldsymbol{\Phi}^{\mathrm{T}} \cdot \boldsymbol{u} = \begin{bmatrix} \boldsymbol{\Phi}_r^{\mathrm{T}} \\ \boldsymbol{\Phi}_\perp^{\mathrm{T}} \end{bmatrix} \cdot \boldsymbol{u}. \tag{3.8}$$

The LIS and CS in local and ambient coordinates are illustrated in Fig. 1. Due to orthogonality

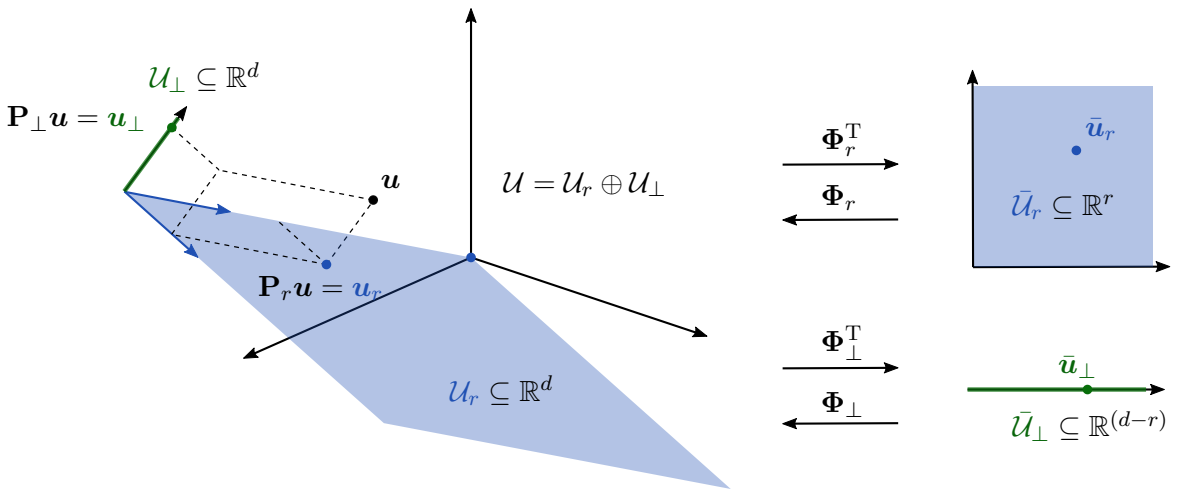

Figure 1. *Ambient space $\mathbb{R}^d$ along with the LIS $\mathcal{U}_r$ and CS $\mathcal{U}_\perp$ as defined by $\mathbf{P}_r$ and $\mathbf{P}_\perp$ (left) and their local counterparts $\bar{\mathcal{U}}_r$ and $\bar{\mathcal{U}}_\perp$ (right).*

of $\boldsymbol{u}_r$ and $\boldsymbol{u}_\perp$ and the rotatinal symmetry of the standard-normal PDF, we may factorize the prior as $\varphi_d(\boldsymbol{u}) = \varphi_r(\bar{\boldsymbol{u}}_r)\varphi_{d-r}(\bar{\boldsymbol{u}}_\perp)$. With this local coordinate prior, the reduced tempered posterior (3.4) reads

$$p^{(r,\star)}_{\boldsymbol{y},\beta}(\bar{\boldsymbol{u}}) \propto \underbrace{\mathbb{E}_{\varphi_d}[\tilde{L}^{\beta_t}(\boldsymbol{U}) | \boldsymbol{\Phi}_r \bar{\boldsymbol{u}}_r] \varphi_r(\bar{\boldsymbol{u}}_r)}_{\text{reduced tempered posterior}} \underbrace{\varphi_{d-r}(\bar{\boldsymbol{u}}_\perp)}_{\text{complementary prior}}. \tag{3.9}$$

By analogy with CEBU, in the t-th step of CEBUred, we approximate the reduced tempered posterior in (3.9) with a parametric model $q^{(r)}(\bar{\boldsymbol{u}}_r, \boldsymbol{v}_{r,t}) : \bar{\mathcal{U}}_r \to \mathbb{R}_{>0}$. The parametric, tempered posterior is

$$q(\bar{\boldsymbol{u}}, \boldsymbol{v}_{r,t}) = q^{(r)}(\bar{\boldsymbol{u}}_r, \boldsymbol{v}_{r,t})\varphi_{d-r}(\bar{\boldsymbol{u}}_\perp). \tag{3.10}$$

Following [58], we select a Gaussian model for $q^{(r)}(\bar{\boldsymbol{u}}_r, \boldsymbol{v}_{r,t})$, although more complicated PDFs such as mixture models may be used as well here. The parameter set $\boldsymbol{v}_{r,t} = \{\boldsymbol{\mu}_{r,t} \in \mathbb{R}^r, \boldsymbol{\Sigma}_{r,t} \in \mathbb{R}^{r\times r}\}$ contains the mean vector $\boldsymbol{\mu}_{r,t}$ and covariance matrix $\boldsymbol{\Sigma}_{r,t}$ of the Gaussian model.

In the t-th step of CEBUred, the new temperature β_t is computed according to (2.9). If $t > 1$, the likelihood in ambient space is evaluated by plugging samples from the previous' step's reduced biasing density $\bar{\boldsymbol{u}}_{r,k} \sim q^{(r)}(\bar{\boldsymbol{u}}_r, \boldsymbol{v}_{r,t})$ and the complementary prior $\bar{\boldsymbol{u}}_{\perp,k} \sim \varphi_{d-r}(\bar{\boldsymbol{u}}_\perp)$ in $\boldsymbol{u}_k = \boldsymbol{\Phi}_{r,t-1}\bar{\boldsymbol{u}}_{r,k} + \boldsymbol{\Phi}_{\perp,t-1}\bar{\boldsymbol{u}}_{\perp,k}$. Thereafter, the gradient covariance matrix $\mathbf{H}$ of the likelihood function with respect to the tempered posterior is estimated to determine the current LIS and CS projections In each step but the first ($t > 1$), a self-normalized IS estimate of $\mathbf{H}_t$ based on samples from the previous biasing density $q(\bar{\boldsymbol{u}}, \widehat{\boldsymbol{v}}_{t-1})$ is computed as

$$\widehat{\mathbf{H}}_t = \frac{\beta_t^2 \sum_{k=1}^{n_\mathbf{H}} \nabla \ln L(\boldsymbol{u}_k)(\nabla \ln L(\boldsymbol{u}_k))^{\mathrm{T}} w_t(\bar{\boldsymbol{u}}_k)}{\sum_{k=1}^{n_\mathbf{H}} w_t(\bar{\boldsymbol{u}}_k)} \qquad \begin{cases} \bar{\boldsymbol{u}}_{r,k} \overset{i.i.d.}{\sim} q^{(r)}(\bar{\boldsymbol{u}}_r, \boldsymbol{v}_{r,t-1}) \\ \bar{\boldsymbol{u}}_{\perp,k} \overset{i.i.d.}{\sim} \varphi_{d-r}(\bar{\boldsymbol{u}}_\perp) \end{cases}. \tag{3.11}$$

If $t = 1$, the weights equal 1 and samples are drawn from the d-dimensional prior in ambient space, $\varphi_d(\boldsymbol{u})$ by setting $\boldsymbol{\Phi}_r = \mathbf{I}_d$ and $\boldsymbol{\Phi}_\perp = \mathbf{0}_d$. For any $t > 1$ the weights are computed as

$$w_t(\bar{\boldsymbol{u}}) = \frac{\tilde{L}^{\beta_t}(\boldsymbol{\Phi}_r\bar{\boldsymbol{u}}_r + \boldsymbol{\Phi}_\perp\bar{\boldsymbol{u}}_\perp)\varphi_r(\bar{\boldsymbol{u}}_r)\varphi_{d-r}(\bar{\boldsymbol{u}}_\perp)}{q(\bar{\boldsymbol{u}}, \boldsymbol{v}_{r,t-1})} = \frac{\tilde{L}^{\beta_t}(\boldsymbol{\Phi}_r\bar{\boldsymbol{u}}_r + \boldsymbol{\Phi}_\perp\bar{\boldsymbol{u}}_\perp)\varphi_r(\bar{\boldsymbol{u}}_r)}{q^{(r)}(\bar{\boldsymbol{u}}_r, \boldsymbol{v}_{r,t-1})}. \tag{3.12}$$

Upon computing the spectrum of $\widehat{\mathbf{H}}_t$, the LIS-dimension r is selected according to (3.7). Once the projections $\boldsymbol{\Phi}_r$ and $\boldsymbol{\Phi}_\perp$ are defined, the parameters of $q^{(r)}(\bar{\boldsymbol{u}}_r, \boldsymbol{v}_{r,t})$ are computed by minimizing the KLD $D_{\mathrm{KL}}(\tilde{p}^{(r,\star)}_{\boldsymbol{y},\beta}(\bar{\boldsymbol{u}})||q(\bar{\boldsymbol{u}}, \boldsymbol{v}_t))$. As in (2.3), this is equivalent to maximizing the negative cross-entropy between the two distributions, i.e.,

$$\begin{aligned}
\boldsymbol{v}_{r,t} &= \arg\min_{\boldsymbol{v}\in\mathcal{V}} D_{\mathrm{KL}}\left(\tilde{p}^{(r,\star)}_{\boldsymbol{y},\beta} \| q(\cdot, \boldsymbol{v})\right) = \arg\max_{\boldsymbol{v}\in\mathcal{V}} -H\left(\tilde{p}^{(r,\star)}_{\boldsymbol{y},\beta}, q(\cdot, \boldsymbol{v})\right) \\
&= \arg\max_{\boldsymbol{v}_r\in\mathcal{V}_r} \int_{\bar{\mathcal{U}}_r}\int_{\bar{\mathcal{U}}_\perp} \mathbb{E}_{\varphi_d}[\tilde{L}^{\beta_t}(\boldsymbol{U})|\boldsymbol{\Phi}_r\bar{\boldsymbol{u}}_r] \ln\left(q^{(r)}(\bar{\boldsymbol{u}}_r, \boldsymbol{v}_r)\right) \varphi_r(\bar{\boldsymbol{u}}_r)\varphi_{d-r}(\bar{\boldsymbol{u}}_\perp)\mathrm{d}\bar{\boldsymbol{u}}_\perp\mathrm{d}\bar{\boldsymbol{u}}_r \\
&= \arg\max_{\boldsymbol{v}_r\in\mathcal{V}_r} \int_{\bar{\mathcal{U}}_r} \mathbb{E}_{\varphi_d}[\tilde{L}^{\beta_t}(\boldsymbol{U})|\boldsymbol{\Phi}_r\bar{\boldsymbol{u}}_r] \ln\left(q^{(r)}(\bar{\boldsymbol{u}}_r, \boldsymbol{v}_r)\right) \varphi_r(\bar{\boldsymbol{u}}_r)\mathrm{d}\bar{\boldsymbol{u}}_r \\
&= \arg\max_{\boldsymbol{v}_r\in\mathcal{V}_r} \mathbb{E}_{\varphi_d}\left[\tilde{L}^{\beta_t}(\boldsymbol{\Phi}_r\bar{\boldsymbol{U}}_r + \boldsymbol{\Phi}_\perp\bar{\boldsymbol{U}}_\perp) \ln\left(q^{(r)}(\bar{\boldsymbol{u}}_r, \boldsymbol{v}_r)\right)\right].
\end{aligned} \tag{3.13}$$

Throughout (3.13) the normalization constant $\tilde{Z}_t$ has been dropped as it is irrelevant for solving the optimization problem. The final equality in (3.13) is a consequence of the factorized

prior in standard-normal space, i.e.,

$$\begin{aligned}\mathbb{E}_{\varphi_d}[\tilde{L}^{\beta_t}(\boldsymbol{U})|\boldsymbol{\Phi}_r\bar{\boldsymbol{u}}_r] &= \int_{\bar{\mathcal{U}}_\perp} \tilde{L}^{\beta_t}(\boldsymbol{\Phi}_r\bar{\boldsymbol{u}}_r + \boldsymbol{\Phi}_\perp\bar{\boldsymbol{u}}_\perp)\frac{\varphi_r(\bar{\boldsymbol{u}}_r)\varphi_{d-r}(\bar{\boldsymbol{u}}_\perp)}{\int_{\mathbb{R}^{d-r}}\varphi_r(\bar{\boldsymbol{u}}_r)\varphi_{d-r}(\bar{\boldsymbol{u}}'_\perp)\mathrm{d}\bar{\boldsymbol{u}}'_\perp}\mathrm{d}\bar{\boldsymbol{u}}_\perp \\ &= \int_{\bar{\mathcal{U}}_\perp} \tilde{L}^{\beta_t}(\boldsymbol{\Phi}_r\bar{\boldsymbol{u}}_r + \boldsymbol{\Phi}_\perp\bar{\boldsymbol{u}}_\perp)\varphi_{d-r}(\bar{\boldsymbol{u}}_\perp)\mathrm{d}\bar{\boldsymbol{u}}_\perp.\end{aligned} \tag{3.14}$$

