Context-aware surrogate modeling for balancing approximation and sampling costs in multi-fidelity importance sampling and Bayesian inverse problems

Terrence Alsup* Benjamin Peherstorfer*

Abstract

Multi-fidelity methods leverage low-cost surrogate models to speed up computations and make occasional recourse to expensive high-fidelity models to establish accuracy guarantees. Because surrogate and high-fidelity models are used together, poor predictions by surrogate models can be compensated with frequent recourse to high-fidelity models. Thus, there is a trade-off between investing computational resources to improve the accuracy of surrogate models versus simply making more frequent recourse to expensive high-fidelity models; however, this trade-off is ignored by traditional modeling methods that construct surrogate models that are meant to replace high-fidelity models rather than being used together with high-fidelity models. This work considers multi-fidelity importance sampling and theoretically and computationally trades off increasing the fidelity of surrogate models for constructing more accurate biasing densities and the numbers of samples that are required from the high-fidelity models to compensate poor biasing densities. Numerical examples demonstrate that such context-aware surrogate models for multi-fidelity importance sampling have lower fidelity than what typically is set as tolerance in traditional model reduction, leading to runtime speedups of up to one order of magnitude in the presented examples.

1 Introduction

Surrogate models provide low-cost approximations of computationally expensive high-fidelity models and so are widely used to make tractable a variety of outer-loop applications such as control, optimization, and uncertainty quantification [34]. Typical examples of surrogate models are simplified-physics models [30, 28, 8], data-fit and machine-learning models [17, 36], and projection-based reduced models [4, 35, 5, 20, 11]. Multi-fidelity methods combine surrogate models for speedups and high-fidelity models for accuracy guarantees [34, 29]. Recourse to the high-fidelity model enables compensation for poor surrogate accuracy, in stark contrast to traditional single-fidelity methods that use surrogate models alone. The opportunity of multi-fidelity methods, which we exploit in the following, is that it is unnecessary that surrogate models achieve tight accuracy guarantees because high-fidelity models are occasionally evaluated to correct results. Rather, it can be beneficial to use surrogate models with very low accuracy in favor of very cheap training and evaluation costs. Clearly, there is a limit of how low the accuracy of surrogate models can be in favor of costs before surrogate models become useless. Thus, in multi-fidelity approaches, there is a trade-off between increasing the accuracy of surrogate models with expensive training methods versus making more frequent recourse to the expensive high-fidelity model to compensate less accurate, but cheaper, surrogate models. Surrogate models that exploit this trade-off are called context-aware models [31]. This work derives context-aware surrogate models for multi-fidelity importance sampling (MFIS) estimators [32], where the surrogate model is used for constructing a Laplace approximation as a biasing density. Our numerical results show that such context-aware surrogate models for MFIS can achieve an error reduction of more than one order of magnitude compared to using a single model alone.

We review related literature. First, there is work on adaptive discretizations for multi-level Monte Carlo methods and stochastic collocation methods [21, 22, 23] that adaptively refine meshes and time steps to obtain a non-uniform hierarchy of surrogate models. Additionally, there is work on continuous multi-level

*Courant Institute of Mathematical Sciences, New York University (alsup@cims.nyu.edu, pehersto@cims.nyu.edu)

Monte Carlo [15] that adapts the model hierarchy in a non-uniform fashion. In contrast to coarse-grid discretizations, we will consider surrogate models for constructing biasing densities, which incur training (offline) costs that we trade off with surrogate-model fidelity and frequency of recourse to the high-fidelity model. The work [12] learns data-fit surrogate models for solving Bayesian inverse problems, without building on multi-fidelity methods and thus without deriving the trade-off between model accuracy and costs. Second, the works [31, 16] explore the trade-off between surrogate-model fidelity and number of times to make recourse to the high-fidelity for multi-fidelity Monte Carlo estimation with control variates, which is in contrast to using importance sampling for variance reduction as in this work. In [13], the authors consider local, data-fit approximations and balance the decay rate of the bias due to the approximation with the variance of sampling with Markov chain Monte Carlo methods. Third, there is a large body of work on using surrogate models and multi-fidelity methods that build on importance sampling without explicitly exploiting the trade-off given by surrogate-model fidelity and frequency of recourse to the high-fidelity model. The work [27, 26] develops a principled strategy to switch between sampling from a surrogate model and from the high-fidelity model to speedup failure and rare event probability estimation. In [10], the authors build on *a posteriori* error estimators to decide if either a surrogate model or the high-fidelity model is evaluated. The authors of [19, 18] develop a multi-fidelity method for importance sampling to efficiently estimate risk-measures such as the conditional value-at-risk. Another line of work considers multi-level sequential Monte Carlo methods such as [6, 24] for reducing the costs of finding biasing densities.

We build on MFIS introduced in [32]. In particular, we develop bounds of the error of MFIS estimator that depends on the surrogate-model fidelity and then derive a trade-off between surrogate-model fidelity and computational costs. The first key ingredient is that we use a Laplace approximation computed with the surrogate model as biasing density. The quality of Laplace approximations has been studied in [14] in terms of the Kullback-Leibler (KL) divergence and in [39] in terms of the Hellinger distance when the noise level approaches zero. Instead, we consider the χ^2 divergence [42] due to its natural interpretation as the variance of the importance weights. There is a large body of work on adaptive importance sampling that studies minimizing the χ^2 divergence to derive an optimal biasing density [3, 37, 2], but these works do not consider the cost of surrogate models during training. The second key ingredient is bounding the error of the importance sampling estimator such as introduced in [9, 1, 38]. These error bounds take the form of a probability divergence between the target distribution and the biasing distribution, which we will use to separate the error due to sampling from the error due to the quality of the biasing density that corresponds to the surrogate-model fidelity.

This manuscript is structured as follows. In Section 2 we outline importance sampling in the multi-fidelity setting along with the bound on the mean-squared error (MSE) in terms of the χ^2 divergence as presented in [1]. Section 3 is the main contribution of this work and derives a bound on the χ^2 divergence of the target from the biasing distribution in terms of the surrogate-model fidelity that leads to the formulation of an optimization problem for finding a trade-off. In Section 4, we apply the results from Section 3 in the case where the target distribution is a posterior distribution arising from a Bayesian inverse problem. In Section 5, we demonstrate our method on three numerical examples. The proposed MFIS estimators with context-aware surrogate models achieve more than one order of magnitude error reduction compared to traditional importance sampling that uses the high-fidelity model alone with the same costs.

2 Importance sampling and problem formulation

Section 2.1 describes the setup of our problem. Section 2.2 is a brief overview of importance sampling and Section 2.3 overviews how the quality of a biasing density influences importance sampling estimators in terms of the χ^2 divergence. Section 2.4 illustrates the multi-fidelity approach to importance sampling and Section 2.5 formulates the trade-off between fidelity and number of samples that we are interested in.

2.1 Notation and problem setting

Let $(\Theta, \mathcal{B}(\Theta), p)$ denote a probability space where $\Theta = \mathbb{R}^d$ is the domain for parameters $\boldsymbol{\theta}$, $\mathcal{B}(\Theta)$ is the Borel σ-algebra of Θ, and p is a probability distribution on Θ. Let p be absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^d$ and refer to both the measure and the density function as p. Furthermore, the density p may only be known up to a normalizing constant $p = \frac{1}{Z}\tilde{p}$, where $\tilde{p} \geq 0$ is the un-normalized

density and $Z = \int_{\boldsymbol{\theta}} \tilde{p}(\boldsymbol{\theta})\ \mathrm{d}\boldsymbol{\theta}$ is the normalizing constant. In the following, we consider situations where the density p and the un-normalized density $\tilde{p}$ are expensive to evaluate. The goal is to compute quantities of interest with respect to the target distribution p which take the form of expectations

$$\mathbb{E}_p\left[f\right] = \int_{\Theta} f(\boldsymbol{\theta}) p(\boldsymbol{\theta})\ \mathrm{d}\boldsymbol{\theta}, \tag{1}$$

where f is a bounded measurable test function, i.e., $\|f\|_{L^\infty} < \infty$ where $\|f\|_{L^\infty} = \operatorname{ess\,sup}_{\boldsymbol{\theta}\in\Theta} |f(\boldsymbol{\theta})|$ under the measure p.

2.2 Importance sampling

Let q be another probability distribution on the Borel space
$(\Theta, \mathcal{B}(\Theta))$ that is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^d$ and is such that p is absolutely continuous with respect to q. We let q refer to both the probability distribution and the density function with respect to the Lebesgue measure. If sampling directly from p is impossible and the normalizing constant Z is unknown, then self-normalized importance sampling can be used with q as the biasing distribution to estimate the expectation (1). Draw m independent and identically distributed samples $\{\boldsymbol{\theta}^{(i)}\}_{i=1}^m \overset{\text{i.i.d.}}{\sim} q$ from the biasing distribution q and re-weight them with the target distribution p to obtain the self-normalized importance sampling estimator

$$\hat{f}_m = \frac{\sum_{i=1}^m f(\boldsymbol{\theta}^{(i)}) w(\boldsymbol{\theta}^{(i)})}{\sum_{i=1}^m w(\boldsymbol{\theta}^{(i)})} \tag{2}$$

of $\mathbb{E}_p\left[f\right]$, where the importance weights $w(\boldsymbol{\theta}^{(i)})$ are given by evaluating the un-normalized likelihood ratio $w(\boldsymbol{\theta}) = \frac{\tilde{p}(\boldsymbol{\theta})}{q(\boldsymbol{\theta})}$ at the samples $\boldsymbol{\theta}^{(i)}$. If all $w(\boldsymbol{\theta}^{(i)}) = 0$, then we define $\hat{f}_m = 0$. The estimator (2) is a consistent estimator of $\mathbb{E}_p\left[f\right]$ as the sample size $m \to \infty$.

2.3 Error of the importance sampling estimator

Theorem 2.1 of [1] gives the following bound on the MSE of the self-normalized importance sampling estimator (2): if p is absolutely continuous with respect to q, then

$$\mathbb{E}\left[\left(\hat{f}_m - \mathbb{E}_p\left[f\right]\right)^2\right] \ \leq\ \frac{4\|f\|_{L^\infty}^2}{m}\left(\chi^2\left(p\ ||\ q\right) + 1\right) \tag{3}$$

holds, with the χ^2 divergence of p from q defined as

$$\chi^2\left(p\ ||\ q\right) + 1 = \mathrm{Var}_q\left[\frac{p}{q}\right] + 1 = \int_{\Theta}\left(\frac{p(\boldsymbol{\theta})}{q(\boldsymbol{\theta})}\right)^2 q(\boldsymbol{\theta})\ \mathrm{d}\boldsymbol{\theta} = \int_{\Theta}\frac{p(\boldsymbol{\theta})}{q(\boldsymbol{\theta})}\ p(\boldsymbol{\theta})\ \mathrm{d}\boldsymbol{\theta}. \tag{4}$$

Note that the inequality (3) holds if $\mathbb{E}_q\left[w^2\right] = \infty$. Since f is bounded, it holds $(\hat{f}_m - \mathbb{E}_p\left[f\right])^2 \leq 4\|f\|_{L^\infty}^2$, which means that the bound (3) is only useful if $m \geq \chi^2\left(p\ ||\ q\right) + 1$. The bound (3) motivates setting the effective sample size to

$$m_{\text{eff}} = \frac{m}{\chi^2\left(p\ ||\ q\right) + 1}, \tag{5}$$

so that a large χ^2 divergence corresponds to a large variance of the weights, meaning more samples are needed to reduce the MSE of the estimator (2). The effective sample size (5) motivates finding a biasing density q that is close to p with respect to the χ^2 divergence. The χ^2 divergence is related to other probability divergences such as the Kullback-Leibler (KL) divergence

$$\mathrm{KL}\left(p\ ||\ q\right) = \int_{\Theta}\log\left(\frac{p(\boldsymbol{\theta})}{q(\boldsymbol{\theta})}\right) p(\boldsymbol{\theta})\ \mathrm{d}\boldsymbol{\theta}$$

and the Hellinger distance

$$d_H(p,\ q) = \left(\frac{1}{2}\int_\Theta \left(\sqrt{p(\boldsymbol{\theta})} - \sqrt{q(\boldsymbol{\theta})}\right)^2 \ \mathrm{d}\boldsymbol{\theta}\right)^{1/2}.$$

The relation is a lower bound given by Jensen's inequality

$$\mathrm{e}^{2d_H(p,\ q)^2} \leq \mathrm{e}^{\mathrm{KL}(p\ ||\ q)} \leq \chi^2\left(p\ ||\ q\right) + 1\,,$$

see [42] for more general information regarding these probability divergences.

2.4 Finding a biasing density

Let $(p_h)_{h>0}$ be a sequence of probability measures on $(\Theta, \mathcal{B}(\Theta))$, where the distributions p_h are approximations to p and the index $h > 0$ denotes the fidelity of the approximation. For each h, let p_h be absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^d$ and use p_h to denote both the density function and the distribution. Let the density functions converge pointwise so $p_h(\boldsymbol{\theta}) \to p(\boldsymbol{\theta})$ as $h \to 0$ for every $\boldsymbol{\theta} \in \Theta$. Define $C > 0$ as the cost of evaluating the un-normalized high-fidelity density $\tilde{p}$ and $c(h) > 0$ as the cost of evaluating the un-normalized surrogate density $\tilde{p}_h$. The un-normalized surrogate densities $\tilde{p}_h$ can be used instead of $\tilde{p}$ to find a biasing density q_h resulting in the MFIS [32] estimator

$$\hat{f}_{h,m} = \frac{\sum_{i=1}^m f(\boldsymbol{\theta}^{(i)}) w_h(\boldsymbol{\theta}^{(i)})}{\sum_{i=1}^m w_h(\boldsymbol{\theta}^{(i)})} \quad \text{where} \quad \{\boldsymbol{\theta}^{(i)}\}_{i=1}^m \overset{\text{i.i.d.}}{\sim} q_h, \tag{6}$$

of $\mathbb{E}_p[f]$ with the importance weights $w_h(\boldsymbol{\theta}^{(i)}) = \tilde{p}(\boldsymbol{\theta}^{(i)})/\tilde{q}_h(\boldsymbol{\theta}^{(i)})$ given by the ratio of the un-normalized densities $\tilde{p}$ and $\tilde{q}_h$ at $\boldsymbol{\theta}^{(i)}$. Note that the un-normalized surrogate densities $\tilde{p}_h$ are not evaluated in computing the estimator (6) and are only evaluated when deriving the biasing density q_h. The bound (3) shows that the quality of the biasing density with respect to the MSE is determined by the variance of the weights $w_h(\boldsymbol{\theta}^{(i)})$ and thus that the number of samples needed to achieve an error tolerance depends directly on the fidelity h of the surrogate density.

2.5 Problem formulation

Multi-fidelity importance sampling gives rise to the following two-step process of estimating $\mathbb{E}_p[f]$ for test functions f: (i) finding the biasing density q_h from $\tilde{p}_h$ and (ii) evaluating the un-normalized densities $\tilde{q}_h$ and $\tilde{p}$ at m samples to obtain an estimate (6) of $\mathbb{E}_p[f]$. Notice that q_h is independent of the test function f and thus can be re-used for many different test functions. The first step incurs training costs to derive q_h using $\tilde{p}_h$, and the second step incurs online costs of evaluating the un-normalized surrogate and expensive high-fidelity densities. The two steps give rise to a trade-off: investing high training costs to find a good biasing density that keeps the χ^2 divergence low means that fewer evaluations of the expensive high-fidelity density are required in the online step and vice versa. Traditional model reduction [35, 5] typically targets computations where the surrogate model replaces the high-fidelity, where such a trade-off does not exist, instead of combining surrogate and high-fidelity models as in multi-fidelity methods such as MFIS. Thus, traditional model reduction provides little guidance on the mathematical formulation of this trade-off and the total costs.

3 Context-aware surrogate models for multi-fidelity importance sampling

We consider the following trade-off: given an error tolerance ϵ, what is the optimal fidelity h of the surrogate model that minimizes the total computational costs subject to the mean-squared error of the multi-fidelity importance sampling estimator (6) being below or equal to the tolerance ϵ. We refer to such surrogate models as *context-aware* because the fidelity is determined specifically for the online computations of the problem

(context) at hand [31], rather than being prescribed without taking the specific context of multi-fidelity computations into account as in traditional model reduction [35, 5].

Section 3.1 revisits the notion of a sub-Gaussian distribution which is used in Section 3.2 to derive an upper bound for $\chi^2(p \mid\mid q_h)$ that depends on the fidelity h. Section 3.3 introduces a Laplace approximation q_h of the low-fidelity surrogate density p_h to be used as the biasing density and discusses its properties. Section 3.4 uses the bound (9) on the χ^2 divergence to formulate an optimization problem that selects a fidelity h^* based on the online stage of MFIS and derives the overall cost complexity of the corresponding estimator. Section 3.5 summarizes the entire computational procedure in algorithmic form.

3.1 Sub-Gaussian distributions

For importance sampling without a fixed test function f, it is imperative that the importance weights have finite variance (i.e., finite χ^2 divergence) which means that the tails of the biasing density cannot be significantly lighter than the tails of the target density p. Sub-Gaussian distributions are characterized by their fast tail decay. A useful norm for quantifying the tail decay of a real-valued random variable, X, is the Orlicz norm defined as

$$\|X\|_{\psi_2} = \inf\left\{t > 0 \mid \mathbb{E}\left[\exp(X^2/t^2)\right] \leq 2\right\},$$

see [43, Sec. 2.5, Sec. 3.4] for other equivalent definitions. For a real random vector $\boldsymbol{x} = (x_1, \ldots, x_d)$, the Orlicz norm is defined to be

$$\|\boldsymbol{x}\|_{\psi_2} = \sup_{\boldsymbol{v} \in S^{d-1}} \left\|\boldsymbol{v}^T \boldsymbol{x}\right\|_{\psi_2},$$

where $S^{d-1} \subset \mathbb{R}^d$ is the unit sphere defined as $S^{d-1} = \{\boldsymbol{v} \in \mathbb{R}^d : \|\boldsymbol{v}\|_2 = 1\}$. A probability distribution π is said to be sub-Gaussian if any random variable $\boldsymbol{x} \sim \pi$ has $\|\boldsymbol{x}\|_{\psi_2} < \infty$. Two examples of sub-Gaussian distributions are multivariate Gaussians and distributions with compact support. If $\boldsymbol{x} \sim N(0, \sigma^2 \boldsymbol{I})$ then $\|\boldsymbol{x}\|_{\psi_2} \leq \sqrt{2}\sigma$. In the following Lemma 1 we give a characterization of sub-Gaussian distributions that will be used in the following sections. The lemma is a multi-dimensional version of Proposition 2.5.2 (iv) in [43]. We did not find this specific result in the literature and so we provide a proof in Appendix A, even though it is a technical auxiliary result for us only.

