Adamのバイアス補正

- Notation
  - パラメータ θ
  - 1次モーメント（の推定値） m
    - ただし，$m_0 = 0$
  - 2次モーメント（の推定値） v
    - ただし，$v_0 = 0$
  - 損失 L
  - 初期学習率 η
  - 減衰率$\rho_1, \rho_2 \in [0, 1)$
- ある時点 t における1次モーメント

  $m_t = \rho_1 m_{t-1} + (1-\rho_1)\frac{\partial L}{\partial \theta_{t-1}} \qquad (1)$

  の期待値 $\mathbb{E}[m_t]$ と勾配の期待値 $\mathbb{E}\left[\frac{\partial L}{\partial \theta_t}\right]$ （＝本当に知りたい勾配の平均値）にどれだけの違いがあるかを確認する．
- そのため，まずは右辺から1次モーメントmに関する項を削除することから始める．
- 1ステップ前の1次モーメント $m_{t-1} = \rho_1 m_{t-2} + (1-\rho_1)\frac{\partial L}{\partial \theta_{t-2}}$ を式(1)に代入すると，

  $$\begin{aligned} m_t &= \rho_1\left(\rho_1 m_{t-2} + (1-\rho_1)\frac{\partial L}{\partial \theta_{t-2}}\right) + (1-\rho_1)\frac{\partial L}{\partial \theta_{t-1}} \\ &= \rho_1^2 m_{t-2} + (1-\rho_1)\left(\frac{\partial L}{\partial \theta_{t-1}} + \rho_1\frac{\partial L}{\partial \theta_{t-2}}\right) \qquad (2) \end{aligned}$$

  となる．同様に，式(2)に2ステップ前の1次モーメント $m_{t-2} = \rho_1 m_{t-3} + (1-\rho_1)\frac{\partial L}{\partial \theta_{t-3}}$ を代入すると，

  $$\begin{aligned} m_t &= \rho_1^2\left(\rho_1 m_{t-3} + (1-\rho_1)\frac{\partial L}{\partial \theta_{t-3}}\right) + (1-\rho_1)\left(\frac{\partial L}{\partial \theta_{t-1}} + \rho_1\frac{\partial L}{\partial \theta_{t-2}}\right) \\ &= \rho_1^3 m_{t-3} + (1-\rho_1)\left(\frac{\partial L}{\partial \theta_{t-1}} + \rho_1\frac{\partial L}{\partial \theta_{t-2}} + \rho_1^2\frac{\partial L}{\partial \theta_{t-3}}\right) \qquad (3) \end{aligned}$$

  となる．式(2)や式(3)より，過去の1次モーメントを代入する操作を i 回繰り返す，つまり， i ステップ前の1次モーメント$m_{t-i} = \rho_1 m_{t-(i+1)} + (1-\rho_1)\frac{\partial L}{\partial \theta_{t-(i+1)}}$ まで代入すると，

  $$\begin{aligned} m_t &= \rho_1^{i+1} m_{t-(i+1)} + (1-\rho_1)\left(\frac{\partial L}{\partial \theta_{t-1}} + \rho_1\frac{\partial L}{\partial \theta_{t-2}} + \rho_1^2\frac{\partial L}{\partial \theta_{t-3}} + \cdots + \rho_1^i\frac{\partial L}{\partial \theta_{t-(i+1)}}\right) \\ &= \rho_1^{i+1} m_{t-(i+1)} + (1-\rho_1)\left(\rho_1^0\frac{\partial L}{\partial \theta_{t-1}} + \rho_1^1\frac{\partial L}{\partial \theta_{t-2}} + \rho_1^2\frac{\partial L}{\partial \theta_{t-3}} + \cdots + \rho_1^i\frac{\partial L}{\partial \theta_{t-(i+1)}}\right) \\ &= \rho_1^{i+1} m_{t-(i+1)} + (1-\rho_1)\sum_{k=0}^{i}\rho_1^k\frac{\partial L}{\partial \theta_{t-(k+1)}} \qquad (4) \end{aligned}$$

  が得られる．式(4)に $i = t-1$ を代入すると，$t-1$回までの過去の1次モーメントを代入した結果，つまり$m_1 = \rho_1 m_0 + (1-\rho_1)\frac{\partial L}{\partial \theta_0}$ まで代入した結果が得られる．よって，

$$
\begin{aligned}
m_t &= \rho_1^t m_0 + (1-\rho_1)\sum_{k=0}^{t-1}\rho_1^k \frac{\partial L}{\partial \theta_{t-(k+1)}} \\
&= (1-\rho_1)\sum_{k=0}^{t-1}\rho_1^k \frac{\partial L}{\partial \theta_{t-(k+1)}} \quad (\text{※初期値}\ \ m_0 = 0\ \text{より}) \qquad (5)
\end{aligned}
$$

