applied sciences

MDPI

Article

Image Reconstruction in Diffuse Optical Tomography Using Adaptive Moment Gradient Based Optimizers: A Statistical Study

Nada Chakhim †, Mohamed Louzar †, Abdellah Lamnii *,† and Mohammed Alaoui †

MISI Laboratory, Faculty of Science and Technology, University Hassan First, Settat 26000, Morocco; n.chakhim@uhp.ac.ma (N.C.); mohamed.louzar@uhp.ac.ma (M.L.); mohammed.alaoui@uhp.ac.ma (M.A.)

* Correspondence: abdellah.lamnii@uhp.ac.ma

† These authors contributed equally to this work.

Received: 4 December 2020; Accepted: 16 December 2020; Published: 20 December 2020

Abstract: Diffuse optical tomography (DOT) is an emerging modality that reconstructs the optical properties in a highly scattering medium from measured boundary data. One way to solve DOT and recover the quantities of interest is by an inverse problem approach, which requires the choice of an optimization algorithm for the iterative approximation of the solution. However, the well-established and proven fact of the no free lunch principle holds in general. This paper aims to compare the behavior of three gradient descent-based optimizers on solving the DOT inverse problem by running randomized simulation and analyzing the generated data in order to shade light on any significant difference—if existing at all—in performance among these optimizers in our specific context of DOT. The major practical problems when selecting or using an optimization algorithm in a production context for a DOT system is to be confident that the algorithm will have a high convergence rate to the true solution, reasonably fast speed and high quality of the reconstructed image in terms of good localization of the inclusions and good agreement with the true image. In this work, we harnessed carefully designed randomized simulations to tackle the practical problem of choosing the right optimizer with the right parameters in the context of practical DOT applications, and derived statistical results concerning rate of convergence, speed, and quality of image reconstruction. The statistical analysis performed on the generated data and the main results for convergence rate, reconstruction speed, and quality between three optimization algorithms are presented in the paper at hand.

Keywords: diffuse optical tomography; inverse problem; adaptive moment; image reconstruction; statistical simulation

1. Introduction

In recent years, the problem of DOT is becoming more attractive since it presents many advantages. It is a non-invasive, non-ionizing, and an inexpensive technique compared to other imaging modalities such as Magnetic Resonance Imaging (MRI) and X-ray [1–3]. DOT has been applied to detect breast tumors [4–7], brain injuries [8,9], imaging newborn infants' heads [10], and providing some important information about tissue metabolism. Solving the DOT problem involves addressing the radiative transfer equation (RTE) that describes the light propagation in biological tissues [11,12]. However, the RTE does not have an analytical close form solution for complex geometries, and its numerical alternative is computationally expensive. Since the diffusion approximation (DA) of the RTE is easy to implement, we will use it as the forward model throughout this work.

It is a well known fact that the inverse problem in DOT is nonlinear and severely ill-posed. Gradient-based methods are commonly used to solve minimization problems in optical tomography [11].

In recent years, a number of new optimizers have been proposed to tackle the problem of convergence when there is insufficient prior knowledge to elect a good learning rate. One of the most popular and practical techniques used to control the distance of each step. The Adaptive moment estimation (Adam) is one of the first adaptive moment optimizers proposed in literature and was presented by Diedriek Kingma and Jimmy Ba [13]. It is a combination of adaptive gradient algorithm (AdaGrad) [14] and Root Mean Square propagation with momentum (RMS prop) [15]. Adam is an efficient optimizer that only requires first order gradients and uses square gradients to scale the learning rate implementing momentum by using the moving average of the gradient rather than the gradient itself. To cope with the shortcoming of Adam, mainly the lack of convergence guarantees, a number of variants of Adam algorithm have been derived lately such as Nesterov-accelerated Adaptive Moment Estimation (Nadam) [16] and the AmsGrad optimizer [17]. For more details, we refer the reader to [18,19].

In this work, we will examine the convergence behavior of Adam, Nadam, and AmsGrad optimizers when applied to the problem of DOT. A comparison between these optimizers will be investigated and discussed. We will characterize the performance of these algorithms with respect to the choice of some hyperparameters and the initial guess error. To evaluate the quality of reconstructed images by the algorithms in quantitative manner, we use quality metrics, such as the structural similarity index (SSIM) and the peak signal-to-noise ratio (PSNR) on the reconstructed images.

The structure of this paper is as follows: In Section 2, we give an overview of the mathematical formulation of the diffusion approximation in continuous wave (CW) cases. In Section 3, we describe the inverse problem and the algorithms we use to reconstruct the absorption coefficient of DOT. In Section 4, we show the results of our statistical analysis of the simulation data. We present conclusions in Section 5.

2. Forward Problem

In this section, we describe the mathematical formulation of the diffusion approximation (DA).

Let $\Omega \subset R^n$, $n = 2, 3$ be our domain of interest, and $\partial\Omega$ the boundary of Ω. Then, the DA inside the domain Ω satisfies the partial differential equation

$$-\nabla[.D(r)\nabla\Phi(r)] + \mu_a(r)\Phi(r) = 0 \qquad r \in \Omega \tag{1}$$

with the Robin-boundary condition

$$\Phi(r) + 2a\, D(r)\frac{\partial\Phi(r)}{\partial\hat{n}} = S(r) \qquad r \in \partial\Omega \tag{2}$$

where $\Phi(r)$ is the photon density, and $D(r)$ is the diffusion coefficient defined by $D(r) = \frac{1}{3(\mu_a+\mu_s')}$. a is the Fresnel reflection coefficient, which depends on the mismatch between the refractive indices, μ_a and μ_s the absorption and scattering coefficient, respectively, and μ_s' the reduced scattering coefficient expressed as $\mu_s' = (1-g)\mu_s$, where g is the anisotropic factor. $S(r)$ describes the boundary condition for the incoming radiation and $\hat{n}$ is the outward normal vector to Ω.

We assume that the medium is highly scattering such that $\mu_a \ll \mu_s$.The forward model (1) and (2) is solved by using the finite element method as described in [20].

3. Inverse Problem

The inverse problem we are interested in consists of determining the couple (μ_a, μ_s) from the set of true data y_i such that

$$F_i(\mu_a, \mu_s) = y_i \qquad 1 \le i \le s \tag{3}$$

where we denote by F_i the forward operator which is assumed to be Fréchet differentiable, and y_i the approximate measured data. In this study, we restrict our attention to the reconstruction of the

absorption coefficient, and we assume that the distribution of the scattering coefficient is known. Then, the objective function can be written as follows:

$$J(\mu_a) = \frac{1}{2}\sum_{i=1}^{s}(F_i(\mu_a) - y_i)^2 \tag{4}$$

Then, this problem can be stated in terms of an optimization problem

$$\mu_a^* = argminJ(\mu_a) \tag{5}$$

Since the inverse problem is ill-posed, it requires regularization. A Total Variation regularization is applied [21]. By adding a regularization term, the cost function is formulated as

$$J_R(\mu_a; \lambda, \mu_a^0) = \frac{1}{2}\sum_{i=1}^{s}(F_i(\mu_a) - y_i)^2 + \lambda R(\mu_a) \tag{6}$$

where $R(\mu_a) = \|\mu_a - \mu_a^0\|^2$ is the regularization operator that enforces smoothness conditions in the solution, and λ is the regularization parameter. μ_a^0 denoted the initial guess error.The forward operator F_i is linearized around some initial guess μ_a^0.

$$F_i(\mu_a) = F_i(\mu_a^0) + F_i'(\mu_a^0)(\mu_a - \mu_a^0) + W(\mu_a^0, i) \tag{7}$$

where F_i' is the Fréchet derivative of the forward operator F_i, and W denotes the Taylor remainder for the linearization around μ_a^0.

The gradient of the objective functional can be written as follows:

$$\nabla J_R(\mu_a; \lambda, \mu_a^0) = \sum_{i=1}^{s} F_i'(\mu_a)^*(F_i(\mu_a) - y_i) + \lambda R'(\mu_a) \tag{8}$$

where $R'(\mu_a)$ is the Fréchet derivative of regularization operator with respect to μ_a.

4. Iterative Inverse Problem Solution

We consider an iterative optimization algorithm denoted by Q. The statement of our problem can be reduced to the iterative form:

$$\begin{cases} \mu_0 = \mu_a^0 \\ \mu_{n+1} = Q(\mu_n; J_R) \end{cases} \tag{9}$$

Naturally, the promise of the algorithm is to get us closer to the solution after each step in an iterative manner. The proof of convergence of any specific algorithm ensures that

$$\lim_{n \to +\infty} Q(\mu_n; J_R) = \mu_a^* \tag{10}$$

and can give even more information on the speed of convergence by deriving a theoretical formula of $\|Q(\mu_n; J_R) - \mu_a^*\|$ as a bounded formula of n.

In general, this is a hard formula to derive, and it is even more difficult when dealing with complex problems like DOT, with many multidimensional parameters. In practice, the convergence speed is influenced by many factors, related to the algorithm itself, and the configuration of the problem (physical reality and constraints). A numerical approach based on simulation and statistical analysis will prove to be very useful in tackling these kinds of hard situations, and can help us to gain more insight in the choice of optimization algorithm and all other practical purpose. As we consider in our study, a family of optimization algorithms based on gradient descent, we can point out the learning rate hyper parameter as the main factor of interest in this context.

From a practical point of view, J_R depends on the structure of the problem, and, consequently, J_R depends on different factors like the nature of inclusions (their number, form, distribution ...), the properties of the medium, and all other parameters that shape the above forward problem as stated in the previous section. In addition, it depends on the choice of the regularization and initial guess. Table 1 below gives an example of the parameters and hyperparameters that can be of interest in studying the practical optimization problem (including the iterative algorithm hyperparameters).

Table 1. Parameters and hyperparameters of interest.

Parameters of the Problem	Hyperparameters of the Optimization Algorithm
n: number of inclusions.	β: learning rate.
D: distribution shape of the inclusion.	μ_a^0: initial guess.
N_d : number of detectors.	λ: regularization coefficient.
N_s : number of sources.	

A more focused statement of the iterative optimization algorithm, for the following study in the present paper, can be formulated as

$$\begin{cases} \mu_0 = \mu_a^0 \\ \mu_{n+1} = Q_{AM}(\mu_n; n, \beta, \Theta) \end{cases} \tag{11}$$

where Q_{AM} describes the adaptive moment algorithm, μ_a^0 the initial guess, n the number of inclusions, β denotes the learning rate hyperparameter, and Θ represents all the remaining parameters.

Hereafter, we address our attention only to the number of inclusions (n), the learning rate β, and the initial guess μ_a^0. In our implementation of the optimization problem, we used an objective function C defined as

$$C(l; \beta, n, \mu_a^0) = \frac{1}{2}(J_R(Q_{AM}(\mu_l; n, \beta)) + \epsilon_0 + |J_R(Q_{AM}(\mu_l; n, \beta) - \epsilon_0|)) \tag{12}$$

We can easily show that

$$C(l; \beta, n, \mu_a^0) = \begin{cases} J_R(Q_{AM}(\mu_l; n, \beta)) & if J_R(Q_{AM}(\mu_l; n, \beta) > \epsilon_0 \\ \epsilon_0 & if J_R(Q_{AM}(\mu_l; n, \beta) \leq \epsilon_0 \end{cases} \tag{13}$$

We define the number of iterations to convergence by

$$N_{Q_{AM}} = min(argmax_l(C(l; \beta, n, \mu_a^0)), L_{max}) \tag{14}$$

This formulation guarantees that our optimization algorithm will stop whenever $J_R(Q_{AM}(\mu_l; n, \beta)$ is lower than ϵ_0 or l is greater than L_{max}, where $\epsilon_0 > 0$ and $L_{max} \geq 1$, are parameters used in iteration stopping criteria, which is explicitly set in this study to be either when the cost function is lower than ϵ_0 or the number of iteration exceeds L_{max}.

The aim of our numerical statistical study of convergence speed can then be brought down to the study of properties of the $N_{Q_{AM}}$ probability distribution $P(N_{Q_{AM}}|n, \beta, \mu_a^0)$ using simulation tools. In the following study, we restrict our attention to the comparison of three algorithms based on the adaptive moment procedure. For more details, we refer the reader to the next section.

5. Simulation and Data Generation

As mentioned before, only the absorption coefficient is reconstructed and discussed. The distribution of the scattering coefficient is assumed to be known. To generate synthetic data, we use the Toast++

software [22], which solves the forward problem (1) and (2) described above, using the finite element method. In all the numerical simulations, a circular domain of radius 20 mm which contains different inclusion sizes and shapes is performed. To avoid inverse crime [23], we use different meshes in the forward and inverse problem. In all cases, we use a circular mesh with 22,011 nodes and 43,400 tetrahedral elements for the forward problem and 15,408 nodes and 30,308 tetrahedral elements for the inverse problem. Sixteen sources and 16 detectors are located on the boundary of the domain with equal distance. The location, size, and number of anomalies in μ_a are chosen randomly with a background $\mu_a^{bkg} = 0.01$ mm^{-1} and $\mu_s'^{bkg} = 2$ mm^{-1}. We consider that there is no change in the anisotropic factor g, which is taken to be equal to 0.9. The regularization parameter λ is set to be equal to 10^{-8}. To solve the minimization problem (5), we use Algorithms 1–3, as described in pseudo codes below [18], where β is the learning rate, and ρ_1 and ρ_2 are the exponential decay rates for the moment estimates. The parameter of stabilization ϵ is set to be equal to 10^{-10}.

To control all the parameters of our simulation, we first control the error of the initial guess of reconstruction μ_a^0 by taking it to be

$$\mu_a^0 = \mu_a^{real} + \alpha \tag{15}$$

where μ_a^{real} is the original image matrix used to solve the forward problem, and α is a random matrix variable sampled uniformly such as $\| \alpha \|_{\text{inf}} = \delta$, where δ itself is a uniformly random number taken in range [0,0.2]. We define δ as the initial guess error.

Algorithm 1: Pseudo-code of Adam.

Require: μ_a^0, β, ρ_1, ρ_2, and ϵ with $\rho_1, \rho_2 \in [0, 1)$
Ensure: μ_a^k
  while J not converged **do**
    $k \leftarrow k + 1$
    $g_k \leftarrow \nabla J_{\mu_a}(\mu_a^{k-1})$
    $m_k \leftarrow \rho_1 . m_{k-1} + (1 - \rho_2) . g_k$
    $v_k \leftarrow \rho_1 . v_{k-1} + (1 - \rho_2) . g_k^2$
    $\hat{m}_k \leftarrow \frac{m_k}{(1-\rho_1^k)}$
    $\hat{v}_k \leftarrow \frac{v_k}{(1-\rho_2^k)}$
    $\mu_a^k \leftarrow \mu_a^{k-1} - \beta \frac{\hat{m}_k}{\sqrt{\hat{v}_k} + \epsilon}$
  end while

Algorithm 2: Pseudo-code of Nadam

Require: μ_a^0, β, ρ_1, ρ_2, and ϵ with $\rho_1, \rho_2 \in [0, 1)$
Ensure: μ_a^k
  while J not converged **do**
    $k \leftarrow k + 1$
    $g_k \leftarrow \nabla J_{\mu_a}(\mu_a^{k-1})$
    $m_k \leftarrow \rho_1 . m_{k-1} + (1 - \rho_2) . g_k$
    $v_k \leftarrow \rho_1 . v_{k-1} + (1 - \rho_2) . g_k^2$
    $\hat{m}_k \leftarrow \frac{m_k}{(1-\rho_1^k)}$
    $\hat{v}_k \leftarrow \frac{v_k}{(1-\rho_2^k)}$
    $\mu_a^k \leftarrow \mu_a^{k-1} - \frac{\beta}{\sqrt{\hat{v}_k} + \epsilon} (\rho_1 \hat{m}_k + \frac{1-\rho_1}{1-\rho_1^k} g_k)$
  end while

Algorithm 3: Pseudo-code of AmsGrad

Require: μ_a^0, β, ρ_1, ρ_2, and ϵ with $\rho_1, \rho_2 \in [0, 1)$
Ensure: μ_a^k
  while Jnot converged **do**
    $k \leftarrow k + 1$
    $g_k \leftarrow \nabla J_{\mu_a}(\mu_a^{k-1})$
    $m_k \leftarrow \rho_1.m_{k-1} + (1 - \rho_2).g_k$
    $v_k \leftarrow \rho_1.v_{k-1} + (1 - \rho_2).g_k^2$
    $\hat{v_k} \leftarrow \max(v_k, \hat{v_{k-1}})$
    $\mu_a^k \leftarrow \mu_a^{k-1} - \frac{\beta}{\sqrt{\hat{v_k}+\epsilon}} m_k$
  end while

The choice of the learning rate is very important. For this purpose, a preliminary study has been conducted where we experimented with the learning rate of all optimizers in the range [0, 0.5], and noticed that, when the learning rate is out of the interval [0.001, 0.3], the minimization of the objective function does not converge or take a very long time to do. For the sake of this simulation, the learning rate is constrained to be chosen uniformly from the range [0.001, 0.3]. We fixed all the other hyper parameters for all optimizers, to the recommended values from the corresponding literature (momentum ρ_1 and ρ_2 are 0.9 and 0.999, respectively). The number of anomalies is taken among 1, 2, and 3 equi-proportionally randomly.

