A Riemannian BFGS Method for Nonconvex Optimization Problems*

Wen Huang[1], P.-A. Absil[1], and Kyle A. Gallivan[2]

[1] Department of Mathematical Engineering, ICTEAM Institute, Université catholique de Louvain, Belgium
[2] Department of Mathematics, Florida State University, FL, USA

Abstract. In this paper, a Riemannian BFGS method is defined for minimizing a smooth function on a Riemannian manifold endowed with a retraction and a vector transport. The method is based on a Riemannian generalization of a cautious update and a weak line search condition. It is shown that the Riemannian BFGS method converges (i) globally to a stationary point without assuming that the objective function is convex and (ii) superlinearly to a nondegenerate minimizer. The weak line search condition removes completely the need to consider the differentiated retraction. The joint diagonalization problem is used to demonstrate the performance of the algorithm with various parameters, line search conditions, and pairs of retraction and vector transport.

1 Introduction

In the Euclidean setting, the BFGS method is widely viewed as the best quasi-Newton method for solving smooth unconstrained optimization problems [DS83,NW06]. Its global and superlinear local convergence is well understood for convex problems (see [DS83] and references therein). However, for nonconvex problems, its convergence properties are more intricate. Recently, Dai [Dai13] has produced a nonconvex cost function for which the standard BFGS method does not converge. Modified BFGS methods exist that converge globally to critical points of nonconvex cost functions [LF01a,LF01b].

Many Riemannian versions of the BFGS method have appeared, e.g., [Gab82,SL10,RW12,SKH13,HGA15], but complete global and local convergence analyses that are not restricted to a specific cost function or a manifold are only given in two of them [RW12,HGA15]. The analyses of both methods require the cost function to satisfy a Riemannian version of convexity for global and superlinear local convergence.

In this paper, we generalize to manifolds the approach in [LF01b] for nonconvex problems by using a Riemannian version of the cautious update

* This paper presents research results of the Belgian Network DYSCO (Dynamical Systems, Control, and Optimization), funded by the Interuniversity Attraction Poles Programme initiated by the Belgian Science Policy Office. This work was supported by grant FNRS PDR T.0173.13.

of the Hessian approximation, and additionally a weak line search condition [BN89, (3.2), (3.3)]. Global and local superlinear convergence results are stated and the joint diagonalization problem [TCA09] is used as an example to demonstrate numerical performance.

A key advantage of the proposed method over those in [RW12,HGA15] is that it offers more leeway on the choice of the vector transport. The version in [RW12] requires vector transport by differentiated retraction, which may not be available to users or may be too expensive. The version in [HGA15] requires only the action of the differentiated retraction along a particular direction. In fact, any method that uses the Riemannian second Wolfe condition will require at least the action of the differentiated retraction along some particular direction. The proposed method is even less demanding: it no longer requires the second Wolfe condition, and the differentiated retraction can be completely avoided.

This paper is organized as follows. Section 2 presents notation used in this paper. Section 3 defines the Riemannian version of BFGS. Global and local convergence results are stated in Section 4. Numerical experiments are reported in Section 5.

2 Notation

The underlying concepts of Riemannian geometry can be found, e.g., in [Boo86,AMS08]. We follow the notation of [AMS08]. Let f denote a cost function defined on a d-dimensional Riemannian manifold $\mathcal{M}$ with the Riemannian metric $g : (\eta_x, \xi_x) \mapsto g_x(\eta_x, \xi_x) \in \mathbb{R}$. $\mathrm{T}_x\,\mathcal{M}$ denotes the tangent space of $\mathcal{M}$ at x and $\mathrm{T}\,\mathcal{M}$ denotes the tangent bundle, i.e., the set of all tangent spaces. For any $\eta_x \in \mathrm{T}_x\,\mathcal{M}$, $\eta_x^\flat$ denotes the function such that $\eta_x^\flat : \mathrm{T}_x\,\mathcal{M} \to \mathbb{R} : \xi_x \mapsto g_x(\eta_x, \xi_x)$.

