Efficient inversion strategies for estimating optical properties with Monte Carlo radiative transport models

Callum M. Macdonald,[a,*] Simon Arridge,[b] and Samuel Powell[c]
[a]University College London, Department of Medical Physics and Biomedical Engineering, London, United Kingdom
[b]University College London, Department of Computer Science, London, United Kingdom
[c]University of Nottingham, Faculty of Engineering, Nottingham, United Kingdom

Abstract

Significance: Indirect imaging problems in biomedical optics generally require repeated evaluation of forward models of radiative transport, for which Monte Carlo is accurate yet computationally costly. We develop an approach to reduce this bottleneck, which has significant implications for quantitative tomographic imaging in a variety of medical and industrial applications.

Aim: Our aim is to enable computationally efficient image reconstruction in (hybrid) diffuse optical modalities using stochastic forward models.

Approach: Using Monte Carlo, we compute a fully stochastic gradient of an objective function for a given imaging problem. Leveraging techniques from the machine learning community, we then adaptively control the accuracy of this gradient throughout the iterative inversion scheme to substantially reduce computational resources at each step.

Results: For example problems of quantitative photoacoustic tomography and ultrasound-modulated optical tomography, we demonstrate that solutions are attainable using a total computational expense that is comparable to (or less than) that which is required for a single high-accuracy forward run of the same Monte Carlo model.

Conclusions: This approach demonstrates significant computational savings when approaching the full nonlinear inverse problem of optical property estimation using stochastic methods.

 [DOI: 10.1117/1.JBO.25.8.085002]

Keywords: Monte Carlo; radiative transport; optical tomography; machine learning; stochastic-gradient descent.

Paper 200101R received Apr. 14, 2020; accepted for publication Jul. 23, 2020; published online Aug. 14, 2020.

1 Introduction

Inverse problems arise in many areas within biomedical optics, both for global characterization of optical properties of media and for image reconstruction, among other applications.[1] Inverse problems are often considered as optimization problems, solved by deriving the gradient of an objective function and iteratively descending through the solution space. This process requires repeated solutions of forward and corresponding adjoint problems that are often computationally demanding in their own right. If the forward problem is given by the solution to a partial differential equation (PDE), then one appealing approach is to solve the forward and inverse problems simultaneously so that the forward problem is only approximately solved at intermediate stages in the algorithm (i.e., before it has finally converged); this approach (which has its basis in optimal control) is known as PDE-constrained optimization.[2–4] In this work, we seek an equivalent

*Address all correspondence to Callum M. Macdonald, E-mail: callum.macdonald@ucl.ac.uk

framework for the case where approximate noisy solutions to the forward model can (or must) be sought by stochastic methods.

The application of stochastic methods for the solution of PDEs is particularly pertinent in problems involving diffuse optics, since the "gold standard" method of solving the radiative transfer equation (RTE)—which is the most generally applicable description of the underlying physics—is to use stochastic (Monte Carlo) techniques;[5] their use in such applications parallels their extensive employment in other fields such as neutron physics.[6] While approximations to the RTE (such as diffusion) which permit deterministic solutions are available, these are often not valid in many cases such as in small domains, close to sources and boundaries, and in regions with weak scattering or strong absorption. Analytical solutions to the RTE itself are known for some geometries, such as infinite space[7] and layered media,[8] but such expressions are not readily available for general domains. The practicality of Monte Carlo techniques has been significantly boosted by recent advances in computational hardware developments, particularly in the application of parallelization.[9,10] Other approaches to improve their computational performance have been explored, such as the introduction of perturbation techniques[11] or variance reduction techniques.[12,13] Consequently, even when the aforementioned approximations to the RTE are reasonable, Monte Carlo solutions may offer an attractive alternative to the use of deterministic techniques such as the finite element method, when the complexity of the geometry or probe requires a high-density discretization of the spatial domain.

With both deterministic and stochastic solvers, the computational cost of the forward model typically remains the limiting factor in image reconstruction procedures. However, stochastic methods have a particular quality distinct from deterministic methods: one may arbitrarily trade computational expense against noise in the estimated solution without bias. In the case of diffuse optics, this trade-off is mediated through the number of virtual photons simulated by the Monte Carlo model for a given problem. This fact naturally leads one to consider how much noise can be tolerated during the solution of the inverse problem, and if a strategy can be found by which to approach this solution with the least work.

Parallels can be drawn between this problem and large-scale machine learning, where the requirement is to find the global minimum of a loss function expressed through fitting a model to a very large set of training data. The recent growth in this field has led to significant developments in optimization methods using stochastic subsets, especially the use of approximate gradients at intermediate steps, which is a technique known as stochastic-gradient descent (SGD). At the heart of this issue is the interplay between optimization and randomness, and the fact that attaining highly accurate estimations of the gradient at each step in SGD can come at a high cost when dealing with large datasets. However, if we can accept certain levels of randomness in our gradient computation, then each step in the gradient descent (GD) can be achieved at a lower computational cost. Returning to the context of biomedical optics, we may be able to accept a "noisy" low-cost forward model computation (which would otherwise be undesirable in the PDE-constrained approach) and simulate fewer photon trajectories during the earlier stages of the inversion process, leading to an overall accuracy versus computation time benefit. Thus, the topic of how to most efficiently utilize finite-sized datasets in machine learning is relevant to the deployment of Monte Carlo-based solvers in biomedical optics.

In this study, we attempt to translate these recent insights from SGD in machine learning into practical suggestions to improve the use of Monte Carlo methods in inverse problems that arise in biomedical optics. To do this, we employ a fully stochastic computation of an objective gradient using forward and adjoint models of the RTE solved by the Monte Carlo method. This allows for the full nonlinear inverse problem to be approached. In our demonstration problems, the inverse problem can be approximately solved using a total computational expenditure which is similar or less than that which would typically be dedicated to a single high-quality (low variance) solution of the forward imaging problem.

This paper is organized as follows. First, we outline some key aspects of GD in Sec. 2, including what appropriate metrics can be used to quantify acceptable levels of variance in the computation of subgradients via a stochastic process (i.e., Monte Carlo), and what step sizes to use to allow convergence. In Sec. 3, we describe the example problems for evaluating the improvements of SGD in a biomedical context, including details of the Monte Carlo forward/adjoint models and gradient calculations. In Sec. 4, we apply these ideas to two different

coupled-physics imaging (CPI) modalities, namely, quantitative photoacoustic tomography (QPAT) and ultrasound-modulated optical tomography (UMOT).[14,15] Both of these problems are nonlinear and entail the RTE for an accurate description, but exhibit different degrees of ill-posedness and resolution; thus they serve to demonstrate the generality of our approach. We evaluate the performance of various Monte Carlo inversions using simulated QPAT and UMOT data in Sec. 4, and discuss what practical lessons can be taken from this in Sec. 5.

2 Modeling and Inversion Problems in Optical Tomography

A common problem in biomedical optics involves finding the internal distribution of some optical properties x within a medium using various measurements made around and/or within the medium y^{obs}. To do this, we can employ some forward model of the underlying physical problem A, which produces an output y, given some estimate of the internal properties x,

$$y = Ax, \tag{1}$$

where in this case, the forward model A could represent the RTE and all relevant aspects of the optical setup (geometry of sources and detectors). In cases where A is not directly invertible, then to solve for an unknown distribution of properties x, we can formulate a cost function as a measure of the quality of an estimate. This could for example be the $L2$-norm of the residual between the real measured data, y^{obs}, and our forward modeled data, y.

$$F(x) = \frac{1}{2}\|y^{\text{obs}} - y\|^2 = \frac{1}{2}\|y^{\text{obs}} - Ax\|^2. \tag{2}$$

From this point, the problem now becomes one of minimization, where we will qualify our solution x^* as that which minimizes the cost function, $x^* = \arg\min_x F(x)$. It is worthy to note that the ground-truth parameters x^{true} may differ from the minimizer x^* leading to reconstruction error. This minimization problem can be approached via iterative GD, where we start with some estimate x_0, and each successive iterate, x_n, is determined by subtracting a (scaled) gradient of our cost function ∇F (relative to the internal optical properties) from the previous iterate

$$x_n = x_{n-1} - \alpha_n \nabla F(x_{n-1}), \tag{3}$$

where α_n is the step size which scales the update term. If we have access to some computation or set of computations (sometimes referred to as a "first-order oracle") which we can call to compute $F(x_{n-1})$ and $\nabla F(x_{n-1})$, then this algorithm can be implemented and is said to converge if $\lim_{n\to\infty} F(x_n) = 0$. In practice, the descent may be terminated early once the cost function reaches some acceptable value, for example, when the norm of the difference between observed and model data is of the same order as measurement noise, a criterion known as the discrepancy principle.[16]

2.1 *Stochastic-Gradient Descent*

In a stochastic setting, for instance, when our forward model A is a Monte Carlo model of radiative transport, then the true cost F and gradient ∇F given in Eq. (3) are not directly available. Instead, we may only have access to estimates of the cost function and gradient (provided by a stochastic first-order oracle). In Sec. 3, we detail the nature of these stochastic Monte Carlo computations in the radiative transport setting. In the interest of generality, for now, we simply assume such models exist, and we can make a call to a "stochastic oracle" to attain F_{S_n} and ∇F_{S_n} which we assume are nonbiased approximations., i.e.,

$$\mathbb{E}[F_{S_n}(x_n)] = F(x_n), \quad \mathbb{E}[\nabla F_{S_n}(x_n)] = \nabla F(x_n), \tag{4}$$

where $\mathbb{E}$ denotes the mean (expected) value for scalar quantities or the mean (expected) vector for vector quantities such as the gradient. Here, S_n denotes the n'th "sample" used in the computation. The meaning of sample here depends on the application. For example, in machine

learning, this may refer to a particular training example (or group of training examples) to be used during one learning iteration.[17] In Monte Carlo modeling of radiative transport, the sample refers to the set of virtual photons (and their associated random number seeds) that are initiated in the simulation to represent an optical source, which are subsequently used to estimate $F(x_n)$ and $\nabla F(x_n)$. The stochastic version of gradient descent (SGD) thus attempts to minimize a sampled objective function, F_{S_n}, by updating the previous iterate with a scaled sampled gradient.

$$x_n = x_{n-1} - \alpha_n \nabla F_{S_n}(x_{n-1}). \tag{5}$$

As with any computation, a call to a stochastic oracle at each iteration comes with a certain computational cost. The particular cost may depend on a number of factors, including the sample size, $|S_n|$. This is one of the reasons why the study of SGD is of such importance in modern machine learning, where training datasets may be of an enormous size, which means that computing a gradient based on all available data at each iteration could be very costly. Rather, individual samples ($|S_n| = 1$) or batches of samples ($|S_n| > 1$) may be used instead at each iteration. While this degrades the quality of any individual gradient estimate compared to using all available data, if the variance of these estimates is maintained below an acceptable value, the overall trade-off may be net positive. What this means in a Monte Carlo radiative transport context is that we may be able to allow low-quality gradient estimations (simulating only a small number photons) for a large part of the inversion process when estimating optical properties, saving on per iteration computational resources, and leading to an overall efficiency improvement. This is in contrast to typical implementations of iterative Monte Carlo solvers in the biomedical optics community, where each iteration is computed with large numbers of photons that are deemed sufficient to produce "stable" and "smooth" (low variance) forward model data.[18–24] In some cases, where a linearized approximation is assumed for the inverse problem, the cost of rerunning the forward model can be avoided using techniques such as perturbation Monte Carlo (PMC) methods.[11,25,26] However, for the full nonlinear problem, although PMC can be used for calculation of the problem Jacobian, this has to be recomputed at each iteration of, for example, a Gauss–Newton optimization scheme.[27]

In this study, if we are to accept a level of variance and imperfection in our forward/adjoint models, this of course raises the question of how much variance is acceptable in order for SGD to be successful? Furthermore, what measure of the variance is the best indicator in terms of efficiency/performance for common Monte Carlo solvers? To begin to answer this, it is important to first note that fixed-step SGD does not in general converge to a solution.[28,29] That is, if α_n is fixed for all n, eventually there will come a point where the next update of the estimate [with the term $\alpha_n \nabla F_{S_n}(x_{n-1})$] will reliably "undo" the work of the prior step, which will effectively halt the descent. The point at which this occurs depends on the variance of ∇F_{S_n}. We can see this by rewriting the sampled stochastic gradient estimate as

$$\nabla F_{S_n}(x_n) = \nabla F(x_n) + \epsilon_{S_n}(x_n), \tag{6}$$

where ϵ is a random vector with $\mathbb{E}[\epsilon_{S_n}(x_n)] = 0$ for all n. As GD progresses successfully, the "true" gradient ∇F will eventually begin reducing in size as we near the minimum. Once the magnitude of the true gradient reduces to a point at which it is comparable to the randomness of ϵ_{S_n}, the problem arises. The larger the expected magnitude of ϵ_{S_n}, the sooner the minimization of the cost function reaches this limiting scenario, where further iterations will only lead to a random walk about this point.

