数理解析研究所講究録 第2094巻 2018年 55-64

幾何学的な最適化アルゴリズムとその応用

京都大学・白眉センター/大学院情報学研究科・数理工学専攻　佐藤 寛之

Hiroyuki Sato[1]

The Hakubi Center for Advanced Research/

Department of Applied Mathematics and Physics, Graduate School of Informatics,

Kyoto University

概 要

本稿では，幾何学的な最適化アルゴリズムの概要および最近の研究について解説する．基本的なアルゴリズムである最急降下法，および大規模問題に有効な共役勾配法と確率的分散縮小勾配法を取り上げる．

1 はじめに

実数値関数の最小化問題または最大化問題を最適化問題といい，最小化または最大化すべき問題を目的関数という．関数 f の最大化問題は $-f$ の最小化問題と等価であるため，以下では最小化問題のみを議論する．最適化問題は工学をはじめとする多くの分野において現れる重要な研究対象である．目的関数の変数が連続的である連続最適化問題は制約なし問題と制約付き問題に大別され，制約条件を満たす変数全体の集合を実行可能領域という．制約なし問題に対する解法としては最急降下法や共役勾配法，ニュートン法などがある一方で，これらの手法は制約条件の情報を用いないため，制約付き問題にそのまま適用することはできない．しかしながら，実行可能領域がリーマン多様体 $\mathcal{M}$ であるような制約付き最適化問題は，$\mathcal{M}$ 上の制約なし問題と見なすこともできる．したがって，ユークリッド空間における制約なし最適化問題に対する解法を $\mathcal{M}$ 上に拡張することで，こうした問題に対する解法を与えることができる．さらに，共役勾配法の大域的収束性やニュートン法の局所的2次収束性など，ユークリッド空間における元の手法をその性質が引き継がれるように $\mathcal{M}$ 上に拡張することで，効率的な解法を導出することができる．

最適化理論をリーマン多様体上に拡張する過程は多くの困難を伴うが，指数写像や平行移動といった概念を用いることで，ニュートン法や共役勾配法など，理論的にはユークリッド空間のものと同様のアルゴリズムがリーマン多様体でも考えられることが [13] で示された．また，[4] では，シュティーフェル多様体およびグラスマン多様体上の最適化問題に対してニュートン法や共役勾配法などが現実の問題や数値計算例とともに議論され，こうした幾何学的な最適化の実用性が示された．さらに，一般のリーマン多様体上の最適化問題に対する議論がなされ，より効率的な最適化アルゴリズムの提案およびその解析が [1] にまとめられている．その後の約10年では，機械学習などにおける大規模最適化問題を解く必要性から，ユークリッド空間における確率的最適化手法が盛んに研究されて

[1]e-mail: hsato[AT]amp.i.kyoto-u.ac.jp

きたと同時に，それらのリーマン多様体への拡張の研究も多くなされている．本稿では，リーマン多様体上の最適化の概要を説明するとともに，当該分野の近年の発展について述べる．とくに，著者が関わってきた共役勾配法および確率的分散縮小勾配法について詳細を述べる．

本稿の構成は次の通りである．2 節ではリーマン多様体上の最適化の一般論を概説し，レトラクションなど，最適化において必要となる幾何学的な概念を導入する．3 節ではユークリッド空間における線形共役勾配法および非線形共役勾配法を復習した後，リーマン多様体上の共役勾配法について述べる．4 節ではリーマン多様体上の確率的分散縮小勾配法の概要を述べる．5 節で本稿のまとめを行う．

2 リーマン多様体上の最適化について

$(\mathcal{M}, g)$ をリーマン多様体とする．ここで，g は $\mathcal{M}$ のリーマン計量であり，$\mathcal{M}$ 上の $x \in \mathcal{M}$ における接空間 $T_x\mathcal{M}$ の g から定まる内積を $\langle \cdot, \cdot \rangle_x$ と書く．以降では $(\mathcal{M}, g)$ を単に $\mathcal{M}$ と表記する．本稿では，次のリーマン多様体 $\mathcal{M}$ 上の制約なし最適化問題を扱う．

問題 2.1.

$$\begin{aligned} &\text{minimize} \quad f(x), \\ &\text{subject to} \quad x \in \mathcal{M}. \end{aligned}$$

ユークリッド空間 $\mathbb{R}^n$ 上の目的関数 f の無制約最小化については，直線探索法とよばれるアプローチがあり，初期点 $x_0 \in \mathbb{R}^n$ および

$$x_{k+1} = x_k + \alpha_k \eta_k, \quad k = 0, 1, 2, \ldots \tag{2.1}$$

によって点列 $\{x_k\}$ を生成する．ここで，$\alpha_k > 0$ はステップ幅，$\eta_k \in \mathbb{R}^n$ は探索方向とよばれる．探索方向は目的関数 f の勾配やヘッセ行列などの情報を用いて計算される．

