5.3.1 の計算について、一応まとめておく。

まず、(5.75) は

$$\tilde{x}_n = \tilde{m}^{(1)} \odot x_n = diagm(\tilde{m}^{(1)})x_n \tag{1}$$

最後の等号は本にはないが、要素を見比べることでわかる。実際は $\tilde{x}_n$ が 1 層目の活性になる。

$$\tilde{a}_n^{(1)} = W^{(1)}\tilde{x}_n = W^{(1)}diagm(\tilde{m}^{(1)})x_n \tag{2}$$

ここから、1 層目の出力がわかり、(5.76) になる。

$$z_n = \phi.(\tilde{a}_n^{(1)}) = \phi.(W^{(1)}\tilde{x}_n) = \phi.(W^{(1)}diagm(\tilde{m}^{(1)})x_n) \tag{3}$$

2 層目にもドロップアウトを適用して、

$$\tilde{z}_n = \tilde{m}^{(2)} \odot z_n = diagm(\tilde{m}^{(2)})z_n = diagm(\tilde{m}^{(2)})\phi.(W^{(1)}diagm(\tilde{m}^{(1)})x_n) \tag{4}$$

(5.78) は出力層の活性化関数がそのままの $z = a$ になるので、

$$\tilde{a}_n = W^{(2)}\tilde{z}_n \tag{5}$$

よって、

$$\tilde{a}_n = W^{(2)}diagm(\tilde{m}^{(2)})\phi.(W^{(1)}diagm(\tilde{m}^{(1)})x_n) = \tilde{W}^{(2)}\phi.(\tilde{W}^{(1)}x_n) \tag{6}$$

(5.82) を考慮する。E_n は 2 乗誤差として、

$$J_{DO}^{(S)}(W) = \frac{1}{M}\sum_{n\in S} E_n(W) + \sum_{l=1}^{2}\lambda_l||W^{(l)}||^2 = \frac{1}{M}\sum_{n\in S}\frac{1}{2}(y - f(x_n;W))^T(y - f(x_n;W)) + \sum_{l=1}^{2}\lambda_l||W^{(l)}||^2 \tag{7}$$

となる。なお、ミニバッチなので、一般の 2 乗誤差を M で割っているものと考えられる。λ_l は適当な定数なので M で割っても大勢に影響がない。

そもそも、事後分布を考えると、

$$p(W|Y,X) = \frac{p(Y,W|X)}{p(Y)} \propto p(Y,W|X) = p(Y|X,W)p(W) = \prod_n p(y_n|x_n,W)\prod_i\prod_j\prod_l p(w_{i,j}^{(l)}) \tag{8}$$

これを最大化したい。そのまま最大化しても良いが、セオリーとして、対数を取って、最大化する。($p(w)$ は各レイヤの各重みとする。それに対して掛け合わせる。)

まず、(5.83) のように確率を決める。

$$\begin{aligned} p(y_n|x_n,W) = \mathcal{N}(y_n|f(x_n;W),\gamma_y^{-1}I) &= \frac{1}{\sqrt{2\pi^{H_0}|\gamma_y^{-1}I|}}exp(-\frac{1}{2}(y - f(x_n;W))^T\gamma_y I(y - f(x_n;W))) \\ &= \frac{1}{\sqrt{2\pi^{H_0}\gamma_y^{-H_0}}}exp(-\frac{\gamma_y}{2}(y - f(x_n;W))^T(y - f(x_n;W))) \end{aligned} \tag{9}$$

同様に、$p(w)$ を考慮する。

$$p(w_{i,j}^{(l)}) = \mathcal{N}(w_{i,j}^{(l)}|0,{\lambda_{tmp}^{(l)}}^{-1}) = \frac{1}{\sqrt{2\pi{\lambda_{tmp}^{(l)}}^{-1}}}exp(-\frac{1}{2}\lambda_{tmp}^{(l)}{w_{i,j}^{(l)}}^2) \tag{10}$$

上記に記載したように事後分布の負の対数を考え、それを最小化する。

$$
\begin{aligned}
ln\, p(W|Y,X) &= -\sum_n ln\, p(y_n|x_n,W) - \sum_i \sum_j \sum_l ln\, p(w_{i,j}^{(l)}) + c \\
= -(\sum_n \frac{\gamma_y}{2}(y - f(x_n;W))^T(y - f(x_n;W)) &+ \sum_i \sum_j \sum_l \frac{1}{2}\lambda_{tmp}^{(l)} {w_{i,j}^{(l)}}^2 + c \\
= -(\sum_n \frac{\gamma_y}{2}(y - f(x_n;W))^T(y - f(x_n;W)) &+ \sum_l \frac{1}{2}\lambda_{tmp}^{(l)} ||w_{i,j}^{(l)}||^2 + c
\end{aligned} \tag{11}
$$

(行列のノルムは初出？) 評価式について、定数で割っても、最小値のときのパラメータ (argmin) は変わらないので、この式を $\gamma_y M$ で割り、定数部分を無視して J とすると以下のようになる。

$$
J = -(\frac{1}{M}\sum_n \frac{1}{2}(y - f(x_n;W))^T(y - f(x_n;W)) + \sum_l \frac{1}{2\gamma_y M}\lambda_{tmp}^{(l)} ||w_{i,j}^{(l)}||^2 \tag{12}
$$

この式と (7) を比較すると、同じ形になっていることがわかる。λ たちは任意の定数なので、$\lambda_l = \frac{1}{2\gamma_y M}\lambda_{tmp}^{(l)}$ で評価式が等しくなる。

また、上記の式展開より、

$$
E_n(W) = \frac{1}{\gamma_y} ln\, p(y_n|x_n,W) \tag{13}
$$

とすることもでき、これを (5.82) 式に代入することで、(5.85) 式を得る。