# 微分不可能な正則化項を含む多目的最適化問題に対する準ニュートン型近接勾配法について

桑原 寛大 [1], 成島 康史 [1]
[1] 慶應義塾大学
e-mail : kandai_kwbr32@keio.jp

## 1　はじめに

本研究では，以下の制約の無い多目的最適化問題について考える：

$$\min_{x\in\mathbb{R}^n} F(x). \tag{1}$$

ただし，$F:\mathbb{R}^n\to\mathbb{R}^m$ は $F(x):=(F_1(x),...,F_m(x))^T$ であり，$F_i:\mathbb{R}^n\to\mathbb{R}$ $(i=1,...,m)$ は $F_i(x)=f_i(x)+g_i(x)$ とする．また，$f_i:\mathbb{R}^n\to\mathbb{R}$ は $L$–平滑な関数であり，$g_i:\mathbb{R}^n\to\mathbb{R}$ は必ずしも微分可能とは限らない凸関数である．

一般の多目的最適化問題 ( つまり，$g_i(x)=0$ の場合 ) に対する最適化手法として，スカラー化手法やヒューリスティック解法が存在するが，これらの手法は収束性について理論的な考察が困難なため，収束性について議論可能な単一目的最適化問題に対する最急降下法やニュートン法 [1] などの降下法を拡張した手法が注目されている．一方，近年，微分不可能な正則化項を含む単一目的最適化問題に対する最適化手法が盛んに研究されており，特に近接勾配法やニュートン型近接勾配法が注目されている．また，単一目的最適化問題に対する近接勾配法を多目的最適化問題に拡張した手法 [2] も提案されている．そこで本研究ではニュートン型近接勾配法の概念を取り入れ，多目的最適化問題 (1) に対するニュートン型近接勾配法を提案する．

## 2　提案手法とその収束性

本研究の提案手法は反復法であり，点列 $\{x_k\}$ を $x_{k+1}=x_k+\alpha_k d_k$ に従って更新する．ただし，$d_k$ は探索方向，$\alpha_k>0$ はステップ幅である．提案手法では以下の問題の解を探索方向 $d_k$ とする：

$$\min_{d\in\mathbb{R}^n}\max_{i=1,...,m}\nabla f_i(x_k)^Td+\frac{1}{2}d^TB_{k,i}d+g_i(x_k+d)-g_i(x_k). \tag{2}$$

ただし，$B_{k,i}\in\mathbb{R}^{n\times n}$ は $\nabla^2 f_i(x_k)$ の近似行列である．このとき，$\nabla f_i(x_k)^Td+\frac{1}{2}d^TB_{k,i}d$ は点 $x_k$ における関数 $f_i$ の二次近似となっているため，提案手法は準ニュートン型近接勾配法と捉えることができ，各反復において $B_{k,i}=\nabla^2 f_i(x_k)$ とすれば，提案手法はニュートン型近接勾配法と捉えることができる．また，近似行列 $B_{k,i}$ が正定値対称行列であると仮定すると，$\max_{i=1,...,m}\nabla f_i(x_k)^Td+\frac{1}{2}d^TB_{k,i}d+g_i(x_k+d)-g_i(x_k)$ は強凸関数であるため，部分問題 (2) は一意解を持つ．以降では部分問題 (2) の最適値を $t(x_k)$ で表すこととする．次に部分問題 (2) の max 関数の上界を変数 $t\in\mathbb{R}$ と置き，$m$ 本の制約式を持つ最適化問題に変形する．この変形した最適化問題に対し，ラグランジュ乗数 $\lambda\in\mathbb{R}^m$ を用いて，ラグランジュ緩和問題を考える．ここで，$\zeta_i\in\partial_d g_i(x_k+d)$ とすると，ラグランジュ緩和問題の最適性条件は以下のように表される：

$$\sum_{i=1,...,m}\lambda_i=1,\quad \sum_{i=1,...,m}\lambda_i(\nabla f_i(x_k)+B_{k,i}d+\zeta_i)=0, \tag{3}$$

$$\lambda_i\geq 0,\quad \nabla f_i(x_k)^Td+\frac{1}{2}d^TB_{k,i}d+g_i(x_k+d)-g_i(x_k)\leq t,\qquad i=1,...,m, \tag{4}$$

