Preconditioned Stochastic Gradient Descent Optimisation for Monomodal Image Registration

Stefan Klein[1,⋆], Marius Staring[2], Patrik Andersson[3], and Josien P.W. Pluim[3]

[1] Biomedical Imaging Group Rotterdam, Erasmus MC, Rotterdam, The Netherlands
[2] LKEB, Leiden University Medical Center, Leiden, The Netherlands
[3] Image Sciences Institute, University Medical Center Utrecht, The Netherlands
s.klein@erasmusmc.nl

Abstract. We present a stochastic optimisation method for intensity-based monomodal image registration. The method is based on a Robbins-Monro stochastic gradient descent method with adaptive step size estimation, and adds a preconditioning matrix. The derivation of the preconditioner is based on the observation that, after registration, the deformed moving image should approximately equal the fixed image. This prior knowledge allows us to approximate the Hessian at the minimum of the registration cost function, without knowing the coordinate transformation that corresponds to this minimum. The method is validated on 3D fMRI time-series and 3D CT chest follow-up scans. The experimental results show that the preconditioning strategy improves the rate of convergence.

Keywords: image registration, optimisation, stochastic gradient descent, preconditioner.

1 Introduction

Image registration is an important technique in medical imaging applications. It refers to the process of spatially aligning images. Extensive surveys on registration methods are presented in [2, 6] for example.

In this article, we focus on intensity-based image registration with a parameterised coordinate transformation. Let $F(\boldsymbol{x}) : \Omega_F \mapsto \mathbb{R}$ and $M(\boldsymbol{x}) : \Omega_M \mapsto \mathbb{R}$ denote the *fixed* and *moving* image, respectively, with $\Omega_F, \Omega_M \subset \mathbb{R}^D$, and D the dimension of the images. Define the parameterised coordinate transformation $\boldsymbol{T}(\boldsymbol{x};\boldsymbol{\mu}) : \Omega_F \times \mathbb{R}^Q \mapsto \Omega_M$ with the parameter vector $\boldsymbol{\mu}$ of dimension Q. Examples of parameterisations include rigid, affine, and B-spline models. The aim of image registration is to find transformation parameters $\hat{\boldsymbol{\mu}}$ such that the deformed moving image $M(\boldsymbol{T}(\boldsymbol{x};\hat{\boldsymbol{\mu}}))$ resembles the fixed image $F(\boldsymbol{x})$. This can be formulated as a minimisation problem:

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

⋆ Corresponding author.

G. Fichtinger, A. Martel, and T. Peters (Eds.): MICCAI 2011, Part II, LNCS 6892, pp. 549–556, 2011.

where $\mathcal{C}(\boldsymbol{\mu})$ represents the cost function. Examples of intensity-based cost functions are mean squared difference (MSD) and mutual information (MI).

In [3–5] it was demonstrated that a Robbins-Monro (RM) type stochastic gradient descent [9] method efficiently solves the minimisation problem (1). The basic RM method uses the following iterative scheme:

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

where $\tilde{\boldsymbol{g}}_k$ denotes an approximation of the true derivative $\boldsymbol{g}_k \equiv \partial\mathcal{C}/\partial\boldsymbol{\mu}(\boldsymbol{\mu}_k)$, $\gamma(k)$ is a scalar gain factor that determines the step size, and K is the number of iterations. The approximated derivative $\tilde{\boldsymbol{g}}_k$ is obtained by computing $\boldsymbol{g}_k$ using only a small subset of voxels $\boldsymbol{x} \in \Omega_F$, randomly selected in every iteration k [5]. The step size $\gamma(k)$ is defined as a slowly decaying function of k:

$$\gamma(k) = a/(k + A)^{\alpha}, \tag{3}$$

with user-specified constants $a > 0$, $A \geq 1$, and $0 < \alpha \leq 1$. It is important to set proper values for these constants. The optimal settings depend on the choice of $\mathcal{C}$, the transformation model, and the image content. To address this issue, an adaptive stochastic gradient descent (ASGD) method was proposed in [3]:

