3A

Machine Learning II
ID: 5684926 Tristan Scheidemann

3A-1.

We compute $\operatorname{argmax}_\lambda \prod_{n=1}^{N} p(k_n;\lambda)$ by finding the roots of the first derivative.

$$\frac{d}{d\lambda}\prod_{n=1}^{N} p(k_n;\lambda) = \frac{d}{d\lambda}\prod_{n=1}^{N}\frac{\lambda^{k_n}}{k_n!}e^{-\lambda} = 0.$$

Ergo:

$$\begin{aligned}
\frac{d}{d\lambda}\prod_{n=1}^{N} p(k_n;\lambda) \quad &= \frac{d}{d\lambda}\prod_{n=1}^{N}\frac{\lambda^{k_n}}{k_n!}e^{-\lambda} \\
&\propto \frac{d}{d\lambda}\prod_{n=1}^{N}\lambda^{k_n}e^{-\lambda} \\
&= \left(\sum_{i=1}^{N} k_i\right)\lambda^{\sum_{i=1}^{N} k_i - 1}e^{-N\lambda} + \left[-N\lambda^{\sum_{i=1}^{N} k_i}e^{-N\lambda}\right] \\
&= e^{-N\lambda}\lambda^{\sum_{i=1}^{N} k_i}\left[\left(\sum_{i=1}^{N} k_i\right) - N\lambda\right] \\
&\Rightarrow \lambda_1 = 0, \lambda_2 = \frac{\left(\sum_{i=1}^{N} k_i\right)}{N}.
\end{aligned}$$

3A-2.

Auxiliary calculation

Reduced formula for completing the square in D dimensions:

$2\mathbf{b}^{\mathbf{T}}\mathbf{x} + \mathbf{x}^{T}C\mathbf{x} = (\mathbf{x}-\mathbf{m})^{T}M(\mathbf{x}-\mathbf{m}),$

with

$M = C,$

$\mathbf{m} = -\left(\frac{1}{2}C + \frac{1}{2}C^{T}\right)^{-1}\mathbf{b}\ (general\ C),$

$\mathbf{m} = -\mathrm{C}^{-1}\boldsymbol{b}$ (C *symmetric*).

Let $c \in \mathbb{R}^+$ be an arbitrary constant.

$$\begin{aligned}
p(\mathbf{w}|\mathbf{t}, \mathbf{X}, \sigma_\epsilon^2) &= \frac{p(\mathbf{t}|\mathbf{w}, \mathbf{X}, \sigma_e^2) \cdot p(\mathbf{w}|\sigma_0^2)}{p(\mathbf{t})} \\
&\propto p(\mathbf{t}|\mathbf{w}, \mathbf{X}, \sigma_e^2) \cdot p(\mathbf{w}|\sigma_0^2) \\
&= \prod_{n=1}^{N} N\left(t_n|\mathbf{w}^\mathrm{T}\mathbf{x}_n, \sigma_e^2\right) \cdot N(\mathbf{w}|\mathbf{0}, \sigma_0^2\mathbf{I}) \\
&\propto e^{-\frac{1}{2\sigma_e^2}\sum_{n=1}^{N}(t_n - \mathbf{w}^\mathrm{T}\mathbf{x}_n)^2} \cdot e^{-\frac{1}{2\sigma_0^2}\|\mathbf{w}\|} \\
&= e^{-\frac{1}{2}\left[\frac{1}{\sigma_\epsilon^2}\sum_{n=1}^{N}\left(\underbrace{t_n^2}_{constant} - 2t_n\mathbf{w}^\mathrm{T}\mathbf{x}_n + (\mathbf{w}^\mathrm{T}\mathbf{x}_n)^2\right) + \frac{\|\mathbf{w}\|}{\sigma_0^2}\right]} \\
&\propto e^{-\frac{1}{2}\left[\underbrace{\frac{1}{\sigma_\epsilon^2}\sum_{n=1}^{N}\left(-2t_n\mathbf{w}^\mathrm{T}\mathbf{x}_n + (\mathbf{w}^\mathrm{T}\mathbf{x}_n)^2\right) + \frac{\|\mathbf{w}\|}{\sigma_0^2}}_{completing\ the\ square}\right]}.
\end{aligned}$$

Completing of the square requires conversion of the exponent into

$(*)$ $2\mathbf{b}^\mathbf{T}\mathbf{w} + \mathbf{w}^T C\mathbf{w}$.

