§ 4.4　Momentum Method

伊 冉 (Andre YI)

2026 年 2 月 4 日

PDF 版はここ↑

1　概要

最急降下法（Gradient Descent）では、コスト関数の局所解（Local Minimum）に陥る可能性がある。これを回避し、かつ収束を加速させるために考案されたのが**モーメンタム法（Momentum Method）**である。物理的な直感として、勾配降下法は「質点（ボール）が斜面を転がり落ちる運動」とみなせます。モーメンタム法では、このモデルに**速度（運動量）**と**摩擦（Friction）**の概念を導入する。

- **速度（Velocity）**：過去の勾配の値を積んで、慣性を持ってエネルギー障壁（Energy Barrier）を乗り越える助けとなる。
- **摩擦（Friction）**：システムのエネルギーを徐々に散逸させ、ボールが最終的に平衡点（最小値）に止まるようにする。

Polyak (1964) による古典的なモーメンタム法の更新式は以下の通りである。

$$x^{n+1} = x^n + v^{n+1} \tag{1}$$

$$v^{n+1} = \mu v^n - \eta \nabla f(x^n) \tag{2}$$

ここで、$\eta > 0$ は学習率、$\mu \in [0, 1]$ はモーメンタム係数（摩擦係数とも呼ばれる）である。

2　Kinematic Interpretation
運動学的解釈

目的関数 $f(x)$ によって形成される「器（cup）」の中を転がる質量 $m = 1$ のボールを考える。ボールには以下の 2 つの力が作用する。

1. **勾配力**: $-\nabla f(x)$（重力のようにポテンシャルを下る力）
2. **摩擦力**: $F_f = -^{1}\rho \dot{x}(t)$（速度に比例し、運動を妨げる力。$\rho > 0$ は減衰係数）

よって、$F_{\text{合力}} = F_{\text{勾配力}} + F_{\text{摩擦力}} = -\rho \dot{x}(t) - \nabla f(x)$。
ニュートンの運動方程式（$\mathbf{F} = \mathbf{ma}$）より、以下の微分方程式が得られる。

$$a = {}^{2}\ddot{x}(t) = -\rho \dot{x}(t) - \nabla f(x(t)) \tag{3}$$

---

[1] ρ はローと言う

[2] $\dot{v} = \frac{d}{dt}v = \frac{d}{dt}\left(\frac{d}{dt}x\right) = \frac{d^2}{dt^2}x = \ddot{x} = -\nabla f(x)$. ここで、$\ddot{x}$ は加速度であり、エックス・ツー・ドットと呼ぶ。

2.1 Energy Dissipation Analysis
エネルギー散逸の解析

Ex 1: $\frac{d}{dt}E_{tot}(t) < 0$ の証明

このシステムにおいて、総エネルギー $E_{tot}(t)$ が保存されず、時間とともに減少することを証明する。

Proof.
総エネルギーは運動エネルギーとポテンシャルエネルギーの和で定義される。

$$E_{tot}(t) = \frac{1}{2}\|\dot{x}(t)\|^2 + f(x(t))$$

これを時間 t で微分する。

$$\begin{aligned}\frac{d}{dt}E_{tot}(t) &= \frac{d}{dt}\left(\frac{1}{2}\dot{x}(t)^T\dot{x}(t) + f(x(t))\right)\\ &= \frac{1}{2}((\frac{d}{dt}\dot{x}(t))^T\dot{x}(t) + \dot{x}(t)^T\frac{d}{dt}\dot{x}(t)) + \nabla f(x(t))^T\frac{d}{dt}(x(t))\\ &= \dot{x}(t)^T\ddot{x}(t) + \dot{x}(t)^T\nabla f(x(t))\\ &= \dot{x}(t)^T\left(\ddot{x}(t) + \nabla f(x(t))\right)\end{aligned}$$

ここで、運動方程式 **[3]** より、$\ddot{x}(t) + \nabla f(x(t)) = -\rho\dot{x}(t)$ なので、これを代入する。

$$\begin{aligned}\frac{d}{dt}E_{tot}(t) &= \dot{x}(t)^T\left(-\rho\dot{x}(t)\right)\\ &= -\rho\|\dot{x}(t)\|^2\\ &= -\rho\|v(t)\|^2 < 0\end{aligned}$$

$\rho > 0$ であるため、速度がゼロでない限りエネルギーは常に減少する。これにより、システムはエネルギーの低い状態（平衡点）へと収束する。 ■

2.2 Phase Space Analysis and Divergence
相空間解析と発散率 div

運動方程式 **[3]** は、位置 x と速度 v に関する 1 階連立微分方程式 **ODEs** として記述できる。

$$\begin{cases}\dot{x}(t) = v(t)\\ \dot{v}(t) = -\rho v(t) - \nabla f(x(t))\end{cases}$$

このベクトル場 $X = (\dot{x}, \dot{v})$ の発散（**Divergence**）を計算する。

$$\begin{aligned}\operatorname{div}(\dot{x},\dot{v}) &= \frac{\partial}{\partial x}(\dot{x}) + \frac{\partial}{\partial v}(\dot{v}) \\ &= \frac{\partial}{\partial x}(v) + \frac{\partial}{\partial v}(-\rho v - \nabla f(x)) \\ &= 0 + (-\rho) \\ &= -\rho < 0\end{aligned}$$

