Newton-like methods for numerical optimization on manifolds

Knut Hüper
National ICT Australia Ltd.
Systems Engineering and Complex Systems Program
Locked Bag 8001
Canberra ACT 2601, Australia
Email: Knut.Hueper@nicta.com.au

Jochen Trumpf
Department of Information Engineering
The Australian National University
Canberra ACT 0200
and
National ICT Australia Ltd.
Email: Jochen.Trumpf@anu.edu.au

Abstract—**Many problems in signal processing require the numerical optimization of a cost function which is defined on a smooth manifold. Especially, orthogonally or unitarily constrained optimization problems tend to occur in signal processing tasks involving subspaces. In this paper we consider Newton-like methods for solving these types of problems. Under the assumption that the parameterization of the manifold is linked to so-called Riemannian normal coordinates our algorithms can be considered as intrinsic Newton methods. Moreover, if there is not such a relationship, we still can prove local quadratic convergence to a critical point of the cost function by means of analysis on manifolds. Our approach is demonstrated by a detailed example, i.e., computing the dominant eigenspace of a real symmetric matrix.**

I. Introduction

Recently, differential geometric ideas were combined with Newton-type methods to solve optimization problems in medicine [1], signal processing [2], geometric means in statistics [3], computer vision [4],[5], and robotics [6],[7]. Current research in the field is focused on (i) a unifying convergence theory, (ii) generalizations to the non-Riemannian setting and (iii) development of new application areas. In this paper we contribute to (i) and (ii).

II. Preliminaries

A. Parametrizations

In this section we briefly recall the notion of local parametrization for smooth manifolds. For further details we refer to [8]. Let M be a smooth n-dimensional real manifold then for every point $p \in M$ there exists a smooth map

$$\mu_p : \mathbb{R}^n \longrightarrow M, \; \mu_p(0) = p$$

which is a local diffeomorphism around $0 \in \mathbb{R}^n$. Such a map will be called a *local parametrization around* p.

If there exists an open neighborhood $U \subset M$ of p^* and a smooth map

$$\mu : U \times \mathbb{R}^n \longrightarrow M$$

such that $\mu(p, x) = \mu_p(x)$ for all $p \in U$ and $x \in \mathbb{R}^n$ we will call $\{\mu_p\}_{p \in M}$ a *locally smooth family of parametrizations around* p^*.

B. Newton's method on $\mathbb{R}^n$

Let $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ be a smooth function and let $x^* \in \mathbb{R}^n$ be a nondegenerate critical point of f, i.e. the Hessian $H_f(x^*)$ is invertible. Newton's method for f is the iteration

$$x_0 \in \mathbb{R}^n, \; x_{k+1} = N_f(x_k) := x_k - (H_f(x_k))^{-1} \nabla_f(x_k) \quad (1)$$

where $\nabla_f(x)$ denotes the gradient with respect to the standard Euclidean inner product. Note that the iteration (1) is only defined if $H_f(x_k)$ is invertible for all $k \in \mathbb{N}_0$. However, since f is smooth, there exists an an open neighborhood of x^* in which the Hessian of f is invertible.

It is well known that the point sequence $\{x_k\}_{k \in \mathbb{N}_0}$ generated by (1) is defined and converges locally quadratic to x^* provided that x_0 is sufficiently close to x^*. For more information see e.g. [9].

III. Main result

In the following we propose a class of Newton-like algorithms to compute a nondegenerate critical point of a smooth cost function $f : M \longrightarrow \mathbb{R}$ which is defined on a smooth manifold M. The main result establishes local quadratic convergence of the proposed algorithm.

Let M be a smooth manifold and let $p^* \in M$ be a nondegenerate critical point of the smooth function $f : M \longrightarrow \mathbb{R}$. Let $\{\mu_p\}_{p \in M}$ and $\{\nu_p\}_{p \in M}$ be two locally smooth families of parametrizations around p^*. Consider the following iteration on M

$$p_0 \in M, \; p_{k+1} = \nu_p \left(N_{f \circ \mu_p}(0) \right) \quad (2)$$

where $N_{f \circ \mu_p}$ is defined in (1).