An IS estimate of $\boldsymbol{v}_{r,t}$ based on samples from $q(\bar{\boldsymbol{u}}, \widehat{\boldsymbol{v}}_{r,t-1})$ reads

$$\widehat{\boldsymbol{v}}_{r,t} = \underset{\boldsymbol{v}_r\in\mathcal{V}_r}{\arg\max}\frac{1}{n}\sum_{k=1}^{n}\ln\left(q^{(r)}(\bar{\boldsymbol{u}}_{r,k}, \boldsymbol{v}_r)\right) w_{t,\mathrm{adj}}(\bar{\boldsymbol{u}}_{r,k}, \bar{\boldsymbol{u}}_{\perp,k}), \qquad \begin{cases}\bar{\boldsymbol{u}}_{r,k} \overset{i.i.d.}{\sim} q^{(r)}(\bar{\boldsymbol{u}}_r, \boldsymbol{v}_{r,t-1}) \\ \bar{\boldsymbol{u}}_{\perp,k} \overset{i.i.d.}{\sim} \varphi_{d-r}(\bar{\boldsymbol{u}}_\perp)\end{cases} \tag{3.15}$$

and requires the computation of the *adjusted weights* $w_{t,\mathrm{adj}}$

$$w_{t,\mathrm{adj}}(\bar{\boldsymbol{u}}_r, \bar{\boldsymbol{u}}_\perp) = \frac{\tilde{L}^{\beta_t}(\boldsymbol{\Phi}_{r,t}\bar{\boldsymbol{u}}_r + \boldsymbol{\Phi}_{\perp,t}\bar{\boldsymbol{u}}_\perp)\varphi_r(\bar{\boldsymbol{u}}_r)\varphi_{d-r}(\bar{\boldsymbol{u}}_\perp)}{q^{(d)}(\bar{\boldsymbol{u}}, \widehat{\boldsymbol{v}}_{t,\mathrm{adj}})}. \tag{3.16}$$

Therein, $\widehat{\boldsymbol{v}}_{t,\mathrm{adj}} = \{\boldsymbol{\mu}_{t,\mathrm{adj}} \in \mathbb{R}^d, \boldsymbol{\Sigma}_{t,\mathrm{adj}} \in \mathbb{R}^{d\times d}\}$ represents the parameters of the d-dimensional Gaussian density $q(\bar{\boldsymbol{u}}, \widehat{\boldsymbol{v}}_{t-1})$ expressed with respect to the updated orthogonal basis $\boldsymbol{\Phi}_t$. Computing adjusted weights with the transformed parameters is necessary to address non-matching bases in the numerator and denominator of (3.16). That is, if the new basis $\boldsymbol{\Phi}_t$ differs from the basis $\boldsymbol{\Phi}_{t-1}$, the complementary prior will no longer be standard-normal with respect to $\boldsymbol{\Phi}_t$. The transformation from $\boldsymbol{\Phi}_{t-1}$ to $\boldsymbol{\Phi}_t$ is linear whereby $q(\bar{\boldsymbol{u}}, \widehat{\boldsymbol{v}}_{t,\mathrm{adj}})$ is Gaussian again and its parameters with respect to $\boldsymbol{\Phi}_t$ can be expressed as

$$\boldsymbol{\mu}_{t,\mathrm{adj}} = \boldsymbol{\Phi}_t^{\mathrm{T}}\underbrace{\boldsymbol{\Phi}_{r,t-1}\boldsymbol{\mu}_{r,t-1}}_{\boldsymbol{\mu}_{t-1}}, \quad \boldsymbol{\Sigma}_{t,\mathrm{adj}} = \boldsymbol{\Phi}_t^{\mathrm{T}}\underbrace{[\boldsymbol{\Phi}_{r,t-1}\boldsymbol{\Sigma}_{r,t-1}\boldsymbol{\Phi}_{r,t-1}^{\mathrm{T}} + \boldsymbol{\Phi}_{\perp,t-1}\boldsymbol{\Phi}_{\perp,t-1}^{\mathrm{T}}]}_{\boldsymbol{\Sigma}_{t-1}}\boldsymbol{\Phi}_t. \tag{3.17}$$

$\boldsymbol{\mu}_{t-1}$ and $\boldsymbol{\Sigma}_{t-1}$ are the mean and covariance vector in ambient space that are subsequently transformed to the reduced spaces given the novel basis $\boldsymbol{\Phi}_t$. This transformation between local and global and subsequent subspaces in steps $t-1$ and t is illustrated in Fig. 2.

3.5. Choosing $n_{\mathbf{H}}$ and n adaptively. $n_{\mathbf{H}}$ is the number of log-likelihood gradient evaluations used to compute $\widehat{H}_t$ in (3.11). n on the other hand is the number of direct evaluations of the tempered likelihood used to estimate the parameters of the t-th biasing density in (3.15). In the absence of f-solvers that are specifically geared towards efficient gradient evaluation such as adjoint solvers [1], computing $\nabla \ln L(\boldsymbol{u})$ is considerably more expensive than evaluating $L(\boldsymbol{u})$.

[6] suggests a heuristic for determining $n_{\mathbf{H}}$ when estimating the second-moment matrix of the gradient ∇f of a Lipschitz-continuous function f, i.e., $\nabla f \leq a$, in order to discover an *active subspace* of f. They use work of [23] on the spectrum of sums of $(n_{\mathbf{H}})$ random matrices to establish bounds on the relative accuracy of the estimated spectrum of $\mathbf{C} = \mathbb{E}_p[\nabla f(\nabla f)^{\mathrm{T}}]$. In the context of CEBUred, we have $f = \beta \log \tilde{L}$ and $p = \tilde{p}_{\boldsymbol{y},\beta}$.

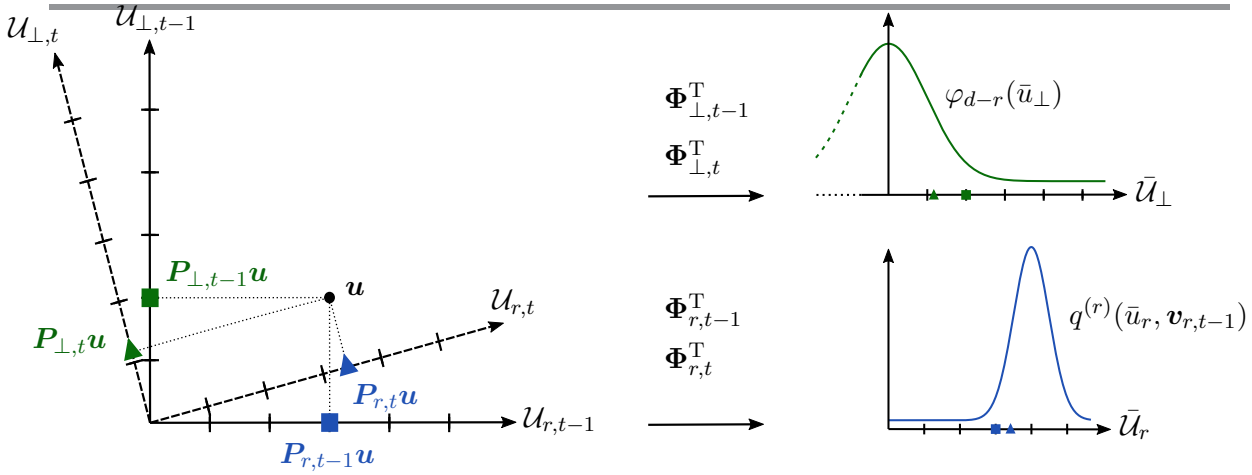

Figure 2. *Left: Mapping between two subsequent orthogonal bases $\boldsymbol{\Phi}_{t-1}$ and $\boldsymbol{\Phi}_t$ in ambient space. Right: Mapping from the two ambient bases to the local LIS (upper right) and CS (lower right).*

[6, Corollary 3.5] states that for $\varepsilon \in (0,1]$, $\mathbb{P}[|\widehat{\lambda}_r - \lambda_r|/\lambda_r \leq \varepsilon] \leq 2d^{-b}$ if the spectrum of $\widehat{\mathbf{H}}$, $\{\widehat{\lambda}_i\}_{i=1}^d$, is computed with $n_{\mathbf{H}} \geq 4(b+1)a\lambda_1 \ln(d)/(\lambda_r\varepsilon)^2$ log-likelihood gradient samples. Drawing on a matrix Bernstein inequality in [56], [6, Corollary 3.8] states that for $\varepsilon \in (0,1]$, $\mathbb{P}[\mathbf{H} - \widehat{\mathbf{H}}\|_2/\|\mathbf{H}\|_2 \leq \varepsilon] \leq 2m^{1-3c/8}$ (the 2-norm of a matrix here is its spectral norm, which also corresponds to its largest singular value) when estimating $\widehat{\mathbf{H}}$ with at least $n_{\mathbf{H}} \geq ca^2 \ln(d)/(\lambda_1\varepsilon^2)$ samples. Finally, choosing ε such that $\varepsilon \leq (\lambda_r - \lambda_{r+1})/(5\lambda_1)$ and using this last lower bound on $n_{\mathbf{H}}$, the distance between the image of the local estimated and true LIS projections is bounded with high probability as well: the distance as measured with the spectral norm $d(\mathrm{Im}(\boldsymbol{\Phi}_r), \mathrm{Im}(\widehat{\boldsymbol{\Phi}}_r)) = \|\boldsymbol{\Phi}_r\boldsymbol{\Phi}_r^{\mathrm{T}} - \widehat{\boldsymbol{\Phi}}_r\widehat{\boldsymbol{\Phi}}_r^{\mathrm{T}}\|_2 = \|\boldsymbol{\Phi}_r^{\mathrm{T}}\widehat{\boldsymbol{\Phi}}_\perp\|_2$ is bounded as $\mathbb{P}[\|\boldsymbol{\Phi}_r^{\mathrm{T}}\widehat{\boldsymbol{\Phi}}_\perp\|_2 \leq 4\lambda_1\varepsilon/(\lambda_r - \lambda_{r+1})] \leq 2m^{1-3c/8}$ according to [6, Corollary 3.10]. [6] translates this bound into a heuristic on account of c, a and the true spectrum $\{\lambda_i\}_{i=1}^d$ being unknown in many use cases involving numerical/simulation models f. The heuristic emerges by summarizing all unknown constants in a fudge factor $\alpha_{\mathbf{H}}$ resulting in

$$n_{\mathbf{H}} = \alpha_{\mathbf{H}} r \ln(d), \tag{3.18}$$

where [6] recommends $\alpha_H \in [2, 10]$ and the target rank r corresponds to the smallest eigenvalue λ_r that shall be estimated with the desired relative accuracy ε. Remarkably, the effort scales logarithmically with the ambient space dimension d suggesting that we can hope to estimate $\mathbf{H}$ with relatively few log-likelihood gradient samples even in very high-dimensional settings. As the target rank r is not known a priori, we detail an iterative procedure to jointly determine r and $n_{\mathbf{H}}$ in Alg. 3.1.

The required number of samples in each level of the CEBU procedure in turn depends on the adaptively selected LIS-rank r through the number of parameters that have to be fitted in the Gaussian reduced biasing density n_{par}. In particular, an r-variate Gaussian requires fitting $n_{\mathrm{par}} = r(r+3)/2$ parameters. To select the number of samples required to accurately

estimate $\widehat{\boldsymbol{v}}_{r,t}$ in (3.15), we use the following heuristic:

$$n(r) = \underbrace{\alpha_{\text{par}}}_{\substack{\text{fudge}\\ \text{factor}}} \underbrace{\frac{1}{2}r(r+3)}_{\substack{\text{number of}\\ \text{parameters}}} \underbrace{(1+\delta_w^2)}_{\substack{\text{inverse}\\ \text{nESS}}}, \tag{3.19}$$

where [6] recommend to chose $\alpha_{\text{par}} \in [2, 10]$. In case an adjoint solver is used for f, the estimation of $\widehat{H}_t$ as in (3.11) will return $n_{\mathbf{H}}$ likelihood evaluations as a byproduct that can be utilizied in estimating $\widehat{\boldsymbol{v}}_{r,t}$ so that only an effective $n(r) - n_{\mathbf{H}}$ new samples need to be drawn and evaluated at each level. The CEBUred algorithm is summarized in Alg. 3.2.