発散が負であることは、相空間内の体積が時間とともに**収縮（contracting）**することを意味する。これにより、解軌道は一点（平衡点）に収束する。

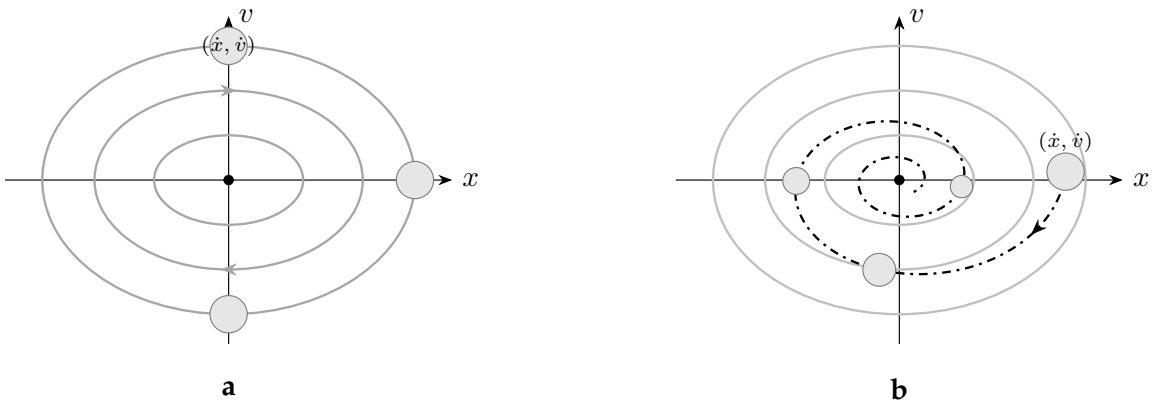

図 1: Solution in phase space.
a. Without friction (div $= 0$), trajectory orbits forever.
b. With friction (div < 0), trajectory spirals into the equilibrium point.

3 Discretization and Algorithm Derivation
離散化とアルゴリズムの導出

コンピュータ上で計算を実行するために、連続時間の **ODEs** システムを有限差分法を用いて離散化する。

まず、等間隔の時間分割 $0 = t_0 < t_1 < \cdots < t_n < \infty$ を考える。時間刻み幅を $\Delta t = t_{n+1} - t_n$ **（定数）**とする。システムの n ステップ目の状態を $(x^n, v^n) = (x(t_n), v(t_n))$ と表記する。

微分 $\dot{x}(t), \dot{v}(t)$ を前進差分（Forward Difference）で近似する

$$\dot{x}(t) \approx \frac{x^{n+1} - x^n}{\Delta t}, \quad \dot{v}(t) \approx \frac{v^{n+1} - v^n}{\Delta t}$$

これらを元の **ODEs** に代入すると

$$\frac{x^{n+1}-x^n}{\Delta t} = v^n$$
$$\frac{v^{n+1}-v^n}{\Delta t} = -\rho v^n - \nabla f(x^n)$$

となり、これを整理すると、以下の差分方程式が得られる。

$$\begin{cases} x^{n+1} - x^n = v^n \Delta t & \textbf{[4]} \\ v^{n+1} - v^n = -\rho v^n \Delta t - \nabla f(x^n)\Delta t & \textbf{[5]} \end{cases}$$

次に、物理のパラメータをアルゴリズムのハイパーパラメータに変換する。

- 時間刻みを $\epsilon = \Delta t$ と置く。
- 摩擦項をまとめるため、$\mu = 1 - \rho\epsilon$ と定義する。ここで $\rho > 0, \epsilon > 0$ より $\mu < 1$ である。

式 **[5]** の右辺を変形する：

$$v^{n+1} = v^n - \rho v^n \Delta t - \nabla f(x^n)\Delta t = (1-\rho\Delta t)v^n - \Delta t \nabla f(x^n)$$

パラメータ ϵ, μ を代入すると、中間的なシステムが得られる。

$$x^{n+1} = x^n + \epsilon v^n \quad \textbf{[6]}$$
$$v^{n+1} = \mu v^n - \epsilon \nabla f(x^n) \quad \textbf{[7]}$$

最後、物理的な速度 v とアルゴリズム上の更新量を整合させるため、速度変数の再スケーリング（Rescaling）を行う。新しい速度変数 $\tilde{v}$ を以下のように定義する。

$$\tilde{v}^n = \epsilon v^n \quad \left(\iff v^n = \frac{\tilde{v}^n}{\epsilon} \right)$$

これを式 **[6]** と **[7]** に代入して計算を進める

位置の更新式:

$$x^{n+1} = x^n + \epsilon \left(\frac{\tilde{v}^n}{\epsilon} \right) = x^n + \tilde{v}^n$$

速度の更新式:

$$\frac{\tilde{v}^{n+1}}{\epsilon} = \mu \left(\frac{\tilde{v}^n}{\epsilon} \right) - \epsilon \nabla f(x^n)$$

この両辺に ϵ を掛けると、

$$\tilde{v}^{n+1} = \mu \tilde{v}^n - \epsilon^2 \nabla f(x^n)$$

ここで、学習率を $\eta = \epsilon^2$ とすると、最終的なアルゴリズムの更新式が得られる。

$$x^{n+1} = x^n + \tilde{v}^n \quad \textbf{[8]}$$
$$\tilde{v}^{n+1} = \mu \tilde{v}^n - \eta \nabla f(x^n) \quad \textbf{[9]}$$

これは Polyak の古典的なモーメンタム法（式 [1], [2]）と形式的に一致する。

4 Analysis of Example 4.4.1
例 4.4.1 の解析

2 次関数に対するモーメンタム法の応用を、線形代数で解析する。

_Ex_ 2: Quadratic Function Analysis

実変数 x の 2 次関数 $f(x) = \frac{1}{2}(ax - b)^2$ （ただし $a \neq 0$）を考える。このとき、モーメンタム法の更新式 **[8]-[9]** は線形式 $s_{n+1} = Ms_n + \beta$ として記述できる。ここで、$s_n = \begin{pmatrix} x^n \\ v^n \end{pmatrix}$, $M = \begin{pmatrix} 1 & \epsilon \\ -\epsilon a^2 & \mu \end{pmatrix}$, $\beta = \begin{pmatrix} 0 \\ \epsilon ab \end{pmatrix}$。

