An Investigation into the Distribution of Ratios of Particle Solver-based Likelihoods

Emil Løvbak and Sebastian Krumscheid

Karlsruhe Institute of Technology, Zirkel 2, 76133 Karlsruhe, Germany
{emil.loevbak,sebastian.krumscheid}@kit.edu

Abstract. We investigate the use of the Metropolis-Hastings algorithm to sample posterior distribution in a Bayesian inverse problem, where the likelihood function is random. Concretely, we consider the case where one has full field observations of a PDE solution, in case a one-dimensional diffusion equation, subject to a Gaussian observation error. Assuming one uses a particle-based Monte Carlo simulation when approximating the likelihood function, one gets an approximate likelihood with additive Gaussian noise in the log-likelihood. We study how these two Gaussian distributions affect the distribution of ratios of approximate likelihood evaluations, as required when evaluating acceptance probabilities in the Metropolis-Hastings algorithm. We do so through both theoretical analysis and numerical experiments.

Keywords: Bayesian inversion, Metropolis-Hastings, random likelihoods, Monte Carlo simulation

1 Introduction

We consider a diffusion equation with homogeneous diffusion coefficient on the one-dimensional spatial domain $x \in [0, L)$ and periodic boundaries, i.e.,

$$\partial_t \rho(x,t) = D \partial_{xx} \rho(x,t), \tag{1}$$

$$\rho(0,t) = \rho(L,t), \qquad \partial_x \rho(0,t) = \partial_x \rho(L,t), \qquad \rho(x,0) = \rho_0(x) \geq 0. \tag{2}$$

Here, $t \in \mathbb{R}_{\geq 0}$ is the time variable and $\rho_0(x)$ is a given initial condition. We approximate the solution to this equation using a Monte Carlo simulation. Such simulations discretize the equation at hand using an ensemble of particles. The particle dynamics are chosen such that the ensemble statistics converge to a (potentially biased) approximation of the true solution of (1)–(2) as the particle ensemble size $P \to \infty$. Although Monte Carlo simulation can not compete with deterministic approaches, such as finite differences [?] or spectral methods [?] for the problem at hand, we are motivated by their indispensable character in various high-impact domains, including nuclear fusion research [?, ?], financial mathematics [?] and computer graphics [?], considering the diffusion equation as a straightforward, low-cost example.

We aim to solve a discretized statistical inverse problem of the form

$$\rho = G[D] + \eta, \qquad \eta \sim \mathcal{N}(0, \Sigma_\eta), \tag{3}$$

where, $G[\cdot] : \mathbb{R} \mapsto \mathbb{R}^N$ is the solution map of (1)–(2) for a given diffusion coefficient $D \in \mathbb{R}$ on a discrete spatial grid. The observation error $\eta \in \mathbb{R}^N$ is an unknown multivariate random quantity, assumed normally distributed with a strictly positive definite covariance matrix $\Sigma_\eta = \Sigma_\eta^\mathsf{T} \in \mathbb{R}_{\succ 0}^{N \times N}$. Although η is modeled as a random variable, it takes a fixed, but unknown, value for a given observation ρ.

We wish to determine a reasonable value of D, given an observation $\rho \in \mathbb{R}^N$. However, we assume that we only have an approximate solution map

$$\hat{G}[D] = G[D] + \delta, \qquad \delta \sim \mathcal{N}(\mu(D), \Sigma(D)), \tag{4}$$

resulting from the use of Monte Carlo simulation. Here, $\delta \in \mathbb{R}^N$ follows a normal distribution with mean $\mu : \mathbb{R} \mapsto \mathbb{R}^N$ and covariance matrix $\Sigma = \Sigma^\mathsf{T} : \mathbb{R} \mapsto \mathbb{R}_{\succ 0}^{N \times N}$, parameterized by the parameter D. We motivate the assumed distribution in Section 2. As $\hat{G}[\cdot]$ is not easily invertible, solving the inverse problem (3) requires multiple evaluations of (1)–(2) [?]. This fact, combined with the high computational cost of Monte Carlo simulation means a cost-accuracy trade-off must be made in the value of P. This trade-off is further exacerbated when multiple evaluations of the model are needed, such as when solving problems of the form (3). One must then take into account both the stochastic nature of the approximation error, as well as the generally high simulation cost incurred.

Due to this computational cost, most efforts to solve inverse problems using Monte Carlo simulation have relied on neglecting the observation error η, to solve the inverse problem as an optimization problem, minimizing the residual $G[D] - y$ under a suitable norm. The main challenge in the setting given by (4), where the forward model contains a solver with stochastic errors, is computing gradients. Various approaches have been applied to tackle this challenge, notably adjoint methods, see e.g. [?, ?, ?, ?, ?], and algorithmic differentiation, see e.g. [?, ?, ?]. However, in this work, we consider a Bayesian approach, where we approximate a posterior distribution of D conditioned on the observation ρ. The Bayesian posterior is viewed as the solution to the inverse problem as it encodes all of the available information about the unknown D given the data ρ.

Using Bayes' rule, we find that the probability density function (PDF) of the posterior for D and a given ρ is

$$\pi_{\text{post}}(D|\rho) = \frac{\pi_0(D)\pi_1(\rho|D)}{\pi_{\text{obs}}(\rho)}, \tag{5}$$

where $\pi_0(\cdot)$ is the PDF of an assumed prior distribution on D, $\pi_1(\rho|D)$ is the likelihood of ρ given D, and $\pi_{\text{obs}}(\cdot)$ is the marginal PDF of the given data ρ. We note that $\pi_{\text{obs}}(\rho)$ is generally unknown.

In our setting, the posterior given by (5) will depend nonlinearly on the random error δ, due to using the approximate solution map (4) in the log-likelihood.

