第 35 卷 第 4 期 应 用 数 学 学 报 Vol. 35 No. 4
2012 年 7 月 ACTA MATHEMATICAE APPLICATAE SINICA July, 2012

Stiefel 流形上的梯度下降法 *

吴秋峰

(东北农业大学理学院, 哈尔滨 150030)

(E-mail: neauqfwu@gmail.com)

刘振忠 †

(东北农业大学理学院, 哈尔滨 150030)

(E-mail: lzz00@126.com)

摘 要 基于 Stiefel 流形上算法的几何框架, 本文提出了 Stiefel 流形上的梯度下降法. 理论上给出了算法收敛性定理. 三个数值仿真算例表明算法是有效的, 与其他方法相比具有更快的收敛速度.

关键词 约束非线性优化问题; 梯度下降法; Stiefel 流形

MR(2000) 主题分类 46T10; 14M15; 90C52

中图分类 O221.2; O229; O189.3+3

1 引言

约束优化问题是在自变量满足约束条件的情况下目标函数最小化的问题. 约束条件不同, 就会得到不同的约束优化问题. 它在机器视觉、信号处理、机器学习、参数识别与自动控制理论等领域都有着重要应用 [1,2]. 例如: 最近邻分类中距离度量学习问题就可归结为凸半正定规划问题 [3]; 控制系统中 PID 控制参数优化设计问题也可归结为约束优化问题 [4]. 正是这些领域提出的不同类型问题刺激了约束优化问题理论的快速发展, 使得约束优化问题成为当今最优化领域中研究和讨论的重要方向之一.

在信号和图像处理中, 采用主成分分析解决数据压缩、特征提取和维数缩减等问题时, 常将该问题归结为带有正交约束的非线性优化问题 [5]; 在信息检索中, 采用非负矩阵分解方法解决非负数据的多变量分析问题时, 也常将该问题归结为带有正交约束的非线性优化问题 [6]. 本文主要研究求解此类带有正交约束的非线性优化问题, 即 Stiefel

本文 2011 年 7 月 9 日收到. 2011 年 10 月 17 日收到修改稿.

† 通讯作者

* 东北农业大学科学技术资助项目 (2011RCA01).

流形上的非线性优化问题. 该问题的一般形式可以表述为 [7]:

$$\min_{X\in V_{n,p}} F(X), \tag{1.1}$$

其中, $F: V_{n,p}\to R^1$ 是凸函数, 且 $F\in C^1$; $V_{n,p}$ 表示 Stiefel 流形, 即:

$$V_{n,p}=\{X_{n\times p}|X^TX=I_p,\ n\geq p\}. \tag{1.2}$$

目前为止, 经过国内外学者的不断探讨, Stiefel 流形上的非线性优化问题的研究已取得了一系列成果. 1998 年 Edelman 等人提出了 Stiefel 流形上算法列式的几何框架, 奠定了 Stiefel 流形上算法设计的基础, 并在此框架下设计了 Stiefel 流形上的 Newton 法和共轭梯度法, 但 Stiefel 流形上的 Newton 法所要计算的目标函数 Hessian 矩阵复杂性, 决定了算法高复杂性, 极大降低了算法的收敛速度 [8]. 之后, 学者们的研究主要围绕求解具有不同的特定目标函数的 Stiefel 流形上的非线性优化问题. 2006 年范金燕等人分析了 Stiefel 流形上的二次规划问题解的特性, 采用解的特性解决 Steifel 流形上二次指派问题 [9]. 2008 年 Yoo 等人提出了正交非负矩阵分解方法, 解决了带有正交约束的最小二乘误差问题 [6]. 2009 年 Dodig 等人将 Stiefel 流形上的二次规划问题转化为半正定规划问题, 采用 Matlab 的半正定规划工具箱加以求解 [10].

本文提出了 Stiefel 流形上梯度下降法, 理论上证明了算法全局收敛性, 数值的收敛性分析、参数固定与自适应调整对比分析和不同算法对比分析仿真表明算法是有效的, 且具有较快的收敛速度.

