Multiscale Modeling of Fluid Materials
TUM School of Engineering and Design
Technical University of Munich

LECTURE NOTES

Machine Learning Essentials

TUM, WiSe 2025/26

Prof. Dr. Julija Zavadlav

Last update: October 30, 2025

Disclaimers

Supervised learning: $\mathcal{D} = \{(\vec{x}_n, \vec{y}_n)\}_{n=1}^{N}$
Regression: $\vec{y}_n$ is continuous
Linear Basis Function model: $f(\vec{x}, \vec{w}) = \vec{w}\,\vec{\phi}(\vec{x})$
Maximum Likelihood: $\vec{\theta}_{ML} = \arg\max_{\vec{\theta}} p(\mathcal{D}|\vec{\theta})$

LECTURE 3

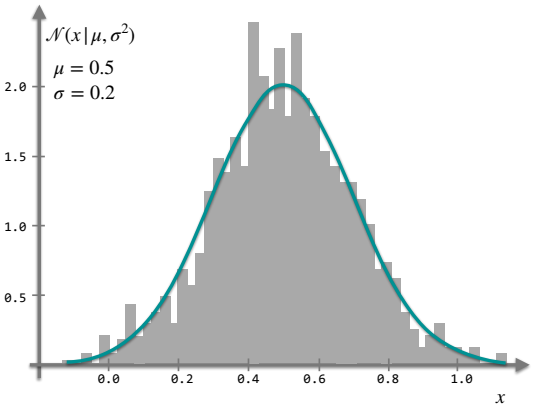
Regression Problem

In this lecture, we will look at the regression problem. We will use the three steps we have introduced last time: (1) Specification of the Model Architecture, (2) Parameter Estimation/ Training, and (3) Prediction of Novel Inputs. Our goal is to derive the loss function for a regression problem by using the maximum likelihood "measure of best" (frequentist approach) and assuming Gaussian noise. We will then see how this loss function has a closed-form solution when we further assume a linear basis function model. This solution is, in fact, equal to the Linear Least Squares - a method employed when you run a "fit" command in MATLAB or other software.

3.1 Gaussian Probability Distribution

The Gaussian distribution, also known as the normal distribution, is one of the most widely used distributions in probability theory and appears in numerous contexts. It was invented by Carl Friedrich Gauss, who also demonstrated that least-squares can be derived under the assumption of Gaussian errors (which we will show in this lecture).

Consider, for example, the problem of determining the resistance by measuring the current at different voltages. The measurement device (multimeter) has some intrinsic, random error (inherent to the device itself). Therefore, if we measure the current at the same applied voltage, we will not always obtain the same result. By recording the results in a histogram, we would obtain a distribution that tends toward a Gaussian distribution. Thus, we can often assume that the random noise in the system is Gaussian distributed.



The Gaussian distribution also arises in the central limit theorem (also known as the law of large numbers), which states that the sum of random variables has a distribution that becomes increasingly Gaussian as the number of terms in the sum increases.

Univariate Gaussian

The univariate Gaussian distribution is determined by two parameters μ and σ^2. The probability density function of the scalar variable x is given by

$$\mathcal{N}(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(x-\mu)^2\right] \tag{3.1}$$

Univariate Gaussian

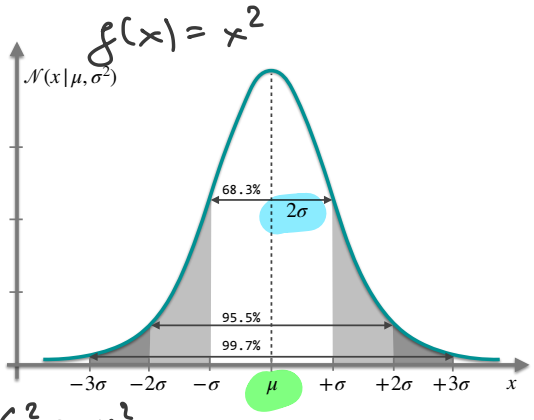
We can see that the Gaussian distribution satisfies the properties of a probability distribution function, i.e.,

$$\mathcal{N}\left(x \mid \mu, \sigma^2\right) > 0$$

$$\int_{-\infty}^{\infty} \mathcal{N}\left(x \mid \mu, \sigma^2\right) \mathrm{d}x = 1 \quad \leftarrow \text{normalization: factor } \frac{1}{\sqrt{2\pi\sigma^2}}$$