To prevent this from happening, we may take one of two actions (or a combination thereof): (i) reduce the step size at each iteration such that we can avoid "backtracking" in the descent, more on this in Sec. 2.3 or (ii) gradually improve the accuracy of our sampled gradient such that the variance of the sampled gradient remains below some threshold value compared to the norm of the true gradient ∇F. In other words, we may wish to ensure the inequality

$$V_{\text{tot}}^2(x_n) := \frac{\mathbb{E}[\|\epsilon_{S_n}(x_n)\|^2]}{\|\nabla F(x_n)\|^2} \leq \gamma_{\text{tot}}^2, \quad \gamma_{\text{tot}} > 0, \tag{7}$$

where γ_{tot} is a positive coefficient describing the acceptable threshold. The aforementioned inequality is known as the "norm test."[30] It is worthy to note that, since for any vector of random variables the variance of its length is the sum of the variances parallel and orthogonal to any fixed vector, this test equally penalizes the components of randomness parallel and perpendicular to the true gradient. Recent studies, however, have demonstrated that controlling the component of randomness parallel to ∇F is potentially a more relevant objective, as the component of the sampled gradient orthogonal to the true gradient is zero in expectation. An alternative measure of acceptable variance in ∇F_{S_n} has thus been introduced as the "inner product test,"[30] which only aims to restrict the component of variance in the sampled gradient parallel to the true gradient ∇F.

$$V_{\|}^2(x_n) := \frac{\mathbb{E}[\langle \epsilon_{S_n}(x_n), \nabla F(x_n)\rangle^2]}{\|\nabla F(x_n)\|^4} \leq \gamma_{\|}^2, \quad \gamma_{\|} > 0. \tag{8}$$

This inner product test imposes a less restrictive limitation of the overall variance in the sampled gradients, particularly in cases where the variance may be higher in directions orthogonal to the true gradient than in the direction parallel to ∇F. However, either of these metrics will be able to exploit the fact that an increased expected error, $\mathbb{E}[\|\epsilon_{S_n}\|]$, will correlate to a cheaper computation of the estimated gradient. Thus, setting larger values of γ_{tot} or $\gamma_{\|}$ in the inequalities will correspond to cheaper computational requirements for each step, but also a more pronounced random walk component to the GD. In many cases, it may be found that the penalty paid by increasing the random walk component is acceptable (up to a point) compared to the penalty paid in computational cost for reducing the expected norm of ϵ to a negligible value. For example, using Monte Carlo RTE simulations to compute ∇F with a negligible level of variance (i.e., setting $\gamma_{\text{tot}} \ll 1$) may take billions of simulated photons at each step. Whereas, it may be possible to compute a gradient that passes the norm test or inner product with larger values of γ_{tot} or $\gamma_{\|}$ with many orders of magnitude less photons, particularly during the early stages of GD, where we may be far from the minimum. The ideal choice of γ_{tot} or $\gamma_{\|}$ will depend on the specific application.

2.2 *Adaptive Sample Size*

We have discussed two different measures of the variance in the sampled gradient ∇F_{S_n} that we wish to investigate in the context of Monte Carlo estimation of media properties, viz., the norm test [Eq. (7)] and the inner product test [Eq. (8)]. To satisfy the inequalities defining these tests as the GD progresses, we will be required to reduce the variance in the sampled gradients ∇F_{S_n} whenever the norm test or inner product test fail. This can be done by increasing the sample size (number of photons used $|S_n|$) when making a call to the stochastic oracle. Two practical considerations are still required: first, how to compute the "true" gradient ∇F, which is needed to evaluate the norm test and inner product test; and second, by how much we should increase the sample size in a situation where one of the tests fails?

The true gradient ∇F is only calculable in the limit that an infinite number of photons are used in the Monte Carlo model. This limit can equivalently be represented as an average over independent repeated outputs of the sampled gradient

$$\nabla F(x_n) = \lim_{N_{\text{rep}} \to \infty} \frac{1}{N_{\text{rep}}} \sum_{j=1}^{N_{\text{rep}}} \nabla F_{S_j}(x_n), \quad |S_j| = |S_n| \quad \forall\ j. \tag{9}$$

Using a finite value of N_{rep} in the evaluation of Eq. (9) provides an approximation to the true gradient, and when this is used to compute the norm and inner product tests [Eqs. (7) and (8)], the inequalities will fail before they would if $N_{\text{rep}} = \infty$, thus acting as a conservative approximation. It is noted that this method of approximating the true gradient is computationally taxing. However, in practice, the inner product test and norm tests can still be conducted efficiently if they are only computed occasionally (not at every iteration) of the descent. For example, using $N_{\text{rep}} = 100$ repeated computations of the sampled gradient to conduct the tests once every

100 iterations (thus only updating our sample size every 100 iterations) would only double the total number of simulated photons required for the inversion. In this study, we evaluate these metrics once every 10 iterations using $N_{\text{rep}} = 100$ repeated sampled gradients. While this is a significant computational burden, we do so in this study as we are interested in assessing the best case scenarios for such methods. It is worthy to note that although we compute the aforementioned approximation to the true gradient to evaluate the inner product and norm tests, we only ever update our estimate using the sampled gradient.

In terms of increasing the sample size in the event where the inner product and/or norm tests fail, this can be done in a number of ways. A simple method we will employ in this study is to scale the current sample size by some factor $\kappa(n)$ to increase the number of photons used in the next iteration

$$|S_{n+1}| = \kappa(n)|S_n|. \tag{10}$$

One option for $\kappa(n)$ is to use the same factor by which the variance exceeds our imposed limit at a given point in the descent. For instance, upon failure of the inner product test for a chosen value of $\gamma_{\|}$, we can increase the sample size on the next iteration using $\kappa(n) = V_{\|}^2(x_n)/\gamma_{\|}^2$. However, we also investigate other forms of $\kappa(n)$ in Sec. 4, which better cope with statistical variations that can lead to overestimating the required sample size increase.

2.3 *Step Size*

In cases where we are not taking actions to bound the error in the sampled gradient (such as enforcing successful outcomes of an inner product test or norm test), fixed-step SGD may only converge to a region around the solution. Reducing the step size sufficiently at each step is usually required to allow convergence.[31] However, it can be shown that if we are bounding the error in the sampled gradient, e.g., by increasing the sample size, then fixed-step SGD may converge so long as the following is satisfied for all n:[30]

$$\alpha_n \leq \frac{1}{(1+\gamma_{\text{tot}}^2)L}, \tag{11}$$

where L is the Lipschitz constant for F. The Lipschitz constant for a functional F is a measure of its rate of change with respect to its parameter and can be defined, for example, as the smallest constant such that $\nabla^2 F \leq L\text{Id}$, where Id is the identity matrix, and we assume that F is twice continuously differentiable. It can also be interpreted as the largest eigenvalue of the Hessian of F.[32] As intuition may indicate, when the sample size (e.g., number of simulated photons) increases toward the maximum number of samples $|S_n| \to |S_{\text{max}}|$ ($|S_{\text{max}}| = \infty$ in the case of Monte Carlo RTE simulations), the expected error in the sampled gradient approaches zero, $|\epsilon_{S_n}| \to 0$, as do the measures of variance in the sampled gradients ($V_{\text{tot}}^2 \to 0$, $V_{\|}^2 \to 0$), as defined in Eqs. (7) and (8). In other words, as the stochasticity in the problem reduces to zero, we approach the classical step size of the deterministic problem given by $\alpha = \frac{1}{L}$.[32]

In this study, we aim to satisfy the aforementioned step size criteria for an assumed value of the Lipschitz constant L, which we will choose conservatively depending on the particular scenario. However, as we are primarily interested in reaching the best possible solution for a given allocation of computational resources, convergence to a region around the unique solution may be sufficient for our purposes. For this reason, we will also investigate larger step size criteria, which violate Eq. (11), yet exhibit good performance in our scenarios of interest.

Taking the above considerations into account, we present a basic method for SGD using adaptive sample sizes in Algorithm 1 (simplified from Ref. 30). The algorithm imposes a limit on the total number of photons to be simulated using Monte Carlo transport models throughout the entire descent, N_{ph}.

Algorithm 1 Inversion using Monte Carlo sampled gradients with adaptive sample size.

Choose initial photon sample size $|S_1|$, and desired value of $\gamma_{\|}$ or γ_{tot}

while $\sum_{i=1}^{n} |S_i| < N_{\text{ph}}$ **do**

  if run test? **then**

    compute sampled gradient, ∇F_{S_n}, and approximate true gradient, ∇F [using Eq. (9)]

    check norm test (or) inner product test is satisfied

    if test fail **then**

      increase sample size on next iteration $|S_{n+1}| = \kappa(n)|S_n|$

    else

      set $|S_{n+1}| = |S_n|$

    end if

  else

    compute sampled gradient only ∇F_{S_n}

    set $|S_{n+1}| = |S_n|$

  end if

  update $x_{n+1} = x_n - \alpha_n \nabla F_{S_n}$

end while

3 Stochastic Forward and Adjoint Models

In this section, we cover the computation of the stochastic forward model and stochastic gradient approximation, referred to as the first-order stochastic oracle. We will cover the basic radiative transport forward problem, and the gradient computations involved in our example problems of absorption estimation in QPAT and UMOT. The specific details of these models are not required to understand the main premise of this paper, but serve as a demonstration in a context familiar to many in the biomedical optics community, where Monte Carlo models of optical transport are employed to estimate medium properties.