2 Stiefel 流形上的梯度下降法

2.1 基本梯度下降法

对于非约束优化问题, 假定目标函数 $F(X)\in C^1$, 梯度下降法基本思想是: 从当前点 X_k 出发, 取 $F(X)$ 在 X_k 处下降的方向作为搜索方向.

梯度下降法计算步骤 [11]:

步骤 1 给定初始点 X_k, 终止控制常数 $\varepsilon>0$ 和步长参数 t, 令 $k=0$;

步骤 2 计算 $\nabla F(X_k)$, 若 $\|\nabla F(X_k)\|\leq\varepsilon$, 停止迭代, 输出 X_k, 否则进行下一步;

步骤 3 取 $p_k=-\nabla F(X_k)$, $X_{k+1}=X_k+tp_k$, $k=k+1$, 转步骤 2.

2.2 梯度下降法改进规则

对于 Stiefel 流形上的非线性优化问题, 梯度下降法并不能保证每次迭代使 X 满足 Stiefel 流形性质. 对梯度下降法的改进之一就是保证每次迭代使 X 满足 Stiefel 流形性质. 为了满足上述条件, X 应在 Stiefel 流形的测地线上加以迭代, 要想求得 Stiefel 流形上 X 处的测地线, 必须求在 Stiefel 流形上 $F(X)$ 在 X 处的切向量 $\widetilde{\nabla}F$, 为了便于求得 $\widetilde{\nabla}F$, 定理 2.1 给出了欧氏空间的 $F(X)$ 在 X 处的梯度 ∇F 与 $\widetilde{\nabla}F$ 的关系.

定理 2.1[8] 设 $F(X)$ 是 Stiefel 流形上的光滑函数, 且 $F(X)\in C^1$, $X\in V_{n,p}$ 是 Stiefel 流形上的点, $\nabla F=\left(\frac{\partial F}{\partial x_{ij}}\right)$ 为在欧氏空间的 $F(X)$ 在 X 处的梯度, $\widetilde{\nabla}F$ 在 Stiefel

流形上 $F(X)$ 在 X 处的切向量， ∇F 与 $\widetilde{\nabla} F$ 的关系：

$$\widetilde{\nabla} F = \nabla F - X(\nabla F)^T X. \tag{2.1}$$

在 Stiefel 流形上在 X 处沿着 $H = -\widetilde{\nabla} F$ 的测地线方程为 $X(t) = \exp_X(tH)$, 为了减少计算量， [8] 给出一种简便计算方法.

定理 2.2[8] 设 X 和 H 均为 $n \times p$ 矩阵，满足 $X^T X = I_p$ 和 $A = X^T H$ 为反对称矩阵，则在 Stiefel 流形上在 X 处沿着 $H = -\widetilde{\nabla} F$ 的测地线方程：

$$X(t) = XM(t) + QN(t), \tag{2.2}$$

其中

$$\begin{pmatrix} M(t) \\ N(t) \end{pmatrix} = \exp t \begin{pmatrix} A & -R^T \\ R & 0 \end{pmatrix} \begin{pmatrix} I_p \\ 0 \end{pmatrix}, \tag{2.3}$$

其中 $Q_{n\times p}$ 和 $R_{p\times p}$ 为 $(I - XX^T)H$ 的 QR 分解所得矩阵.

在梯度下降法中步长参数 t 选取对算法收敛速度起到决定性作用，参数 t 过大，虽然速度提高，但导致收敛不到最优解；参数 t 过小，导致收敛速度过慢，因此，对梯度下降法的改进之二是参数 t 的自适应调整. 实验表明自适应调整参数，有效加快算法的收敛. 在迭代过程中自适应调整参数的规则如下：

(1) 若 $F(X_k) > F(X_{k+1})$, 则 $t_{k+1} = \alpha t_k,\ \alpha > 1$, 如 $\alpha = 1.01$;

(2) 若 $F(X_k) < F(X_{k+1})$, 则 $t_k = \alpha t_k,\ \alpha < 1$, 如 $\alpha = 0.1$.

