Online learning in optical tomography: a stochastic approach

IOP Publishing Inverse Problems

Inverse Problems **34** (2018) 075010 (26pp) https://doi.org/10.1088/1361-6420/aac220

Online learning in optical tomography: a stochastic approach

Ke Chen[1], **Qin Li**[2] **and Jian-Guo Liu**[3]

[1] Mathematics Department, University of Wisconsin-Madison, 480 Lincoln Dr., Madison, WI 53705, United States of America
[2] Mathematics Department and Wisconsin Institutes for Discovery, University of Wisconsin-Madison, 480 Lincoln Dr., Madison, WI 53705, United States of America
[3] Department of Mathematics and Department of Physics, Duke University, Durham, NC 27708, United States of America

E-mail: ke@math.wisc.edu, qinli@math.wisc.edu and jliu@phy.duke.edu

Received 27 October 2017, revised 2 April 2018
Accepted for publication 3 May 2018
Published 29 May 2018

Abstract

We study the inverse problem of radiative transfer equation (RTE) using stochastic gradient descent method (SGD) in this paper. Mathematically, optical tomography amounts to recovering the optical parameters in RTE using the incoming–outgoing pair of light intensity. We formulate it as a PDE-constraint optimization problem, where the mismatch of computed and measured outgoing data is minimized with same initial data and RTE constraint. The memory and computation cost it requires, however, is typically prohibitive, especially in high dimensional space. Smart iterative solvers that only use partial information in each step is called for thereafter. Stochastic gradient descent method is an online learning algorithm that randomly selects data for minimizing the mismatch. It requires minimum memory and computation, and advances fast, therefore perfectly serves the purpose. In this paper we formulate the problem, in both nonlinear and its linearized setting, apply SGD algorithm and analyze the convergence performance.

Keywords: stochastic gradient descent, radiative transfer equation, optical tomography, online learning

(Some figures may appear in colour only in the online journal)

1. Introduction

Optical tomography is a form of computed tomography that extracts tomographic images of objects to be studied using information of light transmitted and scattered through it. It has been vastly used in many applications: in medical imaging near infrared light (NIR) is sent

into biological tissues for tumor or bone structure [23, 24]; in outer space studies: during Galileo's travel around Jupiter, pictures are taken by the near infrared mapping spectrometer (NIMS), and scientists recover components of atmosphere on each satellite [11]. Typically scientists inject a certain amount of light into a bulk of material, and measure the outgoing light intensities at the boundaries. By collecting many such incoming and outgoing light intensity pairs, scientists infer for the optical information of the material.

Mathematically, light is typically characterized by the radiative transfer equation (RTE). It characterizes photon particles that scatter and get absorbed in materials with various optical properties. Optical tomography, therefore is formulated as the inverse problem of the radiative transfer equation. The equation reads:

$$v \cdot \nabla_x f + \sigma(x) f = \int_{\mathbb{V}} k(x, v, v') f(x, v') \mathrm{d}v', \tag{1}$$

where $f(x, v)$, defined on phase space, is the distribution of particles at location x with velocity v. Here $x \in \Omega \subset \mathbb{R}^d$ with $d = 2, 3$, and $v \in \mathbb{V} = \mathbb{S}^{d-1}$, the unit sphere in $\mathbb{R}^d$. $k(x, v, v')$ is the scattering coefficient and it shows the probability of particles moving in direction v' changing to direction v at location x, and $\sigma(x)$ is the total absorption coefficient that represents certain amount of photon particles being absorbed by the material. The equation has a unique solution with the following boundary condition:

$$f|_{\Gamma_-} = \phi(x, v), \tag{2}$$

where Γ_- collects the coordinates on $\partial\Omega$ with incoming velocities (and Γ_+ collects the outgoings):

$$\Gamma_\pm = \{(x, v) : x \in \partial\Omega, \pm v \cdot n_x > 0\}.$$

Here n_x stands for the normal direction pointing out of the domain at point $x \in \partial\Omega$. The wellposedness of the equation in the general L_p space has been studied in [12]. Define the albedo operator that maps the incoming boundary condition to the outgoing data:

$$\mathcal{A} : \phi(x, v) \to (n_x \cdot v)\, f(x, v)|_{\Gamma_+}. \tag{3}$$

In the forward problem setting, the optical properties σ and k are known and one computes $(n_x \cdot v)\, f|_{\Gamma_+}$ for arbitrarily given ϕ. In the inverse setting, one obtains all possible $(\phi, (n_x \cdot v)\, f|_{\Gamma_+})$ pairs and uses them ($\mathcal{A}$ information) to recover σ and k. Note there are multiple ways to define $\mathcal{A}$ depending on the measurements. For example, $\mathcal{A}$ could map ϕ to $f|_{\Gamma_+}$ or the angular averaged measurements $\int (n_x \cdot v) f \mathrm{d}v$.

The problem, due to its large application, has been extensively studied from many aspects. On the analytical side people concern the wellposedness and the stability. More precisely, we ask: 1. Does $\mathcal{A}$ contain enough information to extract all coefficients; 2. How sensitive the recovery is towards the measurements. The first question was initially addressed by a pioneer paper in [10], in which the authors used the singular decomposition technique to prove the uniqueness of the recovery in 3D if σ has no v dependence. This technique was later extended to study angular average data [3, 4] and the case where σ has the v dependence [35]. The second question was looked at as early as in [39], and the bad conditioning was addressed by increasing the modulation frequency in the time-harmonic case [5]. In [8], the authors studied the stability's dependence on the Knudsen number and recover, to some extent, the ill-conditioning of the Calderón type problem in the diffusion limit. See [2] for a review.

On the computation side, different application setups provide different types of measurements, and it drives the development of various numerical techniques [9, 13, 18, 29, 32, 37, 38]. A very general descriptions are found in influential books [14, 26]. Generally speaking

people regard it as an optimization problem with PDE constraints. More precisely, one tries to minimize the mismatch between the measurements and the numerical results assuming the RTE is satisfied. In this process, L_2, L_1 or TV norm of the coefficients are added as penalties to fit certain a prior knowledge. The biggest challenge here, of course, is the size of the problem: in every iteration a forward solver is called, and this deals with the distribution function f that lives on phase space and has N^5 degrees of freedom in 3D (assuming each direction takes N points). Some techniques have been applied to reduce the cost. This includes using the linearization as an approximation [31], applying gradient-based instead of the Jacobian [34] etc. An early review was given by Arridge and Ren [1, 31].

None of the algorithms, however, is online. With traditional approaches, one typically assumes that many experiments are done, and a large number of pairs of $(\phi, (n_x \cdot v)f|_{\Gamma_+})$ are collected ahead of time. These data points are stored and used all-together in the computation as a whole batch. An immediate disadvantage is the run-time memory and computational cost: in each iteration, all experiment measurements are called for to adjust the parameter. We develop online algorithms for inverting RTE in this paper. In particular, we apply the stochastic gradient-descent method. It is a standard online algorithm: we start with one data point (one incoming–outgoing pair), and gradually adjust by incorporating new ones randomly selected from the data pool. This way, in each iteration, only very few data points are required, significantly accelerating the optimization. We stop once error tolerance is achieved. This online routine minimally requires data points, and avoids experiment waste. As will be shown later, numerically it is drastically more efficient too. We have to mention that we are not the very first group to explore the possibility of incorporating the random sampling techniques to inverse problems. The randomized version of the Kaczmarz's method (originally extensively studied in [16]) was proposed in [19] for elliptic equations with Dirichlet-to-Neumann map as the data.

In the following, we review the stochastic gradient descent method in section 2, and show the formulation of the inverse problem in both the linearized and the original nonlinear setting in section 3. Section 4 collects our numerical experiments.

2. Stochastic gradient descent method

We briefly review the stochastic gradient descent method in a general setting. The notation is consistent within this section, and will be adjusted accordingly in later sections.

Stochastic gradient descent (SGD) algorithm and many of its variants are often used to solve optimization problems of the form

$$\min \mathcal{J}(\sigma) = \frac{1}{N}\sum_{k=1}^{N} \mathcal{J}_k(\sigma), \tag{4}$$

where $\mathcal{J}$ is average of all $\mathcal{J}_k$, which maps the trainable parameters $\sigma \in \mathbb{R}^d$ to $\mathbb{R}$. N is the training sample size and could be very large depending on applications. To solve the problem using the standard gradient descent method, one updates σ_n for each step, the parameter at nth step, using:

$$\sigma_{n+1} = \sigma_n - \eta \nabla_x \mathcal{J}(\sigma_n) = \sigma_n - \frac{\eta}{N}\sum_{k=1}^{N} \nabla_\sigma \mathcal{J}_k(\sigma_n). \tag{5}$$

Here η is the gradient descent time step, or sometimes termed learning rate. This method requires derivative with respect to σ for all $\mathcal{J}_k$ evaluated at σ_n and the computation could be prohibitively expensive for big N.

SGD method is a stochastic alternative of gradient descent method (GD). It replaces the full gradient $\nabla_\sigma \mathcal{J}$ by only one sampled version in each iteration. In its simplest form, the SGD iteration is written as

$$\sigma_{n+1} = \sigma_n - \eta_n \nabla_\sigma \mathcal{J}_{\gamma_n}(\sigma_n), \tag{6}$$

where η_n is still the learning rate which may or may not vary in n. The learning direction is no longer the gradient of the whole cost function but is replaced by that of one sample $\mathcal{J}_{\gamma_n}$ randomly chosen from the sample pool ($\{\gamma_n\}$ is a random variable evenly chosen from $\{1, 2, \cdots, N\}$). Per iteration, SGD requires only one sample's derivative in σ at σ_n. Since the computational complexity is much reduced compared with GD, SGD is of favor for many large scale problems [6, 7].

There are many works addressing the performance of SGD. Studies were done on quantifying the convergence rate, choosing optimal learning rate, checking condition number dependence, and extending to nonconvex objectives. Many different variants (large batch training, stochastic average gradient, problem in the linear setting, and semi-stochastic method etc) [17, 22, 30, 33, 36, 40] have been studied too for various of purposes. The convergence in the most general setting is still unknown, and several techniques have been employed to explain it [6, 25, 27]. Among them we specifically mention the technique that links SGD algorithm with stochastic partial differential equations (SDEs). The computation of SDE itself also attracts some studies [28].

In fact, if one rewrites SGD as:

$$\sigma_{n+1} - \sigma_n = -\eta \nabla_\sigma \mathcal{J}(\sigma_n) + \eta \nabla_\sigma (\mathcal{J} - \mathcal{J}_{\gamma_n})(\sigma_n), \tag{7}$$

with η independent on n, it could be explained as the discretization for the following SDE:

$$\mathrm{d}X_t = b(X_t)\mathrm{d}t + a(X_t)\mathrm{d}W_t, \tag{8}$$

with η being the time step, $b(\sigma) = -\nabla_\sigma \mathcal{J}(\sigma)$ being the drift, and $a(x) = (\eta\Sigma)^{1/2}$ is the Brownian motion with the covariance defined by:

$$\Sigma = \frac{1}{N} \sum_k \left(\nabla \mathcal{J}(\sigma) - \nabla \mathcal{J}_k(\sigma)\right) \left(\nabla \mathcal{J}(\sigma) - \nabla \mathcal{J}_k(\sigma)\right)^\top. \tag{9}$$

This observation was made rigorous in [20], and we cite the theorem here:

Theorem 1. *Let $T > 0$ and define Σ as in (9). Assume $\mathcal{J}$, $\mathcal{J}_k$ are Lipschitz continuous, have at most linear asymptotic growth and have sufficiently high derivatives. Then, the stochastic process X_t with $t \in [0, T]$ satisfying*

$$\mathrm{d}X_t = -\nabla \mathcal{J}(X_t)\mathrm{d}t + (\eta\Sigma(X_t))^{1/2}\mathrm{d}W_t \tag{10}$$

is an order 1 weak approximation of the SGD, meaning: for every g of polynomial growth, there exists $C > 0$, independent of η, such that for all $n = 0, 1, ..., n_T = T/\eta$,

$$|\mathbb{E}g(X_{n\eta}) - \mathbb{E}g(\sigma_n)| < C\eta. \tag{11}$$

Here $X_{n\eta}$ is the solution to the SDE (8) evaluated at $n\eta$ and σ_n is the nth iteration solution to the SGD algorithm 6.

