### 5.3.2 単純な例

5.3.1 では、誤差逆伝播法の具体的な手続きについて学んだ。その手続きとは次のようなものであった。

1. 入力ベクトルをネットワークに入れ順伝搬させ、全てのユニットの出力を求める。
2. 出力ユニットにおける誤差 $\delta_k = y_k - t_k$ を算出する。
3. 隠れユニットにおける誤差 $\delta_j = h'(a_j) \sum_k w_{kj} \delta_k$ を逆伝搬により求める。
4. 全ユニットで微分値 $\frac{\partial E_n}{\partial w_{ji}} = \delta_j z_i$ を評価する。

本節では、活性化関数に $\tanh(\cdot)$ を用いた2層ネットワークについて具体的に示している。ただ、この節は前節の内容と重複する部分があるため、後に実際のコードでロジックを確かめるのみとする。

### 5.3.3 逆伝搬の効率

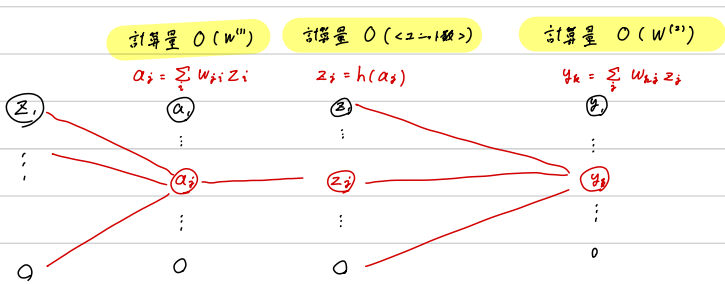
本節は、次節につながる話題ではなく、誤差逆伝播法にまつわる「コラム」のような立ち位置であると理解して読み進める。特に、誤差逆伝播法が計算量 $O(W)$ で誤差の微分値を評価するアルゴリズムであり、後に述べる数値微分の計算量 $O(W^2)$ よりも小さく、正確であることを示す。

まず、誤差の微分が $O(W)$ で計算できることを示すために、2層ネットワークを例に、各ステップごとの計算量を見ていくこととする。

1. 順伝搬により全ユニットの出力を求める。

$W^{(1)}$ : 1層目のパラメータ数 , $W^{(2)}$ : 2層目のパラメータ数 とする。

下図より、計算量は $O(W^{(1)} + W^{(2)}) = O(W)$ となる。 ( <ユニット数> $\ll W$ のため)

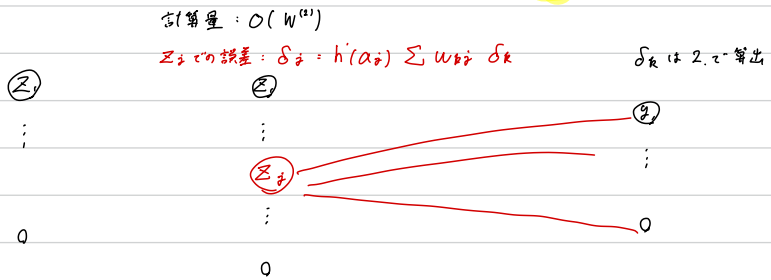


2. 出力ユニットの誤差を算出する

$\delta_k = y_k - t_k$ を評価するだけなので、計算量は $O(<\text{出力ユニット数}>)$

3. 隠れユニットの誤差を誤差逆伝播で算出する。

計算量は、入力層を除く層のパラメータ数なので、$O(W)$




4. 誤差関数の微分を算出する

$\frac{\partial E_n}{\partial w_{ji}} = \delta_j z_i$ を全パラメータで算出するため、計算量は $O(W)$ 。

1−4の結果をまとめると、誤差逆伝播法により誤差関数の微分値を求めるための全体の計算量は $O(W)$ となる。（ユニット数はパラメータ数よりも十分小さいとする）

一方で、微分値を数値計算によって算出する方法がある。これは計算量が $O(W^2)$ であり、誤差逆伝播法の計算量よりも非常に大きいことを示す。この方法は精度においても計算量においても劣った手法ではあるが、アルゴリズムのテストとして実用できる。（後にコードで確かめる）

