Adaptive stochastic gradient descent optimisation for image registration

Int J Comput Vis (2009) 81: 227–239
DOI 10.1007/s11263-008-0168-y

Adaptive Stochastic Gradient Descent Optimisation for Image Registration

Stefan Klein · Josien P.W. Pluim · Marius Staring · Max A. Viergever

Received: 21 March 2008 / Accepted: 4 August 2008 / Published online: 28 August 2008

Abstract We present a stochastic gradient descent optimisation method for image registration with adaptive step size prediction. The method is based on the theoretical work by Plakhov and Cruz (J. Math. Sci. 120(1):964–973, 2004). Our main methodological contribution is the derivation of an image-driven mechanism to select proper values for the most important free parameters of the method. The selection mechanism employs general characteristics of the cost functions that commonly occur in intensity-based image registration. Also, the theoretical convergence conditions of the optimisation method are taken into account. The proposed adaptive stochastic gradient descent (ASGD) method is compared to a standard, non-adaptive Robbins-Monro (RM) algorithm. Both ASGD and RM employ a stochastic subsampling technique to accelerate the optimisation process. Registration experiments were performed on 3D CT and MR data of the head, lungs, and prostate, using various similarity measures and transformation models. The results indicate that ASGD is robust to these variations in the registration framework and is less sensitive to the settings of the user-defined parameters than RM. The main disadvantage of RM is the need for a predetermined step size function. The ASGD method provides a solution for that issue.

Keywords Image registration · Optimisation · Stochastic gradient descent · Adaptive step sizes · Parameter selection

S. Klein (✉) · J.P.W. Pluim · M. Staring · M.A. Viergever
Image Sciences Institute, University Medical Center Utrecht, Q.S.459, P.O. Box 85500, 3508 GA Utrecht, The Netherlands
e-mail: s.klein@erasmusmc.nl

1 Introduction

Image registration is a frequently used technique in the fields of remote sensing and medical imaging. Given a pair of images, image registration is the task of finding a coordinate transformation that spatially aligns the two images. Extensive surveys of registration methods can be found in the literature (Maintz and Viergever 1998; Pluim et al. 2003; Zitová and Flusser 2003). In this article, we focus on intensity-based registration methods, using a parameterised coordinate transformation.

Intensity-based image registration is usually treated as a nonlinear optimisation problem. Define the fixed image $F(\boldsymbol{x}) : \Omega_F \subset \mathbb{R}^D \mapsto \mathbb{R}$, the moving image $M(\boldsymbol{x}) : \Omega_M \subset \mathbb{R}^D \mapsto \mathbb{R}$, and a parameterised coordinate transformation $\boldsymbol{T}(\boldsymbol{x}, \boldsymbol{\mu}) : \Omega_F \times \mathbb{R}^P \mapsto \Omega_M$, where $\boldsymbol{\mu} \in \mathbb{R}^P$ represents the vector of transformation parameters. The following minimisation problem is considered:

$$\hat{\boldsymbol{\mu}} = \arg\min_{\boldsymbol{\mu}} \mathcal{C}(F, M \circ \boldsymbol{T}), \tag{1}$$

where $\mathcal{C}$ is the cost function (or "similarity measure") that measures the similarity of the fixed image and the deformed moving image. The solution $\hat{\boldsymbol{\mu}}$ is the parameter vector that minimises that cost function. Henceforth, we use the short notation $\mathcal{C}(\boldsymbol{\mu}) \equiv \mathcal{C}(F, M \circ \boldsymbol{T})$.

In Klein et al. (2007) it has been shown that a Robbins-Monro (RM) stochastic gradient descent method (Robbins and Monro 1951; Kushner and Yin 2003) is in many applications the best choice for solving the minimisation problem (1). The method uses the following iterative scheme:

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

$$\tilde{\boldsymbol{g}}_k = \boldsymbol{g}(\boldsymbol{\mu}_k) + \boldsymbol{\varepsilon}_k, \tag{3}$$

where $\tilde{\boldsymbol{g}}_k$ denotes an approximation of the true derivative $\boldsymbol{g} \equiv \partial \mathcal{C}/\partial \boldsymbol{\mu}$ at $\boldsymbol{\mu}_k$, and $\boldsymbol{\varepsilon}_k$ is the approximation error. If $\boldsymbol{\varepsilon}_k = \mathbf{0}$, (2) boils down to a common, deterministic gradient descent method. The approximation of $\boldsymbol{g}$ is realised by computing $\boldsymbol{g}$ using not all voxels, but only a small subset of voxels, randomly selected in every iteration. In this way, the computational costs per iteration are greatly reduced, while convergence properties are still similar to those obtained by deterministic gradient descent. The scalar gain factor γ_k, the "step size", is determined by a predefined decaying function of the iteration number k. An often used choice is:

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

with user-specified constants $a > 0$, $A \geq 1$, and $0 < \alpha \leq 1$. A choice of $\alpha = 1$ gives a theoretically optimum rate of convergence when $k \to \infty$ (Kushner and Yin 2003). In practice, the algorithm is stopped after a specified maximum number of iterations, and, therefore, it sometimes makes sense to choose $\alpha < 1$, which causes the step size to decay less fast. The need for setting a, A, and α complicates the usage of RM for image registration. The factor a is especially difficult, since it has no unit, and heavily depends on the choice of the cost function. For example, when we multiply $\mathcal{C}$ by an arbitrary constant c, the value of a would need to be divided by c in order to get the same sequence $\{\boldsymbol{\mu}_k\}$. When a is set too small, the RM method suffers from slow convergence. When a is set too large, the process may become unstable.

The present study concerns a stochastic optimisation method with adaptive step size prediction: adaptive stochastic gradient descent (ASGD). The mechanism to adapt the step size γ_k is based on the inner product of the gradient $\tilde{\boldsymbol{g}}_k$ and the previous gradient $\tilde{\boldsymbol{g}}_{k-1}$. Intuitively, if the gradients in two consecutive iterations point in (almost) the same direction, it is expected that larger steps can be taken. If the gradients point in opposite directions, the step size is reduced. The theoretical convergence properties of the method in one-dimensional ($P = 1$) optimisation problems were studied by Plakhov and Cruz (2004). Cruz (2005a) extended the analysis to multidimensional ($P > 1$) problems. Some numerical experiments are described in (Cruz 2005b), using artificial test functions with $\boldsymbol{\varepsilon}_k$ generated according to a normal distribution. Only two cases ($P = 1$ and $P = 2$) were investigated. No other applications of the method were found in the literature.

Using the theoretical convergence conditions given in (Cruz 2005a), we derive an image-driven selection mechanism for the method's free parameters. The derivation is based on general characteristics of the cost functions that commonly occur in intensity-based image registration problems. A key result is the replacement of a by a new user-defined parameter, δ, which has a more intuitive meaning and is constructed to be independent of the choice of $\mathcal{C}$. The method is validated on several registration problems, with different image modalities, similarity measures, and transformation models, with P ranging from 6 to 4000.

2 Method

First, in Sect. 2.1, the basic ASGD method is explained and a summary is given of the theoretical convergence results. After that, in Sect. 2.2, we describe the first steps towards application of ASGD in image registration. A procedure to set the free parameters of ASGD is derived in Sects. 2.3–2.5. Section 2.6 gives an overview of the entire algorithm.

2.1 Summary of ASGD

In Cruz (2005a) the ASGD method is presented in the context of a general multidimensional root-finding problem[1]: find $\hat{\boldsymbol{\mu}}$ such that $\boldsymbol{\varphi}(\hat{\boldsymbol{\mu}}) = \mathbf{0}$, for some function $\boldsymbol{\varphi}(\boldsymbol{\mu}) : \mathbb{R}^P \mapsto \mathbb{R}^P$. Our minimisation problem is a specific case of this, where $\boldsymbol{\varphi}$ equals $\boldsymbol{g} \equiv \partial \mathcal{C}/\partial \boldsymbol{\mu}$. The ASGD algorithm is then defined as:

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

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

where $[x]^+$ means $\max(x, 0)$, f denotes a sigmoid function, and $\boldsymbol{\mu}_0$, t_0 and t_1 are user-defined initial conditions. For the γ function, the same definition as in (4) can be used. However, in ASGD, the γ function is not evaluated at the iteration number k, as in (2), but at the 'time' t_k. The time is adapted depending on the inner product of the gradient $\tilde{\boldsymbol{g}}_k$ and the previous gradient $\tilde{\boldsymbol{g}}_{k-1}$. If the gradients in two consecutive steps point in the same direction, the inner product is positive, and therefore the time is reduced, which leads to a larger step size $\gamma(t_{k+1})$, since γ is a monotone decreasing function. In this way, the ASGD method implements an adaptive step size mechanism. Note that if we would use $f(x) = 1$, the original RM method is obtained.

The article by Cruz (2005a) provides a proof of "almost-sure" convergence and a proof of asymptotical normality. The proof of almost-sure convergence implies that

$$\lim_{k \to \infty} \boldsymbol{\mu}_k = \hat{\boldsymbol{\mu}}, \tag{7}$$

"with probability 1". The proof of asymptotical normality tells us something about the rate of convergence:

$$\sqrt{k}(\boldsymbol{\mu}_k - \hat{\boldsymbol{\mu}}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{V}), \tag{8}$$

where $\xrightarrow{d}$ indicates convergence in distribution and $\mathcal{N}(\mathbf{0}, \boldsymbol{V})$ denotes a multivariate normal distribution with mean $\mathbf{0}$ and

[1] Note that our notation is somewhat different from Cruz (2005a).