1. 状態 s_n の一般項を初期状態 s_0 を用いて導出せよ。
2. 行列 M の固有値を解析し、システムが平衡点へ収束することを証明せよ。

Proof.
まず、勾配 $f'(x) = a(ax - b) = a^2x - ab$ を更新式に代入し、行列形式に整理する。

$$\begin{aligned} x^{n+1} &= x^n + \epsilon v^n \\ v^{n+1} &= \mu v^n - \epsilon(a^2 x^n - ab) = -\epsilon a^2 x^n + \mu v^n + \epsilon ab \end{aligned}$$

状態ベクトルを $s_n = \begin{pmatrix} x^n \\ v^n \end{pmatrix}$ と置くと、以下の線形漸化式が得られる。

$$s_{n+1} = Ms_n + \beta, \quad \text{where, } M = \begin{pmatrix} 1 & \epsilon \\ -\epsilon a^2 & \mu \end{pmatrix}, \quad \beta = \begin{pmatrix} 0 \\ \epsilon ab \end{pmatrix} \tag{10}$$

次、漸化式 **[10]** を繰り返し計算し、一般項 s_n を導出する。

- $n = 1$:
$$s_1 = Ms_0 + \beta$$
- $n = 2$:
$$s_2 = Ms_1 + \beta = M(Ms_0 + \beta) + \beta = M^2 s_0 + M\beta + \beta$$
- $n = 3$:
$$s_3 = Ms_2 + \beta = M(M^2 s_0 + M\beta + \beta) + \beta = M^3 s_0 + (M^2 + M + I)\beta$$

よって、これを n 回繰り返すと、以下の形式となる。

$$s_n = M^n s_0 + \left(\sum_{k=0}^{n-1} M^k \right) \beta \tag{11}$$

ここで、括弧内の項は行列の等比級数である。$S_n = \sum_{k=0}^{n-1} M^k$ と置き、等式の両辺に $I - M$ を掛けると、

$$(\mathbb{I} - M)S_n = (\mathbb{I} - M)(\mathbb{I} + M + \cdots + M^{n-1}) = \mathbb{I} - M^n$$

ここで、$\mathbb{I}$ は単位行列 **(Unit Matrix)** 行列 $\mathbb{I} - M$ が正則（逆行列が存在）であると仮定すると、

$$^3S_n = (\mathbb{I} - M^n)(\mathbb{I} - M)^{-1}$$

したがって、状態 s_n の一般項は以下のように表される。

$$s_n = M^n s_0 + (^4\mathbb{I}_2 - M^n)(\mathbb{I}_2 - M)^{-1}\beta \quad \textbf{[12]}$$

あと、収束先を求めるために必要な $(\mathbb{I}_2 - M)^{-1}$ を計算する。

$$\mathbb{I}_2 - M = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} - \begin{pmatrix} 1 & \epsilon \\ -\epsilon a^2 & \mu \end{pmatrix} = \begin{pmatrix} 0 & -\epsilon \\ \epsilon a^2 & 1 - \mu \end{pmatrix}$$

2x2 行列の逆行列公式より、行列式は $\det = 0 - (-\epsilon)(\epsilon a^2) = \epsilon^2 a^2$ なので、

$$(\mathbb{I}_2 - M)^{-1} = \frac{1}{\epsilon^2 a^2} \begin{pmatrix} 1 - \mu & \epsilon \\ -\epsilon a^2 & 0 \end{pmatrix} \quad \textbf{[13]}$$

となる。

このシステムが収束させるためには、$\lim_{n\to\infty} M^n = 0$ である必要がある。***i.e.*** M のすべての固有値の絶対値が 1 未満であることと同値である。

次に、行列 M の特性方程式（**Characteristic Equation**）を解く

$$\det(M - \lambda\mathbb{I}) = \det \begin{pmatrix} 1 - \lambda & \epsilon \\ -\epsilon a^2 & \mu - \lambda \end{pmatrix} = 0$$

$$(1 - \lambda)(\mu - \lambda) + \epsilon^2 a^2 = 0 \implies \lambda^2 - (1 + \mu)\lambda + (\mu + \epsilon^2 a^2) = 0$$

この 2 次方程式の解 λ_1, λ_2 について分析する（**7 ページの図 2 参照**）

1. $\epsilon = 0$ のとき：解は $\lambda_1(0) = \mu$ と $\lambda_2(0) = 1$ である。ここで $0 < \mu < 1$ である。
2. $\epsilon > 0$ （`for small enough`）のとき：定数項が増加し、放物線が上方にシフトする。ϵ が十分小さければ、連続性により解は実数のまま区間 $(\mu, 1)$ の内部にある。

よって、$0 < \lambda_1(\epsilon) < \lambda_2(\epsilon) < 1$ が成立するため、

$$\lim_{n\to\infty} M^n = \mathbf{O_{2\times 2}} \quad \text{(ゼロ行列)}$$

[3] 具体的な証明は **Appendix A** を参照してください

[4] $\mathbb{I}_2$ とは 2×2 の単位行列である。

が証明された。

極限 $n \to \infty$ を式 **[12]** に適用する。第一項 $M^n s_0$ は 0 に収束し、第二項の M^n も 0 になる。

$$\begin{aligned} s^* = \lim_{n\to\infty} s_n &= (\mathbb{I}_2 - O)(\mathbb{I}_2 - M)^{-1}\beta \\ &= (\mathbb{I}_2 - M)^{-1}\beta \end{aligned}$$

式 **[13]** の結果と $\beta = (0, \epsilon ab)^T$ を代入して計算する。

$$s^* = \frac{1}{\epsilon^2 a^2} \begin{pmatrix} 1-\mu & \epsilon \\ -\epsilon a^2 & 0 \end{pmatrix} \begin{pmatrix} 0 \\ \epsilon ab \end{pmatrix}$$

