BIMC: The Bayesian Inverse Monte Carlo method for goal-oriented uncertainty quantification

Siddhant Wahal* and George Biros*

Abstract. We consider the problem of estimating rare event probabilities, focusing on systems whose evolution is governed by differential equations with uncertain input parameters. If the system dynamics is expensive to compute, standard sampling algorithms such as the Monte Carlo method may require infeasible running times to accurately evaluate these probabilities. We propose an importance sampling scheme (which we call *"BIMC"*) that relies on solving an auxiliary, *"fictitious"* Bayesian inverse problem. The solution of the inverse problem yields a posterior PDF, a local Gaussian approximation to which serves as the importance sampling density. We apply BIMC to several problems and demonstrate that it can lead to computational savings of several orders of magnitude over the Monte Carlo method. We delineate conditions under which BIMC is optimal, as well as conditions when it can fail to yield an effective IS density.

Key words. Monte Carlo method, Bayesian inference, rare events, importance sampling, uncertainty quantification

AMS subject classifications. 65C05, 62F15, 62P30

1. Introduction.

We consider the following goal-oriented uncertainty quantification (UQ) problem. Let $f(\boldsymbol{x}) : \mathbb{R}^m \to \mathbb{R}$ be a *smooth nonlinear operator*, and $p(\boldsymbol{x})$ a given probability density function (PDF) for $\boldsymbol{x}$. Given a *target interval* $\mathbb{Y} \subset \mathbb{R}$, our goal is to compute $\mu = \mathbb{P}(f(\boldsymbol{x}) \in \mathbb{Y})$. Equivalently, μ is the expectation under $p(\boldsymbol{x})$ of the *indicator function* $\mathbf{1}_{\mathbb{Y}}(f(\boldsymbol{x}))$. [1] We focus on the case when $\mu \ll 1$, *i.e.,* the event $f(\boldsymbol{x}) \in \mathbb{Y}$ is rare.

In our context, $f(\boldsymbol{x})$ is a map from some random finite dimensional parameter space to a quantity-of-interest (QoI). Such parameter-to-QoI maps are often a composition of the solution of a differential equation for a state variable, and an operator that extracts the QoI from the state. The parameters $\boldsymbol{x}$ represent uncertain parameters in the physical model. This uncertainty can arise from a variety of sources, such as lack of knowledge, measurement errors, or noise. Here, we assume that the uncertainty is described by a known PDF, $p(\boldsymbol{x})$. A *Monte Carlo* (MC) method can be used to compute μ by sampling $\boldsymbol{x}$ from $p(\boldsymbol{x})$ and then checking whether $f(\boldsymbol{x}) \in \mathbb{Y}$. But such an approach can be prohibitively expensive if the operator f is expensive to evaluate, especially when $\mu \ll 1$.

Summary of the methodology. We propose a variance reduction scheme based on importance sampling (IS). In IS, samples are drawn from a new distribution, say $q(\boldsymbol{x})$, in order to increase the occurrences of the rare event. We construct our IS density as follows. We begin by setting up an auxiliary *inverse problem*. First we select a $y \in \mathbb{Y}$, and then we find $\boldsymbol{x}$ such that $f(\boldsymbol{x}) \approx y$. This is an ill-posed or inverse problem since given a scalar y we want to reconstruct the vector $\boldsymbol{x}$. A simple counting argument shows that this is impossible unless we use some kind of regularization. To address this ill-posedness we adopt a Bayesian perspective,

* Oden Institute for Computational Engineering and Sciences, The University of Texas at Austin, Austin, TX, 78712, (siddhant@ices.utexas.edu, gbiros@acm.org).

[1] The indicator function, $\mathbf{1}_{\mathbb{Y}}(y)$ assumes the value 1 if $y \in \mathbb{Y}$, and 0 otherwise.

that is, the solution of the inverse problem is not a specific point estimate $\boldsymbol{x}$ but a "posterior distribution", $p(\boldsymbol{x}|y)$, a PDF on the parameters $\boldsymbol{x}$ conditioned on y. We will use a Gaussian approximation of this posterior around the *Maximum A Posteriori* (MAP) point as the importance sampling distribution. The mean of the approximating Gaussian is the MAP point itself, and its covariance is the inverse of the Gauss-Newton Hessian, $\mathbf{H}_{\text{GN}}^{-1}$, of $-\log(p(\boldsymbol{x}|y))$ at the MAP point.

Contributions. In summary, our contributions are the following.

- We introduce the concept of solving *inverse* problems for *forward* uncertainty quantification.
- To our knowledge, this is the first algorithm that exploits derivatives of the forward operator f to arrive at an IS density for simulating rare events.
- We offer a thorough theoretical analysis of the affine-Gaussian inverse problem. This analysis establishes conditions for optimality of our algorithm, as well as guides the tuning of various algorithmic "knobs".
- We apply our methodology to several real and synthetic problems and demonstrate orders-of-magnitude speedup over a vanilla MC implementation.

Limitations.

- The success of our algorithms depends strongly upon the quality (both in terms of accuracy and speed) of the inverse problem solution. When the operator f involves differential equations, efficiently solving the inverse problem requires adjoint operators and perhaps sophisticated PDE-constrained optimization solvers and preconditioners.
- Our methodology has several failure mechanisms. These are described in detail in Section 4. In light of these failure mechanisms, the question of *a priori* assessing the applicability of BIMC to a given problem (*i.e.*, a given combination of $f(\boldsymbol{x})$, $p(\boldsymbol{x})$, and $\mathbb{Y}$) has also been left unexplored.

Related work. The literature on goal-oriented techniques, importance sampling, rare-event probability estimation, and Bayesian inference is quite extensive. Here, we review work that is most relevant.

Goal-oriented methods. The idea of goal-oriented techniques for UQ isn't new (see [16, 17, 28]). However, most of these works focus on dimensionality reduction, and not rare events. Also pertinent is the measure-theoretic approach to inverse problems [5, 7].

Rare-event probability estimation. A large body of work on rare events has been motivated by the problem of assessing the reliability of systems. In such problems, the task is to compute the probability of failure of a system, which occurs when $f(\boldsymbol{x}) < 0$ (or in our framework, when $\mathbb{Y} = (-\infty, 0)$).

Analytical approaches to approximate this failure probability include the First and Second Order Reliability Methods (see [23] for a review). These methods are based on approximating f with a truncated Taylor series expansion around a "design" point. A drawback of these methods is that they have no means of estimating the error in the computed failure probability. We would like to note that the concept of "design" points here is similar to the MAP point in our algorithm, but they are not exactly identical. The design point, say $\boldsymbol{x}^*$, is always constrained to satisfy $f(\boldsymbol{x}^*) = 0$. That is, it lies at the edge of the pre-image $f^{-1}(\mathbb{Y})$. The MAP point, on the other hand, is expected to lie in the interior of the region $f^{-1}(\mathbb{Y})$. Moreover, the fact that $\boldsymbol{x}_{\text{MAP}}$ lies in the interior of $f^{-1}(\mathbb{Y})$ is accounted for, and in fact, exploited, when we choose

tunable parameters of our algorithm.

Statistical approaches to evaluate the failure probability have received considerable attention (see [26] for a review). As opposed to analytical methods, these methods have well-understood convergence properties, and they come with a natural error estimate. In this context, a simple Monte Carlo method is usually inefficient, and some form of variance reduction is usually required. Several importance sampling methods have been proposed to this effect. We refer the reader to [20] for a general introduction to importance sampling. Several IS algorithms ([27, 6, 19, 3]) reuse the concept of design points by placing normal distributions centered there. In [6, 27], the covariance of the IS distribution is either set equal to that of $p(\boldsymbol{x})$, or evaluated heuristically, for example, from samples. In our method, approximating the posterior via a Gaussian yields a natural covariance for the IS density.

Within reliability analysis, another class of algorithms uses surrogate models to reduce the computational effort required to build an IS density [21, 22, 15]. A different approach involves simulating a sequence of relatively higher frequency events to arrive at the rare event probability. This idea is used in the Cross Entropy algorithm [8] to arrive at an optimal IS distribution within a parametric family. It has also been coupled with Markov Chain Monte Carlo methods for high-dimensional reliability problems [2, 14, 4].

A common feature of all these algorithms is that they only use pointwise evaluations of the forward model f (or its low-fidelity surrogates) to arrive at the IS distribution. Hence, these methods are "non-intrusive". On the other hand, the manner in which we construct our IS density naturally endows it with information from derivatives of f. To our knowledge, the only other algorithms that utilize derivative information to construct IS densities are IMIS and LIMIS [24, 9]. However, these aren't tailored for rare-event simulation. Directly substituting the zero-variance (zero-error) IS density (see Subsection 2.2) for the target distribution in these algorithms wouldn't work, since the zero-variance density is non-differentiable, owing to the presence of the characteristic function.

PDE-constrained optimization and Bayesian inverse problems. In BIMC, we rely on adjoints to compute gradients and Hessians of $-\log p(\boldsymbol{x}|y)$. We refer to [10] for an introduction to the method of adjoints. Computing the MAP point is a PDE-constrained optimization problem which can require sophisticated algorithms [1]. Scalable algorithms for characterizing the Hessian of $-\log(p(\boldsymbol{x}|y))$ are described in [12]. Because we construct our IS density through the solution of an inverse problem, our approach can be easily built on top of existing scalable frameworks for solving Bayesian inverse problems, such as [29].

Outline of the paper. The rest of this paper is organized as follows - Table 1 introduces the notation adopted in this paper. Section 2 provides introductions to the Monte Carlo method, importance sampling, as well as Bayesian inference. In Section 3, we describe our algorithm. including analysis that governs the choice of tunable parameters that arise in the algorithm. Section 4 contains numerical experiments and their results, as well as a description of the failure mechanisms of our method. Finally, we summarize our conclusions in Section 5.

2. Background.

2.1. The Monte Carlo Method. One way to compute the rare-event probability, μ, is using the Monte Carlo method. The forward operator is applied on N independent, identically distributed (i.i.d.) samples from $p(\boldsymbol{x})$, $\{\boldsymbol{x}_i\}_{i=1}^N$. Then, an unbiased estimate of μ is:

Symbol	Meaning
f	The input-output, or the forward, map
$\boldsymbol{x}$	Vector of input parameters to f
$p(\boldsymbol{x})$	Input probability density for $\boldsymbol{x}$
$\mathbb{Y}$	Target interval for $f(\boldsymbol{x})$
$\mathbb{P}(f(\boldsymbol{x}) \in \mathbb{Y})$	Probability of the event $f(\boldsymbol{x}) \in \mathbb{Y}$
μ	$\mathbb{P}\left(f(\boldsymbol{x}) \in \mathbb{Y}\right)$
$\mathcal{N}\left(\boldsymbol{x}_0, \boldsymbol{\Sigma}_0\right)$	Normal distribution with mean $\boldsymbol{x}_0$ and covariance $\boldsymbol{\Sigma}_0$
N	Number of Monte Carlo (MC) or Importance Sampling (IS) samples
$\hat{\mu}^N$	MC estimate for μ computed using N samples
$\tilde{\mu}^N$	IS estimate for μ computed using N samples
$\hat{e}_{\text{RMS}}$	Root Mean Square (RMS) error in $\hat{\mu}^N$
$\tilde{e}_{\text{RMS}}$	RMS error in $\tilde{\mu}^N$
$p(y\|\boldsymbol{x})$	The likelihood density
$p(\boldsymbol{x}\|y)$	The posterior density
$\boldsymbol{x}_{\text{MAP}}$	The Maximum *A Posteriori* (MAP) point of $p(\boldsymbol{x}\|y)$
$\mathbf{H}_{\text{GN}}$	The Gauss-Newton Hessian of $-\log p(\boldsymbol{x}\|y)$
$D_{\text{KL}}(p\|\|q)$	The Kullback-Leibler divergence between densities p and q

Table 1: *Summary of key notation used in this paper.*

$$\hat{\mu}^N = \frac{\sum_{i=1}^N \mathbf{1}_{\mathbb{Y}}(f\left(\boldsymbol{x}_i\right))}{N}. \tag{1}$$

The law of large numbers guarantees that in the limit $N \to \infty$, $\hat{\mu}$ converges to μ [25]. The relative Root Mean Square Error (RMSE) in $\hat{\mu}$ is:

$$\hat{e}_{\text{RMS}} = \frac{1}{\mu}\sqrt{\mathbb{E}_p\Big((\hat{\mu}^N - \mu)^2\Big)} = \frac{1}{\mu}\sqrt{\frac{\mathbb{V}_p\Big(\mathbf{1}_{\mathbb{Y}}(f(\boldsymbol{x}))\Big)}{N}} = \sqrt{\frac{\sigma_p^2}{\mu^2 N}}.$$

Since $\mathbf{1}_{\mathbb{Y}}(f(\boldsymbol{x}))$ is a binary random variable, its variance is $\sigma_p^2 = \mu(1-\mu)$. This implies that the relative RMSE is approximately $\sqrt{1/\mu N}$ when $\mu \ll 1$. In order to achieve a specified relative accuracy threshold, the number of samples must then scale as $N \sim 1/\mu$. This is problematic since it can render evaluating extremely rare probabilities virtually impossible if $f(\boldsymbol{x})$ is expensive to evaluate. The evaluation of rare probabilities can be made tractable by reducing the variance of the MC estimate. In BIMC, we aim to achieve variance reduction through importance sampling, which is briefly introduced in the next section.

2.2. Importance Sampling. Importance sampling biases samples towards regions which trigger the rare event (or in our context, where $f(\boldsymbol{x}) \in \mathbb{Y}$) with the help of a new probability density q. The contribution from each sample, however, must be weighed to account for the fact that one is no longer sampling from the original distribution p. Thus,

$$\mu = \int_{\mathbb{R}^m} \mathbf{1}_{\mathbb{Y}}\left(f(\boldsymbol{x})\right) p(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = \int_{\mathbb{R}^m} \mathbf{1}_{\mathbb{Y}}\left(f(\boldsymbol{x})\right) \frac{p(\boldsymbol{x})}{q(\boldsymbol{x})} q(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = \mathbb{E}_q\left(\mathbf{1}_{\mathbb{Y}}\left(f(\boldsymbol{x})\right) \frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}\right). \tag{2}$$

Then, q is called the importance distribution and $p(\boldsymbol{x})/q(\boldsymbol{x})$ is the likelihood ratio. The importance sampling estimate for μ is:

$$\tilde{\mu}^N = \frac{1}{N} \sum_{i=1}^{N} \frac{\mathbf{1}_{\mathbb{Y}}\big(f(\boldsymbol{x}_i)\big) p(\boldsymbol{x}_i)}{q(\boldsymbol{x}_i)}, \quad \boldsymbol{x}_i \sim q(\boldsymbol{x}). \tag{3}$$

The relative RMSE in estimating μ using importance sampling is:

$$\begin{aligned} \tilde{e}_{\mathrm{RMS}} &= \frac{1}{\mu} \sqrt{\mathbb{E}_q\big((\tilde{\mu}^N - \mu)^2\big)} = \sqrt{\frac{\sigma_q^2}{\mu^2 N}}, \text{ where,} \\ \sigma_q^2 &= \mathbb{V}_q\left(\mathbf{1}_{\mathbb{Y}}\big(f(\boldsymbol{x})\big) \frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}\right). \end{aligned} \tag{4}$$

If σ_q^2 is smaller than σ_p^2, the importance sampling estimate of μ is more accurate than the one obtained using simple MC. The main challenge in importance sampling is selecting an importance density q such that $\sigma_q < \sigma_p$. The IS density that minimizes σ_q is known to be $q^* = \mathbf{1}_{\mathbb{Y}}(f(\boldsymbol{x}))p(\boldsymbol{x})/\mu$ (see [13]). That is, the optimal density for importance sampling is just $p(\boldsymbol{x})$ truncated over regions where $f(\boldsymbol{x}) \in \mathbb{Y}$, and then appropriately renormalized. However, q^* cannot be sampled from, since the renormalization constant μ is exactly the probability we set out to compute in the first place. Nevertheless, it defines characteristics desirable of a good importance density - it must have most of its mass concentrated over regions where $f(\boldsymbol{x}) \in \mathbb{Y}$ and resemble $p(\boldsymbol{x})$ in those regions.

So the first step in constructing an effective IS density is identifying regions where $f(\boldsymbol{x}) \in \mathbb{Y}$. As mentioned in Section 1, this is done by solving a Bayesian inverse problem. Before describing the BIMC methodology in detail, we first provide a brief introduction to Bayesian inference in a generalized setting.

2.3. Bayesian inference. In a general setting where inference must be performed, the problem is slightly different. Here the goal is to infer input parameters $\boldsymbol{x}$ from a (possibly noisy) real-world observation of the output, say y. In the Bayesian approach, this problem is solved in the statistical sense. The solution of a Bayesian inference problem is a probability density over the space of parameters that takes into account any prior knowledge about the parameters as well as uncertainties in measurement and/or modeling. This probability density, known as the posterior, expresses how likely it is for a particular estimate to be the true parameter corresponding to the observation.

In addition to the observation y, assume the following quantities have been specified - i) a suitable probability density $p(\boldsymbol{x})$ that captures prior knowledge about the parameters $\boldsymbol{x}$, and ii) the conditional probability density of observing the data y given the parameters $\boldsymbol{x}$, $p(y|\boldsymbol{x})$. Then, from Bayes' theorem, the posterior is given by:

$$p(\boldsymbol{x}|y) \propto p(y|\boldsymbol{x})p(\boldsymbol{x}). \tag{5}$$

The posterior can also be interpreted to be updated beliefs once the data and errors have been assimilated. We would like to emphasize here that in an actual inverse problem, the observation y, as well as the likelihood density $p(y|\boldsymbol{x})$ are physically meaningful. The former corresponds to real-world measurements of the output of the forward model. The latter describes a model for errors arising out due to modeling inadequacy or measurement.