Algorithm 3.1 adapt_H

Input Likelihood and log-gradient $\tilde{L}$ and $\nabla \ln \tilde{L}$, reduced biasing density $q^{(r)}(\bar{\boldsymbol{u}}_r, \boldsymbol{v}_r)$, local LIS & CS projections $\boldsymbol{\Phi}_r$ & $\boldsymbol{\Phi}_\perp$, fudge factor $\alpha_{\mathbf{H}}$, error tolerance ϵ, temperature β_0

Output Subspace samples $\bar{\mathbf{U}}_r, \bar{\mathbf{U}}_\perp$, Likelihood samples $\boldsymbol{\ell}$, local LIS & CS projections $\boldsymbol{\Phi}_r$ & $\boldsymbol{\Phi}_\perp$, temperature β, LIS-rank r, # of $\mathbf{H}$-samples $n_{\mathbf{H}}$

Set $r \leftarrow 1$, $d \leftarrow \text{rowdim}(\boldsymbol{\Phi}_r)$, $n_{\mathbf{H}} \leftarrow \alpha_{\mathbf{H}} \ln(d)$, $\Delta n \leftarrow n_{\mathbf{H}}$, $\bar{\mathbf{U}}_r, \bar{\mathbf{U}}_\perp, \boldsymbol{\ell}, d\boldsymbol{\ell} \leftarrow [\,]$
while $\Delta n > 0$ **do**
  Sample $\bar{\mathbf{U}}_{r,\text{add}} \in \mathbb{R}^{\Delta n \times r} \leftarrow \{\bar{\boldsymbol{u}}_{r,k} \overset{i.i.d.}{\sim} q^{(r)}(\bar{\boldsymbol{u}}_r, \boldsymbol{v}_r)\}_{k=1}^{\Delta n}$
  Sample $\bar{\mathbf{U}}_{\perp,\text{add}} \in \mathbb{R}^{\Delta n \times (d-r)} \leftarrow \{\bar{\boldsymbol{u}}_{\perp,k} \overset{i.i.d.}{\sim} \varphi_{d-r}(\bar{\boldsymbol{u}}_\perp)\}_{k=1}^{\Delta n}$
  Append vertically $\bar{\mathbf{U}}_r \leftarrow [\bar{\mathbf{U}}_r, \bar{\mathbf{U}}_{r,\text{add}}]$, $\bar{\mathbf{U}}_\perp \leftarrow [\bar{\mathbf{U}}_\perp, \bar{\mathbf{U}}_{\perp,\text{add}}]$
  Compute $\boldsymbol{\ell}_{\text{add}} \leftarrow \tilde{L}(\bar{\boldsymbol{U}}_{r,\text{add}}^{\mathrm{T}} \boldsymbol{\Phi}_r + \bar{\boldsymbol{U}}_{\perp,\text{add}}^{\mathrm{T}} \boldsymbol{\Phi}_\perp)$ and $d\boldsymbol{\ell}_{\text{add}} \leftarrow \nabla \tilde{L}(\bar{\boldsymbol{U}}_{r,\text{add}}^{\mathrm{T}} \boldsymbol{\Phi}_r + \bar{\boldsymbol{U}}_{\perp,\text{add}}^{\mathrm{T}} \boldsymbol{\Phi}_\perp)$
  Append vertically $\boldsymbol{\ell} \leftarrow [\boldsymbol{\ell}, \boldsymbol{\ell}_{\text{add}}]$, $d\boldsymbol{\ell} \leftarrow [d\boldsymbol{\ell}, d\boldsymbol{\ell}_{\text{add}}]$
  Evaluate β and $\mathbf{w}$ in function of β_0, $\bar{\boldsymbol{U}}_r$, $\bar{\boldsymbol{U}}_\perp$, $\boldsymbol{\ell}$, $q^{(r)}(\bar{\boldsymbol{U}}_r, \boldsymbol{v}_r)$ with (2.9) & (3.12)
  Evaluate $\widehat{\mathbf{H}}$ in function of β, $\mathbf{w}$ and $d\boldsymbol{\ell}$ with (3.11)
  Evaluate $\{\boldsymbol{\phi}_i, \lambda_i\}_{i=1}^d \leftarrow \text{solve}\{\boldsymbol{\phi} \in \mathbb{R}^d, \lambda \in \mathbb{R} : \widehat{\mathbf{H}}\boldsymbol{\phi} = \boldsymbol{\phi}\lambda\}$
  Select r in function of ϵ and $\{\boldsymbol{\phi}_i, \lambda_i\}_{i=1}^d$ with (3.7)
  Set $\Delta n = \alpha_H r \ln(d) - n_{\mathbf{H}}$
end while
Define $\boldsymbol{\Phi}_r \leftarrow [\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_r]$, $\boldsymbol{\Phi}_\perp \leftarrow [\boldsymbol{\phi}_{r+1}, \ldots, \boldsymbol{\phi}_d]$
return $\bar{\mathbf{U}}_r$, $\bar{\mathbf{U}}_\perp$, $\boldsymbol{\ell}$, $\boldsymbol{\Phi}_r$, $\boldsymbol{\Phi}_\perp$, β, r, $n_{\mathbf{H}}$

4. Experimental results. We perform two numerical examples to demonstrate the capability and test for potential limitations of CEBUred. In the first example, we compare the computational cost and accuracy of CEBUred and CEBU in dependency of the ambient space dimension and verify the results with an analytical solution. In the second example, we examine the performance of CEBUred for different error thresholds as defined by (3.7). Both methods are implemented with a Gaussian model as parametric IS density. In both examples, we infer a material parameter random field based on model output observations. We measure the quality of posterior random field approximations $\widehat{Y}$ against a reference solution Y (either analytical or numerical) in terms of the following spatially averaged relative mean and variance errors:

$$\varepsilon_{\mu_Y} = \frac{\|\mu_Y(\boldsymbol{x}) - \hat{\mu}_Y(\boldsymbol{x})\|_2}{\|\mu_Y(\boldsymbol{x})\|_2} \quad \text{and} \quad \varepsilon_{\sigma_Y^2} = \frac{\|\sigma_Y^2(\boldsymbol{x}) - \hat{\sigma}_Y^2(\boldsymbol{x})\|_2}{\|\sigma_Y^2(\boldsymbol{x})\|_2}. \tag{4.1}$$

Algorithm 3.2 CE-BU-red

Input Likelihood and log-gradients $\tilde{L}$ and $\nabla \ln \tilde{L}$, transform T, parameters $n_{\text{eff}}^*, \alpha_{\mathbf{H}}, \alpha_{\text{par}}, \epsilon$, # post. samples N,

Output posterior samples $\boldsymbol{\Theta}_{\text{post}}$, estimated evidence $\widehat{Z}$

Set $t \leftarrow 0$, $\beta_0 \leftarrow 0$, $\boldsymbol{\Phi}_{r,0} \leftarrow \mathbf{I}_{d\times d}$, $\boldsymbol{\Phi}_{\perp,0} \leftarrow 0$, $\widehat{\boldsymbol{v}}_{r,0} = \{\mathbf{0}_d, \mathbf{I}_{d\times d}\}$,
while $\beta_t < 1$ **do**
  $t \leftarrow t + 1$
  $\bar{\mathbf{U}}_{r,0}, \bar{\mathbf{U}}_{\perp,0}, \boldsymbol{\ell}_0, \boldsymbol{\Phi}_{r,t}, \boldsymbol{\Phi}_{\perp,t}, \beta_t, r_t, n_{\mathbf{H}}$
  $\leftarrow$ adapt_H$(\tilde{L}, \nabla \ln \tilde{L}, q^{(r)}(\bar{\boldsymbol{u}}_r, \boldsymbol{v}_{r,t-1}), \boldsymbol{\Phi}_{r,t-1}, \boldsymbol{\Phi}_{\perp,t-1}, \beta_{t-1}, \alpha_{\mathbf{H}})$
  Compute the required number of samples n in function of r_t with (3.19)
  Sample $\bar{\mathbf{U}}_{r,\text{add}} \in \mathbb{R}^{(n-n_{\mathbf{H}})\times r_t} \leftarrow \{\bar{\boldsymbol{u}}_{r,k} \overset{i.i.d.}{\sim} q^{(r)}(\bar{\boldsymbol{u}}_r, \boldsymbol{v}_{r,t-1})\}_{k=1}^{n-n_{\mathbf{H}}}$
  Sample $\bar{\mathbf{U}}_{\perp,\text{add}} \in \mathbb{R}^{(n-n_{\mathbf{H}})\times(d-r_t)} \leftarrow \{\bar{\boldsymbol{u}}_{\perp,k} \overset{i.i.d.}{\sim} \varphi_{d-r_t}(\bar{\boldsymbol{u}}_\perp)\}_{k=1}^{n-n_{\mathbf{H}}}$
  Compute $\boldsymbol{\ell}_{\text{add}} \in \mathbb{R}^{(n-n_{\mathbf{H}})\times 1} \leftarrow \tilde{L}(\bar{\boldsymbol{U}}_{r,\text{add}}^{\mathrm{T}} \boldsymbol{\Phi}_{r,t} + \bar{\boldsymbol{U}}_{\perp,\text{add}}^{\mathrm{T}} \boldsymbol{\Phi}_{\perp,t})$
  Join vertically $\bar{\mathbf{U}}_r \leftarrow [\bar{\mathbf{U}}_{r,0}, \bar{\mathbf{U}}_{r,\text{add}}]$, $\bar{\mathbf{U}}_\perp \leftarrow [\bar{\mathbf{U}}_{\perp,0}, \bar{\mathbf{U}}_{\perp,\text{add}}]$, $\boldsymbol{\ell} \leftarrow [\boldsymbol{\ell}_0, \boldsymbol{\ell}_{\text{add}}]$
  Compute $\widehat{\boldsymbol{v}}_{t,\text{adj}}(\widehat{\boldsymbol{v}}_{r,t-1}, \boldsymbol{\Phi}_{r,t-1}, \boldsymbol{\Phi}_{r,t}, \boldsymbol{\Phi}_{\perp,t-1}, \boldsymbol{\Phi}_{\perp,t})$ according to (3.17)
  Compute the adjusted weights $\mathbf{w}_{t,\text{adj}} \in \mathbb{R}^{n\times 1} \leftarrow w_{t,\text{adj}}(\bar{\mathbf{U}}_r, \bar{\mathbf{U}}_\perp, \boldsymbol{\ell}, \widehat{\boldsymbol{v}}_{t,\text{adj}})$ with (3.16)
  Compute $\widehat{\boldsymbol{v}}_{r,t}$ with (3.15)
end while
Evaluate $\mathbf{w}_{\text{final}} \in \mathbb{R}^{n\times 1} \leftarrow \left\{ \frac{\tilde{L}(\bar{\mathbf{U}}_r^{\mathrm{T}} \boldsymbol{\Phi}_{r,t} + \bar{\mathbf{U}}_{\perp,t}^{\mathrm{T}} \boldsymbol{\Phi}_\perp) \varphi_{r,t}(\bar{\mathbf{U}}_r^{\mathrm{T}} \boldsymbol{\Phi}_{r,t})}{q^{(r)}(\bar{\mathbf{U}}_r^{\mathrm{T}} \boldsymbol{\Phi}_{r,t}, \widehat{\boldsymbol{v}}_{r,t})} \right\}_{k=1}^{n}$
Estimate evidence $\widehat{Z} \leftarrow \frac{1}{n} \sum_{k=1}^{n} \mathrm{w}_{\text{final},k}$
Normalize weights $\bar{\mathbf{w}}_{\text{final}} \leftarrow \mathbf{w}_{\text{final}}/(n\widehat{Z})$
$\mathbf{U}_{\text{post}} \leftarrow$ Resample (with replacement) N times from $\bar{\mathbf{U}}_r^{\mathrm{T}} \boldsymbol{\Phi}_{r,t} + \bar{\boldsymbol{U}}_{\perp,t}^{\mathrm{T}} \boldsymbol{\Phi}_\perp$ with weighting $\bar{\mathbf{w}}_{\text{final}}$
$\boldsymbol{\Theta}_{\text{post}} \leftarrow T^{-1}(\mathbf{U}_{\text{post}})$
return $\boldsymbol{\Theta}_{\text{post}}$, $\widehat{Z}$

4.1. 1D Cantilever beam.

