2005 International Symposium on Nonlinear
Theory and its Applications (NOLTA2005)
Bruges, Belgium, October 18-21, 2005

A Note on Riemannian Optimization Methods on the Stiefel and the Grassmann Manifolds

Yasunori Nishimori

National Institute of Advanced Industrial Science and Technology (AIST),
AIST Tsukuba Cental 2, 1-1-1, Umezono, Tsukuba, Ibaraki 305-8568, Japan
Email: y.nishimori@aist.go.jp

Abstract—We prove the updating rules given by the Riemannian optimization methods on the Stiefel and the Grassmann manifolds coincide if the target function for optimization on the Stiefel manifold has a symmetry so that it is regarded as a function on the Grassmann manifold. The Grassmann condition is encapsulated in this symmetry. Therefore we do not need the formulas for the Grassmann manifold separately; all of them, the natural gradient method, the conjugate gradient method, and the Newton method reduce to the counterparts for the Stiefel manifold.

1. Introduction

Riemannian optimization methods on the Stiefel and the Grassmann manifolds based on geodesics attracted attention recently and have been used among several research communities such as neural networks [7],[8],[9], pattern recognition [5], computer vision [6], numerical analysis [1],[3], and so on. Although most authors concentrate on either manifold, the Stiefel manifold [7],[8],[9], or the Grassmann manifold [1],[5], in their seminal paper [3], Edelman-Arias-Smith developed formulas for both manifolds. The expressions of their formulas for the Grassmann manifold are, however, very different from the counterparts for the Stiefel manifold. The main aim of this paper is to illustrate the updating rules given by the Riemannian optimization methods on the Stiefel and the Grassmann manifolds actually coincide if the target function for optimization on the Stiefel manifold has a symmetry so that it is regarded as a function on the Grassmann manifold. The fact that the Stiefel manifold is a principal bundle over the Grassmann manifold is exploited. As far as we know, this result has not been stated in the previous literature.

1.1. Riemannian Optimization Method

In this paper we are concerned about the two manifolds: the Stiefel manifold and the Grassmann manifold. The Stiefel manifold is described by orthogonal rectangular matrices of the following form:

$$\{W \in \mathbb{R}^{n\times p} | W^\top W = I_p\}, n \geq p. \tag{1}$$

We denote this set by $\mathrm{St}(n,p)$. The case where $n = p$ is called the orthogonal group and is denoted as $O(n)$. In contrast, the Grassmann manifold only pays attention to the subspace spanned by the column vectors of W. By introducing the following equivalence relation $\sim$

$$W_2 \sim W_1 \Longleftrightarrow \exists R \in O(p), \text{ s.t. } W_2 = W_1 R, \tag{2}$$

the Grassmann manifold $\mathrm{Gr}(n,p)$ can be regarded as the quotient space $\mathrm{St}(n,p)/\sim$. The Riemannian optimization methods were proposed to solve optimization problems posed on a manifold M such as $\mathrm{St}(n,p)$ or $\mathrm{Gr}(n,p)$:

$$\begin{aligned}&\text{iteratively find } \arg\min_M f(W),\\ &\text{where } f \text{ is a real-valued smooth function on } M.\end{aligned} \tag{3}$$

Conventional optimization techniques seek for updated points additively. For instance, the natural gradient method [2] proceeds as follows.

$$W_{k+1} = W_k - \eta \,\mathrm{grad}_W f(W_k), \tag{4}$$

where $\mathrm{grad}_W f(W_k)$ is the Riemannian (natural) gradient defined by a Riemannian metric g on M, and η is a learning constant. This updated point W_{k+1} does not always stay on the manifold M, while the Riemannian optimization methods update a point on a manifold along a geodesic, therefore the updated points always satisfy the manifold constraint and the trajectory of the updated points is stable. A geodesic is an extension of a straight line in the Euclidean space to a manifold M and it is determined by a Riemannian metric g on M. We denote the geodesic equation on M emanating from $W \in M$ in direction to $V \in T_W M$ by $\varphi_M(W,V,t)$, which is a solution to the following equation:

$$\frac{d^2\gamma^i(t)}{dt^2} + \Gamma^i_{jk}\frac{d\gamma^j(t)}{dt}\frac{d\gamma^k(t)}{dt} = 0 (1 \leq i \leq \dim M), \tag{5}$$

where $\varphi_M(W,V,t) = (\gamma^1(t), ...\gamma^{\dim M}(t))$ in local coordinates of M, $\Gamma^i_{jk} = \frac{1}{2}\sum_{l=1}^{\dim M} g^{il}(\partial_j g_{kl} + \partial_k g_{lj} - \partial_l g_{jk})$, $g_{ij} = g(\frac{\partial}{\partial x_i}, \frac{\partial}{\partial x_j})$, g^{ij} is the inverse of g_{ij}. By using this notation, the updating rules for three Riemannian optimization methods over a manifold M are described as follows [3].

- gradient descent method

$$W_{k+1} = \varphi_M(W_k, -\operatorname{grad}_{W_k} f, \eta)$$

- Newton's method

$$W_{k+1} = \varphi_M(W_k, \operatorname{Hess}_{W_k} f^{-1}(-\operatorname{grad}_{W_k} f), 1)$$

- conjugate gradient method (due to Fletcher-Reeves)

$$\begin{aligned} H_0 &= -\operatorname{grad}_{W_0} f, \\ t^k_{\min} &= \arg\min_t f(\varphi_M(W_k, H_k, t)) \\ W_{k+1} &= \varphi_M(W_k, H_k, t^k_{\min}), \\ H_{k+1} &= -\operatorname{grad}_{W_{k+1}} f + \gamma_{k+1} \Pi H_k, \end{aligned} \quad (6)$$

where ΠH_k is the parallel transportation vector of H_k to W_{k+1} along $\varphi_M(W_k, H_k, t)$.

$$\gamma_k = \frac{g(\operatorname{grad}_{W_k} f, \operatorname{grad}_{W_k} f)}{g(\operatorname{grad}_{W_{k-1}} f, \operatorname{grad}_{W_{k-1}} f)}. \quad (7)$$

Here we summarize the geodesic formulas and relevant facts necessary to describe the updating rules for the Stiefel and the Grassmann manifolds. We use the following Riemannian metrics:

$$g^{\mathrm{St}(n,p)}_W(X,Y) = \operatorname{tr}\{X^\top (I - \frac{1}{2} W W^\top) Y\} \quad (8)$$

$$g^{\mathrm{Gr}(n,p)}_W(X,Y) = \operatorname{tr}\{X^\top Y\}, \text{ Euclidean metric.} \quad (9)$$

Together with the tangent space structures:

$$V_1 \in T_W \mathrm{St}(n,p) \Longleftrightarrow W^\top V_1 \text{ is skew symmetric} \quad (10)$$

$$V_2 \in T_W \mathrm{Gr}(n,p) \Longleftrightarrow W^\top V_2 = 0, \quad (11)$$

the natural gradients are expressed as:

$$\operatorname{grad}^{\mathrm{St}(n,p)}_W f = \nabla f(W) - W \nabla f(W)^\top W \quad (12)$$

$$\operatorname{grad}^{\mathrm{Gr}(n,p)}_W f = \nabla f(W) - W W^\top \nabla f(W). \quad (13)$$

First, we present the geodesic formulas obtained by Edelman-Arias-Smith [3].

The geodesic emanating from $W \in \mathrm{St}(n,p)$ in direction $V \in T_W \mathrm{St}(n,p)$ is given by the curve

$$W(t) = WM(t) + QN(t), \quad (14)$$

where

$$QR := K = (I - WW^\top)V \quad (15)$$

is the compact QR decomposition of K ($Q : n \times p$, $R : p \times p$) and $M(t)$ and $N(t)$ are $p \times p$ matrices given by the matrix exponential

$$\begin{pmatrix} M(t) \\ N(t) \end{pmatrix} = \exp t \begin{pmatrix} A & -R^\top \\ R & O \end{pmatrix} \begin{pmatrix} I_p \\ O \end{pmatrix}, \quad (16)$$

where $A = W^\top V$. The geodesic on the Grassmann manifold starting from $W \in \mathrm{Gr}(n,p)$ with $V \in T_W \mathrm{Gr}(n,p)$ is expressed as:

$$W(t) = (WS \ R) \begin{pmatrix} \cos \Sigma t \\ \sin \Sigma t \end{pmatrix} S^\top, \quad (17)$$

where $R\Sigma S^\top$ is the compact singular value decomposition of V.

