Proceedings of the 17th World Congress
The International Federation of Automatic Control
Seoul, Korea, July 6-11, 2008

A Globally Convergent Conjugate Gradient Method for Minimizing Self-Concordant Functions On Riemannian Manifolds ⋆

Huibo Ji * **Jonathan H. Manton** ** **John B. Moore** ***

* *ARC Centre of Excellence in Vision Science, Research School of Biological Sciences, The Australian National University, ACT 0200, Australia; National ICT Australia, 7 London Circuit Canberra ACT 0200, Australia(e-mail: hui-bo.ji@ anu.edu.au).*
** *Department of Electrical and Electronic Engineering, The University of Melborne, Parkville VIC 3010 Australia (e-mail:j.manton@ieee.org).*
*** *Department of Information Engineering, Research School of Information Sciences and Engineering, The Australian National University, ACT 0200, Australia (e-mail: John.Moore@anu.edu.au).*

Abstract: Self-concordant functions are a special class of convex functions in Euclidean space introduced by Nesterov. They are used in interior point methods, based on Newton iterations, where they play an important role in solving efficiently certain constrained optimization problems. The concept of self-concordant functions has been defined on Riemannian manifolds by Jiang et al. and a damped Newton method developed for this context. As a further development, this paper proposes a damped conjugate gradient method, which is an ordinary conjugate gradient method but with a novel step-size selection rule which is proved to ensure the algorithm converges to the global minimum. The advantage of the damped conjugate gradient algorithm over the damped Newton method is that the former has a lower computational complexity. To illustrate the advantages, the algorithm is applied to find the center of mass of given points on a hyperboloid model, known as the Karcher mean.

Keywords: Self-concordant; damped conjugate gradient method; Riemannian manifold; Karcher mean.

1. INTRODUCTION

Recently, algorithms for optimization on smooth manifolds have been developed with application in such areas as medicine Adler et al. [2002], signal processing Manton [2002], machine learning Nishimori et al. [2005], computer vision Ma et al. [2001], Helmke et al. [2004], and robotics Helmke et al. [2002], Helmke, Hüper et al. [2002].

A typical approach to optimization on smooth manifolds is to endow the manifold with a metric structure, to achieve a Riemannian manifold, see Smith et al. [1994], Edelman et al. [1998]. The simplest method for the optimization on Riemannian manifolds is the steepest descent method and although it has good convergence properties it has a slow linear convergence rate. Using second-order information on the cost function, the conjugate gradient method achieves super-linear convergence and the Newton method local quadratic convergence.

Self-concordant functions in Euclidean space were proposed by Nesterov and Nemirovskii Nesterov [2004] and used to construct the barrier function for the interior point method. A damped Newton method was developed for the optimization of such functions. This method is an ordinary Newton method with an explicit step-size choice which guarantees convergence. Building on this work, Jiang et al. [2007], develop and exploit the notion of self-concordance for optimization on Riemannian manifolds. A corresponding damped Newton method was proposed. As a result, there is guidance for the construction of efficient interior-point methods on smooth manifolds. For the Newton-based method, on Riemannian manifolds as well as in Euclidean space, the Newton descent direction is calculated by solving a linear system at each iteration. At potentially less computational cost, the conjugate gradient method can converge to the solution super-linearly without solving a linear system. In Smith et al. [1994], Smith generalized the conjugate gradient method on Riemannian manifolds, using a geodesic search method to find the step-size, generalizing the line search approach in Euclidean space. However, the geodesic search is often hard to compute in practice. Our goal here is to exploit the nice properties of self-concordant functions, and thereby develop a damped conjugate gradient method for the optimization of such functions on Riemannian manifolds. A key contribution is to introduce a novel simply calculated step-size selection to ensure convergence.

⋆ This research is partially supported by National ICT Australia, Ltd. (NICTA). National ICT Australia is funded through the Australian Government's *Backing Australia's Ability* initiative, in part through Australian Research Council.

We propose a conjugate gradient algorithm with an explicit step-size rule guaranteed to converge to the optimal solution of a self-concordant function. The main advantage of our approach is that it only uses the first and second covariant derivatives of the cost function without the need to compute a linear system, yet converges super-linearly. In each step, the complexity of our method is $O(n^2)$ instead of $O(n^3)$ for the damped Newton method, where n is the dimension of the Riemannian manifold.