The posterior by itself is of little use. Often, the task is to evaluate integrals involving the posterior. This might be the case, for example, when trying to characterize uncertainty in the inferred parameters by evaluating moments (mean, covariance) of the posterior. Analytical evaluation of these integrals is often out of the question and a sample based estimate must be used. Except in certain cases, the posterior is an arbitrary PDF in $\mathbb{R}^m$ and generating samples from it requires sophisticated methods such as Markov Chain Monte Carlo. For easy sample generation, the posterior can be locally approximated by a Gaussian around its mode (also known as the Maximum *A Posteriori* point). By linearizing f around the MAP point, it can be shown that the mean of the approximating Gaussian is the MAP point, and its covariance is the inverse of the Gauss-Newton Hessian matrix of $-\log p(\boldsymbol{x}|y)$ at the MAP point [12].

As a concrete example, consider the case when the likelihood density represents Gaussian additive error of magnitude σ, $p(y|\boldsymbol{x}) = \mathcal{N}(f(\boldsymbol{x}), \sigma^2)$. Then, $p(\boldsymbol{x}|y) \propto \exp\left(-\frac{(y-f(\boldsymbol{x}))^2}{2\sigma^2}\right) p(\boldsymbol{x})$, and we have (up to an additive constant),

$$-\log p(\boldsymbol{x}|y) = \frac{1}{2\sigma^2}\left(y - f(\boldsymbol{x})\right)^2 - \log p(\boldsymbol{x}), \tag{6}$$

and, $\boldsymbol{x}_{\text{MAP}}$ can be found as:

$$\boldsymbol{x}_{\text{MAP}} = \underset{\boldsymbol{x}\in\mathbb{R}^m}{\arg\min} \frac{1}{2\sigma^2}\left(y - f(\boldsymbol{x})\right)^2 - \log p(\boldsymbol{x}). \tag{7}$$

Then, the Gauss-Newton Hessian matrix of $-\log(p(\boldsymbol{x}|y))$ can be written as

$$\begin{aligned} \mathbf{H}_{\text{GN}} &= -\nabla^2_{\boldsymbol{x}} \log p(\boldsymbol{x}|y) \\ &= \frac{1}{\sigma^2}(\nabla_{\boldsymbol{x}} f)(\nabla_{\boldsymbol{x}} f)^T - \nabla^2_{\boldsymbol{x}} \log p(\boldsymbol{x}). \end{aligned} \tag{8}$$

Note that, the Gauss-Newton Hessian has the attractive property of being positive-definite. These expressions show that $\boldsymbol{x}_{\text{MAP}}$ can be interpreted as that point in parameter space that minimizes mismatch with the observation but is also highly likely under the prior. So sampling from a Gaussian approximation of the posterior can be thought of as drawing samples in the vicinity of a point that is consistent with the data as well as the prior. In addition, the covariance or spread of the samples is informed by the derivatives of the forward model. While constructing the IS density in BIMC, this feature of the Gaussian approximation of the posterior in a general, real-world setting will be used in conjunction with the knowledge of the shape of the ideal IS density. This completes the presentation of the necessary theoretical background and we are ready to describe the BIMC methodology.

3. Methodology. Recall that the forward UQ problem is to compute $\mathbb{P}\left(f(\boldsymbol{x}) \in \mathbb{Y}\right)$ when $\boldsymbol{x} \sim p(\boldsymbol{x})$. In BIMC, we use the ingredients of the forward UQ problem to construct a *fictitious* Bayesian inverse problem as follows. We

1. select some $y \in \mathbb{Y}$ as a surrogate for real-world observation,
2. use $p(\boldsymbol{x})$ as the prior, and,
3. concoct a likelihood density $p(\boldsymbol{x}|y)$.

This enables us to define a pseudo-posterior $p(\boldsymbol{x}|y)$, and subsequently, a Gaussian approximation to it. We call this inverse problem fictitious because both the observation y and the likelihood density $p(\boldsymbol{x}|y)$ are arbitrarily chosen by us. Neither is y a real-world measurement of a physical quantity, nor does $p(\boldsymbol{x}|y)$ correspond to an actual error model. From here on, we will refer to these artificial quantities as the pseudo-data and the pseudo-likelihood respectively.

We propose using the Gaussian approximation to the posterior as an IS density. As outlined in the previous section, in the real-world setting, the mean of the Gaussian approximation of the posterior (the MAP point) is that point in parameter space that is consistent with the data as well as the prior. So by solving the fictitious Bayesian inverse problem defined earlier, we expect the mean of the IS density to be a point that is consistent with some $y \in \mathbb{Y}$ as well as the nominal PDF $p(\boldsymbol{x})$. This ensures the IS density is centered around regions where $f(\boldsymbol{x}) \in \mathbb{Y}$. Further, the covariance matrix of the Gaussian approximation, and hence the IS density, contains first-order derivative information. This approach is illustrated in Figure 1.

Since a Gaussian likelihood model has been assumed, the pseudo-posterior is proportional to $\exp\left(-(y - f(\boldsymbol{x}))^2/2\sigma^2\right) p(\boldsymbol{x})$. Thus, an alternative interpretation of the pseudo-posterior in this case is as a "mollified" approximation of the ideal IS density, $\mathbf{1}_{\mathbb{Y}}\left(f(\boldsymbol{x})\right) p(\boldsymbol{x})$, where the mollification has been achieved by smudging the sharply defined characteristic function $\mathbf{1}_{\mathbb{Y}}(f(\boldsymbol{x}))$ into a Gaussian, $\exp\left(-(y - f(\boldsymbol{x}))^2/2\sigma^2\right)$ The advantage of doing this lies in the fact that the mollified ideal IS density has well-defined derivatives and can be explored via derivative-aware methods, unlike the true ideal IS density, which isn't differentiable. Algorithms like IMIS [24], and LIMIS [9] can now be employed for rare-event probability estimation by plugging in the pseudo-posterior as the target.

Irrespective of the interpretation, this methodology introduces two tunable parameters—the pseudo-data $y \in \mathbb{Y}$ and variance of the pseudo-likelihood density, σ^2. These parameters can have a profound effect on the accuracy of the importance sampler and must be tuned with care. The tuning strategy depends on the nature of $f(\boldsymbol{x})$ as well as $p(\boldsymbol{x})$. Next, we discuss possible cases as well as the corresponding tuning strategy.

3.1. Affine f, Gaussian p. Although for a affine $f(\boldsymbol{x})$ and Gaussian $p(\boldsymbol{x})$, the probability $\mu = \mathbb{P}(f(\boldsymbol{x}) \in \mathbb{Y})$ may be analytically computed, the availability of analytical expressions for $\boldsymbol{x}_{\text{MAP}}$ and $\mathbf{H}_{\text{GN}}$ in this case illustrates how our importance sampler achieves variance reduction. Let $f(\boldsymbol{x}) = \boldsymbol{v}^T\boldsymbol{x} + \beta$, $p(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}_0, \boldsymbol{\Sigma}_0)$ for some $\boldsymbol{v}$, $\boldsymbol{x}_0 \in \mathbb{R}^m$, $\beta \in \mathbb{R}$ and $\boldsymbol{\Sigma}_0 \in \mathbb{R}^{m\times m}$. Then, μ is given analytically as

$$\mu = \Phi\left(\frac{y_{\max} - \nu}{\gamma}\right) - \Phi\left(\frac{y_{\min} - \nu}{\gamma}\right), \tag{9}$$

where $\nu = \boldsymbol{v}^T\boldsymbol{x}_0 + \beta$, $\gamma^2 = \boldsymbol{v}^T\boldsymbol{\Sigma}_0\boldsymbol{v}$, and Φ is the standard Normal CDF.

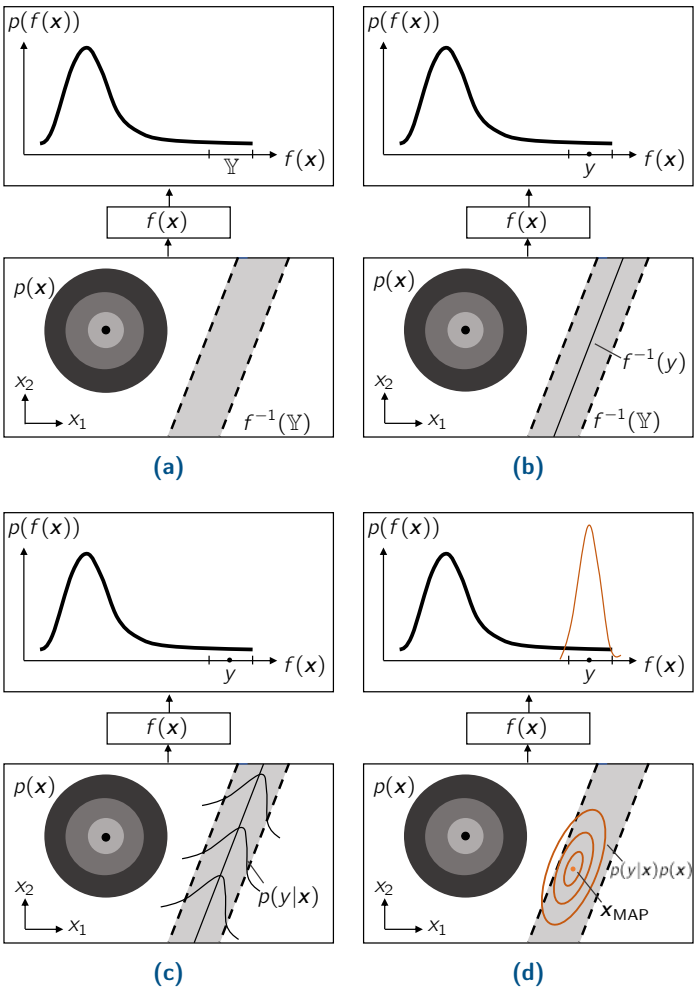


Figure 1: *Summary of the BIMC methodology. In (a), the problem statement is summarized - we need to compute the probability of the pre-image of the target interval, $f^{-1}(\mathbb{Y})$. In (b), we introduce the pseudo-data point $y \in \mathbb{Y}$ and redefine $p(\boldsymbol{x})$ to be the pseudo-prior in a fictitious Bayesian inverse problem. The true inverse of the pseudo-data point y is a straight line in $\mathbb{R}^2$. Next, in (c), we select a pseudo-likelihood density, $p(y|\boldsymbol{x})$. The pseudo-likelihood density can also be viewed as a mollified approximation of the characteristic function. Part (d) shows the contours of the posterior, $p(\boldsymbol{x}|y)$, which is proportional to $p(y|\boldsymbol{x})p(\boldsymbol{x})$. We use a Gaussian approximation to this posterior as an IS density.*

Now, suppose the pseudo-data is some $y \in \mathbb{Y}$ and the variance of the pseudo-likelihood is some $\sigma^2 \in \mathbb{R}$. Then the pseudo-posterior $p(\boldsymbol{x}|y)$ is also a Gaussian and no approximations are necessary. Hence, the IS density is given by $q(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}_{\text{MAP}}, \mathbf{H}_{\text{GN}}^{-1})$, where,

$$\boldsymbol{x}_{\text{MAP}} = \boldsymbol{x}_0 + \frac{y - f(\boldsymbol{x}_0)}{\sigma^2 + \boldsymbol{v}^T \boldsymbol{\Sigma}_0 \boldsymbol{v}} \boldsymbol{\Sigma}_0 \boldsymbol{v}, \quad \mathbf{H}_{\text{GN}}^{-1} = \boldsymbol{\Sigma}_0 - \frac{1}{\sigma^2 + \boldsymbol{v}^T \boldsymbol{\Sigma}_0 \boldsymbol{v}} \big(\boldsymbol{\Sigma}_0 \boldsymbol{v}\big)\big(\boldsymbol{\Sigma}_0 \boldsymbol{v}\big)^T. \tag{10}$$

The expressions for $\boldsymbol{x}_{\text{MAP}}$ and $\mathbf{H}_{\text{GN}}^{-1}$ expose how our importance sampler achieves variance reduction. The MAP point identifies the region in parameter space where $f(\boldsymbol{x}) \approx y$. The spread of the importance sampler, as encapsulated in its covariance, is reduced over its nominal value $\boldsymbol{\Sigma}_0$ in a direction informed by $\boldsymbol{v}$, the gradient of $f(\boldsymbol{x})$. Note that the reduction in variance occurs in just one direction, $\boldsymbol{\Sigma}_0\boldsymbol{v}$; the variance of $p(\boldsymbol{x})$ is retained in all other directions. The parameter σ^2 controls how much $q(\boldsymbol{x})$ is updated over $p(\boldsymbol{x})$- a small value for σ results in a larger shift from $\boldsymbol{x}_0$ and a larger reduction in its spread. These claims become more transparent by noticing that the pushforward of $q(\boldsymbol{x})$ under f is another Gaussian distribution in $\mathbb{R}$ (the pushforward density represents how $f(\boldsymbol{x})$ will be distributed if $\boldsymbol{x}$ is distributed according to $q(\boldsymbol{x})$). The mean χ and variance ξ^2 of this pushforward density are-

$$\chi = (1-\rho^2)y + \rho^2 f(\boldsymbol{x}_0),\ \xi^2 = \rho^2 \boldsymbol{v}^T\boldsymbol{\Sigma}_0\boldsymbol{v},\ \text{where},\rho^2 = \frac{\sigma^2}{\sigma^2 + \boldsymbol{v}^T\boldsymbol{\Sigma}_0\boldsymbol{v}} < 1. \tag{11}$$

A small σ implies small ρ, which means χ is closer y and ξ^2 is small.

Since our goal is importance sampling, we wish to select those values for the tunable parameters that deliver just the right amount of update over p. We do this by minimizing the Kullback-Leibler (KL) divergence between $q(\boldsymbol{x})$ and the ideal IS distribution $q^*(\boldsymbol{x})$. Although not a true metric, the KL divergence between two probability densities is a measure of the distance between them. It is defined as:

$$D_{\text{KL}}(p||q) = \int p(\boldsymbol{x}) \log \frac{p(\boldsymbol{x})}{q(\boldsymbol{x})} \mathrm{d}\boldsymbol{x}. \tag{12}$$

Then, the optimal pseudo-data point and the optimal variance of the pseudo-likelihood density can be obtained as:

$$\begin{pmatrix} \sigma^* \\ y^* \end{pmatrix} = \arg\min_{\sigma,y} D_{\text{KL}}(q^*||q;\sigma,y). \tag{13}$$

Analytic expressions for D_{KL}, y^*, and σ^* are derived for the affine Gaussian case in the supplement in Section SM1. Selecting the tunable parameters in this way, in fact, allows us to make the following claim regarding the resulting IS distribution:

Claim 3.1 (BIMC optimality). *In the affine-Gaussian case, the importance sampling density that results from the BIMC procedure is equivalent to the Gaussian distribution closest in KL divergence to $q^*(\boldsymbol{x})$.*

Proof. Proof given in the supplement in Section SM1: ∎

Hence, BIMC is implicitly searching for the best Gaussian approximation of $q^*(\boldsymbol{x})$. A Gaussian distribution in m dimensions has $m(m+1)/2$ free variables, so a naive search for the best Gaussian approximation of $q^*(\boldsymbol{x})$ will optimize over all $\mathcal{O}(m^2)$ free variables. However, BIMC accomplishes this task be optimizing just 2 free variables. This can be attributed to the similar structure of the pseudo-posterior $p(y|\boldsymbol{x})p(\boldsymbol{x})/p(y)$, and the ideal IS density, $\mathbf{1}_{\mathbb{Y}}(f(\boldsymbol{x}))p(\boldsymbol{x})/\mu$.