 

$$\sum_{i=1,\ldots,m} \lambda_i \left( \nabla f_i(x_k)^T d + \frac{1}{2} d^T B_{k,i} d + g_i(x_k + d) - g_i(x_k) - t \right) = 0. \tag{5}$$

このとき，$d = d_k$，$t = t(x_k)$ とすると，最適性条件 (3)–(5) を満たすようなラグランジュ乗数 $\lambda_k \in \mathbb{R}^m$ が存在する．

部分問題 (2) を解くことによって，探索方向 $d_k$ を求めたのち，点列 $\{x_k\}$ を更新するためにステップ幅 $\alpha_k$ を決定する直線探索を行う．直線探索では定数 $\rho, \gamma \in (0,1)$ を用いて，$\left\{1, \gamma, \gamma^2, \cdots\right\}$ の中から，以下の直線探索条件を満たすような最大のステップ幅 $\alpha_k$ をバックトラッキング法により選択する：

$$F_i(x_k + \alpha_k d_k) \leq F_i(x_k) + \alpha_k \rho t(x_k), \quad i = 1, ..., m.$$

次に提案手法の大域的収束性に関する定理を示す．

**定理 1**　すべての $k \geq 0$，及び $i = 1, ..., m$ について，任意の $v \in \mathbb{R}^n$ に対して，$M_1 \|v\|^2 \leq v^T B_i v \leq M_2 \|v\|^2$ を満たすような正の定数 $M_1, M_2$ が存在し，準位集合 $L = \{x \in \mathbb{R}^n | F(x) \leq F(x_0)\}$ は有界であると仮定する．このとき，$\lim_{k\to\infty} \|d_k\| = 0$ が成立する，さらに，提案手法によって生成された点列 $\{x_k\}$ の任意の集積点は最適化問題 (1) の最適性条件を満たす点（パレート停留点）$x^*$ となる．

最後に提案手法の局所収束性に関する定理を示す．

**定理 2**　定理 1 の仮定が成立しているとし，さらに，$f_i$ は 2 回連続微分可能な強凸関数であり，$\nabla^2 f_i$ はリプシッツ連続であると仮定する．また，近似行列 $\{B_{k,i}\}$ は Dennis-Moré 条件：

$$\lim_{k\to\infty} \frac{\|(\nabla^2 f_i(x^*) - B_{k,i})(x_{k+1} - x_k)\|}{\|x_{k+1} - x_k\|} = 0, \quad i = 1, ..., m$$

を満たすとする．さらに，部分問題 (2) において $B_{k,i} = \nabla^2 f_i(x_k)$ としたときの最適解を $d_k^{\mathrm{nt}}$ とし，対応する最適性条件 (3)–(5) のラグランジュ乗数を $\mu_k \in \mathbb{R}^m$ としたとき，十分大きな $k$ について以下が成り立つと仮定する：

$$\|\lambda_k - \mu_k\| \leq o(\|d_k^{\mathrm{nt}} - d_k\|).$$

このとき，十分大きな $k$ について単位ステップ幅 $\alpha_k = 1$ が選択される．さらに，生成された点列 $\{x_k\}$ は最適解 $x^*$ に超 1 次収束する．

## 参考文献


[1] J. Fliege, L.M.G. Drummond, and B.F. Svaiter, Newton's method for multiobjective optimization, SIAM Journal on Optimization, 20 (2009), 602–626.

[2] H. Tanabe, E.H. Fukuda, and N. Yamashita, Proximal gradient methods for multiobjective optimization and their applications, Computational Optimization and Applications, 72 (2019), 339–361.


 

# Randomized submanifold method for optimization on the Stiefel manifold

Andi HAN[1], Pierre-Louis POIRION[1], Akiko TAKEDA[1,2]
[1] 理化学研究所-AIP, [2] 東京大学
e-mail : andi.han@riken@jp;pierre-louis.poirion@riken@jp;takeda@mist.i.u-tokyo.ac.jp

## 1 Introduction

Let us consider the following optimization on the Stiefel Manifold:

$$\min_{X\in\mathbb{R}^{n\times p}:X^\top X=I_p} F(X), \tag{1}$$

where $F$ is a smooth real-valued function. The problem is defined on the widely known Stiefel manifold $\mathrm{St}(n,p) := \{X \in \mathbb{R}^{n\times p} : X^\top X = I_p\}$, where $p \le n$.