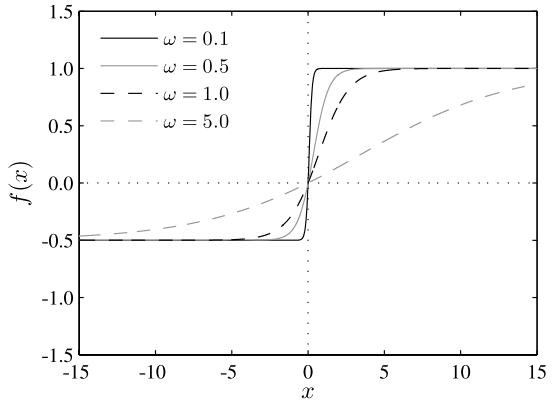

Fig. 1 Examples of the sigmoid function f, with $f_{\text{MAX}} = 1$ and $f_{\text{MIN}} = -0.5$

covariance matrix $\boldsymbol{V}$. To prove the convergence and asymptotical normality, five sets of assumptions are required. The assumptions impose conditions on γ and f, depending on characteristics of the cost function $\mathcal{C}$ and the distribution of gradient approximation errors $\boldsymbol{\varepsilon}_k$.

2.2 Application of ASGD

To apply the ASGD method in practice, we have to specify the γ and f functions. They should be chosen such that the theoretical convergence conditions given in Cruz (2005a) are satisfied.

For the step size function γ we choose the following expression:

$$\gamma(t) = a/(t + A), \tag{9}$$

with $a > 0$ and $A \geq 1$. Compared to (4) the α term is omitted, i.e. $\alpha = 1$, which is the theoretically optimal setting (Kushner and Yin 2003). For f we define a general sigmoid shape with $f(0) = 0$:

$$f(x) = f_{\text{MIN}} + \frac{f_{\text{MAX}} - f_{\text{MIN}}}{1 - (f_{\text{MAX}}/f_{\text{MIN}})e^{-x/\omega}}, \tag{10}$$

with $f_{\text{MAX}} > 0$, $f_{\text{MIN}} < 0$, and $\omega > 0$. Examples of f are shown in Fig. 1. If $\omega \downarrow 0$, the sigmoid approaches a step function.

The ASGD algorithm still requires setting a and A. Moreover, the expression for the sigmoid function f introduces three new parameters: f_{MAX}, f_{MIN}, and ω. Yet, we expect that the adaptive step size mechanism makes the algorithm robust for wider ranges of a and A, compared with RM.

As mentioned in Sect. 2.1, five sets of assumptions are used in Cruz (2005a) to prove convergence and asymptotical normality of the ASGD algorithm. The assumptions impose conditions on γ and f, and are thus important for determining proper values for a, A, f_{MIN}, f_{MAX}, and ω. We now study the assumptions after substitution of the above choices for γ and f. Like in Cruz (2005a) the sets of assumptions needed to prove convergence are numbered B1–B4. The set of assumptions used to prove asymptotical normality is called B5. In comparison with Cruz (2005a), some conditions have been simplified using $\boldsymbol{\varphi} = \partial\mathcal{C}/\partial\boldsymbol{\mu} \equiv \boldsymbol{g}$ (see Sect. 2.1). Also, technical details that are not relevant for this article are omitted. Our comments on the assumptions are written in *italic*.

Assumption B1 (Properties of $\boldsymbol{\varepsilon}_k$) The approximation errors $\boldsymbol{\varepsilon}_k$ are independent identically distributed random vectors with zero mean $\text{E}\boldsymbol{\varepsilon}_k = 0$ and finite covariance matrix $\boldsymbol{\Sigma} \equiv \text{Var}\boldsymbol{\varepsilon}_k$.

Based on characteristics of the cost function $\mathcal{C}$, we postulate in Sect. 2.3 that $\boldsymbol{\varepsilon}_k$ has a normal distribution: $\boldsymbol{\varepsilon}_k \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$.

Assumptions B2 (Properties of $\gamma(t)$)

1. The gain function $\gamma(t)$ is a positive monotone decreasing function defined on $[0, \infty)$. Consequently, $\gamma(0)$ is the maximum gain factor.
2. $\int_0^\infty \gamma(t)dt = \infty$.
3. $\int_0^\infty [\gamma(t)]^2 dt < \infty$.

With $\gamma(t)$ defined by (9) it is easily verified that these assumptions are satisfied and that $\gamma(0) = a/A$.

Assumptions B3 (Conditions depending on $\mathcal{C}$)

1. Provided that
   a) the function $\mathcal{C}(\boldsymbol{\mu})$ has no other extrema than $\hat{\boldsymbol{\mu}}$,
   b) $\mathcal{C}(\boldsymbol{\mu})$ is continuous and twice differentiable everywhere,
   c) there exists a constant $\lambda > 0$ such that the maximum eigenvalue of the Hessian $\boldsymbol{H} \equiv \partial^2\mathcal{C}/\partial\boldsymbol{\mu}\partial\boldsymbol{\mu}^T$ is smaller than or equal to λ for all $\boldsymbol{\mu}$,

   the minimisation problem (1) can be solved with the following deterministic gradient descent method:

   $$\boldsymbol{\mu}_{k+1} = \boldsymbol{\mu}_k - \hat{\gamma}\boldsymbol{g}(\boldsymbol{\mu}_k), \tag{11}$$

   for each $\hat{\gamma} < \gamma(0)$, and for each $\boldsymbol{\mu}_0$.

   Provided that Assumptions B3.1(a)–(c) indeed hold, the choice $\gamma(0) = 2/\lambda$ satisfies the last condition (Shi and Shen 2005). *This assumption thus relates the maximum step size $\gamma(0)$ to the Hessian of the cost function.*
2. There exist $R > 0$ and $\beta_0 > 0$ such that[2]

   $$\|\boldsymbol{g}(\boldsymbol{\mu})\|^2 \geq \frac{1}{2}\gamma(0)\lambda(\|\boldsymbol{g}(\boldsymbol{\mu})\|^2 + \text{tr}(\boldsymbol{\Sigma})) + \beta_0, \tag{12}$$

   for all $\boldsymbol{\mu}$ that satisfy $\|\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}\| \geq R$.

[2] $\text{tr}(\cdot)$ stands for the matrix trace.

This condition relates the maximum step size $\gamma(0)$ to the covariance matrix $\mathbf{\Sigma}$ of the approximation errors. In Sect. 2.4, we use Assumptions B3.1 and B3.2 to choose a value for a.

Assumptions B4 (Properties of $f(x)$)

1. $f(x): \mathbb{R} \mapsto \mathbb{R}$ is a monotone increasing, continuous and bounded function, for which:

$$f_{\text{MAX}} = \lim_{x \to +\infty} f(x) > 0 \quad \text{and} \quad f_{\text{MIN}} = \lim_{x \to -\infty} f(x) \tag{13}$$

The expression for $f(x)$ defined in (10) has been constructed such that Assumption B4.1 is satisfied.

2. Define $E_0 \equiv \mathrm{E} f(\boldsymbol{\varepsilon}_k^T \boldsymbol{\varepsilon}_{k-1})$. The constant E_0 must be positive.

The condition $E_0 > 0$ is satisfied when $f(x) > -f(-x)$ for all $x \neq 0$, provided that $\mathbf{\Sigma} \neq \mathbf{0}$. Combined with (10), this imposes that $f_{\text{MAX}} > -f_{\text{MIN}}$. Furthermore, if $\boldsymbol{\varepsilon}_k \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ and $\omega \downarrow 0$, then $E_0 \uparrow \frac{1}{2}(f_{\text{MAX}} + f_{\text{MIN}})$. In Sect. 2.5 this is used to choose values for f_{MAX} and f_{MIN}. Also, a value for ω is determined, such that indeed $E_0 \approx \frac{1}{2}(f_{\text{MAX}} + f_{\text{MIN}})$.

Assumptions B5 (Asymptotic normality) The following conditions are used to prove asymptotic normality:

1. $\gamma(t) = 1/t$.
2. Define the matrix $\boldsymbol{W}$:

$$\boldsymbol{W} = \frac{1}{2}\boldsymbol{I} - \frac{1}{E_0}\boldsymbol{H}(\hat{\boldsymbol{\mu}}), \tag{14}$$

with $\boldsymbol{I}$ the identity matrix. All eigenvalues of $\boldsymbol{W}$ must be negative.

3. $f(x)$ is a continuous and differentiable function.

Assumption B5.3 is obviously satisfied when $\omega \neq 0$. Our choice for $\gamma(t)$ breaks with Assumption B5.1. However, the proof of asymptotic normality can be easily extended to take our choice of $\gamma(t)$ into account (for this, note that A is a finite constant, which plays no role anymore when $t \to \infty$). The first two assumptions are then modified to:

1. $\gamma(t) = a/(t + A)$.
2. Define the matrix $\boldsymbol{W}$:

$$\boldsymbol{W} = \frac{1}{2}\boldsymbol{I} - \frac{a}{E_0}\boldsymbol{H}(\hat{\boldsymbol{\mu}}), \tag{15}$$

with $\boldsymbol{I}$ the identity matrix. All eigenvalues of $\boldsymbol{W}$ must be negative.

Assumptions B4.2 and B5.2 are used in Sect. 2.5 to choose a value for f_{MAX} and f_{MIN}.

In the following subsections, estimates are derived for the distributions of $\boldsymbol{g}$, $\tilde{\boldsymbol{g}}_k$, $\boldsymbol{\varepsilon}_k$, and the voxel displacements between two iterations. Based on these results and some of the Assumptions B1–B5 mentioned above, settings for a, f_{MIN}, f_{MAX}, and ω are proposed. The value of A is left unspecified. The parameter a is replaced by a new user-defined parameter δ, which, unlike a, has a unit (mm), and an intuitive meaning. Also, it is constructed to be independent of the choice of $\mathcal{C}$.

2.3 Distribution Estimates

In this section we devise expressions for the distributions of $\boldsymbol{g}$, $\tilde{\boldsymbol{g}}_k$ and $\boldsymbol{\varepsilon}_k$, based on the characteristics of the cost function in image registration problems. Using the distribution of $\boldsymbol{g}$, the distribution of voxel displacements per iteration of a deterministic gradient descent process is calculated. The results of this subsection are needed in Sects. 2.4 and 2.5.