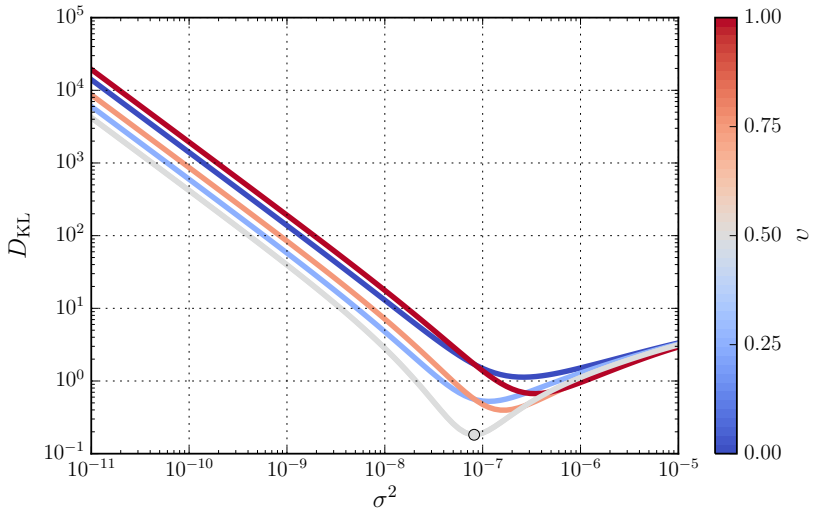


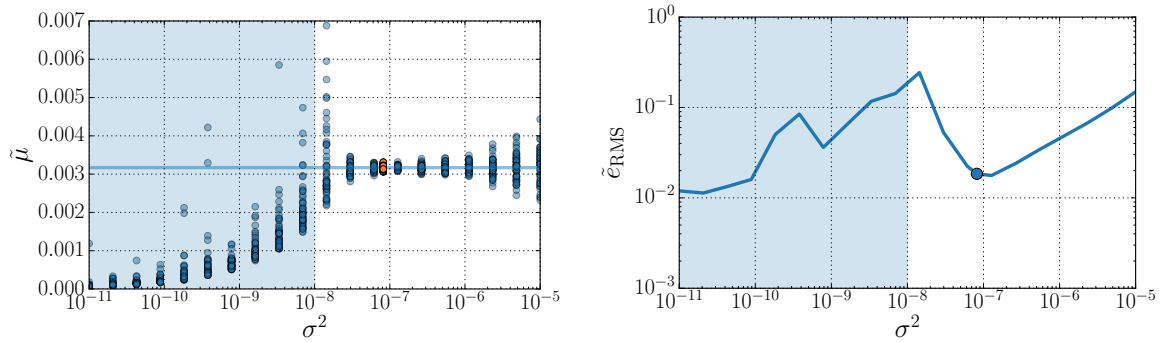
Figure 2: *This figure shows the variation of D_{KL} at various values of the data point y and pseudo-likelihood variance σ^2. Here, the data point y has been normalized to $\upsilon = (y - y_{\min})/(y_{\max} - y_{\min})$ and $f(\boldsymbol{x})$ is an affine transformation from $\mathbb{R}^{100}$ to $\mathbb{R}$. The marker shows the (σ, υ) combination that resulted from numerical minimization of D_{KL}.*

To verify whether minimizing D_{KL} to obtain parameters actually translates to improved performance of our IS density, we synthesized a affine map from $\mathbb{R}^{100}$ to $\mathbb{R}$ (implementation details are provided in the supplement in Section SM2). We measure performance by the relative RMSE in the probability estimate, $\tilde{e}_{\mathrm{RMS}}$, and we expect $\tilde{e}_{\mathrm{RMS}}$ to be small when $y = y^*$ and $\sigma = \sigma^*$. In addition, in this case, μ is available to us analytically. This provides yet another indicator of performance- the absolute difference between the analytical value and the IS estimate must be small when the optimal parameters are being used.

Figure 2 shows the variation of D_{KL} with σ^2 at various y in addition to the optimal (σ, y) combination that results from numerical optimization. We conclude the following from the figure:

- The optimal pseudo-data point lies almost exactly at the mid-point of $\mathbb{Y}$.
- D_{KL} is extremely sensitive to the spread of the pseudo-likelihood probability density σ, much more so than the pseudo-data y. Intuitively, a large value for σ emphasizes the pseudo-prior over the data so that sampling from $q(\boldsymbol{x})$ is akin to sampling from $p(\boldsymbol{x})$. On the other hand, too small a value for σ^2 results in $q(\boldsymbol{x})$ not having enough spread to cover the region where $f(\boldsymbol{x}) \in \mathbb{Y}$, which could result in significant bias when the number of samples is small.

In Figure 3, we fix $y = y^*$ and plot the variation of the probability estimate, $\tilde{\mu}$, and the relative RMSE, $\tilde{e}_{\mathrm{RMS}}$, with σ^2. For each value of σ^2, we performed several independent runs. Figure 3a plots $\tilde{\mu}$ obtained from each run. In Figure 3b, we plot the ensemble average of $\tilde{e}_{\mathrm{RMS}}$ over all simulations at fixed σ^2. Both figures demonstrate that when σ^2 is small, both the probability estimate and the associated RMSE have significant bias (shaded region in the



(a) *The IS probability estimate $\tilde{\mu}$ against σ^2 at $y = y^*$. Each marker indicates $\tilde{\mu}$ from an individual run. The solid line indicates the true value of μ. The markers in orange denote runs at $\sigma = \sigma^*$.*

(b) *The relative RMSE $\tilde{e}_{\mathrm{RMS}}$ against σ^2 at $y = y^*$. Only the average $\tilde{e}_{\mathrm{RMS}}$ over all runs at fixed σ^2 is plotted. The marker is at $\sigma = \sigma^*$.*

Figure 3: *This figure shows the variation of the probability estimate $\tilde{\mu}$ and the relative RMSE $\tilde{e}_{\mathrm{RMS}}$ with the likelihood variance σ^2. At fixed σ^2, we performed 50 independent runs using $N = 1000$ samples. For $\tilde{\mu}$, we plot the probability estimate obtained from each run, whereas for $\tilde{e}_{\mathrm{RMS}}$, we plot the ensemble average at each σ.*

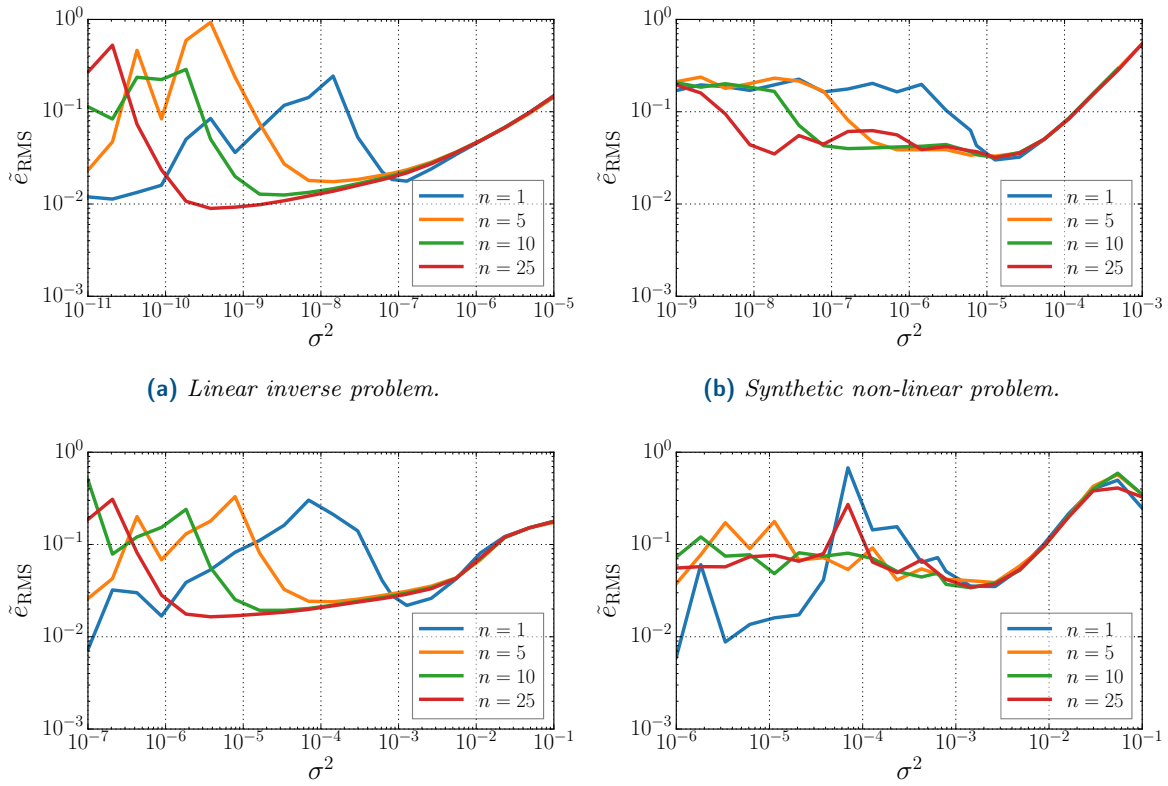
figures). When σ^2 is large, error increases with σ^2 since the emphasis on pseudo-data decreases. There lies an optimal σ^2 somewhere in between, and indeed, minimizing D_{KL} helps identify it. So far we've been using just one pseudo-data point $y \in \mathbb{Y}$. However, it is also possible to use multiple pseudo-data points, $\{y_i\}_{i=1}^n, y_i \in \mathbb{Y}$. In this case, using the same pseudo-likelihood density for all y_i, a posterior $p(\boldsymbol{x}|y_i)$ and its corresponding Gaussian approximation can be obtained for each y_i. These Gaussians can then be collected into a mixture distribution to form the IS density. So, a possibility is to use the following IS density:

$$q(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n} \mathcal{N}\left(\boldsymbol{x}_{\mathrm{MAP}}^{(i)}, \left(\mathbf{H}_{\mathrm{GN}}^{(i)}\right)^{-1}\right). \tag{14}$$

where $\boldsymbol{x}_{\mathrm{MAP}}^{(i)}$ is the MAP point corresponding to y_i and $\mathbf{H}_{\mathrm{GN}}^{(i)}$ is the Hessian matrix of $-\log p(\boldsymbol{x}|y_i)$ at $\boldsymbol{x}_{\mathrm{MAP}}^{(i)}$. Next, we investigate whether using $n > 1$ pseudo-data points in $\mathbb{Y}$ leads to better performance than using just one pseudo-data point, i.e., $n = 1$.

To ensure a fair comparison between the two cases, they must each be run using their respective optimal parameters. When $n > 1$, the tunable parameters are the number of pseudo-data points, n, their values, $\{y_i\}_{i=1}^n$, and the variance of the common pseudo-likelihood density, σ^2. However, minimizing the Kullback-Leibler distance between q and q^* to obtain parameters is no longer possible. This is because the Kullback-Leibler distance between two Gaussian mixtures doesn't have a closed form expression [11]. To proceed, given $n > 1$, we fix $\{y_i\}$ to be n evenly spaced points in $\mathbb{Y}$. We then sweep over several values of n and σ^2 to investigate whether increasing n has any advantages.

Empirical evidence seems to suggest no. In Figure 4, we plot the variation of the ensemble



(a) *Linear inverse problem.*

(b) *Synthetic non-linear problem.*

(c) *Single step reaction.*

(d) *Elliptic PDE.*

Figure 4: *This figure shows how $\tilde{e}_{\mathrm{RMS}}$ varies with σ^2 at different values of n for various forward models. At fixed σ^2, we perform 50 independent simulations and report the ensemble averaged $\tilde{e}_{\mathrm{RMS}}$. When $n = 1$, we use the optimal data point obtained employing the appropriate tuning strategy described in the text. When $n > 1$, we select n evenly spaced points in $\mathbb{Y}$. Similar to Figure 3, IS estimates are biased when σ^2 is small. The extent of the biased regions depends on n and appears to decrease as n increases.*

averaged $\tilde{e}_{\mathrm{RMS}}$ with σ^2 at various values of n and using various forward models, both linear and non-linear (details of the forward models are provided in the supplement in Section SM2). While the error decreases with increasing n for some cases, we believe the decrease isn't large enough to justify the increased computational cost of solving additional inverse problems.

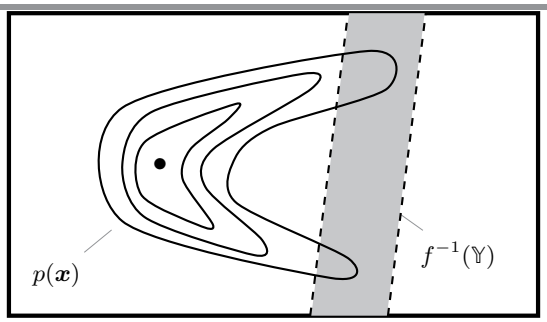


Figure 5: *Arbitrary unimodal p. In this case, the true posterior and the ideal IS density are multimodal. The optimizer used for computing $\boldsymbol{x}_{\text{MAP}}$ will only find one of these modes. Our IS density will then only sample around that mode, leading to incorrect estimates.*

3.2. Affine f, mixture-of-Gaussians p. If $p(\boldsymbol{x})$ is a mixture of Gaussians, $p(\boldsymbol{x}) = \sum_{i=1}^{k} w_i p_i(\boldsymbol{x})$, then, notice that,

$$\mu = \int \mathbf{1}_{\mathbb{Y}}\left(f(\boldsymbol{x})\right) p(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = \sum_{i=1}^{k} w_i \int \mathbf{1}_{\mathbb{Y}}\left(f(\boldsymbol{x})\right) p_i(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = \sum_{i=1}^{k} w_i \mu_i \tag{15}$$

The contribution to μ from each component $p_i(\boldsymbol{x})$, μ_i, can then be calculated using BIMC as described above. If $\tilde{\mu}_i$ is the estimate from each component, μ can be estimated as $\tilde{\mu} = \sum_{i=1}^{k} w_i \tilde{\mu}_i$.

3.3. Affine f, arbitrary unimodal p. In this case, even though p is unimodal, q^*, and the pseudo-posterior $p(\boldsymbol{x}|y)$, can be multi-modal (see Figure 5). Then, depending on the initial guess provided, the optimizer used for computing the MAPs may converge to only one of the modes. A local Gaussian characterization of the pseudo-posterior will only sample near this mode and all the other modes will be ignored. This will cause μ to be underestimated. To avoid this, we propose approximating p with a mixture of Gaussians and then proceeding with the methodology outlined in the previous section. This will lead to an estimate whose accuracy is as good as the accuracy in approximating p with a mixture of Gaussians.

3.4. Non-linear f, Gaussian p. When $f(\boldsymbol{x})$ is non-linear, the KL divergence may not have a tractable closed-form expression even when only one pseudo-data point is used. Although a sample based estimate of the KL divergence can be obtained, it would require evaluating $f(\boldsymbol{x})$ for each sample, increasing the cost of constructing the IS density. To compute the optimal parameters in this case, we instead propose linearizing $f(\boldsymbol{x})$ around the MAP point corresponding to an initial pseudo-data point, $y^{\text{mid}} = \text{mid}\,\mathbb{Y}$, which we denote $\boldsymbol{x}_{\text{MAP}}^{\text{mid}}$. This necessitates solving another optimization problem (as in Equation (7)), for which we require σ^2, a quantity we set out to tune in the first place. However, this σ^2 is only used to construct the linearization and has little bearing on subsequent sampling. We recommend setting $\sigma = 0.1(y_{\max} - y_{\min})$. Once we have $\boldsymbol{x}_{\text{MAP}}^{\text{mid}}$, we linearize $f(\boldsymbol{x})$ as follows:

$$f(\boldsymbol{x}) \approx f(\boldsymbol{x}_{\text{MAP}}^{\text{mid}}) + \mathbf{J}^{\text{mid}}(\boldsymbol{x} - \boldsymbol{x}_{\text{MAP}}^{\text{mid}}) \tag{16}$$

Here, $\mathbf{J}^{\text{mid}} \in \mathbb{R}^{1\times m}$ is the Jacobian matrix of the $f(\boldsymbol{x})$ evaluated at $\boldsymbol{x}^{\text{mid}}_{\text{MAP}}$. From here on, we can proceed to obtain the optimal parameters as in the affine case by identifying $\boldsymbol{v}^T \equiv \mathbf{J}^{\text{mid}}$ and $\beta \equiv f(\boldsymbol{x}^{\text{mid}}_{\text{MAP}}) - \mathbf{J}^{\text{mid}}\boldsymbol{x}^{\text{mid}}_{\text{MAP}}$. Note that such a procedure will not reveal the true optimal parameters that correspond to the non-linear forward model. It only provides an estimate, but allows us to use analytically derived expressions and keep computational costs low. Another consequence of linearizing $f(\boldsymbol{x})$ is that it allows for the analytical computation of the rare event probability associated with the linearized map (Equation (9)). We will refer to this estimate of μ as the linearized probability estimate, μ_{lin}.

3.5. Non-linear f, mixture-of-Gaussian p. This case is similar to Subsection 3.2. Recall that μ is just the weighted sum of probability corresponding to each component mixtures, μ_i. Each μ_i can be estimated by the method outlined above, and then weighed and summed to obtain an estimate for μ.

