演習問題 6.27 解答

高橋優輝

$$
\begin{aligned}
\ln p(\boldsymbol{t}_N \mid \boldsymbol{\theta}) &\approx \ln p(\boldsymbol{t}_N \mid \boldsymbol{a}_N^\star) + \ln p(\boldsymbol{a}_N^\star \mid \boldsymbol{\theta}) - \frac{1}{2}\ln|\boldsymbol{H}| + \frac{N}{2}\ln(2\pi) \\
&= \ln p(\boldsymbol{t}_N \mid \boldsymbol{a}_N^\star) + \ln p(\boldsymbol{a}_N^\star \mid \boldsymbol{\theta}) - \frac{1}{2}\ln|\boldsymbol{W}_N(\boldsymbol{a}_N^\star) + \boldsymbol{C}_N^{-1}| + \frac{N}{2}\ln(2\pi)
\end{aligned} \tag{6.90}
$$

について，式変形を進める．

今までの議論の中で，(6.80) の成立が分かっている．

$$
\ln p(\boldsymbol{a}_N) + \ln p(\boldsymbol{t}_N \mid \boldsymbol{a}_N) = -\frac{1}{2}\boldsymbol{a}_N^\top \boldsymbol{C}_N^{-1}\boldsymbol{a}_N - \frac{N}{2}\ln(2\pi) - \frac{1}{2}\ln|\boldsymbol{C}_N| + \boldsymbol{t}_N^\top \boldsymbol{a}_N - \sum_{n=1}^{N}\ln(1+e^{a_n}) \tag{6.80}
$$

(6.90) に (6.80) を代入すると，

$$
\begin{aligned}
\ln p(\boldsymbol{t}_N \mid \boldsymbol{\theta}) &\approx -\frac{1}{2}(\boldsymbol{a}_N^\star)^\top \boldsymbol{C}_N^{-1}\boldsymbol{a}_N^\star - \frac{N}{2}\ln(2\pi) - \frac{1}{2}\ln|\boldsymbol{C}_N| + \boldsymbol{t}_N^\top \boldsymbol{a}_N^\star - \sum_{n=1}^{N}\ln(1+e^{a_n^\star}) \\
&\quad - \frac{1}{2}\ln|\boldsymbol{W}_N + \boldsymbol{C}_N^{-1}| + \frac{N}{2}\ln(2\pi) \\
&= -\frac{1}{2}(\boldsymbol{a}_N^\star)^\top \boldsymbol{C}_N^{-1}\boldsymbol{a}_N^\star - \frac{1}{2}\ln|\boldsymbol{C}_N| + \boldsymbol{t}_N^\top \boldsymbol{a}_N^\star - \sum_{n=1}^{N}\ln(1+e^{a_n^\star}) - \frac{1}{2}\ln|\boldsymbol{W}_N + \boldsymbol{C}_N^{-1}|
\end{aligned} \tag{$\ast$}
$$

対数尤度関数の最大化のために，$\boldsymbol{\theta}$ による微分を求めたい．f を $\boldsymbol{\theta}$ と $\boldsymbol{\theta}$ に依存する $\boldsymbol{a}_N^\star$ の関数とすると，

$$
\frac{df}{d\theta_j} = \frac{\partial f}{\partial \theta_j} + \frac{\partial f}{\partial \boldsymbol{a}_N^\star} \cdot \frac{d\boldsymbol{a}_N^\star}{d\theta_j}
$$

が成り立つ．

$\boxed{\frac{\partial f}{\partial \theta_j} \text{ の計算}}$

$\boldsymbol{a}_N^\star$ は $\boldsymbol{\theta}$ に依存することに注意して，$\boldsymbol{a}_N^\star$ を固定して，θ_j で偏微分する．ここで，次の (C.21), (C.22) を用いる．

$$\frac{\partial}{\partial x}(\boldsymbol{A}^{-1}) = -\boldsymbol{A}^{-1}\frac{\partial \boldsymbol{A}}{\partial x}\boldsymbol{A}^{-1} \tag{C.21}$$

$$\frac{\partial}{\partial x}\ln|\boldsymbol{A}| = \mathrm{Tr}\left(\boldsymbol{A}^{-1}\frac{\partial \boldsymbol{A}}{\partial x}\right) \tag{C.22}$$

<u>第 1 項</u>

$$\frac{\partial}{\partial \theta_j}\left((\boldsymbol{a}_N^\star)^\top \boldsymbol{C}_N^{-1}\boldsymbol{a}_N^\star\right) = -(\boldsymbol{a}_N^\star)^\top \boldsymbol{C}_N^{-1}\frac{\partial \boldsymbol{C}_N}{\partial \theta_i}\boldsymbol{C}_N^{-1}\boldsymbol{a}_N^\star$$