Theorem 1: Under the condition

$$\mathrm{D}\, \mu_{p^*}(0) = \mathrm{D}\, \nu_{p^*}(0) \quad (3)$$

there exists an open neighborhood $V \subset M$ of p^* such that the point sequence $\{p_k\}_{k \in \mathbb{N}_0}$ generated by (2) converges quadratically to p^* provided $p_0 \in V$.

National ICT Australia Limited is funded by the Australian Government's Department of Communications, Information Technology and the Arts and the Australian Research Council through Backing Australia's Ability and the ICT Centre of Excellence Program.

Proof: Let $\mu, \nu : U \times \mathbb{R}^n \longrightarrow M$ be smooth and such that $\mu(p,x) = \mu_p(x)$ and $\nu(p,x) = \nu_p(x)$ for all $p \in U$ and $x \in \mathbb{R}^n$, where U is a neighborhood of p^*.

In the following we compute the derivative of the algorithm map

$$s : M \longrightarrow M,\ p \mapsto \nu\left(p, -\left(H_{f\circ\mu(p,\cdot)}(0)\right)^{-1} \nabla_{f\circ\mu(p,\cdot)}(0)\right)$$

at p^*.

Let $h \in \mathrm{T}_{p^*}M$ then

$$\begin{aligned} &\mathrm{D}_1\,\nu\left(p^*, -\left(H_{f\circ\mu(p^*,\cdot)}(0)\right)^{-1} \nabla_{f\circ\mu(p^*,\cdot)}(0)\right) h \\ &= \mathrm{D}_1\,\nu(p^*,0)h = h \end{aligned} \tag{4}$$

where the first equality holds since p^* is a critical point of f. Furthermore, we have using the standard Euclidean inner product $\langle\cdot,\cdot\rangle$ on $\mathbb{R}^n$

$$\langle \nabla_{f\circ\mu(p,\cdot)}(0), x\rangle = \mathrm{D}\,f(p)\,\mathrm{D}_2\,\mu(p,0)x$$

for all $p \in U$ and $x \in \mathbb{R}^n$, and hence by differentiating with respect to p at p^*

$$\left\langle \mathrm{D}_p\left(\nabla_{f\circ\mu(p^*,\cdot)}(0)\right)h, x\right\rangle = \mathrm{D}^2\,f(p^*)(\mathrm{D}_2\,\mu(p^*,0)x, h)$$

for all $x \in \mathbb{R}^n$ and $h \in \mathrm{T}_{p^*}M$ where the second term of the product rule vanishes because p^* is critical. For

$$x = \left(H_{f\circ\mu(p^*,\cdot)}(0)\right)^{-1} y \quad \text{and} \quad h = \mathrm{D}_2\,\mu(p^*,0)z\,,$$

where $y, z \in \mathbb{R}^n$ are arbitrary, we get using the symmetry of the Hessian and its inverse

$$\begin{aligned} &\left\langle \left(H_{f\circ\mu(p^*,\cdot)}(0)\right)^{-1} \mathrm{D}_p\left(\nabla_{f\circ\mu(p^*,\cdot)}(0)\right) \mathrm{D}_2\,\mu(p^*,0)z, y\right\rangle \\ &= \mathrm{D}^2\,f(p^*)(\mathrm{D}_2\,\mu(p^*,0)\left(H_{f\circ\mu(p^*,\cdot)}(0)\right)^{-1} y, \mathrm{D}_2\,\mu(p^*,0)z) \\ &= \left\langle H_{f\circ\mu(p^*,\cdot)}(0)\left(H_{f\circ\mu(p^*,\cdot)}(0)\right)^{-1} y, z\right\rangle \\ &= \langle z, y\rangle \end{aligned}$$

for all $y, z \in \mathbb{R}^n$. The next to last equality follows from

$$\left\langle H_{f\circ\mu(p^*,\cdot)}(0)u, v\right\rangle = \mathrm{D}^2\,f(p^*)\left(\mathrm{D}_2\,\mu(p^*,0)u, \mathrm{D}_2\,\mu(p^*,0)v\right)$$

for all $u, v \in \mathbb{R}^n$. We conclude using $z = (\mathrm{D}_2\,\mu(p^*,0))^{-1}h$

$$\left(H_{f\circ\mu(p^*,\cdot)}(0)\right)^{-1} \mathrm{D}_p\left(\nabla_{f\circ\mu(p^*,\cdot)}(0)\right) h = (\mathrm{D}_2\,\mu(p^*,0))^{-1}h$$

and hence using (4) and condition (3)

$$\mathrm{D}\,s(p^*) = h - \mathrm{D}_2\,\nu(p^*,0)(\mathrm{D}_2\,\mu(p^*,0))^{-1}h = 0$$

where the second term of the product rule in the second summand vanishes since p^* is a critical point.