一方，問題 2.1 の実行可能領域 $\mathcal{M}$ では (2.1) の右辺が一般には定義されないので，直線探索を行うことはできない．また，多様体から飛び出すような方向に探索をするのは望ましくない．そこで，点 $x_k \in \mathcal{M}$ を用いて次の点 $x_{k+1} \in \mathcal{M}$ を計算する際に，まず探索方向 η_k を x_k での接ベクトルとして選ぶことにする．さらに，x_k から $\eta_k \in T_{x_k}\mathcal{M}$ の方向に伸びる曲線 γ_{x_k,η_k} を考える．すなわち，曲線 γ_{x_k,η_k} に条件

$$\gamma_{x_k,\eta_k}(0) = x_k, \quad \dot{\gamma}_{x_k,\eta_k}(0) = \eta_k \tag{2.2}$$

を課す．この曲線 γ_{x_k,η_k} 上の $t = \alpha_k$ での f の値 $f(\gamma_{x_k,\eta_k}(\alpha_k))$ が $f(x_k)$ から十分減少するようにステップ幅 $\alpha_k > 0$ を求めることで，x_{k+1} を $x_{k+1} = \gamma_{x_k,\eta_k}(\alpha_k)$ により計算する．曲線 γ_{x_k,η_k} は各点 x_k および探索方向 η_k ごとに異なるものであるが，$\mathcal{M}$ 全体で共通の生成方法があれば便利である．そこで，点 x および x での接ベクトル η の対 (x, η) を引数とする写像，すなわち接バンドル $T\mathcal{M}$ から $\mathcal{M}$ への写像 R を考えよう．$R(x, \eta)$ を $R_x(\eta)$ と書くことにし，x における接空間 $T_x\mathcal{M}$ における零ベクトルを 0_x とする．曲線 $\gamma_{x_k,\eta_k}(t) := R_{x_k}(t\eta_k)$

を考えると，$\gamma_{x_k,\eta_k}(0)=R_{x_k}(0_{x_k})$, $\dot{\gamma}_{x_k,\eta_k}(0)=\mathrm{D}R_{x_k}(0_{x_k})[\eta_k]$ であるから，(2.2) が成り立つための条件は，

$$R_{x_k}(0_{x_k})=x_k,\quad \mathrm{D}R_{x_k}(0_{x_k})[\eta_k]=\eta_k$$

である．この条件が点 x_k や探索方向 η_k に依らずに成り立つとき，R をレトラクションという．すなわち，レトラクションは次のように定義される．

定義 2.1. 次の性質を満たす写像 $R:T\mathcal{M}\to\mathcal{M}$ を $\mathcal{M}$ 上のレトラクションという．

1. $R_x(0_x)=x,\quad x\in\mathcal{M}.$
2. $\mathrm{D}R_x(0_x)[\eta]=\eta,\quad \eta\in T_x\mathcal{M},\quad x\in\mathcal{M}.$

レトラクションの定義から，$\gamma_{x_k,\eta_k}(t):=R_{x_k}(t\eta_k)$ によって曲線 γ_{x_k,η_k} を定めれば，γ_{x_k,η_k} は上記の性質を満たす曲線となる．したがって，x_k から x_{k+1} を計算するための更新式は

$$x_{k+1}=R_{x_k}(\alpha_k\eta_k) \tag{2.3}$$

となる．

ユークリッド空間の場合と同様に，探索方向の計算では目的関数の勾配やヘシアンを用いる．ここで，リーマン多様体 $\mathcal{M}$ 上の勾配やヘシアンは，リーマン計量から定まる量であり，それぞれ $\mathcal{M}$ 上の接ベクトルおよび接空間における線形変換であることに注意されたい．たとえば f の $\mathcal{M}$ 上の点 x における勾配 $\operatorname{grad} f(x)$ は，

$$\langle \operatorname{grad} f(x),\xi\rangle_x=\mathrm{D}f(x)[\xi],\quad \xi\in T_x\mathcal{M}$$