2.3 Stiefel 流形上的梯度下降法

Stiefel 流形上的梯度下降法 (Gradient Descent on the Stiefel Manifold, GDSM) 计算步骤：

步骤 1 给定初始点 X_0, 终止控制常数 $\varepsilon > 0$ 和步长参数初值 t_0, 令 $k = 0$;

步骤 2 计算 $\widetilde{\nabla} F = \nabla F - X(\nabla F)^T X$, 若 $\|\widetilde{\nabla} F(X_k)\|^2 \le \varepsilon$, 停止迭代，输出 X_k, 否则进行下一步;

步骤 3 在 X_k 处沿着 $H = -t_k \widetilde{\nabla} F$ 的迭代公式：

$$X_{k+1} = X_k M + QN,$$

其中 M 和 N 由下式求得

$$\begin{pmatrix} M \\ N \end{pmatrix} = \exp \begin{pmatrix} A & -R^T \\ R & 0 \end{pmatrix} \begin{pmatrix} I_p \\ 0 \end{pmatrix},$$

其中上式 $Q_{n\times p}$ 和 $R_{p\times p}$ 为 $(I - XX^T)H$ 的 QR 分解所得矩阵.

步骤 4 若 $F(X_k) > F(X_{k+1})$, 则 $k = k + 1,\ \ t_{k+1} = \alpha t_k, \alpha > 1$, 转步骤 2. 否则若 $F(X_k) < F(X_{k+1})$, 则 $t_k = \alpha t_k,\ \alpha < 1$, 转步骤 3.

3 算法的收敛性分析

由于 Stiefel 流形是嵌入在欧氏空间中的微分流形，欧氏空间的一些性质和定理自然适用于 Stiefel 流形 [12]. Stiefel 流形上在 X 处正则度量为 [8]:

$$g_c(A,A)=\mathrm{tr}\Big(A^T\Big(I-\frac{1}{2}XX^T\Big)A\Big). \tag{3.1}$$

连接 X_k, X_{k+1} 点的测地线为 $X(t)$, 其中 $X(0)=X_k,\ X(1)=X_{k+1}$; 在 GDSM 算法中迭代的切向量方向为 $-\widetilde{\nabla}F$.

定理 3.1 设 $F: V_{n,p}\to R^1$ 是定义在 Stiefel 流形上的强凸函数，且 $F\in C^1$, 对 GDSM 算法，若下列条件成立：

(1) 初始点 $X_0\in D$;

(2) $S=\{X\in V_{n,p}|F(X)<F(X_0)\}$ 是开凸集 D 的一个子集;

(3) 存在 $L\in R$, 使得 $F(X)\geq L,\ \forall X\in S$;

(4) 存在常数 $b>0$, 使得 HessF 的特征值满足 $\lambda_j(X)\leq b,\ j=1,2,\cdots,n,\ \forall X\in S$;

(5) 存在 $t_k>0$, 使得 $F(X_{k+1})-F(X_k)\leq 0$.

则有：

(1) $\{F(X_k)\}$ 是收敛的;

(2) $\lim\limits_{k\to\infty}\big(\|\widetilde{\nabla}F(X_k)\|_F^2\big)=0$;

(3) $\{X_k\}$ 的收敛点 X^* 是 $F(X)$ 的最小值点.

证 (1) 考察第 k 次迭代的结果. 设连接 X_k, X_{k+1} 点的测地线为 $X(t), X(0)=X_k,\ X(t_k)=X_{k+1}$. 由 Taylor 公式 [12], 有

$$\begin{aligned}F(X_k)-F(X_{k+1})\approx&-t_kg_c(\widetilde{\nabla}F(X(0)),-\widetilde{\nabla}F(X(0)))\\&-\frac{1}{2}t_k^2\mathrm{Hess}\,F(-\widetilde{\nabla}F(X(s_0)),-\widetilde{\nabla}F(X(s_0))),\qquad s_0\in[0,1],\end{aligned}$$

其中

$$\widetilde{\nabla}F=\nabla F-X(\nabla F)^TX;$$

$$\mathrm{Hess}F(\triangle,\triangle)=-\mathrm{tr}\left((\nabla F)^T(\triangle\triangle^TX+X\triangle^T(I-XX^T)\triangle)\right)+\nabla^2F(\triangle,\triangle);$$

且 $\nabla^2F=\big(\frac{\partial^2F}{\partial x_{ij}\partial x_{kl}}\big)$[8].

