凸最適化問題に対する一次法とその理論
—加速勾配法とその周辺—

伊藤　勝

一次法は，この十数年ほどで信号・画像処理や統計学，機械学習などの分野に現れる大規模な凸最適化問題への有用性から活発に研究が行われるようになった．本稿では，一次法研究の一端を担う基礎理論を，加速勾配法とその周辺の観点から解説する．

キーワード：凸最適化問題，一次法，収束率，反復計算量，加速勾配法，近接勾配法，Frank–Wolfe 法

1. はじめに

凸最適化問題に対する一次法は，近接勾配法といった問題の構造を利用した手法によって，2005 年頃から信号・画像処理や統計学，機械学習などの分野における有用性が注目され，現在では加速勾配法などの一次法の存在が広く認識されるようになった．本稿では，現在多様な発展を遂げている一次法の研究について，その一端を担う基礎理論に焦点を当てて，加速勾配法と関連する一次法を取り上げて概論を述べる．具体的には，最急降下法とその一般化である射影劣勾配法を 3 節で導入し，4 節で平滑関数に対する加速化である加速勾配法を解説する．その後，問題構造に特化した手法として注目を集めた近接勾配法（5 節）や Frank–Wolfe 法（6 節）を挙げる．最後に 7 節で，近年の発展的な理論研究につながる話題をいくつか紹介する．

実数値関数の最小化問題に対する一次法は，目的関数の勾配または劣勾配といった一次の情報を用いて近似解の列を生成する反復的アルゴリズムである．最急降下法や共役勾配法は代表的な一次法としてよく知られている．目的関数の二次の情報を用いるニュートン法や内点法と比較すると，一次法は近似解の収束は遅いものの，一反復の計算コストを抑えられるという特徴がある．一次法の性能を決定づけるのは，一反復の計算コストと，近似解の近似誤差の収束率である．一次法の性能は，問題の構造をうまく利用できるかどうかにも大きく依存する．さらに，一次法の内部パラメータであるステップ幅は，収束率に強い影響を与えるため，ステップ幅をどう決めるかという問題も，一次法における興味の対象である．

いとう　まさる
日本大学理工学部数学科
〒 101–8308　東京都千代田区神田駿河台 1–8–14
ito.m@math.cst.nihon-u.ac.jp

2. 凸最適化問題

2.1 準備

本稿では，n 次元実ベクトル空間 $\mathbb{R}^n$ 上の凸最適化問題を対象とし，通常の内積 $\langle x, y\rangle = x^\top y$ およびユークリッドノルム $\|x\|_2 = \sqrt{x^\top x}$ を用いる．

まず，凸解析についていくつかの準備を行う．凸解析についてより詳しくは文献 [1] を参照されたい．集合 $X \subset \mathbb{R}^n$ が凸集合であるとは，任意の $x, y \in X$ と $\lambda \in [0,1]$ に対して $\lambda x + (1-\lambda)y \in X$ となることをいう．閉凸集合 X に対して $x \in \mathbb{R}^n$ から X への距離 $\mathrm{dist}(x, X) := \min_{y\in X} \|x - y\|_2$ が定義できる．

関数 $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ が任意の $x, y \in \mathbb{R}^n$ と $\lambda \in [0,1]$ に対して

$$f(\lambda x + (1-\lambda)y) \le \lambda f(x) + (1-\lambda) f(y)$$

を満たすとき f を凸関数という．また，任意の $\alpha \in \mathbb{R}$ に対してレベル集合 $\{x \in \mathbb{R}^n \mid f(x) \le \alpha\}$ が閉集合となるとき，f は下半連続であるという．

凸関数 f と $x \in \mathbb{R}^n$ に対して，f の x における劣微分を $\partial f(x) = \{g \in \mathbb{R}^n \mid f(y) \ge f(x) + \langle g, y - x\rangle\,,\ \forall y \in \mathbb{R}^n\}$ によって定める．$\partial f(x) \ne \emptyset$ であるとき f は x において劣微分可能といい，$\partial f(x)$ の各元は f の x における劣勾配と呼ばれる．劣勾配は勾配の一般化である．特に，f が x において微分可能であるとき，$\partial f(x) = \{\nabla f(x)\}$ が成り立つ．

2.2 凸最適化問題

凸関数 $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ の閉凸集合 $X \subset \mathbb{R}^n$ 上での最小化問題

$$\min_{x\in X} f(x)$$

を凸最適化問題という．本稿では，この問題の最適値

を f^*，最適解集合を X^* と書く．また，少なくとも一つの最適解 $x^* \in X^*$ が存在すると仮定する．許容誤差 $\varepsilon > 0$ に対して $f(x) - f^* \leq \varepsilon$ を満たす実行可能解 $x \in X$ を ε-近似解と呼ぶ．

一次法の構築・解析のためには，目的関数 f にいくつかの仮定が付加される．これにより問題のクラスが分類され，一次法の性能を大きく左右する．また，一次法の各反復では射影または近接写像といった補助最適化問題を解く必要があるため，制約集合 X や目的関数 f には，この補助最適化問題を効率的に解くことができるような構造が想定される．

2.3 一次法の評価指標

凸最適化問題 $\min_{x\in X} f(x)$ に対して，一次法は各反復で劣勾配や勾配を評価して近似解を更新していく反復的アルゴリズムである．一次法が生成する近似解の点列 $\{x_k\}$ に対して

$$\min_{0\leq i\leq k} f(x_i) \to f^* \quad (k \to \infty) \tag{1}$$

が成り立つとき，その一次法は目的関数値に関して収束するという．ここでは，一次法の性能の評価指標として，近似値 $f_k := \min_{0\leq i\leq k} f(x_i)$ の f^* への収束率に着目する．