Procedure 1 BIMC

Input: $f(\boldsymbol{x})$, $p(\boldsymbol{x})$, $\mathbb{Y}$, N
Output: $\tilde{\mu}$

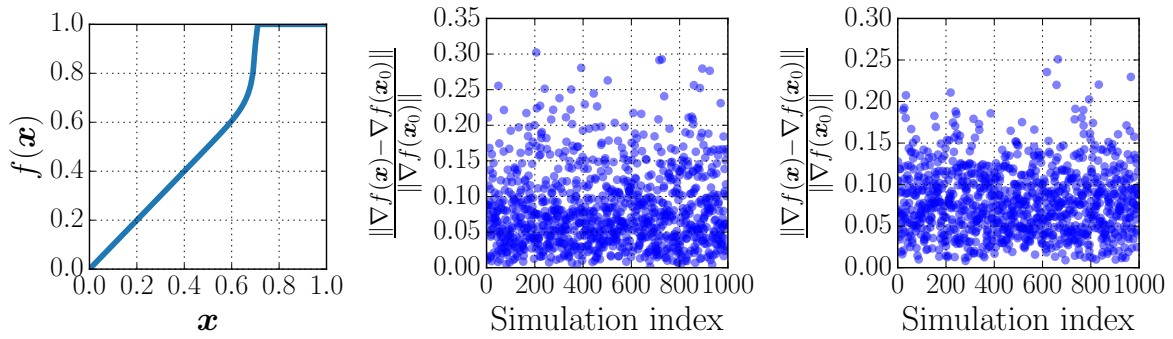
% Select optimal parameters, y^*, σ^*
$y^{\max} \leftarrow \max \mathbb{Y}$, $y^{\min} \leftarrow \min \mathbb{Y}$, $y^{\text{mid}} \leftarrow 0.5(y^{\min} + y^{\max})$
$\sigma_0 \leftarrow 0.1(y^{\max} - y^{\min})$
$\boldsymbol{x}^{\text{mid}}_{\text{MAP}} \leftarrow \texttt{getMAP}(y_{\text{mid}}, \sigma_0)$ % Minimize Equation (6) using $y = y^{\text{mid}}, \sigma = \sigma_0$
$\boldsymbol{v}^T \leftarrow \left.\frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{x}}\right|_{\boldsymbol{x}=\boldsymbol{x}_{\text{MAP}}}$
$\beta \leftarrow f(\boldsymbol{x}^{\text{mid}}_{\text{MAP}}) - \boldsymbol{v}^T\boldsymbol{x}^{\text{mid}}_{\text{MAP}}$
$(y^*, \sigma^*) \leftarrow \texttt{minimizeKLDiv}(\boldsymbol{v}, \beta, p)$ % Minimize $D_{\text{KL}}(q^*||q)$ as in Equation (SM6)

% Build IS density using optimal parameters
$\boldsymbol{x}_{\text{MAP}} \leftarrow \texttt{getMAP}(y^*, \sigma^*)$ % Minimize Equation (6) using $y = y^*, \sigma = \sigma^*$
$\mathbf{H}_{\text{GN}} \leftarrow \texttt{getHessian}(\boldsymbol{x}_{\text{MAP}}, y^*, \sigma^*)$ % Compute Hessian of Equation (6) at $\boldsymbol{x}_{\text{MAP}}$ using $y = y^*, \sigma = \sigma^*$
$q(\boldsymbol{x}) \leftarrow \mathcal{N}(\boldsymbol{x}_{\text{MAP}}, \mathbf{H}^{-1}_{\text{GN}})$

% Sample from q to estimate μ
for $i = 1, \ldots, N$ **do**
    $\boldsymbol{x}_i \sim q(\boldsymbol{x})$
    $w_i \leftarrow \mathbf{1}_{\mathbb{Y}}(f(\boldsymbol{x}_i))p(\boldsymbol{x}_i)/q(\boldsymbol{x}_i)$
end for
$\tilde{\mu} \leftarrow \sum_{i=1}^{N} w_i/N$
return $\tilde{\mu}$

3.6. Summary. To summarize, in this section we described how a fictitious Bayesian inverse problem can be constructed from the components of the forward UQ problem. The solution of this fictitious inverse problem yields a posterior whose Gaussian approximation is our IS density. The parameters on which the IS density depends can be tuned by minimizing an analytical expression for its Kullback-Leibler divergence with respect to the ideal IS den-

sity. A drawback of our method is that we're restricted to nominal densities that are Gaussian mixtures or easily approximated by one. The overall algorithm for arbitrary, non-linear f is given in Algorithm 1. Next, we present and discuss results of our numerical experiments.



(a) *Single step reaction* **(b)** *Autoignition* **(c)** *Synthetic non-linear*

Figure 6: *Non-linearity of f. For the single step reaction problem, we plot the full forward map f for all possible values of the input $\boldsymbol{x}$. For the autoignition and synthetic non-linear problems, we demonstrate how $\nabla f(\boldsymbol{x})$ varies. We draw samples, $\boldsymbol{x}_i$, from $p(\boldsymbol{x})$, and evaluate $\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{x}_0)\|/\|\nabla f(\boldsymbol{x}_0)\|$, where $\boldsymbol{x}_0$ is the prior mean. Departure from 0 of this quantity signifies the degree of non-linearity.*

4. Experiments. In this section, we present results that demonstrate the efficacy of our method. We also report cases where our method fails (detailed discussion about failure mechanisms of BIMC is postponed to the end of this section). The forward models we used in our experiments are briefly summarized below. A detailed description of the models and the problem setup is given in the supplement in Section SM2. Figure 6 shows the variation of f for select models and demonstrates that it is indeed non-linear.

- Affine case: In this case $f(\boldsymbol{x})$ is a affine map from $\mathbb{R}^m$ to $\mathbb{R}$. We choose $m = 2$ for illustration, and $m = 100$ for comparison with MC.
- Synthetic non-linear problem: In this case, $f(\boldsymbol{x})$ is defined to be the following map from $\mathbb{R}^m$ to $\mathbb{R}$.

$$f(\boldsymbol{x}) = \boldsymbol{o}^T\boldsymbol{u}, \text{ where } \left(\mathbf{S} + \varepsilon\boldsymbol{x}\boldsymbol{x}^T\right)\boldsymbol{u} = \boldsymbol{b}. \tag{17}$$

  Here, $\varepsilon \in \mathbb{R}$, $\boldsymbol{o}, \boldsymbol{u}, \boldsymbol{b} \in \mathbb{R}^m$, and $\mathbf{S} \in \mathbb{R}^{m\times m}$. Again, $m = 2$ was chosen for illustration and $m = 10$ for comparison with MC.
- Single step reaction: The forward model here describes a single step chemical reaction using an Arrhenius type rate equation. A progress variable $u \in [0, 1]$ is used to describe the reaction. The parameter $\boldsymbol{x}$ is the initial value of progress variable $u(0)$ and the observable $f(\boldsymbol{x})$ is the value of the progress variable at some final time $t_f, u(t_f)$. Thus $f(\boldsymbol{x})$ is a map from $\mathbb{R}$ to $\mathbb{R}$.
- Autoignition: Here, we allow a mixture of hydrogen and air to undergo autoignition in a constant pressure reactor. A simplified mechanism with 5 elementary involving 8 chemical species is used to describe the chemistry. The parameter $\boldsymbol{x}$ is the vector of

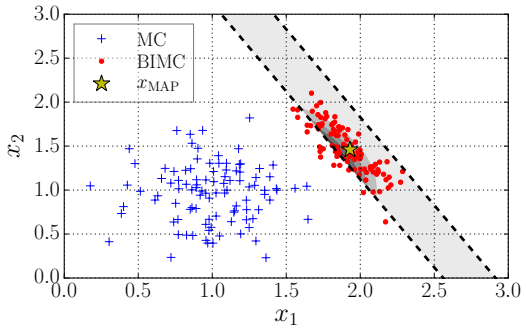
the initial equivalence ratio, initial temperature and the initial pressure in the reactor and the observable is the amount of heat released so that $f(\boldsymbol{x})$ is a map from from $\mathbb{R}^3$ to $\mathbb{R}$.

- Elliptic PDE: In this system, we invert for the discretized log-permeability field in some spatial domain given an observation of the pressure at some point. The forward problem, that is, obtaining the pressure from the log-permeability field, is governed by an elliptic PDE. A finite element discretization results in $f(\boldsymbol{x})$ being a map from $\mathbb{R}^{4225}$ to $\mathbb{R}$.
- The Lorenz system: Here, the forward problem is governed by the chaotic Lorenz equations [18]. The parameter $\boldsymbol{x}$ is the initial condition of the system while the observable is value of the first component of the state vector at some final time t_f. We simulate the Lorenz system over three time horizons, $t_f = 0.1$s, $t_f = 5$s, and $t_f = 15$s. BIMC fails over longer time horizons, i.e., when $t_f = 5$s and $t_f = 15$s.
- Periodic case: Here, $f(\boldsymbol{x})$ is a periodic function in $\mathbb{R}^2$, $f(\boldsymbol{x}) = \sin(x_1)\cos(x_2)$. This is another case when BIMC fails.

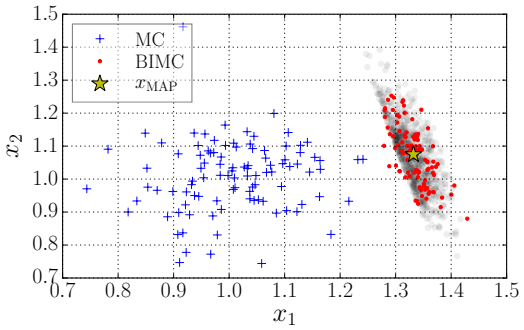
Sampling illustration. We begin by presenting examples in low-dimensions that illustrate the quality of samples from BIMC. In Figure 7, we compare samples generated using MC and BIMC. We also depict the ideal IS density q^* in the figures, either using contours, or through samples. As expected, the variance of the IS density in our method is only decreased in one data-informed direction. The extent of this decrease depends on the variance of the pseudo-likelihood density, $p(y|\boldsymbol{x})$, and a tuning algorithm based on minimizing the Kullback-Leibler distance leads to a good fit between the spread of $q^*(\boldsymbol{x})$ and $q(\boldsymbol{x})$ in this direction. In all other directions, the spread of $q(\boldsymbol{x})$ is same as that of $p(\boldsymbol{x})$. This is because the pseudo-data y does not inform these directions.

As a quantitative estimate of the quality of samples, we report the acceptance ratio, defined as the fraction of samples that evaluate inside $\mathbb{Y}$. The acceptance ratio resulting from BIMC is plotted in Figure 8 (the acceptance ratio from MC on the other hand is $\hat{\mu}$ by definition). We observe that $n = 1$ consistently leads to an acceptance ratio of around 90% irrespective of μ (except in the Periodic and Lorenz, $t_f = 5$s cases; these are failure cases and will be discussed at the end of this section). The slight dip in the acceptance ratio when $n > 1$ can be attributed to the effect of always having $y_{\min}$ and $y_{\max}$ as data points. Because these points lie at the edge of the interval $\mathbb{Y}$, they lead to an increased number of samples that are close to these limit points, but don't actually evaluate inside $\mathbb{Y}$. As n increases however, the number of samples drawn from mixture components corresponding to these two points decreases and the acceptance ratio shows an upward trend.

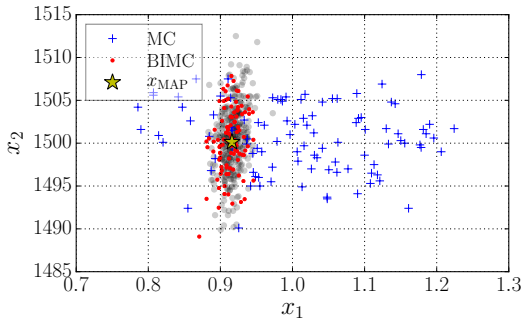
Convergence with number of samples. Next, we compare the relative RMS error, e_{RMS}, from MC and BIMC in Figure 9. BIMC offers the same accuracy using far fewer number of samples and results in an order of magnitude or more of speedup. The exact speedup achieved depends on the magnitude of the probability. In addition, there is little asymptotic effect of using $n > 1$. The corresponding probability estimates are presented in the supplement in Section SM3.



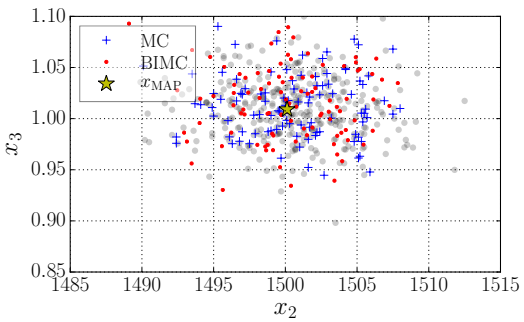
(a) *The affine case.*



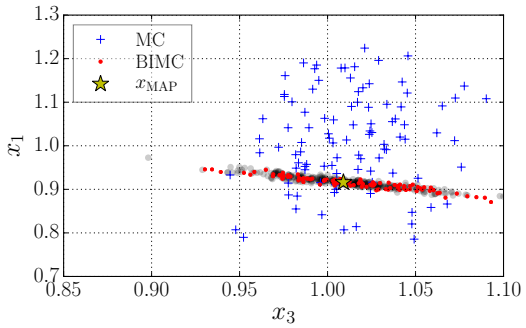
(b) *The synthetic non-linear case.*



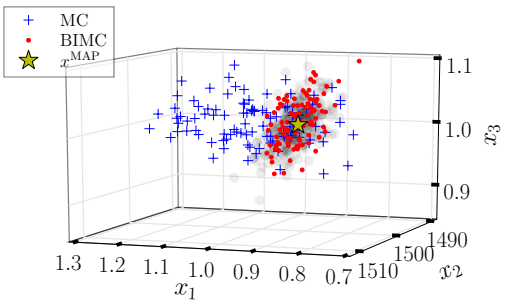
(c) *Autoignition, x_1 - x_2 plane.*



(d) *Autoignition, x_2 - x_3 plane.*



(e) *Autoignition, x_3 - x_1 plane.*



(f) *Autoignition, 3D view.*

Figure 7: *Sampling illustration. In this figure, we plot 100 samples from $p(\boldsymbol{x})$ (which corresponds to vanilla MC) as well as $q(\boldsymbol{x})$ (which corresponds to BIMC) with $n = 1$ for the affine, synthetic non-linear, and the autoignition problems. For the affine case (a), the region in $\mathbb{R}^2$ that evaluates inside $\mathbb{Y}$ is analytically available and is plotted between the thick, dashed lines. Also analytically available is the ideal IS density q^* whose contours are plotted. For all other forward models, a scatter plot of samples drawn from q^* is used to represent its magnitude.*

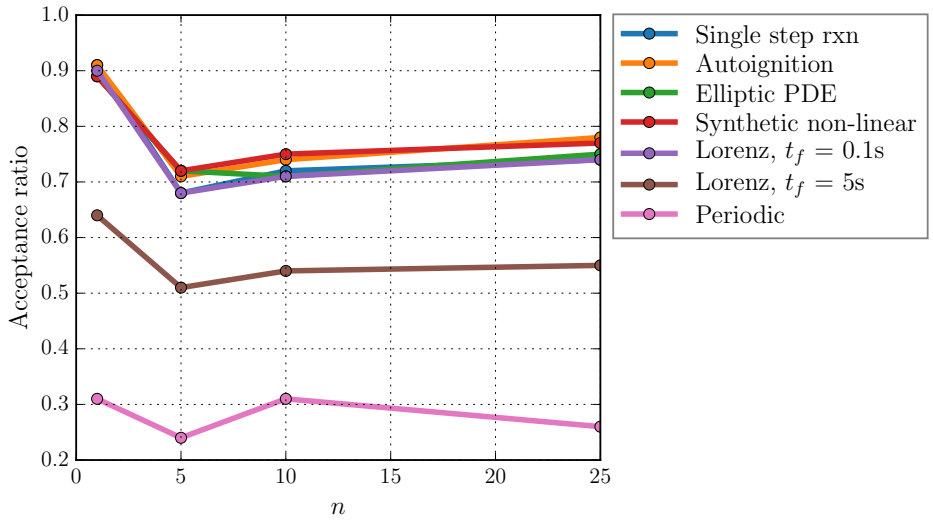
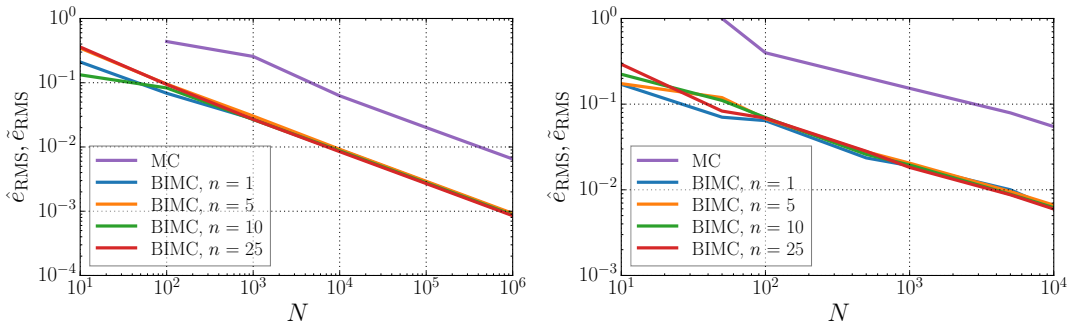
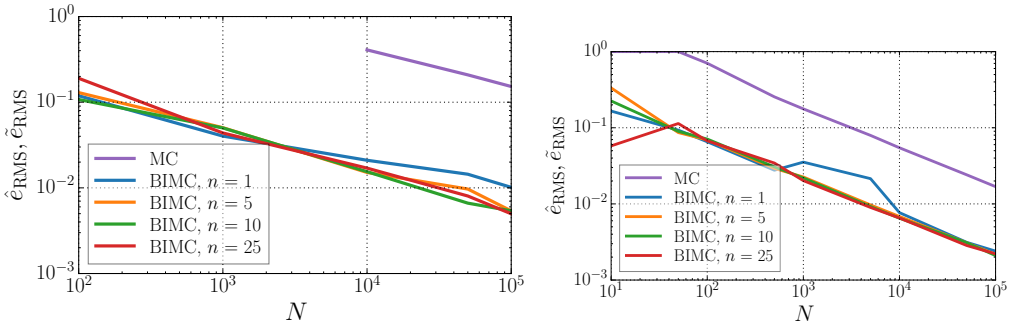


Figure 8: *Fraction of samples that evaluate inside $\mathbb{Y}$ for the different forward models at various values of n. In this experiment $N = 1000$ and μ spans two orders of magnitude, from $\mathcal{O}(10^{-2})$ to $\mathcal{O}(10^{-4})$. The BIMC methodology fails for the periodic and Lorenz, $t_f = 5s$ cases, hence the lower acceptance ratio.*



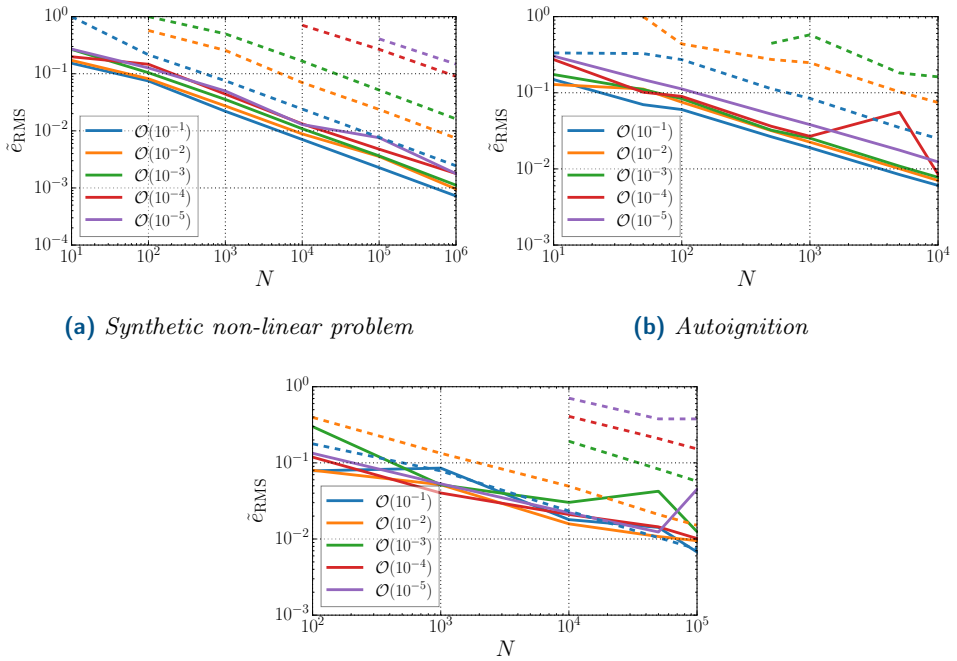
(a) *Single step reaction, $\mu \approx 2.30 \times 10^{-2}$.* **(b)** *Autoignition, $\mu \approx 3.24 \times 10^{-2}$.*



(c) *Elliptic PDE, $\mu \approx 3.91 \times 10^{-4}$.* **(d)** *Lorenz, $t_f = 0.1s.$, $\mu \approx 3.28 \times 10^{-2}$.*

Figure 9: *Comparison of performance of MC and BIMC. The variation of the relative RMSE, e_{RMS}, is plotted against the number of samples N. For reference, the most accurate probability estimate is also reported.*

Effect of probability magnitude. In Figure 10, we study the effect of the probability magnitude on the relative RMSE, $\tilde{e}_{\text{RMS}}$. We notice that BIMC is only weakly dependent on the probability magnitude. This is because selecting parameters by minimizing D_{KL} leads to an IS density that is optimally adapted for sampling around $\mathbb{Y}$.