A retraction is a C^1 map $R : \mathrm{T}\,\mathcal{M} \to \mathcal{M}$ such that $R(0_x) = x$ for all $x \in \mathcal{M}$ and $\frac{d}{dt}R(t\xi_x)|_{t=0} = \xi_x$ for all $\xi_x \in \mathrm{T}_x\,\mathcal{M}$. The domain of R does not have to be the entire tangent bundle, however, it is usually the case in practice. In this paper, we assume that R is well-defined whenever needed. R_x denotes the restriction of R to $\mathrm{T}_x\,\mathcal{M}$. A vector transport $\mathcal{T} : \mathrm{T}\,\mathcal{M} \oplus \mathrm{T}\,\mathcal{M} \to \mathrm{T}\,\mathcal{M}, (\eta_x, \xi_x) \mapsto \mathcal{T}_{\eta_x}\xi_x$ with associated retraction R is a mapping[1] such that, for all (x, η_x) in the domain of R and all $\xi_x, \zeta_x \in \mathrm{T}_x\,\mathcal{M}$, it holds that (i) $\mathcal{T}_{\eta_x}\xi_x \in \mathrm{T}_{R(\eta_x)}\,\mathcal{M}$, (ii) $\mathcal{T}_{\eta_x}$ is a linear map. An isometric vector transport $\mathcal{T}_\mathrm{S}$ additionally satisfies $g_{R_x(\eta_x)}(\mathcal{T}_{\mathrm{S}_{\eta_x}}\xi_x, \mathcal{T}_{\mathrm{S}_{\eta_x}}\zeta_x) = g_x(\xi_x, \zeta_x)$. The vector transport by differentiated retraction $\mathcal{T}_R$ is defined to be $\mathcal{T}_{R_{\eta_x}}\xi_x := \frac{d}{dt}R_x(\eta_x + t\xi_x)|_{t=0}$.

3 Riemannian BFGS Method with Cautious Update

[1] This mapping is not even required to be continuous in the definition. The smoothness is imposed in the convergence analyses.

The proposed Riemannian generalization of the BFGS method with cautious update is stated in Algorithm 1.

Algorithm 1 Cautious RBFGS method

Input: Riemannian manifold $\mathcal{M}$ with Riemannian metric g; a retraction R; isometric vector transport $\mathcal{T}_{\mathrm{S}}$, with R as the associated retraction; continuously differentiable real-valued function f on $\mathcal{M}$, bounded below; initial iterate $x_0 \in \mathcal{M}$; initial Hessian approximation $\mathcal{B}_0$ that is symmetric positive definite with respect to the metric g; convergence tolerance $\varepsilon > 0$; constants $\chi_1 > 0$ and $\chi_2 > 0$ in the line search condition;

1: $k \leftarrow 0$;
2: **while** $\|\operatorname{grad} f(x_k)\| > \varepsilon$ **do**
3: Obtain $\eta_k \in \mathrm{T}_{x_k}\mathcal{M}$ by solving $\mathcal{B}_k\eta_k = -\operatorname{grad} f(x_k)$;
4: Set $x_{k+1} = R_{x_k}(\alpha_k\eta_k)$, where $\alpha_k > 0$ is computed from a line search procedure to satisfy either

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

or

$$h_k(\alpha_k) - h_k(0) \leq \chi_2 h_k'(0), \tag{2}$$

where $h_k(t) = f(R_{x_k}(t\eta_k))$.
5: Define the linear operator $\mathcal{B}_{k+1} : \mathrm{T}_{x_{k+1}}\mathcal{M} \to \mathrm{T}_{x_{k+1}}\mathcal{M}$ by (4);
6: $k \leftarrow k + 1$;
7: **end while**

When $\mathcal{M}$ is a Euclidean space, the line search condition in Step 4 of Algorithm 1 is weak since it has been shown in [BN89, Sections 3 and 4] that many line search conditions, e.g., the Curry-Altman condition, the Goldstein condition, the Wolfe condition and the Armijo-Goldstein condition, imply either (1) or (2) if the gradient of the function is Lipschitz continuous. In the Riemannian setting, note that the function $f \circ R_x : \mathrm{T}_x\mathcal{M} \to \mathbb{R}$ is defined on a linear space. It follows that the Euclidean results about line search are applicable, i.e., the above conditions also imply either (1) or (2) when the gradient of the function satisfies the Riemannian Lipschitz continuous condition [AMS08, Definition 7.4.1].