- $f_k - f^* \leq c\exp(-rk)$ $(\forall k \geq k_0)$ となる $c, r, k_0 > 0$ が存在するとき，その一次法は一次収束するという．r が大きいほど収束が速い．
- $f_k - f^* \leq ck^{-s}$ $(\forall k \geq k_0)$ となる $c, s, k_0 > 0$ が存在するとき，その一次法は劣一次収束するという．s が大きいほど収束が速い．

もちろん，一次収束は劣一次収束よりも優秀である．

一次法によっては収束性 (1) は保証しないが ε-近似解を得ることは保証できる場合があり，この場合は収束率の代わりに反復計算量によって一次法の性能を測る．許容誤差 ε に対する，ある一次法の反復計算量とは，生成される近似解の点列 $\{x_k\}$ が $f(x_k) - f^* \leq \varepsilon$ を満たす最小の反復回数 k として定義される．したがって，対象とする一次法が収束性 (1) を満たす場合には，収束率を考えることと反復計算量を考えることは本質的に同じである．

3. 最急降下法とその一般化

無制約最適化問題における最急降下法は，一次法のなかでも最も素朴なものの一つであろう．すなわち，制約集合を $X = \mathbb{R}^n$ として微分可能な凸関数 f の無制約最小化問題 $\min_{x\in\mathbb{R}^n} f(x)$ を考えたとき，最急降下法は初期点 $x_0 \in \mathbb{R}^n$ に対して次の反復を繰り返す．

$$x_{k+1} = x_k - \lambda_k \nabla f(x_k), \quad k = 0, 1, 2, \ldots.$$

ここで $\lambda_k > 0$ はステップ幅と呼ばれる内部パラメータである．より一般には，劣微分可能な目的関数 f に対する劣勾配法

$$x_{k+1} = x_k - \lambda_k g_k, \quad g_k \in \partial f(x_k) \tag{2}$$

が考えられる．制約付きの凸最適化問題 $\min_{x\in X} f(x)$ に対して上記の劣勾配法は，閉凸集合 X への直交射影 $\pi_X(x) = \operatorname{argmin}_{z\in X} \|z - x\|_2$ を用いて射影劣勾配法として一般化される．

アルゴリズム 1（射影劣勾配法）**.** 初期点 $x_0 \in X$ をとり，$k = 0, 1, 2, \ldots,$ に対して以下の反復を繰り返す．

$$x_{k+1} := \pi_X(x_k - \lambda_k g_k), \quad g_k \in \partial f(x_k), \ \lambda_k > 0.$$

射影劣勾配法の各反復は，目的関数 f を劣勾配 g_k を用いて二次関数で次のように“近似”して，その最小点を x_{k+1} と更新することと解釈できる．

$$x_{k+1} = \operatorname*{argmin}_{x\in X} \left\{ f(x_k) + \langle g_k, x - x_k\rangle + \frac{1}{2\lambda_k}\|x - x_k\|_2^2 \right\}. \tag{3}$$

射影劣勾配法は，X への直交射影が効率的に計算できるという問題構造を必要とする．

ステップ幅の選択は射影劣勾配法の性能に影響を与える．たとえば，以下の不等式 (4) は射影劣勾配法が満たすよく知られた近似誤差の上界であり，後述する命題 2 などにおいてステップ幅を決定するうえで参考になる [2]（ただし $D := \operatorname{dist}(x_0, X^*)$ とする）．

$$\min_{0\leq i\leq k} f(x_i) - f^* \leq \frac{D^2 + \sum_{i=0}^k \lambda_i^2 \|g_i\|_2^2}{2\sum_{i=0}^k \lambda_i}, \quad \forall k \geq 0. \tag{4}$$

3.1 リプシッツ関数に対する射影劣勾配法

ここでは，凸最適化問題のクラスとして，目的関数がリプシッツ関数であるものを考え，射影劣勾配法の基本的な評価を述べよう．凸関数 f が集合 S と定数 $M > 0$ に対して

$$|f(x) - f(y)| \leq M\|x - y\|_2, \quad \forall x, y \in S$$

を満たすとき，f は S 上で M-リプシッツであるという．S が開集合であるとき，f が S 上で M-リプシッ

ツであることと，以下が成り立つことは同値である．

$$\|g\|_2 \le M, \quad \forall x \in S,\ \forall g \in \partial f(x). \tag{5}$$

命題 2. 凸最適化問題 $\min_{x\in X} f(x)$ について，f は X を含む開集合上で劣微分可能かつ M-リプシッツな凸関数とする．このとき，許容誤差 $\varepsilon > 0$ に対する，ステップ幅 $\lambda_k := \varepsilon/\|g_k\|_2^2$ を用いた射影劣勾配法の反復計算量は高々

$$\frac{M^2D^2}{\varepsilon^2} \tag{6}$$

である．ただし $D = \mathrm{dist}(x_0, X^*)$ とする．

証明. 不等式 (4) の右辺は，$\lambda_k = \varepsilon/\|g_k\|_2^2$ を代入して不等式 (5) を用いれば，次のように上から評価される．

$$\frac{D^2}{2\varepsilon\sum_{i=0}^k \|g_i\|_2^{-2}} + \frac{\varepsilon}{2} \le \frac{M^2D^2}{2\varepsilon(k+1)} + \frac{\varepsilon}{2}.$$

この右辺は，$k+1 \ge \frac{M^2D^2}{\varepsilon^2}$ ならば ε 以下になる． □

反復計算量の上界 (6) の重要な点は，それが本質的にこれ以上は改善できないという事実である．より具体的には，X が球 $\{x \in \mathbb{R}^n \mid \|x - x_0\|_2 \le R\}$ を含むと仮定したとき，任意の“劣勾配法[1]” $\mathcal{A}$ に対してある M-リプシッツな凸関数 f が存在して，$\mathcal{A}$ は問題 $\min_{x\in X} f(x)$ に対して少なくとも $\min\{n, M^2R^2/\varepsilon^2\}$ の反復計算量をもつ [3]．この意味で，反復計算量の上界 (6) はリプシッツ凸関数のクラスに対する“最適な反復計算量”である．