The expectation of x under the Gaussian distribution is $\mathbb{E}[f(x)] = \int f(x)\, p(x)\, dx$

$$\mathbb{E}[x] = \int_{-\infty}^{\infty} \mathcal{N}\left(x \mid \mu, \sigma^2\right) x \mathrm{d}x = \mu \tag{3.2}$$



Thus μ is the average value of x and thus referred to as the mean. Graphically, the mean corresponds to the position of the peak. If we compute the expectation of x^2, we obtain

$$\mathbb{E}\left[x^2\right] = \int_{-\infty}^{\infty} \mathcal{N}\left(x \mid \mu, \sigma^2\right) x^2 \mathrm{d}x = \mu^2 + \sigma^2. \tag{3.3}$$

Thus, the variance of x equals $\mu^2+\sigma^2 - \mu^2$

$$\mathrm{var}[x] = \mathbb{E}\left[x^2\right] - \mathbb{E}[x]^2 = \sigma^2, \tag{3.4}$$

and the σ^2 is reffered to as the variance parameter. The square root of variance σ is called the standard deviation. Graphically, the standard deviation encodes the width of the curve with full width at half maximum $\mathrm{FWHM} = 2.355\sigma$.

Multivariate Gaussian

We can generalize the distribution for a D-dimensional vector $\boldsymbol{x} \in \mathbb{R}^D$ of continous variables, i.e.,

$$\mathcal{N}(\vec{x} \mid \vec{\mu}, \Sigma) = \frac{1}{(2\pi)^{D/2}} \frac{1}{\sqrt{|\Sigma|}} \exp\left[-\frac{1}{2}(\vec{x}-\vec{\mu})^T \Sigma^{-1} (\vec{x}-\vec{\mu})\right] \tag{3.5}$$

Multivariate Gaussian

where the determinant of a matrix Σ is denoted by $|\Sigma|$, and its inverse by Σ^{-1}. $\boldsymbol{\mu} \in \mathbb{R}^D$ is a vector of means and $\Sigma \in \mathbb{R}^{DxD}$ is the covariance matrix since the expected values are now

$$\mathbb{E}[\boldsymbol{x}] = \boldsymbol{\mu} \quad \text{and} \quad \mathbb{E}\left[\boldsymbol{x}\boldsymbol{x}^T\right] = \boldsymbol{\mu}\boldsymbol{\mu}^T + \Sigma \quad \rightarrow \quad \mathrm{var}[\boldsymbol{x}] = \Sigma \tag{3.6}$$

The diagonal and off-diagonal components of the covariance matrix are

$$\Sigma_{ii} = \mathrm{var}[x_i] \quad \text{and} \quad \Sigma_{ij} = \mathrm{cov}[x_i, x_j]. \tag{3.7}$$

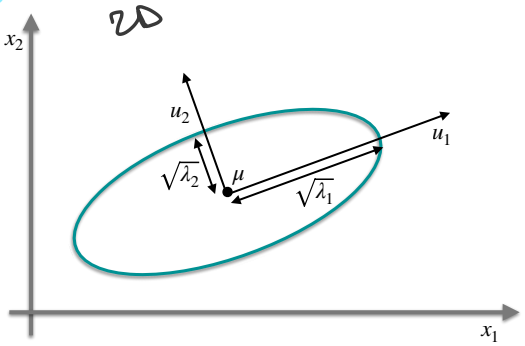
Geometric Interpretation

The Gaussian distribution is constant on surfaces for which the exponent is constant

$$\Delta^2 = (\boldsymbol{x}-\boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{x}-\boldsymbol{\mu}) \tag{3.8}$$