To illustrate the convergence properties of our method, we apply it to find the Karcher mean of a set of given data points on a hyperboloid model. That is, given some points $p_1, \ldots, p_k$ on the hyperboloid model I_n, the problem is to find the point on I_n which minimizes the mean squared intrinsic distance to every point of $p_1, \ldots, p_k$. This Karcher mean was first introduced in Karcher [1977] as the center of mass on a Riemannian manifold. Methods to find this mean on Riemannian manifolds have been well studied, see Manton [2006], Absil et al. [2004]. However, the optimization tasks in Manton [2006], Absil et al. [2004] are defined on Riemannian manifolds with positive curvatures. Even though these methods can still be used to find the Karcher mean on Riemannian manifolds with negative curvatures, until now, we are not aware of any particular method exploiting the property of negative curvatures. In Karcher [1977], it is shown that the Karcher mean cost function defined on the Riemannian manifolds with negative curvatures is convex. Using this result together with the property that the hyperboloid model has constant negative curvature, it is here proved that the Karcher mean cost function defined on this model is self-concordant. Simulation results show our method converges to the Karcher mean of given points on the hyperboloid model super-linearly.

In Section 2, we will provide some preliminaries associated with self-concordant functions defined on Riemannian manifolds. Then in Section 3, the damped conjugate gradient method is proposed for the optimization of such functions and it is proved that this method converges to the minimum of the self-concordant function. In Section 4, an example is included to illustrate the convergence properties of our method.

2. SELF-CONCORDANT FUNCTIONS ON RIEMANNIAN MANIFOLDS

In this section, we review properties of self-concordant functions on Riemannian manifolds set out in Jiang et al. [2007], and provide motivation for introducing our proposed damped conjugate gradient method.

2.1 Notation for Riemannian manifolds

Let M denote a smooth n-dimensional geodesically complete Riemannian manifold. Recall that C^k smooth means derivatives of the order k exist and are continuous. For convenience, by smooth, we mean C^∞, that is, derivatives of all orders exist. Let T_pM denote the tangent space at the point $p \in M$. Since M is a Riemannian manifold, it comes with an inner product $\langle \cdot, \cdot \rangle_p$ on T_pM for each $p \in M$.

This induces the norm $\| \cdot \|_p$ given by $\|X\|_p = \langle X, X \rangle_p^{\frac{1}{2}}$ for $X \in T_pM$.

There is a natural way (precisely, the Levi-Civita connection) of defining acceleration on a Riemannian manifold which is consistent with the metric structure.

A curve with zero acceleration at every point is called a geodesic. Since M is geodesically complete, given a point $p \in M$ and a tangent vector $X \in T_pM$, there exists a unique geodesic $\gamma_X : \mathbb{R} \to M$ such that $\gamma_X(0) = p$ and $\dot{\gamma}_X(0) = X$. We therefore define an exponential map $\mathrm{Exp}_p : T_pM \to M$ by $\mathrm{Exp}_p(X) = \gamma_X(1)$ for all $X \in T_pM$. Note that $\mathrm{Exp}_p tX$ is the geodesic emanating from p in the direction X. Another consequence of M being geodesically complete is that any two points on M can be joined by a geodesic of shortest length. The distance $d(p, q)$ between two points $p, q \in M$ is defined to be the length of this minimizing geodesic. Since the length of the curve $\gamma : [0, 1] \to M$, $\gamma(t) = \mathrm{Exp}_p tX$, is $\|X\|_p$, it follows that if $q = \mathrm{Exp}_p X$ then $d(p, q) \leq \|X\|_p$, where the inequality is possible if there exists a shorter geodesic connecting p and q.

If $\gamma : [0, 1] \to M$ is a smooth curve from $p = \gamma(0)$ to $q = \gamma(1)$, there is an associated linear isomorphism $\tau_{pq} : T_pM \to T_qM$ called parallel transport. One of its properties is that lengths of vectors and angles between vectors are preserved, i.e. $\forall X, Y \in T_pM$, $\langle \tau_{pq}X, \tau_{pq}Y \rangle_q = \langle X, Y \rangle_p$.

Let N be an open subset of M. Consider the function $f : N \to \mathbb{R}$. Given $p \in N$ and $X \in T_pN$, the first, second and third covariant derivatives of f are defined as follows:

$$\nabla_X f(p) = \left.\frac{d}{dt}\right|_{t=0} \{f(\mathrm{Exp}_p tX)\}, \tag{1}$$

$$\nabla_X^2 f(p) = \left.\frac{d^2}{dt^2}\right|_{t=0} \{f(\mathrm{Exp}_p tX)\}, \tag{2}$$

$$\nabla_X^3 f(p) = \left.\frac{d^3}{dt^3}\right|_{t=0} \{f(\mathrm{Exp}_p tX)\}. \tag{3}$$

The gradient of f at $p \in N$, denoted by $\mathrm{grad}_p f$, is defined as the unique tangent vector in T_pN such that $\nabla_X f(p) = \langle \mathrm{grad}_p f, X \rangle$ for all $X \in T_pN$.