となる．これで目的の表現が得られた．

- 式(5)の両辺の期待値を取ると，

$$
\begin{aligned}
\mathbb{E}[m_t] &= \mathbb{E}\left[(1-\rho_1)\sum_{k=0}^{t-1}\rho_1^k \frac{\partial L}{\partial \theta_{t-k}}\right] \\
&= (1-\rho_1)\sum_{k=0}^{t-1}\rho_1^k \mathbb{E}\left[\frac{\partial L}{\partial \theta_{t-k}}\right]
\end{aligned}
$$

となる．ここで，勾配の期待値はすべて等しい，つまり$\mathbb{E}\left[\frac{\partial L}{\partial \theta_t}\right] = \mathbb{E}\left[\frac{\partial L}{\partial \theta_{t-1}}\right] = \cdots = \mathbb{E}\left[\frac{\partial L}{\partial \theta_1}\right]$ と仮定する（これが仮定できない場合でも，減衰率をρ_1を1以下に設定していれば問題ない）．このとき，

$$
\begin{aligned}
\mathbb{E}[m_t] &= (1-\rho_1)\sum_{k=0}^{t-1}\rho_1^k \mathbb{E}\left[\frac{\partial L}{\partial \theta_t}\right] \\
&= \mathbb{E}\left[\frac{\partial L}{\partial \theta_t}\right](1-\rho_1)\sum_{k=0}^{t-1}\rho_1^k \quad \left(\text{※}\mathbb{E}\left[\frac{\partial L}{\partial \theta_t}\right]\text{が}k\text{に依存しないため}\Sigma\text{の外に出すことができる}\right) \\
&= \mathbb{E}\left[\frac{\partial L}{\partial \theta_t}\right](1-\rho_1)\left(1+\rho_1+\rho_1^2+\cdots+\rho_1^{t-1}\right) \qquad (6)
\end{aligned}
$$

となる．ここで，$1+\rho_1+\rho_1^2+\cdots+\rho_1^{t-1}$は初項1，公比$\rho_1$，項数が$t$の等比数列の和であるため，等比数列の和の公式（参考： https://mathtrain.jp/sumtouhi ）より

$$
\begin{aligned}
1+\rho_1+\rho_1^2+\cdots+\rho_1^{t-1} &= \frac{1\times(1-\rho_1^t)}{1-\rho_1} \\
&= \frac{1-\rho_1^t}{1-\rho_1} \qquad (7)
\end{aligned}
$$

となる．式(6)に式(7)を代入すると，

$$
\begin{aligned}
\mathbb{E}[m_t] &= \mathbb{E}\left[\frac{\partial L}{\partial \theta_t}\right](1-\rho_1)\frac{1-\rho_1^t}{1-\rho_1} \\
&= \mathbb{E}\left[\frac{\partial L}{\partial \theta_t}\right](1-\rho_1^t)
\end{aligned}
$$

となるため，1次モーメントの期待値$\mathbb{E}[m_t]$は勾配の期待値$\mathbb{E}\left[\frac{\partial L}{\partial \theta_t}\right]$の$(1-\rho_1^t)$倍ズレていることがわかる（このズレが**バイアス**）．よって，1次モーメントm_tを$\frac{1}{1-\rho_1^t}$倍すれば，期待値を勾配の期待値と一致させることができる．そのため，バイアスを修正した1次モーメント

$$
\hat{m}_t = \frac{m_t}{1-\rho_1^t}
$$

を更新に使用する．

- 2次モーメントの期待値$\mathbb{E}[v_t]$と勾配の二乗の期待値$\mathbb{E}\left[\frac{\partial L}{\partial \theta_t}^2\right] = \mathbb{E}\left[\frac{\partial L}{\partial \theta_t}\odot\frac{\partial L}{\partial \theta_t}\right]$についても同様に

$$\begin{aligned}\mathbb{E}[v_t] &= \mathbb{E}\left[\frac{\partial L}{\partial \theta_t}^2\right](1-\rho_2)\frac{1-\rho_2^t}{1-\rho_2} \\ &= \mathbb{E}\left[\frac{\partial L}{\partial \theta_t}^2\right](1-\rho_2^t)\end{aligned}$$

となるため，バイアス補正した2次モーメント

$$\hat{v}_t = \frac{v_t}{1-\rho_2^t}$$

を更新に使用する.

- **Note**：勾配の分散 $\mathrm{Var}\left[\frac{\partial L}{\partial \theta}\right]$ は，分散の性質より $\mathrm{Var}\left[\frac{\partial L}{\partial \theta}\right] = \mathbb{E}\left[\frac{\partial L}{\partial \theta}^2\right] - \mathbb{E}\left[\frac{\partial L}{\partial \theta}\right]^2$ （参考： https://mathtrain.jp/variance ）であることから，

  $$\mathrm{Var}\left[\frac{\partial L}{\partial \theta}\right] \leq \mathbb{E}\left[\frac{\partial L}{\partial \theta}^2\right]$$

  である．つまり，分散は2次モーメントで上から抑えることができるため，2次モーメントは厳密には分散ではない．二次モーメントは通常の分散よりも大きくなるが，ばらつきを評価した指標として「分散」と表現しているときもある.