1 Curve fitting / Linear regression

Remember: Linear does **not** mean linear in x but linear in w!
Given $D = \{(\mathbf{x}_1, y_1), ..., (\mathbf{x}_N, y_N)\}$, $\mathbf{x}_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$ choose $f : \mathbb{R}^D \to \mathbb{R}$ to predict $\hat{y}$ from $\mathbf{x}$

General: $$f(\mathbf{x}) = \mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}) = \sum_{d=1}^{D} w_d \phi_d(\mathbf{x})$$

Identity (vector): $\boldsymbol{\varphi}(\mathbf{x}) = \mathbf{x}$ → $$f(\mathbf{x}) = \mathbf{w}^T \mathbf{x} = \sum_{d=1}^{D} w_d x_d$$

Polynomial (scalar): $\boldsymbol{\varphi}(x) = (1, x, x^2, x^3, \dots)$ → $$f(x) = \mathbf{w}^T \mathbf{x} = \sum_{j=0}^{M} w_j x^j$$

You could also use Fourier basis, Wavelets etc.

Task 1.1 Least Squares

Given are $N = 10$ observations of the process

$$y(x) = \sin(2\pi x) + e \qquad \text{with} \qquad e \sim \mathcal{N}\left(0, \sigma_e^2 = 0.18\right)$$

as training data, where x is evenly spaced in the interval $[0, 1]$. Estimate the coefficients $\mathbf{w} = (w_0, \dots, w_M)^{\mathrm{T}}$ of the polynomial $\hat{y}(x, \mathbf{w})$. Our goal is to use the polynomial to predict an observation of y given a new input x as follows:

$$\hat{y}(x, \mathbf{w}) = \sum_{j=0}^{M} w_j x^j = \mathbf{w}^{\mathrm{T}} \boldsymbol{\varphi}(x)$$
$$\boldsymbol{\varphi}(x) = \begin{bmatrix} x^0 & x^1 & \dots & x^M \end{bmatrix}^{\mathrm{T}}$$

Determine a system of equations for the coefficients $\mathbf{w}$ by minimizing the error function

$$\begin{aligned} E(\mathbf{w}) &= \frac{1}{2} \sum_{n=1}^{N} \left(\hat{y}(x_n, \mathbf{w}) - y_n\right)^2 \\ &= \frac{1}{2} \left\|\boldsymbol{\Phi}(\mathbf{x})\mathbf{w} - \mathbf{y}\right\|^2 . \end{aligned}$$

- How do you find a possible minimum?
- How is $\mathbf{x}$, $\boldsymbol{\Phi}(\mathbf{x})$ and $\mathbf{y}$ defined?

Hints:

- The result can be expressed using matrices: $\mathbf{A}\mathbf{w} = \mathbf{v}$
- Can you express the solution such that it depends on the data matrix $\boldsymbol{\Phi} = \boldsymbol{\Phi}(\mathbf{x})$?

Solution

> To minimize $E(\mathbf{w})$ we have to calculate $\mathbf{w}$ so that $\frac{\partial E(\mathbf{w})}{\partial w_i} \overset{!}{=} 0$. Substituting $\hat{y}(x, \mathbf{w})$ into $E(\mathbf{w})$,

differentiating with respect to w_i, with $i \in \{0, \dots, M\}$ and rearranging results into

$$\begin{aligned}
\frac{\partial}{\partial w_i}\left(\frac{1}{2}\sum_{n=1}^{N}\left(\sum_{j=0}^{M} w_j x_n^j - y_n\right)^2\right) &\overset{!}{=} 0 \\
\sum_{n=1}^{N}\left(\sum_{j=0}^{M} w_j x_n^j - y_n\right) x_n^i &\overset{!}{=} 0 \\
\sum_{n=1}^{N}\sum_{j=0}^{M} w_j x_n^j x_n^i - \sum_{n=1}^{N} y_n x_n^i &\overset{!}{=} 0 \\
\sum_{j=0}^{M}\underbrace{\sum_{n=1}^{N} x_n^{i+j}}_{A_{ij}} w_j &= \underbrace{\sum_{n=1}^{N} y_n x_n^i}_{v_i} \\
\sum_{j=0}^{M} A_{ij} w_j &= v_i
\end{aligned}$$