$$\boldsymbol{\mu}_{k+1} = \boldsymbol{\mu}_k - \gamma(t_k)\tilde{\boldsymbol{g}}_k, \quad t_{k+1} = [t_k + \text{sigmoid}(-\tilde{\boldsymbol{g}}_k'\tilde{\boldsymbol{g}}_{k-1})]^+, \tag{4}$$

where $[x]^+$ stands for $\max(x, 0)$, the prime denotes the transpose operation, t_0 and t_1 are user-defined constants, and γ as above. The "time" variable t_k realises an adaptive behaviour, in which the step sizes are increased when consecutive gradients $\tilde{\boldsymbol{g}}_k$ point in a similar direction, and vice versa. Based on the theoretical convergence conditions, reasonable values for a, A and α were estimated.

Both RM and ASGD are gradient descent type methods, which typically expose a low rate of convergence on badly scaled cost functions, characterised by a high ($\gg 1$) condition number of the Hessian $\boldsymbol{H} \equiv \partial^2\mathcal{C}/\partial\boldsymbol{\mu}\partial\boldsymbol{\mu}$ at $\hat{\boldsymbol{\mu}}$ [8]. In this paper, we propose a preconditioning strategy for RM and ASGD, specifically designed for monomodal image registration. The preconditioning is demonstrated to accelerate convergence in both rigid and nonrigid registration problems.

2 Method

2.1 Preconditioned Stochastic Gradient Descent

The use of a preconditioning matrix is a well-known technique to accelerate optimisation methods [8]. Based on the standard RM method, we define the following preconditioned stochastic gradient descent (PSGD) method:

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

where the preconditioner $\boldsymbol{P}$ is a positive definite $Q \times Q$ matrix. It serves to scale the derivative $\tilde{\boldsymbol{g}}_k$, and should be chosen such that larger steps are taken in

directions where the cost function is flat, and smaller steps in directions where the cost function has a high curvature. The theoretically optimal choice $\boldsymbol{P} = \boldsymbol{H}(\hat{\boldsymbol{\mu}})^{-1}$ makes (5) similar to the Newton-Raphson method. Unfortunately, since $\hat{\boldsymbol{\mu}}$ is unknown before registration, this choice of $\boldsymbol{P}$ is impossible to compute. It is however possible to compute an approximation, as follows.

2.2 A Preconditioner for Monomodal Image Registration

In this subsection, a preconditioning matrix for monomodal image registration problems is derived. For explanation, MSD is used as a cost function, but the derivation is similar for other cost functions. The MSD cost function is given by:

$$\mathcal{C}(\boldsymbol{\mu}) = \frac{1}{V} \sum_{\boldsymbol{x} \in \Omega_F} \left(F(\boldsymbol{x}) - M(\boldsymbol{T}(\boldsymbol{x}; \boldsymbol{\mu}))\right)^2, \tag{6}$$

with V the number of $\boldsymbol{x} \in \Omega_F$. For the derivative $\boldsymbol{g}(\boldsymbol{\mu})$ and the Hessian $\boldsymbol{H}(\boldsymbol{\mu})$ we have:

$$\boldsymbol{g}(\boldsymbol{\mu}) \equiv \frac{\partial \mathcal{C}}{\partial \boldsymbol{\mu}} = -\frac{2}{V} \sum_{\boldsymbol{x} \in \Omega_F} (F - M \circ \boldsymbol{T}) \frac{\partial \boldsymbol{T}'}{\partial \boldsymbol{\mu}} \frac{\partial M}{\partial \boldsymbol{x}}, \tag{7}$$

$$\begin{aligned} \boldsymbol{H}(\boldsymbol{\mu}) \equiv \frac{\partial^2 \mathcal{C}}{\partial \boldsymbol{\mu} \partial \boldsymbol{\mu}} = \frac{2}{V} \sum_{\boldsymbol{x} \in \Omega_F} \Bigg[& \frac{\partial \boldsymbol{T}'}{\partial \boldsymbol{\mu}} \frac{\partial M}{\partial \boldsymbol{x}} \frac{\partial M}{\partial \boldsymbol{x}}' \frac{\partial \boldsymbol{T}}{\partial \boldsymbol{\mu}} \\ & - (F - M \circ \boldsymbol{T}) \left(\frac{\partial^2 \boldsymbol{T}'}{\partial \boldsymbol{\mu} \partial \boldsymbol{\mu}} \frac{\partial M}{\partial \boldsymbol{x}} + \frac{\partial \boldsymbol{T}'}{\partial \boldsymbol{\mu}} \frac{\partial^2 M}{\partial \boldsymbol{x} \partial \boldsymbol{x}} \frac{\partial \boldsymbol{T}}{\partial \boldsymbol{\mu}} \right) \Bigg], \end{aligned} \tag{8}$$