Among several possible Riemannian generalizations of the BFGS update formula [RW12,SKH13,HGA15], we opt here for

$$\mathcal{B}_{k+1} = \tilde{\mathcal{B}}_k - \frac{\tilde{\mathcal{B}}_k s_k (\tilde{\mathcal{B}}_k^* s_k)^\flat}{(\tilde{\mathcal{B}}_k^* s_k)^\flat s_k} + \frac{y_k y_k^\flat}{y_k^\flat s_k}, \tag{3}$$

where $\tilde{\mathcal{B}}_k = \mathcal{T}_{\mathrm{S}_{\alpha_k\eta_k}} \circ \mathcal{B}_k \circ \mathcal{T}^{-1}_{\mathrm{S}_{\alpha_k\eta_k}}$, $y_k = \beta_k^{-1} \operatorname{grad} f(x_{k+1}) - \mathcal{T}_{\mathrm{S}_{\alpha_k\eta_k}} \operatorname{grad} f(x_k)$, $s_k = \mathcal{T}_{\mathrm{S}_{\alpha_k\eta_k}} \alpha_k\eta_k$, and β_k is an arbitrary number satisfying $|\beta_k - 1| \leq L_\beta \|\alpha_k\eta_k\|$ and $L_\beta > 0$ is a constant. The motivation for introducing β_k

is to make this update subsume the update in [HGA15], which uses $\beta_k = \frac{\|\alpha_k \eta_k\|}{\|\mathcal{T}_{R_{\alpha_k \eta_k}} \alpha_k \eta_k\|}$.

If $y_k^\flat s_k > 0$, then the symmetric positive definiteness of $\tilde{\mathcal{B}}_k$ implies the symmetric positive definiteness of $\mathcal{B}_{k+1}$ [HGA15]. The positive definiteness of the sequence $\{\mathcal{B}_k\}$ is important in the sense that it guarantees that the search direction is a descent direction. However, not all line search conditions imply $y_k^\flat s_k > 0$. In the existing papers [RW12,HGA15], the Wolfe condition with information about $\mathcal{T}_R$ is used to guarantee $y_k^\flat s_k > 0$. In this paper, instead of enforcing $y_k^\flat s_k > 0$ by the Wolfe condition, we guarantee symmetric positive definiteness of $\mathcal{B}_{k+1}$ by resorting to the following cautious update formula

$$\mathcal{B}_{k+1} = \begin{cases} \tilde{\mathcal{B}}_k - \frac{\tilde{\mathcal{B}}_k s_k (\tilde{\mathcal{B}}_k^* s_k)^\flat}{(\tilde{\mathcal{B}}_k^* s_k)^\flat s_k} + \frac{y_k y_k^\flat}{y_k^\flat s_k}, \text{ if } \frac{y_k^\flat s_k}{\|s_k\|^2} \geq \vartheta(\|\operatorname{grad} f(x_k)\|) \\ \tilde{\mathcal{B}}_k, \qquad\qquad \text{otherwise,} \end{cases} \tag{4}$$

where ϑ is a monotone increasing function satisfying $\vartheta(0) = 0$ and ϑ strictly increasing at 0. Formula (4) reduces to the cautious update formula of [LF01b] when $\mathcal{M}$ is a Euclidean space. Using update (4) does not require the Wolfe condition, which yields more leeway for choosing a line search condition. When $\frac{y_k^\flat s_k}{\|s_k\|^2} \ngeq \vartheta(\|\operatorname{grad} f(x_k)\|)$, $\mathcal{B}_{k+1}$ can be set to be any given constant matrix, e.g., id, rather than $\tilde{\mathcal{B}}_k$. The choice does not affect the theoretical results given later.

4 Convergence Analysis

Due to length limitations, we only state the convergence results without proofs. The proofs will be given in a forthcoming paper. Theorems 1 and 2 state the global and local convergence results respectively.

Theorem 1. *Let $\{x_k\}$ be a sequence generated by Algorithm 1. Assume that the level set $\Omega = \{x \in \mathcal{M} \mid f(x) \leq f(x_0)\}$ is compact, that there exists $L_1 > 0$ such that $\|\mathcal{T}_\eta \operatorname{grad} f(x) - \operatorname{grad} f(R_x(\eta))\| \leq L_1 \|\eta\|$ for all $x \in \mathcal{U}$ and η such that $R_x(\eta) \in \mathcal{U}$, and that the function $\hat{f} = f \circ R$ is radially L-C^1 function [AMS08, Definition 7.4.1] for all $x \in \Omega$. Then $\liminf_{k\to\infty} \|\operatorname{grad} f(x_k)\| = 0$.*

Theorem 2. *Let $\{x_k\}$ be a sequence generated by Algorithm 1 that converges to a nondegenerate minimizer x^* of f. Suppose there exists a neighborhood $\tilde{\Omega}$ of x^* such that*