Given a deterministic discretization error δ, one can apply standard textbook results to bound the resulting error on the posterior (5) in, e.g., Hellinger distance [?, Ch. 1]. Similar results on posterior convergence were proven in [?] for random errors δ, under suitable conditions, for the limit $\delta \to 0$. However, we consider the case of non-negligible stochastic errors. To the best of our knowledge, no theoretical results on well-posedness exist in this setting. We focus on algorithmic analysis in this work, rather than the posterior itself.

Markov chain Monte Carlo methods (MCMC) are designed to sample the posterior given by (5), despite the unknown denominator. Given ρ, these methods construct a Markov chain of correlated samples of D where each transition in the chain evaluates ratios of posterior probabilities, hence making the denominators cancel when deciding to accept or reject a proposed sample. The most straightforward approach is the Metropolis-Hastings algorithm [?, ?]; however, many notable variants and extensions have been developed, e.g., Hamiltonian Monte Carlo [?] that incorporates gradient information.

In the considered setting with stochastic errors (4), we note that the use of MCMC for sampling the posterior (5) results in a nested sampling problem as each MCMC sample requires a full Monte Carlo simulation. Our interest lies in understanding the effect of non-negligible noise in the forward map. For inherently random but unbiased likelihoods, one can rely on pseudo-marginal MCMC methods [?]. Such likelihoods are often the result of a product or sum of individual contributions from different data points [?] or simulation outputs [?]. Pseudo-marginal methods consider the combined probability space of the parameter to be inferred and the stochastic dimension of the forward map. The aim is to approximate the marginal distribution of the inferred parameter D, i.e., integrating out the forward map's stochasticity, using samples from this combined probability space. Although these methods are well studied, the assumption of unbiasedness means that they cannot be directly applied in our setting.

The goal of this work is to consider the non-asymptotic sampling regime in P, i.e., simulations with non-negligible stochastic error. We study how the Monte Carlo simulation error affects the ratio of computed approximate likelihood values, as a first step towards understanding the acceptance-rejection behavior of the Metropolis-Hastings algorithm in this setting. Notably, we observe that this ratio becomes a random variable, whose distribution we study in terms of its moments. This line of work complements that presented in [?], applying MCMC to Monte Carlo simulations in photoacoustic imaging.

The remainder of this paper is structured as follows. In Section 2 we introduce a simple Monte Carlo discretization for the one-dimensional diffusion equation and make observations on the structure of the resulting discretization error. Next, in Section 3 we present a theoretical analysis of the ratio of approximate likelihoods evaluated using such solvers, in terms of the moments of the ratio's distribution. This analysis is then further supported by experimental results in Section 4. Finally, in Section 5 we draw some conclusions and discuss some their potential for producing future algorithmic developments.

2   Monte Carlo simulation of the diffusion equation

In this Section, we derive a Monte Carlo discretization of the model (1)–(2) and comment on the structure of the resulting error. This derivation is purposefully brief. For a more detailed mathematical derivation of Monte Carlo simulation for parabolic PDEs, we refer to textbooks, such as [**?**]. We assume, for simplicity, that $\rho(x,t) \geq 0$. As such, we can interpret $\rho(x,t)$ as a probability density over x at any given time t. We then make use of the fact that (1) is the Fokker-Planck equation to the stochastic differential equation (SDE)

$$\mathrm{d}X(t) = \sqrt{2D}\mathrm{d}W(t), \qquad X(0) \stackrel{\text{i.i.d.}}{\sim} \rho_0, \tag{6}$$

with $W(t)$ a standard 1D Wiener process, and the initial condition $X(0)$ is distributed according to the density ρ_0.

Given (6), we construct an ensemble of P particles, defined by their positions

$$\{X_{p,k}\}_{p=1}^P \tag{7}$$

at time $t = k\Delta t$. Using an Euler-Maruyama discretization, we propagate each particle p in (7) as a discretized realization of (6), i.e., for $p = 1, \ldots, P$

$$X_{p,k+1} = X_{p,k} + \sqrt{2D\Delta t}W_{p,k}, \qquad W_{p,k} \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0,1), \qquad k \geq 0, \tag{8}$$

and $X_{p,0} \stackrel{\text{i.i.d.}}{\sim} \rho_0$. Simulating such an ensemble, produces an ensemble of particles, where the values $X_{p,k}$ are distributed according to $\rho(x,t)$ for any $t = k\Delta t$.

Once we have the ensemble of particles (7), we apply a binning strategy to compute a solution on a grid with $N = \frac{L}{\Delta x}$ cells We estimate the density at the center of the n-th grid cell $x_n = \left(n + \frac{1}{2}\right)\Delta x$ as

$$\rho(x_n, k\Delta t) \approx \hat{\rho}_{n,k} = \frac{1}{P\Delta x}\sum_{p=1}^P \mathcal{I}_n(X_{p,k}), \qquad n = 0, \ldots, N-1, \tag{9}$$

where $\mathcal{I}_n$ is the indicator function for the grid cell with index n. Binning straight-forwardly produces a vector of length N representing the solution at a given moment in time as a histogram, but introduces two sources of error: (i) the approximation of the solution by a piecewise constant function introduces a bias due to spatial discretization; (ii) each particle can only contribute to a single cell at each time step, hence the computed results will have a high variance.