$$\cos \Sigma = \begin{pmatrix} \cos\sigma_1 & & 0 \\ & \ddots & \\ 0 & & \cos\sigma_p \end{pmatrix}, \quad \text{where } \Sigma = \begin{pmatrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_p \end{pmatrix}. \quad (18)$$

Stating their formulas, we instead utilize our formulas obtained in [8], which give geometrically simpler interpretations and are easier to analyze .

$$\varphi_{\mathrm{St}(n,p)}(W, -\operatorname{grad}^{\mathrm{St}(n,p)}_W f, t) = \exp(-t(\nabla f(W) W^\top - W \nabla f(W)^\top)) W \quad (19)$$

$$\varphi_{\mathrm{St}(n,p)}(W, V, t) = \exp(t(DW^\top - WD^\top)) W, \quad (20)$$

$$\text{where } D = (I - \frac{1}{2} W W^\top) V \quad (21)$$

$$\varphi_{\mathrm{Gr}(n,p)}(W, -\operatorname{grad}^{\mathrm{Gr}(n,p)}_W f, t) = \exp(-t(\nabla f(W) W^\top - W \nabla f(W)^\top)) W \quad (22)$$

$$\varphi_{\mathrm{Gr}(n,p)}(W, V, t) = \exp(t(VW^\top - WV^\top)) W \quad (23)$$

2. Reduction of the Grassmann formula to the Stiefel formula

In this section we prove the main result:
if $f : \mathrm{St}(n,p) \to \mathbb{R}$ has a $O(p)$-symmetry, i.e.

$$f(WR) = f(W) \text{ for all } W \in \mathrm{St}(n,p), R \in O(p), \quad (24)$$

the updating rules given by the gradient descent, the conjugate gradient method (due to Fletcher-Reeves), and the Newton method over the Stiefel manifold coincide with those over the Grassmann manifold respectively.

Proof: Let us denote the smooth curve over the orthogonal group $O(p)$ passing through the identity by $c(t)$.

$$c(t) : [-a, a] \longmapsto O(p), \qquad c(0) = I. \quad (25)$$

Differentiating the relation $c(t)c(t)^\top = I_p$ at I_p, we get

$$c'(0)c(0)^\top + c(0)(c(0)')^\top = 0. \quad (26)$$

Thus $c'(0) = X$ is a skew symmetric matrix. We next consider $Wc(t)$, which is a curve on $\mathrm{St}(n,p)$ passing

through W. It follows from the definition of gradient that

$$\langle Wc^{'}(0), \nabla_W f\rangle = (Wc^{'}(0))f^1 \tag{27}$$

Because of the condition (24), f takes the same value at $Wc(t)$:

$$f(Wc(t)) = f(W). \tag{28}$$

Therefore, $Wc^{'}(t)f = \frac{d}{dt}f(Wc(t))|_{t=0} = 0 \Longleftrightarrow$

$$\langle WX, \nabla_W f\rangle = \mathrm{tr}((WX)^\top \nabla_W f) = \mathrm{tr}\, X^\top W^\top \nabla_W f = 0 \tag{29}$$

for all X: skew symmetric matrix. It follows that $W^\top \nabla_W f$ is symmetric. Hence

$$W(\nabla_W f)^\top W = WW^\top \nabla_W f, \tag{30}$$

which leads to $\mathrm{grad}_W^{\mathrm{Gr}(n,p)} f = \mathrm{grad}_W^{\mathrm{St}(n,p)} f$. Thus

$$\varphi_{\mathrm{St}(n,p)}(W, -\mathrm{grad}_W^{\mathrm{St}(n,p)} f, t) = \varphi_{\mathrm{Gr}(n,p)}(W, -\mathrm{grad}_W^{\mathrm{Gr}(n,p)} f, t). \tag{31}$$

For analyzing the Newton's method, the notion of Hessian is important.