1. *the objective function f is twice continuously differentiable in $\tilde{\Omega}$ and there exists positive constants a_{10} and a_{11} such that for all $y \in \tilde{\Omega}$, $\|\operatorname{Hess} f(y) - \mathcal{T}_{S_\eta} \operatorname{Hess} f(x^*) \mathcal{T}_{S_\eta}^{-1}\| \leq a_{10}\|\eta\|$, where $\eta = R_{x^*}^{-1} y$;*
2. *the retraction R is twice continuously differentiable in $\tilde{\Omega}$ and there is a constant a_5 such that for all $x, y \in \tilde{\Omega}$, $\max_{t\in[0,1]} \operatorname{dist}(R_x(t\eta), x^*) \leq a_9 \max(\operatorname{dist}(x, x^*), \operatorname{dist}(y, x^*))$, where $\eta = R_x^{-1} y$;*

3. *the isometric vector transport $\mathcal{T}_S$ with associated retraction R is continuous and satisfies $\mathcal{T}_{0_x}\xi_x = \xi_x$ for all $\xi_x \in \mathrm{T}_x\mathcal{M}$, $\|\mathcal{T}_{\mathrm{S}_\eta} - \mathcal{T}_{R_\eta}\| \leq \tilde{L}\|\eta\|$ and $\|\mathcal{T}_{\mathrm{S}_\eta}^{-1} - \mathcal{T}_{R_\eta}^{-1}\| \leq \tilde{L}\|\eta\|$ for some constant $\tilde{L}$.*

Then there exists an index k_0 such that $\alpha_k = 1$ satisfies either (1) *or* (2) *for $k \geq k_0$. Moreover, if $\alpha_k = 1$ is used for all $k \geq k_0$, then x_k converges to x^* superlinearly, i.e., $\lim_{k\to\infty} \frac{\mathrm{dist}(x_{k+1},x^*)}{\mathrm{dist}(x_k,x^*)} = 0$.*

It is shown in [Hua13, Theorem 5.2.4] that $\alpha_k = 1$ eventually satisfies the two frequently used line search conditions, i.e., the Wolfe condition

$$h_k(\alpha_k) \leq h_k(0) + c_1\alpha_k h_k'(0) \tag{5}$$
$$h_k'(\alpha_k) \geq c_2 h_k'(0) \tag{6}$$

where $0 < c_1 < 0.5 < c_2 < 1$ and the Armijo-Goldstein condition

$$h_k(\alpha_k) \leq h_k(0) + \sigma\alpha_k h_k'(0), \tag{7}$$

where α_k is the largest value in the set $\{t^{(i)} | t^{(i)} \in [\varrho_1 t^{(i-1)}, \varrho_2 t^{(i-1)}], t^{(0)} = 1\}$, $0 < \varrho_1 < \varrho_2 < 1$ and $0 < \sigma < 0.5$. Therefore, if $\alpha_k = 1$ is attempted first using one of the line search conditions, then the superlinear convergence of Algorithm 1 is obtained. At present, no conditions on χ_1 and χ_2 in (1) and (2) that guarantee a similar result are known.

If $h'(t)$ must be evaluated at $t \neq 0$ in line search conditions, such as the Wolfe condition, then the action of vector transport by differentiated retraction is required only in a particular direction. More specifically, the term

$$h'(t) = g_{R_{x_k}(t\eta_k)}(\mathrm{grad}\, f(R_{x_k}(t\eta_k)), \mathcal{T}_{R_{t\eta_k}}\eta_k)$$

requires the action of vector transport by differentiated retraction, $\mathcal{T}_{R_\eta}\xi$, with η and ξ on a same direction. This is discussed in [HGA15] and one approach to resort to as little information on the differentiated retraction as possible is also proposed. If $h'(t)$ is not required at $t \neq 0$, such as in the Armijo-Goldstein condition, then the differentiated retraction can be completely avoided since $\mathcal{T}_{R_{0\eta_k}}\eta_k = \eta_k$.

5 Experiments

In this section, we investigate numerically the impact of choosing the Wolfe versus the Armijo-Goldstein condition in Step 4 of on Algorithms 1.

5.1 Problem, Retraction, Vector Transport and Step Size

The joint diagonalization (JD) problem on the Stiefel manifold [TCA09] is used to illustrate the numerical performance:

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

where $\mathrm{St}(p,n) = \{X \in \mathbb{R}^{n\times p} | X^T X = I_p\}$, matrices $C_1, \ldots, C_N$ are given symmetric matrices, $\mathrm{diag}(M)$ denotes the vector formed by the diagonal entries of matrix M, and $\|\cdot\|_2$ denotes the 2-norm.