行列の積を計算すると：

- x 成分: $\frac{1}{\epsilon^2 a^2}\left((1-\mu)\cdot 0 + \epsilon \cdot \epsilon ab\right) = \frac{\epsilon^2 ab}{\epsilon^2 a^2} = \frac{b}{a}$
- v 成分: $\frac{1}{\epsilon^2 a^2}\left((-\epsilon a^2)\cdot 0 + 0 \cdot \epsilon ab\right) = 0$

となる。よって、

$$s^* = \begin{pmatrix} b/a \\ 0 \end{pmatrix}$$

これは $f(x)$ の最小値を与える点 $x^* = b/a$（勾配 $a(ax-b) = 0$ の解）と完全に一致する。以上より、モーメンタム法は初期値に関わらず正しい最適解へ収束することが示された。

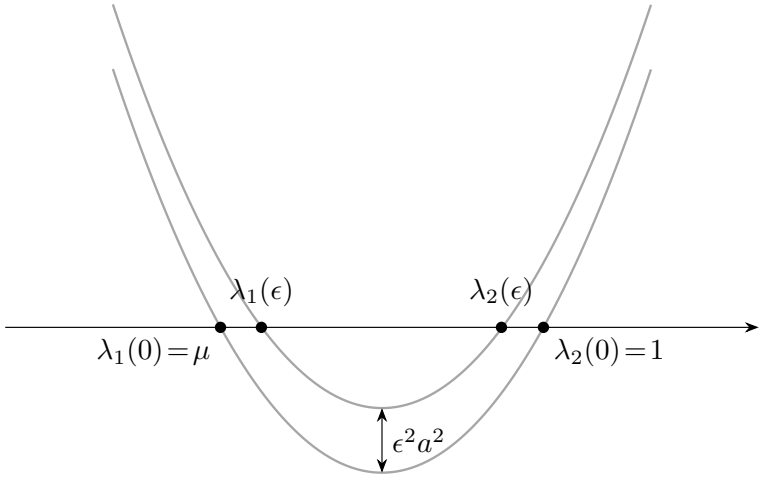

図 2: For small upward shifts the roots λ_i remain real and situated between μ and 1.

■

◆◆◆補足◆◆◆

モーメンタム法の命名について: 通常、物理学における「運動量（Momentum）」は物体を前へ押し進めるものであるが、この手法では実際には摩擦力（$\mu < 1$）を用いて粒子を**減衰（Damp）**させている。これは平衡点を行き過ぎる（**Overshoot**）のを防ぐためである。もし局所（**local**）解から脱出するために「勢い」をつけたい場合は、$\mu > 1$（負の摩擦＝加速）を設定する必要がある。

5 Convergence Conditions 収束条件

モーメンタム法によって生成される数列 $(x^n)_n$ および $(v^n)_n$ の収束性を議論するために、まずこれらの数列の閉じた形式（Closed-form expression）、すなわち一般項を導出する。

5.1 Exact Formulas for Sequences 数列の厳密な公式

位置 x^n と速度 v^n の漸化式を繰り返し展開（Iterating）することで、一般項を求める。

■**位置 x^n の導出**　式 **[8]** $x^n = x^{n-1} + v^n$ を繰り返し適用する。

$$\begin{aligned} x^n &= x^{n-1} + v^n \\ x^{n-1} &= x^{n-2} + v^{n-1} \\ &\vdots \\ x^1 &= x^0 + v^1 \end{aligned}$$

これらを辺々加えると、中間項がなくされ、以下の式が得られる。

$$x^n = x^0 + \sum_{k=1}^{n} v^k \qquad \textbf{[14]}$$

■**速度 v^n の導出**　簡単のため、$b_n = \nabla f(x^n)$ と置く。式 **[9]** $v^n = \mu v^{n-1} - \eta b_{n-1}$ を展開する。

$$\begin{aligned} v^n &= \mu v^{n-1} - \eta b_{n-1} \\ &= \mu(\mu v^{n-2} - \eta b_{n-2}) - \eta b_{n-1} \quad \cdots (\text{where, } v^{n-1} = \mu v^{n-2} - \eta b_{n-2}) \\ &= \mu^2 v^{n-2} - \mu\eta b_{n-2} - \eta b_{n-1} \\ &= \mu^2(\mu v^{n-3} - \eta b_{n-3}) - \mu\eta b_{n-2} - \eta b_{n-1} \\ &= \mu^3 v^{n-3} - \eta(\mu^2 b_{n-3} + \mu b_{n-2} + b_{n-1}) \end{aligned}$$

このように繰り返すと、

$$v^n = \mu^n v^0 - \eta \sum_{j=0}^{n-1} \mu^{n-1-j} b_j$$

インデックスを整理（v^{n+1} の形にし、和の添字を調整）すると、以下の形式が得られる。

$$v^{n+1} = \mu^{n+1} v^0 - \eta \sum_{i=0}^{n} \mu^{n-i} b_i \tag{15}$$

この右辺第 2 項は、勾配 b_i と減衰係数 μ のべき乗との**畳み込み（Convolution）**の形になっている。

5.2 Mathematical Tools for Convergence
収束解析のための数学的道具

v^{n+1} を解析するために、次の畳み込み級数に関する定義と定理を導入する。

Def. 5.1 (Convolution Series):
2 つの数列 $(a_n)_{n\geq 0}$ と $(b_n)_{n\geq 0}$ に対し、その畳み込み級数 $(c_n)_{n\geq 0}$ は以下のように定義される。

$$c_n = \sum_{i=0}^{n} a_i b_{n-i}$$

また、無限級数としての畳み込みは $\sum_{n\geq 0} c_n$ で表される。

Prop. 5.2 (Limit of Convolution Term): [5]
実数列 (a_n) が 0 に収束し、級数 $\sum b_n$ が絶対収束すると仮定する。このとき、畳み込み項の極限は 0 になる。

$$\lim_{n\to\infty} \sum_{i=0}^{n} a_i b_{n-i} = 0$$

Thm. 5.3 (Convergence of Convolution Series): [6]
級数 $\sum a_n$ が収束し、級数 $\sum b_n$ が絶対収束するとする。このとき、それらの畳み込み級数 $\sum c_n$ も収束し、その和は各級数の和の積に等しい。

$$\sum_{n\geq 0} \left(\sum_{i=0}^{n} a_i b_{n-i} \right) = \left(\sum_{n\geq 0} a_n \right) \left(\sum_{n\geq 0} b_n \right)$$

5.3 Main Convergence Results
主要な収束結果

以上の道具を用いて、モーメンタム法の収束条件を記述する以下の重要な命題を証明する。

Prop. 5.4 (Convergence of Momentum Method):

[5] 畳み込み級数の極限については、**Appendix B. B.1** を参照してください。
[6] 畳み込み級数の収束については、**Appendix B. B.2** を参照してください。

(a) もし勾配 $\nabla f(x^n)$ が 0 に収束するならば、速度列 (v^n) も $n \to \infty$ で 0 に収束する。
(b) もし勾配のノルムの級数 $\sum_{n\geq 0} \|\nabla f(x^n)\|$ が収束するならば、数列 (x^n) および (v^n) は共に収束する。

■(a) の証明

Proof.

式 **[15]** を考える。

$$v^{n+1} = \mu^{n+1} v^0 - \eta \sum_{i=0}^{n} \mu^{n-i} b_i$$

ここで $0 < \mu < 1$ であるため、μ^n は $n \to \infty$ で 0 に収束する。よって $\mu^{n+1} v^0 = 0$。
第 2 項について：

- 数列 μ^n は幾何級数であり、$\sum |\mu^n|$ は絶対収束する。
- 仮定より $b_n = \nabla f(x^n) \to 0$ である。

これらを **Prop. 5.2** に適用（a_i を b_i、b_{n-i} を μ^{n-i} とみなす）すると、畳み込み項の極限は 0 となる。したがって、

$$v^* = \lim_{n\to\infty} v^{n+1} = v^0 \cdot 0 - \eta \cdot 0 = 0$$

■

■(b) の証明

Proof.

仮定より $\sum \|\nabla f(x^n)\|$ が収束している。級数が収束するための必要条件は一般項が 0 に収束することであるため[7]、$\|\nabla f(x^n)\| \to 0$、***i.e.*** $\nabla f(x^n) \to 0$。よって、(a) の結果より直ちに v^n **は 0 に収束する**。

次に x^n の収束性を示す。式 **[14]** より $x^{n+1} = x^0 + \sum_{k=0}^{n} v^{k+1}$ であるから、級数 $\sum v^{k+1}$ の収束を示せば良い。式 **[15]** を代入して整理する。

$$\begin{aligned} x^{n+1} &= x^0 + \sum_{k=0}^{n} \left(\mu^{k+1} v^0 - \eta \sum_{i=0}^{k} \mu^{k-i} b_i \right) \\ &= x^0 + v^0 \sum_{k=0}^{n} \mu^{k+1} - \eta \sum_{k=0}^{n} \left(\sum_{i=0}^{k} \mu^{k-i} b_i \right) \end{aligned}$$

[7] 級数が収束するための必要条件（一般項が 0 に収束する）の証明については、**Appendix C** を参照してください。

ここで $n \to \infty$ の極限を取る。

- 第 2 項：等比級数の和 $\sum \mu^{k+1}$ は収束する（和は $\frac{\mu}{1-\mu}$）。
- 第 3 項：$\sum \mu^n$ は絶対収束し、$\sum \|b_n\|$ も収束（仮定より）するため、$\sum b_n$ も絶対収束する。よって **Thm.5.3** より、この畳み込み級数も収束する。

したがって、極限 $x^* = \lim_{n\to\infty} x^{n+1}$ が存在し、以下のように書ける。

$$x^* = x^0 + v^0 \frac{\mu}{1-\mu} - \frac{\eta}{1-\mu} \sum_{n\geq 0} \nabla f(x^n)$$

x^* は閉じた形式の解があることにより、x^n も収束することが示された。 ■

6 Nesterov Accelerated Gradient (NAG)

Rem. 6.1: Nesterov Accelerated Gradientnag 標準的なモーメンタム法の改良版として、Nesterov (1983) によって提案された **NAG** がある。これは勾配を計算する位置を変更する手法である。現在の位置 x^n で勾配を計算する代わりに、慣性項によって「次に到達するであろう位置」$x^n + \mu v^n$ で勾配を評価する。

$$x^{n+1} = x^n + v^{n+1} \quad \textbf{[16]}$$

$$v^{n+1} = \mu v^n - \eta \nabla f(x^n + \mu v^n) \quad \textbf{[17]}$$

この「先読み（Look-ahead）」により、NAG は振動を抑制し、収束率を向上させることができる。

以下に、NAG のアルゴリズム的な詳細と、なぜこの手法が標準的なモーメンタム法より優れているのか、その**更新ロジック**と**その改善点**に分けて分析する。

6.1 Update Logic of NAG
NAG の更新ロジック

数式 **[16]** と **[17]** は理論的に美しいが、実装においては「勾配を計算する点」と「実際に更新する点」が異なるため、手順を明確に理解する必要がある。NAG は以下の 2 段階のプロセスと解釈できる。

