Bayesian Inverse Problems with Monte Carlo Forward Models

Guillaume Bal,[*] Ian Langmore,[†] Youssef Marzouk[‡]

December 7, 2011

Abstract

The full application of Bayesian inference to inverse problems requires exploration of a posterior distribution that typically does not possess a standard form. In this context, Markov chain Monte Carlo (MCMC) methods are often used. These methods require many evaluations of a computationally intensive forward model to produce the equivalent of one independent sample from the posterior. We consider applications in which approximate forward models at multiple resolution levels are available, each endowed with a probabilistic error estimate. These situations occur, for example, when the forward model involves Monte Carlo integration. We present a novel MCMC method called MC^3 that uses low-resolution forward models to approximate draws from a posterior distribution built with the high-resolution forward model. The acceptance ratio is estimated with some statistical error; then a confidence interval for the true acceptance ratio is found, and acceptance is performed correctly with some confidence. The high-resolution models are rarely run and a significant speed up is achieved.

Our multiple-resolution forward models themselves are built around a new importance sampling scheme that allows Monte Carlo forward models to be used efficiently in inverse problems. The method is used to solve an inverse transport problem that finds applications in atmospheric remote sensing. We present a path-recycling methodology to efficiently vary parameters in the transport equation. The forward transport equation is solved by a Monte Carlo method that is amenable to the use of MC^3 to solve the inverse transport problem using a Bayesian formalism.

Keywords: linear transport; perturbation Monte Carlo; Bayesian; importance sampling; inverse problems; Markov chain Monte Carlo

1 Introduction

The Bayesian methodology has proven to be a convenient framework to solve inverse problems from available data with limited information. Sampling the posterior distribution is the major computational difficulty associated with Bayesian inversion. This

[*] Department of Applied Physics and Applied Mathematics, Columbia University, gb2030@columbia.edu

[†] Corresponding author. Department of Applied and Applied Mathematics, Columbia University, 415-272-6321, ianlangmore@gmail.com

[‡] Department of Aeronautics and Astronautics, Massachusetts Institute of Technology, ymarz@mit.edu

distribution is often high dimensional and does not possess a standard form in most applications of interest. Several authors have found ways to use low-resolution forward models to speed MCMC simulation. In particular, we note the "two-level" chains developed in [14, 7, 5, 15]. Two-level chains use low- and high-resolution forward models in a Metropolis-Hastings scheme and may be summarized as follows: Proposals to sample the posterior distribution are generated as usual, and then pre-accepted using an acceptance ratio built with a low-resolution model. If pre-accepted, they are then accepted or rejected using an acceptance ratio that measures a misfit between low and high resolution models. As a result, the high-resolution forward model is rarely run—and when it is, acceptance is almost assured.

Using these two-level chains as a starting point, we introduce a multi-level MCMC scheme adapted to Monte Carlo forward models, called MC^3 . Since our forward model output is itself a Monte Carlo estimate (hence the name $(MC)^3 = MC \times MCMC$), we can control resolution by adjusting the number of random samples drawn. This gives rise to multiple resolution levels and corresponding forward models $F_1, F_2, \ldots$, converging to F_∞. Since the Monte Carlo estimate is a sum of independent random variables, the error is well approximated by a Gaussian random variable and we can obtain *a posteriori* estimates of its variance. Thus, when we utilize a low-resolution model we are able to accurately calibrate its effect on the posterior at that level. These *finite-resolution posteriors* π_i (6) are used in the first stage of a two-level scheme as described above. For the second stage we do not fix an upper resolution level. Instead, we progressively increase the resolution j until the second-level acceptance probability β_{ij} has a desired level of accuracy. The desired "accuracy" corresponds to making the correct decision with confidence $\lambda \in (0, 1)$. By "correct decision" we mean the acceptance/rejection choice corresponding to an evaluation of $\beta_{i\infty}$, where $\beta_{i\infty}$ uses the exact forward model F_∞. This modification necessarily introduces some distortion of the resultant posterior. We obtain a stationary distribution π_λ for an idealized version of our chain and characterize its deviation from π_∞ in Theorem 2.1 below. This deviation can be made arbitrarily small and depends on λ and the initial resolution level. Very little distortion is present if (i) λ is close to 1 independent of the accuracy at the initial resolution level; or (ii) if the initial resolution level is chosen to be reasonably accurate.

Since sampling of the posterior distribution requires a large number of forward solves (at different resolutions in the MC^3 scheme) for different values of the inversion parameters, it is important to reduce the computational cost of these forward solves as well. In applications to remote sensing, one of the main reasons Monte Carlo forward models are slow is that they are measuring the probability of a rare event (a photon from the sun reaching a small detector). This rare event sampling can be accelerated using an importance sampling scheme similar to *perturbation Monte Carlo* developed for medical imaging [10, 9, 4]. This *path recycling* scheme sends many paths once through a reference atmosphere and then stores only those that hit the detector. Then, to simulate detector hits in another atmosphere an importance sampling scheme is used whereby the original paths are re-weighed to account for changes in atmospheric absorption/scattering (provided these changes result in an equivalent measure, as is the case in photon propagation modeled by a transport equation). We differ from standard perturbation Monte Carlo schemes in that our path recycling scheme also provides multiple-resolution forward models that are used to speed up the sampling of the posterior distribution. To that end, our finite-resolution forward model F_j recycles a $j-$dependent number H_j of all available paths, but uses information from many more stored paths H_{max} in order to reduce variance (see (24)). Although our paper focuses on the MC^3 algorithm, we obtain a significant variance reduction using our path-recycling scheme, which appears to be novel. The theoretical variance reduction results are summarized in Theorems 3.1 and 3.2.

Both the MC^3 and path-recycling methods were motivated by an inverse problem in atmospheric imaging. Here we are presented with a passive remote sensing problem involving the identification of an absorbing plume (e.g., a pollution cloud) using detected sunlight. The detector is spatially small and measures incoming light from fifteen different angles. Our atmosphere and plume are parameterized with five unknown parameters (plume absorption cross-section α, plume center (x_p, z_p), plume radius ρ, and atmospheric constant c). Since the measurements are sparse, introducing prior information in the Bayesian setting is useful. Since the number of parameters to recover is small, full exploration of the posterior distribution is both feasible and desirable. Unfortunately, because the detector is small, most Monte Carlo forward paths do not reach it. Importance sampling schemes can increase the probability of a detector hit [2], but we instead use the path recycling scheme described above, as it is better suited to the multi-resolution inverse problem. On its own, path recycling reduces forward model run-time by a factor of thousands, but this is still too slow for posterior exploration. We therefore use path-recycling in conjunction with MC^3 and present results in Section 4. We stress again that because of the enormous computational cost involved in transport simulations and the Bayesian framework, we would not be able to produce such results on desktop architectures without the huge variance reductions afforded by path-recycling and MC^3 .

While we will demonstrate the MC^3 scheme on an atmospheric imaging problem, we note that many other applications use forward model predictions that are Monte Carlo estimates. These include more general problems of radiation transport, molecular dynamics simulations, floating random walk methods for capacitance extraction in electronic design, determination of pricing measures in finance, and DSMC (Direct simulation Monte Carlo) methods in fluid flow.

In Section 2 we describe a MCMC scheme for Bayesian inverse problems where the forward model is available at multiple resolution levels and endowed with tight (probabilistic) error estimates. In Section 3 we develop our path-recycling forward model. In Section 4 we combine techniques from Sections 2 and 3 to solve a five-dimensional Bayesian inverse problem.

2 Multi-Level Metropolis-Hastings

Here we describe Metropolis-Hastings schemes that use multiple-resolution forward models (along with accurate error estimates) to significantly speed up a Metropolis Hastings scheme. In particular, MC^3 is presented as an extension of the two-level schemes developed in [7, 5, 15]. The novelty of MC^3 is evaluation of an acceptance ratio up to some confidence (the $\lambda-$*acceptance method*); this aspect of MC^3 may be used with standard Metropolis or any scheme where an acceptance ratio must be evaluated. We refer the reader to [11] for an introduction to Bayesian inverse problems.

2.1 Basic setup

The *Bayesian* viewpoint models the unknown quantities as random variables. Our unknown is a random vector $x \in \Gamma \subset \mathbb{R}^n$ with *prior* probability density $f_{prior}(x)$. This is the distribution we *assume* (from prior knowledge) on x before any data are collected.

We assume our *data* $\mathbf{d} \in \mathbb{R}^m$ are given by an *infinite-resolution forward model* $F_\infty : \mathbb{R}^n \to \mathbb{R}^m$ plus an additive independent (Gaussian) noise term E.

$$\mathbf{d} := F_\infty(X) + E, \qquad E \sim \mathcal{N}(0, \Sigma_E), \qquad E \perp\!\!\!\perp X. \tag{1}$$

The methods presented here are independent of the choice of prior density. The additive Gaussian noise model will simplify the algebra involved in the $\lambda-$acceptance evaluation, but is not necessary either.

Our *infinite resolution posterior* is thus

$$\pi_\infty(x \,|\, \mathbf{d}) \propto f_{prior}(x) \frac{1}{|\Sigma_E|^{1/2}} \exp\left\{ -\frac{1}{2} \|\mathbf{d} - F_\infty(x)\|^2_{\Sigma_E^{-1}} \right\}, \tag{2}$$

where for vectors $v \in \mathbb{R}^m$ and matrices $A \in \mathbb{R}^{m \times m}$ we define $\|v\|^2_A := v^T A v$.

We do not have access to F_∞, but instead a sequence of approximate models $F_1, F_2, \cdots$. In our framework, the approximate models are an unbiased sum of i.i.d. random variables, and so we are justified using a Gaussian error model

$$F_j(x) \sim \mathcal{N}(F_\infty(x), \operatorname{Var}\{F_j(x)\}). \tag{3}$$

We assume that $\operatorname{Var}\{F_j(x)\} =: \Sigma_j(x)$ can be estimated accurately. This is the case when the forward model is solved by Monte Carlo.

Equation (3) leads to an *enhanced noise model* (see e.g. [1, 12], or [11] as well as [16] for more discussion of model error and inverse problems) at resolution level j:

$$\mathbf{d} := F_j(x) + E_j(x) + E, \qquad E_j(x) \sim \mathcal{N}(0, \Sigma_j(x)), \quad E_j(x) \perp\!\!\!\perp E, \quad x \perp\!\!\!\perp E, \tag{4}$$

and a likelihood at resolution level j,

$$\pi_j(\mathbf{d} \,|\, x) = \frac{1}{(2\pi)^{m/2} |\Sigma_E + \Sigma_j(x)|^{1/2}} \exp\left\{ -\frac{1}{2} \|\mathbf{d} - F_j(x)\|^2_{(\Sigma_E + \Sigma_j(x))^{-1}} \right\}. \tag{5}$$

Instead of one posterior, we have a sequence of *finite resolution posteriors* $\{\pi_j(x \,|\, \mathbf{d})\}_{j=1}^\infty$:

$$\pi_j(x \,|\, \mathbf{d}) \propto f_{prior}(x) \pi_j(\mathbf{d} \,|\, x). \tag{6}$$

Assume that for a.e. fixed $x \in \operatorname{supp}(f_{prior})$, $F_j(x) \to F_\infty(x)$ and $\operatorname{Var}\{F_j(x)\} \to 0$. Then, as a consequence of dominated convergence, we can approximate expectations against π_∞ by using π_j in the sense that for all f such that $\int |f(x)| \pi(x) \, \mathrm{d}x < \infty$,

$$\int f(x) \pi_j(x \,|\, \mathbf{d}) \, \mathrm{d}x \to \int f(x) \pi_\infty(x \,|\, \mathbf{d}) \, \mathrm{d}x. \tag{7}$$

From here on we omit the explicit conditioning on $\mathbf{d}$ and simply write $\pi_j(x)$, $j = 1, 2, \ldots, \infty$.

2.2 The algorithms

Here we present three MCMC chains that take advantage of our multiple resolution models F_j and the error estimates.

2.2.1 Metropolis Hastings at resolution level j

As a starting point, we first present the standard Metropolis-Hastings algorithm (algorithm 1) [17], using our finite resolution posterior π_j from (6).