Lemma 1. *A random vector $\boldsymbol{x}$ with density π is sub-Gaussian if and only if there exists a symmetric positive-definite matrix $\boldsymbol{A}$ such that for all vectors $\boldsymbol{\mu} \in \mathbb{R}^d$*

$$\mathbb{E}_\pi\left[\exp\left((\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{A} (\boldsymbol{x} - \boldsymbol{\mu})\right)\right] < \infty.$$

Remark 1. *In the case where π is a Gaussian with covariance $\boldsymbol{\Sigma}$, the matrix $\boldsymbol{A}$ must be such that $\frac{1}{2}\boldsymbol{\Sigma}^{-1} - \boldsymbol{A}$ is symmetric positive definite, in which case Lemma 1 is closely related to Fernique's theorem about the tail decay of Gaussian densities. This constraint on $\boldsymbol{A}$ will translate to a constraint on the biasing density for non-Gaussian target densities as will be made precise in the next section.*

3.2 Bounding the χ^2 divergence

In this section we derive the dependence of the MSE of the estimator (6) with respect to $\mathbb{E}_p[f]$ on the fidelity h used to find the biasing density q_h. We bound $\chi^2(p \mid\mid q_h)$ with respect to h and we want this bound to factor into a part depending only on the ratio p/p_h and a part depending only on the ratio p_h/q_h. The following example demonstrates that such a decomposition is not straightforward: let

$$p(x) = a\mathrm{e}^{-ax}, \quad p_h(x) = b\mathrm{e}^{-bx}, \quad q_h(x) = c\mathrm{e}^{-cx} \qquad x \geq 0,$$

for $a, b, c > 0$, be three exponential distributions. Then

$$\chi^2(p \mid\mid p_h) = \int_0^\infty \frac{a^2}{b} \mathrm{e}^{-(2a-b)x} \mathrm{d}x = \frac{a^2}{b(2a-b)}$$

if $a > b/2$ and ∞ otherwise. By taking $a = 2$, $b = 3/2$ and $c = 1$, we have that

$$\chi^2\left(p \mid\mid p_h\right) < \infty, \quad \chi^2\left(p_h \mid\mid q_h\right) < \infty,$$

but that

$$\chi^2\left(p \mid\mid q_h\right) = \infty\,,$$

which means that we cannot directly decompose the χ^2 divergence into the product of χ^2 divergences with an intermediate distribution (namely p_h). In contrast, the Cauchy-Schwarz inequality gives

$$\chi^2\left(p \mid\mid q_h\right) + 1 = \left\| \frac{p}{q_h} \right\|_{L^1(p)} = \left\langle \frac{p}{p_h}, \frac{p_h}{q_h} \right\rangle_{L^2(p)} \leq \left\| \frac{p}{p_h} \right\|_{L^2(p)} \left\| \frac{p_h}{q_h} \right\|_{L^2(p)}, \tag{7}$$

which requires the likelihood ratios p/p_h and p_h/q_h to be in $L^2(p)$ as opposed to $L^1(p)$, which is required for the bound (3) to hold and be finite. We note that, while we restrict ourselves to bounded test functions f in this work, other bounds similar to (3) exist [1, Theorem 2.3] but it remains unclear how to write them as a probability divergence between the target and biasing distributions and they do not necessarily admit a clear decomposition between accuracy of the surrogate density and accuracy of the approximation to the surrogate density as in (7). We also note that we can allow more general test functions $f \in L^2$ as long as we place stronger assumptions on the ratio of densities so that $\|p/p_h\|_{L^\infty(p)}, \|p_h/q_h\|_{L^\infty(p)} < \infty$ as in [39]. Here we choose to loosen these assumptions at the cost of only considering bounded test functions. The next four assumptions and the theorem that follows are sufficient for the likelihood ratios p/p_h and p_h/q_h to be in $L^2(p)$ and to decompose the χ^2 divergence as in the right-hand side of Equation (7).

Assumption 1 (Exponential form of the densities)**.** *The densities p, p_h, and q_h have the form*

$$p(\boldsymbol{\theta}) = \frac{1}{Z}\mathrm{e}^{-\Phi(\boldsymbol{\theta})}, \quad p_h(\boldsymbol{\theta}) = \frac{1}{Z_h}\mathrm{e}^{-\Phi_h(\boldsymbol{\theta})}, \quad q_h(\boldsymbol{\theta}) = \frac{1}{\tilde{Z}_h}\mathrm{e}^{-\tilde{\Phi}_h(\boldsymbol{\theta})},$$

with potentials $\Phi, \Phi_h, \tilde{\Phi}_h \in \mathcal{C}^2(\Theta)$ that are twice continuously differentiable, normalizing constants $Z, Z_h, \tilde{Z}_h$, and $\Phi_h(\boldsymbol{\theta}) \to \Phi(\boldsymbol{\theta})$ for all $\boldsymbol{\theta} \in \Theta$ as $h \to 0$.

Assumption 2 (Decay of the target density)**.** *The target density p is sub-Gaussian with matrix $\boldsymbol{A}$; see Lemma 1.*

Assumption 3 (Error of the surrogate potentials)**.** *There exists an error function $\delta(h) > 0$ and a function $\tau(\boldsymbol{\theta}) \geq 0$, such that*

$$\Phi_h(\boldsymbol{\theta}) \leq \Phi(\boldsymbol{\theta}) + \delta(h)\tau(\boldsymbol{\theta})$$

for all $\boldsymbol{\theta} \in \Theta$, where $\delta(h) \to 0$ as $h \to 0$.

Assumption 4 (Biasing densities)**.** *There exists a function $\gamma(h) > 0$ and a function $\omega(\boldsymbol{\theta}) \geq 0$ such that for all h*

$$\tilde{\Phi}_h(\boldsymbol{\theta}) \leq \Phi_h(\boldsymbol{\theta}) + \gamma(h)\omega(\boldsymbol{\theta})$$

for all $\boldsymbol{\theta} \in \Theta$.

Assumption 2 is independent of the surrogate densities and is necessary to avoid heavy tailed target distributions for which importance sampling can fail. Note that we make the sub-Gaussian assumption specifically because we use a Laplace approximation as the biasing density, which is Gaussian, in Section 3.3. Assumptions 3 and 4 are each controlling one of the terms on the right-hand side of Equation (7): Assumption 3 ensures that the surrogate densities are sufficiently accurate with respect to the target density while Assumption 4 ensures that the choice of approximation to the surrogate density is sufficiently close. In both cases we only assume the asymmetric inequality of the form $\Phi_h(\boldsymbol{\theta}) \leq \Phi(\boldsymbol{\theta}) + \delta(h)\tau(\boldsymbol{\theta})$ as opposed to $|\Phi_h(\boldsymbol{\theta}) - \Phi(\boldsymbol{\theta})| \leq \delta(h)\tau(\boldsymbol{\theta})$ (and similarly for Assumption 4) because importance sampling can fail if the tails of the biasing density are lighter than the tails of the target density, but will still converge even if they are heavier. Note that the restrictions on the tails of the biasing and target distribution are inherited by importance sampling rather than being a restrictions imposed by the proposed approach of trading off surrogate fidelity and costs.

Remark 2. *Assumption* 4 *does* not *assume that* $\gamma(h) \to 0$ *as* $h \to 0$. *Starting with Section* 3.3, *we will choose the density* q_h *to be a Laplace approximation of* p_h, *which does not necessarily converge to* p_h *as* $h \to 0$.

Theorem 1 gives the decomposition and bound depending on the fidelity h.

Theorem 1. *Let Assumptions* 1, 2, 3, *and* 4 *hold and assume there exist constants* $\tau_0, \omega_0 > 0$ *such that*

$$\tau(\boldsymbol{\theta}) \leq \|\boldsymbol{\theta}\|^2 + \tau_0, \quad \omega(\boldsymbol{\theta}) \leq \|\boldsymbol{\theta}\|^2 + \omega_0 \,.$$

Let $h_{\max}$ *be such that for all* $h \leq h_{\max}$

$$\gamma(h) \leq \frac{1}{4}\lambda_{\min}^{\boldsymbol{A}} \,, \tag{8}$$

with $\boldsymbol{A}$ *being the matrix from Assumption* 2 *and* $\lambda_{\min}^{\boldsymbol{A}}$ *being its smallest eigenvalue, then for all* h *sufficiently small we have that*

$$\chi^2(p \mid\mid q_h) + 1 \leq K_0 \mathrm{e}^{K_1\delta(h) + K_2\gamma(h)} \tag{9}$$

where K_0, K_1, K_2 *are all positive constants independent of* h.

By the assumption in Theorem 1 that $\gamma(h) \leq \lambda_{\min}^{\boldsymbol{A}}/4$, the bound (9) can be written in the form

$$\chi^2(p \mid\mid q_h) + 1 \leq \tilde{K}_0 \mathrm{e}^{K_1\delta(h)} \tag{10}$$

where the constant $\tilde{K}_0$ now absorbs the dependency on the approximation q_h

$$\tilde{K}_0 = K_0 \mathrm{e}^{K_2\lambda_{\min}^{\boldsymbol{A}}/4} \geq K_0 \mathrm{e}^{K_2\gamma(h)} \,. \tag{11}$$

In the limit as the fidelity $h \to 0$, the upper bound (10) remains bounded by the constant $\tilde{K}_0$, which is determined entirely by the choice of biasing densities q_h. Notice that $\tilde{K}_0 > 1$ because $K_0 \geq 1$ (see last line of proof of Theorem 1) and the argument $K_2\gamma(h) > 0$ of the exponential function is positive by Assumption 4.

Proof of Theorem 1. By Assumption 2, p is sub-Gaussian with matrix $\boldsymbol{A} \succ 0$ so that by Lemma 1

$$\frac{1}{Z}\int_\Theta \exp\left(\boldsymbol{\theta}^T\boldsymbol{A}\boldsymbol{\theta} - \Phi(\boldsymbol{\theta})\right) \mathrm{d}\boldsymbol{\theta} < \infty \,.$$

Recall that Z is the normalizing constant from Assumption 1.
Part 1: Bounding high-fidelity to surrogate ratio

The first term on the right-hand-side of Equation (7) can be bounded using Assumption 3:

$$\begin{aligned}
\left\|\frac{p}{p_h}\right\|_{L^2(p)}^2 &= \frac{1}{Z}\left(\frac{Z_h}{Z}\right)^2 \int_\Theta \exp\left\{2\left(\Phi_h(\boldsymbol{\theta}) - \Phi(\boldsymbol{\theta})\right) - \Phi(\boldsymbol{\theta})\right\} \mathrm{d}\boldsymbol{\theta} \\
&\leq \frac{1}{Z}\left(\frac{Z_h}{Z}\right)^2 \int_\Theta \exp\left\{2\delta(h)\left(\|\boldsymbol{\theta}\|^2 + \tau_0\right) - \Phi(\boldsymbol{\theta})\right\} \mathrm{d}\boldsymbol{\theta} \,.
\end{aligned}$$

Re-writing this last line gives

$$\left\|\frac{p}{p_h}\right\|_{L^2(p)}^2 \leq \frac{1}{Z}\left(\frac{Z_h}{Z}\right)^2 \exp\left(2\tau_0\delta(h)\right) \int_\Theta \exp\left\{2\delta(h)\|\boldsymbol{\theta}\|^2 - \Phi(\boldsymbol{\theta})\right\} \mathrm{d}\boldsymbol{\theta} \,. \tag{12}$$

Now the two dependencies of the right-hand side of (12) on the fidelity h are through the ratio Z_h/Z and through $\delta(h)$. For now we just bound the integral on the right-hand side of (12), which is finite since $\boldsymbol{A} \succ 2\delta(h)\boldsymbol{I}$ for all h sufficiently small. Adding and subtracting $\boldsymbol{\theta}^T\boldsymbol{A}\boldsymbol{\theta}$ in (12) gives

$$\begin{aligned}
\left\|\frac{p}{p_h}\right\|_{L^2(p)}^2 &\leq \frac{1}{Z}\left(\frac{Z_h}{Z}\right)^2 \exp\left(2\tau_0\delta(h)\right) \int_\Theta \exp\left\{2\delta(h)\|\boldsymbol{\theta}\|^2 - \Phi(\boldsymbol{\theta})\right\} \mathrm{d}\boldsymbol{\theta} \\
&= \frac{1}{Z}\left(\frac{Z_h}{Z}\right)^2 \exp\left(2\tau_0\delta(h)\right) \int_\Theta \exp\left\{-\boldsymbol{\theta}^T\left(\boldsymbol{A} - 2\delta(h)\boldsymbol{I}\right)\boldsymbol{\theta} + \boldsymbol{\theta}^T\boldsymbol{A}\boldsymbol{\theta} - \Phi(\boldsymbol{\theta})\right\} \mathrm{d}\boldsymbol{\theta} \,.
\end{aligned}$$

Putting this together with the fact that $\boldsymbol{A} - 2\delta(h)\boldsymbol{I} \succ 0$ gives

$$\left\| \frac{p}{p_h} \right\|_{L^2(p)}^2 \leq \frac{1}{Z} \left(\frac{Z_h}{Z} \right)^2 \exp\left(2\tau_0 \delta(h)\right) \int_\Theta \exp\left\{ \boldsymbol{\theta}^T \boldsymbol{A} \boldsymbol{\theta} - \Phi(\boldsymbol{\theta}) \right\} \mathrm{d}\boldsymbol{\theta} \tag{13}$$

to complete the bound of the first term on the right-hand side of Equation (7).
Part 2: Bounding surrogate to biasing density ratio

The second term on the right-hand side of Equation (7) is bounded in a similar fashion. By Assumption 4 we can bound

$$\begin{aligned}
\left\| \frac{p_h}{q_h} \right\|_{L^2(p)}^2 &= \frac{1}{Z} \left(\frac{\tilde{Z}_h}{Z_h} \right)^2 \int_\Theta \exp\left\{ 2\left(\tilde{\Phi}_h(\boldsymbol{\theta}) - \Phi_h(\boldsymbol{\theta}) \right) - \Phi(\boldsymbol{\theta}) \right\} \mathrm{d}\boldsymbol{\theta} \\
&\leq \frac{1}{Z} \left(\frac{\tilde{Z}_h}{Z_h} \right)^2 \int_\Theta \exp\left\{ 2\gamma(h) \left(\|\boldsymbol{\theta}\|^2 + \omega_0 \right) - \Phi(\boldsymbol{\theta}) \right\} \mathrm{d}\boldsymbol{\theta} \\
&= \frac{1}{Z} \left(\frac{\tilde{Z}_h}{Z_h} \right)^2 \exp\left(2\omega_0 \gamma(h)\right) \int_\Theta \exp\left\{ 2\gamma(h) \|\boldsymbol{\theta}\|^2 - \Phi(\boldsymbol{\theta}) \right\} \mathrm{d}\boldsymbol{\theta} \,.
\end{aligned}$$

Again we add and subtract $\boldsymbol{\theta}^T \boldsymbol{A} \boldsymbol{\theta}$ to obtain

$$\left\| \frac{p_h}{q_h} \right\|_{L^2(p)}^2 \leq \frac{1}{Z} \left(\frac{\tilde{Z}_h}{Z_h} \right)^2 \exp\left(2\omega_0 \gamma(h)\right) \int_\Theta \exp\left\{ -\boldsymbol{\theta}^T \left(\boldsymbol{A} - 2\gamma(h)\boldsymbol{I} \right) \boldsymbol{\theta} + \boldsymbol{\theta}^T \boldsymbol{A} \boldsymbol{\theta} - \Phi(\boldsymbol{\theta}) \right\} \mathrm{d}\boldsymbol{\theta} \,.$$

Using this with the fact that $\boldsymbol{A} - 2\gamma(h)\boldsymbol{I} \succeq 0$ for all $h \leq h_{\max}$ gives

$$\left\| \frac{p_h}{q_h} \right\|_{L^2(p)}^2 \leq \frac{1}{Z} \left(\frac{\tilde{Z}_h}{Z_h} \right)^2 \exp\left(2\omega_0 \gamma(h)\right) \int_\Theta \exp\left\{ \boldsymbol{\theta}^T \boldsymbol{A} \boldsymbol{\theta} - \Phi(\boldsymbol{\theta}) \right\} \mathrm{d}\boldsymbol{\theta} \,. \tag{14}$$

Multiplying the right-hand sides of the bounds (13) and (14) and then taking the square root gives together with (7) that

$$\left\| \frac{p}{q_h} \right\|_{L^1(p)} \leq \frac{1}{Z} \left(\frac{\tilde{Z}_h}{Z} \right) \exp\left\{ \delta(h)\tau_0 + \gamma(h)\omega_0 \right\} \int_\Theta \exp\left\{ \boldsymbol{\theta}^T \boldsymbol{A} \boldsymbol{\theta} - \Phi(\boldsymbol{\theta}) \right\} \mathrm{d}\boldsymbol{\theta} \tag{15}$$

holds. The integral is independent of h, so it remains to bound the ratio of normalizing constants.
Part 3: Bounding ratio of normalizing constants

In general, if p_h is not in the family of biasing densities then we may have $\tilde{Z}_h \neq Z_h$, and thus,

$$\frac{\tilde{Z}_h}{Z} \not\to 1$$

as $h \to 0$. Instead we just give a constant upper bound on $\tilde{Z}_h$ that is independent of the fidelity h. By Assumption 1, the normalizing constant $\tilde{Z}_h$ satisfies

$$\begin{aligned}
\tilde{Z}_h &= \int_\Theta \exp\left\{ -\tilde{\Phi}_h(\boldsymbol{\theta}) \right\} \mathrm{d}\boldsymbol{\theta} \\
&= \int_\Theta \exp\left\{ -\tilde{\Phi}_h(\boldsymbol{\theta}) + \Phi_h(\boldsymbol{\theta}) - \Phi_h(\boldsymbol{\theta}) + \Phi(\boldsymbol{\theta}) - \Phi(\boldsymbol{\theta}) \right\} \mathrm{d}\boldsymbol{\theta} \\
&= Z \int_\Theta \exp\left\{ -\tilde{\Phi}_h(\boldsymbol{\theta}) + \Phi_h(\boldsymbol{\theta}) - \Phi_h(\boldsymbol{\theta}) + \Phi(\boldsymbol{\theta}) \right\} \; p(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} \,.
\end{aligned}$$