The Hessian of f at $p \in N$ is the unique symmetric bilinear form $\mathrm{Hess}_p f$ defined by the property

$$\mathrm{Hess}_p f(X, X) = \nabla_X^2 f(p), \quad X \in T_pN. \tag{4}$$

We say a subset N of M is convex if for any $p, q \in N$, out of all the geodesics connecting p and q, there is precisely one which is contained in N. Note that this is a weaker condition than that used extensively in Udriste [1994]. A function $f : N \subset M \to \mathbb{R}$ is said to be convex if N is a convex set and for any geodesic $\gamma : [0, 1] \to N$, the function $f \circ \gamma : [0, 1] \to \mathbb{R}$ satisfies the usual definition of convexity, namely

$$f(\gamma(t)) \leq (1 - t)f(\gamma(0)) + tf(\gamma(1)), \quad t \in [0, 1]. \tag{5}$$

If $f : N \to \mathbb{R}$ is $\mathcal{C}^\infty$-smooth and N is convex, then f is convex if and only if $\nabla_X^2 f(p) \geq 0$ for all $p \in N$ and $X \in T_pN$.

The epigraph $\mathrm{epi}(f)$ of f is defined by

$$\text{epi}(f) = \{(p,t) \in N \times \mathbb{R} | f(p) \leq t\}. \tag{6}$$

A function f is said to be closed if its epigraph epi(f) is closed in $M \times \mathbb{R}$.

2.2 Self-Concordant Functions

The definition of self-concordance was generalized to manifolds in Jiang et al. [2007] and is repeated below.

Definition 1. Let M be a smooth n-dimensional geodesically complete Riemannian manifold. Let $f : N \subset M \to \mathbb{R}$ be a $\mathcal{C}^3$-smooth closed function. Then f is self-concordant if

(1) N is an open convex subset of M;
(2) f is convex on N;
(3) there exists a constant $M_f > 0$ such that the inequality
$$|\nabla_X^3(f(p))| \leq M_f(\nabla_X^2 f(p))^{3/2} \tag{7}$$
holds for all $p \in N$ and $X \in T_pN$.

The reason why f is required to be closed in Definition 1 is to ensure that f behaves nicely on the boundary of N; this is shown in the following proposition.

Proposition 2. Let $f : N \to R$ be self-concordant. Let $\partial(N)$ denote the boundary of N. Then for any $\bar{p} \in \partial(N)$ and any sequence of points $p_k \in N$ converging to $\bar{p}$ we have $f(p_k) \to \infty$.

Self concordant functions have interesting properties which facilitate our further analysis. For details, see Jiang et al. [2007]. Let $f : N \to \mathbb{R}$ be self-concordant. Since second order covariant derivatives of self-concordant functions are always nonnegative, they can be used to define a Dikin-type ellipsoid $W(p;r) \subset T_pN$: for any $p \in N$, and $r > 0$,

$$W(p;r) := \{X_p \in T_pN \mid [\nabla_{X_p}^2 f(p)]^{1/2} < r\}. \tag{8}$$

Mapping all the elements in $W(p;r)$ by the exponential map Exp_p yields a subset $Q(p;r)$ of M where

$$Q(p;r) = \{q \in M \mid q = \text{Exp}_p X_p,\ X_p \in W(p;r)\}. \tag{9}$$

Then the following interesting properties hold.