The quantity Δ is called the Mahalanobis distance from $\boldsymbol{x}$ to $\boldsymbol{\mu}$ and reduces to Euclidean distance when Σ is the identity matrix. To understand how Δ looks in $\boldsymbol{x}$ space, we can consider the eigenvector equation for the covariance matrix

$$\Sigma \boldsymbol{u}_i = \lambda_i \boldsymbol{u}_i \quad \text{or} \quad \Sigma U = U\Lambda, \tag{3.9}$$



where U is a matrix whose columns are eigenvectors $\boldsymbol{u}_i$ and $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_D)$. Since Σ is symmetric, the eigenvalues are real and eigenvectors can be chosen to form an orthonormal set, i.e., $U^T = U^{-1}$. Right-hand multiplication of the eigenvector equation by U^T gives

$$\Sigma U = U\Lambda \;/\; U^T$$

$$\Sigma U U^T = U\Lambda U^T$$

$$\Sigma = U\Lambda U^T = \sum_{i=1}^{D} \lambda_i \boldsymbol{u}_i \boldsymbol{u}_i^T$$

The inverse covariance matrix can be expressed as

$$\Sigma^{-1} = (U\Lambda U^T)^{-1} = (U\Lambda U^{-1})^{-1} = U\Lambda^{-1}U^{-1} = U\Lambda^{-1}U^T = \sum_{i=1}^{D} \frac{1}{\lambda_i} \boldsymbol{u}_i \boldsymbol{u}_i^T \tag{3.10}$$

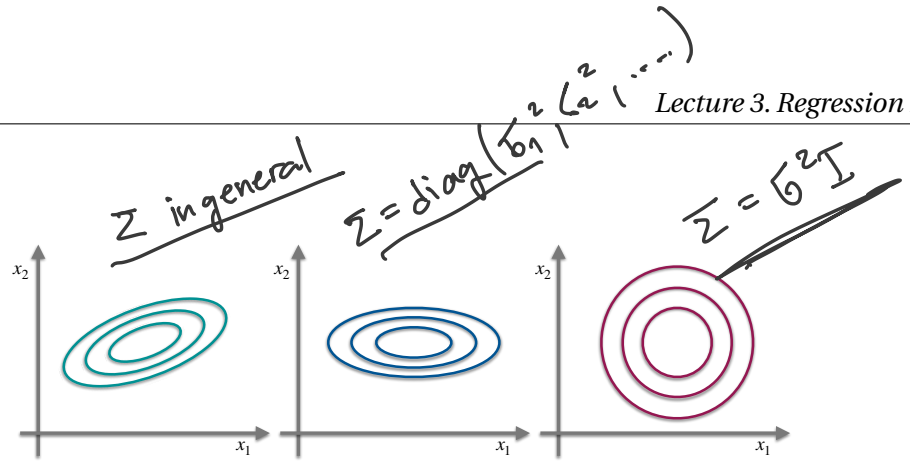
Substituting this result into the Mahalanobis distance, we obtain

$$\Delta^2 = (\boldsymbol{x}-\boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{x}-\boldsymbol{\mu})$$

$$= (\boldsymbol{x}-\boldsymbol{\mu})^T \left(\sum_{i=1}^{D} \frac{1}{\lambda_i} \boldsymbol{u}_i \boldsymbol{u}_i^T \right) (\boldsymbol{x}-\boldsymbol{\mu}) = \sum_{i=1}^{D} \frac{(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{u}_i \boldsymbol{u}_i^T (\boldsymbol{x}-\boldsymbol{\mu})}{\lambda_i}$$

$$= \sum_{i=1}^{D} \frac{y_i^2}{\lambda_i} \quad \text{where} \quad y_i = \boldsymbol{u}_i^T (\boldsymbol{x}-\boldsymbol{\mu}) \quad \text{or} \quad \boldsymbol{y} = U(\boldsymbol{x}-\boldsymbol{\mu}),$$