Conversion of each term into matrices:

$$\frac{1}{\sigma_\epsilon^2}\sum_{n=1}^{N}\left(-2t_n\mathbf{w}^\mathrm{T}\mathbf{x}_n\right) = \frac{1}{\sigma_\epsilon^2}[-2\mathbf{t}^T\mathbf{X}\mathbf{w}] = 2\left[-\frac{1}{\sigma_\epsilon^2}\mathbf{t}^T\mathbf{X}\right]\mathbf{w},$$

$$\frac{1}{\sigma_\epsilon^2}\sum_{n=1}^{N}\left(\mathbf{w}^\mathrm{T}\mathbf{x}_n\right)^2 = \frac{1}{\sigma_\epsilon^2}[\mathbf{w}^T\mathbf{X}^T(\mathbf{w}^T\mathbf{X}^T)^T] = \frac{1}{\sigma_\epsilon^2}\mathbf{w}^T\mathbf{X}^T\mathbf{X}\mathbf{w},$$

$$\frac{\|\mathbf{w}\|}{\sigma_0^2} = \frac{1}{\sigma_0^2}\mathbf{w}^T\mathbf{w} = \mathbf{w}^T\frac{1}{\sigma_0^2}\mathbf{I}\mathbf{w}.$$

Our square can only be completed, if $(*)$ contains the above matrices. This is fulfilled by the following assignment:

$$2\mathbf{b}^\mathbf{T}\mathbf{w} = 2\left[-\frac{1}{\sigma_\epsilon^2}\mathbf{t}^T\mathbf{X}\right]\mathbf{w},$$

$$\mathbf{w}^T C\mathbf{w} = \mathbf{w}^T\left(\frac{1}{\sigma_0^2}\mathbf{I} + \frac{1}{\sigma_\epsilon^2}\mathbf{X}^T\mathbf{X}\right)\mathbf{w}.$$

$$\begin{aligned}
\Rightarrow \quad \mathbf{b}^\mathbf{T} &= \left[-\frac{1}{\sigma_\epsilon^2}\mathbf{t}^T\mathbf{X}\right], \\
C &= \left(\frac{1}{\sigma_0^2}\mathbf{I} + \frac{1}{\sigma_\epsilon^2}\mathbf{X}^T\mathbf{X}\right).
\end{aligned}$$

We note that C is symmetric by extension of $\mathbf{X}^T\mathbf{X}$'s symmetry.

Ergo

$M = C = \Sigma_\mathrm{N}^{-1}$,

$$\mathbf{m} = -\mathrm{C}^{-\mathbf{1}}\boldsymbol{b} = -\Sigma\left[-\frac{1}{\sigma_\epsilon^2}\mathbf{t}^T\mathbf{X}\right]^T = \Sigma\left[\frac{1}{\sigma_\epsilon^2}\mathbf{X}^T\mathbf{t}\right] = \boldsymbol{\mu}_\mathrm{N}.$$

It follows that $p(\mathbf{w}|\mathbf{t},\mathbf{X},\sigma_\epsilon^2) \sim N\left(\Sigma\left[\frac{1}{\sigma_\epsilon^2}\mathbf{X}^T\mathbf{t}\right], \left(\frac{1}{\sigma_0^2}\mathbf{I} + \frac{1}{\sigma_\epsilon^2}\mathbf{X}^T\mathbf{X}\right)^{-1}\right)$.

3A-3.

Note that $\mathbf{t}$ is now a variable and $\mathbf{w}$ a hyperparameter and assumed constant, which forbids us from completing the square of

$$\frac{1}{\sigma_\epsilon^2}\sum_{n=1}^N\left(-2t_n\mathbf{w}^\mathrm{T}\mathbf{x}_n + \left(\mathbf{w}^\mathrm{T}\mathbf{x}_n\right)^2\right) + \frac{\|\mathbf{w}\|}{\sigma_0^2}$$

like in the exercise before.

For any fixed $\mathbf{w}$ we have:

$$\begin{aligned}
p(\mathbf{t}|\mathbf{X}) &= \frac{p(\mathbf{t}|\mathbf{w},\mathbf{X})\,p(\mathbf{w})}{p(\mathbf{w}|\mathbf{t},\mathbf{X})} \\
&= \frac{\prod_{n=1}^N\left[(2\pi\sigma_\epsilon^2)^{-\frac{1}{2}}\,e^{-\frac{1}{2\sigma_\epsilon^2}(t_n-\mathbf{w}^T\mathbf{x}_n)^2}\right]\cdot(2\pi\sigma_0^2)^{-\frac{D}{2}}e^{-\frac{1}{2\sigma_0^2}\|w\|^2}}{(2\pi)^{-\frac{D}{2}}\,(\det\Sigma_N)^{-\frac{1}{2}}\,e^{-\frac{1}{2}(\mathbf{w}-\boldsymbol{\mu}_N)^T\Sigma_N(\mathbf{w}-\boldsymbol{\mu}_N)}} \\
&= (2\pi\sigma_\epsilon^2)^{-\frac{N}{2}}\,(2\pi\sigma_0^2)^{-\frac{D}{2}}\,(2\pi)^{\frac{D}{2}}\,(\det\Sigma_N)^{\frac{1}{2}}\,e^{-\frac{1}{2\sigma_\epsilon^2}\sum_{n=1}^N(t_n-\mathbf{w}^T\mathbf{x}_n)^2-\frac{1}{2\sigma_0^2}\|w\|^2+\frac{1}{2}(\mathbf{w}-\boldsymbol{\mu}_N)^T\Sigma_N(\mathbf{w}-\boldsymbol{\mu}_N)} \\
&= (2\pi r_\epsilon^{-1})^{-\frac{N}{2}}\,(2\pi r_0^{-1})^{-\frac{D}{2}}\,(2\pi)^{\frac{D}{2}}\,(\det\Sigma_N)^{\frac{1}{2}}\,e^{\underbrace{-\frac{r_\epsilon}{2}\sum_{n=1}^N(t_n-\mathbf{w}^T\mathbf{x}_n)^2-\frac{r_0}{2}\|w\|^2+\frac{1}{2}(\mathbf{w}-\boldsymbol{\mu}_N)^T\Sigma_N(\mathbf{w}-\boldsymbol{\mu}_N).}_{k(\mathbf{t})}}
\end{aligned}$$