We define the error $\delta_k \in \mathbb{R}^N$ of the Monte Carlo simulation, at time $t = k\Delta t$, using the elementwise notation

$$\delta_{n,k} = \rho(x_n, t) - \hat{\rho}_{n,k}, \qquad n = 0, \ldots, N-1, \qquad k \geq 0.$$

As the discretization error δ_k is stochastic and independent for every realization, we use mean-squared error (MSE) as an error metric. For notational simplicity, we do so at the level of individual cells. We decompose the error into the squared bias and variance, i.e.,

$$\mathbb{E}[\delta_{n,k}^2] = \mathbb{E}^2[\delta_{n,k}] + \mathbb{V}[\delta_{n,k}]. \tag{10}$$

Bias. For the homogeneous diffusion equation, the discretized dynamics (8) produced unbiased trajectories corresponding with realizations of the SDE (6), i.e., the time discretization does not introduce an error. We remark, however, that the discretization of more general SDEs will result in biased approximations. Here, the only sources of error lie in the estimation step. Specifically, (9) is a biased estimator, due to producing a histogram with bins of width Δx. By construction, the expectation of (9) is given by

$$\mathbb{E}\left[\hat{\rho}_{n,k}\right] = \frac{1}{\Delta x} \int_{n\Delta x}^{(n+1)\Delta x} \rho(x,t)\mathrm{d}x = \frac{F((n+1)\Delta x, t) - F(n\Delta x, t)}{\Delta x},$$

with $F(x,t)$ the primitive of $\rho(x,t)$, i.e., the cumulative distribution function of $X(t)$. Taking the Taylor series around an arbitrary point $x \in [n\Delta x, (n+1)\Delta x]$, we get that

$$\mathbb{E}\left[\hat{\rho}_{n,k}\right] = \rho(x,t) + \left(\left(n + \frac{1}{2}\right)\Delta x - x\right)\frac{\partial \rho}{\partial x}(x,t) + \mathcal{O}(\Delta x^2).$$

Hence, the values at the x_n have a binning bias that decays with order two in $\Delta x \to 0$. For general values of x, we note that the function is approximated with order one in Δx as

$$\left(n + \frac{1}{2}\right)\Delta x - x \leq \frac{\Delta x}{2}.$$

Variance. The finite ensemble size P induces a sampling error that scales with the number of particles in a given cell. As each particle in the ensemble (7) is fully independent and produces a single contribution to the binning estimator (9), we observe through the law of large numbers that the variance of the individual grid cells scales with $\mathcal{O}\left(\frac{1}{P\Delta x}\right)$ as $P\Delta x \to \infty$. Moreover, the distribution of δ_k will converge to a multivariate Gaussian in this limit, with mean and variance corresponding to the bias and variance in the error decomposition (10). Hence, motivating the assumed distribution in (4). Despite the trajectory independence, we note that the grid cells themselves are correlated due to the fact that each trajectory can only score in one cell at a given time step.

3   Analysis of approximate likelihood ratios

We now take a deeper dive into how a stochastic approximation error $\delta \in \mathbb{R}^N$, as specified in (4), affects ratios of approximate likelihood evaluations. In what follows, we drop the subscripts used in Section 2, for notational simplicity. Throughout this section we make use of inner products and norms induced by symmetric strictly positive definite matrices. For example, given two arbitrary vectors $\alpha, \beta \in \mathbb{R}^N$ and a symmetric strictly positive definite matrix $Q = Q^\mathsf{T} \in \mathbb{R}_{\succ 0}^{N \times N}$, we have that $\langle \alpha, \beta \rangle_Q = \alpha^\mathsf{T} Q \beta$ and $\|\alpha\|_Q^2 = \alpha^\mathsf{T} Q \alpha$.

Given a discretization error (4), we introduce an approximate likelihood function by plugging the difference $\rho - \hat{G}[D]$ into the PDF of the normal distribution modeling the additive observation error in (3), i.e.,

$$\hat{l}(D) = \frac{1}{\sqrt{(2\pi)^N |\Sigma_\eta|}} \exp\left(-\frac{1}{2}\left\|\rho - \hat{G}[D]\right\|_{\Sigma_\eta^{-1}}^2\right). \tag{11}$$

We now present this paper's core theoretical result, relating the covariance matrices of the distributions of the random variables δ and η to the existence of moments of the distribution of ratios of approximate likelihoods (11).

Theorem 1. *Given two parameter values $D_1, D_2 \in \mathbb{R}$ and an approximate forward map $\hat{G} : \mathbb{R} \mapsto \mathbb{R}^N$ such that $\hat{G}[D] = G[D] + \delta$ with $\delta \overset{i.i.d.}{\sim} \mathcal{N}(\mu(D), \Sigma(D))$ where $\mu : \mathbb{R} \mapsto \mathbb{R}^N$ and $\Sigma = \Sigma^\mathsf{T} : \mathbb{R} \mapsto \mathbb{R}^{N \times N}_{\succ 0}$. The approximate likelihood ratio under the assumption of Gaussian observation error* (3), *denoted by*

$$\frac{\hat{l}(D_1)}{\hat{l}(D_2)}, \tag{12}$$

is a random variable, who's p-th moment exists iff $p\Sigma_\eta \succ \Sigma(D_2)$, with $\succ$ denoting the Löwner order.

Proof. We introduce notation $\mu_i = \mu(D_i)$, $\Sigma_i = \Sigma(D_i)$, $\Delta\rho_i = \rho - G[D_i]$ and $\delta_i \sim \mathcal{N}(\mu_i, \Sigma_i)$ for $i = 1, 2$. Subsequently, we write

$$\frac{\hat{l}(D_1)}{\hat{l}(D_2)} = \exp\left(\frac{\|\Delta\rho_2 - \delta_2\|_{\Sigma_\eta^{-1}}^2 - \|\Delta\rho_1 - \delta_1\|_{\Sigma_\eta^{-1}}^2}{2}\right), \tag{13}$$

taking into account that the normalizing constants in (11) are independent of δ_1 and δ_2. As noted in Section 2, δ_1 and δ_2 are independent. Hence, the p-th raw moment of (13) is given by the product

$$\mathbb{E}_{\delta_1}\left[\exp\left(-p\frac{\|\Delta\rho_1 - \delta_1\|_{\Sigma_\eta^{-1}}^2}{2}\right)\right] \mathbb{E}_{\delta_2}\left[\exp\left(p\frac{\|\Delta\rho_2 - \delta_2\|_{\Sigma_\eta^{-1}}^2}{2}\right)\right],$$

which we rewrite using the substitution $M^{-1} = p\Sigma_\eta^{-1}$ to get

$$\mathbb{E}_{\delta_1}\left[\exp\left(-\frac{\|\delta_1 - \Delta\rho_1\|_{M^{-1}}^2}{2}\right)\right] \mathbb{E}_{\delta_2}\left[\exp\left(\frac{\|\delta_2 - \Delta\rho_2\|_{M^{-1}}^2}{2}\right)\right]. \tag{14}$$

