Medical Image Analysis
Manuscript Draft

Manuscript Number: MEDIA-D-16-00587

Title: Averaged stochastic optimization for medical image registration based on variance reduction

Article Type: Research Paper

Keywords: Image registration; Stochastic optimization; Polyak averaging; Variance reduction

Abstract: In image registration the optimal transformation parameters of a given transformation model are typically obtained by minimizing a cost function. Stochastic gradient descent (SGD) is an efficient optimization algorithm for image registration. In SGD optimization, stochastic approximations of the cost function derivative are used in each iteration to update the transformation parameters. The stochastic approximation error leads to large variance in the parameters. To enforce convergence nonetheless, SGD methods are typically implemented in combination with a gradually decreasing update step size. However, selecting a proper sequence of step sizes is a major challenge in practice. An alternative strategy in numerical optimization is to use a constant step size and enforce convergence by averaging the parameters obtained by SGD over several iterations. It was proven mathematically that the highest possible rate of convergence is achieved in this way. Inspired by this work, we propose an averaged SGD (Avg-SGD) method for efficient image registration. In the Avg-SGD approach, a constant step size is used, in combination with an exponentially weighted iterate averaging scheme. The Avg-SGD method is suitable for both rigid and nonrigid registration problems. Experiments on simulated 2D brain MRI data and real 3D lung CT scans demonstrate the effectiveness of the Avg-SGD method in terms of convergence rate, accuracy and precision.

Averaged stochastic optimization for medical image registration based on variance reduction

Wei Sun[a,d,*], Dirk H.J. Poot[a,b], Xuan Yang[c], Wiro J. Niessen[a,b], Stefan Klein[a]

[a] *Biomedical Imaging Group Rotterdam, Departments of Radiology and Medical Informatics, Erasmus MC, Rotterdam, The Netherlands*
[b] *Department of Image Science and Technology, Faculty of Applied Sciences, Delft University of Technology, Delft, The Netherlands*
[c] *College of Computer Science and Software Engineering, Shenzhen University, China*
[d] *Laboratory of Neuro Imaging (LONI), Keck School of Medicine of University of Southern California, Los Angeles, USA*

Abstract

In image registration the optimal transformation parameters of a given transformation model are typically obtained by minimizing a cost function. Stochastic gradient descent (SGD) is an efficient optimization algorithm for image registration. In SGD optimization, stochastic approximations of the cost function derivative are used in each iteration to update the transformation parameters. The stochastic approximation error leads to large variance in the parameters. To enforce convergence nonetheless, SGD methods are typically implemented in combination with a gradually decreasing update step size. However, selecting a proper sequence of step sizes is a major challenge in practice. An alternative strategy in numerical optimization is to use a constant step size and enforce convergence by averaging the parameters obtained by SGD over several iterations. It was proven mathematically that the highest possible rate of convergence is achieved in this way. Inspired by this work, we propose an averaged SGD (Avg-SGD) method for efficient image registration. In the Avg-SGD approach, a constant step size is used, in combination with an exponentially weighted iterate averaging scheme. The Avg-SGD method is suitable for both rigid and nonrigid registration problems. Experiments on simulated 2D brain MRI data and real 3D lung CT scans demonstrate the effectiveness of the Avg-SGD method in terms of convergence rate, accuracy and precision.

Keywords:
Image registration, Stochastic optimization, Polyak averaging, Variance reduction

1. Introduction

Image registration is an indispensable technique in medical image analysis (Maintz and Viergever, 1998; Hill et al., 2001; Toga and Thompson, 2001; Crum et al., 2004; Sotiras et al., 2013; Viergever et al., 2016). To solve a registration problem, a cost function that measures the dissimilarity between images is usually defined, and then minimized by a numerical optimization routine. During the optimization procedure, the transformation parameters of image registration are estimated iteratively to relate the image coordinates of fixed and moving images. The choice of numerical optimization procedure has a major impact on the performance of the image registration algorithm.

Klein et al. (2007) evaluated the performance of eight optimization methods for image registration: two gradient descent algorithms (with different step size selections) (Nocedal and Wright, 1999), quasi-Newton (Dennis and Moré, 1977), nonlinear conjugate gradient (Dai, 2003), Kiefer-Wolfowitz (Kiefer and Wolfowitz, 1952), simultaneous perturbation (Spall, 1992), Robbins-Monro (RM) (Robbins and Monro, 1951), and evolution strategy (Hansen and Ostermeier, 2001). Most of these methods are gradient-based techniques, in which the derivative of the cost function with respect to the transformation parameters is used to define the search direction in parameter space. The first four of these optimization methods belong to deterministic category, which means that the gradient of the cost function is calculated in a deterministic way. The second three methods are SGD optimizers, which need only approximated gradients of the cost function to compute the search direction. The last method can be considered as

[*] Corresponding author. Address: Biomedical Imaging Group Rotterdam, Erasmus Medical Center, 3015 GE Rotterdam, the Netherlands. Tel.: +31 107044078.
Email address: `sunwei@ieee.org` (Wei Sun)

a stochastic optimizer, but it does not depend on gradient information of the cost function. Through a systematic comparison, the RM optimizer achieved promising performance among deterministic and stochastic candidates in terms of computation time, registration accuracy and robustness. In this work we focus on the RM method with explicit first-order derivatives of the cost function. So far, SGD methods based on the RM approach have been widely applied in many registration problems (Viola and Wells III, 1997; Klein et al., 2008; Bhagalia et al., 2009; Murphy et al., 2011; Metz et al., 2011; Smal et al., 2012; de Groot et al., 2013; Sun et al., 2013, 2017).

In SGD optimization the issue of selecting a good sequence of step sizes γ is a major challenge in practice (Kushner and Yin, 2003). SGD methods are typically implemented in combination with a gradually decreasing step size. The selection of γ is critical for the performance of the optimization process. If the step size γ is chosen too small, the process of minimizing the cost function will be too slow and easily get stuck at an early stage. If γ is selected too large, the noise during the stochastic optimization will become too large and the reliability of the optimization cannot be guaranteed. In both cases, the convergence rates of optimization are not optimal. To estimate γ automatically in image registration problems, the ASGD optimizer was proposed in Klein et al. (2009). In the ASGD method, an image-driven mechanism is used to predict a reasonable value of γ, which satisfies several theoretical conditions for convergence. However, the ASGD optimizer has a relatively high computation cost during the estimation process when the number of transformation parameters becomes high. To tackle this issue, a fast ASGD optimizer was proposed in Qiao et al. (2016). Although the ASGD method provides a reasonable choice of γ, it does not claim to achieve an optimal rate of convergence.

The convergence rate of SGD can be accelerated by using the second-order information (Hessian matrix) of the cost function (Bousquet and Bottou, 2008). When the second-order information is adopted, the former scalar γ becomes a matrix which is employed to control the step size according to the curvature of the optimization landscape. Because the second-order information is not known in advance, various methods for predicting the Hessian matrix were proposed in Bottou and Le Cun (2005). However, the computational cost of determining the full Hessian matrix is usually too high to maintain. Therefore, different approaches for estimating an approximated Hessian matrix were proposed in Klein et al. (2011) and Qiao et al. (2015). However, those approximated estimations cannot guarantee a convergence rate which is as good as when using the full Hessian matrix in practice.

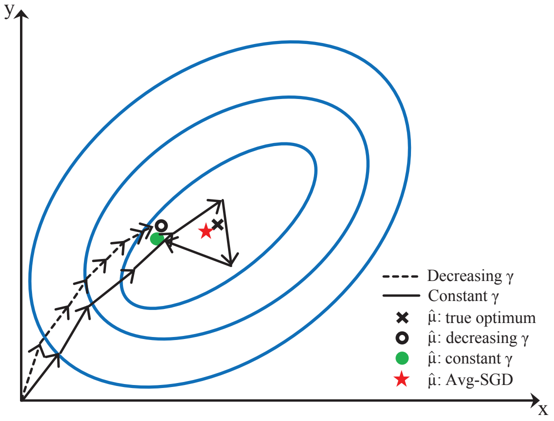


Figure 1: Illustration of the principle of the Avg-SGD method on a two-dimensional optimization landscape. Blue lines indicate isocontours of the optimization landscape. The method with decreasing step size γ (dashed line) may approach the optimum too slowly, and thus may terminate far from the true optimum. The method with constant γ (solid line) reaches the neighourhood of the optimum quickly, but then keeps on jumping around the optimum. The Avg-SGD approach in combination with constant γ would average out these fluctuations around the true optimum and thus achieve convergence to a point near the true optimum.

A different strategy to accelerate the convergence rate was proposed by Polyak and Juditsky (1992). They proposed to average the iterates of a SGD optimization, and proved that the averaged sequence of the estimated parameters converges in an optimal rate, which is as good as full second-order SGD. We refer to this method as Avg-SGD. With this approach, we may use a larger than usual step size γ and let the averaging take care of the increased noise effects that are due to the larger step size. In this way we can substantially improve the overall convergence speed and make the choice of γ less critical. Compared with the second-order SGD, the averaging technique is easy to implement and more attractive in practice. Figure 1 illustrates the principle of the Avg-SGD optimizer.