4.1.1. Problem description. We consider an Euler-Bernoulli beam with one clamped and one free end. It has length $L = 5$ m and a point load of $P = 20$ kN acting on the free end (Fig. 3). Its bending moment $M(x)$ can be obtained from the Euler-Bernoulli equation and reads [5]

$$M(x) = -E(x)I(x)\frac{d^2 w(x)}{dx^2}, \tag{4.2}$$

where $E(x)$ is the beam's Young's modulus and $I(x)$ is its moment of inertia. Both can be summarized as the beam's axial flexibility $F(x) = 1/E(x)I(x)$. The bending moment of the cantilever beam is computed as $M(x) = -P(L - x)$. Hence, the vertical deformation is given by

$$w(x, F(x)) = P \int_0^x \int_0^s (L - x)F(x)\mathrm{d}t\mathrm{d}s. \tag{4.3}$$

The axial flexibility $F(x)$ is considered uncertain and spatially variable along the beam axis. We assign a homogeneous Gaussian prior random field with mean $\mu_F = 10^{-4}\text{kN}^{-1}\text{m}^{-2}$, standard deviation $\sigma_F = 3.5 \cdot 10^{-5}\text{kN}^{-1}\text{m}^{-2}$ and exponential autocorrelation kernel

$$\rho(x, x'; l) = \exp\left(-\frac{\|x - x'\|_1}{l}\right), \tag{4.4}$$

where l is the correlation length. The correlation length of the random field of the axial flexibility is $l_F = 2\text{m}$. The forward model is given by a finite element (FE) model employing 100

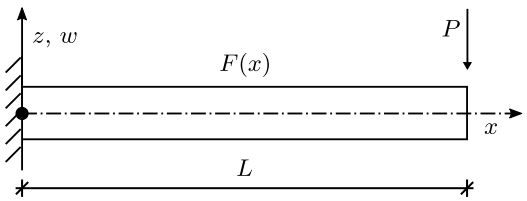

Figure 3. *Cantilever beam with point load.*

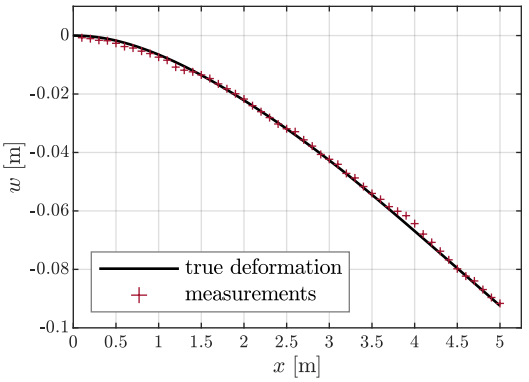

Figure 4. *True deformation and measurements.*

Euler-Bernoulli beam elements with cubic shape functions. The goal is to obtain samples of the posterior distribution of the axial flexibility given $n_{\text{obs}} = 50$ equally spaced measurements $\{x_{\text{meas},i}\}_{i=1}^{n_{\text{obs}}}$ of the vertical deformation along the beam axis (see Fig. 4). Adjoint methods [1] are a computationally efficient tool for obtaining the model gradients required to compute $\nabla\tilde{L}$ as long as the number of model outputs of which derivatives are computed (n_{obs}) is smaller than the number of model inputs (d) with respect to which derivatives are computed. In the context of this example, the adjoint method is thus used for the setting $d = 100$ only and the direct method is used in all other settings.

We assume the measurements to be corrupted by the additive, centered Gaussian noise vector $\boldsymbol{\eta} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\eta\eta})$. The noise covariance matrix is defined as $[\boldsymbol{\Sigma}_{\eta\eta}]_{ij} = \sigma_\eta^2 \rho(x_{\text{meas},i}, x_{\text{meas},j})$ with noise standard deviation $\sigma_\eta = 0.001\text{m}$, exponential correlation kernel $\rho(\cdot, \cdot; l_\eta)$ and correlation length $l_\eta = 1\text{m}$. The random vector describing the vertical deformations in data space $\mathbb{R}^{n_{\text{obs}}}$, i.e., at the n_{obs} measurement locations $\{x_{\text{meas},i}\}_{i=1}^{n_{\text{obs}}}$ is defined as

$$\tilde{\boldsymbol{w}} = \boldsymbol{w} + \boldsymbol{\eta}. \tag{4.5}$$

Given a set of realizations of $\tilde{\boldsymbol{w}}$, i.e., observational data $\tilde{\boldsymbol{y}}$, the likelihood function reads

$$L(\boldsymbol{F}) = \frac{1}{\sqrt{(2\pi)^{n_{\text{obs}}} \det(\boldsymbol{\Sigma}_{\eta\eta})}} \exp\left([\tilde{\boldsymbol{y}} - \mathcal{G}(\boldsymbol{F})]\boldsymbol{\Sigma}_{\eta\eta}^{-1}[\tilde{\boldsymbol{y}} - \mathcal{G}(\boldsymbol{F})]^{\mathrm{T}}\right), \tag{4.6}$$

where $\mathcal{G}(\cdot)$ represents the FE-model and returns the vertical deformations of the beam at the n_{obs} measurement locations.

The measurements for this example are obtained by generating a single random realization of the prior random field of the axial flexibility, solving (4.3) numerically at 1001 equally spaced discretization points and then adding randomly generated noise according to (4.5) to the solutions at the locations of the measurements $\{x_{\text{meas},i}\}_{i=1}^{n_{\text{obs}}}$. By using the analytical expression instead of the FE-model for the generation of the measurements, we avoid the so-called 'inverse crime' [29].

4.1.2. Analytical posterior. The following derivations closely follow [57] where the example is investigated as well. Since $F(x)$ is Gaussian and $w(x, F(x))$ is a linear function of $F(x)$ (4.2), the prior distribution of $w(x)$ is also Gaussian. Its mean and covariance read

$$\mu_w(x) = P \int_0^x \int_0^s (L-t)F(t)\mathrm{d}t\mathrm{d}s = \frac{P\mu_F}{6}x^2(3L-x) \quad \text{and} \tag{4.7a}$$

$$\Sigma_{ww}(x,x') = P \int_0^{x'} \int_0^x \int_0^{s'} \int_0^s (L-t)(L-t')\Sigma_{FF}(t,t')\mathrm{d}t\mathrm{d}t'\mathrm{d}s\mathrm{d}s'. \tag{4.7b}$$

The explicit expression of (4.7b) is obtained using a computer algebra system and omitted here due to its tedious form.

An analytical solution of the posterior of the axial flexibility can be derived, since both the prior and the likelihood are Gaussian [45]. To this end, the Gaussian random vector $\boldsymbol{F}' = [\boldsymbol{F}; \tilde{\boldsymbol{w}}]$ is considered, which contains the discretized random flexibility field, $\boldsymbol{F} \in \mathbf{R}^n$ and the n_{obs} deformation measurements $\tilde{\boldsymbol{w}} \in \mathbf{R}^{n_{\text{obs}}}$. The mean vector and covariance matrix of $\boldsymbol{F}'$ may be partitioned as

$$\boldsymbol{\mu}_{F'} = \begin{bmatrix} \boldsymbol{\mu}_F \\ \boldsymbol{\mu}_{\tilde{w}} \end{bmatrix} \quad \text{and} \quad \boldsymbol{\Sigma}_{F'F'} = \begin{bmatrix} \boldsymbol{\Sigma}_{FF} & \boldsymbol{\Sigma}_{F\tilde{w}} \\ \boldsymbol{\Sigma}_{F\tilde{w}}^{\mathrm{T}} & \boldsymbol{\Sigma}_{\tilde{w}\tilde{w}} \end{bmatrix}. \tag{4.8}$$

As $\boldsymbol{F}'$ is jointly Gaussian, the posterior $\boldsymbol{F}|\tilde{\boldsymbol{y}}$ is Gaussian as well and has PDF

$$p(\boldsymbol{f}|\tilde{\boldsymbol{y}}) = \frac{1}{\sqrt{(2\pi)^n \det(\boldsymbol{\Sigma}_{FF|\tilde{y}})}} \exp\left(-\frac{1}{2}[\boldsymbol{f}-\boldsymbol{\mu}_{F|\tilde{y}}]^{\mathrm{T}}\boldsymbol{\Sigma}_{FF|\tilde{y}}^{-1}[\boldsymbol{f}-\boldsymbol{\mu}_{F|\tilde{y}}]\right). \tag{4.9}$$

The posterior mean and covariance matrix are equal to the following conditional mean $\boldsymbol{\mu}_{F|\tilde{y}}$ and covariance matrix $\boldsymbol{\Sigma}_{FF|\tilde{y}}$:

$$\boldsymbol{\mu}_{F|\tilde{y}} = \boldsymbol{\mu}_F + \boldsymbol{\Sigma}_{F\tilde{w}}\boldsymbol{\Sigma}_{\tilde{w}\tilde{w}}^{-1}(\tilde{\boldsymbol{y}} - \boldsymbol{\mu}_{\tilde{w}}) \quad \text{and} \quad \boldsymbol{\Sigma}_{FF|\tilde{y}} = \boldsymbol{\Sigma}_{FF} - \boldsymbol{\Sigma}_{F\tilde{w}}\boldsymbol{\Sigma}_{\tilde{w}\tilde{w}}^{-1}\boldsymbol{\Sigma}_{F\tilde{w}}^{\mathrm{T}}. \tag{4.10}$$

All quantities in (4.10) are computed within the partition in (4.8) except from $\boldsymbol{\mu}_{\tilde{w}}$, which is obtained by $\mathbb{E}[\tilde{\boldsymbol{w}}] = \mathbb{E}[\boldsymbol{w} + \boldsymbol{\eta}] = \boldsymbol{\mu}_w$.

4.1.3. Parameters of numerical study. The flexibility random field is discretized in space using a midpoint method [12] with d collocation points. d is therefore the ambient space dimension of the Bayesian inverse problem, where scenarios $d = \{10, 25, 50, 100\}$ are investigated. We use CEBU and CEBUred to obtain samples from the d-dimensional posterior distribution of the axial flexibility given a set of $n_{\text{obs}} = 50$ measurements. We use $\delta_{v,target} = 1.5$ and $n = 1000$ samples per level for all CEBU runs. For CEBUred we chose $\delta_{v,target} = 1.5$, $\alpha_{\text{par}} = 4$, $\alpha_{\mathbf{H}} = 6$ and $\varepsilon = 1.0$. Results are averaged over 54 repeated runs of both CEBU and CEBUred.

4.1.4. Discussion of results. Fig. 5 shows the posterior flexibility fields obtained with both CEBU and CEBUred at varying ambient dimension. At $d = 5$, the results obtained with both CEBU and CEBUred coincide with the analytical reference posterior as indicated by the almost congruent scatter points in the top left panel of Fig. 5. However, discretizing the flexibility field with only 5 subparts does not allow for an accurate representation of the posterior field at the clamping. There, the axial flexibility exerts the strongest influence on the beam deformation thus requiring a finer discretization.