(a) *Synthetic non-linear problem* **(b)** *Autoignition*

(c) *Elliptic PDE*

Figure 10: *Effect of varying probability levels. In this figure, we plot the variation of ensemble averaged relative $\tilde{e}_{\text{RMS}}$ with the number of samples N for various forward models with varying levels of probability. The varying probability levels are selected by moving $\mathbb{Y}$ to the tail regions of $p(\boldsymbol{x})$. The dashed lines indicate errors associated with MC while the solid lines indicate errors in BIMC.*

Extremely rare events. In our final experiment, we push BIMC to compute probabilities of extremely rare events. The rare events were constructed by shifting $\mathbb{Y}$ further and further into the tail region of the push forward of $p(\boldsymbol{x})$ under f. BIMC is able to compute extremely small probabilities using a modest number of samples. This experiment also corroborates our claim that the accuracy of our method is only weakly dependent on the probability magnitude μ. We also report the probability estimate resulting from linearizing $f(\boldsymbol{x})$ around $\boldsymbol{x}^{\text{mid}}$ and conclude that the linearized probability estimate is a good indicator of the order of magnitude of the true probability.

Table 2: *Extremely rare events, $N = 1000$.*

BIMC, $n = 1$		BIMC, $n = 5$		Linearized
$\tilde{\mu}^N$	$\tilde{e}^N_{\text{RMS}}$	$\tilde{\mu}^N$	$\tilde{e}^N_{\text{RMS}}$	μ_{lin}
3.6214×10^{-3}	3.24×10^{-2}	3.7892×10^{-3}	3.87×10^{-2}	5.9770×10^{-3}
2.3938×10^{-6}	6.00×10^{-2}	2.1421×10^{-6}	6.10×10^{-2}	1.8252×10^{-6}
5.4224×10^{-8}	6.64×10^{-2}	5.4310×10^{-8}	7.08×10^{-2}	4.2072×10^{-8}
5.7578×10^{-10}	1.04×10^{-1}	5.6271×10^{-10}	6.79×10^{-2}	3.8026×10^{-10}

(a) *Synthetic non-linear problem*

$n = 1$		$n = 5$		Linearized
$\tilde{\mu}^N$	$\tilde{e}^N_{\text{RMS}}$	$\tilde{\mu}^N$	$\tilde{e}^N_{\text{RMS}}$	μ_{lin}
4.3626×10^{-3}	2.47×10^{-2}	4.3688×10^{-3}	3.52×10^{-2}	4.5667×10^{-3}
1.1158×10^{-5}	3.91×10^{-2}	1.1278×10^{-5}	4.52×10^{-2}	8.4646×10^{-5}
7.6348×10^{-7}	5.86×10^{-2}	8.0428×10^{-7}	7.76×10^{-2}	3.8022×10^{-6}
3.5977×10^{-10}	9.69×10^{-2}	3.8106×10^{-10}	1.54×10^{-1}	2.6634×10^{-10}

(b) *Autoignition*

$n = 1$		$n = 5$		Linearized
$\tilde{\mu}^N$	$\tilde{e}^N_{\text{RMS}}$	$\tilde{\mu}^N$	$\tilde{e}^N_{\text{RMS}}$	μ_{lin}
2.6422×10^{-3}	5.05×10^{-2}	2.7045×10^{-3}	3.90×10^{-2}	2.2526×10^{-3}
5.6726×10^{-6}	9.44×10^{-2}	5.1764×10^{-6}	4.76×10^{-2}	4.1409×10^{-6}
8.4630×10^{-9}	4.94×10^{-2}	8.6889×10^{-9}	5.58×10^{-2}	8.7048×10^{-9}
8.2730×10^{-10}	4.99×10^{-2}	8.0669×10^{-10}	7.36×10^{-2}	9.1534×10^{-10}

(c) *Elliptic PDE*

Failure cases. Here, we report cases which caused BIMC to fail. Figure 11 shows MC and BIMC samples for the periodic forward problem. Because $f(\boldsymbol{x})$ has circular contours, the ideal IS density q^* has support over a circular region in $\mathbb{R}^2$. This is also evident from how the samples from q^* are spread. Using a single Gaussian distribution to approximate this complicated density results in a poor fit, and hence, failure of the BIMC method. The nature of the poor fit is noteworthy. The IS density approximates $q^*(\boldsymbol{x})$ well in the direction that is informed by the data. In the directions orthogonal to this data-informed direction, it inherits the covariance of $p(\boldsymbol{x})$, and as such, cannot approximate q^* as it curves around.

Also, notice that the pre-image $f^{-1}(\mathbb{Y})$ is the union of two disconnected regions in parameter space. As a result, the ideal IS density, q^*, has two modes, one near $[1, 1]^T$, and a weaker one near $[-1, 2.5]^T$. Which mode is discovered depends on the initial guess provided to the numerical optimization routine. Currently, there exists no robust mechanism in BIMC to discover all the modes of q^*. This is also the cause of failure when the Lorenz system is inverted over $t_f = 5$s.

Another route to failure occurs if the optimal parameters based on an analysis of the linearized inverse problem aren't appropriate for the full non-linear problem. While we don't expect the two to be exactly equal, we implicitly assume that they will be close enough, and serious problems may occur if they're not. For instance, if the pseudo-likelihood variance from the linearized analysis is much smaller than the (unknown) optimal pseudo-likelihood variance for the full non-linear problem, then large IS weights may be observed, leading to biased estimates of the failure probability.

Finally, BIMC can also fail when the solution of the inverse problem cannot be computed. This happens when the Lorenz problem is simulated over a much longer time horizon, $t_f = 15$s. In this case, the optimizer failed to identify a descent direction and converge to a minimum. Physically, this happens because of the chaotic nature of the problem. Since all trajectories of the Lorenz system eventually settle on the attractor, going from a point on the attractor back in time is a highly ill-conditioned problem.

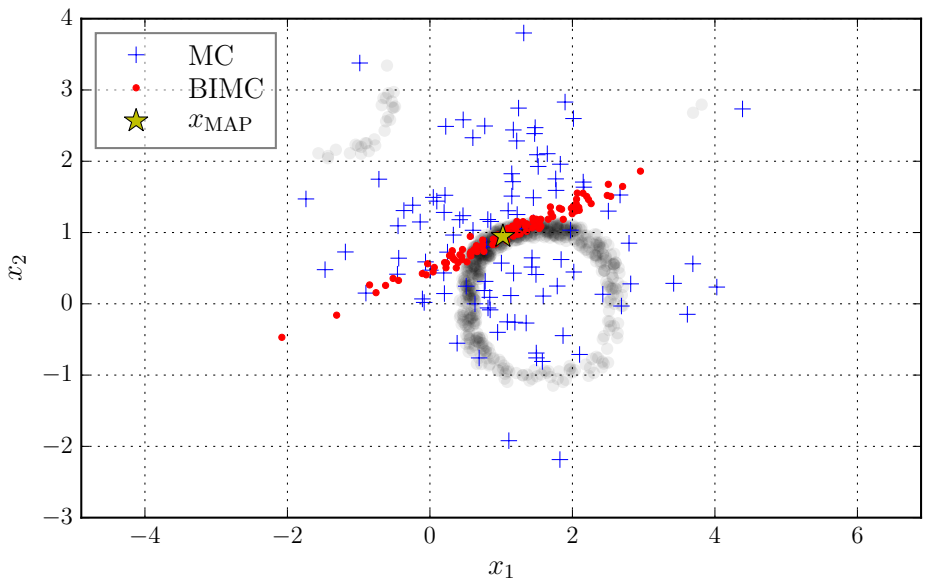


Figure 11: *A failure case. Here, $f(\boldsymbol{x}) = \sin(x_1)\cos(x_2)$ is a periodic function in $\mathbb{R}^2$. Gray markers depict samples from the ideal IS density q^* in this case.*

Summary. In summary, the effectiveness of BIMC depends on the interplay between the directions *not* informed by the pseudo-data point, and the variation of the forward map in these directions. If, at the scale of the covariance of the nominal density p, $f(\boldsymbol{x})$ varies too quickly in these directions (like the Periodic example), the PDF constructed in BIMC will make for a poor IS density. On the other hand, if $f(\boldsymbol{x})$ varies slowly enough (as in the synthetic non-linear, and autoignition examples) or not at all (the affine case), then BIMC is effective. Thus, we conclude that BIMC is best suited to forward maps that are weakly non-linear at the scale of the covariance of the nominal density p. Physically, this means that the uncertainties in the input parameters must small enough that f appears almost linear. Note that f can still

be highly non-linear at larger scales.

Apart from the forward map being only weakly non-linear, there are two additional requirements. The regions in parameter space that evaluate inside $\mathbb{Y}$ should not be disjoint. The final and perhaps the most important requirement is that the solution of the inverse problem must be computable.

5. Conclusion and future work. In this article, we addressed the problem of efficiently computing rare-event probabilities in systems with uncertain input parameters. Our approach, called BIMC, employs importance sampling in order to achieve efficiency. Noting the structural similarity between the (theoretical) ideal importance sampling density and the posterior distribution of a fictitious inference problem, our importance sampling distribution is constructed by approximating such a fictitious posterior via a Gaussian distribution. The approximation process allows the incorporation of the derivatives of the input-output map into the importance sampling distribution, which is how our scheme achieves parsimonious sampling. Our theoretical analysis establishes that this procedure is optimal in the setting where the input-output map is affine and the nominal density is Gaussian. Hence, BIMC is best applied to maps that appear nearly affine at the scale of the covariance of the nominal distribution. Our numerical experiments support this conclusion and demonstrate that when this is the case, BIMC can lead to speedups of several orders-of-magnitude. Experiments also reveal several drawbacks in BIMC. We will concern ourselves with fixing these drawbacks in part II of this paper.

Acknowledgments. We would like to acknowledge Umberto Villa's assistance in setting up the Elliptic PDE example.

REFERENCES

[1] V. Akçelik, G. Biros, O. Ghattas, J. Hill, D. Keyes, and B. van Bloemen Waanders, *Parallel algorithms for PDE-constrained optimization*, in Parallel Processing for Scientific Computing, SIAM, 2006, pp. 291–322.

[2] S. Au and J. Beck, *A new adaptive importance sampling scheme for reliability calculations*, Structural Safety, 21 (1999), pp. 135 – 158, https://doi.org/https://doi.org/10.1016/S0167-4730(99)00014-4, http://www.sciencedirect.com/science/article/pii/S0167473099000144.

[3] S. Au, C. Papadimitriou, and J. Beck, *Reliability of uncertain dynamical systems with multiple design points*, Structural Safety, 21 (1999), pp. 113 – 133, https://doi.org/https://doi.org/10.1016/S0167-4730(99)00009-0, http://www.sciencedirect.com/science/article/pii/S0167473099000090.

[4] S.-K. Au and J. L. Beck, *Estimation of small failure probabilities in high dimensions by subset simulation*, Probabilistic Engineering Mechanics, 16 (2001), pp. 263 – 277, https://doi.org/https://doi.org/10.1016/S0266-8920(01)00019-4, http://www.sciencedirect.com/science/article/pii/S0266892001000194.

[5] J. Breidt, T. Butler, and D. Estep, *A measure-theoretic computational method for inverse sensitivity problems I: Method and analysis*, SIAM Journal on Numerical Analysis, 49 (2011), pp. 1836–1859, https://doi.org/10.1137/100785946, https://doi.org/10.1137/100785946, https://arxiv.org/abs/https://doi.org/10.1137/100785946.

[6] C. G. Bucher, *Adaptive sampling - an iterative fast monte carlo procedure*, Structural Safety, 5 (1988), pp. 119 – 126, https://doi.org/https://doi.org/10.1016/0167-4730(88)90020-3, http://www.sciencedirect.com/science/article/pii/0167473088900203.

[7] T. Butler, D. Estep, and J. Sandelin, *A computational measure theoretic approach to inverse sensitivity problems II: A posteriori error analysis*, SIAM Journal on Numerical Analysis, 50 (2012), pp. 22–

45, https://doi.org/10.1137/100785958, https://doi.org/10.1137/100785958, https://arxiv.org/abs/ https://doi.org/10.1137/100785958.

[8] P.-T. De Boer, D. P. Kroese, S. Mannor, and R. Y. Rubinstein, *A tutorial on the cross-entropy method*, Annals of Operations Research, 134 (2005), pp. 19–67.

[9] M. Fasiolo, F. E. de Melo, and S. Maskell, *Langevin incremental mixture importance sampling*, Statistics and Computing, 28 (2018), pp. 549–561.

[10] M. Gunzburger, *Perspectives in Flow Control and Optimization*, Society for Industrial and Applied Mathematics, 2002, https://doi.org/10.1137/1.9780898718720, https://epubs.siam.org/doi/abs/10.1137/1.9780898718720, https://arxiv.org/abs/https://epubs.siam.org/doi/pdf/10.1137/1.9780898718720.

[11] J. R. Hershey and P. A. Olsen, *Approximating the Kullback Leibler Divergence Between Gaussian Mixture Models*, in 2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07, vol. 4, April 2007, pp. IV–317–IV–320, https://doi.org/10.1109/ICASSP.2007.366913.

[12] T. Isaac, N. Petra, G. Stadler, and O. Ghattas, *Scalable and efficient algorithms for the propagation of uncertainty from data through inference to prediction for large-scale problems, with application to flow of the antarctic ice sheet*, Journal of Computational Physics, 296 (2015), pp. 348 – 368, https://doi.org/https://doi.org/10.1016/j.jcp.2015.04.047, http://www.sciencedirect.com/science/article/pii/S0021999115003046.

[13] H. Kahn and A. W. Marshall, *Methods of reducing sample size in Monte Carlo computations*, Journal of the Operations Research Society of America, 1 (1953), pp. 263–278.

[14] L. Katafygiotis and K. Zuev, *Estimation of small failure probabilities in high dimensions by adaptive linked importance sampling*, COMPDYN 2007, (2007).

[15] B. Kramer, A. N. Marques, B. Peherstorfer, U. Villa, and K. Willcox, *Multifidelity probability estimation via fusion of estimators*, Tech. Report ACDL TR-2017-3, Massachusetts Institute of Technology, 2017.

[16] C. Lieberman and K. Willcox, *Goal-oriented inference: Approach, linear theory, and application to advection diffusion*, SIAM Review, 55 (2013), pp. 493–519, https://doi.org/10.1137/130913110, https://doi.org/10.1137/130913110, https://arxiv.org/abs/https://doi.org/10.1137/130913110.

[17] C. Lieberman and K. Willcox, *Nonlinear goal-oriented Bayesian inference: Application to carbon capture and storage*, SIAM Journal on Scientific Computing, 36 (2014), pp. B427–B449.

[18] E. N. Lorenz, *Deterministic nonperiodic flow*, Journal of the Atmospheric Sciences, 20 (1963), pp. 130–141.

[19] R. Melchers, *Importance sampling in structural systems*, Structural Safety, 6 (1989), pp. 3 – 10, https://doi.org/https://doi.org/10.1016/0167-4730(89)90003-9, http://www.sciencedirect.com/science/article/pii/0167473089900039.

[20] A. B. Owen, *Monte Carlo theory, methods and examples*, 2013.

[21] B. Peherstorfer, T. Cui, Y. Marzouk, and K. Willcox, *Multifidelity importance sampling*, Computer Methods in Applied Mechanics and Engineering, 300 (2016), pp. 490 – 509, https://doi.org/http://dx.doi.org/10.1016/j.cma.2015.12.002, http://www.sciencedirect.com/science/article/pii/S004578251500393X.