Step 1. 予測（**Prediction**）：まず、勾配を無視して、慣性（Momentum）だけで移動した場合の仮の位置 $\hat{x}^n$ を計算する。

$$\hat{x}^n = x^n + \mu v^n$$

Step 2. 修正（**Correction**）：この「先読みした位置 $\hat{x}^n$」における勾配 $\nabla f(\hat{x}^n)$ を計算し、それを用いて現在の速度ベクトルを修正し、最終的な位置更新を行う。

$$v^{n+1} = \mu v^n - \eta \nabla f(\hat{x}^n)$$

$$x^{n+1} = x^n + v^{n+1}$$

このロジックを直感的に言えば、**「盲目のハイカー」と「先を見通すハイカー」の違い**である。標準的な Momentum 法は、現在の斜度だけを見て勢いよく飛び出すが、NAG は「このまま進んだらどうなるか」を一旦確認してから、足の踏み出し方を微調整するのである。

6.2 Major Improvements
主要な改善点

NAG が標準的な勾配降下法（Gradient Descent, GD）や Momentum 法よりも優れている理由は、以下である。

6.2.1 Correction of Overshooting
Overshoot（行き過ぎ）の抑制

谷底（最適解）に向かってボールが転がり落ちる状況を想像してほしい。

- **Standard Momentum**: 谷底に近づいても、過去の加速（μv^n）が残っているため、勢い余って谷底を通り過ぎてしまい、反対側の斜面を登ってしまう（オーバーシュート）。現在の位置での勾配は減速を指示するかもしれないが、過去の勢いが勝る場合がある。
- **NAG (Look-ahead)**: NAG は「もし慣性で進んだら、谷底を通り過ぎて反対側の斜面（上り坂）に達する」ことを**移動する前に**検知する[8]。先読み点での勾配 $\nabla f(x^n + \mu v^n)$ は「戻れ（逆方向）」という情報を指しているため、速度更新式において慣性項 μv^n を強力に打ち消すブレーキとして作用する。

[8] この先読み効果により、パラメータの更新ベクトルは常に「曲率（Curvature）」の情報を間接的に取り込むことになる。

6.3 Visualizing the Vector Update Difference
ベクトル更新の幾何学的比較

最後に、標準的なモーメンタム法と NAG のベクトル更新の違いは以下のようである。

Standard Momentum **Nesterov (NAG)**

図 3: Vector summation comparison.
Left: Standard Momentum adds gradient at x^n and momentum.
Right: NAG moves by momentum first, measures gradient at the look-ahead point, then corrects.

Appendix A　補足証明：行列の等比級数の和と逆行列

Ex 3: 行列の等比級数の和

$S_n = \sum_{k=0}^{n-1} M^k$ と置き、等式の両辺に $I - M$ を掛けると、

$$(\mathbb{I} - M)S_n = (\mathbb{I} - M)(\mathbb{I} + M + \cdots + M^{n-1}) = \mathbb{I} - M^n$$

ここで、$\mathbb{I}$ は単位行列 **(Unit Matrix)**, $M = \begin{pmatrix} 1 & \epsilon \\ -\epsilon a^2 & \mu \end{pmatrix}$

行列 $\mathbb{I} - M$ が正則（逆行列が存在）であると仮定すると、

$$S_n = (\mathbb{I} - M^n)(\mathbb{I} - M)^{-1}$$

となることを示せ

Proof.

■行列の多項式の可換性

一般に行列の積は非可換（$AB \neq BA$）であるが、行列 M とその多項式 $P(M)$ は可換である。$S_n = I + M + \cdots + M^{n-1}$ は M の多項式であるため、M および $(I - M)$ と可換になる。まず、$(\mathbb{I} - M)$ を S_n の右側から掛けた場合を展開する。

$$\begin{aligned} S_n(\mathbb{I} - M) &= (\mathbb{I} + M + M^2 + \cdots + M^{n-1})(\mathbb{I} - M) \\ &= (\mathbb{I} + M + \cdots + M^{n-1}) - (M + M^2 + \cdots + M^n) \\ &= \mathbb{I} + (M - M) + (M^2 - M^2) + \cdots + (M^{n-1} - M^{n-1}) - M^n \\ &= \mathbb{I} - M^n \end{aligned}$$

同様に、左側から掛けても $(\mathbb{I} - M)S_n = \mathbb{I} - M^n$ が成立する。

■M と $(\mathbb{I} - M)$ が可換となる証明

行列 M と $(\mathbb{I} - M)$ が可換であることを示す。

$$M(\mathbb{I} - M) = M - M^2$$
$$^{9}(\mathbb{I} - M)M = M - M^2$$

したがって、$M(\mathbb{I} - M) = (\mathbb{I} - M)M$ が成立する。

[9] $\mathbb{I}M = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} 1 & \epsilon \\ -\epsilon a^2 & \mu \end{pmatrix} = \begin{pmatrix} 1 & \epsilon \\ -\epsilon a^2 & \mu \end{pmatrix} = \begin{pmatrix} 1 & \epsilon \\ -\epsilon a^2 & \mu \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = M\mathbb{I} = M$

■M と $(\mathbb{I}-M)^{-1}$ が可換となる証明
行列 M と $(\mathbb{I}-M)^{-1}$ が可換であることを示す。

$$M(\mathbb{I}-M)^{-1}=(\mathbb{I}-M)^{-1}M$$

両辺に $(\mathbb{I}-M)$ を掛けると、

$$\begin{aligned}M&=(\mathbb{I}-M)^{-1}M(\mathbb{I}-M)\\&=(\mathbb{I}-M)^{-1}(M-M^2)\\&=(\mathbb{I}-M)^{-1}(\mathbb{I}-M)M\\&=M\end{aligned}$$