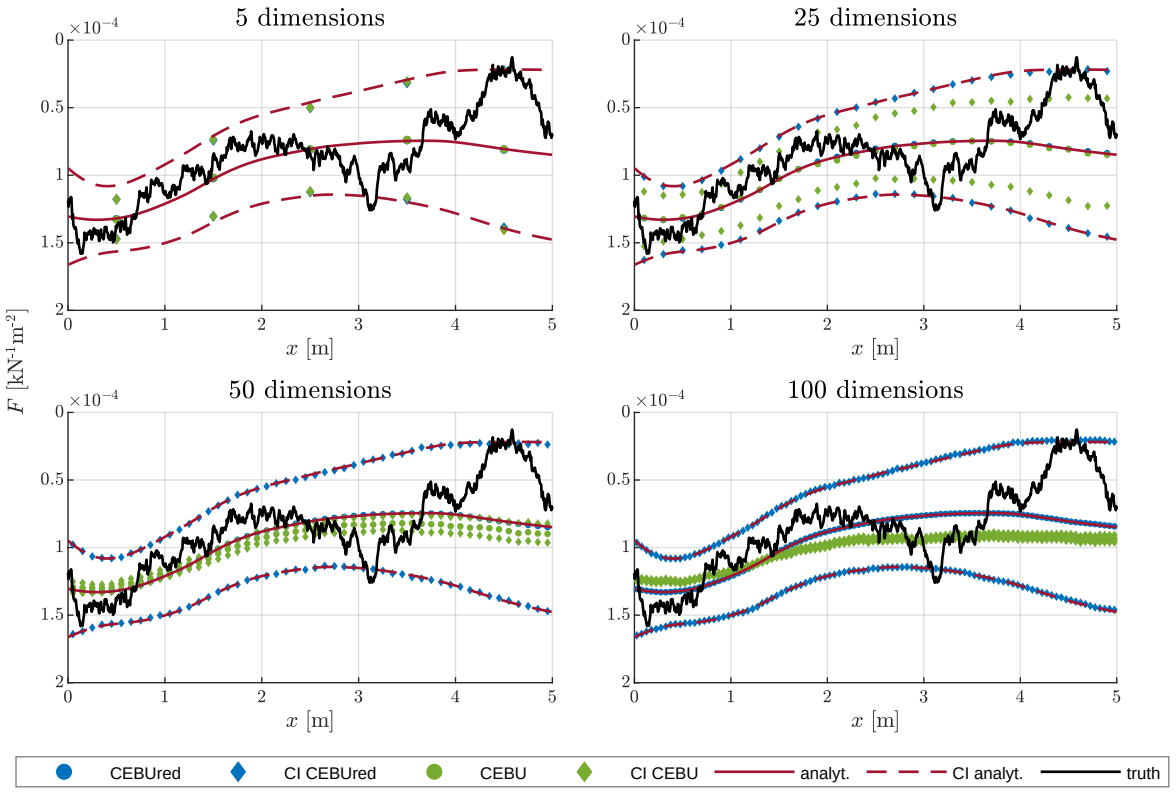

Figure 5. *Axial flexibility posterior field: mean and 95% credible intervals. CEBU and CEBUred solutions are plotted at the random field collocation points (midpoints).*

As d increases, the results obtained by CEBU deteriorate, as indicated by both an increasing deviation of the CEBU solution from its analytical counterpart in both mean and 95% posterior credible bounds. At $d > 11$, CEBU has too few samples available to accurately fit all biasing density parameters in ambient space. CEBUred, on the other hand, agrees closely with the analytical solution if the chosen discretization is fine enough.

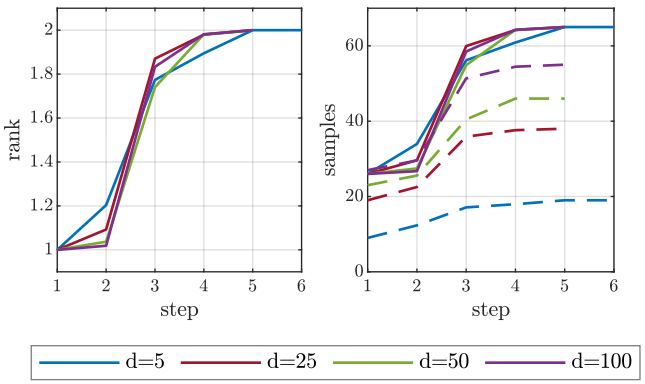

Figure 6. *Left: Average number of selected LIS dimension with increasing CEBUred step index. Right: Average number of evaluated samples per step with increasing CEBUred step index. The solid and the dashed lines represent the total number of model and model gradient evaluations, respectively.*

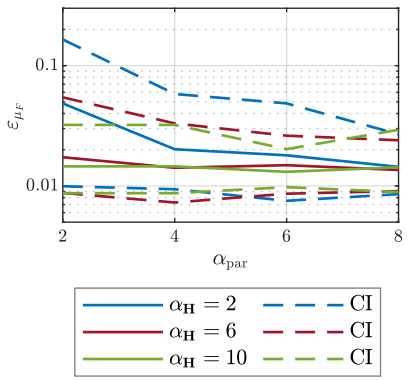

Figure 7. *Relative posterior mean error for different combinations of $\alpha_{\mathbf{H}}$ and α_{par} at $d = 100$.*

Fig. 6 shows the number of selected ranks r (corresponds to the dimension of the effectively used subspace (LIS) in CEBUred) and number of samples plotted over the CEBUred step index. The number of LIS dimensions reduces to $r = 1$ within the first step for all tested d. The number of runs per number of steps for different d are shown in Table 1. At $d = 5$, one of the 54 runs terminated after 6 steps, whereas all other simulations terminated after a maximum of 5 steps. Table 1 suggests that for the given FE-discretization, finer random field discretizations tend to stabilize the simulation in the sense that most runs require the same number of steps.

Table 1

Number of runs broken down according to required number of CEBUred steps at varying d.

# of steps	$d = 5$	$d = 25$	$d = 50$	$d = 100$
2	1	0	0	0
3	15	5	1	1
4	29	43	42	46
5	8	6	11	7
6	1	0	0	0

For all investigated ambient dimensions d and in each CEBUred step, the beam problem possesses very low-dimensional (likelihood-informed) subspaces, in which the inverse problem

Table 2

Likelihood and gradient evaluations for the beam example per run of CEBU and CEBUred (CEBUred: averaged over 54 runs).

Problem dimension	CEBU (Likelihood calls)	CEBUred (Likelihood calls)	CEBUred (Gradient calls)
$d = 5$	3796	170.0	54.3
$d = 10$	3870	181.0	115.8
$d = 25$	3833	184.6	143.5
$d = 100$	3833	182.7	168.6

can be solved efficiently ($r = 1 - 2$). In this LIS, significantly less samples are required to accurately characterize biasing densities compared to CEBU, which operates in d-dimensional ambient space. Fig. 6 (solid lines) illustrates the correspondence of the number of required samples in CEBUred with the reduced space dimension (rank r). As d increases, more gradients are evaluated (Fig. 6, right, dashed lines). This is due to the factor $\ln d$ in (3.18).

In Table 2 we list the average number of required likelihood and likelihood gradient evaluations for both CEBU and CEBUred. The number of required likelihood evaluations within CEBUred remains approximately constant across all investigated dimensions and is more than an order of magnitude lower compared to number of evaluations required by CEBUred. The numbe of likelihood gradient evaluations grows with d but remains below the number of likelihood evaluations. Depending on the method of evaluating these gradients, a gradient call may however be considerably more expensive than a likelihood call.

Table 3

Relative posterior mean and variance error at varying d averaged over 54 repeated runs of CEBUred.

error	$d = 5$	$d = 25$	$d = 50$	$d = 100$
ε_{μ_F}	0.0588	0.0219	0.0167	0.0142
$\varepsilon_{\sigma_F^2}$	0.2168	0.1039	0.1090	0.0907

$\mu_F(\boldsymbol{x})$ and $\sigma^2(\boldsymbol{x})$ are the analytical posterior mean and variance, respectively, evaluated at the discretization points and $\hat{\mu}_F(\boldsymbol{x})$ and $\hat{\sigma}_F^2(\boldsymbol{x})$ are their sample-based counterparts obtained with CEBUred. Table 3 shows the relative error of the mean and the variance for the different dimensions d. What is not immediately obvious from the plots in Fig. 5 is that a finer discretization indeed leads to smaller relative errors. However, the decrease slows down from $d = 25$ and is rather small between $d = 50$ and $d = 100$. Fig. 7 shows the relative posterior mean error ε_{μ_F} for different combinations of the fudge factors α_H and α_{par} at $d = 100$. The relative posterior mean error decreases signifcantly between $\alpha_{\mathbf{H}} = 2$ and $\alpha_{\mathbf{H}} = 6$ and remains constant as $\alpha_{\mathbf{H}}$ is increased from 6 to 10. This is likely due to the fixed error threshold of $\epsilon = 1.0$, which prescribes the approximation quality of the optimal projector. Once the relevant part of the $\mathbf{H}$-spectrum (corresponding to the choice of ϵ) is estimated accurately, increasing $\alpha_{\mathbf{H}}$ bears no further effect. In this case, larger values of α_{par} only lead to a larger number of model evaluations, which in turn can lead to a better fit of the biasing density. At

$\alpha_{\mathbf{H}} = 2$, the error decreases with increasing α_{par}, whereas it remains approximately constant when increasing α_{par} at $\alpha_{\mathbf{H}} \geq 6$. In the latter case, the large number of gradient evaluations (each of which also yields a model evaluation) are already sufficient to accurately estimate the parameters of the reduced biasing density such that increasing α_{par} will not further reduce the error.

4.2. 2D plate in plane stress. The example was first presented in [33] in the context of uncertainty quantification. We consider the adapted version from [57]. Through this example, we investigate how the accuracy of the resulting posterior improves by using different error thresholds ϵ as defined by (3.7).

4.2.1. Problem description. We consider a 2D square steel plate in plane-stress with side length 32 cm, thickness $t = 1$ cm and a hole with radius $r = 2$ cm located at its center (Fig. 8). The plate is clamped at the left-hand side and loaded with a constant line load $q = 6$ kN cm^{-2} acting on its right-hand side. The plate has density $\rho = 7850$ kg m^{-3}, which is required to account for body forces (oriented in negative x_2-direction), and the Poisson ratio is $\nu = 0.29$.

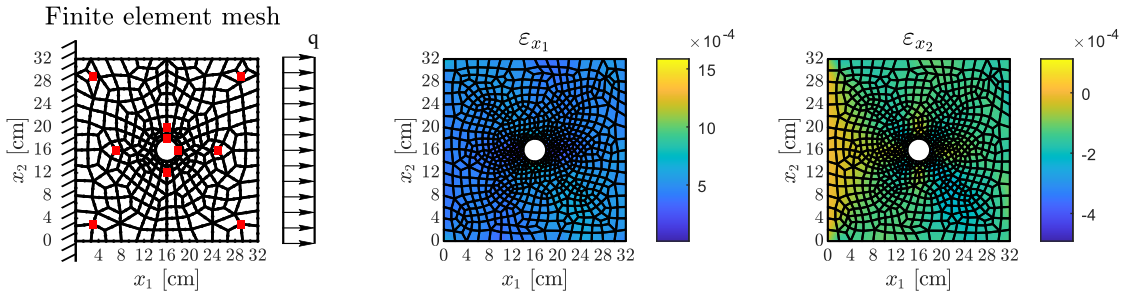

Figure 8. *Left: FE-model of the plate. The red squares mark the positions of the strain gauges. Center and right: True fields of the strains in x_1- and x_2-direction.*

Assuming plane stress, the displacement field $\boldsymbol{u}(x_1, x_2) = [u_{x_1}(x_1, x_2), u_{x_2}(x_1, x_2)]^{\mathrm{T}}$ can be computed implicitly based on elasticity theory through a set of elliptic PDEs (Cauchy-Navier equations) [28]:

$$G(x_1, x_2)\nabla^2 \boldsymbol{u}(x_1, x_2) + \frac{E(x_1, x_2)}{2(1-\nu)} \nabla(\nabla \cdot \boldsymbol{u}(x_1, x_2)) + \boldsymbol{B} = 0. \tag{4.11}$$

$G(x_1, x_2) := E(x_1, x_2)/(2(1+\nu))$ is the shear modulus, $E(x_1, x_2)$ is Young's modulus, and $\boldsymbol{B} = [b(x_1), b(x_2)]^{\mathrm{T}}$ is the vector of body forces acting on the plate. In order to solve (4.11), an FE model with 282 eight-noded quadrilateral finite elements is used (Fig. 8).

In this example, the plate's Young's modulus is considered uncertain and spatially variable. We assign a homogeneous random field prior with log-normal marginal distributions with mean $\mu_E = 2 \cdot 10^4$ kN cm^{-2} and standard deviation $\sigma_E = 3 \cdot 10^3$ kN cm^{-2}. The mean and standard deviation of the underlying Gaussian field $\ln E(x_1, x_2)$ follow as $\mu_{\ln E} = 9.89$ and $\sigma_{\ln E} = 0.15$, respectively, and its correlation structure is modelled with the exponential kernel of (4.4) and correlation length $l_{\ln E} = 10$ cm.