[22] B. Peherstorfer, B. Kramer, and K. Willcox, *Multifidelity preconditioning of the cross-entropy method for rare event simulation and failure probability estimation*, SIAM/ASA Journal on Uncertainty Quantification, (2018).

[23] R. Rackwitz, *Reliability analysis - a review and some perspectives*, Structural Safety, 23 (2001), pp. 365 – 395, https://doi.org/https://doi.org/10.1016/S0167-4730(02)00009-7, http://www.sciencedirect.com/science/article/pii/S0167473002000097.

[24] A. E. Raftery and L. Bao, *Estimating and projecting trends in HIV/AIDS generalized epidemics using incremental mixture importance sampling*, Biometrics, 66 (2010), pp. 1162–1173.

[25] C. P. Robert and G. Casella, *Monte Carlo Integration*, Springer New York, New York, NY, 2004, pp. 79–122, https://doi.org/10.1007/978-1-4757-4145-2_3, http://dx.doi.org/10.1007/978-1-4757-4145-2_3.

[26] G. Rubino and B. Tuffin, *Rare event simulation using Monte Carlo methods*, John Wiley & Sons, 2009.

[27] G. Schuëller and R. Stix, *A critical appraisal of methods to determine failure probabilities*, Structural

Safety, 4 (1987), pp. 293 – 309, https://doi.org/https://doi.org/10.1016/0167-4730(87)90004-X, http://www.sciencedirect.com/science/article/pii/016747308790004X.

[28] A. Spantini, T. Cui, K. Willcox, L. Tenorio, and Y. Marzouk, *Goal-oriented optimal approximations of Bayesian linear inverse problems*, SIAM Journal on Scientific Computing, 39 (2017), pp. S167–S196, https://doi.org/10.1137/16M1082123, https://doi.org/10.1137/16M1082123, https://arxiv.org/abs/https://doi.org/10.1137/16M1082123.

[29] U. Villa, N. Petra, and O. Ghattas, *hIPPYlib: an extensible software framework for large-scale deterministic and linearized bayesian inversion*, (2016), http://hippylib.github.io.

SUPPLEMENTARY MATERIALS: BIMC: The Bayesian Inverse Monte Carlo method for goal-oriented uncertainty quantification

Siddhant Wahal* and George Biros*

SM1. Analysis of the affine-Gaussian case. In this section, we offer detailed derivations of claims made in the main text. In what follows we assume:

- $f(\boldsymbol{x})$ is affine and defined as $f(\boldsymbol{x}) = \boldsymbol{v}^T\boldsymbol{x} + \beta$ for some $\boldsymbol{v} \in \mathbb{R}^m$, $\beta \in \mathbb{R}$,
- the nominal density $p(\boldsymbol{x})$ is Gaussian with mean $\boldsymbol{x}_0$ and covariance $\boldsymbol{\Sigma}_0$, $p(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}_0, \boldsymbol{\Sigma}_0)$,
- the pseudo-data is some $y \in \mathbb{Y}$ and pseudo-likelihood density $p(y|\boldsymbol{x}) = \mathcal{N}(f(\boldsymbol{x}), \sigma^2)$.

SM1.1. Kullback Leibler divergence. Here, we derive an analytical expression for the Kullback Leibler divergence between the ideal IS distribution, $q^*(\boldsymbol{x})$, for this problem, and the IS density, $q(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}_{\text{MAP}}, \mathbf{H}_{\text{GN}}^{-1})$.

By definition of the KL divergence, we have,

$$\begin{aligned} D_{\text{KL}}\left(q^*||q\right) &= \int_{\mathbb{R}^m} \frac{\mathbf{1}_{\mathbb{Y}}(f(\boldsymbol{x}))p(\boldsymbol{x})}{\mu} \log \frac{\mathbf{1}_{\mathbb{Y}}\left(f(\boldsymbol{x})\right)p(\boldsymbol{x})}{\mu q(\boldsymbol{x})} \mathrm{d}\boldsymbol{x} \\ &= \frac{1}{\mu}\int_{f^{-1}(\mathbb{Y})} p(\boldsymbol{x})\Big(\log\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})} - \log\mu\Big)\mathrm{d}\boldsymbol{x} \\ &= \frac{1}{\mu}\int_{f^{-1}(\mathbb{Y})} p(\boldsymbol{x})\log\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}\mathrm{d}\boldsymbol{x} - \log\mu. \end{aligned} \tag{SM1}$$

where $f^{-1}(\mathbb{Y}) = \{\boldsymbol{x} \in \mathbb{R}^m : \boldsymbol{v}^T\boldsymbol{x} + \beta \in \mathbb{Y}\}$.

Note that the pushforward density of $p(\boldsymbol{x})$ under $f(\boldsymbol{x})$ is also a Gaussian with mean $\nu = \boldsymbol{v}^T\boldsymbol{x}_0 + \beta$, and variance $\gamma^2 = \boldsymbol{v}^T\boldsymbol{\Sigma}_0\boldsymbol{v}$. Let $\rho^2 = \sigma^2/(\sigma^2+\gamma^2)$. It can be shown that

$$\log\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})} = \frac{\left(y - f(\boldsymbol{x})\right)^2}{2\sigma^2} + \log\rho - \frac{(y-\nu)^2}{2(\sigma^2+\gamma^2)}. \tag{SM2}$$

Therefore,

$$\begin{aligned} \int_{f^{-1}(\mathbb{Y})} p(\boldsymbol{x})\log\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}\mathrm{d}\boldsymbol{x} &= \int_{f^{-1}(\mathbb{Y})} \left(\frac{\left(y-f(\boldsymbol{x})\right)^2}{2\sigma^2} + \log\rho - \frac{(y-\nu)^2}{2(\sigma^2+\gamma^2)}\right) p(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\ &= \int_{f^{-1}(\mathbb{Y})} \frac{\left(y-f(\boldsymbol{x})\right)^2}{2\sigma^2} p(\boldsymbol{x})\mathrm{d}\boldsymbol{x} + \left(\log\rho - \frac{(y-\nu)^2}{2(\sigma^2+\gamma^2)}\right)\mu. \end{aligned} \tag{SM3}$$

To continue further, we change variables to $z = f(\boldsymbol{x})$ and take advantage of our knowledge of the probability density for z, p_Z. This probability density is nothing but the push forward of $p(\boldsymbol{x})$ under f. Thus, $p_Z = \mathcal{N}\left(\nu, \gamma^2\right)$.

* Oden Institute for Computational Engineering and Sciences, The University of Texas at Austin, Austin, TX, 78712, (siddhant@ices.utexas.edu, gbiros@acm.org).

$$\begin{aligned}\int_{f^{-1}(\mathbb{Y})} p(\boldsymbol{x}) \log \frac{p(\boldsymbol{x})}{q(\boldsymbol{x})} \mathrm{d}\boldsymbol{x} &= \int_{\mathbb{Y}} \frac{(y-z)^2}{2\sigma^2} p_Z(z) \mathrm{d}z + \left(\log \rho - \frac{(y-\nu)^2}{2(\sigma^2+\gamma^2)} \right) \mu \\ &= \frac{y^2}{2\sigma^2}\mu + \frac{1}{2\sigma^2} \int_{\mathbb{Y}} z^2 p_Z(z) \mathrm{d}z - \frac{y}{\sigma^2} \int_{\mathbb{Y}} z p_Z(z) \mathrm{d}z \\ &\quad + \left(\log \rho - \frac{(y-\nu)^2}{2(\sigma^2+\gamma^2)} \right) \mu.\end{aligned} \tag{SM4}$$

Note that the two integrals in the final equation can be related to the mean and variance of a truncated normal distribution. Let $p_T(z)$ be the truncated distribution when p_Z only has support on the interval $\mathbb{Y}$. This truncated distribution is in fact the push forward under $f(\boldsymbol{x})$ of $q^*(\boldsymbol{x})$. Thus, $p_T = \mathbf{1}_{\mathbb{Y}}(z) p_Z / \mu$. We denote the mean and variance of p_T by ν_T and ${\gamma_T}^2$ respectively.

Finally,

$$\int_{f^{-1}(\mathbb{Y})} p(\boldsymbol{x}) \log \frac{p(\boldsymbol{x})}{q(\boldsymbol{x})} \mathrm{d}\boldsymbol{x} = \left(\frac{(y-\nu_T)^2 + {\gamma_T}^2}{2\sigma^2} + \log \rho - \frac{(y-\nu)^2}{2(\sigma^2+\gamma^2)} \right) \mu. \tag{SM5}$$

Computing ν_T and γ_T is prone to catastrophic cancellation. We recommend using the scaled complementary error function, `erfcx`, for this purpose, following [SM1]. The final expression for the KL divergence is,

$$D_{\mathrm{KL}}(q^* || q) = \frac{(y-\nu_T)^2 + {\gamma_T}^2}{2\sigma^2} - \frac{(y-\nu)^2}{2(\sigma^2+\gamma^2)} + \log \rho - \log \mu. \tag{SM6}$$

This expression is minimized when

$$\begin{aligned} y^* &= \frac{\nu_T \gamma^2 - \nu \gamma_T^2}{\gamma^2 - \gamma_T^2} \\ \sigma^{*2} &= \frac{\gamma_T^2 \gamma^2}{\gamma^2 - \gamma_T^2} \end{aligned} \tag{SM7}$$

SM1.2. Establishing BIMC optimality. In this sub-section, we provide proof of Claim 3.1. We assume without loss of generality that $\boldsymbol{x}_0 = \mathbf{0}$ and $\boldsymbol{\Sigma}_0 = \mathbf{I}$. Further, denote by $\hat{\boldsymbol{v}} = \boldsymbol{v}/\|\boldsymbol{v}\|$. Now, select $\hat{\mathbf{V}}$ so that it's columns form an orthonormal basis for $\mathbb{R}^m \backslash \operatorname{span}(\hat{\boldsymbol{v}})$. Then, we have, for any $\boldsymbol{x}$:

$$\begin{aligned} \boldsymbol{x} &= \hat{\boldsymbol{v}}\hat{\boldsymbol{v}}^T \boldsymbol{x} + \hat{\mathbf{V}}\hat{\mathbf{V}}^T \boldsymbol{x} \\ &= \hat{\boldsymbol{v}} x_1 + \hat{\mathbf{V}} \boldsymbol{x}_\perp. \end{aligned} \tag{SM8}$$

where $x_1 := \hat{\boldsymbol{v}}^T\boldsymbol{x} \in \mathbb{R}$ and $\boldsymbol{x}_\perp := \hat{\mathbf{V}}^T\boldsymbol{x} \in \mathbb{R}^{m-1}$.

Using this decomposition of the parameter space, notice that

$$\begin{aligned} f(\boldsymbol{x}) &= \boldsymbol{v}^T\boldsymbol{x} + \beta \\ &= \|\boldsymbol{v}\|x_1 + \beta \\ &= \hat{f}(x_1), \end{aligned}$$

and,

$$p(\boldsymbol{x}) = p_1(x_1)p_\perp(\boldsymbol{x}_\perp), \tag{SM9}$$

where, $p_1(x_1) = \mathcal{N}(0,1)$ and $p_\perp = \mathcal{N}(\mathbf{0}_\perp, \mathbf{I}_\perp)$. By $\mathbf{0}_\perp$ and $\mathbf{I}_\perp$, we refer to the $m-1$ dimensional zero vector and identity matrix respectively. Hence, such a decomposition of the parameter space enables one to express ideal IS density, q^*, as

$$\begin{aligned} q^*(\boldsymbol{x}) &= \frac{\mathbf{1}_{\mathbb{Y}}(f(\boldsymbol{x}))p(\boldsymbol{x})}{\mu} \\ &= \frac{\mathbf{1}_{\mathbb{Y}}(f(x_1))p_1(x_1)}{\mu}p_\perp(\boldsymbol{x}_\perp) \\ &= q_1^*(x_1)p_\perp(\boldsymbol{x}_\perp). \end{aligned} \tag{SM10}$$

Before we establish the optimality of the BIMC algorithm, we will require the following result on the structure of the optimal Gaussian approximation of $q^*(\boldsymbol{x}_1, \boldsymbol{x}_\perp)$.

Proposition SM1.1. *The Gaussian distribution closest in KL divergence to* $q^*(x_1, \boldsymbol{x}_\perp) = \mathbf{1}_{\mathbb{Y}}(\hat{f}(x_1))p_1(x_1)p_\perp(\boldsymbol{x}_\perp)/\mu$ *must be of the form* $q_{1,\text{opt}}(x_1)p_\perp(\boldsymbol{x}_\perp)$*, where* $q_{1,\text{opt}}(x_1)$ *is the closest Gaussian approximation of* $\mathbf{1}_{\mathbb{Y}}(\hat{f}(x_1))p_1(x_1)/\mu$.

Proof. Let $\mathcal{G}^m$ denote the set of all Gaussian distributions over $\mathbb{R}^m$. Any Gaussian distribution in $\mathbb{R}^m$ can be expressed as a joint density over the variables $x_1, \boldsymbol{x}_\perp$, $q(x_1, \boldsymbol{x}_\perp)$. Then, the closest Gaussian approximation of $q^*(x_1, \boldsymbol{x}_\perp)$ is given by:

$$q_{\text{opt}}(x_1, \boldsymbol{x}_\perp) = \underset{q \in \mathcal{G}^m}{\arg\min}\, D_{\text{KL}}\left(q^*(x_1, \boldsymbol{x}_\perp)||q(x_1, \boldsymbol{x}_\perp)\right).$$

Using the definition of KL divergence, we have,

$$D_{\text{KL}}\left(q^*(x_1, \boldsymbol{x}_\perp)||q(x_1, \boldsymbol{x}_\perp)\right) = \int q^*(x_1, \boldsymbol{x}_\perp)\log\frac{q^*(x_1, \boldsymbol{x}_\perp)}{q(x_1, \boldsymbol{x}_\perp)}\mathrm{d}x_1\mathrm{d}\boldsymbol{x}_\perp. \tag{SM11}$$

To simplify notation, denote $\mathbf{1}_{\mathbb{Y}}(\hat{f}(x_1))p_1(x_1)/\mu$ with $q_1^*(\boldsymbol{x}_1)$. Now, using the fact that $q(x_1, \boldsymbol{x}_\perp) = q_1(x_1)q_\perp(\boldsymbol{x}_\perp|x_1)$, the chain rule of KL divergence gives:

$$D_{\mathrm{KL}}(q^*(x_1, \boldsymbol{x}_\perp)||q(x_1, \boldsymbol{x}_\perp)) \\ = D_{\mathrm{KL}}(q_1^*(x_1)||q_1(x_1)) + \mathbb{E}_{q_1^*}\left[D_{\mathrm{KL}}(p_\perp(\boldsymbol{x}_\perp)||q_\perp(\boldsymbol{x}_\perp|x_1))\right].$$

Then, because both terms are positive,

$$\begin{aligned} q_{\mathrm{opt}} &= \arg\min_{q\in\mathcal{G}^m} D_{\mathrm{KL}}(q^*(x_1, \boldsymbol{x}_\perp)||q(x_1, \boldsymbol{x}_\perp)) \\ &= \arg\min_{q_1(x_1)q_\perp(\boldsymbol{x}_\perp|x_1)\in\mathcal{G}^m} D_{\mathrm{KL}}(q_1^*(x_1)||q_1(x_1)) \\ &\quad + \arg\min_{q_1(x_1)q_\perp(\boldsymbol{x}_\perp|x_1)\in\mathcal{G}^m} \mathbb{E}_{q_1^*}\left[D_{\mathrm{KL}}(p_\perp(\boldsymbol{x}_\perp)||q_\perp(\boldsymbol{x}_\perp|x_1))\right]. \end{aligned}$$

Since the KL divergence is a positive quantity, $\mathbb{E}_{q_1^*}\left[D_{\mathrm{KL}}(p_\perp(\boldsymbol{x}_\perp)||q_\perp(\boldsymbol{x}_\perp|x_1))\right] \geq 0$. The equality is achieved iff $q_\perp(\boldsymbol{x}_\perp|x_1) = p_\perp(\boldsymbol{x}_\perp)$. Hence, $D_{\mathrm{KL}}(p_\perp(\boldsymbol{x}_\perp)||q_\perp(\boldsymbol{x}_\perp|x_1))$ is minimized when $q_\perp(\boldsymbol{x}_\perp|x_1) = p_\perp(\boldsymbol{x}_\perp)$. By inspection, the first term is minimized by $q_{1,\mathrm{opt}}(x_1)$, the closest Gaussian approximation of $q_1^*(x_1)$. Thus, $q_{\mathrm{opt}} = q_{1,\mathrm{opt}}(x_1)p_\perp(\boldsymbol{x}_\perp)$. ∎

Proposition SM1.1 states that for the affine-Gaussian case, q_{opt} must marginalize to the corresponding marginal of $p(\boldsymbol{x})$ in all directions in parameter space to which $f(\boldsymbol{x})$ is insensitive. Note that in the proof, the affine property of $f(\boldsymbol{x})$ was never explicitly used. Hence, this result can be easily extended to arbitrary non-linear $f(\boldsymbol{x})$ in the following sense. If there are *global* directions in parameter space to which $f(\boldsymbol{x})$ is insensitive, an optimal Gaussian or Gaussian mixture approximation of q^* must marginalize to the corresponding marginal of $p(\boldsymbol{x})$ in those directions.

Next, we prove Claim 3.1.

SM1.2.1. Proof of Claim 3.1.