Written with matrix $\mathbf{A}$ and vectors $\mathbf{w}$ and $\mathbf{v}$ we can solve for $\mathbf{v}$:

$$\begin{aligned}
\mathbf{A}\mathbf{w} &= \mathbf{v} \\
\mathbf{w} &= \mathbf{A}^{-1}\mathbf{v}
\end{aligned}$$

Alternative solution: Using the data matrix

$$\boldsymbol{\Phi}(\mathbf{x}) = \begin{bmatrix} \boldsymbol{\varphi}(x_1)^{\mathrm{T}} \\ \vdots \\ \boldsymbol{\varphi}(x_N)^{\mathrm{T}} \end{bmatrix} = \begin{bmatrix} x_1^0 & \dots & x_1^M \\ \vdots & & \vdots \\ x_N^0 & \dots & x_N^M \end{bmatrix} = \begin{bmatrix} 1 & x_1 & x_1^2 & \dots & x_1^M \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_N & x_N^2 & \dots & x_N^M \end{bmatrix}$$

in the objective, the calculation is:

$$\begin{aligned}
E(\mathbf{w}) &= \frac{1}{2}\left(\mathbf{w}^{\mathrm{T}}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\mathbf{w} - 2\mathbf{w}^{\mathrm{T}}\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{y} - \mathbf{y}^{\mathrm{T}}\mathbf{y}\right) \\
\frac{\partial E(\mathbf{w})}{\partial \mathbf{w}} &= \boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\mathbf{w} - \boldsymbol{\Phi}^{\mathrm{T}}\mathbf{y} \overset{!}{=} \mathbf{0} \\
\mathbf{w} &= \left(\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{y} \\
&= \boldsymbol{\Phi}^{+}\mathbf{y}
\end{aligned}$$

where $\boldsymbol{\Phi} = \boldsymbol{\Phi}(\mathbf{x})$ and $\boldsymbol{\Phi}^{+}$ is the Moore-Penrose inverse or pseudoinverse of $\mathbf{X}$. The advantage of the pseudoinverse is, that this inverse is more stable, when using a numeric solver. In numpy you can use `np.linalg.lstsq` or `np.linalg.pinv` to solve a linear system that is written with a pseudoinverse.

Task 1.2 Least Squares (Code)

We now want to visualize the regression for different polynomial orders. The code to plot the result is already written for you. However, it calls a function `get_weight_vector` which must be written by you.

a) Figure out, how you can determine the data matrix $\mathbf{X}$.

b) Write the code for the function `get_weight_vector`. What do the arguments mean of this function? Which shape do they have?

c) Is it possible to estimate a curve that hits every training point? When should this happen? Does this happen? If not, do you have an idea why? Test this also with $N = 15$.

Hints:

- If you are not familiar with python, this, this and/or this tutorial might help you to get started
- Look at the documentation of the `polyvander` function (doc)
  - Try to relate it to the ϕ in the introduction
- To invert the matrix, you can use `np.linalg.inv` (doc)
- To get the matrix-matrix product, you need to use the `matmul` function (doc)
- To get the transposed matrix, you can simply access its `T` property

Task 1.3 RMSE (Code)

Generate 100 new test samples for x in the range 0 to 1 and calculate the RMSE between those samples. Calculate the RMSE for the training and test data. Let the linear regression order M start from 0. Choose the highest regression order such, that the RMSE goes to zero. Does the RMSE goes to zero for the training or test data?

Note: The coefficients should still be estimated using the old number of samples.

Task 1.4 Regularization

Determine a system of equations for the coefficients $\mathbf{w}$ by minimizing the error function

$$\widetilde{E}\left(\mathbf{w}\right) = \frac{1}{2}\sum_{n=1}^{N}\left(\hat{y}\left(x_n, \mathbf{w}\right) - y_n\right)^2 + \frac{\lambda}{2}\|\mathbf{w}\|^2.$$

Hint: Compared to the previous task, we just have a slightly different cost function. The steps are basically the same.