3.1 *Forward Model*

For any optical source $Q(\mathbf{r}, \hat{\mathbf{s}})$, either incident on a medium or present within it, we wish to model the resulting radiance, $\phi(\mathbf{r}, \hat{\mathbf{s}})$, describing the radiant flux at each position $\mathbf{r}$, and in each direction $\hat{\mathbf{s}}$. This can be achieved using the RTE.

$$\underbrace{[\hat{\mathbf{s}} \cdot \nabla + \mu_a(\mathbf{r}) + \mu_s(\mathbf{r})]}_{\mathcal{T}_{\mu_a,\mu_s}} \phi(\mathbf{r}, \hat{\mathbf{s}}) = \underbrace{\mu_s(\mathbf{r}) \int_{\mathcal{S}^2} p(\hat{\mathbf{s}}, \hat{\mathbf{s}}') \phi(\mathbf{r}, \hat{\mathbf{s}}') \mathrm{d}\hat{\mathbf{s}}'}_{\mathcal{S}_{\mu_s}} + Q(\mathbf{r}, \hat{\mathbf{s}}), \tag{12}$$

where $\mathcal{T}$ and $\mathcal{S}$ denote the attenuation and scattering operators, respectively, which together compose the RTE operator, $\mathcal{L}$. For notational convenience, we assume that Eq. (12) is combined with appropriate boundary conditions, which we do not write explicitly here; see Ref. 33 for more details. Here, μ_a is the absorption coefficient, μ_s is the scattering coefficient, and p is the scattering phase function. Using the defined operators, Eq. (12) can be rewritten in a more compact form:

$$\mathcal{L}_{\mu_a,\mu_s} \phi = (\mathcal{T}_{\mu_a,\mu_s} - \mathcal{S}_{\mu_s}) \phi = Q. \tag{13}$$

To obtain (stochastic) estimates of the radiance resulting from a given source, and thus to obtain an estimate of any derived data function $y(\phi)$, we can implement a Monte Carlo solver, $\mathcal{L}_{\mathrm{MC}}^{-1}$. In this study, we have adapted a hardware-accelerated version (utilizing graphics processing units) of the commonly employed "Monte Carlo multilayer" program used to simulate radiative transport within a layered planar medium.[34,35] The basic operation of this program is unchanged from the original release. Simulated photons are initiated by sampling from a given source function, Q, and scattering/absorption events are pseudorandomly generated along each photon's trajectory until either: (i) the photon leaves the domain or (ii) the photon drops its weight below some threshold value. In this study, the scattering directions are sampled from the Henyey–Greenstein scattering phase function. The expected accuracy of the computed radiance using Monte Carlo solvers $\mathcal{L}_{\mathrm{MC}}^{-1}$ depends on the total number of photons used, i.e., the sample size $|S_n|$. As $|S_n| \to \infty$, the radiance approaches the deterministic solution of the RTE. Importantly, however, Monte Carlo models allow an estimate of the radiance to be achieved with any number of photons with $|S_n| \geq 1$. The expected computational requirements (number of floating point operations) of the Monte Carlo solver $\mathcal{L}_{\mathrm{MC}}^{-1}$ also scales with the number of photons simulated, and it is this trade-off between accuracy of the forward model (and corresponding adjoint model) and computational cost that we will be investigating.

3.2 *Gradient Computation: Adjoint Model*

To compute the gradient of our cost function ∇F with respect to the optical properties of the medium, we make use of an adjoint RTE model. Although direct methods of finding the derivative of a Monte Carlo method can also be developed,[12] adjoint methods have more applicability in general, and also allow closer comparison with optimization techniques used in machine learning. For further details of forward and adjoint methods in the RTE, we refer to Ref. 36; for specific details of CPI problems, we refer to Ref. 37. We first consider a change to Eq. (12) where $\mu_{\mathrm{a}} \to \mu_{\mathrm{a}} + \mu_{\mathrm{a}}^{\delta}$, $\mu_{\mathrm{s}} \to \mu_{\mathrm{s}} + \mu_{\mathrm{s}}^{\delta}$, for the same source Q, which results in a change in radiance $\phi \to \phi + \phi^{\delta}$. This implies

$$\begin{aligned}(\mathcal{T}_{\mu_{\mathrm{a}}+\mu_{\mathrm{a}}^{\delta},\mu_{\mathrm{s}}+\mu_{\mathrm{s}}^{\delta}} - \mathcal{S}_{\mu_{\mathrm{s}}+\mu_{\mathrm{s}}^{\delta}})(\phi+\phi^{\delta}) &= (\mathcal{T}_{\mu_{\mathrm{a}},\mu_{\mathrm{s}}} - \mathcal{S}_{\mu_{\mathrm{s}}})\phi,\\ \Rightarrow (\mathcal{T}_{\mu_{\mathrm{a}},\mu_{\mathrm{s}}} - \mathcal{S}_{\mu_{\mathrm{s}}})\phi^{\delta} &= -(\mu_{\mathrm{a}}^{\delta} + \mu_{\mathrm{s}}^{\delta} - \mathcal{S}_{\mu_{\mathrm{s}}^{\delta}})\phi, \qquad (14)\end{aligned}$$

$$\mathcal{L}_{\mu_{\mathrm{a}},\mu_{\mathrm{s}}}\phi^{\delta} = -\underbrace{(\mu_{\mathrm{a}}^{\delta} + \mu_{\mathrm{s}}^{\delta} - \mathcal{S}_{\mu_{\mathrm{s}}^{\delta}})}_{\mathcal{L}^{\delta}_{\mu_{\mathrm{a}}^{\delta},\mu_{\mathrm{s}}^{\delta}}}\phi. \qquad (15)$$

We also define the fluence, Φ, as the angular integral of the radiance:

$$\Phi(\mathbf{r}) = \int_{\mathcal{S}^2} \phi(\mathbf{r},\hat{\mathbf{s}})\mathrm{d}\hat{\mathbf{s}}. \qquad (16)$$

To proceed beyond this point, we must now consider the specific form of the data function relevant to a particular modality of interest. We begin with the first of our two example modalities, QPAT.

3.2.1 *QPAT case*

In QPAT, the medium is illuminated with a pulsed optical source, Q (see Fig. 1). The distributed optical energy is absorbed at various points within the sample, giving rise to internal acoustic waves. These acoustic waves can be detected at the surface of the medium by a sensor and processed to locate the initial pressure distribution p_0 within the medium.[38–40] This internal pressure distribution is related to the spatial distribution of absorbed optical energy, h, where

$$h(\mathbf{r}) = \mu_{\mathrm{a}}(\mathbf{r})\Phi(\mathbf{r}), \qquad (17)$$

and where Φ is the optical fluence of Eq. (16). We have omitted the Grüneisen parameter for clarity of exposition, though this parameter can be included in practice. Assuming that we can

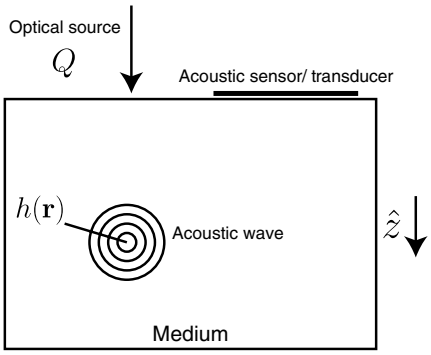

Fig. 1 Setup for QPAT.

recover the absorbed optical energy, h, the problem remains to find the distribution of $\mu_a(\mathbf{r})$ within the medium.[41,42] It is worth noting that although the optical source is pulsed, it is acceptable to use a continuous-wave (time-independent) model to describe ϕ and Φ because the time scale of the acoustic wave propagation is orders of magnitude slower than the optical propagation.[43] First, restating our cost function in terms of the QPAT data function, h, we have

$$F^{\mathrm{QPAT}} = \frac{1}{2}\int_{\Omega}(h^{\mathrm{obs}} - h)^2 \mathrm{d}r = \frac{1}{2}\langle h^{\mathrm{obs}} - h, h^{\mathrm{obs}} - h\rangle_{L^2(\Omega)}. \tag{18}$$

We then write the Fréchet derivative of F^{QPAT} as

$$DF^{\mathrm{QPAT}} = -\langle h^{\mathrm{obs}} - h, Dh\mu_a^{\delta}\rangle_{L^2(\Omega)}, \tag{19}$$

where μ_a^{δ} is a small change in absorption. In this paper, we will neglect changes in scattering, however, the below formalism is still general for the gradient with respect to absorption. The gradient term with respect to scattering coefficient is described in Ref. 42 and will be included in future investigations. Writing the Fréchet derivative of h as

$$Dh = \Phi + \mu_a \cdot D\Phi, \tag{20}$$

and defining $\Phi^{\delta} = D\Phi\mu_a^{\delta}$, we arrive at

$$DF^{\mathrm{QPAT}} = -\langle \Phi(h^{\mathrm{obs}} - h), \mu_a^{\delta}\rangle_{L^2(\Omega)} - \langle \mu_a(h^{\mathrm{obs}} - h), \Phi^{\delta}\rangle_{L^2(\Omega)}. \tag{21}$$

Next, we define the adjoint radiance, ϕ^*, as the solution to

$$\mathcal{L}^*\phi^* = \mu_a(h^{\mathrm{obs}} - h) \tag{22}$$

where the right-hand side describes the "adjoint source" which is isotropic in $\hat{\mathbf{s}}$. We then substitute the above into Eq. (21) to give

$$DF^{\mathrm{QPAT}} = -\langle \Phi(h^{\mathrm{obs}} - h), \mu_a^{\delta}\rangle_{L^2(\Omega)} - \langle \mathcal{L}^*\phi^*, \phi^{\delta}\rangle_{L^2(\Omega\times S^{n-1})}, \tag{23}$$

where we exploited the fact that the right-hand side of Eq. (22) does not depend on direction. Using the definition of the adjoint operator, and the fact that the change in radiance is zero on the boundary $\partial\Omega$ yields

$$DF^{\mathrm{QPAT}} = -\langle \Phi(h^{\mathrm{obs}} - h), \mu_a^{\delta}\rangle_{L^2(\Omega)} - \langle \phi^*, \mathcal{L}\phi^{\delta}\rangle_{L^2(\Omega\times S^{n-1})}. \tag{24}$$

Finally, we make use of the perturbation expression Eq. (15), while again, here, we neglect any change in scattering. This gives

$$DF^{\mathrm{QPAT}} = -\langle \Phi(h^{\mathrm{obs}} - h), \mu_a^{\delta}\rangle_{L^2(\Omega)} + \langle \phi^*\phi, \mu_a^{\delta}\rangle_{L^2(\Omega\times S^{n-1})}, \tag{25}$$

allowing us to define the (absorption) gradient as in Eq. (33) of Ref. 42

$$\frac{\partial F^{\mathrm{QPAT}}}{\partial \mu_{\mathrm{a}}} = \nabla F^{\mathrm{QPAT}} = -\Phi(h^{\mathrm{obs}} - h) + \int_{S^{n-1}} \phi^* \phi \, \mathrm{d}\hat{\mathbf{s}} \tag{26}$$

To compute a stochastic approximation of this gradient, we can thus use the forward model Monte Carlo solver $\mathcal{L}_{\mathrm{MC}}^{-1}$ to provide estimates of ϕ and Φ, and an adjoint Monte Carlo solver $\mathcal{L}_{\mathrm{MC}}^{-1*}$ to produce ϕ^* from an adjoint source term $Q_{\mathrm{adj}} = \mu_{\mathrm{a}}(h^{\mathrm{obs}} - h)$, as defined in Eq. (22). Due to the symmetry of the problem, the adjoint solver is identical to the forward solver and follows the same basic operating principles. The only difference is that here the adjoint source $Q_{\mathrm{adj}} = \mu_{\mathrm{a}}(h^{\mathrm{obs}} - h)$ may in fact be negative in some locations. This is handled by splitting the source term into two parts, one purely positive, Q_{adj}^{+}, and one purely negative, Q_{adj}^{-}. Two simulations are then run (where the total number of photons to be used is split between the two simulations accordingly), and the results summed to produce ϕ^*. Algorithm 2 describes the basic operation for computing a sampled gradient, ∇F_{S_n}, for QPAT using the above derivation. This will be used in conjunction with Algorithm 1 to conduct an inversion with adaptive sample size for each iterate, $|S_n|$.