Proof. We begin by showing that BIMC produces a Gaussian distribution that indeed satisfies the form prescribed by Proposition SM1.1. Let the IS distribution produced by BIMC be denoted $q_{\mathrm{BIMC}}(x)$. Because q_{BIMC} is a Bayesian posterior distribution, it has the following form:

$$q_{\mathrm{BIMC}}(\boldsymbol{x}) = \frac{p(y|\boldsymbol{x})p(\boldsymbol{x})}{p(y)}, \tag{SM12}$$

where,

$$\begin{aligned} p(y|\boldsymbol{x}) &\propto \exp\left(-\frac{1}{2\sigma^2}(y - f(\boldsymbol{x}))^2\right) \\ &\propto \exp\left(-\frac{1}{2\sigma^2}(y - \hat{f}(x_1))^2\right) \\ &\propto p(y|x_1). \end{aligned}$$

Also, since $p(\boldsymbol{x})$ is a standard Gaussian, it can be expressed as $p(\boldsymbol{x}) = p_1(\boldsymbol{x}_1)p_\perp(\boldsymbol{x}_\perp)$. Therefore,

$$q_{\text{BIMC}}(\boldsymbol{x}) \propto p(y|x_1)p_1(x_1)p_\perp(\boldsymbol{x}_\perp). \tag{SM13}$$

Hence, the marginal distribution of $\boldsymbol{x}_\perp$ when $x \sim q_{\text{BIMC}}(\boldsymbol{x})$ is $p_\perp(\boldsymbol{x}_\perp)$, as required by Proposition SM1.1. Additionally, since $\hat{f}(x_1)$ is affine, $p(y|x_1)p_1(x_1)$ is Gaussian. All that remains to be shown is that (13) results in parameters such that $p(y|x_1)p_1(x_1)$ is the optimal Gaussian approximation of $q_1^*(x_1) = \mathbf{1}_{\mathbb{Y}}(\hat{f})(\boldsymbol{x}_1)p_1(x_1)/\mu$.

Now, notice that $q_1^*(x_1) = \mathbf{1}_{\mathbb{Y}}(\hat{f}(x_1))p_1(x_1)/\mu$ is a truncated normal distribution. That is because

$$\begin{aligned} q_1^*(x_1) &= \mathbf{1}_{\mathbb{Y}}(\hat{f}(x_1))p_1(x_1)/\mu \\ &= \mathbf{1}_{[y_{\min},y_{\max}]}(\|\boldsymbol{v}\|x_1 + \beta)p_1(x_1)/\mu \\ &= \mathbf{1}_{[\frac{y_{\min}-\beta}{\|\boldsymbol{v}\|},\frac{y_{\max}-\beta}{\|\boldsymbol{v}\|}]}(x_1)p_1(x_1)/\mu. \end{aligned}$$

Thus, q_1^* is a standard normal distribution truncated over $\left[\frac{y_{\min}-\beta}{\|\boldsymbol{v}\|}, \frac{y_{\max}-\beta}{\|\boldsymbol{v}\|}\right]$. Denote the mean and variance of q_1^* by ν_T and γ_T^2. Then the optimal Gaussian approximation of q_1^* is $\mathcal{N}(\phi_T, \omega_T^2)$. Hence, we have to demonstrate that the minimizer of (13) are such that $p(y|x_1)p_1(x_1)$ is the same as $\mathcal{N}(\phi_T, \omega_T^2)$.

The expression for $D_{\text{KL}}(q^*(x_1, \boldsymbol{x}_\perp)||q(x_1, \boldsymbol{x}_\perp)$ in this case is:

$$\begin{aligned} D_{\text{KL}}(q^*(x_1, \boldsymbol{x}_\perp)||q(x_1, \boldsymbol{x}_\perp)) = \text{const.} \; &+ \frac{\|\boldsymbol{v}\|^2}{2\sigma^2}\left(\omega_T^2 + \left(\frac{y-\beta}{\|\boldsymbol{v}\|} - \phi_T\right)^2\right) \\ &- \frac{1}{2}\frac{(y-\beta)^2}{\sigma^2 + \|\boldsymbol{v}\|^2} + \frac{1}{2}\log\frac{\sigma^2}{\sigma^2 + \|\boldsymbol{v}\|^2}. \end{aligned}$$

This expression is minimized in the BIMC algorithm. The minimum occurs at

$$\begin{aligned} y^* &= \frac{\|\boldsymbol{v}\|}{1-\omega_T^2}\phi_T + \beta \\ \sigma^{*2} &= \frac{\|\boldsymbol{v}\|^2\omega_T^2}{1-\omega_T^2}. \end{aligned} \tag{SM14}$$

Since $\omega_T^2 < 1$ (Remark 2.1 in [SM6]), σ^{*2} is a valid variance for the likelihood distribution. Now, it can be shown that,

$$\frac{p(y^*|x_1)p_1(x_1)}{p(y^*)} = \mathcal{N}\left(\frac{\|\boldsymbol{v}\|}{\|\boldsymbol{v}\|^2 + \sigma^{*2}}(y^* - \beta), \frac{\sigma^{*2}}{\sigma^{*2} + \|\boldsymbol{v}\|^2}\right).$$

Plugging in expressions for y^* and σ^{*2} from (SM14), we obtain,

$$\frac{p(y^*|x_1)p_1(x_1)}{p(y^*)} = \mathcal{N}(\phi_T, \omega_T^2).$$

Therefore, $p(y^*|x_1)p_1(x_1)/p(y^*)$ is the Gaussian closest in KL divergence to $\mathbf{1}_{\mathbb{Y}}(\hat{f}(\boldsymbol{x}_1)p_1(\boldsymbol{x}_1)/\mu$. Hence, by Proposition SM1.1, $q_{\text{BIMC}}(\boldsymbol{x}) = p(y^*|x_1)p_1(x_1)p_\perp(\boldsymbol{x}_\perp)/p(y^*)$ must be the closest Gaussian distribution to $q^*(\boldsymbol{x}_r, \boldsymbol{x}_\perp)$ ∎

SM2. Implementation details.

SM2.1. The affine case. We construct a affine map from $\mathbb{R}^m$ to $\mathbb{R}$. The map is defined as

$$f(\boldsymbol{x}) = \boldsymbol{o}^T \mathbf{A} \boldsymbol{x}, \tag{SM15}$$

where, $\boldsymbol{o} = \left(\frac{1}{m}, \frac{1}{m}, \ldots, \frac{1}{m}\right)^T \in \mathbb{R}^m$ is an observation operator and

$$\mathbf{A} = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & \frac{1}{2} & 0 & \cdots & 0 \\ 0 & 0 & \frac{1}{3} & & \vdots \\ \vdots & \vdots & & \ddots & 0 \\ 0 & 0 & \cdots & 0 & \frac{1}{m} \end{pmatrix}. \tag{SM16}$$

SM2.1.1. Implementation. We assume $p(\boldsymbol{x})$ is a Gaussian distribution with mean $\boldsymbol{x}_0$ and covariance $\boldsymbol{\Sigma}_0$, where $\boldsymbol{x}_0 = (1, 1, \ldots, 1)^T \in \mathbb{R}^m$ and $\boldsymbol{\Sigma_0} = 0.1\mathbf{I} \in \mathbb{R}^{m \times m}$. Here, $\mathbf{I}$ is the m dimensional identity matrix. We chose $\mathbb{Y} = [1.2803, 1.4571]$ when $m = 2$ and $\mathbb{Y} = \left[6.2 \times 10^{-2}, 6.3 \times 10^{-3}\right]$ when $m = 100$. The MAP point was computed using MATLAB's `fminunc` routine.

SM2.2. Synthetic non-linear case. For quick testing, we construct the following non-linear problem from $\mathbb{R}^m$ to $\mathbb{R}$:

$$f(\boldsymbol{x}) = \boldsymbol{o}^T \boldsymbol{u}, \text{ where,} \tag{SM17}$$

$$\left(\mathbf{S} + \varepsilon \boldsymbol{x}\boldsymbol{x}^T\right) \boldsymbol{u} = \boldsymbol{b}. \tag{SM18}$$

SM2.2.1. Implementation. Again, $\boldsymbol{o} \in \mathbb{R}^m$ is an observation operator. For this problem, we chose $\boldsymbol{o} = [1, 0, \ldots, 0]^T$. $\mathbf{S}$ is a randomly chosen symmetric positive definite matrix, while $\boldsymbol{b}$ is a randomly chosen vector whose entries are distributed according to the standard normal distribution. We set $\varepsilon = 0.01\|\mathbf{S}\|_2$. The nominal probability density in this case is $p(\boldsymbol{x})$ is a Gaussian with mean $\boldsymbol{x}_0 = (1, 1, \ldots, 1)^T$ and covariance $\boldsymbol{\Sigma_0} = 0.01\mathbf{I}$. The target interval $\mathbb{Y}$ is chosen to be $[1.24, 1.25]$ when $m = 10$ and $[0.919, 0.923]$ when $m = 2$. MATLAB's `fminunc` routine was used for optimization again.

SM2.3. Single step reaction. The single step reaction is described by the following ODE:

$$\begin{aligned} &\frac{\mathrm{d}u}{\mathrm{d}t} = \frac{S^*(u)}{\tau_{\mathrm{R}}}, \quad 0 < t < t_f, \\ &S^*(u) = Bu(1-u)\exp\left(\frac{-T_{\mathrm{Act}}}{T_{\mathrm{u}} + (T_{\mathrm{b}} - T_{\mathrm{u}})u}\right), \end{aligned} \tag{SM19}$$

and we define $x = u(0)$ and $f(x) = u(t_f)$.

This equation uses the Arrhenius equation to describe the rate of a chemical reaction in terms of a progress variable u. The progress variable is routinely employed in the analyses of turbulent flames and is 0 in regions of pure reactants and 1 in pure products [SM2].

Here, τ_{R} is the time scale of the reaction and $S^*(u)$ is the normalized source term. The numerical constant in $S^*(u)$, B, ensures that it integrates to unity, T_{Act} is the activation temperature, T_{u} is the temperature of the unburnt reactants and T_{b} the temperature of the burnt products.

Since u is always bounded between 0 and 1, the differential operator defined in (SM19) is a map from $[0, 1]$ to $[0, 1]$.

SM2.3.1. Implementation. For our numerical experiments, we set, $T_{\mathrm{u}} = 300$ K, $T_{\mathrm{b}} = 2100$ K, $T_{\mathrm{Act}} = 30{,}000$ K, $B = 6.11 \times 10^7$, $\tau_{\mathrm{R}} = 1$s.

Further, we choose $\mathbb{Y} = [0.7, 0.8]$ and $p(x) = \mathcal{N}(0.5, 0.01)$. We used MATLAB's `fmincon` routine to perform constrained optimization in order to compute the MAP point.

SM2.4. Hydrogen Autoignition. We observe the heat released, Q, during autoignition of a hydrogen-air mixture in an adiabatic, constant pressure, fixed mass reactor. To describe the chemistry, we use a reduced mechanism that involves 5 elementary reactions among 8 chemical species - $\mathrm{H_2}, \mathrm{O_2}, \mathrm{N_2}, \mathrm{HO_2}, \mathrm{H}, \mathrm{O}, \mathrm{OH}, \mathrm{H_2O}$ [SM5]. We assume-

- reactants are ideal gases,
- there are no spatial gradients of temperature or species concentrations,
- the volume of the reactor can change to keep the pressure constant,
- only $\mathrm{H_2}, \mathrm{O_2}$, and $\mathrm{N_2}$ are present in the reactor initially.

Then specifying the pressure (P), temperature (T), and equivalence ratio (ϕ, defined as $\frac{[\mathrm{H_2}]}{2[\mathrm{O_2}]}$) is sufficient to completely define the initial state of the system. It is this triad that we define as our parameter vector, $\boldsymbol{x} = (\phi, T, P)^T$. As stated earlier, the observable is the total heat released Q.

SM2.4.1. Implementation. We assume $p(\boldsymbol{x})$ is a Gaussian with mean $\boldsymbol{x}_0$ and covariance Σ_0, where,

$$\begin{aligned} \boldsymbol{x}_0 &= \begin{bmatrix} 1 \\ 1500 \\ 1.01325 \end{bmatrix}, \\ \Sigma &= \begin{bmatrix} 0.01 & 0 & 0 \\ 0 & 15.0 & 0 \\ 0 & 0 & 0.00101 \end{bmatrix}. \end{aligned} \tag{SM20}$$

In addition, we select $\Upsilon = [21000, 22000]$. The initial volume of the reactor is set to $V_0 = 1\text{m}^3$. We use MATLAB's inbuilt `fminunc` algorithm to perform the non-linear optimization.

Table SM1: *Reduced chemistry for hydrogen autoignition. Reaction rate constant* $k = AT^b e^{-E/RT}$. *Units are mol, cm, s, K, kJ. Chaperone efficiencies are* 2.5 *for* H_2, 16.0 *for* H_2O *and* 1.0 *for all other species. Troe falloff with* $F_{cent} = 0.5$ *is assumed for reaction 5.*

No.	Reaction		A	b	E
1	$H_2 + O_2 \longrightarrow H + HO_2$		2.69×10^{12}	0.36	231.86
2	$H + O_2 \longrightarrow OH + O$		3.52×10^{16}	-0.7	71.4
3	$O + H_2 \longrightarrow H + OH$		5.06×10^{4}	2.7	26.3
4	$OH + H_2 \longrightarrow H + H_2O$		1.17×10^{9}	1.3	15.2
5	$H + O_2 + M \longrightarrow HO_2 + M$	k_0	5.75×10^{19}	-1.4	0.0
		k_∞	4.65×10^{12}	0.4	0.0

SM2.4.2. Summary of equations. Williams ([SM5]) identified a set of 5 elementary steps to study autoiginition of hydrogen-air mixtures. These elementary steps are given in Table SM1.

The net rates of production of each species, $\dot{S}_i$, are given below:

$$\begin{aligned}
\dot{S}_{\text{H}_2} &= -\dot{\omega}_1 - \dot{\omega}_3 - \dot{\omega}_4, \\
\dot{S}_{\text{O}_2} &= -\dot{\omega}_1 - \dot{\omega}_2 - \dot{\omega}_5, \\
\dot{S}_{\text{O}} &= \dot{\omega}_2 - \dot{\omega}_3, \\
\dot{S}_{\text{H}} &= \dot{\omega}_1 - \dot{\omega}_2 + \dot{\omega}_3 + \dot{\omega}_4 - \dot{\omega}_5, \\
\dot{S}_{\text{OH}} &= \dot{\omega}_2 + \dot{\omega}_3 - \dot{\omega}_4, \\
\dot{S}_{\text{HO}_2} &= \dot{\omega}_1 + \dot{\omega}_5, \\
\dot{S}_{\text{H}_2\text{O}} &= \dot{\omega}_4 \\
\dot{S}_{\text{N}_2} &= 0,
\end{aligned} \tag{SM21}$$

where ω_i is the rate of the ith reaction in Table SM1:

$$\begin{aligned}
\dot{\omega}_1 &= k_{f,1}[H_2][O_2] \\
\dot{\omega}_2 &= k_{f,2}[H][O_2] \\
\dot{\omega}_3 &= k_{f,3}[O][H_2] \\
\dot{\omega}_4 &= k_{f,4}[OH][H_2] \\
\dot{\omega}_5 &= k_{f,5}[H][O_2].
\end{aligned} \tag{SM22}$$

Here,

$$k_{f,i} = A^{(i)} T^{b^{(i)}} e^{-E^{(i)}/RT}, \; i = 1 \ldots 4, \tag{SM23}$$

R, the universal gas constant has the value 8.314 J/K/mol. $A^{(i)}$, $b^{(i)}$ and $E^{(i)}$ are as specified in Table SM1. $k_{f,5}$ is calculated using the Lindemann form as:

$$\begin{aligned}
&P_r = \frac{k_0[M]}{k_\infty},\\
&[M] = 2.5[H_2] + [O_2] + [H] + [O] + [OH] + [HO_2] + 16[H_2O] + [N_2],\\
&k_{f,5} = k_\infty\Big(\frac{P_r}{1+P_r}\Big),\\
&k_0 = A^{(0)}T^{b^{(0)}}e^{-E^{(0)}/RT},\\
&k_\infty = A^{(\infty)}T^{b^{(\infty)}}e^{-E^{(\infty)}/RT},
\end{aligned} \tag{SM24}$$

with the Arrhenius parameters as in Table SM1.

The molar enthalpies ($\bar{h}$) and specific heats ($\bar{c}_p$) for each species are given by:

$$\begin{aligned}
&\frac{\bar{c}_p(T)}{R} = a_0 + a_1T + a_2T^2 + a_3T^3 + a_4T^4,\\
&\frac{\bar{h}(T)}{RT} = a_0 + \frac{a_1}{2}T + \frac{a_2}{3}T^2 + \frac{a_3}{4}T^3 + \frac{a_4}{5}T^4 + \frac{a_5}{T}.
\end{aligned} \tag{SM25}$$

where the coefficients $a_0 \dots a_5$ are tabulated in ref. [SM3] for each species.