Consider the connection between SDE and the Fokker–Planck equation, the rewrite of the scheme (7) can also be regarded as the discretization for:

$$\partial_t u = b(x) \cdot \nabla u + \frac{1}{2}\eta\Sigma : \nabla^2 u \tag{12}$$

and this was made rigorous in [15] by using a small jump approximation in Markov process.

These results essentially claim that the SGD results can be interpreted by the solution to the SDE and the Fokker–Planck. Once the connection is drawn, the analysis to the SDE could be carried to understand the convergence behavior of SGD. Indeed, the equation contains a drift term and a diffusion term, in charge of bringing two types of behaviors. Suppose the initial guess is far away from the optimal and $\nabla_\sigma \mathcal{J}$ is very big, then the drift term will dominate. The solution therefore will firstly move according to the direction given by the drift term and quickly converge to a state to have $\nabla_\sigma \mathcal{J} = 0$. Once the drift term is small enough, the diffusion term will dominate, and this gives a Brownian motion like oscillating behavior. The two phases are termed the descent phase and the fluctuations phase, and the transition time is usually determined by setting $\mathbb{E}(X_t) = \sqrt{\mathrm{Var}(X_t)}$.

The solution to the SDE could be made more explicit when η, the learning rate is small. In the zero limit of η, the diffusion term shrinks. By performing the standard asymptotic expansion in η to (8), the solution to the SDE, in the leading order, becomes:

$$X_t \sim \mathcal{N}(X_{0,t}, \eta S_t), \tag{13}$$

a Gaussian process centers at $X_{0,t}$, a deterministic process that satisfies:

$$\frac{\mathrm{d}}{\mathrm{d}t} X_{0,t} = -\nabla \mathcal{J}(X_{0,t}),$$

with fluctuation S_t governed by:

$$\frac{\mathrm{d}}{\mathrm{d}t} S_t = -S_t H_t - H_t S_t + \Sigma_t. \tag{14}$$

Here $H_t = \nabla^2 \mathcal{J}(X_{0,t})$ is the Hessian of $\mathcal{J}$ evaluated at $X_{0,t}$, and $\Sigma_t = \Sigma(X_{0,t})$, with Σ defined in (9). The interested readers are referred to [20] for more details.

3. Inverting for optical properties of RTE

We apply SGD to the inverse problem in RTE. We first unify the notation. We focus on the critical case in this paper, meaning the absorption and the scattering term have the same intensities. The method takes minimum changes when the two terms are different. The calculation will be presented in remark 1 and numerical experiments will be demonstrated in section 4. The equation writes, in 2D:

$$\begin{cases} v \cdot \nabla f = \sigma(x_1, x_2)\mathcal{L}[f], & x = (x_1, x_2) \in [0,1]^2, v \in \mathbb{S} \\ f|_{\Gamma_-} = \phi(x_1, x_2, v) \end{cases},$$

where $\mathcal{L}[f]$ is the collision term:

$$\mathcal{L}[f] = \int_{\mathbb{S}} f \mathrm{d}v - f = \langle f \rangle_v - f.$$

Here $\mathrm{d}v$ is a normalized measure. If we write $v = (\cos\theta, \sin\theta)$ then:

$$\begin{cases} \cos\theta \partial_{x_1} f + \sin\theta \partial_{x_2} f = \sigma(x_1, x_2)\mathcal{L}[f], & (x_1, x_2, \theta) \in [0,1]^2 \times [-\pi, \pi] \\ f|_{\Gamma_-} = \phi(x_1, x_2, \theta) \end{cases}. \tag{15}$$

In the equation Γ_- collects coordinates on the four boundary lines with velocities pointing into the domain:

$$\begin{aligned}\Gamma_- = &\{x_1 = 0, x_2 \in [0,1], \cos\theta > 0\} \cup \{x_1 = 1, x_2 \in [0,1], \cos\theta < 0\} \\ &\cup \{x_1 \in [0,1], x_2 = 0, \sin\theta > 0\} \cup \{x_1 \in [0,1], x_2 = 1, \sin\theta < 0\},\end{aligned}$$

and Γ_+ collects the rest.

For every run of the experiment, one turns on light supported on Γ_- with prescribed intensities, termed $\phi^{(k)}$ and collects outgoing intensities, termed $\psi^{(k)}$. We note that $\psi^{(k)}$ contains pollution in the measuring procedure. The superindex k labels the round of experiment.

Throughout the section we may encounter the following norms:

$$\|f\|_{\pm}^2 = \int_{\Gamma_\pm} |f|^2 \mathrm{d}x\mathrm{d}v, \quad \|f\|_2^2 = \int_{\Omega\times\mathbb{S}} |f|^2 \mathrm{d}x\mathrm{d}v.$$

The following two subsections are devoted to nonlinear and linearized versions of the inverse problem, both of which employ dual problems for extracting information.

3.1. Nonlinear version

We look for the scattering coefficient $\sigma(x_1, x_2)$ in the nonlinear setting in this section. This is achieved by matching the result of the albedo operator acting on the incoming data $\phi^{(k)}$ and the measured data $\psi^{(k)}$. More precisely we perform the PDE-constraint optimization. Define the cost function:

$$\mathcal{J}_k = \frac{1}{2}\|(n\cdot v) f^{(k)} - \psi^{(k)}\|_+^2 + \frac{\alpha}{2}\|\sigma\|_2^2 \tag{16}$$

and the PDE constraint:

$$(v\cdot\nabla - \sigma\mathcal{L}) f^{(k)} = 0, \quad f^{(k)}|_{\Gamma_-} = \phi^{(k)}, \tag{17}$$

then we minimize:

$$\begin{cases} \min_\sigma & \frac{1}{N}\sum_k \mathcal{J}^{(k)} = \frac{1}{N}\sum_k \left(\frac{1}{2}\|(n\cdot v) f^{(k)} - \psi^{(k)}\|_+^2 + \frac{1}{2}\alpha\|\sigma\|_2^2\right) \\ \text{s.t.} & v\cdot\nabla f^{(k)} = \sigma(x_1,x_2)\mathcal{L}[f^{(k)}], \quad f^{(k)}|_{\Gamma_-} = \phi^{(k)} \end{cases}. \tag{18}$$

A more compact form of the problem writes:

$$\min_\sigma \quad \frac{1}{N}\sum_k \left(\frac{1}{2}\|\mathcal{A}(\sigma)[\phi^{(k)}] - \psi^{(k)}\|_+^2 + \frac{1}{2}\alpha\|\sigma\|_2^2\right) \tag{19}$$

where $\mathcal{A}$ is the albedo operator determined by σ that maps the incoming data ϕ to the outgoing data $(n\cdot v) f|_{\Gamma_+}$ with f satisfying (17). A Kolmogorov regularizer $\|\sigma\|_2$ is added. Both the mismatch term and regularization term are measured in L_2 norm. Note that the data is of the form of $(n\cdot v) f|_{\Gamma_+}$ but not $f|_{\Gamma_+}$.

The update formula given by SGD is straightforward:

$$\sigma_{n+1} = \sigma_n - \eta_n \frac{\mathrm{d}}{\mathrm{d}\sigma}\mathcal{J}_{\gamma_n}(\sigma_n), \tag{20}$$

with γ_n randomly selected from $\{1, \cdots, N\}$. This means in each iteration, to update σ from time step n to $n+1$, one randomly select a incoming–outgoing pair $(\phi^{(\gamma_n)}, \psi^{(\gamma_n)})$ and use the corresponding Fréchet derivative $\frac{\mathrm{d}}{\mathrm{d}\sigma}\mathcal{J}_{\gamma_n}$ evaluated at the previous data σ_n. To compute

the Fréchet derivative, however, we need to employ the dual problem. We now derive it, and ignore sub-index γ_n for conciseness of the notation.

We use the Lagrangian formulation. For all independent f, σ and the duals g and λ, we define the Lagrangian:

$$\mathsf{L}(\sigma,f,g,\lambda) = \mathcal{J}(\sigma,f) + \langle g,(v\cdot\nabla_x - \sigma\mathcal{L})f\rangle_2 + \langle \lambda,f|_{\Gamma_-} - \phi\rangle_-, \tag{21}$$

with the last two terms coming from multiplying the two constraints (the equation and the boundary condition) by the Lagrangian multiplier (g,λ). If the two constraints in (17) are satisfied, f and σ are no longer independent, and the last two terms disappear. On this special manifold, the Lagrangian is equivalent to $\mathcal{J}$. We denote such f by f_σ. On $f = f_\sigma$ manifold:

$$\mathsf{L}(\sigma,f_\sigma,g,\lambda) = \mathcal{J}(\sigma,f_\sigma). \tag{22}$$

Take derivative with respect to σ:

$$\frac{\mathrm{d}\mathcal{J}}{\mathrm{d}\sigma} = \frac{\partial \mathsf{L}}{\partial\sigma} + \frac{\partial \mathsf{L}}{\partial f}\frac{\partial f}{\partial\sigma}.$$

Suppose g and λ are selected properly to make $\frac{\partial \mathsf{L}}{\partial f} = 0$, then:

$$\frac{\mathrm{d}\mathcal{J}}{\mathrm{d}\sigma} = \frac{\partial \mathsf{L}}{\partial\sigma} = \frac{\partial\mathcal{J}}{\partial\sigma} - \int_{\mathbb{S}} g\mathcal{L}[f]\mathrm{d}v = \alpha\sigma - \int_{\mathbb{S}} g\mathcal{L}[f]\mathrm{d}v, \tag{23}$$

a formulation that could be explicitly computed.

To have $\frac{\partial \mathsf{L}}{\partial f} = 0$, we note that

$$\begin{aligned}
\frac{\partial \mathsf{L}}{\partial f} &= \frac{\partial\mathcal{J}}{\partial f} + \frac{\partial}{\partial f}\langle g,(v\cdot\nabla_x - \sigma\mathcal{L})f\rangle_{\Omega\times\mathbb{S}} + \frac{\partial}{\partial f}\langle\lambda,f|_{\Gamma_-} - \phi\rangle_- \\
&= \frac{\partial}{\partial f}\left[\frac{1}{2}\langle(v\cdot n)f - \psi,(v\cdot n)f-\psi\rangle_+ + \langle g,(v\cdot\nabla_x - \sigma\mathcal{L})f\rangle_{\Omega\times\mathbb{S}} + \langle\lambda,f|_{\Gamma_-} - \phi\rangle_-\right] \\
&= \frac{\partial}{\partial f}\left[\frac{1}{2}\langle(v\cdot n)f - \psi,(v\cdot n)f-\psi\rangle_+ + \langle(v\cdot n)g,f\rangle_+ + \langle(-v\cdot\nabla_x - \sigma\mathcal{L})g,f\rangle_{\Omega\times\mathbb{S}}\right. \\
&\quad \left. + \langle\lambda,f|_{\Gamma_-} - \phi\rangle_- + \langle(v\cdot n)g,f\rangle_-\right]
\end{aligned}$$

where in the last equation we have used:

$$\langle g,(v\cdot\nabla_x - \sigma\mathcal{L})f\rangle_{\Omega\times\mathbb{S}} = \langle(-v\cdot\nabla_x - \sigma\mathcal{L})g,f\rangle_{\Omega\times\mathbb{S}} + \int_{\Gamma_+\cup\Gamma_-}(n\cdot v)fg\mathrm{d}x\mathrm{d}v.$$

We combine terms supported in different domains, and let them vanish:

$$\begin{cases}(-v\cdot\nabla_x - \sigma\mathcal{L})g = 0, & (x,v)\in\Omega\times\mathbb{S}\\ (n\cdot v)f - \psi + g = 0, & (x,v)\in\Gamma_+\\ \lambda + (n\cdot v)g = 0, & (x,v)\in\Gamma_-\end{cases}. \tag{24}$$

The first two equations combined provide the restriction of g, i.e. g satisfies the dual problem:

$$\begin{cases}-v\cdot\nabla g = \sigma\mathcal{L}[g]\\ g|_{\Gamma_+} = -(n\cdot v)f|_{\Gamma_+} + \psi\end{cases}. \tag{25}$$

In each iteration, to update (20), we compute (25) with the current guess σ_n for g using the mismatch being the boundary condition, and then generate the Fréchet derivative using (23). We summarize the procedure in algorithm 1.

Algorithm 1. Find solution to the minimization problem (18).

Data: N experiments with

1. incoming data $\{\phi^{(k)}\}$;
2. outgoing measurements: $\{\psi^{(k)}\}$
3. error tolerance ε;
4. initial guess σ_0.

