Exercise 4

Machine Learning I

4A-1.

Prerequisites

Leibnitz rule for constants $-\infty < a, b < +\infty$:

$$\frac{d}{dx}\left(\int_a^b f(x,t)dt\right) = \int_a^b \frac{\partial}{\partial x} f(x,t)dt$$

Additionally, let at least one of the following conditions hold:

(i) $f(x,t)$ is measurable and nonnegative

(ii) $\int_a^b |f(x,t)|dt$ is finite

then we can switch the order of integration according to Tonelli/Fubini, respectively.

Note: In our case, at least one of (i), (ii) is nearly always satisfied. Think about why.

Lastly, we need $(*)$ *Theorem 1* concerning uniform convergence from these notes:

http://www.math.ucla.edu/~tao/resource/general/131bh.1.03s/week45.pdf

Let D be the domain of $\mathbf{x}$.

$$\begin{aligned} E[y(\mathbf{x}) - t] &= \int_{t_1}^{t_2}\int_D (y(\mathbf{x}) - t)^2 p(\mathbf{x}, t) d\mathbf{x}\, dt \\ &= \underbrace{\int_D \int_{t_1}^{t_2} (y(\mathbf{x}) - t)^2 p(\mathbf{x}, t) dt\, d\mathbf{x}}_{Fubini/Tonelli} \\ &= \int_D p(\mathbf{x}) \int_{t_1}^{t_2} (y(\mathbf{x}) - t)^2 p(t|\mathbf{x}) dt\, d\mathbf{x} \end{aligned}$$

Minimizing the loss cumulative loss for all **t** equals minimizing the loss for each t_i separately$(**)$. Note: Inside the interior integral, $\mathbf{x}$ is constant. Let $y(\mathbf{x}) = z$:

$$\begin{aligned} \frac{\partial}{\partial z}\int_{t_1}^{t_2} (z - t)^2 p(t|\mathbf{x}) dt &= \underbrace{\int_{t_1}^{t_2} \frac{\partial}{\partial z}(z - t)^2 p(t|\mathbf{x}) dt}_{Leibnitz\ rule} \\ &= 2\int_{t_1}^{t_2} (z - t) p(t|\mathbf{x}) dt \end{aligned}$$

We can now solve for z:

$$
\begin{aligned}
& 2\int_{t_1}^{t_2}(z-t)p(t|\mathbf{x})dt && = 0 \\
\Leftrightarrow\quad & \underbrace{z\int_{t_1}^{t_2}p(t|\mathbf{x})dt}_{=1} && = \int_{t_1}^{t_2}tp(t|\mathbf{x})dt \\
\Leftrightarrow\quad & y(\mathbf{x}) && = E[t|\mathbf{x}]
\end{aligned}
$$

According to the Leibnitz rule, this only holds for finite limits t_1, t_2. To extend this proof to the infinite domain, we construct the sequence:

$$f_n' = \frac{\partial}{\partial z}\int_{-n}^{n}(z-t)^2 p(t|\mathbf{x})dt$$

Because probabilities sum to one, if n tends to infinity, $p(t|\mathbf{x})$ vanishes for most t_i. The $(z-t)^2$ will not compensate that, as we required $E[y(\mathbf{x})-t]$ to be finite earlier.

Accordingly, $\lim\limits_{n\to\infty} f_n{}'$ converges to g uniformly. Due to $(*)$, this means the functions f_n converge uniformly to f, with $f' = g$.

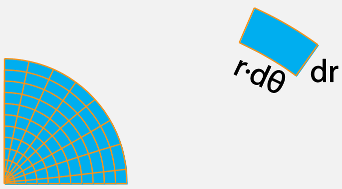
Spoken plainly, this means if we have an infinite domain $\mathbb{R}$, we can approximate the solution arbitrarily close by increasing $[t_1, t_2]$.

$(**)$ Only because the x_i are independent. In our situation this is the case, otherwise we would also have to integrate over all possibilities $p(x|x_i, \ldots x_0)$.

4A-2.

Auxiliary calculation

The area of a single infinitesimal d-dimensional piece of $f(r, \boldsymbol{\theta})$ is $r^{d-1} d\theta_1 \cdot \ldots \cdot d\theta_{d-1} \cdot dr$. This is trivially an d-dimensional extension of the two-dimensional case shown below:



.

Additionally, to convert a function $f(r, \boldsymbol{\theta})$ from hyperspherical coordinates into cartesian coordinates $f(\mathbf{x})$, we use the following trigonometric conversion:

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \ldots \\ x_{d-1} \\ x_d \end{pmatrix} = \begin{pmatrix} r\cos\theta_1 \\ r\sin\theta_1\cos\theta_2 \\ \ldots \\ r\sin\theta_1\sin\theta_2\sin\theta_3\ldots\sin\theta_{d-3}\sin\theta_{d-2}\cos\theta_{d-1} \\ r\sin\theta_1\sin\theta_2\sin\theta_3\ldots\sin\theta_{d-3}\sin\theta_{d-2}\sin\theta_{d-1} \end{pmatrix},$$

i.e.

$$f(\mathbf{x}) = f(r\cos\theta_1, r\sin\theta_1\cos\theta_2, \ldots, r\sin\theta_1\sin\theta_2\sin\theta_3\ldots\sin\theta_{d-3}\sin\theta_{d-2}\sin\theta_{d-1}).$$

Lastly, let

$$K(\boldsymbol{\theta}) = \cos\theta_1{}^2 + (\sin\theta_1\cos\theta_2)^2 + \cdots \\ + (\sin\theta_1\sin\theta_2\sin\theta_3\ldots\sin\theta_{d-3}\sin\theta_{d-2}\sin\theta_{d-1})^2.$$

Note: $\boldsymbol{\theta}$ describe points on the unit d-sphere, so it is no surprise that $\|K(\boldsymbol{\theta})\|^2 = 1$ for all $\boldsymbol{\theta}$, because the radius of the unit sphere is 1.