Dividing by Z and using Assumptions 3 and 4 we have

$$\frac{\tilde{Z}_h}{Z} \leq \int_\Theta \exp\left\{ -\delta(h)(\|\boldsymbol{\theta}\|^2 + \tau_0) - \gamma(h)(\|\boldsymbol{\theta}\|^2 + \omega_0) \right\} p(\boldsymbol{\theta}) \; d\boldsymbol{\theta} \leq 1 \,, \tag{16}$$

because the term inside the exponential is less than or equal to 0 and p is a density. Finally, combining the bounds (13), (14), and (16) gives the result

$$\chi^2\left(p \mid\mid q_h\right) + 1 = \left\| \frac{p}{q_h} \right\|_{L^1(p)} \leq \exp\left\{\delta(h)\tau_0 + \gamma(h)\omega_0\right\} \mathbb{E}_p\left[\exp\left(\boldsymbol{\theta}^T \boldsymbol{A}\boldsymbol{\theta}\right)\right],$$

where the expectation is independent of h. Here

$$K_0 = \mathbb{E}_p\left[\exp\left(\boldsymbol{\theta}^T \boldsymbol{A}\boldsymbol{\theta}\right)\right], \quad K_1 = \tau_0, \quad K_2 = \omega_0$$

are all independent of the fidelity h. □

Remark 3. *The assumption that $\tau(\boldsymbol{\theta}) \leq \|\boldsymbol{\theta}\|^2 + \tau_0$ holds is similar to the pointwise Assumption 4.8 in Theorem 4.6 of [40]. In [40], the pointwise bound can grow faster with respect to $\boldsymbol{\theta}$ than in our case because there the Hellinger distance, which is upper-bounded by the χ^2 divergence, is considered. In our case, under Assumption 2, the potential Φ grows at least quadratically with respect to $\|\boldsymbol{\theta}\|$. For a similar reason we require that $\gamma(h) \leq \frac{1}{4}\lambda^A_{\min}$, even though $\gamma(h)$ may not converge to zero. This inherently places restrictions on what approximations may be used as biasing densities for importance sampling and is analogous to assumptions made on biasing densities in importance sampling in general such as [1, Theorem 2.1].*

3.3 Laplace approximation

In the following, we use a Laplace approximation of a surrogate density p_h as a specific choice of biasing density q_h. A Laplace approximation q_h is a Gaussian approximation to the density p_h whose mean is a mode of p_h

$$\boldsymbol{\mu}_h^{\mathrm{LAP}} = \operatorname*{argmin}_{\boldsymbol{\theta}\in\Theta} \; -\log \tilde{p}_h(\boldsymbol{\theta}) = \operatorname*{argmin}_{\boldsymbol{\theta}\in\Theta} \; \Phi_h(\boldsymbol{\theta}), \tag{17}$$

and whose covariance is the negative inverse Hessian of the log-likelihood evaluated at the mode

$$\boldsymbol{\Sigma}_h^{\mathrm{LAP}} = -\left[\nabla\nabla^T \log \tilde{p}_h\left(\boldsymbol{\mu}_h^{\mathrm{LAP}}\right)\right]^{-1} = \left[\nabla\nabla^T \Phi_h\left(\boldsymbol{\mu}_h^{\mathrm{LAP}}\right)\right]^{-1}. \tag{18}$$

A Laplace approximation may not exist for certain distributions where the covariance matrix $\boldsymbol{\Sigma}_h^{\mathrm{LAP}}$ or Hessian at the mode is not full-rank. If the following proposition applies, then a Laplace approximation exists and is a suitable biasing distribution; we refer to [39] for in-depth discussions about Laplace approximations as biasing distributions if the covariance matrix is singular. More generally, we are interested in finding optimal biasing densities and types of biasing densities than Laplace approximations. Two notable examples are parametrized transport maps and Gaussian mixture models for greater flexibility. While such biasing densities may result in better approximations than the Laplace approximation, they are computationally more challenging to fit and we are unaware of results that provide similar guarantees on the potential that we prove in the following proposition.

Proposition 1. *Let Assumption 1 hold and assume there exists a $\sigma^2_{\min} > 0$, independent of h, such that*

$$\boldsymbol{\theta}^T \boldsymbol{\Sigma}_h^{\mathrm{LAP}} \boldsymbol{\theta} \geq \sigma^2_{\min} \|\boldsymbol{\theta}\|^2, \tag{19}$$

for all $\boldsymbol{\theta} \in \Theta$. Further, assume there exist constants $V \in \mathbb{R}$ and $v > 0$ such that

$$\Phi_h(\boldsymbol{\theta}) \geq V - v\|\boldsymbol{\theta}\|^2 \tag{20}$$

for all h. Finally, let $B_R = \{\boldsymbol{\theta} : \|\boldsymbol{\theta}\| \leq R\}$ be the ball of radius R centered at 0, and assume that for all $D > 0$, there exists an $R(D) > 0$ such that for all $\boldsymbol{\theta} \notin B_{R(D)}$ and all $h > 0$

$$\Phi_h(\boldsymbol{\theta}) \geq D. \tag{21}$$

Then, the Laplace approximation satisfies Assumption 4 for all h sufficiently small.

Proof. By Assumption 1, a Laplace approximation

$$\tilde{\Phi}_h(\boldsymbol{\theta}) = \Phi_h\left(\boldsymbol{\mu}_h^{\mathrm{LAP}}\right) + \frac{1}{2}\left(\boldsymbol{\theta} - \boldsymbol{\mu}_h^{\mathrm{LAP}}\right)^T \left[\nabla\nabla^T \Phi_h(\boldsymbol{\mu}_h^{\mathrm{LAP}})\right]^{-1} \left(\boldsymbol{\theta} - \boldsymbol{\mu}_h^{\mathrm{LAP}}\right)$$

is the second-order Taylor expansion of Φ_h around one of the modes $\boldsymbol{\mu}_h^{\mathrm{LAP}}$. The first derivative is zero since it is expanded around a minimizer. Therefore,

$$\tilde{\Phi}_h(\boldsymbol{\theta}) - \Phi_h(\boldsymbol{\theta}) = -R_h(\boldsymbol{\theta})\,,$$

where $R_h(\boldsymbol{\theta})$ is the remainder of higher order terms from the Taylor expansion. The bound (19) implies that

$$\boldsymbol{\theta}^T \left(\boldsymbol{\Sigma}_h^{\mathrm{LAP}}\right)^{-1} \boldsymbol{\theta} \leq \frac{1}{\sigma_{\min}^2}\|\boldsymbol{\theta}\|^2\,,$$

and when combined with the bound (20) gives

$$\begin{aligned}\tilde{\Phi}_h(\boldsymbol{\theta}) - \Phi_h(\boldsymbol{\theta}) &\leq \tilde{\Phi}_h(\boldsymbol{\theta}) - V + v\|\boldsymbol{\theta}\|^2 \\ &\leq \Phi_h\left(\boldsymbol{\mu}_h^{\mathrm{LAP}}\right) + \frac{1}{2\sigma_{\min}^2}\|\boldsymbol{\theta} - \boldsymbol{\mu}_h^{\mathrm{LAP}}\|^2 - V + v\|\boldsymbol{\theta}\|^2\,.\end{aligned}$$

Combining this with the fact that $\|\boldsymbol{x} - \boldsymbol{y}\|^2 \leq 2\|\boldsymbol{x}\|^2 + 2\|\boldsymbol{y}\|^2$ yields

$$\tilde{\Phi}_h(\boldsymbol{\theta}) - \Phi_h(\boldsymbol{\theta}) \leq \Phi_h\left(\boldsymbol{\mu}_h^{\mathrm{LAP}}\right) + \left(\frac{1}{\sigma_{\min}^2} + v\right)\|\boldsymbol{\theta}\|^2 + \frac{1}{\sigma_{\min}^2}\|\boldsymbol{\mu}_h^{\mathrm{LAP}}\|^2 - V\,.$$

Now we claim that the terms $\Phi_h(\boldsymbol{\mu}_h^{\mathrm{LAP}})$ and $\|\boldsymbol{\mu}_h^{\mathrm{LAP}}\|^2$ can be bounded independent of h. Let $D = \Phi(0)+1$ and consider that, by assumption, there exists a ball $B_{R(D)}$ such that

$$\Phi_h(\boldsymbol{\theta}) \geq \Phi(0) + 1\,, \quad \forall \boldsymbol{\theta} \notin B_{R(D)}\,.$$

By Assumption 1, we know that $\Phi_h(0) \to \Phi(0)$ and so that for all h sufficiently small, there exist points $\boldsymbol{\theta}_h'$, such that $\Phi_h(\boldsymbol{\theta}_h') \leq \Phi(0) + 1$. Hence, the minimizers $\boldsymbol{\mu}_h^{\mathrm{LAP}} \in B_R$ for all h sufficiently small. Thus, there are constants $B_1, B_2 > 0$ independent of h such that $\Phi_h\left(\boldsymbol{\mu}_h^{\mathrm{LAP}}\right) \leq B_1$ and $\|\boldsymbol{\mu}_h^{\mathrm{LAP}}\|^2 \leq B_2$. Thus, by setting

$$\gamma(h) = \frac{1}{\sigma_{\min}^2} + v, \quad \omega(\boldsymbol{\theta}) = \|\boldsymbol{\theta}\|^2 + \omega_0, \quad \omega_0 = \frac{B_1 + B_2/\sigma_{\min}^2 - V}{\sigma_{\min}^{-2} + v}$$

Assumption 4 holds. □

If Proposition 1 applies, then it is guaranteed that there exists a Laplace approximation and that its covariance matrix remains non-singular as the fidelity h is reduced: Condition (19) ensures that the covariance matrix $\boldsymbol{\Sigma}_h^{\mathrm{LAP}}$ is positive definite and hence that a Laplace approximation q_h of p_h exists for all $h > 0$. The requirement that $\sigma_{\min}^2$ is independent of h prevents the sequence of covariance matrices from approaching a singular matrix in the limit $h \to 0$. Condition (20) is related to Assumption 2.6(i) from [40]. A pointwise bound is used to satisfy Assumption 4 and ensure the integrability from Theorem 1. Condition (21) implies that $\Phi_h(\boldsymbol{\theta}) \to \infty$ as $\|\boldsymbol{\theta}\| \to \infty$ uniformly in h, and so we know that a global minimizer exists for each potential Φ_h; however, it is not necessarily unique. In the scenario where multiple global minima exist, we may choose any $\boldsymbol{\mu}_h^{\mathrm{LAP}}$ from the set of global minimizers. In particular, we allow for multi-modal target and surrogate densities p and p_h and allow for the Laplace approximation to be a local approximation of one of the local optimum as long as the covariance matrix satisfies the assumptions of Proposition 1 so that $p, p_h, \tilde{p}_h$ all satisfy Assumption 3 and 4, which are necessary for the importance sampling estimator (6) to converge. In particular we note that we do not need to find all global minimizers.

Remark 4. *If Proposition* 1 *holds, then the Laplace approximation serves as a suitable biasing density for importance sampling in the sense that Assumption* 4 *holds, which is needed for Theorem* 1. *Recall that as the fidelity $h \to 0$ we may* not *have $\gamma(h) \to 0$ and so $\chi^2(p \,||\, q_h)$ may not go to zero. We note that Proposition* 1 *implies the existence of a Laplace approximation and Assumption* 4 *but not necessarily a $\gamma(h)$ that satisfies condition* (8) *in Theorem* 1.

3.4 Trading off fidelity and costs of surrogate model for MFIS

We now consider the trade-off between selecting a fidelity h to construct a Laplace approximation and the number of samples m in the MFIS estimator (6).

3.4.1 Offline and online costs of MFIS with Laplace approximation as biasing density

The total computational costs of estimating $\mathbb{E}_p[f]$ with the MFIS estimator $\hat{f}_{h,m}$ defined in Equation (6) can be decomposed into training (offline) costs to fit the biasing density q_h and the online costs to sample and re-weight; cf. Section 2.5.

In the training phase, the biasing density is constructed. In the following, we consider a Laplace approximation q_h of the surrogate density p_h as the biasing density. The Laplace approximation is constructed from M evaluations of the un-normalized surrogate density $\tilde{p}_h$ and so the training costs are $Mc(h)$ in our case. Recall that $c(h)$ is the cost of evaluating the un-normalized surrogate density $\tilde{p}_h$. For example, in Section 5, M will be the total number of surrogate-density evaluations used in Newton's method until machine precision is reached, where both the gradient and Hessian are computed using either finite differences or the adjoint method as well as computing the Hessian at the mode.

In the online phase, the weights of the MFIS estimator are obtained by evaluating the target density and the biasing density at m samples. We model the online costs as mC, where C denotes the cost of a single evaluation of the un-normalized target density $\tilde{p}$. No evaluations of the surrogate density are necessary in the online phase because only the biasing density (Laplace approximation in our case) is evaluated, which has costs that typically are independent of h and negligible compared to evaluating the target density $\tilde{p}$. However, notice that the online costs depend implicitly on the fidelity h because the number of samples m to reach an MSE below a threshold depends on the quality of the biasing distribution in the sense of the divergence $\chi^2(p||q_h)$; cf. Section 2.3.

We obtain as the total costs of the MFIS estimator

$$\text{cost}(\hat{f}_{h,m}) = mC + Mc(h)\,, \tag{22}$$

which depends on the number of samples m and on the fidelity h of the surrogate.

3.4.2 Cost complexity bounds of MFIS

The following theorem provides cost-complexity bounds for the MFIS estimator under assumptions of the surrogate densities cost and error. We define the context-aware MFIS estimator to be the estimator (6) with fidelity h^* and sample size m^* given by the following theorem.

Theorem 2. *Suppose that Theorem 1 and Proposition 1 apply. Consider a tolerance $0 < \epsilon \leq 1$ and set $K_0' = 4\|f\|_{L^\infty}^2 \tilde{K}_0 + 1$, where $\tilde{K}_0$ is the constant in Equation (11). If the surrogate density evaluation costs grow as $c(h) = \beta^{1/h}$ with the fidelity h and the surrogate error decays as $\delta(h) = \alpha^{-1/h}$ in Assumption 3, with $\alpha, \beta > 1$, and we restrict $h \in [0, \log(\alpha)/2]$, then there exist $h^* \in [0, \log(\alpha)/2]$ and $m^* \in \mathbb{N}$ such that the MFIS estimator $\hat{f}_{h^*,m^*}$ achieves an MSE less than the tolerance ϵ and the costs are bounded as*

$$\text{cost}(\hat{f}_{h^*,m^*}) \leq \overline{\text{cost}}(\hat{f}_{h^*,m^*}) = \frac{CK_0'}{\epsilon} \mathrm{e}^{K_1 \epsilon^{1/(1+\log_\alpha \beta)}} + M\epsilon^{-1/(1+\log_\beta \alpha)}\,.$$

If instead $c(h) = h^{-\beta}$ and $\delta(h) = h^\alpha$ with $\alpha, \beta > 0$, then the costs are bounded as

$$\text{cost}(\hat{f}_{h^*,m^*}) \leq \overline{\text{cost}}(\hat{f}_{h^*,m^*}) = \frac{CK_0'}{\epsilon} \mathrm{e}^{K_1 \epsilon^{\alpha/(\alpha+\beta)}} + M\epsilon^{-\beta/(\alpha+\beta)}\,.$$

Here we use the notation $\overline{\text{cost}}$ to denote the upper bound to $\text{cost}(\hat{f}_{h^*,m^*})$ as in Theorem 2 above.

Remark 5. *For $\delta(h) = \alpha^{-1/h}$ we require that $h \leq \log(\alpha)/2$ to satisfy the convexity assumption in Lemma 2. Note that as $\epsilon \to 0$, the fidelity h must also go to zero by Equation (26). In particular, taking ϵ sufficiently small, smaller than*

$$\epsilon \leq \frac{4\|f\|_{L^\infty}^2 \tilde{K}_0 K_1 C \log \alpha}{M \log \beta} \mathrm{e}^{K_1 \alpha^{-2/\log \alpha}} (\alpha\beta)^{-2/\log \alpha}\,,$$

will ensure that $h \leq \log(\alpha)/2$ *as required.*

The rates on the error $\delta(h)$ and the cost $c(h)$ can arise, for example, in the Bayesian inverse problem setting in Section 4, where surrogate models are used to construct the surrogate densities p_h. Several concrete examples will be given in Section 5. Notice that $\gamma(h)$ from Assumption 4 influences implicitly the constant $\tilde{K}_0$ as shown in (11), which amplifies Remark 2 that it is unnecessary that $\gamma(h)$ goes to 0 for $h \to 0$ for Theorem 2 to hold.

Before we prove Theorem 2, we state the following lemma that solves an auxiliary optimization problem highlighting the trade-off between the costs and fidelity of the surrogate model.

