ニューラルネットワーク実習04　最適化

浅川伸一

最適化

- **Adagrad** (Duchi, Hazan, and Singer 2011)
- **Adadelta** (Zeiler 2012a)
- **RMSProp** (Tieleman and Hinton 2012)
- **Adam** (Kingma and Ba 2015)
- **自然勾配法** (Amari 1998)

勾配降下法とは

盲目の登山者アナロジー blind hiker analogy で語られる。

$$w_{t+1} = w_t - \alpha_t \mathbf{g}(w) \tag{1}$$

$$\mathbf{g}(w) = \frac{\partial l}{\partial w}$$

```
import torch.optim

optimizer = optim.SGD(model.parameters(), lr=0.01,
        momentum=0.9)
optimizer = optim.Adam(model.parameters(), lr=0.0001)
optimizer = optim.Adadelta(model.paramters())
optimizer = optim.Adagrad(model.paramters(),
        weight_decay=0)
optimizer = optim.RMSprop(model.paramters())
```

確率的学習 SGD

(Bottou and Bousquet 2007)

損失関数 $l(w)$ を**ランダムサンプリング** random sampling して置き換える。

$$w_{t+1} = w_t - \alpha_t \mathbf{g}(w) \tag{2}$$

$$w_{t+1} = w_t - \alpha_t \mathbb{E}[\mathbf{g}(w)] \tag{3}$$

- ここで α は学習係数
- $\mathbb{E}$ は**期待値**をとる演算

$$\mathbb{E}[\mathbf{g}(w)] = \frac{1}{N} \sum_{i=1}^{N} \mathbf{g}(w) \tag{4}$$

モーメント

慣性項 と訳されます。一つ前の勾配と現在の勾配とを利用してパラメータを更新します。

$$\Delta w_{t+1} = m\Delta w_t - \alpha \mathbf{g}(w_t) \tag{5}$$

ニュートン法

勾配降下法では，1 次微分だけを用いて，パラメータを更新します。

これに対して，2 次微分を用いてパラメータを更新する手法をニュートン法と総称します。

wikipeida のニュートン法では
https://en.wikipedia.org/wiki/Newton%27s_method_in_optimization

$$\mathbf{x}_{n+1} = \mathbf{x}_n - \gamma[\mathbf{H}f(\mathbf{x}_n)]^{-1}\nabla f(\mathbf{x}_n). \tag{6}$$

表記をニューラルネットワーク風に書き換えると

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \gamma\frac{\mathbf{g}(w_t)}{\mathbf{H}(\mathbf{w}_t)}. \tag{7}$$

- 上式で $\mathbf{H}$ は 2 階微分を表します。1 階微分である勾配 $\mathbf{g}$ がベクトルで表されているのであれば，　各パラメータについてもう一度微分をするので，H は行列となります。

- ここでは，そのことを考慮して，ベクトルを $\mathbf{g}$ のごとく太字の小文字で表記し，行列を $\mathbf{H}$ のように太字の大文字で表記してあります。
- 学習係数を固定したパラメータであると捉えるのではなく，その都度，ヘシアン行列を用いて調整すること。
- ニュートン法は計算コストがかかるという問題がある。そのため，ニュートン法を 1 階微分で代用できないかという研究が多数行われています。

Adagrad (Duchi, Hazan, and Singer 2011)

$$\Delta w_{t+1,i} = -\frac{\alpha}{\sqrt{G_{t,ii} + \epsilon}} \nabla J\left(w_{t,}\right).$$

- ここで $G_{t,ii}$ は $\Delta w_{t,i}$ を対角成分とする**対角行列** (の自乗), すなわち H の対角近似, かつ `Adagrad` は**各要素** i **について個別に考慮**
- $\epsilon = 1e-8$ は分母が 0 にならないための保険
- ヘシアン行列を対角要素だけで近似する

Adadelta (Zeiler 2012b)

$$\mathbb{E}\left[\mathbf{g}^2\right]_t = \gamma \mathbb{E}\left[\mathbf{g}^2\right]_{t-1} + [1-\gamma]\,\mathbf{g}$$

- Adagrad が行っていたヘシアン行列の近似を，その都度計算するのではなく，履歴を保持しておいて平均値で近似することで精度向上を意図した手法

$$\Delta w_t = -\frac{\alpha}{\sqrt{\mathbb{E}\left[g^2\right]_t + \epsilon}}\mathbf{g}\left(w_{t,i}\right)$$

ここで $\sqrt{\mathbb{E}\left[g^2\right]_t}$ は平均自乗和の開平 root mean squared :RMS と見なせるので

$$\Delta w_{t,i} = -\frac{\alpha}{\mathrm{RMS}[g]_t}\mathbf{g}\left(w_{t,i}\right)$$

- 実際，繰り返し回数を t とすると $\frac{1}{t} = \gamma$ とすれば $1-\gamma = \frac{t-1}{t}$ となるので，平均を 計算していることになる。

RMSprop

(Tieleman and Hinton 2012)

$$\mathbb{E}\left[g^2\right]_t = 0.9\mathbb{E}\left[g^2\right]_{t-1} + 0.1g_t^2$$

$$\Delta w_{t,i} = -\frac{\alpha}{\sqrt{\mathbb{E}\left[g^2\right] + \epsilon}}\mathbf{g}\left(w_{t,i}\right)$$

$\gamma = 0.9$, $\alpha = 0.001$ が使われる。

Adam

(Kingma and Ba 2015)

