目的関数の平滑化とディープニューラルネットワークの汎化性能におけるモーメンタム法の慣性項の役割

会員　明治大学　*佐藤 尚樹　SATO Naoki
会員　明治大学　飯塚 秀明　IIDUKA Hideaki

1. はじめに

ディープニューラルネットワークを含む非凸目的関数の最適化において、確率的勾配降下法 (SGD) に慣性項を加えたモーメンタム法は SGD と比べて収束が速く、優れた汎化性能をもたらすことが知られているが、その理論的な説明は不足している。本発表は、モーメンタム法の探索方向と最急降下方向の間の確率的ノイズに注目し、慣性項、確率的ノイズの大きさ、汎化性能の関係を明らかにすることで、なぜ、どのような条件でモーメンタム法が SGD よりも優れるのかを解明する。

2. モーメンタム法

確率的勾配降下法 (SGD) は、ミニバッチ $\mathcal{S}_t$ で全勾配 ∇f を推定したミニバッチ確率的勾配 $\nabla f_{\mathcal{S}_t}$ を利用して、次のように点列を更新する手法である。ただし、$\eta > 0$ は学習率である。

$$\boldsymbol{x}_{t+1} := \boldsymbol{x}_t - \eta \nabla f_{\mathcal{S}_t}(\boldsymbol{x}_t)$$

モーメンタム法は、時刻 t での探索方向に、時刻 $t-1$ での SGD の探索方向の情報を取り入れた手法で、様々な種類がある。本発表では、それらの中で最も基本的な 2 つのモーメンタム法に着目する。Stochastic Heavy Ball method (SHB) は、次のように定義されるモーメンタム法である。ただし、$\beta \in [0, 1)$ は慣性係数であり、$\beta = 0$ のとき、SGD と一致する。

$$\begin{aligned}\boldsymbol{m}_t &:= \nabla f_{\mathcal{S}_t}(\boldsymbol{x}_t) + \beta \boldsymbol{m}_{t-1} \\ \boldsymbol{x}_{t+1} &:= \boldsymbol{x}_t - \eta \boldsymbol{m}_t\end{aligned}$$

PyTorch や TensorFlow が提供するモーメンタム法は SHB であるため、数値実験では非常によく利用される一方で、理論的に解析されることは少ない。多くの先行研究で理論的に解析されてきたのは、次のように定義される Normalized Stochastic Heavy Ball method (NSHB) である。同様に $\beta = 0$ のとき、SGD と一致する。

$$\begin{aligned}\boldsymbol{d}_t &:= (1-\beta)\nabla f_{\mathcal{S}_t}(\boldsymbol{x}_t) + \beta \boldsymbol{d}_{t-1} \\ \boldsymbol{x}_{t+1} &:= \boldsymbol{x}_t - \eta \boldsymbol{d}_t\end{aligned}$$

逆に NSHB は、数値実験では優れず、滅多に使われることはない。

3. モーメンタム法の平滑化特性

まず、一般に関数の平滑化は、正規分布や一様分布に従う確率変数で関数を畳み込むことで実現される。

定義 1 (関数の平滑化) *L_f-Lipschitz* 関数 f を平滑化して得られる関数 $\hat{f}_\delta \colon \mathbb{R}^d \to \mathbb{R}$ は、

$$\hat{f}_\delta(\boldsymbol{x}) := \mathbb{E}_{\boldsymbol{u} \sim B(\boldsymbol{0};1)}\left[f(\boldsymbol{x} - \delta \boldsymbol{u})\right]$$

と表される。ここで、$\delta \in \mathbb{R}$ は平滑化の度合いを表し、$\boldsymbol{u} \in \mathbb{R}^d$ は中心が $\boldsymbol{0}$ で半径が 1 の閉球 $B(\boldsymbol{0};1)$ から一様にサンプリングされたベクトルである。

SHB の探索方向と最急降下方向の間には、各反復で $\boldsymbol{\omega}_t^{\mathrm{SHB}} := \boldsymbol{m}_t - \nabla f(\boldsymbol{x}_t)$ だけ確率的ノイズが生じている。同様に NSHB も、各反復で $\boldsymbol{\omega}_t^{\mathrm{NSHB}} := \boldsymbol{d}_t - \nabla f(\boldsymbol{x}_t)$ だけ確率的ノイズが生じている。このとき、それらの大きさは次のような上界を持つことが示せる。