We discretize $\ln E$ by means of a Karhunen- Loève-expansion (KL-expansion). To this end, we solve the following homogeneous Fredholm integral equation of the second kind [21]

$$\sigma^2_{\ln E} \int_D \rho(\boldsymbol{x}, \boldsymbol{x}'; l_{\ln E}) \phi_k(\boldsymbol{x}') \mathrm{d}\boldsymbol{x}' = \lambda_k \phi_k(\boldsymbol{x}) \tag{4.12}$$

for the covariance kernel's set of eigenpairs $\{\lambda_k, \phi_k\}$. Consequently, we can express the log-normal Young's modulus prior as the exp of a KL-expansion [21] like

$$E(x_1, x_2; \boldsymbol{\theta}) = \exp\left[\mu_E + \sum_{k=1}^{\infty} \sqrt{\lambda_k} \phi_k(x_1, x_2) \theta_k\right], \tag{4.13}$$

where the coefficients θ_k are pairwise independent standard-normal Gaussian random variables.

We estimate the set of eigenpairs $\{\lambda_k, \phi_k\}$ for the KL-expansion by solving (4.12) using the Nyström method on a grid of 160×160 Gauss-Legendre quadrature points. The eigenfunctions are interpolated at the numerical integration points of the elements of the FE-model [44].

Truncating the KL-expansion (4.13) after M terms results in an M-order KL-approximation of E, which we denote as $\hat{E}(\boldsymbol{x}; \boldsymbol{\theta})$. This approximation recovers the random field mean exactly, however is associated with an under-representation of its variance $\ln \sigma^2_E$. This under-representation is often measured with the global relative variance error of the M-order KL-approximation:

$$\bar{\epsilon}_{\ln \sigma^2} = \frac{1}{|D|} \int_D \left| \frac{\mathbb{V}[E(\boldsymbol{x}; \boldsymbol{\theta})] - \mathbb{V}[\hat{E}(\boldsymbol{x}; \boldsymbol{\theta})]}{\mathbb{V}[E(\boldsymbol{x}; \boldsymbol{\theta})]} \right| \mathrm{d}\boldsymbol{x} = 1 - \frac{1}{|D| \cdot \ln \sigma^2_E} \sum_{k=1}^{M} \lambda_k. \tag{4.14}$$

Therein, D is the spatial domain of the random field E. The inference task for the example consists in learning the Young's modulus' posterior distribution based on strain measurements at $n_{\text{obs}} = 10$ positions on the plate. At each position, two gauges measure the strain in x_1- and x_2-direction (Fig. 8, left, red squares), respectively. Hence, a set of 20 measurements are available to solve the inference task. We generate the measurements artificially by using a FE-model on a finer mesh of 779 elements, in order to once again avoid the 'inverse crime' [29]. The true strains are depicted in the center and right plot in Fig. 8.

The strain measurements are generated by solving the forward problem based on a single Young's modulus prior random field realization. This realization is generated using a midpoint method discretized at the numerical integration (Gauss) points of the plate FE model rather than a KL-approximation. Consequently, noise is added to the computed strains at the measurement locations. We model the noise as a centered Gaussian random vector $\boldsymbol{\eta} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\eta\eta})$. The noise standard deviation is set to $\sigma_\eta = 10^{-4}$ and the autocorrelation of both x_1- and x_2-strain measurements is modelled with the exponential kernel (4.4) using a correlation length of $l_\eta = 10$ cm. The cross-correlation function between x_1- and x_2-strain measurements is taken as the autocorrelation function multiplied by a cross-correlation coefficient of 0.25.

4.2.2. Numerical reference posterior. We use *adaptive Bayesian Updating with Subset Simulation* (aBUS-SuS) to verify the solution obtained with CEBUred. aBUS-SuS has been tested on a variety of engineering applications, e.g., in [52, 53, 4, 27]. aBUS-SuS recasts the Bayesian inverse problem as a structural reliability problem [57]. *Subset simulation* (SuS) [2] is a robust and efficient method for solving such structural reliability problems and within aBUS-SuS, SuS is employed to solve general Bayesian inverse problems. SuS itself requires carrying out an MCMC sampling step for which we use a pCN sampler [8] with adaptive scaling [42].

4.2.3. Parameters of numerical study. We discretize the Young's modulus random field by a KL-approximation (4.13) with 879 terms producing a Bayesian inverse problem with ambient dimension $d = 879$. The chosen number of terms accounts for at least 97% of the spatial variance of the random field meaning the average variance error (4.14) is $\leq 3\%$.

The inference task is solved by using CEBUred with different error tolerances ϵ. We choose $\epsilon = \{1.0, 10^{-1}, 10^{-2}, 10^{-3}\}$. The remaining parameters are set as $\delta_{v,target} = 1.5$, $\alpha_{\text{par}} = 3$, and $\alpha_{\mathbf{H}} = 4$. The gradients of the likelihood function required at each step of CEBUred are evaluated with the adjoint method [1] (derived for this particular problem in [58, Appendix A]). In the final step, we draw $N = 10000$ samples from the approximate posterior distribution. The parameters for the numerical reference posterior generated with aBUS-SuS are $n = 20000$ for both samples per subset level and final samples of the approximated posterior and intermediate conditional probability $p = 0.1$. Except for the reference posterior, which is computed once only, we repeat each analysis with CEBUred 40 times and average all results over the individual runs.

4.2.4. Discussion of results. Fig. 9 compares the posterior fields obtained with CEBUred using an error threshold of $\epsilon = 10^{-2}$ and the reference posterior obtained with aBUS-SuS along six sections across the plate. Along each section, the CEBUred-based posterior means are in good agreement with the reference posterior mean. For all other tested error thresholds ($\epsilon \in \{1, 10^{-1}, 10^{-2}\}$), similar results are obtained for the average posterior means and variances taken over 40 repeated CEBUred runs (see Table 5). The coefficients of variation of the posterior mean and variance estimates are rather large for $\epsilon \geq 10^{-1}$, but decrease significantly between $\epsilon = 10^{-1}$ and $\epsilon = 10^{-2}$ based on the results given in Table 5.

The numbers of required likelihood and likelihood gradient evaluations for aBUS-SuS and CEBUred at the four tested error threshold are listed in Table 4. At $\epsilon = 10^{-2}$, CEBUred reduces the number of required likelihood calls by roughly two orders of magnitude compared to the reference aBUS-SuS run, which comes at the cost of 642 additional gradient calls.

Fig. 10 shows the mean ranks and the corresponding number of model and gradient evaluations. To ensure an accurate construction of the LIS, the eigenvectors corresponding to the r largest eigenvalues of $\mathbf{H}$ must be reasonably well estimated. The number of samples required for the estimation is determined with the heuristic formula given by (3.18). According to this formula, the number of samples for the estimation of $\mathbf{H}$ linearly depends on r. Therefore, the lines in the left plot and the dashed lines in the right plot in Fig. 10 are linearly dependent.

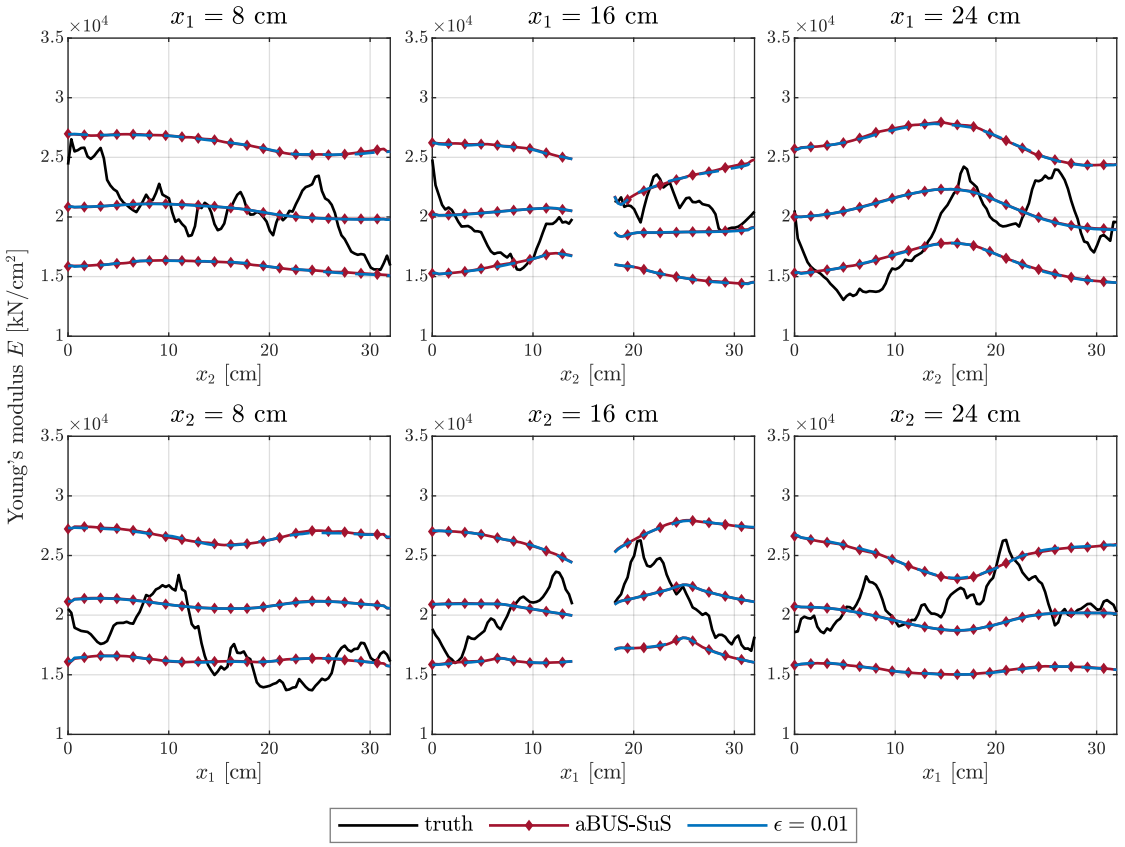

Figure 9. *Posterior Young's modulus random field obtained with aBUS-SuS and with CEBUred using $\epsilon = 10^{-2}$: Means (solid lines) and 95% credible intervals (CI, dashed lines) at three vertical (top row) and three horizontal (bottom row) cross-sections.*

Table 4

Likelihood and gradient evaluations for the plate example per run of aBUS-SuS and CEBUred (CEBUred: averaged over 40 runs).

	aBUS-SUS	CEBUred ($\epsilon = 1$)	CEBUred ($\epsilon = 10^{-1}$)	CEBUred ($\epsilon = 10^{-2}$)	CEBUred ($\epsilon = 10^{-3}$)
Llikelihood calls	120000	76	276	1313	10621
Gradient calls	-	105.3	254.5	642.0	1975.1

For $\epsilon \geq 10^{-1}$, we obtain LIS dimensions of $r \leq 5$ in all steps.

Fig. 10 (right) depicts the number of gradient and model evaluations per CEBUred step for varying ϵ exposing that for $\epsilon = 1$, the number of gradient evaluations performed to estimate the LIS exceeds the overall number model evaluations required to perform the parameter update of the biasing density. This is due to the fact that the number of required model evaluations quadratically depends on the LIS dimension r through (3.19), where in turn r decreases

Table 5

Relative posterior mean and variance error at varying ϵ averaged over 40 runs with coefficient of variation in round brackets

error	$\epsilon = 1$	$\epsilon = 10^{-1}$	$\epsilon = 10^{-2}$	$\epsilon = 10^{-3}$
ε_{μ_E} (CoV)	0.017(1.66)	0.019(0.98)	0.003(0.32)	0.002(0.17)
$\varepsilon_{\sigma_E^2}$ (CoV)	0.174(0.52)	0.201(0.75)	0.039(0.33)	0.027(0.10)

with increasing ϵ.