Result: The minimizer σ to the optimization problem (18) that is within ε accuracy in residue.

while $\|\frac{\mathrm{d}}{\mathrm{d}\sigma}\mathcal{J}_{\gamma_n}(\sigma_n)\| > \varepsilon$ **do**

Step I: randomly pick $\gamma_n \in \{1, \cdots, N\}$;

Step II: compute the forward problem (17) using boundary $\phi = \phi^{(\gamma_n)}$ with $\sigma = \sigma_n$ for $f^{(\gamma_n)}$;

Step III: compute the dual problem (25) using boundary $-(v \cdot n) f^{(\gamma_n)}|_{\Gamma_+} + \psi^{(\gamma_n)}$ with $\sigma = \sigma_n$ for $g^{(\gamma_n)}$;

Step IV: compute the Fréchet derivative (23): $\frac{\mathrm{d}}{\mathrm{d}\sigma}\mathcal{J}_{\gamma_n}(\sigma_n) = \alpha\sigma_n - \int_{\mathbb{S}^1} \mathcal{L}[f^{(\gamma_n)}] g^{(\gamma_n)} \mathrm{d}v$;

Step V: update using (20): $\sigma_{n+1} = \sigma_n - \eta \frac{\mathrm{d}}{\mathrm{d}\sigma}\mathcal{J}_{\gamma_n}(\sigma_n)$.

$n = n + 1$.

end

We emphasize that for clinic interests, N data points $\{\phi^{(k)}, \psi^{(k)}\}$ do not need to be prepared beforehand. Before converging, in each step, an NIR laser is randomly placed on Γ_- to generate $\phi^{(k)}$ and recerivers are placed on Γ_+ to collect $\psi^{(k)}$. Experiments are stopped once the algorithm gives convergence. In this way, no redundant information is collected and this online algorithm maximally saves the experimenting time.

Remark 1. It is of clinical interests that sometimes the equation (15) is not in the critical case and the total absorption term is different from the scattering case. For simplicity we set the scattering being 1 and study here how to recover the absorption term. The equation writes

$$v \cdot \nabla_x f = \mathcal{L} f - \sigma f \tag{26}$$

with boundary condition

$$f|_{\Gamma_-} = \phi.$$

And the goal is to use the information of $\mathcal{A}$ to recover σ. The minimization form writes as:

$$\begin{cases} \min_\sigma & \frac{1}{N}\sum_k \mathcal{J}_k = \frac{1}{N}\sum_k \left(\frac{1}{2}\|(n \cdot v) f^{(k)} - \psi^{(k)}\|_+^2 + \frac{1}{2}\alpha\|\sigma\|_2^2\right) \\ \text{s.t.} & v \cdot \nabla f^{(k)} = \mathcal{L}[f^{(k)}] - \sigma(x_1, x_2) f^{(k)}, \quad f^{(k)}|_{\Gamma_-} = \phi^{(k)} \end{cases}.$$

Following the same procedure, for all k, the Lagrangian is defined:

$$\mathsf{L}(\sigma, f, g, \lambda) = \mathcal{J}(\sigma, f) + \langle g, (v \cdot \nabla_x - \mathcal{L} + \sigma) f \rangle_2 + \langle \lambda, f|_{\Gamma_-} - \phi \rangle_-,$$

with the last two terms coming from the Lagrangian multiplier (g, λ). On $f = f_\sigma$ manifold, the two terms drop and the Lagrangian is equivalent to $\mathcal{J}$, and:

$$\mathsf{L}(\sigma, f_\sigma, g, \lambda) = \mathcal{J}(\sigma, f_\sigma). \tag{27}$$

Table 1. Numerical cost comparison: we compare the number of RTEs needs to be computed per iteration, the number of iterations needed for convergence, and the total amout of RTEs requried using SGD and GD. The last column shows the cost ratio. Larger sample size N provides bigger savings.

	SGD			GD			
N	RTE per iteration	Iteration	Total RTEs	RTE per iteration	Iteration	Total RTEs	Ratio (%)
100	2	2000	4000	200	100	20 000	20.0
200	2	1047	2094	400	87	34 800	6.02
400	2	935	1870	800	85	68 000	2.75

Take derivative with respect to σ:

$$\frac{\mathrm{d}\mathcal{J}}{\mathrm{d}\sigma} = \frac{\partial \mathsf{L}}{\partial \sigma} + \frac{\partial \mathsf{L}}{\partial f}\frac{\partial f}{\partial \sigma} = \alpha\sigma + \int_{\mathbb{S}} g f \,\mathrm{d}v.$$

In the second equation we purposely select g and λ to have $\frac{\partial \mathsf{L}}{\partial f} = 0$. This requires:

$$\begin{cases} (-v\cdot\nabla_x - \mathcal{L})g + \sigma g = 0, & (x,v)\in\Omega\times\mathbb{S} \\ (n\cdot v)f - \psi + g = 0, & (x,v)\in\Gamma_+ \\ \lambda + (n\cdot v)g = 0, & (x,v)\in\Gamma_- \end{cases}.$$

Once again the first two equations combined provide the restriction of g, the dual equation:

$$\begin{cases} -v\cdot\nabla g = \mathcal{L}[g] - \sigma g \\ g|_{\Gamma_+} = -(n\cdot v)f|_{\Gamma_+} + \psi \end{cases}. \tag{28}$$

In conclusion, to use SGD, we use the following in each iteration:

$$\sigma_{n+1} = \sigma_n - \eta_n \frac{\mathrm{d}}{\mathrm{d}\sigma}\mathcal{J}_{\gamma_n} = \sigma_n - \eta_n\left(\alpha\sigma_n + \int_{\mathbb{S}} g f \,\mathrm{d}v\right),$$

where f solves (26) with $\phi^{(\gamma_n)}$ being the boundary condition and σ_n being the media, and g solves (28) with $\psi^{(\gamma_n)}$ and σ_n.

3.2. Linearized procedure

In this section we describe the SGD applied on the linearized problem. The linearization is conducted upon σ_0, a background scattering coefficient believed to be very close to the true σ. The equation reads:

$$\begin{cases} v\cdot\nabla_x f &= \sigma\mathcal{L}f, \quad (x,v)\in\Omega\times\mathbb{S}, \\ f|_{\Gamma_-} &= \phi \end{cases}, \tag{29}$$

and its linearization is conducted assuming:

$$\tilde{\sigma}(x) = \sigma(x) - \sigma_0(x) \quad \text{and} \quad |\tilde{\sigma}| \ll |\sigma| \quad \text{(a.e.)}.$$

Then the linearized problem with the same inflow boundary condition reads as

$$\begin{cases} v\cdot\nabla_x f_0 = \sigma_0 \mathcal{L} f_0, & (x,v)\in\Omega\times\mathbb{S}, \\ f_0|_{\Gamma_-} = \phi \end{cases}. \tag{30}$$

Let

$$\tilde{f}(x,v) = f(x,v) - f_0(x,v)$$

be the fluctuation, we subtract the two equations (29) and (30) for:

$$\begin{cases} v\cdot\nabla_x \tilde{f} = \sigma_0\mathcal{L}\tilde{f} + \tilde{\sigma}\mathcal{L}f_0 \\ \tilde{f}|_{\Gamma_-} = 0 \end{cases}, \tag{31}$$

where we have omitted the higher order term $\tilde{\sigma}\mathcal{L}\tilde{f}$. To extract information to match the given data, we once again use the dual problem. Suppose we would like to find the information at $(x_*, v_*)\in\Gamma_+$, then we assign a delta function at the point for g to use as the boundary condition:

$$\begin{cases} -v\cdot\nabla_x g = \sigma_0\mathcal{L}g \\ g|_{\Gamma_+} = \delta_{x_*,v_*}(x,v) \end{cases}. \tag{32}$$

Multiply (32) by $\tilde{f}$ and multiply (31) by g and subtract them, we get

$$(n_*\cdot v_{x_*})\tilde{f}(x_*,v_*) = \int_\Omega \tilde{\sigma}\int_{\mathbb{S}^1}\mathcal{L}[f_0]g\,\mathrm{d}v\mathrm{d}x. \tag{33}$$

Note the left hand side is known since:

$$(n_*\cdot v_{x_*})\tilde{f}(x_*,v_*) = (n_*\cdot v_{x_*})f(x_*,v_*) - (n_*\cdot v_{x_*})f_0(x_*,v_*) \tag{34}$$

with the first term being a measurement $\psi(x_*,v_*)$, and the second computed from (30). We denote it by:

$$b(x_*,v_*;\phi) := (v_*\cdot n_{x_*})\tilde{f}(x_*,v_*) = \psi(x_*,v_*) - (v_*\cdot n_{x_*})f_0(x_*,v_*;\phi), \tag{35}$$

with f_0 implicitly depend on the inflow ϕ. We also denote the Fredholm kernel on the right hand side:

$$\beta(x,x_*,v_*;\phi) := \int_{\mathbb{S}^1}\mathcal{L}[f_0](x,v;\phi)g(x,v;\delta_{x_*,v_*})\mathrm{d}v, \tag{36}$$

as a function of x, x_*, v_* implicitly depend on ϕ. Then the equation rewrites:

$$\int_\Omega \beta(x,x_*,v_*;\phi)\tilde{\sigma}(x)\mathrm{d}x = b(x_*,v_*;\phi). \tag{37}$$

This formulation shows that to recover $\tilde{\sigma}$ amounts to invert the first type Fredholm integral. Note that this equation holds true for every $(x_*,v_*)\in\Gamma_+$.

The equal sign rarely holds true in reality due to the data pollution. Numerically each experiment prepares one specific incoming and outgoing pair $(\phi^{(k)},\psi^{(k)})$, which uniquely defines $b^{(k)}$ and $\beta^{(k)}$ according to (35) and (36). We then seek for σ that minimizes the following cost:

$$\min_\sigma \frac{1}{N}\sum_k \mathcal{J}_{(k)} = \frac{1}{N}\sum_k\left(\frac{1}{2}\|\int_\Omega \beta^{(k)}\sigma\mathrm{d}x - b^{(k)}\|_+^2 + \frac{\alpha}{2}\|\sigma\|_2^2\right) \tag{38}$$

where we abuse the notation σ to denote $\tilde{\sigma}$. The first term in $\mathcal{J}$ is the mismatch in (37) and the second term is the regularizer with a hyper-parameter α. Both terms are measured in L_2. In a compact form, it writes as:

$$\min_{\sigma} \frac{1}{N} \sum_{k} \left(\frac{1}{2} \|\mathcal{A}_0(\sigma)[\phi^{(k)}] - \psi^{(k)}\|_+^2 + \frac{\alpha}{2} \|\sigma\|_2^2 \right),$$

where $\mathcal{A}_0$ is the linearized albedo operator that maps the incoming flow ϕ supported on Γ_- to an outgoing flow measured at $(x_*, v_*) \in \Gamma_+$.

$$\mathcal{A}_0(\sigma)[\phi] = \int_\Omega \beta(x, x_*, v_*; \phi)\sigma(x)\mathrm{d}x$$

On this formulation, the application of SGD is straightforward:

$$\sigma_{n+1}(x) = \sigma_n(x) - \eta_n \left(\int_{\Gamma_+} \beta^{(\gamma_n)}(x, x_*, v_*) \left(\int_\Omega \beta^{(\gamma_n)}(\tilde{x}, x_*, v_*)\sigma_n(\tilde{x})\mathrm{d}\tilde{x} - b^{(\gamma_n)}(x_*, v_*) \right) \mathrm{d}x_*\mathrm{d}v_* + \alpha\sigma_n(x) \right) \tag{39}$$

with γ_n randomly selected from $\{1, \cdots, N\}$ at every step. We summarize the algorithm:

Algorithm 2. SGD applied on the minimization problem (38).

Data: N experiments with

1. incoming data $\phi^{(k)}$ for $\{k = 1, \cdots, N\}$;
2. outgoing measurements $\psi^{(k)}$ for $\{k = 1, \cdots, N\}$;
3. error tolerance ε;
4. initial guess σ_0.

Result: The minimizer σ to the optimization problem (38) that is within ε accuracy.

Step I: compute the dual problem (32) using δ_{x_*, v_*} for all $(x_*, v_*) \in \Gamma_+^{\mathrm{d}}$;

while $\|\frac{\mathrm{d}}{\mathrm{d}\sigma}\mathcal{J}_{\gamma_n}(\sigma_n)\| > \varepsilon$ **do**

Step II: randomly pick $\gamma_n \in \{1, \cdots, N\}$;

Step III: compute the background problem (30) using $\phi^{(\gamma_n)}$ for $f_0^{(\gamma_n)}$;

Step IV: compute $\beta^{(\gamma_n)}$ by (36);

Step V: compute $b^{(\gamma_n)}$ using (35) with $\psi^{(\gamma_n)}$ and $f_0^{(\gamma_n)}$;

Step VI: update using (39).