where the compact notation $M \circ \boldsymbol{T} \equiv M(\boldsymbol{T}(\boldsymbol{x}; \boldsymbol{\mu}))$ was introduced, and all function arguments were omitted. Our aim is to find an approximation $\widetilde{\boldsymbol{H}}$ to $\boldsymbol{H}(\hat{\boldsymbol{\mu}})$, whose inverse can be used as a preconditioning matrix $\boldsymbol{P}$.

When F and M are images of the same modality, we can exploit the fact that $M \circ \boldsymbol{T}$ will be approximately equal to F after successful registration: $F(\boldsymbol{x}) \approx M(\boldsymbol{T}(\boldsymbol{x}; \hat{\boldsymbol{\mu}}))$. With this approximation, the following two identities are derived:

$$F - M \circ \boldsymbol{T} = 0, \quad \frac{\partial M}{\partial \boldsymbol{x}} = \left[\frac{\partial \boldsymbol{T}'}{\partial \boldsymbol{x}} \right]^{-1} \frac{\partial F}{\partial \boldsymbol{x}}. \tag{9}$$

Substituting (9) in (8) yields the following approximation of the Hessian at $\hat{\boldsymbol{\mu}}$:

$$\widetilde{\boldsymbol{H}} = \frac{2}{V} \sum_{\boldsymbol{x} \in \Omega_F} \frac{\partial \boldsymbol{T}'}{\partial \boldsymbol{\mu}} \left[\frac{\partial \boldsymbol{T}'}{\partial \boldsymbol{x}} \right]^{-1} \frac{\partial F}{\partial \boldsymbol{x}} \frac{\partial F'}{\partial \boldsymbol{x}} \left[\frac{\partial \boldsymbol{T}}{\partial \boldsymbol{x}} \right]^{-1} \frac{\partial \boldsymbol{T}}{\partial \boldsymbol{\mu}}. \tag{10}$$

Since $\hat{\boldsymbol{\mu}}$ is unknown, we approximate the terms $\partial \boldsymbol{T}/\partial \boldsymbol{\mu}(\boldsymbol{x}; \hat{\boldsymbol{\mu}})$ and $\partial \boldsymbol{T}/\partial \boldsymbol{x}(\boldsymbol{x}; \hat{\boldsymbol{\mu}})$ by $\partial \boldsymbol{T}/\partial \boldsymbol{\mu}(\boldsymbol{x}; \boldsymbol{\mu}_0)$ and $\partial \boldsymbol{T}/\partial \boldsymbol{x}(\boldsymbol{x}; \boldsymbol{\mu}_0)$, respectively. By setting $\partial \boldsymbol{T}/\partial \boldsymbol{x} \approx \boldsymbol{I}$ (assuming small deformations) the following expression is finally obtained:

$$\widetilde{\boldsymbol{H}} = \frac{2}{V} \sum_{\boldsymbol{x} \in \Omega_F} \frac{\partial \boldsymbol{T}'}{\partial \boldsymbol{\mu}}(\boldsymbol{x}; \boldsymbol{\mu}_0) \frac{\partial F}{\partial \boldsymbol{x}}(\boldsymbol{x}) \frac{\partial F'}{\partial \boldsymbol{x}}(\boldsymbol{x}) \frac{\partial \boldsymbol{T}}{\partial \boldsymbol{\mu}}(\boldsymbol{x}; \boldsymbol{\mu}_0). \tag{11}$$