$$\mathrm{Hess}_W f(V,V) = \left.\frac{d^2}{dt^2} f(W(t))\right|_{t=0}, \tag{32}$$

where $W(t)$ is the geodesic starting from W in direction to $W'(0) = V$. From this definition, we get the following formulas [3]:

$$\mathrm{Hess}_W^{\mathrm{St}} f(V_1, V_2) = f_{WW}(V_1, V_2) + \frac{1}{2}\mathrm{tr}((\nabla_W f)^\top V_1 W^\top + W^\top V_1 (\nabla_W f)^\top) V_2) - \frac{1}{2}\mathrm{tr}((W^\top \nabla_W f + (\nabla_W f)^t W) V_1^t (I - WW^t) V_2). \tag{33}$$

$f_{WW}(V_1, V_2)$ denotes $\sum_{ij,kl} (f_{WW})_{ij,kl} (V_1)_{ij} (V_2)_{kl}$, where

$$(f_{WW})_{ij,kl} = \frac{\partial^2 f}{\partial W_{ij} W_{kl}}.$$

$$\mathrm{Hess}_W^{\mathrm{Gr}(n,p)} f(V_1, V_2) = f_{WW}(V_1, V_2) - \mathrm{tr}(V_1^\top V_2 W^\top \nabla_W f). \tag{34}$$

The assumption (24) allows us to set $W^\top \nabla_W f$ is symmetric, therefore for all $V_1, V_2 \in T_W \mathrm{Gr}(n,p)$,

$$\mathrm{Hess}_W^{\mathrm{St}(n,p)} f(V_1, V_2) = f_{WW}(V_1, V_2) - \mathrm{tr}(W^\top \nabla_W f V_1^\top V_2) = \mathrm{Hess}_W^{\mathrm{Gr}(n,p)} f(V_1, V_2). \tag{35}$$

[1]Here the tangent vector $Wc^{'}(0)$ is regarded as a differential operator acting on a function f.

In addition, for all $X \in T_W \mathrm{Gr}(n,p)$

$$\mathrm{tr}((-\mathrm{grad}_W^{\mathrm{St}(n,p)} f)^\top (I - \frac{1}{2} WW^\top) X) = \mathrm{tr}((-\mathrm{grad}_W^{\mathrm{Gr}(n,p)} f)^\top X) = g_{\mathrm{Gr}(n,p)}(-\mathrm{grad}_W^{\mathrm{Gr}(n,p)} f, X), \tag{36}$$

because $W^\top X = O$. Thus the formula for the inverse of the Hessian for the Grassmann manifold is reproduced from the one for the Stiefel manifold;

if V_1 is the solution of

$$\mathrm{Hess}_W^{\mathrm{St}(n,p)} f(V_1, X) = g_{\mathrm{St}(n,p)}(-\mathrm{grad}_W^{\mathrm{St}(n,p)} f, X). \quad \text{for all } X \in T_W \mathrm{St}(n,p), \tag{37}$$

it coincides with the solution V_2 of

$$\mathrm{Hess}_W^{\mathrm{Gr}(n,p)} f(V_2, Y) = g_{\mathrm{Gr}(n,p)}(-\mathrm{grad}_W^{\mathrm{Gr}(n,p)} f, Y), \quad \text{for all } Y \in T_W \mathrm{Gr}(n,p). \tag{38}$$

(We assume $\mathrm{Hess}_W^{\mathrm{St}(n,p)}$ and $\mathrm{Hess}_W^{\mathrm{Gr}(n,p)}$ are nondegenerate.) Therefore

$$\varphi_{\mathrm{St}(n,p)}(W_k, \mathrm{Hess}_{W_k}^{\mathrm{St}(n,p)} f^{-1}(-\mathrm{grad}_{W_k}^{\mathrm{St}(n,p)} f), 1) = \varphi_{\mathrm{Gr}(n,p)}(W_k, \mathrm{Hess}_{W_k}^{\mathrm{Gr}(n,p)} f^{-1}(-\mathrm{grad}_{W_k}^{\mathrm{Gr}(n,p)} f), 1). \tag{39}$$