where $\boldsymbol{y} = (y_1, \ldots, y_D)^T$. We interpret $\{y_i\}$ as a new coordinate system defined by orthonormal vectors $\boldsymbol{u}_i$ that are shifted and rotated with respect to the original $\{x_i\}$ coordinates. The Gaussian distribution is constant on ellipsoids (distribution can only be properly normalized with positive eigenvalues) with the center at $\boldsymbol{\mu}$ and axes oriented along $\boldsymbol{u}_i$. The scaling factors in the directions of the axes are given by $\sqrt{\lambda_i}$. If the covariance matrix is diagonal, i.e., $\Sigma = \text{diag}(\sigma_1^2, \ldots, \sigma_D^2)$, then $U = I$ and constant density contours are axis-aligned ellipsoids. If the covariance matrix is isotropic, $\Sigma = \sigma^2 I$, the contours are spherical.



3.2 Frequentist Approach to Regression

Statistical Model

Suppose we have observed some data $\mathcal{D}$, which consists of N observations of the form $\{(\boldsymbol{x}_n, y_n)\}_{n=1}^{N}$. For simplicity, we have assumed a vector input $\boldsymbol{x}_n$ and a scalar output variable y_n, but the problem can be generalized for a vector output. We postulate the following statistical model

$$y = f(\vec{x}, \vec{w}) + \varepsilon \tag{3.11}$$

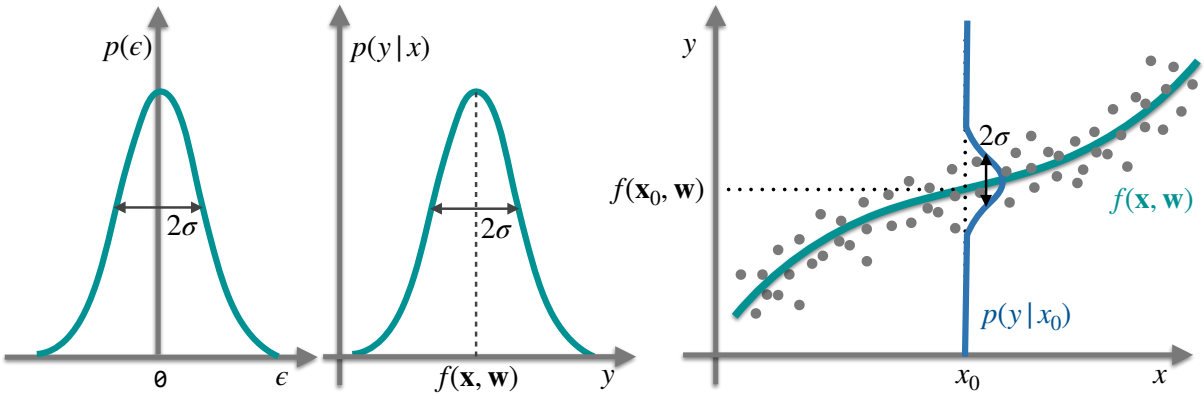
where $f(\boldsymbol{x}, \boldsymbol{w})$ is a deterministic function and ϵ is a noise/error term that describes everything that the model cannot capture. We assume that ϵ is a zero-mean Gaussian random variable with variance σ^2, i.e.,

$$p(\varepsilon) = \mathcal{N}(\varepsilon \mid 0, \sigma^2) \tag{3.12}$$

From Eqs 3.11 and 3.12, it follows that the probability distribution $p(y \mid \boldsymbol{x})$ is also a Gaussian distribution with a mean equal to $f(\boldsymbol{x}, \boldsymbol{w})$ and variance σ^2, i.e.

$$p(y \mid \vec{x}) = \mathcal{N}(y \mid f(\vec{x}, \vec{w}), \sigma^2) \tag{3.13}$$

The parameters of the statistical model are $\vec{\theta} = [\vec{w}, \sigma]$. While a Gaussian random variable is a reasonable choice in many situations, it may not be the best choice in specific settings. Examples where this assumption should be replaced include discrete data, small datasets (Student's t-distribution), and rare events, such as the time until an earthquake occurs (Exponential distribution).