We then define the preconditioning matrix by $\boldsymbol{P} \equiv [\widetilde{\boldsymbol{H}} + \beta\lambda\boldsymbol{I}]^{-1}$, with $0 \leq \beta \leq 1$ a user-defined factor, and $\lambda > 0$ the maximum eigenvalue of $\widetilde{\boldsymbol{H}}$, which can be estimated using an iterative block-Lanczos method. By adding the identity matrix, the condition number of $\boldsymbol{P}$ is limited to $(\beta+1)/\beta$, as a safeguard in case of ill-conditioned $\widetilde{\boldsymbol{H}}$, which may arise in nonrigid registration problems. In rigid registration problems, $\beta \downarrow 0$ is a valid choice. Instead of explicitly computing the matrix inverse, a Cholesky decomposition $\widetilde{\boldsymbol{H}} + \beta\lambda\boldsymbol{I} = \boldsymbol{L}\boldsymbol{L}'$ is used, allowing fast application of $\boldsymbol{P} \equiv \boldsymbol{L}^{'-1}\boldsymbol{L}^{-1}$ in each iteration. Note that $\widetilde{\boldsymbol{H}}$ is independent of $\boldsymbol{\mu}_k$, so $\widetilde{\boldsymbol{H}}$ and the Cholesky decomposition only need to be computed once.

2.3 Preconditioned Adaptive Stochastic Gradient Descent

A preconditioned version of ASGD, called PASGD, is derived in this subsection, which combines the adaptive step size mechanism of (4) with the preconditioner.

First we show that the PSGD method (5) is in fact a standard RM algorithm performed in a different parameter space. Let us introduce a new parameter vector $\boldsymbol{\nu} = \boldsymbol{L}'\boldsymbol{\mu}$. The original minimisation problem (1) is equivalent to:

$$\hat{\boldsymbol{\nu}} = \arg\min_{\boldsymbol{\nu}} \mathcal{D}(\boldsymbol{\nu}), \quad \text{with } \mathcal{D}(\boldsymbol{\nu}) \equiv \mathcal{C}(\boldsymbol{L}^{'-1}\boldsymbol{\nu}). \tag{12}$$

Define $\boldsymbol{h}_k \equiv \partial\mathcal{D}/\partial\boldsymbol{\nu}(\boldsymbol{\nu}_k) = \boldsymbol{L}^{-1}\partial\mathcal{C}/\partial\boldsymbol{\mu}(\boldsymbol{\mu}_k) = \boldsymbol{L}^{-1}\boldsymbol{g}_k$. The basic RM scheme (2) in terms of $\boldsymbol{\nu}$ reads $\boldsymbol{\nu}_{k+1} = \boldsymbol{\nu}_k - \gamma(k)\tilde{\boldsymbol{h}}_k$. Substituting $\tilde{\boldsymbol{h}}_k = \boldsymbol{L}^{-1}\tilde{\boldsymbol{g}}_k$ and $\boldsymbol{\nu} = \boldsymbol{L}'\boldsymbol{\mu}$ yields (13), and multiplying both sides of the equation by $\boldsymbol{L}^{'-1}$ yields (14):

$$\boldsymbol{L}'\boldsymbol{\mu}_{k+1} = \boldsymbol{L}'\boldsymbol{\mu}_k - \gamma(k)\boldsymbol{L}^{-1}\tilde{\boldsymbol{g}}_k, \tag{13}$$

$$\boldsymbol{\mu}_{k+1} = \boldsymbol{\mu}_k - \gamma(k)\boldsymbol{L}^{'-1}\boldsymbol{L}^{-1}\tilde{\boldsymbol{g}}_k, \tag{14}$$

in which we can recognise the preconditioner $\boldsymbol{P} \equiv \boldsymbol{L}^{'-1}\boldsymbol{L}^{-1}$.

Doing the same for the ASGD scheme (4) results in the PASGD method:

$$\boldsymbol{\mu}_{k+1} = \boldsymbol{\mu}_k - \gamma(t_k)\boldsymbol{P}\tilde{\boldsymbol{g}}_k, \quad t_{k+1} = [t_k + \text{sigmoid}(-\tilde{\boldsymbol{g}}_k'\boldsymbol{P}\tilde{\boldsymbol{g}}_{k-1})]^+. \tag{15}$$