This problem can be solved by the Riemannian gradient descent method that iteratively updates

$$X_{k+1} = \mathrm{Retr}_{X_k}\big(-\eta_k \mathrm{grad}F(X_k)\big)$$

for some stepsize $\eta_k$. This involves computation of Riemannian gradient, $\mathrm{grad}F(X_k)$, and the use of a retraction map, Retr, that ensures the feasibility of the iterates. However, the computational bottleneck is the computation of the retraction, $\mathrm{Retr}_{X_k}$, which requires typically $O(np^2)$ operations. In this talk, we aim to introduce a subspace method to solve (1). In the Euclidean setting, subspace methods consist in introducing, at each iteration, a function $F_k$ consisting of the function $F$ restricted to a subspace generated by a $r \times n$ (with $r < n$) matrix $P_k^\top$:

$$\forall u \in \mathbb{R}^r,\ F_k(u) = F(x_k + P_k^\top u),$$

where $\{x_k\}_{k\in\mathbb{N}}$ is the sequence of iterates. By computing a descent direction for $F_k$ at $u = 0$, we compute the next iterates. For references, see [1] or [2] for example.

Given a known point $X_k \in \mathrm{St}(n,p)$, Let us re-parameterize the next iterate as $X_{k+1} = UX_k$ for $U \in \mathcal{O}(n)$, where $\mathcal{O}(n)$ denotes the orthogonal group of dimension $n$, and consider

$$X_{k+1} = \operatorname*{arg\,min}_{U\in\mathcal{O}(n)} F(UX_k).$$

We introduce the random subgroup of $\mathcal{O}(n)$ parameterized as follows. For any $P \in \mathcal{O}(n)$,

$$U_P(Y) = P^\top \begin{pmatrix} Y & 0 \\ 0 & I_{n-r} \end{pmatrix} P.$$

Hence, given a matrix $P \in \mathcal{O}(n)$, we consider the following optimization problem

$$X_{k+1} = \operatorname*{arg\,min}_{Y\in\mathcal{O}(r)} \widetilde{F}_P(Y) := F(U_P(Y)X_k),$$

over $\mathcal{O}(r)$ where $1 \ge r < n$ is a parameter. The aim is to approximately minimize $\widetilde{F}_P(Y)$ in terms of $Y$ for every iteration around $X_k$.

Let $\mathcal{P}(n)$ be the set of $n\times n$ permutation matrices. The framework is summarized in Algorithm 1.

 

**Algorithm 1** Random subgroup method

1: Initialize $X_0 \in \mathrm{St}(n,p)$.
2: **for** $k = 0, ..., K-1$ **do**
3: Randomly sample $P_k \in \mathcal{P}(n)$ and let $\widetilde{F}(Y) = F(U_{P_k}(Y)X_k)$
4: Compute Riemannian gradient $\mathrm{grad}\widetilde{F}_{P_k}(I_r)$.
5: Update $Y_k = \mathrm{Retr}_{I_r}(-\eta\,\mathrm{grad}\widetilde{F}_{P_k}(I_r))$.
6: Set $X_{k+1} = U_{P_k}(Y_k)X_k$.
7: **end for**

## 2 Analysis

**Assumption 1.** *$F(X)$ has bounded gradient and Hessian in the ambient space, i.e., $\|\nabla F(X)\| \le C_0$, $\|\nabla^2 F(X)[U]\| \le C_1\|U\|$ for any $X \in \mathrm{St}(n,p)$, $U \in \mathbb{R}^{n\times p}$.*

**Theorem 1.** *Under Assumption 1 and select $\eta = \frac{1}{L}$ with $L = C_0 + C_1$, we obtain that for all $k \ge 1$,*

$$\mathbb{E}\left[\min_{i\le k}\|\mathrm{grad}F(X_i)\|\right] \le \frac{1}{\sqrt{k+1}}\sqrt{\frac{4}{L}\frac{n(n-1)}{r(r-1)}(F(X_0)-F^*)}. \tag{2}$$

## 3 Experiments

To test our method, we perform numerical experiments on the following problems:

- The Procrustes problem
$$\min_{X\in\mathrm{St}(n,p)} F(X) = \|XA - B\|^2$$
for some matrix $A \in \mathbb{R}^{p\times p}, B \in \mathbb{R}^{n\times p}$.

- The PCA problem
$$\min_{X\in\mathrm{St}(n,p)} F(X) = -trace(X^\top AX)$$
where $A \in \mathbb{R}^{n\times n}$.