3.2.2 *UMOT case*

Referring to Fig. 2, in UMOT, we have an optical light source Q_q incident on a medium, as well as an optical detector J_m. In addition, an ultrasound source is incident on the medium, where the focus $\eta(\mathbf{r})$ is scanned through the sample.[44,45] Assuming for simplicity an ideal (delta function) ultrasound focus, the data of interest in this case are found to be of the form[46]

$$b(\mathbf{r}) = \eta(\mathbf{r})\Phi_q(\mathbf{r})\Phi_m(\mathbf{r}), \tag{27}$$

where Φ_q is the fluence resulting from the optical source Q_q, and Φ_m is the resulting fluence from a virtual source Q_m which is reciprocal to the detector J_m.[46] From this point, we proceed in similar fashion as in Sec. 3.2.1, where now our data fitting error is given by

$$F^{\mathrm{UMOT}} = \frac{1}{2}\int_{\Omega} (b^{\mathrm{obs}} - b)^2 \mathrm{d}r = \frac{1}{2}\langle b^{\mathrm{obs}} - b, b^{\mathrm{obs}} - b\rangle_{L^2(\Omega)}, \tag{28}$$

and its Fréchet derivative as

$$DF^{\mathrm{UMOT}} = -\langle b^{\mathrm{obs}} - b, Db\mu_{\mathrm{a}}^{\delta}\rangle_{L^2(\Omega)}. \tag{29}$$

In this case, the Fréchet derivative of b becomes

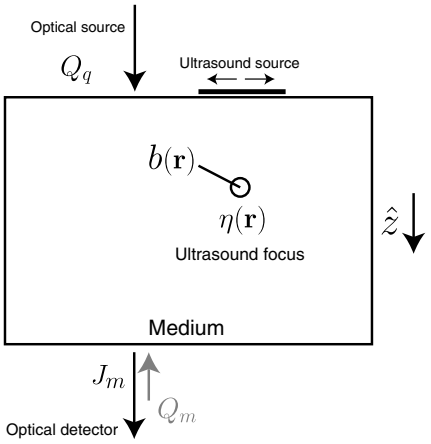

Fig. 2 Setup for UMOT in the transmission geometry.

$$Db = \eta\Phi_q \cdot D\Phi_m + \eta\Phi_m \cdot D\Phi_q, \tag{30}$$

leading to

$$DF^{\mathrm{UMOT}} = -\langle \eta\Phi_q(b^{\mathrm{obs}} - b), \Phi_m^{\delta}\rangle_{L^2(\Omega)} - \langle \eta\Phi_m(b^{\mathrm{obs}} - b), \Phi_q^{\delta}\rangle_{L^2(\Omega)}. \tag{31}$$

Here, we need to define two adjoint radiances, $\phi^{*,1}$ and $\phi^{*,2}$, as the solution to

$$\mathcal{L}^*\phi^{*,1} = \eta\Phi_q(b^{\mathrm{obs}} - b), \tag{32}$$

$$\mathcal{L}^*\phi^{*,2} = \eta\Phi_m(b^{\mathrm{obs}} - b), \tag{33}$$

and substituting into Eq. (31) to give

$$DF^{\mathrm{UMOT}} = -\langle \mathcal{L}^*\phi^{*,1}, \phi_m^{\delta}\rangle_{L^2(\Omega\times S^{n-1})} - \langle \mathcal{L}^*\phi^{*,2}, \phi_q^{\delta}\rangle_{L^2(\Omega\times S^{n-1})}; \tag{34}$$

by the same arguments as for QPAT we get

$$DF^{\mathrm{UMOT}} = -\langle \phi^{*,1}, \mathcal{L}\phi_m^{\delta}\rangle_{L^2(\Omega\times S^{n-1})} - \langle \phi^{*,2}, \mathcal{L}\phi_q^{\delta}\rangle_{L^2(\Omega\times S^{n-1})}. \tag{35}$$

Again using the perturbation expression [Eq. (15)], we have

$$DF^{\mathrm{UMOT}} = \langle \phi^{*,1}\phi_m, \mu_{\mathrm{a}}^{\delta}\rangle_{L^2(\Omega\times S^{n-1})} + \langle \phi^{*,2}\phi_q, \mu_{\mathrm{a}}^{\delta}\rangle_{L^2(\Omega\times S^{n-1})}, \tag{36}$$

allowing us to define the (absorption) gradient as

$$\frac{\partial F^{\mathrm{UMOT}}}{\partial \mu_{\mathrm{a}}} = \nabla F^{\mathrm{UMOT}} = \int_{S^{n-1}} (\phi^{*,1}\phi_m + \phi^{*,2}\phi_q)\mathrm{d}\hat{\mathbf{s}}. \tag{37}$$

Thus, similar to the QPAT case, here we are able to compute a stochastic approximation of this gradient using the forward model Monte Carlo solver $\mathcal{L}_{\mathrm{MC}}^{-1}$ to provide ϕ_q and ϕ_m from our two sources, and an adjoint Monte Carlo solver $\mathcal{L}_{\mathrm{MC}}^{-1*}$ to produce $\phi^{*,1}$ and $\phi^{*,2}$ from the adjoint source terms $Q_{\mathrm{adj}}^1 = \eta\phi_q(b^{\mathrm{obs}} - b)$ and $Q_{\mathrm{adj}}^2 = \eta\phi_m(b^{\mathrm{obs}} - b)$, as defined in Eqs. (32) and (33). Here as well, adjoint source terms are split into two parts, one purely positive, Q_{adj}^{+}, and one purely negative, Q_{adj}^{-}, with the photon budget being split accordingly. Algorithm 3 describes the basic operation for computing a sampled gradient, ∇F_{S_n}, for UMOT using the above derivation. This will be used in conjunction with Algorithm 1 to conduct an inversion with adaptive sample size for each iterate, $|S_n|$.

3.3 *Fluence Monte Carlo*

It should be noted that numerous Monte Carlo radiative transport solvers do not explicitly output the radiance, as this requires additional programming to store the angular ordinates at each location. Commonly, only the fluence will be available, which is the angular integral of the radiance Eq. (16). In such cases, the aforementioned integrals for the gradients of interest Eqs. (26) and (37) can be computed under the assumption of approximately angularly isotropic radiances,

Algorithm 2 Monte Carlo sampled QPAT gradient.

1. Compute $\mathcal{L}_{\mathrm{MC}}^{-1}Q \mapsto \phi$, Φ, using $|S_n|/2$ photons
2. Construct internal adjoint source $Q_{\mathrm{adj}} = \mu_{\mathrm{a}}(h^{\mathrm{obs}} - h)$
3. Compute $\mathcal{L}_{\mathrm{MC}}^{-1*}Q_{\mathrm{adj}} \mapsto \phi^*$, Φ^*, using $|S_n|/2$ photons
4. Use Eq. (26) to compute gradient ∇F_{S_n}

where for example $\int \phi^* \phi d\hat{\mathbf{s}}$ becomes $\Phi^* \Phi$. The accuracy of this approximation of course depends on the true angular dependence of the radiances, where the approximation is poorest in regions close to directional light sources, but improves further away. The higher the scattering asymmetry g of the medium, the slower the approximation improves as a function of distance from these sources. In many cases, however, this is a satisfactory assumption and is employed in the below example cases.

4 Results

In this section, we present the results of a number of investigations using our two example problems of QPAT and UMOT. We will demonstrate the implementation of the forward-adjoint Monte Carlo solvers described above, along with adaptive sampling strategies to estimate the absorption coefficient of a medium via SGD. Here, we investigate media with a semi-infinite slab geometry, with numerous layers in the z direction having different optical properties, but otherwise homogeneous in the x and y directions. The application to layered geometry in this demonstration was chosen for simplicity to provide an easily recognizable setting to test these adaptive sampling methods. Furthermore, while apparently simplistic, layered geometries are still of practical interest for applications including instrument calibration and validation, and the imaging of biological structures with small curvature but significant heterogeneity in depth. The latter example includes studies such as functional (cognitive) imaging when localized to small activation regions. Application of these new methods in more complicated 3D geometries will be carried out in future work. Each of the medium layers can be described in terms of thickness, scattering coefficient, absorption coefficient, (background) refractive index, and scattering asymmetry parameter. We will assume all parameters of the layered medium are known *a priori* with the exception of the absorption coefficient, which we will attempt to solve for. For the examples in this study, we set the total slab thickness to 2 cm, and the inversion is conducted with a resolution of 0.25 mm, (80 layers). The true "measured" data in all problems are generated using a single forward model Monte Carlo simulation using a large sample size of 10^9 photons. With this sample size, the variance of the measured forward data h^{obs}, b^{obs} is found to be negligible in this setup, and as such can be treated as effectively equivalent to the deterministic solution of the RTE.

To conduct an inversion, we stipulate a total photon budget, N_{ph}, for which all combined sample sizes in the descent must not exceed, i.e., $\sum_n |S_n| \leq N_{\text{ph}}$. Once the total photon budget is expended, we terminate the descent. This is to emulate an imposed restriction on computational resources required to reach a solution. While each iteration (involving forward and adjoint runs of the Monte Carlo) has a nonzero computational overhead, optimization of these Monte Carlo programs for repeated iteration (such as employed in Ref. 23) allows this overhead to become negligibly small. This means that the required computational resources of the inversion (and therefore required computation time) are proportional to the total number of simulated photons used throughout the descent, i.e., the photon budget N_{ph}. The inversions are carried out using Algorithm 1, along with Algorithms 2 and 3 to compute the gradients for QPAT and UMOT, respectively. In Algorithm 1, we will compute the metrics V_{tot}^2 and $V_{\parallel}^2$ and conduct the norm test and inner product test once every 10 iterations to evaluate the quality of our computed gradients (using $N_{\text{rep}} = 100$ independent repeated samples of the gradient), and to update the step size and sample size. It is worthy to note that as this is an investigation of how such methods might

Algorithm 3 Monte Carlo sampled UMOT gradient.

1. Compute $\mathcal{L}_{\text{MC}}^{-1} Q_q \mapsto \phi_q$, Φ_q, and $\mathcal{L}_{\text{MC}}^{-1} Q_m \mapsto \phi_m$, Φ_m, each using $|S_n|/4$ photons
2. Construct internal adjoint sources $Q_{\text{adj}}^1 = \eta \Phi_q (b^{\text{obs}} - b)$ and $Q_{\text{adj}}^2 = \eta \Phi_m (b^{\text{obs}} - b)$
3. Compute $\mathcal{L}_{\text{MC}}^{-1*} Q_{\text{adj}}^1 \mapsto \phi^{*,1}$, $\Phi^{*,1}$, and $\mathcal{L}_{\text{MC}}^{-1*} Q_{\text{adj}}^1 \mapsto \phi^{*,2}$, $\Phi^{*,2}$, each using $|S_n|/4$ photons
4. Use Eq. (37) to compute ∇F_{S_n}

Table 1 Table showing the different inversion strategies used. Strategy 1 has a constant step size, with adaptive sample size. Strategies 2 and 3 both have adaptive step sizes and adaptive sample sizes. It is worthy to note that in accordance with Algorithm 1, the sample size is only increased upon a failure of the relevant test. If the test passes, then $|S_{n+1}| = |S_n|$.