(1) For any $p \in N \subseteq M$,
$$Q(p;1) \subseteq N. \tag{10}$$
(2) For any $p \in N$ and $X_p \in T_pN$, such that for $t \in [0,1]$ the geodesic $\text{Exp}_p tX_p$ is contained in N. Let $q = \text{Exp}_p X_p$. Then we have
$$f(q) \ \geq \ f(p) + \nabla_{X_p} f(p) + \omega([\nabla_{X_p}^2 f(p)]^{1/2}) \tag{11}$$
where $\omega(t) = t - \ln(1+t)$. Note that if $p \neq q$, then $\omega([\nabla_{X_p}^2 f(p)]^{1/2}) > 0$, hence (11) gives a useful lower bound on $f(q)$.
(3) For any $p \in N$ and $X_p \in W(p;1)$, let $q = \text{Exp}_p X_p$. Then we have
$$\begin{aligned}(1 - [\nabla_{X_p}^2 f(p)]^{1/2})^2 \nabla_{X_p}^2 f(p) &\leq \nabla_{\tau_{pq} X_p}^2 f(q) \\ &\leq \frac{\nabla_{X_p}^2 f(p)}{(1 - [\nabla_{X_p}^2 f(p)]^{1/2})^2},\end{aligned} \tag{12}$$
$$f(q) \ \leq \ f(p) + \nabla_{X_p} f(p) + \omega_*([\nabla_{X_p}^2 f(p)]^{1/2}) \tag{13}$$
where $\omega_*(t) = -t - \ln(1-t)$. Similarly, if $p \neq q$, then $\omega_*([\nabla_{X_p}^2 f(p)]^{1/2})$ is also strictly positive, hence (13) gives a useful upper bound on $f(q)$.

In Jiang et al. [2007], a damped Newton method is proposed for optimization of self-concordant functions on Riemannian manifolds based on its properties. This method is shown to quadratically converge to the minimum of a self-concordant function. However, since the damped Newton method is a Newton-based method, each of its step requires solving a linear system. Consequently, it increases the computational complexity. Therefore, we are motivated to find the gradient-based method for optimization of self-concordant functions on Riemannian manifolds. Before we give our damped conjugate gradient method in the next section, we need the following properties of self-concordance on Riemannian manifolds.

Proposition 3. Let $f_i : N \to R$ be self-concordant with constants M_{f_i}, $i = 1,2$ and let $\alpha, \beta > 0$. Then the function $f(x) = \alpha f_1(x) + \beta f_2(x)$ is self-concordant with the constant

$$M_f = \max\left\{\frac{1}{\sqrt{\alpha}} M_{f_1}, \frac{1}{\sqrt{\beta}} M_{f_2}\right\}. \tag{14}$$

Proposition 4. For any $p \in N$, and $X_p \in T_pN$, if $r = \sqrt{\nabla_{X_p}^2 f(p)} < 1$ we have

$$\begin{aligned}(1 - r + \frac{r^2}{3}) \nabla_{X_p}^2 f(p) &\leq \int_0^1 \nabla_{\tau_{p\text{Exp}_p tX_p} X_p}^2 f(\text{Exp}_p tX_p) dt \\ &\leq \frac{\nabla_{X_p}^2 f(p)}{1-r}.\end{aligned} \tag{15}$$

3. DAMPED CONJUGATE GRADIENT METHOD

In this section, a damped conjugate gradient method is presented for optimization of self-concordant functions.

Let $N \subset M$ be a convex open set. Then we consider optimization problems of the form

$$\min_{x \in N} f : N \subset M \to R. \tag{16}$$

In general, it is hard to solve (16) since N is an open subset of a Riemannian manifold. In Jiang et al. [2007], the case when f is self-concordant on Riemannian manifolds is considered.

This paper will concentrate on the special case of the optimization problem (16) when f satisfies the following assumption.

Assumption 1. The function f in (16) is self-concordant, has a minimum and $\nabla_X^2 f(p) > 0$, $\forall p \in N$, $X \in T_pN$. By scaling f if necessary, it is assumed without loss of generality that f satisfies (7) with $M_f = 2$.

Assumption 1 guarantees that f has a unique minimum on N. Let $K = \{p \in N \Big| f(p) \leq f(p_1),\ p_1 \in N\}$. Then $\forall p \in K$, Assumption 1 implies that there exist $\alpha,\ \theta > 0$ such that Smith et al. [1994]

$$\theta \|X\|_p^2 \leq \nabla_X^2 f(p) \leq \alpha \|X\|_p^2 \quad X \in T_p N. \tag{17}$$

Since self-concordant functions on Riemannian manifolds have nice properties, it is possible to define a damped conjugate gradient method to solve (16) when f in (16) satisfies Assumption 1.

Suppose we are at a point p_k at time k. Given an appropriate step-size t_k and conjugate gradient direction H_k, the conjugate gradient method sets $p_{k+1} = \mathrm{Exp}_{p_k} t_k H_k$.