In image registration, the cost function usually takes the following form:

$$\mathcal{C}(\boldsymbol{\mu}) = \Psi\left(\frac{1}{|\Omega'_F|}\sum_{\boldsymbol{x}_i \in \Omega'_F} \xi(F(\boldsymbol{x}_i), M(\boldsymbol{T}(\boldsymbol{x}_i, \boldsymbol{\mu})))\right), \tag{16}$$

with $\Psi(u): \Xi \mapsto \mathbb{R}$ and $\xi(u, v): \mathbb{R} \times \mathbb{R} \mapsto \Xi$ continuous, differentiable functions, $\Omega'_F \subset \Omega_F$ the discrete set of voxel coordinates $\boldsymbol{x}_i$ of the fixed image, and $|\Omega'_F|$ the cardinality of this set. The domain Ξ may be simply equal to $\mathbb{R}$, but may also be of a multidimensional nature: $\mathbb{R}^P$ or $\mathbb{R}^{P \times Q}$, for example. An example that is covered by (16) is the sum of squared differences: $\Xi = \mathbb{R}$, $\Psi(u) = u$, $\xi(u, v) = (u - v)^2$. Another example is mutual information (Collignon et al. 1995; Viola and Wells III 1995; Thévenaz and Unser 2000; Hermosillo et al. 2002), for which:

$$\Xi = \mathbb{R}^{P \times Q}, \tag{17}$$

$$\Psi(u) = \sum_{p=1}^{P}\sum_{q=1}^{Q} u_{pq} \log \frac{u_{pq}}{(\sum_p u_{pq})(\sum_q u_{pq})}, \tag{18}$$

$$\xi_{pq}(u, v) = \beta(p - u)\beta(q - v), \tag{19}$$

with $P \times Q$ the joint histogram size, and $\beta(u): \mathbb{R} \mapsto \mathbb{R}$ a Parzen window function.

We now take the derivative of (16). For clarity of notation we consider the case $\Xi = \mathbb{R}$:

$$\boldsymbol{g} \equiv \frac{\partial \mathcal{C}}{\partial \boldsymbol{\mu}} = \frac{1}{|\Omega'_F|}\sum_{\boldsymbol{x}_i \in \Omega'_F} \frac{\partial \boldsymbol{T}}{\partial \boldsymbol{\mu}}^T \frac{\partial M}{\partial \boldsymbol{x}} \frac{\partial \xi}{\partial v} \frac{\partial \Psi}{\partial u}. \tag{20}$$

We would like to estimate the distribution of $\boldsymbol{g}$ in a neighbourhood $\Upsilon \subset \mathbb{R}^P$ around $\hat{\boldsymbol{\mu}}$, containing $\boldsymbol{\mu}_0$. The idea is that this distribution predicts the gradients that will be measured during optimisation. The following two assumptions are needed:

Assumption A1 ($\partial \boldsymbol{T}/\partial \boldsymbol{\mu}$ is independent of $\boldsymbol{\mu}$) For each $\boldsymbol{x}_i \in \Omega'_F$ the following holds:

$$\boldsymbol{J}_i \equiv \frac{\partial \boldsymbol{T}}{\partial \boldsymbol{\mu}}(\boldsymbol{x}_i, \boldsymbol{\mu}_0) = \frac{\partial \boldsymbol{T}}{\partial \boldsymbol{\mu}}(\boldsymbol{x}_i, \boldsymbol{\mu}), \quad \forall \boldsymbol{\mu} \in \Upsilon. \tag{21}$$

This assumption holds when the transformation model is parameterised such that $\partial^2 \boldsymbol{T}/\partial \boldsymbol{\mu} \partial \boldsymbol{\mu}^T = \boldsymbol{0}$. The B-spline transformation (Rueckert et al. 1999) *is an example of such a parameterisation. Also an affine transformation, parameterised by the affine matrix elements, satisfies the assumption. For a rigid transformation parameterised by Euler angles the assumption is violated, since $\boldsymbol{T}$ then becomes a non-linear function of $\boldsymbol{\mu}$, but it holds approximately if Υ is not too large.*

Assumption A2 (Distribution of z_i) Based on (20), define:

$$z_i \equiv \frac{\partial M}{\partial \boldsymbol{x}} \frac{\partial \xi}{\partial v} \frac{\partial \Psi}{\partial u}. \tag{22}$$

Then, $\{z_i\}$ are mutually independent random vectors, identically distributed according to:

$$z_i \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I}), \tag{23}$$

with σ some constant.

This assumption is a simplification of reality. Any results based on this assumption must be validated.

Combining (20)–(23) gives us an estimate of the distribution of $\boldsymbol{g}$:

$$\begin{aligned} \boldsymbol{g} = \frac{1}{|\Omega'_F|} \sum_{\boldsymbol{x}_i \in \Omega'_F} \boldsymbol{J}_i^T z_i &\sim \mathcal{N}\left(\boldsymbol{0}, \frac{\sigma^2}{|\Omega'_F|^2} \sum_{\boldsymbol{x}_i \in \Omega'_F} \boldsymbol{J}_i^T \boldsymbol{J}_i\right) \\ &= \mathcal{N}\left(\boldsymbol{0}, \frac{\sigma^2}{|\Omega'_F|} \boldsymbol{C}\right), \end{aligned} \tag{24}$$

where we introduced:

$$\boldsymbol{C} \equiv \frac{1}{|\Omega'_F|} \sum_{\boldsymbol{x}_i \in \Omega'_F} \boldsymbol{J}_i^T \boldsymbol{J}_i. \tag{25}$$

The same approach can be followed for the approximated derivative $\tilde{\boldsymbol{g}}_k$. Approximation is realised by stochastic subsampling:

$$\tilde{\boldsymbol{g}}_k = \frac{1}{|S_k|} \sum_{\boldsymbol{x}_i \in S_k} \frac{\partial \boldsymbol{T}}{\partial \boldsymbol{\mu}}^T \frac{\partial M}{\partial \boldsymbol{x}} \frac{\partial \xi}{\partial v} \frac{\partial \Psi}{\partial u}, \tag{26}$$

with $S_k \subset \Omega'_F$ a set of samples, randomly selected in every iteration k. The distribution for $\tilde{\boldsymbol{g}}_k$ is estimated in the same way as above:

$$\tilde{\boldsymbol{g}}_k \sim \mathcal{N}\left(\boldsymbol{0}, \frac{\sigma^2}{|S_k|^2} \sum_{\boldsymbol{x}_i \in S_k} \boldsymbol{J}_i^T \boldsymbol{J}_i\right). \tag{27}$$

The following approximation is proposed:

$$\frac{1}{|S_k|} \sum_{\boldsymbol{x}_i \in S_k} \boldsymbol{J}_i^T \boldsymbol{J}_i \approx \frac{1}{|\Omega'_F|} \sum_{\boldsymbol{x}_i \in \Omega'_F} \boldsymbol{J}_i^T \boldsymbol{J}_i. \tag{28}$$

The approximation becomes more accurate for increasing $|S_k|$, and when $\boldsymbol{J}_i$ varies more gradually over the image domain Ω_F. Using this approximation, the expression for the distribution of $\tilde{\boldsymbol{g}}_k$ becomes:

$$\tilde{\boldsymbol{g}}_k \sim \mathcal{N}\left(\boldsymbol{0}, \frac{\sigma^2}{|S_k|} \boldsymbol{C}\right). \tag{29}$$

The distribution of the approximation errors $\boldsymbol{\varepsilon}_k = \boldsymbol{g} - \tilde{\boldsymbol{g}}_k$ is computed in a similar way, by subtracting (26) from (20), and using the approximation (28):

$$\boldsymbol{\varepsilon}_k \sim \mathcal{N}\left(\boldsymbol{0}, \sigma^2 \left(\frac{1}{|\Omega'_F|} - \frac{1}{|S_k|}\right) \boldsymbol{C}\right). \tag{30}$$

Note that when the number of samples $|S_k|$ is independent of k, the distributions of $\tilde{\boldsymbol{g}}_k$ and $\boldsymbol{\varepsilon}_k$ are also independent of k.

We turn our attention to Assumption A2. The assumptions states that z_i are independent random variables. In images, the assumption of independency between neighbouring voxels $\boldsymbol{x}_i$ is usually not satisfied. The image is a discretisation of a continuous signal, and thus observed (sampled) at a certain scale, possibly inducing dependencies between the grey values of neighbouring voxels. Consequently, the corresponding z_i may also be related. The impact of this on the distribution of $\boldsymbol{g}$, see (24), can be demonstrated by an imaginary experiment. Suppose we have experimentally estimated the distribution of $\boldsymbol{g}$ in some region Υ. Then, we resample the fixed image $F(\boldsymbol{x})$ on a twice as dense grid, using for example linear interpolation to interpolate between voxels, and repeat the experiment. Intuitively, we would not expect a different distribution of $\boldsymbol{g}$. However, the number of voxels in the fixed image, $|\Omega'_F|$, has increased with a factor 2^D, with D the dimension of the fixed image. According to (24), the variance of the distribution should therefore be divided by a factor 2^D, which is clearly wrong. We must conclude that the dependency of the variance on the number of voxels only holds when the z_i are truly independent. Since this is hard to verify, we propose to use the following distribution estimates, instead of (24), (29), and (30):

$$\boldsymbol{g} \sim \mathcal{N}(\boldsymbol{0}, \sigma_1^2 \boldsymbol{C}), \tag{31}$$

$$\tilde{\boldsymbol{g}}_k \sim \mathcal{N}(\boldsymbol{0}, \sigma_2^2 \boldsymbol{C}), \tag{32}$$

$$\boldsymbol{\varepsilon}_k \sim \mathcal{N}(\boldsymbol{0}, \sigma_3^2 \boldsymbol{C}) = \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}), \tag{33}$$

with σ_1, σ_2, and σ_3 unknown scalar constants, unrelated to each other, which will be experimentally determined. The last equality refers to the comment on Assumption B1. To estimate the constants σ_i, we perform N evaluations of $\boldsymbol{g}$,

$\tilde{\boldsymbol{g}}_k$, and $\boldsymbol{\epsilon}_k = \boldsymbol{g} - \tilde{\boldsymbol{g}}_k$, and fit σ_i such that the empirical average vector magnitudes equal the theoretical expectations of the vector magnitudes. For example, σ_1 is determined such that:

$$\frac{1}{N}\sum_{n=1}^{N} \|\boldsymbol{g}(\boldsymbol{\mu}_n)\|^2 = \sigma_1^2 \mathrm{tr}(\boldsymbol{C}), \tag{34}$$

where the right-hand side equals $\mathrm{E}\|\boldsymbol{g}\|^2$, in accordance with (31). The $\boldsymbol{\mu}_n$ vectors are randomly sampled around $\boldsymbol{\mu}_0$, using a normal distribution with diagonal covariance matrix:

$$\boldsymbol{\mu}_n \sim \mathcal{N}(\boldsymbol{\mu}_0, \sigma_4^2 \boldsymbol{I}), \tag{35}$$

where σ_4 is a scalar constant, chosen such that the voxel displacements $\|\boldsymbol{T}(\boldsymbol{x}_j, \boldsymbol{\mu}_n) - \boldsymbol{T}(\boldsymbol{x}_j, \boldsymbol{\mu}_0)\|$ caused by the parameter change from $\boldsymbol{\mu}_0$ to $\boldsymbol{\mu}_n$ remain with high probability (≈ 0.95) below a user-defined value δ. The exact procedure is explained at the end of Sect. 2.4. The user-defined constant N, introduced in (34), should be chosen high enough such that $\frac{1}{N}\sum_n \|\boldsymbol{g}(\boldsymbol{\mu}_n)\|^2$ is a good estimate of the true expectation $\mathrm{E}\|\boldsymbol{g}\|^2$. When $\boldsymbol{C}$ equals the identity matrix, the average squared gradient magnitude $\frac{1}{N}\sum_n \|\boldsymbol{g}(\boldsymbol{\mu}_n)\|^2$ has a χ^2_{NP} distribution. The ratio between the standard deviation and the expectation of a χ^2_{NP} distribution equals $\sqrt{2}/\sqrt{NP}$. We can thus expect that, with increasing P, N can be lowered. For arbitrary $\boldsymbol{C}$, the ratio between standard deviation and expectation can be shown to have an upper bound of $\sqrt{2/N}$. From this, it is clear that a value $N \approx 10$ is a reasonable choice.

Having estimated the distribution of $\boldsymbol{g}$, we can calculate the distribution of voxel displacements per iteration of a deterministic gradient descent process. The deterministic gradient descent procedure mentioned in Assumption B3.1, (11), is considered: $\boldsymbol{\mu}_{k+1} = \boldsymbol{\mu}_k - \hat{\gamma}\boldsymbol{g}(\boldsymbol{\mu}_k)$. The displacement $\boldsymbol{d}_k$ of voxel $\boldsymbol{x}_j$ between iteration k and $k+1$ is defined by:

$$\boldsymbol{d}_k(\boldsymbol{x}_j) \equiv \boldsymbol{T}(\boldsymbol{x}_j, \boldsymbol{\mu}_{k+1}) - \boldsymbol{T}(\boldsymbol{x}_j, \boldsymbol{\mu}_k). \tag{36}$$

Our goal is to estimate the distribution of $\boldsymbol{d}_k(\boldsymbol{x}_j)$ for some $\boldsymbol{\mu}_k \in \Upsilon$. This result is used in Sect. 2.4 to estimate λ (see Assumption B3.1), which is used to select a such that Assumptions B3 are satisfied. According to the Taylor expansion of $\boldsymbol{T}$ around $\boldsymbol{\mu}_k$:

$$\boldsymbol{d}_k(\boldsymbol{x}_j) \approx \frac{\partial \boldsymbol{T}}{\partial \boldsymbol{\mu}}(\boldsymbol{x}_j, \boldsymbol{\mu}_k) \cdot (\boldsymbol{\mu}_{k+1} - \boldsymbol{\mu}_k) = \boldsymbol{J}_j(\boldsymbol{\mu}_{k+1} - \boldsymbol{\mu}_k), \tag{37}$$

where the last equality follows from Assumption A1. Substitution of (11) gives:

$$\boldsymbol{d}_k(\boldsymbol{x}_j) \approx -\hat{\gamma}\boldsymbol{J}_j\boldsymbol{g}(\boldsymbol{\mu}_k). \tag{38}$$

Using (31) we obtain:

$$\boldsymbol{d}_k(\boldsymbol{x}_j) \sim \mathcal{N}\left(\boldsymbol{0}, \hat{\gamma}^2\sigma_1^2\boldsymbol{J}_j\boldsymbol{C}\boldsymbol{J}_j^T\right). \tag{39}$$

Note that the estimated distribution of $\boldsymbol{d}_k(\boldsymbol{x}_j)$ is independent of k.

The distribution estimates that have been derived in this subsection are used in the following subsections. In Sect. 2.4, (31), (33), and (39) are used to select a. Equation (33) is used also in Sect. 2.5 to select ω.

2.4 Selection of a

In this subsection an appropriate value of a is estimated, using Assumptions B3 and (31), (33), and (39). The value of A is considered a user-defined constant. The method consists of two steps. First, a deterministic gradient descent method is considered. The maximum value of a that still ensures convergence is estimated, based on Assumption B3.1, (39), and an additional user input: the maximum allowed voxel displacement δ. After that, Assumption B3.2 is combined with (31) and (33), to derive an expression for a that takes the stochastic approximation errors into account.

As mentioned in Sect. 2.2, Assumption B3.1, the maximum value for $\gamma(0) = a/A$ that ensures convergence of the deterministic gradient process (11) equals $2/\lambda$, provided that conditions B3.1(a)–(c) hold. Condition B3.1(a) is often not satisfied in image registration problems. This is a general problem of image registration, which will not be further addressed in this article. Henceforth, we simply assume that $\boldsymbol{\mu}_0$ is chosen within the capture range of the desired local minimum $\hat{\boldsymbol{\mu}}$. The value of λ, which is defined by condition B3.1(c), is generally unknown. We propose to estimate λ based on an additional user input parameter δ: the maximum allowed magnitude of the voxel displacements $\boldsymbol{d}_k(\boldsymbol{x}_j)$. The problem becomes thus to compute a λ such that

$$\|\boldsymbol{d}_k(\boldsymbol{x}_j)\| < \delta, \quad \forall k, j, \tag{40}$$

when $\hat{\gamma} < 2/\lambda$. According to (39) the voxel displacement $\boldsymbol{d}_k(\boldsymbol{x}_j)$ has a normal distribution, independent of k, with variance depending on $\hat{\gamma}$. The criterion given in (40) must therefore be weakened to, for example:

$$\Pr\left(\|\boldsymbol{d}_k(\boldsymbol{x}_j)\| > \delta\right) < \rho, \quad \forall j, \tag{41}$$

with ρ some small value, say 0.05. We approximate (41) by:

$$\mathrm{E}\|\boldsymbol{d}_k(\boldsymbol{x}_j)\|^2 + 2\sqrt{\mathrm{Var}\|\boldsymbol{d}_k(\boldsymbol{x}_j)\|^2} < \delta^2, \quad \forall j. \tag{42}$$

This approximation is justified by the Vysochanskij-Petunin inequality (Vysochanskij and Petunin 1980). For the expectation and variance the following expressions can be derived using (39):

$$\mathrm{E}\|\boldsymbol{d}_k(\boldsymbol{x}_j)\|^2 = \hat{\gamma}^2\sigma_1^2\mathrm{tr}\left(\boldsymbol{J}_j\boldsymbol{C}\boldsymbol{J}_j^T\right), \tag{43}$$

$$\mathrm{Var}\|\boldsymbol{d}_k(\boldsymbol{x}_j)\|^2 = 2\hat{\gamma}^4\sigma_1^4\|\boldsymbol{J}_j\boldsymbol{C}\boldsymbol{J}_j^T\|_F^2, \tag{44}$$

with $\|\cdot\|_F$ denoting the Frobenius norm. Substitution in (42) gives:

$$\hat{\gamma}^2 < \min_{\boldsymbol{x}_j\in\Omega'_F} \frac{\delta^2/\sigma_1^2}{\mathrm{tr}(\boldsymbol{J}_j\boldsymbol{C}\boldsymbol{J}_j^T) + 2\sqrt{2}\|\boldsymbol{J}_j\boldsymbol{C}\boldsymbol{J}_j^T\|_F}. \tag{45}$$

Setting the right-hand side equal to $(2/\lambda)^2$ results in the desired estimate of λ. The maximum value of a for a deterministic gradient descent method can then be computed using: $\gamma(0) = a/A = 2/\lambda$. We denote this maximum by a_{MAX}:

$$a_{\mathrm{MAX}} \equiv \frac{2A}{\lambda} \tag{46}$$

$$= \frac{A\delta}{\sigma_1} \min_{\boldsymbol{x}_j\in\Omega'_F} \left[\mathrm{tr}\left(\boldsymbol{J}_j\boldsymbol{C}\boldsymbol{J}_j^T\right) + 2\sqrt{2}\|\boldsymbol{J}_j\boldsymbol{C}\boldsymbol{J}_j^T\|_F\right]^{-\frac{1}{2}}. \tag{47}$$