- The quadratic assignment problem solves
$$\min_{X\in\mathrm{St}(n,n)} F(X) = trace(A^\top(X\odot X)B(X\odot X)^\top)$$

## 参考文献


[1] Gower R, Kovalev D, Lieder F, Richtárik P (2019) Rsn: Randomized subspace newton. *Advances in Neural Information Processing Systems* 32.

[2] Kozak D, Becker S, Doostan A, Tenorio L (2021) A stochastic subspace approach to gradient-free optimization in high dimensions. *Computational Optimization and Applications* 79(2):339–368.


 

# ヘッセ行列を含む連続力学系モデルとそれに対応する最適化手法の収束率の改善の試みについて

田部井 淳志 [1], 田中 健一郎 [1]
[1] 東京大学 大学院情報理工学系研究科
e-mail : atsushi-tabei2001@g.ecc.u-tokyo.ac.jp

## 1　背景：最適化手法と連続力学系における収束レートの対応

本研究においては，特定のクラスに属する関数 $f:\mathbb{R}^n \to \mathbb{R}$ について，次の最小化問題を考える．

$$\min_{x\in\mathbb{R}^n} f(x)$$

これに対する基本的な方法として，次の更新式で表される最急降下法を用いたものがある．

$$x_{k+1} = x_k - \tau\nabla f(x_k)$$

さらに，これを次のように「連続版」に対応させて解析する研究がなされている．

$$\dot{x}(t) = -\nabla f(x(t))$$

この対応の利点として，次の表に示すように，ステップ $k$ に対する関数値 $f(x_k)$ や時刻 $t$ に対する関数値 $f(x(t))$ と最適値 $f(x_*)$ の差の $k$ や $t$ に対する縮まり方（以後，これを収束レートと呼ぶ）のオーダーの対応が指摘されている（表の対応は目的関数 $f$ が凸関数の場合のものである）．

| 更新式 | $x_{k+1} = x_k - \tau\nabla f(x_k)$ | $\dot{x}(t) = -\nabla f(x(t))$ |
|---|---|---|
| 収束レート | $O(\frac{1}{k})$ | $O(\frac{1}{t})$ |

このような対応関係を前提として，より良い収束レートをもつ連続力学系が考えられている．

## 2　背景：リアプノフ関数による収束レートの評価とリアプノフ関数の機械的導出

連続力学系に対する収束レートの評価方法として，リアプノフ関数を用いたものが存在する．

**定義 1（リアプノフ関数）**　ある常微分方程式に従う連続力学系に対して定められる関数 $\mathcal{E}(t)$ に対して，$\mathcal{E}(t)$ が単調非増加で下に有界な時，$\mathcal{E}(t)$ をその連続力学系のリアプノフ関数と呼ぶ．

**補助定理 2（リアプノフ関数を用いた収束レートの導出**, cf.Kamijima et al.(2024)[2]**）**　ある連続力学系に対し，式 (1) の形のリアプノフ関数 $\mathcal{E}(t)$ が存在すれば式 (2) の収束レートが得られる．

$$\mathcal{E}(t) = \mathcal{F}(t) + e^{\gamma(t)}(f - f_*), \mathcal{F}(t) \geq 0 \tag{1}$$

$$f - f_* = O(e^{-\gamma(t)}) \tag{2}$$

この方法を利用する場合，基本的な方針としては式 (1) の形を満たしつつ，$e^{\gamma(t)}$ がなるべく大きくなるようなリアプノフ関数 $\mathcal{E}(t)$ を得ることを目指すことになる．しかし，このリアプノフ関数 $\mathcal{E}(t)$ は従来発見的に作られており，一般の目的関数 $f$ に対してリアプノフ関数を与えることは困難だった．これを受け，Suh et al.(2022)[1] では，リアプノフ関数 $\mathcal{E}(t)$ を機械的に導出する方法を提案し，一次法に対応する（$\nabla f$ までしか含まれない）いくつかの連続力学系に対して実際に導出した．Kamijima et al.(2024)[2] はこれを発展させ，二次法に対応する（$\nabla^2 f$ が含まれる）連続力学系にもこの手法が適用できることを確認し，得られたリアプノフ関数を用いた収束レートの解析を行った．

 