Since the PASGD scheme is essentially an ordinary ASGD method in the domain of $\boldsymbol{\nu}$, all convergence conditions given in [3] remain valid, but should be interpreted in the space of $\boldsymbol{\nu}$. To estimate a, A, and α, a similar procedure as presented in [3] can therefore be used. The main idea is explained here, but the exact derivation is omitted for brevity. As suggested in [3], we set $\alpha = 1$ and $A = 20$. For a, the following expression was proposed in [3]:

$$a \equiv a_{\text{MAX}}\eta \equiv a_{\text{MAX}}\text{E}||\boldsymbol{g}||^2 / \left(\text{E}||\boldsymbol{g}||^2 + \text{E}||\boldsymbol{g} - \tilde{\boldsymbol{g}}||^2\right), \tag{16}$$

where $a_{\text{MAX}} \equiv 2A/\lambda$, with λ defined as the maximum eigenvalue of the Hessian of the cost function, and E denotes the expectation. The η factor intuitively takes into account that the step size should be reduced with increasing approximation error of $\tilde{\boldsymbol{g}}$. Whereas in [3] λ was unknown and had to be estimated from a user-defined maximum allowed voxel displacement per iteration, in this work

we can simply use $\lambda = 1$, as it can be derived that $\frac{\partial^2 \mathcal{D}}{\partial \boldsymbol{\nu} \partial \boldsymbol{\nu}} \approx \boldsymbol{I}$. Applying the reparametrisation to the definition of η changes the $\mathrm{E}||\boldsymbol{g}||^2$ terms to $\mathrm{E}\boldsymbol{g}'\boldsymbol{P}\boldsymbol{g}$ (and similar for $\boldsymbol{g} - \tilde{\boldsymbol{g}}$). Like in [3], the expectations in the definition of η are replaced by their empirical estimates. Since evaluating $\boldsymbol{g}$ (the exact cost function derivative) is an expensive operation, we apply the approximation $||\boldsymbol{g} - \tilde{\boldsymbol{g}}||^2 \approx ||\boldsymbol{g}||^2 + ||\tilde{\boldsymbol{g}}||^2$ to decouple the $\boldsymbol{g}$ and $\tilde{\boldsymbol{g}}$ terms. The term $\mathrm{E}\boldsymbol{g}'\boldsymbol{P}\boldsymbol{g}$ is estimated by a single measurement of $\boldsymbol{g}'\boldsymbol{P}\boldsymbol{g}$ (evaluated at $\boldsymbol{\mu}_0$), and $\mathrm{E}\tilde{\boldsymbol{g}}'\boldsymbol{P}\tilde{\boldsymbol{g}}$ is estimated by averaging a few (N) measurements at random positions $\boldsymbol{\mu}_n$ generated according to $\boldsymbol{\mu}_n - \boldsymbol{\mu}_0 \sim L^{'-1}\mathcal{N}(\boldsymbol{0}, \sigma_1^2 \boldsymbol{I})$, with $\sigma_1^2 = \mathrm{E}\boldsymbol{g}\boldsymbol{P}\boldsymbol{g}/Q$.

3 Experiments and Results

The proposed PASGD method was compared to the standard RM method, ASGD, and to a deterministic LBFGS quasi-Newton (QN) method [8]. All algorithms were integrated in `elastix` [4], an open source software package for image registration. The Cholesky decomposition was implemented using the CHOLMOD library [1]. Two applications were considered: rigid registration of 3D functional MR images (fMRI) and nonrigid registration of 3D CT chest scans.

3.1 Rigid Registration of fMRI Series

Eight fMRI time-series were acquired in the context of research on brain-computer interfaces (BCI). Seven series were recorded with a 2D EPI sequence; one with 3D PRESTO. Each time-series consisted of $\tau \approx$ 200-400 scans. The image size was $64 \times 64 \times$ [20-40], with $4 \times 4 \times 4$ mm voxels. In the BCI experiments, real-time rigid registration of each scan to the first scan is required [7]. For our experiment, scans at time points $t = 0, 1, 100, 200, (300,)$ and τ were selected. All scans with $t > 0$ were registered to the scan at $t = 0$, which resulted in a total of 37 registrations. Since the head's motion was small in most cases, the experiments were repeated with an extra initial offset to make the registration problem more challenging. The applied translations and rotations were drawn from a uniform distribution between ± 8 mm and $\pm 6°$, respectively.