由条件 (4) 和测地线性质知 $\|-\widetilde{\nabla}F(X(s_0))\|_F=\|-\widetilde{\nabla}F(X(0))\|_F$, 因此

$$F(X_k)-F(X_{k+1})\geq-t_kg_c(\widetilde{\nabla}F(X(0)),-\widetilde{\nabla}F(X(0)))-\frac{t_k^2}{2}b\|-\widetilde{\nabla}F(X(0))\|_F^2. \tag{3.2}$$

在 Stiefel 流形上有

$$g_c(\widetilde{\nabla}F,-\widetilde{\nabla}F)=-\|-\widetilde{\nabla}F\|_F\|\nabla F\|_F, \tag{3.3}$$

其中 $\|\cdot\|_F$ 为矩阵的 Frobenias 范数.

将 (3.3) 式代入 (3.2) 式有

$$F(X_k)-F(X_{k+1})\geq t_k\|\widetilde{\nabla}F(X_k)\|_F^2-\frac{t_k^2}{2}b\|\widetilde{\nabla}F(X_k)\|_F^2=\Big(t_k-\frac{t_k^2}{2}b\Big)\|\widetilde{\nabla}F(X_k)\|_F^2. \quad (3.4)$$

由条件 (5) 知 $\{F(X_k)\}$ 是单调递减的，由条件 (3) 知 $\{F(X_k)\}$ 有界，故 $\{F(X_k)\}$ 是收敛的.

(2) 对 (3.4) 式两边取极限得

$$\lim_{k\to\infty}\big(\|\widetilde{\nabla}F(X_k)\|_F^2\big)=0. \quad (3.5)$$

(3) 由于 $\{X_k\}$ 的收敛点 X^*, 则 X^* 为 $F(X)$ 的驻点，即满足 $\|\widetilde{\nabla}F(X^*)\|_F^2=0$, 若 $\|\widetilde{\nabla}F(X^*)\|_F^2\neq 0$, 则 $\lim\limits_{k\to\infty}\big(\|\widetilde{\nabla}F(X_k)\|_F^2\big)=\|\widetilde{\nabla}F(X^*)\|_F^2\neq 0$ 与 (3.5) 式矛盾.

由 $F(X)$ 是强凸函数知，在任意点处 Hess$F$ 是正定的，故 $\{X_k\}$ 的收敛点 X^* 是 $F(X)$ 的最小值点.

4　数值仿真算例

本文的 3 个算例环境均为： CPU 为 Q9400@2.66GHz, 内存为 4G, 操作系统为 Windows XP, 编程环境为 Matlab 7.0.

4.1　收敛性分析

考察以下优化问题：

$$\min_{X\in V_{n,p}}\frac{1}{2}\|AX-B\|_F^2, \quad (4.1)$$

其中 $A_{n\times n}, B_{n\times p}$ 已知，问题 (4.1) 表述为：

$$F(X)=\frac{1}{2}\big[\mathrm{tr}\,(X^TA^TAX)-2\mathrm{tr}\,(B^TAX)+\mathrm{tr}\,(B^TB)\big].$$

因此，在欧氏空间的 $F(X)$ 在 X 处的梯度为：

$$\nabla F=A^TAX-A^TB. \quad (4.2)$$

在 Stiefel 流形上 $F(X)$ 在 X 处的切向量为：

$$\widetilde{\nabla}F=(A^TAX-A^TB)-X(A^TAX-A^TB)^TX. \quad (4.3)$$

取 $n=5,\ p=3,\ A$ 为在 $[0,1]$ 随机产生 5×5 矩阵，而 $B=AI_{5,3}$, 因此此问题的最优解为 $X=I_{5,3}$, 步长参数初始值 $t_0=0.1$, 所采用的收敛性评价指标为 $\|X-I_{5,3}\|_F$. 针对问题 (4.1), GDSM 的收敛性分析如图 1 所示.