Now the result follows from a Taylor expansion of s around p^*. □

A few remarks are in order. Geometrically, algorithm (2) does the following. The current iteration point p_k is pulled back to Euclidean space via the local parametrization μ_{p_k} around p_k, then one Euclidean Newton step is performed for the function expressed in local coordinates, followed by a projection back onto the manifold using the local parametrization ν_{p_k} around p_k.

In the case of a Riemannian manifold M and for the special choice $\{\mu_p\}_{p\in M} = \{\nu_p\}_{p\in M}$, both Riemannian normal coordinates (cf. [10]), algorithm (2) is precisely the so-called Newton method along geodesics of D. Gabay [11], more recently also referred to as the intrinsic Newton method.

M. Shub [12] discusses a Newton method to compute a zero of a smooth vector field on a smooth manifold endowed with a connection. His algorithm allows for smooth projections from the tangent bundle which have derivative equal to the identity at the base point. In the case of a gradient vector field on a Riemannian manifold endowed with the Levi-Civita connection, Shub's algorithm coincides with ours when $\{\mu_p\}_{p\in M}$ are Riemannian normal coordinates.

In [13],[14] the Newton method along geodesics of [11] was rediscovered and variants with different projections were proposed. However, the convergence proofs given do not apply to this more general situation, except for one particular case where the algorithm obtained coincides with the classical Rayleigh quotient iteration on the sphere.

P.-A. Absil et al. [15] further discuss the Newton method along geodesics and derive a cubic convergence result.

J.H. Manton [2] proposes a similar algorithm for a specific cost function on a specific manifold, but gives no convergence proof.

IV. Application

In the sequel we apply algorithm (2) in order to compute the dominant eigenspace of a real symmetric matrix.

A. Problem formulation

Let $N \in \mathbb{R}^{n\times n}$ be a symmetric matrix, i.e. $N = N^\top$. Let $\lambda_1 \geq \cdots \geq \lambda_k > \lambda_{k+1} \geq \cdots \geq \lambda_n$ be the eigenvalues of N. The *k-dimensional dominant eigenspace $\mathcal{V}_N$ of* N is the subspace of $\mathbb{R}^n$ spanned by the eigenvectors corresponding to the k largest eigenvalues $\lambda_1, \ldots, \lambda_k$. The requirement $\lambda_k > \lambda_{k+1}$ makes this subspace unique.

The problem we want to solve is given N, how to compute a basis for $\mathcal{V}_N$.

B. The Graßmann manifold

Since our objective is to compute a certain k-dimensional subspace of $\mathbb{R}^n$, the natural manifold to consider is the *Graßmann manifold* $\mathrm{G}(k,n)$ of all k-dimensional subspaces of $\mathbb{R}^n$. This manifold can be considered as a quotient manifold of the *Stiefel manifold* $\mathrm{St}(k,n)$ of all matrices $X \in \mathbb{R}^{n\times k}$ with $X^\top X = I_k$ with respect to the right action

$$\phi : \mathrm{St}(k,n) \times \mathrm{O}(k) \longrightarrow \mathrm{St}(k,n),\ (X,Q) \mapsto X \cdot Q \tag{5}$$

of the *orthogonal group* $\mathrm{O}(k)$ of all matrices $Q \in \mathbb{R}^{k\times k}$ with $Q^\top Q = I_k$ (see e.g. [16]). We denote an element of $\mathrm{G}(k,n)$ by $[X]$, meaning the equivalence class of $X \in \mathrm{St}(k,n)$ with respect to the equivalence relation generated by (5). The columns of X can be thought of as an orthogonal basis of a k-dimensional subspace, where the action (5) describes an orthogonal change of basis.