Fig. 11 displays the nESS of the final posterior sample along with the target nESS, $n_{\text{eff}}^* =$

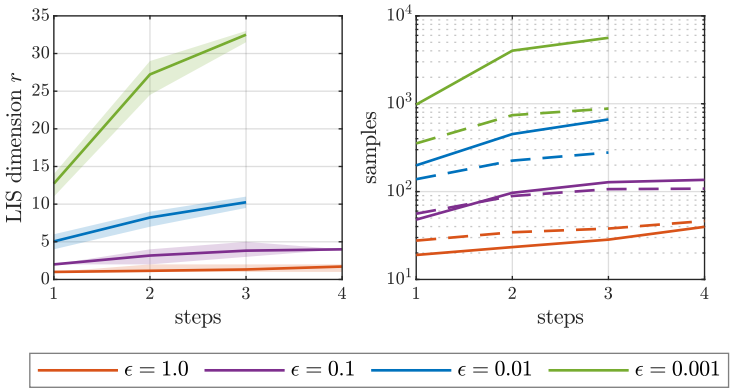

Figure 10. *Left: Average number of selected LIS dimension with increasing CEBUred step index (95% CI indicated as shaded area). Right: Average number of evaluated samples per step with increasing CEBUred step index. The **solid** and the **dashed** lines represent the total number of **model** and **model gradient** evaluations, respectively.*

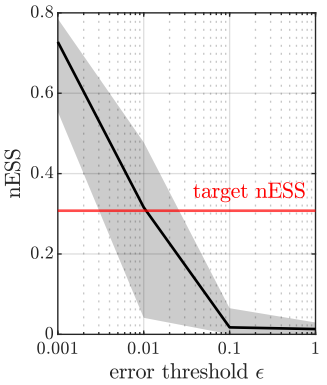

Figure 11. *Mean of the normalized effective sample size (nESS) (95% CI indicated as shaded area) plotted over the tested error thresholds along with the target nESS (solid red line).*

0.3077, that is related to the target coefficient of variation of the IS weights $\delta_w = 1.5$ through (2.8). The simulations with $\epsilon = 1.0$ and $\epsilon = 10^{-1}$ exhibit small nESS well below the target. In these cases, the dimensionality of the LIS is too low to accurately represent the posterior of the random field and, consequently, the IS weights have large variance. This leads to larger posterior mean and variance errors and larger associated coefficients of variation of these error measures for $\epsilon = 1.0$ and $\epsilon = 10^{-1}$ as documented in Table 5.

Compared to the beam application, where n_{eff}^* was achieved for all investigated ambient dimensions with $\epsilon = 1.0$, the plate obviously requires a stricter error threshold. According to Table 5 and Fig. 11, $\epsilon = 10^{-2}$ is a good choice for the present example. However, if the threshold is chosen very small, e.g., $\epsilon = 10^{-3}$, no significant improvement is observed. In this case, the marginally increased accuracy will not justify the additional computational expenses incurred by reducing the threshold.

Fig. 12 (top) depicts the spectrum of $\mathbf{H}$ at each step of CEBUred and each predefined

Table 6

Number of runs broken down according to required number of CEBUred steps at varying ϵ.

# of steps	$\epsilon = 1$	$\epsilon = 10^{-1}$	$\epsilon = 10^{-2}$	$\epsilon = 10^{-3}$
2	3	0	0	0
3	30	39	40	40
4	7	1	0	0

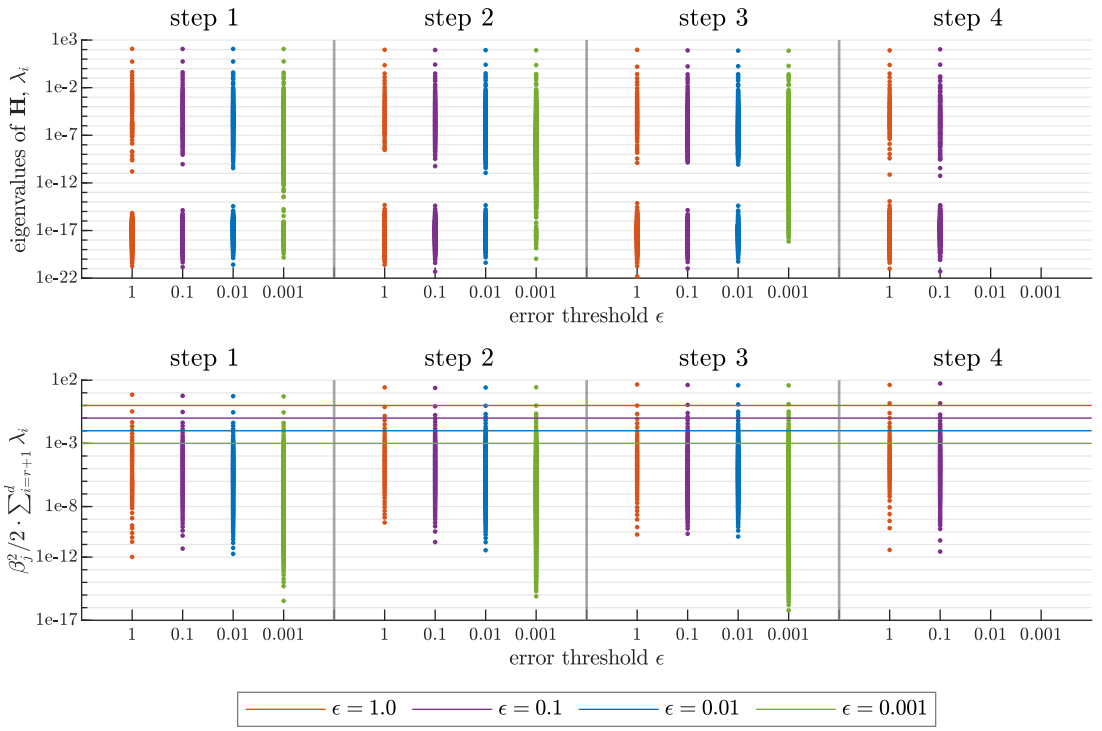

Figure 12. *Top: $\mathbf{H}$-spectra plotted versus CEBUred step number for each error threshold ϵ. Bottom: Upper bound of the KLD between the tempered, full posterior and the tempered, optimally reduced posterior as defined by (3.7). For each ϵ and CEBUred step, the largest scatter point corresponds to $r = 1$, the second-largest corresponds to $r = 2$ and so on to $r = d - 1$. Solid lines indicate the four tested values of the error threshold ϵ.*

error threshold. As shown in Table 6, all runs with $\epsilon \leq 10^{-2}$ required three steps, whereas only few runs with $\epsilon \geq 10^{-1}$ required four steps. For this reason, no spectra appear in the rightmost panels of Fig. 12 at step 4 for $\epsilon \leq 10^{-2}$. All displayed spectra share two dominant eigenvalues of comparable magnitude followed by a sharp decay. The main difference amongst spectra associated with different ϵ is the number of samples used to estimate $\mathbf{H}$, $n_{\mathbf{H}}$. While $n_{\mathbf{H}}$ has no influence on the dominant eigenvalues, it bears some effect on a gap in the center of the spectrum, that becomes narrower and eventually closes as $n_{\mathbf{H}}$ increases. This effect, however, is negligible for the LIS construction as it takes place at eigenvalue magnitudes well

below that of the smallest eigenvalue whose corresponding eigenvector is included in the LIS for any choice of ϵ.

In Fig. 12 (bottom), we show the criterion (3.7) for each possible choice of $r = 1 \dots d-1$ along with the tested error thresholds $\epsilon = \{1.0, 10^{-1}, 10^{-2}, 10^{-3}\}$. At any given step t, we can directly compare the scatter points belonging to different error thresholds since they have equal β_t on average. The number of scatter points that lie above the horizontal solid lines indicating ϵ determine the LIS dimension r. The upper scatter points are approximately equal for any choice of ϵ and any t as the corresponding summations are dominated by their leading term.

Using the heuristic given in (3.18) leads to a good estimate of the desired first eigenvalues and vectors. This is important because the accuracy of the LIS depends on these eigenvectors. However, for the computation of the upper bound of the KLD between the full posterior and the optimal reduced posterior (3.7) and the subsequent determination of the rank, all eigenvalues are needed. For problems with rapidly decaying eigenvalue spectra of their $\mathbf{H}$ matrix, as we see in this example, this is not a concern in practice, since their smallest eigenvalues have little effect on the computation of (3.7).

5. Concluding remarks.

We present CEBUred (**C**ross-**E**ntropy-based IS method for **B**ayesian **U**pdating in **red**uced space), an algorithm for approximating posterior distributions that are the solutions of nonlinear Bayesian inverse problems. Such problems often arise in the context of finding inverse solutions to computationally expensive numerical models and solvers. Thus, computational efficiency is of the essence, which translates to minimizing the number of required samples (evaluations of the numerical model) to approximate the sought posterior distribution at a prescribed accuracy. We address high-dimensional problem settings that arise, e.g., if the inference target is represented by random fields or processes, by identifying low-dimensional linear subspaces [60] in which we perform cross-entropy-based importance sampling [15]. These subspaces are obtained as truncated eigenspaces of the second-moment matrix of the gradient of the log-likelihood $\mathbf{H}$.

We investigate CEBUred using two benchmark problems from engineering mechanics. In the first example, the material parameter random field of a cantilever beam subject to a point load is inferred from noise-distorted deflection measurements. We examine the performance of CEBU versus CEBUred versus a known analytical posterior reference solution at varying dimension of the material parameter random field discretization. We find that the dimensionality reduction is vital to ensure the posterior approximation accuracy is independent of the problem dimension by comparing CEBU and CEBUred. The second problem consists of inferring the material parameter random field of a clamped steel plate under load from a strain measurement at 10 locations on the plate. The inference problem is set in a 879-dimensional space as the parameter random field is discretized with a 879-term KLE. We find that the quality of the posterior approximation produced by CEBUred is closely connected to the choice of the error threshold ϵ that controls the number of dimensions retained in reduced space. As ϵ decreases, CEBUred is able to recover the reference posterior solution accurately

both in mean and credible intervals. We further compare the performance of CEBUred to that of aBUS-SUS (adaptive Bayesian updating with subset simulation, [4]), which is a well-established method for nonlinear BIPs in high-dimensions.

The results of our numerical investigations show that CEBUred is a powerful method for solving high-dimensional nonlinear BIPs if the underlying computational model admits a low-dimensional representation, i.e., if the spectrum of $\mathbf{H}$ exhibits fast decay. From a computational perspective, CEBUred is particularly useful if the model allows for the cheap evaluation of gradients, e.g., if an adjoint solver is used and the the number of BIP inputs exceeds the number of available observations. In these cases, CEBUred achieves the same accuracy as aBUS-SuS at considerably lower computational expense.

Conversely, if an adjoint solver is not available or not efficient in the sense described above, the gain in computational efficiency provided by dimensionality reduction may be overcompensated by expensive gradient evaluations. In such case, the number of required gradient evaluations could be significantly reduced by using 'data-free likelihood-informed dimension reduction' as recently proposed in [10]. There, $\mathbf{H}$ is constructed in expectation over the data such that no knowledge about the posterior density is needed. Consequently, the upper bound between the exact and approximated posterior is controlled in expectation over the data. Following this method, the algorithm of CEBUred could be modified such that $\mathbf{H}$ is constructed at the beginning and therefore, only at this stage model gradient information would be required. Alternatively, one may turn to gradient-free supervised dimension reduction methods in order to identify sutaible subspaces to solve the Bayesian inverse problem.

Acknowledgments. We acknowledge support by the German Research Foundation (DFG) through Grants STR 1140/11-1 and PA 2901/1-1.

REFERENCES

[1] J. S. Arora and E. J. Haug, *Methods of design sensitivity analysis in structural optimization*, AIAA Journal, 17 (1979), pp. 970–974, https://doi.org/10.2514/3.61260.
[2] S.-K. Au and J. L. Beck, *Estimation of small failure probabilities in high dimensions by subset simulation*, Probabilistic engineering mechanics, 16 (2001), pp. 263–277.
[3] A. Beskos, G. Roberts, A. Stuart, and J. Voss, *MCMC methods for diffusion bridges*, Stochastics and Dynamics, 08 (2008), pp. 319–350, https://doi.org/10.1142/S0219493708002378.
[4] W. Betz, I. Papaioannou, J. L. Beck, and D. Straub, *Bayesian inference with subset simulation: Strategies and improvements*, Computer Methods in Applied Mechanics and Engineering, 331 (2018), pp. 72–93, https://doi.org/https://doi.org/10.1016/j.cma.2017.11.021.
[5] A. F. Bower, *Applied mechanics of solids*, CRC press, 2009.
[6] P. G. Constantine, *Active Subspaces*, Society for Industrial and Applied Mathematics, Philadelphia, PA, 2015, https://doi.org/10.1137/1.9781611973860, https://arxiv.org/abs/https://epubs.siam.org/doi/pdf/10.1137/1.9781611973860.
[7] P. G. Constantine, C. Kent, and T. Bui-Thanh, *Accelerating markov chain monte carlo with active subspaces*, SIAM Journal on Scientific Computing, 38 (2016), pp. A2779–A2805, https://doi.org/10.1137/15M1042127.
[8] S. L. Cotter, G. O. Roberts, A. M. Stuart, and D. White, *MCMC Methods for Functions: Modifying Old Algorithms to Make Them Faster*, Statistical Science, 28 (2013), pp. 424 – 446, https://doi.org/10.1214/13-STS421.

