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 $\mathbf{t}$ 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, $\lim_{n\to\infty} f_n{}'$ converges to g uniformly. Due to $(*)$, 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 $p(x|x_i, \dots x_0)$.

4A-2.

Auxiliary calculation

Using the Jacobian integral substitution, the area of an infinitesimal d-dimensional volume element is

$$d^n V = \left|\det\frac{\partial(x_i)}{\partial(r,\phi_j)}\right| dr d\theta_1 \dots d\theta_{n-1}$$

$$\begin{aligned} d^n V &= \left|\det\frac{\partial(x_i)}{\partial(r,\phi_j)}\right| dr d\theta_1 \dots d\theta_{n-1} \\ &= r^{n-1} \underbrace{\sin^{n-2}\theta_1 \sin^{n-3}\theta_2 \cdot \dots \cdot \sin\theta_{n-2}}_{g(\theta_1,\dots,\theta_{n-1})} dr d\theta_1 \dots d\theta_{n-1} \\ &= r^{n-1} g(\boldsymbol{\theta}) dr d\theta_1 \dots d\theta_{n-1} \end{aligned}$$

Above can be seen here:

https://en.wikipedia.org/wiki/N-sphere#Spherical_coordinates

Furthermore: We have $||\mathbf{x}||^2 = r^2$ for all $\boldsymbol{\theta}$, because the squared length of a coordinate point is r^2.

The surface area $S_D = S_{n-1}$ of an n dimensional sphere with radius r is denoted by

$$S_{n-1} = \frac{dV_n(R)}{dR} = nC_n R^{n-1}$$

where $V_n(R)$ denotes the volume of a sphere of radius R.
This can be seen here:
http://scipp.ucsc.edu/~haber/ph116A/volume_11.pdf

There we can also find the identity (Eq.7):

$$\begin{aligned} nC_n &= \int \dots \int d\Omega_{n-1} \\ (*) \qquad &= \int \dots \int g(\boldsymbol{\theta}) d\theta_1 d\theta_2 \dots d\theta_{n-1} \end{aligned}$$

$$\begin{aligned} p(r,\theta_1,\theta_2,\dots,\theta_{n-1}) &= \int_0^r \int_0^{2\pi} \dots \int_0^{\pi} \underbrace{(2\pi)^{-\frac{D}{2}} e^{-\frac{1}{2}r^2}}_{pdf\ normal\ dist.} \underbrace{r^{d-1}\sin^{d-2}\theta_1 \sin^{d-3}\theta_2 \cdot \dots \cdot \sin\theta_{d-2}\, d\theta_1 \dots d\theta_{d-1} dr}_{infinitismal\ area} \\ &= \int_0^r (2\pi)^{-\frac{D}{2}} e^{-\frac{1}{2}r^2} r^{d-1} \int_0^{2\pi} \dots \int_0^{\pi} \underbrace{g(\boldsymbol{\theta}) d\theta_1 \dots d\theta_{n-1}}_{=nC_n \cdot 1, acc. to. (*)} dr \\ &= \int_0^r (2\pi)^{-\frac{D}{2}} e^{-\frac{1}{2}r^2} r^{d-1} S_D\, dr \end{aligned}$$

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

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

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:

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

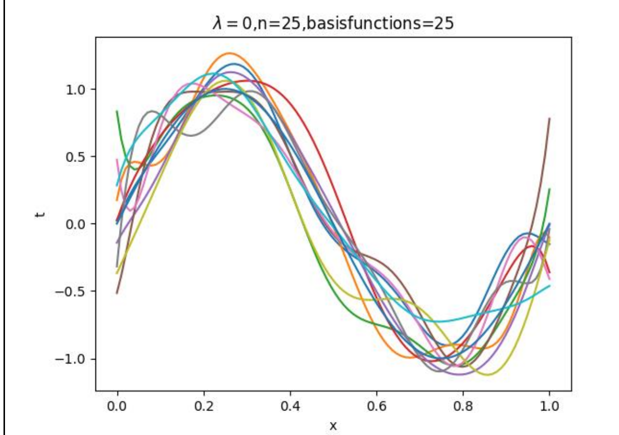
Generalized for all w:

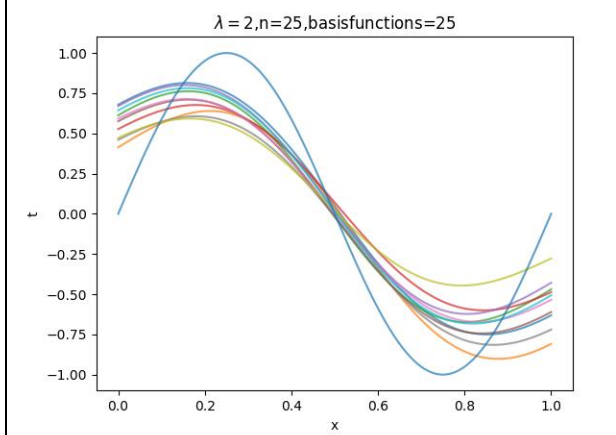
$$\frac{\partial}{\partial \mathbf{w}}\ln L(\mathbf{w}) = -\boldsymbol{\Phi}^T(\mathbf{t} - \boldsymbol{\Phi}\mathbf{w}) + \lambda\mathbf{w}$$

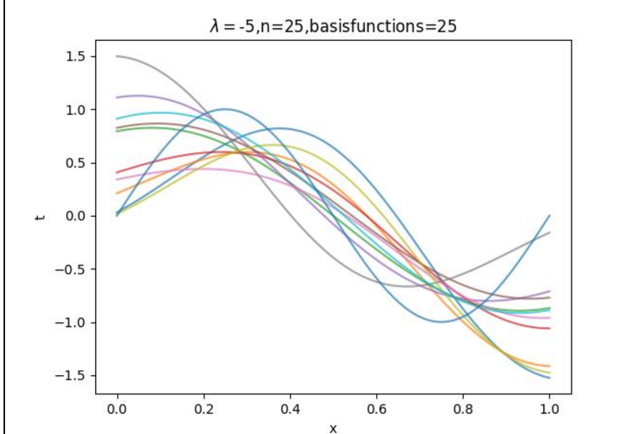
Setting zero and solving for $\mathbf{w}$:

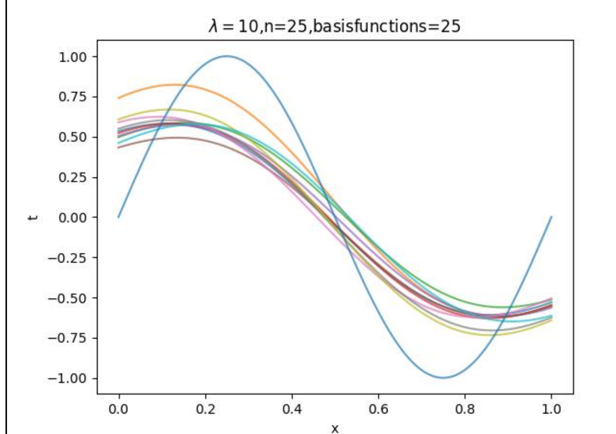
$$\begin{aligned}
& -\boldsymbol{\Phi}^T(\mathbf{t} - \boldsymbol{\Phi}\mathbf{w}) + \lambda\mathbf{w} && = 0 \\
\Leftrightarrow\quad & -\boldsymbol{\Phi}^T\mathbf{t} + (\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \lambda\mathbf{I})\mathbf{w} && = 0 \\
\Leftrightarrow\quad & \mathbf{w} && = (\lambda\mathbf{I}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

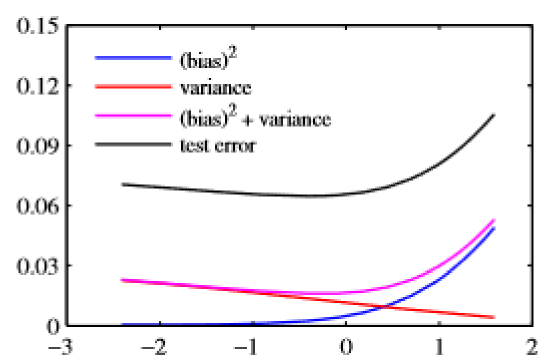
As usual, $\mathbf{A}^{+}$ denotes the Penrose pseudo inverse.











Our results coincide with this picture from the Bishop Book, showing that much regularization leads to a worse fit (higher bias) but less variance.l