The family $\{\mu_p\}_{p\in\mathrm{G}(k,n)}$ of parametrizations is given by

$$p = \left[Q\begin{pmatrix} I \\ 0\end{pmatrix}\right],\ \mu_p(Z) = \left[Q\exp\begin{pmatrix} 0 & Z \\ -Z^\top & 0\end{pmatrix}\begin{pmatrix} I \\ 0\end{pmatrix}\right] \quad (6)$$

where $Q \in \mathrm{O}(n)$ and $Z \in \mathbb{R}^{k\times(n-k)}$. Here we have used the fact that the action of $\mathrm{O}(n)$ on $\mathrm{St}(k,n)$ by left multiplication is transitive. Furthermore, the exponential map $\exp : \mathfrak{so}(n) \longrightarrow \mathrm{SO}(n)$ is a local diffeomorphism around 0 on the subspace

$$\left\{\begin{pmatrix} 0 & Z \\ -Z^\top & 0\end{pmatrix} \,\middle|\, Z \in \mathbb{R}^{k\times(n-k)}\right\}$$

of the Liealgebra $\mathfrak{so}(n) = \{\Omega \in \mathbb{R}^{n\times n} \,|\, \Omega^\top = -\Omega\}$ of all skew-symmetric matrices. Since left multiplication by an invertible matrix is also a local diffeomorphism and $\exp(0_n) = I_n$ it follows that μ_p is a local diffeomorphism around p with $\mathrm{D}\,\mu_p(0) = \mathrm{id}$.

The second family $\{\nu_p\}_{p\in\mathrm{G}(k,n)}$ of parametrizations we consider is given by

$$p = \left[Q\begin{pmatrix} I \\ 0\end{pmatrix}\right],\ \nu_p(Z) = \left[QQ_Z\begin{pmatrix} I \\ 0\end{pmatrix}\right] \quad (7)$$

where $Q_Z \in \mathrm{O}(n)$ comes from the Gram-Schmidt process applied to the columns of

$$\begin{pmatrix} I & 0 \\ -Z^\top & I\end{pmatrix} \in \mathbb{R}^{n\times n}.$$

Computations similar to those in [17] show that $\mathrm{D}\,\nu_p(0) = \mathrm{id}$ and hence condition (3) is satisfied.

C. The generalized Rayleigh quotient

It is well known that the global maximum of the *generalized Rayleigh quotient*

$$f : \mathrm{G}(k,n) \longrightarrow \mathbb{R},\ [X] \mapsto \mathrm{tr}(X^\top NX) \quad (8)$$

is the k-dimensional dominant eigenspace $\mathcal{V}_N$ with maximum value $f(\mathcal{V}_N) = \lambda_1 + \cdots + \lambda_k$, see e.g. [16],[18].

An easy computation shows that

$$\nabla_{f\circ\mu_p}(0) = \begin{pmatrix} 0 & -N_{12} \\ N_{12}^\top & 0\end{pmatrix} \quad (9)$$

and

$$H_{f\circ\mu_p}(0)Z = \begin{pmatrix} 0 & ZN_{22} - N_{11}Z \\ Z^\top N_{11} - N_{22}Z^\top & 0\end{pmatrix} \quad (10)$$

where

$$Q^\top NQ = \begin{pmatrix} N_{11} & N_{12} \\ N_{12}^\top & N_{22}\end{pmatrix}.$$

Hence $N_{f\circ\mu_p}(0) = -\left(H_{f\circ\mu_p}(0)\right)^{-1}\nabla_{f\circ\mu_p}(0)$ is given by the solution to the Sylvester equation

$$ZN_{22} - N_{11}Z = N_{12}\,. \quad (11)$$

D. The algorithm

```
Step 1.
    Pick $Q_0 \in \mathrm{O}(n)$, corresponding to $p_0 \in \mathrm{G}(k,n)$,
    and set $j = 0$.
Step 2.
    Compute
        $\begin{pmatrix} N_{11} & N_{12} \\ {N_{12}}^\top & N_{22}\end{pmatrix} = Q_j^\top NQ_j$
Step 3.
    Solve the Sylvester equation $ZN_{22} - N_{11}Z = N_{12}$
    for $Z$.
Step 4.
    Orthonormalize the columns of
        $\begin{pmatrix} I_k & 0 \\ -Z^\top & I_{n-k}\end{pmatrix}$
    by the Gram-Schmidt process to obtain $Q_Z$.
Step 5.
    Set $Q_{j+1} = Q_jQ_Z$, set $j = j + 1$ and goto Step 2.
```

It follows from Theorem 1 that the first k columns of Q_j converge quadratically to a basis of $\mathcal{V}_N$ provided Q_0 was chosen properly. An initial guess for Q_0 can for example be computed by applying a steepest ascent method with step size control, see [18], or by exploiting prior knowledge.