ステップ幅 $\lambda_k := \varepsilon/\|g_k\|_2^2$ を用いても射影劣勾配法は必ずしも収束性 (1) を保証しないことに注意する．X の有界性を仮定すれば，最適な収束率を保証するステップ幅の選択規則がある（たとえば [4]）．

3.2 リプシッツ関数に対する双対平均化法

ここでは最急降下法の別の一般化として Nesterov の双対平均化法 (dual averaging) [5] を紹介する．リプシッツ関数に対する射影劣勾配法について，命題 2 のステップ幅の取り方は反復計算量の意味で最適性を実現するものの，収束性 (1) が保証されず，最適な収束率の実現には X の有界性を仮定する必要があった．一方で，双対平均化法は有界性の仮定がなくとも最適な収束率を実現する．

[1] ここには劣微分の選び方について制限が加わる．

アルゴリズム 3（双対平均化法）. 初期点 $x_0 \in X$ をとり，各 $k = 0, 1, 2, \ldots,$ に対して

$$x_{k+1} := \pi_X\left(x_0 - \frac{1}{\beta_k}\sum_{i=0}^k \lambda_i g_i\right),\ g_k \in \partial f(x_k)$$

とする．ここで，$\lambda_k > 0$ はステップ幅，$\beta_k > 0$ はスケーリングパラメータと呼ばれる．

双対平均化法は特殊ケースとして最急降下法を含む．すなわち，無制約 $X = \mathbb{R}^n$ のときを考えると射影 $\pi_X(\cdot)$ は恒等写像となるから，スケーリングパラメータ $\beta_k \equiv 1$ を用いた双対平均化法の反復は $x_{k+1} := x_0 - \sum_{i=0}^k \lambda_i g_i$ となる．これは更新式 (2) と同等である．

スケーリングパラメータの導入により，次の劣一次収束性が得られる．

命題 4. f は X を含む開集合上で劣微分可能かつ M-リプシッツな凸関数とする．このとき，パラメータ

$$\lambda_k \equiv 1, \quad \beta_k = \gamma\sqrt{k+1}, \quad k = 0, 1, 2, \ldots \quad (\gamma > 0)$$

による双対平均化法は，任意の $k \ge 0$ に対して以下を満たす．ただし，$D = \mathrm{dist}(x_0, X^*)$ とする．

$$\min_{0\le i\le k} f(x_i) - f^* \le \frac{1}{\sqrt{k+1}}\left(\frac{\gamma D^2}{2} + \frac{M^2}{\gamma}\right).$$

この命題から，双対平均化法は $\mathrm{O}(1/\sqrt{k})$ の収束率をもつ．言い換えると，任意の許容誤差 ε に対して $\mathrm{O}(1/\varepsilon^2)$ の（最適な）反復計算量を保証する．ただし，ほかのパラメータ M や D に関しては射影勾配法の反復計算量 (6) に劣る．双対平均化法に対しても (6) を得るには $\gamma = \sqrt{2}M/D$ とする必要がある．

双対平均化法が最適な収束率を実現したことは，スケーリングパラメータの導入によって恩恵を受けた部分が大きい．同様の手法は，射影劣勾配法にも利用できる [6]．

3.3 平滑関数に対する射影勾配法

ここでは，凸最適化問題のクラスとして，目的関数が平滑であるものを考える．凸関数 f が X 上で連続的微分可能であって，勾配の L-リプシッツ連続性

$$\|\nabla f(x) - \nabla f(y)\|_2 \le L\|x - y\|_2, \quad \forall x, y \in X$$

が成り立つとき，f は X 上で L-平滑であるという．X 上の L-平滑な凸関数 f は，任意の $x, y \in X$ に対して次の不等式を満たす．

$$f(x) \le f(y) + \langle \nabla f(y), x-y \rangle + \frac{L}{2}\|x-y\|_2^2 . \quad (7)$$

特に，f が $\mathbb{R}^n$ 上で二階連続的微分可能であるとき，f が $\mathbb{R}^n$ 上で L-平滑であることと，任意の $x \in \mathbb{R}^n$ に対してヘッセ行列 $\nabla^2 f(x)$ の最大固有値が L 以下であることは同値である．

平滑関数に対しては劣勾配と勾配は同じであるから，アルゴリズム 1 を射影勾配法と呼んで，以下のように反復を行う．初期点 $x_0 \in X$ として，

$$x_{k+1} := \pi_X(x_k - \lambda_k \nabla f(x_k)), \quad k = 0, 1, 2, \ldots .$$

以下に示すように射影勾配法は $\mathrm{O}(1/k)$ の収束率を保証する (cf. [7, 8]).