From (13), provided $p_{k+1} \in W^0(p_k; 1)$, we have

$$f(p_k) - f(p_{k+1}) \geq -\nabla_{t_k H_k} f(p_k) + [\nabla_{t_k H_k}^2 f(p_k)]^{1/2} + \ln(1 - [\nabla_{t_k H_k}^2 f(p_k)]^{1/2}). \tag{18}$$

We propose choosing t_k to maximize the right side hand in (18). Later in Theorem 5, it is proved that such a strategy guarantees convergence to the minimum of the cost function. Initially, we assume that $\nabla_{H_k} f(p_k) < 0$. Later, in Lemma 1, it is proved that this assumption is correct. Hence, t_k is required to be positive.

The right side of (18) is of the form $\psi(t_k)$ where $\psi(t) = \alpha t + \ln(1 - \beta t)$ with $\alpha = -\nabla_{H_k} f(p_k) + \sqrt{\nabla_{H_k}^2 f(p_k)}$ and $\beta = \sqrt{\nabla_{H_k}^2 f(p_k)}$. Note that β will be strictly positive if we are not at the minimum of f. Therefore, ψ is defined on the interval $[0, 1/\beta)$. If $t_k \in [0, 1/\beta)$, $p_{k+1} \in W^0(p_k)$ as required for (18) to be a valid bound.

Differentiating $\psi(t)$ yields

$$\psi'(t) = \alpha - \frac{\beta}{1 - \beta t}, \tag{19}$$

$$\psi''(t) = -\frac{\beta^2}{(1 - \beta t)^2} < 0, \tag{20}$$

showing that $\psi(t)$ is concave on its domain $[0, 1/\beta)$. It achieves its maximum at

$$t = \frac{\alpha - \beta}{\alpha \beta}. \tag{21}$$

Let $\lambda_k = \frac{-\nabla_{H_k} f(p_k)}{\sqrt{\nabla_{H_k}^2 f(p_k)}}$. Substituting α and β into t, we obtain

$$t_k = \frac{\lambda_k}{(1 + \lambda_k)\sqrt{\nabla_{H_k}^2 f(p_k)}}. \tag{22}$$

Therefore, the proposed damped conjugate gradient algorithm for (16) is as follows.

Algorithm 1: Damped Conjugate Gradient Algorithm

step 0: Select an initial point $p_0 \in N$, compute $H_0 = G_0 = -\mathrm{grad}_{p_0} f$, and set $k = 0$.

step k: If $\mathrm{grad}_{p_k} f = 0$, then terminate. Otherwise, compute

$$\lambda_k = \frac{-\nabla_{H_k} f(p_k)}{\sqrt{\nabla_{H_k}^2 f(p_k)}}, \tag{23}$$

$$t_k = \frac{\lambda_k}{(1 + \lambda_k)\sqrt{\nabla_{H_k}^2 f(p_k)}}, \tag{24}$$

$$p_{k+1} = \mathrm{Exp}_{p_k} t_k H_k, \tag{25}$$

$$G_{k+1} = -\mathrm{grad}_{p_{k+1}} f, \tag{26}$$

$$\gamma_{k+1} = \frac{\langle G_{k+1}, G_{k+1} \rangle_{p_{k+1}}}{\langle G_k, H_k \rangle_{p_k}}, \tag{27}$$

$$H_{k+1} = G_{k+1} + \gamma_{k+1} \tau_{p_k p_{k+1}} H_k, \tag{28}$$

where $\tau_{p_k p_{k+1}}$ is the parallel translation with respect to the geodesic from p_k to p_{k+1}. If $k + 1 \mod n - 1 = 0$, set $H_{k+1} = G_{k+1}$. Increment k and repeat until convergence.

The convergence of Algorithm 1 is demonstrated in Theorem 5 with the help of Lemma 1, 2 and 3.

Lemma 1. Let the cost function $f : N \to \mathbb{R}$ in (16) satisfy Assumption 1. Assume p_0 is such that $\mathrm{grad}_{p_0} f \neq 0$. Then either 1) Algorithm 1 terminates after a finite number iterations if $\mathrm{grad}_{p_k} f = 0$ at a certain k, or 2) Algorithm 1 generates an infinite sequence $\{p_k\}$ of points (That is, there are no divisions by zeros) if zero gradient never encountered in the iteration and moreover, $\forall k$, $\nabla_{H_k} f(p_k) = \langle \mathrm{grad}_{p_k} f, H_k \rangle_{p_k} < 0$.

Lemma 2. Let $\{p_k\}$ be an infinite sequence of points generated by Algorithm 1 where the cost function $f : N \to R$ satisfies Assumption 1. Then:

(1) $\forall k$, $p_k \in N$.
(2) If $\mathrm{grad}_{p_{k+1}} f \neq 0$, then $\lambda_k > 0$.
(3) If $\mathrm{grad}_{p_{k+1}} f \neq 0$, then $f(p_{k+1}) \leq f(p_k) + \omega(\lambda_k) < f(p_k)$ where $\omega(t) = t - \ln(1 + t)$.