The Stiefel manifold $\mathrm{St}(p,n)$ can be viewed as a submanifold of $\mathbb{R}^{n\times p}$. The chosen Riemannian metric g on $\mathrm{St}(p,n)$ is the metric endowed from its embedding space, i.e., $g(\eta_X, \xi_X) = \mathrm{tr}(\eta_X^T \xi_X)$. With this Riemannian metric g, the gradient is given in [TCA09, Section 2.3]. As discussed in [HAG15, Section 2.2], a tangent vector $\eta_X \in \mathrm{T}_X \mathcal{M}$ can be represented by a vector in the embedding space $\mathbb{R}^{n\times p}$ or a d-dimensional coefficient vector of a basis of $\mathrm{T}_X \mathcal{M}$, where $d = np - p(p+1)/2$ is the dimension of $\mathrm{St}(p,n)$. In our experiments, we use a d-dimensional representation of tangent vectors. By varying the basis and fixing the coefficients, one can define the vector transport by parallelization [HAG15, Section 2.3.1 and 5]. The implementation of vector transport is then simply an identity [Hua13, Section 9.5].

The retraction is chosen to be qf retraction [AMS08, (4.7)]

$$R_X(\eta_X) = \mathrm{qf}(X + \eta_X), \tag{8}$$

where qf denotes the Q factor of the QR decomposition with nonnegative elements on the diagonal of R.

5.2 Tests and Results

The C_i matrices are selected as $C_i = R_i + R_i^T$, where the elements of $R_i \in \mathbb{R}^{n\times n}$ are independently drawn from the standard normal distribution. The initial iterate X_0 is given by applying Matlab's function *orth* to a matrix whose elements are drawn from the standard normal distribution using Matlab's *randn*. The code can be found in `http://www.math.fsu.edu/~whuang2/papers/ARBMNOP.htm`.

Let RBFGS-W and RBFGS-A denote Algorithm 1 with the Wolfe condition and the Armijo-Goldstein condition respectively. Since the Wolfe condition requires the evaluation of $h'(t)$ at $t \neq 0$, we use the locking condition proposed in [HGA15], which restricts the retraction R and the isometric vector transport $\mathcal{T}_{\mathrm{S}}$:

$$\mathcal{T}_{\mathrm{S}_\xi} \xi = \beta \mathcal{T}_{R_\xi} \xi, \quad \beta = \frac{\|\xi\|}{\|\mathcal{T}_{R_\xi} \xi\|}. \tag{9}$$

Let RV1 denote retraction (8) and the vector transport by parallelization, which does not satisfy the locking condition (9); RV2 denote retraction (8) and the vector transport using the approach of [HGA15, Section 4.2], which does satisfy the locking condition (9) but the vector transport is not smooth and relatively expensive.

The experimental results with various parameters and algorithms are reported in Table 1. Note that there is no result for RBFGS-W with RV1

Table 1. An average of 1000 random runs of RBFGS. $n = 12$, $p = 8$, $c_1 = \sigma = 10^{-4}$. The subscript $-k$ indicates a scale of 10^{-k}. $iter$, nf, ng, nV and t denote the number of iterations, number of function evaluations, number of gradient evaluations, number of vector transport and computational time (millisecond) respectively.