This intuitive but effective averaging technique has drawn lots of attention in machine learning field in recent years (Bottou, 2014). For large-scale machine learning tasks, Bottou (2010) compared the performances of a normal SGD, a second-order SGD and the averaged SGD algorithms. Their results show that the averaged method achieved a better convergence than the second-order SGD on training a linear SVM for the ALPHA task (Bordes et al., 2009). For large-scale visual recognition, Ushiku et al. (2014) suggested that the averaging technique accelerates not only the first-order

SGD optimization but also the second-order algorithms.

In this paper we investigate the potential of the Avg-SGD method for image registration. In theoretical analyses, it is typically assumed that the iteration number $k \to \infty$ (Kushner and Yin, 2003). However, this assumption is impractical for image registration where we only have limited computation time. Given finite k, it is preferable to skip or alleviate the effect of iterates in the early phase of optimization because the estimated transformation parameters may change dramatically. In this research, we present two versions of the Avg-SGD method: *postponed* and *exponential*. For the postponed version the first k_0 iterates are skipped in computing the averaged transformation parameters. In the exponential version, the effect of the early iterations is exponentially decreased, avoiding the need to set a hard threshold k_0. We compared the new Avg-SGD method with the state-of-the-art ASGD optimizer in extensive medical registration experiments on simulated and real imaging data, testing both rigid and nonrigid registrations.

But you'd still have to choose exp params.

The remainder of the manuscript is organized as follows. Section 2 provides an overview of stochastic optimization for image registration and explains the proposed Avg-SGD approach. Section 3 describes the experiments which are used to evaluated the proposed approaches. The experimental results are presented in Section 4. In Section 5 the experimental results are interpreted and discussed. Then, the conclusions are drawn in the last section.

2. Method

2.1. Stochastic optimization for image registration

Let $F(\boldsymbol{x}) : \Omega_F \subset \mathbb{R}^D \to \mathbb{R}$ and $M(\boldsymbol{x}) : \Omega_M \subset \mathbb{R}^D \to \mathbb{R}$ denote the D-dimensional fixed and moving images where $\boldsymbol{x}$ represents an image coordinate, and Ω_F and Ω_M are the fixed and moving image domains, respectively. Suppose $\mathbf{T}(\boldsymbol{\mu}, \boldsymbol{x}) : \mathbb{R}^P \times \Omega_F \to \Omega_M$ is a coordinate transformation where $\boldsymbol{\mu} \in \mathbb{R}^P$ represents the vector of transformation parameters. $\mathbf{T}(\boldsymbol{\mu}, \boldsymbol{x})$ could be a translation, rigid, affine or nonrigid (e.g., B-spline) transformation model. Then, the registration problem is defined as:

$$\hat{\boldsymbol{\mu}} = \arg\min_{\boldsymbol{\mu}} C(\boldsymbol{\mu}, \Omega_F), \tag{1}$$

where $C(\boldsymbol{\mu}, \Omega_F)$ calculates the dissimilarity between the original fixed image $F(\boldsymbol{x})$ and the deformed moving image $M(\mathbf{T}(\boldsymbol{\mu}, \boldsymbol{x}))$ on the domain $\boldsymbol{x} \in \Omega_F$. Examples of C are mutual information (Viola and Wells III, 1997; Maes et al., 1997), the sum of squared differences (SSD), and normalized correlation coefficient. For instance, the cost function C of SSD is defined as:

$$C(\boldsymbol{\mu}, \Omega_F) = \frac{1}{|\Omega_F|} \sum_{\boldsymbol{x}_i \in \Omega_F} \left(F(\boldsymbol{x}_i) - M(\mathbf{T}(\boldsymbol{\mu}, \boldsymbol{x}_i))\right)^2 . \tag{2}$$

In image registration, an iterative optimization strategy is applied to solve Eq. (1) and determines the optimal set of parameters $\hat{\boldsymbol{\mu}}$. In the evaluation of different optimizers (Klein et al., 2007), the SGD optimizer turned out to be a competitive alternative to deterministic algorithms in nonrigid registration problems. SGD optimization is based on the following iterative update strategy:

$$\boldsymbol{\mu}_{k+1} = \boldsymbol{\mu}_k - \gamma_k \tilde{\boldsymbol{g}}_k, \quad k = 0, 1, 2 \ldots K, \tag{3}$$

where $\tilde{\boldsymbol{g}}_k$ represents a stochastic approximation of the cost function derivative $\partial C / \partial \boldsymbol{\mu}$, evaluated at the current estimated transformation parameters $\boldsymbol{\mu}_k$, and γ_k is a scalar gain factor that controls the step size along $\tilde{\boldsymbol{g}}_k$. To guarantee the convergence of SGD optimization, a common choice of γ_k is

$$\gamma_k = a/(k + A)^{\alpha}, \tag{4}$$

where $a > 0$, $A \geq 1$, and $0 < \alpha \leq 1$ are user-defined parameters. Kushner and Yin (2003) proved that $\alpha = 1$ gives a theoretically optimal rate of convergence when $k \to \infty$. Due to having no unit and heavily depending on the selected cost function, the value of a is difficult to choose. Klein et al. (2009) presented an image-driven estimation mechanism for estimating a by substituting it with a new parameter δ. The estimation, which only runs one time prior to the iterative optimization, is based on general characteristics of the cost functions that are commonly used in intensity-based image registration problems. The user-specified parameter δ is independent of the choice of C and represents the maximum allowed voxel displacement (mm) during optimization. Next to a, the constant A can be set in order to control the sensitivity of γ_k to k in the first iterations. For a very high A and a finite number of iterations, γ_k will effectively become a constant. In addition, an adaptive strategy to choose step size γ_k was proposed in Klein et al. (2009). Equation (3) is reformulated as:

$$\boldsymbol{\mu}_{k+1} = \boldsymbol{\mu}_k - \gamma(t_k) \tilde{\boldsymbol{g}}_k, \quad k = 0, 1, 2 \ldots K, \tag{5}$$

$$t_{k+1} = [t_k + f(-\tilde{\boldsymbol{g}}_k^T \tilde{\boldsymbol{g}}_{k-1})]^+, \tag{6}$$

where $[x]^+$ denotes $\max(x, 0)$, f is a sigmoid function, t_0 and t_1 are user-defined initial conditions. In Eq. (5), the γ function is not evaluated at the integer iteration number k, but at the real number t_k. If the gradients in two successive iterations have similar directions, the inner product of $\tilde{\boldsymbol{g}}_k$ and $\tilde{\boldsymbol{g}}_{k-1}$ is positive, and thus t_k is reduced. As presented in Eq. (4), γ is a monotone decreasing function. Therefore, $\gamma(t_{k+1})$ adapts to a larger step size. In this way, the ASGD method in Klein et al. (2009) implements a dynamic mechanism to determine the step size. If $f(x) = 1$, the ASGD becomes the original RM method. In previous studies, the ASGD method has been applied in many image registration tasks (Metz et al., 2011; Murphy et al., 2011; Smal et al., 2012; de Groot et al., 2013; Sun et al., 2013, 2017). In our experiments, we will compare the performances of the proposed Avg-SGD method and the ASGD optimizer.

In the SGD optimizer (Klein et al., 2007), the stochastic approximation $\tilde{\boldsymbol{g}}_k$ is calculated by evaluating $\partial C/\partial \boldsymbol{\mu}$ on a small random subset $\widetilde{\Omega}_F^k \subset \Omega_F$ with S image samples, thus reducing the computation time per iteration. This subset $\widetilde{\Omega}_F^k$ should be randomly refreshed at each iteration k, to make the approximation stochastic. We thus can write:

$$\tilde{\boldsymbol{g}}_k = \frac{\partial C}{\partial \boldsymbol{\mu}}(\boldsymbol{\mu}_k, \widetilde{\Omega}_F^k) \approx \frac{\partial C}{\partial \boldsymbol{\mu}}(\boldsymbol{\mu}_k, \Omega_F). \tag{7}$$

For example, if we choose SSD as cost function C (see Eq. (2)), $\tilde{\boldsymbol{g}}_k$ is computed as:

$$\tilde{\boldsymbol{g}}_k = \frac{2}{|\widetilde{\Omega}_F^k|} \sum_{\boldsymbol{x}_i \in \widetilde{\Omega}_F^k} \left(F(\boldsymbol{x}_i) - M\left(\mathbf{T}(\boldsymbol{\mu}_k, \boldsymbol{x}_i)\right)\right) \tag{8}$$
$$\times \left(\frac{\partial \mathbf{T}}{\partial \boldsymbol{\mu}}\bigg|_{(\boldsymbol{\mu}_k, \boldsymbol{x}_i)} \right)^T \left(\frac{\partial M}{\partial \boldsymbol{x}}\bigg|_{\mathbf{T}(\boldsymbol{\mu}_k, \boldsymbol{x}_i)} \right).$$

In this work, we use the same approach for computing $\tilde{\boldsymbol{g}}_k$.