命題 5. X 上の L-平滑関数 f に対する射影勾配法は，ステップ幅を $\lambda_k \equiv \frac{1}{L}$ $(k \ge 0)$ と選択することで，以下の劣一次収束性を保証する．ただし $D = \mathrm{dist}(x_0, X^*)$ とする．

$$f(x_{k+1}) - f^* \le \frac{LD^2}{2(k+1)}, \quad \forall k \ge 0. \quad (8)$$

ステップ幅 $\lambda_k = 1/L$ の選択には目的関数のリプシッツ定数 L を必要とするが，L が未知である場合でも“直線探索法”の活用によってこれを推定しつつ収束率 $\mathrm{O}(1/k)$ を達成できる．

3.3.1 一次収束性

射影勾配法は，目的関数に強凸性を仮定すれば一次収束性を保証する．定数 $\mu \ge 0$ に対して，凸関数 f が凸集合 S 上で μ-強凸であるとは，任意の $x, y \in S$ と $\lambda \in [0,1]$ に対して

$$f(\lambda x + (1-\lambda)y) \le \lambda f(x) + (1-\lambda) f(y) - \frac{\mu}{2}\lambda(1-\lambda)\|x-y\|_2^2$$

が成り立つことをいう．通常の凸性と 0-強凸性は同等である．f が S 上で連続的微分可能であれば，f が S 上で μ-強凸であることと

$$f(x) \ge f(y) + \langle \nabla f(y), x-y \rangle + \frac{\mu}{2}\|x-y\|_2^2 \quad (9)$$

が任意の $x, y \in S$ に対して成り立つことは同値である．特に，S が開凸集合かつ f が S 上で二階連続的微分可能であるとき，f が S 上で μ-強凸であることと，任意の $x \in S$ に対してヘッセ行列 $\nabla^2 f(x)$ の最小固有値が μ 以上であることは同値である．

$\mu > 0$ であれば，μ-強凸関数 f の閉凸集合 X 上の最小化問題 $\min_{x \in X} f(x)$ には最適解 x^* が必ず一意に存在する．μ-強凸関数に対する射影勾配法は，以下の一次収束性をもつ (cf. [7, 9]).

命題 6. 目的関数 f が X 上で L-平滑かつ μ-強凸であるとすれば，固定ステップ幅 $\lambda_k \equiv 1/L$ による射影勾配法は次の一次収束性を満たす．ただし $D := \|x_0 - x^*\|_2$ である．

$$f(x_{k+1}) - f^* \le \frac{LD^2}{2}\exp\left(-k\frac{\mu}{L}\right), \quad \forall k \ge 0. \quad (10)$$

ここで興味深いことに，定数 μ はステップ幅の選択に使われずともこの一次収束率が保証される．

射影劣勾配法はリプシッツ関数に対しては最適な反復計算量 (6) を保証したが，平滑関数に対する収束率 (8) や (10) は最適ではない．次の節で，平滑関数に対して最適な収束率をもつ一次法を紹介する．

4. 加速勾配法

平滑関数に対する加速勾配法は，射影勾配法を上回る収束率を保証する一次法として Nesterov [10] が確立したのち，さまざまなバリエーションが提案されてきた（たとえば，[8, 11–13]）．中でも Beck and Teboulle [8] による FISTA (Fast Iterative Shrinkage-Thresholding Algorithm) は画像・信号処理などの分野において加速勾配法を広めるのに大きく貢献した．これらのバリエーションの理論解析は，本質的に Nesterov による “estimating sequence” を用いるアプローチが基礎になっており，ほかのアルゴリズムの解析にも応用される有用な概念である [6, 12, 14]．ここでは Nesterov [11] による加速勾配法と estimating sequence のアプローチの要点を解説する．

Nesterov の加速勾配法は，x_k とは別の点 y_k を作り，点 y_k から射影勾配法のステップ $x_{k+1} = \pi_X(y_k - \lambda_k \nabla f(y_k))$ を行うという点が特徴である．そしてこの y_k を決めるのに estimating seqeunce という，以下で定義される二次関数の列 $\{\varphi_k(x)\}$ の最小化問題 $\min_{x \in X} \varphi_k(x)$ を各反復で解く必要がある．

$$\varphi_k(x) := \frac{1}{S_k}\sum_{i=0}^{k} \lambda_i \Big[f(x_i) + \langle \nabla f(x_i), x - x_i \rangle + \frac{\mu}{2}\|x - x_i\|_2^2 \Big] + \frac{1}{2S_k}\|x - x_0\|_2^2, \quad (11)$$

ただし $S_k = \sum_{i=0}^{k} \lambda_i$ はステップ幅 λ_i の和である．最小化問題 $\min_{x\in X} \varphi_k(x)$ の最適解は，以下のベクトル v_k に対して $\pi_X(v_k)$ によって計算できる．

$$v_k := \frac{1}{1+\mu S_k}\left[x_0 - \sum_{i=0}^{k} \lambda_i(\nabla f(x_i) - \mu x_i)\right]. \quad (12)$$

この estimating sequence $\{\varphi_k(x)\}$ の重要な性質は，その最小値と元の問題の最適値 f^* との関係である：

$$\min_{x\in X} \varphi_k(x) \le f^* + \frac{\mathrm{dist}(x_0, X^*)^2}{2S_k}. \quad (13)$$

不等式 (13) は強凸性 (9) よりただちにわかる．

以下に Nesterov の加速勾配法 [11] を述べる（強凸な平滑関数に当てはめた場合の記述である）．

アルゴリズム 7（加速勾配法）． f は X 上で L-平滑かつ μ-強凸であるとする．点 $x_0 \in X$ をとり $S_0 = 0$ とおく．$k = 0, 1, 2, \ldots$ に対して以下の反復を繰り返す．

(a) ステップ幅 λ_{k+1} を，λ の 2 次方程式 $\frac{\lambda^2}{S_k+\lambda} = 2\frac{1+\mu S_k}{L}$ の正の解として定める．$S_{k+1} = S_k + \lambda_{k+1}$ とする．

(b) $z_k := \pi_X(v_k)$ $(= \mathrm{argmin}_{x\in X} \varphi_k(x))$ を計算する．ただし $\varphi_k(x)$ と v_k はそれぞれ (11) と (12) で定める．

(c) $y_k := (1-\tau_k)x_k + \tau_k z_k$ と定める．ただし $\tau_k := \lambda_{k+1}/S_{k+1}$ である．