The parameter vector $\boldsymbol{\mu}$ was formed by $\boldsymbol{t}$ and $\boldsymbol{S}\boldsymbol{\theta}$, where $\boldsymbol{t}$ is the translation vector, $\boldsymbol{\theta}$ represents the Euler angles, and $\boldsymbol{S}$ is a diagonal matrix with elements:

$$s_{ii} = \left(\int_{\Omega_F} \left|\left| \frac{\partial \boldsymbol{T}}{\partial \theta_i}(\boldsymbol{x}; \boldsymbol{\mu}_0) \right|\right|^2 d\boldsymbol{x} \Big/ \int_{\Omega_F} d\boldsymbol{x} \right)^{-\frac{1}{2}}. \tag{17}$$

Matrix $\boldsymbol{S}$ scales the rotation parameters by the average voxel displacement caused by a small perturbation of the rotation angle. This brings the values of the elements of $\boldsymbol{\mu}$ approximately in the same range, thus avoiding a very badly scaled cost function. For PASGD, the rescaling step was omitted ($\boldsymbol{S} = \boldsymbol{I}$), since the preconditioning matrix already should take care of this. To compute $\widetilde{\boldsymbol{H}}$, $V = 50\,000$ samples were used. To compute $\tilde{\boldsymbol{g}}_k$, $V = 2000$ random samples were used (except for QN, which used all voxels). For $\boldsymbol{P}$, $\beta = 10^{-7}$ was used.

Table 1. Results of the experiments with fMRI series

	No additional offset		With additional offset	
	$\|\|\Delta\boldsymbol{T}\|\|$ [mm] avg ± sd	nr. of it. avg ± sd	$\|\|\Delta\boldsymbol{T}\|\|$ [mm] avg ± sd	nr. of it. avg ± sd
QN	0.00 ± 0.00	3 ± 2	0.01 ± 0.02	8 ± 2
RM $a = 0.0025$	0.14 ± 0.19	48 ± 68	1.15 ± 1.48	216 ± 63
RM $a = 0.005$	0.14 ± 0.14	25 ± 39	0.30 ± 0.40	156 ± 82
RM $a = 0.01$	0.13 ± 0.13	17 ± 30	0.14 ± 0.14	80 ± 54
RM $a = 0.02$	0.14 ± 0.13	40 ± 52	0.14 ± 0.13	63 ± 41
RM $a = 0.04$	0.41 ± 0.61	134 ± 83	0.38 ± 0.55	134 ± 84
ASGD	0.14 ± 0.14	15 ± 30	0.15 ± 0.15	102 ± 53
PASGD	0.14 ± 0.13	11 ± 33	0.13 ± 0.13	19 ± 36

The number of iterations was set to $K = 250$. The RM method was tested for $a \in \{0.0025, 0.005, 0.01, 0.02, 0.04\}$, with $A = 50$ and $\alpha = 0.6$.

The result of QN without additional offset was treated as gold standard. The differences between the gold standard's transformation $\boldsymbol{T}(\boldsymbol{x}; \hat{\boldsymbol{\mu}}_{\text{gold}})$ and the other methods' transformations were computed to verify that all methods converged to the same solution. Table 1 presents for each method the average and standard deviation of $||\Delta\boldsymbol{T}|| \equiv ||\boldsymbol{T}(\boldsymbol{x}; \hat{\boldsymbol{\mu}}_{\text{gold}}) - \boldsymbol{T}(\boldsymbol{x}; \hat{\boldsymbol{\mu}})||$ over all $\boldsymbol{x}$ in all images. Both with and without additional offset, all differences were smaller than a voxel.

To compare the convergence rates, we chose ASGD as a baseline, and measured the performance improvement with respect to that method. For each method, we counted the number of iterations before $\mathcal{C}(\boldsymbol{\mu}_k) \leq 1.01 \cdot \mathcal{C}(\hat{\boldsymbol{\mu}}_{\text{ASGD}})$ occurred for the first time for at least 5 consecutive iterations. Note that $\mathcal{C}(\boldsymbol{\mu}_k)$ was calculated based on *all* voxels of the fixed image (not only the $V = 2000$ random samples that were used to compute $\tilde{\boldsymbol{g}}_k$). The results are summarised in Table 1 (nr. of it.) by the average and standard deviation over all images. The QN method required the least number of iterations, as expected, since it uses all voxels in each iteration (which makes it more expensive per iteration). RM performed best with $a \approx 0.01$, which gave similar results as ASGD. The PASGD method outperformed RM and ASGD.