Lemma 3. Let $\{p_k\}$ and $\{H_k\}$ be infinite sequences generated by Algorithm 1 where the cost function $f : N \to R$ satisfies Assumption 1. If $\mathrm{grad} f_{p_k} \neq 0$, then for all k

$$\frac{\|H_{k+1}\|_{p_{k+1}}^2}{\|\mathrm{grad}_{p_{k+1}} f\|_{p_{k+1}}^4} \leq \frac{\|H_k\|_{p_k}^2}{\|\mathrm{grad}_{p_k} f\|_{p_k}^4} + \frac{3}{\|\mathrm{grad}_{p_k} f\|_{p_k}^2}. \tag{29}$$

Theorem 5. Consider the optimization problem in (16). If the cost function $f : N \to \mathbb{R}$ in (16) satisfies Assumption 1, then Algorithm 1 converges to the unique minimum of f.

Proofs of Lemma 1, 2, 3 and Theorem 5 are omitted to save space. Interested readers may ask the authors for details.

4. ILLUSTRATIVE EXAMPLE

In this section, we consider the problem of computing the center of mass of a set of given points defined on the hyperboloid model. Before defining this problem, we first introduce the geometric properties of the hyperboloid model. In $\mathbb{R}^{n+1}$, consider the following quadratic form Q,

$$Q(x) = -\sum_{i=1}^{n} x_i^2 + x_{n+1}^2. \tag{30}$$

Let $A = \mathrm{diag}(\underbrace{-1,-1,\ldots,-1}_{n},1)$. Then Q can be represented in terms of A by

$$Q(x) = x^T Ax. \tag{31}$$

Given this quadratic form, the upper fold I_n of the hyperboloid is determined by the formula

$$I_n = \{x \in \mathbb{R}^{n+1} | Q(x) = 1,\ x_{n+1} > 0\}. \tag{32}$$

The set I_n can be regarded as a differentiable hypersurface in $\mathbb{R}^{n+1}$ since it is an open subset of the pre-image of a regular value of a differentiable function. In particular, it inherits from $\mathbb{R}^{n+1}$ a differentiable structure of dimension n. For any $x \in I_n$, the tangent space $T_x I_n$ is

$$T_x I_n = \{X \in \mathbb{R}^{n+1} | x^T AX = 0\}. \tag{33}$$

For any $x \in I_n$, we define a Riemannian metric on the tangent space of x by

$$\langle X, Y\rangle = X^T(-A)Y,\ X, Y \in T_x I_n. \tag{34}$$

Given a point $x \in I_n$ and a non-zero tangent vector $X \in T_x I_n$, the geodesic emanating from x in the direction X is given by Benedetti et al. [1992]

$$\mathrm{Exp}_x tX = x\cosh(\theta t) + \frac{1}{\theta} X \sinh(\theta t) \tag{35}$$

where $\theta = \sqrt{X^T(-A)X}$.

The intrinsic distance between x and y on this hyperboloid model is given by William [1993]

$$d(x, y) = \mathrm{arccosh}(x^T Ay). \tag{36}$$

Recall that $\mathrm{arccosh}(t) = \ln(t + \sqrt{t^2 - 1})$ for $t > 1$. Since $d(x, y) \geq 0$, we can expect that $x^T Ay \geq 1$ holds for all $x,\ y \in I_n$.

Consider the following optimization problem

$$\arg\min_{p\in I_n} f(p) = \frac{1}{2}\sum_{i=1}^{n} d^2(p, p_i) = \frac{1}{2}\sum_{i=1}^{n} \mathrm{arccosh}^2(p^T Ap_i) \tag{37}$$

where $p_i \in I_n,\ i = 1, \ldots, m$. The solution of (37) is called the Karcher mean Karcher [1977] of the given points $p_1,\ \ldots,\ p_m$.