The second assumption that imposes a constraint on a is Assumption B3.2. Using $\gamma(0) = a/A$, $\mathrm{tr}(\boldsymbol{\Sigma}) = \mathrm{tr}(\sigma_3^2\boldsymbol{C}) = \mathrm{E}\|\boldsymbol{\varepsilon}_k\|^2$, and the definition of a_{MAX} in (46), we rewrite (12) as:

$$a \leq \frac{2A}{\lambda}\frac{\|\boldsymbol{g}(\boldsymbol{\mu})\|^2 - \beta_0}{\|\boldsymbol{g}(\boldsymbol{\mu})\|^2 + \mathrm{E}\|\boldsymbol{\varepsilon}_k\|^2}$$

$$= a_{\mathrm{MAX}}\frac{\|\boldsymbol{g}(\boldsymbol{\mu})\|^2 - \beta_0}{\|\boldsymbol{g}(\boldsymbol{\mu})\|^2 + \mathrm{E}\|\boldsymbol{\varepsilon}_k\|^2}. \tag{48}$$

When the expected approximation error $\mathrm{E}\|\boldsymbol{\varepsilon}_k\|^2$ goes to zero, and $\beta_0 \downarrow 0$, this condition equals $a < a_{\mathrm{MAX}}$. The condition corresponds to the intuition that a lower gain should be used when the approximation error increases. Exact verification of (48) for all $\|\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}\| \geq R$, as Assumption B3.2 demands, seems not feasible. We therefore propose to use the following estimate of a:

$$a = a_{\mathrm{MAX}}\frac{\mathrm{E}\|\boldsymbol{g}\|^2}{\mathrm{E}\|\boldsymbol{g}\|^2 + \mathrm{E}\|\boldsymbol{\varepsilon}_k\|^2} \equiv a_{\mathrm{MAX}}\eta, \tag{49}$$

with $0 < \eta \leq 1$. For $\mathrm{E}\|\boldsymbol{g}\|^2$ and $\mathrm{E}\|\boldsymbol{\varepsilon}_k\|^2$ their empirical estimates can be used directly, see the left-hand side of (34).

Summarising, we have replaced the original parameter a by a new user-defined parameter, δ. Unlike a, the new parameter δ has a unit (mm), and an intuitive meaning. In Sect. 3, the sensitivity of ASGD to the values of δ and A is experimentally investigated.

As announced in Sect. 2.3, we also use δ to select the value of σ_4, which occurs in (35). The voxel displacement caused by the parameter change from $\boldsymbol{\mu}_0$ to $\boldsymbol{\mu}_n$ is considered:

$$\boldsymbol{d}_{0,n}(\boldsymbol{x}_j) \equiv \boldsymbol{T}(\boldsymbol{x}_j, \boldsymbol{\mu}_n) - \boldsymbol{T}(\boldsymbol{x}_j, \boldsymbol{\mu}_0). \tag{50}$$

Following a similar approach as in Sect. 2.3, the distribution of $\boldsymbol{d}_{0,n}(\boldsymbol{x}_j)$ can be estimated, given the distribution of $\boldsymbol{\mu}_n$, which was defined in (35). The result is:

$$\boldsymbol{d}_{0,n}(\boldsymbol{x}_j) \sim \mathcal{N}\left(\boldsymbol{0}, \sigma_4^2\boldsymbol{J}_j\boldsymbol{J}_j^T\right). \tag{51}$$

With similar reasoning as earlier in this section, we select σ_4 such that:

$$\Pr\left(\|\boldsymbol{d}_{0,n}(\boldsymbol{x}_j)\| > \delta\right) < \rho, \quad \forall j, \tag{52}$$

with ρ some small value, say 0.05. This condition is approximated by:

$$\mathrm{E}\|\boldsymbol{d}_{0,n}(\boldsymbol{x}_j)\|^2 + 2\sqrt{\mathrm{Var}\|\boldsymbol{d}_{0,n}(\boldsymbol{x}_j)\|^2} < \delta^2, \quad \forall j, \tag{53}$$

which, using (51), gives the following solution for σ_4^2:

$$\sigma_4^2 = \min_{\boldsymbol{x}_j\in\Omega'_F} \frac{\delta^2}{\|\boldsymbol{J}_j\|_F^2 + 2\sqrt{2}\|\boldsymbol{J}_j\boldsymbol{J}_j^T\|_F}. \tag{54}$$

2.5 Selection of Sigmoid Parameters

The selection of the sigmoid function parameters f_{MAX} and f_{MIN} is based on the condition for asymptotic normality: Assumption B5.2. This assumption constrains the value of E_0, which is directly related to f_{MAX} and f_{MIN}, according to Assumption B4.2. The third parameter ω, which defines the scale of the sigmoid, is chosen as a small fraction of the standard deviation of $\boldsymbol{\varepsilon}_k^T\boldsymbol{\varepsilon}_{k-1}$, such that $E_0 \approx \frac{1}{2}(f_{\mathrm{MAX}} + f_{\mathrm{MIN}})$.

Assumption B5.2 states that matrix $\boldsymbol{W} = \frac{1}{2}\boldsymbol{I} - \frac{a}{E_0}\boldsymbol{H}(\hat{\boldsymbol{\mu}})$ is assumed to have negative eigenvalues only. Let $\lambda^* > 0$ denote the minimum positive eigenvalue of $\boldsymbol{H}(\hat{\boldsymbol{\mu}})$ (Assumption B5.2 can never be satisfied with negative eigenvalues of $\boldsymbol{H}(\hat{\boldsymbol{\mu}})$). The condition then becomes:

$$E_0 < 2a\lambda^*. \tag{55}$$

Combining (55), (49), and (46) gives:

$$E_0 < 4A\frac{\lambda^*}{\lambda}\eta. \tag{56}$$

So, given fixed A and cost function properties λ and λ^*, the maximum allowed value of E_0 is directly proportional to the ratio η. Following Assumption B4.2 and assuming that ω is small, we have $E_0 \approx \frac{1}{2}(f_{\mathrm{MAX}} + f_{\mathrm{MIN}})$. Substitution in (56) gives:

$$f_{\mathrm{MIN}} < 8A\frac{\lambda^*}{\lambda}\eta - f_{\mathrm{MAX}}. \tag{57}$$

A reasonable choice for the maximum of the sigmoid function is $f_{\mathrm{MAX}} = 1$. This implies that the forward time step

Table 1 Overview of data sets and experiments

Anatomy	Brain	Prostate	Right lung
Modality	CT and 1.5T MR T1	3T MR SSFP	CT
Dimensions	CT: $512 \times 512 \times 50$	$200 \times 200 \times 70$	$120 \times 160 \times 200$
	MR: $256 \times 256 \times 50$		
Voxel size [mm]	CT: $0.45 \times 0.45 \times 3$	$0.5 \times 0.5 \times 1$	$2 \times 2 \times 2$
	MR: $0.85 \times 0.85 \times 3$		
Nr. of patients	9	6 (2 scans/person)	5 (2 scans/person)
Registration	CT with MR	Day 1 with day 2	Day 1 with day 2
Similarity measure	MI	MI	MSD, NC, MI, NMI
Transformation	Rigid	B-spline	Affine, B-spline
Nr. of parameters P	6	2000	12, 4000
B-spline control point			
Grid spacing [mm]	–	$16 \times 16 \times 16$	$40 \times 40 \times 40$
Evaluation measure	MSE	DSC	DSC
Section	3.2	3.2	3.3 and 3.4

$t_{k+1} - t_k$ equals at most the time step made by the RM method. Demanding $-f_{\text{MAX}} < f_{\text{MIN}} < 0$, we propose:

$$f_{\text{MIN}} = \eta - f_{\text{MAX}} = \eta - 1, \tag{58}$$

where we assumed that A can be chosen such that $8A\lambda^*/\lambda > 1$, in order to satisfy (57). If A is chosen too low, the consequence is that asymptotic normality can not be guaranteed anymore. By choosing A very high this risk is avoided, but one has to keep in mind that the property of asymptotic normality is not always relevant in practice. In practical applications, the number of iterations K, see (5), is finite due to limited available computation time. Choosing $A \to \infty$ would result in a nearly constant gain sequence $\gamma(t_k)$ for all iterations $0 \leq k \leq K$. The adaptive behaviour of ASGD would, consequently, be eliminated completely. In Sect. 3, the sensitivity of ASGD to the value of A is experimentally investigated.

For the selection of f_{MIN} and f_{MAX} we assumed that $E_0 \equiv \text{E}f(\boldsymbol{\varepsilon}_k^T \boldsymbol{\varepsilon}_{k-1}) \approx \frac{1}{2}(f_{\text{MAX}} + f_{\text{MIN}})$. The approximation only holds if ω is much smaller than $|\boldsymbol{\varepsilon}_k^T \boldsymbol{\varepsilon}_{k-1}|$, with high probability. According to (33), $\boldsymbol{\varepsilon}_k$ and $\boldsymbol{\varepsilon}_{k-1}$ are independent normally distributed variables with mean $\mathbf{0}$ and variance $\sigma_3^2 \boldsymbol{C}$. The expected value of the inner product $\boldsymbol{\varepsilon}_k^T \boldsymbol{\varepsilon}_{k-1}$ is zero. We propose to choose ω as a small fraction ζ of the standard deviation of $\boldsymbol{\varepsilon}_k^T \boldsymbol{\varepsilon}_{k-1}$:

$$\omega = \zeta \sqrt{\text{Var}\left(\boldsymbol{\varepsilon}_k^T \boldsymbol{\varepsilon}_{k-1}\right)}, \tag{59}$$

with $\zeta \approx \frac{1}{10}$ for example. For the variance it can be shown that:

$$\text{Var}\left(\boldsymbol{\varepsilon}_k^T \boldsymbol{\varepsilon}_{k-1}\right) = \sigma_3^4 \|\boldsymbol{C}\|_F^2. \tag{60}$$

An alternative strategy might be to actually set $\omega \downarrow 0$ (as small as machine precision allows), but this would start to interfere with Assumption B5.3.

