A Riemannian BFGS Method for Nonconvex Optimization Problems

W. Huang[1], P.-A. Absil[1] and K. A. Gallivan[2]

[1]Université catholique de Louvain

[2]Florida State University

14 September 2015

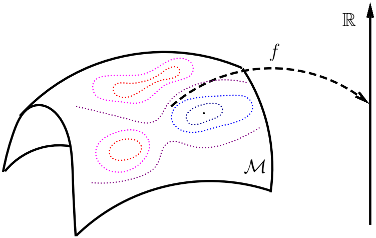
Riemannian Optimization

Constrained Problem: Given $f(x) : \mathcal{M} \to \mathbb{R}$, solve

$$\min_{x \in \mathcal{M}} f(x)$$

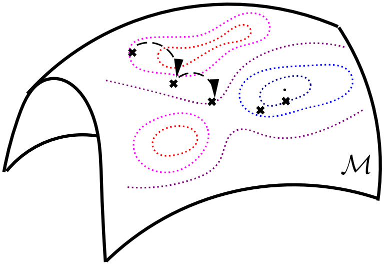
where $\mathcal{M}$ is a Riemannian manifold.

Goal Adapt unconstrained Euclidean algorithms to function on $\mathcal{M}$.

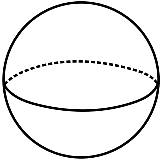


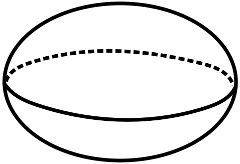
Comparison with Constrained Optimization

- All iterates on the manifold
- Convergence properties of unconstrained optimization algorithms
- No need to consider Lagrange multipliers or penalty functions
- Explore the structure of the constrained set



Examples of Manifolds

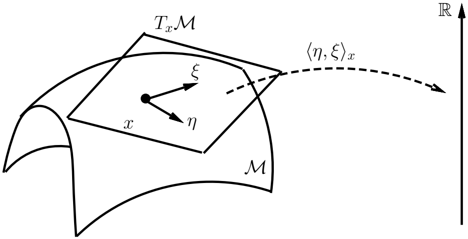




- Stiefel manifold: $\mathrm{St}(p, n) = \{X \in \mathbb{R}^{n \times p} | X^T X = I_p\}$
- Grassmann manifold: Set of all p-dimensional subspaces of $\mathbb{R}^n$
- Set of fixed rank matrices
- So on

Riemannian Manifolds

Roughly, a Riemannian manifold $\mathcal{M}$ is a smooth set with a smoothly-varying inner product on the tangent spaces.



An Application: Independent Component Analysis

- Determine a few independent components from a large number of samples
- Joint diagonalization on the Stiefel manifold [TCA09]

$$\min_{X \in \mathrm{St}(p,n)} f(X) = \min_{X \in \mathrm{St}(p,n)} -\sum_{i=1}^{N} \|\mathrm{diag}(X^T C_i X)\|_2^2,$$

where $\mathrm{St}(p,n) = \{X \in \mathbb{R}^{n \times p} | X^T X = I_p\}$, $\mathrm{diag}(M)$ denotes a vector formed by the diagonal entry of matrix M and $\| \cdot \|_2$ denotes the 2-norm

- $C_1, \ldots, C_N$ are covariance matrices.

Line search-based Methods

Consider the following generic update for a Euclidean line search-based optimization algorithm:

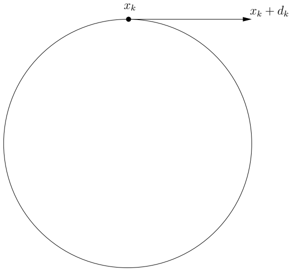
$$x_{k+1} = x_k + \alpha_k d_k \ .$$

This iteration is implemented in numerous ways, e.g.:

- Steepest descent: $d_k = -\nabla f(x_k)$
- Newton's method: $d_k = -\left[\nabla^2 f(x_k)\right]^{-1} \nabla f(x_k)$

Riemannian Manifolds Provide

- Riemannian concepts describing directions and movement on the manifold
- Riemannian analogues for gradient and Hessian

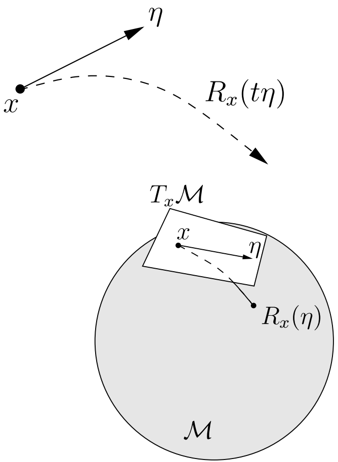


Retractions

Definition

A retraction is a mapping R from TM to M satisfying the following:

- R is continuously differentiable
- $R_x(0) = x$
- $\mathrm{D}\, R_x(0)[\eta] = \eta$



- maps tangent vectors back to the manifold
- defines curves in a direction

Euclidean	Riemannian
$x_{k+1} = x_k + \alpha_k d_k$	$x_{k+1} = R_{x_k}(\alpha_k \eta_k)$

Riemannian line search-based Methods

Riemannian Optimization Algorithm

1. At iterate $x \in M$
2. Find $\eta \in T_x M$ which satisfies certain condition.
3. Choose new iterate $x_+ = R_x(\alpha\eta)$.
4. Goto step 1.

- Riemannian steepest descent [AMS08]: $\eta = -\mathrm{grad}\, f(x)$
- Riemannian Newton [AMS08]: $\eta = -\mathrm{Hess}\, f(x)^{-1} \mathrm{grad}\, f(x)$

Riemannian Methods

In different tangent spaces

- Conjugate gradient method: $\eta_{x_{k-1}}$ and $\mathrm{grad}\, f(x_k)$
- Quasi-Newton method:
  $\mathrm{grad}\, f(x_{k-m}), \mathrm{grad}\, f(x_{k-m+1}), \ldots, \mathrm{grad}\, f(x_k)$

Vector Transport

- Vector transport: Transport a tangent vector from one tangent space to another
- $\mathcal{T}_{\eta_x}\xi_x$, denotes transport of ξ_x to tangent space of $R_x(\eta_x)$. R is a retraction associated with $\mathcal{T}$
- Isometric vector transport $\mathcal{T}_{\mathrm{S}}$ preserve the length of tangent vector

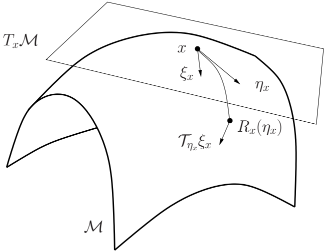


Figure: Vector transport.

BFGS methods

Results of Euclidean BFGS method:

- Converge globally to a stationary point for a strongly convex cost function
- Superlinearly converge to the non-degenerate minimizer, i.e.,

$$\lim_{k \to \infty} \frac{\|x_{k+1} - x_*\|}{\|x_k - x_*\|} = 0.$$

Riemannian BFGS methods

RBFGS method by Qi [Qi11]
specific pair of retraction and vector transport
RBFGS method by Ring and Wirth [RW12]
the expression for $\mathcal{T}_{R_{\eta_x}}$
RBFGS method by Huang et. al [HGA15]
$\mathcal{T}_{R_{\eta_x}} \eta_x$ given η_x

require

- a convex cost function
- Isometric vector transport

Vector transport by differentiated retraction

Vector transport by differentiated retraction denoted by $\mathcal{T}_R$ is defined by $\mathcal{T}_{R_{\eta_x}} \xi_x := \frac{d}{dt} R_x(\eta_x + t\xi_x)|_{t=0}$.

A New Riemannian BFGS Method

Goal

Develop a Riemannian BFGS method that does not need vector transport by differentiated retraction and convexity of the cost function.

Riemannian BFGS framework

1. At iterate $x \in M$
2. Find $\eta = -\mathcal{B}_k^{-1} \mathrm{grad}\, f(x_k) \in T_x M$, where $\mathcal{B}_k$ is updated by a formula $\mathcal{B}_{k+1} = \psi(\mathcal{B}_k, y_k, s_k)$ (See [HGA15] for details), where $y_k, s_k \in T_{x_k}\mathcal{M}$. (Euc: $y_k = \mathrm{grad}\, f(x_{k+1}) - \mathrm{grad}\, f(x_k)$, $s_k = x_{k+1} - x_k$)
3. Choose new iterate $x_+ = R_x(\alpha\eta)$, where α is a step size satisfying certain condition.
4. Goto step 1.

A New Riemannian BFGS Method

- Why is $\mathcal{T}_R$ used?

  Line search: $h(t) = f(R_x(t\eta_x))$, $h'(t) = \langle \mathrm{grad}\, f(R_x(t\eta_x)), \mathcal{T}_{R_{t\eta_x}} \eta_x \rangle$

- Line search: [BN89, (3.2), (3.3)] require the step size α_k satisfy either

$$h_k(\alpha_k) - h_k(0) \leq -\chi_1 \frac{h_k'(0)^2}{\|\eta_k\|^2} \quad (1)$$

or

$$h_k(\alpha_k) - h_k(0) \leq \chi_2 h_k'(0), \quad (2)$$

where $h_k(t) = f(R_{x_k}(t\eta_k))$, χ_1 and χ_2 are positive constants.