Solution

> To minimize $\widetilde{E}\left(\mathbf{w}\right)$ we have to calculate $\mathbf{w}$ so that $\frac{\partial \widetilde{E}(\mathbf{w})}{\partial w_i} \overset{!}{=} 0$. Substituting $y\left(x, \mathbf{w}\right)$ into $\widetilde{E}\left(\mathbf{w}\right)$,

differentiating with respect to w_i, with $i = 0 \dots M$ and rearranging results into

$$\begin{aligned}
\frac{\partial}{\partial w_i}\left(\frac{1}{2}\sum_{n=1}^{N}\left(\sum_{j=0}^{M} w_j x_n^j - y_n\right)^2 + \frac{\lambda}{2}\|\mathbf{w}\|^2\right) &\overset{!}{=} 0\\
\sum_{n=1}^{N}\left(\sum_{j=0}^{M} w_j x_n^j - y_n\right) x_n^i + \lambda w_i &\overset{!}{=} 0\\
\sum_{n=1}^{N}\sum_{j=0}^{M} w_j x_n^j x_n^i - \sum_{n=1}^{N} y_n x_n^i + \lambda w_i &\overset{!}{=} 0\\
\sum_{j=0}^{M}\underbrace{\sum_{n=1}^{N} x_n^{i+j}}_{A_{ij}} w_j + \lambda w_i &= \underbrace{\sum_{n=1}^{N} y_n x_n^i}_{v_i}\\
\sum_{j=0}^{M} A_{ij} w_j + \lambda w_i &= v_i
\end{aligned}$$

Written with matrix $\mathbf{A}$, unity matrix $\mathbf{I}$ and vectors $\mathbf{w}$ and $\mathbf{v}$ we can solve for $\mathbf{v}$:

$$\begin{aligned}
\mathbf{A}\mathbf{w} + \lambda\mathbf{I}\mathbf{w} &= \mathbf{v}\\
(\mathbf{A} + \lambda\mathbf{I})\,\mathbf{w} &= \mathbf{v}\\
\mathbf{w} &= (\mathbf{A} + \lambda\mathbf{I})^{-1}\,\mathbf{v}
\end{aligned}$$

Task 1.5 Regularization (Code)

Add the regularization to the function you have written before and add an additional slider to control it. What are reasonable values to try for λ (Linear range?)?

Additional notes:

- `lambda` is a reserved keyword in python (for an anonymous function) so please use another name for the variable
- Can you now increase the polynomial order?
- How does the regularization influence stability of the matrix inversion?

Task 1.6 RMSE (Code)

Use the RMSE function to plot the RMSE for different M as a function of $\ln \lambda$. Can you reproduce the plot from the lecture?

Task 1.7

Show that maximum likelihood (ML) estimation of $\mathbf{w}$ is equal to the least squares solution. Assume that the training data points are i.i.d. and the probability for one point is $p(y_n|x_n, \mathbf{w}) = \mathcal{N}\left(y_n; \hat{y}(x_n, \mathbf{w}), \sigma_e^2\right)$

Solution

The probability of one training data point (y_n, x_n) given $\mathbf{w}$ is

$$p(y_n|x_n, \mathbf{w}) = \mathcal{N}\left(y_n; y(x_n, \mathbf{w}), \sigma_e^2\right). \tag{1}$$

For the likelihood of the training data $\mathbf{y} = (y_1, \dots, y_N)^{\mathrm{T}}$ and $\mathbf{x} = (x_1, \dots, x_N)^{\mathrm{T}}$ we then get

$$p\left(\mathbf{y}|\mathbf{x}, \mathbf{w}\right) = \prod_{n=1}^{N} \mathcal{N}\left(y_n; y\left(x_n, \mathbf{w}\right), \sigma_e^2\right). \tag{2}$$

Taking the logarithm gives (Note: the logarithm of a product of values equals to the the sum of the logarithms of the values)

$$\begin{aligned}\ln\left(p\left(\mathbf{y}|\mathbf{w}\right)\right) &= \sum_{n=1}^{N} \ln\left(\mathcal{N}\left(y_n; \hat{y}\left(x_n, \mathbf{w}\right), \sigma_e^2\right)\right)\\ &= \sum_{n=1}^{N}\left(-\frac{1}{2}\ln\left(\sigma_e^2\right) - \frac{1}{2}\ln\left(2\pi\right) - \frac{1}{2\sigma_e^2}\left(y_n - \hat{y}\left(x_n, \mathbf{w}\right)\right)^2\right)\\ &= -\frac{N}{2}\ln\left(\sigma_e^2\right) - \frac{N}{2}\ln\left(2\pi\right) - \frac{1}{2\sigma_e^2}\sum_{n=1}^{N}\left(y_n - \hat{y}\left(x_n, \mathbf{w}\right)\right)^2\end{aligned}$$