2.6 Overview of the Algorithm

The following steps describe the entire algorithm:

1. Compute $\boldsymbol{C}$ using (25).
2. Compute σ_4 using (54).
3. Generate N instances of $\boldsymbol{\mu}_n$ according to (35). Compute for each $\boldsymbol{\mu}_n$ the exact cost function derivative $\boldsymbol{g}$, the approximated derivative $\tilde{\boldsymbol{g}}_k$, and the approximation error $\boldsymbol{\varepsilon}_k = \boldsymbol{g} - \tilde{\boldsymbol{g}}_k$. Note that, to compute $\tilde{\boldsymbol{g}}_k$, a new set of voxels S_k must be selected for each $\boldsymbol{\mu}_n$.
4. Compute σ_1 using (34). Compute σ_3 similarly.
5. Compute a_{MAX} using (47).
6. Compute η and a using (49).
7. Set $f_{\text{MAX}} = 1$ and compute f_{MIN} using (58).
8. Compute ω using (59) and (60).
9. Start the optimisation defined by (5), (6), (9), and (10). Convergence is assumed after K iterations: $\hat{\boldsymbol{\mu}} = \boldsymbol{\mu}_K$.

Steps 1–8 serve to estimate a, f_{MAX}, f_{MIN}, and ω. Note that this has to be done only once, before starting the actual optimisation routine in step 9. The required user settings are t_0, t_1, K, δ, A, N, and ζ. The initial conditions t_0 and t_1 will probably have a minor influence on the performance as long as they are chosen much smaller than the number of iterations: $t_0, t_1 \ll K$. The meaning of δ is explained in Sect. 2.4 and the influence of A is discussed in Sect. 2.5. For N, a value ≈ 10 is suggested in Sect. 2.3. For ζ, a value $\approx \frac{1}{10}$ is recommended in Sect. 2.5.

3 Experiments and Results

3.1 Experiment Setup

The ASGD method has been evaluated on three medical image registration problems. Table 1 gives an overview of the data sets that were used, the type of registration experiments, and the evaluation measures for quantifying registration accuracy. The bottom row indicates in which subsection the experiments are described.

The ASGD method was implemented as a part of the `elastix` package (elastix.isi.uu.nl). Rigid, affine, and non-rigid B-spline transformation models were tested. A three-level multiresolution framework was used in all experiments. The images were smoothed with a Gaussian filter with a standard deviation of 2, 1, and 0.5 (voxel units), in each level respectively. For the B-spline transform, the B-spline control point grid spacing was halved in each resolution level, such that in the final resolution the grid spacing reported in Table 1 was reached. Four similarity measures were used: mean squared intensity difference (MSD), normalised correlation (NC), mutual information (MI), and normalised mutual information (NMI). Both MI and NMI were implemented using cubic B-spline Parzen windows, as in Thévenaz and Unser (2000), with a 32×32 joint histogram. For the rigid registrations, the transformation was parameterised using the translation vector $\boldsymbol{t} = (t_1, t_2, t_3)^T$ and the Euler angles $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3)^T$. Since the Euler angles can have an entirely different range than the translations, we used the following reparameterisation:

$$\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{I} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{S} \end{bmatrix} \begin{bmatrix} \boldsymbol{t} \\ \boldsymbol{\theta} \end{bmatrix}, \tag{61}$$

with $\boldsymbol{S}$ a diagonal scaling matrix, with on the diagonal:

$$s_{ii} = \left(\int_{\Omega_F} \left\| \frac{\partial \boldsymbol{T}}{\partial \theta_i}(\boldsymbol{x}, \boldsymbol{\mu}_0) \right\|^2 d\boldsymbol{x} \Big/ \int_{\Omega_F} d\boldsymbol{x} \right)^{-\frac{1}{2}}. \tag{62}$$

The rotation parameters are thus scaled by the average voxel displacement caused by a small perturbation of the rotation angle. In case of an affine transformation we used the same strategy for the matrix elements. In case of a B-spline transformation the control point coefficients directly formed the parameters $\boldsymbol{\mu}$. Note that the rigid transformation with Euler angles does not satisfy Assumption A1. However, it is included in the experiments in order to demonstrate that ASGD still works when the rotations are reasonably small.

For the brain images, the ground truth CT-MR registrations were available. The scans were acquired using a stereotactic frame, which was later erased from the images by post-processing, in the context of the "Retrospective Image Registration Evaluation" project (West et al. 1997). In our experiments we quantified the registration accuracy by computing the mean square error of the transformation at the eight corner points of the image:

$$\text{MSE} \equiv \frac{1}{8} \sum_{c=1}^{8} \| \boldsymbol{T}(\boldsymbol{x}_c, \hat{\boldsymbol{\mu}}) - \boldsymbol{T}(\boldsymbol{x}_c, \hat{\boldsymbol{\mu}}^G) \|, \tag{63}$$

with $\hat{\boldsymbol{\mu}}^G$ the ground truth.

For the MR prostate scans, expert manual segmentations of the prostate were available. The Dice similarity coefficient (DSC) (Dice 1945) of the segmentation S_F of the fixed image and the segmentation S_M of the deformed moving image was used for evaluation:

$$\text{DSC} \equiv \frac{2|S_F \cap S_M|}{|S_F| + |S_M|}. \tag{64}$$

The DSC measures overlap of the two segmentations and thus gives an indication of the registration quality. A value of 1 means perfect registration. A value of 0 means that the segmentations have no overlap at all.

For the CT lung images, we used the DSC of the lung airways as an evaluation measure. The lung airways were segmented using an automatic region-growing algorithm, described in (Hu et al. 2001; Sluimer et al. 2005).

In all experiments we used the initial conditions $t_0 = t_1 = 0$. As suggested in Sect. 2, we used $N = 10$ and $\zeta = \frac{1}{10}$. The number of voxels used to compute $\tilde{\boldsymbol{g}}_k$, denoted by $|S_k|$ in (26), was set to 2000, as recommended in Klein et al. (2007). For the remaining free parameters, δ, A, and K, the settings are reported in the following subsections. In all experiments, the extra computation time required by ASGD to perform steps 1–8, see Sect. 2.6, was comparable to the time spent in step 9.

In Sect. 3.2, the ASGD method is compared with the standard RM method. The brain and prostate data are used for this purpose. In Sect. 3.3, the lung images are used to test ASGD with different similarity metrics. In Sect. 3.4, the relation between δ and the maximum voxel displacement is verified.

3.2 Adaptive vs. Non-Adaptive

In this subsection, we test the effect of the step size adaptation. The ASGD method is compared to the standard RM method in a series of experiments on the brain and prostate data, for a range of values of δ, A, and K.

The RM method, see (2), requires definition of the step size sequence $\{\gamma_k\}$. For fair comparison with ASGD, we use the following function:

$$\gamma_k = a/(E_0 k + A), \tag{65}$$

with a, A, and E_0 as computed for the ASGD method. With this choice γ_0 equals $\gamma(t_0)$, so RM and ASGD start with the

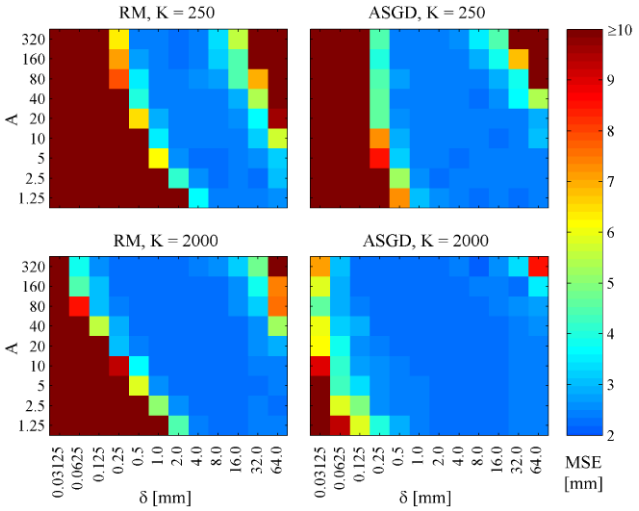

Fig. 2 RM vs. ASGD for rigid registration of brain scans. A low MSE indicates better registration

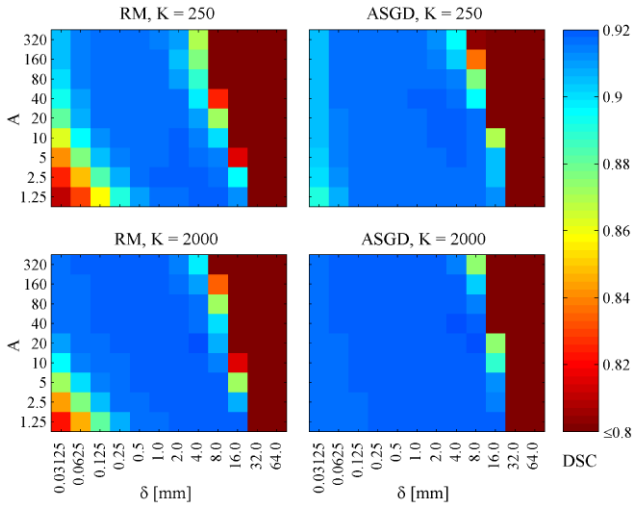

Fig. 3 RM vs. ASGD for nonrigid registration of prostate scans. A high DSC indicates better registration

same step size. Also, it can be shown (Cruz 2005a) that γ_k and $\gamma(t_k)$ converge to the same value as $k \to \infty$. For this it is necessary to see that, with ASGD, E_0 equals the expected value of the time increment $t_{k+1} - t_k$ when $\boldsymbol{g}(\boldsymbol{\mu}_k) \approx \mathbf{0}$.

The registration experiments were performed for all possible combinations of $\delta \in \{0.03125, 0.0625, \ldots, 64\}$ (in mm), $A \in \{1.25, 2.5, \ldots, 320\}$, and $K \in \{250, 2000\}$. The brain data were registered using a rigid transformation model. For each (δ, A, K) combination the mean MSE over the 9 CT-MR registrations was calculated. The prostate scans were registered using a nonrigid B-spline transformation. After registration, the mean DSC over the 6 image pairs was computed. The measured computation time per registration on an AMD Opteron 2600 MHz was approximately 5 min.