2.2. Averaged stochastic gradient descent (Avg-SGD)

Let us define

$$\lim_{k \to \infty} \boldsymbol{\mu}_k = \boldsymbol{\mu}^*. \tag{9}$$

The asymptotic normality under certain assumptions of the rate of convergence of SGD can be defined as Klein et al. (2009):

$$\sqrt{k}(\boldsymbol{\mu}_k - \boldsymbol{\mu}^*) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{V}), \tag{10}$$

where $\xrightarrow{d}$ represents the convergence in distribution as $k \to \infty$, and $\mathcal{N}(\mathbf{0}, \boldsymbol{V})$ is a multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{V}$.

To accelerate the convergence of stochastic optimization, Polyak and Juditsky (1992) and Ruppert (1988) proposed to average the trajectory of stochastic optimization. In contrast to other accelerating techniques (e.g., second-order SGD), the averaging technique is simple and requires no prior information about the cost function. The original Polyak averaging can be formulated as:

$$\bar{\boldsymbol{\mu}}_k = \frac{1}{k+1} \sum_{i=0}^{k} \boldsymbol{\mu}_i, \tag{11}$$

where $\bar{\boldsymbol{\mu}}_k$ is the sequence of averaged parameters. It is worth noting that this iterate averaging process does not interfere with the original SGD algorithm; the estimates $\boldsymbol{\mu}_k$ are unaffected by $\bar{\boldsymbol{\mu}}_k$. The basic idea is that $\bar{\boldsymbol{\mu}}_k$ converges faster to $\boldsymbol{\mu}^*$ than $\boldsymbol{\mu}_k$ does. Polyak and Juditsky (1992) presented a proof that $\bar{\boldsymbol{\mu}}_k$ converges to $\boldsymbol{\mu}^*$ as good as the full second-order algorithm. In later work, Yin (1992) showed that

$$\sqrt{k}(\bar{\boldsymbol{\mu}}_k - \boldsymbol{\mu}^*) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \bar{\boldsymbol{V}}), \tag{12}$$

where $\bar{\boldsymbol{V}}$ is the smallest possible covariance matrix when an asymptotically optimal matrix-valued step size (e.g., second-order SGD) is adopted.

If $\gamma_k \to 0$ slower than $O(1/k)$, Kushner and Yin (2003) extended the averaging theory to a window definition:

$$\bar{\boldsymbol{\mu}}_k^{win} = \frac{\gamma_k}{\tau} \sum_{i=k-\tau/\gamma_k+1}^{k} \boldsymbol{\mu}_i, \tag{13}$$

where $k \geq \tau/\gamma_k - 1$ and the number of iterates in the window of averaging is $\tau/\gamma_k = o(k)$ for arbitrary real $\tau > 0$. They proved that

$$(\bar{\boldsymbol{\mu}}_k^{win} - \boldsymbol{\mu}^*) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \frac{\gamma_k \bar{\boldsymbol{V}}}{\tau}). \tag{14}$$

Eq. (14) shows that the covariance matrix $\bar{\boldsymbol{V}}$ can be reduced proportionally to the size of averaging window τ/γ_k. Kushner and Yin (2003) confirmed that this desirable property of averaging holds both for constant and decreasing step size γ_k.

In practice, we only have a finite number of iterations K which is different from the theoretical analysis under the assumption $k \to \infty$. Because $\boldsymbol{\mu}_i$ may change substantially in the initial phase of optimization, it would be preferable to skip or alleviate the effect of first k_0 iterations in Eq. (11). In this work, we propose two averaging methods. The first one is the *postponed* averaging method as an alternative to Eq. (13),

$$\bar{\boldsymbol{\mu}}_k^{post} = \frac{1}{k-k_0+1} \sum_{i=k_0}^{k} \boldsymbol{\mu}_i, \tag{15}$$

Table 1: Summary of registration methods.

Method	Step size(γ)	Optimization
SGD-constant	Constant	SGD
SGD-adaptive	Adaptively decreasing	SGD
Avg-constant	Constant	Postponed/exponential Avg-SGD
Avg-adaptive	Adaptively decreasing	Postponed/exponential Avg-SGD

where $0 \leq k_0 \leq k$. In the above definition, the iterations before k_0 are neglected, and only the iterates from k_0 to k are averaged to compute $\bar{\boldsymbol{\mu}}_k^{post}$. However, there is no prior information to properly choose k_0 for a practical problem. If k_0 is selected too small, the 'premature' $\boldsymbol{\mu}_i$ where $i \leq k_0$ could be involved in the averaging calculation. If k_0 is selected too large, the number of averaged iterates will become too low to achieve substantial acceleration of convergence. We will illustrate this issue in the later experiments.

To remedy the issue of choosing k_0, we also define an *exponential* moving average:

$$\bar{\boldsymbol{\mu}}_{k+1}^{ex} = (1-\epsilon)\bar{\boldsymbol{\mu}}_k^{ex} + \epsilon\boldsymbol{\mu}_{k+1} \tag{16}$$

$$= (1-\epsilon)^{k+1}\boldsymbol{\mu}_0 + \epsilon\sum_{j=1}^{k+1}(1-\epsilon)^{k+1-j}\boldsymbol{\mu}_j, \tag{17}$$

where $0 < \epsilon < 1$ and $\bar{\boldsymbol{\mu}}_0^{ex} = \boldsymbol{\mu}_0$. In this exponential averaging method, the influence of previous iterates is decreased exponentially. In this way, we avoid the need to set a hard threshold k_0. The influence of initial iterations is gradually decayed. The weighing factor ϵ needs to be set by the user. In this work, we fixed ϵ to a constant parameter for all experiments.

Note that for the choice $\epsilon = 1/(k+2)$, Eq. (16) boils down to a recursive formulation of the original Polyak averaging method (Eq. (11)). Also the postponed averaging as defined in Eq. (15) can be written in a recursive formulation, by setting $\epsilon = 1/(k-k_0+2)$, resulting in:

$$\bar{\boldsymbol{\mu}}_{k+1}^{post} = \left(1 - \frac{1}{k-k_0+2}\right)\bar{\boldsymbol{\mu}}_k^{post} + \frac{1}{k-k_0+2}\boldsymbol{\mu}_{k+1}. \tag{18}$$

The recursive formulation has the advantage that it requires less memory than the direct formulation, since only the current $\bar{\boldsymbol{\mu}}_k^{post}$ (or $\bar{\boldsymbol{\mu}}_k^{ex}$) needs to be stored during the optimization.

2.3. Step size selection

The use of Avg-SGD may allow us to use a constant step size γ instead of a decaying step size, which may benefit the convergence rate in non-asymptotic settings, as was illustrated in Figure 1. The increased noise on $\boldsymbol{\mu}_k$ due to the large step sizes in later iterations could be attenuated by the averaging process in the Avg-SGD method. In this work, we will therefore compare different step size selection schemes in combination with the Avg-SGD and conventional SGD approaches.

As described in Section 2.1, the ASGD optimizer has two major advantages. The first one is that ASGD uses a new parameter δ to replace a in Eq. (4), which is independent of the choice of cost function C and represents the maximum allowed voxel displacement. The second one is an adaptive mechanism in Eqs. (5) and (6) to adjust step size γ during the optimization. In this work we implemented the Avg-SGD method by using the benefit of δ to set initial step size, and treat the adaptive γ sequence as a state-of-the-art technique of the SGD optimization for comparison. Table 1 summarizes the registration methods evaluated in this work. As suggested in Klein et al. (2009), $A = 20$ and $\alpha = 1$ were chosen in all experiments for adaptively decreasing γ. In Table 1, the SGD-constant method represents the ordinary SGD optimizer using constant γ. The SGD-adaptive method is the standard ASGD optimizer equipped with adaptively decreasing γ. The Avg-constant method represents the proposed Avg-SGD approach using constant γ. The Avg-adaptive approach is the Avg-SGD method combined with the adaptively decreasing γ. The Avg-constant and Avg-adaptive methods are evaluated using either the postponed or the exponential averaging method.

2.4. Transformation models

In this paper, the performances of candidate registration approaches are evaluated on the rigid and B-spline transformation models.

2.4.1. Rigid transformation

A rigid transformation is defined as:

$$\mathbf{T}(\boldsymbol{\mu}, \boldsymbol{x}) = \boldsymbol{R}(\boldsymbol{x} - \boldsymbol{c}) + \boldsymbol{t} + \boldsymbol{c}, \tag{19}$$

where $\boldsymbol{R}$ represents a rotation matrix, $\boldsymbol{c}$ is the center of rotation located at the center of the field of view of the fixed image, and $\boldsymbol{t}$ is the vector of translations. The transformation of the rigid model is parameterized using $\boldsymbol{\mu} = (\boldsymbol{\theta}^T, \boldsymbol{t}^T)^T$ where $\boldsymbol{\theta}$ represents the vector of Euler angles. For example transformation parameters $\boldsymbol{\mu}$ are expressed as $(\theta, t_x, t_y)^T$ in 2D registration.