Likelihood function

likelihood

Now, we would like to determine the maximum likelihood estimate $\hat{\boldsymbol{\theta}}_{\mathrm{ML}} = \operatorname{argmax}_{\boldsymbol{\theta}} p(\mathcal{D}|\boldsymbol{\theta})$ of this statistical model. We start by writing the likelihood function

$$\mathcal{D} = \{(\vec{x}_n, y_n)\}_{n=1}^N \qquad p(\mathcal{D}|\vec{\theta}) = p(\vec{y} \mid X, \vec{w}, \sigma^2) \tag{3.14}$$

where we have grouped together N observations into a column vector of inputs $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N]^T$ and the corresponding outputs $\boldsymbol{y} = [y_1, \ldots, y_N]^T$.

$p(A,B) = p(A)p(B)$

We assume that these N observations are independent and identically distributed (i.i.d.). Each observation is distributed according to Eq. (3.13). Since the joint probability of independent events is equal to the product of individual events, we can write the likelihood function as a product of Gaussian functions

$$p(\mathcal{D}|\vec{\theta}) = \prod_{n=1}^N \mathcal{N}(y_n \mid f(\vec{x}_n, \vec{w}), \sigma^2) \tag{3.15}$$

We now switch to the log-likelihood function, as applying the logarithm does not change the location of the maximum (although the value at the maximum will be different). Employing the logarithm product and power rules, we obtain

log-trick!

$\ln(ab) = \ln a + \ln b$

$\ln a^b = b \ln a$

$$\begin{aligned}
\ln p(\mathcal{D}|\vec{\theta}) &= \ln \prod_{n=1}^N \mathcal{N}(y_n \mid f(\vec{x}_n, \vec{w}), \sigma^2) \\
&= \sum_{n=1}^N \ln \mathcal{N}(y_n \mid f(\vec{x}_n, \vec{w}), \sigma^2) \\
&= \sum_{n=1}^N \ln \left\{ \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2} (y_n - f(\vec{x}_n, \vec{w}))^2 \right] \right\} \\
&= \ln \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)^N - \frac{1}{2\sigma^2} \sum_{n=1}^N (y_n - f(\vec{x}_n, \vec{w}))^2 \\
&= \frac{N}{2} \ln(2\pi) - N \ln \sigma - \frac{1}{2\sigma^2} \sum_{n=1}^N (y_n - f(\vec{x}_n, \vec{w}))^2
\end{aligned} \tag{3.16}$$

To find the $\hat{\boldsymbol{w}}_{\mathrm{ML}}$, we need to determine the maximum of the log-likelihood function, or equivalently, the minimum of the negative log-likelihood (loss) function.

$$\hat{\boldsymbol{w}}_{ML} = \operatorname*{argmin}_{\boldsymbol{w}} \left[\frac{N}{2} \ln(2\pi) + N \ln \sigma + \underbrace{\frac{1}{2\sigma^2} \sum_{n=1}^N [y_n - f(\boldsymbol{x}_n, \boldsymbol{w})]^2}_{L(\vec{w})} \right] = \operatorname*{argmin}_{\boldsymbol{\theta}} L(\boldsymbol{w}), \tag{3.17}$$

We see that only the last term depends on $\boldsymbol{w}$, thus we can disregard the other two terms. The prefactor $\frac{1}{2\sigma^2}$ will not change the position of the maximum and can be skipped or replaced by $1/2$. With these considerations in mind, we end up with the loss function called the sum-of-squares error function

$$L(\vec{w}) = \sum_{n=1}^{N} \left[y_n - f(\vec{x}_n, \vec{w}) \right]^2 \tag{3.18}$$