By computation, we obtain the first, second and third covariant derivatives of the cost function at $p \in I_n$ in the direction $H \in T_p I_n$ by

$$\nabla_H f(p) = \theta \sum_{i=1}^{m} \mathrm{arccosh}(p^T Ap_i) \frac{X^T Ap_i}{\sqrt{(p^T(-A)p_i)^2 - 1}},$$

$$\nabla_H^2 f(p) = \theta^2 \sum_{i=1}^{m} \Bigg[\frac{(X^T Ap_i)^2}{(p^T AP_i)^2 - 1} + \mathrm{arccosh}(p^T Ap_i) \frac{p^T Ap_i[(p^T Ap_i)^2 - (X^T Ap_i)^2 - 1]}{(p^T AP_i)^2 - 1)^{\frac{3}{2}}} \Bigg],$$

$$\nabla_H^3 f(p) = \theta^3 \sum_{i=1}^{m} \left[\frac{X^T Ap_i}{((p^T AP_i)^2 - 1)^{\frac{3}{2}}} - \frac{(X^T Ap_i)^3}{((p^T AP_i)^2 - 1)^{\frac{5}{2}}} \right] (3p^T Ap_i(p^T Ap_i^2 - 1)^{\frac{1}{2}} - \mathrm{arccosh}(p^T Ap_i)(2p^T Ap_i^2 + 1)),$$

where $\theta = \sqrt{H^T(-A)H}$ and $X = H/\theta$.

In view of Theorem 2.1 in Page 111 in Udriste [1994], since I_n is simply connected, complete with negative sectional curvature, the function $f(p)$ in (37) is strictly convex. Hence, this implies that for all $p \in I_n$ and non-zero tangent vector $H \in T_p I_n$

$$\nabla_H^2 f(p) > 0. \tag{38}$$

For any $p \in I_n$, the gradient $\mathrm{grad}_p f$ of f is given by

$$\mathrm{grad}_p f = (k^T Ap)p - k \tag{39}$$

where $k = \sum_{i=1}^{m} \frac{\arccos(p^T Ap_i)p_i}{\sqrt{(p^T(-A)p_i)^2 - 1}}$.

The following lemma shows that the function f in (37) is self-concordant.

Lemma 4. The function f in (37) is a self-concordant function defined on the n-dimensional hyperboloid model with the constant $M_f = \sqrt{\frac{16}{27}}$.

Since the function f in (37) is self-concordant, we are able to apply our damped conjugate gradient algorithms to find the minimum of f on I_n. The proposed damped conjugate gradient method for solving (37) is given as follows.

Algorithm 2: Damped Conjugate Gradient Algorithm for (37)

step 0: Select an initial point $p_0 \in I_n$, compute $H_0 = G_0 = -\mathrm{grad}_{p_0} f$ by (43), and set $k = 0$.

step k: If $\mathrm{grad}_{p_k} f = 0$, then terminate. Otherwise, compute

$$\lambda_k = \frac{-\nabla_{H_k} f(p_k)}{\sqrt{\nabla_{H_k}^2 f(p_k)}}, \tag{40}$$

$$t_k = \frac{\lambda_k}{(1 + \lambda_k)\sqrt{\nabla_{H_k}^2 f(p_k)}}, \tag{41}$$

$$p_{k+1} = p_k \cosh\left(\sqrt{H_k(-A)H_k} t_k\right) + \frac{H_k}{\sqrt{H_k(-A)H_k}} \sinh\left(\sqrt{H_k(-A)H_k} t_k\right), \tag{42}$$

$$G_{k+1} = -\mathrm{grad}_{p_{k+1}} f, \tag{43}$$

$$\gamma_{k+1} = \frac{\langle G_{k+1}, G_{k+1}\rangle_{p_{k+1}}}{\langle G_k, H_k\rangle_{p_k}}, \tag{44}$$

$$H_{k+1} = G_{k+1} + \gamma_{k+1} \tau_{p_k p_{k+1}} H_k, \tag{45}$$

where $\tau_{p_k p_{k+1}}$ is the parallel transport with respect to the geodesic from p_k to p_{k+1}. If $k + 1 \mod n - 1 = 0$, set $H_{k+1} = G_{k+1}$. Increment k and repeat until convergence.

We applied the above algorithm to task (37) and compared its performance against the damped Newton method Jiang et al. [2007]. In particular, we take $n = 19$. All the simulations were implemented in MATLAB 7.0 and all results were obtained using a 3.06 GHZ Pentium 4 machine, with 1Gb of memory, running Windows XP Professional.

Figure 1 illustrates the result of the damped Newton method and conjugate gradient method on function f in (37). Table 1 shows the simulation time and accuracy using the damped conjugate gradient and Newton methods. From Figure 1, it can be see that the damped Newton method converges to the minimum quadratically, whereas the damped conjugate gradient method converges super-linearly. Although the damped conjugate gradient method requires more steps, since it avoids computing a linear system, it takes less time than the damped Newton method, seen from Table 1.