を満たす，$T_x\mathcal{M}$ の一意的なベクトルとして定義される．

更新式 (2.3) における探索方向を $\eta_k=-\operatorname{grad} f(x_k)$ として反復を行うアルゴリズムを最急降下法という．ベクトル $-\operatorname{grad} f(x_k)/\|\operatorname{grad} f(x_k)\|_{x_k}$ は，ノルムが 1 であるようなあらゆる $\eta\in T_{x_k}\mathcal{M}$ の中で

$$\left.\frac{d}{dt}f(R_{x_k}(t\eta))\right|_{t=0}=\mathrm{D}f(R_{x_k}(0_{x_k}))[\mathrm{D}R_{x_k}(0_{x_k})[\eta]]=\mathrm{D}f(x_k)[\eta]=\langle \operatorname{grad} f(x_k),\eta\rangle_{x_k}$$

を最小にするものであることから，$-\operatorname{grad} f(x_k)$ は x_k における最急降下方向とよばれる．この意味で，最急降下法は自然な最適化アルゴリズムではあるが，その収束は遅いことが知られており，多くの改良アルゴリズムが研究されている．

次節以降では，そのような研究の中からリーマン多様体上の共役勾配法と確率的分散縮小勾配法を取り上げ議論する．なお，ユークリッド空間における共役勾配法では，反復の最初の探索方向は最急降下法と同じく最急降下方向とするが，その次以降の反復では，最急降下方向と 1 つ前の探索方向のスカラー倍の和として探索方向を計算する．リーマン多様体 $\mathcal{M}$ 上では x_k における勾配 $\operatorname{grad} f(x_k)$ や探索方向 η_k は $T_{x_k}\mathcal{M}$ の元であるため，x_k における勾配 $\operatorname{grad} f(x_k)\in T_{x_k}\mathcal{M}$ と，1 つ前の探索方向 $\eta_{k-1}\in T_{x_{k-1}}\mathcal{M}$ は異なる接空間に属している．したがって，これらを足し合わせることはできず，η_{k-1} を $T_{x_k}\mathcal{M}$ に写す必要がある．これを実現するのが，平行移動の概念を一般化した vector transport とよばれる写像である [1]．

定義 2.2. 次の性質を満たす写像 $\mathcal{T}:\bigcup_{x\in\mathcal{M}}(T_x\mathcal{M}\times T_x\mathcal{M})\to T\mathcal{M}$ を $\mathcal{M}$ 上の vector transport という．ただし，$\xi,\eta\in T_x\mathcal{M}$ に対して，$\mathcal{T}(\xi,\eta)$ のことを $\mathcal{T}_\xi(\eta)$ と書くことにする．

1. レトラクション R が存在して，$\mathcal{T}_\xi(\eta)\in T_{R_x(\xi)}\mathcal{M},\ \xi,\eta\in T_x\mathcal{M},\ x\in\mathcal{M}$.
2. $\mathcal{T}_{0_x}(\eta)=\eta,\ \eta\in T_x\mathcal{M},\ x\in\mathcal{M}$.
3. $\mathcal{T}_\xi(a\eta+b\zeta)=a\mathcal{T}_\xi(\eta)+b\mathcal{T}_\xi(\zeta),\ a,b\in\mathbb{R},\ \xi,\eta,\zeta\in T_x\mathcal{M},\ x\in\mathcal{M}$.

共役勾配法は vector transport が有効にはたらく最たる例であるが，確率的分散縮小勾配法など，他の最適化手法でも同様の困難を解決するために vector transport が用いられる．

3 共役勾配法

3.1 線形共役勾配法と非線形共役勾配法

ニュートン法は局所的 2 次収束性をもつ最適化手法であり，その収束性はリーマン多様体 $\mathcal{M}$ 上でも成り立つことが証明されている [1, 13]．しかしながら，ニュートン法の各反復では $\mathcal{M}$ の次元 d と同じ次元の線形方程式を解く必要があり，d が非常に大きいような大規模な最適化問題では，ニュートン方程式を繰り返し解くのは困難である．ニュートン方程式を現実的な時間で解くことができる場合でも，ニュートン法が大域的収束性をもたないことから，事前に何らかの別の方法で最適解に十分近い近似解を得る必要がある．

このような問題を解決するために，共役勾配法や準ニュートン法が用いられる．本節で議論する共役勾配法は大域的収束性をもち，1 反復における計算コストも最急降下法とあまり変わらない一方で，収束は最急降下法に比べて非常に速いという性質をもつ．