$n = n + 1$.

end

3.2.1. Discretization. We briefly describe the discrete version of (38). This is to replace the integration by its numerical version, and σ and $b^{(k)}$ are replaced by their discrete counterparts as well. To be precise,

$$\int_\Omega \beta^{(k)}\sigma\mathrm{d}x \quad \to \quad \mathsf{A}^{(k)}\sigma,$$

where $\mathsf{A}^{(k)}$ is a matrix of size $n_+ \times n_x$, where n_+ is the number of coordinates in Γ_+ and n_x is the number of coordinates in Ω. Its entries are defined by:

$$\mathsf{A}_{mn}^{(k)} = \beta^{(k)}(x_n, x_{*,m}, v_{*,m})\Delta x_n,$$

with Δx_n being the volume grid point x_n represents. For evenly distributed grids in $2D$, $\Delta x_n = \Delta x^2$ where Δx is the mesh size. In this way:

$$\left(\mathsf{A}^{(k)}\sigma\right)_m = \sum_n \beta^{(k)}(x_n, x_{*,m}, v_{*,m})\sigma(x_n)\Delta x_n,$$

numerically approximates $\int \beta^{(k)}\sigma \mathrm{d}x$ evaluated at $(x_{*,m}, v_{*,m})$, the mth pair on Γ_+. Notations σ and $b^{(k)}$ are abused to denote both continuous and discrete versions.

Now the objective function becomes:

$$\mathcal{J}_k(\sigma) = \frac{1}{2}\|\mathsf{A}^{(k)}\sigma - b^{(k)}\|_2^2 + \frac{\alpha}{2}\|\sigma\|_2^2. \tag{40}$$

Typically when rewritten in this way, α needs to be adjusted to incorporate the constant in the numerical integration, but we abuse the notation and still use α.

Numerically to update in each step, one needs to take gradient of $\mathcal{J}_k$ with respect to σ. Given the simple form we are studying here, it is simply, denoted by $G^{(k)}$:

$$G^{(k)}(\sigma) = \nabla_\sigma \mathcal{J}_k = \mathsf{A}^{(k)\top}\mathsf{A}^{(k)}\sigma - \mathsf{A}^{(k)\top}b^{(k)} + \alpha\sigma.$$

Denote

$$\mu_k := \mathsf{A}^{(k)^\top}\mathsf{A}^{(k)}, \quad \text{and} \quad \nu_k := -\mathsf{A}^{(k)\top}b^{(k)}, \tag{41}$$

then it has a simpler form:

$$G^{(k)}(\sigma) = (\mu_k + \alpha)\sigma + \nu_k.$$

Note α is a number and μ_k is a matrix of size $n_x \times n_x$ and ν_k is a vector of n_x length. We also immediately have:

$$G(\sigma) = \nabla_\sigma \mathcal{J} = \frac{1}{N}\sum_{k=1}^N \nabla_\sigma \mathcal{J}_k = \frac{1}{N}\sum_{k=1}^N G^{(k)}. \tag{42}$$

Define

$$\mu_{\mathsf{A}} := \mathbb{E}[\mathsf{A}^{(\gamma_k)^\top}\mathsf{A}^{(\gamma_k)}] = \frac{1}{N}\sum_{k=1}^N \mathsf{A}^{(k)^\top}\mathsf{A}^{(k)}, \quad \text{and} \quad \nu_{\mathsf{A}} := -\frac{1}{N}\sum_{k=1}^N \mathsf{A}^{(k)\top}b^{(k)}, \tag{43}$$

then (42) has a simpler form:

$$G(\sigma) = (\mu_{\mathsf{A}} + \alpha)\,\sigma + \nu_{\mathsf{A}}. \tag{44}$$

To update from n to $n+1$ step, one randomly pick γ_n and update σ_n using the gradient information of $\nabla_\sigma \mathcal{J}_{\gamma_n}$:

$$\sigma_{n+1} = \sigma_n - \eta G^{(\gamma_n)}(\sigma_n) = \sigma_n - \eta\left((\mu_{\gamma_n} + \alpha)\sigma_n + \nu_{\gamma_n}\right). \tag{45}$$

4. Error analysis

In this section we analyze the convergence of SGD on the linearized problem (38). Recall the minimization:

$$\min_\sigma \mathcal{J} = \min_\sigma \frac{1}{N}\sum_{k=1}^N \mathcal{J}^{(k)} = \frac{1}{N}\sum_k \left(\frac{1}{2}\|\int_\Omega \beta^{(k)}\sigma \mathrm{d}x - b^{(k)}\|_+^2 + \frac{\alpha}{2}\|\sigma\|_2^2\right), \tag{46}$$

where $\int_\Omega \beta^{(k)}\sigma \mathrm{d}x$, upon integrating over $x \in \Omega$ provides a function supported on $(x_*, v_*) \in \Gamma_+$, and the update formula (45). Denote σ^* the true solution to the minimization problem, meaning $G(\sigma^*) = 0$, and subtract it from the equation (45), we get the updating formula for the error. Denote $e_n = \sigma_n - \sigma^*$, the error at nth step, then:

$$\begin{aligned} e_{n+1} &= e_n - \eta G^{(\gamma_n)}(\sigma_n) \\ &= e_n - \eta\left(G(\sigma_n) - G(\sigma^*)\right) + \eta\left(G(\sigma_n) - G^{(\gamma_n)}(\sigma_n)\right) \\ &= e_n - \eta\left(\mu_{\mathsf{A}} + \alpha\right) e_n + \eta\left(G(\sigma_n) - G^{(\gamma_n)}(\sigma_n)\right) \qquad (47) \\ &= \underbrace{e_n - \eta\left(\mu_{\mathsf{A}} + \alpha\right) e_n}_{\text{decay}} + \underbrace{\eta\left((\mu_{\mathsf{A}} - \mu_{\gamma_n})\sigma_n + \nu_{\mathsf{A}} - \nu_{\gamma_n}\right)}_{\text{fluctuation}}. \qquad (48) \end{aligned}$$

From the first to the second line, we used the fact that $G(\sigma^*) = 0$, and from the second to the third line, we use the fact that G is linear on σ as seen in (44), and definitions in (41) and (43).

We further denote

$$B = \mathbb{I} - \eta\mu_{\mathsf{A}} - \eta\alpha, \quad \text{and} \quad d_n = \eta\left[(\mu_{\mathsf{A}} - \mu_{\gamma_n})\sigma_n + \nu_{\mathsf{A}} - \nu_{\gamma_n}\right], \qquad (49)$$

then the update formula becomes:

$$e_{n+1} = Be_n + d_n. \qquad (50)$$

According to this formula, we immediately see that the decay of e_n is controlled by two pieces: the first term provides the iterative decay while the second term gives fluctuation that represents the randomness from sampling γ_n. The key of error analysis is to:

1. find appropriate η so that $B = \mathbb{I} - \eta(\mu_{\mathsf{A}} + \alpha)$ has smaller than 1 spectrum, leading to convergence;
2. show the fluctuation term has mean zero, and thus it is not producing extra error on average;
3. show the fluctuation term has very small variance, and thus the chance of producing extra error is small.

The first argument is relatively straightforward, and the latter two amount to analyze the behavior of d_n. We first summarize it in lemma 1 and collect error analysis on the mean and the variance in theorems 2 and 3, respectively.

Lemma 1. *Assume*

$$\mathbb{E}\left(\|\mu_{\mathsf{A}} - \mu_{\gamma_n}\|_2^2\right) < C_\mu, \text{and} \quad \mathbb{E}\left(\|\nu_{\mathsf{A}} - \nu_{\gamma_n}\|_2^2\right) < C_\nu, \quad \forall n.$$

Using the definition in (49) we have:

1. $\mathbb{E}(d_n) = 0$ *for all* n;
2. $\mathrm{Cov}[d_i, d_j] = 0$ *for all* $1 \leqslant i < j \leqslant n$;
3. $\mathrm{Cov}[d_n, d_n] \leqslant C\eta^2(\mathbb{E}\left(\|\sigma_n\|_2^2\right) + 1)$.

Proof.

1. According to the definition:

$$\frac{1}{\eta}\mathbb{E}(d_n) = \mathbb{E}((\mu_{\mathsf{A}} - \mu_{\gamma_n})\sigma_n) + \mathbb{E}(\nu_{\mathsf{A}} - \nu_{\gamma_n}).$$

The second term is zero due to equations (41) and (43). To study the first term we first realize that the randomness comes from both γ_n and σ_n. Due to (20), σ_n only depends on $\{\gamma_1, \ldots, \gamma_{n-1}\}$, and thus it is independent of γ_n. Therefore:

$$\mathbb{E}((\mu_{\mathsf{A}} - \mu_{\gamma_n})\sigma_n) = \mathbb{E}(\mu_{\mathsf{A}} - \mu_{\gamma_n})\mathbb{E}(\sigma_n).$$

Given (41) and (43), we see $\mathbb{E}((\mu_{\mathsf{A}} - \mu_{\gamma_n})\sigma_n) = 0$ and thus d_n is mean zero.

2. Since d_i is mean zero:

$$\mathrm{Cov}[d_i, d_j] = \mathbb{E}(d_i d_j^\top) = \mathbb{E}\left(d_i \mathbb{E}(d_j^\top)\right) = 0.$$

The first equation comes from d_i and d_j being mean zero. The second equation holds true because $i < j$.

3. For the third covariance:

$$\begin{aligned}\frac{1}{\eta^2}\mathrm{Cov}[d_n, d_n] =& \frac{1}{\eta^2}\mathbb{E}(d_n d_n^\top)\\ =& \mathbb{E}\left((\mu_{\mathsf{A}} - \mu_{\gamma_n})\sigma_n\sigma_n^\top(\mu_{\mathsf{A}} - \mu_{\gamma_n})^\top\right) + \mathbb{E}\left((\nu_{\mathsf{A}} - \nu_{\gamma_n})(\nu_{\mathsf{A}} - \nu_{\gamma_n})^\top\right)\\ &+ \mathbb{E}\left((\mu_{\mathsf{A}} - \mu_{\gamma_n})\sigma_n(\nu_{\mathsf{A}} - \nu_{\gamma_n})^\top\right) + \mathbb{E}\left((\nu_{\mathsf{A}} - \nu_{\gamma_n})\sigma_n^\top(\mu_{\mathsf{A}} - \mu_{\gamma_n})^\top\right).\end{aligned}$$

Take arbitrary $x \in \mathbb{R}^{N_x}$ with $\|x\|_2 = 1$ and multiply on both sides, we have

$$\begin{aligned}\frac{1}{\eta^2}x^\top\mathrm{Cov}[d_n, d_n]x =& x^\top\mathbb{E}\left((\mu_{\mathsf{A}} - \mu_{\gamma_n})\sigma_n\sigma_n^\top(\mu_{\mathsf{A}} - \mu_{\gamma_n})^\top\right)x + x^\top\mathbb{E}\left((\nu_{\mathsf{A}} - \nu_{\gamma_n})(\nu_{\mathsf{A}} - \nu_{\gamma_n})^\top\right)x\\ &+ 2x^\top\mathbb{E}\left((\mu_{\mathsf{A}} - \mu_{\gamma_n})\sigma_n(\nu_{\mathsf{A}} - \nu_{\gamma_n})^\top\right)x\\ \leqslant& 2C_\mu\mathbb{E}\left(\|\sigma_n\|_2^2\right) + 2C_\nu.\end{aligned}$$

To obtain the inequality we used the fact that

$$\begin{aligned}x^\top\mathbb{E}\left((\mu_{\mathsf{A}} - \mu_{\gamma_n})\sigma_n\sigma_n^\top(\mu_{\mathsf{A}} - \mu_{\gamma_n})^\top\right)x =& x^\top\mathbb{E}\left((\mu_{\mathsf{A}} - \mu_{\gamma_n})\mathbb{E}\left(\sigma_n\sigma_n^\top\right)(\mu_{\mathsf{A}} - \mu_{\gamma_n})^\top\right)x\\ \leqslant& C_\mu\mathbb{E}\left(\|\sigma_n\|_2^2\right)\end{aligned}$$

and that

$$\begin{aligned}2x^\top\mathbb{E}\left((\mu_{\mathsf{A}} - \mu_{\gamma_n})\sigma_n(\nu_{\mathsf{A}} - \nu_{\gamma_n})^\top\right)x =& 2\mathbb{E}\left(x^\top(\mu_{\mathsf{A}} - \mu_{\gamma_n})\mathbb{E}(\sigma_n)(\nu_{\mathsf{A}} - \nu_{\gamma_n})^\top x\right)\\ \leqslant& \mathbb{E}\left[\left(x^\top(\mu_{\mathsf{A}} - \mu_{\gamma_n})\mathbb{E}(\sigma_n)\right)^2 + \left((\nu_{\mathsf{A}} - \nu_{\gamma_n})^\top x\right)^2\right]\\ \leqslant& C_\mu\mathbb{E}\left(\|\sigma_n\|_2^2\right) + C_\nu.\end{aligned}$$

We achieve the conclusion by multiplying η^2 on both sides and choose $C = 2\max\{C_\mu, C_\nu\}$. □

With this lemma we study the mean and the variance of the error in the following two theorems.