A New Riemannian BFGS Method: Avoid $\mathcal{T}_R$

- If the gradient of f is Lipschitz continuous [AMS08, Definition 7.4.1], then above line search condition is implied by, e.g.,
  - The Goldstein condition
  - The Wolfe condition
  - The Armijo-Goldstein condition

A New Riemannian BFGS Method: For nonconvex problem

- Is a modification on Riemannian BFGS method required for nonconvex problem?

  Yes, an example is given in [Dai13], which shows that BFGS method fails to converge for a nonconvex cost function.

- How to make Riemannian BFGS method work for nonconvex problem?

  Multiple ideas exist for Euclidean BFGS, e.g., [LF01a, LF01b].

A New Riemannian BFGS Method: For nonconvex problem

- For BFGS method, a Riemannian version of [Pow76, Lemma 2] holds, i.e.,

$$\frac{\|y_k\|^2}{\langle y_k, s_k \rangle} \text{ is bounded above} \Rightarrow \liminf_{k \to \infty} \|\mathrm{grad}\, f(x_k)\| = 0.$$

- For nonconvex problem, $\frac{\|y_k\|^2}{\langle y_k, s_k \rangle}$ is not bounded above in general.
- Cautious BFGS update: [LF01b]

$$\mathcal{B}_{k+1} = \begin{cases} \psi(\mathcal{B}_k, y_k, s_k), & \text{if } \frac{\langle y_k, s_k \rangle}{\|s_k\|^2} \geq \vartheta(\|\mathrm{grad}\, f(x_k)\|) \\ \mathcal{T}_{S_{s_k}} \circ \mathcal{B}_k \circ \mathcal{T}_{S_{s_k}}^{-1} \text{ or } \mathrm{id}, & \text{otherwise}, \end{cases} \tag{3}$$

where ϑ is a monotone increasing function satisfying $\vartheta(0) = 0$ and ϑ is strictly increasing at 0.

A New Riemannian BFGS Method: Theoretical Results

Global convergence

Under some reasonable assumptions, the iterates $\{x_k\}$ generated by the RBFGS method satisfies

$$\liminf_{k \to \infty} \|\operatorname{grad} f(x_k)\| = 0.$$

Superlinear convergence

Under some reasonable assumptions, the cautious BFGS update reduces to ordinary BFGS update around a non-degenerate minimizer and the superlinear convergence holds, i.e.,

$$\lim_{k \to \infty} \frac{\operatorname{dist}(x_{k+1}, x_*)}{\operatorname{dist}(x_k, x_*)} = 0.$$

A New Riemannian BFGS Method

Summary:

- Vector transport by differentiated retraction is not required
- Global convergence is guaranteed for nonconvex cost function
- Superlinear convergence is guaranteed

Limited memory version of this Riemannian BFGS (LRBFGS) can be obtained.

Experiments

Problem: Joint diagonalization on the Stiefel manifold [TCA09]

$$\min_{X \in \mathrm{St}(p,n)} -\sum_{i=1}^{N} \|\mathrm{diag}(X^T C_i X)\|_2^2.$$

- Riemannian BFGS with the Armijo-Goldstein condition

$$h_k(\alpha_k) \leq h_k(0) + \sigma \alpha_k h_k'(0),$$

  where α_k is the largest value in the set $\{1, \varrho, \varrho^2, \varrho^3, \ldots\}$, $0 < \varrho < 1$ and $0 < \sigma < 0.5$.

  - An isometric vector transport is sufficient.
  - isometric vector transport [HGA15, Section 4.1]
  - The implementation of this vector transport is identity

Experiments

- Riemannian BFGS with the Wolfe condition [HGA15]

$$h_k(\alpha_k) \leq h_k(0) + c_1 \alpha_k h_k'(0)$$
$$h_k'(\alpha_k) \geq c_2 h_k'(0)$$

where $0 < c_1 < 0.5 < c_2 < 1$.