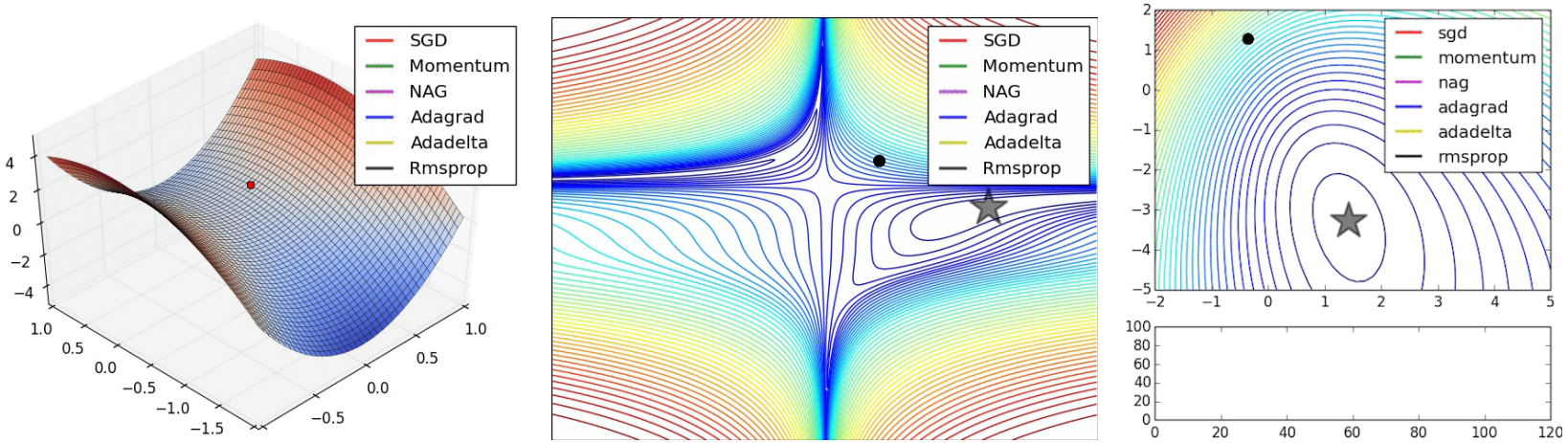
$$m_t = \beta_1 m_{t-1} + [1 - \beta_1] J(w_t)$$
$$v_t = \beta_2 v_{t-1} + [1 - \beta_2] g^2$$

$$\Delta w_{t+1} = -\frac{\alpha}{\sqrt{v_t} + \epsilon} m_t.$$

ここで $\beta_1 = 0.9999$, $\beta_2 = 1e-8$

- Adam の名前は適応モーメント推定（adaptive moment estimation）に由来
- Adam は 適応的手法とモーメンタム法の両方の利点を持っています。

Gif アニメ



出典 アレックス ラッドフォード YouTube <https://imgur.com/a/Hqolp>

AdaDelta や RMSprop と同様 Adam は過去の勾配のスライディングウィンドウに基づいて学習率をパラメータごとに調整し ます。 さらに時間に沿って道筋を滑らかにするためのモーメンタムの成分もあります。

他にも多くの手法が存在します。 派生形や実用的なヒントを含めたより包括的な議論は Sebastian Ruder の ブログ記事を参照してください。

自然勾配法 Natural gradient method

(Amari 1998)

- ヘシアン行列をフィッシャーの情報行列 I で置き換える

$$\Delta \mathbf{w}_{t+1} = -\frac{\alpha}{\mathbf{I}(\mathbf{w}_t)} \mathbf{g}(\mathbf{w}_t).$$

- フィッシャーの情報量行列は，目標関数の対数尤度の 2 回微分です

$$\mathbf{I} = \mathbb{E}\left[\log\left(\frac{\partial^2 w}{\partial w_i \partial w_j}\right)\right]$$

- 情報幾何学の基本にもなっている

まとめ

- 最適化手法として，Adagrad，AdaDelta，RMSprop，Adam，自然勾配法 を紹介しました。
- いずれの手法も，ニュートン法におけるヘシアン行列の近似を考えています。どのようなアイデアで近似を行うのかで手法ごとに特徴があります

クイズ

勾配降下法を最適化する手法の中で，フィッシャーの情報行列を用いる手法の名称は何だったでしょうか

文献

Amari, Shun-ichi. 1998. “Natural Gradient Works Efficiently in Learning.” *Neural Computation* 10: 251–76.

Bottou, Leon, and Olivier Bousquet. 2007. “The Tradeoffs of Large Scale Learning.” In *Advances in Neural Information Processing Systems*. Vol. 20. Cambridge, MA, USA: MIT Press.

Duchi, John, Elad Hazan, and Yoram Singer. 2011. “Adaptive Subgradient Methods for Online Learning and Stochastic Optimization.” *Journal of Machine Learning Research* 12: 2121-59.

Kingma, Diederik P., and Jimmy Lei Ba. 2015. “ADAM: A Method for Stochastic Optimization.” *arXiv Preprint*, 1–13.

Tieleman, T., and G. Hinton. 2012. “Lecture 6.5 - RMSProp, COURSERA: Neural Networks for Machine Learning.” COURSEA.

Zeiler, Matthew D. 2012a. “ADADELTA: An Adaptive Learning Rate Method.” *CoRR*.

———. 2012b. “ADADELTA: An Adaptive Learning Rate Method.” *ArXiv Preprint*. https://arxiv.org/abs/1212.5701.