今回は出来が悪いです。先に謝ります。

1　ヘッセ行列

まずは前回の振り返りとして、そもそもヘッセ行列を計算するモチベーションについて再確認する。前回レジュメ（0327 輪講.pdf）の p.5 にも記載の通り、後に 5.7 節においてヘッセ行列はベイズニューラルネットワークの重みの事後分布を考える際、その分布の共分散行列に顔を出す量となる。したがいモチベーションとしては、後に現れる表式内の量について、その計算アルゴリズムを頭に入れておこうというものであった[*1]。

以下ではまず誤算関数が特定の形で与えられる場合について対応するヘシアンを計算し、そのうえで近似表現を得ることとする（5.4.2-5.4.4）[*2]。そのあとで、近似を用いず厳密な評価を行う場合についての表式も得る（5.4.5）。そして最後に、ヘシアンの計算を実行するにあたって、計算負荷のオーダー及び計算速度の観点から効率的な手法について紹介する。

1.1　外積による近似

誤差関数を二乗和誤差で与えたときに、ヘシアンがベクトル（出力の grad）の外積を用いて近似的に表されることを示す。外積の形で求めておくと、あとで逆行列が計算しやすいという嬉しさがある。なお、ここではニューラルネットワークを回帰問題に使うことを前提に議論する。

まずヘシアンを得るために誤差関数に nabla 演算子を 2 回かましてやる。

$$
\begin{aligned}
\boldsymbol{H} &= \nabla\nabla E \\
&= \nabla\nabla \frac{1}{2}\sum_{n=1}^{N}(y_n - t_n)^2 \ (\because (5.82)) \\
&= \nabla \frac{\partial}{\partial y_n}\frac{\partial y_n}{\partial \boldsymbol{w}}\frac{1}{2}\sum(y_n - t_n)^2 \\
&= \nabla \frac{\partial y_n}{\partial \boldsymbol{w}}\frac{1}{2}\cdot 2\sum(y_n - t_n) \\
&= \nabla \frac{\partial y_n}{\partial \boldsymbol{w}}\sum(y_n - t_n) \\
&= \nabla\nabla y_n \sum(y_n - t_n) + \frac{\partial y_n}{\partial \boldsymbol{w}}\sum(y_n - t_n) \\
&= \sum(y_n - t_n)\nabla\nabla y_n + \sum \nabla y_n (\nabla y_n)^T = (5.83)
\end{aligned}
\tag{1}
$$

ここから（5.83）の第二項を消去する近似を考える。この項を消去して良いという考え方を確率変数が連続な場合について示す（対応；演習 5.17）。そのあと離散的なサンプリングに議論を変えてやると（5.84）が導かれる。

[*1] だったら 5.7 節の直前でやれよという話だが...5.4 と 5.5 の繋がりはよくわからないわけだし。

[*2] p.282 サブセクション 5.7 において（5.166）内のヘシアンを求める際には近似された表現を用いる。この後すぐ学ぶように近似手法はいくつかあって（対角近似/外積近似/有限幅による近似）、その中のどれを使っていくのかは明示されていないが、計算負荷や速度を考慮に入れて適当な手法を選択していくのだと推察される。

まず二乗和誤差が（5.193）で与えられるとする。これを w_s で Chain Rules を使いながら偏微分してやると

$$\begin{aligned}\frac{\partial E}{\partial w_s} &= \frac{\partial}{\partial w_s}\frac{1}{2}\int\int (y(\boldsymbol{x},\boldsymbol{w})-t)^2 p(\boldsymbol{x},t)d\boldsymbol{x}dt \ (\because\ (5.193)) \\ &= \frac{\partial y}{\partial w_s}\frac{\partial}{\partial y}\frac{1}{2}\int\int (y(\boldsymbol{x},\boldsymbol{w})-t)^2 p(\boldsymbol{x},t)d\boldsymbol{x}dt \\ &= \frac{\partial y}{\partial w_s}\int\int (y(\boldsymbol{x},\boldsymbol{w})-t) p(\boldsymbol{x},t)d\boldsymbol{x}dt \end{aligned} \tag{2}$$

