§ 4.4　Momentum Method

伊 冉 (Andre YI)

2026 年 1 月 29 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）に静止するようにする。

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**: $F_f = -\rho \dot{x}(t)$[1]（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** として記述できます。

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

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。

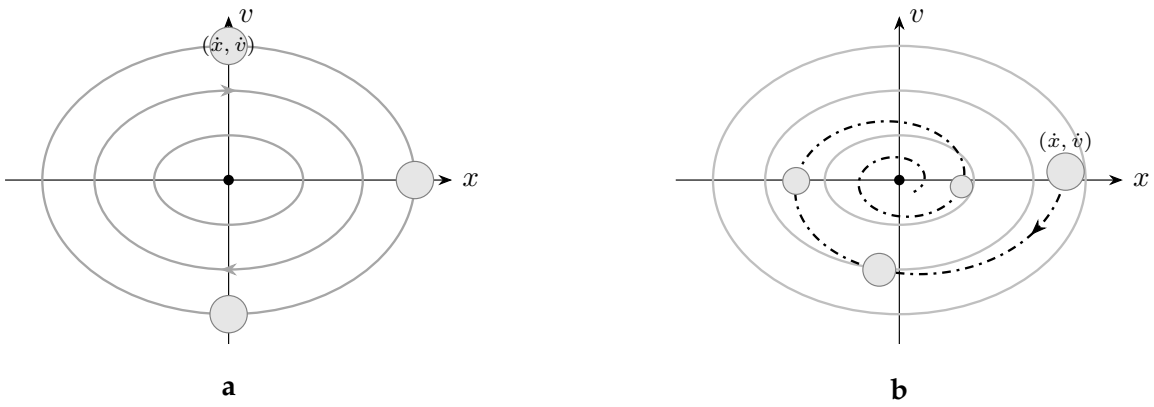

図 1: Solution in phase space.
a. Without friction (div $= 0$), trajectory orbits forever.
b. With friction (div < 0), trajectory spirals into the equilibrium point.

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]** の右辺を変形します：

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

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ます。

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

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

1. 位置の更新式:

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

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

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

したがって、状態 s_n の一般項は以下のように表される。

$$s_n = M^n s_0 + ({}^3\mathbb{I}_2 - M^n)(\mathbb{I}_2 - M)^{-1}\beta \qquad \textbf{[12]}$$

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、

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

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

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

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

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 $0 < \mu < 1$ です。
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] $\mathbb{I}_2$ 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 になります。

$$\begin{aligned} s^* = \lim_{n\to\infty} s_n &= (I_2 - O)(I_2 - M)^{-1}\beta \\ &= (I_2 - M)^{-1}\beta \end{aligned}$$

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

したがって、

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

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。

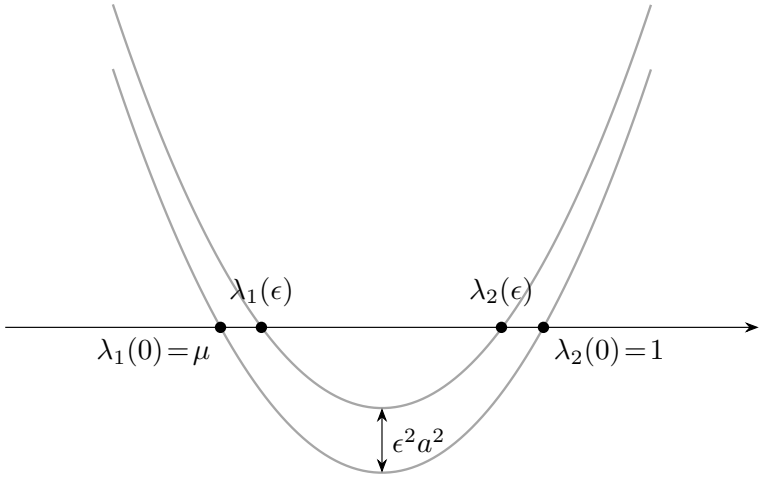

図 2: For small upward shifts the roots λ_i remain real and situated between μ and 1.

■

◆◆◆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。