日本オペレーションズ・リサーチ学会
2023 年春季研究発表会

パラメータ調整不要な非凸加速勾配法

05000504　東京大学　*丸茂直貴　MARUMO Naoki
01308490　東京大学　武田朗子　TAKEDA Akiko

1. はじめに

一般的な非凸最適化問題

$$\min_{x\in\mathbb{R}^d} \ f(x) \tag{1}$$

を考える．関数 $f\colon \mathbb{R}^d \to \mathbb{R}$ は下に有界とし，以下の Lipschitz 連続性を仮定する．

仮定 1. 定数 $L_f, M_f > 0$ が存在し，任意の $x, y \in \mathbb{R}^d$ に対し，

(a) $\|\nabla f(x) - \nabla f(y)\| \le L_f \|x - y\|$

(b) $\|\nabla^2 f(x) - \nabla^2 f(y)\| \le M_f \|x - y\|$

が成り立つ．

問題 (1) のように一般的な問題に対するアルゴリズムとして, 計算効率の観点から f の関数値と勾配の情報のみを用いる**一次法**が盛んに研究されてきた．一次法に関する古典的な結果に，仮定 1(a) の下で ε-停留点[1] が最急降下法により $O(\varepsilon^{-2})$ 回の関数値・勾配評価で求まるというものがある．この計算量は仮定 1(a) の下では最良であることも知られている [4].

仮定 1(a) に加えて仮定 1(b) も課し，計算量 $O(\varepsilon^{-7/4})$ もしくは $\tilde{O}(\varepsilon^{-7/4})$[2] を達成するより高度な一次法も提案されている [1, 3, 5, 6, 9]．しかしこれらは，Lipschitz 定数 L_f，M_f や許容誤差 ε を既知としてアルゴリズム中で用いている．現実の最適化問題においてはこれらの値は未知であることが多く，アルゴリズムを適用しようとすると精緻なパラメータ調整が必要となる．この問題を解決するため，本研究では以下の特長を持つ一次法を提案する．

- 仮定 1 の下で計算量 $O(\varepsilon^{-7/4})$ を達成する．
- Lipschitz 定数 L_f，M_f を自動的に推定する．
- 許容誤差 ε を入力として必要としない．

2. M_f の推定を実現する不等式

上述の特長を持つ一次法の設計における最大の課題は，関数値と勾配の情報のみから Hesse 行列の Lipschitz 定数 M_f を推定することである．本研究では M_f の推定のために，以下の 2 つの補題に基づく，Hesse 行列を用いない新しい解析手法を導入する．

補題 1. 仮定 1(b) が成り立つとする．このとき，任意の $x, y \in \mathbb{R}^d$ に対し，

$$f(x) - f(y) \le \frac{1}{2}\langle \nabla f(x) + \nabla f(y), x - y\rangle + \frac{M_f}{12}\|x - y\|^3$$

が成り立つ．

補題 2. 仮定 1(b) が成り立つとする．任意の $z_1, \dots, z_n \in \mathbb{R}^d$ と $\sum_{i=1}^n \lambda_i = 1$ なる任意の $\lambda_1, \dots, \lambda_n \ge 0$ に対し $\bar{z} := \sum_{i=1}^n \lambda_i z_i$ とおく．このとき，

$$\left\|\nabla f(\bar{z}) - \sum_{i=1}^n \lambda_i \nabla f(z_i)\right\| \le \frac{M_f}{2}\sum_{i=1}^n \lambda_i \|z_i - \bar{z}\|^2$$

が成り立つ．

補題 1 は数値解析の分野では台形則の誤差評価式として知られている不等式 [2, 式 (5.1.4)] の多次元版と見なせる．補題 2 は勾配に対する Jensen 型の不等式[3] と言える．これらの不等式は M_f を含むが Hesse 行列そのものは含まない．これにより M_f の推定が可能となる．なお，既存の計算量解析 [1, 3, 5, 6, 9] は M_f と Hesse 行列をともに含む不等式を用いており，これを M_f が未知の場合に拡張することは困難である．

3. アルゴリズム

提案アルゴリズムをアルゴリズム 1 に示す．本アルゴリズムは，加速勾配法 [8] に 2 種類の再始動機構（行 9，11）を付加したものである．アルゴリズム中の L と M_k はそれぞれ L_f と M_f の推定値である．L は再始動時にしか更新されないが，M_k は各反復で更新されることに注意する．

L は古典的なバックトラッキングに似た手法により計算される．降下条件 $f(x_k) \le f(x_0) - \frac{LS_k}{2(k+1)}$ が満たされないとき，現在の推定値 L が小さいと

[1] $\|\nabla f(x)\| \le \varepsilon$ なる点 $x \in \mathbb{R}^d$.
[2] $\tilde{O}(\varepsilon^{-\alpha})$ は $O(\varepsilon^{-\alpha}\,\mathrm{polylog}(\varepsilon^{-1}))$ を意味する．
[3] Jensen の不等式は，凸関数 $\phi\colon \mathbb{R}^d \to \mathbb{R}$ に対し $\phi(\bar{z}) \le \sum_{i=1}^n \lambda_i \phi(z_i)$ が成り立つというものである．

アルゴリズム 1 再始動機構を備えた加速勾配法

Input: $x_{\text{init}} \in \mathbb{R}^d$; $L_{\text{init}}, M_0 > 0$

$(x_0, y_0) \leftarrow (x_{\text{init}}, x_{\text{init}})$, $L \leftarrow L_{\text{init}}$, $k \leftarrow 0$
repeat
  $k \leftarrow k + 1$
  $x_k \leftarrow y_{k-1} - \frac{1}{L}\nabla f(y_{k-1})$
  $y_k \leftarrow x_k + \frac{k}{k+1}(x_k - x_{k-1})$
  $\bar{y}_k \leftarrow \frac{2}{k(k+1)} \sum_{i=0}^{k-1} (i+1) y_i$
  $S_k \leftarrow \sum_{i=1}^{k} \|x_i - x_{i-1}\|^2$
  M_k を式 (2) により計算
  if $f(x_k) > f(x_0) - \frac{LS_k}{2(k+1)}$ **:**
    $(x_0, y_0) \leftarrow (x_{k-1}, x_{k-1})$, $L \leftarrow 2L$, $k \leftarrow 0$
  else if $(k+1)^5 M_k^2 S_k > L^2$ **:**
    $(x_0, y_0) \leftarrow (x_k, x_k)$, $k \leftarrow 0$
until convergence
return $\bar{y}_k$

判断し，L をより大きな値に再設定して加速勾配法を再始動する．この降下条件は Armijo 条件の類似物である．ただしここでは，一反復での目的関数値の減少量ではなく，k 反復での減少量を参照する．

より非自明なのが M_k の計算である．M_k は更新式

$$M_k = \max\left\{ M_{k-1}, 4\frac{(k+1)^2\|\nabla f(\bar{y}_k)\| - 4L\|x_k - x_{k-1}\|}{(k-1)(k+5)^2 S_k}, \right.$$
$$12\frac{f(y_k) - f(x_k) - \frac{1}{2}\langle \nabla f(y_k) + \nabla f(x_k), y_k - x_k\rangle}{\|y_k - x_k\|^3},$$
$$\left.\frac{\|(k+1)\nabla f(y_k) + k\nabla f(x_{k-1}) - (2k+1)\nabla f(x_k)\|}{k\|x_k - x_{k-1}\|^2}\right\} \quad (2)$$

により計算される．この式は，補題 1，2 に基づいて設計されている．

4. 計算量解析

次の命題は，Lipschitz 定数の推定値 L，M_k の上界を与える．

命題 1. 仮定 1 が成り立つとする．このとき，アルゴリズム 1 において

(a) $L \le \max\{L_{\text{init}}, 2L_f\}$

(b) $M_k \le \max\{M_0, M_f\}$

が常に成り立つ．

この上界は，2 つの再始動条件と M_k の更新式，そして補題 1，2 を組み合わせることで示せる．この上界をそれぞれ $\bar{L}$，$\bar{M}$ とおく．提案アルゴリズムの反復数について，以下の結果が示せる．

定理 1. 仮定 1 が成り立つとする．このとき，アルゴリズム 1 において初めて $\|\nabla f(\bar{y}_k)\| \le \varepsilon$ が満たされるまでの合計反復数は

$$\frac{70\bar{L}^{1/2}\bar{M}^{1/4}}{\varepsilon^{7/4}}\Big(f(x_{\text{init}}) - \inf_{x\in\mathbb{R}^d} f(x)\Big) + O(\varepsilon^{-1/4})$$

以下である．

アルゴリズム 1 は一反復あたり 2 点 x_k，y_k で関数値評価を，3 点 x_k，y_k，$\bar{y}_k$ で勾配評価をする．そのため，合計評価回数も定理 1 で与えられた反復数と同じ $O(\varepsilon^{-7/4})$ となる．

5. おわりに

数値実験の結果は口頭発表で紹介する．本発表はプレプリント [7] に基づく．

参考文献

[1] Z. Allen-Zhu and Y. Li. NEON2: Finding local minima via first-order oracles. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

[2] K. E. Atkinson. *An Introduction to Numerical Analysis*. John Wiley & Sons, 2nd edition, 1989.

[3] Y. Carmon, J. C. Duchi, O. Hinder, and A. Sidford. "Convex until proven guilty": Dimension-free acceleration of gradient descent on non-convex functions. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pages 654–663, 2017.

[4] C. Cartis, N. I. M. Gould, and P. L. Toint. On the complexity of steepest descent, Newton's and regularized Newton's methods for nonconvex unconstrained optimization problems. *SIAM Journal on Optimization*, 20(6):2833–2852, 2010.

[5] C. Jin, P. Netrapalli, and M. I. Jordan. Accelerated gradient descent escapes saddle points faster than gradient descent. In *Proceedings of the 31st Conference On Learning Theory*, volume 75, pages 1042–1085, 2018.

[6] H. Li and Z. Lin. Restarted nonconvex accelerated gradient descent: No more polylogarithmic factor in the $O(\varepsilon^{-7/4})$ complexity. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pages 12901–12916, 2022.

[7] N. Marumo and A. Takeda. Parameter-free accelerated gradient descent for nonconvex minimization, 2022. URL `https://arxiv.org/abs/2212.06410`.

[8] Y. Nesterov. A method for solving a convex programming problem with convergence rate $O(1/k^2)$. In *Soviet Mathematics Doklady*, volume 27, pages 372–376, 1983.

[9] Y. Xu, R. Jin, and T. Yang. NEON+: Accelerated gradient methods for extracting negative curvature for non-convex optimization, 2017. URL `https://arxiv.org/abs/1712.01033`.