Strategy	Step size, α_n	Sample size, $\lvert S_{n+1}\rvert = \kappa(n)\lvert S_n\rvert$
1	$\frac{1}{(1+\gamma_{\text{tot}}^2)L}$	$\lvert S_{n+1}\rvert = \frac{V_{\text{tot}}^2}{\gamma_{\text{tot}}^2}\lvert S_n\rvert$
2	$\frac{1}{(1+V_{\text{tot}}^2)L}$	$\lvert S_{n+1}\rvert = \frac{V_{\parallel}^2}{\gamma_{\parallel}^2}\lvert S_n\rvert$
3	$\frac{1}{(1+V_{\text{tot}})L}$	$\lvert S_{n+1}\rvert = \frac{V_{\parallel}}{\gamma_{\parallel}}\lvert S_n\rvert$

perform in best case scenarios, we do not include the photons used to compute these metrics as counting against the total allowed photon budget.

There are three different strategies we have employed to control the step size and sample size as the inversion progresses, see Table 1 for a summary. Strategy 1 uses a fixed-step size as described in Eq. (11) for a chosen value of γ_{tot}. The sample size is adaptive and attempts to enforce successful outcomes of the norm test ($V_{\text{tot}}^2 \leq \gamma_{\text{tot}}^2$), by increasing the sample size when the norm test is violated. In the event of a violation of this inequality, the fractional increase in the sample size is equivalent to the factor by which the norm test fails, $V_{\text{tot}}^2/\gamma_{\text{tot}}^2$. Strategy 2 uses an adaptive step size which still satisfies Eq. (11); however, it selects the largest step size possible for this criterion each time the metrics are evaluated. In this strategy, the sample size is also adaptive and attempts to enforce successful outcomes of the inner product test ($V_{\parallel}^2 \leq \gamma_{\parallel}^2$) by increasing the sample size when the inner product test is violated. In the event of a violation of this inequality, the fractional increase in the sample size is equivalent to the factor by which the inner product test fails, $V_{\parallel}^2/\gamma_{\parallel}^2$. In strategy 3, we attempt to accelerate the descent using a larger adaptive step size with V_{tot} in the denominator in place of V_{tot}^2. Upon failure of the inner product test, the sample size is increased by fraction $V_{\parallel}/\gamma_{\parallel}$, and differs from strategy 2 to reduce the speed at which the photon budget is depleted. This is an attempt to reduce premature increase of the sample size caused by volatility in the computation of the norm and inner product metrics.

Finally, we introduce an error function for the estimated absorption distribution, μ_{a}, as

$$F_{\mu_{\text{a}}} = \frac{1}{2}\|\mu_{\text{a}}^{\text{true}} - \mu_{\text{a}}\|^2. \tag{38}$$

This metric would not be available under normal circumstances (as we do not know the ground truth $\mu_{\text{a}}^{\text{true}}$), however it is useful to monitor in terms of the underlying performance of each strategy. Furthermore, as we will see, the sampled data cost function F_{S_n} is itself heavily dependent on the number of photons (sample size) used in the forward Monte Carlo and is thus not an ideal indicator of proximity to the true solution.

4.1 *QPAT*

We begin with our example QPAT problem. The starting sample size in all cases shown is $|S_1| =$ 200 photons per iteration (100 for each forward run, and 100 for each adjoint run in accordance with Algorithm 2), and the total photon budget for the inversion was set to $N_{\text{ph}} = 2 \times 10^6$ photons. The Lipschitz constant was set at $L = 2.5$, as this displayed stable descent in our test problems using large photon budgets (low-variance case). The initial estimate of the absorption distribution in the medium is $\mu_{\text{a}} = 0.2\ \text{cm}^{-1}$ in all layers. The scattering coefficient of all layers was set to $\mu_{\text{s}} = 40\ \text{cm}^{-1}$, and the scattering asymmetry parameter was set to $g = 0.9$. The ground-truth absorption $\mu_{\text{a}}^{\text{true}}$ is shown in Fig. 3(a), along with the final retrieved absorption distributions obtained via strategies 1, 2, and 3 using the stated values of γ_{tot} and $\gamma_{\parallel}$. Figure 3(b) shows the corresponding measured data and the final forward modeled data for each

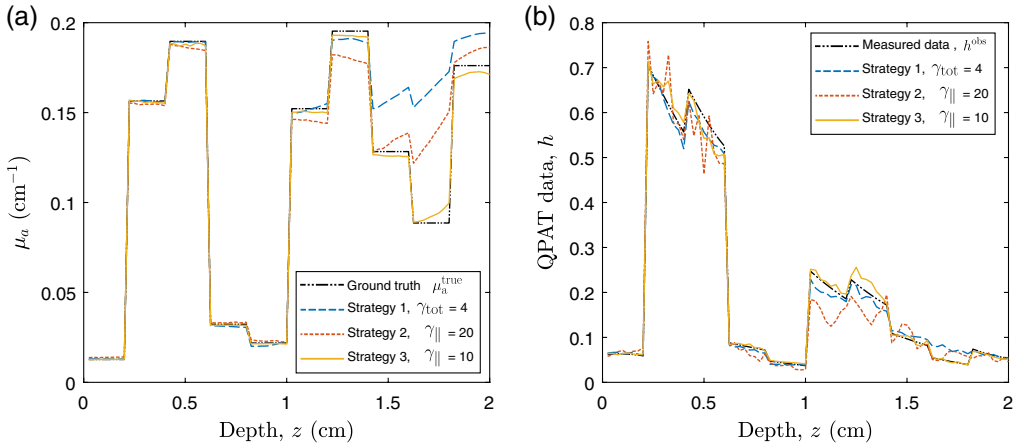

Fig. 3 QPAT inversion: (a) ground-truth absorption distribution, μ_a^{true}, and estimated absorption distribution, μ_a, at the point where the photon budget is expended, using each of the three strategies with the stated values of γ_{tot} or $\gamma_{\|}$. (b) Associated measured data from ground-truth medium and simulated forward data at the end of the inversion using each strategy.

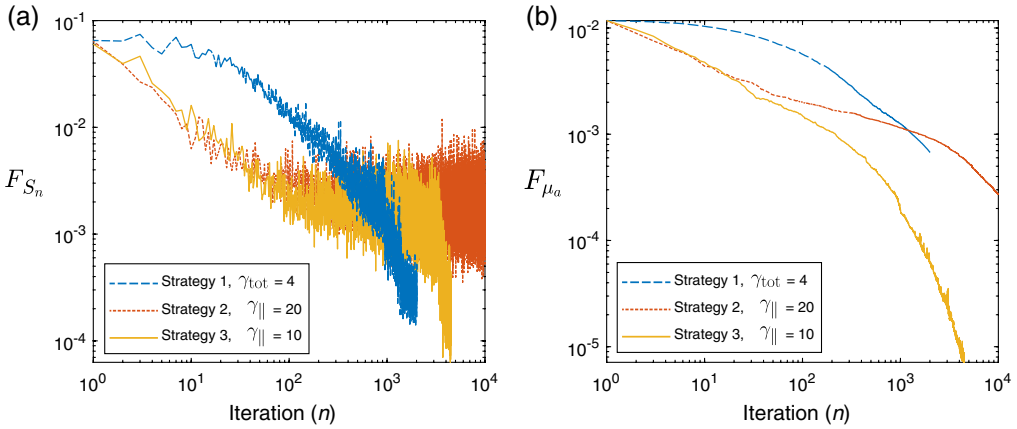

Fig. 4 QPAT inversion: (a) sampled cost function, F_{S_n}, as a function of iteration, n. (b) Error in absorption estimate, F_{μ_a}, as a function of iteration, n.

of the strategies. Figure 4 shows the outcome of each strategy in terms of the sampled data cost function F_{S_n} and the absorption error function F_{μ_a}. It can be seen that the ranking of these methods in terms of the lowest achieved value of the sampled cost function F_{S_n} does not correlate directly to the best outcomes in terms of the error in the estimated absorption F_{μ_a}. This is due to the above-mentioned dependence of the sampled data cost function on the sample size used in the forward model, where for example the case of strategy 2 only appears to perform poorly in terms of F_{S_n} due to its small sample size used throughout the inversion. This is more clearly shown in Fig. 3(b), where the final forward modeled data from strategy 2 are noisier than the other strategies due to the low sample size at the end of the inversion, where this noise would clearly impact the sampled cost function. It is worthy to note that the relevant step sizes and sample sizes for each of these three examples are shown in Fig. 5. Before finding the best parameter for strategy 1, we trialed a range of values of γ_{tot} over the range (0.1, 200). With lower values, the adaptive sample size was required to increase rapidly to maintain high-quality (low variance) sample gradients. This resulted in the photon budget being depleted early, terminating the descent after around 100 iterations, which did not perform well. Too large a value of $\gamma_{\|}$ and the norm test never failed, meaning the sample size was never required to increase and the inversion progressed for the maximum 10,000 iterations permitted by the photon budget. However, as strategy 1 has a fixed value of γ_{tot}^2 in the denominator of the step size, large values also result in

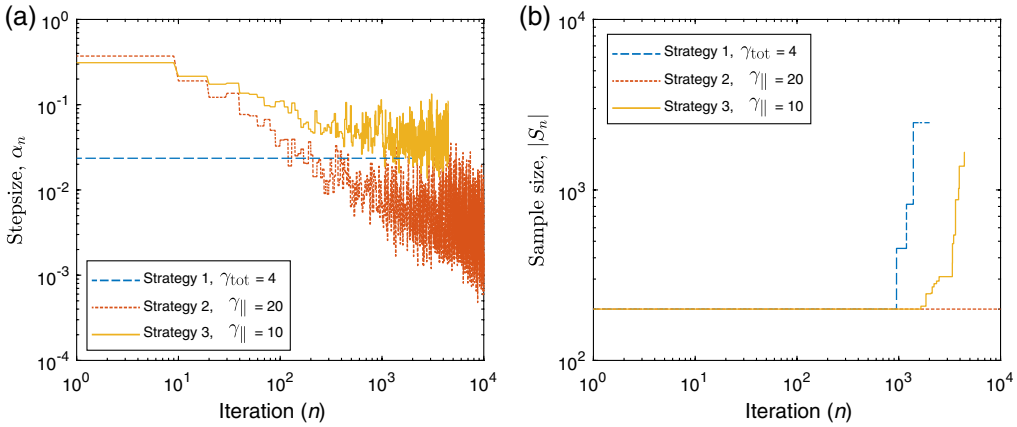

Fig. 5 QPAT inversion: (a) step sizes, α_n, as a function of iteration, n. (b) Adaptive sample size, $|S_n|$, as a function of iteration.

step sizes that were too small to perform well. A value of $\gamma_{\text{tot}} = 4$ was found to strike a balance between these two extremes and was the best performer using strategy 1. Strategy 2 has an adaptive step size which selects the largest possible step size that still satisfies Eq. (11), instead of selecting a constant step size that accounts for the worst case scenario, as in strategy 1. For this reason, we found that the largest value of $\gamma_{\|} = 20$ was the best performer for this strategy, where the photon budget remained at 200 photons for each of the 10,000 iterations. For strategy 3, the best performer was a value of $\gamma_{\|} = 10$, where larger values appeared to allow too much variance in the gradient, leading to unstable descents. In all strategies, the recovered absorption distribution matched the ground-truth absorption more closely in the regions of the sample closest to the light source at $z = 0$. This is due to the decay of the fluence as a function of depth, as we can see the QPAT signal is highest at shallow depths in Fig. 3(b). The deeper regions of the sample were the last to approach the ground truth in each of the three strategies.