Figure 1 shows the resulting distributions (histograms) from running the simulation, for the three parameters of the study, the learning rate β, the number of anomalies, and the perturbation coefficient δ. To be fair, we use the same parameters for all optimizers in each simulation instance.

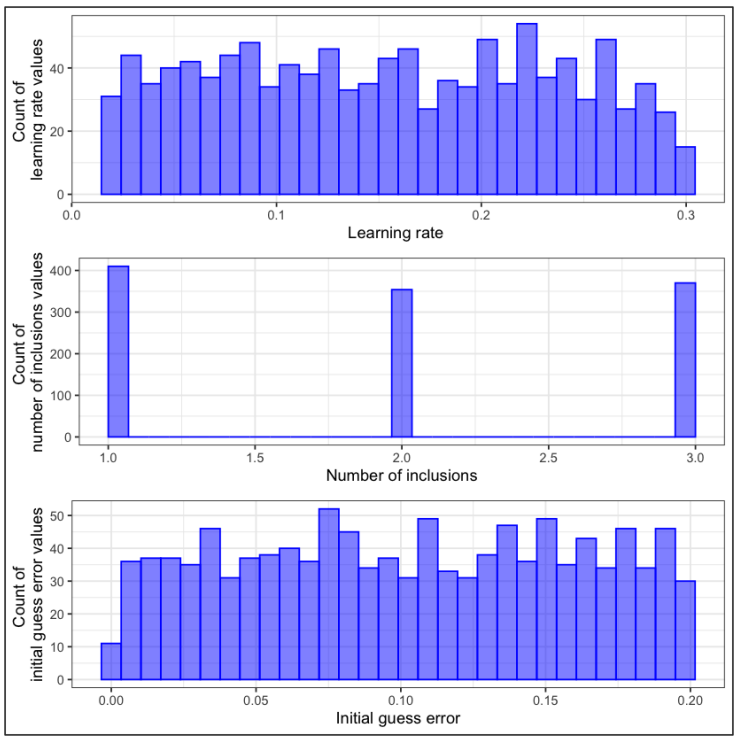


Figure 1. Distribution of the learning rate hyper parameter β on the top, number of inclusions n in the middle, and the initial guess error δ on the bottom.

6. Results

In this part of our study, we will characterize the convergence rate of the three algorithms, and compare the convergence/divergence behavior in relation to the parameters of simulation, and, finally, we will examine the quality of the resulting reconstructions of the three optimizers.

First of all, we choose in the context of this present analysis, the definition of divergence to be the state of the running optimization when the error minimization didn't improve for longer than 200 iterations in total.

The first subject of focus is the convergence rate of each algorithm. Let X_{AD}, X_{NAD}, and X_{AMS} be three random variables representing the state of convergence for Adam, Nadam, and AmsGrad optimizers. These variables take values 0 or 1, depending on either the corresponding algorithm diverges or converges, such that:

$$\begin{cases} X_{AD} = 0 \text{ if Adam diverge} \\ X_{AD} = 1 \text{ if Adam converge} \end{cases}$$

$$\begin{cases} X_{NAD} = 0 \text{ if Nadam diverge} \\ X_{NAD} = 1 \text{ if Nadam converge} \end{cases}$$

$$\begin{cases} X_{AMS} = 0 \text{ if AmsGrad diverge} \\ X_{AMS} = 1 \text{ if AmsGrad converge} \end{cases}$$

The simulation provided us with three samples of independent and identically distributed $(X_{AD,n})_{n\leq 1340}$, $(X_{NAD,n})_{n\leq 1340}$, and $(X_{AMS,n})_{n\leq 1340}$. To statistically estimate the rates of convergence, namely $P(X_{AD} = 1) = p_{AD}$, $P(X_{NAD} = 1) = p_{NAD}$, and $P(X_{AMS} = 1) = p_{AMS}$.

We use the three estimators

$$\begin{cases} \hat{p}_{AD} = \frac{\sum \#(X_{AD}=1)}{N} \\ \hat{p}_{NAD} = \frac{\sum \#(X_{NAD}=1)}{N} \\ \hat{p}_{AMS} = \frac{\sum \#(X_{AMS}=1)}{N} \end{cases}$$

where # denotes the count function, in order to construct 95% confidence intervals based on the large number normal approximation, as presented in Table 2.

Table 2. The 95% non-parametric confidence intervals.

Target Proportion	Confidence Interval Lower Bound	Confidence Interval Upper Bound
$\hat{p}_{AD}$	0.87	0.9
$\hat{p}_{NAD}$	0.87	0.91
$\hat{p}_{AMS}$	0.88	0.9

To shed light on the influence of simulation parameters on convergence rates, we run a logistic regression to estimate the conditional distributions $P(X|\beta, n, \delta)$, where $X \in \{X_{AD}, X_{NAD}, X_{AMS}\}$, n denotes the number of anomalies in the image, and the remaining variables β and δ are as mentioned earlier. The result of this procedure is depicted in Table 3 presenting p-values for the statistical significance of each regression parameter.

From Table 3, we conclude that the main parameter that also has a significant influence on convergence of these algorithms is the learning rate hyper-parameter. Since the logistic coefficient forβ is positive for Adam and Nadam, the larger the learning rate, the more guarantee there is for the algorithm to converge. This statement is reversed for AmsGrad, as we observe a negative coefficient for learning rate. The AmsGrad is also impacted negatively with the number of anomalies in the image.

Table 3. Logistic regression coefficients and corresponding p-values.

	Adam	**Nadam**	**AmsGrad**
β	$(1.17, < 2 \times 10^{-16})$	$(1.22, < 2 \times 10^{-16})$	$(-0.35, < 1.5 \times 10^{-5})$
n	$(-0.005, 0.6)$	$(-0.007, 0.44)$	$(-0.044, < 7.6 \times 10^{-16})$
δ	$(0.07, 0.6)$	$(0.1, 0.5)$	$(-0.016, 0.91)$

Running our experiment simulation provided us with 1340 convergent instances in total (this means where $(X_{AD} = 1, X_{NAD} = 1, X_{AMS} = 1)$), to evaluate the comparative performance between optimizers in terms of the speed of convergence as measured by the number of iterations taken by each optimizer to reach the solution, We will conduct a statistical analysis on the generated data, comparing first the speed globally between optimizers and then relating it to the variables of simulation such as the initial guess error, the choice of learning rate and the number of anomalies in the image. In addition, the influence of these variables on reconstructed image quality will be discussed, and PSNR and SSIM score values are calculated for each simulation instance; we kindly refer the reader to later discussions about reconstruction quality in this paper for more information on these scores.

A number of statistical methods have been applied and results are examined to describe the convergence speed behavior of each algorithm when applied to the inverse problem of DOT.

Image reconstruction in optical tomography is an ill-posed nonlinear inverse problem, the algorithms based on the gradient descent present no guarantee to converge to the global minima when there are local minima in the optimization problem at hand, the convergence point depends heavily on the choice of the starting point of the optimization, and, generally, these algorithms converge (depending also on the learning rate) to the nearest local minima to the initial starting point.

Image reconstruction in optical tomography is an ill-posed nonlinear inverse problem, the algorithms based on gradient descent present no guarantee to converge to the global minima when there are local minima in the optimization problem at hand, the convergence point depends heavily on the choice of the starting point of the optimization, and, generally, these algorithms converge (depending also on the learning rate) to the nearest local minima to the initial starting point.