したがって、$M(\mathbb{I}-M)^{-1}=(\mathbb{I}-M)^{-1}M$ が成立する。

■$(\mathbb{I}-M^n)$ と $(\mathbb{I}-M)^{-1}$ が可換となる証明
行列 $(\mathbb{I}-M^n)$ と $(\mathbb{I}-M)^{-1}$ が可換であることを示す。

$$\begin{aligned}(\mathbb{I}-M^n)(\mathbb{I}-M)^{-1}&=\mathbb{I}(\mathbb{I}-M)^{-1}-M^n(\mathbb{I}-M)^{-1}\\&=(\mathbb{I}-M)^{-1}\mathbb{I}-(\mathbb{I}-M)^{-1}M^n\\&=(\mathbb{I}-M)^{-1}(\mathbb{I}-M^n)\end{aligned}$$

したがって、$(\mathbb{I}-M^n)(\mathbb{I}-M)^{-1}=(\mathbb{I}-M)^{-1}(\mathbb{I}-M^n)$ が成立する。

以上より、行列 M、$(\mathbb{I}-M)$、$(\mathbb{I}-M)^{-1}$、$(\mathbb{I}-M^n)$ はすべて互いに可換であることが示された。したがって、等式の両辺に $(\mathbb{I}-M)^{-1}$ を掛ける操作は、左側からでも右側からでも同じ結果をもたらす。

$$S_n=(\mathbb{I}-M)^{-1}(\mathbb{I}-M^n)=(\mathbb{I}-M^n)(\mathbb{I}-M)^{-1}$$

となる。 ■

Appendix B　補足証明：畳み込み級数の収束性解析

B.1　命題 5.2 の証明：畳み込み項の極限

_Ex_ 4: Limit of Convolution Term

実数列 (a_n) が 0 に収束し、級数 $\sum_{n=0}^{\infty} b_n$ が絶対収束すると仮定する。このとき、以下の極限が成立する。

$$\lim_{n\to\infty} c_n = 0, \quad \text{where } c_n = \sum_{i=0}^{n} a_i b_{n-i}$$

Proof.
目的は、任意の $\epsilon > 0$ に対して、ある N が存在し、すべての $n > N$ において $|c_n| < \epsilon$ となることを示すことである。(***i.e.*** $\forall \epsilon > 0, \exists N \in \mathbb{N}$ s.t. $\forall \mathbb{N} \ni n > N \implies |c_n| < \epsilon$)
まず、項の順序を入れ替えて整理する（$j = n - i$ とおく）。

$$c_n = \sum_{i=0}^{n} a_i b_{n-i} = \sum_{j=0}^{n} b_j a_{n-j}$$

したがって、以下を使って証明する。

$$|c_n| = \left| \sum_{j=0}^{n} b_j a_{n-j} \right| \leq \sum_{j=0}^{n} |b_j||a_{n-j}|$$

■定数の準備

1. 仮定より $\sum b_n$ は絶対収束するため、その総和を $\mathbf{B} = \sum_{n=0}^{\infty} |b_n|$ と置く。$B = 0$ ならば自明であるため、$B > 0$ とする。
2. 仮定より $a_n \to 0$ である。収束する数列は有界であるため、ある定数 $M > 0$ が存在して、すべての n について $|a_n| < M$ が成り立つ。

■収束の定義

1. $\forall \epsilon > 0, \sum |b_n| < \infty : \exists N_1 \in \mathbb{N}$ s.t. $\forall k \in \mathbb{N},\ k > N_1 \implies \sum_{j=k+1}^{\infty} |b_j| < \dfrac{\epsilon}{2M} \quad (M > 0)$
2. $a_n \to 0\ (n \to \infty)$ より、$\forall \epsilon > 0, \exists N_2 \in \mathbb{N}$ s.t. $\forall \mathbb{N} \ni m > N_2 \implies |a_m| < \dfrac{\epsilon}{2B} \quad (B > 0)$

■和の分割

任意の $\epsilon > 0$ を固定する。和を「前半部分」と「後半部分」の 2 つに分割して計算する。整数

$K = N_1$ を用いて、和を分ける。

$$|c_n| \leq \underbrace{\sum_{j=0}^{K} |b_j||a_{n-j}|}_{\text{第 1 項}} + \underbrace{\sum_{j=K+1}^{n} |b_j||a_{n-j}|}_{\text{第 2 項}}$$

■第 1 項の評価（a の収束性を使用）

$n \to \infty$ とすると、各項の a_{n-j} のインデックス $n-j$ は無限大に近づく。$a_n \to 0$ であるため、すべての $0 \leq j \leq K$ に対して以下を満たすことができる。

$$|a_{n-j}| < \frac{\epsilon}{2B}$$

このとき、$n - j > N_2$ where, $0 \leq j \leq K$ が成り立つため、$n > N_2 + K$ (ここで、$j = K$) が必要である。そのため、$m = n - j \geq n - K > N_2$ となる。したがって、第 1 項は以下のように評価できる。

$$\text{第 1 項} = \sum_{j=0}^{K} |b_j||a_{n-j}| < \frac{\epsilon}{2B} \sum_{j=0}^{K} |b_j| \leq \frac{\epsilon}{2B} \cdot B = \frac{\epsilon}{2}$$

■第 2 項の評価（b の絶対収束性を使用）

$j = K + 1 > K = N_1$ かつ $|a_{n-j}| \leq M$ であるため、第 2 項は以下のように評価できる。

$$\text{第 2 項} = \sum_{j=K+1}^{n} |b_j||a_{n-j}| \leq M \sum_{j=K+1}^{n} |b_j| \leq M \sum_{j=K+1}^{\infty} |b_j| < M \cdot \frac{\epsilon}{2M} = \frac{\epsilon}{2}$$

以上より、十分大きな n に対して、

$$|c_n| \leq \sum_{j=0}^{n} |b_j||a_{n-j}| < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon$$