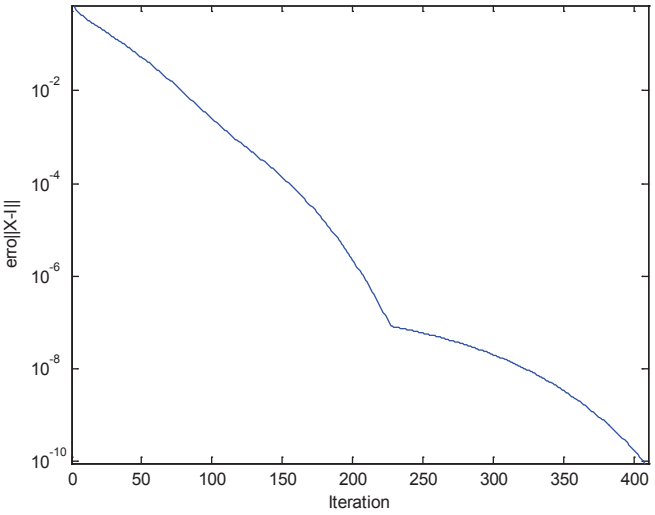

图 1 算法的收敛性

从图 1 可知， GDSM 算法收敛到最优解，且收敛速度较快.

4.2 参数固定与自适应调整对比分析

针对问题 (4.1), 考察在 GDSM 算法中步长参数固定 (Fixed Parameters, GDSM-FP) 与自适应调整 (Adaptive Adjustment Parameters, GDSM-AAP) 两种情况下, 对算法的收敛速度影响的对比分析如表 1 所示.

从表 1 可以看出 GDSM-FP 算法随着参数 t_0 选取的不同，迭代次数不同，参数 t_0 选取不恰当极大影响算法的收敛速度；GDSM-AAP 算法随着参数 t_0 选取的不同，迭代次数差别不明显，表明在 GDSM-AAP 中，参数 t_0 初值的选取对收敛速度的影响不大，都能很快收敛到最优解.

表 1 GDSM-FP 与 GDSM-AAP 收敛速度分析

	步长参数 t_0 初值			
迭代次数	0.15	0.10	0.01	0.001
GDSM-FP	759	1143	11518	115301
GDSM-AAP	1171	1079	1223	1448

4.3 不同算法对比性分析

针对问题 (4.1), 考察 Stiefel 流形上的梯度下降法 (GDSM), Stiefel 流形上的牛顿法

[8] (Newton on the Stiefel Manifold, NSM) 和 Stiefel 流形上的共轭梯度法 [8] (Conjugate Gradient on the Stiefel Manifold, CGSM), 随着问题规模的增大，比较 GDSM, NSM 和 CGSM 的运行速度，如图 2 所示.

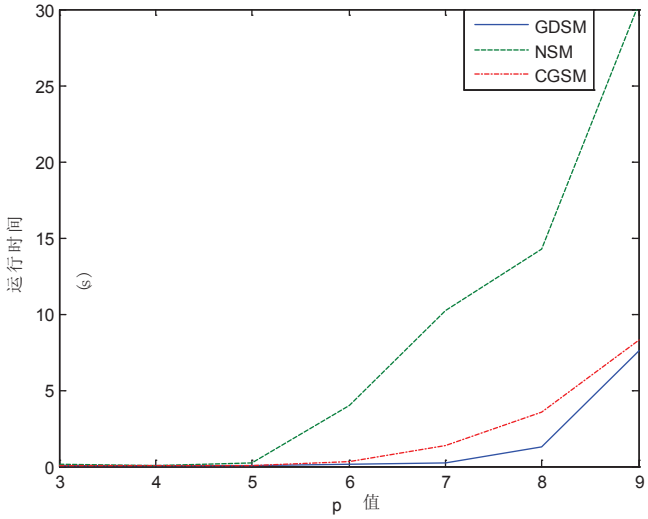

图 2　GDSM, NSM 和 CGSM 速度比较分析

从图 2 可以看出，随着问题规模的增大，p (即 $A_{n\times p}$, $p=(n+1)/2$) 值增大，GDSM 相对于 NSM 和 CGSM 收敛速度变化比较平缓. NSM 随着问题规模的增大， NSM 和 CGSM 收敛速度慢原因在于 NSM 需要计算 Hessian 矩阵，而 CGSM 需要重新合成迭代方向. 而 GDSM 只需计算 $\widetilde{\nabla} F$ 即可，因此， GDSM 比 NSM 和 CGSM 具有较快的收敛速度.