Sum-of-Squares Error

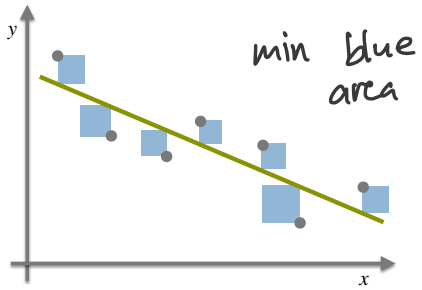
which we need to minimize. We could also normalize the loss by the number of data points, which would again not change the position of the minimum. In this case, we obtain the so-called Mean Squared Error (MSE)

$$L(\vec{w}) = \frac{1}{N} \sum_{n=1}^{N} \left[y_n - f(\vec{x}_n, \vec{w}) \right]^2 \tag{3.19}$$

Mean Squared Error (MSE)

Note that we discuss the loss function here because it is used to optimize the parameters. We could also use the same functional form to evaluate the model, and then we would talk about error functions.

The ~~Least Squares~~ regression method, therefore, is minimizing the squared L2-norm ($\|\boldsymbol{x}\|_2^2 = \sum_i x_i^2$) of the squared distance between the observed value y_n and the output of the deterministic function $f(\boldsymbol{x}_n, \boldsymbol{w})$ (visualized as squares). The optimal parameters are those for which the amount of blue color is minimized.



We can also determine the variance σ^2 parameter of the statistical model. Differentiating Eq. (3.16) with respect to σ and equating to zero we obtain

$$0 = -\frac{N}{\sigma} + \frac{1}{\sigma^3} \sum_{n=1}^{N} \left[y_n - f(\boldsymbol{x}_n, \boldsymbol{w}) \right]^2$$

$$\hat{\sigma}^2_{\mathrm{ML}} = \frac{1}{N} \sum_{n=1}^{N} \left[y_n - f(\boldsymbol{x}_n, \hat{\boldsymbol{w}}_{\mathrm{ML}}) \right]^2 .$$

Prediction of Novel Inputs

Once we have determined the parameters of the statistical model $\hat{\boldsymbol{\theta}}_{\mathrm{ML}} = [\hat{\boldsymbol{w}}_{\mathrm{ML}}, \hat{\sigma}_{\mathrm{ML}}]^T$ we make a prediction for a new input $\boldsymbol{x}_*$ with

$$y_* = \mathcal{N}\left(y \,|\, f(\boldsymbol{x}_*, \hat{\boldsymbol{w}}_{\mathrm{ML}}), \hat{\sigma}^2_{\mathrm{ML}} \right). \tag{3.20}$$

Note that the prediction is now a probability distribution.

Demo: Least Squares & Outliers

To better understand the Least Squares method, see demo at https://mless.pythonanywhere.com/leastsquares where we investigate the impact of adding an outlier to the training data set. Given a training dataset $\{(x_n, y_n)\}_{n=1}^{N=10}$ we aim to find $f(x, \boldsymbol{w}) = w_0 + w_1 x + w_2 x^2 + w_3 x^3$ which is an approximation of a true function $g(x) = \sin(2\pi x)$. Data points are drawn from $y = \sin(2\pi x) + \mathcal{N}\left(0, \sigma^2 = 0.01\right)$ while the outlier data point is drawn from $y = \sin(2\pi x) + \mathcal{N}(0, \sigma^2 = 0.25)$. Answer the following questions:

- How does the addition of an outlier affect the Least Squares fit? Is (or when is) the Least Squares method sensitive to outliers?
  the result is different, critical for small datasets
- How can we modify the "Measure of Best" to obtain a method that is less sensitive to outliers?
  $L(\vec{w}) = \sum \| y_n - f(x_n, w) \|$ L_1-norm
- In what situation could we have an outlier in the dataset?
  sim / experiment gone wrong,
- How can we avoid outliers in the dataset?
  try to visualise! look at data points with large errors

Summary

- assumptions: gaussian noise, i.i.d data
- we did not say anything about $f(\vec{x}, \vec{w})$ → a NN
- If $f(\vec{x}, \vec{w})$ is LBFM → closed-form solution = linear LS method