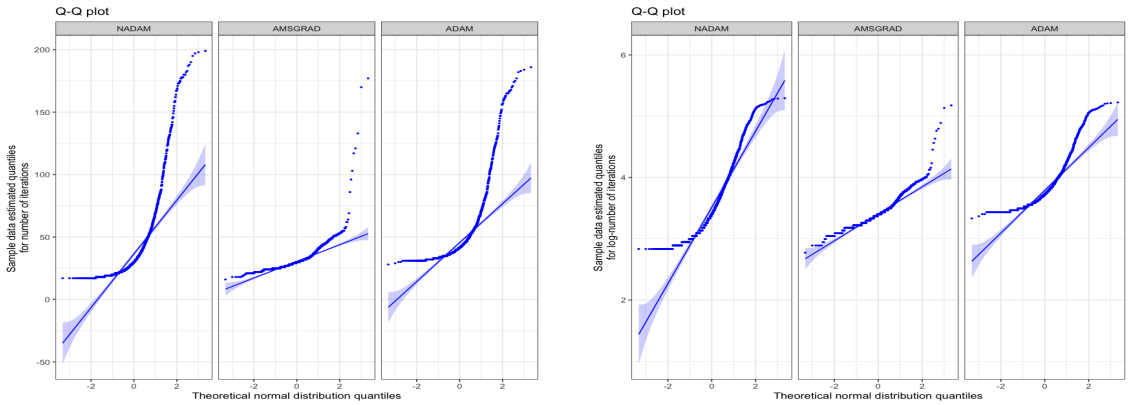
In this section, we address the optimization problem (image reconstruction) from the perspective of the speed of convergence (as one of the very important matters in practical use of DOT in clinical applications) rather than sensitivity of the algorithms to the choice of the initial guess with respect to their efficiency to find global minima (which is the other important practical issue in applying DOT); this last perspective is equally relevant and without a doubt needs particular attention and further analysis, but in the scope of our current paper remains an open question to follow up, as our randomized simulation design was focused on controlling the factors that influence speed of convergence. We can use the same approach as in this work to quantify (statistically speaking) the efficiency and sensitivity to reach the global minima depending on problem factors, but this obviously needs to redesign the simulation to generate the appropriate data suitable for this substantially different analysis objective.

The "blindness" toward the globality/locality character of the reached optimum for the gradient descent-based algorithms is an inherent property because the gradient is a local concept, and by itself carries only local information about the objective function which makes these algorithms very sensitive to the choice of learning rate and initialization. The adaptive moment included features does not add to the picture but some amount of "memory" of the recent gradients.

A rough observation that can be mentioned here is the fact that, in our generated sample data, most of the time the convergent instances for Nadam and Adam were to the global minima, but we can't really draw any statistical evidence from this naïve observation because our randomized simulation design does not support this analysis.

First of all, we check the distributions of number of iterations (speed) for normality, in the hope to be able to harness the large and powerful available parametric statistical approaches, from literature heavily relying on this (workhorse) normal distribution.

Probability distributions of speed of convergence and the log of speed of convergence are shown in the QQ plot described in Figure 2a,b, respectively. From these two graphs, it clearly appears that these distributions are very far from being reasonably considered normally or log-normally distributed. This is not a surprising fact indeed, knowing that these distributions are not symmetric to begin with, and look (strongly) skewed, but we wanted to exclude the possibilities of any approximate (left truncated) normal distributions.



(**a**) (**b**)

Figure 2. QQ-plot of speed of convergence and log of speed of convergence from left to right, for different optimizers.

Confirming this visual observation, the results of running Shapiro–Wilk normality tests on the three data samples are listed in Table 4. From Table 4, we conclude that the number of iterations for different optimizers significantly deviate from being normally distributed, and there is very little evidence, if none at all, that supports the normality. We did not test the goodness of fit for other density functions like Gumbel, Fréchet, and Weibul, even though the look of the distributions may suggest this family of extreme value distribution (EVD), mainly for two reasons:

First, those EVDs, even if approximately fitted to our empirical distribution, will not provide us, following our best judgment, with any advantage, considering the fact that the nature of exact distribution is not our main goal in itself, but rather is the distributions' locations, while all of the well known available parametric statistical methods for this purpose are based on the assumption that the samples come from (approximate) normal distribution.

Second, since we stopped the optimization iterations at 200 as mentioned above, we automatically lost information about the distribution in the extreme left part of the tail (which is almost 10% of the population according to the estimates in Table 2, for the three algorithms). This fact would certainly impact (heavily) the estimation of any EVD parameter, and, consequently, would reduce the power of any parametric test based on those inherently biased, and grossly approximate fits, which will minimize the comparative advantage of the eventual parametric over a non-parametric alternative method.

Table 4. Shapiro–Wilk normality test results.

Optimizers	Shapiro–Wilk [%]	*p*-Value
Nadam	0.71168	$<2.2 \times 10^{-16}$
Adam	0.68959	$<2.2 \times 10^{-16}$
AmsGrad	0.63334	$<2.2 \times 10^{-16}$

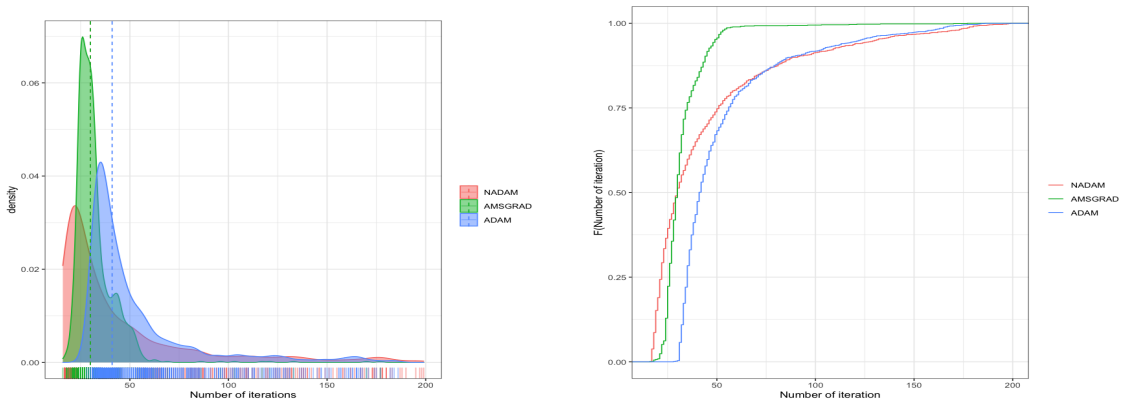
Following the arguments discussed above, we will use non-parametric statistical approaches to recover further information about the three optimizer performances from data, and, since the exact distributions are not well defined, we will use the empirical cumulative distribution as a legitimate approximation.

From the superposition of the three optimizers' empirical densities and cumulative densities functions of speed of convergence, as shown in Figure 3a,b, respectively, we note the differences in the central tendencies of the speed of convergence for the three optimizers, and we remark that the minimization of the objective function converges faster in the case of AmsGrad algorithm in comparison to the other two algorithms. To gain more credible evidence about these preliminary raw observations, we conducted a Kruskal–Wallis paired test [24] to elicit any significant difference of means among the three optimizers. Results of the tests are included in a box plot shown in Figure 4 with p-values. We can conclude with high confidence that there is a significant difference ($p < 0.05$) between the speed of convergence for the three optimizers. Comparing the means of number of iterations between each two algorithms individually, and especially between Adam and AmsGrad that look very close (mean wise), we conclude that there is a significant difference between these two groups too.

To frame these differences in speed between the three algorithms, we generate the 95% confidence intervals for the median differences using the bootstrap method with 10,000 replicates each. Normal, Percentile, and pivotal 95% confidence intervals have been calculated. Results are summarized in Table 5. From this table, we spot a clear advantage of Nadam and AmsGrad over Adam in the speed of convergence (on average) while the difference between AmsGrad and Nadam is around just four steps.

Table 5. The 95% non-parametric confidence intervals based on the bootstrap method.

Groups	Point Estimation	Standard Error	Normal	Percentile	Pivotal
AmsGrad vs. Nadam	4	0.36	(3.29, 4.71)	(3, 4)	(4, 5)
Adam vs. Nadam	12	0.5	(11.04, 12.96)	(11, 12)	(12, 13)
AmsGrad vs. Adam	15	0.44	(14.13, 15.87)	(15, 16)	(14, 15)



(**a**) (**b**)

Figure 3. (**a**) Densities of number of iteration for each optimizer and (**b**) the empirical cumulative density functions.

Following the logic of our study, we investigate the relationship between speed of convergence and each of the three factors of the simulation, namely the number of anomalies in image, the initial guess error, and the choice of the learning rate. To verify the impact of number of anomalies on the speed of convergence, the Kruskal–Wallis test is applied on each algorithm speed of convergence sample data, as grouped by the number of inclusions. Kruskal–Wallis test results are presented in Figure 5, and we can conclude (by failing to reject the Kruskal–Wallis null hypothesis) that the number

of anomalies present in the image is not significantly affecting the speed of convergence ($p > 0.05$) for different optimizers.