Algorithm 1 Metropolis-Hastings at resolution level j

```
1: Given $x_k$, draw $y \sim q(\cdot \,|\, x_k)$
2: Put
   $$\alpha_j(x_k, y) := \min\left\{1, \frac{q(x_k \,|\, y)\pi_j(y)}{q(y \,|\, x_k)\pi_j(x_k)}\right\}$$
3: With probability $\alpha$, **accept** and set $x_{k+1} = y$. Otherwise set $x_{k+1} = x_k$
```

Note that the transition kernel associated with standard Metropolis-Hastings is

$$K(x, y) : = \alpha_j(x, y)q(y \,|\, x) + \delta_x(y)\left(1 - \int \alpha_j(x, y')q(y' \,|\, x)\,\mathrm{d}y'\right),$$

where $\delta_x(\cdot)$ is the Dirac measure on $\mathbb{R}^n$ concentrated at x.

2.2.2 Two-level Metropolis-Hastings

The two-level chain (algorithm 2, as developed in [7, 5, 15]) is as follows:

Algorithm 2 Two-Level Metropolis Hastings at resolution levels-ij

```
1: Given $x_k$, draw $y \sim q(\cdot \,|\, x_k)$
2: Put
   $$\alpha_i(x_k, y) := \min\left\{1, \frac{q(x_k \,|\, y)\pi_i(y)}{q(y \,|\, x_k)\pi_i(x_k)}\right\}$$
3: With probability $\alpha_i(x_k, y)$, **pre-accept** $y$. Otherwise set $y \leftarrow x_k$.
4: The second-level proposal is now $y$, effectively drawn from
   $$q_i(y \,|\, x_k) = \alpha_i(x_k, y)q(y \,|\, x_k) + \delta_{x_k}(y)\left(1 - \int \alpha_i(x_k, y)q(y \,|\, x_k)\,\mathrm{d}y\right).$$
   Set
   $$\beta_{ij}(x_k, y) : = \min\left\{1, \frac{q_i(x_k \,|\, y)\pi_j(y)}{q_i(y \,|\, x_k)\pi_j(x_k)}\right\}.$$
5: With probability $\beta_{ij}(x_k, y)$, **accept** and set $x_{k+1} = y$. Otherwise set $x_{k+1} = x_k$.
```

Note that there is no need to compute the integral defining q_i since

$$\beta_{ij}(x_k, y) := \min\left\{1, \frac{q_i(x_k \mid y)\pi_j(y)}{q_i(y \mid x_k)\pi_j(x_k)}\right\} = \min\left\{1, \frac{\pi_i(x_k)\pi_j(y)}{\pi_i(y)\pi_j(x_k)}\right\}.$$

The equality of the two minimizers above is shown in (16). One can show that under reasonable assumptions on the proposal q, $\pi_j(\cdot)$ is the stationary distribution of the two-level chain (see [7, 5]).

Remark 2.1. The two-level chain (algorithm 2) can be viewed in two ways:

(a) If we are "pre-rejected" ($y \leftarrow x_k$) in line 3 of algorithm 2, then we are guaranteed $\beta_{ij} = 1$ and we need not evaluate β_{ij} (or F_j). In this sense the pre-acceptance stage filters out poor draws. This in turn allows one to make more "courageous" proposals (e.g., from the prior) and not waste time evaluating the high-resolution forward model on them.

(b) In algorithm 1, the "ideal" proposal is π_j. Since $\pi_i \approx \pi_j$ we can also view the two-level chain as a low-resolution chain producing high-quality proposals for a high-resolution chain.

2.2.3 The multi-stage algorithm MC^3

The MC^3 chain is now introduced. We will re-use α_i from algorithm 2 and replace β_{ij} with

$$\beta_{i\infty}(x_k, y) := \min\left\{1, \frac{q_i(x_k \mid y)\pi_\infty(y)}{q_i(y \mid x_k)\pi_\infty(x_k)}\right\} = \min\left\{1, \frac{\pi_i(x_k)\pi_\infty(y)}{\pi_i(y)\pi_\infty(x_k)}\right\}.$$

The equality of the two minimizers above is similar to the case of β_{ij}, which is shown in (16).

If we used $\beta_{i\infty}$ in place of β_{ij} in the two-level chain (or in place of α_i in the one-level chain) then the invariant distribution would be the infinite resolution posterior $\pi_\infty(\cdot)$. Since in practice we do not have access to $\pi_\infty(\cdot)$, we instead use an asymptotic formula to approximate draws from π_∞ using a level of confidence $\lambda \in (0, 1)$.

The two-stage algorithm 2 performs its second-level acceptance step as follows:

Step 1: Draw $u \sim \mathcal{U}[0, 1]$

Step 2: If $u < \beta_{ij}$ accept, otherwise reject.

The multi-stage algorithm MC^3 replaces step 2 with the λ-*acceptance step*:

Step 2a: Using data generated at resolution level j, determine if, with confidence λ, we can say $\beta_{i\infty} > u$ or $\beta_{i\infty} < u$

Step 2b: If $\beta_{i\infty} > u$ with confidence λ, then accept. If $\beta_{i\infty} < u$ with confidence λ, then reject. If our confidence in $\beta_{i\infty}$ versus u is less than λ, then increase resolution j and goto step 2a.

The mechanics of step 2a and what we mean by the λ-*acceptance step* are explained in section 2.3. Roughly speaking, the ratio $\pi_\infty(y)/\pi_\infty(x_k)$ in $\beta_{i\infty}$ is replaced by $\pi_j(y)/\pi_j(x_k)$ plus a Gaussian error term.

Algorithm 3 MC^3 at initial level i, final level j_{max}, and confidence λ

```
Given $x_k$, draw $y \sim q(\cdot \,|\, x_k)$
Put
    $$\alpha_i(x_k, y) := \min\left\{1, \frac{q(x_k \,|\, y)\pi_i(y)}{q(y \,|\, x_k)\pi_i(x_k)}\right\}$$
With probability $\alpha_i(x_k, y)$, **pre-accept** $y$. Otherwise set $y \leftarrow x_k$.
The second-level proposal is now $y$, effectively drawn from
    $$q_i(y \,|\, x_k) = \alpha_i(x_k, y)q(y \,|\, x_k) + \delta_{x_k}(y)\left(1 - \int \alpha_i(x_k, y)q(y \,|\, x_k)\,\mathrm{d}y\right).$$
Draw $u \sim \mathcal{U}[0, 1]$, set $j \leftarrow i$, $test \leftarrow indeterminate$
while ($test$==$indeterminate$) AND ($j \leq j_{max}$) do
  if with confidence $\lambda$ at $j^{th}$ level, we have $\beta_{i\infty} > u$ then
    Accept and set $x_{k+1} \leftarrow y$.    Set $test \leftarrow determinate$
  else if with confidence $\lambda$ at $j^{th}$ level, we have $\beta_{i\infty} < u$ then
    Reject and set $x_{k+1} \leftarrow x$.    Set $test \leftarrow determinate$
  end if
  $j \leftarrow j + 1$
end while
```

This algorithm requires the definition of intervals of confidence parameterized by the confidence level $\lambda \in (0, 1)$. Assuming that the Monte Carlo forward simulations are sufficiently accurate, then the statistical errors are accurately described by a Gaussian approximation as an application of the central limit theorem. The resulting intervals of confidence are analyzed in the following section, which is an intrinsic part of the definition of MC^3. More precisely, the notion of inequalities with confidence λ at j^{th} level is described in (14) below.

2.3 The asymptotic confidence interval

We use the *delta method* to derive an asymptotic confidence interval for the acceptance test used in algorithm 3. Note that in our numerical results we will have the following isotropic model for the likelihood function: $\Sigma_E = \sigma_E^2 I_m$. Since this simplifies the presentation in this section significantly, we only consider this special case here. Generalizations to non-iid (or even non-Gaussian) additive error are straightforward but not considered for concreteness.

We first reduce the test $\beta_{i\infty} > u$ to a form more amenable to an interval test. First recall,

$$\beta_{i\infty}(x, y) : = \min\left\{1, \frac{q_i(x \,|\, y)\pi_\infty(y)}{q_i(y \,|\, x)\pi_\infty(x)}\right\} = \min\left\{1, \frac{\pi_i(x)\pi_\infty(y)}{\pi_i(y)\pi_\infty(x)}\right\}.$$

Define

$$\bar{X} = \mathbf{d} - F_\infty(x), \qquad \bar{Y} := \mathbf{d} - F_\infty(y), \qquad \varphi(v, w) := |v|^2 - |w|^2, \tag{8}$$

so that

$$\frac{\pi_\infty(y)}{\pi_\infty(x)} = \frac{f_{prior}(y)}{f_{prior}(x)} \exp\left\{ \frac{1}{2\sigma_E^2} \varphi(\bar{X}, \bar{Y}) \right\}.$$

Therefore, our "ideal" acceptance criteria can be written

$$\text{accept} \Leftrightarrow \beta_{i\infty} > u \Leftrightarrow \varphi(\bar{X}, \bar{Y}) > 2\sigma_E^2 \log\left(u \frac{f_{prior}(x)\pi_i(y)}{f_{prior}(y)\pi_i(x)} \right). \tag{9}$$

Second, we derive an asymptotic relation between $\varphi(\bar{X}, \bar{Y})$ and data available at resolution level j. Defining

$$X_j := \mathbf{d} - F_j(x), \qquad Y_j := \mathbf{d} - F_j(y), \tag{10}$$

we have

$$\varphi(X_j, Y_j) - \varphi(\bar{X}, \bar{Y}) \approx \nabla\varphi(\bar{X}, \bar{Y})^T (X_j - \bar{X}, Y_j - \bar{Y}).$$

This approximation is accurate for small $|\bar{X} - X_j|$ and $|\bar{Y} - Y_j|$, which is the case so long as $F_j(x) \approx F_\infty(x)$ and $F_j(y) \approx F_\infty(y)$. Our assumption (3) means that

$$(X_j - \bar{X}, Y_j - \bar{Y}) \sim \mathcal{N}(0, \mathrm{Cov}\,(X_j, Y_j)).$$

Note that we define the covariance and variance of random vectors X, Y by

$$\mathrm{Cov}\,(X, Y) := \mathbb{E}\left\{ (X - \mathbb{E}\{X\})\,(Y - \mathbb{E}\{Y\})^T \right\}, \qquad \mathrm{Var}\{X\} := \mathrm{Cov}\,(X, X). \tag{11}$$

We therefore approximate

$$\varphi(X_j, Y_j) \approx \mathcal{N}(\varphi(\bar{X}, \bar{Y}), \mu_j^2), \quad \mu_j^2 := \nabla\varphi(\bar{X}, \bar{Y})^T \mathrm{Cov}\,(X_j, Y_j)\, \nabla\varphi(\bar{X}, \bar{Y}). \tag{12}$$

This argument may be made rigorous, see e.g., [3] for a proof of the delta method in one dimension. Assuming (12), (9) becomes

$$\text{accept} \Leftrightarrow \beta_{i\infty} > u \Leftrightarrow \mathbb{E}\{\varphi(X_j, Y_j)\} > 2\sigma_E^2 \log\left(u \frac{f_{prior}(x)\pi_i(y)}{f_{prior}(y)\pi_i(x)} \right). \tag{13}$$

Third, we use (12) to derive a confidence interval around a random variable $\varphi(X_j, Y_j)$ that contains the mean $\mathbb{E}\{\varphi(X_j, Y_j)\}$ with probability $\lambda \in (0, 1)$. We do not quite know μ_j in (12) and must content ourselves with an estimate $\hat{\mu}_j$, which can be easily computed using the random variables defining X_j, Y_j and is therefore obtained at resolution level j; see the description in section 3.3.2 for the inverse transport problem. Briefly, $F_j(x)$, $F_j(y)$ (and hence X_j, Y_j) are sums of i.i.d. random variables, so we use the empirical covariance matrix as an estimate of $\mathrm{Cov}\,(X_j, Y_j)$. Then we use X_j, Y_j in place of $\bar{X}$, $\bar{Y}$ in $\nabla\varphi(\bar{X}, \bar{Y})$. This is put together to give $\hat{\mu}_j^2$.

With $z_\lambda > 0$ such that $\mathrm{P}[-z_\lambda < \mathcal{N}(0, 1) < z_\lambda] = \lambda$, we define our confidence interval

$$C_\lambda := (\varphi(X_j, Y_j) - z_\lambda \hat{\mu}_j\,,\, \varphi(X_j, Y_j) + z_\lambda \hat{\mu}_j).$$

We say that with confidence λ, we have $\varphi(\bar{X}, \bar{Y}) \in C_\lambda$. See [3] for an introduction to confidence intervals.