(d) $x_{k+1} := \pi_X(y_k - \lambda_{k+1}\nabla f(y_k))$ を計算する．

上記の加速勾配法は，任意の $k \ge 1$ に対して $f(x_k) \le \min_{x\in X} \varphi_k(x)$ が成り立つようにうまく設計されている．ゆえに，(13) から

$$f(x_k) - f^* \le \frac{\mathrm{dist}(x_0, X^*)^2}{2S_k}$$

である．この上界において，ステップ幅の和 S_k の増大する速度が収束率を決定する．手順 (a) におけるステップ幅の選択方法に着目すれば S_k の増大率を解析することができ，加速勾配法の収束率は以下のようにして得られる．

定理 8 [11]． 閉凸集合 X 上で L-平滑かつ μ-強凸な目的関数 f の最小化問題 $\min_{x\in X} f(x)$ に対する加速勾配法（アルゴリズム 7）について以下の劣一次収束性が成り立つ．ただし $D := \mathrm{dist}(x_0, X^*)$ である．

$$f(x_k) - f^* \le \frac{LD^2}{k^2}, \quad \forall k \ge 1.$$

特に $\mu > 0$ であれば，$k \ge 1$ に対して以下の一次収束性が成り立つ．

$$f(x_k) - f^* \le LD^2 \exp\left(-k\sqrt{\frac{2\mu}{L}}\right).$$

この結果から，加速勾配法は射影勾配法の収束率 (8) に比べて $\mathrm{O}(1/k)$ から $\mathrm{O}(1/k^2)$ へと“加速”されたことがわかる．さらに強凸関数である場合にも $\mathrm{O}(\exp(-k\mu/L))$ から $\mathrm{O}(\exp(-k\sqrt{\mu/L}))$ に改善される．

加速勾配法が保証する収束率は，定数倍の違いを除いてこれ以上改善できないことが無制約の場合に示されており，加速勾配法は L-平滑な凸関数の最小化に対する一次法の中での“最適性”をもつ．

命題 9 [15]． 任意の $1 \le k \le (n-1)/2$ と $x_0 \in \mathbb{R}^n$ に対して，ある L-平滑な凸関数 f が存在して次を満たす．任意の $x_{i+1} \in x_0 + \mathrm{span}\{\nabla f(x_0), \ldots, \nabla f(x_i)\}$，$i = 0, 1, 2, \ldots$ を満たす点列 $\{x_i\}$ に対して

$$f(x_k) - f(x^*) \ge \frac{3L\|x_0 - x^*\|_2^2}{32(k+1)^2}.$$

ただし $x^* = \mathrm{argmin}_{x\in\mathbb{R}^n} f(x)$ とし，$\mathrm{span}\{v_1, \ldots, v_i\}$ はベクトル $v_1, \ldots, v_i$ が張る線型部分空間である．

また，L-平滑かつ μ-強凸な目的関数のクラスに対しても同様の結果によって，加速勾配法が最適性をもつ．

5. 近接勾配法

目的関数が平滑でないとしても，平滑関数と“単純な”構造をもつ凸関数の和として表されるのであれば，“近接点法”と組み合わせることで射影勾配法や加速勾配法を応用することができる．

今，最適化問題

$$\min_{x\in\mathbb{R}^n} [f(x) + g(x)]$$

について f は L-平滑な凸関数であり，$g : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ は下半連続な凸関数であるとする．このような分離構造をもつ問題において，f に対しては一次の情報を用いて近似を試みるが，g は近似せずにそのまま扱うことを考えよう．すなわち，一次法の各反復で計算していた射影の代わりに，以下の近接写像を用いる．

$$\mathrm{prox}_{\lambda g}(x) := \operatorname*{argmin}_{y\in\mathbb{R}^n}\left\{g(y)+\frac{1}{2\lambda}\|x-y\|_2^2\right\}.$$

これが，制約付き凸最適化問題 $\min_{x\in X} f(x)$ における直交射影 $\pi_X(x)$ を一般化していることは次のようにしてわかる．$g(x)$ を閉凸集合 X の標示関数とする，すなわち，

$$g(x)=\begin{cases}0 & (x\in X)\\ +\infty & (x\notin X)\end{cases}$$

とする．このとき，$\min_{x\in\mathbb{R}^n}[f(x)+g(x)] = \min_{x\in X} f(x)$ であり $\mathrm{prox}_{\lambda g}(x)=\pi_X(x)$ となる．

射影勾配法や加速勾配法の射影演算を近接写像で置き換えることで，問題 $\min_{x\in\mathbb{R}^n}[f(x)+g(x)]$ に対する“近接勾配法”やその加速化が得られる．このようなアプローチは，画像・信号処理や機械学習などに多様な応用をもち，近接勾配法などの一次法の有用性が着目された．ここでは近接勾配法のアルゴリズムについて簡易な導入に留めるが，この観点からの一次法の理論や事例については，[11, 16–18] や小野氏の記事 [19] を参照されたい．

近接勾配法は，射影勾配法の反復を次のように一般化したものである．

$$x_{k+1} := \mathrm{prox}_{\lambda_k g}(x_k-\lambda_k\nabla f(x_k)),\ k=0,1,2,\dots$$

近接勾配法は，射影勾配法と全く同じ収束率の評価を保つ．すなわち，凸関数 f が L-平滑であれば，ステップ幅 $\lambda_k\equiv 1/L$ によって劣一次収束 (8) が保証され，さらに f が μ-強凸であれば一次収束性 (10) もまた成り立つ（各不等式で f を $f+g$ に置き換えよ）．

加速勾配法についても同様に一般化が得られる．すなわち，アルゴリズム 7 において射影計算 (b) と (d) を次のように一般化する．

(b) $z_k := \mathrm{prox}_{\gamma_k g}(v_k)$ とする．ただし $\gamma_k := S_k/(\mu S_k+1)$ である．

(d) $x_{k+1}:=\mathrm{prox}_{\lambda_{k+1}g}(y_k-\lambda_{k+1}\nabla f(y_k))$ とする．

この一般化に対しても，凸関数 f が L-平滑（および μ-強凸）であるとき，定理 8 の二つの不等式が（f を $f+g$ で置き換えると）成り立つ．

6. 射影を用いない一次法

これまでに解説した一次法は，射影計算や近接写像といった補助最適化を各反復で解く必要があった．Frank–Wolfe 法 [20] は，二次計画法とその一般化に対して提案された古典的な一次法であり，収束率は加速勾配法に劣るものの，線形な補助最適化を用いることで各反復の計算コストの削減が期待できる．近年，機械学習などの分野において再考察され，注目を集めるようになった [14, 21, 22].