Theorem 2. *Denote σ^* the minimizer of problem (46) and the expected value of error:*

$$u_n = \mathbb{E}(e_n) = \mathbb{E}(\sigma_n - \sigma^*).$$

Assume that μ_{A} (defined in (43) has a bounded spectrum, meaning there exists C_A such that:

$$\|\mu_{\mathsf{A}}\|_2 \leqslant C_A, \tag{51}$$

then for $0 < \eta < \frac{2}{C_A + \alpha}$, the expected value of error decays to zero exponentially fast:

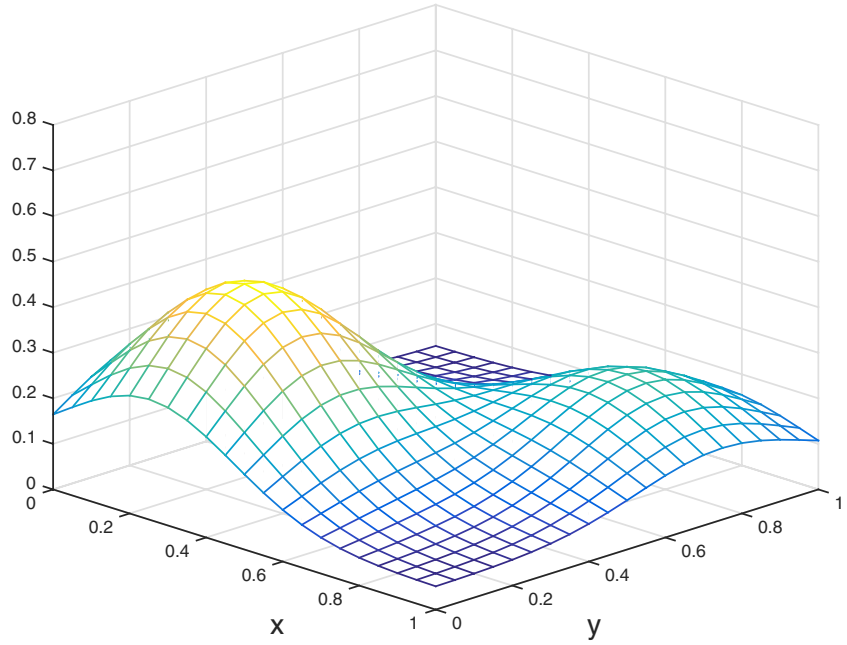

Figure 1. Real scattering coefficient.

$$\|u_n\|_2 \leqslant \lambda^n \|u_0\|_2, \tag{52}$$

where $|\lambda| < 1$ *will be defined in (54).*

Proof. We start from the iteration formula for e_n in (20). Take expectation on both sides:

$$u_{n+1} = u_n - \eta(\mu_A + \alpha)u_n + \eta\mathbb{E}(d_n). \tag{53}$$

Since d_n is mean zero according to the previous lemma, (53) becomes:

$$u_{n+1} = (\mathbb{I} - \eta\mu_A - \eta\alpha)u_n.$$

With $0 < \eta < \frac{2}{C_A+\alpha}$ and define

$$\lambda := \|\mathbb{I} - \eta\mu_A - \eta\alpha\|_2, \tag{54}$$

λ is guaranteed to be controlled by 1 and we achieve the conclusion. □

Theorem 3. *With small learning rate* η*, the error of SGD algorithm has bounded covariance:*

$$\mathrm{Cov}[e_n, e_n] \lesssim \eta, \quad \forall n.$$

Proof. We once again use:

$$e_{n+1} = Be_n + d_n,$$

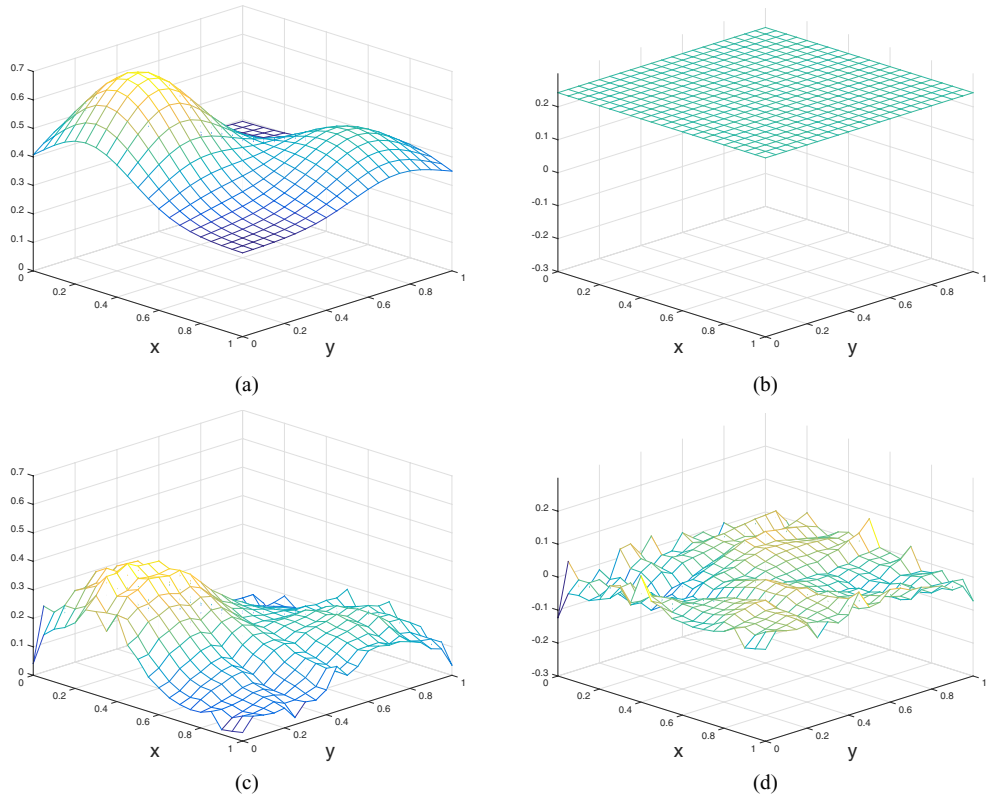

Figure 2. Nonlinear setting with initial guess being a constant deviation from the true media. (a) Initial guess σ_0 is a constant deviation from the true media. (b) Difference of σ_0 and the true media. (c) σ_{2000}. (d) Difference of σ_{2000} and the true media.

with

$$B = \mathbb{I} - \eta\mu_{\mathsf{A}} - \eta\alpha, \quad \text{and} \quad d_n = (\mu_{\mathsf{A}} - \mu_{\gamma_n})\sigma_n + \nu_{\mathsf{A}} - \nu_{\gamma_n}.$$

By induction,

$$e_n = B^n e_0 + \sum_{j=1}^{n-1} B^{n-j} d_j.$$

Take covariance of both sides and recall $\mathrm{Cov}[d_i, d_j] = 0$ for all $i \neq j$:

$$\mathrm{Cov}[e_n, e_n] = \sum_{i,j} \mathrm{Cov}[B^{n-i} d_i, B^{n-j} d_j] = \sum_i B^{n-i} \mathrm{Cov}[d_i, d_i] (B^{n-i})^\top.$$

Take arbitrary $x \in \mathbb{R}^{N\times 1}$ with $\|x\|_2 = 1$ and multiply on both sides, we have

$$x^\top \mathrm{Cov}[e_n, e_n] x = \sum_i (x^\top B^{n-i}) \mathrm{Cov}[d_i, d_i] (x^\top B^{n-i})^\top \leqslant \sum_i C\eta^2 \lambda^{2(n-i)} (\mathbb{E}[\|\sigma_i\|_2^2] + 1)$$

where the inequality incorporates the previous lemma. Further notice that $\mathbb{E}[\|\sigma_i\|_2^2] \leqslant \mathbb{E}[\|e_i\|_2^2] + \|\sigma^*\|_2^2$, we absorb the constant:

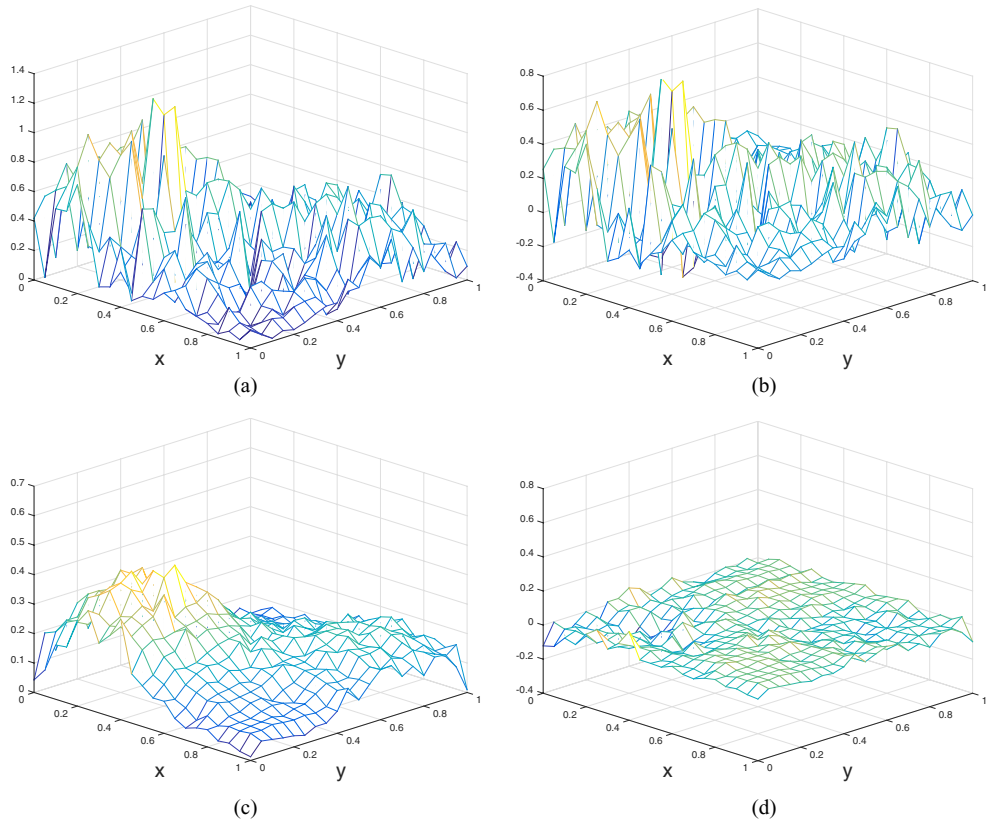

Figure 3. Nonlinear setting with initial guess being a random field. (a) Initial guess σ_0 is a random field. (b) Difference of σ_0 and the true media. (c) σ_{2000}. (d) Difference of σ_{2000} and the true media.

$$x^\top \mathrm{Cov}[e_n, e_n]x \leqslant \sum_i \tilde{C}\eta^2\lambda^{2(n-i)}(\mathbb{E}[\|e_i\|_2^2] + 1), \tag{55}$$

where $\tilde{C} = C + \|\sigma^*\|_2$. This inequality only serves as a iterative formula. Upon assuming $\mathbb{E}[\|e_i\|_2^2]$ is uniformly bounded by M, then:

$$x^\top \mathrm{Cov}[e_n, e_n]x \leqslant \tilde{C}\eta^2(M+1)\frac{1-\lambda^{2n}}{1-\lambda^2} \lesssim \eta. \tag{56}$$

The last inequality comes from the definition of $\lambda = \|\mathbb{I} - \eta\mu_{\mathsf{A}} - \eta\alpha\|_2 = \mathcal{O}(1-\eta)$. Since x is arbitrary, we achieve the conclusion.