Lemma 2. *Let* $\hat{c}(\hat{h})$ *and* $\hat{e}(\hat{h})$ *be continuous non-negative convex functions, where one of them is strictly convex. Let further* $\hat{c}(\hat{h})$ *decrease monotonically and* $\hat{e}(\hat{h})$ *increases monotonically as* $\hat{h} \to \infty$*. Let* $\epsilon > 0$ *be a tolerance and* $\hat{M} \in \mathbb{N}$ *be a constant independent of* $\hat{h}$*. Then, there exists a unique solution* $(\hat{h}^*, \hat{m}^*)$ *of*

$$\begin{aligned} \underset{\hat{m}\in\mathbb{R},\hat{h}\geq 0}{\textit{minimize}} \quad & \hat{m}C + \hat{M}\hat{c}(\hat{h}) \\ \textit{subject to} \quad & \frac{1}{\hat{m}}\hat{e}(\hat{h}) \leq \epsilon\,. \end{aligned} \tag{23}$$

Proof of Lemma 2. We proceed as follows: first we show that if a solution exists it cannot occur at zero or infinity (i.e. too high or low fidelity), then we show that a solution exists over a compact interval, and finally show its uniqueness. For any $\hat{h}$, the optimal $\hat{m}$ is the one that achieves equality in the constraint

$$\hat{m} = \frac{\hat{e}(\hat{h})}{\epsilon}\,.$$

Plugging this into the objective function gives the minimization problem over $\hat{h}$ only.

$$\underset{\hat{h}\geq 0}{\text{minimize}} \quad C\frac{\hat{e}(\hat{h})}{\epsilon} + \hat{M}\hat{c}(\hat{h})\,. \tag{24}$$

We first show that the infimum of the objective function cannot occur as $\hat{h} \to \infty$ or as $\hat{h} \to 0$. Since $\hat{c}(\hat{h})$ is non-negative and decreasing we know that $\hat{c}(\hat{h}) \to c_0$ for some constant $c_0 \geq 0$. Moreover, $\hat{e}(\hat{h})$ is increasing, so we know that there exists an $\hat{h}_{\max} < \infty$, such that any optimal solution $\hat{h}^*$ must satisfy $\hat{h}^* \leq \hat{h}_{\max}$. Similarly, since $\hat{e}(\hat{h})$ is non-negative and decreasing as $\hat{h} \to 0$ we know that $\hat{e}(\hat{h}) \to e_0$ for some constant $e_0 \geq 0$ as $\hat{h} \to 0$. Moreover, $\hat{c}(\hat{h})$ is increasing as $\hat{h} \to 0$, and since the objective function (24) is monotonically increasing as $\hat{h} \to 0$, we know that there exists an $\hat{h}_{\min} > 0$, such that any optimal solution $\hat{h}^*$ must satisfy $\hat{h}^* \geq \hat{h}_{\min}$. Hence

$$\underset{\hat{h}\geq 0}{\text{minimize}} \quad C\frac{\hat{e}(\hat{h})}{\epsilon} + \hat{M}\hat{c}(\hat{h}) = \underset{\hat{h}\in[\hat{h}_{\min},\hat{h}_{\max}]}{\text{minimize}} \quad C\frac{\hat{e}(\hat{h})}{\epsilon} + \hat{M}\hat{c}(\hat{h})$$

Since the objective function is continuous over a compact set, we know that a minimizer exists.

Finally, the sum of a strictly convex function and a convex function is strictly convex, so we know that this objective function is strictly convex in h, and therefore the minimizer is unique. □

Proof of Theorem 2. Combining the result of Theorem 1 in Equation (10) with the bound (3), let

$$e(h) = 4\|f\|_{L^\infty}^2 \tilde{K}_0 \mathrm{e}^{K_1\delta(h)}\,.$$

Because the composition of the convex function $\delta(h)$ and the strictly convex and increasing function $x \mapsto \mathrm{e}^x$ is strictly convex, we know that $e(h)$ must be strictly convex and therefore satisfies the assumptions of Lemma 2, meaning that a unique solution $\hat{h}^*, \hat{m}^* \in \mathbb{R}$ exists.

Consider $c(h) = \beta^{1/h}$ and $\delta(h) = \alpha^{-1/h}$ with $\alpha, \beta > 1$. We can remove the constraint to instead minimize

$$\underset{h\geq 0}{\text{minimize}} \quad \frac{4\|f\|_{L^\infty}^2 \tilde{K}_0 C}{\epsilon}\mathrm{e}^{K_1\delta(h)} + Mc(h), \tag{25}$$

which is analogous to (24). By setting the derivative of (25) with respect to h to zero, the optimal solution satisfies

$$\frac{4\|f\|_{L^\infty}^2 \tilde{K}_0 K_1 C \log \alpha}{M \log \beta} \mathrm{e}^{K_1 \alpha^{-1/h}} = \epsilon (\alpha\beta)^{1/h}\,, \tag{26}$$

meaning that $1/\hat{h}^* \in \mathcal{O}(\log_{\alpha\beta} \epsilon^{-1})$ as $\epsilon \to 0$ since the left-hand-side must approach a constant. Motivated by this observation, we set $1/h^* = \log_{\alpha\beta} \epsilon^{-1}$ exactly and then the number of samples needed is

$$m^* = \lceil \hat{m}^* \rceil = \left\lceil \frac{4\|f\|_{L^\infty}^2 \tilde{K}_0}{\epsilon} \mathrm{e}^{K_1 \epsilon^{1/(1+\log_\alpha \beta)}} \right\rceil \leq \frac{4\|f\|_{L^\infty}^2 \tilde{K}_0}{\epsilon} \mathrm{e}^{K_1 \epsilon^{1/(1+\log_\alpha \beta)}} + 1\,.$$

where we have used that $\log_{\alpha\beta} \epsilon = \frac{\log_\alpha \epsilon}{1+\log_\alpha \beta} = \frac{\log_\beta \epsilon}{1+\log_\beta \alpha}$. Since $\epsilon \leq 1$ we know that $\mathrm{e}^{K_1 \epsilon^{1/(1+\log_\alpha \beta)}}/\epsilon > 1$, and so

$$m^* \leq K_0' \frac{\mathrm{e}^{K_1 \epsilon^{1/(1+\log_\alpha \beta)}}}{\epsilon}\,.$$

Plugging this in for m into the objective function, gives an upper bound on the total computational costs

$$\mathrm{cost}(\hat{f}_{h^*,m^*}) \leq \frac{CK_0'}{\epsilon} \mathrm{e}^{K_1 \epsilon^{1/(1+\log_\alpha \beta)}} + M\epsilon^{-1/(1+\log_\beta \alpha)}\,.$$

Now consider $c(h) = h^{-\beta}$ and $\delta(h) = h^\alpha$ with $\alpha, \beta \geq 1$. Set again the derivative to zero to find that the optimal solution satisfies

$$\frac{4\|f\|_{L^\infty}^2 C \tilde{K}_0 K_1}{M} \left(\frac{\alpha}{\beta}\right) \mathrm{e}^{K_1 h^\alpha} h^{\alpha+\beta} = \epsilon\,,$$

so that $\hat{h}^* \in \mathcal{O}(\epsilon^{1/(\alpha+\beta)})$ as $\epsilon \to 0$. If we set $h^* = \epsilon^{1/(\alpha+\beta)}$, then the number of samples needed is

$$m^* = \lceil \hat{m}^* \rceil = \left\lceil \frac{4\|f\|_{L^\infty}^2 \tilde{K}_0}{\epsilon} \mathrm{e}^{K_1 \epsilon^{\alpha/(\alpha+\beta)}} \right\rceil \leq \frac{4\|f\|_{L^\infty}^2 \tilde{K}_0}{\epsilon} \mathrm{e}^{K_1 \epsilon^{\alpha/(\alpha+\beta)}} + 1 \leq \frac{K_0'}{\epsilon} \mathrm{e}^{K_1 \epsilon^{\alpha/(\alpha+\beta)}}\,,$$

with total computational cost bounded as

$$\mathrm{cost}(\hat{f}_{h^*,m^*}) \leq \frac{CK_0'}{\epsilon} \mathrm{e}^{K_1 \epsilon^{\alpha/(\alpha+\beta)}} + M\epsilon^{-\beta/(\alpha+\beta)}\,.$$

□

Remark 6. *Although we have assumed that training costs correspond to fitting the Laplace approximation, Lemma 2 shows that the results will extend more generally to any approximation where the costs of fitting the biasing density with respect to the fidelity h satisfies the assumption of Lemma 2.*

3.4.3 Discussion of cost complexity bounds of context-aware MFIS

We now compare the cost bounds of the context-aware MFIS estimators $\hat{f}_{h^*,m^*}$ derived in Theorem 2 with the costs of fixed-fidelity MFIS estimators $\hat{f}_{\bar{h},\bar{m}}$, where the fidelity $\bar{h}$ is fixed independent of ϵ. The number of samples $\bar{m}$ is selected depending on the tolerance ϵ as

$$\bar{m} = \inf \left\{ m \in \mathbb{N} : \frac{e(\bar{h})}{m} \leq \epsilon \right\}\,,$$

analogously to the context-aware MFIS estimator. Note that the sample size depends as well on the fidelity $\bar{h}$. The costs of the fixed-fidelity MFIS estimator are

$$\mathrm{cost}(\hat{f}_{\bar{h},\bar{m}}) = \bar{m}C + Mc(\bar{h})\,.$$

If $\delta(h) = \alpha^{-1/h}$ and $c(h) = \beta^{1/h}$, then the costs of the fixed-fidelity estimator are bounded as

$$\mathrm{cost}(\hat{f}_{\bar{h},\bar{m}}) \leq \overline{\mathrm{cost}}(\hat{f}_{\bar{h},\bar{m}}) = \frac{CK_0'}{\epsilon} \mathrm{e}^{K_1 \alpha^{-1/\bar{h}}} + M\beta^{1/\bar{h}}\,,$$

and if $\delta(h) = h^\alpha$ and $c(h) = h^{-\beta}$ then the costs are bounded as

$$\text{cost}(\hat{f}_{\bar{h},\bar{m}}) \leq \overline{\text{cost}}(\hat{f}_{\bar{h},\bar{m}}) = \frac{CK_0'}{\epsilon} \mathrm{e}^{K_1 \bar{h}^\alpha} + M\bar{h}^{-\beta} \,.$$

We now compare the costs of the context-aware MFIS and the fixed-fidelity MFIS estimators by comparing their cost upper bounds $\overline{\text{cost}}$ as $\epsilon \to 0$. First, consider the case where $\delta(h) = \alpha^{-1/h}$ and $c(h) = \beta^{1/h}$. As $\epsilon \to 0$, we have that

$$\lim_{\epsilon \to 0} \frac{\overline{\text{cost}}(\hat{f}_{\bar{h},\bar{m}})}{\overline{\text{cost}}(\hat{f}_{h^*,m^*})} = \lim_{\epsilon \to 0} \frac{\frac{CK_0'}{\epsilon} \mathrm{e}^{K_1 \alpha^{-1/\bar{h}}} + M\beta^{1/\bar{h}}}{\frac{CK_0'}{\epsilon} \mathrm{e}^{K_1 \epsilon^{1/(1+\log_\alpha \beta)}} + M\epsilon^{-1/(1+\log_\beta \alpha)}} \,.$$

Multiply the numerator and denominator by ϵ to get

$$\lim_{\epsilon \to 0} \frac{CK_0' \mathrm{e}^{K_1 \alpha^{-1/\bar{h}}} + \epsilon M\beta^{1/\bar{h}}}{CK_0' \mathrm{e}^{K_1 \epsilon^{1/(1+\log_\alpha \beta)}} + M\epsilon^{1-1/(1+\log_\beta \alpha)}} \,.$$

As $\epsilon \to 0$, the numerator goes to $CK_0' \mathrm{e}^{K_1 \alpha^{-1/\bar{h}}}$ and the denominator goes to CK_0' since $\alpha > 1$. Therefore, the speedup obtained with the context-aware MFIS estimator in the limit of $\epsilon \to 0$ is

$$\lim_{\epsilon \to 0} \frac{\overline{\text{cost}}(\hat{f}_{\bar{h},\bar{m}})}{\overline{\text{cost}}(\hat{f}_{h^*,m^*})} = \mathrm{e}^{K_1 \alpha^{-1/\bar{h}}} > 1 \,.$$

Now consider the other case where $\delta(h) = h^\alpha$ and $c(h) = h^{-\beta}$. We have that

$$\lim_{\epsilon \to 0} \frac{\overline{\text{cost}}(\hat{f}_{\bar{h},\bar{m}})}{\overline{\text{cost}}(\hat{f}_{h^*,m^*})} = \lim_{\epsilon \to 0} \frac{\frac{CK_0'}{\epsilon} \mathrm{e}^{K_1 \bar{h}^\alpha} + M\bar{h}^{-\beta}}{\frac{CK_0'}{\epsilon} \mathrm{e}^{K_1 \epsilon^{\alpha/(\alpha+\beta)}} + M\epsilon^{-\beta/(\alpha+\beta)}} \,.$$

Multiplying both the numerator and denominator by ϵ gives

$$\lim_{\epsilon \to 0} \frac{CK_0' \mathrm{e}^{K_1 \bar{h}^\alpha} + \epsilon M\bar{h}^{-\beta}}{CK_0' \mathrm{e}^{K_1 \epsilon^{\alpha/(\alpha+\beta)}} + M\epsilon^{1-\beta/(\alpha+\beta)}} \,.$$

As $\epsilon \to 0$, the numerator converges to $CK_0' \mathrm{e}^{K_1 \bar{h}^\alpha}$, and since $\beta/(\alpha+\beta) < 1$, the denominator converges to CK_0'. Hence, the speedup obtained with the proposed context-aware MFIS estimator in the limit $\epsilon \to 0$ is

$$\lim_{\epsilon \to 0} \frac{\overline{\text{cost}}(\hat{f}_{\bar{h},\bar{m}})}{\overline{\text{cost}}(\hat{f}_{h^*,m^*})} = \mathrm{e}^{K_1 \bar{h}^\alpha} > 1 \,.$$

In both cases we observe that as the tolerance $\epsilon \to 0$, the dominant term for the MFIS estimator cost approaches order $\mathcal{O}(1/\epsilon)$ and the bulk of the cost shifts to the online sampling cf. Section 2.5. We see that the speedup as $\epsilon \to 0$ depends on the rate of the error $\delta(\bar{h})$ going to zero.

3.5 Computational procedure

Algorithm 1 summarizes the context-aware importance sampling procedure. Given constants $\tilde{K}_0, K_1, C, M, \|f\|_{L^\infty}$, and the tolerance ϵ as well as the cost and accuracy functions c and δ, the context-aware importance sampling Algorithm 1 first solves the optimization problem (23) for $(\hat{h}^*, \hat{m}^*)$. A Laplace approximation to the surrogate density p_{h^*} is then computed using Newton's method. The Hessian at the mode is then inverted directly to obtain the covariance of the Laplace approximation or can alternatively be stored as a precision matrix to avoid the matrix inversion. This concludes the offline phase of finding the biasing density. For the online phase we draw $m^* = \lceil \hat{m}^* \rceil$ samples from the Laplace approximation q_{h^*} and re-weight using the un-normalized high-fidelity density $\tilde{p}$ using the estimator (6).

Algorithm 1 requires the constants $\tilde{K}_0, K_1, C, M, \|f\|_{L^\infty}$. Similar to other multi-level and multi-fidelity methods, we propose to first perform a pilot study to estimate these constants before using them in the computational procedure. Such pilot studies may be expensive; however, since the test function f is independent of the constants, we only need to estimate these constants once and can then re-use them to compute a variety of statistics with respect to the target distribution p. This makes the context-aware importance sampling procedure appealing for estimating families of expectations or probabilities such as cumulative density functions or survival functions.

Algorithm 1 Context-aware importance sampling

1: Constants $\tilde{K}_0, K_1, C, \epsilon, M, \|f\|_{L^\infty}$ and functions c, δ
2: Solve the optimization problem (23) for $(h^*, \hat{m}^*)$ using $\|f\|_{L^\infty}, \tilde{K}_0, K_1, C, M, \epsilon, c, \delta$
3: Compute a Laplace approximation q_{h^*} of p_{h^*} with M evaluations of $\tilde{p}_{h^*}$
4: Draw $m^* = \lceil \hat{m}^* \rceil$ i.i.d. samples $\{\boldsymbol{\theta}^{(i)}\}_{i=1}^{m^*}$ from q_{h^*}
5: Compute $\hat{f}_{h^*,m^*}$ using (6) **return** Estimate $\hat{f}_{h^*,m^*}$

4 Bayesian inverse problems

We now apply the context-aware MFIS estimator for inference in Bayesian inverse problems where the target p is a posterior distribution and we are interested in expectations $\mathbb{E}_p[f]$ of this distribution. Section 4.1 describes the general setup of a Bayesian inverse problem and Section 4.2 applies the results of Section 3 to the case where p is a posterior distribution.

4.1 Setup of a Bayesian inverse problem

Let data $\boldsymbol{y} \in \mathbb{R}^{d'}$ be generated by an unknown parameter $\boldsymbol{\theta}_{\text{truth}} \in \mathbb{R}^d$ with a Gaussian noise model,

$$\boldsymbol{y} = \mathcal{F}(\boldsymbol{\theta}_{\text{truth}}) + \boldsymbol{\eta},$$

where $\boldsymbol{\eta} \sim N(0, \boldsymbol{\Gamma})$, $\boldsymbol{\Gamma} \in \mathbb{R}^{d' \times d'}$ is the covariance matrix (symmetric and positive definite) of the added noise, and $\mathcal{F} : \Theta \to \mathbb{R}^{d'}$ is the high-fidelity parameter-to-observable map. Let π_{pr} denote a prior distribution over the parameter $\boldsymbol{\theta}$, so that the negative log-posterior has the form

$$-\log p(\boldsymbol{\theta}) = \Phi(\boldsymbol{\theta}) = \frac{1}{2}\|\boldsymbol{y} - \mathcal{F}(\boldsymbol{\theta})\|^2_{\boldsymbol{\Gamma}^{-1}} - \log \pi_{\text{pr}}(\boldsymbol{\theta}).$$

The norm is defined as $\|\boldsymbol{v}\|^2_{\boldsymbol{\Gamma}^{-1}} = \langle \boldsymbol{\Gamma}^{-1}\boldsymbol{v},\ \boldsymbol{v}\rangle$. While it is possible to use the prior distribution as a biasing density, if the posterior contracts around the data then the χ^2 divergence of the posterior from the prior may be very large resulting in a high variance estimator with a low effective sample size.

Let $\mathcal{F}_h$ denote the surrogate parameter-to-observable map with fidelity h and let it be such that the sequence $\mathcal{F}_h(\boldsymbol{\theta}) \to \mathcal{F}(\boldsymbol{\theta})$ converges pointwise for each $\boldsymbol{\theta} \in \Theta$. Additionally, we assume that $\mathcal{F}, \mathcal{F}_h \in \mathcal{C}^2(\Theta)$. In many cases the parameter-to-observable map $\mathcal{F}$ is a function of an intermediate state variable u, such as the full solution to a parametrized partial differential equation (PDE) depending on the parameters $\boldsymbol{\theta}$. The surrogate parameter-to-observable map $\mathcal{F}_h$ is given by approximating this state variable u with an approximation u_h. The approximation for the state variable u_h could be given by finite elements [7], finite difference [25], a different time step for an ordinary differential equation [25], finitely many terms in a Karhunen-Loève expansion [41], and others.