アルゴリズム 10（Frank–Wolfe 法）**.** 有界な閉凸集合 X 上の凸最適化問題 $\min_{x\in X} f(x)$ を考える．初期点 $x_0\in X$ をとり，各 $k=0,1,2,\dots,$ に対して以下の反復を繰り返す．

(a) y_k を次の最適化問題の一つの最適解とする：

$$\zeta_k := \min_{x\in X}[f(x_k)+\langle\nabla f(x_k), x-x_k\rangle]. \quad (14)$$

(b) $x_{k+1} := x_k+\lambda_k(y_k-x_k)$ とする．ただし $\lambda_k\in(0,1]$ とする．

手順 (a) の補助問題は目的関数が線形であり，最適解 y_k の存在を保証するため，X の有界性が仮定されていることに注意する．パラメータ $\lambda_k\in(0,1]$ はステップ幅を表し，X の凸性により x_{k+1} は実行可能解となる．Frank–Wolfe 法は y_k-x_k を探索方向としており，この探索方向は補助問題 (14) により $\min\{\langle\nabla f(x_k), z\rangle \mid z\in X-\{x_k\}\}$ (≤ 0) の最適解として選ばれている．

Frank–Wolfe 法に対しては，$\mathrm{O}(1/k)$ の収束率が知られている [20, 22].

定理 11. 目的関数 f は X 上で L-平滑であるとする．このとき，ステップ幅 $\lambda_k:=\frac{2}{k+2}$ による Frank–Wolfe 法は，任意の $k\ge 1$ に対して以下を満たす．

$$f(x_{k+1})-f^*\le f(x_{k+1})-\zeta_k\le\frac{2L}{k+4}.$$

この結果から，Frank–Wolfe 法の興味深い特徴がわかる．まず，$f(x_{k+1})-\zeta_k$ は各反復で計算可能であるから，これを近似誤差の上界としてアルゴリズムの終了判定に利用できる．また，ステップ幅の定義 $\lambda_k=\frac{2}{k+2}$ はリプシッツ定数 L を必要としない．収束率は $\mathrm{O}(1/k)$ であり，加速勾配法の $\mathrm{O}(1/k^2)$ には劣るが，各反復の補助問題は線形な最適化であり，射影よりも少ない計算コストが期待される．

このほかの特徴として，線形な補助最適化を解く利点が応用上の側面から現れることがある．たとえば，制約集合 X として ℓ_1 ノルムの球 $X=\{x\in\mathbb{R}^n \mid \|x\|_1\le\tau\}$ を考えるとき，Frank–Wolfe 法における

線形最適化 (14) の解は多面体 X の $2n$ 個の端点 $\{\pm\tau e_i \mid i = 1,\dots,n\}$（$e_1,\dots,e_n$ は $\mathbb{R}^n$ の単位座標ベクトル）の中に存在する．端点から選んだ解 y_k を用いると，近似解 x_{k+1} は非零要素が高々一つしか増加しない．このような近似解の疎性は，スパースベクトル推定において有意義な性質である．

7. その他の手法・発展的な話題

7.1 Bregman 関数を用いた一次法

本稿で解説した射影勾配法や加速勾配法は，各反復で直交射影を補助問題として計算していたが，これを Bregman 関数で一般化した形で一次法が議論されることもよくある．これにより問題構造によってはうまく Bregman 関数を選んで補助問題求解の効率化を図ることができる．今，$\|\cdot\|$ を $\mathbb{R}^n$ の任意のノルムとし，このノルムに関して X 上で 1-強凸かつ連続的微分可能な関数 $\psi(x)$ をとる（一般のノルムに対する強凸性の定義はユークリッドノルムの場合と全く同様である）．このとき，$x, y \in X$ に対して

$$D_\psi(x,y) = \psi(x) - \psi(y) - \langle \nabla\psi(y), x-y\rangle$$

を，ψ に関する Bregman 関数という．Bregman 関数には $D_\psi(x,y) \geq 0$ かつ $D_\psi(x,y) = 0$ となるのは $x = y$ のときに限るという距離的な性質がある．

射影劣勾配法の更新式は (3) で与えられていたが，ここで項 $\frac{1}{2}\|x - x_k\|^2$ を $D_\psi(x, x_k)$ に置き換えて得られる反復法

$$x_{k+1} = \operatorname*{argmin}_{x\in X}\left\{ f(x_k) + \langle g_k, x - x_k\rangle + \frac{1}{\lambda_k} D_\psi(x, x_k)\right\} \quad (15)$$

のことを鏡像降下法 (mirror descent)[2, 3] という．この鏡像降下法が射影劣勾配の一般化であることは $\psi(x) = \frac{1}{2}\|x\|_2^2$ ととればわかる．特にこのとき，$D_\psi(x,y) = \frac{1}{2}\|x-y\|_2^2$ である．