To show that there exists a constant $M > 0$ such that $\mathbb{E}[\|e_i\|_2^2]$ is truly uniformly bounded by M, we use mathematical induction. It is easy to prove the argument is true for $i = 0$ by choosing $M = 2\max\{1, \mathbb{E}[\|e_0\|_2^2]\} = \max\{2, 2\|e_0\|_2^2\}$. Then we assume the argument is true for all $i < n$ and we want to show that $\mathbb{E}[\|e_n\|_2^2] \leqslant M$. We notice that

$$\mathbb{E}[\|e_n\|_2^2] = \mathrm{Tr}(\mathrm{Cov}[e_n, e_n]) \leqslant N\|\mathrm{Cov}[e_n, e_n]\|_2,$$

then since (55) is true for any x, we have

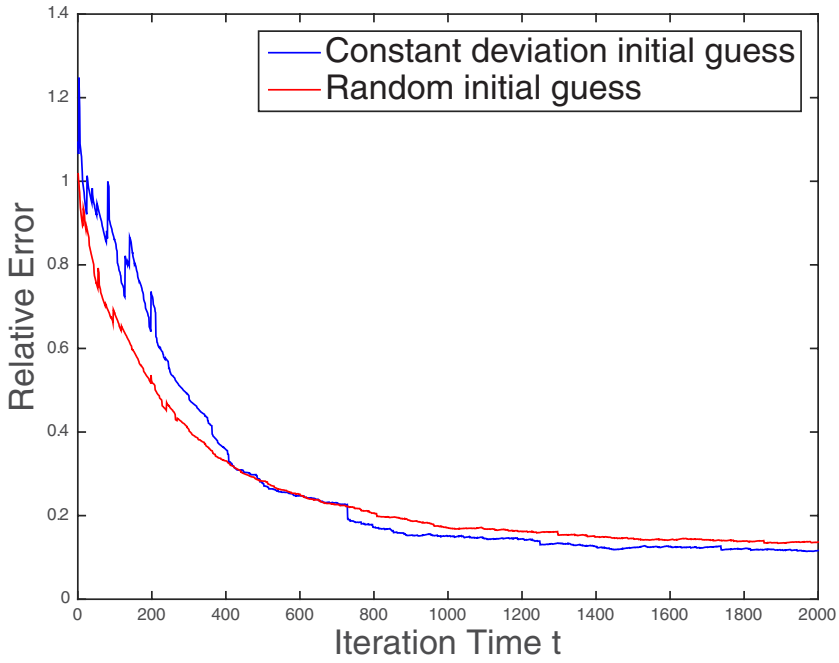

Figure 4. Nonlinear setting. The convergence of relative error in time. We see that the error decays almost exponentially fast at the beginning with small fluctuations and gradually saturate. The learning rate η_n is extremely small after 1000 times steps and the decay significantly slows down.

$$\|\mathrm{Cov}[e_n, e_n]\|_2 \leqslant \sum_i \tilde{C}\eta^2\lambda^{2(n-i)}(\mathbb{E}[\|e_i\|_2^2] + 1).$$

Combine the above two inequalities and our induction assumption for $i < n$, we derive that

$$\mathbb{E}[\|e_n\|_2^2] \leqslant N\left(\tilde{C}\eta^2(M+1)\frac{1-\lambda^{2(n-1)}}{1-\lambda^2} + \tilde{C}\eta^2(\mathbb{E}[\|e_n\|_2^2] + 1)\right).$$

For small enough η, this leads to:

$$\mathbb{E}[\|e_n\|_2^2] \leqslant 2N\left(\tilde{C}\eta^2(M+1)\frac{1-\lambda^{2(n-1)}}{1-\lambda^2} + \tilde{C}\eta^2\right).$$

Use the fact $\lambda = \mathcal{O}(1-\eta)$, we can further choose η small such that

$$\mathbb{E}[\|e_n\|_2^2] \leqslant \frac{1}{2}(M+1) + \frac{1}{2} = \frac{M}{2} + 1 \leqslant M,$$

which finishes the mathematical induction. □

We finally comment that the two theorems above in fact resonate the analysis in the general setting as stated in section 2. There are two main pieces in the error: the iterative decaying term, and the fluctuation term. If the initial guess gives an order 1 error, then the decaying term dominates first, and one simply see the error converging to zero exponentially fast. Once the

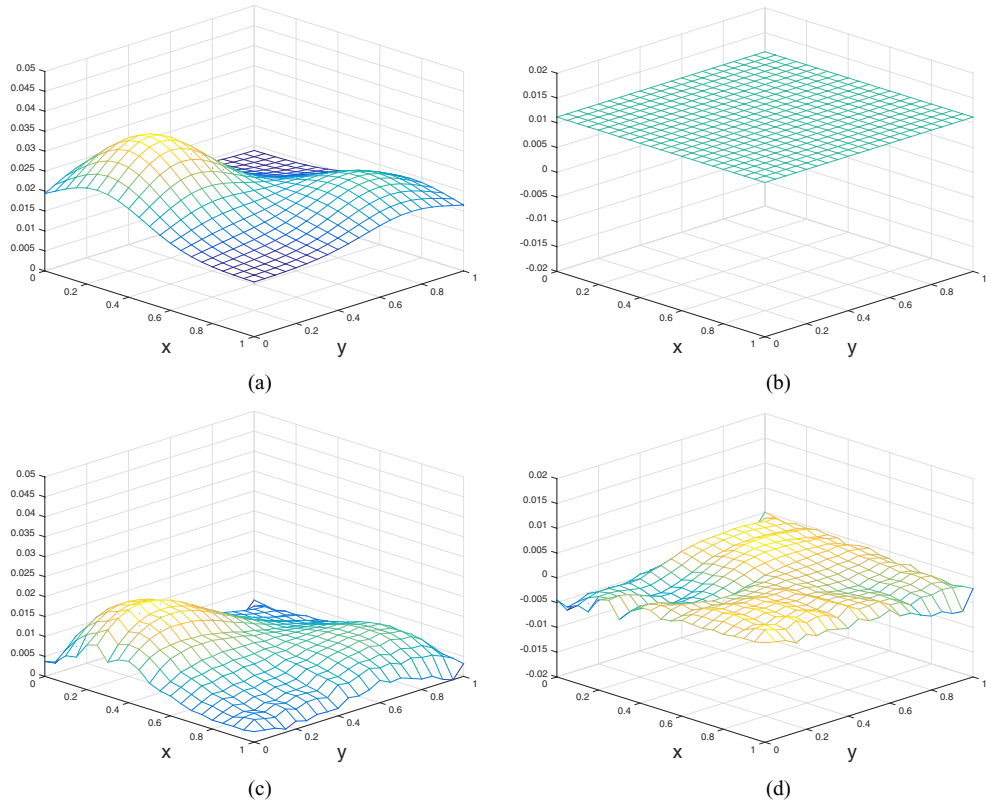

Figure 5. Linearized setting with initial guess being a constant deviation from the true media. (a) Initial guess $\tilde{\sigma}_0$ is a constant deviation. (b) Difference of $\tilde{\sigma}_0$ with the true media $\tilde{\sigma}$. (c) $\tilde{\sigma}_{20000}$. (d) Difference of $\tilde{\sigma}_{20000}$ with the true media $\tilde{\sigma}$.

error becomes as small as the variance (which is at η level), the fluctuation term dominates. To force the error converging to zero, numerically one could gradually decrease η so that the error fluctuates around zero with smaller and smaller variance. The result will be seen in our numerical results too.

5. Numerical test

To illustrate our theoretical results, we present a few numerical test below. The computational space domain is a unit square $\Omega = [0,1]^2$ with mesh size $\mathrm{d}x = 1/20 = 0.05$, and the velocity domain a unit circle $\mathbb{S}$ with mesh size $\mathrm{d}\theta = \frac{2\pi}{40}$. Therefore in the discrete setting:

$$\Omega^d \times \mathbb{S}^d = \{(x_m, \theta_n) = (m_1\mathrm{d}x, m_2\mathrm{d}x, -\pi + n\mathrm{d}\theta) : \text{with} \quad m_1, m_2 = 0, \cdots, 20, n = 0, \cdots, 40\},$$

and

$$\Gamma_-^d = \{(x_1 = m_1\mathrm{d}x, x_2 = m_2\mathrm{d}x, \theta = n\mathrm{d}\theta)\}$$

with

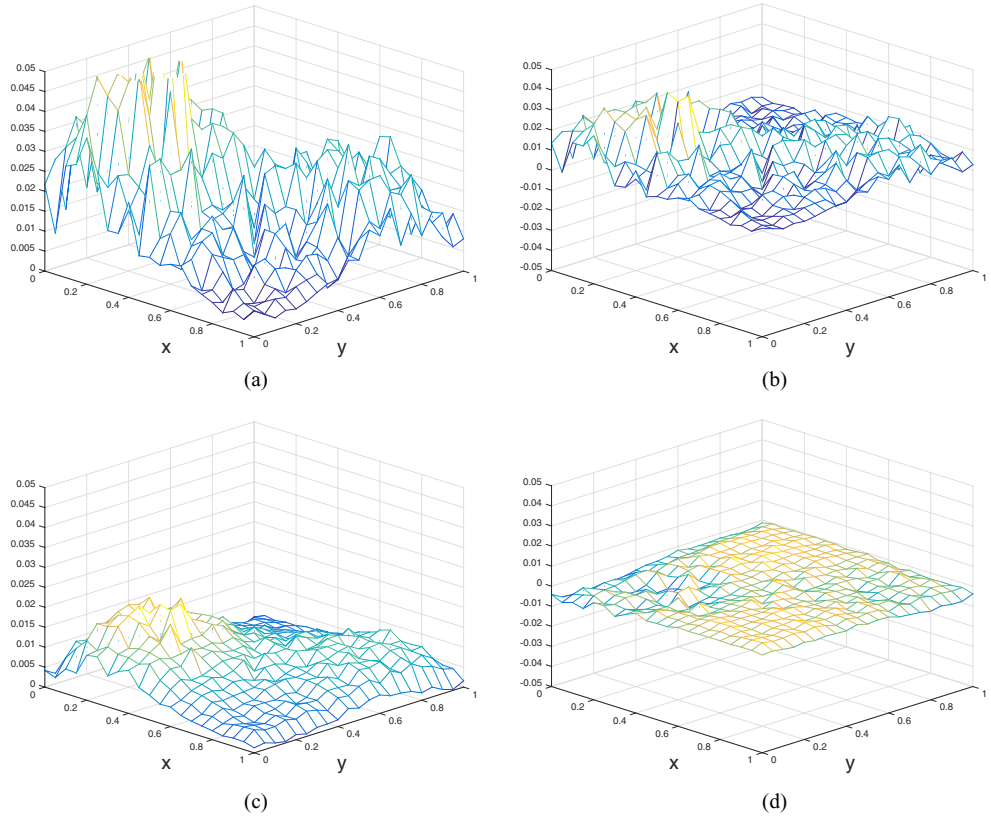

Figure 6. Linearized setting with initial guess being a random field. (a) Initial guess $\tilde{\sigma}_0$ is a random field. (b) Difference of $\tilde{\sigma}_0$ with the true media $\tilde{\sigma}$. (c) $\tilde{\sigma}_{20\,000}$. (d) Difference of $\tilde{\sigma}_{20\,000}$ with the true media $\tilde{\sigma}$.

$$\{m_1 = 0, m_2 \in [0, 20], n \in [10, 30]\} \cup \{m_1 = 20, m_2 \in [0, 20], n \in ([0, 10] \cup [30, 40])\}$$
$$\cup \{m_1 \in [0, 20], m_2 = 0, n \in [20, 40]\} \cup \{m_1 \in [0, 20], m_2 = 20, n \in [0, 20]\}.$$

We use GMRES [21] to solve the forward problem (15) with tolerance 10^{-12}. The scattering coefficient in our experiment is set to be

$$\sigma(x_1, x_2) = \frac{1}{20}\left[1 + 8\exp\left(-10(x_1 - \frac{1}{4})^2 - 10(x_2 - \frac{1}{4})^2\right) + 4\exp\left(-10(x_1 - \frac{3}{4})^2 - 10(x_2 - \frac{3}{4})^2\right)\right]. \tag{57}$$

Its evaluation in Ω ranges from 0.05 to 0.45, as plotted in figure 1.