が成り立つ。よって、$\lim_{n\to\infty} c_n = 0$ である。 ■

B.2 定理 5.3 の証明：畳み込み級数の収束

Ex 5: Convergence of Convolution Series

級数 $\sum a_n$ が値 A に収束し、級数 $\sum b_n$ が値 B に**絶対収束**するとする。このとき、畳み込み級数 $C_n = \sum_{k=0}^{n} c_k$ （ただし $c_k = \sum_{i=0}^{k} a_i b_{k-i}$）は収束し、その和は AB に等しい。

$$\sum_{n=0}^{\infty} c_n = \left(\sum_{n=0}^{\infty} a_n\right)\left(\sum_{n=0}^{\infty} b_n\right)$$

Proof.
各級数の部分和を以下のように定義する。

$$A_n = \sum_{i=0}^{n} a_i, \quad B_n = \sum_{i=0}^{n} b_i, \quad C_n = \sum_{k=0}^{n} c_k$$

目的は、$n \to \infty$ のとき $C_n \to AB$ を示すことである。

■部分和 C_n の変形
C_n の定義を展開し、和の順序を入れ替える（下図の三角形領域での和をイメージすると分かりやすい）。

$$\begin{aligned} C_n &= \sum_{k=0}^{n} c_k = \sum_{k=0}^{n}\sum_{i=0}^{k} a_i b_{k-i} \\ &= a_0 b_0 \\ &+ (a_0 b_1 + a_1 b_0) \\ &+ \dots \\ &+ (a_0 b_n + \dots + a_n b_0) \end{aligned}$$

これを b_j についてまとめ直す（a ではなく b に注目するのは、$\sum b_n$ が絶対収束するという強い条件を持っているためである）。b_k が係数となる項を集めると、

$$C_n = \sum_{k=0}^{n} b_k (a_0 + a_1 + \dots + a_{n-k}) = \sum_{k=0}^{n} b_k A_{n-k}$$

となる。

■誤差項への分解
$A_n \to A$ であるから、誤差項 δ_n を定義する。

$$A_n = A + \delta_n \quad (\text{ただし } \lim_{n\to\infty} \delta_n = 0)$$

これを C_n の式に代入する。

$$\begin{aligned} C_n &= \sum_{k=0}^{n} b_k(A + \delta_{n-k}) \\ &= \sum_{k=0}^{n} b_k A + \sum_{k=0}^{n} b_k \delta_{n-k} \\ &= A \underbrace{\sum_{k=0}^{n} b_k}_{B_n} + \underbrace{\sum_{k=0}^{n} b_k \delta_{n-k}}_{E_n\text{(誤差項)}} \\ &= AB_n + E_n \end{aligned}$$

■極限の評価

$n \to \infty$ のときの計算を確認する。

- 第 1 項: AB_n について B_n は B に収束するため、$\lim_{n\to\infty} AB_n = AB$ である。
- 第 2 項: $E_n = \sum_{k=0}^{n} b_k \delta_{n-k}$ について、これは数列 (b_k) と数列 (δ_k) の畳み込みになっている。
  ここで、以下の条件が満たされている。
  1. 数列 (δ_n) は 0 に収束する（定義より）。
  2. 級数 $\sum b_n$ は絶対収束する（定理の仮定より）。

  これはまさに、前述の **Prop.4** の条件と完全に一致する。したがって、命題 5.2 を適用すると、
  $$\lim_{n\to\infty} E_n = 0$$

以上より、

$$\lim_{n\to\infty} C_n = \lim_{n\to\infty} (AB_n + E_n) = AB + 0 = AB = \left(\sum_{n=0}^{\infty} a_n\right)\left(\sum_{n=0}^{\infty} b_n\right)$$

となり、定理は証明された。 ■

Appendix C　補足証明：級数収束の必要条件

Prop. Appendix C.1 (Necessary Condition for Series Convergence):
無限級数 $\sum_{n=0}^{\infty} a_n$ が収束するならば、その一般項 a_n は 0 に収束する。

$$\sum_{n=0}^{\infty} a_n = S \implies \lim_{n\to\infty} a_n = 0$$

Proof.
級数 $\sum a_n$ が値 S に収束するということは、その第 n 部分和 S_n の極限が存在し、それが S であることを意味する。

$$S_n = \sum_{k=0}^{n} a_k, \quad \text{and} \quad \lim_{n\to\infty} S_n = S$$

ここで、一般項 a_n （ただし $n \geq 1$）は、隣り合う部分和の差として表現できる。

$$a_n = S_n - S_{n-1}$$

この等式の両辺について $n \to \infty$ の極限をとる。

- $n \to \infty$ のとき、当然ながら $n-1 \to \infty$ である。
- したがって、数列 (S_n) が S に収束するならば、1 つ項をずらした数列 (S_{n-1}) も同じ値 S に収束する。

$$\lim_{n\to\infty} S_{n-1} = S$$

極限の線形性（差の極限は極限の差）を用いて計算すると、

$$\begin{aligned}\lim_{n\to\infty} a_n &= \lim_{n\to\infty} (S_n - S_{n-1}) \\ &= \lim_{n\to\infty} S_n - \lim_{n\to\infty} S_{n-1} \\ &= S - S \\ &= 0\end{aligned}$$

よって、$\lim_{n\to\infty} a_n = 0$ が示された。 ■

Rem. Appendix C.2 (Note on Converse):
この命題の逆は真ではないである。***i.e.*** 「一般項 $a_n \to 0$ であっても、級数 $\sum a_n$ が収束するとは限らない」。

代表的な反例は**調和級数** $\sum \frac{1}{n}$ である。

$$\lim_{n\to\infty} \frac{1}{n} = 0 \quad \text{であるが、} \quad \sum_{n=1}^{\infty} \frac{1}{n} = \infty \quad \text{(発散)}$$

したがって、$\nabla f(x^n) \to 0$ が確認できたとしても、それだけで x^n が収束する（位置が止まる）保証にはならず、より強い条件（例：Prop 4.4.6 (b) の絶対収束など）が必要となる。