ユークリッド空間における最初の共役勾配法は $x\in\mathbb{R}^n$ についての線形方程式

$$Ax=b \tag{3.1}$$

の解法として提案された手法である [6]．ここで，A は n 次正定値対称行列，b は n 次元ベクトルである．A の正定値対称性から

$$f(x):=\frac{1}{2}x^\top Ax-b^\top x \tag{3.2}$$

は凸関数であり，f の極小点は f の最小点でもある．したがって，f の最小点 x_* は，ユークリッド勾配 ∇f について

$$\nabla f(x_*)=Ax_*-b=0$$

を満たすので，方程式 (3.1) の解でもある．ゆえに，(3.1) を解くためには 2 次関数 (3.2) を最小化すればよい．反復アルゴリズムによって x_* に収束するような点列 $\{x_k\}$ を生成することが目標になるが，共役勾配法における点 x_k での探索方向 η_k としては，$k=0$ については $\eta_0=-\nabla f(x_0)$, その後は最急降下方向 $-\nabla f(x_k)$ を改良したものとして，

$$\eta_k=-\nabla f(x_k)+\beta_k\eta_{k-1},\quad k=1,2,\dots \tag{3.3}$$

が用いられる．ここで，β_k はスカラーである．また，A を $A = R^\top R$ とコレスキー分解し，変数変換 $y = Rx$ を行うと，

$$f(x) = \frac{1}{2} y^\top y - \left(R^{-\top} b\right)^\top y =: g(y)$$

となる．$g(y)$ については，標準内積の下で直交する n 個の方向ベクトル $\eta'_0, \eta'_1, \ldots, \eta'_{n-1}$ に対して逐次的に正確な直線探索によって最小化を行うことで，高々n 反復で最適解が求まる．$f(x)$ に議論を戻し，探索方向を $\eta_i := R^{-1}\eta'_i,\ i = 0, 1, 2, \ldots, n-1$ とすると，直交性 $(\eta'_i)^\top \eta'_j = 0, i \neq j$ は

$$\eta_i^\top A \eta_j = 0, \quad i \neq j \tag{3.4}$$

と等価である．探索方向がこの条件を満たすように，すなわち A に関して共役となるように β_k を決める．そのための必要条件 $\eta_k^\top A \eta_{k-1} = 0$ から

$$\beta_k = \frac{\nabla f(x_k)^\top A \eta_{k-1}}{\eta_{k-1}^\top A \eta_{k-1}} \tag{3.5}$$

と定まるが，逆に，(3.5) が成り立てば (3.4) が成り立つことを示すことができる．さらに，(3.5) の β_k について，

$$\beta_k = \frac{\nabla f(x_k)^\top \nabla f(x_k)}{\nabla f(x_{k-1})^\top \nabla f(x_{k-1})} \tag{3.6}$$

や

$$\beta_k = \frac{\nabla f(x_k)^\top \nabla f(x_k)}{\eta_{k-1}^\top \left(\nabla f(x_k) - \nabla f(x_{k-1})\right)} \tag{3.7}$$

が成り立つことも証明できる．また，点 x_k と探索方向 η_k が定まったとき，正確な直線探索によるステップ幅 α_k は

$$\alpha_k = \mathop{\arg\min}_{t>0} f(x_k + t\eta_k) = -\frac{\eta_k^\top \nabla f(x_k)}{\eta_k^\top A \eta_k} = \frac{\nabla f(x_k)^\top \nabla f(x_k)}{\eta_k^\top A \eta_k} \tag{3.8}$$

と陽に書ける．式 (2.1), (3.3), (3.6), (3.8) による反復法が共役勾配法である．これは，後に述べる非線形共役勾配法と区別するため，とくに線形共役勾配法ともよばれる．

さて，ここまでは線形方程式 (3.1) を解くために (3.2) で定義される目的関数 f の最小化を考えたが，式 (2.1), (3.3), (3.6) は f が 2 次関数でなくとも，f がユークリッド空間で定義される関数であり勾配 ∇f が存在すればそのまま用いることができる．そこで，線形共役勾配法に基づいて，一般の目的関数 f の最小化手法である非線形共役勾配法を考えよう．以下では非線形共役勾配法を単に共役勾配法ともよぶことにする．

(3.8) は目的関数が 2 次関数であることを利用して得られた式であるためそのまま用いることはできないが，x_k におけるステップ幅 α_k は $\phi(t) := f(x_k + t\eta_k)$ をある程度小さくするような t として選ぶことにする．ここで，定数 c_1, c_2 を $0 < c_1 < c_2 < 1$ を満たすものとする．$\phi(t)$ が $\phi(0) = f(x_k)$ より十分小さくなっていることを課すアルミホ条件