Taking the logarithm:

$$\begin{aligned}
\log p(\mathbf{t}|\mathbf{X}) &= \log(2\pi r_\epsilon^{-1})^{-\frac{N}{2}} + \log\ (2\pi r_0^{-1})^{-\frac{D}{2}} + \log(2\pi)^{\frac{D}{2}} + \log(\det\Sigma_N)^{\frac{1}{2}} + k(\mathbf{t}) \\
&= -\frac{N}{2}\log 2\pi + \frac{N}{2}\log r_\epsilon - \frac{D}{2}\log 2\pi + \frac{D}{2}\log r_0 + \frac{D}{2}\log 2\pi + \frac{1}{2}\log\det\Sigma_N + k(\mathbf{t}) \\
&= -\frac{N}{2}\log 2\pi + \frac{N}{2}\log r_\epsilon + \frac{D}{2}\log r_0 + \frac{1}{2}\log\det\Sigma_N + k(\mathbf{t}).
\end{aligned}$$

Since above holds for any $\mathbf{w}$, we arbitrarily set $\mathbf{w} = \boldsymbol{\mu}_\mathrm{N}$.

$$\begin{aligned}
\mathbf{w} = \boldsymbol{\mu}_\mathrm{N} \quad &\Rightarrow -\frac{N}{2}\log 2\pi + \frac{N}{2}\log r_\epsilon + \frac{D}{2}\log r_0 + \frac{1}{2}\log\det\Sigma_N + k(\mathbf{t}) \\
&\Leftrightarrow -\frac{N}{2}\log 2\pi + \frac{N}{2}\log r_\epsilon + \frac{D}{2}\log r_0 + \frac{1}{2}\log\det\Sigma_N \\
&+\left[-\frac{r_\epsilon}{2}\sum_{n=1}^N(t_n - \boldsymbol{\mu}_\mathrm{N}{}^T\mathbf{x}_n)^2 - \frac{r_0}{2}\|\boldsymbol{\mu}_\mathrm{N}\|^2 + \frac{1}{2}(\boldsymbol{\mu}_\mathrm{N} - \boldsymbol{\mu}_N)^T\Sigma_N(\boldsymbol{\mu}_\mathrm{N} - \boldsymbol{\mu}_N)\right] \\
&\Leftrightarrow -\frac{N}{2}\log 2\pi + \frac{N}{2}\log r_\epsilon + \frac{D}{2}\log r_0 + \frac{1}{2}\log\det\Sigma_N \\
&+\left[-\frac{r_\epsilon}{2}\sum_{n=1}^N(t_n - \boldsymbol{\mu}_\mathrm{N}{}^T\mathbf{x}_n)^2 - \frac{r_0}{2}\|\boldsymbol{\mu}_\mathrm{N}\|^2\right]
\end{aligned}$$

To express $k(\mathbf{t})$ completely in terms of matrices, we use the following conversion:

$$\sum_{n=1}^{N}(t_n - \boldsymbol{\mu}_{\mathrm{N}}{}^T\mathbf{x}_n)^2 = (\mathbf{t} - \boldsymbol{\mu}_N^T\mathbf{X}^T)^T(\mathbf{t} - \boldsymbol{\mu}_N^T\mathbf{X}^T) = (\mathbf{X}^T\boldsymbol{\mu}_N - \mathbf{t})^T(\mathbf{X}^T\boldsymbol{\mu}_N - \mathbf{t}).$$

Ergo

$$\log p(\mathbf{t}|\mathbf{X}) \quad = -\frac{N}{2}\log 2\pi + \frac{N}{2}\log r_\epsilon + \frac{D}{2}\log r_0 + \frac{1}{2}\log\det\Sigma_N + \left[-\frac{r_\epsilon}{2}(\mathbf{X}^T\boldsymbol{\mu}_N - \mathbf{t})^T(\mathbf{X}^T\boldsymbol{\mu}_N - \mathbf{t}) - \frac{r_0}{2}\|\boldsymbol{\mu}_{\mathrm{N}}\|^2\right].$$