Maximizing the log-likelihood with respect to $\mathbf{w}$ is equal to minimizing the negative log-likelihood with respect to $\mathbf{w}$. Also the first and second term are independent of $\mathbf{w}$ and the factor $\frac{1}{\sigma_e^2}$ in the third term will not change the position (in $\mathbf{w}$) of the minimum. As a result we have to minimize the same function as the least squares error function:

$$\frac{1}{2}\sum_{n=1}^{N}\left(y_n - \hat{y}\left(x_n, \mathbf{w}\right)\right)^2. \tag{3}$$

Task 1.8

Show that maximum a posteriori (MAP) estimation of $\mathbf{w}$ is equal to the regularized least squares solution when using the following prior distribution for $\mathbf{w}$:

$$p\left(\mathbf{w}\right) = \mathcal{N}\left(\mathbf{w}; \mathbf{0}, \sigma_w^2\mathbf{I}\right). \tag{4}$$

Hint: The posterior distribution $p(\mathbf{w}|\mathbf{x}, \mathbf{y})$ can be computed using Bayes' theorem

$$p(\mathbf{w}|\mathbf{x}, \mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{w}, \mathbf{x})p(\mathbf{w}|\mathbf{x})}{p(\mathbf{y}|\mathbf{x})} = \frac{p(\mathbf{y}|\mathbf{w}, \mathbf{x})p(\mathbf{w})}{p(\mathbf{y}|\mathbf{x})}. \tag{5}$$

As the denominator is independent of $\mathbf{w}$ we only have to maximize

$$p(\mathbf{w}|\mathbf{x}, \mathbf{y}) \propto p(\mathbf{y}|\mathbf{w}, \mathbf{x})p(\mathbf{w}). \tag{6}$$

Addition

This solution is identical to the regularized least squares solution. Can you identify λ in this solution?

Which prior is implicitly assumed for $\mathbf{w}$ in task 1.1? Explain why the weights explode as the order increases and why we get a flat line if we choose a big λ in task 1.5

Solution

Taking the logarithm gives

$$\begin{aligned} l(\mathbf{w}|\mathbf{y}) &= \ln\left(p(\mathbf{y}|\mathbf{w})p(\mathbf{w})\right) && (7)\\ &= \ln\left(p(\mathbf{y}|\mathbf{w})\right) + \ln\left(p(\mathbf{w})\right) && (8)\\ &= \ln\left(p(\mathbf{y}|\mathbf{w})\right) - \frac{M}{2}\ln\left(\sigma_w\right) - \frac{M}{2}\ln\left(2\pi\right) - \frac{1}{2\sigma_w^2}\mathbf{w}^{\mathrm{T}}\mathbf{w}. && (9) \end{aligned}$$

Again we minimize the negative log-likelihood with respect to $\mathbf{w}$. So we have to minimize

$$\frac{1}{2\sigma_e^2}\sum_{n=1}^{N}\left(y_n - y\left(x_n, \mathbf{w}\right)\right)^2 + \frac{1}{2\sigma_w^2}\mathbf{w}^{\mathrm{T}}\mathbf{w} \tag{10}$$

where the first term is the same as for the ML estimation. Multiplying by σ_e^2 will not change the position (in $\mathbf{w}$) of the minimum and therefore we have to minimize the same function as for regularized least squares with $\lambda = \frac{\sigma_e^2}{\sigma_w^2}$

$$\frac{1}{2}\sum_{n=1}^{N}\left(y_n - y\left(x_n, \mathbf{w}\right)\right)^2 + \frac{\sigma_e^2}{2\sigma_w^2}\|\mathbf{w}\|^2. \tag{11}$$

Task 1.9

Compute the predictive distribution $p\left(y|x, \mathbf{x}, \mathbf{y}\right)$ for t given a new value of x and given the training data $\mathbf{y} = \left(y_1, \ldots, y_N\right)^{\mathrm{T}}$ and $\mathbf{x} = \left(x_1, \ldots, x_N\right)^{\mathrm{T}}$.