しかし，その中で扱われたものの一つである連続力学系 $\dot{x}+a\nabla^2 f\dot{x}+\nabla f=0$ について，Kamijima et al.(2024)[2] では目的関数 $f$ が $\mu$-強凸であるならば $a=0$ とした場合の収束レート $O(e^{-\mu t})$ が最速であるとされた．しかし，この連続力学系とその収束レートは最急降下法に対応するものである．$\nabla^2 f$ の使用を許している（対応する最適化手法としては二次法の枠に入ることを許している）にも関わらず，それを利用しない最急降下法に対応する連続力学系が最善の収束レートを示すということである．この結論の不自然さに着目し，本研究では連続力学系 $\dot{x}+a\nabla^2 f\dot{x}+\nabla f=0$ を用いて最急降下法に対応する連続力学系の収束レート $O(e^{-\mu t})$ よりも良い収束レートを得られないか検討した．

## 3　連続力学系 $\dot{x}+a\nabla^2 f\dot{x}+\nabla f=0$ に対するリアプノフ関数

本研究においては，Kamijima et al.(2024)[2] の定理 10 と同じリアプノフ関数を用いつつ，$a$ が満たすべき条件を $a\geq 0$ から $a>-\frac{1}{L}$ に緩和し，それにより収束レートを改善させた．

**定理 3（連続力学系 $\dot{x}+a\nabla^2 f\dot{x}+\nabla f=0$ のリアプノフ関数）** $\mu$-強凸かつ $L$-平滑である関数 $f$ に対して，関数

$$\mathcal{E}(t)=\frac{1}{2}(1+\mu a)\dot{\gamma}e^{\gamma}\left\|x-x_*\right\|^2+a\dot{\gamma}e^{\gamma}\left(f_*-f-\langle\nabla f,x_*-x\rangle-\frac{\mu}{2}\left\|x-x_*\right\|^2\right)+e^{\gamma}(f-f_*)$$

は連続力学系 $\dot{x}+a\nabla^2 f\dot{x}+\nabla f=0$ のリアプノフ関数となっていて，特に式 (1) の形をしている．ただし，$a$ 及び $\gamma$ は以下の 4 式を満たすように取る．

$$\dot{\gamma}\geq 0,\ a>-\frac{1}{L},\ \ddot{\gamma}+\dot{\gamma}^2+\mu\left(\dot{a}\dot{\gamma}+a\ddot{\gamma}+a\dot{\gamma}^2-\dot{\gamma}\right)\leq 0,\ \dot{a}\dot{\gamma}+a\ddot{\gamma}+a\dot{\gamma}^2-\dot{\gamma}\leq 0$$

**定理 4（連続力学系 $\dot{x}+a\nabla^2 f\dot{x}+\nabla f=0$ の収束レート）** $\mu$-強凸かつ $L$-平滑である関数 $f$ に対して，定理 3 を用いて証明可能な最速の収束レートは微小な $\varepsilon>0$ を用いて，$O\left(e^{-\mu\frac{1}{1-\frac{\mu}{L+\varepsilon}}t}\right)$ である．

## 4　まとめ

本研究においては，Kamijima et al.(2024)[2] から発展して連続力学系 $\dot{x}+a\nabla^2 f\dot{x}+\nabla f=0$ に対して目的関数 $f$ が $\mu$-強凸かつ $L$-平滑であれば最急降下法よりも速い収束レート $O\left(e^{-\mu\frac{1}{1-\frac{\mu}{L+\varepsilon}}t}\right)$ を取りうることを示した．これは最急降下法に対応する連続力学系に比べ，ヘッセ行列の導入によって収束レートが（条件数が大きい場合は特に）改善されることを意味している．したがって，この連続力学系に対応する最適化手法についても，最急降下法と比べた収束レートの改善が期待できる．

一方，ニュートン法などと比べると収束レート自体は悪いので，ヘッセ行列を勾配で近似して一次法の枠に収めるなどの工夫が必要である．また，今回は強凸かつ平滑な関数を対象にしたが，制約を凸のみに緩和するなど，多くの課題が残っている．

## 参考文献


[1] Suh, Jaewook J and Roh, Gyumin and Ryu, Ernest K, Continuous-time analysis of accelerated gradient methods via conservation laws in dilated coordinate systems, International Conference on Machine Learning, 2022, 206420–20667.

[2] Tomoya Kamijima and Shun Sato and Kansei Ushiyama and Takayasu Matsuo and Ken'ichiro Tanaka, Analysis of continuous dynamical system models with Hessians derived from optimization methods, JSIAM Letters, 16. (2024), 29-32.


 