パラメータ $w_{ji}$ による誤差関数の微分値は次の式で評価できる。

$$\frac{\partial E_n}{\partial w_{ji}} = \frac{E_n(w_{ji}+\varepsilon) - E_n(w_{ji})}{\varepsilon} + O(\varepsilon)$$

$E(w_{ji})$ の評価は通常の順伝搬を一度実行すれば評価できる。$E_n(w_{ji}+\varepsilon)$ の評価は、パラメータ $w_{ji}$ に摂動を加えたもので順伝搬を実行すると評価できる。よって、パラメータによる微分の計算量は $O(W)$ である。

これを $w_{ji}$ 以外のそれぞれのパラメータについて算出すると、全体の計算量は $O(W^2)$ となる。

### 5.3.4 ヤコビ行列

これまで扱ってきた逆伝播の手続きが、他の微分計算にも応用できることを示す。

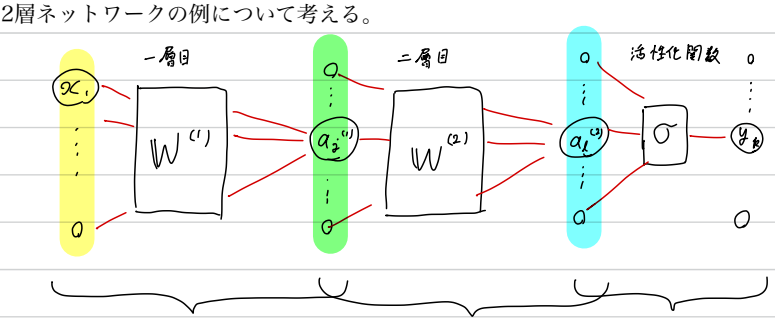
（前節・次節とのストーリーのつながりはほぼないと考えられる）

特に次式で示されるヤコビ行列が逆伝播で求まることを示す。

$$J_{ki} \equiv \frac{\partial y_k}{\partial x_i}$$

ヤコビ行列が求まると何が嬉しいかについては特に言及されていない。しかし式の形から考察すると、クラス k に対する予測においてどの特徴量を重要視しているのかを確かめることができそう。

2層ネットワークの例について考える。




①、②、③の順番で式を導出する。実際に値を算出する際は、③で求まったものを②に代入、②で求まったものを①に代入、というように逆伝播の枠組みで算出する。

①、②、③の式はそれぞれ次の式に相当する。（式の導出は 5.3.1 の誤差関数微分の評価とほぼ同じなので詳細は割愛する。）

$$① \quad J_{ki} = \frac{\partial y_k}{\partial x_i} = \sum_j \frac{\partial y_k}{\partial a_j^{(1)}} \cdot \frac{\partial a_j^{(1)}}{\partial x_i} \quad (\text{chain rule})$$

$$= \sum_j w_{ji}^{(1)} \frac{\partial y_k}{\partial a_j^{(1)}}$$

② $$\frac{\partial y_k}{\partial a_j^{(1)}} = \sum_l \frac{\partial y_k}{\partial a_l^{(2)}} \frac{\partial a_l^{(2)}}{\partial a_j^{(1)}} = \sum_l w_{lj}^{(2)} \cdot \frac{\partial y_k}{\partial a_l^{(2)}} = h'(a_j^{(1)}) \sum_l w_{lj}^{(2)} \frac{\partial y_k}{\partial a_l^{(2)}}$$

③ ．＜ $\sigma$ が シグモイド関数の場合 ＞

$$\frac{\partial y_k}{\partial a_l^{(2)}} = \frac{\partial y_l}{\partial a_l^{(2)}} = \frac{\partial \sigma(a_l^{(2)})}{\partial a_l^{(2)}} = \sigma'(a_l^{(2)})$$ （教科書では $\delta_{kl}\,\sigma'(a_l^{(2)})$）

出力ユニットは1つだけ

＜ $\sigma$ が ソフトマックス関数の場合 ＞

$$\frac{\partial y_k}{\partial a_l^{(2)}} = y_k(\delta_{kl} - y_l)$$ （ref. ソフトマックス関数の微分 : p 208 (4.106)）