Notes:

- Use the following transformations:

$$\begin{aligned} &\text{given}\\ p(\mathbf{u}) &= \mathcal{N}\left(\mathbf{u}; \boldsymbol{\mu}, \boldsymbol{\Sigma}\right)\\ p(\mathbf{v}|\mathbf{u}) &= \mathcal{N}\left(\mathbf{v}; \mathbf{A}\mathbf{u} + \mathbf{b}, \mathbf{C}\right)\\ &\text{we can calculate}\\ p(\mathbf{v}) &= \mathcal{N}\left(\mathbf{v}; \mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{C} + \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^{\mathrm{T}}\right)\\ p(\mathbf{u}|\mathbf{v}) &= \mathcal{N}\left(\mathbf{u}; \mathbf{S}\left(\mathbf{A}^{\mathrm{T}}\mathbf{C}^{-1}\left(\mathbf{v} - \mathbf{b}\right) + \boldsymbol{\Sigma}^{-1}\mu\right), \mathbf{S}\right)\\ &\text{where}\\ \mathbf{S} &= \left(\boldsymbol{\Sigma}^{-1} + \mathbf{A}^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{A}\right)^{-1} \end{aligned}$$

- Use the same prior as in task 1.8.
- $p\left(y|x, \mathbf{x}, \mathbf{y}\right) = \int p(y|\mathbf{w})p(\mathbf{w}|\mathbf{y})\mathrm{d}\mathbf{w}$

Solution

Before starting with the solution, a small recap about the assumptions and connections between the random variables. The target y and the observation $\mathbf{x}$ are modeled as

$$y = \mathbf{w}^{\mathrm{T}}\boldsymbol{\varphi}(x) + e \tag{12}$$

where the parameters $\mathbf{w}$ and the noise e are random with a Gaussian distribution:

$$\begin{aligned} \mathbf{w} &\sim \mathcal{N}\left(\mathbf{w}; \mathbf{0}, \sigma_w\mathbf{I}\right) && (13)\\ e &\sim \mathcal{N}\left(e; 0, \sigma_e\right). && (14) \end{aligned}$$

The parameters $\mathbf{w}$ are obtained from N training data pairs of y_n and $\mathbf{x}_n$.
The distribution $p\left(y|x,\mathbf{x},\mathbf{y}\right)$ can be expressed as the marginalization of $p(y,\mathbf{w}|\mathbf{y})$ over $\mathbf{w}$

$$\begin{align}
p\left(y|x,\mathbf{x},\mathbf{y}\right) &= \int p(y,\mathbf{w}|x,\mathbf{x},\mathbf{y})\mathrm{d}\mathbf{w} \tag{15}\\
&= \int p(y|\mathbf{w},x,\mathbf{x},\mathbf{y})p(\mathbf{w}|x,\mathbf{x},\mathbf{y})\mathrm{d}\mathbf{w} \tag{16}\\
&= \int p(y|\mathbf{w},x)p(\mathbf{w}|\mathbf{x},\mathbf{y})\mathrm{d}\mathbf{w}. \tag{17}
\end{align}$$

For $p(y|\mathbf{w},x) = p(y|\mathbf{w},x,\mathbf{x},\mathbf{y})$ we used the independence of y and $\mathbf{x},\mathbf{y}$ when $\mathbf{w}$ is known. This independence comes from the fact, that the information only travels form $\mathbf{x},\mathbf{y}$ to $\mathbf{w}$ and then from $\mathbf{w}$ to y.
The posterior distribution $p(\mathbf{w}|\mathbf{x},\mathbf{y})$ can be computed using Bayes' theorem the same way as for the MAP estimation although in contrast to MAP estimation this time we have to compute the complete distribution instead of the mean:

$$\begin{align}
p(\mathbf{w}|\mathbf{x},\mathbf{y}) &= \frac{p(\mathbf{y}|\mathbf{w},\mathbf{x})p(\mathbf{w})}{p(\mathbf{y}|\mathbf{x})} \tag{18}\\
&\propto p(\mathbf{y}|\mathbf{w},\mathbf{x})p(\mathbf{w}) \tag{19}\\
&= \prod_n p(y_n|\mathbf{w},x_n)p(\mathbf{w}) \tag{20}\\
&\propto \exp\left(-\frac{1}{2\sigma_e^2}\sum_n(y_n-\boldsymbol{\varphi}^{\mathrm{T}}\mathbf{w})^2-\frac{1}{\sigma_w^2}\mathbf{w}^{\mathrm{T}}\mathbf{w}\right) \tag{21}\\
&= \exp\left(-\frac{1}{2\sigma_e^2}\|\mathbf{y}-\boldsymbol{\Phi}\mathbf{w}\|^2-\frac{1}{2\sigma_w^2}\mathbf{w}^{\mathrm{T}}\mathbf{w}\right) \tag{22}\\
&= \exp\left(-\frac{1}{2\sigma_e^2}\left(\mathbf{y}^{\mathrm{T}}\mathbf{y}-2\mathbf{y}^{\mathrm{T}}\boldsymbol{\Phi}\mathbf{w}+\mathbf{w}^{\mathrm{T}}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\mathbf{w}\right)-\frac{1}{2\sigma_w^2}\mathbf{w}^{\mathrm{T}}\mathbf{w}\right) \tag{23}\\
&= \exp\left(-\frac{1}{2}\left(\frac{\mathbf{y}^{\mathrm{T}}\mathbf{y}}{\sigma_e^2}-\underbrace{\frac{2\mathbf{y}^{\mathrm{T}}\boldsymbol{\Phi}}{\sigma_e^2}\mathbf{w}}_{=2\mathbf{m}^{\mathrm{T}}\mathbf{S}^{-1}\mathbf{w}}+\mathbf{w}^{\mathrm{T}}\underbrace{\left(\frac{\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}}{\sigma_e^2}+\frac{\mathbf{I}}{\sigma_w^2}\right)}_{=\mathbf{S}^{-1}}\mathbf{w}\right)\right) \tag{24}
\end{align}$$

This if the form of a Gaussian distributed variable, so $p(\mathbf{w}|\mathbf{x},\mathbf{y})$ have to be Gaussian distributed and we can look in the exponent to identify the mean $\mathbf{m}$ and the covariance matrix $\mathbf{S}$:

$$\begin{align}
\mathbf{S} &= \left(\frac{\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}}{\sigma_e^2}+\frac{\mathbf{I}}{\sigma_w^2}\right)^{-1} \tag{25}\\
\mathbf{S}^{-1}\mathbf{m} &= \boldsymbol{\Phi}^{\mathrm{T}}\mathbf{y} \tag{26}\\
\mathbf{m} &= \mathbf{S}\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{y} = \left(\frac{\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}}{\sigma_e^2}+\frac{\mathbf{I}}{\sigma_w^2}\right)^{-1}\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{y} \tag{27}\\
p(\mathbf{w}|\mathbf{x},\mathbf{y}) &= \mathcal{N}(\mathbf{w};\mathbf{m},\mathbf{S}) \tag{28}
\end{align}$$

To obtain $p\left(y|x,\mathbf{x},\mathbf{y}\right)$, we take a look at

$$y = \mathbf{w}^{\mathrm{T}}\boldsymbol{\varphi}(x)+e. \tag{29}$$

When $\mathbf{x}$ and $\mathbf{y}$ are given, $\mathbf{w}$ is Gaussian distributed. e is also Gaussian distributed. $\boldsymbol{\varphi}(x)$ is not random, since x is given. So when $\mathbf{x}$, $\mathbf{y}$ and x are given y is a linear combination of Gaussian distributed random variables, so y is also Gaussian distributed and we only need to estimate the mean and variance to get the distribution of $p\left(y|x,\mathbf{x},\mathbf{y}\right)$.

$$\begin{aligned}
\mu_{y|x,\mathbf{x},\mathbf{y}} &= \mathbb{E}\left[y|x,\mathbf{x},\mathbf{y}\right] \tag{30}\\
&= \mathbb{E}\left[(\mathbf{w}^{\mathrm{T}}\boldsymbol{\varphi}(x)+e)|x,\mathbf{x},\mathbf{y}\right] \tag{31}\\
&= \mathbb{E}\left[\mathbf{w}^{\mathrm{T}}|\mathbf{x},\mathbf{y}\right]\boldsymbol{\varphi}(x)+\mathbb{E}\left[e\right] \tag{32}\\
&= \mathbf{m}\boldsymbol{\varphi}(x) \tag{33}\\
\sigma_{y|x,\mathbf{x},\mathbf{y}} &= \mathrm{var}(y|x,\mathbf{x},\mathbf{y}) \tag{34}\\
&= \mathrm{var}((\mathbf{w}^{\mathrm{T}}\boldsymbol{\varphi}(x)+e)|x,\mathbf{x},\mathbf{y}) \tag{35}\\
&= \mathrm{var}((\mathbf{w}^{\mathrm{T}}\boldsymbol{\varphi}(x)|x,\mathbf{x},\mathbf{y})+\mathrm{var}(e) \tag{36}\\
&= \boldsymbol{\varphi}(x)^{\mathrm{T}}\mathrm{var}((\mathbf{w}|x,\mathbf{x},\mathbf{y})\boldsymbol{\varphi}(x)+\mathrm{var}(e) \tag{37}\\
&= \boldsymbol{\varphi}(x)^{\mathrm{T}}\mathbf{S}\boldsymbol{\varphi}(x)+\sigma_e \tag{38}\\
p\left(y|x,\mathbf{x},\mathbf{y}\right) &= \mathcal{N}(y;\mu_{y|x,\mathbf{x},\mathbf{y}},\sigma_{y|x,\mathbf{x},\mathbf{y}}) \tag{39}
\end{aligned}$$