$$\phi(t) \leq \phi(0) + c_1 t \phi'(0) \tag{3.9}$$

や，(3.9) に加えて，t における ϕ の接線の傾き $\phi'(t)$ が $t=0$ での接線の傾き $\phi'(0)$ より十分大きいという条件

$$\phi'(t) \geq c_2\phi'(0) \tag{3.10}$$

を課すウルフ条件が提案されている．なお，(3.10) が意味をもつのは η_k が降下方向であるとき，すなわち $\phi'(0) = \nabla f(x_k)^\top \eta_k$ が負であるときであるが，β_k を適切に定めれば $\phi'(0) < 0$ は保証される．また，接線の傾きが 0 に近くなるようなステップ幅を選ぶほど正確な直線探索に近づくので，(3.10) の代わりに

$$|\phi'(t)| \leq c_2|\phi'(0)| \tag{3.11}$$

も課すこともあり，(3.9) と (3.11) をあわせて強ウルフ条件という．これらを f について書き直すことで，直線探索の際のステップ幅が満たすべき条件を書き下すことができる．

なお，線形共役勾配法の場合，すなわち目的関数が (3.2) の場合は β_k を (3.6) としても (3.7) としても理論的には同じであるが，一般の目的関数の場合はこれらは異なる値を与える．非線形共役勾配法では，(3.6) を Fletcher–Reeves の β, (3.7) を Dai–Yuan の β とよぶ．β の計算方法は他にも多く提案されており，それぞれについて理論的な解析の研究が行われている [5].

3.2 リーマン多様体上の共役勾配法

2 節で述べたように，リーマン多様体 $\mathcal{M}$ 上の反復法ではレトラクション R を用いて式 (2.3) により点列を生成する．また，探索方向の計算ではユークリッド空間の場合の (3.3) において勾配 $\nabla f(x_k)$ を $\mathcal{M}$ 上の勾配 $\operatorname{grad} f(x_k)$ に変更する必要があるだけでなく，$\operatorname{grad} f(x_k) \in T_{x_k}\mathcal{M}$ および $\eta_{k-1} \in T_{x_{k-1}}\mathcal{M}$ が異なるベクトル空間に属していて足し合わせることができないという問題がある．これを解決するため，2 節で導入した $\mathcal{M}$ 上の vector transport $\mathcal{T}$ を用いて η_{k-1} を $T_{x_k}\mathcal{M}$ のベクトルに写す．すなわち，η_k を

$$\eta_k = -\operatorname{grad} f(x_k) + \beta_k \mathcal{T}_{\alpha_{k-1}\eta_{k-1}}(\eta_{k-1}) \tag{3.12}$$

によって計算する．

また，ステップ幅についてはレトラクション R が定める曲線上の探索を考える．すなわち，$\phi(t) := f(R_{x_k}(t\eta_k))$ について (3.9)–(3.11) を f について書き下すことで，条件

$$f(R_{x_k}(t\eta_k)) \leq f(x_k) + c_1 t\langle \operatorname{grad} f(x_k), \eta_k\rangle_{x_k}, \tag{3.13}$$

$$\langle \operatorname{grad} f(R_{x_k}(t\eta_k)), \mathrm{D}R_{x_k}(t\eta_k)[\eta_k]\rangle_{R_{x_k}(t\eta_k)} \geq c_2\langle \operatorname{grad} f(x_k), \eta_k\rangle_{x_k}, \tag{3.14}$$

$$|\langle \operatorname{grad} f(R_{x_k}(t\eta_k)), \mathrm{D}R_{x_k}(t\eta_k)[\eta_k]\rangle_{R_{x_k}(t\eta_k)}| \leq c_2|\langle \operatorname{grad} f(x_k), \eta_k\rangle_{x_k}| \tag{3.15}$$

を得る．$\mathbb{R}^n$ の場合と同様に，(3.13) をアルミホ条件，(3.13) と (3.14) をあわせてウルフ条件，(3.13) と (3.15) をあわせて強ウルフ条件という．

β_k の計算については，(3.6) を多様体に拡張するのは容易である．すなわち，内積をリーマン計量から定まるものにし，

$$\beta_k = \frac{\langle \operatorname{grad} f(x_k), \operatorname{grad} f(x_k)\rangle_{x_k}}{\langle \operatorname{grad} f(x_{k-1}), \operatorname{grad} f(x_{k-1})\rangle_{x_{k-1}}} \tag{3.16}$$

とすればよい．これは Fletcher–Reeves の β を $\mathcal{M}$ 上へ拡張したものである．[8] によってこの β を用いた共役勾配法の大域的収束性がある仮定の下で示された．ただし，以下では目的関数 f やレトラクション R は十分滑らかであるとする．

定理 3.1 ([8])**.** R を $\mathcal{M}$ 上のレトラクションとし，R を用いて

$$\mathcal{T}_{\xi}(\eta) := \mathrm{D}R_x(\xi)[\eta], \quad \xi, \eta \in T_x\mathcal{M}, \quad x \in \mathcal{M} \tag{3.17}$$