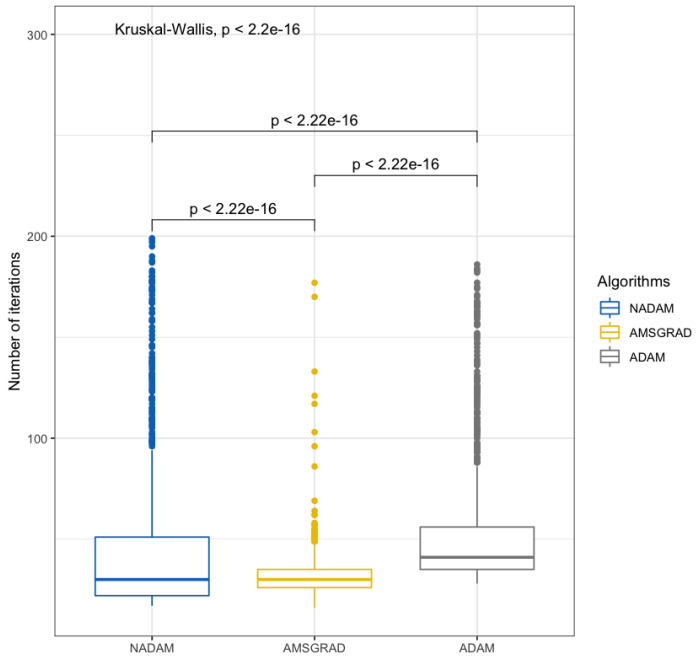


Figure 4. Kruskal–Wallis paired test for the number of iterations between different groups of optimizers with the resulting *p*-value of the test for each group.

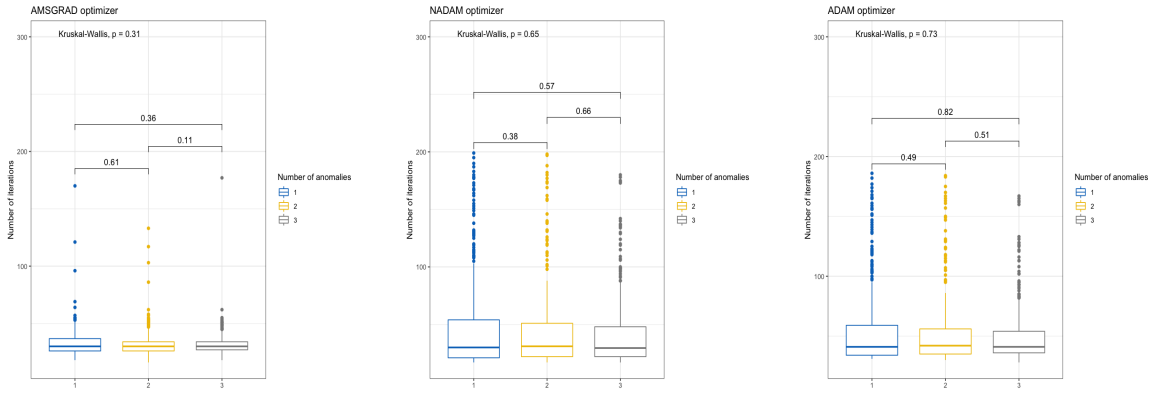
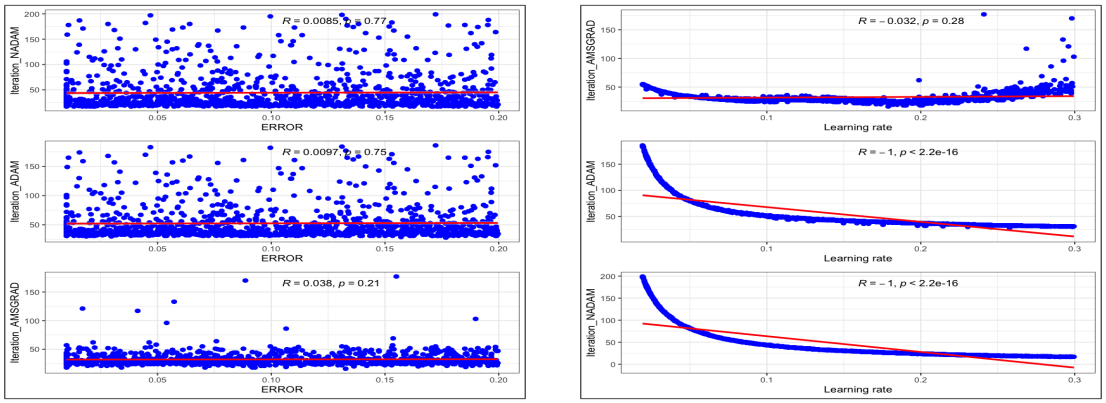


Figure 5. Kruskal–Wallis test on number of anomalies for different optimizers with the resulting *p*-value of the test for each group.

To fulfill our investigation, we discuss the impact of initial guess error and learning rate parameter over number of iterations as shown in Figure 6a and Figure 6b, respectively. The Spearman's coefficient of correlation is used due to its robustness against outliers which appears in data. Scatter plots in Figure 6a,b show the relationship between the initial guess error and the learning rate parameter on the speed of convergence, respectively. Spearman's coefficient of correlation R and *p*-value are mentioned at the top of each graph. From Figure 6b, we notice that, when the learning rate ranges in [0.001, 0.2], Nadam and Adam algorithms take more iterations than the AmsGrad algorithm. In addition, we note that the AmsGrad algorithm presents some robustness toward the learning rate in this range and presents some outliers in the range [0.001, 0.2]. According to the Spearman's correlation coefficient, we observe a very strong correlation ($R = -1$) between learning rate parameter and number of

iterations for Adam and Nadam optimizers and a negligible correlation for the case of AmsGrad optimizer and presents some outliers when the learning rate ranges in [0.2, 0.3]. On the other hand, Figure 6a shows the relationship between the initial guess error and number of iterations taken by each optimizer to reach convergence of cost functional. We note that the error has the same impact on Adam and Nadam algorithms, when comparing their p-value and coefficient of correlation. However, we observe that the AmsGrad is more efficient than the other two optimizers even if the error is far from the real image.



(**a**) (**b**)

Figure 6. Correlation between number of iteration of each optimizer and (**a**) initial guess error δ (**b**) learning rate hyper parameter β.

To assess the quality performance in reconstructed images between these optimizers, we performed statistical tests for differences of means on PSNR and SSIM as measured for reconstructed images, between the optimizers. These two scores are defined as follows:

$$PSNR = 10\log_{10}\left(\frac{max^2(\mu_a^{true})}{\frac{1}{N}\sum_{i=1}^{N}(\mu_a^{recon}(i) - \mu_a^{true}(i))^2}\right),$$

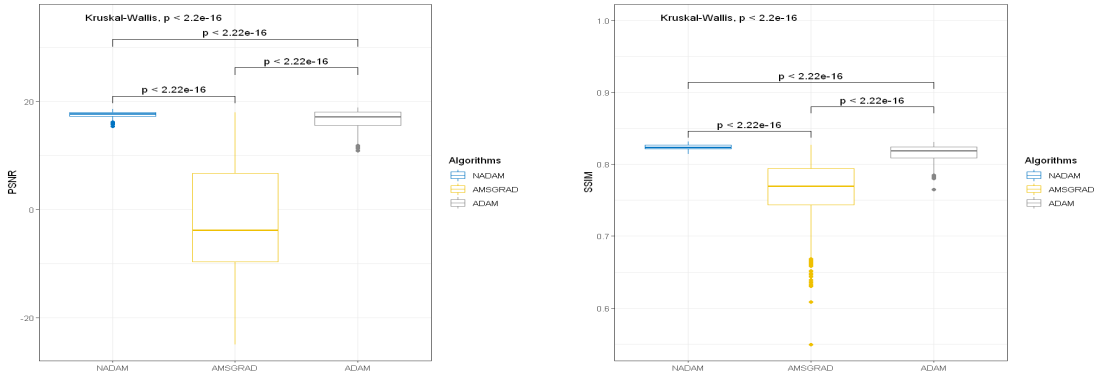
$$SSIM(\mu_a^{true}, \mu_a^{recon}) = [l(\mu_a^{true}, \mu_a^{recon})]^x + [c(\mu_a^{true}, \mu_a^{recon})]^y + [s(\mu_a^{true}, \mu_a^{recon}]^z,$$

$$\begin{cases} l(\mu_a^{true}, \mu_a^{recon}) = \frac{(2\bar{m}_{\mu_a^{true}}\bar{m}_{\mu_a^{recon}} + C_1)}{(\bar{m}^2_{\mu_a^{true}} + \bar{m}^2_{\mu_a^{recon}} + C_1)} \\ c(\mu_a^{true}, \mu_a^{recon}) = \frac{(2\sigma_{\mu_a^{true}}\sigma_{\mu_a^{recon}} + C_1)}{(\sigma^2_{\mu_a^{true}} + \sigma^2_{\mu_a^{recon}} + C_2)} \\ s(\mu_a^{true}, \mu_a^{recon}) = \frac{(\sigma_{\mu_a^{true}\mu_a^{recon}} + C_3)}{(\sigma_{\mu_a^{true}}\sigma_{\mu_a^{recon}} + C_3)} \end{cases}$$

where $l(\mu_a^{true}, \mu_a^{recon})$, $c(\mu_a^{true}, \mu_a^{recon})$, and $s(\mu_a^{true}, \mu_a^{recon})$ are the luminance, contrast, and structure variations between the true image μ_a^{true} and reconstructed image μ_a^{recon}, respectively, and $x > 0$, $y > 0$, and $z > 0$ are three parameters used to adjust relative importance of the three components of the similarity measure. $\bar{m}_{\mu_a^{true}}$ and $\bar{m}_{\mu_a^{recon}}$ are the means of pixel values of μ_a^{true} and μ_a^{recon}, respectively. We denote by $\sigma_{\mu_a^{true}}$, $\sigma_{\mu_a^{recon}}$, and $\sigma_{\mu_a^{true}\mu_a^{recon}}$ the standard deviation of μ_a^{true} and μ_a^{recon}, and the covariance of image μ_a^{true} and μ_a^{recon}, respectively. C_1, C_2, and C_3 are constants.