<u>第 2 項</u>

$$\frac{\partial}{\partial \theta_j}\ln|\boldsymbol{C}_N| = \mathrm{Tr}\left(\boldsymbol{C}_N^{-1}\frac{\partial \boldsymbol{C}_N}{\partial \theta_j}\right)$$

<u>第 5 項</u>

$$\begin{aligned}
\frac{\partial}{\partial \theta_j}\ln|\boldsymbol{W}_N + \boldsymbol{C}_N^{-1}| &= \mathrm{Tr}\left((\boldsymbol{W}_N + \boldsymbol{C}_N^{-1})^{-1}\frac{\partial}{\partial \theta_j}(\boldsymbol{W}_N + \boldsymbol{C}_N^{-1})\right) \\
&= -\mathrm{Tr}\left((\boldsymbol{W}_N + \boldsymbol{C}_N^{-1})^{-1}\boldsymbol{C}_N^{-1}\frac{\partial \boldsymbol{C}_{\boldsymbol{N}}}{\partial \theta_j}\boldsymbol{C}_N^{-1}\right) \\
&= -\mathrm{Tr}\left(\boldsymbol{C}_N^{-1}(\boldsymbol{W}_N + \boldsymbol{C}_N^{-1})^{-1}\boldsymbol{C}_N^{-1}\frac{\partial \boldsymbol{C}_{\boldsymbol{N}}}{\partial \theta_j}\right) \quad (\because \text{トレース作用素の循環性})
\end{aligned}$$

よって,

$$
\begin{aligned}
\frac{\partial}{\partial\theta_i}\ln p(\boldsymbol{t}_N \mid \boldsymbol{\theta}) &= \frac{1}{2}\left((\boldsymbol{a}_N^\star)^\top \boldsymbol{C}_N^{-1}\frac{\partial \boldsymbol{C}_N}{\partial\theta_i}\boldsymbol{C}_N^{-1}\boldsymbol{a}_N^\star\right) \\
&\quad -\frac{1}{2}\left(\mathrm{Tr}\left(\boldsymbol{C}_N^{-1}\frac{\partial \boldsymbol{C}_N}{\partial\theta_j}\right) - \mathrm{Tr}\left(\boldsymbol{C}_N^{-1}(\boldsymbol{W}_N+\boldsymbol{C}_N^{-1})^{-1}\boldsymbol{C}_N^{-1}\frac{\partial \boldsymbol{C}_{\boldsymbol{N}}}{\partial\theta_j}\right)\right) \\
&= \frac{1}{2}\left((\boldsymbol{a}_N^\star)^\top \boldsymbol{C}_N^{-1}\frac{\partial \boldsymbol{C}_N}{\partial\theta_i}\boldsymbol{C}_N^{-1}\boldsymbol{a}_N^\star\right) \\
&\quad -\frac{1}{2}\mathrm{Tr}\left(\boldsymbol{C}_N^{-1}(\boldsymbol{I}-(\boldsymbol{W}_N+\boldsymbol{C}_N^{-1})^{-1}\boldsymbol{C}_N^{-1})\frac{\partial \boldsymbol{C}_N}{\partial\theta_j}\right) \\
&= \frac{1}{2}\left((\boldsymbol{a}_N^\star)^\top \boldsymbol{C}_N^{-1}\frac{\partial \boldsymbol{C}_N}{\partial\theta_i}\boldsymbol{C}_N^{-1}\boldsymbol{a}_N^\star\right) \\
&\quad -\frac{1}{2}\mathrm{Tr}\left(\boldsymbol{C}_N^{-1}(\boldsymbol{W}_N+\boldsymbol{C}_N^{-1})^{-1}((\boldsymbol{W}_N+\boldsymbol{C}_N^{-1})-\boldsymbol{C}_N^{-1})\frac{\partial \boldsymbol{C}_N}{\partial\theta_j}\right) \\
&= \frac{1}{2}\left((\boldsymbol{a}_N^\star)^\top \boldsymbol{C}_N^{-1}\frac{\partial \boldsymbol{C}_N}{\partial\theta_i}\boldsymbol{C}_N^{-1}\boldsymbol{a}_N^\star\right) \\
&\quad -\frac{1}{2}\mathrm{Tr}\left(\boldsymbol{C}_N^{-1}(\boldsymbol{W}_N+\boldsymbol{C}_N^{-1})^{-1}\boldsymbol{W}_N\frac{\partial \boldsymbol{C}_N}{\partial\theta_j}\right) \\
&= \frac{1}{2}\left((\boldsymbol{a}_N^\star)^\top \boldsymbol{C}_N^{-1}\frac{\partial \boldsymbol{C}_N}{\partial\theta_i}\boldsymbol{C}_N^{-1}\boldsymbol{a}_N^\star\right) \\
&\quad -\frac{1}{2}\mathrm{Tr}\left(((\boldsymbol{W}_N+\boldsymbol{C}_N^{-1})\boldsymbol{C}_N)^{-1}\boldsymbol{W}_N\frac{\partial \boldsymbol{C}_N}{\partial\theta_j}\right) \quad (\because (\boldsymbol{AB})^{-1}=\boldsymbol{B}^{-1}\boldsymbol{A}^{-1}) \\
&= \frac{1}{2}\left((\boldsymbol{a}_N^\star)^\top \boldsymbol{C}_N^{-1}\frac{\partial \boldsymbol{C}_N}{\partial\theta_i}\boldsymbol{C}_N^{-1}\boldsymbol{a}_N^\star\right) \\
&\quad -\frac{1}{2}\mathrm{Tr}\left((\boldsymbol{W}_N\boldsymbol{C}_N+\boldsymbol{I})^{-1}\boldsymbol{W}_N\frac{\partial \boldsymbol{C}_N}{\partial\theta_j}\right) \\
&= \frac{1}{2}\left((\boldsymbol{a}_N^\star)^\top \boldsymbol{C}_N^{-1}\frac{\partial \boldsymbol{C}_N}{\partial\theta_i}\boldsymbol{C}_N^{-1}\boldsymbol{a}_N^\star\right) \\
&\quad -\frac{1}{2}\mathrm{Tr}\left((\boldsymbol{C}_N\boldsymbol{W}_N+\boldsymbol{I})^{-1}\boldsymbol{W}_N\frac{\partial \boldsymbol{C}_N}{\partial\theta_j}\right) \quad (\because \boldsymbol{W}_N:\text{対角行列}) \quad (6.91)
\end{aligned}
$$

$\boxed{\frac{\partial f}{\partial \boldsymbol{a}_N^\star} \cdot \frac{d\boldsymbol{a}_N^\star}{d\theta_j} \text{ の計算}}$

今，

$$\Psi(\boldsymbol{a}_N) \coloneqq \ln p(\boldsymbol{a}_N) + \ln p(\boldsymbol{t}_N \mid \boldsymbol{a}_N) \tag{6.80}$$

はラプラス近似によって，

$$\frac{\partial}{\partial \boldsymbol{a}_N^\star} \Psi(\boldsymbol{a}_N^\star) = \boldsymbol{0}$$

となるように近似されているので，

$$\left(\frac{\partial}{\partial \boldsymbol{a}_N^\star} \ln p(\boldsymbol{t}_N \mid \boldsymbol{\theta})\right) \cdot \frac{d\boldsymbol{a}_N^\star}{d\theta_j} = -\frac{1}{2}\sum_{n=1}^{N} \frac{\partial \ln |\boldsymbol{W}_N + \boldsymbol{C}_N^{-1}|}{\partial a_n^\star} \frac{\partial a_n^\star}{\partial \theta_j} \tag{$\ast\ast$}$$

となる．ここで，

$$\begin{aligned}\frac{\partial \ln |\boldsymbol{W}_N + \boldsymbol{C}_N^{-1}|}{\partial a_n^\star} &= \mathrm{Tr}\left((\boldsymbol{W}_N + \boldsymbol{C}_N^{-1})^{-1} \frac{\partial}{\partial a_n^\star}(\boldsymbol{W}_N + \boldsymbol{C}_N^{-1})\right) \\ &= \mathrm{Tr}\left((\boldsymbol{W}_N + \boldsymbol{C}_N^{-1})^{-1} \frac{\partial \boldsymbol{W}_N}{\partial a_n^\star}\right) \quad (\because \boldsymbol{C}_N \colon \boldsymbol{a}_N \text{に非依存}) \end{aligned} \tag{$\ast\ast\ast$}$$

が成り立つ．さらに，

$$(\boldsymbol{W}_N + \boldsymbol{C}_N^{-1})^{-1} = (\boldsymbol{I} + \boldsymbol{C}_N \boldsymbol{W}_N)^{-1} \boldsymbol{C}_N$$

であり，

$$\begin{aligned}\frac{\partial \boldsymbol{W}_N}{\partial a_n^\star} &= \frac{\partial}{\partial a_n^\star} \mathrm{diag}\left(\begin{bmatrix}\sigma(a_1^\star)(1-\sigma(a_1^\star)) & \cdots & \sigma(a_n^\star)(1-\sigma(a_n^\star)) & \cdots & \sigma(a_N^\star)(1-\sigma(a_N^\star))\end{bmatrix}\right) \\ &= \mathrm{diag}\left(\begin{bmatrix}0 & \cdots & \sigma(a_n^\star)(1-\sigma(a_n^\star))(1-2\sigma(a_n^\star)) & \cdots & 0\end{bmatrix}\right) \quad \left(\because \frac{d\sigma}{da} = \sigma(1-\sigma) \quad (4.88)\right)\end{aligned}$$