2.4.2. B-spline transformation

The B-spline free-form deformation (FFD) transformation model (Rueckert et al., 1999) is defined as:

$$\mathbf{T}(\boldsymbol{\mu}, \boldsymbol{x}) = \boldsymbol{x} + \sum_{\boldsymbol{\xi}\in\Xi}\mathbf{c}_{\boldsymbol{\xi}}\Phi_D(\boldsymbol{x}/\eta - \boldsymbol{\xi}), \tag{20}$$

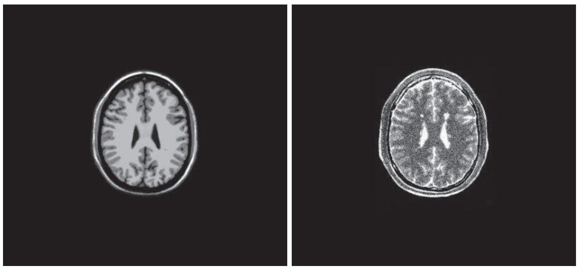


Figure 2: Example of 2D brain MRI data: (a) T1-weighted image; (b) T2-weighted image with visible multiple-sclerosis lesions.

where $\Phi_D(\boldsymbol{x}) : \mathbb{R}^D \to \mathbb{R}$ represents the D-dimensional B-spline function, $\Xi \subset \mathbb{Z}^D$ is a D-dimensional control-point grid, η is the grid spacing, $\mathbf{c}_{\boldsymbol{\xi}}$ represents the coefficient vector for a control point $\boldsymbol{\xi}$, and the parameter vector $\boldsymbol{\mu}$ is formed by the elements of all coefficient vectors ($\boldsymbol{\mu} = \{\mathbf{c}_{\boldsymbol{\xi}} \mid \boldsymbol{\xi} \in \Xi\}$). For a given position $\boldsymbol{x}$, the summation goes effectively inside the support of $\Phi_D(\boldsymbol{x})$.

3. Experiments

Two sets of experiments are performed: 1) Multimodal rigid registration on 2D simulated T1-weighted and T2-weighted brain MRI data. 2) Monomodal nonrigid registration on 3D lung CT data. Section 3.1 describes the general design of the experiments. Section 3.2 presents evaluation measures. Sections 3.3 and 3.4 give specific details for the experiments on brain MRI and lung CT.

3.1. Experimental settings

All methods were implemented as part of the open source image registration package `elastix` (Klein et al., 2010). For the multimodal rigid registration on brain data, we employed the MI metric presented in Mattes et al. (2003) where a B-spline Parzen window was used to model the joint probability. Registration experiments on lung CT images were performed with B-spline FFD transformation using SSD as the dissimilarity term. For rigid transformation, we set $\delta = 4\nu$ in this work, where ν represents the average pixel/voxel size (arithmetic average over all dimensions). For nonrigid B-spline transformation, a more conservative setting $\delta = \nu$ was used. For both optimizers, the number of random samples S was set to 2000. The number of iterations K of the optimizer was set to 2000 for all experiments. Trilinear interpolation was utilized to interpolate the moving image. For the postponed Avg-SGD in Eq. (18), we chose $k_0 \in \{0, 399, 799, 1199, 1599\}$ to start the averaging process. For the exponential Avg-SGD in Eq. (16), we fixed $\epsilon = 0.01$ in all experiments.

To investigate the effect of different starting points, we used a multilevel optimization, in which the transformation parameters $\hat{\boldsymbol{\mu}}$ estimated at level l were used to initialize $\boldsymbol{\mu}_0$ at level $l + 1$. For this purpose, we used the result $\hat{\boldsymbol{\mu}} = \boldsymbol{\mu}_K$ obtained by the conventional SGD-adaptive method, to ensure that in each level all methods start from the same point $\boldsymbol{\mu}_0$. No image blurring was performed. Three levels were used for the rigid registration experiments. Two levels were used for the nonrigid registration experiments. In the nonrigid registration experiments, the B-spline control point spacing η was set to 64 and 32 mm in the two levels, respectively.

3.2. Evaluation measures

To quantify different aspects of convergence behavior, we use three measures: accuracy curves, reproducibility curves, and fluctuation.

Let $\Gamma(k)$ represent registration accuracy as a function of the iteration number k. The exact definitions of Γ are provided in Sections 3.3 and 3.4. We compute the mean of $\Gamma(k)$ over multiple registration cases, and plot this mean as a function of k, to obtain accuracy curves.

The reproducibility measure assesses the change of the registration accuracy caused by intrinsic randomness of SGD optimization. To quantify this, the registrations were repeated with $R = 20$ random seeds. The random seed affects the selection of the random subsets $\widetilde{\Omega}_F^k$ in Eq. (7). The standard deviation of registration accuracy was calculated over those seeds to measure the reproducibility of each method. Thus, we can define the reproducibility as: $\mathrm{Std}(\Gamma_r(k))$ with $\Gamma_r(k)$ the accuracy at iteration k using random seed r. In the experiments, the reproducibility is evaluated at multiple iterations k and is plotted as a function of k, in order to obtain reproducibility curves.

The fluctuation of the accuracy curve over the last 100 iterations was computed to assess the variability of the registration result near termination of the optimization procedure. Large fluctuation of $\Gamma(k)$ in the final iterations would indicate the optimization has not fully converged yet. The second-order derivative of the curve was adopted to measure the fluctuation. The second-order derivative at iteration k can be approximated by finite difference:

$$\Gamma''(k) \approx \Gamma(k+1) - 2\Gamma(k) + \Gamma(k-1), \qquad (21)$$

where $\Gamma(k)$ represents the registration accuracy at iteration k. For each test case, we calculate the root mean square of these second-order derivatives over the last

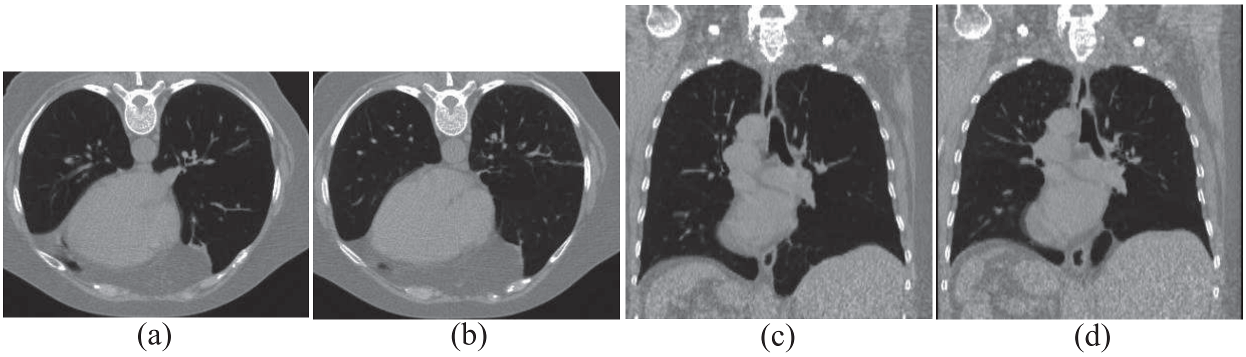


Figure 3: Example of 3D lung CT data of one patient from the DIR-Lab data: (a) transverse view of inhale phase; (b) transverse view of exhale phase; (c) coronal view of inhale phase; (d) coronal view of exhale phase.

100 iterations, and we summarize the results of all test cases using box plots.

3.3. Rigid registration on 2D brain MRI

Multimodal rigid registration was performed on 2D slices of simulated brain MRI data. We used the T1- and T2-weighted images provided by the Simulated Brain Database (SBD) from Brainweb (Cocosco et al., 1997). To generate image pairs for registration, we utilized the options provided by BrainWeb to add multiple-sclerosis lesions and 9% noise to the T2-weighted phantom. As shown in Figures 2 (a) and (b), we randomly selected one typical T1-weighted slice and its corresponding T2-weighted slice sharing the same slice number. The dimension and pixel sizes for each 2D brain image are 400 × 400 pixels and 1 × 1mm. Pairwise registrations were carried out between the image pair. Random rigid transformations were created and used as initial transformations $\boldsymbol{T}_{init}$. For $\boldsymbol{T}_{init}$, 20 initial rigid transformations, which served as the ground truth, were randomly generated with uniform range $|\theta| \in [0.2, 0.4]$rad, $|t_x| \in [15, 60]$pixels and $|t_y| \in [15, 60]$pixels. The active transformation $\boldsymbol{T}_{\mu}$ was precomposed with $\boldsymbol{T}_{init}$. The average residual deformation $Residual(\boldsymbol{T}_{init}, \boldsymbol{T}_{\hat{\mu}})$ measured inside brain mask was used to measure registration accuracy Γ. The residual metric measures the average Euclidean distance between the recovered $\boldsymbol{T}_{init}$, i.e., $\boldsymbol{T}_{\hat{\mu}}(\boldsymbol{T}_{init}(\boldsymbol{x}_i))$ and the original location $\boldsymbol{x}_i$:

$$Residual(\boldsymbol{T}_{init}, \boldsymbol{T}_{\hat{\mu}}) = \frac{1}{|\Omega_F|} \sum_{\boldsymbol{x}_i \in \Omega_F} \left\| \boldsymbol{T}_{\hat{\mu}}(\boldsymbol{T}_{init}(\boldsymbol{x}_i)) - \boldsymbol{x}_i \right\| . \tag{22}$$

With 20 $\boldsymbol{T}_{init}$, 20 random seeds, and two registration directions between the fixed and moving images, there are 20 × 20 × 2 test cases for each approach.