以上より、ヤコビ行列についても誤差逆伝播で算出可能であることがわかった。

## 5.4 ヘッセ行列

5.4 節ではヘッセ行列を効率よく算出するためのいくつかの方法を学ぶ。ヘッセ行列を学ぶ理由については p251 の中段に箇条書きで記してあるので読み合わせとする。4. に記載の通り5.7 節で使用する。その際は、ベイズニューラルネットワークのパラメータの事後分布の分散として登場する。p282 (5.166)

ここで求めるヘッセ行列は、誤差関数のパラメータによる2階微分のことである。

$$H = \frac{\partial^2 E}{\partial w_{ji} \partial w_{lk}}$$

のちの項で学ぶ方法のまとめを先取りして下記表にまとめる。

| | 計算量 | 精度 | その他 |
|---|---|---|---|
| 5.4.1 対角近似 | $O(W)$ | △ | |
| 5.4.2 外積による近似 | $O(W^2)$ | ○ | 逆行列も効率良く求まる。 |
| 5.4.4 有限幅の差分による近似 | $O(W^3)$ | △ | |
| 有限幅の差分による近似（改） | $O(W^2)$ | △ | |
| 5.4.5 厳密な評価（誤差逆伝播の利用） | $O(W^2)$ | ◎ | |

どの手法が良いか、については明確に言及されていないが、日本語から汲み取るに、5.4.2 の方法か、5.4.5 の方法が良く使われていそうである。

ひとつずつ手法をみていく。

### 5.4.1 対角近似

ヘッセ行列を使用する際は、ヘッセ行列自身よりもヘッセ行列の逆行列を用いることが多いらしい。$W \times W$ 行列の逆行列を算出するための計算量は（ググったところ）$O(W^3)$ らしい。対角近似の方法は、ヘッセ行列が $O(W)$ で算出でき、その逆行列も $O(W)$ で算出できる点で優れている。一方で近似の精度は良くない。

ヘッセ行列の対角成分 $H_{(i,j),(i,j)}$ は次のように算出される。

$$H_{(i,j),(i,j)} = \frac{\partial^2 E_n}{\partial w_{ji}^2} = \frac{\partial^2 E_n}{\partial a_j^2}\left(\frac{\partial a_j}{\partial w_{ji}}\right)^2 + \frac{\partial E_n}{\partial a_j}\cdot\frac{\partial^2 a_j}{\partial w_{ji}^2} \quad (*)$$

$$= \frac{\partial^2 E_n}{\partial a_j^2}\left\{\frac{\partial}{\partial w_{ji}}\left(\sum_i w_{ji} z_j\right)\right\}^2 + \frac{\partial E_n}{\partial a_j}\cdot\frac{\partial^2}{\partial w_{ji}^2}\left(\sum_i w_{ji} z_j\right)$$

$$= \frac{\partial^2 E_n}{\partial a_j^2}\cdot z_j^2 + \frac{\partial E_n}{\partial a_j}\cdot 0$$

$$= \frac{\partial^2 E_n}{\partial a_j^2} z_j^2$$

（＊）は2階微分に対して Chain rule を適用することによって導出できる。

$$\frac{\partial^2 E_n}{\partial w_{ji}^2} = \frac{\partial}{\partial w_{ji}}\left(\frac{\partial E_n}{\partial a_j}\cdot\frac{\partial a_j}{\partial w_{ji}}\right)$$

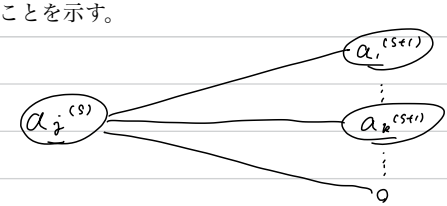
$$= \frac{\partial^2 E_n}{\partial w_{ji}\,\partial a_j}\cdot\frac{\partial a_j}{\partial w_{ji}} + \frac{\partial E_n}{\partial a_j}\cdot\frac{\partial^2 a_j}{\partial w_{ji}^2}$$