であるので，$(\ast\ast\ast)$ に代入して，

$$\frac{\partial \ln |\boldsymbol{W}_N + \boldsymbol{C}_N^{-1}|}{\partial a_n^\star} = \left((\boldsymbol{I} + \boldsymbol{C}_N \boldsymbol{W}_N)^{-1} \boldsymbol{C}_N\right)_{nn} \sigma(a_n^\star)(1-\sigma(a_n^\star))(1-2\sigma(a_n^\star))$$

が成り立つ．これを $(\ast\ast)$ に代入すると，

$$\left(\frac{\partial}{\partial \boldsymbol{a}_N^\star} \ln p(\boldsymbol{t}_N \mid \boldsymbol{\theta})\right) \cdot \frac{d\boldsymbol{a}_N^\star}{d\theta_j} = -\frac{1}{2}\sum_{n=1}^{N} \left((\boldsymbol{I} + \boldsymbol{C}_N \boldsymbol{W}_N)^{-1} \boldsymbol{C}_N\right)_{nn} \sigma(a_n^\star)(1-\sigma(a_n^\star))(1-2\sigma(a_n^\star)) \frac{\partial a_n^\star}{\partial \theta_j} \tag{6.92}$$

が得られる．

また，

$$
\begin{aligned}
\frac{\partial \boldsymbol{a}_N^\star}{\partial \theta_j} &= \frac{\partial}{\partial \theta_j}(\boldsymbol{C}_N(\boldsymbol{t}_N - \boldsymbol{\sigma}_N)) \\
&= \frac{\partial \boldsymbol{C}_N}{\partial \theta_j}(\boldsymbol{t}_N - \boldsymbol{\sigma}_N) + \boldsymbol{C}_N \frac{\partial}{\partial \theta_j}(\boldsymbol{t}_N - \boldsymbol{\sigma}_N) \\
&= \frac{\partial \boldsymbol{C}_N}{\partial \theta_j}(\boldsymbol{t}_N - \boldsymbol{\sigma}_N) - \boldsymbol{C}_N \frac{\partial \boldsymbol{\sigma}_N}{\partial \theta_j} \quad (\because \boldsymbol{t}_N \text{ は } \theta_j \text{ に非依存})
\end{aligned}
$$

ここで，

$$
\begin{aligned}
\frac{\partial \boldsymbol{\sigma}_N}{\partial \theta_j} &= \begin{bmatrix} \frac{\partial \sigma(a_1^\star)}{\partial \theta_j} \\ \vdots \\ \frac{\partial \sigma(a_N^\star)}{\partial \theta_j} \end{bmatrix} \\
&= \begin{bmatrix} \frac{\partial \sigma(a_1^\star)}{\partial a_1^\star} \frac{\partial a_1^\star}{\partial \theta_j} \\ \vdots \\ \frac{\partial \sigma(a_N^\star)}{\partial a_N^\star} \frac{\partial a_N^\star}{\partial \theta_j} \end{bmatrix} \\
&= \begin{bmatrix} \sigma(a_1^\star)(1-\sigma(a_1)^\star)\frac{\partial a_1^\star}{\partial \theta_j} \\ \vdots \\ \sigma(a_N^\star)(1-\sigma(a_N)^\star)\frac{\partial a_N^\star}{\partial \theta_j} \end{bmatrix} \\
&= \mathrm{diag}(\begin{bmatrix} \sigma(a_1^\star)(1-\sigma(a_1)^\star) & \cdots & \sigma(a_N^\star)(1-\sigma(a_N)^\star) \end{bmatrix}) \frac{\partial \boldsymbol{a}_N^\star}{\partial \theta_j} \\
&= \boldsymbol{W}_N \frac{\partial \boldsymbol{a}_N^\star}{\partial \theta_j}
\end{aligned}
$$

であるので，

$$
\frac{\partial \boldsymbol{a}_N^\star}{\partial \theta_j} = \frac{\partial \boldsymbol{C}_N}{\partial \theta_j}(\boldsymbol{t}_N - \boldsymbol{\sigma}_N) - \boldsymbol{C}_N \boldsymbol{W}_N \frac{\partial \boldsymbol{a}_N^\star}{\partial \theta_j} \tag{6.93}
$$

が成り立つ．これを並び替えると，

$$
\frac{\partial \boldsymbol{a}_N^\star}{\partial \theta_j} = (\boldsymbol{I} + \boldsymbol{W}_N \boldsymbol{C}_N)^{-1} \frac{\partial \boldsymbol{C}_N}{\partial \theta_j}(\boldsymbol{t}_N - \boldsymbol{\sigma}_N) \tag{6.94}
$$

が得られる．以上，(6.91), (6.92), (6.94) を組み合わせることで，対数尤度関数の勾配を求めることができる．