Lastly, we consider the conjugate gradient method. Let us denote the k-th updated point and the k-th updated search direction for the Stiefel and the Grassmann manifold by $W_k^{\mathrm{St}(n,p)}, H_k^{\mathrm{St}(n,p)}$, and $W_k^{\mathrm{Gr}(n,p)}, H_k^{\mathrm{Gr}(n,p)}$ respectively. Then

$$H_0^{\mathrm{St}(n,p)} = -\mathrm{grad}_{W_0}^{\mathrm{St}(n,p)} f = -\mathrm{grad}_{W_0}^{\mathrm{Gr}(n,p)} f = H_0^{\mathrm{Gr}(n,p)}. \tag{40}$$

Our geodesic formulas (19)-(23) verify

$$W_1^{\mathrm{St}(n,p)} = W_1^{\mathrm{Gr}(n,p)} = \exp(t_{\min}^0 (H_0^{\mathrm{Gr}(n,p)} W_0^\top - W_0 (H_0^{\mathrm{Gr}(n,p)})^\top) W_0. \tag{41}$$

We show the following: if $W_k^{\mathrm{St}(n,p)} = W_k^{\mathrm{Gr}(n,p)}$, and $H_k^{\mathrm{St}(n,p)} = H_k^{\mathrm{Gr}(n,p)} \in T_{W_k}\mathrm{Gr}(n,p)$, then $H_{k+1}^{\mathrm{St}(n,p)} = H_{k+1}^{\mathrm{Gr}(n,p)}$ holds. We can easily get from our geodesic formulas (19)-(23) that

$$W_{k+1}^{\mathrm{St}(n,p)} = W_{k+1}^{\mathrm{Gr}(n,p)} = \exp(t_{\min}^k (H_k^{\mathrm{Gr}(n,p)} W_k^\top - W_k (H_k^{\mathrm{Gr}(n,p)})^\top)) W_k. \tag{42}$$

The assumption (24) gives

$$\mathrm{grad}_{W_k}^{\mathrm{St}(n,p)} f = \mathrm{grad}_{W_k}^{\mathrm{Gr}(n,p)} f \text{ and } \gamma_k^{\mathrm{St}(n,p)} = \gamma_k^{\mathrm{Gr}(n,p)}.$$

Moreover, since

$$\varphi_{\mathrm{St}(n,p)}(W_k, H_k^{\mathrm{St}(n,p)}, t) = \varphi_{\mathrm{Gr}(n,p)}(W_k, H_k^{\mathrm{Gr}(n,p)}, t)$$

are geodesics, the parallel transportation vector of $H_k^{\mathrm{St}(n,p)}$ along $\varphi_{\mathrm{St}(n,p)}(W_k, H_k^{\mathrm{St}(n,p)}, t)$ to W_{k+1} is equal to the velocity vector of $\varphi_{\mathrm{St}(n,p)}(W_k, H_k^{\mathrm{St}(n,p)}, t)$ at $t = t_{\min}^k$. Therefore

$$\begin{aligned}
&\Pi_{\mathrm{St}(n,p)} H_k^{\mathrm{St}(n,p)} \\
&\quad = (H_k^{\mathrm{Gr}(n,p)} W_k^\top - W_k (H_k^{\mathrm{Gr}(n,p)})^\top) W_{k+1}^{\mathrm{Gr}(n,p)} \\
&\quad = \Pi_{\mathrm{Gr}(n,p)} H_k^{\mathrm{Gr}(n,p)}, \qquad (43)
\end{aligned}$$

and so $H_{k+1}^{\mathrm{St}(n,p)} = H_{k+1}^{\mathrm{Gr}(n,p)}$ holds. Thus, by induction, it follows that $W_s^{\mathrm{St}(n,p)} = W_s^{\mathrm{Gr}(n,p)}$, and $H_s^{\mathrm{St}(n,p)} = H_s^{\mathrm{Gr}(n,p)}$, $s = 0, 1, 2, \ldots$ □

Note that the formulas for $\Pi_{\mathrm{St}(n,p)} H_k^{\mathrm{St}(n,p)}$, $\Pi_{\mathrm{Gr}(n,p)} H_k^{\mathrm{Gr}(n,p)}$ obtained in [3] are much more complicated and obscure the geometrical meaning.