5　结论

基于 Steifel 流形上算法的几何框架, 本文提出了一种 Stiefel 流形上的梯度下降法. 给出了算法收敛性定理，三个数值仿真算例表明算法是有效的，且具有较快的收敛速度.

参 考 文 献

[1] Champagne B. Adaptive Eigendecomposition of Data Covariance Matrices Based on First- order Perturbations. *IEEE Transaction on Signal Processing*, 1994, 42(10): 2758–2770

[2] Favaro P, Soatto S. A Geometric Approach to Shape from Defocus. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2005, 27(3): 406–417

[3] Weinberger K Q, Saul L K. Distance Metric Learning for Large Margin Nearest Neighbor Classification. *Journal of Machine Learning Research*, 2009, 10: 207–244

[4] Li D H, Gao F R, Xue Y L, et al. Optimization of Decentralized PI/PID Controllers Based on Genetic Algorithm. *Asian Journal of Control*, 2007, 9(3): 306–316

[5] 段玲，黄建国．主成分分析的一个黎曼几何随机算法．上海交通大学学报, 2004, 38(1): 71–74
(Duan L, Huang J. A Riemannian Geometry Underlying Stochastic Algorithm for Adaptive Principal Component Analysis. *Journal of Shang Hai Jiao Tong University*, 2004, 38(1): 71–74)

[6] Yoo J, Choi S. Orthogonal Nonnegative Matrix Factorization: Multiplicative Updates on Stiefel Manifolds. 9th International Conference on Intelligent Data Engineering and Automated Learning, Daejeon, South Korea, 2008, 140–147

[7] Uschmajew A. Well-posedness of Convex Maximization Problems on Stiefel Manifolds and Orthognal Tensor Product Approximations. *Numerische Mathematik*, 2010, 115(2): 309–331

[8] Edelman A, Arias T, Smith S T. The Geometry of Algorithms with Orthogonality Constraints. *SIAM Journal on Matrix Analysis and Applications*, 1998, 20: 303–353

[9] Fan J Y, Nie P Y. Quadratic Programs over the Stiefel Manifold. *Operations Research Letters*, 2006, 34: 135–141

[10] Dodig M, Stosic M, Xavier J. On the Minimizing a Quadratic Function on Stiefel Manifolds. Technical Report, Instituto de Sistemas e Robotica, 2009. Available at http://users.isr.ist.utl.pt /?jx avier/ ctech.pdf

[11] 袁亚湘，孙文瑜．最优化理论与方法．北京：科学出版社， 1997: 108–109
(Yuan Y, Sun W. Optimization Theory and Methods. Beijing: Science Press, 1997: 108–109)

[12] 黄建国，孙连山，叶中行．黎曼流形上带 Armijo 步长准则优化算法．上海交通大学学报, 2002, 36(2): 267–271
(Huang J, Sun L, Ye Z. Optimization Algorithm with Armijo Rule on Riemann Manifold. *Journal of Shanghai Jiao Tong University*, 2002, 36(2): 267–271)

Gradient Descent on the Stiefel Manifold

Wu Qiufeng

(*College of Science, Northeast Agricultural University, Harbin* 150030)

(*E-mail: neauqfwu@gmail.com*)

Liu Zhenzhong

(*College of Science, Northeast Agricultural University, Harbin* 150030)

Abstract This paper presents gradient descent on the Stiefel manifolds based on algorithmic geometrical framework on the Stiefel manifolds. Theoretically, convergence theorem of this algorithm is given. Three numerical simulations are shown to verify the efficiency of the proposed algorithm, and to have faster convergence rate compared with other methods.

Key words constrained nonlinear optimization problem; gradient descent; Stiefel manifold

MR(2000) Subject Classification 46T10; 14M15; 90C52

Chinese Library Classification O221.2; O229; O189.3+3