[9] T. Cui, J. Martin, Y. M. Marzouk, A. Solonen, and A. Spantini, *Likelihood-informed dimension reduction for nonlinear inverse problems*, Inverse Problems, 30 (2014), p. 114015, https://doi.org/10.1088/0266-5611/30/11/114015, https://doi.org/10.1088/0266-5611/30/11/114015.

[10] T. Cui and O. Zahm, *Data-free likelihood-informed dimension reduction of bayesian inverse problems*, Inverse Problems, 37 (2021), p. 045009, https://doi.org/10.1088/1361-6420/abeafb.

[11] P. Del Moral, A. Doucet, and A. Jasra, *Sequential monte carlo samplers*, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 68 (2006), pp. 411–436, https://doi.org/https://doi.org/10.1111/j.1467-9868.2006.00553.x.

[12] A. Der Kiureghian and J.-B. Ke, *The stochastic finite element method in structural reliability*, Probabilistic Engineering Mechanics, 3 (1988), pp. 83–91, https://doi.org/https://doi.org/10.1016/0266-8920(88)90019-7, https://www.sciencedirect.com/science/article/pii/0266892088900197.

[13] A. Doucet, A. Smith, N. de Freitas, and N. Gordon, *Sequential Monte Carlo Methods in Practice*, Information Science and Statistics, Springer New York, 2001.

[14] V. Elvira, L. Martino, D. Luengo, and M. F. Bugallo, *Novel weighting and resampling schemes in Population Monte Carlo*, in XXVIème Colloque GRETSI, Juan-les-Pins, France, Sept. 2017, https://hal.archives-ouvertes.fr/hal-01684862.

[15] M. Engel, O. Kanjilal, I. Papaioannou, and D. Straub, *Bayesian updating and marginal likelihood estimation by cross entropy based importance sampling.* Manuscript, 2021.

[16] H. P. Flath, L. C. Wilcox, V. Akelik, J. Hill, B. van Bloemen Waanders, and O. Ghattas, *Fast algorithms for bayesian uncertainty quantification in large-scale linear inverse problems based on low-rank partial hessian approximations*, SIAM Journal on Scientific Computing, 33 (2011), pp. 407–432, https://doi.org/10.1137/090780717.

[17] A. Gelman, J. Carlin, H. Stern, D. Dunson, A. Vehtari, and D. Rubin, *Bayesian Data Analysis, Third Edition*, Chapman & Hall/CRC Texts in Statistical Science, Taylor & Francis, 2013.

[18] A. Gelman and Y. Yao, *Holes in bayesian statistics*, Journal of Physics G: Nuclear and Particle Physics, 48 (2020), p. 014002, https://doi.org/10.1088/1361-6471/abc3a5, https://doi.org/10.1088/1361-6471/abc3a5.

[19] J. Geweke, *Bayesian inference in econometric models using monte carlo integration*, Econometrica, 57 (1989), pp. 1317–39.

[20] S. Geyer, I. Papaioannou, and D. Straub, *Cross entropy-based importance sampling using gaussian densities revisited*, Structural Safety, 76 (2019), pp. 15–27.

[21] R. G. Ghanem and P. D. Spanos, *Spectral stochastic finite-element formulation for reliability analysis*, Journal of Engineering Mechanics, 117 (1991), pp. 2351–2372.

[22] W. Gilks, S. Richardson, and D. Spiegelhalter, *Markov Chain Monte Carlo in Practice*, Chapman & Hall/CRC Interdisciplinary Statistics, Taylor & Francis, 1 ed., 1995.

[23] A. Gittens and J. A. Tropp, *Tail bounds for all eigenvalues of a sum of random matrices*, 2011, https://arxiv.org/abs/1104.4513.

[24] A. Jasra, D. A. Stephens, A. Doucet, and T. Tsagarakis, *Inference for Lévy-driven stochastic volatility models via adaptive sequential monte carlo*, Scandinavian Journal of Statistics, 38 (2011), pp. 1–22.

[25] E. Jaynes, *On the rationale of maximum-entropy methods*, Proceedings of the IEEE, 70 (1982), pp. 939–952, https://doi.org/10.1109/PROC.1982.12425.

[26] H. Jeffreys, *Theory of Probability*, Oxford, Oxford, England, third ed., 1961.

[27] S.-H. Jiang, I. Papaioannou, and D. Straub, *Bayesian updating of slope reliability in spatially variable soils with in-situ measurements*, Engineering Geology, 239 (2018), pp. 310 – 320.

[28] C. Johnson, *Numerical Solution of Partial Differential Equations by the Finite Element Method*, Dover Books on Mathematics Series, Dover Publications, Incorporated, 2012, https://books.google.de/books?id=PYXjyoqy5qMC.

[29] J. Kaipio and E. Somersalo, *Statistical and Computational Inverse Problems*, Springer, 2005, https://doi.org/10.1007/b138659.

[30] P. Koutsourelakis, *A multi-resolution, non-parametric, bayesian framework for identification of spatially-varying model parameters*, Journal of Computational Physics, 228 (2009), pp. 6184–6211, https://doi.org/https://doi.org/10.1016/j.jcp.2009.05.016, https://www.sciencedirect.com/science/article/pii/S0021999109002708.

[31] N. Kurtz and J. Song, *Cross-entropy-based adaptive importance sampling using gaussian mixture*, Structural Safety, 42 (2013), pp. 35 – 44.
[32] J. Latz, I. Papaioannou, and E. Ullmann, *Multilevel sequential2 monte carlo for bayesian inverse problems*, Journal of Computational Physics, 368 (2018), pp. 154–178, https://doi.org/https://doi.org/10.1016/j.jcp.2018.04.014, https://www.sciencedirect.com/science/article/pii/S0021999118302286.
[33] P. Liu and K. Liu, *Selection of random field mesh in finite element reliability analysis*, Journal of Engineering Mechanics, 119 (1993), pp. 667–680, https://doi.org/10.1061/(ASCE)0733-9399(1993)119:4(667).
[34] P.-L. Liu and A. Der Kiureghian, *Multivariate distribution models with prescribed marginals and covariances*, Probabilistic Engineering Mechanics, 1 (1986), pp. 105–112, https://doi.org/https://doi.org/10.1016/0266-8920(86)90033-0, https://www.sciencedirect.com/science/article/pii/0266892086900330.
[35] Y. M. Marzouk and H. N. Najm, *Dimensionality reduction and polynomial chaos acceleration of bayesian inference in inverse problems*, Journal of Computational Physics, 228 (2009), pp. 1862–1902, https://doi.org/https://doi.org/10.1016/j.jcp.2008.11.024.
[36] J. C. Mattingly, N. S. Pillai, and A. M. Stuart, *Diffusion limits of the random walk Metropolis algorithm in high dimensions*, The Annals of Applied Probability, 22 (2012), pp. 881 – 930, https://doi.org/10.1214/10-AAP754.
[37] R. M. Neal, *Annealed importance sampling*, Statistics and Computing, 11 (2001), p. 125139, https://doi.org/10.1023/A:1008923215028, https://doi.org/10.1023/A:1008923215028.
[38] R. M. Neal, *MCMC using Hamiltonian dynamics*, in Handbook of Markov Chain Monte Carlo, Chapman; Hall/CRC, 2010, pp. 113–162.
[39] R. B. Nelsen, *An Introduction to Copulas*, Springer Publishing Company, Incorporated, 2010.
[40] A. B. Owen, *Monte Carlo theory, methods and examples*, 2013.
[41] A. OHagan, *Expert knowledge elicitation: Subjective but scientific*, The American Statistician, 73 (2019), pp. 69–81, https://doi.org/10.1080/00031305.2018.1518265, https://doi.org/10.1080/00031305.2018.1518265, https://arxiv.org/abs/https://doi.org/10.1080/00031305.2018.1518265.
[42] I. Papaioannou, W. Betz, K. Zwirglmaier, and D. Straub, *Mcmc algorithms for subset simulation*, Probabilistic Engineering Mechanics, 41 (2015), pp. 89–103.
[43] I. Papaioannou, S. Geyer, and D. Straub, *Improved cross entropy-based importance sampling with a flexible mixture model*, Reliability Engineering & System Safety, 191 (2019), p. 106564.
[44] W. H. Press, W. T. Vetterling, S. A. Teukolsky, and B. P. Flannery, *Numerical recipes*, vol. 818, Cambridge university press Cambridge, 1986.
[45] C. Robert, *The Bayesian choice: from decision-theoretic foundations to computational implementation*, Springer Science & Business Media, 2007.
[46] G. O. Roberts and J. S. Rosenthal, *Optimal scaling for various Metropolis-Hastings algorithms*, Statistical Science, 16 (2001), pp. 351 – 367, https://doi.org/10.1214/ss/1015346320.
[47] M. Rosenblatt, *Remarks on a multivariate transformation*, The annals of mathematical statistics, 23 (1952), pp. 470–472.
[48] R. Y. Rubinstein, *Optimization of computer simulation models with rare events*, European Journal of Operational Research, 99 (1997), pp. 89–112.
[49] R. Y. Rubinstein and P. W. Glynn, *How to deal with the curse of dimensionality of likelihood ratios in monte carlo simulation*, Stochastic Models, 25 (2009), pp. 547–568, https://doi.org/10.1080/15326340903291248.
[50] R. Y. Rubinstein and D. P. Kroese, *Simulation and the Monte Carlo method*, Springer, 3 ed., 2017.
[51] A. Spantini, A. Solonen, T. Cui, J. Martin, L. Tenorio, and Y. Marzouk, *Optimal low-rank approximations of bayesian linear inverse problems*, SIAM Journal on Scientific Computing, 37 (2015), pp. A2451–A2487, https://doi.org/10.1137/140977308.
[52] D. Straub and I. Papaioannou, *Bayesian updating with structural reliability methods*, Journal of Engineering Mechanics, 141 (2015), https://doi.org/10.1061/(ASCE)EM.1943-7889.0000839, http://ascelibrary.org/doi/abs/10.1061/%28ASCE%29EM.1943-7889.0000839.
[53] D. Straub, I. Papaioannou, and W. Betz, *Bayesian analysis of rare events*, Journal of Computational Physics, 314 (2016), pp. 538–556.
[54] A. M. Stuart, *Inverse problems: A bayesian perspective*, Acta Numerica, 19 (2010), p. 451559, https:

//doi.org/10.1017/S0962492910000061.

[55] E. Torre, S. Marelli, P. Embrechts, and B. Sudret, *A general framework for data-driven uncertainty quantification under complex input dependencies using vine copulas*, Probabilistic Engineering Mechanics, 55 (2019), pp. 1–16, https://doi.org/https://doi.org/10.1016/j.probengmech.2018.08.001.

[56] J. A. Tropp, *User-friendly tail bounds for sums of random matrices*, Foundations of Computational Mathematics, 12 (2012), pp. 389–434, https://doi.org/10.1007/s10208-011-9099-z.

[57] F. Uribe, I. Papaioannou, W. Betz, and D. Straub, *Bayesian inference of random fields represented with the Karhunen-Loève expansion*, Computer Methods in Applied Mechanics and Engineering, 358 (2020), p. 112632, https://doi.org/https://doi.org/10.1016/j.cma.2019.112632.

[58] F. Uribe, I. Papaioannou, Y. M. Marzouk, and D. Straub, *Cross-entropy-based importance sampling with failure-informed dimension reduction for rare event simulation*, SIAM/ASA Journal on Uncertainty Quantification, 9 (2021), pp. 818–847, https://doi.org/10.1137/20M1344585.

[59] Z. Wang and J. Song, *Cross-entropy-based adaptive importance sampling using von mises-fisher mixture for high dimensional reliability analysis*, Structural Safety, 59 (2016), pp. 42–52.

[60] O. Zahm, T. Cui, K. Law, A. Spantini, and Y. Marzouk, *Certified dimension reduction in nonlinear bayesian inverse problems*, 2021, https://arxiv.org/abs/1807.03712.