Finally, to use this confidence interval in the second stage of algorithm 3, we start from (13) and then (with $\gtrapprox$, $\lessapprox$ denoting $>$, $<$ with confidence λ at the j^{th} level, i.e., for the $j-$dependent confidence interval C_λ):

$$\begin{aligned} 2\sigma_E^2 \log\left(u\frac{f_{prior}(x)\pi_i(y)}{f_{prior}(y)\pi_i(x)}\right) &< \varphi(X_j, Y_j) - z_\lambda\hat{\mu}_j \quad \Rightarrow \quad \beta_{i\infty} \gtrapprox u \quad \Rightarrow \quad \text{accept} \\ 2\sigma_E^2 \log\left(u\frac{f_{prior}(x)\pi_i(y)}{f_{prior}(y)\pi_i(x)}\right) &> \varphi(X_j, Y_j) + z_\lambda\hat{\mu}_j \quad \Rightarrow \quad \beta_{i\infty} \lessapprox u \quad \Rightarrow \quad \text{reject.} \end{aligned} \tag{14}$$

The above formulas summarize the multi-stage algorithm MC^3 . At a given level j, we either accept or reject according to the above probabilities, and go to the next level $j+1$ otherwise.

2.4 Chain analysis

The two-level and multi-level chains do not have as limiting distribution π_∞ (this would require access to F_∞). Instead, the two-level chain limits to π_j (a broader version of π_∞ defined explicitly in (6)) and MC^3 has π_λ (a distorted version of π_∞ defined implicitly in Theorem 2.1).

2.4.1 Two-level chain analysis

For the two-level chain, following [7, 5] we have that (for $x_k \neq y$)

$$q_i(x_k \,|\, y) = \alpha_i(y, x_k) q(x_k \,|\, y) = \frac{\pi_i(x_k)}{\pi_i(y)} q(y \,|\, x_k) \alpha_i(x_k, y) = \frac{\pi_i(x_k)}{\pi_i(y)} q_i(y \,|\, x_k). \tag{15}$$

As a result, we have

$$\beta_{ij}(x_k, y) = \min\left\{1, \frac{\pi_i(x_k)\pi_j(y)}{\pi_i(y)\pi_j(x_k)}\right\}. \tag{16}$$

The transition kernel of algorithm 2 is thus

$$K_{ij}(x_k, y) = \beta_{ij}(x_k, y) q_i(y \,|\, x_k) + \delta_{x_k}(y)\left(1 - \int \beta_{ij}(x_k, y) q_i(y \,|\, x_k)\, \mathrm{d}y\right).$$

It satisfies the detailed balance equation

$$\pi_j(x_k) K_{ij}(x_k, y) = \pi_j(y) K_{ij}(y, x_k).$$

Under standard assumptions, π_j may be shown to be the limiting distribution [7, 5].

2.4.2 Approximate MC^3 chain analysis

We present here an approximation of the MC^3 chain for which we are able to obtain an invariant distribution.

To construct the approximation, first assume that every time the model F_j is used, a new set of i.i.d. random variables are drawn to compute $F_j(x)$. So for example, if we visit a particular x twice (so that $F_j(x)$ is computed twice), then the two computations are independent. This assumption is not met by our forward model described in section 3. There, the same set of paths is recycled for all x.

As a second assumption, we idealize the implementation of the accept/reject conditions in (14). Our idealization assumes that we accept/reject once we know, with confidence exactly equal to λ, that $\beta_{i\infty} < u$ or $\beta_{i\infty} > u$. While this is a correct interpretation of a confidence interval, this does not strictly hold (in our MC^3 implementation) for two reasons:

(a) The normal approximation (delta method) is only asymptotically correct.

(b) We consider the $\lambda-$acceptance test (14) for multiple lower resolution levels before finally resolving the question at a high enough level. In detail: we increase j until we can evaluate (9) with confidence greater than or equal to λ. We always initially take $j = i$, and most of the time this results in a confidence interval C_κ for some $\kappa_j > 0$. So long as $\lambda > \kappa_j$, we keep increasing resolution, until at some point $\kappa_j > \lambda$ and we stop.

Given these two assumptions, MC^3 becomes an algorithm (the λ-approximate algorithm) that acts like two-level MCMC but makes the wrong decision at every step with probability λ. We present this as algorithm 4 and analyze the error made on the invariant measure.

We are not able to find the invariant measure associated to MC^3 (or even prove that one exists). The results of this section should then be interpreted as advisory. Indeed, we use them to set the levels i and λ later in section 4, but conclude that our method works for the problem at hand only after obtaining numerical results.

Note that algorithm 4 makes use of the functions q_i, $\beta_{i\infty}$ from algorithm 3.

Algorithm 4 λ-approximate algorithm

```
Starting from $x_k$, draw $y \sim q_i(\cdot \,|\, x_k)$.
With probability $\beta_{i\infty}(x_k, y)$, set $z \leftarrow y$ and $z^c \leftarrow x_k$,
else set $z \leftarrow x_k$ and $z^c \leftarrow y$
With probability $\lambda$, set $x_{k+1} \leftarrow z$
else set $x_{k+1} \leftarrow z^c$
```

Using algorithm 4 we accept y (recall that we could have $y = x$ at this point) with probability

$$R_\lambda(x) = \lambda r_{i\infty}(x) + (1-\lambda)(1 - r_{i\infty}(x)), \qquad r_{i\infty}(x) := \int q_i(y \,|\, x)\beta_{i\infty}(x, y)\,\mathrm{d}y.$$

If reject y, it must then be done with probability $1 - R_\lambda(x)$. So algorithm 4 has transition kernel

$$K_\lambda(x, y) := R_\lambda(x)\frac{q_i(y \,|\, x)\beta_{i\infty}(x, y)}{r_{i\infty}(x)} + (1 - R_\lambda(x))\delta_x(y).$$

We have the following theorem, whose proof we relegate to the appendix.

Theorem 2.1. *If $\lambda \in (0, 1]$ the transition kernel K_λ associated to algorithm 4 has invariant density*

$$\pi_\lambda(x) \propto \frac{\pi_\infty(x)}{m_\lambda(x)}, \qquad m_\lambda(x) := \frac{R_\lambda(x)}{r_{i\infty}(x)} = \lambda + \frac{(1-\lambda)(1 - r_{i\infty}(x))}{r_{i\infty}(x)}.$$

If we assume further that $E_+ := \{x : \pi_\infty(x)\}$ is open and connected, and for every bounded $G \subset E_+$ there exist constants C_j, C_q, D_j, δ such that for $x, y \in G$ with $|x-y| < \delta$ and $j = 1, 2, \ldots$

(i) $C_j \pi_\infty(x) \leq \pi_j(x) < D_j$

(ii) $C_q \pi_\infty(y) \leq q(y \,|\, x)$

then we have the following convergence result: If $f \in L^1(\pi_\lambda)$, then

$$\lim_{N\to\infty} \frac{1}{N} \sum_{k=1}^{N} f(x^k) = \int f(x)\pi_\lambda(x)\, dx,$$

and with $\|\cdot\|_{TV}$ the total variation *distance (see* (17)*), μ an arbitrary initial distribution, $K_\lambda^n(x,\cdot)$ the measure for n iterations of the transition kernel K_λ, and Π_λ the distribution associated with π_λ*

$$\left\| \int K_\lambda^n(x,\cdot)\mu(\,dx) - \Pi_\lambda \right\|_{TV} \to 0, \qquad \textit{monotonically.}$$

Remark 2.2. Some heuristics are evident:

(a) When λ is smaller, $m_\lambda(x)$ is more dependent on $r_i(x)$, hence will vary more with x, and hence more distortion is present.

(b) If one wishes to maintain a fixed amount of distortion (measured in some metric), i can be decreased only if λ is simultaneously increased.

2.4.3 Distortion example

To investigate the distortion of the measure and verify the heuristics in remark 2.2 we compute explicitly the distortion in a simple one-dimensional case. The target distribution and the proposal are both Gaussian.

$$\pi_\infty(x) \propto \exp\left\{-x^2/(2\sigma_E^2)\right\}, \qquad q(x \,|\, y) = q(x) \propto \exp\left\{-x^2/(2\sigma_q^2)\right\}.$$

This roughly corresponds to a forward+noise model $\mathbf{d} = X + E$ where $\mathbf{d} = 0$, $E \sim \mathcal{N}(0, \sigma_E^2)$ and $X \sim \mathcal{U}[-M, M]$ for some $M \gg 1$ so that the prior has negligible impact on the posterior. To approximate (6) we give π_i broadened likelihood with variance $\sigma_E^2 + \sigma_i^2$. In other words,

$$\pi_i(x) \propto \exp\left\{ -\frac{x^2}{2\left(\sigma_E^2 + \sigma_i^2\right)} \right\}, \qquad \delta > 0.$$

Using this model, we can numerically compute m_λ and then the distance between the distributions Π_∞, Π_λ (corresponding to π_∞, π_λ) can be measured with the *total variation* distance:

$$\|\Pi_\infty - \Pi_\lambda\|_{TV} := \sup_{A\subset\mathbb{R}} |\Pi_\infty(A) - \Pi_\lambda(A)| = \frac{1}{2}\int_{\mathbb{R}} |\pi_\infty(x) - \pi_\lambda(x)|\, \mathrm{d}x. \tag{17}$$

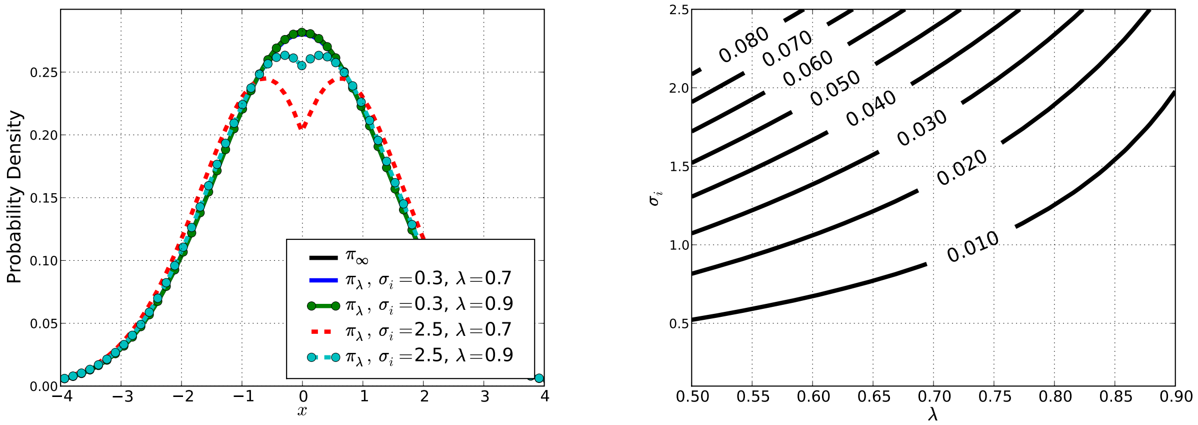

Figure 1: Left: Plot of π_λ for different values of λ, σ_i. Maximum distortion occurs near the peak of the distribution. Right: Contour plot of lines of constant total variation error $\|\Pi_\infty - \Pi_\lambda\|_{TV}$.

The first equality is a definition and the second can be seen by integrating $(\pi_\infty - \pi_\lambda)\mathbf{1}_A + (\pi_\lambda - \pi_\infty)\mathbf{1}_{\mathbb{R}\setminus A}$ where $A = \{x : \pi_\infty(x) \geq \pi_\lambda(x)\}$.

In Figure 1 we visualize this simple experiment. We set $\sigma_E = 1$, $\sigma_q = 3$, and sweep λ and σ_i. We see that maximum distortion occurs indeed for small λ and large σ_i. While decreasing σ_q (still holding $\sigma_E^2 < \sigma_E^2 + \sigma_i^2 < \sigma_q^2$) increases error, the same contour-line shape remains.

2.5 Numerical comparison of chains

We use the one-level, two-level, and MC^3 chains with a variety of parameter values to generate samples. The posterior distribution of interest fits the framework in section 2.1; it is in fact the atmospheric imaging problem alluded to in the introduction and described in section 4. Also see section 4 for a plot of the posterior. The purpose of the present section, however, is to compare chain performance for different choices of initial resolution level i, final level j_{max}, and confidence λ.