3.3 Linear Least Squares

If we assume a linear basis function model for the deterministic function $f(\boldsymbol{x}, \boldsymbol{w}) = \boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x})$, we can find a closed-form solution for the regression problem. For each data sample in the dataset $\mathcal{D} = \{(\boldsymbol{x}_n, y_n)\}_{n=1}^{N}$, we have

$$\begin{aligned} y_n &= f(\boldsymbol{x}_n, \boldsymbol{w}) + \epsilon_n \quad \forall n = 1, \ldots, N \\ &= \boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x}_n) + \epsilon_n \\ &= w_0 + \sum_{m=1}^{M-1} w_m \phi_m(\boldsymbol{x}_n) + \epsilon_n, \end{aligned}$$

N equations

where $\boldsymbol{w}$ is an M-dimensional vector.

Matrix Formulation

We can rewrite the above N equations jointly using a matrix formulation

$$\boldsymbol{y} = \Phi \boldsymbol{w} + \boldsymbol{\epsilon} \tag{3.21}$$

or

$$\underbrace{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ \vdots \\ y_N \end{bmatrix}}_{\boldsymbol{y}\in\mathbb{R}^N} = \underbrace{\begin{bmatrix} \phi_0(\boldsymbol{x}_1) & \phi_1(\boldsymbol{x}_1) & \dots & \phi_{M-1}(\boldsymbol{x}_1) \\ \phi_0(\boldsymbol{x}_2) & \phi_1(\boldsymbol{x}_2) & \dots & \phi_{M-1}(\boldsymbol{x}_2) \\ \vdots & & & \vdots \\ \vdots & & & \vdots \\ \phi_0(\boldsymbol{x}_N) & \phi_1(\boldsymbol{x}_N) & \dots & \phi_{M-1}(\boldsymbol{x}_N) \end{bmatrix}}_{\Phi\in\mathbb{R}^{N\times M}} \underbrace{\begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_{M-1} \end{bmatrix}}_{\boldsymbol{w}\in\mathbb{R}^M} + \underbrace{\begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \vdots \\ \epsilon_N \end{bmatrix}}_{\boldsymbol{\epsilon}\in\mathbb{R}^N}, \tag{3.22}$$

where Φ is called a design matrix. Here, each row corresponds to one data sample.

Example: Matrix Formulation for Quadratic Polynomial

Let us consider a polynomial basis set, i.e., $\phi_m(x) = x^m$ and $M = 3$. Furthermore, we consider the following dataset:

	x_n	y_n
n=1	1	2
n=2	2	9

$$\begin{bmatrix} 2 \\ 9 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 4 \end{bmatrix} \begin{bmatrix} w_0 \\ w_1 \\ w_2 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \end{bmatrix}$$

Closed-form Solution

We want to find the Maximum Likelihood Estimate for this problem. We saw that assuming Gaussian noise and independent and identically distributed (i.i.d.) observations boils down to minimizing the

sum-of-squares error function, i.e.

$$\begin{aligned} E(\boldsymbol{w}) &= \sum_{n=1}^{N} [y_n - f(\boldsymbol{x}_n, \boldsymbol{w})]^2 \\ &= \sum_{n=1}^{N} [y_n - \boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x}_n)]^2 \\ &= (\boldsymbol{y} - \Phi \boldsymbol{w})^T (\boldsymbol{y} - \Phi \boldsymbol{w}) \\ &= (\boldsymbol{y}^T - \boldsymbol{w}^T \Phi^T)(\boldsymbol{y} - \Phi \boldsymbol{w}) \\ &= \left[\boldsymbol{y}^T \boldsymbol{y} - \boldsymbol{y}^T \Phi \boldsymbol{w} - \boldsymbol{w}^T \Phi^T \boldsymbol{y} + \boldsymbol{w}^T \Phi^T \Phi \boldsymbol{w} \right] \\ &= \left[\boldsymbol{y}^T \boldsymbol{y} - 2\boldsymbol{w}^T \Phi^T \boldsymbol{y} + \boldsymbol{w}^T \Phi^T \Phi \boldsymbol{w} \right]. \end{aligned} \tag{3.23}$$