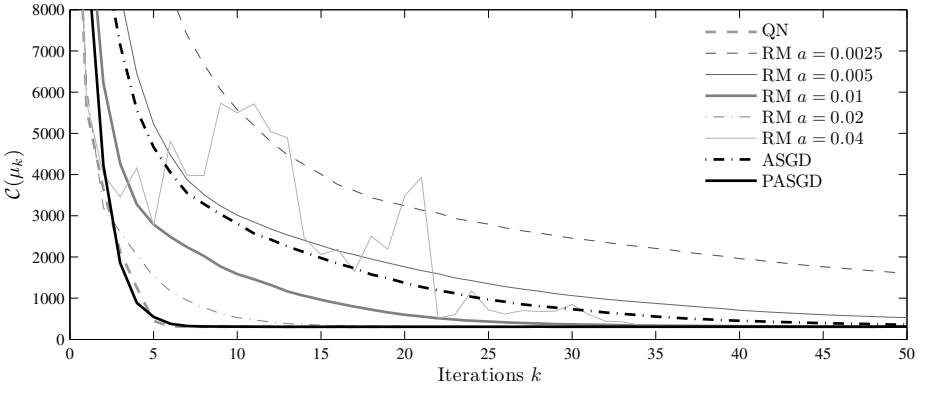

Fig. 1. Cost function plot for one of the fMRI experiments with additional offset

In Fig. 1, for one example image pair, the cost function $\mathcal{C}(\boldsymbol{\mu}_k)$ is plotted as a function of k for all methods. All 37 graphs were visually inspected and the pattern was fairly consistent. In a few cases the RM methods with $a \in \{0.02, 0.04\}$ suffered from instabilities (heavily fluctuating cost function values), indicating that the step sizes were too large.

3.2 Nonrigid Registration of CT Chest Scans

CT chest scans of five patients were obtained from the Department of Radiology, UMC Utrecht. For each patient a baseline and a follow-up scan, taken 3-9 months later, were available. Each scan was manually cropped around the right lung, and downsampled by a factor of two, which gave images of about $120 \times 160 \times 220$ voxels, with voxel size approximately $1.4 \times 1.4 \times 1.4$ mm. As a region of interest for registration, a dilated (kernel radius 10) lung segmentation was used.

Each follow-up scan was registered to the baseline scan using a B-spline transformation model [10]. Initial experiments showed that a regularisation term needs to be added to the cost function, to avoid foldings. The sum of second order spatial derivatives of the deformation field was used as a regularisation term, with a weighting factor of $5 \cdot 10^7$. The Hessian at $\boldsymbol{\mu}_0$ of this term was also included in the preconditioner. A three-level multiresolution strategy was employed. The distance between the B-spline control points was halved in each resolution level, such that at the final level the control points were spaced 20 mm in each direction. The

Table 2. Results of the experiments with CT chest scans. R3 is the finest resolution.

	$\|\|\Delta \boldsymbol{T}\|\|$ [mm] avg ± sd	R1 nr. of it. avg ± sd	R2 nr. of it. avg ± sd	R3 nr. of it. avg ± sd
QN	0.00 ± 0.00	11 ± 0	5 ± 3	3 ± 2
ASGD	1.26 ± 2.77	242 ± 3	214 ± 9	155 ± 28
PASGD $\beta = 1$	0.72 ± 1.47	52 ± 26	22 ± 11	25 ± 18
PASGD $\beta = 0.1$	0.36 ± 0.51	28 ± 36	5 ± 3	13 ± 10
PASGD $\beta = 0.01$	0.36 ± 0.38	39 ± 55	5 ± 3	13 ± 11