We separately work out the two expectations. The first factor of (14) becomes

$$\int_{\mathbb{R}^N} \exp\left(-\frac{\|\delta_1 - \Delta\rho_1\|_{M^{-1}}^2}{2}\right) \frac{1}{\sqrt{|2\pi\Sigma_1|}} \exp\left(-\frac{\|\delta_1 - \mu_1\|_{\Sigma_1^{-1}}^2}{2}\right) d\delta_1, \tag{15}$$

which we interpret as a product of Gaussian PDFs for δ_1. We now show that this product is proportional to a single Gaussian PDF with precision matrix $S_1^{-1} = M^{-1} + \Sigma_1^{-1}$. Note that S_1^{-1} is strictly positive definite as the sum of strictly positive definite matrices. Hence, we introduce $\alpha_1 = S_1 \left(M^{-1} \Delta\rho_1 + \Sigma_1^{-1} \mu_1 \right)$ as the mean of the Gaussion proportional to the integrand and work out that

$$\|\delta_1 - \Delta\rho_1\|_{M^{-1}}^2 + \|\delta_1 - \mu_1\|_{\Sigma_1^{-1}}^2$$

$$= \|\delta_1\|_{M^{-1}}^2 - 2\langle\delta_1, \Delta\rho_1\rangle_{M^{-1}} + \|\Delta\rho_1\|_{M^{-1}}^2 + \|\delta_1\|_{\Sigma_1^{-1}}^2 - 2\langle\delta_1, \mu_1\rangle_{\Sigma_1^{-1}} + \|\mu_1\|_{\Sigma_1^{-1}}^2$$

$$= \|\delta_1\|_{S_1^{-1}}^2 - 2\langle\delta_1, \alpha_1\rangle_{S_1^{-1}} + \left(\|\alpha_1\|_{S_1^{-1}}^2 - \|\alpha_1\|_{S_1^{-1}}^2 \right) + \|\Delta\rho_1\|_{M^{-1}}^2 + \|\mu_1\|_{\Sigma_1^{-1}}^2$$

$$= \|\delta_1 - \alpha_1\|_{S_1^{-1}}^2 - \|\alpha_1\|_{S_1^{-1}}^2 + \|\Delta\rho_1\|_{M^{-1}}^2 + \|\mu_1\|_{\Sigma_1^{-1}}^2$$

$$\triangleq \|\delta_1 - \alpha_1\|_{S_1^{-1}}^2 + 2C_1,$$

where we use $\triangleq$ to denote that we have defined the δ_1-independent constant C_1.

We now write (15) as

$$\int_{\mathbb{R}^N} \exp\left(-\frac{\|\delta_1 - \alpha_1\|_{S_1^{-1}}^2}{2} \right) \frac{1}{\sqrt{|2\pi\Sigma_1|}} \exp\left(-C_1\right) \mathrm{d}\delta_1$$

$$= \int_{\mathbb{R}^N} \exp\left(-\frac{\|\delta_1 - \alpha_1\|_{S_1^{-1}}^2}{2} \right) \frac{1}{\sqrt{|2\pi\Sigma_1|}} \exp\left(-C_1\right) \mathrm{d}\delta_1$$

$$= \sqrt{\frac{|S_1|}{|\Sigma_1|}} \exp\left(-C_1\right).$$

The second factor in (14) requires a bit more care due to the differing signs. We first write the factor as

$$\int_{\mathbb{R}^N} \exp\left(\frac{\|\delta_2 - \Delta\rho_2\|_{M^{-1}}^2}{2} \right) \frac{1}{\sqrt{|2\pi\Sigma_2|}} \exp\left(-\frac{\|\delta_2 - \mu_2\|_{\Sigma_2^{-1}}^2}{2} \right) \mathrm{d}\delta_2. \qquad (16)$$

We define $S_2^\dagger = \Sigma_2^{-1} - M^{-1}$ and $\gamma_2 = \left(\Sigma_2^{-1} \mu_2 - M^{-1} \Delta\rho_2 \right)$. Note that we consider $S_2^\dagger$ as the Moore-Penrose pseudoinverse of a not yet introduced matrix S_2, as it may be singular. We write $\delta_2 = \delta_{2,0} + \delta_{2,\perp}$ and $\gamma_2 = \gamma_{2,0} + \gamma_{2,\perp}$, with $\delta_{2,0}, \gamma_{2,0} \in \ker(S_2^\dagger)$ and $\delta_{2,\perp}, \gamma_{2,\perp} \in \mathrm{range}(S_2^\dagger)$. We then define $\alpha_2 = S_2\gamma_2 = S_2\gamma_{2,\perp}$. For compactness, we allow some abuse of notation, still using norms and inner product notation weighted by the potentially indefinite matrix $S_2^\dagger$.