3.4. Nonrigid registration on 3D lung CT

The publicly available DIR-Lab 3D chest CT data set facilitates a rigorous and objective assessment of the spatial accuracy of registration methods (Castillo et al., 2009). The DIR-Lab data set contains 10 pairs of scans with 300 manually annotated landmarks on the lungs, which allows us to evaluate the registration accuracy. The voxel sizes and dimensions of these scans are around 1.0×1.0×2.5mm and around 256×256×110 voxels. To focus on the lung region, lung masks were created to restrict the registration. The masks were created by thesholding, 3D-6-neighborhood connected component analysis, and morphological closing operation using a spherical kernel with a diameter of 9 voxels. In the experiments, the exhale phase (moving image) was registered to the inhale phase (fixed image). Figure 3 shows examples of transverse and coronal views of the inhale and exhale phases of one patient. The mean of target registration errors (TRE), which measure the distances between the transformed and ground truth landmarks, was used to measure the registration accuracy Γ. Here, each test was also repeated with 20 random seeds. Therefore, there are in total 10 × 20 test cases for 10 patients over 20 random seeds.

4. Results

First, we will compare the performance of the postponed and exponential averaging considering only the results for the first level of rigid brain MRI registration. Figure 4 presents the registration accuracy $\Gamma(k)$ at each iteration $k = 0, 100, 200, \ldots K$ for a single registration experiment. Figure 4 (a) shows that the SGD-constant approach converges much faster than the SGD-adaptive method at the early stage of optimization. Nevertheless, it starts to fluctuate around the bot-

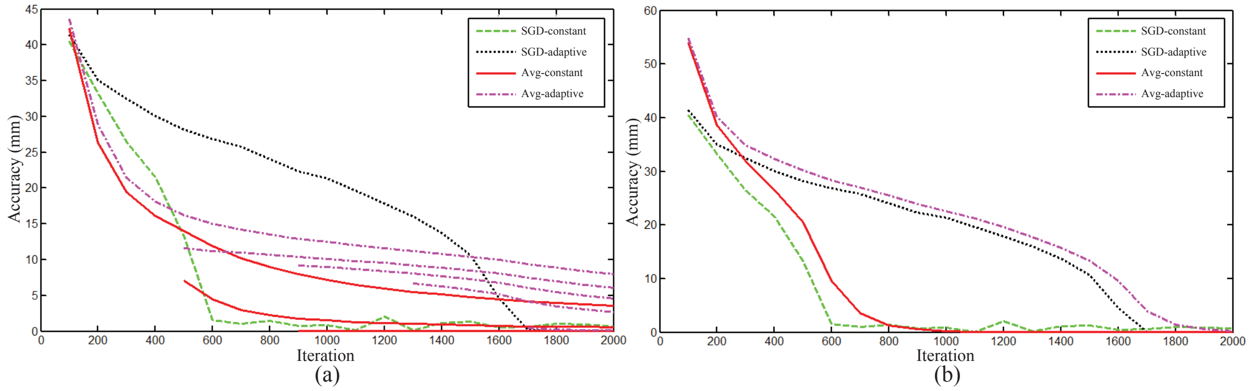


Figure 4: Registration accuracy Γ as a function of k, for a single registration experiment. (a) Convergence curves by using constant and adaptively decreasing γ combined with the conventional SGD and the postponed Avg-SGD approaches, with $k_0 \in \{0, 399, 799, 1199, 1599\}$ for postponed Avg-SGD; (b) Convergence curves by using constant and adaptively decreasing γ combined with the conventional SGD and the exponential Avg-SGD approaches.

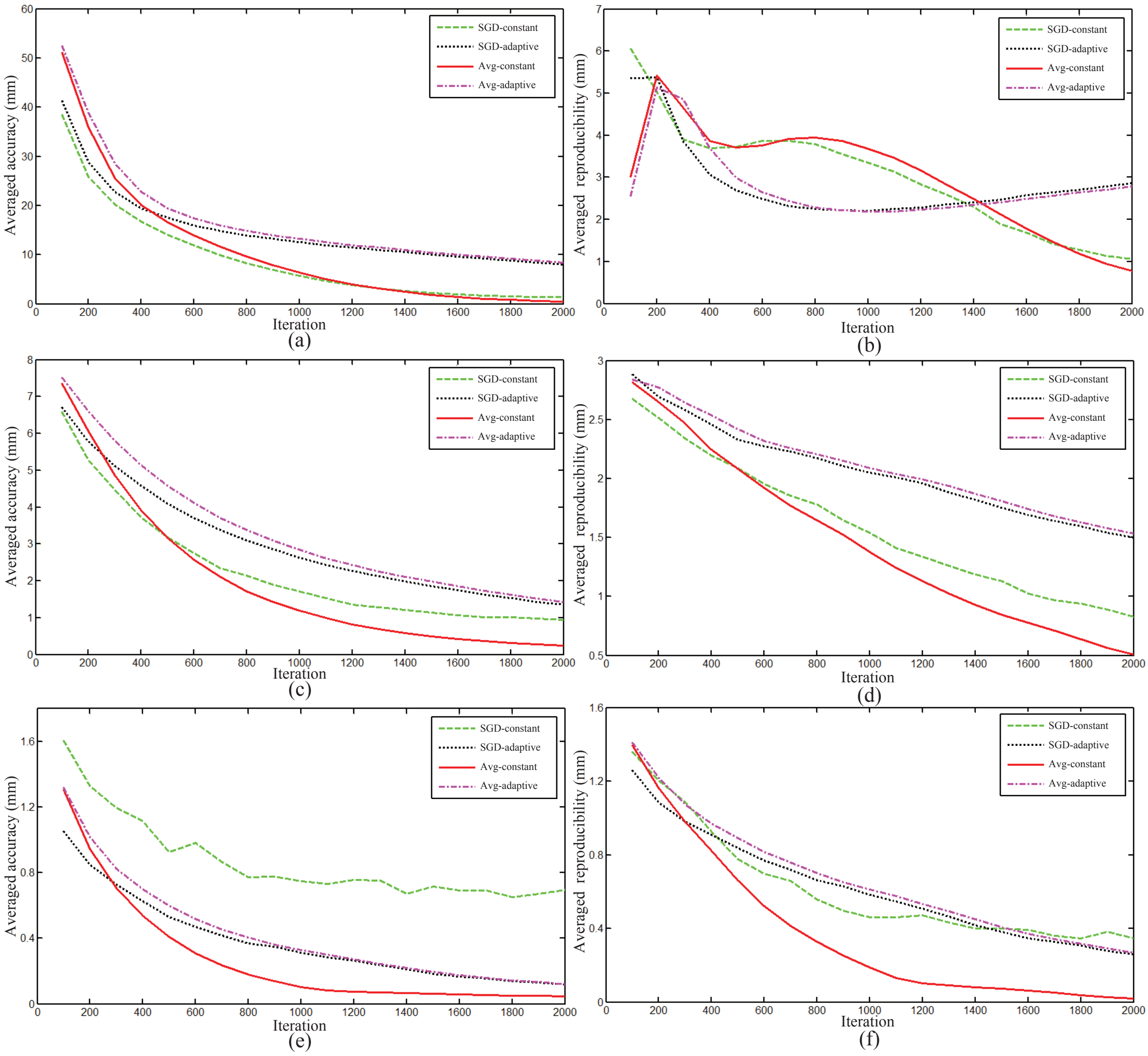
tom (optimum) after 600 iterations. The SGD-adaptive method reaches the optimum after around 1700 iterations. The Avg-constant and Avg-adaptive methods each resulted in five curves, representing different values of k_0. For the postponed Avg-constant method, the fluctuation caused by constant γ could be alleviated by choosing a suitable value for k_0. However, the acceleration could not be fully realized when using a too small or too large value for k_0. Since there is no prior information about k_0, it may therefore be hard to use the postponed Avg-constant approach in practice. The postponed Avg-adaptive method neither led to a convincing acceleration of convergence. Figure 4 (b) shows the results of the exponential Avg-SGD methods. It can be observed that the Avg-constant method reduces the fluctuations existing in the curve of the SGD-constant method. In addition, the Avg-constant method converges much faster than the SGD-adaptive and Avg-adaptive approaches. From Figure 4 we may conclude that the exponential Avg-SGD approach (Avg-constant in Figure 4 (b)) achieved the best performance. Compared with the postponed Avg-SGD method, the same $\epsilon = 0.01$ of the exponential Avg-SGD method works for both constant and decreasing γ and shows stable performance. Therefore, we will only use the exponential definition to represent the Avg-SGD method in the following experiments.