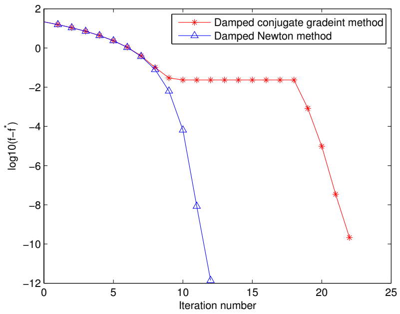

Fig. 1. The result of damped conjugate gradient method for (37)

algorithm	time(second)	accuracy
damped conjugate gradient method	0.062	10^{-5}
damped Newton method	0.313	10^{-5}

Table 1. Simulation time and accuracy

5. CONCLUSIONS

In this paper, we propose a damped conjugate gradient method for optimization of self-concordant functions on Riemannian manifolds. Such a method is a conjugate gradient method with a novel step-size selection rule, which ensures that this algorithm converges to the global minimum. The advantage of the damped conjugate gradient method over the quadratically damped Newton method is that the former has a lower computational complexity yet converges super-linearly. Both methods are applied to examples and shown to converge to the minimum of a self-concordant function.

REFERENCES

R. Adler, J.-P. Dedieu, J. Margulies, M. Martens and M. Shub. Newton Method on Riemannian Manifolds and a Geometric Model for the Human Spine. *IMA Journal on Numerical Analysis*, vol.22, pp. 359-390, 2002.

J. H. Manton. Optimization algorithms exploiting unitary constraints. *IEEE Transactions on Signal Processing*, vol. 50, no. 3, pp. 635-650, 2002.

Y. Nishimori and S. Akaho. Learning algorithms utilizing quasi-geodesic flows on the Stiefel manifold. *Neurocomputing*, vol. 67, pp. 106-135, 2005.

Y. Ma, J.Kosecka and S. Sastry. Optimization criteria and geometric algorithms for motion and structure estimation. *International Journal of Computer Vision*, vol. 44, no. 3, pp. 219-249, 2001.

U. Helmke, K.Hüper, P. Lee and J. Moore. Essential matrix estimation via Newton-type methods. *Proceedings of the MTNS*, Leuven, 2004.

U. Helmke, S. Ricardo and S. Yoshizawa. Newton's algorithm in Euclidean Jordan algebras, with applications to robotics. *Commun. Inf. Syst.*, vol. 2, no. 3, pp. 283-297, 2002.

U. Helmke, K. Hüper and J. Moore. Quadratically convergent algorithms for optimal dextrous hand grasping. *IEEE Transactions on Robotics and Automation*, vol. 18, no. 2, pp. 138-146, April 2002.

D. Jiang, J.B. Moore and H. Ji. Self-Concordant Functions for Optimization on Smooth Manifolds. *Journal of Global Optimization*, vol. 38, no. 3, pp. 437-457, July 2007.

Y. Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course.* Kluwer Academic Publishers, 2004.

S. T. Smith. Optimization techniques on Riemannian manifolds. *Faelds Institute Communications*, 3:113-135, 1994.

A. Edelman, T. A. Arias, and S. T. Smith. The geometry of algorithms with orthogonality constraints. *SIAM Journal on Matrix Analysis and Applications*, 20(2):303-353, 1998.

J. H. Manton. A centroid (Karcher mean) approach to the joint approximate diagonalisation problem: The real symmetric case. *Digital Signal Processing*, vol.16, pp. 468-478, 2006.

P. -A. Absil, R. Mahony and R. Sepulchre. Riemannian geometry of Grassmann manifolds with a view on algorithmic computation. *Acta Applicandae Mathematicae*, vol.80, no.2, pp. 199-220, January 2004.

R. Godement. *Analysis I: Convergence, Elementary Functions.* Springer, 2004.

S. Boyd and L. Vandenberghe. *Convex Optimization.* Cambridge University Press, 2004.

F. R. William. Hyperbolic Geometry on a Hyperboloid. *The American Mathematical Monthly*, 100:442-455, 1993.

R. Benedetti, and C. Petronio. Lectures on Hyperbolic Geometry. Springer-Verlag, Berlin, 1992.

C. Udriste. Convex Funtions and Optimization Methods on Riemannian Manifolds. Kluwer Academic Publishers, Dordrecht, 1994.

H. Karcher. Riemannian center of mass and mollifier smoothing. *Communications on Pure and Applied Mathematics*, 30:509-541, 1977.