物理学とディープラーニング (ゼミ)

B4 柳瀬調知

2023 年 5 月 10 日

4.3.2 最適化アルゴリズム 後半

勾配降下法での学習における問題点は、振動と局所的極小値に落ちることの 2 つあった。このうち 2 つ目の問題には、ランダム性を取り入れた確率的勾配降下法を用いて対処されることを見てきた。一方で、深い谷がある場合に、更新幅が大きくなりすぎて、いつまでも収束しないことになりうるという問題もある。

1. モーメンタム法

対処法としては、「モーメンタム法」がある。これは、物体の慣性をとりいれた方法で、前の時刻での勾配の影響も取り入れることで、振動を防ぐ方法である。

$$\Delta\boldsymbol{\theta}_t = \eta\boldsymbol{v}_t \tag{1}$$

$$\boldsymbol{v_t} = -(1-\alpha)\nabla L(\boldsymbol{\theta}_t) + \alpha\boldsymbol{v}_{t-1} \tag{2}$$

この式と離散化された運動方程式を変形したものを見比べると、似た形になっているとわかる。

$$\Delta\boldsymbol{X}_t = \Delta t\frac{\boldsymbol{P}_t}{m} \tag{3}$$

$$\frac{\boldsymbol{P}_t}{m} = -\frac{\Delta t}{m}\nabla U(\boldsymbol{X}_t) + \alpha\frac{\boldsymbol{P}_{t-1}}{m} \tag{4}$$

損失関数が急激に落ち込む特異点があっても、慣性があることで滑らかに落ちていく。これにより、意図しないパラメータ $\boldsymbol{\theta}$ の大きなジャンプを防げる。ギザギザ凹凸も慣性があればおかしな挙動が抑制される。

2. AdaGrad

ここまで、計算時間については考えなかった。しかし実際には、プラトー (勾配の小さい広い領域) を抜けるまで多くの時間を要することなどがある。また、座標方向による勾配の違いも問題になる。そこで、今度は学習率を改良する。

$$\Delta\boldsymbol{\theta}_t = -\frac{\eta}{\sqrt{\sum_t \{\nabla L(\boldsymbol{\theta}_t)\}^2}}\nabla L(\boldsymbol{\theta}_t) \tag{5}$$

漸化式を使ってかくと、

$$\Delta\boldsymbol{\theta}_t = -\frac{\eta}{\sqrt{w_t}}\nabla L(\boldsymbol{\theta}_t) \tag{6}$$

$$w_t = w_{t-1} + \{\nabla L(\boldsymbol{\theta}_t)\}^2 \tag{7}$$

となる。分母が時間についての累積となっているのがポイント。すでに大きな勾配をとってきた方向については学習率を減衰させ、いままで勾配が小さかった方向へは学習率を増大させる。

この方法は、学習の初期に勾配が大きいと、更新量が小さくなって戻らないことが弱点となっている。

3. RMSProp

AdaGrad の弱点を克服したものが RMSProp である。AdaGrad のように過去すべての勾配の情報をとりいれるのではなく、十分過去の勾配情報を指数的な減衰因子により消滅させる。

$$\Delta\boldsymbol{\theta}_t = -\frac{\eta}{\sqrt{w_t}}\nabla L(\boldsymbol{\theta}_t) \tag{8}$$

$$w_t = \rho w_{t-1} + (1-\rho)\{\nabla L(\boldsymbol{\theta}_t)\}^2 \tag{9}$$

(9) の漸化式から決まる数列を初項 $w_0 = 0$ として第 4 項まで求めると、

$$w_1 = (1-\rho)\nabla L(\boldsymbol{\theta}_1) \tag{10}$$

$$w_2 = \rho(1-\rho)\nabla L(\boldsymbol{\theta}_1) + (1-\rho)\nabla L(\boldsymbol{\theta}_2) \tag{11}$$

$$w_3 = \rho^2(1-\rho)\nabla L(\boldsymbol{\theta}_1) + \rho(1-\rho)\nabla L(\boldsymbol{\theta}_2) + (1-\rho)\nabla L(\boldsymbol{\theta}_3) \tag{12}$$

となる。過去の情報になるほど減衰されているのがわかる。例えば、初めに勾配が急な区間があり、w_t が非常に大きくなったとしても、フラットな領域を進むうちに w_t は小さくなり学習が進む。ρ=0.99 だと 100 ステップ後にはおよそ 0.37 倍小さくなる。

4. Adam

モーメンタム法と RMSProp を組み合わせたもの。使用頻度が最も高い。

$$\Delta\boldsymbol{\theta}_t = -\frac{\eta}{\sqrt{\hat{w}_t}}\hat{\boldsymbol{v}}_i \tag{13}$$

$$\boldsymbol{v_t} = -(1-\alpha)\nabla L(\boldsymbol{\theta}_t) + \alpha\boldsymbol{v}_{t-1} \tag{14}$$

$$w_t = \rho w_{t-1} + (1-\rho)\{\nabla L(\boldsymbol{\theta}_t)\}^2 \tag{15}$$

・リスケールの詳細

(10)〜(12) 式からの類推により、一般項は

$$w_t = (1-\rho)\sum_{s=1}^{t}\rho^{t-s}\{\nabla L(\boldsymbol{\theta}_t)\}^2 \tag{16}$$