Figure 5 shows the registration accuracy and reproducibility obtained by candidate approaches with three registration levels on brain MRI data. Figures 5 (a), (c) and (e) show the mean of $\Gamma(k)$ over $20 \times 20 \times 2$ registration cases, at each iteration $k = 0, 100, 200, \ldots K$ from the first to the third registration level, respectively. For each subsequent level, we used the results obtained by the SGD-adaptive method as the initialization for all approaches. It can be observed that the initial $\Gamma(k)$ is reduced at successive levels. At the first level, both the original SGD and Avg-SGD methods using constant γ achieved substantially better convergence than the approaches using adaptively decreasing γ. Among the methods using constant γ, it can be found that the Avg-constant method converges a bit slowly in early phase due to the recursive averaging. However, it achieved slightly better accuracy than the SGD-constant method in the end. The difference could be caused by the too large noise of the SGD-constant method around the final solution $\hat{\boldsymbol{\mu}}$. In Figure 5 (c), approaches start from the same medium initial $\Gamma(k)$ after the first registration level. At the second level, the SGD-constant and Avg-constant methods still performed better than the SGD-adaptive and Avg-adaptive approaches. In addition, the difference between the convergence curves achieved by the SGD-constant and Avg-constant methods becomes significant. At the last registration level (Figure 5 (e)), the SGD-constant method obtained the worst convergence due to too large noise. In contrast, the Avg-constant method still achieved the best convergence rate.

Table 2: Registration accuracy (mm) on brain MRI data using rigid transformation. The best result at each level is marked bold.

	SGD-constant	SGD-adaptive	Avg-constant	Avg-adaptive
Level 1	1.35±1.06	8.04±4.83	**0.53±1.04**	8.41±4.86
Level 2	0.94±0.35	1.35±2.20	**0.23±0.49**	1.42±2.31
Level 3	0.70±0.07	0.12±0.25	**0.05±0.01**	0.12±0.27

Table 2 presents the final registration accuracy achieved by candidate approaches at each registration level. As explained before, we guarantee the same



(a)

(b)

(c)

(d)

(e)

(f)

Figure 5: Convergence curve and reproducibility of registration methods on brain MRI data using rigid transformation. (a), (c) and (e) averaged convergence curves from the first to the third registration levels over $20 \times 20 \times 2$ test cases; (b), (d) and (f) averaged reproducibility curves from the first to the third registration levels over $20 \times 20 \times 2$ test cases.

initial deformation at each level by using the SGD-adaptive method as the initialization. It can be found that the proposed Avg-constant approach constantly yielded the best registration accuracy for large, medium and small initial displacements at all three levels. For example, Avg-constant achieved even better accuracy with only one registration level than the other approaches using two levels.

Figures 5 (b), (d) and (f) show the averaged reproducibility $\mathrm{Std}(\Gamma_r(k))$ achieved by candidate methods at the same registration levels as Figures 5 (a), (c) and (e), respectively. At the first resolution level (Figure 5 (b)), at intermediate number of iterations, the difference between the methods using constant and adaptively decreasing γ is caused by relatively large changes of optimization routes due to the different random seeds in combination with constant γ and large initial transformations. It can also be observed that the SGD-constant and Avg-constant approaches outperform the methods using adaptively decreasing γ in the later iterations. For the approaches using constant γ, the Avg-constant method achieved slightly better final $\mathrm{Std}(\Gamma_r(k))$. At the second registration level (Figure 5 (d)), the methods using constant γ consistently outperform the SGD-adaptive and Avg-adaptive approaches. The Avg-constant method obtained the best reproducibility in all cases. At the final registration level in Figure 5 (f), the SGD-constant method obtained similar $\mathrm{Std}(\Gamma_r(k))$ as the methods using adaptively decreasing γ due to the large noise. Nevertheless, the Avg-constant approach still achieved the best reproducibility at this level.

Figure 6 summarizes the degree of fluctuation at the final registration level. Each box plot represents the root mean square of the second-order derivatives computed by Eq. (21) over the last 100 iterations on in total 20 × 20 × 2 test cases for each registration method. It can be observed that the SGD-constant method has the largest fluctuation. This fluctuation was reduced substantially using the Avg-constant approach although the same constant step sequence was adopted.

Figure 7 shows the results of candidate approaches using two-level registration on lung CT data. Figures 7 (a) and (c) present the mean of $\Gamma(k)$ over 10 × 20 registration cases, at each iteration $k = 0, 100, 200, \ldots K$ for the first and second registration levels, respectively. As shown in Figure 7 (a) the methods using constant γ converged faster than the approaches using adaptively decreasing γ at the first registration level. It can also be found that the SGD-constant and Avg-constant method obtained similar final accuracy. At the second registration level (Figure 7 (c)), the methods using constant γ still outperformed the approaches using adaptively decreasing γ. In addition, the proposed Avg-constant method achieved a better convergence rate than the SGD-constant approach.

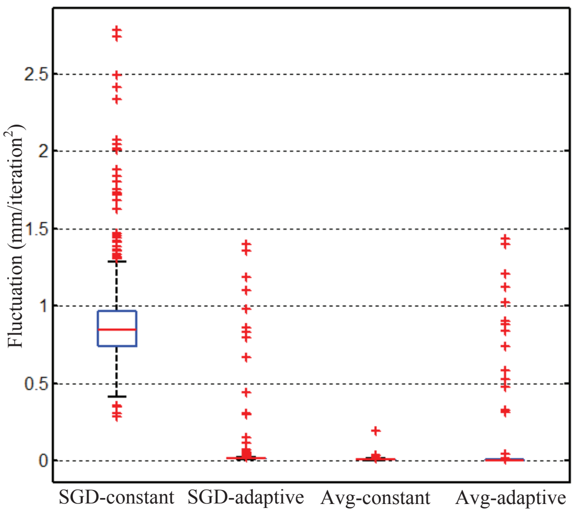


Figure 6: Degrees of fluctuations of registration methods on brain MRI data.

Table 3: Registration accuracy (mm) on lung CT data using B-spline transformation. The best result at each level is marked bold.

	SGD-constant	SGD-adaptive	Avg-constant	Avg-adaptive
Level 1	1.66±0.37	1.92±0.77	**1.65±0.32**	1.92±0.78
Level 2	1.45±0.28	1.46±0.35	**1.43±0.28**	1.46±0.35

The results of final registration accuracy achieved by candidate approaches at each registration level are presented in Table 3. It can be observed that the proposed Avg-constant method constantly generated the best registration accuracy at each level. In comparison with the previous research on the same data (Papież et al., 2014), the averaged registration accuracy reported in Table 3 is better.

Figures 7 (b) and (d) show the reproducibility $\mathrm{Std}(\Gamma_r(k))$ over 20 random seeds. The reproducibilities are averaged over 10 patients. As shown in Figure 7 (b), the final values of $\mathrm{Std}(\Gamma_r(k))$ of the SGD-constant and Avg-constant approaches are better than the methods using adaptively decreasing γ. It can be noticed that the Avg-constant method achieved the best reproducibility. At the second registration level (Figure 7 (d)), the SGD-constant method was the least reproducible among all methods. In contrast, the reproducibility was improved by using the Avg-constant method.

Figure 8 shows the degree of fluctuation at the final registration level on lung data. It can be observed that the SGD-constant method produced the largest fluctua-

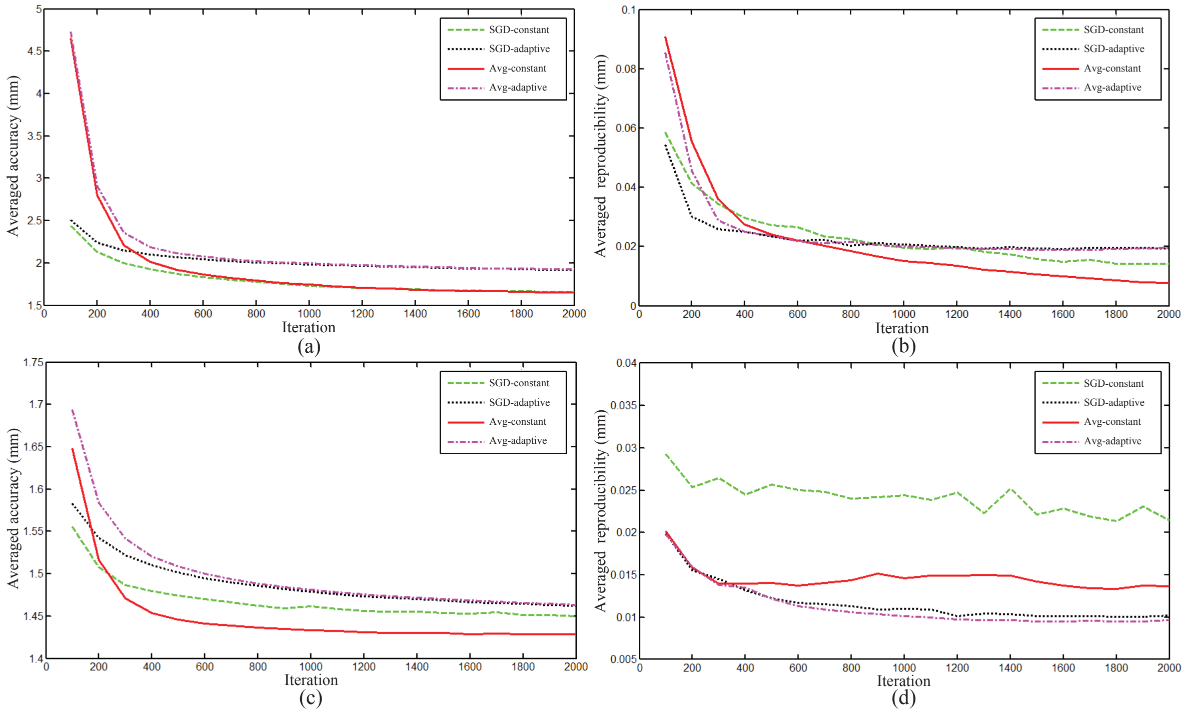


Figure 7: Convergence curve and reproducibility curves of registration methods on lung CT data using B-spline transformation. (a) and (c) averaged convergence curves on from the first to the second registration levels over 10 × 20 test cases; (b) and (d) averaged reproducibility curves from the first to the second registration levels over 10 × 20 test cases.