Alternative solution:
To compute $p\left(y|x,\mathbf{x},\mathbf{y}\right)$ we have to start with $p(y|x,\mathbf{w})$ and marginalize over $\mathbf{w}$ using the posterior distribution $p(\mathbf{w}|\mathbf{x},\mathbf{y})$ of $\mathbf{w}$ given the training data $\mathbf{x}$ and $\mathbf{y}$ (Note: by the given model when knowing the posterior distribution of $\mathbf{w}$ given $\mathbf{y}$ then $p(y|x,\mathbf{w},\mathbf{x},\mathbf{y})$ only depends on $\mathbf{w}$ because the information only travels form $\mathbf{y}$ to $\mathbf{w}$ and then from $\mathbf{w}$ to y so we can simply write $p(y|\mathbf{w})$)

$$\begin{aligned}
p\left(y|x,\mathbf{x},\mathbf{y}\right) &= \int p(y|\mathbf{w})p(\mathbf{w}|\mathbf{y})\mathrm{d}\mathbf{w} \tag{40}\\
&= \int p(y,\mathbf{w}|\mathbf{y})\mathrm{d}\mathbf{w}. \tag{41}
\end{aligned}$$

The posterior distribution $p(\mathbf{w}|\mathbf{x},\mathbf{y})$ can be computed using Bayes' theorem the same way as for the MAP estimation although in contrast to MAP estimation this time we have to compute the complete distribution.

$$p(\mathbf{w}|\mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{w})p(\mathbf{w})}{p(\mathbf{y})} = \frac{p(\mathbf{y}|\mathbf{w})p(\mathbf{w})}{\int p(\mathbf{y}|\mathbf{w})p(\mathbf{w})\mathrm{d}\mathbf{w}}. \tag{42}$$

To use the given transformations we first rewrite the polynomial $\hat{y}\left(x,\mathbf{w}\right)$ in vector matrix notation. For one value x we get

$$\begin{aligned}
\hat{y}\left(x,\mathbf{w}\right) &= \sum_{j=0}^{M} w_j x^j \tag{43}\\
&= \boldsymbol{\varphi}(x)^{\mathrm{T}}\mathbf{w} \text{ with } \boldsymbol{\varphi}(x) \qquad = \left(1,x,x^2,\ldots,x^M\right)^{\mathrm{T}}. \tag{44}
\end{aligned}$$

For the vector $\mathbf{x}$ we get

$$\hat{y}\left(\mathbf{x},\mathbf{w}\right) = \boldsymbol{\Phi}(\mathbf{x})\mathbf{w} \text{ with } \boldsymbol{\Phi}(\mathbf{x}) \qquad = \left(\boldsymbol{\varphi}(x_1),\ldots,\boldsymbol{\varphi}(x_N)\right)^{\mathrm{T}}. \tag{45}$$

Using this result we can rewrite the distribution $p(\mathbf{y}|\mathbf{w})$ of $\mathbf{y}$ given the coefficients $\mathbf{w}$ of the polynomial. Recalling the prior distribution $p\left(\mathbf{w}\right)$ for $\mathbf{w}$ we can now compute the posterior distribution $p(\mathbf{w}|\mathbf{y})$ using the given transformations