V. Conclusion

We have presented a locally quadratically convergent algorithm to compute a basis for the dominant eigenspace of a real symmetric matrix. Each iteration requires only a finite number of operations, which is in contrast to the algorithms available in the literature. This is due to our choice of the backprojection onto the manifold. We do not need to evaluate matrix exponentials nor do we compute singular value decompositions. Theorem 1 guided us in the choice of the backprojection since it spells out the degree of freedom we have while still preserving local convergence properties. Furthermore, the dimension of the Sylvester equation (11) is smaller than in the algorithms that are available in the literature. This is because we exploit the geometry of the Graßmann manifold instead of working merely in the coordinates of the embedding matrix space.

Acknowledgment

The authors would like to thank Uwe Helmke and Jonathan H. Manton for many fruitful discussions. Special thanks go to Mike Shub for sending us a hardcopy of his excellent paper [12].

References

[1] R. Adler, J.-P. Dedieu, J. Margulies, M. Martens, and M. Shub, “Newton’s method on Riemannian manifolds and a geometric model for the human spine,” *IMA J. of Numerical Analysis*, vol. 22, pp. 359–390, 2002.

[2] J. Manton, “Optimization algorithms exploiting unitary constraints,” *IEEE Transactions on Signal Processing*, vol. 50, no. 3, pp. 635–650, 2002.

[3] I. Brace, K. Hüper, and J. Manton, “The Karcher mean of points on the special orthogonal group,” *unpublished*, 2004.

[4] Y. Ma, J. Košecká, and S. Sastry, “Optimization criteria and geometric algorithms for motion and structure estimation,” *Int. J. of Computer Vision*, vol. 44, no. 3, pp. 219–249, 2001.

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

[6] 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.

[7] U. Helmke, K. Hüper, and J. Moore, “Qadratically convergent algorithms for optimal dextrous hand grasping,” *IEEE Transactions on Robotics and Automation*, vol. 18, no. 2, pp. 138–146, April 2002.

[8] S. Lang, *Fundamentals of Differential Geometry*. New York: Springer, 1999.

[9] D. G. Luenberger, *Linear and nonlinear programming*, 2nd ed. Reading: Addison-Wesley, 1984.

[10] J. Jost, *Riemannian Geometry and Geometric Analysis*, 2nd ed., ser. Universitext. Berlin: Springer, 1998.

[11] D. Gabay, “Minimizing a differentiable function over a differentiable manifold,” *J. of Optimization Theory and Applications*, vol. 37, no. 2, pp. 177–219, 1982.

[12] M. Shub, “Some remarks on dynamical systems and numerical analysis,” in *Dynamical systems and partial differential equations (Caracas, 1984)*. Caracas: Univ. Simon Bolivar, 1986, pp. 69–91.

[13] S. Smith, “Optimization techniques on Riemannian manifolds,” in *Hamiltonian and gradient flows, algorithms and control*, ser. Fields institute communications, A. Bloch, Ed. Providence: American Math. Soc., 1994, pp. 113–136.

[14] A. Edelman, T. Arias, and S. Smith, “The geometry of algorithms with orthogonality constraints,” *SIAM J. Matrix Anal. Appl.*, vol. 20, no. 2, pp. 303–353, 1998.

[15] P.-A. Absil, R. Mahony, and R. Sepulchre, “Riemannian geometry of Grassmann manifolds with a view on algorithmic computation,” *Acta Applicandae Mathematicae*, vol. 80, pp. 199–220, 2004.

[16] U. Helmke and J. Moore, *Optimization and Dynamical Systems*, ser. CCES. London: Springer, 1994.

[17] W. W. Symes, “The QR algorithm and scattering for the finite nonperiodic Toda lattice,” *Physica 4D*, pp. 275–280, 1982.

[18] R. Mahony, U. Helmke, and J. Moore, “Gradient algorithms for principal component analysis,” *J. Austr. Math. Soc. Ser. B*, vol. 37, pp. 430–450, 1996.