次に上式を w_r で同じように偏微分してやる。積の微分公式も適当に使えば良い。

$$\begin{aligned}\frac{\partial}{\partial w_r}\frac{\partial E}{\partial w_s} &= \frac{\partial}{\partial w_r}\frac{\partial y}{\partial w_s}\int\int (y(\boldsymbol{x},\boldsymbol{w})-t) p(\boldsymbol{x},t)d\boldsymbol{x}dt \\ &= \frac{\partial^2 y}{\partial w_r \partial w_s}\int\int (y(\boldsymbol{x},\boldsymbol{w})-t) p(\boldsymbol{x},t)d\boldsymbol{x}dt + \frac{\partial y}{\partial w_s}\frac{\partial y}{\partial w_r}\frac{\partial}{\partial y}\int\int (y(\boldsymbol{x},\boldsymbol{w})-t) p(\boldsymbol{x},t)d\boldsymbol{x}dt \\ &= \frac{\partial^2 y}{\partial w_r \partial w_s}\int\int (y(\boldsymbol{x},\boldsymbol{w})-t) p(\boldsymbol{x},t)d\boldsymbol{x}dt + \frac{\partial y}{\partial w_s}\frac{\partial y}{\partial w_r}\int\int p(\boldsymbol{x},t)d\boldsymbol{x}dt \\ &= \frac{\partial^2 y}{\partial w_r \partial w_s}\int\int (E(t|\boldsymbol{x})-t) p(\boldsymbol{x},t)d\boldsymbol{x}dt + \frac{\partial y}{\partial w_s}\frac{\partial y}{\partial w_r}\int p(\boldsymbol{x})d\boldsymbol{x} \ (\because\ (1.89)) \\ &= \frac{\partial^2 y}{\partial w_r \partial w_s}\int (E(t|\boldsymbol{x})-E(t|\boldsymbol{x})) p(\boldsymbol{x})d\boldsymbol{x} + \frac{\partial y}{\partial w_s}\frac{\partial y}{\partial w_r}\int p(\boldsymbol{x})d\boldsymbol{x} \\ &= \frac{\partial y}{\partial w_s}\frac{\partial y}{\partial w_r}\int p(\boldsymbol{x})d\boldsymbol{x} = (5.194) \end{aligned} \tag{3}$$

以上の計算から出力について二階微分の項が消えており、離散化の場合についても近似的に二階微分の項を消去できると考えても良い。したがって結局ヘシアンは次のようになる。

$$\boldsymbol{H} \simeq \sum \nabla y_n (\nabla y_n)^T = \sum \boldsymbol{b}_n \boldsymbol{b}_n^T \tag{4}$$

$\boldsymbol{b}_n = \nabla y_n$ と約束している。ヘシアンがベクトルによる外積の和で表されていることから、この近似方法は外積による近似と呼ばれている。

1.2 ヘッセ行列の逆行列

ここでは先のサブセクションで求めた外積表示によるヘシアンについてその逆行列を考える。

まず、本格的なヘシアンの逆行列を求める計算に入る前にそもそも逆行列を考えるモチベーションを説明する。この逆行列は p.282 においてベイズニューラルネットワークの事後分布、予測分布を考えるにあたり、それらの共分散行列の中に姿を表す[*3]。したがいその共分散行列を求めるための準備として、ここでヘシアンの逆行列について考えておこういうことになる[*4]。

また、このヘシアンの逆行列の計算方法は効率的な方法であることが知られていて、それゆえに広く用いられる手法であると考えられる[*5]。

計算方法については読み合わせ。

[*3] 例えば（5.167）を見よ。

[*4] 何度でも言いますが教科書では頁が離れすぎです。もう少し工夫して議論できなかったのでしょうか...

[*5] 定量的な効率の良さはわかりませんが...

1.3　有限幅の差分による近似

読み合わせ。

1.4　ヘッセ行列の厳密な評価

逆伝播のテクニックを用いることでヘシアンを厳密に求めることが可能であり、ここではその手法について学ぶ。この手法は任意の微分可能な誤算関数、活性化関数について適用できる手法であることが知られている[*6]。

ここでは 2 層の重みを持つネットワークの場合について考える。入力、ユニット、そして出力に関する添字のルールは p.255 の中部に従うものとする。この時 Chain Rules を用いることで（5.93）～（5.95）を導出し、ネットワークについてのヘシアンを求めていく。求める手順は教科書に従うものとする。

1. 両方の重みが第二層にある：

$$
\begin{aligned}
\frac{\partial^2 E_n}{\partial w_{kj}^{(2)} w_{k'j'}^{(2)}} &= \frac{\partial}{\partial w_{kj}^{(2)}} \frac{\partial}{\partial w_{k'j'}^{(2)}} E_n \\
&= \frac{\partial}{\partial w_{kj}^{(2)}} \frac{\partial a_{k'}}{\partial w_{k'j'}^{(2)}} \frac{\partial E_n}{\partial a_{k'}} \\
&= \frac{\partial}{\partial w_{kj}^{(2)}} z_{j'} \frac{\partial E_n}{\partial a_{k'}} \\
&= z_{j'} \frac{\partial}{\partial a_{k'}} \frac{\partial a_k}{\partial w_{kj}^{(2)}} \frac{\partial E_n}{\partial a_k} \\
&= z_{j'} \frac{\partial}{\partial a_{k'}} z_j \frac{\partial E_n}{\partial a_k} \\
&= z_j z_{j'} \frac{\partial^2 E_n}{\partial a_k \partial a_{k'}} \\
&= z_j z_{j'} M_{kk'} \quad (\because (5.92)) \qquad (5)
\end{aligned}
$$