We compute the *autocorrelation time*, for each one-dimensional marginal of samples (call this AT_ℓ, $\ell = 1, \dots, n$). The autocorrelation time for a stationary sequence $\{X_\ell^k\}_{k=0}^\infty$ is defined as

$$AT_\ell := 1 + 2\sum_{k=1}^\infty \frac{\mathbb{E}\left\{X_\ell^0 X_\ell^k\right\}}{\mathbb{E}\left\{(X_\ell^0)^2\right\}}.$$

It follows then that

$$N \,\mathrm{Var}\left\{\frac{1}{N}\sum_{k=1}^N X_\ell^k\right\} \to \mathrm{Var}\left\{X_\ell^0\right\} AT_\ell.$$

Asymptotically, N/AT_ℓ correlated samples have the same variance-reducing power as the N uncorrelated samples. See the discussion of *effective sample size* in [17]. Note that calculating autocorrelation time is non-trivial, and straightforward methods may have non-vanishing variance in the limit $N \to \infty$. For this reason, we use the *initial positive sequence estimator* described in section 3.3 of [8].

In all cases we used 20 different resolution levels, 0 through 19, with level $j+1$ recycling roughly twice as many random variables as level j. Runtime is almost linearly proportional to the number of recycled random variables. We therefore let the total number of random variables recycled in a simulation run serve as a proxy for *forward-model-time.* Our performance metric is

$$\text{fwd-model-time/1000-effective-samples} := \frac{\text{\#Recycled-random-variables}}{1000N\left(\frac{1}{n}\sum_{\ell=1}^{n} AT_\ell\right)^{-1}}.$$

Simple estimates of the error incurred by using a finite resolution posterior at level j are shown in Table 1. We consider $j = 6$ a very low resolution and $j = 15$ a moderate resolution.

	j=6	j=15	j=19		j=6	j=15	j=19
$\max_x \sqrt{\text{Trace}\left(\Sigma_j(x)\Sigma_E^{-1}\right)}$	0.75	0.05	0.01	$\max_x \sqrt{\text{Cond}(\Sigma_j(x)\Sigma_E^{-1})}$	3.0	0.15	0.05

Table 1: Comparison of finite-resolution error in the posterior. The trace gives an indication of the average error across the m dimensions of data. The condition number gives an indication of the maximal error in any one direction.

We compared various schemes with various parameter choices. One-level results were obtained at level $j = 6$; two-level results were obtained with $i = 6$ and $j = 15$; while MC^3 results were obtained with $i = 6$ and $j_{max} = 19$. See Table 2 and also Figure 2 for a performance comparison. We chose $i = 6$ since this resulted in the lowest *fwd-model-time/1000-effective-samples* for two-level and MC^3 . At this level, a sufficient number of paths interact with the plume to satisfy (27). In other words, we trust the central limit approximation involved in our $\lambda-$acceptance step. Table 1 indicates that we are around the $\sigma_i = 2.0$ line in Figure 1, and therefore $\lambda \approx 0.90$ should give small distortion. Increasing i results would be a safer choice, but this increases forward model time. When moderate resolution is desired $(j, j_{max} = 15)$, MC^3 is about twice as fast at producing uncorrelated random variables as the two-level scheme. When high resolution is desired, the improvement increases to nine times. Figure 3 shows how forward model time may be reduced by decreasing λ.

The dynamics of the MC^3 chain are partially explained by Figure 3. Here we see that lower values of λ result in fewer uses of high-resolution forward models. We also see that the high-resolution forward models, despite being used infrequently, still account for a significant fraction of forward model evaluation time.

3 Path Recycling for Monte Carlo Transport

Here we describe a new *path-recycling* scheme that allows Monte Carlo forward models to be used efficiently in inverse problems. After the transport models of photon propagation are recalled, the path-recycling scheme is presented in the construction of the Monte Carlo estimator $T_n[\gamma]$ in (23) below, which measures the probability of a photon

1-Level	1-Level	1-level	2-Level	2-level	MC^3	MC^3
j=6	j=15	j=19	(i,j)=(6,15)	(i,j)=(6,19)	(i,j$_{max}$,λ)=(6,15,0.90)	(i,j$_{max}$,λ)=(6,19,0.90)
59	10,694	103,584	2,067	10,078	1,035	1,125

Table 2: Comparison of performance. The last row of the table reports *fwd-model-time/1000-effective-samples*. Results for the one-level scheme at $j = 15$ and $j = 19$ are estimated. On one 2.6 GHz Intel core, MC^3 with $(i, j_{max}, \lambda) = (6, 15, 0.90))$ generates one effective sample approximately every 101 seconds. See Figure 3 for additional MC^3 results.

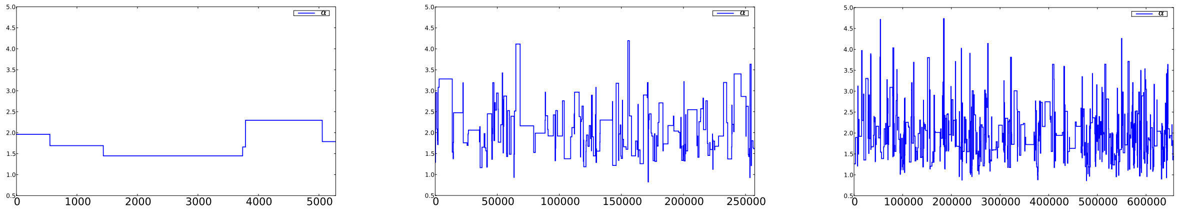

Figure 2: Left: 5hrs, 1-level. Center: 5hrs, 2-level. Right: 5hrs, MC^3 . Parameters are the same as table 2 with $j, j_{max} = 15$.

reaching a given detector in a given environment γ. This estimator is very similar to those used in *perturbation Monte Carlo* [10, 9, 4]. When very accurate information is available in a reference environment, a significantly more efficient (i.e., with much lower variance) unbiased estimator $F_j(\gamma)$ is then introduced in (24) below. We then combine the path recycling scheme for inverse transport calculations with the $(MC)^3$ algorithm and characterize our model errors and intervals of confidence.

3.1 Transport of photons in the atmosphere

We consider photon flux at a single wavelength (all velocities $v \in \mathbb{S}$) in a connected domain $R \subset \mathbb{R}^2$ with boundary ∂R. Denote by $\Gamma_\pm$ the incoming/outgoing boundary flux. That is, with $\hat{n}(r)$ the outgoing normal at $r \in \partial R$, $\Gamma_\pm := \{(r, v) \in \partial R \times \mathbb{S} : \pm\hat{n}(r) \cdot v > 0\}$.

We model photon flux density u by

$$v \cdot \nabla_r u(r, v) + \sigma(r) u(r, v) = Ku(r), \qquad u\big|_{\Gamma_-}(r) = \frac{K(u|_{\Gamma_+})(r)}{|\hat{n}(r) \cdot v|} + \frac{S(r)}{|\hat{n}(r) \cdot v|}, \tag{18}$$

where S is the photon source and

$$\begin{aligned} Kf(r, v) &= \sigma_s(r) \int_{\mathbb{S}^{d-1}} \theta(r, v' \to v) f(r, v') \, \mathrm{d}v', & r \in R \setminus \partial R \\ Kf(r, v) &= \sigma_s(r) \int_{\hat{n}(r) \cdot v' > 0} \Theta(r, v' \to v) |\hat{n}(r) \cdot v'| f(r, v') \, \mathrm{d}v' & r \in \partial R. \end{aligned} \tag{19}$$

The functions θ, Θ, σ_s, and $\sigma = \sigma_s + \sigma_a$ account for scattering, absorption, and scattering+absorption in the domain. See [13].

We also assume the presence of a purely absorbing disk-shaped plume with center (x_p, z_p), radius ρ. Away from this plume the absorption/scattering cross sections σ_a,

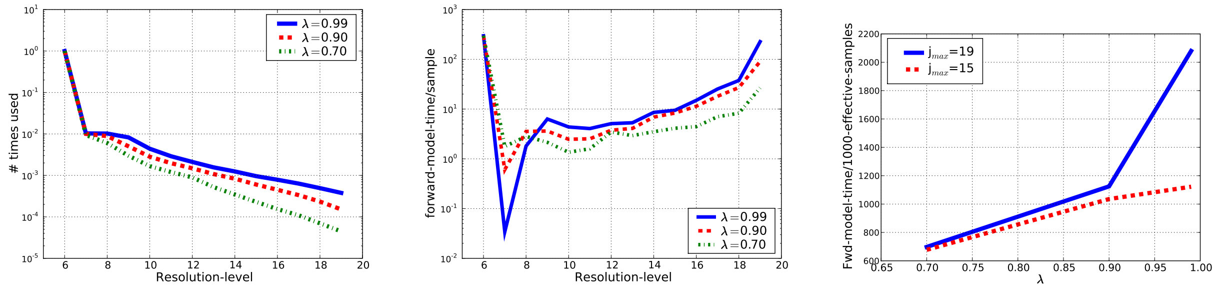

Figure 3: Comparison of MC^3 for different values of λ. In all cases $i = 6$ and $j_{max} = 19$. Left: Raising λ increases the number of uses of higher resolution-level forward models. Center: Although $j = 19$ was used the least, the time spent in it is significant since forward-model-time $\propto 2^j$. Right: Higher λ and/or j_{max} results in more forward-model-time/1000-effective-samples. We see that if less accuracy is desired, one may use lower λ and save time.

σ_s are given by

$$\sigma_a(r) := \sigma_{a,0} e^{-(c_0+c)z}, \qquad \sigma_s(r) = \sigma_{s,0} e^{-(c_0+c)z}. \tag{20}$$

In the support of the plume σ_a is modified by the addition of the constant α.

We group the unknown parameters into the vector $\gamma = (\alpha, x_p, z_p, \rho, c)$. In the name of realism many constraints must be invoked.

$$0 \le \alpha < \infty, \quad \{(x,z) : |(x,z) - (x_p, z_p)| \le \rho\} \subset R, \quad -c_0 \le c < c_{max}.$$

Thus $\gamma \in \Gamma \subset \mathbb{R}^5$ where

$$\Gamma := [0, \infty) \times [x_{min}, x_{max}] \times [z_{min}, z_{max}] \times [\rho_{min}, \rho_{max}] \times [-c_0, c_{max}]. \tag{21}$$

All other parameters/constants are known and the inverse problem involves the recovery of γ. In particular, the detector location and size are fixed.

3.2 Path measures and path recycling

Here we describe the viewpoint that different atmospheric parameters correspond to different measures on the space of possible photon paths. Our forward model uses one fixed set of paths to compute measurements in many different atmospheres. This allows significant time savings since we need not re-simulate paths that miss the detector. Probabilistic error estimates are also obtained.

Collectively, the parameter γ describes a *world* in which we simulate photon travel. In determining unknown parameters we will have to simulate many different worlds. This path measure is induced by the Markov chain corresponding to the chosen atmospheric parameters. We give below a heuristic description of the Markov chain, and refer the reader to [13] for details.

1. A starting position and direction (x_0, v_0) are drawn from the source probability distribution S.

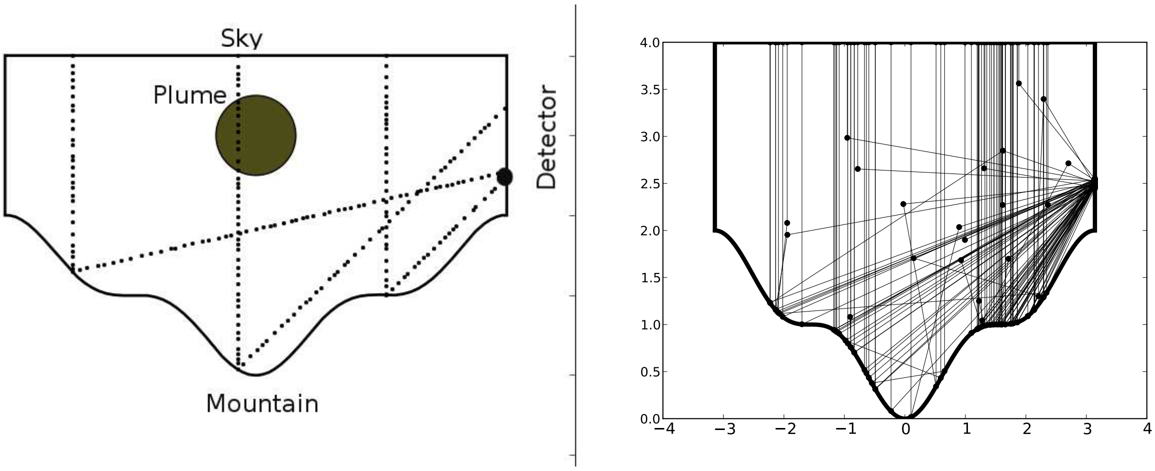

Figure 4: Left: An atmosphere with a sky above and a reflecting mountain surface below. The inverse problem is to characterize the round plume. Right: Traces of 75 paths from sky to the detector. A typical forward model uses 5-300 thousand paths.