$$\begin{aligned} p(\mathbf{y}|\mathbf{w}) &= \mathcal{N}\left(\mathbf{y}; \boldsymbol{\Phi}(\mathbf{x})\mathbf{w}, \sigma_e^2\mathbf{I}\right) && (46)\\ p\left(\mathbf{w}\right) &= \mathcal{N}\left(\mathbf{w}; \mathbf{0}, \sigma_w^2\mathbf{I}\right) && (47)\\ \Rightarrow p(\mathbf{w}|\mathbf{y}) &= \mathcal{N}\left(\mathbf{w}; \mathbf{m}, \mathbf{S}\right) && (48)\\ \mathbf{m} &= \frac{1}{\sigma_e^2}\mathbf{S}\boldsymbol{\Phi}(\mathbf{x})^{\mathrm{T}}\mathbf{y} && (49)\\ \mathbf{S} &= \left(\frac{1}{\sigma_w^2}\mathbf{I} + \frac{1}{\sigma_e^2}\boldsymbol{\Phi}(\mathbf{x})^{\mathrm{T}}\boldsymbol{\Phi}(\mathbf{x})\right)^{-1} && (50) \end{aligned}$$

using for the transformation from $p(\mathbf{u}) \leftarrow p(\mathbf{w})$ and $p(\mathbf{v}|\mathbf{u}) \leftarrow p(\mathbf{y}|\mathbf{w})$ to $p(\mathbf{u}|\mathbf{v}) \leftarrow p(\mathbf{w}|\mathbf{y})$

$$\begin{aligned} &\mathbf{u} = \mathbf{w}, \boldsymbol{\mu} = \mathbf{0}, \boldsymbol{\Sigma} = \sigma_w^2\mathbf{I} && (51)\\ &\mathbf{v} = \mathbf{y}, \mathbf{A} = \boldsymbol{\Phi}(\mathbf{x}), \mathbf{b} = \mathbf{0}, \mathbf{C} = \sigma_e^2\mathbf{I} && (52) \end{aligned}$$

Note: Comparing $\mathbf{m}$ to the the MAP estimate for $\mathbf{w}$ (or when using regularized least squares) we can see that the mean of the posterior distribution of $\mathbf{w}$ given $\mathbf{y}$ is in fact the map estimate for $\mathbf{w}$.

Using the the resulting posterior distribution $p(\mathbf{w}|\mathbf{y})$ and recalling the distribution $p(t|\mathbf{w})$ for t given a new value of x we finally can compute the predictive distribution $p(t|\mathbf{y})$

$$\begin{aligned} p(y|\mathbf{w}) &= \mathcal{N}\left(y; \boldsymbol{\varphi}(x)^{\mathrm{T}}\mathbf{w}, \sigma_e^2\right) && (53)\\ \Rightarrow p\left(y|x, \mathbf{x}, \mathbf{y}\right) &= \mathcal{N}\left(y; \boldsymbol{\varphi}(x)^{\mathrm{T}}\mathbf{m}, \sigma_e^2 + \boldsymbol{\varphi}(x)^{\mathrm{T}}\mathbf{S}\boldsymbol{\varphi}(x)\right) && (54) \end{aligned}$$

using for the transformation from $p(\mathbf{u}) \leftarrow p(\mathbf{y}|\mathbf{w})$ and $p(\mathbf{v}|\mathbf{u}) \leftarrow p(y|\mathbf{w})$ to $p(\mathbf{v}) \leftarrow p\left(y|x, \mathbf{x}, \mathbf{y}\right)$

$$\begin{aligned} &\mathbf{u} = \mathbf{w}, \boldsymbol{\mu} = \mathbf{m}, \boldsymbol{\Sigma} = \mathbf{S} && (55)\\ &\mathbf{v} = y, \mathbf{A} = \boldsymbol{\varphi}(x)^{\mathrm{T}}, \mathbf{b} = 0, \mathbf{C} = \sigma_e^2 && (56) \end{aligned}$$

Note: Comparing the distribution $p(y|\mathbf{w})$ and $p\left(y|x, \mathbf{x}, \mathbf{y}\right)$ we can see that both distributions are very similar where for $p\left(y|x, \mathbf{x}, \mathbf{y}\right)$ the true value for $\mathbf{w}$ is replaced by the MAP estimate $\mathbf{m}$ and the additional variance is added to the original variance. This result should be intuitive because by using an estimate for $\mathbf{w}$ we increase the variance of the resulting distribution.

Task 1.10 (Code)

Write a function to compute the mean and variance of $p\left(y|x, \mathbf{x}, \mathbf{y}\right)$ and use the plot to observe the influence of the model parameters