により vector transport $\mathcal{T}$ を定義する．また，x_k におけるステップ幅 α_k を，$0 < c_1 < c_2 < 1/2$ とした強ウルフ条件 (3.13), (3.15) を満たす t として選ぶことにする．このとき，Fletcher–Reeves の β を用いた $\mathcal{M}$ 上の共役勾配法 (2.3), (3.12), (3.16) によって生成される点列 $\{x_k\}$ が

$$\|\mathcal{T}_{\alpha_k\eta_k}(\eta_k)\|_{x_{k+1}} \leq \|\eta_k\|_{x_k}, \quad k = 0, 1, 2, \ldots \tag{3.18}$$

を満たすとき，

$$\liminf_{k\to\infty} \|\mathrm{grad}\, f(x_k)\|_{x_k} = 0 \tag{3.19}$$

が成り立つ．

しかしながら，仮定 (3.18) は自然なものではなく，一般には成り立たない．そこで，アルゴリズムに工夫を加え，探索方向の計算を (3.12) の代わりに

$$\eta_k = -\,\mathrm{grad}\, f(x_k) + \beta_k \mathcal{T}^{(k-1)}_{\alpha_{k-1}\eta_{k-1}}(\eta_{k-1}) \tag{3.20}$$

とする方法が [10] で提案された．ここで，

$$\mathcal{T}^{(k)}_{\alpha_k\eta_k}(\eta_k) = \begin{cases} \mathcal{T}_{\alpha_k\eta_k}(\eta_k) & (\|\mathcal{T}_{\alpha_k\eta_k}(\eta_k)\|_{x_{k+1}} \leq \|\eta_k\|_{x_k} \text{ のとき}), \\ \dfrac{\|\eta_k\|_{x_k}}{\|\mathcal{T}_{\alpha_k\eta_k}(\eta_k)\|_{x_{k+1}}} \mathcal{T}_{\alpha_k\eta_k}(\eta_k) & (\text{それ以外のとき}) \end{cases} \tag{3.21}$$

である．このアルゴリズムについて次の結果が示されている．

定理 3.2 ([10])**.** R を $\mathcal{M}$ 上のレトラクションとし，R を用いて (3.17) により vector transport $\mathcal{T}$ を定義する．また，x_k におけるステップ幅 α_k を，$0 < c_1 < c_2 < 1/2$ とした強ウルフ条件 (3.13), (3.15) を満たす t として選ぶことにする．このとき，$\mathcal{M}$ 上の共役勾配法 (2.3), (3.16), (3.20), (3.21) によって生成される点列 $\{x_k\}$ について，(3.19) が成り立つ．

定理 3.1 における仮定 (3.18) が定理 3.2 では不要であることに注意されたい．

Fletcher–Reeves の β を用いる共役勾配法ではユークリッド空間上，リーマン多様体上のどちらの場合も，大域的収束性を保証するにはステップ幅が強ウルフ条件を満たす必要がある．しかし，ユークリッド空間において (3.7) で定義される Dai–Yuan の β を用いた共役勾配法においては，ステップ幅がより弱い条件であるウルフ条件 (3.9), (3.10) を満たせば大域的収束性が保証される [3]．式 (3.7) の $\mathcal{M}$ 上への拡張の仕方はいくつかのものが考えられるが，ユークリッド空間における収束性解析とうまく対応付けるには

$$\beta_k = \frac{\langle \mathrm{grad}\, f(x_k), \mathrm{grad}\, f(x_k)\rangle_{x_k}}{\langle \mathrm{grad}\, f(x_k), \mathcal{T}^{(k-1)}_{\alpha_{k-1}\eta_{k-1}}(\eta_{k-1})\rangle_{x_k} - \langle \mathrm{grad}\, f(x_{k-1}), \eta_{k-1}\rangle_{x_{k-1}}} \tag{3.22}$$

とすればよいことが [9] によって示された．ここで，$\mathcal{T}^{(k)}$ は (3.21) で定義されるものである．

定理 3.3 ([10]). R を $\mathcal{M}$ 上のレトラクションとし，R を用いて (3.17) により vector transport $\mathcal{T}$ を定義する．また，x_k におけるステップ幅 α_k を，$0 < c_1 < c_2 < 1$ としたウルフ条件 (3.13), (3.14) を満たす t として選ぶことにする．このとき，Dai–Yuan の β を用いた $\mathcal{M}$ 上の共役勾配法 (2.3), (3.20)–(3.22) によって生成される点列 $\{x_k\}$ について，(3.19) が成り立つ．