$$= \frac{\partial}{\partial a_j}\left(\frac{\partial E_n}{\partial a_j}\frac{\partial a_j}{\partial w_{ji}}\right)\frac{\partial a_j}{\partial w_{ji}} + \frac{\partial E_n}{\partial a_j}\cdot\frac{\partial^2 a_j}{\partial w_{ji}^2}$$

$$= \left\{\frac{\partial^2 E_n}{\partial a_j^2}\cdot\frac{\partial a_j}{\partial w_{ji}} + \frac{\partial E_n}{\partial a_j}\cdot\underbrace{\frac{\partial^2 a_j}{\partial a_j\,\partial w_{ji}}}_{=\frac{\partial 1}{\partial w_{ji}}=0}\right\}\frac{\partial a_j}{\partial w_{ji}} + \frac{\partial E_n}{\partial a_j}\cdot\frac{\partial^2 a_j}{\partial w_{ji}^2}$$

$$= \frac{\partial^2 E_n}{\partial a_j^2}\cdot\left(\frac{\partial a_j}{\partial w_{ji}}\right)^2 + \frac{\partial E_n}{\partial a_j}\cdot\frac{\partial^2 a_j}{\partial w_{ji}^2}$$

$\frac{\partial^2 E_n}{\partial a_j^2}$ は誤差逆伝播法によって $O(W)$ で求めることができる。これを示すために、$a_j^{(s)}$ を1つ出力側の層のユニット $a_k^{(s+1)}$ の1階微分 $\frac{\partial E_n}{\partial a_k^{(s+1)}}$ と2階微分 $\frac{\partial^2 E_n}{\partial a_k^{(s+1)2}}$ を使って表すことを示す。



$$\frac{\partial^2 E_n}{\partial a_i^{(s)2}} = \text{（2階微分の Chain rule を適用）}$$

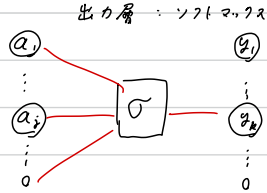
$$= h'(a_i^{(s)})^2 \sum_k \sum_{k'} w_{ki} w_{k'i} \frac{\partial^2 E_n}{\partial a_k^{(s+1)} \partial a_{k'}^{(s+1)}} + h''(a_i^{(s)}) \sum_k w_{ki} \frac{\partial E_n}{\partial a_k^{(s+1)}}$$

対角成分 以外 0

$$= h'(a_i^{(s)})^2 \sum_k w_{ki}^2 \frac{\partial^2 E_n}{\partial a_k^{(s+1)2}} + h''(a_i^{(s)}) \sum_k w_{ki} \frac{\partial E_n}{\partial a_k^{(s+1)}}$$

隠れ層については伝播の式が分かったが、出力層の部分だけは別個で考慮する必要がある。

出力層：ソフトマックス関数




$$\frac{\partial E_n}{\partial a_i} = \sum_k \frac{\partial E_n}{\partial y_k} \frac{\partial y_k}{\partial a_i} = \sum_k \frac{\partial}{\partial y_k} \left\{ \frac{1}{2} \sum_s (y_s - t_s)^2 \right\} \cdot y_k (\delta_{ki} - y_i)$$

$$= \sum_k \frac{1}{2} \left\{ \sum_s \delta_{ks} \cdot 2 (y_s - t_s) \right\} \cdot y_k (\delta_{ki} - y_i)$$

$$= \sum_k (y_k - t_k) y_k (\delta_{ki} - y_i)$$

$$\frac{\partial^2 E_n}{\partial a_i^2} = \frac{\partial}{\partial a_i} \left( \frac{\partial E_n}{\partial a_i} \right) = \left( \begin{array}{l} \text{あとは頑張れば計算できそう．} \\ \frac{\partial y_k}{\partial a_i} = y_k (\delta_{ki} - y_i) \text{ を使う} \end{array} \right)$$

出力層における微分も導出できたので、逆伝播によってヘッセ行列の対角成分が求まる。誤差関数の1階微分の逆伝播と2階微分の逆伝播をそれぞれ $O(W)$ で実行できるため、ヘッセ行列を算出するための計算量は $O(W)$ となる。