Following a similar argument to before, and making use of the fact that $\delta_{2,0} \perp \gamma_{2,\perp}$, we get

$$\|\delta_2 - \mu_2\|_{\Sigma_2^{-1}}^2 - \|\delta_2 - \Delta\rho_2\|_{M^{-1}}^2$$

$$= \|\delta_2\|_{\Sigma_2^{-1}}^2 - 2\langle\delta_2, \mu_2\rangle_{\Sigma_2^{-1}} + \|\mu_2\|_{\Sigma_2^{-1}}^2 - \|\delta_2\|_{M^{-1}}^2 + 2\langle\delta_2, \Delta\rho_2\rangle_{M^{-1}} - \|\Delta\rho_2\|_{M^{-1}}^2$$

$$= \|\delta_{2,\perp}\|_{S_2^\dagger}^2 - 2\langle\delta_{2,\perp}, \alpha_2\rangle_{S_2^\dagger} - 2\langle\delta_{2,0}, \gamma_{2,0}\rangle + \|\mu_2\|_{\Sigma_2^{-1}}^2 - \|\Delta\rho_2\|_{M^{-1}}^2$$

$$= \|\delta_{2,\perp} - \alpha_2\|^2_{S_2^\dagger} - 2\langle \delta_{2,0}, \gamma_{2,0}\rangle - \|\alpha_2\|^2_{S_2^\dagger} + \|\mu_2\|^2_{\Sigma_2^{-1}} - \|\Delta\rho_2\|^2_{M^{-1}}$$

$$\triangleq \|\delta_{2,\perp} - \alpha_2\|^2_{S_2^\dagger} - 2\langle \delta_{2,0}, \gamma_{2,0}\rangle + 2C_2$$

where, similar to C_1, C_2 is a constant term, independent of δ_2.

We now work out (16) as

$$\int_{\mathbb{R}^N} \exp\left(-\frac{\|\delta_{2,\perp} - \alpha_2\|^2_{S_2^\dagger}}{2}\right) \exp\left(\langle \delta_{2,0}, \gamma_{2,0}\rangle\right) \frac{1}{\sqrt{|2\pi\Sigma_2|}} \exp\left(-C_2\right) d\delta_2. \quad (17)$$

We observe that the integrand is a smooth function that only vanishes at infinity if (i) $\delta_{2,0}$ remains bounded and (ii) the nonzero eigenvalues of $S_2^\dagger$ are positive. Condition (i) can only be met if $\delta_{2,0} = 0$ and $S_2^\dagger$ must be invertible and thus strictly positive definite. We therefore state $\forall \delta_2 \in \mathbb{R}^N : \delta_2 = \delta_{2,\perp}$. Hence, (17) evaluates to

$$\sqrt{\frac{|S_2|}{|\Sigma_2|}} \exp\left(-C_2\right).$$

We combine integrability condition (ii) for (17) with the fact that (15) is strictly positive and the definition of M to conclude that the p-th moment only exists iff $\Sigma_2^{-1} \succ p\Sigma_\eta^{-1}$. From the definition of the Löwner order, one can show that this condition is equivalent to $p\Sigma_\eta \succ \Sigma(D_2)$. $\qquad\square$

During the proof of Theorem 1, we make extensive use of the properties of normal distributions to compute bounds for the moments of the distribution of the approximate likelihood ratio (12). However, we note that this ratio itself will not be normally distributed as it only has finitely many moments.

4   Numerical experiments

We now perform a numerical experiment to confirm the observations in Theorem 1. To this end, we use the model problem (1)–(2) and consider transitions between fixed values for D_1 and D_2. We produce synthetic observations, using a finite difference discretization whose solution is perturbed with noise distributed according to (3). Throughout this section we set $N = 100$, $L = 10$, $t = 10$ and $D^* = 0.1$. The reference solution to produce the observation is computed with $\Delta t = 0.1$, while the Monte Carlo simulations use $\Delta t = t$ to cut on computational cost, thanks to the unbiasedness of (8). For D_1 and D_2, we consider differing combinations of these parameters from the set $\{0.08, 0.1\}$, i.e., the true value and a similar, but less probable, value. The code used to generate the results in this section can be found at github.com/UQatKIT/ FrontUQ-2024-proceedings-likelihood-ratios.

To enable visualization, we simplify the covariance matrices of the distributions of η and δ as follows. We fix $\Sigma_\eta = \sigma_\eta^2 I$, with I the identity matrix.

We simplify the Monte Carlo covariance to a scalar variance σ_δ^2, by taking the maximum over all histogram cells, i.e.,

$$\sigma_\delta^2 = \max(\mathrm{diag}(\Sigma(D))), \tag{18}$$

for the given value D. Taking the maximum is motivated by the resulting value being robust to the effects of bins with zero or very few particles. All expectations in this section are computed using Monte Carlo sampling with 1000 random realizations of the observation ρ. To generate each curve in the plots in this section, we take $P = 10^2, \ldots, 10^6$ increasing in multiples of 10 and estimating σ_δ as specified in (18). In some cases, in particular for small P, we get invalid results for the quantities being plotted. These invalid results result from the exponential in (11) evaluating to zero due to highly unlikely solver outputs. Taking the ratio of expectations then produces NaN values. As both zero and NaN values cannot be sensibly plotted in a log scale, these points are omitted from the figures.

Before considering expectations of ratios of approximate likelihoods, we first present the expectations of the approximate likelihoods themselves in Figure 1. We observe that the expected approximate likelihood for $D_1 = 0.1$ (Figure 1a) is consistently larger than that for $D_2 = 0.08$ (Figure 1b), as can be seen in Figure 1c that shows the ratio of the expectations of these approximate likelihoods. We also note the sharp drop in the expectation of the approximate likelihood for both D_1 and D_2 once $\sigma_\eta < \sigma_\delta$. As one would expect, for larger values of P, the ratio of expectations in Figure 1c converges to one as σ_η increases.