ユークリッド空間における共役勾配法も研究が続けられているため，そうした新しいアルゴリズムについてもリーマン多様体上への拡張が望まれる．

4 リーマン多様体上の確率的分散縮小勾配法

本節では，目的関数が

$$f(x) = \frac{1}{N}\sum_{n=1}^{N} f_n(x) \tag{4.1}$$

の形で書かれている場合の最適化問題 2.1 を扱う．共役勾配法は $\mathcal{M}$ の次元が大きい問題に対して有効であることを 3 節で述べたが，本節で扱う確率的な最適化手法は N が非常に大きい場合の目的関数 (4.1) の最小化問題に対して有効な手法である．

目的関数 (4.1) の勾配は N 個の f_n の勾配の和となるから，N が大きいとき，最急降下法における f の勾配の計算コストも大きくなる．そこで，f の勾配の代わりに，ランダムに選んだ番号 n に対して f_n の勾配だけを用いて更新を行うのが確率的勾配降下法である．この考え方はユークリッド空間，リーマン多様体 $\mathcal{M}$ の場合のどちらにも適用することができる．すなわち，$\mathcal{M}$ 上での確率的勾配降下法の反復は，$1, 2, \ldots, n$ のいずれかの値をランダムにとる確率変数 i_k を用いて

$$x_{k+1} = R_{x_k}\left(-\alpha_k \operatorname{grad} f_{i_k}(x_k)\right)$$

となる．$i_k = n$ となる確率が各 $n \in \{1, 2, \ldots, N\}$ について一様に $1/N$ であるとき，$\operatorname{grad} f_{i_k}(x_k)$ の i_k についての期待値は

$$\mathbb{E}[\operatorname{grad} f_{i_k}(x_k)] = \frac{1}{N}\sum_{n=1}^{N} \operatorname{grad} f_n(x_k) = \operatorname{grad} f(x_k)$$

となり，f の勾配 $\operatorname{grad} f(x_k)$ と一致する．したがって，$\operatorname{grad} f_{i_k}(x_k)$ は $\operatorname{grad} f(x_k)$ の確率的近似である．

確率的勾配降下法は確率的最適化の基本的な手法であるが，最急降下法と同様，多くの改良の余地がある．その一つの試みとして，[7] によって提案されたユークリッド空間上での確率的分散縮小勾配法（SVRG）では，$\nabla f_{i_k}(x_k)$ に補正項を付け加えて分散が小さくなるように工夫する．[14] では SVRG が指数写像や平行移動を用いてリーマン多様体上に拡張されているが，実際の数値計算ではレトラクションや vector transport を用いることが望ましい．これを実現した [11] によるアルゴリズム R-SVRG を次ページに示す．

二重のループ構造からなるやや複雑なアルゴリズムであるが，計算コストが小さいが分散の大きい確率的勾配 $\operatorname{grad} f_{i_t^s}$ に，分散が 0 である元の f の勾配 $\operatorname{grad} f$ の情報を取り

アルゴリズム 1 R-SVRG [11]

Require: $m_s > 0$ とステップ幅の列 $\{\alpha_t^s\} > 0$.

$\tilde{x}^0$ を初期化する.

for $s = 1, 2, \ldots$ **do**

  勾配 $\operatorname{grad} f(\tilde{x}^{s-1})$ を計算し，$x_0^s = \tilde{x}^{s-1}$ を保存する.

  for $t = 1, 2, \ldots, m_s$ **do**

    $i_t^s \in \{1, 2, \ldots, N\}$ を一様にランダムに選択する.

    $\tilde{x}^{s-1}$ から x_{t-1}^s へ向かう接ベクトル ζ を $\zeta = R_{\tilde{x}^{s-1}}^{-1}(x_{t-1}^s)$ により計算する.

    ξ_t^s を vector transport を用いて

$$\xi_t^s = \operatorname{grad} f_{i_t^s}(x_{t-1}^s) - \mathcal{T}_\zeta \left(\operatorname{grad} f_{i_t^s}(\tilde{x}^{s-1}) - \operatorname{grad} f(\tilde{x}^{s-1})\right) \tag{4.2}$$

    により計算する.