5.1. Nonlinear case

In the nonlinear case (18), we use 1000 data points $\{(\phi^{(j)}, \psi^{(j)}) : 1 \leqslant j \leqslant 1000\}$, where $\phi^{(j)}(x, v)$ is a Dirac delta function centered at a random boundary point and pointing to a random inflow direction. $\psi^{(j)}(x, v)$ is the corresponding measurement on the outflow boundary, i.e.

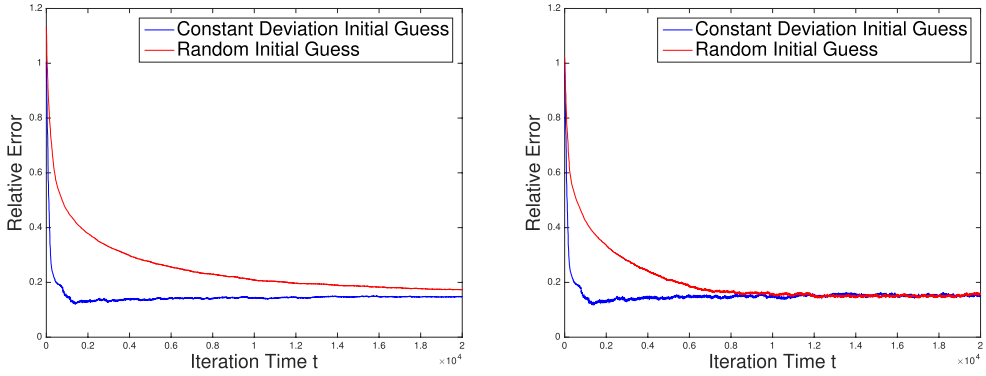

Figure 7. Linearized setting. The convergence of relative error in time. The error decays almost exponentially fast at the beginning with small fluctuations and gradually saturate. The two panels are for changing-in-time learning rate and the constant learning rate respectively.

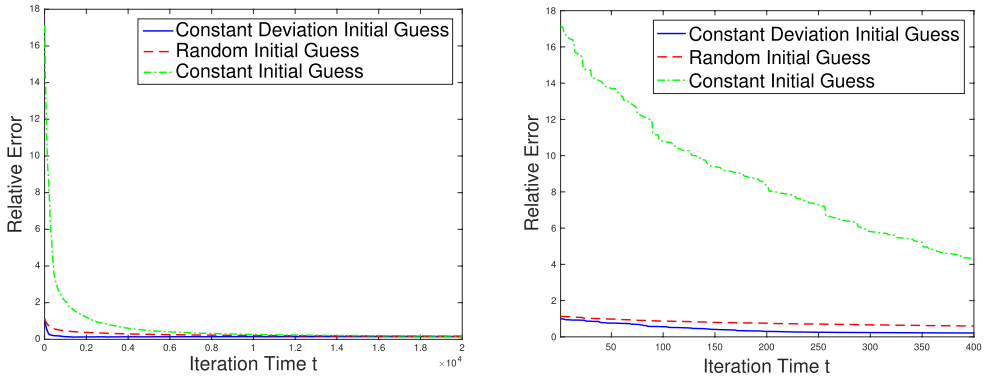

Figure 8. Linearized setting. The green dashed line shows the convergence of relative error with the initial guess far away from the true solution. The plot on the right is the zoom-in to the first 400 steps. The oscillation introduced from the stochasticity in the algorithm is obvious.

$$\phi^{(j)}(x,v) = \delta(x - x_{(j)})\delta(v - v_{(j)}), (x_{(j)}, v_{(j)}) \in \Gamma_- \quad \text{and} \quad \psi^{(j)}(x,v) = (n_x \cdot v) f(x,v;\phi^{(j)})|_{\Gamma_+}. \tag{58}$$

For our numerical experiments, we set the regularization parameter $\alpha = 1$ and learning rate $\eta_n = \frac{\eta_0}{1+\eta_0 \alpha n}$ with $\eta_0 = 0.0044$. Note that the learning rate is a hyperparameter that can be adjusted according to users' preferences. We choose the recommended $\frac{1}{n}$ from [7]. We test our algorithm with two different initial guesses: 1. Initial guess is a constant deviation from the real scattering coefficient $\sigma_0 = \sigma + 0.18$; 2. Initial guess is the product of the scattering coefficient and a random field: $\sigma_0 = \sigma R$, where $R \in \mathbb{R}^{21\times 21}$ has i.i.d. random variable components drew from uniform distribution $U([0.1, 3.1])$. In each iteration, two forward problems (one original and one dual) are solved to compute the gradient and we run SGD algorithm for 2000 steps.

We present the numerical solutions in figures 2 and 3 for constant deviation and random deviation as the initial guess respectively. In both, the upper left plot shows the initial guess

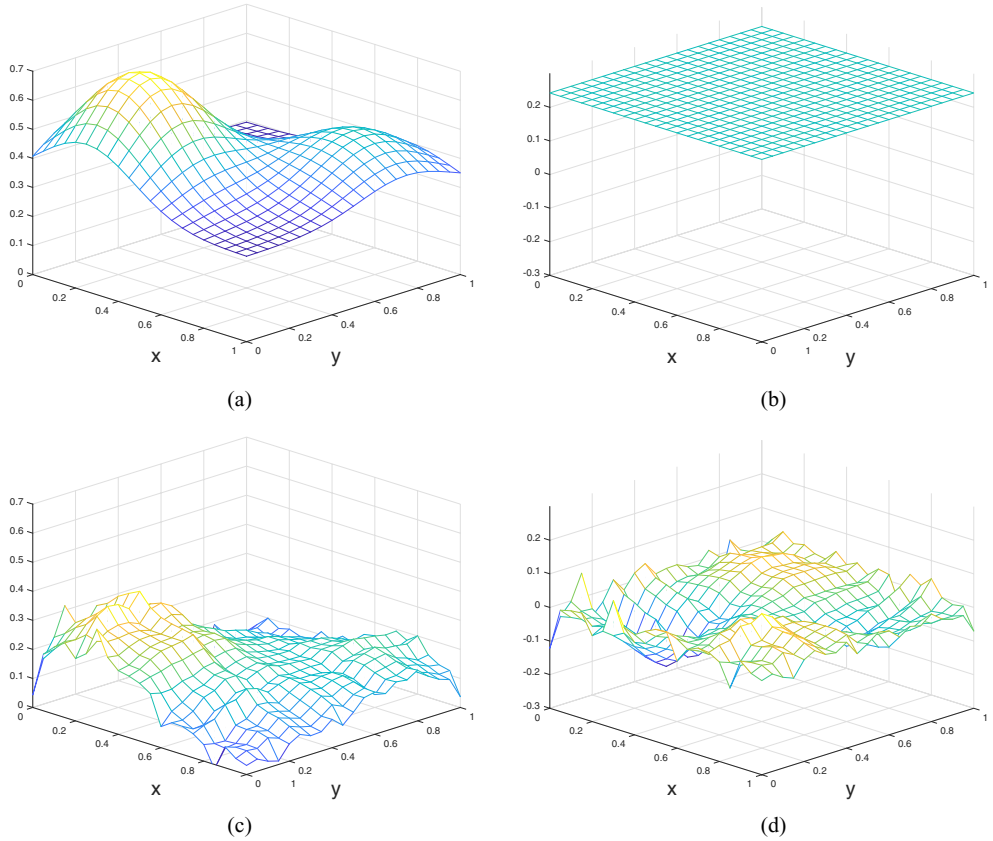

Figure 9. The plots show the absorption coeffient recovery. The two plots on the left panel show the media at initial time step and after 2000 iterations. The errors are shown in the two plots on the right. (a) Initial guess σ_0 is a constant deviation from the true media. (b) Difference of σ_0 and the true media. (c) σ_{2000}. (d) Difference of σ_{2000} and the true media.

σ_0, and the difference compared with the true media is plotted in the upper right. The lower left and lower right plots show the numerical solution after 2000 iterations and its difference from the true media. We also record the relative error between σ_n and σ and plot the decay in figure 4. Note that due to the nontrivial regularization term, we cannot expect the solution converging to the true media. As seen in figure 4 the error saturates at 0.2. It does provide very good recovery visually as seen in figures 2 and 3.

5.2. Linear case

We use the same data set in the linearized setting. The background state is given as proportional to the real media $\sigma_0 = 0.95\sigma$, and thus the to-be-recovered perturbed media $\tilde{\sigma}$, by definition (3.2) ranges from 0.0025 to 0.0225. We choose same regularization coefficient $\alpha = 1$. We also test the problem using the constant learning rate $\eta_0 = 0.0002$ and the learning rate recommended in [6]: $\eta_n = \frac{\eta_0}{1+\eta_0 \alpha n}$ with $\eta_0 = 0.0002$.

We once again use constant deviation and random deviation as the initial guess for the SGD algorithm. For constant deviation initial guess we set $\tilde{\sigma}_0 = \tilde{\sigma} + 0.0111$ whereas for random

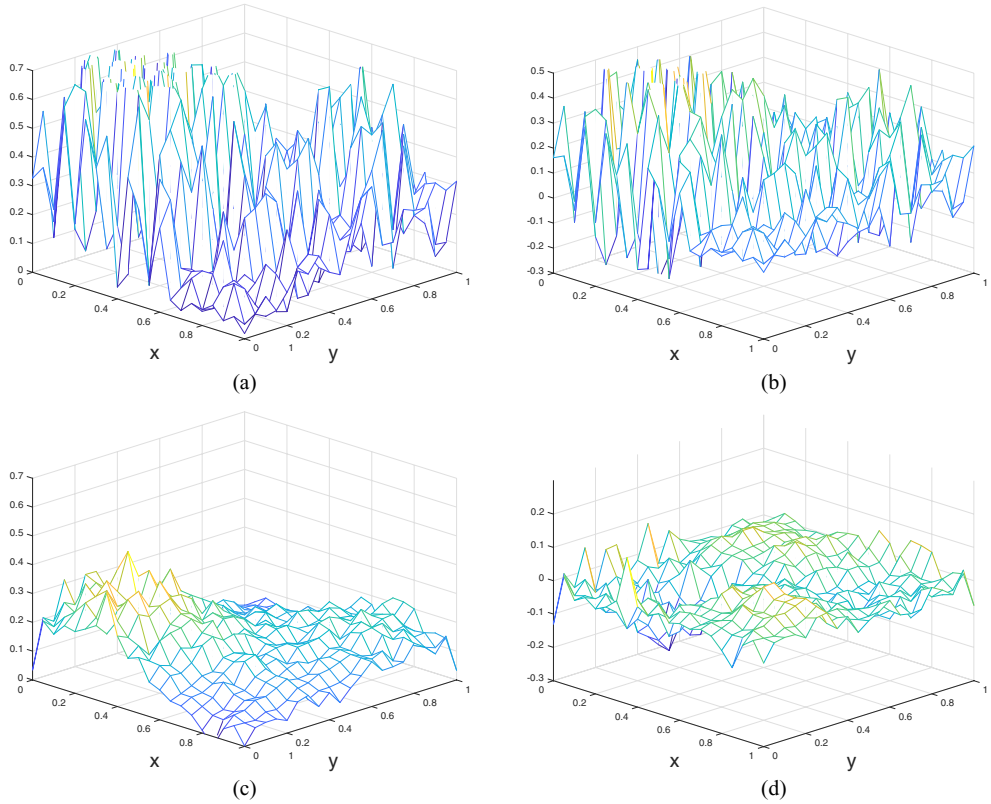

Figure 10. SGD is used to recover the absorption coeffient with initial guess being a random field. The media given at the initial step and after 2000 iterations are plotted, together with the errors. (a) Initial guess σ_0 is a random field. (b) Difference of σ_0 and the true media. (c) σ_{2000}. (d) Difference of σ_{2000} and the true media.

initial guess we set $\tilde{\sigma}_0 = \tilde{\sigma}R$ with $R \in \mathbb{R}^{21\times 21}$ drew its components from uniform distribution $U([1,3])$.

As presented in algorithm 2, several *offline* adjoint problems are pre-computed using background state σ_0 with Dirac delta outflow boundary conditions. In each iteration, only one forward problem is solved using background state σ_0 and random input $\phi^{(\gamma^n)}$ for $f_0(x, v; \phi^{(\gamma^n)})$. We run SGD algorithm with 20000 iterations. The numerical results are demonstrated in figures 5 and 6. They have constant and random deviation as the initial guess respectively. The decay of the relative error for both types of learning rates are shown in figure 7. In figure 8 we plot and compare the convergence of the error when the initial guess largely deviates from the true solution: $\sigma_0 = 0.2000$. The initial relative error is as large as 17.12.