From Fig. 6, we see the values of our two metrics $V_{\|}$ and V_{tot}. In all cases, both measures of the variance begin at low values, indicating that even with low numbers of photons being simulated, the computed gradients are of reasonable quality, likely due to the poor initial first guess being far from the true solution. Each of the measures of variance increase as the inversion progresses until they begin to violate the norm test or inner product test depending on the strategy. It is seen that the strategy 1 example attempts to keep $V_{\text{tot}} \leq 4$, however, due to some level of variation in the metrics themselves, this condition can be seen to be violated regularly, requiring regular updates to the sample size. For strategy 2, the imposed limit of $V_{\|} \leq 20$ is never violated,

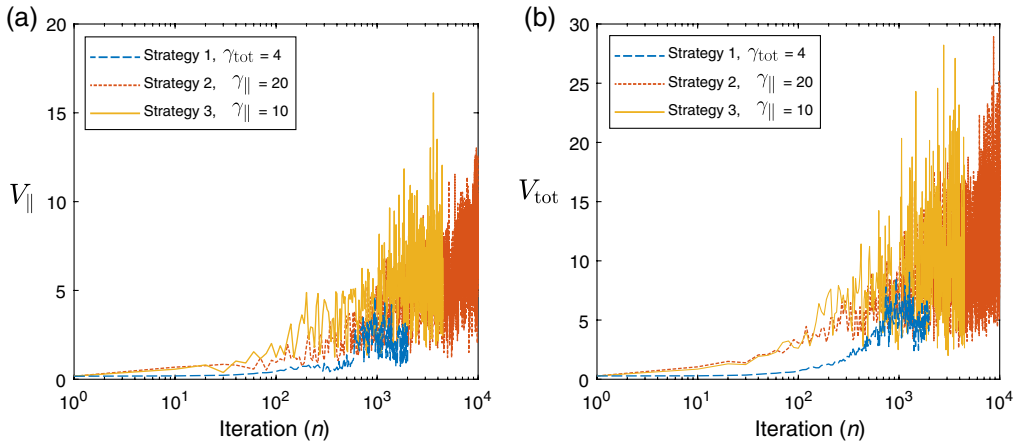

Fig. 6 QPAT inversion: (a) $V_{\|}$ as a function of iteration and (b) V_{tot} as a function of iteration.

Table 2 Final outcomes of QPAT inversions with various medium optical properties and starting values of μ_a. Values of F_{S_n} and F_{μ_a} are the final values at the end of each inversion after the stated number of iterations. In each case, strategy 3 was employed, with a starting sample size of $|S_1| = 200$ photons per iteration, and a total photon budget of $N_{ph} = 2 \times 10^6$ photons. Slab thickness is 2 cm in all cases, with the same ground-truth μ_a^{true} distribution as shown in Fig. 3(a).

		Starting μ_a (cm^{-1})		
		0.01	0.2	1.0
Medium properties	$g = 0.9$ $\mu_s = 40$ cm^{-1}	$\gamma_\parallel = 20$, 10,000 iterations $F_{S_n} = 2.35 \times 10^{-3}$ $F_{\mu_a} = 3.26 \times 10^{-5}$	$\gamma_\parallel = 10$, 4476 iterations $F_{S_n} = 4.20 \times 10^{-4}$ $F_{\mu_a} = 7.65 \times 10^{-6}$	$\gamma_\parallel = 10$, 6533 iterations $F_{S_n} = 3.38 \times 10^{-4}$ $F_{\mu_a} = 1.01 \times 10^{-5}$
	$g = 0.9$ $\mu_s = 4$ cm^{-1}	$\gamma_\parallel = 5$, 3819 iterations $F_{S_n} = 5.30 \times 10^{-4}$ $F_{\mu_a} = 2.3 \times 10^{-7}$	$\gamma_\parallel = 5$, 2579 iterations $F_{S_n} = 9.31 \times 10^{-5}$ $F_{\mu_a} = 3.56 \times 10^{-7}$	$\gamma_\parallel = 5$, 2834 iterations $F_{S_n} = 1.06 \times 10^{-4}$ $F_{\mu_a} = 2.19 \times 10^{-7}$
	$g = 0$ $\mu_s = 4$ cm^{-1}	$\gamma_\parallel = 20$, 10,000 iterations $F_{S_n} = 4.01 \times 10^{-3}$ $F_{\mu_a} = 8.36 \times 10^{-5}$	$\gamma_\parallel = 5$, 2056 iterations $F_{S_n} = 1.39 \times 10^{-4}$ $F_{\mu_a} = 3.44 \times 10^{-5}$	$\gamma_\parallel = 10$, 10,000 iterations $F_{S_n} = 2.92 \times 10^{-3}$ $F_{\mu_a} = 8.72 \times 10^{-5}$

and thus the sample size is never required to increased. We also see that strategy 3 manages to keep $V_\parallel \leq 10$ for the majority of the descent.

In addition to these experiments shown in Figs. 3–6, we also trialed a number of other conditions including media with isotropic scatterers (i.e., with $g = 0$), various scattering coefficients, and various initial estimates of the absorption. In all cases explored, the methods showed similar behavior as above, but with some differences in the ideal values of γ_{tot} and $\gamma_\parallel$ for each strategy. The outcomes of a range of these experiments are shown in Table 2 for various problem parameters. Strategy 3 was used in all cases in the table, with the same Lipschitz constant ($L = 2.5$), starting sample size ($|S_1| = 200$ photons), photon budget ($N_{ph} = 2 \times 10^6$ photons), and ground-truth absorption distribution μ_a^{true} as used in the above examples. The final attained values of the sampled data cost function F_{S_n} and absorption error F_{μ_a} are similar in all cases with the exception of the high asymmetry and low scattering case ($g = 0.9$ and $\mu_s = 4$ cm^{-1}). In this case, the reduced scattering coefficient is only $\mu_s' = \mu_s(1 - g) = 0.4$ cm^{-1}, meaning much lower overall attenuation of the light through the sample. This results in a more uniform data function, h, where the simulated photons probe the domain more uniformly, and allows the problem to converge significantly faster than in the higher attenuating cases demonstrated in Figs. 3–6. It is also worth noting that in the regime with low scattering and high scattering asymmetry, it is generally problematic for the performance of approximate transport models such as the diffusion approximation, and the results here highlight the flexibility of RTE based approaches, and the efficiency of the proposed adaptive sampling techniques.

Finally, interesting behavior was observed when using certain initial guesses of the absorption. An example of this is shown in Fig. 7, where we show the resulting cost functions for a starting estimate of $\mu_a = 1$ cm^{-1} (significantly overestimating the absorption at all depths), and medium properties of $g = 0.9$ and $\mu_s = 40$ cm^{-1}. In this case, we see that the descent appears to encounter local minima in the data cost function F_{S_n} at various points during the descent, depending on the particular strategy used. However, the algorithm manages to escape these local minima and converge to a better solution. This is seen to be the case for all three strategies shown in Fig. 7.

4.2 *UMOT*

Next, we demonstrate similar experiments performed using the UMOT modality described in Sec. 3.2.2 for the transmission geometry. In this setup, we used the same medium slab size as the QPAT example, and the same optical properties apart from the absorption distribution. The starting sample size in all cases shown is $|S_1| = 4000$ photons per iteration, 1000 for each forward

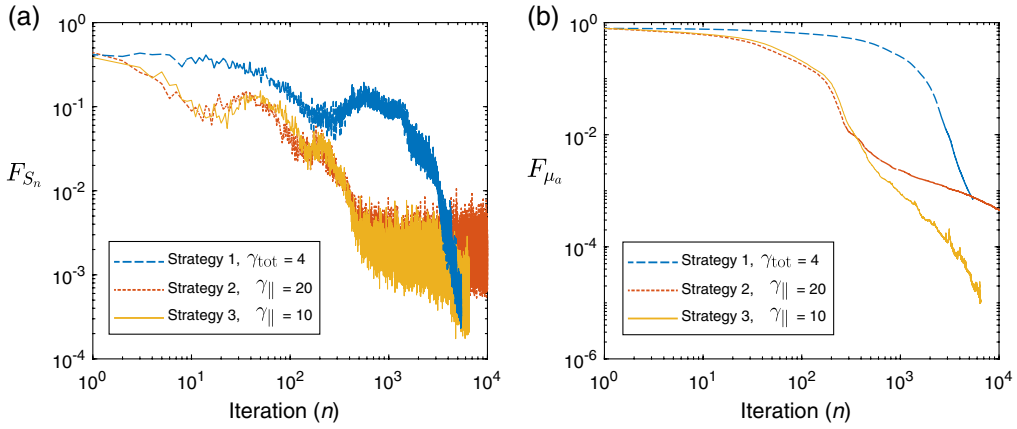

Fig. 7 QPAT inversion: with initial estimate of $\mu_a = 1.0$ cm^{-1}: (a) sampled cost function, F_{S_n}, as a function of iteration, n. (b) Error in absorption estimate, F_{μ_a}, as a function of iteration, n.

run (per each of the two sources), and 1000 for each of the two adjoint sources as outlined in Algorithm 3. The total photon budget for the inversion was set to $N_{\text{ph}} = 4 \times 10^8$ photons. The Lipschitz constant was set at $L = 50$, as this displayed stable descent in our test problems using large photon budgets (low-variance case). The initial estimate of the absorption distribution in the medium is $\mu_a = 0.1$ cm^{-1} in all layers. The ground-truth absorption μ_a^{true} is shown in Fig. 8(a), along with the final retrieved absorption distributions obtained via strategies 1, 2, and 3 using the stated values of γ_{tot} and $\gamma_{\|}$. Figure 8(b) shows the true measured UMOT data, b^{obs}, along with the forward modeled data from the final estimated medium for each strategy. Figure 9 shows the outcome of each strategy in terms of the sampled data cost function, F_{S_n}, and the absorption error function, F_{μ_a}. The relevant step sizes and sample sizes for each of these three examples are shown in Fig. 10, and the values of the metrics measuring the variance in the sampled gradients are presented in Fig. 11. Similar to the QPAT modality, we found that strategy 3 performed the best in terms of the final achieved value of the error in the absorption estimate F_{μ_a}.

In addition to the results shown in Figs. 8–11, in Table 3 we present a summary of results for a range of different medium optical parameters and starting estimates of the absorption. In all cases, strategy 3 was used, and the starting photon budget was the same as in the previous UMOT examples ($|S_1| = 4000$ photons), with a total photon budget of $N_{\text{ph}} = 4 \times 10^8$ photons. For each

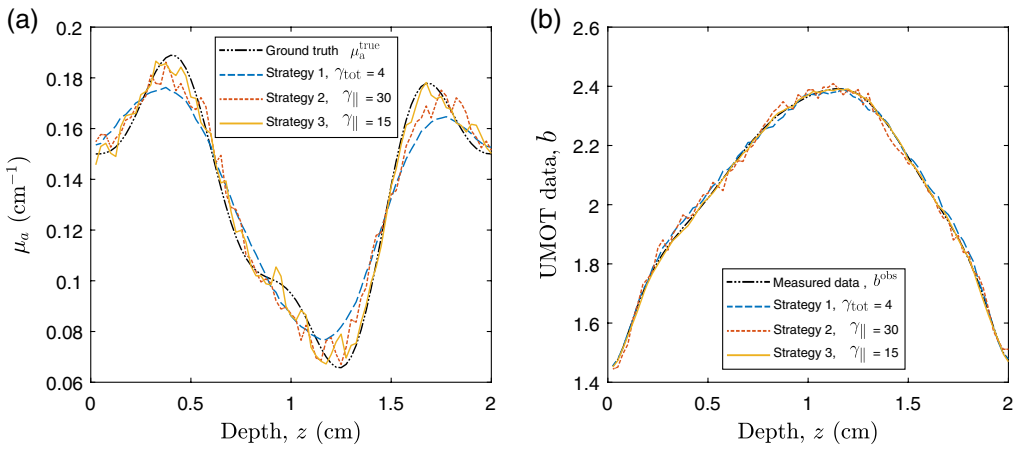

Fig. 8 UMOT inversion: (a) ground-truth absorption distribution, μ_a^{true}, and recovered absorption distribution μ_a using each of the three strategies with the stated values of γ_{tot} or $\gamma_{\|}$. (b) Associated measured data from ground-truth medium and simulated forward data at the end of the inversion using each strategy.