In the last line we have used the fact that $\boldsymbol{y}^T \Phi \boldsymbol{w}$ is a $(1 \times N)(N \times M)(M \times 1)$ so (1×1) matrix which is always symmetric. To obtain a minimum, we compute the derivative with respect to $\boldsymbol{w}$

$$\begin{aligned} \frac{\partial}{\partial \boldsymbol{w}} \left[\boldsymbol{y}^T \boldsymbol{y} - 2\boldsymbol{w}^T \Phi^T \boldsymbol{y} + \boldsymbol{w}^T \Phi^T \Phi \boldsymbol{w} \right] &= 0 \\ -2\Phi^T \boldsymbol{y} + \left[\Phi^T \Phi + (\Phi^T \Phi)^T \right] \boldsymbol{w} &= 0 \\ -\Phi^T \boldsymbol{y} + \Phi^T \Phi \boldsymbol{w} &= 0, \end{aligned}$$

where we have used the matrix differentiation rules $\frac{\partial \boldsymbol{x}^T \boldsymbol{a}}{\partial \boldsymbol{x}} = \boldsymbol{a}$ and $\frac{\partial \boldsymbol{x}^T A \boldsymbol{x}}{\partial \boldsymbol{x}} = (A + A^T)\boldsymbol{x}$. We obtain the so-called normal equations

$$\Phi^T \Phi \boldsymbol{w} = \Phi^T \boldsymbol{y}. \tag{3.24}$$

If the basis functions ϕ are linearly independent $\Phi^T \Phi$ is a symmetric, positive definite matrix[1]. Thus, it can be inverted, and we can write the least squares solution $\hat{\boldsymbol{w}}$ as

$$\hat{\boldsymbol{w}}_{ML} = (\Phi^T \Phi)^{-1} \Phi^T \boldsymbol{y} \tag{3.25}$$

Least Squares Solution

Is the obtained solution truly a minimum? Mathematically, we check if an extremum is a minimum by checking if the second derivative is positive. To see this, we write the Taylor series expansion around $\hat{\boldsymbol{w}}$:

$$\begin{aligned} E(\hat{\boldsymbol{w}} + \tilde{\boldsymbol{w}}) &= E(\hat{\boldsymbol{w}}) + \tilde{\boldsymbol{w}}^T \frac{\partial E(\hat{\boldsymbol{w}})}{\partial \boldsymbol{w}} + \frac{1}{2} \tilde{\boldsymbol{w}}^T \frac{\partial^2 E(\hat{\boldsymbol{w}})}{\partial^2 \boldsymbol{w}} \tilde{\boldsymbol{w}} \\ &= E(\hat{\boldsymbol{w}}) + \frac{1}{2} \tilde{\boldsymbol{w}}^T \Phi^T \Phi \tilde{\boldsymbol{w}} > E(\hat{\boldsymbol{w}}). \end{aligned}$$

We used the fact that $\frac{\partial E(\hat{\boldsymbol{w}})}{\partial \boldsymbol{w}} = 0$ and $\frac{\partial^2 E(\hat{\boldsymbol{w}})}{\partial^2 \boldsymbol{w}} = \Phi^T \Phi$. We conclude that the least square solution $\hat{\boldsymbol{w}}$ is a minimum since adding a small vector $\tilde{\boldsymbol{w}}$ will increase E.

[1] Symmetry is straight forward $(\Phi^T \Phi)^T = \Phi^T (\Phi^T)^T = \Phi^T \Phi$. Recall that a matrix B is positive definite if for any nonzero vector $\boldsymbol{y}$: $\boldsymbol{y}^T B \boldsymbol{y} > 0$. To see that this is true for matrix $\Phi^T \Phi$ we write: $\boldsymbol{y}^T \Phi^T \Phi \boldsymbol{y} = (\Phi \boldsymbol{y})^T (\Phi \boldsymbol{y}) = \|\Phi \boldsymbol{y}\|_2^2 > 0$.