2. The photon travels along the path $x_0 + tv_0$, $t > 0$, interacting by point $x_0 + tv_0$ with probability $1 - \exp\left\{-\int_0^t \sigma(x_0 + sv_0)ds\right\}$. If the photon does not interact in the volume it will always interact at the surface.
3. At an interaction, the photon is either absorbed or scattered.
   - At a volume interaction, the photon will be absorbed with probability $\sigma_a(r)/\sigma(r)$. If not, it will scatter into a new direction using the probability density $\theta(r, v_0 \to v_1)$.
   - At a surface interaction, the photon will be absorbed with probability $\sigma_a(r)/\sigma(r)$. If not it will choose a new direction using the density $\Theta(r, v_0 \to v_1)$.
4. This continues until the photon is absorbed. Exit from the domain is accounted for by making $\sigma_a = \infty$, $\sigma_s = 0$ outside of R. We note that $\sigma_a = \infty$, $\sigma_s = 0$ at the detector as well.

This induces a measure on the space of finite-length paths

$$\Omega := \{\omega = (r_0, \ldots, r_\tau) : r_j \in R\}.$$

Note that under reasonable conditions the *stopping time* $\tau < \infty$ and so we have a probability measure P_γ. In the special case $\gamma = \gamma_0 := (0, 0, 2, 0, c_0)$ (corresponding to no plume and nominal background) we have our *reference measure* P.

This allows us to define a differential measure dP (and similarly dP_γ) and expectation $\mathbb{E}_\mathrm{P}\{\cdot\}$ by

$$\mathrm{P}[A] := \mathbb{E}_\mathrm{P}\{\mathbf{1}_A\} = \int_\Omega \mathbf{1}_A(\omega)\,\mathrm{dP}(\omega) = \int_A \mathrm{dP}(\omega), \tag{22}$$

where for $A \subset \Omega$, the *indicator function*

$$\mathbf{1}_A(\omega) := \begin{cases} 1, & \omega \in A \\ 0, & \omega \notin A. \end{cases}$$

As an example of a subset of paths consider those that hit (and are necessarily absorbed in) the detector. Denote these by a disjoint union $\mathcal{D} = D_1 \cup \cdots \cup D_m$, meaning that if $\omega \in \mathcal{D}$ then the path ω ended up in the detector, and if $\omega \in D_\nu$ then ω hit the detector with incoming angle in the interval $\left(\frac{(\nu-1)\pi}{m} - \frac{\pi}{2}, \frac{\nu\pi}{m} - \frac{\pi}{2}\right)$. Let

$$\mathbf{D} : = D_1 \times \cdots \times D_m, \qquad \mathbf{1_D} := (\mathbf{1}_{D_1}, \ldots, \mathbf{1}_{D_m}),$$

and thus our measurement is

$$\mathrm{P}_\gamma[\mathbf{D}] : = \mathbb{E}_\gamma\left\{\mathbf{1_D}\right\} := (\mathrm{P}_\gamma[D_1], \ldots, \mathrm{P}_\gamma[D_m]).$$

One can similarly define $\mathbb{E}_\mathrm{P}\left\{\mathbf{1_D}\right\} = \mathrm{P}[\mathbf{D}]$.

3.2.1 The multi-resolution forward models

We are now in position to describe our forward model and obtain probabilistic error estimates. This model uses importance sampling to compute $\mathrm{P}_\gamma[\mathbf{D}]$ from one fixed set of reference paths. This technique is an advancement upon schemes developed in the context of medical imaging (see e.g., [10, 9, 4]).

Choosing $n \in \mathbb{N}$, we generate n paths $\{\omega^1, \ldots, \omega^n\}$. Now for any random variable X,

$$\frac{1}{n}\sum_{j=1}^{n} \mathbf{1_D}(\omega^j) X(\omega^j) \xrightarrow{a.s.} \mathbb{E}_\mathrm{P}\left\{\mathbf{1_D} X\right\}, \qquad n \to \infty.$$

For example, we could generate paths from measure P_γ, and then $n^{-1}\sum_{j=1}^{n} \mathbf{1_D}(\omega^j) \to \mathbb{E}_\gamma\left\{\mathbf{1_D}\right\}$.

It is very important to realize that since we only intend on estimating expectations involving detector hits (e.g., $\mathbb{E}_\gamma\left\{\mathbf{1_D}X\right\}$), we only need store paths that hit the detector. The expected number of detector hits is exactly $n\mathrm{P}[\mathcal{D}] \ll n$.

For every new γ we could generate a new set of paths and repeat the above procedure. This would be costly since path generation involves complicated steps. Instead, consider fixing one set of reference paths $\{\omega^j\}_{j=1}^n$ (in practice storing only those that hit the detector) generated by the reference measure P and then set

$$T_n(\gamma) : = \frac{1}{n}\sum_{k=1}^{n} \mathbf{1_D}(\omega^k) \left|\frac{\mathrm{dP}_\gamma}{\mathrm{dP}}\right|(\omega^k) \approx \int_\Omega \mathbf{1_D} \left|\frac{\mathrm{dP}_\gamma}{\mathrm{dP}}\right| \mathrm{dP} = \int_\Omega \mathbf{1_D}\, \mathrm{dP}_\gamma = \mathbb{E}_\gamma\left\{\mathbf{1_D}\right\}. \quad (23)$$

Computation of T_n requires computing the Radon-Nikodym derivative for the $\approx n\mathrm{P}[\mathcal{D}]$ paths that hit the detector. This is significantly faster than sending n new paths. For details as to this calculation, see [13]. For the present work, it will suffice to assume the following

Assumptions 3.1. Assume that for every $\gamma \in \Gamma$, the Radon-Nikodym derivative $\left|\frac{\mathrm{dP}_\gamma}{\mathrm{dP}}\right|$ exists.

Although fast, T_n can be significantly improved by using (for relatively small n) information from a simulation that used very large n. This is where we depart from the previously mentioned perturbation Monte Carlo schemes.

We first generate n_{max} paths using the reference measure P. Denote by H^ν_{max} the collection of paths $\omega^k \in D_\nu$. That is,

$$H^\nu_{max} := \left\{\omega^1, \ldots, \omega^{n_{max}}\right\} \cap D_\nu.$$

For $\nu = 1, \ldots, m$, let

$$H^\nu_1 \subset H^\nu_2 \subset \cdots \subset H^\nu_{max},$$

be nested subsets of H^ν_{max} of (fixed, deterministic) size $|H^\nu_j|$. Note that H^ν_j, H^ν_{max} consist of i.i.d. draws from $\mathrm{P}[\cdot \,|\, D_\nu]$. Since $|H^\nu_{max}| = n_{max}^{-1} \sum_{k=1}^{n_{max}} \mathbf{1}_{D_\nu}(\omega^k)$, we have

$$\mathrm{Cov}_{\mathrm{P}}\left(|H^\nu_{max}|, |H^\mu_{max}|\right) = \frac{1}{n_{max}} \begin{cases} \mathrm{P}[D_\nu] - \mathrm{P}[D_\nu]^2, & \nu = \mu \\ -\mathrm{P}[D_\nu]\mathrm{P}[D_\mu], & \nu \neq \mu. \end{cases}$$

Above $\mathrm{Cov}_{\mathrm{P}}(X, Y)$ is defined as in (11) and the subscript P makes it clear that expectations are with respect to the measure P. Although $\{|H^\nu_{max}|\}_{\nu=1}^m$ are negatively correlated, so long as $|H^\nu_j|$ may be selected independently of H^ν_{max}, the sets H^ν_j are independent. We will always ensure this condition holds.

Our improvement on T_j is $F_j = (F^1_j, \ldots, F^m_j)$ where

$$F^\nu_j(\gamma) := \frac{|H^\nu_{max}|}{n_{max}} \frac{1}{|H^\nu_j|} \sum_{\omega^k \in H^\nu_j} \left|\frac{d\mathrm{P}_\gamma}{d\mathrm{P}}\right|(\omega^k) \tag{24}$$

Notice that if $\mathrm{P} = \mathrm{P}_\gamma$, then F^ν_j sums $|H^\nu_j|$ i.i.d. draws from $\mathrm{P}[\cdot \,|\, D_\nu]$, and each of them scores a hit $|H^\nu_{max}|/n_{max}$. In other words, up to the approximations $\mathrm{P}_\gamma \approx \mathrm{P}$, and $|H^\nu_{max}|/n_{max} \approx \mathrm{P}[D_\nu]$, $F^\nu_j(\gamma)$ sums $|H^\nu_j|$ random variables, each one recording the exact solution. Hence (up to these approximations) $F^\nu_j(\gamma)$ computes $\mathrm{P}[D_\nu]$ with zero variance.

On the practical side, for $j < j'$, $H^\nu_j \subset H^\nu_{j'}$, and therefore computation of $F_{j'}$ is quicker after computation of F_j is done.

We collect this and other facts in a lemma

Lemma 3.1.

$$Cov_{\mathrm{P}}\left(|H^\nu_{max}|, |H^\mu_{max}|\right) = \frac{1}{n_{max}} \begin{cases} \mathrm{P}[D_\nu] - \mathrm{P}[D_\nu]^2, & \nu = \mu \\ -\mathrm{P}[D_\nu]\mathrm{P}[D_\mu], & \nu \neq \mu. \end{cases}$$

The sets $\{H^\nu_j\}$, are independent, and for all j, the set H^ν_j consists of $|H^\nu_j|$ i.i.d. draws from $d\mathrm{P}[\cdot \,|\, D_\nu] = \mathrm{P}[D_\nu]^{-1}\mathbf{1}_{D_\nu}\, d\mathrm{P}$. Conditioning on $|H^\nu_{max}|$ does not change this:

$$d\mathrm{P}\left[\omega \,|\, D_\nu, |H^\nu_{max}|\right] = d\mathrm{P}\left[\omega \,|\, D_\nu\right] = \frac{\mathbf{1}_{D_\nu}(\omega)\, d\mathrm{P}(\omega)}{\mathrm{P}[D_\nu]}.$$

The next theorem shows that F_j is unbiased. See the appendix A for the proof.

Theorem 3.1.

$$\mathbb{E}_{\mathrm{P}}\left\{F_j(\gamma)\right\} = \mathbb{E}_\gamma\left\{\mathbf{1}_{\mathbf{D}}\right\} = \mathrm{P}_\gamma[\mathbf{D}].$$

Remark 3.1. Since $|H_j^\nu|$ is deterministic, it follows that $F_j^\nu(\gamma)$ is a sum of $|H_j^\nu|$ unbiased random variables and so by the strong law of large numbers, for fixed x, $F_j(x) \to F_\infty(x)$ *a.s.*. It is straightforward then to show that with probability one, $F_j(x) \to F_\infty(x)$ for a.e. x. Equation (7) then follows.

The following theorem shows that in the limit $\mathrm{dP}_\gamma \to \mathrm{dP}$, and $|H_{max}^\nu| \to \infty$, the $F_j^\nu(\gamma)$ are uncorrelated zero-variance estimates of $\mathrm{P}_\gamma[D_\nu]$. See appendix A for a proof.