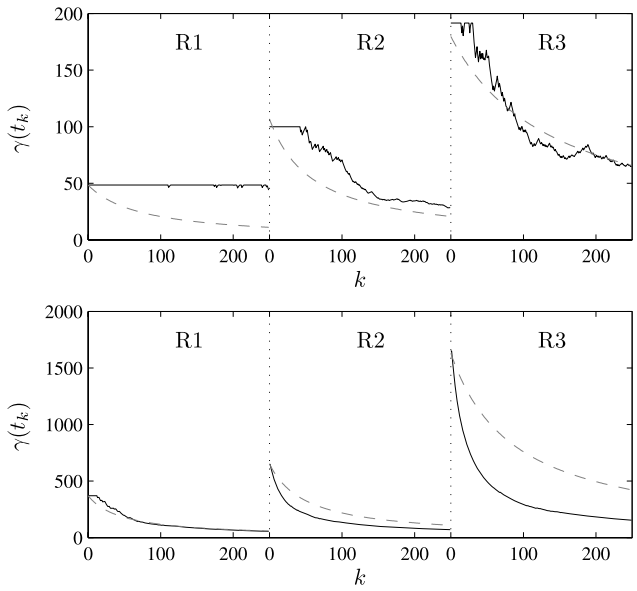

Fig. 4 Example of step size adaptation by ASGD. The *solid black line* is for ASGD; the *dashed grey line* for RM. The *upper graph* shows the result for $\delta = 0.25$ mm. The *lower graph* was created using $\delta = 2$ mm. Note that the *vertical axes* have different scales

In Figs. 2 and 3 the results are visualised on a colour scale. Each pixel represents the mean MSE or DSC for a combination of δ and A. The adaptive step size mechanism clearly improved the robustness with respect to the user-defined parameters A and δ. Increasing the number of iterations from $K = 250$ to $K = 2000$ improved the robustness of both RM and ASGD. However, Fig. 3 shows that the ASGD method with $K = 250$ gave better results still than RM with $K = 2000$.

As an illustration of the step size adaptation by ASGD we plotted the values of $\gamma(t_k)$ during registration of one of the prostate image pairs. Figure 4 shows the result for $\delta = 0.25$ mm (upper graph) and $\delta = 2$ mm (lower graph), both with $A = 20$ and $K = 250$. The labels R1–3 represent the three resolution levels. The solid black line is for ASGD. The dashed grey line shows the predefined step size function that was used for RM, as given by (65). The adaptive step size mechanism of ASGD is clearly observed in each resolution: when the algorithm starts with a small step size ($\delta = 0.25$ mm), the step size decays less fast than with a large initial step size ($\delta = 2$ mm). For example, in resolution R2, with $\delta = 0.25$, the ASGD step size remains nearly constant in the first 50 iterations, whereas with $\delta = 2.0$ the step size immediately starts decaying at $k = 0$.

3.3 ASGD with different similarity measures

In this subsection, we investigate the influence of the similarity measure on the choice of δ. Registration experiments were performed on the CT lung data using different similarity measures, for a range of δ values.

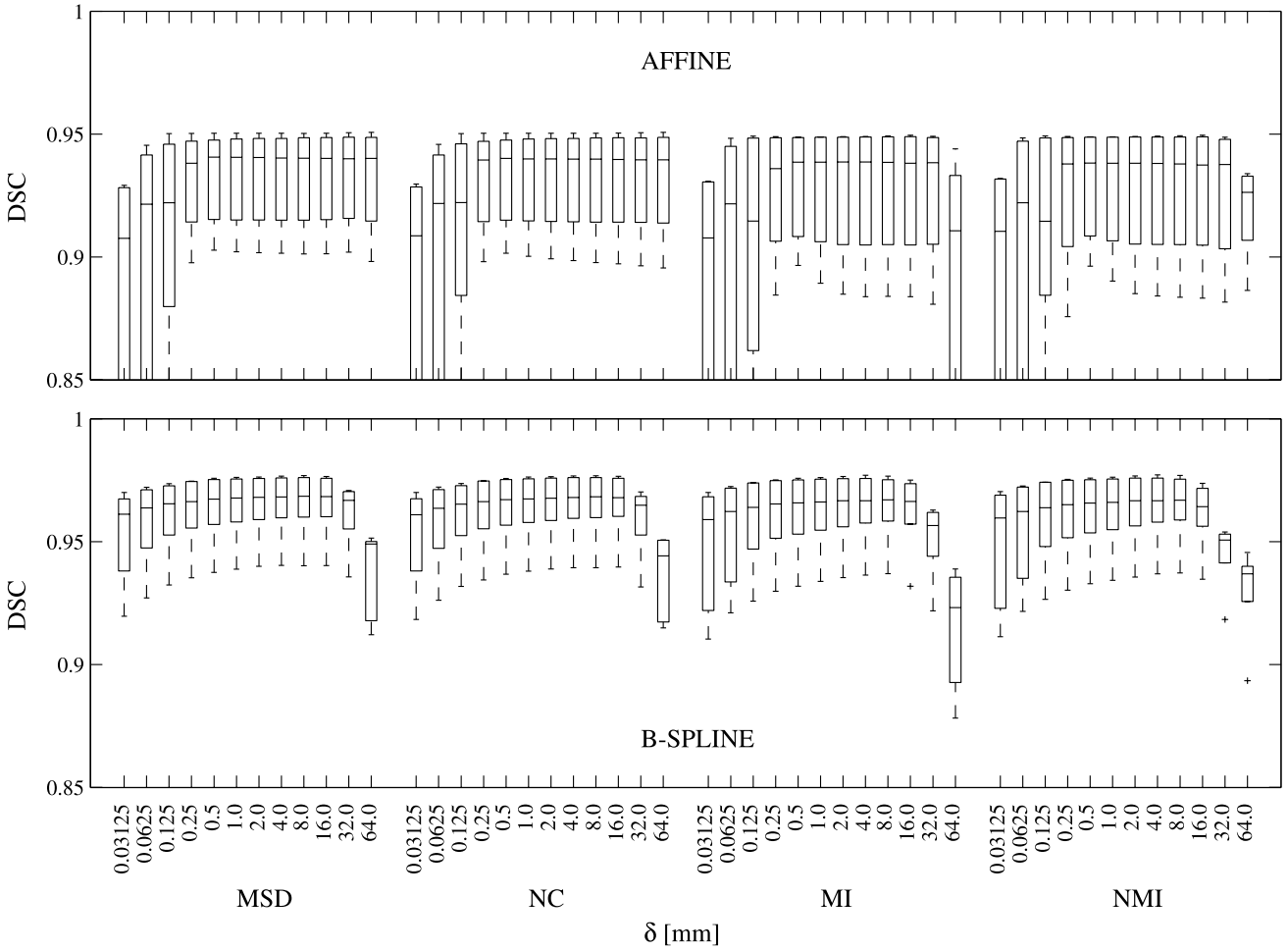

Fig. 5 ASGD with different similarity measures for registration of CT lung scans. The *upper plot* shows the results for an affine transformation. The *lower plot* shows the results for a B-spline transformation

Four similarity measures were tested: MSD, NC, MI, and NMI. For δ the range $\{0.03125, 0.0625, \ldots, 64\}$ (in mm) was used. The entire experiment was done using both an affine and a B-spline transformation. All registrations were done with $A = 20$, which gave good performance in the previous section. A relatively low number of iterations was used, $K = 250$, such that the effect of varying δ becomes more apparent.

The results are summarised in Fig. 5. Each boxplot summarises the distribution of DSC values after registration of the five image pairs. The upper graph shows the results using the affine transformation. The lower graph shows the results obtained with the B-spline transformation. Both for the affine and the B-spline registrations, the optimal value of δ was independent of the choice of the similarity measure. For affine registration, the range $0.5 \leq \delta \leq 32$ mm gave the best results. With the B-spline transformation, the range $1 \leq \delta \leq 16$ mm gave the best results. For $\delta = 1$ mm, the calculated values of a in the finest resolution level are reported in Table 2, averaged over the five image pairs. The large differences between the values show that choosing a manually would not have been a trivial task.

Table 2 Average values of a in the finest resolution level of the lung image registrations, using $\delta = 1$ mm

	MSD	NC	MI	NMI
Affine	0.0017	620	240	780
B-spline	0.73	270000	43000	140000

3.4 Maximum voxel displacement

In Sect. 2.4, δ was introduced as a user setting with an intuitive meaning, being the maximum voxel displacement per iteration of the deterministic gradient descent process $\boldsymbol{\mu}_{k+1} = \boldsymbol{\mu}_k - \hat{\gamma}\, \boldsymbol{g}(\boldsymbol{\mu}_k)$, with constant step size $\hat{\gamma} = a_{\mathrm{MAX}}/A$.

Table 3 The 95% quantiles of the ratio $\|\boldsymbol{d}_k(\boldsymbol{x}_j)\|/\delta$. A value close to 1 is desirable. Each entry in the table is based on 5 image pairs, $K = 100$ iterations, and all voxels $\boldsymbol{x}_j \in \Omega'_F$

Transform	Resolution	δ [mm]											
		0.03125	0.0625	0.125	0.25	0.5	1.0	2.0	4.0	8.0	16.0	32.0	64.0
Affine	1	0.8	0.8	0.8	0.7	0.7	0.6	0.6	1.7	1.8	1.4	1.3	1.0
Affine	2	0.7	0.7	0.7	0.6	0.7	0.9	1.0	0.6	0.8	0.9	1.3	1.1
Affine	3	0.7	0.7	0.7	0.7	0.6	1.2	0.4	0.8	0.8	0.8	1.1	1.2
B-spline	1	0.5	0.4	0.4	0.4	0.3	1.8	2.3	1.9	1.6	1.4	1.4	1.2
B-spline	2	0.7	0.6	0.5	0.4	1.7	4.2	2.3	1.6	1.6	1.6	1.4	–
B-spline	3	0.9	0.8	0.6	0.4	4.3	2.6	2.2	1.7	1.6	1.4	1.4	–

The estimate of a_{MAX} relies on some simplifying assumptions and approximations, most notably Assumption A2. The following experiment serves to verify whether the voxel displacements indeed remain below δ.