問題構造に適合した Bregman 関数をとれる例として単体上の凸最適化問題を挙げよう [2]．制約集合を単体 $X = \Delta := \{x = (x^{(1)},\dots,x^{(n)}) \in \mathbb{R}^n \mid x \geq 0,\ \sum_{i=1}^n x^{(i)} = 1\}$ とし，ℓ_1 ノルム $\|\cdot\| = \|\cdot\|_1$ をとる．このとき関数 $\psi(x) = \sum_{i=1}^n x^{(i)} \log x^{(i)}$ は ℓ_1 ノルムに対して単体 Δ の相対的内部上で 1-強凸となり，対応する Bregman 関数は

$$D_\psi(x,y) = \sum_{i=1}^n x^{(i)} \log \frac{x^{(i)}}{y^{(i)}}, \quad x, y \in \Delta$$

で与えられる．さらに，鏡像降下法の補助問題 (15) の最適解は閉形式をもち，$\mathrm{O}(n)$ で計算できる．

双対平均化法 [5] や加速勾配法 [11] は，もともと Bregman 関数を用いて考察されており，本稿で紹介した収束率と同様の結果が成り立つ（ただしリプシッツ性や平滑性の定義も一般のノルム $\|\cdot\|$ に置き換わる）．たとえば，鏡像降下法は

$$D = \min\{D_\psi(x_0, x^*) \mid x^* \in X^*\}$$

と置き換えることで不等式 (4) や命題 2 が成り立つ．

Bregman 関数を用いた一次法の理論は，最近でも進展が見られる，重要な課題の一つである [23, 24]．

7.2 勾配ノルムを最適性指標とする場合

本稿では，一次法の性能を測るための指標として，$f(x_k) - f^*$ に関しての収束率を対象とした．これは一次法の研究において最も代表的なものであるが，f^* を知らない限り各反復で $f(x_k) - f^*$ を直接計算することはできない．その代わりに，各反復で補助最適化問題を追加で解いて f^* の下界 f_k^* を用いる手法があり，このとき近似誤差の計算可能な上界 $f(x_k) - f_k^*$ もまた $f(x_k) - f^*$ と同じ収束率を保つようにできる [12]．

$f(x_k) - f^*$ の代わりに，近似誤差として勾配のノルム $\|\nabla f(x_k)\|_2$（無制約の場合）またはその制約付き問題への一般化を対象とすることも多い．こちらのほうが各反復で計算できるという利点があるが，理論的にわかっていることが限定される．ここでは無制約の凸最適化問題

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

を考えて，$\|\nabla f(x)\|_2 \leq \varepsilon$ を目指す一次法を紹介しよう．

凸関数 f が $\mathbb{R}^n$ 上で L-平滑であるとすると，一般に次の不等式が成り立つ（不等式 (7) で $x = y - \nabla f(y)/L$ を代入すれば示せる）．

$$\frac{\|\nabla f(x)\|_2^2}{2L} \leq f(x) - f^*, \quad \forall x \in \mathbb{R}^n. \quad (16)$$

今，f が $\mathbb{R}^n$ 上で L-平滑かつ μ-強凸 ($\mu > 0$) であるとすれば，Nesterov の加速勾配法は定理 8 より $f(x_k) - f^* \leq LD^2 \exp\left(-k\sqrt{2\mu/L}\right)$ を満たすから，(16) と合わせて

$$\|\nabla f(x_k)\|_2 \leq \sqrt{2}LD \exp\left(-k\sqrt{\frac{\mu}{2L}}\right), \quad \forall k \geq 0.$$

ゆえに $\mathrm{O}\left(\sqrt{\frac{L}{\mu}}\log\frac{1}{\varepsilon}\right)$ の反復回数で $\|\nabla f(x_k)\|_2 \leq \varepsilon$ が得られる．この反復計算量は最適であることが知ら

れている．

次に強凸性がない場合を考えよう．凸関数 f が $\mathbb{R}^n$ 上で L-平滑であるとする．Nesterov の加速勾配法を使えば上記と同じ議論で $\mathrm{O}(LD/\varepsilon)$ の反復回数で $\|\nabla f(x_k)\|_2 \le \varepsilon$ が得られる．また，Nesterov の加速勾配法を少し変形すると $\mathrm{O}(LD/\varepsilon^{2/3})$ に改良できる [15]．これをさらに改良する方法がある．そのために，f の代わりに以下の正則化 f_μ に対して Nesterov の加速勾配法を適用する．

$$f_\mu(x) := f(x) + \frac{\mu}{2}\|x - x_0\|_2^2, \quad \mu := \frac{\varepsilon}{2D}.$$

このとき生成される点列 $\{x_k\}$ に対して $\|\nabla f(x_k)\|_2 \le \varepsilon$ を得るのに必要な反復回数は高々

$$\mathrm{O}\left(\sqrt{\frac{LD}{\varepsilon}}\log\frac{LD}{\varepsilon}\right). \tag{17}$$

この反復計算量は $\Omega(\sqrt{LD/\varepsilon})$ の下界 [25] が知られているという意味で“準最適”である．上界 (17) において対数の部分をなくして，反復計算量の上界 $\mathrm{O}(\sqrt{LD/\varepsilon})$ を達成する一次法が存在するかどうかは一般の L-平滑関数に対して未解決である（二次関数に限定すれば存在する [15, 25]）．

上記の正則化と Nesterov の加速勾配法を組み合わせる手法については，近年 Catalyst [26] や再出発法 [27] に見られる新しい技法が発表されている．

$f(x_k) - f^*$ や $\|\nabla f(x_k)\|_2$ といった最適性指標について反復計算量の限界を調べる最近の興味深い手法としては，“一次法の性能”を最適化問題として定式化し，それをいくつかの緩和技法により半正定値計画などを通して数値的に解析するものがある [28–30]．

8. おわりに

本稿では，加速勾配法とそれに関連する手法について，一次法の基礎理論に特化して概論を述べた．本稿の内容は，凸最適化問題に対する一次法の理論としては 1980 年代にはすでにその原型が確立されていたものが少なくないが，近接勾配法（5 節）に見たようにこれらの基礎が新しい一次法の構築に役立ち，問題構造に特化した考察を通じて発展してきた．一次法は基礎理論においてもさまざまな面白い進展が続いており，本稿が一次法の理論に興味をもつきっかけや参考となれば幸いである．

参考文献

[1] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, 1970.