Theorem 3.2. *As* $n_{max} \to \infty$,

$$Var_{\mathrm{P}}\left\{F_j(\gamma)\right\}_{\nu\mu} \to \delta_{\mu\nu}\frac{\mathrm{P}[D_\nu]}{|H_j^\nu|}\int_{D_\nu}\left(\left|\frac{\mathrm{dP}_\gamma}{\mathrm{dP}}\right|(\omega) - \frac{\mathrm{P}_\gamma[D_\nu]}{\mathrm{P}[D_\nu]}\right)^2 \mathrm{dP}(\omega).$$

Remark 3.2. A similar calculation shows that

$$\mathrm{Var}_{\mathrm{P}}\left\{T_{n_j}(\gamma)\right\}_{\nu\mu} = \frac{1}{n_j}\begin{cases}\int_{D_\nu}\left(\left|\frac{\mathrm{dP}_\gamma}{\mathrm{dP}}\right| - \mathrm{P}_\gamma[D_\nu]\right)\mathrm{dP}_\gamma, & \nu = \mu\\ -\mathrm{P}_\gamma[D_\nu]\mathrm{P}_\gamma[D_\mu], & \nu \neq \mu.\end{cases}$$

One can replace (in the expression for $\mathrm{Var}_{\mathrm{P}}\left\{F_j\right\}$) $|H_j^\nu|$ with the approximation $n_j\mathrm{P}[D^\nu]$ and see that $\mathrm{Var}_{\mathrm{P}}\left\{F_j(\gamma)\right\} \ll \mathrm{Var}_{\mathrm{P}}\left\{T_{n_j}(\gamma)\right\}$ if $\mathrm{dP} \approx \mathrm{dP}_\gamma$. In other words, the variance of our unbiased estimator F_j is significantly smaller than the estimator T_{n_j} typically used in aforementioned perturbation Monte Carlo schemes.

3.3 Path recycling and MC^3

Here we show that the multiple resolution forward models F_j (24) meet most of the assumptions from section 2. As was established in remark 3.1, lemmas 3.1 and 3.2 establish that $F_j \to F_\infty$ in the sense of (7). The Gaussian error model $F_j(\gamma) \sim \mathcal{N}(F_\infty(\gamma)\,,\,\Sigma_j(\gamma))$ will be established in sections 3.3.1. The covariance estimate used in section 2.3 is then established in section 3.3.2.

Note that since we are recycling paths, different runs of these models are not independent. This is the only way the F_j do not meet the criteria of section 2. This fact turns out to be an advantage for practical reasons: Since we expect $F_j(x)$ and $F_j(y)$ to be positively correlated for x near y, we see that $\hat{\mu}_j^2$ (an estimate of the variance in our acceptance-ratio estimate, see (26)) will be significantly lower. Thus, we are able to use fewer paths to determine acceptance/rejection. If we were using an optimization-based inversion method, a similar phenomenon would occur. The general theme is that if X and Y are random variables (say some functions of $F_j(x)$, $F_j(y)$), then $\mathrm{Var}\left\{X - Y\right\} = \mathrm{Var}\left\{X\right\} + \mathrm{Var}\left\{Y\right\} - 2\mathrm{Cov}\left(X, Y\right)$. In other words, error in the differential measurements is reduced. This is absolutely essential since if x is close to y, we expect $|F_\infty(x) - F_\infty(y)| \ll \mathrm{Var}\left\{F_j(x)\right\}$ unless j is very large.

3.3.1 Characterization of the model error

We use lemma 3.2 and the central limit theorem to characterize the error. We estimate this using standard techniques. The result is that we approximate $F_j(\gamma)$ as following a $\mathcal{N}(\mathrm{P}_\gamma[\mathbf{D}]\,,\,\Sigma_j(\gamma))$ distribution, where

$$\Sigma_j(\gamma) := \operatorname{diag}\left\{\sigma_j^1(\gamma),\ldots,\sigma_j^m(\gamma)\right\},$$
$$\sigma_j^\nu(\gamma) := \left(\frac{|H_{max}^\nu|}{n_{max}}\right)^2 \frac{1}{|H_j^\nu|}\frac{1}{|H_j^\nu|-1}\sum_{\omega^k\in H_j^\nu}\left(\left|\frac{\mathrm{dP}_\gamma}{\mathrm{dP}}\right|(\omega^k) - \frac{F_j^\nu(\gamma)}{|H_{max}^\nu|/n_{max}}\right)^2. \tag{25}$$

The final sum is the standard estimate of the variance of the random variable $\left|\frac{\mathrm{dP}_\gamma}{\mathrm{dP}}\right|(\omega^k)$.

3.3.2 The estimate of μ_j

For our confidence interval in section 2.3 we need to estimate

$$\mu_j^2 := \nabla\varphi(\bar{X},\bar{Y})^T \operatorname{Cov}(X,Y)\,\nabla\varphi(\bar{X},\bar{Y}),$$

where

$$\varphi(X,Y) := |X|^2 - |Y|^2, \qquad \text{and for } \nu = 1,\ldots,m$$
$$X^\nu := \mathbf{d}^\nu - F_j^\nu(\gamma) = \mathbf{d}^\nu - \frac{|H_{max}^\nu|}{n_{max}}\frac{1}{|H_j^\nu|}\sum_{\omega^k\in H_j^\nu}\left|\frac{\mathrm{dP}_\gamma}{\mathrm{dP}}\right|(\omega^k),$$

and Y^ν is defined similarly. We also have $\bar{X} := \mathbb{E}_\mathrm{P}\{X\}$, $\bar{Y} := \mathbb{E}_\mathrm{P}\{Y\}$. This is easy since the variables $\left|\frac{\mathrm{dP}_\gamma}{\mathrm{dP}}\right|(\omega^k)$ are i.i.d. and due to Theorem 3.2 we have an asymptotic variance estimate. As a result (see appendix B), a consistent estimator is

$$\hat{\mu}_j^2 := 4\sum_{\nu=1}^m\left[(X^\nu)^2\operatorname{Var}_\mathrm{P}\{X^\nu\} - 2X^\nu Y^\nu\operatorname{Cov}_\mathrm{P}(X^\nu,Y^\nu) + (Y^\nu)^2\operatorname{Var}_\mathrm{P}\{Y^\nu\}\right]. \tag{26}$$

The question arises, "how large should $|H_j^\nu|$ be before we believe the CLT approximation?" Using a theorem on the remainder in a CLT approximation we arrive at the following scaling criteria (see appendix B)

$$\sqrt{\#\left\{\omega^k\in H_j^\nu : \omega^k \text{ hits the plume}\right\}} \gg 1. \tag{27}$$

In the process of computing the weights $\left|\frac{\mathrm{dP}_\gamma}{\mathrm{dP}}\right|$ we can easily keep track of whether or not the plume was hit. We thus choose our lowest resolution level high enough so that (27) is met.

3.4 Numerical verification/performance of forward model

The accuracy of our forward model (24) was verified. In particular, we conducted a variety of tests where

(i) A Monte Carlo simulation of transport in an atmosphere parameterized by γ was run until the relative mean square error was less than 1/3%. The mean is stored as $M(\gamma)$.

(ii) Another Monte Carlo simulation was run in an atmosphere parameterized by $\gamma_0 = (0,0,1,0,0)$ (no plume and baseline atmosphere). The paths are stored.

(iii) The paths from (ii) are used in the forward model (24) to compute $F_j(\gamma)$.

The forward model "passes" if

- $|F_j(\gamma) - M(\gamma)|/M(\gamma) \leq 0.01$ when j is at the highest resolution level
- As $y \neq \gamma$ becomes sufficiently different than γ, the relative error $|F_j(y) - M(\gamma)|/M(\gamma)$ becomes much worse than 1%

The forward model was seen to pass for a variety of γ.

The performance increase is dramatic. For example, it took 11,727 minutes to generate approximately 231 million paths (of which 1/348 hit the detector). These paths can be recycled in only 30.9 seconds (22,770 times quicker). One might have expected a speed-up of only about 348 times. The dramatic difference is due mostly to the fact that the original paths were cast using complicated Python code that explicitly stepped the photons through their path, while the much simpler recycling could be done using optimized Cython code. In any case, recycling paths only involves computing a ratio of weights, and in many cases may be much quicker than sending the original paths.

4 Use in an Inverse Problem

Here we combine the forward models from section 3 in a Bayesian inverse problem fitting the framework of section 2.

The scene is a sunlit valley with reflecting mountain, a detector located on the right side, and an absorbing plume to be reconstructed. See Figure 4. The plume is parameterized by $\gamma = (\alpha, x_p, z_p, \rho, c)$ where α is the plume absorption, (x_p, z_p) is the plume center, ρ is the plume radius, and c is the atmospheric constant. We assume $\gamma \in \Gamma$ (defined in (21)). Our *prior* density $f_{prior}(\gamma)$ is

$$f_{prior}(\gamma) \propto \left[\mathbf{1}_{[0,\infty)}(\alpha)\alpha^{1/2}e^{-\alpha}\right] \left[\mathbf{1}_{[-0.25,0.25]}(x_p)\right] \left[\mathbf{1}_{[2.2,2.7]}(z_p)\right] \left[\mathbf{1}_{[0,1]}(\rho)\right] \left[\mathbf{1}_{(-c_0,c_0)}(c)\pi\right]. \tag{28}$$

In other words, we assume all components are independent with α being Gamma distributed, and all the others being uniform.

We ran a Monte Carlo forward simulation in an atmosphere containing a plume to generate the "noiseless" data. A separate set of paths was then generated for use in the inverse problem. We added Gaussian noise corresponding to an SNR of 10. We then sampled from the posterior using the MC^3 algorithm using initial resolution level $i = 6$, $j_{max} = 15$, and confidence $\lambda = 0.90$. This is compared with two-level scheme at $i = 6$, $j = 15$ and one-level scheme at $j = 6$. One can see that MC^3 generates virtually the same posterior as the more expensive two-level scheme. The difference could be entirely attributed to the finite number of samples used to construct the histogram. The one-level scheme at $j = 6$ has too low resolution: the likelihood is too broad and hence the posterior π_6 closely follows the prior. We also ran MC^3 simulations with $\lambda = 0.70, 0.99$ and observed results similar to those obtained with $\lambda = 0.90$.

From the practical standpoint, we find that the horizontal position x_p and the plume absorption and plume radius are reasonably reconstructed, in the sense that the posterior marginals are much tighter than the prior marginals for these parameters. The reconstruction of the vertical position z of the plume remains very inaccurate.

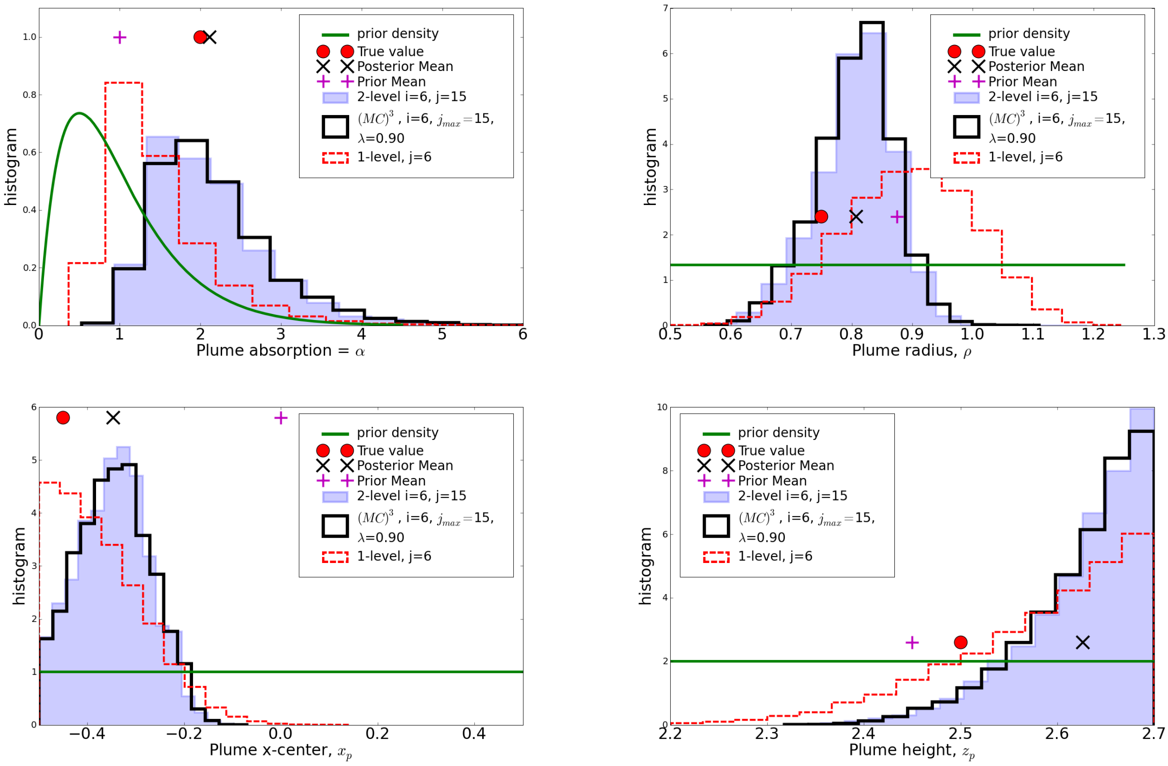

Figure 5: Two marginal posteriors. Plotted is a histogram from an MC^3 run with $(i, j_{max}, \lambda) = (6, 15, 0.99)$. (The two-level and other MC^3 results were almost identical to this result.) Also shown are the prior, the posterior from a one-level $j = 6$ run, and the posterior mean (from the MC^3 run) as well as the prior mean and true value.