	N		Armijo-Goldstien: $[\varrho_1, \varrho_2]$				Wolfe: c_2			
			$[\frac{1}{2}, \frac{1}{2}]$	$[\frac{1}{4}, \frac{3}{4}]$	$[\frac{1}{16}, \frac{15}{16}]$	$[\frac{1}{64}, \frac{63}{64}]$	$\frac{1}{2}$	$\frac{3}{4}$	$\frac{15}{16}$	$\frac{63}{64}$
RV1	128	$iter$	2.39_2	1.94_2	1.91_2	1.91_2	\	\	\	\
		nf	3.06_2	2.13_2	2.06_2	2.06_2	\	\	\	\
		ng	2.40_2	1.95_2	1.92_2	1.92_2	\	\	\	\
		nV	4.77_2	3.89_2	3.81_2	3.81_2	\	\	\	\
		t	3.18_{-2}	2.63_{-2}	2.58_{-2}	2.61_{-2}	\	\	\	\
	512	$iter$	1.96_2	1.91_2	1.91_2	1.91_2	\	\	\	\
		nf	2.15_2	2.08_2	2.08_2	2.07_2	\	\	\	\
		ng	1.97_2	1.92_2	1.92_2	1.92_2	\	\	\	\
		nV	3.92_2	3.82_2	3.83_2	3.83_2	\	\	\	\
		t	9.32_{-2}	8.96_{-2}	8.88_{-2}	8.92_{-2}	\	\	\	\
RV2	128	$iter$	1.46_2	1.64_2	1.67_2	1.47_2	1.23_2	1.32_2	1.36_2	1.42_2
		nf	1.70_2	1.97_2	2.03_2	1.68_2	1.86_2	1.84_2	1.68_2	1.65_2
		ng	1.47_2	1.65_2	1.68_2	1.48_2	1.67_2	1.62_2	1.50_2	1.47_2
		nV	2.93_2	3.27_2	3.35_2	2.94_2	4.13_2	4.22_2	4.20_2	4.26_2
		t	2.64_{-2}	2.89_{-2}	2.94_{-2}	2.64_{-2}	2.80_{-2}	2.76_{-2}	2.60_{-2}	2.56_{-2}
	512	$iter$	1.49_2	1.49_2	1.53_2	1.48_2	1.31_2	1.38_2	1.40_2	1.51_2
		nf	1.69_2	1.69_2	1.75_2	1.66_2	1.97_2	1.89_2	1.71_2	1.80_2
		ng	1.50_2	1.50_2	1.54_2	1.49_2	1.76_2	1.66_2	1.53_2	1.56_2
		nV	2.98_2	2.99_2	3.05_2	2.96_2	4.34_2	4.36_2	4.31_2	4.49_2
		t	7.82_{-2}	7.76_{-2}	7.89_{-2}	7.75_{-2}	8.92_{-2}	8.47_{-2}	7.91_{-2}	8.00_{-2}

since the well-definedness of RBFGS-W requires the locking condition. It can be seen that the performances of the Armijo-Goldstein condition and the Wolfe condition with the chosen algorithms are similar.

RBFGS with RV1 performs worse than RBFGS with RV2 in the sense of number of function and gradient evaluations. This implies that the locking condition, to some extent, reduces the number of function and gradient evaluations in RBFGS with either the Armijo-Goldstein condition or the Wolfe condition. Note that even though $h'(t)$ at $t \neq 0$ is not used in the Armijo-Goldstein line search condition, the locking condition can still reduce the number of function and gradient evaluations. However, due to the low complexities on vector transport, RBFGS-A with RV1 still have competitive performance in the sense of computational time.

6 Conclusion

The results demonstrate the global convergence expected in the algorithm. While the locking condition is no longer required, we see that using it reduces the number of function and gradient evaluations. For problems such as joint diagonalization with large enough N so those evaluations are dominated computationally, a reduction in overall time results.

References

AMS08. P.-A. Absil, R. Mahony, and R. Sepulchre. *Optimization algorithms on matrix manifolds*. Princeton University Press, Princeton, NJ, 2008.

BN89. 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.

Boo86. W. M. Boothby. *An introduction to differentiable manifolds and Riemannian geometry*. Academic Press, second edition, 1986.

Dai13. 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.

DS83. J. E. Dennis and R. B. Schnabel. *Numerical methods for unconstrained optimization and nonlinear equations*. Springer, New Jersey, 1983.

Gab82. D Gabay. Minimizing a differentiable function over a differential manifold. *Journal of Optimization Theory and Applications*, 37(2):177–219, 1982.

HAG15. W. Huang, P.-A. Absil, and K. A. Gallivan. A Riemannian symmetric rank-one trust-region method. *Mathematical Programming*, 150(2):179–216, February 2015.

HGA15. 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.

Hua13. W. Huang. *Optimization algorithms on Riemannian manifolds with applications*. PhD thesis, Florida State University, Department of Mathematics, 2013.

LF01a. 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.

LF01b. 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.

NW06. J. Nocedal and S. J. Wright. *Numerical optimization*. Springer, second edition, 2006.

RW12. 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.

SKH13. M. Seibert, M. Kleinsteuber, and K. Hüper. Properties of the BFGS method on Riemannian manifolds. *Mathematical System Theory - Festschrift in Honor of Uwe Helmke on the Occasion of his Sixtieth Birthday*, pages 395–412, 2013.

SL10. B. Savas and L. H. Lim. Quasi-Newton methods on Grassmannians and multilinear approximations of tensors. *SIAM Journal on Scientific Computing*, 32(6):3352–3393, 2010.

TCA09. 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.