tion among all methods. However, the proposed Avg-constant approach reduced the fluctuation substantially.

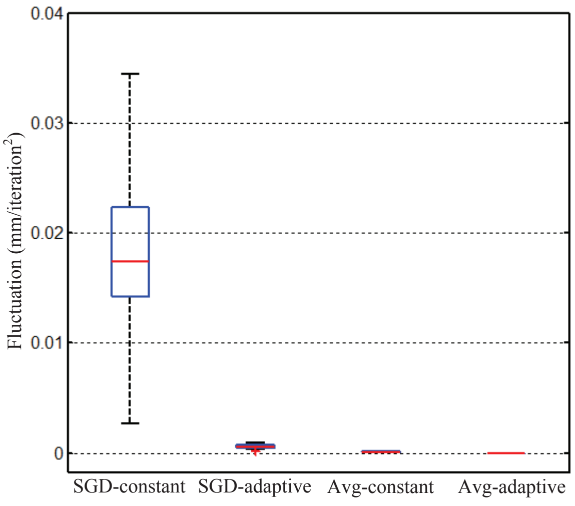


Figure 8: Degrees of fluctuations of registration methods on lung CT data.

5. Discussion

We proposed the Avg-SGD optimization method for image registration. The Avg-SGD method uses a constant instead of a decreasing step size γ, in order to accelerate optimization. The averaging process of the Avg-SGD approach compensates for the increased noise due to constant γ. The performance of the Avg-SGD method was evaluated in comparison to the state-of-the-art ASGD optimizer. The Avg-SGD approach was applied to both multimodal and monomodal registration problems including rigid and B-spline transformation models. The improvements in registration accuracy, registration reproducibility and fluctuation of the convergence curve prove the effectiveness of the Avg-SGD method.

In the experimental results, the postponed Avg-SGD method with constant step size could improve the convergence rate, but it was very sensitive to the choice of k_0. For a too small k_0, the effect of 'premature' $\boldsymbol{\mu}_i$ hampers the convergence rate in a practical situation with finite number of iterations. Meanwhile, the convergence rate could also be slowed down by a too large k_0. As a feasible alternative, the exponential Avg-SGD approach is introduced in this work. The exponential Avg-SGD approach starts the averaging operation at the beginning of optimization. For a constant $\epsilon \in (0, 1)$, the effects of previous iterates are exponentially decreased. In this way, the handicap of the postponed Avg-SGD method is avoided in the exponential Avg-SGD approach. In all experiments, we fixed $\epsilon = 0.01$ and obtained promising convergence rates and registration precision. Nevertheless, estimating an optimal ϵ based on characteristics of the registration problem (image data, similarity measure, transformation model) could be an interesting direction for future research.

For large and intermediate initial deformations, i.e., at the early registration levels, both the SGD-constant and Avg-constant methods achieved faster convergence rate than the approaches using adaptively decreasing γ. However, the constant γ resulted in large fluctuations around the optimum and thus lower reproducibility of the final solution obtained by the SGD-constant approach. The proposed Avg-SGD approach compensates for the fluctuations caused by constant γ, which was confirmed by the improved reproducibility curves. Compared with the methods using constant γ, the approaches using adaptively decreasing γ even obtained lower reproducibility. The reason is that these methods did not really converge yet at $k = K$, which causes a higher variation over different random seeds.

For small initial deformations, i.e., at the last registration level, the noise caused by the constant γ has significant influence on the SGD method. The difference of the convergence rate obtained by the SGD-constant and Avg-constant methods was distinct. The final registration accuracy produced by the SGD-constant approach was similar to or even got worse than the approaches using adaptively decreasing γ. However, the better convergence rate was still preserved by using the new Avg-SGD method. The reproducibility of the result in case of small initial deformation proves that the Avg-SGD method can effectively reduce the noise effect due to constant γ.

The fluctuation of the convergence curve was also evaluated to measure registration precision. The variation of the accuracy over iterations is reflected by the fluctuation measure. The boxplots (Figures 6 and 8) clearly showed that the SGD-constant method results in large fluctuations in the final iterations. This is exactly the reason why SGD methods are usually implemented with decreasing step sizes; due to the stochastic noise on the gradients $\tilde{\boldsymbol{g}}_k$ convergence needs to be enforced in some way. To achieve this, commonly, the step size is decreased with increasing iteration number k. Indeed, the SGD-adaptive method resulted in low fluctuation values. In this work, the Avg-SGD method proved to be another effective way to dampen the fluctuations and enforce convergence, eliminating the need to decrease the step size.

Regarding the computation time, the Avg-SGD approach was implemented in a recursive manner. During the optimization, only one set of averaged transformation parameters needs to be stored. To update the averaged parameters we only need simple vector operations. Therefore, the additional computational cost raised by the Avg-SGD method is trivial in practice.

As future work, it would be interesting to explore other extensions of SGD known from the literature, and study their performance on image registration problems in comparison to the Avg-SGD method. Candidate approaches are for example the momentum-based techniques that are widely used for training of neural networks (Qian, 1999), AdaGrad (Duchi et al., 2011), and distributed SGD methods (Zinkevich et al., 2010; Recht et al., 2011). Finally, it would be interesting to evaluate whether the exponentially weighted iterate averaging scheme is also beneficial for classic deterministic (i.e., non-stochastic) gradient-descent optimization methods.

6. Conclusion

In this work, we developed the Avg-SGD optimization method for image registration. The proposed approach compensates for the stochastic noise inherent to SGD by averaging over iterations. Thanks to the iterative averaging, large step sizes can be maintained throughout the entire optimization process, resulting in accelerated convergence while preserving the registration precision. The improved registration results demonstrate the effectiveness of the Avg-SGD method.

7. Acknowledgements

Funding: This work was supported by the National Natural Science Foundation of China [grant number U1301251].

References

Bhagalia, R., Fessler, J., Kim, B., *et al.*, 2009. Accelerated nonrigid intensity-based image registration using importance sampling. IEEE T. Med. Imaging 28, 1208–1216.

Bordes, A., Bottou, L., Gallinari, P., 2009. SGD-QN: careful quasi-Newton stochastic gradient descent. J. Mach. Learn. Res. 10, 1737–1754.

Bottou, L., 2010. Large-scale machine learning with stochastic gradient descent, in: COMPSTAT. Springer, pp. 177–186.

Bottou, L., 2014. Stochastic Gradient Descent. `http://leon.bottou.org/projects/sgd` (accessed October 13, 2016).

Bottou, L., Le Cun, Y., 2005. On-line learning for very large data sets. Appl. Stoch. Model. Bus. 21, 137–151.

Bousquet, O., Bottou, L., 2008. The tradeoffs of large scale learning, in: NIPS, pp. 161–168.

Castillo, R., Castillo, E., R., G., *et al.*, 2009. A framework for evaluation of deformable image registration spatial accuracy using large landmark point sets. Phys. Med. Biol. 54, 1849–1870.

Cocosco, C.A., Kollokian, V., Kwan, R.K.S., 1997. Brainweb: online interface to a 3D MRI simulated brain database. NeuroImage 5, 425.

Crum, W., Hartkens, T., Hill, D., 2004. Non-rigid image registration: theory and practice. Brit. J. Radiol. 77, S140–S153.

Dai, Y.H., 2003. A family of hybrid conjugate gradient methods for unconstrained optimization. Math. Comput. 72, 1317–1328.

Dennis, Jr, J.E., Moré, J.J., 1977. Quasi-Newton methods, motivation and theory. SIAM Rev. 19, 46–89.

Duchi, J., Hazan, E., Singer, Y., 2011. Adaptive subgradient methods for online learning and stochastic optimization. J. Mach. Learn. Res. 12, 2121–2159.

de Groot, M., Vernooij, M.W., Klein, S., *et al.*, 2013. Improving alignment in tract-based spatial statistics: evaluation and optimization of image registration. Neuroimage 76, 400–411.