定義 2 任意の $t \in \mathbb{N}$ に対して、

$$\mathbb{E}\left[\|\boldsymbol{\omega}_t^{\mathrm{SHB}}\|\right] \leq \sqrt{\frac{C_{\mathrm{SHB}}^2}{b} + \hat{\beta}\left(\frac{C_{\mathrm{SHB}}^2}{b} + K_{\mathrm{SHB}}^2\right)},$$

$$\mathbb{E}\left[\|\boldsymbol{\omega}_t^{\mathrm{NSHB}}\|\right] \leq \sqrt{\frac{1}{1-\beta}\frac{C_{\mathrm{NSHB}}^2}{b}}.$$

が成り立つ。ただし、$\hat{\beta} := \frac{\beta(\beta^2-\beta+1)}{(1-\beta)^2}$ とする。

したがって、$\boldsymbol{\omega}_t^{\mathrm{SHB}}$ は、正規分布に従うベクトル $\boldsymbol{u}_t \sim \mathcal{N}\left(\boldsymbol{0}; \frac{1}{\sqrt{d}} I_d\right)$ を使って、

$$\boldsymbol{\omega}_t^{\mathrm{SHB}} = \sqrt{\frac{C_{\mathrm{SHB}}^2}{b} + \hat{\beta}\left(\frac{C_{\mathrm{SHB}}^2}{b} + K_{\mathrm{SHB}}^2\right)}\boldsymbol{u}_t =: \psi \boldsymbol{u}_t,$$

と表すことができる。ここで、確率的ノイズ $\boldsymbol{\omega}_t^{\mathrm{SHB}}$ がどのような分布に従うかを理論的に導出することは困難である。本発表では、$\boldsymbol{\omega}_t^{\mathrm{SHB}}$ が正規分布に従うという実験的観察をもとに、$\boldsymbol{\omega}_t^{\mathrm{SHB}}$ が正規分布に従うと仮定している。

時刻 t において、最急降下法で点列を更新した先を $\boldsymbol{y}_t$ とし、SHB で点列を更新した先を $\boldsymbol{x}_{t+1}$ とする。すなわち、

$$\begin{aligned}\boldsymbol{y}_t &:= \boldsymbol{x}_t - \eta\nabla f(\boldsymbol{x}_t),\\ \boldsymbol{x}_{t+1} &:= \boldsymbol{x}_t - \eta\boldsymbol{m}_t = \boldsymbol{x}_t - \eta(\nabla f(\boldsymbol{x}_t) + \boldsymbol{\omega}_t^{\mathrm{SHB}})\end{aligned}$$

とすると、

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{\omega}_t^{\mathrm{SHB}}}\left[\boldsymbol{y}_{t+1}\right]\\ &\quad= \mathbb{E}_{\boldsymbol{\omega}_t^{\mathrm{SHB}}}\left[\boldsymbol{y}_t\right] - \eta\nabla\mathbb{E}_{\boldsymbol{\omega}_t^{\mathrm{SHB}}}\left[f\left(\boldsymbol{y}_t - \eta\boldsymbol{\omega}_t^{\mathrm{SHB}}\right)\right]\\ &\quad= \mathbb{E}_{\boldsymbol{\omega}_t^{\mathrm{SHB}}}\left[\boldsymbol{y}_t\right] - \eta\nabla\mathbb{E}_{\boldsymbol{u}_t\sim\mathcal{N}\left(\mathbf{0};\frac{1}{\sqrt{d}}I_d\right)}\left[f(\boldsymbol{y}_t - \psi\boldsymbol{u}_t)\right]\\ &\quad\approx \mathbb{E}_{\boldsymbol{\omega}_t^{\mathrm{SHB}}}\left[\boldsymbol{y}_t\right] - \eta\nabla\mathbb{E}_{\boldsymbol{u}_t\sim B(\mathbf{0};1)}\left[f(\boldsymbol{y}_t - \psi\boldsymbol{u}_t)\right]\\ &\quad= \mathbb{E}_{\boldsymbol{\omega}_t^{\mathrm{SHB}}}\left[\boldsymbol{y}_t\right] - \eta\nabla\hat{f}_{\eta\psi}(\boldsymbol{y}_t),\end{aligned}$$