We consider the case where the prior π_{pr} is Gaussian $N(\boldsymbol{\mu}_{\text{pr}}, \boldsymbol{\Sigma}_{\text{pr}})$, so that we can write the potential from Assumption 1 as

$$\Phi(\boldsymbol{\theta}) = \frac{1}{2}\|\boldsymbol{y} - \mathcal{F}(\boldsymbol{\theta})\|^2_{\boldsymbol{\Gamma}^{-1}} + \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\mu}_{\text{pr}})^T \boldsymbol{\Sigma}_{\text{pr}}^{-1}(\boldsymbol{\theta} - \boldsymbol{\mu}_{\text{pr}}). \tag{27}$$

With a Gaussian prior the resulting posterior distribution is always sub-Gaussian since we can take the matrix $\boldsymbol{A} = \frac{1}{4}\boldsymbol{\Sigma}_{\text{pr}}^{-1}$ in Lemma 1. The potentials Φ_h are defined similarly but with the surrogate maps $\mathcal{F}_h$ replacing $\mathcal{F}$.

4.2 Bounding χ^2 divergence with model error

We now translate bounds on the model error between $\mathcal{F}$ and $\mathcal{F}_h$ to the χ^2 divergence $\chi^2(p \,||\, q_h)$, where q_h is a Laplace approximation to the surrogate posterior p_h. The next two assumptions allow us to make the transition.

Assumption 5. *The high-fidelity parameter-to-observable map $\mathcal{F}$ is globally Lipschitz meaning there exists a constant $B > 0$ such that for all $\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}} \in \Theta$*

$$\|\mathcal{F}(\boldsymbol{\theta}) - \mathcal{F}(\tilde{\boldsymbol{\theta}})\| \leq B\|\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}\|\,.$$

Assumption 5 is almost the Lipschitz Assumption 2.7(ii) from [40] except there the constant B only needs to hold for bounded sets of $\boldsymbol{\theta}$. Assumption 5 is satisfied if the map $\mathcal{F}$ is linear, for example, or if the map is the sum of a linear term and a smooth bounded function. Alternatively, we note that Assumption 5 may also be relaxed so that $\mathcal{F}(\boldsymbol{\theta})$ grows at most linearly asymptotically as $\|\boldsymbol{\theta}\| \to \infty$. Such an assumption will still ensure that the potential does not grow faster than quadratically as needed for the assumptions of Theorem 1.

Assumption 6. *For all $\boldsymbol{\theta} \in \Theta$ and h we have*

$$\|\mathcal{F}_h(\boldsymbol{\theta}) - \mathcal{F}(\boldsymbol{\theta})\| \leq \tilde{\delta}(h)\tilde{\tau}(\boldsymbol{\theta})$$

with $\tilde{\delta}(h) \to 0$ as $h \to 0$ with $\tilde{\tau}(\boldsymbol{\theta})$ independent of h.

Assumption 6 is similar to Assumption (4.11) in Corollary 4.9 of [40], although the pointwise bound is also looser there than here for the same reason as given in Remark 3. Theorem 3 is analogous to Theorem 1 from earlier but now is applied specifically to the Bayesian inverse problem.

Theorem 3. *If Assumptions 5 and 6 are satisfied with $|\tilde{\tau}(\boldsymbol{\theta})| \leq \|\boldsymbol{\theta}\| + \tilde{\tau}_0$ for some $\tilde{\tau}_0 > 0$, then Assumption 3 is also satisfied with*

$$\delta(h) = \left(\frac{2B+1}{\kappa_{\min}}\right)\tilde{\delta}(h)$$

and $\tau(\boldsymbol{\theta})$ a quadratic function of $\|\boldsymbol{\theta}\|$ that is independent of h.

Proof. Using the form of the log-posterior (27) we write

$$|\Phi_h(\boldsymbol{\theta}) - \Phi(\boldsymbol{\theta})| = \left|\|\mathcal{F}_h(\boldsymbol{\theta}) - \boldsymbol{y}\|^2_{\boldsymbol{\Gamma}^{-1}} - \|\mathcal{F}(\boldsymbol{\theta}) - \boldsymbol{y}\|^2_{\boldsymbol{\Gamma}^{-1}}\right|$$

since the prior terms cancel. To simplify notation, set $\Delta(\boldsymbol{\theta}) = \mathcal{F}(\boldsymbol{\theta}) - \mathcal{F}_h(\boldsymbol{\theta})$ and $\zeta(\boldsymbol{\theta}) = \mathcal{F}(\boldsymbol{\theta}) - \boldsymbol{y}$, so that $\zeta(\boldsymbol{\theta}) - \Delta(\boldsymbol{\theta}) = \mathcal{F}_h(\boldsymbol{\theta}) - \boldsymbol{y}$. Now, we can instead write

$$\begin{aligned}
|\Phi_h(\boldsymbol{\theta}) - \Phi(\boldsymbol{\theta})| &= \left|\|\zeta(\boldsymbol{\theta})\|^2_{\boldsymbol{\Gamma}^{-1}} - \|\zeta(\boldsymbol{\theta}) - \Delta(\boldsymbol{\theta})\|^2_{\boldsymbol{\Gamma}^{-1}}\right| \\
&= \left|\|\zeta(\boldsymbol{\theta})\|^2_{\boldsymbol{\Gamma}^{-1}} - \left\langle \boldsymbol{\Gamma}^{-1}\left(\zeta(\boldsymbol{\theta}) - \Delta(\boldsymbol{\theta})\right),\ \zeta(\boldsymbol{\theta}) - \Delta(\boldsymbol{\theta})\right\rangle\right| \\
&= \left|2\langle \Delta(\boldsymbol{\theta}), \boldsymbol{\Gamma}^{-1}\zeta(\boldsymbol{\theta})\rangle - \|\Delta(\boldsymbol{\theta})\|^2_{\boldsymbol{\Gamma}^{-1}}\right|\,.
\end{aligned}$$

Applying the triangle inequality and then the Cauchy-Schwarz inequality to this last line gives

$$|\Phi_h(\boldsymbol{\theta}) - \Phi(\boldsymbol{\theta})| \leq 2\|\Delta(\boldsymbol{\theta})\|\|\boldsymbol{\Gamma}^{-1}\zeta(\boldsymbol{\theta})\| + \|\Delta(\boldsymbol{\theta})\|^2_{\boldsymbol{\Gamma}^{-1}}\,. \tag{28}$$

Using that $\boldsymbol{y} = \mathcal{F}(\boldsymbol{\theta}_{\text{truth}}) + \boldsymbol{\eta}$ and the triangle inequality gives

$$\begin{aligned}
\|\boldsymbol{\Gamma}^{-1}\zeta(\boldsymbol{\theta})\| &= \|\boldsymbol{\Gamma}^{-1}(\mathcal{F}(\boldsymbol{\theta}) - \boldsymbol{y})\| \\
&\leq \|\boldsymbol{\Gamma}^{-1}(\mathcal{F}(\boldsymbol{\theta}) - \mathcal{F}(\boldsymbol{\theta}_{\text{truth}}))\| + \|\boldsymbol{\Gamma}^{-1}\boldsymbol{\eta}\|\,.
\end{aligned}$$

Assumption 5 then gives the bound

$$\|\boldsymbol{\Gamma}^{-1}\zeta(\boldsymbol{\theta})\| \leq \frac{B}{\kappa_{\min}}\|\boldsymbol{\theta} - \boldsymbol{\theta}_{\text{truth}}\| + \|\boldsymbol{\Gamma}^{-1}\boldsymbol{\eta}\|, \tag{29}$$

where $\kappa_{\min} > 0$ is the smallest eigenvalue of the covariance matrix $\boldsymbol{\Gamma}$, i.e., the direction along which the posterior is most peaked. Similarly, we bound

$$\|\Delta(\boldsymbol{\theta})\|^2_{\boldsymbol{\Gamma}^{-1}} = \langle \boldsymbol{\Gamma}^{-1}\Delta(\boldsymbol{\theta}),\ \Delta(\boldsymbol{\theta})\rangle \leq \frac{1}{\kappa_{\min}}\|\Delta(\boldsymbol{\theta})\|^2\,. \tag{30}$$

Substituting bounds (29) and (30) into (28) yields

$$|\Phi_h(\boldsymbol{\theta}) - \Phi(\boldsymbol{\theta})| \leq 2\left(\frac{B}{\kappa_{\min}}\left(\|\boldsymbol{\theta} - \boldsymbol{\theta}_{\text{truth}}\|\right) + \|\boldsymbol{\Gamma}^{-1}\boldsymbol{\eta}\|\right)\|\Delta(\boldsymbol{\theta})\| + \frac{1}{\kappa_{\min}}\|\Delta(\boldsymbol{\theta})\|^2,$$

and the triangle inequality gives

$$|\Phi_h(\boldsymbol{\theta}) - \Phi(\boldsymbol{\theta})| \leq 2\left(\frac{B}{\kappa_{\min}}\left(\|\boldsymbol{\theta}\| + \|\boldsymbol{\theta}_{\text{truth}}\|\right) + \|\boldsymbol{\Gamma}^{-1}\boldsymbol{\eta}\|\right)\|\Delta(\boldsymbol{\theta})\| + \frac{1}{\kappa_{\min}}\|\Delta(\boldsymbol{\theta})\|^2. \tag{31}$$

Assumption 6 along with the assumption that $|\tilde{\tau}(\boldsymbol{\theta})| \leq \|\boldsymbol{\theta}\| + \tilde{\tau}_0$ says $\|\Delta(\boldsymbol{\theta})\| \leq \tilde{\delta}(h)\left(\|\boldsymbol{\theta}\| + \tilde{\tau}_0\right)$, so we get that

$$|\Phi_h(\boldsymbol{\theta}) - \Phi(\boldsymbol{\theta})| \leq 2\left(\frac{B}{\kappa_{\min}}\left(\|\boldsymbol{\theta}\| + \|\boldsymbol{\theta}_{\text{truth}}\|\right) + \|\boldsymbol{\Gamma}^{-1}\boldsymbol{\eta}\|\right)\tilde{\delta}(h)\left(\|\boldsymbol{\theta}\| + \tilde{\tau}_0\right) + \frac{1}{\kappa_{\min}}\tilde{\delta}(h)^2\left(\|\boldsymbol{\theta}\| + \tilde{\tau}_0\right)^2,$$

and thus

$$\begin{aligned}|\Phi_h(\boldsymbol{\theta}) - \Phi(\boldsymbol{\theta})| \leq &\left(\frac{2B}{\kappa_{\min}}\|\boldsymbol{\theta}_{\text{truth}}\| + 2\|\boldsymbol{\Gamma}^{-1}\boldsymbol{\eta}\| + \frac{\tilde{\delta}(h)\tilde{\tau}_0}{\kappa_{\min}}\right)\tilde{\delta}(h)\tilde{\tau}_0 \\ &+ \left(\frac{2B}{\kappa_{\min}}\tilde{\tau}_0 + \frac{2B}{\kappa_{\min}}\|\boldsymbol{\theta}_{\text{truth}}\| + \frac{2}{\kappa_{\min}}\tilde{\delta}(h)\tilde{\tau}_0 + 2\|\boldsymbol{\Gamma}^{-1}\boldsymbol{\eta}\|\right)\tilde{\delta}(h)\|\boldsymbol{\theta}\| \\ &+ \left(\frac{2B}{\kappa_{\min}} + \frac{1}{\kappa_{\min}}\tilde{\delta}(h)\right)\tilde{\delta}(h)\|\boldsymbol{\theta}\|^2.\end{aligned}$$

Using that $\tilde{\delta}(h) \leq 1$ for all h sufficiently small and $\|\boldsymbol{\theta}\| \leq 1 + \|\boldsymbol{\theta}\|^2$ gives

$$|\Phi_h(\boldsymbol{\theta}) - \Phi(\boldsymbol{\theta})| \leq \delta(h)\tau(\boldsymbol{\theta}),$$

where

$$\delta(h) = \left(\frac{2B+1}{\kappa_{\min}}\right)\tilde{\delta}(h)$$

is as in Assumption 3 and $\tau(\boldsymbol{\theta})$ is quadratic in $\|\boldsymbol{\theta}\|$ and is bounded independent of h. □

Corollary 1. *Suppose that Theorem 1 applies with Assumption 3 provided by Theorem 3. Then, together with Proposition 1 this implies that the cost complexity of the context-aware importance sampling estimator with a Laplace approximation biasing density is given by Theorem 2.*

5 Numerical Results

This section demonstrates our context-aware importance sampling approach on three examples. All runtime measurements were performed on compute nodes equipped with Intel Xeon Gold 6148 2.4GHz processors and 192GB of memory using a Python 3.6 implementation.

5.1 Steady-state heat conduction

In the first example we consider a steady-state heat diffusion model with constant heat source and infer a 6-dimensional variable diffusivity.

5.1.1 Problem Setup

Let $\Omega = (0,1) \subset \mathbb{R}$ and $\Theta = \mathbb{R}^6$ and consider the PDE

$$\begin{aligned}-\left(\exp\left(k(x;\boldsymbol{\theta})\right)u_x(x;\boldsymbol{\theta})\right)_x &= 1, \quad x \in \Omega \\ u(0;\boldsymbol{\theta}) &= 0 \\ k(1;\boldsymbol{\theta})u_x(1;\boldsymbol{\theta}) &= 0\end{aligned} \tag{32}$$

where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_6)^T \in \Theta$, $k : \Omega \times \Theta \to \mathbb{R}$ is the log-diffusivity, and $u : \Omega \times \Theta \to \mathbb{R}$ is the temperature function. The log-diffusivity $k(x;\ \boldsymbol{\theta})$ is a smoothed piecewise constant. In particular, let

$$I(x, \alpha) = \left(1 + \exp\left(-\frac{x - \alpha}{0.005}\right)\right)^{-1}$$

and $\alpha_i = (i-1)/6$ for $i = 1, \ldots, 7$. Define

$$\hat{k}_i(x; \boldsymbol{\theta}) = (1 - I(x, \alpha_i))\hat{k}_{i-1}(x; \boldsymbol{\theta}) + I(x, \alpha_i)\theta_i \tag{33}$$

for $i = 2, \ldots, 6$ and $\hat{k}_1(x; \boldsymbol{\theta}) = \theta_1$. Now set the log-diffusivity $k = \hat{k}_6$. We discretize (32) in the spatial domain Ω using linear finite elements with mesh width $h > 0$ (i.e. h^{-1} many elements) and the corresponding sparse (tri-diagonal) linear system is solved using a Cholesky factorization. The parameter-to-observable map $\mathcal{F}_h : \Theta \to \mathbb{R}^{120}$ is the discretized solution u_h with mesh width h evaluated at 120 equally-spaced points on Ω

$$(\mathcal{F}_h(\boldsymbol{\theta}))_i = u_h(i/120), \quad i = 1, \ldots, 120\,.$$

For the high-fidelity parameter-to-observable map we set $H^{-1} = 256$ elements, (i.e. $\mathcal{F} = \mathcal{F}_H$) and for the surrogate maps $\mathcal{F}_h$ we consider $h^{-1} = 8, 12, 16, \ldots, 64$ (multiples of 4 for the number of elements).

5.1.2 Setup of the inverse problem

A single observation $\boldsymbol{y} = \mathcal{F}(\boldsymbol{\theta}_{\text{truth}}) + \boldsymbol{\eta}$ is generated where $\boldsymbol{\theta}_{\text{truth}} = (1, \ldots, 1)^T \in \mathbb{R}^6$ and $\boldsymbol{\eta} \sim N(\mathbf{0},\ 10^{-5}\boldsymbol{I}_{120\times 120})$. The added noise corresponds to approximately 1% of the true solution u at the right endpoint $x = 1$. The prior distribution is taken to be a Gaussian with mean $\boldsymbol{\mu}_{\text{pr}} = (1, \ldots, 1)^T \in \mathbb{R}^6$ and covariance $\boldsymbol{\Sigma}_{\text{pr}} = 10^{-1}\boldsymbol{I}_{6\times 6} \in \mathbb{R}^{6\times 6}$. For the test function let $\boldsymbol{v}_1 \in \mathbb{R}^6$ be the largest eigenvector of $\boldsymbol{\Sigma}^{\text{LAP}}$ and set

$$f(\boldsymbol{\theta}) = 2 \cdot \mathbf{1}\left\{(\boldsymbol{\theta} - \boldsymbol{\mu}^{\text{LAP}}) \cdot \boldsymbol{v}_1 \geq 0\right\} - 1 \tag{34}$$

so that $f(\boldsymbol{\theta}) \in \{\pm 1\}$ for all values of $\boldsymbol{\theta}$. The idea behind this choice of test function is that the asymptotic variance of the MFIS estimator (6) is largest whenever f is not tightly concentrated around its expectation under q_{h^*}. Here the expectation of f under q_{h^*} should be close to zero even though f itself is never close to zero.

5.1.3 Results

A Laplace approximation to each surrogate posterior p_h is fit using the Newton-CG method where the gradient and Hessian matrix are computed using a second-order finite difference scheme with a total of $M = 1150$ model evaluations at each fidelity. The cost function has the form $c(h) = c_0 + c_1/h$, where c_0 is included to model any baseline cost independent of the fidelity, and accuracy has the form $\delta(h) = a_1 h^2$ since we use linear finite elements. The cost is linear in h^{-1} since the system of linear equations is tri-diagonal. We estimate the χ^2 divergence with Monte Carlo estimator

$$\hat{\chi}^2_{h,m} = m\frac{\sum_{i=1}^m \left(\tilde{p}(\boldsymbol{\theta}^{(i)})/q_h(\boldsymbol{\theta}^{(i)})\right)^2}{\left(\sum_{i=1}^m \tilde{p}(\boldsymbol{\theta}^{(i)})/q_h(\boldsymbol{\theta}^{(i)})\right)^2} \longrightarrow \chi^2\left(p \mid\mid q_h\right) + 1, \quad \text{almost surely as } m \to \infty \tag{35}$$

and $\{\boldsymbol{\theta}^{(i)}\}_{i=1}^m$ are i.i.d. samples drawn from q_h. Then the curve $\tilde{K}_0 e^{K_1 h^2}$ is fit using the estimated χ^2 values $\hat{\chi}^2_{h,10^3}$ for each fidelity $h^{-1} = 8, 12, 16, \ldots, 64$ averaged over $N_1 = 500$ independent trials. The measured χ^2 values are

$$\hat{\chi}^2_{\text{meas},h} = \frac{1}{N_1}\sum_{i=1}^{N_1}\left(\hat{\chi}^2_{h,m}\right)^{(i)}$$

with the superscript (i) denoting one of the independent trials. The fitted curve along with the measured values are shown in Figure 1. The χ^2 divergence is large for low fidelities but quickly levels off and then is

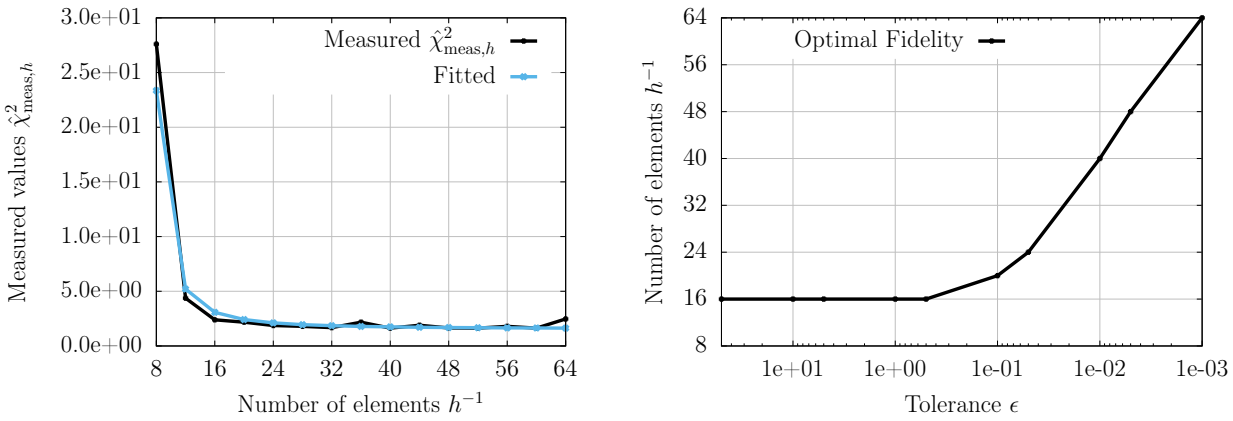

Figure 1: (Left) The measured χ^2 divergences, $\hat{\chi}^2_{\text{meas},h}$, between the high-fidelity posterior p and the Laplace approximation q_h to each surrogate posterior p_h. (Right) The selected fidelity for the number of elements $(h^*)^{-1}$ from the optimization (23) as the tolerance ϵ on the MSE changes.

limited by the restriction of the biasing density to be the Laplace approximation rather than the surrogate density itself.

Since we only consider finitely many surrogate models $h^{-1} = 8, 12, 16, \ldots, 64$, we approximate the solution of the optimization problem (23) with a brute force search to find the best fidelity h^* from the list of fidelities that we consider and set $m^* = \lceil \hat{m}^* \rceil$ with $\hat{m}^*$ corresponding to h^*. Figure 1 shows the selected fidelity as a function of the tolerance ϵ. As the tolerance shrinks we require a higher-fidelity model to fit a Laplace approximation. Using the pair (h^*, m^*), Figure 2 shows the theoretical optimal trade-off between cost in seconds and the MSE (tolerance) of the estimator $\hat{f}_{h^*,m^*}$. We estimated the true value $\mathbb{E}_p[f]$ using $\hat{f}_{H,10^5}$ and averaged the results over $N_2 = 500$ independent trials (again denoted by the superscript (i))

$$\bar{f} = \frac{1}{N_2} \sum_{i=1}^{N_2} \hat{f}^{(i)}_{H,10^5} \,. \tag{36}$$

Next we estimated the MSE of $\hat{f}_{h^*,m^*}$ using $N_3 = 1000$ trials

$$\widehat{\text{MSE}}_\epsilon = \frac{1}{N_3} \sum_{i=1}^{N_3} \left(\hat{f}^{(i)}_{h^*,m^*} - \bar{f} \right)^2 . \tag{37}$$

Here the subscript ϵ denotes the dependence of the pair (h^*, m^*) on the tolerance ϵ. Figure 2 shows the averaged MSE over $N_3 = 1000$ trials for different tolerances ϵ as well as the MSE for the estimators $\hat{f}_{H,m_H}$ and $\hat{f}_{h_0,m_{h_0}}$ where the number of samples is

$$m_h = \left\lceil \frac{\tilde{K}_0}{\epsilon} \exp\left(K_1 h^2\right) \right\rceil$$

and $h_0 = 8$ is the lowest fidelity we consider (for the surrogate only estimator we average only $N_3 = 500$ trials). For moderate error tolerances we can achieve an order of magnitude speedup since most of the cost comes from fitting a Laplace approximation; using a very accurate model is not necessary, but using a very cheap surrogate model is insufficient. As the tolerance shrinks, most of the computation shifts to the online sampling phase which begins to dominate and little speedup is obtained. This matches the theoretical speedup derived in Section 3.4.

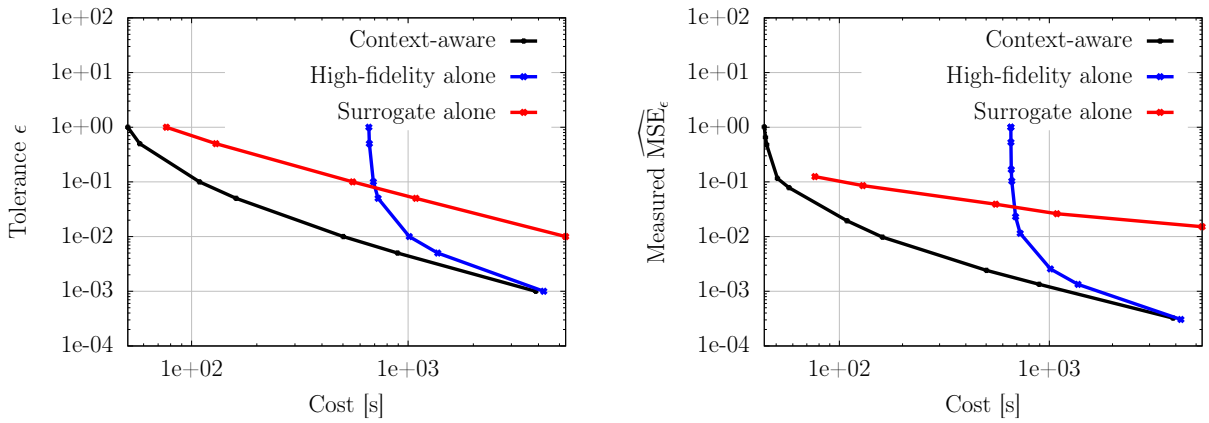

Figure 2: (Left) The theoretical error tolerance ϵ against the total cost (seconds of CPU time) to fit the Laplace approximation q_{h^*} of p_{h^*} and draw m^* samples. (Right) The actual measured $\widehat{\text{MSE}}_\epsilon$ against the total cost. Note that the results shown in the left plot is an upper bound for the results shown in right plot by the bound (3).

5.2 Euler Bernoulli Problem

In the second example we infer the effective stiffness of an Euler Bernoulli beam. The forward-model code is available on GitHub[1] and was developed by Matthew Parno as a part of the 2018 Gene Golub SIAM Summer School on "Inverse Problems: Systematic Integration of Data with Models under Uncertainty". The rest of the setup of this problem is taken from Section 4.2 of [33].

5.2.1 Problem Setup

Let $\Omega = (0,1) \subset \mathbb{R}$ and $\Theta = \mathbb{R}^6$ and consider the PDE

$$\frac{\partial^2}{\partial x^2}\left(E(x;\boldsymbol{\theta})\frac{\partial^2}{\partial x^2}u(x;\boldsymbol{\theta})\right) = g(x), \quad x \in \Omega \tag{38}$$

with boundary conditions

$$u(0;\boldsymbol{\theta}) = 0, \quad \frac{\partial u}{\partial x}(0;\boldsymbol{\theta}) = 0, \quad \frac{\partial^2 u}{\partial x^2}(1;\boldsymbol{\theta}) = 0, \quad \frac{\partial^3 u}{\partial x^3}(1;\boldsymbol{\theta}) = 0$$

where $u : \Omega \times \Theta \to \mathbb{R}$ is the displacement and $E : \Omega \times \Theta \to \mathbb{R}$ is the effective stiffness of the beam. The applied force $g(x)$ is taken to be $g(x) = 1$. The effective stiffness $E(x;\boldsymbol{\theta})$ is a smooth piecewise constant defined in the same way as the log-diffusivity (33) but with θ_i replaced by $|\theta_i|$ for $i = 1, \ldots, 6$. We discretize (38) in the spatial domain Ω with a mesh width $h > 0$ (i.e. $h^{-1} + 1$ grid points) using a second-order finite difference scheme and solve the resulting linear system of equations for the discretized solution u_h at the grid points.

The parameter-to-observable map $\mathcal{F}_h : \Theta \to \mathbb{R}^{40}$ is the linear interpolant of the $h^{-1} + 1$ grid points evaluated at 40 equally spaced points in the spatial domain $(0,1)$

$$(\mathcal{F}_h(\boldsymbol{\theta}))_i = u_h\left(\frac{i-1}{39}\right), \quad i = 1, \ldots, 40$$

Note that we exclude the left end-point at $x = 0$ since it is fixed by the boundary conditions. We set the high-fidelity map to be $\mathcal{F} = \mathcal{F}_H$ with $H^{-1} + 1 = 256$ grid points and for the surrogate maps we again consider $h^{-1} + 1 = 8, 12, 16, \ldots, 64$.

[1] `https://github.cim/g2s3-2018/labs`

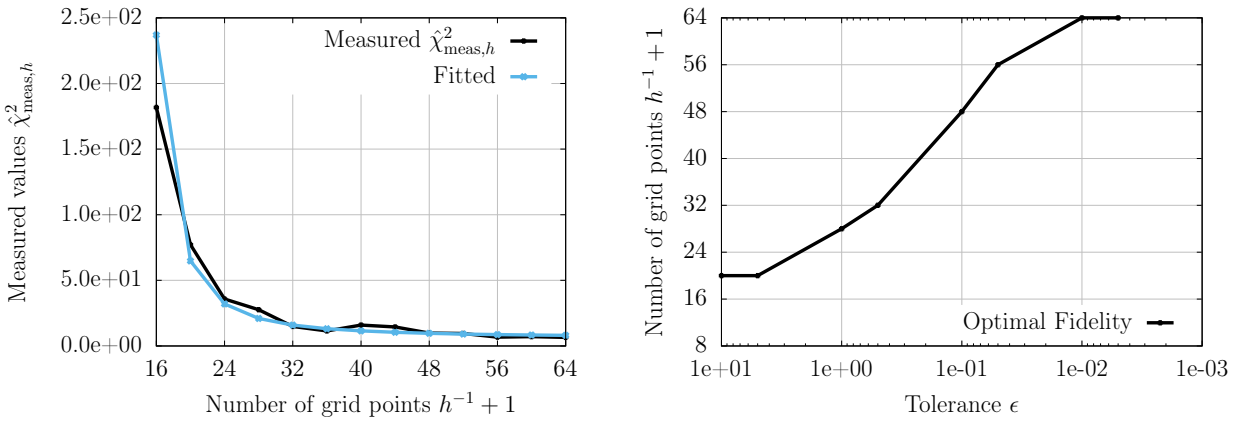

Figure 3: (Left) The measured χ^2 divergences, $\hat{\chi}^2_{\text{meas},h}$, of the high-fidelity posterior p from the Laplace approximation q_h to each surrogate posterior p_h. (Right) The selected fidelity for the number of grid points $(h^*)^{-1}+1$ from the optimization (23) as the tolerance ϵ on the MSE changes.

5.2.2 Setup of the inverse problem

A single observation $\boldsymbol{y} = \mathcal{F}(\boldsymbol{\theta}_{\text{truth}}) + \boldsymbol{\eta} \in \mathbb{R}^{40}$ is generated where $\boldsymbol{\theta}_{\text{truth}} = (1, \dots, 1)^T \in \mathbb{R}^6$ and $\boldsymbol{\eta} \sim N(\mathbf{0},\ \boldsymbol{\Gamma})$ with noise covariance $\boldsymbol{\Gamma} = 5.623 \times 10^{-4} \boldsymbol{I}_{40\times 40}$. The added noise now corresponds to approximately 5% of the true solution u at the right endpoint $x = 1$. The prior is again a Gaussian with mean $\boldsymbol{\mu}_{\text{pr}} = (1, \dots, 1)^T \in \mathbb{R}^6$ and covariance $\boldsymbol{\Sigma}_{\text{pr}} = 1.778 \times 10^{-2} \boldsymbol{I}_{6\times 6} \in \mathbb{R}^{6\times 6}$. For the test function we use the same test function (34) from the steady-state heat problem.

5.2.3 Results

We again fit a Laplace approximation to each surrogate posterior p_h using Newton-CG with the gradient and Hessian computed by second-order finite difference approximations. The total number of model evaluations is $M = 1800$ at each fidelity. The cost function has the form $c(h) = c_0 + c_1/h$ (linear in h^{-1} because the system of linear equations from the discretization is sparse) and the surrogate accuracy has the form $\delta(h) = a_1 h^2$ from the second-order finite difference spatial discretization.

We use the χ^2 divergence estimator $\hat{\chi}^2_{h,10^5}$ from (35) and average the results over $N_1 = 100$ independent trials to obtain the measured value $\hat{\chi}^2_{\text{meas},h}$ as in (35) for each surrogate map $h^{-1}+1 = 8, 12, 16 \dots, 64$. We then use these measured values to fit the curve $\tilde{K}_0 e^{K_1 h^2}$. Figure 3 shows the results. Observe that the χ^2 divergence quickly levels off again.

The fidelity and sample size (h^*, m^*) are found using a brute-force search and Figure 3 shows the selected number of grid points $(h^*)^{-1}+1$ as a function of the MSE tolerance ϵ. When the tolerance is small the selected fidelity is the highest fidelity since we do not consider any surrogate models with $h^{-1}+1$ between 64 and 256. Figure 4 shows the theoretical optimal cost and error trade-off for $\hat{f}_{h^*,m^*}$. We estimated the true value $\mathbb{E}_p[f]$ using $\hat{f}_{H,10^5}$ with $N_2 = 100$ independent trials using equation (36) and the MSE was estimated with $N_3 = 2500$ independent trials using equation (37). Here the lowest-fidelity surrogate model corresponds to $h_0 = 16$. From the plot we can observe an order of magnitude speedup for moderate tolerances where we do not need to use a high-fidelity model to fit the Laplace approximation. Also note that the theoretical trade-off is an upper bound but the shape matches closely with the measured results.

5.3 Advection-Diffusion Problem

In this example, we infer the initial center of a concentration of gas that diffuses throughout a domain with advection. The forward model is a slightly modified version of what is shipped with hIPPYlib [2] [44, 45, 46].

[2] https://hippylib.github.io/tutorials_v3.0.0/4_AdvectionDiffusionBayesian/

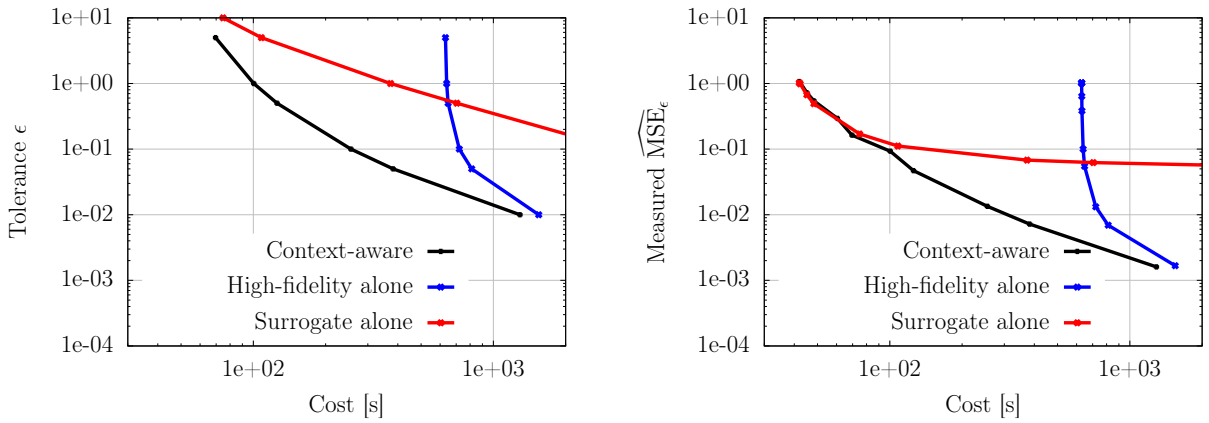

Figure 4: (Left) The theoretical error tolerance ϵ against the total cost (seconds of CPU time) to fit the Laplace approximation q_{h^*} of p_{h^*} and draw m^* samples. (Right) The actual measured $\widehat{\text{MSE}}_\epsilon$ against the total cost. Note that the results in the left plot are upper bounding the results in the right plot by the bound (3).

5.3.1 Problem Setup

Following the setup in hIPPYlib [44, 45, 46], consider the domain

$$\Omega = [0,1]^2 \setminus ([0.25, 0.5] \times [0.15, 0.4] \cup [0.6, 0.75] \times [0.6, 0.85]) \subset \mathbb{R}^2 \,,$$

with two rectangular holes, which is also the parameter domain $\Theta = \Omega$ in this example. Let $u : \Omega \times [0,1] \times \Theta \to \mathbb{R}$ denote the concentration of a gas at position $\boldsymbol{x} \in \Omega$ and time $t \in [0,1]$ and let it be the solution of the following PDE

$$\begin{aligned} \partial_t u(\boldsymbol{x}, t; \boldsymbol{\theta}) - \kappa \Delta u(\boldsymbol{x}, t; \boldsymbol{\theta}) + \boldsymbol{v}(\boldsymbol{x}) \cdot \nabla u(\boldsymbol{x}, t; \boldsymbol{\theta}) = 0, \quad & (\boldsymbol{x}, t) \in \Omega \times [0,1] \,, \\ u(\boldsymbol{x}, 0; \boldsymbol{\theta}) = \mathrm{e}^{-10(x_1 - \theta_1)^2 - 10(x_2 - \theta_2)^2}, \quad & \boldsymbol{x} \in \Omega \,, \\ \kappa \nabla u(\boldsymbol{x}, t; \boldsymbol{\theta}) = \boldsymbol{n}, \quad & (\boldsymbol{x}, t) \in \partial\Omega \times [0,1] \,, \end{aligned} \tag{39}$$

where $\kappa = 10^{-3}$ is the diffusion coefficient, $\boldsymbol{n}$ is the outward unit normal vector from the boundary, and the velocity field $\boldsymbol{v} : \Omega \to \mathbb{R}^2$ is the solution of the steady-state Navier-Stokes equation with the left and right walls driving the flow (see Figure 5)

$$\begin{aligned} -\frac{1}{\mathrm{Re}} \boldsymbol{v}(\boldsymbol{x}) + \nabla q(\boldsymbol{x}) + \boldsymbol{v}(\boldsymbol{x}) \cdot \nabla \boldsymbol{v}(\boldsymbol{x}) = \boldsymbol{0}, \quad & \boldsymbol{x} \in \Omega \,, \\ \nabla \cdot \boldsymbol{v}(\boldsymbol{x}) = 0, \quad & \boldsymbol{x} \in \Omega \,, \\ \boldsymbol{v}(\boldsymbol{x}) = \mathbf{g}(\boldsymbol{x}), \quad & \boldsymbol{x} \in \partial\Omega \,. \end{aligned} \tag{40}$$

Here $\mathrm{Re} = 10^2$ is the Reynold's number of the gas, $q : \Omega \to \mathbb{R}$ is the pressure field, and $\mathbf{g} : \partial\Omega \to \mathbb{R}^2$ is an external force field acting only on the boundary of the domain. In particular, $\mathbf{g}(\boldsymbol{x}) = \mathbf{e}_2$ if $x_1 = 0$ on the left wall, $\mathbf{g}(\boldsymbol{x}) = -\mathbf{e}_2$ if $x_1 = 1$ on the right wall, and $\mathbf{g}(\boldsymbol{x}) = \mathbf{0}$ otherwise, where $\mathbf{e}_2 = (0,1)^\top$ is the second standard basis vector. Note that the parameter dependence is only through the initial condition $u(\boldsymbol{x}, 0; \boldsymbol{\theta})$, where $\boldsymbol{\theta} \in \Omega$ corresponds to the center of the initial concentration. Also note that the velocity field $\boldsymbol{v}$ is independent of the parameters and determined ahead of time.

Because the parameter domain $\Theta = \Omega$ is bounded, Assumptions 2-4 are satisfied. Moreover, the forward model that maps the initial condition to the solution u is differentiable, and so the parameter-to-observable map $\mathcal{F}$ is differentiable as well. Because it is differentiable on a compact domain, $\mathcal{F}$ is also globally Lipschitz and hence the assumptions needed for Theorem 3 apply.

We follow the setup in [44, 45, 46] and discretize (39) in the spatial domain Ω using first order Lagrange finite elements and solve in time using the implicit Euler method to obtain an approximate solution u_h.

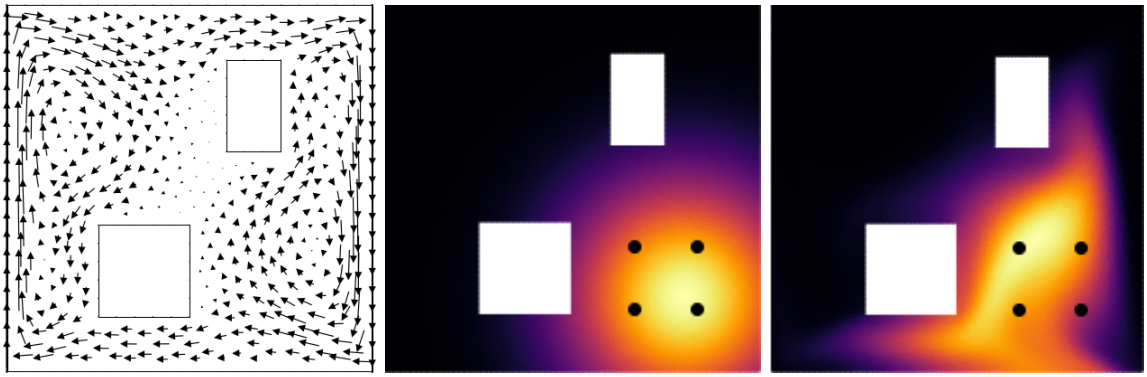

Figure 5: (Left) The velocity field $\boldsymbol{v}$ throughout the domain with the two rectangular barriers. (Middle) The initial concentration $u(\boldsymbol{x}, 0; \boldsymbol{\theta})$ centered at $\boldsymbol{\theta} \in \Theta$. (Right) The concentration $u(\boldsymbol{x}, 1; \boldsymbol{\theta})$ at time $t = 1$ when we observe the solution. In both middle and right plots the observation points are shown as the four black dots in the bottom-right corner.

For the high-fidelity model, the total number of degrees of freedom after discretizing in space is 14,313 and we use a time step size of 10^{-3}. For the surrogate models the total number of degrees of freedom in the discretized system ranges from 20 to 3,661 and we use a time step size of 10^{-2}. The fidelity h here corresponds to the maximum width of a cell in the mesh, which decreases as the number of degrees of freedom (cells) is increased. The parameter-to-observable map is the pointwise observation of the concentration at the final time at four points in the bottom-right quadrant of the domain as in Figure 5, which means that $\mathcal{F}_h(\boldsymbol{\theta}) = (u_h(\boldsymbol{x}_1, 1; \boldsymbol{\theta}) \dots, u_h(\boldsymbol{x}_4, 1; \boldsymbol{\theta}))^\top \in \mathbb{R}^4$ with $\boldsymbol{x}_1 = (2/3,\ 1/6)^\top$, $\boldsymbol{x}_2 = (5/6,\ 1/6)^\top$, $\boldsymbol{x}_3 = (2/3,\ 1/3)$, and $\boldsymbol{x}_4 = (5/6,\ 1/3)^\top$.

5.3.2 Setup of the inverse problem

We generate a single observation $\boldsymbol{y} = \mathcal{F}(\boldsymbol{\theta}_{\text{truth}}) + \eta \in \mathbb{R}^4$, where $\boldsymbol{\theta}_{\text{truth}} = (0.8,\ 0.2)^\top$ and $\eta \sim N(\mathbf{0}, \sigma^2 \boldsymbol{I}_{4\times 4})$ with noise variance $\sigma^2 = 8.876 \times 10^{-3}$; cf. Figure 5. The noise corresponds to 10% of the true solution $u(\boldsymbol{x}, 1; \boldsymbol{\theta})$. The prior is Gaussian with mean $\boldsymbol{\mu}_{\text{pr}} = (0.75,\ 0.25)^\top$ and covariance $\boldsymbol{\Sigma}_{\text{pr}} = 10^{-2} \boldsymbol{I}_{2\times 2}$.

5.3.3 Results

A Laplace approximation to each surrogate posterior p_h is fit using Newton's method where the gradient and Hessian matrix are computed using the chain rule combined with the adjoint method. We select a random initial point in the bottom-right quadrant $\boldsymbol{\theta}_0 \in [0.5,\ 1] \times [0, 0.5]$ and find that 10 Newton iterations, but not fewer, is sufficient for the norm of the gradient of the log posterior to achieve machine precision. Computing the Hessian at each iteration involves 4 linear solves including the forward solve, the adjoint solve, and the forward and adjoint incremental equations. Note that although the PDE (39) is linear, the dependence through the parameters $\boldsymbol{\theta}$ is nonlinear, and thus we must recompute the Hessian at every iteration (i.e. it is not constant). We also note that the forward and adjoint solves are re-used for the computation of the gradient, so that the total number of linear solves needed across the entire offline phase is 40.

To obtain the cost function $c(h)$, we measure the runtime of each surrogate and high fidelity model and average over 10,000 trials. We finally fit a curve of the form $c(h) = c_0 h^{-\beta}$. Similarly for the surrogate error we fit a curve of the form $\delta(h) \propto h^{\alpha}$. We measure the chi-squared divergence using 500,000 samples and then fit the curve of the form $\tilde{K}_0 \mathrm{e}^{K_1 \delta(h)}$ as shown in Figure 6 to be input into the optimization problem (25). Note that the online phase of sampling and re-weighting according to the high-fidelity posterior p is embarrassingly parallel, and so to reduce the computational cost we parallelize over $n_{\text{proc}} = 64$ processors. The optimal fidelity or number of degrees of freedom in this case is given by the solution to the optimization problem

$$\underset{h \geq 0}{\text{minimize}} \quad \frac{4\|f\|_{L^\infty}^2 \tilde{K}_0 C}{\epsilon n_{\text{proc}}} \mathrm{e}^{K_1 \delta(h)} + M c(h)\,, \tag{41}$$

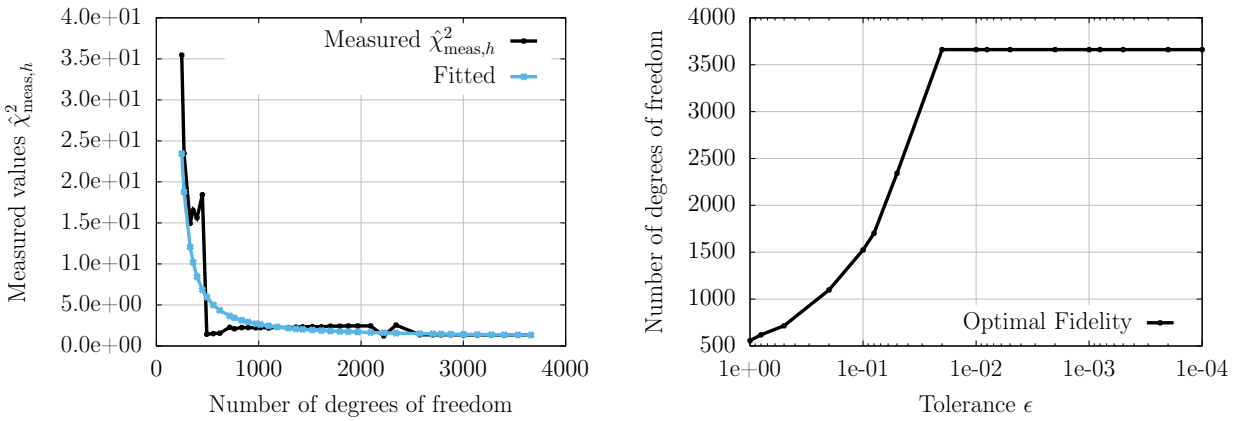

Figure 6: (Left) The measured and fitted values of $\chi^2(q_h \mid\mid p) + 1$ for each Laplace approximation to a surrogate model p_h. (Right) The optimal number of degrees of freedom as given by (41) for different tolerances ϵ.

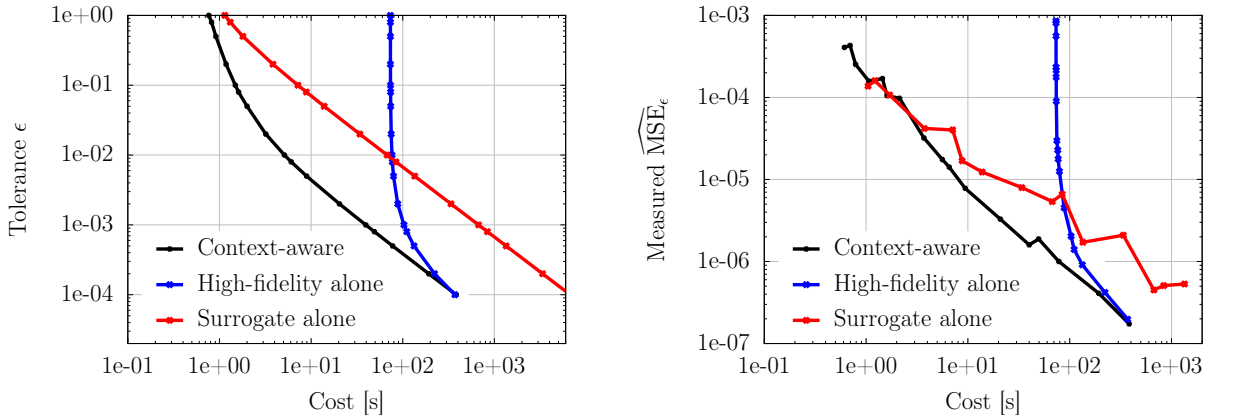

Figure 7: (Left) The upper bound on the mean-squared error vs. the theoretical cost of the entire computational procedure. (Right) The measured mean-squared error of the context-aware importance sampling estimator with the optimal surrogate model vs. cost. For the context-aware estimator the mean-squared error is computed by averaging over 100 independent trials. For the high and low-fidelity estimators the mean-squared error is estimated by averaging over 50 independent trials.

and shown in Figure 6.

Figure 7 shows the speedup predicting by the optimization problem versus the speedup measured numerically after sampling with the computed biasing density and the context-aware importance sampling estimator. The reference value used to compute the mean-squared error was computed using 10^6 samples from the high-fidelity Laplace approximation and then re-weighted. Initially, the context-aware estimator selects a much cheaper surrogate model to achieve a large initial speedup of several orders of magnitude compared to the high-fidelity model for high error tolerances. For smaller tolerances, more accurate surrogate model are selected to optimize the cost-error trade-off. In this regime the context-aware estimator outperforms the estimators that use low-fidelity surrogate models alone due to the chi-squared divergence being lower, allowing for fewer necessary samples. Note that the results in plot on the left correspond an upper bound and are independent of the choice of test function, and that for different test functions the actual MSE may be lower. However, the trend of how the cost-error ratio behaves for different tolerances is comparable, which demonstrates the viability of the proposed approach.

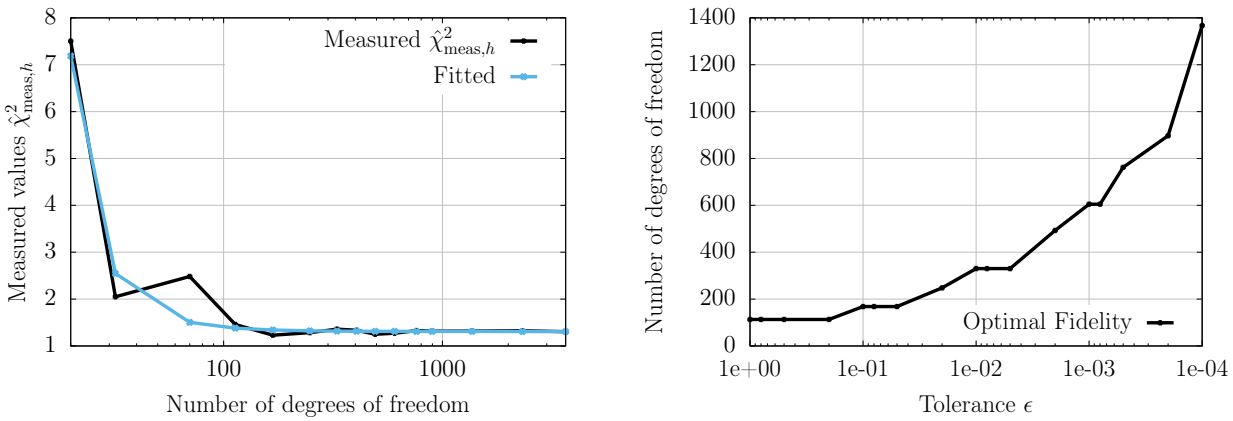

Figure 8: (Left) The measured and fitted values of $\chi^2(q_h \mid\mid p) + 1$ for each Laplace approximation to a surrogate model p_h. (Right) The optimal number of degrees of freedom as given by (41) for different tolerances ϵ.

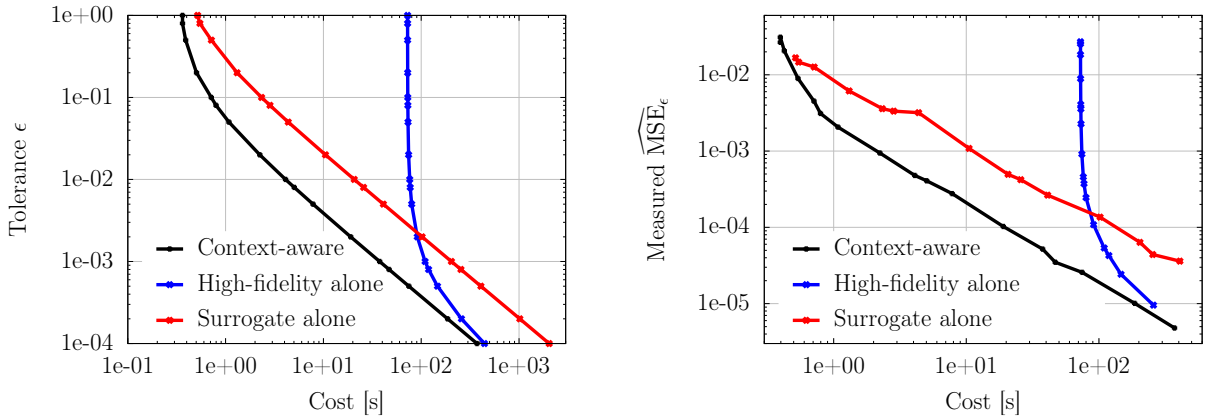

Figure 9: (Left) The upper bound on the mean-squared error vs. the theoretical cost of the entire computational procedure. (Right) The measured mean-squared error of the context-aware importance sampling estimator with the optimal surrogate model vs. cost. For the context-aware estimator the mean-squared error is computed by averaging over 100 independent trials. For the high and low-fidelity estimators the mean-squared error is estimated by averaging over 50 independent trials.

5.3.4 Extension to 12-dimensional parameter

Instead of inferring the origin of a single initial concentration of gas, here we infer the origin of six initial concentrations giving rise to a 12-dimensional parameter. We now have

$$u(\boldsymbol{x}, 0; \boldsymbol{\theta}) = \sum_{i=1}^{6} \mathrm{e}^{-10(x_1-\theta_{2i-1})^2-10(x_2-\theta_{2i})^2}, \quad \boldsymbol{x} \in \Omega\,.$$

The forward model is the same except that we add four additional sensors in the top corner of the domain for observing data so that $\mathcal{F} : \Omega^6 \subset \mathbb{R}^{12} \to \mathbb{R}^8$. We also increase the relative noise in the observations to 20%. We again take a Gaussian prior with covariance $\Sigma_{\text{pr}} = 4 \times 10^{-3}\boldsymbol{I}$. A reference value of the posterior mean was computed using 10^6 importance-weighted samples from the Laplace approximation to the high-fidelity posterior. The rest of the setup is the same as in previous subsections. Figures 8 and 9 show that the context-aware estimator outperforms both the estimator that uses the high-fidelity alone and the estimator that only uses the low-fidelity model for constructing the biasing density.