Next, in Figure 2, we consider a transition from $D_2 = 0.1$ to $D_1 = 0.1$. We plot the expectation of the approximate likelihood ratio in Figure 2a. Here we observe a significant blowup in the expectation at approximately $\sigma_\eta = \sigma_\delta$, as predicted by Theorem 1. For sufficiently small values of σ_δ, we observe that the expectation of the likelihood ratio converges to the true value of 1. To better understand the divergence of the approximate likelihood ratio, we plot the expectation of the minimum of the ratio and 1 in Figure 2b, which in the case of $D_1 = D_2$ directly corresponds to the acceptance probability in the Metropolis-Hastings algorithm, see e.g. [**?**, Ch. 7]. Here, it becomes clear that when the approximate likelihood ratio diverges, the acceptance probability converges to the value 0.5. This behavior gives insight into the precise mechanism behind the divergence. Namely, in this regime, the stochastic error in the approximate likelihood evaluations is sufficiently large that the ratio consists of a numerator and denominator with wildly differing orders of magnitude. Hence, the random variable converges to a distribution with two states, one of which diverges to infinity, the other converges to zero. Each state has probability 0.5.

We now present similar plots for transitions between the values 0.08 and 0.1. In Figure 3, we consider the transition from $D_2 = D^* = 0.1$ to $D_1 = 0.08$. Here we note that Figure 3b now does not in general represent the acceptance probability as we neglect the transition probabilities of the proposal distribution in the Metropolis-Hastings algorithm and the prior distribution. However, we still gain insight into the behavior of the acceptance probability. We also note that this subfigure will correspond to the acceptance probability when considering uninformed prior distributions and symmetric proposal distributions.
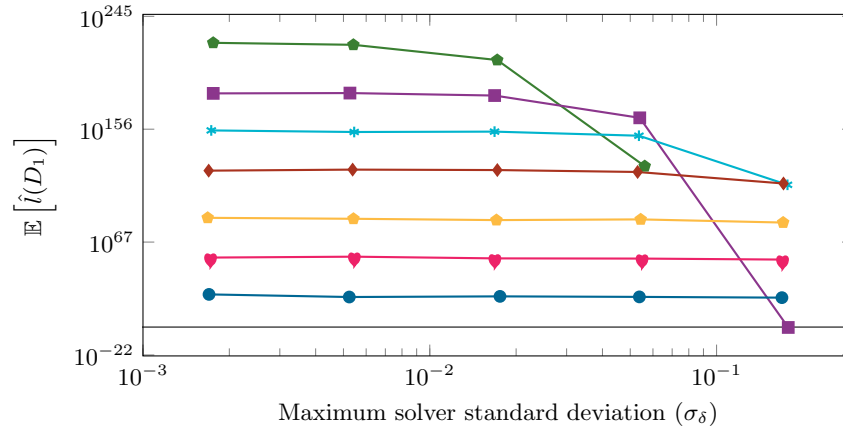
(a) Expectation of the proposal likelihood at $D_1 = 0.1$.
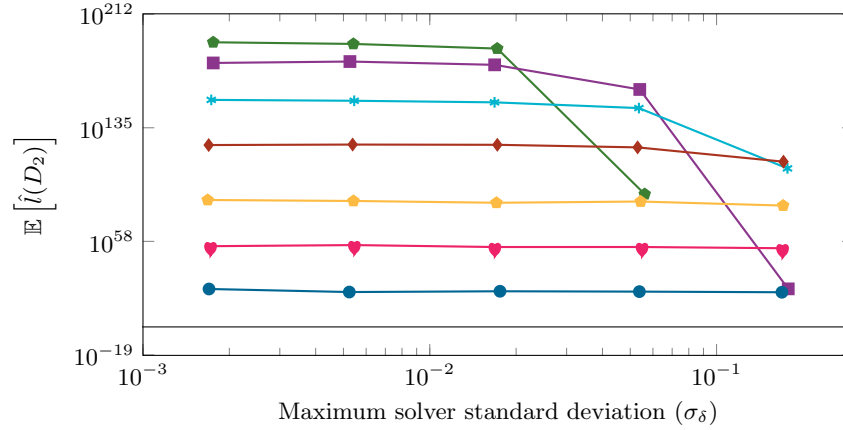
(b) Expectation of the current state's likelihood at $D_2 = 0.08$.
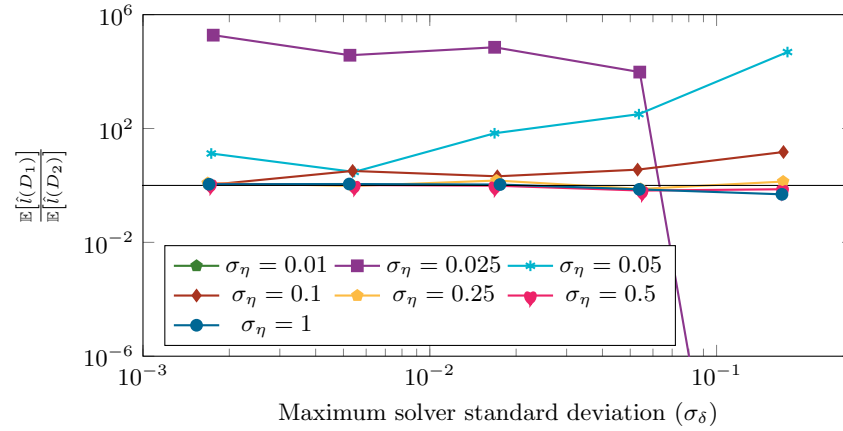
(c) Ratio of the expectations of the proposal and current state's likelihoods. The curve for $\sigma_\eta = 0.01$ is not visible due to taking values between 10^{21} and 10^{37}.