が成り立つ。よって、SHB で関数 f を最適化することは、期待値の意味では大きさ $\eta\psi$ のノイズで f を平滑化した関数 $\hat{f}_{\eta\psi}$ を最急降下法で最適化することと等価であると言える。NSHB にも同様の議論を適用すると、それぞれの最適化手法で、最適化の最中に暗黙のうちに発生し、目的関数の平滑化に寄与している確率的ノイズの大きさ δ^{opt} は、次のように表すことができる。

$$\delta^{\mathrm{SGD}} = \eta\sqrt{\frac{C_{\mathrm{SGD}}^2}{b}}, \tag{1}$$

$$\delta^{\mathrm{SHB}} = \eta\sqrt{\left(1+\hat{\beta}\right)\frac{C_{\mathrm{SHB}}^2}{b} + \hat{\beta}K_{\mathrm{SHB}}^2}, \tag{2}$$

$$\delta^{\mathrm{NSHB}} = \eta\sqrt{\frac{1}{1-\beta}\frac{C_{\mathrm{NSHB}}^2}{b}}, \tag{3}$$

ただし、b はバッチサイズ、C_{opt}^2 は最適化手法ごとの確率的勾配の分散、K_{opt}^2 は最適化手法ごとの全勾配のノルムの上界である。すなわち、$(\boldsymbol{x}_t)_{t\in\mathbb{N}}$ を各最適化手法が生成した点列だとし、$\mathsf{G}_{\boldsymbol{\xi}_t}(\boldsymbol{x}_t)$ を時刻 t における確率的勾配とするとき、$\mathbb{E}_{\boldsymbol{\xi}_t}\left[\|\mathsf{G}_{\boldsymbol{\xi}_t}(\boldsymbol{x}_t) - \nabla f(\boldsymbol{x}_t)\|^2\right] \leq C_{\mathrm{opt}}^2$ と $\mathbb{E}\left[\|\nabla f(\boldsymbol{x}_t)\|^2\right] \leq K_{\mathrm{opt}}^2$ を仮定している。

4. 平滑化の度合いと汎化性能

式 (1)-(3) に注目すると、学習率 η、バッチサイズ b、慣性係数 β はユーザが任意に設定できる値だが、確率的勾配の分散 C_{opt}^2 と全勾配のノルムの上界 K_{opt}^2 は未知数である。これら未知数を数値実験を交えて推定することで、平滑化の度合い δ^{opt} を数値的に導出することに成功した。ResNet18 という深層学習モデルを CIFAR100 データセットで訓練するとき、推定されたこれらの値は、$C_{\mathrm{SGD}}^2 < 1280, C_{\mathrm{SHB}}^2 < 25.3, C_{\mathrm{NSHB}}^2 < 128, K_{\mathrm{SHB}} = 1.77$ となった。未知数の推定の詳細は文献 [1] の 4 章を参照されたい。

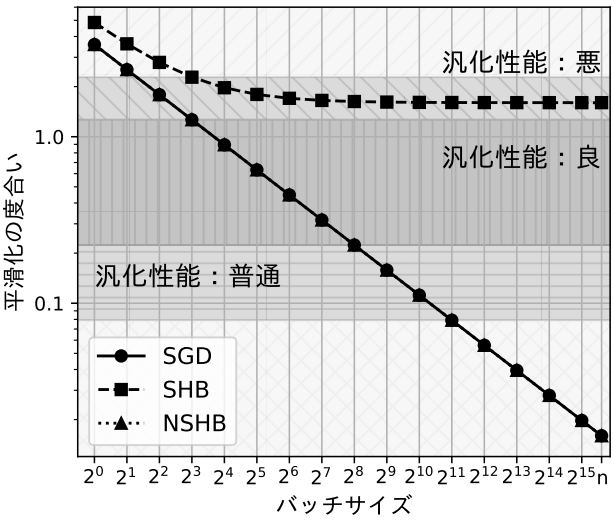

図 1: バッチサイズ b と平滑化の度合い δ^{opt} の関係

図 1 は、$\eta = 0.1, \beta = 0.9$ と、上述の推定値を代入して得られた平滑化の度合い δ^{opt} をプロットしたものである。式 (1)-(3) からも明らかなように、SGD と NSHB はバッチサイズが大きくなればなるほど平滑化の度合いが 0 に近づいていくのに対して、SHB の平滑化の度合いにはバッチサイズと独立な項が存在するため、バッチサイズが大きくなっても平滑化の度合いが小さくなることがない。また、平滑化の度合いはテスト精度とも相関があり、大きすぎず小さすぎない平滑化の度合いがより良い汎化性能をもたらすことが観察された。

参考文献

[1] N. Sato and H. Iiduka, Role of Momentum in Smoothing Objective Function and Generalizability of Deep Neural Networks. arXiv, https://arxiv.org/abs/2402.02325, 2024.