The global comparison of quality of reconstructed images are shown in Figure 7a,b, as the result of running the Kruskal–Wallis for PSNR (we eliminated AmsGrad outliers where $PSNR < -25$ db) and SSIM, grouping each score sample by optimizer. From Figure 7, it appears that the PSNR and the

SSIM of AmsGrad are much lower (worse quality) than those of Nadam and Adam. In addition, we can observe that the means of PSNR and SSIM for Adam and Nadam are very close.



(**a**) (**b**)

Figure 7. Kruskal–Wallis test for (**a**) PSNR and (**b**) SSIM for different groups of optimizers, with the statistical significance p-value for each group. The simulation cases with $PSNR < -25$ db are excluded.

To evaluate the influence of number of inclusions on image quality, we conduct a Wilcoxon test [25]. The test was applied according to different groups of numbers of inclusions. The resulting p-values of this test are summarized in Table 6. The results analysis shows that there is a significant statistical difference between means due to the difference in number of inclusion present in images (p-value < 0.05).

Table 6. Summary of Wilcoxon test between different number of anomalies for each optimizer.

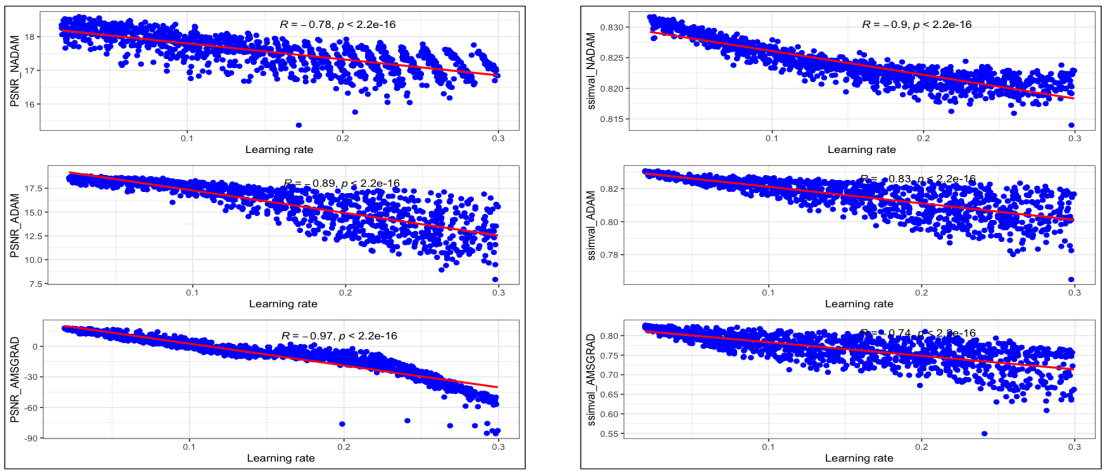
	Adam			**Nadam**			**AmsGrad**		
	1 vs. 2	**2 vs. 3**	**1 vs. 3**	**1 vs. 2**	**2 vs. 3**	**1 vs. 3**	**1 vs. 2**	**2 vs. 3**	**1 vs. 3**
PSNR	0.02	2.5×10^{-4}	1.8×10^{-6}	0.002	0.004	7.8×10^{-11}	0.003	0.031	0.001
SSIM	6.4×10^{-5}	1.2×10^{-4}	1.2×10^{-14}	0.008	0.02	0.03	$<2.2 \times 10^{-16}$	$<2.2 \times 10^{-16}$	$<2.2 \times 10^{-16}$

A similar conclusion is deduced about the influence of learning rate on PSNR and SSIM. Scatter plots in Figure 8a,b clearly show this strong influence of learning rate on PSNR and SSIM, respectively. The resulting Spearman's correlation coefficients by optimizer (and the corresponding p-value) for PSNR and SSIM are mentioned at the top of each graph. As shown in Figure 8a, we notice that there is a strong negative correlation between learning rate hyper-parameter and PSNR for the case of Adam and AmsGrad. For the case of Nadam, we note a moderate negative correlation between the choice of learning rate and PSNR of reconstructed images. From Figure 8b, we observe that there is a strong negative correlation between learning rate parameter and SSIM for the case of Adam and Nadam. In addition, there is a moderate negative correlation between learning rate and SSIM in the case of AmsGrad. The resulting p-values mentioned at the top of each graph indicate that these correlations are statistically significant, and, consequently, we can conclude the same about the significance of the influence of learning rate choice on the quality of the resulting reconstructed image. Thus, a small value of learning rate that ranges between 0.001 and 0.2 is recommended.

Concerning initial guess error, scatter plots in Figure 9 demonstrate the influence of initial guess error on reconstructed image quality. From Figure 9a,b, the obtained results show that there is no significant statistical differences between the initial guess error and resulting quality (PSNR/SSIM). Thus, we can conclude with high confidence ($p > 0.05$) that the image quality is only influenced by the number of anomalies in the image and the choice of the learning rate.

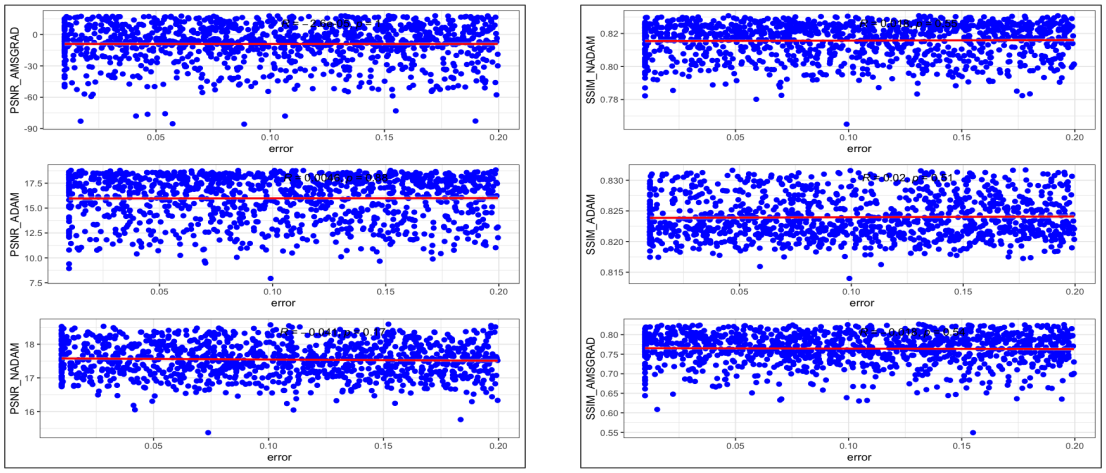
We illustrate some cases from our simulation. Figure 10 shows the reconstructed absorption coefficient μ_a for the case of one inclusion for an initial guess error equal to $\delta = 0.2$. Different values

of learning rate are used. The background of true images are taken equal to $\mu_a^{bck} = 0.01$ mm^{-1} and $\mu_s^{'bck} = 2$ mm^{-1}. The reconstruction using Nadam and Adam showed a good localization of inclusion. In addition, its size is the same compared to the true image with optical properties close to those of true image values. Some artifacts are observed in the borders close to sources and detectors region when the learning rate is higher than 0.1. For the case of AmsGrad reconstruction, we observe that the size of reconstructed image matches those for the true image with some artifacts in the center when the learning rate is lower than 0.1. However, when the learning rate is greater than 0.1, we remark that AmsGrad can localize the inclusion, but with some artifacts in the borders. The size and the shape of inclusion do not match those in the true image. Figure 11 shows the reconstructed absorption coefficient μ_a for the case of two inclusions with different shapes for the same values of initial guess error and optical properties used in the first case of one inclusion. From Figure 11, we notice that we obtain a good localization of both inclusions for the case of Nadam and Adam for different values of learning rates. However, when the learning rate is higher than 0.01, we observe some artifacts near the borders. For the case of AmsGrad reconstruction, it is clear that, when the learning rate exceeds 0.1, the size and the shape of inclusion do not match those figuring in the true image.