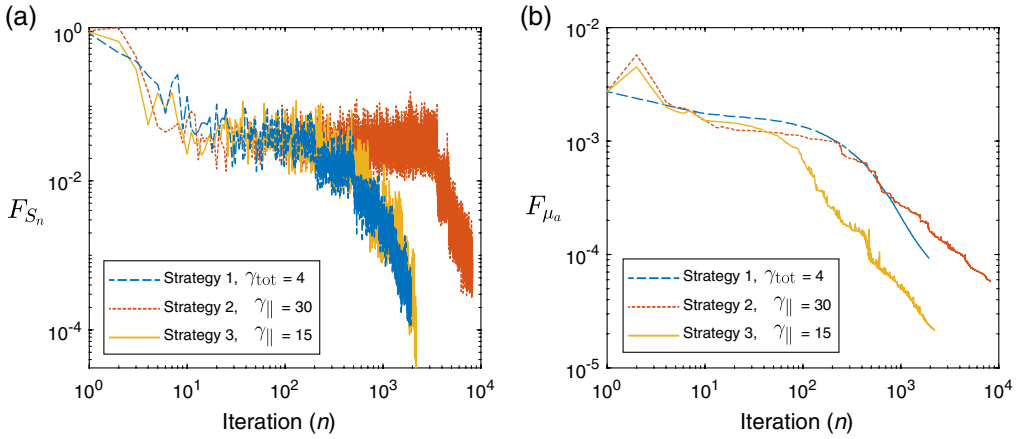

Fig. 9 UMOT inversion: (a) sampled cost function, F_{S_n}, as a function of iteration, n. (b) Error in absorption estimate, F_{μ_a}, as a function of iteration, n.

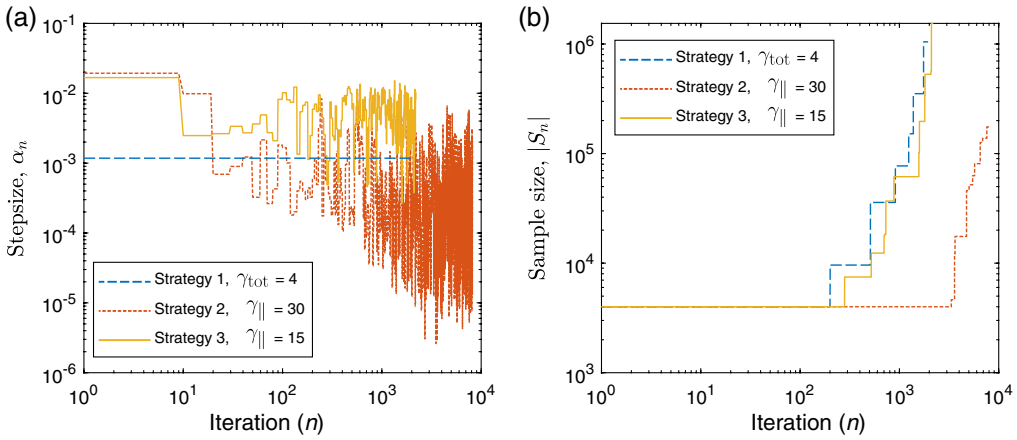

Fig. 10 UMOT inversion: (a) step sizes, α_n, as a function of iteration, n. (b) Adaptive sample size, $|S_n|$, as a function of iteration.

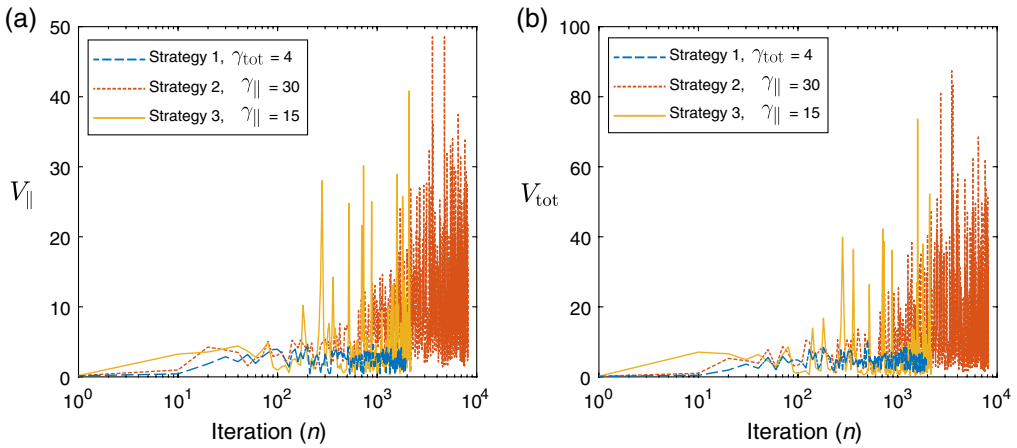

Fig. 11 UMOT inversion: (a) $V_{\|}$ as a function of iteration and (b) V_{tot} as a function of iteration.

Table 3 Final outcomes of UMOT inversions with various medium optical properties and starting values of μ_a. Values of F_{S_n} and F_{μ_a} are the final values at the end of each inversion after the stated number of iterations. In each case, strategy 3 was employed, with a starting sample size of $|S_1| = 4000$ photons per iteration, and a total photon budget of $N_{ph} = 4 \times 10^8$ photons. Slab thickness is 2 cm in all cases, with the same ground-truth μ_a^{true} distribution as shown in Fig. 8(a).

		Starting μ_a (cm^{-1})		
		0.01	0.1	1.0
Medium properties	$g = 0.9$ $\mu_s = 40$ cm^{-1}	$\gamma_\parallel = 15$, 11,412 iterations $F_{S_n} = 1.88 \times 10^{-3}$ $F_{\mu_a} = 4.23 \times 10^{-5}$	$\gamma_\parallel = 10$, 3495 iterations $F_{S_n} = 2.50 \times 10^{-4}$ $F_{\mu_a} = 2.20 \times 10^{-5}$	$\gamma_\parallel = 10$, 7823 iterations $F_{S_n} = 5.69 \times 10^{-4}$ $F_{\mu_a} = 9.37 \times 10^{-5}$
	$g = 0.9$ $\mu_s = 4$ cm^{-1}	$\gamma_\parallel = 15$, 19,017 iterations $F_{S_n} = 4.71 \times 10^{-3}$ $F_{\mu_a} = 8.19 \times 10^{-5}$	$\gamma_\parallel = 15$, 15,813 iterations $F_{S_n} = 5.48 \times 10^{-3}$ $F_{\mu_a} = 2.07 \times 10^{-5}$	$\gamma_\parallel = 15$, 14,249 iterations $F_{S_n} = 4.08 \times 10^{-3}$ $F_{\mu_a} = 3.36 \times 10^{-5}$
	$g = 0$ $\mu_s = 4$ cm^{-1}	$\gamma_\parallel = 15$, 11,594 iterations $F_{S_n} = 1.65 \times 10^{-3}$ $F_{\mu_a} = 7.13 \times 10^{-5}$	$\gamma_\parallel = 15$, 7304 iterations $F_{S_n} = 5.87 \times 10^{-4}$ $F_{\mu_a} = 4.17 \times 10^{-5}$	$\gamma_\parallel = 15$, 13,981 iterations $F_{S_n} = 1.05 \times 10^{-3}$ $F_{\mu_a} = 7.98 \times 10^{-5}$

of the inversions presented in this table, we conducted the inner product test once every 50 iterations, using $N_{rep} = 50$ repeated evaluations of the gradient. The resulting inversions display similar error in these cases to the above examples where we used $N_{rep} = 100$ repeated evaluations of the sampled gradient once every 10 iterations to run the inner product test. This demonstrates that the described methods can still be successful when dedicating fewer computational resources to the inner product or norm test metrics, which control the adaptive sample size and step size.

5 Discussion and Conclusions

The results shown in Sec. 4 demonstrate that the adaptive sampling strategies performed well in both our example problems of QPAT and UMOT. We were able to achieve low error estimates of the medium absorption using a total computational expenditure that was either comparable to or significantly lower than the resources required to simulate a single low-variance run of the forward problem. In each demonstration, the adaptive sampling strategies maintained low photon numbers throughout the early stages of the inversion. Photon numbers were only increased when required to keep the variance in the gradients below the stipulated limits. These adaptive sampling strategies thus enabled significant computational savings compared to a naïve implementation, which might seek to use low-variance (high quality) computations of the gradient at every iteration. For instance, if we were to use a constant step size of $1/L$ and the same number of photons per iteration as that which was used to generate the measured data (10^9 photons), then we find we still required hundreds of iterations to reach a similar quality estimate of the absorption as seen in the above problems. This means that the computational requirements of the low-variance approach would be proportional to $N_{ph} = 10^{11}$ photons. Comparing this to $N_{ph} = 2 \times 10^6$ photons used in the QPAT examples or $N_{ph} = 4 \times 10^8$ photons used in the UMOT examples, the required computational resources/time to attain our solutions with these adaptive sampling methods is multiple orders of magnitude lower compared to the naïve low-variance approach.

In this work, we have emphasized the similarities between our approach and that of SGD, as employed in the context of machine learning. However, it should be noted that there are significant differences between the two settings. In machine learning, the measured data are assumed to consist of a large number of samples to be fit to a deterministic model to minimize a suitable loss function, and each stochastic gradient is generated by a random subset of these data forming the descent direction of a subfunction. The same method has also been applied in

alternative image reconstruction techniques where the data can be more naturally considered as consisting of a large number of random samples from some underlying distribution, for example, in positron emission tomography.[47] By contrast, our image reconstruction approach considers the complete measured data on each iteration, with stochasticity arising from the approximation within the forward model: we are effectively subsampling the gradient in terms of the parameter space, rather than data space. This is to say that at each iteration we utilize a subset of some notionally complete model, rather than of the data. The motivation by which each approach is employed is consistent: stochasticity is intentionally introduced to whichever part of the objective function introduces the greatest computational demand. This suggests a third possible approach, where the computational load of the (sub) gradient computation can be lowered through some stochastic division of both the data and the model; this might be relevant in imaging modalities with discrete counting data, such as time-domain and/or dynamic diffuse optical tomography.