This is to be expected for particles launched from the sky vertically and propagating through an atmosphere with limited scattering: very few photons providing information about the plume height reach the detector. In other words, the available data are not sufficiently informative to accurately reconstruct this parameter. The Bayesian framework allows us to quantify such statements and the MC^3 algorithm allows us to do so at a more reasonable computational cost than other methods.

5 Conclusions and Outlook

A coherent framework was presented for solving Bayesian inverse problems when the forward model involves Monte Carlo. Path-recycling yields efficient multi-resolution forward models for use in the Bayesian framework, though the utility of these models is not restricted to Bayesian inference. In this context, however, MC^3 takes advantage of multi-resolution forward models to quickly draw from the posterior distribution. Its application is not limited to the forward models considered here; MC^3 may be used separately whenever multi-resolution forward models are available with tight probabilistic error estimates.

MC^3 was shown to be robust with respect to the choice of initial resolution level i

and confidence λ: the discrepancy for finite i and $\lambda < 1$ was quantified in a simple toy model, and shown to be small in our transport problem (which used path-recycling). This does not provide strict parameter choice rules for (i, λ) as would be desirable. However, a reasonable procedure was explained and verified in sections 2.5 and 4. In general, the *forward-model-time/1000-effective-samples* for MC^3 at levels i, j_{max} is superior to that of a two-level MCMC scheme at $i, j = j_{max}$ (we experienced two to nine times speedup), but slower than a one-level scheme at resolution i. The latter, however, may be extremely inaccurate. The efficiency gain provided by MC^3 rises significantly when higher resolution is desired.

We see path recycling as a necessary element in solving any inverse problem with a Monte Carlo forward model. We see MC^3 as a great time saver in Bayesian applications. Future work should provide further intuition and ideally some concrete results about parameter choice. We also mention that the MC^3 scheme could be adapted to forward models that are not Monte Carlo, yet that are endowed with tight error estimates.

A Proofs

Proof of Theorem 2.1. The proof of Theorem 2.1 follows standard techniques. We include it since the chain in algorithm 4 is not a Metropolis-Hastings scheme and so some care must be taken.

If $\lambda = 1$ then the result is trivial so assume in the sequel that $\lambda \in (0, 1)$. We first show that the density π_λ is well-defined. Since R_λ is a convex combination of $r_{i\infty}(x)$ and $1 - r_{i\infty}(x)$, we have $0 < R_\lambda < 1$. Therefore, $\lambda \leq m_\lambda$ and thus $0 < \pi_\infty / m_\lambda \leq \pi_\infty / \lambda$ is integrable and we can normalize it to define the density π_λ.

Second, we show

$$K_\lambda(x, y)\pi_\lambda(x) = K_\lambda(y, x)\pi_\lambda(y). \tag{29}$$

In other words, the chain in algorithm 4 satisfies the detailed balance equation with distribution π_λ. To that end we re-write

$$K_\lambda(x, y) = m_\lambda(x) q_i(y \,|\, x) \beta_{i\infty}(x, y) + (1 - R_\lambda(x))\delta_x(y). \tag{30}$$

Also note that

$$\beta_{i\infty}(x, y) := \min \left\{ 1, \frac{q_i(x \,|\, y)\pi_\infty(y)}{q_i(y \,|\, x)\pi_\infty(x)} \right\} = \min \left\{ 1, \frac{q_i(x \,|\, y) m_\lambda(y) \pi_\lambda(y)}{q_i(y \,|\, x) m_\lambda(x) \pi_\lambda(x)} \right\}.$$

(29) then follows. Thus π_λ is an invariant distribution of K_λ.

To show the convergence results we assume (i) and (ii) and proceed to show irreducibility, a-periodicity, and Harris recurrence.

Let $x_0 \in E_+$ and $K^n_\lambda(x_0, \cdot)$ be the transition kernel associated with n steps of the chain starting at x_0. Let $B \subset E_+$ with $\int_B \pi_\infty(y)\,\mathrm{d}y > 0$. Irreducibility will follow once we show $K^n_\lambda(x_0, B) > 0$ for some n. To that end first note that

$$K_\lambda(x, y) \geq \lambda q(y \,|\, x) \alpha_i(x, y) \beta_{i\infty}(x, y). \tag{31}$$

Second, some algebra shows that

$$\alpha_i(x,y)\beta_{ij}(x,y) = \min\left\{1, \frac{\pi_i(x)\pi_j(y)}{\pi_i(y)\pi_j(x)}, \frac{q(x\,|\,y)\pi_i(y)}{q(y\,|\,x)\pi_i(x)}, \frac{q(x\,|\,y)\pi_j(y)}{q(y\,|\,x)\pi_j(x)}\right\}. \tag{32}$$

Third, find a bounded open set G with $x_0 \in G$ and $\int_{G\cap B} \pi_\infty(y)\,\mathrm{d}y > 0$. We then have δ, C_i, D_i, C_q such that for all $x, y \in G$ with $0 < |x-y| < \delta$, the inequalities of (i) and (ii) are satisfied. Then, using (31), (32) we have (for $x, y \in G$, $|x-y| < \delta$) $K_\lambda(x,y) \geq \lambda\pi_\infty(y)C_q \min\left\{1, \frac{C_i}{D_i}, C_i\right\}$.

We finally link x_0 to B with n open balls $B_k \subset G$ of diameter less than δ, $B_i \cap B_{i+1} \neq \emptyset$, $B_n \cap B \neq \emptyset$. It follows that (for a new constant $C > 0$)

$$K_\lambda^n(x_0, B) \geq C \int_{B\cap B_n} \cdots \int_{B_2\cap B_1} \pi_\infty(y_n)\cdots\pi_\infty(y_1)\,\mathrm{d}y_1\cdots\mathrm{d}y_n > 0.$$

Hence the chain is π_∞ irreducible.

A-periodicity follows since by choosing a small enough ball $B_{x_0} \ni x_0$, one can show that $K_\lambda^n(x_0, B_{x_0}) > 0$ for all n.

We now show Harris recurrence. The proof here follows more-or-less lemma 7.3 in [17] or corollary 2 in [18]. and relies on showing that the only bounded harmonic functions are constant. Recall that a function h is harmonic with respect to our chain if $\mathbb{E}\left\{h(X^1)\,|\,X^0 = x_0\right\} = h(x_0)$. Since π_λ is an invariant probability measure of the chain, the chain is recurrent by proposition 6.36 of [17]. As in lemma 7.3 from [17] this implies that h is π_λ almost everywhere equal to a constant $\bar{h}$. To show that h is constant everywhere (on the support of π_λ) we write

$$\begin{aligned} h(x_0) = \mathbb{E}\left\{h(X^1)\,|\,x_0\right\} &= \int K_\lambda(x_0,x_1)h(x_1)\,\mathrm{d}x_1 \\ &= \int m_\lambda(x_0)q_i(x_1\,|\,x_0)\beta_{i\infty}(x_0,x_1)h(x_1)\,\mathrm{d}x_1 + (1 - R_\lambda(x_0))\,h(x_0). \end{aligned}$$

The above integral is unchanged if we replace $h(x_1)$ by $\bar{h}$. Then since $\int m_\lambda q_i \beta_{i\infty}\,\mathrm{d}x_1 = R_\lambda(x_0)$ we have $R_\lambda(x_0)\left(\bar{h} - h(x_0)\right) = 0$. The irreducibility portion of this proof shows that $R_\lambda > 0$, hence $h(x_0) = \bar{h}$ and h is constant.

Having shown the chain is Harris recurrent and a-periodic with invariant measure π_∞, the convergence results follow directly from Theorems 6.63, 6.51, and proposition 6.52 in [17]. □

Proof of Theorem 3.1. Using lemma 3.1, we have that $\mathbb{E}_{\mathrm{P}}\left\{F_j^\nu(\gamma)\right\}$ is equal to

$$\mathbb{E}_{\mathrm{P}}\left\{\frac{|H_{max}^\nu|}{n_{max}}\frac{1}{|H_j^\nu|}\mathbb{E}_{\mathrm{P}}\left\{\sum_{\omega^k\in H_j^\nu}\left|\frac{\mathrm{dP}_\gamma}{\mathrm{dP}}\right|(\omega^k)\;|\;|H_{max}^\nu|\right\}\right\} = \mathbb{E}_{\mathrm{P}}\left\{\frac{|H_{max}^\nu|}{n_{max}}\frac{1}{|H_j^\nu|}|H_j^\nu|\mathbb{E}_{\mathrm{P}}\left\{\left|\frac{\mathrm{dP}_\gamma}{\mathrm{dP}}\right|\,|\,D_\nu\right\}\right\}$$
$$= \mathbb{E}_{\mathrm{P}}\left\{\frac{|H_{max}|}{n_{max}}\frac{\mathrm{P}_\gamma[D_\nu]}{\mathrm{P}[D_\nu]}\right\} = \mathrm{P}_\gamma[D_\nu].$$

The last equality follows since $\mathbb{E}_{\mathrm{P}}\left\{|H_{max}|\right\}$ is the expected number of hits in D_ν, equal to $n_{max}\mathrm{P}[D_\nu]$. □

Proof of Theorem 3.2. Writing $|H^\nu_{max}|/n_{max} = (|H^\nu_{max}|/n_{max} - \mathrm{P}[D_\nu]) + \mathrm{P}[D_\nu]$, and similarly for μ, we can express $\mathrm{Cov}_\mathrm{P}\left(F^\nu_j, F^\nu_j\right)$ as

$$\mathrm{Var}_\mathrm{P}\left\{\frac{\mathrm{P}[D_\nu]}{|H^\nu_j|}\sum_{\omega^k\in H^\nu_j}\left|\frac{\mathrm{dP}_\gamma}{\mathrm{dP}}\right|(\omega^k)\right\}$$

plus terms that tend to zero as $n_{max}\to\infty$ (due to lemma 3.1). We now compute the variance of this term, keeping in mind that the $\omega^k \in H^\nu_j$ are i.i.d. draws from $\mathrm{P}[\cdot\,|\,D_\nu]$.

$$\begin{aligned}\mathrm{Var}_\mathrm{P}\left\{\frac{\mathrm{P}[D_\nu]}{|H^\nu_j|}\sum_{\omega^k\in H^\nu_j}\left|\frac{\mathrm{dP}_\gamma}{\mathrm{dP}}\right|(\omega^k)\right\} &= \frac{\mathrm{P}[D_\nu]^2}{|H^\nu_j|}\int_\Omega\left(\left|\frac{\mathrm{dP}_\gamma}{\mathrm{dP}}\right|(\omega) - \frac{\mathrm{P}_\gamma[D_\nu]}{\mathrm{P}[D_\nu]}\right)^2 \mathrm{dP}[\omega\,|\,D_\nu]\\ &= \frac{\mathrm{P}[D_\nu]}{|H^\nu_j|}\int_{D_\nu}\left(\left|\frac{\mathrm{dP}_\gamma}{\mathrm{dP}}\right|(\omega) - \frac{\mathrm{P}_\gamma[D_\nu]}{\mathrm{P}[D_\nu]}\right)^2 \mathrm{dP}(\omega).\end{aligned} \tag{33}$$