  - The isometric vector transport is required to satisfies $\mathcal{T}_{S_\xi} \xi = \beta \mathcal{T}_{R_\xi} \xi, \quad \beta = \frac{\|\xi\|}{\|\mathcal{T}_{R_\xi} \xi\|}, \forall \xi \in T_x \mathcal{M}$
  - isometric vector transport [HGA15, Sections 2.3.1 and 5]
  - The implementation of this vector transport is a rank-two update

Parameters and Settings

$$\min_{X \in \mathrm{St}(p,n)} -\sum_{i=1}^{N} \|\mathrm{diag}(X^T C_i X)\|_2^2,$$
$$\mathrm{St}(p,n) = \{X \in \mathbb{R}^{n \times p} | X^T X = I_p\}.$$

- $p = 8$, and $n = 12$
- Stopping criterion: $\|\mathrm{grad}\, f(x_k)\| / \|\mathrm{grad}\, f(x_0)\| < 10^{-6}$
- C++ with compiler g++-4.7 on 64 bit Ubuntu platform with 3.6GHz CPU

Results

Table: An average of 1000 random runs of RBFGS. LS denotes line search condition. VT denotes vector transport

N	LS	VT	iter	nf	ng	nV	t (millisecond)
32	Armijo	identity	183	198	184	365	9.89
	Armijo	rank-2	133	150	134	266	12.3
	Wolfe	rank-2	129	146	132	390	13.1
128	Armijo	identity	190	207	191	380	25.7
	Armijo	rank-2	141	159	142	281	25.8
	Wolfe	rank-2	137	156	141	414	25.8
512	Armijo	identity	196	216	197	393	90.9
	Armijo	rank-2	148	167	149	295	77.7
	Wolfe	rank-2	143	164	148	432	75.8

Results

Table: An average of 1000 random runs of RBFGS. LS denotes line search condition. VT denotes vector transport

N	LS	VT	iter	nf	ng	nV	t (millisecond)
	Armijo	identity	183	198	184	365	9.89
32	Armijo	rank-2	133	150	134	266	12.3
	Wolfe	rank-2	129	146	132	390	13.1
	Armijo	identity	190	207	191	380	25.7
128	Armijo	rank-2	141	159	142	281	25.8
	Wolfe	rank-2	137	156	141	414	25.8
	Armijo	identity	196	216	197	393	90.9
512	Armijo	rank-2	148	167	149	295	77.7
	Wolfe	rank-2	143	164	148	432	75.8

The relative efficiency depends on the relative cost on vector transport and cost function evluation

Conclusion

- Combine the line search in [BN89] and the cautious BFGS update in [LF01b] and generalize them to the Riemannian setting.
- Global convergence for a nonconvex cost function
- Superlinear convergence rate
- Vector transport by differentiated retraction not required
- Code: `www.math.fsu.edu/~whuang2/ROPTLIB`

References I

P.-A. Absil, R. Mahony, and R. Sepulchre.
Optimization algorithms on matrix manifolds.
Princeton University Press, Princeton, NJ, 2008.

R. H. Byrd and J. Nocedal.
A tool for the analysis of quasi-newton methods with application to unconstrained minimization.
SIAM Journal on Numerical Analysis, 26(3):727–739, 1989.

Y.-H. Dai.
A perfect example for the BFGS method.
Mathematical Programming, 138(1-2):501–530, March 2013.
doi:10.1007/s10107-012-0522-2.

Wen Huang, K. A. Gallivan, and P.-A. Absil.
A Broyden Class of Quasi-Newton Methods for Riemannian Optimization.
SIAM Journal on Optimization, 25(3):1660–1685, 2015.

D.-H. Li and M. Fukushima.
A modified BFGS method and its global convergence in nonconvex minimization.
Journal of Computational and Applied Mathematics, 129:15–35, 2001.

D.-H. Li and M. Fukushima.
On the global convergence of the BFGS method for nonconvex unconstrained optimization problems.
SIAM Journal on Optimization, 11(4):1054–1064, January 2001.
doi:10.1137/S1052623499354242.

References II

M. J. D. Powell.
Some global convergence properties of a variable metric algorithm for minimization without exact line searches.
Nonlinear Programming, SIAM-AMS Proceedings, 9, 1976.

C. Qi.
Numerical optimization methods on Riemannian manifolds.
PhD thesis, Florida State University, Department of Mathematics, 2011.

W. Ring and B. Wirth.
Optimization methods on Riemannian manifolds and their application to shape space.
SIAM Journal on Optimization, 22(2):596–627, January 2012.
doi:10.1137/11082885X.

F. J. Theis, T. P. Cason, and P.-A. Absil.
Soft dimension reduction for ICA by joint diagonalization on the Stiefel manifold.
Proceedings of the 8th International Conference on Independent Component Analysis and Signal Separation, 5441:354–361, 2009.