Acknowledgements

The authors acknowledge support of the National Science Foundation under Grant No. 1761068 and Grant No. 1901091. The first author was supported in part by the Research Training Group in Modeling and Simulation funded by the National Science Foundation via grant RTG/DMS – 1646339. The second author also acknowledges support from the AFOSR under Award Number FA9550-21-1-0222 (Dr. Fariba Fahroo).

References

[1] S. Agapiou, O. Papaspiliopoulos, D. Sanz-Alonso, and A. Stuart. Importance sampling: Intrinsic dimension and computational cost. *Statist. Sci.*, 32(3):405–431, 2017.

[2] Ö. Akyildiz and J. Míguez. Convergence rates for optimised adaptive importance samplers. *Statistics and Computing*, 31(12), 2021.

[3] W. A. Al-Qaq, M. Devetsikiotis, and J. K. Townsend. Stochastic gradient optimization of importance sampling for the efficient simulation of digital communication systems. *IEEE Transactions on Communications*, 43(12):2975–2985, 1995.

[4] A. Antoulas. *Approximation of Large-Scale Dynamical Systems*. Advances in Design and Control. SIAM, 2005.

[5] P. Benner, S. Gugercin, and K. Willcox. A survery of projection-based model reduction for parametric dynamical systems. *SIAM Rev.*, 57(4):483–531, 2015.

[6] A. Beskos, A. Jasra, K. Law, R. Tempone, and Y. Zhou. Multilevel sequential Monte Carlo samplers. *Stochastic Processes and their Applications*, 127(5):1417–1440, 2017.

[7] S. Brenner and R. Scott. *The Mathematical Theory of Finite Element Methods*. Texts in Applied Mathematics. Springer-Verlag New York, 2008.

[8] Y. Cao, M. Gunzburger, F. Hua, and X. Wang. Analysis and finite element approximation of a coupled, continuum pipe-flow/Darcy model for flow in porous media with embedded conduits. *Numerical Methods Partial Differential Equations*, 27:1242–1252, 2011.

[9] S. Chatterjee and P. Diaconis. The sample size required in importance sampling. *Ann. Appl. Probab.*, 28(2):1099–1135, 2018.

[10] P. Chen and A. Quarteroni. Accurate and efficient evaluation of failure probability for partial differential equations with random input data. *Computer Methods in Applied Mechanics and Engineering*, 267:233–260, 2013.

[11] P. Chen, A. Quarteroni, and G. Rozza. Reduced basis methods for uncertainty quantification. *SIAM/ASA J. Uncertain. Quantif.*, 5:813–869, 2017.

[12] T. Cui, Y. M. Marzouk, and K. E. Willcox. Data-driven model reduction for the Bayesian solution of inverse problems. *International Journal for Numerical Methods in Engineering*, 102(5):966–990, 2015.

[13] A. Davis, Y. Marzouk, A. Smith, and N. Pillai. Rate-optimal refinement strategies for local approximation MCMC. *Statistics and Computing*, 32(60), 2022.

[14] G. P. Dehaene. Computing the quality of the Laplace approximation. In *Neural Information Processing Systems*, 2017.

[15] G. Detommaso, T. Dodwell, and R. Scheichl. Continuous level Monte Carlo and sample-adaptive model hierarchies. *SIAM/ASA Journal on Uncertainty Quantification*, 7(1):93–116, 2019.

[16] I.-G. Farcas. *Context-aware Model Hierarchies for Higher-dimensional Uncertainty Quantification*. Dissertation, Technische Universität München, München, 2020.

[17] A. Forrester and A. Keane. Recent advances in surrogate-based optimization. *Progr. Aerosp. Sci.*, 45:50–79, 2009.

[18] M. Heinkenschloss, B. Kramer, and T. Takhtaganov. Adaptive reduced-order model construction for conditional value-at-risk estimation. *SIAM/ASA Journal on Uncertainty Quantification*, 8(2):668–692, 2020.

[19] M. Heinkenschloss, B. Kramer, T. Takhtaganov, and K. Willcox. Conditional-value-at-risk estimation via reduced-order models. *SIAM/ASA Journal on Uncertainty Quantification*, 6(4):1395–1423, 2018.

[20] J. S. Hesthaven, G. Rozza, and B. Stamm. *Certified Reduced Basis Methods for Parametrized Partial Differential Equations*. SpringerBriefs in Mathmatics. Springer International Publishing, 2016.

[21] H. Hoel, E. von Schwerin, A. Szepessy, and R. Tempone. Adaptive multilevel Monte Carlo simulation. In B. Engquist, O. Runborg, and Y.-H. R. Tsai, editors, *Numerical Analysis of Multiscale Computations*, pages 217–234, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg.

[22] H. Hoel, E. von Schwerin, A. Szepessy, and R. Tempone. Implementation and analysis of an adaptive multilevel Monte Carlo algorithm. *Monte Carlo Methods and Applications*, 20(1):1 – 41, 2014.

[23] J. Lang, R. Scheichl, and D. Silvester. A fully adaptive multilevel stochastic collocation strategy for solving elliptic PDEs with random data. *Journal of Computational Physics*, 419:109692, 2020.

[24] J. Latz, I. Papaioannou, and E. Ullmann. Multilevel sequential^2 Monte Carlo for Bayesian inverse problems. *Journal of Computational Physics*, 368:154–178, 2018.

[25] R. J. LeVeque. *Finite Difference Methods for Ordinary and Partial Differential Equations: Steady-State and Time-Dependent Problems*. SIAM, 2007.

[26] J. Li, J. Li, and D. Xiu. An efficient surrogate-based method for computing rare failure probability. *Journal of Computational Physics*, 230(24):8683–8697, 2011.

[27] J. Li and D. Xiu. Evaluation of failure probability via surrogate models. *Journal of Computational Physics*, 229(3):8966–8980, 2010.

[28] A. J. Majda and G. Gershgorin. Quantifying uncertainty in climate change science through empirical information theory. *Proceedings of the National Academy of Sciences of the United States of America*, 107(34):14958–14963, 2010.

[29] A. Narayan, C. Gittelson, and D. Xiu. A stochastic collocation algorithm with multifidelity models. *SIAM J. Sci. Comput.*, 36:495–521, 2014.

[30] L. W. Ng and K. Willcox. Monte-Carlo information-reuse approach to aircraft conceptual design optimization under uncertainty. *Journal of Aircraft*, 53:427–438, 2016.

[31] B. Peherstorfer. Multifidelity Monte Carlo estimation with adaptive low-fidelity models. *SIAM/ASA Journal on Uncertainty Quantification*, 7:579–603, 2019.

[32] B. Peherstorfer, T. Cui, Y. Marzouk, and K. Willcox. Multifidelity importance sampling. *Computer Methods in Applied Mechanics and Engineering*, 300:490–509, 2016.

[33] B. Peherstorfer and Y. Marzouk. A transport-based multifidelity preconditioner for Markov chain Monte Carlo. *Advances in Computational Mathematics*, 45:2321–2348, 2019.

[34] B. Peherstorfer, K. Willcox, and M. Gunzburger. Survery of multifidelity methods in uncertainty propagation, inference, and optimization. *SIAM Review*, 60(3):550–591, 2018.

[35] A. Quarteroni, G. Rozza, and A. Manzoni. Certified reduced basis approximation for parametrized partial differential equations and applications. *Journal of Mathematics in Industry*, 1(1):1–49, 2011.

[36] C. Rasmussen and C. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2016.

[37] E. Ryu and S. Boyd. Adaptive importance sampling via stochastic convex programming. *pre-print*, 2014. arXiv:1412.4845.

[38] D. Sanz-Alonso. Importance sampling and necessary sample size: An information theory approach. *SIAM/ASA J. Uncertainty Quantification*, 6(2):867–879, 2018.

[39] C. Schillings, B. Sprungk, and P. Wacker. On the convergence of the Laplace approximation and noise-level-robustness of Laplace-based Monte Carlo methods for Bayesian inverse problems. *Numer. Math.*, 145:915–971, 2020.

[40] A. Stuart. Inverse problems: A Bayesian perspective. *Acta Numerica*, 19:451–559, 2010.

[41] T. Sullivan. *Introduction to Uncertainty Quantification*. Springer, 2015.

[42] A. B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer, 2009.

[43] R. Vershynin. *High-dimensional probability: an introduction with applications in data science*. Cambridge University Press, 2018.

[44] U. Villa, N. Petra, and O. Ghattas. Documentation to "hIPPYlib: an Extensible Software Framework for Large-scale Deterministic and Bayesian Inverse Problems", 2016. http://hippylib.github.io.

[45] U. Villa, N. Petra, and O. Ghattas. hIPPYlib: an Extensible Software Framework for Large-scale Deterministic and Bayesian Inverse Problems. *Journal of Open Source Software*, 3(30), 2018.

[46] U. Villa, N. Petra, and O. Ghattas. HIPPYlib: An Extensible Software Framework for Large-Scale Inverse Problems Governed by PDEs: Part I: Deterministic Inversion and Linearized Bayesian Inference. *ACM Trans. Math. Softw.*, 47(2), Apr. 2021.

A Proof of Lemma 1

Proof. Suppose that $\boldsymbol{x}$ is a sub-Gaussian random vector and consider the matrix to be a multiple of the identity, $\boldsymbol{A} = \alpha \boldsymbol{I}$ with $\alpha > 0$. We now only need to show that there exists an $\alpha > 0$ such that for all $\boldsymbol{\mu} \in \mathbb{R}^d$

$$\mathbb{E}_\pi \left[\exp\left(\alpha \|\boldsymbol{x} - \boldsymbol{\mu}\|^2 \right) \right] = \mathbb{E}_\pi \left[\exp\left((\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{A} (\boldsymbol{x} - \boldsymbol{\mu}) \right) \right] < \infty \,.$$

Since $\|\boldsymbol{v} + \boldsymbol{w}\|^2 \leq 2\|\boldsymbol{v}\|^2 + 2\|\boldsymbol{w}\|^2$ by the triangle inequality and the fact that $(a+b)^2 \leq 2a^2 + 2b^2$, we get the upper bound

$$\mathbb{E}_\pi \left[\exp\left(\alpha \|\boldsymbol{x} - \boldsymbol{\mu}\|^2 \right) \right] \leq \mathbb{E}_\pi \left[\exp\left(2\alpha \|\boldsymbol{\mu}\|^2 + 2\alpha \|\boldsymbol{x}\|^2 \right) \right] = \exp\left(2\alpha \|\boldsymbol{\mu}\|^2 \right) \mathbb{E}_\pi \left[\exp\left(2\alpha \|\boldsymbol{x}\|^2 \right) \right] \,.$$

Therefore, we only need to find $\alpha > 0$ such that

$$\mathbb{E}_\pi \left[\exp\left(2\alpha \|\boldsymbol{x}\|^2 \right) \right] < \infty \,.$$

We now use the assumption that $\boldsymbol{x}$ is sub-Gaussian by taking the marginals

$$\begin{aligned}
\mathbb{E}_\pi \left[\exp\left(2\alpha \|\boldsymbol{x}\|^2 \right) \right] &= \mathbb{E}_\pi \left[\exp\left(2\alpha \sum_{i=1}^d x_i^2 \right) \right] \\
&= \mathbb{E}_\pi \left[\exp\left(2\alpha \sum_{i=1}^d |\boldsymbol{e}_i^T \boldsymbol{x}|^2 \right) \right] \\
&= \mathbb{E}_\pi \left[\prod_{i=1}^d \exp\left(2\alpha |\boldsymbol{e}_i^T \boldsymbol{x}|^2 \right) \right] \,,
\end{aligned}$$

where $\boldsymbol{e}_i$ is the i-th canonical unit vector. We proceed by induction on the dimension d and repeatedly use the Cauchy-Schwarz inequality to show that this expectation is finite. When $d = 1$, take α_1 such that $\frac{1}{\sqrt{2\alpha_1}} > \|\boldsymbol{x}\|_{\psi_2}$ so that

$$\mathbb{E}_\pi\left[\exp\left(2\alpha_1|\boldsymbol{e}_1^T\boldsymbol{x}|^2\right)\right] = \mathbb{E}_\pi\left[\exp\left(\frac{|\boldsymbol{e}_1^T\boldsymbol{x}|^2}{(1/\sqrt{2\alpha_1})^2}\right)\right] \leq 2\,.$$

Note that since $\boldsymbol{x}$ is sub-Gaussian $\|\boldsymbol{x}\|_{\psi_2} < \infty$ we can indeed find an $\alpha_1 > 0$ to satisfy the inequality. Now suppose that for dimension $d-1$ there exists an α_{d-1} such that

$$\mathbb{E}_\pi\left[\prod_{i=1}^{d-1}\exp\left(2\alpha_{d-1}|\boldsymbol{e}_i^T\boldsymbol{x}|^2\right)\right] = C_{d-1} < \infty\,.$$

By using the Cauchy-Schwarz inequality, we get that

$$\mathbb{E}_\pi\left[\prod_{i=1}^{d}\exp\left(2\alpha_d|\boldsymbol{e}_i^T\boldsymbol{x}|^2\right)\right] \leq \mathbb{E}_\pi\left[\prod_{i=1}^{d-1}\exp\left(4\alpha_d|\boldsymbol{e}_i^T\boldsymbol{x}|^2\right)\right]^{1/2}\mathbb{E}_\pi\left[\exp\left(4\alpha_d|\boldsymbol{e}_d^T\boldsymbol{x}|^2\right)\right]^{1/2}.$$

Taking $\alpha_d \leq \alpha_{d-1}/2$ gives

$$\mathbb{E}_\pi\left[\prod_{i=1}^{d-1}\exp\left(4\alpha_d|\boldsymbol{e}_i^T\boldsymbol{x}|^2\right)\right]^{1/2} \leq \mathbb{E}_\pi\left[\prod_{i=1}^{d-1}\exp\left(2\alpha_{d-1}|\boldsymbol{e}_i^T\boldsymbol{x}|^2\right)\right]^{1/2} = C_{d-1}^{1/2}\,.$$

Taking α_d such that $\frac{1}{\sqrt{4\alpha_d}} > \|\boldsymbol{x}\|_{\psi_2}$ gives

$$\mathbb{E}_\pi\left[\exp\left(4\alpha_d|\boldsymbol{e}_d^T\boldsymbol{x}|^2\right)\right]^{1/2} \leq \mathbb{E}_\pi\left[\exp\left(\frac{|\boldsymbol{e}_d^T\boldsymbol{x}|^2}{(1/\sqrt{4\alpha_d})^2}\right)\right]^{1/2} \leq \sqrt{2}\,.$$

Thus, take $\alpha_d < \frac{1}{4}\min\{2\alpha_{d-1},\ \|\boldsymbol{x}\|_{\psi_2}^{-2}\}$, so that

$$\mathbb{E}_\pi\left[\prod_{i=1}^{d}\exp\left(2\alpha_d|\boldsymbol{e}_i^T\boldsymbol{x}|^2\right)\right] \leq \sqrt{2C_{d-1}} < \infty\,.$$

Since the dimension is finite, we know that we will always be able to take $\alpha_d > 0$. Setting $\alpha = \alpha_d$, shows the first direction of the lemma.

For the converse suppose that there exists a symmetric positive-definite matrix $\boldsymbol{A} \succ 0$ so that for all vectors $\boldsymbol{\mu}$

$$\mathbb{E}_\pi\left[\exp\left((\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{A}(\boldsymbol{x}-\boldsymbol{\mu})\right)\right] < \infty\,.$$

In particular, for $\boldsymbol{\mu} = 0$

$$\mathbb{E}_\pi\left[\exp\left(\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}\right)\right] = C < \infty\,.$$

For any $\boldsymbol{v} \in S^{d-1}$, we have that

$$\mathbb{E}_\pi\left[\exp\left(\frac{|\boldsymbol{v}^T\boldsymbol{x}|^2}{t^2}\right)\right] \leq \mathbb{E}_\pi\left[\exp\left(\frac{\|\boldsymbol{x}\|^2}{t^2}\right)\right],$$

since $|\boldsymbol{v}^T\boldsymbol{x}| \leq \|\boldsymbol{v}\|\|\boldsymbol{x}\|$. Also, since the minimum eigenvalue satisfies $\lambda_{\min}^{\boldsymbol{A}} \leq \frac{\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}}{\|\boldsymbol{x}\|^2}$ for all $\boldsymbol{x} \neq 0$, we get

$$\mathbb{E}_\pi\left[\exp\left(\frac{\|\boldsymbol{x}\|^2}{t^2}\right)\right] \leq \mathbb{E}_\pi\left[\exp\left(\frac{\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}}{\lambda_{\min}^{\boldsymbol{A}}t^2}\right)\right] = \mathbb{E}_\pi\left[\left\{\exp\left(\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}\right)\right\}^{1/\lambda_{\min}^{\boldsymbol{A}}t^2}\right].$$

If $\lambda_{\min}^{\boldsymbol{A}}t^2 > 1$, then the function

$$g(x) = x^{1/(\lambda_{\min}^{\boldsymbol{A}}t^2)}$$

is concave and increasing in x. By Jensen's inequality, we obtain

$$\mathbb{E}_\pi\left[\left\{\exp\left(\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}\right)\right\}^{1/\lambda_{\min}^{\boldsymbol{A}}t^2}\right] \leq \mathbb{E}_\pi\left[\exp\left(\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}\right)\right]^{1/\lambda_{\min}^{\boldsymbol{A}}t^2} = C^{1/\lambda_{\min}^{\boldsymbol{A}}t^2}\,.$$

Setting $C^{1/\lambda_{\min}^{\boldsymbol{A}}t^2} \leq 2$ and solving for t gives

$$t \geq \sqrt{\frac{\log C}{\lambda_{\min}^{\boldsymbol{A}}\log 2}}\,.$$

Since this inequality holds for every $\boldsymbol{v} \in S^{d-1}$ we know that $\|\boldsymbol{x}\|_{\psi_2} < \infty$ and hence $\boldsymbol{x}$ is sub-Gaussian. □