(**a**) (**b**)

Figure 8. Correlation between learning rate and (**a**) PSNR and (**b**) SSIM for different optimizers. The resulting Spearman's coefficient R and the p-value are shown at the top of each graph.



(**a**) (**b**)

Figure 9. Correlation between initial guess error and (**a**) PSNR and (**b**) SSIM for different optimizers. The resulting Spearman's coefficient R and the p-value are shown at the top of each graph.

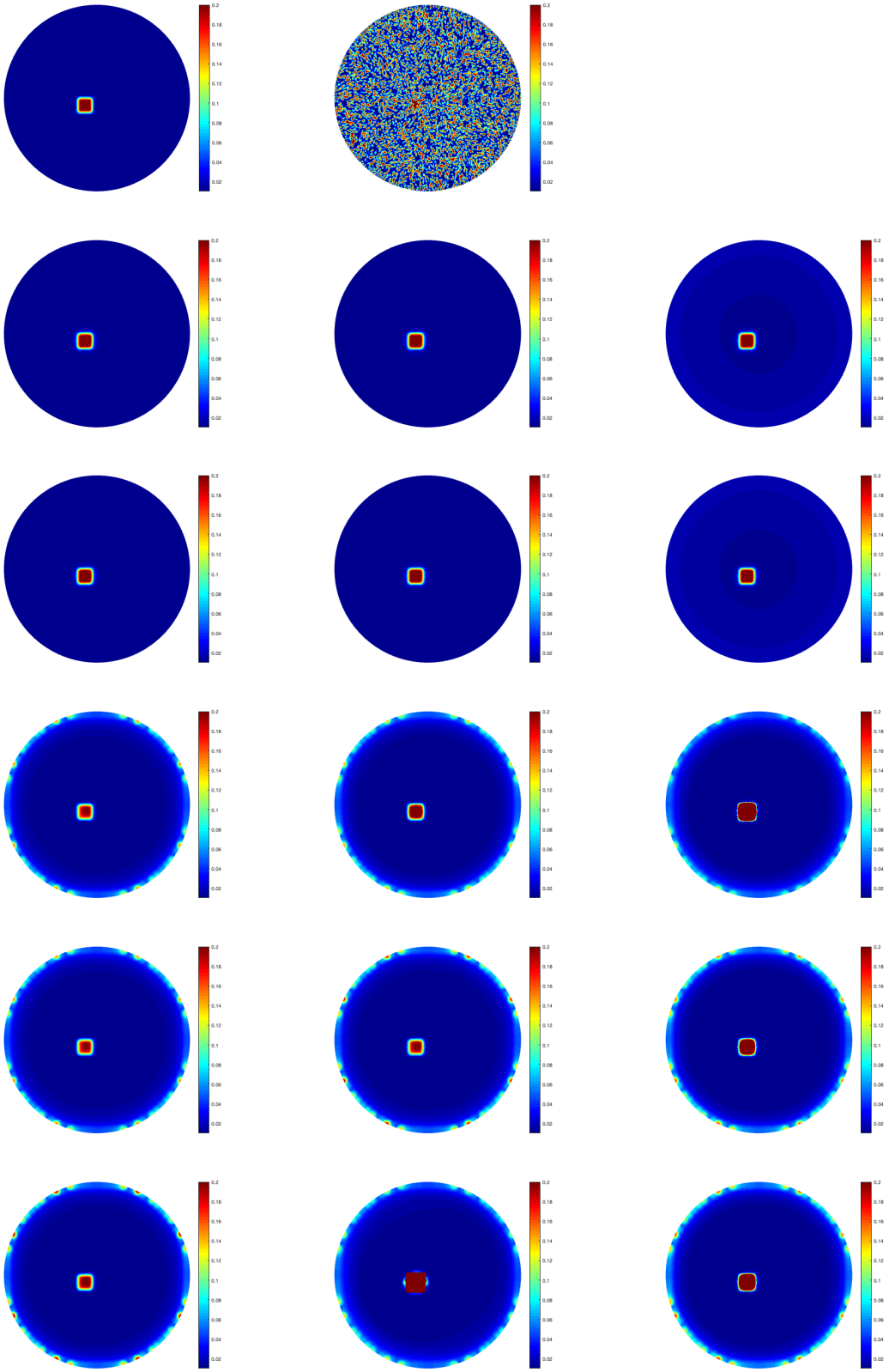


Figure 10. Reconstruction of the absorption coefficient μ_a with one inclusion. The first row presents the true image (**left**) and initial guess image with an initial guess error $\delta = 0.2$ (**right**). The second, third, fourth, and fifth rows present the reconstruction images using the learning rate β taking values equal to 0.001,0.01,0.1, 0.2, and 0.3, respectively, using Nadam, Adam, and AmsGrad, from (**left** to **right**).

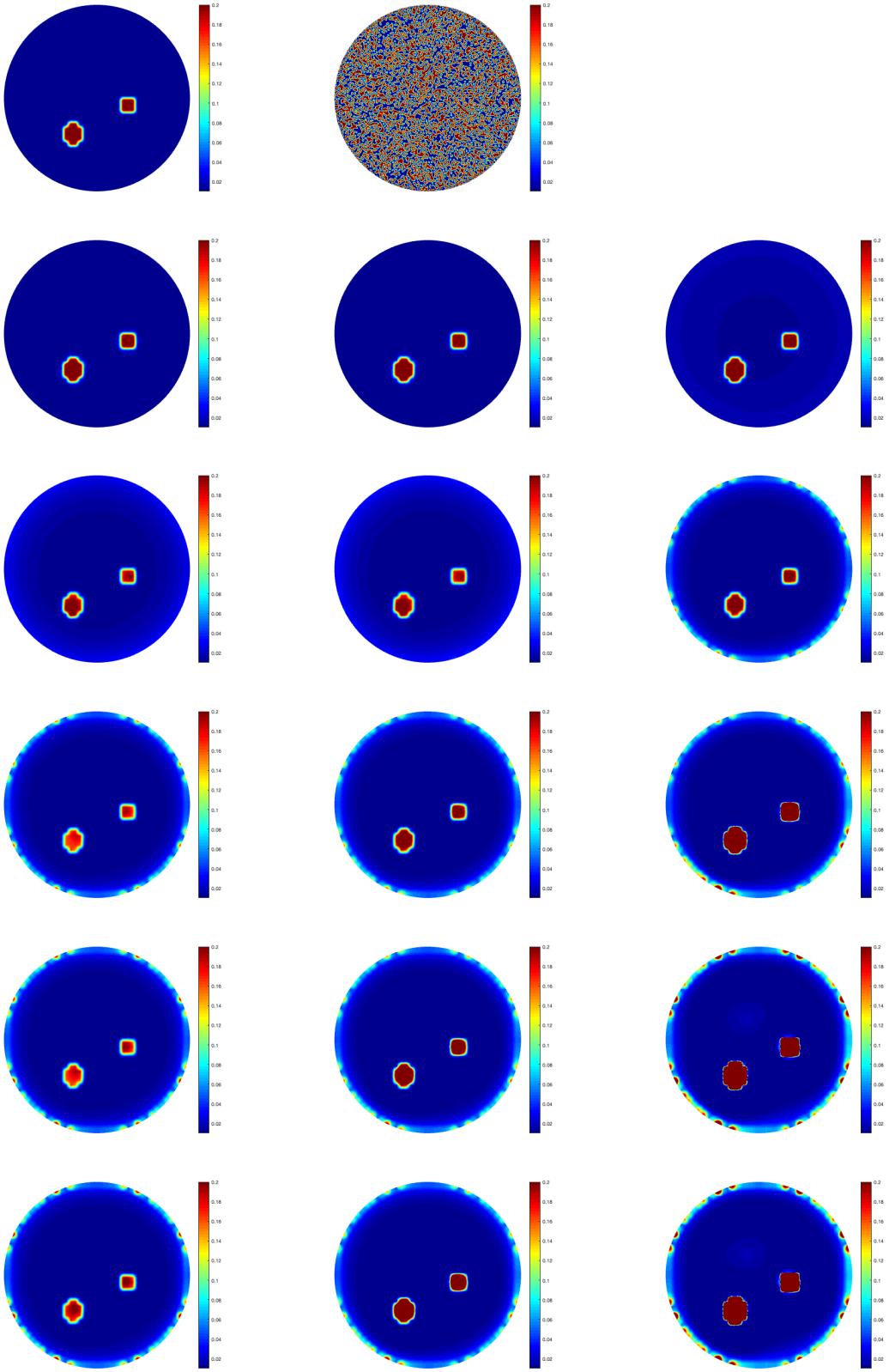


Figure 11. Reconstruction of the absorption coefficient $\mu_a$with two inclusions. The first row presents the true image (**left**) and initial guess image with an initial guess error $\delta = 0.2$ (**right**). The second, third, fourth, and fifth rows present the reconstruction images using the learning rate β taking values equal to 0.001,0.01, 0.1, 0.2, and 0.3, respectively, using Nadam, Adam, and AmsGrad, from (**left** to **right**).