Fig. 1: Computing the expectation of the approximate likelihoods and the ratio of these expectations, given a discrete solution of the problem (1)–(2) perturbed by synthetic observation noise, for $D^* = D_1 = 0.1$ and $D_2 = 0.08$.
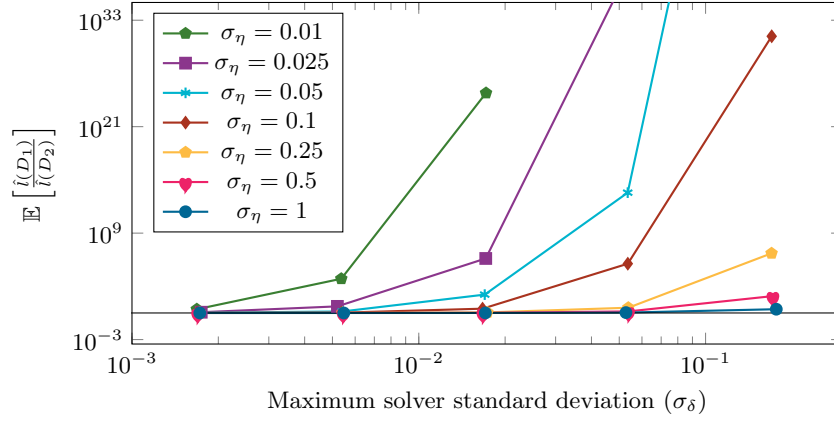
(a) Expectation of the approximate likelihood ratio.
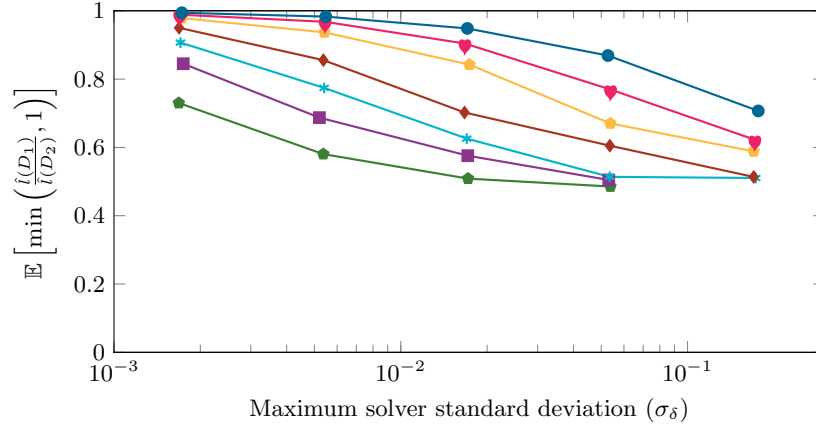
(b) Expectation of the approximate likelihood ratio, truncated to the range $[0, 1]$.

Fig. 2: Computing the expectation of the approximate likelihood ratio given a discrete solution of the problem (1)–(2) perturbed by synthetic observation noise for $D^* = D_1 = D_2 = 0.1$.
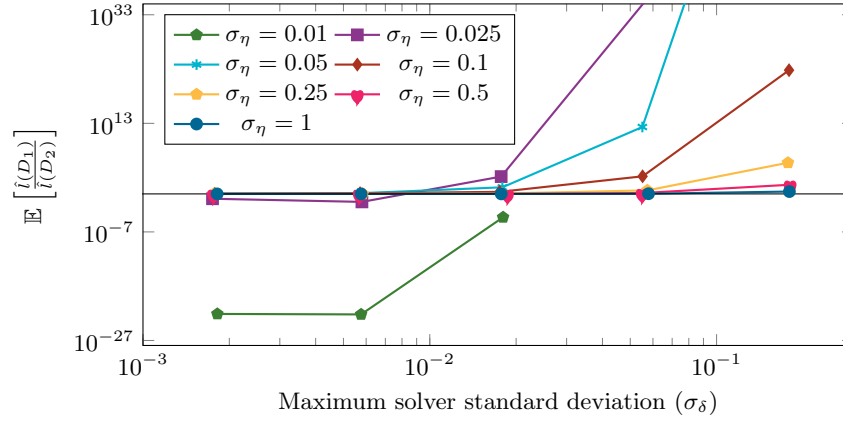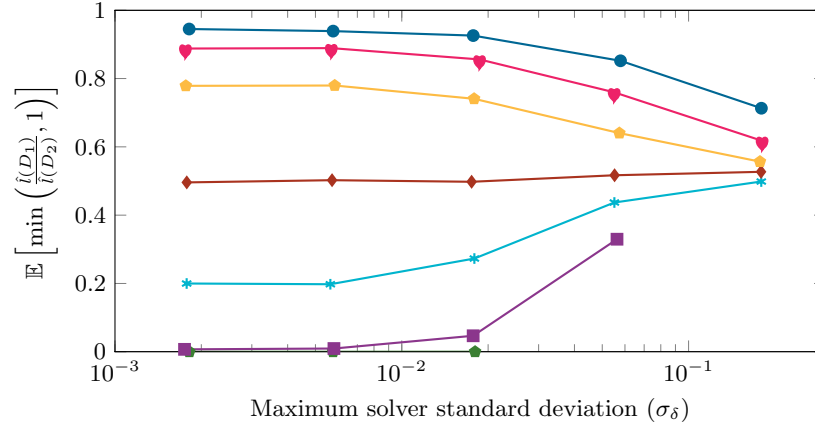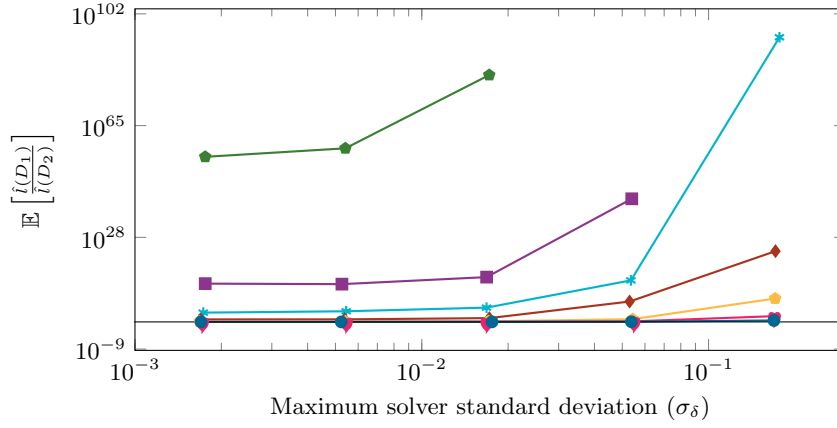
(a) Expectation of the approximate likelihood ratio.