となる。SGD などで訓練サンプルに関する期待値をとると、

$$E[w_t] \simeq E[\{\nabla L(\boldsymbol{\theta}_t)\}^2](1-\rho)\sum_{s=1}^{t}\rho^{t-s} \quad (t \gg 1) \tag{17}$$

$$= E[\{\nabla L(\boldsymbol{\theta}_t)\}^2](1-\rho)\frac{1-\rho^t}{1-\rho} \tag{18}$$

$$= E[\{\nabla L(\boldsymbol{\theta}_t)\}^2](1-\rho^t) \tag{19}$$

となる。ただし、1 行目の近似では十分時間がたって勾配が時間的に一定になったと仮定した。よって、$E[\frac{w_t}{1-\rho^t}] = E[\{\nabla L(\boldsymbol{\theta}_t)\}^2]$ となるので、$\hat{\boldsymbol{w}}_t = \frac{w_t}{1-\rho^t}$ とおいて補正している。

$\hat{\boldsymbol{v}}_t$ についてもほぼ同様になっていて、$\boldsymbol{v}_t$ の一般項は、

$$\boldsymbol{v}_t = -(1-\alpha)\sum_{s=1}^{t}\alpha^{t-s}\nabla L(\boldsymbol{\theta}_t) \tag{20}$$

となる。期待値をとると、

$$E[\boldsymbol{v}_t] \simeq -E[\nabla L(\boldsymbol{\theta}_t)](1-\alpha^t) \quad (t \gg 1) \tag{21}$$

$$E[\frac{\boldsymbol{v}_t}{1-\alpha^t}] = -E[\nabla L(\boldsymbol{\theta}_t)] \tag{22}$$

である。(13) 式に期待値で入れると、

$$\Delta\theta \simeq -\frac{\eta}{\sqrt{E[\{\nabla L(\boldsymbol{\theta}_t)\}^2]}}E[\nabla L(\boldsymbol{\theta}_t)] \quad (t \gg 1) \tag{23}$$

となる。

4.3.3 バックプロパゲーション

損失関数は $\boldsymbol{\theta}$ の非常に複雑な関数。差分近似で数値微分を計算するよりも、解析的に計算できるところはしてしまい、なるべく離散化せずに計算する (自動微分)。

1. フォワードモード

$$\dot{w} = \frac{\partial w}{\partial a}\dot{a} + \frac{\partial w}{\partial b}\dot{b} + \frac{\partial w}{\partial c}\dot{c} \tag{24}$$

$$\dot{\phi} = \frac{\partial \phi}{\partial w}\dot{w} + \frac{\partial \phi}{\partial c}\dot{c} \tag{25}$$

$$\dot{L} = \frac{\partial L}{\partial \phi}\dot{\phi} \tag{26}$$

例えば、a で微分するときは $\dot{w} = \frac{\partial w}{\partial a}, \dot{\phi} = \frac{\partial \phi}{\partial a}, \dot{L} = \frac{\partial L}{\partial a}$、$(\dot{a}, \dot{b}, \dot{c}) = (1, 0, 0)$ となる。

2. リヴァースモード

$$\bar{L} := \frac{\partial L}{\partial L} \tag{27}$$

$$\bar{\phi} := \frac{\partial L}{\partial \phi} = \frac{\partial L}{\partial \phi}\bar{L} \tag{28}$$

$$\bar{w} := \frac{\partial L}{\partial w} = \frac{\partial \phi}{\partial w}\bar{\phi} \tag{29}$$

$$\bar{a} := \frac{\partial L}{\partial a} = \frac{\partial w}{\partial a}\bar{w} \tag{30}$$

初めの 3 式は a, b, c によらず共通だから、1 回の初期値で計算すれば済むので効率的。ニューラルネットでは、アウトプットのニューロン数よりも a, b, c のようなパラメータの方が多いので、リヴァースモードを使った方が効率よい。

・勾配消失問題

2 層のニューラルネット風に $L = L(\phi_2), \phi_2 = \phi_2(w_2), w_2 = w_2(\phi_1), \phi_1 = \phi_1(w_1, c), w_1 = w_1(a, b, c)$ を考える。このとき、

$$\bar{a} := \frac{\partial L}{\partial a} = \frac{\partial w_1}{\partial a}\frac{\partial \phi_1}{\partial w_1}\frac{\partial w_2}{\partial \phi_1}\frac{\partial \phi_2}{\partial w_2}\frac{\partial L}{\partial \phi_2} \tag{31}$$

$\phi_i(w_i)$ がシグモイド関数のとき、$\phi_i^{'}(w_i) = \phi_i(1 - \phi_i(w_i)) \leq \frac{1}{4}$ より、大きく見積もっても $(\frac{1}{4})^2$ の因子がつく。一般に、多層化された m 層では $(\frac{1}{4})^m$ の因子が付くと類推できるので、なかなか更新されなくなる (勾配消失問題)。しかし、ReLU のようなユニットを使えば、微分が 1 で何度掛け合わされても 1 である。