7. Discussion and Conclusions

This research work analyzed the behavior of three optimizers when applied to the inverse problem of DOT regarding the speed of convergence and quality of reconstruction. The three optimizers under study, namely Nadam, Adam, and AmsGrad, are enhanced versions of the simple gradient descent algorithm, and have proved to perform very well in solving optimization problems in other areas of applications, especially in Deep Learning model search. The study we performed is based on a carefully designed randomized numerical simulation that aimed to gain credible statistical evidence on the actual performance of these optimizers when applied to solving the DOT inverse problem. We focused our attention on the impact of number of inclusions, the learning rate choice, and initial guess error on the speed of convergence. We also considered the impact of these same parameters on the quality of image reconstruction. The results derived using mainly non-parametric statistical approaches provide a scientifically credible quantification of the actual performance of these optimizers, with respect to the choice of learning rate, and under the constraint of the true numbers of inclusions and the arbitrariness of the initial starting point of the optimization.

The study provided valuable guidelines in terms of statistical evidence of the importance of the good choice of the learning rate for the three algorithms, and statistically proved the robustness of Nadam and Adam to the initial guess and the number of inclusions; these results can help improve and promote further the application of DOT in practical medical applications. However, we did not study the impact of these parameters on the simultaneous reconstruction of the absorption and scattering coefficients.

In future work, we aim to combine Nadam and AmsGrad and construct an algorithm that switches back and forth between AmsGrad and Nadam in a controlled fashion, where the AmsGrad optimizer will be used to accelerate the speed of convergence and Nadam to obtain a good quality of reconstructed images.

Author Contributions: N.C. conceived the idea, executed the simulations, analyzed the data, and wrote the paper. M.L., A.L. and M.A. supervised the findings of this work, provided ideas and directions, analyzed the data, and reviewed the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Dai, X.; Zhang, T.; Yang, H.; Tang, J.; Carney, P.R.; Jiang, H. Fast noninvasive functional diffuse optical tomography for brain imaging. *J. Biophotonics* **2018**, *11*, e201600267. [CrossRef] [PubMed]
2. Zimmermann, B.B.; Deng, B.; Singh, B.; Martino, M.; Selb, J.J.; Fang, Q.; Sajjadi, A.Y.; Cormier, J.A.; Moore, R.H.; Kopans, D.B.; et al. Multimodal breast cancer imaging using coregistered dynamic diffuse optical tomography and digital breast tomosynthesis. *J. Biomed. Opt.* **2017**, *22*, 046008. [CrossRef] [PubMed]
3. Yoo, J.; Sabir, S.; Heo, D.; Kim, K.H.; Wahab, A.; Choi, Y.; Lee, S.I.; Chae, E.Y.; Kim, H.H.; Bae, Y.M.; et al. Deep learning diffuse optical tomography. *IEEE Trans. Med. Imaging* **2019**, *39*, 877–887. [CrossRef] [PubMed]
4. Cochran, J.M.; Busch, D.R.; Lin, L.; Minkoff, D.L.; Schweiger, M.; Arridge, S.; Yodh, A.G. Hybrid time-domain and continuous-wave diffuse optical tomography instrument with concurrent, clinical magnetic resonance imaging for breast cancer imaging. *J. Biomed. Opt.* **2019**, *24*, 051409. [CrossRef] [PubMed]
5. Vavadi, H.; Mostafa, A.; Zhou, F.; Uddin, K.S.; Althobaiti, M.; Xu, C.; Bansal, R.; Ademuyiwa, F.; Poplack, S.; Zhu, Q. Compact ultrasound-guided diffuse optical tomography system for breast cancer imaging. *J. Biomed. Opt.* **2018**, *24*, 021203. [CrossRef] [PubMed]
6. Taroni, P.; Pifferi, A.; Quarto, G.; Spinelli, L.; Torricelli, A.; Abbate, F.; Villa, A.M.; Balestreri, N.; Menna, S.; Cassano, E.; et al. Noninvasive assessment of breast cancer risk using time-resolved diffuse optical spectroscopy. *J. Biomed. Opt.* **2010**, *15*, 060501. [CrossRef] [PubMed]
7. Zhu, Q.; Hegde, P.U.; Ricci, A., Jr.; Kane, M.; Cronin, E.B.; Ardeshirpour, Y.; Xu, C.; Aguirre, A.; Kurtzman, S.H.; Deckers, P.J.; et al. Early-stage invasive breast cancers: Potential role of optical tomography with US localization in assisting diagnosis. *Radiology* **2010**, *256*, 367–378. [CrossRef]

8. Ferradal, S.L.; Liao, S.M.; Eggebrecht, A.T.; Shimony, J.S.; Inder, T.E.; Culver, J.P.; Smyser, C.D. Functional imaging of the developing brain at the bedside using diffuse optical tomography. *Cereb. Cortex* **2016**, *26*, 1558–1568. [CrossRef]
9. Hernandez-Martin, E.; Gonzalez-Mora, J.L. Diffuse Optical Tomography Using Bayesian Filtering in the Human Brain. *Appl. Sci.* **2020**, *10*, 3399. [CrossRef]
10. Lee, C.W.; Cooper, R.J.; Austin, T. Diffuse optical tomography to investigate the newborn brain. *Pediatr. Res.* **2017**, *82*, 376–386. [CrossRef]
11. Arridge, S.R.; Schotland, J.C. Optical tomography: Forward and inverse problems. *Inverse Probl.* **2009**, *25*, 123010. [CrossRef]
12. Klose, A.D.; Hielscher, A.H. Iterative reconstruction scheme for optical tomography based on the equation of radiative transfer. *Med. Phys.* **1999**, *26*, 1698–1707. [CrossRef] [PubMed]
13. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
14. Duchi, J.; Hazan, E.; Singer, Y. Adaptive subgradient methods for online learning and stochastic optimization. *J. Mach. Learn. Res.* **2011**, *12*. [CrossRef]
15. Tieleman, T.; Hinton, G. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA Neural Netw. Mach. Learn.* **2012**, 4, 26–31.
16. Reddi, S.J.; Hefny, A.; Sra, S.; Poczos, B.; Smola, A. Stochastic variance reduction for nonconvex optimization. In Proceedings of the International Conference on Machine Learning, New York, NY, USA, 20–22 June 2016; pp. 314–323.
17. Tran, P.T.; Phong, L.T. On the convergence proof of amsgrad and a new version. *IEEE Access* **2019**, *7*, 61706–61716. [CrossRef]
18. Ruder, S. An overview of gradient descent optimization algorithms. *arXiv* **2016**, arXiv:1609.04747.
19. Zhou, D.; Tang, Y.; Yang, Z.; Cao, Y.; Gu, Q. On the convergence of adaptive gradient methods for nonconvex optimization. *arXiv* **2018**, arXiv:1808.05671.
20. Arridge, S.; Schweiger, M.; Hiraoka, M.; Delpy, D. A finite element approach for modeling photon transport in tissue. *Med Phys.* **1993**, *20*, 299–309. [CrossRef]
21. Rodrıguez, J.A.O. Regularization Methods for Inverse Problems. Ph.D. Thesis, University of Minnesota, Minneapolis, MN, USA, March 2011.
22. Schweiger, M.; Arridge, S.R. The Toast++ software suite for forward and inverse modeling in optical tomography. *J. Biomed. Opt.* **2014**, *19*, 040801. [CrossRef]
23. Colton, D.; Kress, R. *Inverse Acoustic and Electromagnetic Scattering Theory*; Springer Nature: New York, NY, USA, 2019; Volume 93.
24. Kruskal, W.H.; Wallis, W.A. Use of ranks in one-criterion variance analysis. *J. Am. Stat. Assoc.* **1952**, *47*, 583–621. [CrossRef]
25. Gehan, E.A. A generalized two-sample Wilcoxon test for doubly censored data. *Biometrika* **1965**, *52*, 650–653. [CrossRef] [PubMed]

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).