    レトラクション R を用いて x_t^s を $x_t^s = R_{x_{t-1}^s}\left(-\alpha_{t-1}^s \xi_t^s\right)$ と更新する.

  end for

  $\tilde{x}^s = x_{m_s}^s$.

end for

入れることを目標にしている．そこで，$\operatorname{grad} f$ を毎回の反復で計算するのではなく，回数 $m_1, m_2, \ldots$ を定め，m_s 回の反復が終わって $\tilde{x}^s$ が得られたときに $\operatorname{grad} f(\tilde{x}^s)$ を計算する．これがアルゴリズムにおける外部反復に相当し，点列 $\{\tilde{x}^1, \tilde{x}^2, \ldots\}$ を生成する．一方，外部反復の番号 s を固定したとき，内部反復では点列 $\{x_1^s, x_2^s, \ldots, x_{m_s}^s\}$ を生成するが，この過程を通して $\operatorname{grad} f(\tilde{x}^{s-1})$ を $\operatorname{grad} f(x_{t-1}^s)$, $t = 1, 2, \ldots, m_s$ の近似であると見なし，(4.2) によって $\operatorname{grad} f_{i_t^s}(x_{t-1}^s)$ より分散の小さい ξ_t^s を計算している．容易にわかるように ξ_t^s の平均は $\operatorname{grad} f(x_{t-1}^s)$ に一致する一方で，ξ_t^s の分散は $\operatorname{grad} f_{i_t^s}(x_{t-1}^s)$ の分散より小さくなるため，確率的勾配降下法より収束が速くなる．

本稿では割愛するが，R-SVRG の収束の理論的な結果とともに，良好な数値実験の結果が [11] で示されている．

5 結論

本稿では，リーマン多様体上の最適化理論を概説した後，多様体上の最急降下法とその改良版である共役勾配法，および，確率的勾配降下法とその改良版である確率的分散縮小勾配法について議論した．共役勾配法は多様体の次元が大きい問題，確率的分散縮小勾配法は目的関数が多数の項の和の形であるような問題に対して有効である．

リーマン多様体上の最適化問題として定式化できる問題は数多くあり，固有値問題や特異値分解などの数値線形代数の問題をはじめとして，統計手法や制御理論など応用は多岐にわたる [2,12]．本稿で述べたようなアルゴリズムの理論的性質の研究とともに，多くの実問題への応用も期待される．

謝辞

本研究は JSPS 科研費 JP16K17647 の助成を受けている．

参考文献

[1] P.-A. Absil, R. Mahony and R. Sepulchre, *Optimization Algorithms on Matrix Manifolds*, Princeton University Press, 2008.

[2] K. Aihara and H. Sato, A matrix-free implementation of Riemannian Newton's method on the Stiefel manifold, *Optimization Letters*, **11** (2017), 1729–1741.

[3] Y. H. Dai and Y. Yuan, A nonlinear conjugate gradient method with a strong global convergence property, *SIAM Journal on Optimization*, **10** (1999), 177–182.

[4] A. Edelman, T. A. Arias and S. T. Smith, The geometry of algorithms with orthogonality constraints, *SIAM Journal on Matrix Analysis and Applications*, **20** (1998), 303–353.

[5] W. W. Hager and H. Zhang, A survey of nonlinear conjugate gradient methods, *Pacific Journal on Optimization*, **2** (2006), 35–58.

[6] M. R. Hestenes and E. Stiefel, Methods of conjugate gradients for solving linear systems, *Journal of Research of the National Bureau of Standards*, **49** (1952), 409–436.

[7] R. Johnson and T. Zhang, Accelerating stochastic gradient descent using predictive variance reduction, *Advances in Neural Information Processing Systems (NIPS)*, **26** (2013), 315–323.

[8] W. Ring and B. Wirth, Optimization methods on Riemannian manifolds and their application to shape space, *SIAM Journal on Optimization*, **22** (2012), 596–627.

[9] H. Sato, A Dai–Yuan-type Riemannian conjugate gradient method with the weak Wolfe conditions, *Computational Optimization and Applications*, **64** (2016), 101–118.

[10] H. Sato and T. Iwai, A new, globally convergent Riemannian conjugate gradient method, *Optimization*, **64** (2015), 1011–1031.

[11] H. Sato, H. Kasai and B. Mishra, Riemannian stochastic variance reduced gradient, arXiv preprint, arXiv:1702.05594 (2017).

[12] H. Sato and K. Sato, Riemannian optimal system identification algorithm for linear MIMO systems, *IEEE Control Systems Letters*, **1** (2017), 376–381.

[13] S. T. Smith, Optimization techniques on Riemannian manifolds, *Fields Institute Communications*, **3** (1994), 113–135.

[14] H. Zhang, S. J. Reddi and S. Sra, Riemannian SVRG: Fast stochastic optimization on Riemannian manifolds, *Advances in Neural Information Processing Systems (NIPS)*, **29** (2016), 4592–4560.