Our work suggests a number of interesting future developments:

- In the examples shown here the "observed" data were effectively "noise free" by virtue of running the forward Monte Carlo on a very large number of photons. Thus an interesting topic for further study will be to evaluate these methods on noisy forward data, wherein the data fitting term should not be iterated to convergence, but where regularization should be introduced either by early stopping (i.e., by setting a minimum threshold for the data error) or by adding an explicit penalty term.
- Related to the previous point, we further note that our objective function employed a least-squares data fitting term in this study. Depending upon the nature of the noise in the data and that of the stochastic forward model, more suitable metrics may include the Kullback–Leibler discrepancy (for Poisson likelihood) or a generalized measure of the distance between samples of probability distributions (Wasserstein distance[48]).
- Our results demonstrate a consistent tendency for the adaptive sampling method to exhibit a geometric increase in the sample size as the descent progresses. This suggests that our adaptive approach could be employed to find a particular set of sampling parameters that perform well in a given regime, including the starting photon budget $|S_1|$, rate of increase of the sample size $\kappa(n)$, and rate of change of the step size α_n. If a suitable set of such parameters could be found, they could help determine a fully prescribed sampling strategy. Once calibrated for a given problem of choice, this would avoid the need to explicitly compute the variance of the sampled gradients during the descent, and lead to even greater efficiency and speed in the inverse problem.
- Further topics of interest include more advanced methods of variance reduction (e.g., recursive gradients[49]); adaptive estimates of the Lipschitz constant as described in Ref. 30; alternative optimization strategies such as back-tracking line-search, or primal dual methods;[50] the use of preconditioning and/or second-order optimization methods;[51] and an in-depth comparison of these nonlinear adaptive models to the alternative approaches such as PMC.[27]

In summary, we have successfully demonstrated a means by which stochastic forward models, not directly amenable to standard variational methods for optimization, can be employed efficiently in nonlinear image reconstruction. We expect this concept to lead to many new directions of research in optical image reconstruction.

Disclosures

No conflicts of interest, financial or otherwise, are declared by the authors.

Acknowledgments

This work was funded by the Engineering and Physical Sciences Research Council under Grant Nos. EP/N032055/1 and EP/N025946/1. S. Powell further acknowledges the Royal Academy of

Engineering fellowship, RF1516/15/33. The authors would like to thank Robert Twyman and Kris Thielemans for helpful discussions on stochastic gradient and data subset methods.

References

1. S. R. Arridge and J. Schotland, "Optical tomography: forward and inverse problems," *Inverse Prob.* **25**(12), 123010 (2009).
2. L. T. Biegler et al., "Large-scale PDE-constrained optimization: an introduction," in *Large-Scale PDE-Constrained Optimization*, L. T. Biegler et al., Eds., pp. 3–13, Springer, Berlin, Heidelberg (2003).
3. K. Bredies et al., *Control and Optimization with PDE Constraints*, Springer, Basel (2013).
4. J. C. De los Reyes, *Numerical PDE-Constrained Optimization*, Springer Briefs in Optimization, Springer, Heidelberg (2015).
5. C. Zhu and Q. Liu, "Review of Monte Carlo modeling of light transport in tissues," *J. Biomed. Opt.* **18**(5), 050902 (2013).
6. I. Lux and L. Koblinger, *Monte Carlo Particle Transport Methods: Neutron and Photon Calculations*, CRC Press, Boca Raton (1991).
7. A. Liemert and A. Kienle, "Analytical solution of the radiative transfer equation for infinite-space fluence," *Phys. Rev. A* **83**, 015804 (2011). Erratum Phys. Rev. A 83, 039903 (2011).
8. A. Liemert, D. Reitzle, and A. Kienle, "Analytical solutions of the radiative transport equation for turbid and fluorescent layered media," *Sci. Rep.* **7**, 3819 (2017).
9. W. C. Lo et al., "Hardware acceleration of a Monte Carlo simulation for photodynamic treatment planning," *J. Biomed. Opt.* **14**(1), 014019 (2009).
10. N. Ren et al., "GPU-based Monte Carlo simulation for light propagation in complex heterogeneous tissues," *Opt. Express* **18**(7), 6811–6823 (2010).
11. C. K. Hayakawa et al., "Perturbation Monte Carlo methods to solve inverse photon migration problems in heterogeneous tissues," *Opt. Lett.* **26**(17), 1335–1337 (2001).
12. R. Graaff et al., "Condensed Monte Carlo simulations for the description of light transport," *Appl. Opt.* **32**(4), 426–434 (1993).
13. I. T. Lima, A. Kalra, and S. S. Sherif, "Improved importance sampling for Monte Carlo simulation of time-domain optical coherence tomography," *Biomed. Opt. Express* **2**(5), 1069–1081 (2011).
14. H. Ammari, *Mathematical Modeling in Biomedical Imaging II: Optical, Ultrasound, and Opto-Acoustic Tomographies*, Springer, Heidelberg (2011).
15. S. R. Arridge and O. Scherzer, "Imaging from coupled physics," *Inverse Prob.* **28**(8), 080201 (2012).
16. V. A. Morozov, "On the solution of functional equations by the method of regularization," *Sov. Math. Doklady* **7**, 414–417 (1966).
17. R. Johnson and T. Zhang, "Accelerating stochastic gradient descent using predictive variance reduction," in *Adv. Neural Inf. Process. Syst.*, pp. 315–323 (2013).
18. M. Friebel et al., "Determination of optical properties of human blood in the spectral range 250 to 1100 nm using Monte Carlo simulations with hematocrit-dependent effective scattering phase functions," *J. Biomed. Opt.* **11**(3), 034021 (2006).
19. I. Yaroslavsky et al., "Inverse hybrid technique for determining the optical properties of turbid media from integrating-sphere measurements," *Appl. Opt.* **35**(34), 6797–6809 (1996).
20. Q. Fang and D. Boas, "Monte Carlo simulation of photon migration in 3D turbid media accelerated by graphics processing units," *Opt. Express* **17**(22), 20178–20190 (2009).
21. Q. Fang, "Mesh-based Monte Carlo method using fast ray-tracing in Plücker coordinates," *Biomed. Opt. Express* **1**, 165 (2010).
22. J. Buchmann et al., "Three-dimensional quantitative photoacoustic tomography using an adjoint radiance Monte Carlo model and gradient descent," *J. Biomed. Opt.* **24**(6), 066001 (2019).

23. R. Hochuli et al., "Quantitative photoacoustic tomography using forward and adjoint Monte Carlo models of radiance," *J. Biomed. Opt.* **21**(12), 126004 (2016).
24. A. A. Leino, A. Pulkkinen, and T. Tarvainen, "ValoMC: a Monte Carlo software and MATLAB toolbox for simulating light transport in biological tissue," *OSA Continuum* **2**(3), 957–972 (2019).
25. U. Tricoli et al., "Reciprocity relation for the vector radiative transport equation and its application to diffuse optical tomography with polarized light," *Opt. Lett.* **42**(2), 362–365 (2017).
26. R. Yao, X. Intes, and Q. Fang, "A direct approach to compute Jacobians for diffuse optical tomography using perturbation Monte Carlo-based photon 'replay'," *Biomed. Opt. Express* **9**(10), 4588–4603 (2018).
27. A. Leino et al., "Perturbation Monte Carlo method for quantitative photoacoustic tomography," *IEEE Trans. Med. Imaging* (2020).
28. L. Bottou, F. E. Curtis, and J. Nocedal, "Optimization methods for large-scale machine learning," *SIAM Rev.* **60**(2), 223–311 (2018).
29. D. Newton, F. Yousefian, and R. Pasupathy, "Stochastic gradient descent: recent trends," in *Recent Advances in Optimization and Modeling of Contemporary Problems*, pp. 193–220, INFORMS, Catonsville, Maryland (2018).
30. R. Bollapragada, R. Byrd, and J. Nocedal, "Adaptive sampling strategies for stochastic optimization," *SIAM J. Optim.* **28**(4), 3312–3343 (2018).
31. H. Robbins and S. Monro, "A stochastic approximation method," *Ann. Math. Stat.* **22**, 400–407 (1951).
32. Y. Nesterov, *Introductory Lectures on Convex Optimization: A Basic Course*, Springer Science & Business Media, New York (2013).
33. S. R. Arridge, "Optical tomography in medical imaging," *Inverse Prob.* **15**(2), R41–R93 (1999).
34. A. Erik, S. Tomas, and A. Stefan, "CUDAMCML program," http://www.atomic.physics.lu.se/biophotonics/research/monte-carlo-simulations/gpu- monte-carlo (2009).
35. L. Wang, S. L. Jacques, and L. Zheng, "MCML—Monte Carlo modeling of light transport in multi-layered tissues," *Comput. Methods Programs Biomed.* **47**(2), 131–146 (1995).
36. G. Bal, "Inverse transport theory and applications," *Inverse Prob.* **25**(5), 053001 (2009).
37. G. Bal, "Hybrid inverse problems and internal functionals," in Inverse Problems and Applications: Inside Out II, Vol. **60**, pp. 325–368 (2013).
38. L. V. Wang, "Multiscale photoacoustic microscopy and computed tomography," *Nat. Photonics* **3**, 503–509 (2009).
39. P. Beard, "Biomedical photoacoustic imaging," *Interface Focus* **1**, 602–631 (2011).
40. L. Nie and X. Chen, "Structural and functional photoacoustic molecular tomography aided by emerging contrast agents," *Chem. Soc. Rev.* **43**(20), 7132–7170 (2014).
41. B. Cox et al., "Quantitative spectroscopic photoacoustic imaging: a review," *J. Biomed. Opt.* **17**(6), 061202 (2012).
42. T. Saratoon et al., "A gradient-based method for quantitative photoacoustic tomography using the radiative transfer equation," *Inverse Prob.* **29**, 075006 (2013).
43. K. Wang and M. Anastasio, "Photoacoustic and thermoacoustic tomography: image formation principles," in *Handbook of Mathematical Methods in Imaging*, O. Scherzer, Ed., pp. 781–815, Springer, New York (2011).
44. H. Ammari et al., "A reconstruction algorithm for ultrasound-modulated diffuse optical tomography," *Proc. Am. Math. Soc.* **142**(9), 3221–3236 (2014).
45. F. J. Chung and J. C. Schotland, "Inverse transport and acousto-optic imaging," *SIAM J. Math. Anal.* **49**, 4704–4721 (2017).
46. S. Powell, S. R. Arridge, and T. S. Leung, "Gradient-based quantitative image reconstruction in ultrasound-modulated optical tomography: first harmonic measurement type in a linearised diffusion formulation," *IEEE Trans. Med. Imaging* **35**(2), 456–467 (2016).
47. K. Thielemans and S. Arridge, "Adaptive adjustment of the number of subsets during iterative image reconstruction," in *IEEE Nuclear Sci. Symp. and Med. Imaging Conf.*, pp. 1–2, IEEE (2016).

48. C. Villani, *Optimal Transport: Old and New*, Grundlehren der mathematischen Wissenschaften, Vol. **338**, Springer, Berlin (2009).
49. L. Nguyen et al., "SARAH: a novel method for machine learning problems using stochastic recursive gradient," in *Int. Conf. Mach. Learn.*, p. 34 (2017).
50. A. Chambolle et al., "Stochastic primal-dual hybrid gradient algorithm with arbitrary sampling and imaging applications," *SIAM J. Optim.* **28**(4), 2783–2808 (2018).
51. P. Moritz, R. Nishihara, and M. I. Jordan, "A linearly-convergent stochastic L-BFGS algorithm," in *Proc. 19th Int. Conf. Artif. Intell. and Stat.*, Vol. 41, pp. 249–258 (2016).

Biographies of the authors are not available.