[2] A. Beck and M. Teboulle, “Mirror descent and nonlinear projected subgradient methods for convex optimization,” *Operations Research Letters*, **31**, pp. 167–175, 2003.

[3] A. S. Nemirovsky and D. B. Yudin, *Problem Complexity and Method Efficiency in Optimization*, Nauka Publishers, 1979 in Russian. (English translation: John Wiley & Sons, New York, 1983.)

[4] A. Nedić and S. Lee, “On stochastic subgradient mirror-descent algorithm with weighted averaging,” *SIAM Journal on Optimization*, **24**, pp. 84–107, 2014.

[5] Y. Nesterov, “Primal-dual subgradient methods for convex problems,” *Mathematical Programming*, **120**, pp. 221–259, 2009.

[6] M. Ito and M. Fukuda, “A family of subgradient-based methods for convex optimization problems in a unifying framework,” *Optimization Methods and Software*, **31**, pp. 952–982, 2016.

[7] E. S. Levitin and B. T. Polyak, “Constrained minimization methods,” *USSR Computational Mathematics and Mathematical Physics*, **6**, pp. 1–50, 1966.

[8] A. Beck and M. Teboulle, “A fast iterative shrinkage-thresholding algorithm for linear inverse problems,” *SIAM Journal on Imaging Sciences*, **2**, pp. 183–202, 2009.

[9] O. Devolder, F. Glineur and Y. Nesterov, “First-order methods with inexact oracle: The strongly convex case,” CORE Discussion Paper, No. 16, 2013.

[10] Y. Nesterov, “A method of solving a convex programming problem with convergence rate $\mathrm{O}(1/k^2)$,” *Soviet Mathematics Doklady*, **27**, pp. 372–376, 1983.

[11] Y. Nesterov, “Gradient methods for minimizing composite functions,” *Mathematical Programming*, **140**, pp. 125–161, 2013.

[12] Y. Nesterov, “Universal gradient methods for convex optimization problems,” *Mathematical Programming*, **152**, pp. 381–404, 2015.

[13] P. Tseng, “On accelerated proximal gradient methods for convex-concave optimization,” Technical Report, University of Washington, 2008.

[14] Y. Nesterov, “Complexity bounds for primal-dual methods minimizing the model of objective function,” *Mathematical Programming*, **171**, pp. 311–330, 2018.

[15] Y. Nesterov, *Lectures on Convex Optimization*, Springer, 2018.

[16] M. Fukushima and H. Mine, “A generalized proximal point algorithm for certain non-convex minimization problems,” *International Journal of Systems Science*, **12**, pp. 989–1000, 1981.

[17] A. Beck and M. Teboulle, “Gradient-based algorithms with applications to signal recovery problems,” *Convex Optimization in Signal Processing and Communications*, D. Palomar and Y. Eldar (eds.), pp. 33–88, Cambridge University Press, 2010.

[18] N. Parikh and S. Boyd, “Proximal algorithms,” *Foundations and Trends in Optimization*, **1**, pp. 123–231, 2014.

[19] 小野峻佑，“近接分離アルゴリズムとその応用—信号処理・画像処理的観点から—，”オペレーションズ・リサーチ：

経営の科学, **64**(6), pp. 316–325, 2019.
[20] M. Frank and P. Wolfe, "An algorithm for quadratic programming," *Naval Research Logistics Quarterly*, **3**, pp. 95–110, 1956.
[21] M. Jaggi, "Revisiting Frank–Wolfe: Projection-free sparse convex optimization," In *Proceedings of the 30th International Conference on Machine Learning*, pp. 427–435, 2013.
[22] R. M. Freund and P. Grigas, "New analysis and results for the Frank–Wolfe method," *Mathematical Programming*, **155**, pp. 199–230, 2016.
[23] H. H. Bauschke, J. Bolte and M. Teboulle, "A descent lemma beyond Lipschitz gradient continuity: First-order methods revisited and applications," *Mathematics of Operations Research*, **42**, pp. 349–376, 2016.
[24] H. Lu, R. M. Freund and Y. Nesterov, "Relatively-smooth convex optimization by first-order methods, and applications," *SIAM Journal on Optimization*, **28**, pp. 333–354, 2018.
[25] A. S. Nemirovsky, "Information-based complexity of linear operator equations," *Journal of Complexity*, **8**, pp. 153–175, 1992.
[26] H. Lin, J. Mairal and Z. Harchaoui, "A universal catalyst for first-order optimization," *Advances in Neural Information Processing Systems*, pp. 7854–7907, 2015.
[27] M. Liu and T. Yang, "Adaptive accelerated gradient converging method under Hö lderian error bound condition," *Advances in Neural Information Processing Systems*, pp. 3104–3114, 2017.
[28] Y. Drori and M. Teboulle, "Performance of first-order methods for smooth convex minimization: A novel approach," *Mathematical Programming*, **145**, pp. 451–482, 2014.
[29] D. Kim and J. A. Fessler, "Optimized first-order methods for smooth convex minimization," *Mathematical Programming*, **159**, pp. 81–107, 2016.
[30] A. B. Taylor, J. M. Hendrickx and F. Glineur, "Smooth strongly convex interpolation and exact worst-case performance of first-order methods," *Mathematical Programming*, **161**, pp. 307–345, 2017.