Hansen, N., Ostermeier, A., 2001. Completely derandomized self-adaptation in evolution strategies. Evol. Comput. 9, 159–195.

Hill, D., Batchelor, P., Holden, M., *et al.*, 2001. Medical image registration. Phys. Med. Biol. 46, R1–R45.

Kiefer, J., Wolfowitz, J., 1952. Stochastic estimation of the maximum of a regression function. Ann. Math. Stat. 23, 462–466.

Klein, S., van der Heide, U.A., Lips, I.M., *et al.*, 2008. Automatic segmentation of the prostate in 3D MR images by atlas matching using localized mutual information. Med. Phys. 35, 1407–1417.

Klein, S., Pluim, J.P.W., Staring, M., *et al.*, 2009. Adaptive stochastic gradient descent optimisation for image registration. Int. J. Comput Vision 81, 227–239.

Klein, S., Staring, M., Andersson, P., *et al.*, 2011. Preconditioned stochastic gradient descent optimisation for monomodal image registration, in: MICCAI. Springer, pp. 549–556.

Klein, S., Staring, M., Murphy, K., *et al.*, 2010. Elastix: a toolbox for intensity-based medical image registration. IEEE T. Med. Imaging 29, 196–205.

Klein, S., Staring, M., Pluim, J.P.W., 2007. Evaluation of optimization methods for nonrigid medical image registration using mutual information and B-splines. IEEE T. Image Process. 16, 2879–2890.

Kushner, H.J., Yin, G., 2003. Stochastic approximation and recursive algorithms and applications. volume 35. Springer Science & Business Media.

Maes, F., Collignon, A., Vandermeulen, D., *et al.*, 1997. Multimodality image registration by maximization of mutual information. IEEE T. Med. Imaging 16, 187–198.

Maintz, J., Viergever, M.A., 1998. A survey of medical image registration. Med. Image Anal. 2, 1 – 36.

Mattes, D., Haynor, D.R., Vesselle, H., *et al.*, 2003. PET-CT image registration in the chest using free-form deformations. IEEE T. Med. Imaging 22, 120–128.

Metz, C., Klein, S., Schaap, M., *et al.*, 2011. Nonrigid registration of dynamic medical imaging data using *n*D+t B-splines and a groupwise optimization approach. Med. Image Anal. 15, 238–249.

Murphy, K., Van Ginneken, B., Reinhardt, J., *et al.*, 2011. Evaluation of registration methods on thoracic CT: The EMPIRE10 challenge. IEEE T. Med. Imaging 30, 1901–1920.

Nocedal, J., Wright, S.J., 1999. Numerical optimization. New York: Springer-Verlag.

Papież, B.W., Heinrich, M.P., Fehrenbach, J., *et al.*, 2014. An implicit sliding-motion preserving regularisation via bilateral filtering for deformable image registration. Med. Image Anal. 18, 1299–1311.

Polyak, B.T., Juditsky, A.B., 1992. Acceleration of stochastic approximation by averaging. SIAM J. Control Optim. 30, 838–855.

Qian, N., 1999. On the momentum term in gradient descent learning algorithms. Neural networks 12, 145–151.

Qiao, Y., van Lew, B., Lelieveldt, B.P.F., *et al.*, 2016. Fast automatic step size estimation for gradient descent optimization of image registration. IEEE T. Med. Imaging 35, 391–403.

Qiao, Y., Sun, Z., Lelieveldt, B.P., *et al.*, 2015. A stochastic quasi-newton method for non-rigid image registration, in: MICCAI. Springer, pp. 297–304.

Recht, B., Re, C., Wright, S., *et al.*, 2011. Hogwild: A lock-free approach to parallelizing stochastic gradient descent, in: NIPS, pp. 693–701.

Robbins, H., Monro, S., 1951. A stochastic approximation method. Ann. Math. Stat. , 400–407.

Rueckert, D., Sonoda, L.I., Hayes, C., *et al.*, 1999. Nonrigid registration using free-form deformations: application to breast MR images. IEEE T. Med. Imaging 18, 712 – 721.

Ruppert, D., 1988. Efficient estimations from a slowly convergent Robbins-Monro process. Technical Report. Cornell University Operations Research and Industrial Engineering.

Smal, I., Carranza-Herrezuelo, N., Klein, S., *et al.*, 2012. Reversible jump MCMC methods for fully automatic motion analysis in tagged MRI. Med. Image Anal. 16, 301–324.

Sotiras, A., Davatzikos, C., Paragios, N., 2013. Deformable medical image registration: a survey. IEEE T. Med. Imaging 32, 1153–1190.

Spall, J.C., 1992. Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. IEEE T. Automat. Contr. 37, 332–341.

Sun, W., Niessen, W., van Stralen, M., *et al.*, 2013. Simultaneous multiresolution strategies for nonrigid image registration. IEEE T. Image Process. 22, 4905–4917.

Sun, W., Poot, D.H., Smal, I., Yang, X., Niessen, W.J., Klein, S., 2017. Stochastic optimization with randomized smoothing for image registration. Medical Image Analysis 35, 146–158.

Toga, A., Thompson, P., 2001. The role of image registration in brain mapping. Image Vision Comput. 19, 3–24.

Ushiku, Y., Hidaka, M., Harada, T., 2014. Three guidelines of online learning for large-scale visual recognition, in: CVPR, IEEE. pp. 3574–3581.

Viergever, M.A., Maintz, J.A., Klein, S., Murphy, K., Staring, M., Pluim, J.P., 2016. A survey of medical image registration c under review. Medical Image Analysis 33, 140 – 144.

Viola, P., Wells III, W.M., 1997. Alignment by maximization of mutual information. Int. J. Comput Vision 24, 137–154.

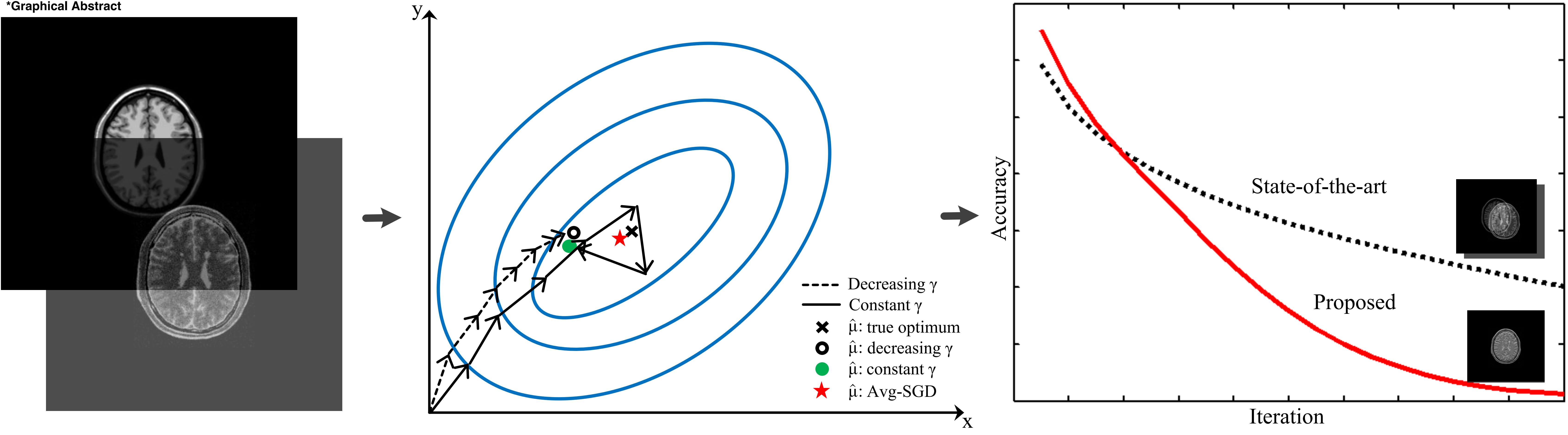
Yin, G., 1992. Stochastic approximation via averaging: the Polyaks approach revisited, in: Simulation and Optimization. Springer, pp. 119–134.

Zinkevich, M., Weimer, M., Li, L., *et al.*, 2010. Parallelized stochastic gradient descent, in: NIPS, pp. 2595–2603.

*Highlights

- We propose an averaged SGD (Avg-SGD) method for efficient image registration.
- A constant step size is used, in combination with an exponentially weighted iterate averaging scheme in the Avg-SGD approach.
- Extensive experiments on simulated 2D brain MRI data and real 3D lung CT scans with both rigid and nonrigid transformations were carried out to evaluate the performance of the Avg-SGD method.
- The new Avg-SGD method achieved better convergence rate and registration accuracy than the state-of-the-art approach while still keeping competitive registration precision.

***Graphical Abstract**



y
x
- - - Decreasing γ
—— Constant γ
✖ µ̂: true optimum
○ µ̂: decreasing γ
● µ̂: constant γ
★ µ̂: Avg-SGD
Accuracy
Iteration
State-of-the-art
Proposed