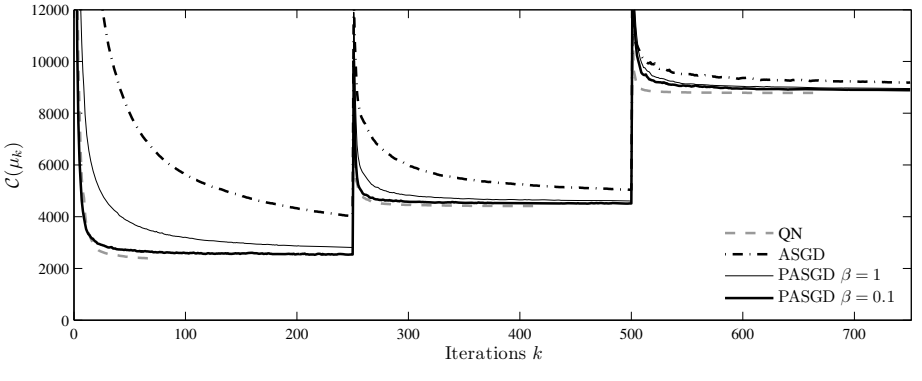

Fig. 2. Convergence results for one of the CT image pairs

images were smoothed using a Gaussian kernel with standard deviation of 2, 1, and 0.5 times the voxel size, for each resolution level respectively. The matrix $\widetilde{\boldsymbol{H}}$ was computed using $V = 100\,000$. PASGD was tested with $\beta \in \{1, 0.1, 0.01\}$. The tests with RM were omitted in this section.

For evaluation the same approach was followed as in the fMRI experiments. Table 2 summarises the evaluation results. The convergence results (nr. of it.) were calculated for each resolution separately (R1-R3). The numerical results in Table 2 indicate that PASGD achieved faster convergence than ASGD. The influence of β was moderate. Figure 2 plots the cost function series for one of the image pairs. PASGD with $\beta = 0.01$ was omitted for clarity, since it was very similar to $\beta = 0.1$.

4 Conclusion

The experiments with fMRI and CT data show that the proposed preconditioning technique has a beneficial effect on the rate of convergence, both in rigid and nonrigid registration problems. The PASGD method is, just as RM and ASGD, designed to work with stochastic estimates of the cost function derivatives, which leads to low computational costs per iteration [5]. The PASGD method couples this with a good rate of convergence by using second order information of the cost function.

References

1. Chen, Y., Davis, T., Hager, W., Rajamanickam, S.: Algorithm 887: CHOLMOD, supernodal sparse cholesky factorization and update/downdate. ACM Trans. Math. Softw. 35(3), 1–14 (2008)
2. Hill, D.L.G., Batchelor, P.G., Holden, M., Hawkes, D.J.: Medical image registration. Phys. Med. Biol. 46(3), R1–R45 (2001)
3. Klein, S., Pluim, J.P.W., Staring, M., Viergever, M.A.: Adaptive stochastic gradient descent optimisation for image registration. Int. J. Comput. Vis. 81(3), 227–239 (2009)
4. Klein, S., Staring, M., Murphy, K., Viergever, M.A., Pluim, J.P.W.: `elastix`: a toolbox for intensity-based medical image registration. IEEE Trans. Med. Imag. 29(1), 196–205 (2010)
5. Klein, S., Staring, M., Pluim, J.P.W.: Evaluation of optimization methods for nonrigid medical image registration using mutual information and B-splines. IEEE Trans. Image Process. 16(12), 2879–2890 (2007)
6. Maintz, J.B.A., Viergever, M.A.: A survey of medical image registration. Med. Image Anal. 2(1), 1–36 (1998)
7. Mathiak, K., Posse, S.: Evaluation of motion and realignment for functional magnetic resonance imaging in real time. Magn. Reson. Med. 45, 167–171 (2001)
8. Nocedal, J., Wright, S.J.: Numerical optimization. Springer, New York (1999)
9. Robbins, H., Monro, S.: A stochastic approximation method. Ann. Math. Stat. 22(3), 400–407 (1951)
10. Rueckert, D., Sonoda, L.I., Hayes, C., Hill, D.L.G., Leach, M.O., Hawkes, D.J.: Nonrigid registration using free-form deformations: Application to breast MR images. IEEE Trans. Med. Imag. 18(8), 712–721 (1999)