Comparing to the nonlinear case, the convergence of relative error requires more iterations as here we aim to recover the small residue $\tilde{\sigma} = \sigma - \sigma_0$, which is much smaller than σ.

In table 1 we record the number of RTEs that need to compute per iteration, the number of iterations needs to achieve convergence, and the total number RTEs computed for all three sample sizes, and both methods. Note that in each iteration, SGD requires computation of one forward RTE $\sim$(17) and one dual RTE (25), while GD requires computation of N forward and N duals. Note also that with N $=$ 100 both SGD and GD fail to converge before achieving the maximum number of allowed iterations.

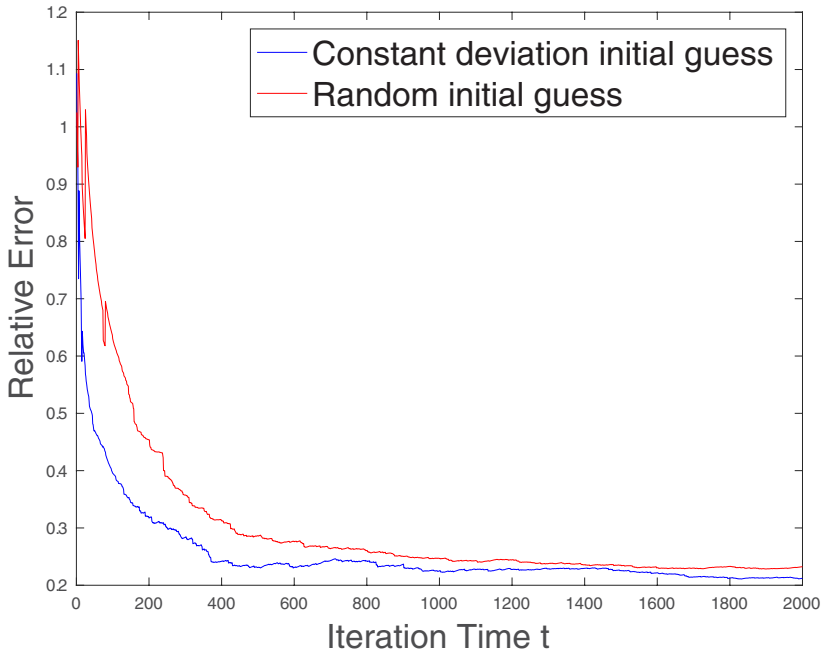

Figure 11. Absorption coefficient recovery. With respect to time steps, the relative error decays almost exponentially in time at the beginning with some fluctuation given by the stochastic nature of the algorithm.

5.3. Numerical cost study

We dedicate this subsection for comparing numerical cost of SGD and the classical GD method. Initial guess is set as $\sigma_0 = \sigma R$ with $R \in \mathbb{R}^{21\times 21}$ drew from uniform distribution $U([0.1, 3.1])$. Regularizer $\alpha = 1$ and learning rate $\eta_0 = 0.0044$. Both SGD and GD are used for the optimizer with the sample size N being 100, 200 and 400. The computation is terminated once error tolerance $\mathrm{TOL} = 0.2$ is reached, or maximum number of iteration is achieved. We set maximum number of iteration 2000 for SGD and 100 for GD.

5.4. Absorption coefficient recovery

We recover the absorption coefficient in this subsection following the strategy in remark 1. The scattering coefficient is set as $\sigma_s(x_1, x_2) = 1$ and the to-be-recovered absorption coefficient is set as:

$$\sigma_a(x_1, x_2) = \frac{1}{20}\left[1 + 8\exp\left(-10(x_1 - \frac{1}{4})^2 - 10(x_2 - \frac{1}{4})^2\right) + 4\exp\left(-10(x_1 - \frac{3}{4})^2 - 10(x_2 - \frac{3}{4})^2\right)\right] \tag{59}$$

as plotted in figure 1. 1000 data points $\{(\phi^{(j)}, \psi^{(j)}) : 1 \leqslant j \leqslant 1000\}$ are prepared. Numerically to run SGD, we set the regularization coefficient $\alpha = 1$, and the learning rate $\eta_n = \frac{\eta_0}{1+\eta_0\alpha n}$ with $\eta_0 = 0.0441$. Two initial guesses are made: one initial guess is a constant away from the true media $\sigma_0 = \sigma + 0.18$, and another being a random initial $\sigma_0 = \sigma R$. The numerical solution after 2000 iterations are presented in figures 9 and 10 for constant deviation and random deviation initial guesses respectively. In figure 11 we show the decay of relative errors with respect to the time steps.

Acknowledgments

The work of KC and QL is supported in part by a start-up fund of QL from UW-Madison and the National Science Foundation under the grant DMS-1619778. The work of JL is supported in part by the National Science Foundation under the grants DMS-1514826 and DMS-1107444: RNMS KI-Net.

ORCID iDs

Ke Chen https://orcid.org/0000-0003-2383-4289
Qin Li https://orcid.org/0000-0001-9210-8948

References

[1] Arridge S 1999 Optical tomography in medical imaging *Inverse Problems* **15** R41–93
[2] Bal G 2009 Inverse transport theory and applications *Inverse Problems* **25** 053001
[3] Bal G and Jollivet A 2009 Time-dependent angularly averaged inverse transport *Inverse Problems* **25** 075010
[4] Bal G, Langmore I and Monard F 2008 Inverse transport with isotropic sources and angularly averaged measurement *Inverse Problems Imaging* **2** 23–42
[5] Bal G and Monard F 2012 Inverse transport with isotropic time-harmonic sources *SIAM J. Math. Anal.* **44** 134–61
[6] Bottou L 2010 *Large-Scale Machine Learning with Stochastic Gradient Descent* (Heidelberg: Physica-Verlag HD) pp 177–86
[7] Bottou L 2012 *Stochastic Gradient Descent Tricks* (Berlin: Springer) pp 421–36
[8] Chen K, Li Q and Wang L 2018 Stability of stationary inverse transport equation in diffusion scaling *Inverse Problems* **34** 025004
[9] Cheng Y, Gamba I M and Ren K 2011 Recovering doping profiles in semiconductor devices with the Boltzmann–Poisson model *J. Comput. Phys.* **230** 3391–412
[10] Choulli M and Stefanov P 1998 An inverse boundary value problem for the stationary transport equation *Osaka J. Math.* **36** 87–104
[11] Dout S, Schmitt B, Lopes-Gautier R, Carlson R, Soderblom L, Shirley J and the Galileo NIMS Team 2001 Mapping So_2 frost on Io by the modeling of nims hyperspectral images *Icarus* **149** 107–32
[12] Egger H and Schlottbom M 2014 An lp theory for stationary radiative transfer *Appl. Anal.* **93** 1283–96
[13] Egger H and Schlottbom M 2015 Numerical methods for parameter identification in stationary radiative transfer *Comput. Optim. Appl.* **62** 67–83
[14] Epstein C 2007 *Introduction to the Mathematics of Medical Imaging* 2nd edn (Philadelphia, PA: SIAM)
[15] Feng Y, Li L and Liu J G 2018 Semi-groups of stochastic gradient descent and online principal component analysis: properties and diffusion approximations *Commun. Math. Sci.* accepted
[16] Haltmeier M, Leitao A and Scherzer O 2007 Kaczmarz methods for regularizing nonlinear ill-posed equations I: convergence analysis *Inverse Problems Imaging* **1** 289
[17] Konecny J, Qu Z and Richtarik P 2017 Semi-stochastic coordinate descent *Optim. Methods Softw.* **32** 993–1005
[18] Lehtikangas O, Tarvainen T, Kim A and Arridge S 2015 Finite element approximation of the radiative transport equation in a medium with piece-wise constant refractive index *J. Comput. Phys.* **282** 345–59
[19] Leitao A and Svaiter B F 2016 On projective landweberkaczmarz methods for solving systems of nonlinear ill-posed equations *Inverse Problems* **32** 025004
[20] Li Q and Tai C 2017 Stochastic modified equations and adaptive stochastic gradient algorithms *Proc. of the 34th Int. Conf. on Machine Learning* (*Proc. of Machine Learning Research,*

International Convention Centre (Sydney, Australia, 6–11 August 2017) vol 70) ed D Precup and Y W Teh (Cham: Springer) pp 2101–10
[21] Li Q and Wang L 2017 Implicit asymptotic preserving method for linear transport equation *Commun. Comput. Phys.* **22** 157–81
[22] Mandt S, Hoffman M D and Blei D M 2016 A variational analysis of stochastic gradient algorithms *Proc. the 33rd Int. Conf. on Int. Conf. on Machine Learning* vol 48 www.JMLR.org
[23] Montejo L D, Jia J, Kim H K, Netz U J, Blaschke S, Muller G A and Hielscher A H 2013 Computer-aided diagnosis of rheumatoid arthritis with optical tomography, part 1: feature extraction *J. Biomed. Opt.* **18** 076001
[24] Montejo L D, Jia J, Kim H K, Netz U J, Blaschke S, Muller G A and Hielscher A H 2013 Computer-aided diagnosis of rheumatoid arthritis with optical tomography, part 2: image classification *J. Biomed. Opt.* **18** 076002
[25] Moulines E and Bach F R 2011 Non-asymptotic analysis of stochastic approximation algorithms for machine learning *Advances in Neural Information Processing Systems 24* ed J Shawe-Taylor *et al* (Red Hook, NY: Curran Associates and Inc.) pp 451–9
[26] Natterer F 2001 *The Mathematics of Computerized Tomography* (Philadelphia, PA: SIAM)
[27] Needell D, Ward R and Srebro N 2014 Stochastic gradient descent, weighted sampling, and the randomized kaczmarz algorithm *Advances in Neural Information Processing Systems 27* ed Z Ghahramani *et al* (Red Hook, NY: Curran Associates and Inc.) pp 1017–25
[28] Ozen H C and Bal G 2016 Dynamical polynomial chaos expansions and long time evolution of differential equations with random forcing *SIAM/ASA J. Uncertain. Quantification* **4** 609–35
[29] Prieto K and Dorn O 2017 Sparsity and level set regularization for diffuse optical tomography using a transport model in 2d *Inverse Problems* **33** 014001
[30] Recht B, Re C, Wright S and Niu F 2011 Hogwild: a lock-free approach to parallelizing stochastic gradient descent *Advances in Neural Information Processing Systems 24* ed J Shawe-Taylor *et al* (Red Hook, NY: Curran Associates and Inc.) pp 693–701
[31] Ren K 2010 Recent developments in numerical techniques for transport-based medical imaging methods *Commun. Comput. Phys.* **8** 1–50
[32] Ren K, Zhang R and Zhong Y 2015 Inverse transport problems in quantitative pat for molecular imaging *Inverse Problems* **31** 125012
[33] Roux N L, Schmidt M and Bach F R 2012 A stochastic gradient method with an exponential convergence rate for finite training sets *Advances in Neural Information Processing Systems 25* ed F Pereira *et al* (Red Hook, NY: Curran Associates and Inc.) pp 2663–71
[34] Saratoon T, Tarvainen T, Cox B and Arridge S 2013 A gradient-based method for quantitative photoacoustic tomography using the radiative transfer equation *Inverse Problems* **29** 075006
[35] Stefanov P and Tamasan A 2009 Uniqueness and non-uniqueness in inverse radiative transfer *Proc. Am. Math. Soc.* 137 2335–44
[36] Strohmer T and Vershynin R 2008 A randomized kaczmarz algorithm with exponential convergence *J. Fourier Anal. Appl.* **15** 262
[37] Tang J, Han W and Han B 2013 A theoretical study for rte-based parameter identification problems *Inverse Problems* **29** 095002
[38] Tarvainen T, Kolehmainen V, Arridge S R and Kaipio J P 2011 Image reconstruction in diffuse optical tomography using the coupled radiative transportdiffusion model *J. Quant. Spectrosc. Radiat. Transfer* **112** 2600–8
[39] Wang J 1999 Stability estimates of an inverse problem for the stationary transport equation *Ann. Inst. Henri Poincare* **70** 473–95
[40] Zhang L, Mahdavi M and Jin R 2013 Linear convergence with condition number independent access of full gradients *Advances in Neural Information Processing Systems 26* ed C J C Burges *et al* (Red Hook, NY: Curran Associates and Inc.) pp 980–8