2. 両方の重みが第一層にある：（> Go to 1 番目に送った写真）[*7]
3. 重みは 1 つの層に 1 つずつある：（> Go to 2 番目に送った写真）[*8]

1.5　ヘッセ行列の積の高速な計算

ここではヘシアンそのものを（計算負荷が高いまま）計算するのではなく、ちょっとした工夫をすることで計算負荷を落とした状態で、しかもヘシアンを計算したことにできる手法を学ぶ。

実際のところヘシアン $\boldsymbol{H}$ は、あるベクトル $\boldsymbol{v}$ との積 $\boldsymbol{v}^T\boldsymbol{H}$ という形でアウトプットされる[*9]。$\boldsymbol{v}^T\boldsymbol{H}$ は要素が W 個であり、元のヘシアンの要素は W^2 個であることから、計算負荷のオーダーを 1 つ落とした状態で

[*6] 適用できることの嬉しさはよくわかっていませんが、まあ「多くの場合においてあまり心配することなく」議論を進めて良いとか、そのくらいの大雑把なイメージを持っています。

[*7] すいません。力尽きたので手書きで... こっちはできてます。

[*8] すいません。力尽きたので手書きで... 計算合わず。写真は送るだけ送ります。

[*9] 例えば p.283（5.173）を見よ。ただし式内の行列 A についてはその逆行列を考えていることに注意されたい。したがい一目で積 $\boldsymbol{v}^T\boldsymbol{H}$ という形と対応しているわけではありません。

計算が可能となる。

以下では 2 層ネットワークの例について、誤差関数の grad を評価するための順伝播と逆伝播の方程式から $\boldsymbol{v}^T\boldsymbol{H}$ を評価するための方程式を構成していく。実は結果的に新しく構成される方程式も標準的な（誤算関数の grad に関する）順伝播と逆伝播の方程式と相似な形式になることが確かめられる[*10]。

計算は読み合わせ。計算結果を見れば確かに（5.98-100）と（5.101-103）が対応し、そして（5.108-109）と（5.110-111）が対応していると言える。

以上でヘシアンの話はひとまず終わり。ここで注意のためにあえて書いておくが、セクション 5.4 と 5.5 は話の繋がりはないので頭を切り替えて欲しい。

2　ニューラルネットワークの正則化

ここでは過学習を防ぐためのモデルパラメタの選び方について学ぶ[*11]。ここは読み合わせ。

2.1　無矛盾なガウス事前分布

正則化項 $\frac{\lambda}{2}\boldsymbol{w}^T\boldsymbol{w}$ を入れ込む荷重減衰の課題としてネットワーク写像のスケリーングの性質が満たされなくなるということがある[*12]。ここでいうスケリーングの性質は「入力 x を線形変換した際に、重み w もそれに応じて線形変換することによって、出力 y を不変にできること」を指しており、ひとまずこの性質について簡単な計算を辿ることで確認を行う。

ここでの本題は正則化項を入れ込む場合についてもネットワークの写像が線形変換のもとで不変であることを要請することである。ここでは実際、正則化項の形を適切にとることでスケーリングについて不変な性質を満たすことができることを示す。

次に、上で新しく作った正則化項が変則事前分布と呼ばれる分布に対応していることを確かめる。最後に変則事前分布についてその超パラメタ（重み/バイアス）がネットワーク訓練にどのような影響を与えるのか解釈を行う。

[*10] 教科書では「方程式をそっくり再現している」と書かれているが、これは言い過ぎだと思う。だって対応する方程式同士を見比べれば分かるように、そもそも項の数が合っていないのだから。だから私は「相似」という日本語を当てることにしている。

[*11] あまりにも前サブセクションとのストーリー的な断絶があると思う。前後の話がつながらないことを前提として、あくまでこのセクション内で閉じた形で議論を進めていくことにする。

[*12] 「無矛盾でない」などという日本語で訳を当てることにセンスのなさを感じる。どうせ inconsistent という単語を直訳したのだろうが...