The CT lung registrations were repeated with the deterministic gradient descent scheme mentioned above, using $K = 100$, and MI as a similarity measure. The voxel displacements $\|\boldsymbol{d}_k(\boldsymbol{x}_j)\|$ were computed for all $\boldsymbol{x}_j \in \Omega'_F$, in each iteration k. Table 3 reports the 95% quantiles of the ratio $\|\boldsymbol{d}_k(\boldsymbol{x}_j)\|/\delta$ for each resolution level separately. Each entry in the table is based on 5 image pairs. Entries with '–' indicate that for at least one of the image pairs the registration failed completely, i.e. the overlap between the fixed and moving image became too small to continue registration (due to very large step sizes). The table shows that with the affine transformation the ratio was close to 1, meaning that most voxel displacements indeed remained below δ. With the B-spline transformation, for $\delta \geq 0.5$ the actual displacements exceeded δ with a factor 2 on average. For $\delta < 0.5$, the actual displacements remained below δ.

4 Discussion

The experiments show that ASGD works for a rather broad range of δ and A. The results in Sect. 3.2 indicate that $A = 20$ works well in general, both for rigid and nonrigid registration. With that setting, for the applications we considered, the optimum value of δ was approximately equal to the size of a voxel. Of course, that relation is not always exactly satisfied, since simply upsampling the images will not lower the optimum value of δ. However, the experiments in Sects. 3.2 and 3.3 show that the registration results are relatively insensitive to the value of δ, as long as $\frac{1}{4}V \leq \delta \leq 4V$, with V the (average) voxel size. For rigid and affine registration somewhat higher δ values tend to work better than for nonrigid, which corresponds to intuition.

The results in Sect. 3.4, Table 3, show that the actually realised voxel displacements were not in all cases lower than δ. This is due to the simplifying assumptions used to estimate a_{MAX} given δ. Especially Assumption A2 may not be satisfied. While the estimate of a_{MAX} still appears to work quite well in practice, further improvements may be obtained by improving the estimate of the distribution of $\boldsymbol{g}$. In our approach, the estimated distribution of $\boldsymbol{g}$ is for a large part based on the model for the covariance matrix, given by (25). This allows us to use a low number N of gradient evaluations, since only one parameter (σ_1) has to be determined, see (31) and below. Another approach would be to use the common maximum likelihood estimate of the covariance matrix: $\boldsymbol{C} = \frac{1}{N}\sum_n \boldsymbol{g}(\boldsymbol{\mu}_n)\boldsymbol{g}(\boldsymbol{\mu}_n)^T$. However, this would require a larger N. An interesting technique that combines the two approaches is the *shrinkage* method described in Schäfer and Strimmer (2005). In that article, a linear combination of the model based estimate of $\boldsymbol{C}$ and the maximum likelihood estimate is employed, with the weighting determined by explicitly minimising the expectation of a squared error loss function.

In all experiments described in this article, the initial conditions t_0 and t_1 were simply set to 0. It might be beneficial to try larger values, such as $t_0 = t_1 = A$. In this way, the method could become more robust to large values of δ.

In our study, we have only considered parametric transformation models. It would be interesting to integrate the ASGD method also in a nonparametric registration framework (Modersitzki 2004). Note that this would require incorporation of a regularisation term in (16).

5 Conclusion

An optimisation method with adaptive step size prediction for image registration has been presented: adaptive stochastic gradient descent (ASGD). The method is designed to work with stochastic approximations of the cost

function derivatives, and, thus, requires little computation time per iteration. In comparison with a standard Robbins-Monro (RM) stochastic gradient descent scheme, the ASGD method is more robust, because of its adaptive step size prediction. The main contribution of this article is the selection mechanism for the method's free parameters. The selection mechanism takes into account the choice of similarity measure, the transformation model, and the image content, in order to estimate proper values for the most important settings. The influences of the remaining free parameters δ, A, and K were experimentally investigated. The experiments showed that ASGD works for a broad range of δ and A. The optimum value of δ appeared to be unaffected by the choice of the similarity measure. In general, a reasonable setting is to use $A = 20$ and δ equal to the average voxel size of the images. Increasing the number of iterations from $K = 250$ to $K = 2000$ improved the robustness of both RM and ASGD, with respect to the choice of δ and A. However, the ASGD method with $K = 250$ gave already better results than RM with $K = 2000$.

In Klein et al. (2007), it was shown for a number of medical image registration applications that RM outperforms several well-known deterministic optimisation methods, such as quasi-Newton and nonlinear conjugate gradient. It was pointed out that the main disadvantage of RM is the need for a predetermined step size function. The ASGD method presented in this article provides a solution for that issue.

Acknowledgements Funding for this research has been provided by the Netherlands Organisation for Scientific Research (NWO). We thank the authors of the "Matrix Cookbook" (Petersen and Pedersen 2007), which has been a valuable resource. The work also benefited from the use of the Insight Segmentation and Registration Toolkit (ITK), an open source software developed as an initiative of the U.S. National Library of Medicine and available at www.itk.org.

The brain images and their ground truth transformations originated from the "Retrospective Image Registration Evaluation" project, National Institutes of Health, Project Number 8R01EB002124-03, Principal Investigator, J. Michael Fitzpatrick, Vanderbilt University, Nashville, TN. The prostate and lung images were acquired at the radiotherapy and radiology departments, respectively, of the University Medical Center Utrecht, The Netherlands. The authors kindly acknowledge Ellen Kerkhof for providing the manual prostate segmentations.

References

Collignon, A., Maes, F., Delaere, D., Vandermeulen, D., Suetens, P., & Marchal, G. (1995). Automated multi-modality image registration based on information theory. In Bizais, Y., Barillot, C., Di Paola, R. (Eds.), *Information Processing in Medical Imaging* (pp. 263–274). Dordrecht: Kluwer Academic.

Cruz, P. (2005a). *Almost sure convergence and asymptotical normality of a generalization of Kesten's stochastic approximation algorithm for multidimensional case* (Technical Report). Cadernos de Matemática, Série de Investigação, Collection of University of Aveiro, Department of Mathematics. http://193.136.81.248/dspace/handle/2052/74.

Cruz, P. (2005b). *Aproximação estocástica com valor do passo adaptativo*. PhD thesis, University of Aveiro, Department of Mathematics. http://193.136.81.248/dspace/handle/2052/103.

Dice, L. R. (1945). Measures of the amount of ecologic association between species. *Ecology*, *26*(3), 297–302.

Hermosillo, G., Chefd'hotel, C., & Faugeras, O. (2002). Variational methods for multimodal image matching. *International Journal of Computer Vision*, *50*(3), 329–343.

Hu, S., Hoffman, E. A., & Reinhardt, J. M. (2001). Automatic lung segmentation for accurate quantitation of volumetric X-Ray CT images. *IEEE Transactions on Medical Imaging*, *20*(6), 490–498.

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 Transactions on Image Processing*, *16*(12), 2879–2890.

Kushner, H. J., & Yin, G. G. (2003). *Stochastic Approximation and Recursive Algorithms and Applications* (2nd edn.) New York: Springer.

Maintz, J. B. A., & Viergever, M. A. (1998). A survey of medical image registration. *Medical Image Analysis*, *2*(1), 1–36.

Modersitzki, J. (2004). *Numerical Methods for Image Registration*. London: Oxford University Press.

Petersen, K. B., & Pedersen, M. S. (2007). The Matrix Cookbook. http://matrixcookbook.com.

Plakhov, A., & Cruz, P. (2004). A stochastic approximation algorithm with step size adaptation. *Journal of Mathematics and Sciences*, *120*(1), 964–973.

Pluim, J. P. W., Maintz, J. B. A., & Viergever, M. A. (2003). Mutual-information-based registration of medical images: a survey. *IEEE Transactions on Medical Imaging*, *22*(8), 986–1004.

Robbins, H., & Monro, S. (1951). A stochastic approximation method. *Annals of Mathematical Statistics*, *22*(3), 400–407.

Rueckert, D., Sonoda, L. I., Hayes, C., Hill, D. L. G., Leach, M. O., & Hawkes, D. J. (1999). Nonrigid registration using free-form deformations: Application to breast MR images. *IEEE Transactions on Medical Imaging*, *18*(8), 712–721.

Schäfer, J., & Strimmer, K. (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology*, *4*(1), article 32.

Shi, Z. J., & Shen, J. (2005). Step-size estimation for unconstrained optimization methods. *Computational & Applied Mathematics*, *24*(3), 399–416.

Sluimer, I., Prokop, M., & van Ginneken, B. (2005). Toward automated segmentation of the pathological lung in CT. *IEEE Transactions on Medical Imaging*, *24*(8), 1025–1038.

Thévenaz, P., & Unser, M. (2000). Optimization of mutual information for multiresolution image registration. *IEEE Transactions on Image Processing*, *9*(12), 2083–2099.

Viola, P., & Wells III, W. M. (1995). Alignment by maximization of mutual information. In Grimson, E., Shafer, S., Blake, A., & Sugihara, K. (Eds.), *International Conference on Computer Vision* (pp. 16–23). Los Alamitos: IEEE Computer Society Press.

Vysochanskij, D. F., & Petunin, Y. I. (1980). Justification of the 3σ rule for unimodal distributions. *Theory of Probability and Mathematical Statistics*, *21*, 25–36.

West, J., et al. (1997). Comparison and evaluation of retrospective intermodality brain image registration techniques. *Journal of Computer Assisted Tomography*, *21*(4), 554–566.

Zitová, B., & Flusser, J. (2003). Image registration methods: a survey. *Image and Vision Computing*, *21*(11), 977–1000.