This is the statement of the theorem. □

B Estimate of μ_j and the scaling criteria

Note that

$$\mathrm{Cov}_\mathrm{P}(X,Y) := \begin{pmatrix}\mathrm{Var}_\mathrm{P}\{X\} & \mathrm{Cov}_\mathrm{P}(X,Y)\\ \mathrm{Cov}_\mathrm{P}(Y,X) & \mathrm{Var}_\mathrm{P}\{Y\}\end{pmatrix},$$

where due to the approximate independence of the components (true as $n_{max}\to\infty$)

$$\begin{aligned}\mathrm{Var}_\mathrm{P}\{X\} &\approx \mathrm{diag}\left(\mathrm{Var}_\mathrm{P}\{X^\nu\}\right), \qquad \mathrm{Var}_\mathrm{P}\{Y\} = \mathrm{diag}\left(\mathrm{Var}_\mathrm{P}\{Y^\nu\}\right)\\ \mathrm{Cov}_\mathrm{P}(X,Y) &\approx \mathrm{Cov}_\mathrm{P}(Y,X) = \mathrm{diag}\left(\mathrm{Cov}_\mathrm{P}(X^\nu,Y^\nu)\right),\end{aligned}$$

Since X, Y are sums are i.i.d. random variables, we can estimate the covariance using Theorem 3.2 (this is similar to (25)).

$$\begin{aligned}\mathrm{Var}_\mathrm{P}\{X^\nu\} &\approx \left(\frac{|H^\nu_{max}}{n_{max}}\right)^2\frac{1}{|H^\nu_j|}\frac{1}{|H^\nu_j|-1}\sum_{\omega^k\in H^\nu_j}\left(\left|\frac{\mathrm{dP}_\gamma}{\mathrm{dP}}\right|(\omega^k) - \frac{F^\nu_j(\gamma)}{|H^\nu_{max}|/n_{max}}\right)^2\\ \mathrm{Cov}_\mathrm{P}(X^\nu,Y^\nu) &\approx \left(\frac{|H^\nu_{max}}{n_{max}}\right)^2\frac{1}{|H^\nu_j|}\frac{1}{|H^\nu_j|-1}\\ &\quad\times\sum_{\omega^k\in H^\nu_j}\left(\left|\frac{\mathrm{dP}_\gamma}{\mathrm{dP}}\right|(\omega^k) - \frac{F^\nu_j(\gamma)}{|H^\nu_{max}|/n_{max}}\right)\left(\left|\frac{\mathrm{dP}_\mathrm{y}}{\mathrm{dP}}\right|(\omega^k) - \frac{F^\nu_j(y)}{|H^\nu_{max}|/n_{max}}\right).\end{aligned} \tag{34}$$

To compute $\hat{\mu}^2_j \approx \mu^2_j$, we note that $\nabla\varphi(X,Y) = 2\left(X^1,\ldots,X^m,-Y_1,\ldots,-Y^m\right)$,

$$\mathrm{Var}\{(X,Y)\} = \begin{pmatrix}\mathrm{diag}\left(\mathrm{Var}_\mathrm{P}\{X^\nu\}\right) & \mathrm{diag}\left(\mathrm{Cov}_\mathrm{P}(X^\nu,Y^\nu)\right)\\ \mathrm{diag}\left(\mathrm{Cov}_\mathrm{P}(X^\nu,Y^\nu)\right) & \mathrm{diag}\left(\mathrm{Cov}_\mathrm{P}(X^\nu,Y^\nu)\right)\end{pmatrix}$$

So we have $\nabla\varphi(\bar{X},\bar{Y})^T\mathrm{Cov}\,(X,Y)\,\nabla\varphi(\bar{X},\bar{Y})$

$$\approx 4\sum_{\nu=1}^{m}\Big[(X^\nu)^2\,\mathrm{Var}_\mathrm{P}\,\{X^\nu\} - 2X^\nu Y^\nu \mathrm{Cov}_\mathrm{P}\,(X^\nu,Y^\nu) + (Y^\nu)^2\,\mathrm{Var}_\mathrm{P}\,\{Y^\nu\}\Big] := \hat{\mu}_j^2.$$

Our estimate (34) satisfies $|\mathrm{Cov}_\mathrm{P}\,(X^\nu,Y^\nu)| \le (\mathrm{Var}_\mathrm{P}\,\{X^\nu\}\,\mathrm{Var}_\mathrm{P}\,\{Y^\nu\})^{1/2}$ (with equality if and only if $\left|\frac{\mathrm{dP}_\gamma}{\mathrm{dP}}\right|(\omega^k) - F_j(\gamma) = \left|\frac{\mathrm{dP}_\mathrm{y}}{\mathrm{dP}}\right|(\omega^k) - F_j(y)$ for all $\omega^k \in H_j^\nu$). This implies

$$\begin{aligned}&(X^\nu)^2\,\mathrm{Var}_\mathrm{P}\,\{X^\nu\} - 2X^\nu Y^\nu \mathrm{Cov}_\mathrm{P}\,(X^\nu,Y^\nu) + (Y^\nu)^2\,\mathrm{Var}_\mathrm{P}\,\{Y^\nu\}\\ &\ge \left(X^\nu\sqrt{\mathrm{Var}_\mathrm{P}\,\{X^\nu\}} - Y^\nu\sqrt{\mathrm{Var}_\mathrm{P}\,\{Y^\nu\}}\right)^2 \ge 0,\end{aligned}$$

so our estimate $\hat{\mu}_j^2$ will be non-negative and will be zero if and only if

$$\left|\frac{\mathrm{dP}_\gamma}{\mathrm{dP}}\right|(\omega^k) - F_j(\gamma) = \left|\frac{\mathrm{dP}_\mathrm{y}}{\mathrm{dP}}\right|(\omega^k) - F_j(y), \qquad \omega^k \in H_j^\nu. \tag{35}$$

We now derive (27). Consider the following heuristic argument: Start by writing $X^\nu = \bar{X}^\nu + \delta X^\nu$, $Y^\nu = \bar{Y}^\nu + \delta Y^\nu$. Next, provided ρ is small enough, most shots do not interact with the plume and so to first approximation $X^\nu \approx Y^\nu$. As a model for this, we write, $Y^\nu \approx X^\nu - aZ^\nu$ where $a \ll 1$ is the effect of plume interaction and $Z^\nu = |H_j^\nu|^{-1}\sum_{k=1}^{|H_j^\nu|} Z_k^\nu$ with $\mathrm{P}[Z_k^\nu = 1] = p^\nu$, and $\mathrm{P}[Z_k^\nu = 0] = 1 - p^\nu$. Now

$$\varphi(X,Y) = |\bar{X}+\delta X|^2 - |\bar{Y}+\delta Y|^2 = \sum_{\nu=1}^{m}|\bar{X}^\nu + \delta X^\nu|^2 - |\bar{Y}^\nu + \delta Y^\nu|^2,$$

and ignoring terms cubic in the small variables a, δ we have

$$\begin{aligned}|\bar{X}^\nu + \delta X^\nu|^2 - |\bar{Y}^\nu + \delta Y^\nu|^2 &= |\bar{X}^\nu + \delta X^\nu|^2 - |\bar{X}^\nu - a\bar{Z}^\nu + \delta X^\nu - a\delta Z^\nu|^2\\ &\approx \left[2a\bar{X}^\nu\bar{Z}^\nu - (a\bar{Z}^\nu)^2\right] + 2a\bar{X}^\nu\delta Z^\nu.\end{aligned}$$

The term in brackets is deterministic, and the next is

$$2a\bar{X}\frac{1}{|H_j^\nu|}\sum_{k=1}^{|H_j^\nu|}\left(Z_k^\nu - \bar{Z}^\nu\right), \qquad Z_k^\nu \sim \mathit{Bernoulli}(p^\nu).$$

So, essentially we are concerned with a CLT approximation for a sum of $|H_j^\nu|$ *Bernoulli*(p^ν) random variables. Consider the following theorem

Theorem B.1 ([6])**.** *Let $Z_1, \dots$ be i.i.d. with $\mathbb{E}Z_i = 0$, $\mathbb{E}|Z_i|^3 < \infty$. If D_n is the distribution of $(Z_1 + \dots + Z_n)/(n\mathbb{E}Z_1^2)^{1/2}$, and $\mathcal{N}(x)$ is that of a standard normal, then*

$$|D_n(x) - \mathcal{N}(x)| \le 3\frac{\mathbb{E}|Z_1|^3}{\sqrt{n}(\mathbb{E}Z_1^2)^{3/2}}.$$

With $Z_1^\nu \sim \mathit{Bernoulli}(p^\nu)$ ($p^\nu \ll 1$), we have $\mathbb{E}|Z_1^\nu - \bar{Z}_1^\nu|^i \approx p^\nu$ for $i = 2, 3$ and so the right hand side in Theorem B.1 is $\approx 3/(p^\nu|H_j^\nu|)^{1/2}$. Since $p^\nu \approx \#\left\{\omega^k \in H_j^\nu : \omega^k \text{ hits the plume}\right\}/|H_j^\nu|$, our scaling criteria is $\sqrt{\#\left\{\omega^k \in H_j^\nu : \omega^k \text{ hits the plume}\right\}} \gg 1$.

Acknowledgments

The work of GB and IL was partially supported by the NSF Grant DMS-0804696 and DOE Grant DE-FG52-08NA28779. The work of YM was supported by a grant from the DOE Early Career Research Program (DE-SC0003908). The authors would also like to thank Lisandro Dalcin at CIMEC for developing and supporting the mpi4py module for Python.

References

[1] Kaipio J. P. Kolehmainen V. Schweiger M. Somersalo E. Tarvainen T. Arridge, S. R. and M. Vauhkonen. Approximation errors and model reduction with an application in optical diffusion tomography. *Inverse Problems*, 22(175), 2006.

[2] G. Bal, A. Davis, and I. Langmore. A hybrid (Monte Carlo/deterministic) approach for multi-dimensional radiation transport. *submitted*, 2011.

[3] G. Casella and R. Berger. *Statistical Inference.* Duxbury, 2002.

[4] J. Chen and X. Intes. Time-gated perturbation Monte Carlo for whole body functional imaging in small animals. *Optics Express*, 17(22), October 2009.

[5] J. Andrés Christen and C. Fox. Markov chain Monte Carlo using an approximation. *Journal of computational and graphical statistics*, 14(4):795–810, 2005.

[6] R. Durrett. *Probability: Theory and examples.* Brooks/Cole, third edition, 2005.

[7] Y. Efendiev, T. Hou, and W. Luo. Preconditioning markov chain Monte Carlo simulations using coarse-scale models. *SIAM J. Sci. Comput*, 28(2):776–803, 2006.

[8] C.J. Geyer. Practical Markov chain Monte Carlo. *Statistical Science*, 7(4):473–511, 1992.

[9] C. Hayakawa and J Spanier. *Perturbation Monte Carlo methods for the solution of inverse problems.* Monte Carlo and quasi Monte Carlo methods. Springer, 2004.

[10] C. K. Hayakawa, J. Spanier, and F. Bevilacqua *et al.* Perturbation Monte Carlo methods to solve inverse photon migration problems in heterogeneous tissues. *Optics Letters*, 26(17):1333–1337, 2001.

[11] J. P. Kaipio and E. Somersalo. *Statistical and Computational inverse problems.* Springer Verlag, New York, 2004.

[12] J. P. Kaipio and E. Somersalo. Statistical inverse problems: discretization, model reduction, and inverse crimes. *Journal of computational and applied mathematics*, 198(2):493–504, 2007.

[13] I. Langmore, A. Davis, and G. Bal. Toward physics-based atmosphere/surface remote sensing in 3d geometry: Proof-of-concept for an absorbing gaseous plume in a deep valley using reflected sunlight. *In preparation.*

[14] J.S Liu. *Monte Carlo strategies in scientific computing.* Springer series in statistics. Springer, 2008. page 194.

[15] J.D. Moulton, C. Fox, and D. Svyatskiy. Multilevel approximations in sample-based inversion from the Dirichlet-to-Neumann map. *Journal of Physics: Conference series*, 124, 2008.

[16] H. K. Pikkarainen. State estimation approach to nonstationary inverse problems: discretization error and filtering problem. *Inverse Problems*, 22(365), 2006.

[17] C. Robert and G. Casella. *Monte Carlo Statistical Methods.* Springer, 2004.

[18] L. Tierney. Markov chains for exploring posterior distributions. *The Annals of Statistics*, 22(4):1701–1762, 1994.