More generally, we can extend the above result to a principal bundle and its base manifold. Let us consider a principal G-bundle P over M, and denote the projection by $\pi : P \to M$. (For details of the principal bundle theory, the readers should refer to [4].) $\mathrm{St}(n,p)$ is a principal $O(p)$-bundle over $\mathrm{Gr}(n,p)$. We introduce a connection θ ($\mathfrak{g}$-valued 1 form) on P. By using θ, we decompose T_pP into the sum of the horizontal space H_p and the vertical space V_p. We endow P with a G-invariant metric g_p so that H_P is orthogonal to V_P, and endow M with the metric g_M induced from P:

$$g_M(u, v) \equiv g_P(u_{H_P}, v_{H_p}), \qquad (44)$$

where u, $v \in T_mM$, u_{H_P}, v_{H_p} are lift of u, v to H_p, $p \in \pi^{-1}(m)$. We assume a function F to be optimized on P has a G-symmetry in the sense that F takes the same value on each fiber: $F(\pi^{-1}(m))$ =const. In other words,

$$F(p \cdot g) = F(p), \text{ for all } p \in P,\ g \in G. \qquad (45)$$

Then the following holds:

$$\pi(\varphi_P(p, -\operatorname{grad}_p^P F, t)) = \varphi_M(m, -\operatorname{grad}_m^M f, t), \quad (46)$$

$$\begin{aligned}
&\pi(\varphi_P(p, \operatorname{Hess}_p^P F^{-1}(-\operatorname{grad}_p^P F), 1)) = \\
&\qquad \varphi_M(m, \operatorname{Hess}_m^M f^{-1}(-\operatorname{grad}_m^M f), 1) \quad (47)
\end{aligned}$$

$$H_k^P = H_k^M, \qquad (48)$$

where H_k^P, H_k^M denote the search directions at k-th iteration for P and M determined by the conjugate gradient method (6). The proof proceeds like the same manner as the Stiefel and the Grassmann cases.

3. Conclusion

Exploiting our previously obtained geodesic formulas for the Stiefel manifold, we gave a unifying view on the Riemannian optimization methods on the Stiefel and the Grassmann manifolds. Instead of investigating the quotient space structure of the Grassmann manifold, we had only to take the symmetry of the target function into account and apply the formulas for the Stiefel manifold. This may give an insight for applying the Riemannian optimization methods to other homogeneous spaces.

Acknowledgments

The author would like to thank Shotaro Akaho, Jun Fujiki, Kazuya Takabatake for discussions, Toshihiro Kamishima for help in typesetting.

References

[1] P-A. Absil, R. Mahony, R. Sepulchre, and P. Van Dooren, Riemannian geometry of Grassmann manifolds with a view on algorithmic computation, *Acta Applicandae Mathematicae*, **80** (2), pp.199-220, 2004.

[2] S. Amari, Natural gradient works efficiently in Learning, *Neural Computation*, **10**, pp.251-276, 1998.

[3] 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), pp.303-353, 1998.

[4] S. Kobayashi, and K. Nomizu, *Foundations of Differential Geometry*, John Wiley & Sons.

[5] X. Liu, A. Srivastava, K. Gallivan, Optimal Linear Representations of Images for Object Recognition, *IEEE Trans. Pattern Analysis and Machine Intelligence*, **26**(5), pp.662-666, 2004.

[6] Y. Ma, J. Kosecka, and S. S. Sastry, Optimization Criteria and Geometric Algorithms for Motion and Structure Estimation. , *International Journal of Computer Vision*, **44**(3), 219-249, 2001.

[7] Y. Nishimori, Learning Algorithm for Independent Component Analysis by Geodesic Flows on Orthogonal Group, *Proceedings of International Joint Conference on Neural Networks (IJCNN1999)*, **2**, pp.1625-1647, 1999.

[8] Y. Nishimori, Learning Algorithms Utilizing Quasi-Geodesic Flows on the Stiefel Manifold, *Neurocomputing*, **67** pp.106-135, 2005.

[9] M. D. Plumbley, Algorithms for non-negative independent component analysis. *IEEE Transactions on Neural Networks*, **14** (3), pp.534-543, 2003.