(b) Expectation of the approximate likelihood ratio, truncated to the range $[0, 1]$.

Fig. 3: Computing the expectation of the approximate likelihood ratio given a discrete solution of the problem (1)–(2) perturbed by synthetic observation noise for $D^* = D_2 = 0.1$ and $D_1 = 0.08$.
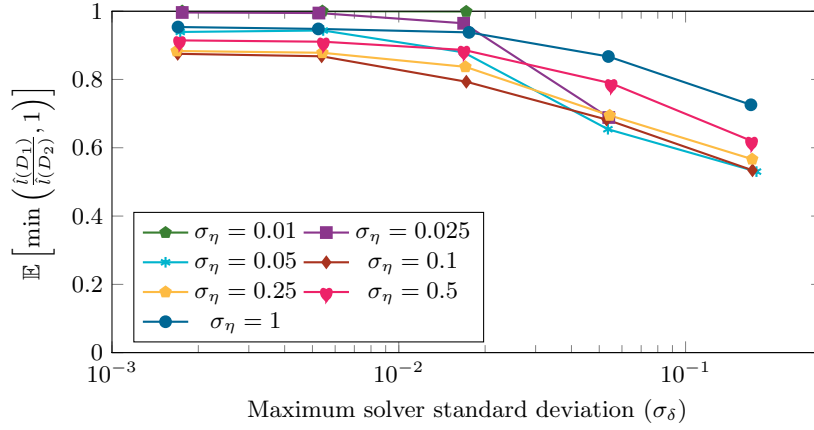
(a) Expectation of the approximate likelihood ratio.

(b) Expectation of the approximate likelihood ratio, truncated to the range $[0, 1]$.

Fig. 4: Computing the expectation of the approximate likelihood ratio given a discrete solution of the problem (1)–(2) perturbed by synthetic observation noise for $D^* = D_1 = 0.1$ and $D_2 = 0.08$.

We observe a similar blow-up pattern in Figure 3a to that in Figure 2a, but now observe that not all curves converge to 1 as σ_δ decreases. For sufficiently small values of σ_η, the computed approximate likelihood ratio correctly takes small values to denote the low likelihood of the transition being accepted. In Figure 3b, we see a similar convergence to the value 0.5 as σ_δ increases. As opposed to Figure 2b, we note that the curves converge to different values for small σ_δ, indicating that we are able to retrieve a sensible acceptance probability.

Similarly, Figure 4 presents equivalent plots for the transition from 0.1 to 0.08. In Figure 4a, we see a similar behavior to the previous figures, in that the curves start to blow up around $\sigma_\eta = \sigma_\delta$. For small values of σ_δ, we see the inverse of the behavior observed in Figure 3a. Namely, the expectation of the ratio converges to one from above as σ_η increases. Considering Figure 4b, we note a similar convergence towards the value 0.5 for larger values of σ_δ as in the previous figures. However, for smaller values of σ_δ we note that the curves are not sequentially ordered in their values of σ_η. The leftmost part of the curves decrease as σ_η increases from 0.01 through 0.1, but then increase again as σ_η further increases to 1. It appears that the initial decrease is due to the decrease of the corresponding sequence of curves in Figure 4a, i.e., when the expectation of the approximate likelihood ratio decreases, the minimum of 1 and this ratio will more often take values smaller than 1. As σ_η increases further, it is likely that the likelihood function becomes sufficiently flat that the variance induced by the Monte Carlo solver no longer has a strong impact on the computed approximate likelihood value. Hence, the variance of the approximate likelihood ratio will decrease, and the truncated ratio converges in expectation to 1.

5   Conclusions

We have conducted an initial study on the effect of using Monte Carlo simulations to evaluate approximate likelihood ratios, as used in Metropolis-Hastings. In Section 3 we presented a theorem, stating that the expectation of this ratio exists if the difference between the covariance matrix of the Monte Carlo error in the denominator δ_2 and that of the observation error η is positive definite. As the variance of the approximation error further decreases, the distribution of this ratio gains an increasing number of moments. We conducted a numerical study of the behavior of this ratio, confirming the existence criterion outlined in the theorem for transitions between two states with differing true likelihood values. When the forward model errors variance grows too large, we observe that the approximate likelihood ratio, truncated to the range $[0, 1]$, converges in expectation to the value 0.5. Hence, the approximate likelihood evaluations no longer provide any useful information for the accept-reject step of the Metropolis-Hastings algorithm.

We conclude that it is likely feasible to use Metropolis-Hastings sampling in settings where one must accept relatively large stochastic errors. However, it is important to assure that the forward model variance does not surpass the limits established in Theorem 1. In addition, there is a clear bias in the approx-

imate likelihood ratio and corresponding acceptance probability that decreases with the the forward model variance. These observation are useful for settings such as Monte Carlo simulation, where one can explicitly control the variance of the forward model. Theorem 1 also provides closed-form expressions for the moments of the distribution of the approximate likelihood ratio. We expect these expressions to be useful for algorithm design. One can use, e.g., estimates for the variance of the approximate likelihood ratio to produce error bounds for the computed acceptance probability and determine how many additional trajectories must be simulated in the forward model to make a clear acceptance or rejection decision in the Metropolis-Hastings algorithm.

Acknowledgments