The rate of heat release per unit volume is:

$$\Delta\dot{q} = -\bar{h}_{H_2}\dot{S}_{H_2} - \bar{h}_{O_2}\dot{S}_{O_2} - \bar{h}_H\dot{S}_H - \bar{h}_O\dot{S}_O - \bar{h}_{OH}\dot{S}_{OH} \tag{SM26}$$

$$- \bar{h}_{HO_2}\dot{S}_{HO_2} - \bar{h}_{H_2O}\dot{S}_{H_2O} - \bar{h}_{N_2}\dot{S}_{N_2}. \tag{SM27}$$

The following ODEs describe the chemistry inside a constant pressure adiabatic reactor:

$$\begin{aligned}
&\frac{\mathrm{d}T}{\mathrm{d}t} = \frac{\Delta\dot{q}}{\beta},\\
&\frac{\mathrm{d}[H_2]}{\mathrm{d}t} = \dot{S}_{H_2} - \alpha[H_2],\\
&\frac{\mathrm{d}[O_2]}{\mathrm{d}t} = \dot{S}_{O_2} - \alpha[O_2],\\
&\frac{\mathrm{d}[H]}{\mathrm{d}t} = \dot{S}_H - \alpha[H],\\
&\frac{\mathrm{d}[O]}{\mathrm{d}t} = \dot{S}_O - \alpha[O],\\
&\frac{\mathrm{d}[OH]}{\mathrm{d}t} = \dot{S}_{OH} - \alpha[OH],\\
&\frac{\mathrm{d}[HO_2]}{\mathrm{d}t} = \dot{S}_{HO_2} - \alpha[HO_2],
\end{aligned}$$

$$
\begin{aligned}
&\frac{\mathrm{d}[H_2O]}{\mathrm{d}t} = \dot{S}_{\mathrm{H_2O}} - \alpha[H_2O],\\
&\frac{\mathrm{d}[N_2]}{\mathrm{d}t} = \dot{S}_{\mathrm{N_2}} - \alpha[N_2],\\
&\frac{1}{V}\frac{\mathrm{d}V}{\mathrm{d}t} = \alpha,\\
&\alpha = \left(\frac{\dot{S}_{\mathrm{H_2}} + \dot{S}_{\mathrm{O_2}} + \dot{S}_{\mathrm{H}} + \dot{S}_{\mathrm{O}} + \dot{S}_{\mathrm{OH}} + \dot{S}_{\mathrm{HO_2}} + \dot{S}_{\mathrm{H_2O}} + \dot{S}_{\mathrm{N_2}}}{[H_2] + [O_2] + [H] + [O] + [OH] + [HO_2] + [H_2O] + [N_2]} + \frac{1}{T}\frac{\mathrm{d}T}{\mathrm{d}t}\right),\\
&\beta = [H_2]\bar{c}_{p,\mathrm{H_2}} + [O_2]\bar{c}_{p,\mathrm{O_2}} + [H]\bar{c}_{p,\mathrm{H}} + [O]\bar{c}_{p,\mathrm{O}} + [OH]\bar{c}_{p,\mathrm{OH}}\\
&+ [HO_2]\bar{c}_{p,\mathrm{HO_2}} + [H_2O]\bar{c}_{p,\mathrm{H_2O}} + [N_2]\bar{c}_{p,\mathrm{N_2}},
\end{aligned}
$$

The net heat released is

$$
Q = \int_0^{t_f} \Delta\dot{q}V(t)\mathrm{d}t. \tag{SM28}
$$

SM2.5. The Lorenz system. The Lorenz system is defined by the following ordinary differential equations:

$$
\begin{aligned}
\frac{\mathrm{d}u_1(t)}{\mathrm{d}t} &= s(u_2 - u_1), \quad 0 < t < t_f,\\
\frac{\mathrm{d}u_2(t)}{\mathrm{d}t} &= u_1(r - u_3) - u_2,\\
\frac{\mathrm{d}u_3(t)}{\mathrm{d}t} &= u_2u_2 - bu_3,
\end{aligned} \tag{SM29}
$$

The parameter vector is the initial condition of the system, $\boldsymbol{x} = \boldsymbol{u}(0)$ while the observable is $u_1(t_f)$.

We set $s = 10$, $r = 28$, $b = 8/3$ and observe u_1 after two time horizons, $t_f = 0.1$s and $t_f = 5$s. The nominal density $p(\boldsymbol{x})$ is selected to be a Gaussian with mean $\boldsymbol{x}_0$ and covariance $\boldsymbol{\Sigma}_0$ where,

$$
\begin{aligned}
\boldsymbol{x}_0 &= \begin{bmatrix} 1.508870 \\ -1.531271 \\ 25.46091 \end{bmatrix}, \text{ and,}\\
\boldsymbol{\Sigma}_0 &= \begin{bmatrix} 0.01508870 & 0 & 0 \\ 0 & 0.01531271 & 0 \\ 0 & 0 & 0.02546091 \end{bmatrix}.
\end{aligned} \tag{SM30}
$$

The target intervals are chosen to be $\mathbb{Y} = [-5, -4]$ when $t_f = 0.1$s and $\mathbb{Y} = [-0.22, -0.21]$ when $t_f = 5$s.

SM2.5.1. Implementation. This system is chaotic with maximal Lyapunov exponent $\lambda \approx 0.906$. We perform the optimization with MATLAB's inbuilt `fminunc` algorithm using analytically derived gradients.

SM2.6. Elliptic PDE. In this experiment, we invert for the log permeability field, g in the following elliptic PDE:

$$\begin{aligned} -\nabla \cdot (e^g \nabla u) &= h \text{ in } \Omega, \\ u &= u_D \text{ on } \partial\Omega_D, \\ e^g \nabla u \cdot \boldsymbol{n} &= u_N \text{ on } \partial\Omega_N, \end{aligned} \tag{SM31}$$

where $\Omega \subset \mathbb{R}^2$ is an open domain with boundary $\partial\Omega = \partial\Omega_D \cup \partial\Omega_N$, $\partial\Omega_D \cap \partial\Omega_N = \emptyset$. $\partial\Omega_D$ and $\partial\Omega_N$ denote Dirichlet and Neumann type boundaries with boundary values u_D and u_N respectively. $\boldsymbol{n}$ is a unit vector normal to $\partial\Omega$ in the outward direction and $h \in L^2(\Omega)$ is the source term.

We assume Ω is a unit square, there is no source term, the left and right walls are no-flux boundaries, and the top and bottom walls are Dirichlet boundaries. That is,

$$\begin{aligned} \Omega &= [0,1] \times [0,1], \\ h &= 0, \\ \partial\Omega_D &= (0,1) \times \{1\} \cup (0,1) \times \{0\}, \\ \partial\Omega_N &= \{0,1\} \times (0,1), \\ u_N &= 0, \\ u_D &= \begin{cases} 1, & \boldsymbol{x} \in (0,1) \times \{1\} \\ 0, & \boldsymbol{x} \in (0,1) \times \{0\} \end{cases}. \end{aligned} \tag{SM32}$$

This is an instance of Bayesian inference in infinite dimensions. The problem can be reduced to finite dimensions by using, for instance, a finite element discretization. The inference is then performed for the vector of coefficients of the finite element basis functions chosen. Here, we use first order Lagrange basis functions with 4225 degrees of freedom. Thus, the parameter vector then is the vector of coefficients $\boldsymbol{x} = (g_1, g_2, \ldots, g_m) \in \mathbb{R}^m, m = 4225$. We define the parameter-to-observable map as the fluid velocity at a particular location in the domain Ω, $f(\boldsymbol{x}) = u(0.1, 0.5)$.

To solve the inverse problem, we use `hIPPYlib` [SM4]. `hIPPYlib` is a scalable software framework to solve large scale PDE constrained inverse problems. It relies on `FEniCS` for the discretization and solution of the PDE and `PETSc` for efficient implementation of linear algebra routines. `hIPPYlib` provides state-of-the-art algorithms for PDE constrained optimization, including an implementation of the Inexact Newton-CG algorithm for computing the MAP point as well as randomized algorithms for constructing a low rank approximation of the Hessian at the MAP point. For full details, we refer the reader to [SM4]. We would like to remark here that this inverse problem appears as a model problem in [SM4] and has been slightly modified for our experiments. For completeness, we reproduce relevant details from [SM4] here.

While constructing $p(\boldsymbol{x})$, it was assumed that the true log-permeability at 5 locations in $\Omega = [0,1] \times [0,1]\,, \boldsymbol{\omega}_1 = (0.1, 0.1), \boldsymbol{\omega}_2 = (0.1, 0.9), \boldsymbol{\omega}_3 = (0.5, 0.5), \boldsymbol{\omega}_4 = (0.9, 0.1), \boldsymbol{\omega}_5 =$

$(0.9, 0.9)$, is known. Let the true log-permeability at these points be $x^1_{\text{true}}, x^2_{\text{true}}, \ldots, x^5_{\text{true}}$. In addition, the following mollifier functions were defined

$$\delta_i(\boldsymbol{\omega}) = \exp\left(-\frac{\gamma^2}{\delta^2}\|\boldsymbol{\omega} - \boldsymbol{\omega}_i\|_{\boldsymbol{\Theta}^{-1}}\right). \tag{SM33}$$

where $\boldsymbol{\Theta}$ is an anisotropic symmetric positive definite tensor of the form

$$\boldsymbol{\Theta} = \begin{pmatrix} \theta_1 \sin^2\alpha & (\theta_1 - \theta_2)\sin\alpha\cos\alpha \\ (\theta_1 - \theta_2)\sin\alpha\cos\alpha & \theta_2\cos^2\alpha \end{pmatrix}. \tag{SM34}$$

The various parameters were set to $\gamma = 0.1, \delta = 0.5, \alpha = \pi/4, \theta_1 = 2, \theta_2 = 0.5$. The covariance of $p(\boldsymbol{x})$ was finally defined as $\boldsymbol{\Sigma}_0 = \mathcal{A}^{-2}$, where,

$$\mathcal{A} = \tilde{\mathcal{A}} + p\sum_{i=1}^{5}\delta_i\mathbf{I}, \tag{SM35}$$

$$= \tilde{\mathcal{A}} + p\mathcal{M}. \tag{SM36}$$

where $\tilde{A}$ is a differential operator of the form $\gamma\nabla\cdot(\theta\nabla) + \delta\mathbf{I}$ and p is a penalization parameter, which was set to $p = 10$.

The mean of the nominal PDF $p(\boldsymbol{x})$, $\boldsymbol{x}_0$, was set to be the solution of the following regularized least squares problem

$$\boldsymbol{x}_0 = \arg\min_{\boldsymbol{x}} = \frac{1}{2}\langle\boldsymbol{x},\boldsymbol{x}\rangle_{\tilde{\mathcal{A}}} + \frac{p}{2}\langle\boldsymbol{x}_{\text{true}} - \boldsymbol{x}, \boldsymbol{x}_{\text{true}} - \boldsymbol{x}\rangle_{\mathcal{M}}. \tag{SM37}$$

The nominal distribution $p(\boldsymbol{x})$ is then defined to be $\mathcal{N}(\boldsymbol{x}_0, \boldsymbol{\Sigma}_0)$ and $\mathbb{Y} = [0.6, 0.7]$.

SM2.7. Periodic map. In this case the input-output map is defined to be:

$$f(\boldsymbol{x}) = \sin(x_1)\cos(x_2). \tag{SM38}$$

SM2.8. Implementation. Again, $p(\boldsymbol{x})$ is assumed to be a Gaussian distribution with mean $\boldsymbol{x}_0$ and covariance $\boldsymbol{\Sigma}_0$, where $\boldsymbol{x}_0 = (1,1)^T$ and $\boldsymbol{\Sigma_0} = \mathbf{I}$. We chose $\mathbb{Y} = [0.4, 0.6]$. The MAP point was computed using MATLAB's `fminunc` routine.

SM3. Results. Here we report the probability estimates corresponding to Figure 9.

Table SM2: *Single step reaction*

N	MC	BIMC, n=1	BIMC, n=5	BIMC, n=10	BIMC, n=25
10	$0 \cdot 10^{0}$	$1.7227 \cdot 10^{-2}$	$2.5837 \cdot 10^{-2}$	$3.8965 \cdot 10^{-2}$	$2.2064 \cdot 10^{-2}$
100	$5 \cdot 10^{-2}$	$2.2673 \cdot 10^{-2}$	$2.0598 \cdot 10^{-2}$	$2.0679 \cdot 10^{-2}$	$1.9208 \cdot 10^{-2}$
1,000	$1.5 \cdot 10^{-2}$	$2.2675 \cdot 10^{-2}$	$2.2016 \cdot 10^{-2}$	$2.3448 \cdot 10^{-2}$	$2.2505 \cdot 10^{-2}$
10,000	$2.49 \cdot 10^{-2}$	$2.3091 \cdot 10^{-2}$	$2.3294 \cdot 10^{-2}$	$2.3089 \cdot 10^{-2}$	$2.3093 \cdot 10^{-2}$
$1 \cdot 10^{5}$	$2.417 \cdot 10^{-2}$	$2.3032 \cdot 10^{-2}$	$2.3011 \cdot 10^{-2}$	$2.3090 \cdot 10^{-2}$	$2.3127 \cdot 10^{-2}$
$1 \cdot 10^{6}$	$2.3121 \cdot 10^{-2}$	$2.3034 \cdot 10^{-2}$	$2.3033 \cdot 10^{-2}$	$2.3010 \cdot 10^{-2}$	$2.3040 \cdot 10^{-2}$

Table SM3: *Autoignition*

N	MC	BIMC, n=1	BIMC, n=5	BIMC, n=10	BIMC, n=25
10	$0 \cdot 10^{0}$	$3.7632 \cdot 10^{-2}$	$3.8167 \cdot 10^{-2}$	$3.0462 \cdot 10^{-2}$	$2.7556 \cdot 10^{-2}$
100	$6 \cdot 10^{-2}$	$3.4138 \cdot 10^{-2}$	$3.1870 \cdot 10^{-2}$	$3.0651 \cdot 10^{-2}$	$3.0507 \cdot 10^{-2}$
1,000	$4.1 \cdot 10^{-2}$	$3.2459 \cdot 10^{-2}$	$3.2642 \cdot 10^{-2}$	$3.2349 \cdot 10^{-2}$	$3.3389 \cdot 10^{-2}$
10,000	$3.24 \cdot 10^{-2}$	$3.2382 \cdot 10^{-2}$	$3.2430 \cdot 10^{-2}$	$3.2464 \cdot 10^{-2}$	$3.2586 \cdot 10^{-2}$

Table SM4: *Lorenz, $t_f = 0.1s$*

N	MC	BIMC, n=1	BIMC, n=5	BIMC, n=10	BIMC, n=25
10	$1 \cdot 10^{-1}$	$3.3462 \cdot 10^{-2}$	$2.1451 \cdot 10^{-2}$	$3.2064 \cdot 10^{-2}$	$4.3372 \cdot 10^{-2}$
100	$2 \cdot 10^{-2}$	$3.2753 \cdot 10^{-2}$	$3.2673 \cdot 10^{-2}$	$3.0035 \cdot 10^{-2}$	$3.1618 \cdot 10^{-2}$
1,000	$3.1 \cdot 10^{-2}$	$3.3606 \cdot 10^{-2}$	$3.2297 \cdot 10^{-2}$	$3.2493 \cdot 10^{-2}$	$3.3008 \cdot 10^{-2}$
10,000	$3.25 \cdot 10^{-2}$	$3.2906 \cdot 10^{-2}$	$3.3250 \cdot 10^{-2}$	$3.3020 \cdot 10^{-2}$	$3.2456 \cdot 10^{-2}$
$1 \cdot 10^{5}$	$3.402 \cdot 10^{-2}$	$3.2853 \cdot 10^{-2}$	$3.3004 \cdot 10^{-2}$	$3.2878 \cdot 10^{-2}$	$3.2969 \cdot 10^{-2}$

Table SM5: *Elliptic PDE*

N	MC	BIMC, n=1	BIMC, n=5	BIMC, n=10	BIMC, n=25
100	$0 \cdot 10^{0}$	$3.8350 \cdot 10^{-4}$	$3.5502 \cdot 10^{-4}$	$4.6057 \cdot 10^{-4}$	$4.3995 \cdot 10^{-4}$
1,000	$0 \cdot 10^{0}$	$3.6220 \cdot 10^{-4}$	$4.1600 \cdot 10^{-4}$	$3.8300 \cdot 10^{-4}$	$3.8608 \cdot 10^{-4}$
10,000	$6 \cdot 10^{-4}$	$3.8517 \cdot 10^{-4}$	$3.8503 \cdot 10^{-4}$	$3.8666 \cdot 10^{-4}$	$3.8213 \cdot 10^{-4}$
$1 \cdot 10^{5}$	$4.3 \cdot 10^{-4}$	$3.9141 \cdot 10^{-4}$	$3.9097 \cdot 10^{-4}$	$3.9021 \cdot 10^{-4}$	$3.8937 \cdot 10^{-4}$

REFERENCES

[1] J. F. de Cossio-Diaz, *Moments of the univariate truncated normal distribution.* https://github.com/cossio/TruncatedNormal.jl/blob/master/notes/normal.pdf.

[2] S. B. Pope and M. S. Anand, *Flamelet and distributed combustion in premixed turbulent flames*, in Symposium (International) on Combustion, vol. 20, Elsevier, 1985, pp. 403–410.

[3] G. P. Smith, D. M. Golden, M. Frenklach, N. W. Moriarty, B. Eiteneer, M. Goldenberg, C. T. Bowman, R. K. Hanson, S. Song, J. William C. Gardiner, V. V. Lissianski, and Z. Qin. http://www.me.berkeley.edu/gri_mech/.

[4] U. Villa, N. Petra, and O. Ghattas, *hIPPYlib: an extensible software framework for large-scale deterministic and linearized bayesian inversion*, (2016), http://hippylib.github.io.

[5] F. A. Williams, *Detailed and reduced chemistry for hydrogen autoignition*, Journal of Loss prevention in the Process Industries, 21 (2008), pp. 131–135.

[6] J. V. Zidek and C. van Eeden, *Uncertainty, entropy, variance and the effect of partial information*, vol. Volume 42 of Lecture Notes–Monograph Series, Institute of Mathematical Statistics, Beachwood, OH, 2003, pp. 155–167, https://projecteuclid.org/euclid.lnms/1215091936.