The centered sphere is described by $B_0(r) \coloneqq \{x_1^2 + \cdots + x_d^2 \le r^2 : x_i \in \mathbb{R}\}$.

Armed with this knowledge, $P\big(B_0(r)\big)$ becomes:

$$\int_0^r \int_0^{2\pi} \ldots \int_0^{\pi} \underbrace{(2\pi)^{-\frac{D}{2}} e^{-\frac{1}{2}\|\mathbf{x}\|^2}}_{pdf\ normal\ dist.} \underbrace{r^{d-1} d\theta_1 \ldots d\theta_{d-1} dr}_{infinitismal\ area} = \int_0^r \int_0^{2\pi} \ldots \int_0^{\pi} (2\pi)^{-\frac{D}{2}} e^{-\frac{1}{2}r^2 \underbrace{K(\boldsymbol{\theta})}_{=1}} r^{d-1} d\theta_1 \ldots d\theta_{d-1} dr,$$

$$= \int_0^r r^{d-1} (2\pi)^{-\frac{D}{2}} e^{-\frac{1}{2}r^2} \underbrace{\int_0^{2\pi} \ldots \int_0^{\pi} d\theta_1 \ldots d\theta_{d-1}}_{Surface\ Area\ unit\ n-sphere\ S_D} dr,$$

$$= \int_0^r S_D r^{d-1} (2\pi)^{-\frac{D}{2}} e^{-\frac{1}{2}r^2} dr.$$

Ergo $p(r)dr = S_D r^{d-1} (2\pi)^{-\frac{D}{2}} e^{-\frac{1}{2}r^2}$.

Now we are looking for the maximum density $\max_r p(r)$:

$$\frac{d}{dr}\left[\log r^{d-1} + \log e^{-\frac{1}{2}r^2}\right] = \frac{(d-1)r^{d-2}}{r^{d-1}} - r = 0.$$

$\Leftrightarrow (d-1) = r^2$.
Because radii are non-negative, we have a maximum at $\sqrt{d-1}$.

Now if we set $\|\mathbf{x}\| = \sqrt{d-1}$, we get

$$\frac{p(\mathbf{x})}{p(0)} = \frac{(2\pi)^{-\frac{D}{2}} e^{-\frac{1}{2}\|\mathbf{x}\|^2}}{(2\pi)^{-\frac{D}{2}} e^{-\frac{1}{2}\|\mathbf{0}\|^2}} = \frac{e^{-\frac{1}{2}(d-1)}}{e^{-\frac{1}{2}}} = e^{-\frac{d}{2}}.$$

4A-3.

Let $L(\mathbf{w})$:

$$L(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\left(t_n - \mathbf{w}^T\boldsymbol{\phi}(x_n)\right)^2 + \frac{\lambda}{2}\mathbf{w}^T\mathbf{w}$$

Derivative of $\ln L(\mathbf{w})$ with respect to w_i:

$$\frac{\partial}{\partial w_i}\ln L(\mathbf{w}) = \sum_{n=1}^{N}\phi_i(x_n)\left(t_n - \mathbf{w}^T\boldsymbol{\phi}(x_n)\right) + \lambda w_i$$

Conversion to matrix/vector operations:

$$\sum_{n=1}^{N}\phi_i(x_n)\left(t_n - \mathbf{w}^T\boldsymbol{\phi}(x_n)\right) + \lambda w_i = \mathrm{col_i}(\mathbf{\Phi})^T(\mathbf{t} - \mathbf{\Phi}\mathbf{w}) + \lambda w_i$$

Generalized for all w:

$$\frac{\partial}{\partial \mathbf{w}}\ln L(\mathbf{w}) = \mathbf{\Phi}^T(\mathbf{t} - \mathbf{\Phi}\mathbf{w}) + \lambda\mathbf{w}$$

Setting zero and solving for $\mathbf{w}$:

$$\begin{aligned}
& \mathbf{\Phi}^T(\mathbf{t} - \mathbf{\Phi}\mathbf{w}) + \lambda\mathbf{w} && = 0 \\
\Leftrightarrow\quad & \mathbf{\Phi}^T\mathbf{t} + (-\mathbf{\Phi}^T\mathbf{\Phi} + \lambda\mathbf{I})\mathbf{w} && = 0 \\
\Leftrightarrow\quad & \mathbf{w} && = (-\lambda\mathbf{I} + \mathbf{\Phi}^T\mathbf{\Phi})^{+}\mathbf{\Phi}^T\mathbf{t}
\end{aligned}$$

As usual, $\mathbf{A}^{+}$ denotes the Penrose pseudo inverse.
Because λ can be an arbitrary normative factor, it is also possible to write:

$$\mathbf{w} = (\lambda\mathbf{I} + \mathbf{\Phi}^T\mathbf{\Phi})^{+}\mathbf{\Phi}^T\mathbf{t}$$

Pictures from Python: