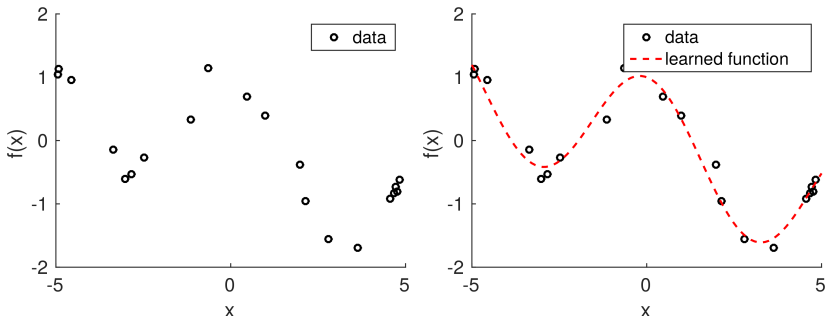
9

Linear Regression

In the following, we will apply the mathematical concepts from Chapters 2, 6,5, 7 to solving linear regression (curve fitting) problems. In regression, we want to find a function f that maps inputs $\boldsymbol{x} \in \mathbb{R}^D$ to corresponding function values $f(\boldsymbol{x}) \in \mathbb{R}$ based on a set of training inputs $\boldsymbol{x}_i$ and corresponding noisy observations $y_i = f(\boldsymbol{x}_i) + \epsilon$, where ϵ is a random variable. An illustration of such a regression problem is given in Figure 9.1. A typical regression problem is given in Figure 9.1(a): For some input values x we observe (noisy) function values $y = f(x) + \epsilon$. The task is to infer the function f that generated the data. A possible solution is given in Figure 9.1(b).

Regression is a fundamental problem in machine learning, and regression problems appear in a diverse range of research areas and applications, including time-series analysis (e.g., system identification), control and robotics (e.g., reinforcement learning, forward/inverse model learning), optimization (e.g., line searches, global optimization), and deep-learning applications (e.g., computer games, speech-to-text translation, image recognition, automatic video annotation). Regression is also a key ingredient of classification algorithms.

Finding a regression function requires solving a variety of problems, including



(a) Regression problem: Observed noisy function values from which we wish to infer the underlying function that generated the data.

(b) Regression solution: Possible function that could have generated the data.

Figure 9.1 (a) Regression problem and (b) possible solution.

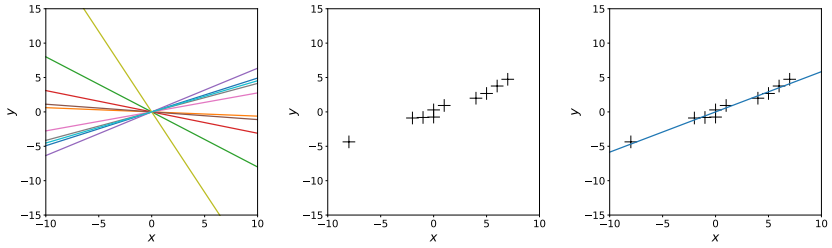
- **Choice of the model (type) and the parametrization** of the regression function. Given a data set, what function classes (e.g., polynomials) are good candidates for modeling the data, and what particular parametrization (e.g., degree of the polynomial) should we choose? Model selection, as discussed in Section 8.1, allows us to compare various models to find the simplest model that explains the training data reasonably well.
- **Finding good parameters.** Having chosen a model of the regression function, how do we find good model parameters? Here, we will need to look at different loss/objective functions (they determine what a "good" fit is) and optimization algorithms that allow us to minimize this loss.
- **Overfitting and model selection.** Overfitting is a problem when the regression function fits the training data "too well" but does not generalize to unseen test data. Overfitting typically occurs if the underlying model (or its parametrization) is overly flexible and expressive, see Section 8.1. We will look at the underlying reasons and discuss ways to mitigate the effect of overfitting.
- **Relationship between loss functions and parameter priors.** Loss functions (optimization objectives) are often motivated and induced by probabilistic models. We will look at a the connection between loss functions and the underlying prior assumptions that induce these losses.
- **Uncertainty modeling.** In any practical setting, we have access to only a finite, potentially large, amount of (training) data for selecting the model class and the corresponding parameters. Given that this finite amount of training data does not cover all possible scenarios, we way want to describe the remaining parameter uncertainty to obtain a measure of confidence of the model's prediction at test time; the smaller the training set the more important uncertainty modeling. Consistent modeling of uncertainty equips model predictions with confidence bounds.

In the following, we will be using the mathematical tools from Chapters 5, 6 and 7 to solve linear regression problems. We will discuss maximum likelihood and maximum a posteriori (MAP) estimation to find optimal model parameters. Using these parameter estimates, we will have a brief look at generalization errors and overfitting. Toward the end of this chapter, we will discuss Bayesian linear regression, which allows us to reason about model parameters at a higher level, thereby removing some of the problems encountered in maximum likelihood and MAP estimation.

9.1 Problem Formulation

We consider the regression problem

$$y = f(\boldsymbol{x}) + \epsilon\,, \tag{9.1}$$



(a) Example functions (straight lines) that can be described using the linear model in (9.2).

(b) Training set.

(c) Maximum likelihood estimate

Figure 9.2 Linear regression without features. (a) Examples functions that fall into this category. (b) Training set. (c) Maximum likelihood estimate.

where $\boldsymbol{x} \in \mathbb{R}^D$ are inputs and $y \in \mathbb{R}$ are observed function values (targets). Furthermore, $\epsilon \sim \mathcal{N}\left(0, \sigma^2\right)$ is independent, identically distributed (i.i.d.) measurement noise. In this particular case, ϵ is Gaussian distributed with mean 0 and variance σ^2. Our objective is to find a function that is close (similar) to the unknown function that generated the data.

In this chapter, we focus on parametric models, i.e., we choose a parametrized function f and find parameters that "work well" for modeling the data. In linear regression, we consider the special case that the parameters appear linearly in our model. An example of linear regression is

$$y = f(\boldsymbol{x}) + \epsilon = \boldsymbol{x}^\top \boldsymbol{\theta} + \epsilon\,, \tag{9.2}$$

where $\boldsymbol{\theta} \in \mathbb{R}^D$ are the parameters we seek, and $\epsilon \sim \mathcal{N}\left(0, \sigma^2\right)$ is i.i.d. Gaussian measurement/observation noise. In (9.2) we chose a parametrization $f(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{\theta}$. For the time being we assume that the noise variance σ^2 is known. The noise model induces the *likelihood*

likelihood

$$p(y|\boldsymbol{x}, \boldsymbol{\theta}) = \mathcal{N}\left(y \,|\, \boldsymbol{x}^\top \boldsymbol{\theta},\, \sigma^2\right)\,, \tag{9.3}$$

which is the probability of observing a target value y given that we know the input location $\boldsymbol{x}$ and the parameters $\boldsymbol{\theta}$. Note that the only source of uncertainty originates from the observation noise (as $\boldsymbol{x}$ and $\boldsymbol{\theta}$ are assumed known in (9.3))—without any observation noise, the relationship between $\boldsymbol{x}$ and y would be deterministic and (9.3) would be a delta function.

For $x, \theta \in \mathbb{R}$ the linear regression model in (9.2) can describe straight lines (linear functions), and the parameter θ would be the slope of the line. Figure 9.2(a) shows some examples. This model is not only linear in the parameters, but also linear in the inputs x. We will see later that $y = \phi(x)\theta$ for nonlinear transformations ϕ is also a linear regression model because "linear regression" refers to models that are "linear in the parameters", i.e., models that describe a function by a linear combination of input features.

Linear regression refers to models that are linear in the parameters.

In the following, we will discuss in more detail how to find good parameters $\boldsymbol{\theta}$ and how to evaluate whether a parameter set "works well".

9.2 Parameter Estimation

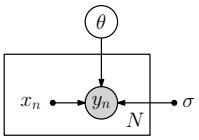
Consider the linear regression setting (9.2) and assume we are given a *training set* $\mathcal{D}$ consisting of N inputs $\boldsymbol{x}_n \in \mathbb{R}^D$ and corresponding observations/targets $y_n \in \mathbb{R}$, $n = 1, \dots, N$. The corresponding graphical model is given in Figure 9.3. Note that y_i and y_j are conditionally independent given their respective inputs $\boldsymbol{x}_i, \boldsymbol{x}_j$, such that the likelihood function factorizes according to

$$p(y_1, \dots, y_N | \boldsymbol{x}_1, \dots, \boldsymbol{x}_N) = \prod_{n=1}^{N} p(y_n | \boldsymbol{x}_n) = \prod_{n=1}^{N} \mathcal{N}\left(y_n \,|\, \boldsymbol{x}_n^\top \boldsymbol{\theta},\, \sigma^2\right). \quad (9.4)$$

The likelihood and the factors $p(y_n|\boldsymbol{x}_n)$ are Gaussian due to the noise distribution.

training set

Figure 9.3 Probabilistic graphical model for linear regression. Observed random variables are shaded, deterministic/known values are without circles. The parameters $\boldsymbol{\theta}$ are treated as unknown/latent quantities.



In the following, we are interested in finding optimal parameters $\boldsymbol{\theta}^* \in \mathbb{R}^D$ for the linear regression model (9.2). Once the parameters are found, we can predict function values by using the estimate $\boldsymbol{\theta}^*$ in the (9.2), such that at an arbitrary test input $\boldsymbol{x}_*$ we predict

$$p(y_*|\boldsymbol{x}_*, \boldsymbol{\theta}^*) = \mathcal{N}\left(y_* \,|\, \boldsymbol{x}_*^\top \boldsymbol{\theta}^*,\, \sigma^2\right). \quad (9.5)$$

9.2.1 Maximum Likelihood Estimation

maximum likelihood estimation

A widely used approach to finding the desired parameters $\boldsymbol{\theta}_{\text{ML}}$ is *maximum likelihood estimation* where we find parameters $\boldsymbol{\theta}_{\text{ML}}$ that maximize the likelihood (9.4). We obtain the maximum likelihood parameters as

$$\boldsymbol{\theta}_{\text{ML}} = \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\theta}), \quad (9.6)$$

where we define $\boldsymbol{X} := [\boldsymbol{x}_1 | \cdots | \boldsymbol{x}_N]^\top \in \mathbb{R}^{N \times D}$ and $\boldsymbol{y} := [y_1, \dots, y_N]^\top \in \mathbb{R}^N$ as the collections of training inputs and targets, respectively.

Remark. Note that the likelihood is not a probability distribution in $\boldsymbol{\theta}$: It is simply a function of the parameters $\boldsymbol{\theta}$ but does usually not integrate to 1 (i.e., it is unnormalized), and may not even be integrable with respect to $\boldsymbol{\theta}$. However, the likelihood in (9.6) is a normalized probability distribution in $\boldsymbol{y}$. ♢

To find the desired parameters $\boldsymbol{\theta}_{\text{ML}}$ that maximize the likelihood, we typically perform gradient ascent (or gradient descent on the negative likelihood). In the case of linear regression we consider here, however, a closed-form solution exists, which makes iterative gradient descent unnecessary. In practice, instead of maximizing the likelihood directly, we apply the log-transformation to the likelihood function and minimize the negative log-likelihood.

Since the logarithm is a (strictly) monotonically increasing function, the optimum of a function f is identical to the optimum of $\log f$.

Remark (Log Transformation). Since the likelihood function is a product of N Gaussian distributions, the log-transformation is useful since a) it does not suffer from numerical underflow, b) the differentiation rules will

turn out simpler. Numerical underflow will be a problem when we multiply N probabilities, where N is the number of data points, since we cannot represent very small numbers, such as 10^{-256}. Furthermore, the log-transform will turn the product into a sum of log-probabilities, such that the corresponding gradient is a sum of individual gradients, instead of a repeated application of the product rule (5.54) to compute the gradient of a product of N terms. ♢

To find the optimal parameters $\boldsymbol{\theta}_{\text{ML}}$ of our linear regression problem, we minimize the negative log-likelihood

$$\theta_{\text{ML}} \in \arg\min_{\boldsymbol{\theta}} \big(-\log p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\theta}) \big) \tag{9.7}$$

$$= \arg\min_{\boldsymbol{\theta}} \left(-\log \prod_{i=1}^{N} p(y_i|\boldsymbol{x}_i, \boldsymbol{\theta}) \right) \tag{9.8}$$

$$= \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{N} -\log p(y_i|\boldsymbol{x}_i, \boldsymbol{\theta}) \,, \tag{9.9}$$

where we exploited that the likelihood (9.4) factorizes over the number of data points due to our independence assumption on the training set.

In the linear regression model (9.2) the likelihood is Gaussian (due to the Gaussian additive noise term), such that we arrive at

$$-\log p(y_i|\boldsymbol{x}_i, \boldsymbol{\theta}) = \frac{1}{2\sigma^2}(y_i - \boldsymbol{x}_i^\top \boldsymbol{\theta})^2 + \text{const} \tag{9.10}$$

where the constant includes all terms independent of $\boldsymbol{\theta}$. Using (9.10) in the negative log-likelihood (9.9) we obtain (ignoring the constant terms)

The negative log-likelihood function is also called *error function*.

$$\mathcal{L}(\boldsymbol{\theta}) := -\log p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\theta}) = \frac{1}{2\sigma^2} \sum_{i=1}^{N} (y_i - \boldsymbol{x}_i^\top \boldsymbol{\theta})^2 \tag{9.11}$$

$$= \frac{1}{2\sigma^2} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta})^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}) = \frac{1}{2\sigma^2} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\|^2 \,, \tag{9.12}$$

where $\boldsymbol{X} = [\boldsymbol{x}_1 | \cdots | \boldsymbol{x}_N]^\top \in \mathbb{R}^{N \times D}$ is called the *design matrix*. In (9.12) we replaced the sum of squared errors between the observations y_i and the corresponding model prediction $\boldsymbol{x}_i^\top \boldsymbol{\theta}$ with the squared norm of the difference term $\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}$.

design matrix

There is some notation overloading: We often summarize the set of training inputs in $\boldsymbol{X}$, whereas in the design matrix we additionally assume a specific "shape".

With (9.12) we have now a concrete form of the negative log-likelihood function we need to optimize. We immediately see that (9.12) is quadratic in $\boldsymbol{\theta}$. This means that we can find a unique global solution $\boldsymbol{\theta}_{\text{ML}}$ for minimizing the negative log-likelihood $\mathcal{L}$. We can find the global optimum by computing the gradient of $\mathcal{L}$, setting it to $\mathbf{0}$ and solving for $\boldsymbol{\theta}$.

Using the results from Chapter 5, we compute the gradient of $\mathcal{L}$ with respect to the parameters as

$$\frac{d\mathcal{L}}{d\boldsymbol{\theta}} = \frac{d}{d\boldsymbol{\theta}} \left(\frac{1}{2\sigma^2} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta})^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}) \right) \tag{9.13}$$

Remember that $\|\boldsymbol{x}\|^2 := \boldsymbol{x}^\top \boldsymbol{x}$ if we choose the dot product as the inner product.

$$= \frac{1}{2\sigma^2}\frac{d}{d\boldsymbol{\theta}}\left(\boldsymbol{y}^\top\boldsymbol{y} - 2\boldsymbol{y}^\top\boldsymbol{X}\boldsymbol{\theta} + \boldsymbol{\theta}^\top\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\theta}\right) \tag{9.14}$$

$$= \frac{1}{\sigma^2}(-\boldsymbol{y}^\top\boldsymbol{X} + \boldsymbol{\theta}^\top\boldsymbol{X}^\top\boldsymbol{X}) \in \mathbb{R}^{1\times D}\,. \tag{9.15}$$

As a necessary condition for $\boldsymbol{\theta}$ being optimal we set this gradient to $\mathbf{0}$ and obtain

$$\frac{\partial\mathcal{L}}{\partial\boldsymbol{\theta}} = \mathbf{0} \overset{(9.15)}{\iff} (\boldsymbol{\theta}_{\text{ML}})^\top\boldsymbol{X}^\top\boldsymbol{X} = \boldsymbol{y}^\top\boldsymbol{X} \tag{9.16}$$

$$\iff \boldsymbol{\theta}_{\text{ML}}^\top = \boldsymbol{y}^\top\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1} \tag{9.17}$$

$$\iff \boldsymbol{\theta}_{\text{ML}} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y} \tag{9.18}$$

We could right-multiply the first equation by $(\boldsymbol{X}^\top\boldsymbol{X})^{-1}$ because $\boldsymbol{X}^\top\boldsymbol{X}$ is positive definite (if we do not have two identical inputs $\boldsymbol{x}_i, \boldsymbol{x}_j$ for $i \neq j$).

Remark. In this case, setting the gradient to $\mathbf{0}$ is a necessary and sufficient condition and we obtain a global minimum since the Hessian $\nabla^2_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta}) = \boldsymbol{X}^\top\boldsymbol{X} \in \mathbb{R}^{D\times D}$ is positive definite. ♢

Example (Fitting Lines)

Let us have a look at Figure 9.2, where we aim to fit a straight line $f(x) = \theta x$, where θ is an unknown slope, to a data set using maximum likelihood estimation. Examples of functions in this model class (straight lines) are shown in Figure 9.2(a). For the data set shown in Figure 9.2(b) we find the maximum likelihood estimate of the slope parameter θ using (9.18) and obtain the maximum likelihood linear function in Figure 9.2(c).

Maximum Likelihood Estimation with Features

Thus far, we considered the linear regression setting described in (9.2), which allowed us to fit straight lines to data using maximum likelihood estimation. However, straight lines are not particularly expressive when it comes to fitting more interesting data. Fortunately, linear regression offers us a way to fit nonlinear functions within the linear regression framework: Since "linear regression" only refers to "linear in the parameters", we are free to perform an arbitrary nonlinear transformation $\phi(\boldsymbol{x})$ of the inputs $\boldsymbol{x}$ and then linearly combine them. The parameters, which we collect in the vector $\boldsymbol{\theta}$, still appear only linearly.

The corresponding linear regression model is given as

$$y = \boldsymbol{\phi}^\top(\boldsymbol{x})\boldsymbol{\theta} + \epsilon = \sum_{k=1}^{K}\theta_k\,,\phi_k(\boldsymbol{x}) + \epsilon \tag{9.19}$$

where $\boldsymbol{\phi} : \mathbb{R}^D \to \mathbb{R}^K$ is a (nonlinear) transformation of the inputs $\boldsymbol{x}$ and ϕ_k is the kth component of the *feature vector* $\boldsymbol{\phi}$.

feature vector

Example (Polynomial Regression)
We are concerned with a regression problems $y = \boldsymbol{\phi}^\top(x)\boldsymbol{\theta} + \epsilon$, where $x \in \mathbb{R}$ and $\boldsymbol{\theta} \in \mathbb{R}^K$. A transformation that is often used in this context is

$$\boldsymbol{\phi}(x) = \begin{bmatrix} \phi_0(x) \\ \phi_1(x) \\ \vdots \\ \phi_{K-1}(x) \end{bmatrix} = \begin{bmatrix} 1 \\ x \\ x^2 \\ x^3 \\ \vdots \\ x^{K-1} \end{bmatrix} \in \mathbb{R}^K . \tag{9.20}$$

This means, we "lift" the original one-dimensional input space into a K-dimensional feature space consisting of all monomials $k = 0, \ldots, K-1$. With these features, we can model polynomials of degree $\leqslant K-1$ within the framework of linear regression: A polynomial of degree $K-1$ is given by

$$f(x) = \sum_{k=0}^{K-1} \theta_k x^k = \boldsymbol{\phi}^\top(x)\boldsymbol{\theta} \tag{9.21}$$

where $\boldsymbol{\phi}$ is defined in (9.20) and $\boldsymbol{\theta} = [\theta_0, \ldots, \theta_{K-1}]^\top \in \mathbb{R}^K$ contains the (linear) parameters θ_k.

Let us now have a look at maximum likelihood estimation of the parameters $\boldsymbol{\theta}$ in the linear regression model (9.19). We consider training inputs $\boldsymbol{x}_i \in \mathbb{R}^D$ and targets $y_i \in \mathbb{R}$, $i = 1, \ldots, N$, and define the *feature matrix* (*design matrix*) as

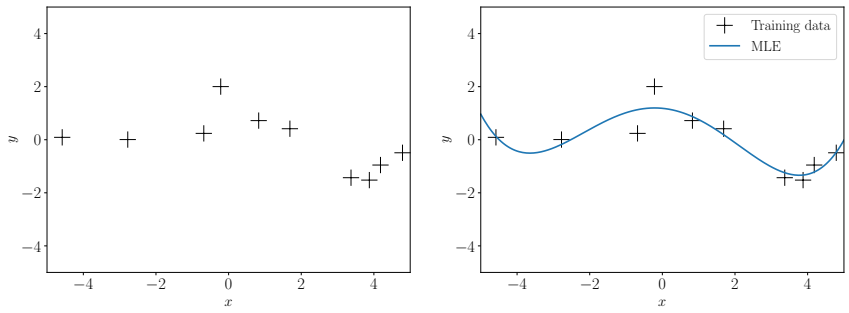
feature matrix
design matrix

$$\boldsymbol{\Phi} := \begin{bmatrix} \boldsymbol{\phi}^\top(\boldsymbol{x}_1) \\ \vdots \\ \boldsymbol{\phi}^\top(\boldsymbol{x}_N) \end{bmatrix} = \begin{bmatrix} \phi_1(\boldsymbol{x}_1) & \cdots & \phi_K(\boldsymbol{x}_1) \\ \phi_1(\boldsymbol{x}_2) & \cdots & \phi_K(\boldsymbol{x}_2) \\ \vdots & & \vdots \\ \phi_1(\boldsymbol{x}_N) & \cdots & \phi_K(\boldsymbol{x}_N) \end{bmatrix} \in \mathbb{R}^{N \times K} , \tag{9.22}$$

where $\Phi_{ij} = \phi_j(\boldsymbol{x}_i)$ and $\phi_j : \mathbb{R}^D \to \mathbb{R}$.

Example (Feature Matrix for Second-order Polynomials)
For a second-order polynomial and N training points $x_i \in \mathbb{R}, i = 1, \ldots, N$, the feature matrix is

$$\boldsymbol{\Phi} = \begin{bmatrix} 0 & x_1 & x_1^2 \\ 0 & x_2 & x_2^2 \\ \vdots & \vdots & \vdots \\ 0 & x_N & x_N^2 \end{bmatrix} . \tag{9.23}$$



(a) Regression data set.

(b) Polynomial of degree 4 determined by maximum likelihood estimation.

Figure 9.4 Polynomial regression. ((a)) Data set consisting of (x_i, y_i) pairs, $i = 1, \dots, 20$. ((b)) Maximum likelihood polynomial of degree 4.

With the feature matrix $\boldsymbol{\Phi}$ defined in (9.22) the negative log-likelihood for the linear regression model (9.19) can be written as

$$-\log p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\theta}) = \frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{\theta})^\top(\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{\theta}) + \text{const}\,. \tag{9.24}$$

Comparing (9.24) with the negative log-likelihood in (9.12) for the "feature-free" model, we immediately see we just need to replace $\boldsymbol{X}$ with $\boldsymbol{\Phi}$. Since both $\boldsymbol{X}$ and $\boldsymbol{\Phi}$ are independent of the parameters $\boldsymbol{\theta}$ that we wish to optimize, we therefore also arrive immediately at the *maximum likelihood estimate*

maximum likelihood estimate

$$\boldsymbol{\theta}_{\text{ML}} = (\boldsymbol{\Phi}^\top\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^\top\boldsymbol{y} \tag{9.25}$$

for the linear regression problem with nonlinear features defined in (9.19).

Example (Maximum Likelihood Polynomial Fit)

Consider the data set in Figure 9.4(a). The data set consists of $N = 20$ pairs (x_i, y_i), where $x_i \sim \mathcal{U}[-5, 5]$ and $y_i = -\sin(x_i/5) + \cos(x_i) + \epsilon$, where $\epsilon \sim \mathcal{N}\left(0,\, 0.2^2\right)$.

We fit a polynomial of degree $K = 4$ using maximum likelihood estimation, i.e., parameters $\boldsymbol{\theta}_{\text{ML}}$ are given in (9.25). The maximum likelihood estimate yields function values $\boldsymbol{\phi}^\top(x_*)\boldsymbol{\theta}_{\text{ML}}$ at any test location x_*. The result is shown in Figure 9.4(b).

Estimating the Noise Variance

Thus far, we assumed that the noise variance σ^2 is known. However, we can also use the principle of maximum likelihood estimation to obtain a point estimate σ^2_{ML} for the noise variance. To do this, we follow the stan-

dard procedure: we write down the log-likelihood, compute the derivative with respect to $\sigma^2 > 0$, set it to 0 and solve:

$$\log p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\theta}, \sigma^2) = \sum_{n=1}^{N} \log \mathcal{N}\big(y_n \,|\, \boldsymbol{\theta}^\top \boldsymbol{\phi}(\boldsymbol{x}_n),\, \sigma^2\big) \tag{9.26}$$

$$= \sum_{n=1}^{N} \left(-\frac{1}{2}\log(2\pi) - \frac{1}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(y_n - \boldsymbol{\theta}^\top \boldsymbol{\phi}(\boldsymbol{x}_n))^2 \right) \tag{9.27}$$

$$= -\frac{N}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\underbrace{\sum_{n=1}^{N}(y_n - \boldsymbol{\theta}^\top \boldsymbol{\phi}(\boldsymbol{x}_n))^2}_{=:\boldsymbol{S}} + \text{const}\,. \tag{9.28}$$

The partial derivative of the log-likelihood with respect to σ^2 is then

$$\frac{\partial \log p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\theta}, \sigma^2)}{\partial \sigma^2} = -\frac{N}{2\sigma^2} + \frac{1}{2\sigma^4}\boldsymbol{S} = 0 \tag{9.29}$$

$$\iff \frac{N}{2\sigma^2} = \frac{1}{2\sigma^4}\boldsymbol{S} \tag{9.30}$$

$$\iff \sigma^2_{\text{ML}} = \frac{1}{N}\boldsymbol{S} = \frac{1}{N}\sum_{n=1}^{N}(y_n - \boldsymbol{\theta}^\top \boldsymbol{\phi}(\boldsymbol{x}_n))^2\,. \tag{9.31}$$

Therefore, the maximum likelihood estimate for the noise variance is the average squared distance between the noise-free function values $\boldsymbol{\theta}^\top \boldsymbol{\phi}(\boldsymbol{x}_n)$ and the corresponding noisy observations y_n at $\boldsymbol{x}_n$, for $n = 1, \ldots, N$.

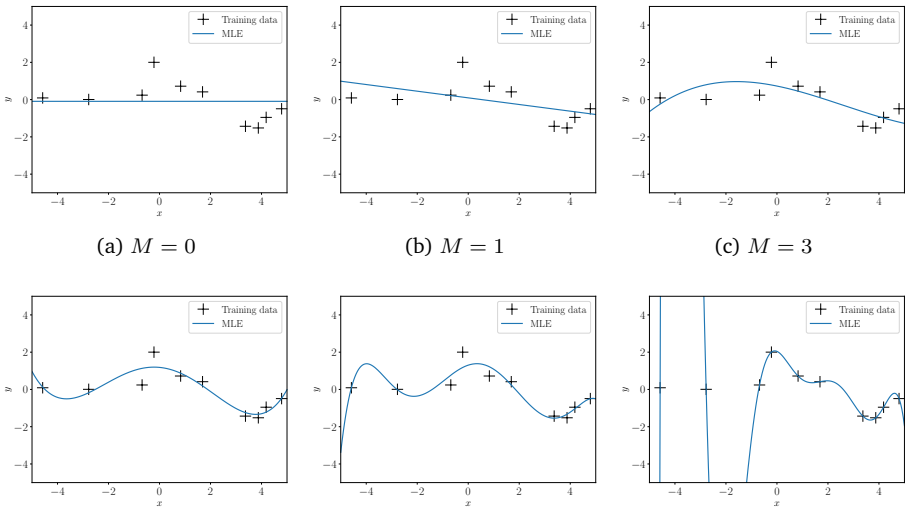
Properties of Maximum Likelihood Estimators

The maximum likelihood estimate $\boldsymbol{\theta}_{\text{ML}}$ possesses the following properties:

- Asymptotic consistency: The MLE converges to the true value in the limit of infinitely many observations, plus a random error that is approximately normal.
- The size of the samples necessary to achieve these properties can be quite large.
- The error's variance decays in $1/N$ where N is the number of data points.
- Especially, in the "small" data regime, maximum likelihood estimation can lead to *overfitting*.

9.2.2 Overfitting in Linear Regression

We have seen that we can use maximum likelihood estimation to fit linear models (e.g., polynomials) to data. We can evaluate the quality of the model by computing the error/loss incurred. One way of doing this is to



(a) $M = 0$ (b) $M = 1$ (c) $M = 3$

(d) $M = 4$ (e) $M = 6$ (f) $M = 9$

Figure 9.5 Polynomial fits for different degrees M.

compute the negative log-likelihood (9.9), which we minimized to determine the MLE. Alternatively, given that the noise parameter σ^2 is not a free parameter, we can ignore the scaling by $1/\sigma^2$, so that we end up with a squared-error-loss function $\|\boldsymbol{y} - \boldsymbol{\Phi\theta}\|^2$. Instead of using this squared loss, we often use the *root mean squared error (RMSE)*

root mean squared error (RMSE)

$$\sqrt{\|\boldsymbol{y} - \boldsymbol{\Phi\theta}\|^2/N} = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(y_n - \boldsymbol{\phi}^\top(\boldsymbol{x}_n)\boldsymbol{\theta})^2}\,, \tag{9.32}$$

which (a) allows us to compare errors of data sets with different sizes and (b) has the same scale and the same units as the observed function values y_i. Note that the division by σ^2 makes the log-likelihood "unit-free".

The RMSE is normalized.

Assume, we fit a model that maps post-codes ($\boldsymbol{x}$ is given in latitude,longitude) to house prices (y-values are GBP). Then, the RMSE is also measured in GBP, whereas the squared error is given in GBP^2.

We can use the RMSE (or the log-likelihood) to determine the best degree of the polynomial by finding the value M, such that the error is minimized. Given that the polynomial degree is a natural number, we can perform a brute-force search and enumerate all (reasonable) values of M.

For a training set of size N it is sufficient to test $0 \leqslant M \leqslant N - 1$.

Figure 9.5 shows a number of polynomial fits determined by maximum likelihood for the dataset from Figure 9.4(a) with $N = 10$ observations. We notice that polynomials of low degree (e.g., constants ($M = 0$) or linear ($M = 1$) fit the data poorly and, hence, are poor representations of the true underlying function. For degrees $M = 3, \ldots, 5$ the fits look plausible and smoothly interpolate the data. When we go to higher-degree polynomials, we notice that they fit the data better and better—in the extreme case of $M = N - 1 = 9$, the function will pass through every single data point. However, these high-degree polynomials oscillate wildly and are a poor representation of the underlying function that generated the data, such that we suffer from *overfitting*.

overfitting

Note that the noise variance $\sigma^2 > 0$.

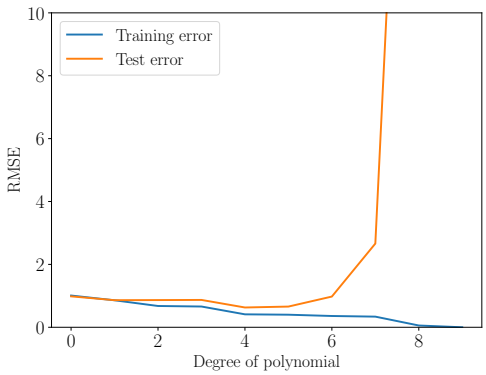


Figure 9.6 Training and test error.

Remember that the goal is to achieve good generalization by making accurate predictions for new (unseen) data. We obtain some quantitative insight into the dependence of the generalization performance on the polynomial of degree M by considering a separate test set comprising 200 data points generated using exactly the same procedure used to generate the training set. As test inputs, we chose a linear grid of 200 points in the interval of $[-5, 5]$. For each choice of M, we evaluate the RMSE (9.32) for both the training data and the test data.

Looking now at the test error, which is a qualitive measure of the generalization properties of the corresponding polynomial, we notice that initially the test error decreases, see Figure 9.6 (red). For fourth-order polynomials the test error is relatively low and stays relatively constant up to degree 5. However, from degree 6 onward the test error increases significantly, and high-order polynomials have very bad generalization properties. In this particular example, this also is evident from the corresponding maximum likelihood fits in Figure 9.5. Note that the training error (blue curve in Figure 9.6) never increases as a function of M. In our example, the best generalization (the point of the smallest test error) is achieved for a polynomial of degree $M = 4$.

9.2.3 Regularization and Maximum A Posteriori Estimation

We just saw that maximum likelihood estimation is prone to overfitting. It often happens that the magnitude of the parameter values becomes relatively big if we run into overfitting (Bishop, 2006). One way to mitigate the effect of overfitting is to penalize big parameter values by a technique called *regularization*. In regularization, we add a term to the log-likelihood that penalizes the magnitude of the parameters $\boldsymbol{\theta}$. A typical example is a regularized "loss function" of the form

regularization

$$-\log p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\theta}) + \lambda \|\boldsymbol{\theta}\|_2^2 , \tag{9.33}$$

where the second term is the regularizer, and $\lambda \geqslant 0$ controls the "strictness" of the regularization.

Remark. Instead of the Euclidean norm $\|\cdot\|_2$, we can choose any p-norm $\|\cdot\|_p$. In practice, smaller values for p lead to sparser solutions. Here, "sparse" means that many parameter values $\theta_i = 0$, which is also useful for variable selection. For $p = 1$, the regularizer is called LASSO (least absolute shrinkage and selection operator) and was proposed by Tibshirani (1996). ♢

From a probabilistic perspective, adding a regularizer is identical to placing a prior distribution $p(\boldsymbol{\theta})$ on the parameters and then selecting the parameters that maximize the posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{X}, \boldsymbol{y})$, i.e., we choose the parameters $\boldsymbol{\theta}$ that are "most probable" given the training data.

In a Bayesian setting, the posterior over the parameters $\boldsymbol{\theta}$, given the training data $\boldsymbol{X}, \boldsymbol{y}$, is obtained by applying Bayes' theorem as

$$p(\boldsymbol{\theta}|\boldsymbol{X}, \boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\boldsymbol{y}|\boldsymbol{X})}. \tag{9.34}$$

The parameter vector $\boldsymbol{\theta}_{\text{MAP}}$ that maximizes the posterior (9.34) is called the *maximum a-posteriori* (*MAP*) estimate.

maximum a-posteriori
MAP

To find the MAP estimate, we follow steps that are similar in flavor to maximum likelihood estimation. We start with the log-transform and compute the log-posterior as

$$\log p(\boldsymbol{\theta}|\boldsymbol{X}, \boldsymbol{y}) = \log p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\theta}) + \log p(\boldsymbol{\theta}) + \text{const}, \tag{9.35}$$

where the constant comprises the terms that are independent of $\boldsymbol{\theta}$. We see that the log-posterior in (9.35) is the sum of the log-likelihood $p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\theta})$ and the log-prior $\log p(\boldsymbol{\theta})$.

Remark (Relation to Regularization). Choosing a Gaussian parameter prior $p(\boldsymbol{\theta}) = \mathcal{N}(\mathbf{0}, b^2\boldsymbol{I})$, $b^2 = \frac{1}{2\lambda}$, the (negative) log-prior term will be

$$-\log p(\boldsymbol{\theta}) = \underbrace{\lambda \boldsymbol{\theta}^\top \boldsymbol{\theta}}_{=\lambda\|\boldsymbol{\theta}\|_2^2} + \text{const}, \tag{9.36}$$

and we recover exactly the regularization term in (9.33). This means that for a quadratic regularization, the regularization parameter λ in (9.33) corresponds to twice the precision (inverse variance) of the Gaussian (isotropic) prior $p(\boldsymbol{\theta})$. Therefore, the log-prior in (9.35) plays the role of a regularizer that penalizes implausible values, i.e., values that are unlikely under the prior. ♢

To find the MAP estimate $\boldsymbol{\theta}_{\text{MAP}}$, we minimize the negative log-posterior with respect to $\boldsymbol{\theta}$, i.e., we solve

$$\boldsymbol{\theta}_{\text{MAP}} \in \arg\min_{\boldsymbol{\theta}} \{-\log p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\theta}) - \log p(\boldsymbol{\theta})\}. \tag{9.37}$$

We compute the gradient of the negative log-posterior with respect to $\boldsymbol{\theta}$ as

$$-\frac{\partial \log p(\boldsymbol{\theta}|\boldsymbol{X},\boldsymbol{y})}{\partial \boldsymbol{\theta}} = -\frac{\partial \log p(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} - \frac{\partial \log p(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\,, \tag{9.38}$$

where we identify the first term on the right-hand-side as the gradient of the negative log-likelihood given in (9.15).

More concretely, with a Gaussian prior $p(\boldsymbol{\theta}) = \mathcal{N}(\mathbf{0},\, b^2\boldsymbol{I})$ on the parameters $\boldsymbol{\theta}$, the negative log-posterior for the linear regression setting (9.19), we obtain the negative log posterior

$$-\log p(\boldsymbol{\theta}|\boldsymbol{X},\boldsymbol{y}) = \frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{\Phi}\boldsymbol{\theta})^\top(\boldsymbol{y}-\boldsymbol{\Phi}\boldsymbol{\theta}) + \frac{1}{2b^2}\boldsymbol{\theta}^\top\boldsymbol{\theta} + \text{const}\,. \tag{9.39}$$

Here, the first term corresponds to the contribution from the log-likelihood, and the second term originates from the log-prior. The gradient of the log-posterior with respect to the parameters $\boldsymbol{\theta}$ is

$$-\frac{\partial \log p(\boldsymbol{\theta}|\boldsymbol{X},\boldsymbol{y})}{\partial \boldsymbol{\theta}} = \frac{1}{\sigma^2}(\boldsymbol{\theta}^\top\boldsymbol{\Phi}^\top\boldsymbol{\Phi} - \boldsymbol{y}^\top\boldsymbol{\Phi}) + \frac{1}{b^2}\boldsymbol{\theta}^\top\,. \tag{9.40}$$

We will find the MAP estimate $\boldsymbol{\theta}_{\text{MAP}}$ by setting this gradient to $\mathbf{0}$:

$$\frac{1}{\sigma^2}(\boldsymbol{\theta}^\top\boldsymbol{\Phi}^\top\boldsymbol{\Phi} - \boldsymbol{y}^\top\boldsymbol{\Phi}) + \frac{1}{b^2}\boldsymbol{\theta}^\top = \mathbf{0} \tag{9.41}$$

$$\iff \boldsymbol{\theta}^\top\left(\frac{1}{\sigma^2}\boldsymbol{\Phi}^\top\boldsymbol{\Phi} + \frac{1}{b^2}\boldsymbol{I}\right) - \frac{1}{\sigma^2}\boldsymbol{y}^\top\boldsymbol{\Phi} = \mathbf{0} \tag{9.42}$$

$$\iff \boldsymbol{\theta}^\top\left(\boldsymbol{\Phi}^\top\boldsymbol{\Phi} + \frac{\sigma^2}{b^2}\boldsymbol{I}\right) = \boldsymbol{y}^\top\boldsymbol{\Phi} \tag{9.43}$$

$$\iff \boldsymbol{\theta}^\top = \boldsymbol{y}^\top\boldsymbol{\Phi}\left(\boldsymbol{\Phi}^\top\boldsymbol{\Phi} + \frac{\sigma^2}{b^2}\boldsymbol{I}\right)^{-1} \tag{9.44}$$

$$\iff \boldsymbol{\theta}_{\text{MAP}} = \left(\boldsymbol{\Phi}^\top\boldsymbol{\Phi} + \frac{\sigma^2}{b^2}\boldsymbol{I}\right)^{-1}\boldsymbol{\Phi}^\top\boldsymbol{y}\,. \tag{9.45}$$

Comparing the MAP estimate in (9.45) with the maximum likelihood estimate in (9.25) we see that the only difference between both solutions is the additional term $\frac{\sigma^2}{b^2}\boldsymbol{I}$ in the inverse matrix. This term ensures that $\boldsymbol{\Phi}^\top\boldsymbol{\Phi} + \frac{\sigma^2}{b^2}\boldsymbol{I}$ is strictly positive definite (i.e., its inverse exists) and plays the role of the *regularizer*.

$\boldsymbol{\Phi}^\top\boldsymbol{\Phi}$ is positive semidefinite and the additional term is strictly positive definite, such that all eigenvalues of the matrix to be inverted are positive.

regularizer

Example (MAP Estimation for Polynomial Regression)

In the polynomial regression example from Section 9.2.1, we place a Gaussian prior $p(\boldsymbol{\theta}) = \mathcal{N}(\mathbf{0},\, \boldsymbol{I})$ on the parameters $\boldsymbol{\theta}$ and determine the MAP estimates according to (9.45). In Figure 9.7, we show both the maximum likelihood and the MAP estimates for polynomials of degree 6 (left) and degree 8 (right). The prior (regularizer) does not play a significant role for the low-degree polynomial, but keeps the function relatively

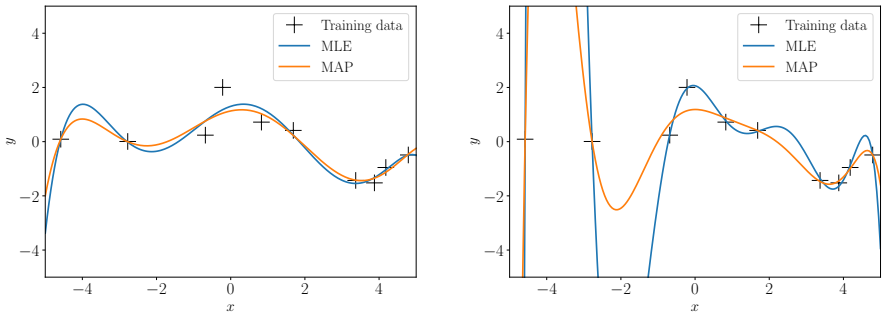


Figure 9.7 Polynomial regression: Maximum likelihood and MAP estimates.

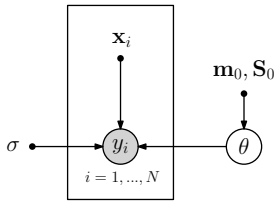


Figure 9.8 Graphical model for Bayesian linear regression.

smooth for higher-degree polynomials. However, the MAP estimate can only push the boundaries of overfitting—it is not a general solution to this problem.

In the following, we will discuss Bayesian linear regression where we average over all plausible sets of parameters instead of focusing on a point estimate.

9.3 Bayesian Linear Regression

Thus far, we looked at linear regression models where we estimated the parameters $\boldsymbol{\theta}$, e.g., by means of maximum likelihood or MAP estimation. We discovered that MLE can lead to severe overfitting, in particular, in the small-data regime. MAP addresses this issue by placing a prior on the parameters that plays the role of a regularizer.

Bayesian linear regression pushes the idea of the parameter prior a step further and does not even attempt to compute a point estimate of the parameters, but instead the full posterior over the parameters is taken into account when making predictions. This means we do not fit any parameters, but we compute an average over all plausible parameters settings (according to the posterior).

9.3.1 Model

In Bayesian linear regression, we consider the following model

$$\begin{aligned} p(y|\boldsymbol{x}, \boldsymbol{\theta}) &= \mathcal{N}\big(y \,|\, \phi^\top(\boldsymbol{x})\boldsymbol{\theta},\, \sigma^2\big) \quad \text{likelihood} \\ p(\boldsymbol{\theta}) &= \mathcal{N}\big(\boldsymbol{m}_0,\, \boldsymbol{S}_0\big) \quad \text{prior}\,, \end{aligned} \tag{9.46}$$

where we now explicitly place a Gaussian prior $p(\boldsymbol{\theta}) = \mathcal{N}\big(\boldsymbol{m}_0, \boldsymbol{S}_0\big)$ on the parameter vector $\boldsymbol{\theta}$. The graphical corresponding graphical model is shown in Fig. 9.8.

Why is a Gaussian prior a convenient choice?

9.3.2 Prior Predictions

In practice, we are usually not so much interested in the parameter values $\boldsymbol{\theta}$. Instead, our focus often lies in the predictions we make with those parameter values. In a Bayesian setting, we take the parameter distribution and average over all plausible parameter settings when we make predictions.

More specifically, to make predictions at an input location $\boldsymbol{x}_*$, we integrate out the parameter distribution and obtain

$$p(y_*|\boldsymbol{x}_*) = \int p(y_*|\boldsymbol{x}_*, \boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta} = \mathbb{E}_{\boldsymbol{\theta}}[p(y_*|\boldsymbol{x}_*, \boldsymbol{\theta})]\,, \tag{9.47}$$

where we compute an average of the predictions of y_* for all plausible parameters $\boldsymbol{\theta}$ according to the parameter distribution $p(\boldsymbol{\theta})$. In our model, we chose a conjugate Gaussian parameter prior so that the predictive distribution is Gaussian as well. If we take the prior distribution $p(\boldsymbol{\theta}) = \mathcal{N}\big(\boldsymbol{m}_0, \boldsymbol{S}_0\big)$, we obtain the predictive distribution as

$$p(y_*|\boldsymbol{x}_*) = \mathcal{N}\big(\boldsymbol{\phi}^\top(\boldsymbol{x}_*)\boldsymbol{m}_0,\, \boldsymbol{\phi}^\top(\boldsymbol{x}_*)\boldsymbol{S}_0\phi(\boldsymbol{x}_*) + \sigma^2\big)\,, \tag{9.48}$$

where we used that (i) the posterior is Gaussian due to conjugacy, (ii), the Gaussian noise is independent so that $\mathbb{V}[y_*] = \mathbb{V}[\boldsymbol{\phi}^\top(\boldsymbol{x}_*)\boldsymbol{\theta}] + \mathbb{V}[\epsilon]$, (iii) y_* is a linear transformation of $\boldsymbol{\theta}$ so that we can apply the rules for computing the mean and covariance of the prediction analytically by using (6.41) and (6.42), respectively.

In (9.48), the term $\boldsymbol{\phi}^\top(\boldsymbol{x}_*)\boldsymbol{S}_0\boldsymbol{\phi}(\boldsymbol{x}_*)$ in the predictive variance explicitly accounts for the uncertainty associated with the parameters $\boldsymbol{\theta}$, whereas σ^2 is the uncertainty contribution due to the measurement noise.

Example (Prior over Functions)

Let us consider a Bayesian linear regression problem with polynomials of degree 5. We choose a parameter prior $p(\boldsymbol{\theta}) = \mathcal{N}\big(\mathbf{0},\, \frac{1}{4}\boldsymbol{I}\big)$. Figure 9.9 visualizes the prior over functions induced by this parameter prior, including some function samples from this prior.

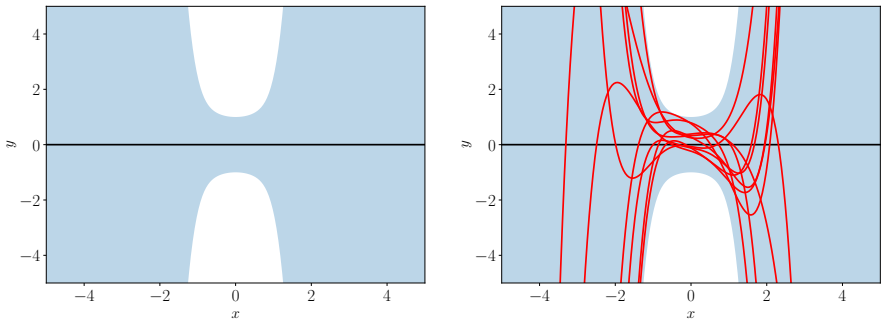


Figure 9.9 Prior over functions. Left: Distribution over functions represented by the mean function (black line) and the marginal uncertainties (shaded), representing the 95% confidence bounds. Right: Samples from the prior over functions which are induced by the samples from the parameter prior.

So far, we looked at computing predictions using the parameter prior $p(\boldsymbol{\theta})$. However, when we have a parameter posterior (given some training data $\boldsymbol{X}, \boldsymbol{y}$), the same principles for prediction and inference hold as in (9.47)—we just need to replace the prior $p(\boldsymbol{\theta})$ with the posterior $p(\boldsymbol{\theta}|\boldsymbol{X}, \boldsymbol{y})$. In the following, we will derive the posterior distribution in detail before using it to make predictions.

9.3.3 Parameter Posterior

Given a training set of inputs $\boldsymbol{x}_n \in \mathbb{R}^D$ and corresponding observations $y_n \in \mathbb{R}$, $n = 1, \ldots, N$, we compute the posterior over the parameters using Bayes' theorem as

$$p(\boldsymbol{\theta}|\boldsymbol{X}, \boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\boldsymbol{y}|\boldsymbol{X})}, \tag{9.49}$$

where $\boldsymbol{X}$ is the collection of training inputs and $\boldsymbol{y}$ the collection of training targets. Furthermore, $p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\theta})$ is the likelihood, $p(\boldsymbol{\theta})$ the parameter prior and

$$p(\boldsymbol{y}|\boldsymbol{X}) = \int p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta} \tag{9.50}$$

marginal likelihood
evidence

the *marginal likelihood/evidence*, which is independent of the parameters $\boldsymbol{\theta}$ and ensures that the posterior is normalized, i.e., it integrates to 1. We can think of the marginal likelihood as the likelihood averaged over all possible parameter settings (with respect to the prior distribution $p(\boldsymbol{\theta})$).

In our specific model (9.46), the posterior (9.49) can be computed in closed form as

$$p(\boldsymbol{\theta}|\boldsymbol{X}, \boldsymbol{y}) = \mathcal{N}\left(\boldsymbol{\theta} \,|\, \boldsymbol{m}_N, \boldsymbol{S}_N\right), \tag{9.51}$$

$$\boldsymbol{S}_N = (\boldsymbol{S}_0^{-1} + \sigma^{-2}\boldsymbol{\Phi}^\top\boldsymbol{\Phi})^{-1}, \tag{9.52}$$

$$\boldsymbol{m}_N = \boldsymbol{S}_N(\boldsymbol{S}_0^{-1}\boldsymbol{m}_0 + \sigma^{-2}\boldsymbol{\Phi}^\top\boldsymbol{y}), \tag{9.53}$$

where the subscript N indicates the size of the training set. In the following, we will detail how we arrive at this posterior.

Bayes' theorem tells us that the posterior $p(\boldsymbol{\theta}|\boldsymbol{X}, \boldsymbol{y})$ is proportional to the product of the likelihood $p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\theta})$ and the prior $p(\boldsymbol{\theta})$:

$$\text{posterior} \qquad p(\boldsymbol{\theta}|\boldsymbol{X}, \boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\boldsymbol{y}|\boldsymbol{X})} \tag{9.54}$$

$$\text{likelihood} \qquad p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\theta}) = \mathcal{N}\big(\boldsymbol{y} \,|\, \boldsymbol{\Phi}\boldsymbol{\theta},\, \sigma^2\boldsymbol{I}\big) \tag{9.55}$$

$$\text{prior} \qquad p(\boldsymbol{\theta}) = \mathcal{N}\big(\boldsymbol{\theta} \,|\, \boldsymbol{m}_0,\, \boldsymbol{S}_0\big) \tag{9.56}$$

We will discuss two approaches that will give us the desired posterior.

Approach 1: Linear Transformation of Gaussian Random Variables

Looking at the numerator of the posterior in (9.54), we know that the Gaussian prior times the Gaussian likelihood (where the parameters on which we place the Gaussian appears linearly in the mean) is an (unnormalized) Gaussian (see Section 6.6). If we want to compute this product by using the results from (6.73)–(6.74) in Section 6.6, we need to ensure the product has the "right" form

If necessary, we can find the normalizing constant of this unnormalized Gaussian, see (6.75).

$$\mathcal{N}\big(\boldsymbol{y} \,|\, \boldsymbol{\Phi}\boldsymbol{\theta},\, \sigma^2\boldsymbol{I}\big)\mathcal{N}\big(\boldsymbol{\theta} \,|\, \boldsymbol{m}_0,\, \boldsymbol{S}_0\big) = \mathcal{N}\big(\boldsymbol{\theta} \,|\, \boldsymbol{\mu},\, \boldsymbol{\Sigma}\big)\mathcal{N}\big(\boldsymbol{\theta} \,|\, \boldsymbol{m}_0,\, \boldsymbol{S}_0\big) \tag{9.57}$$

for some $\boldsymbol{\mu}, \boldsymbol{\Sigma}$. With this form we determine the desired product immediately as

$$\mathcal{N}\big(\boldsymbol{\theta} \,|\, \boldsymbol{\mu},\, \boldsymbol{\Sigma}\big)\mathcal{N}\big(\boldsymbol{\theta} \,|\, \boldsymbol{m}_0,\, \boldsymbol{S}_0\big) \propto \mathcal{N}\big(\boldsymbol{\theta} \,|\, \boldsymbol{m}_N,\, \boldsymbol{S}_N\big) \tag{9.58}$$

$$\boldsymbol{S}_N = (\boldsymbol{S}_0^{-1} + \boldsymbol{\Sigma}^{-1})^{-1} \tag{9.59}$$

$$\boldsymbol{m}_N = \boldsymbol{S}_N(\boldsymbol{S}_0^{-1}\boldsymbol{m}_0 + \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})\,. \tag{9.60}$$

In order to get the "right" form, we need to turn $\mathcal{N}\big(\boldsymbol{y} \,|\, \boldsymbol{\Phi}\boldsymbol{\theta},\, \sigma^2\boldsymbol{I}\big)$ into $\mathcal{N}\big(\boldsymbol{\theta} \,|\, \boldsymbol{\mu},\, \boldsymbol{\Sigma}\big)$ for appropriate choices of $\boldsymbol{\mu}, \boldsymbol{\Sigma}$. We will do this by using a linear transformation of Gaussian random variables (see Section 6.6), which allows us to exploit the property that linearly transformed Gaussian random variables are Gaussian distributed. More specifically, we will find $\boldsymbol{\mu} = \boldsymbol{B}\boldsymbol{y}$ and $\boldsymbol{\Sigma} = \sigma^2\boldsymbol{B}\boldsymbol{B}^\top$ by linearly transforming the relationship $\boldsymbol{y} = \Phi\boldsymbol{\theta}$ in the likelihood into $\boldsymbol{B}\boldsymbol{y} = \boldsymbol{\theta}$ for a suitable $\boldsymbol{B}$. We obtain

$$\boldsymbol{y} = \boldsymbol{\Phi}\boldsymbol{\theta} \overset{\times\boldsymbol{\Phi}^\top}{\iff} \boldsymbol{\Phi}^\top\boldsymbol{y} = \boldsymbol{\Phi}^\top\boldsymbol{\Phi}\boldsymbol{\theta} \overset{\times(\boldsymbol{\Phi}^\top\boldsymbol{\Phi})^{-1}}{\iff} \underbrace{(\boldsymbol{\Phi}^\top\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^\top}_{=:\boldsymbol{B}}\boldsymbol{y} = \boldsymbol{\theta} \tag{9.61}$$

Therefore, we can write $\boldsymbol{\theta} = \boldsymbol{B}\boldsymbol{y}$, and by using the rules for linear transformations of the mean and covariance from (6.41)–(6.42) we obtain

$$\mathcal{N}\big(\boldsymbol{\theta} \,|\, \boldsymbol{B}\boldsymbol{y},\, \sigma^2\boldsymbol{B}\boldsymbol{B}^\top\big) = \mathcal{N}\big(\boldsymbol{\theta} \,|\, (\boldsymbol{\Phi}^\top\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^\top\boldsymbol{y},\, \sigma^2(\boldsymbol{\Phi}^\top\boldsymbol{\Phi})^{-1}\big) \tag{9.62}$$

after some re-arranging of the terms for the covariance matrix.

If we now (9.62) and define its mean as $\boldsymbol{\mu}$ and covariance matrix as $\boldsymbol{\Sigma}$

in (9.60) and (9.59), respectively, we obtain the covariance $\boldsymbol{S}_N$ and the mean $\boldsymbol{m}_N$ of the parameter posterior $\mathcal{N}(\boldsymbol{\theta} \,|\, \boldsymbol{m}_N, \boldsymbol{S}_N)$ as

$$\boldsymbol{S}_N = (\boldsymbol{S}_0^{-1} + \sigma^{-2}\boldsymbol{\Phi}^\top\boldsymbol{\Phi})^{-1} \tag{9.63}$$

$$\boldsymbol{m}_N = \boldsymbol{S}_N(\boldsymbol{S}_0^{-1}\boldsymbol{m}_0 + \underbrace{\sigma^{-2}(\boldsymbol{\Phi}^\top\boldsymbol{\Phi})}_{\boldsymbol{\Sigma}^{-1}}\underbrace{(\boldsymbol{\Phi}^\top\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^\top\boldsymbol{y}}_{\boldsymbol{\mu}}) \tag{9.64}$$

$$= \boldsymbol{S}_N(\boldsymbol{S}_0^{-1}\boldsymbol{m}_0 + \sigma^{-2}\boldsymbol{\Phi}^\top\boldsymbol{y}), \tag{9.65}$$

The posterior mean equals the MAP estimate.

respectively. Note that the posterior mean $\boldsymbol{m}_N$ equals the MAP estimate $\boldsymbol{\theta}_{\text{MAP}}$ from (9.45). This also makes sense since the posterior distribution is unimodal (Gaussian) with its maximum at the mean.

Remark. The posterior precision (inverse covariance) of the parameters (see (9.63))

$$\boldsymbol{S}_N^{-1} = \boldsymbol{S}_0^{-1} + \frac{1}{\sigma^2}\boldsymbol{\Phi}^\top\boldsymbol{\Phi} \tag{9.66}$$

contains two terms: $\boldsymbol{S}_0^{-1}$ is the prior precision and $\frac{1}{\sigma^2}\boldsymbol{\Phi}^\top\boldsymbol{\Phi}$ is a data-dependent (precision) term. Both terms (matrices) are symmetric and positive definite. The data-dependent term $\frac{1}{\sigma^2}\boldsymbol{\Phi}^\top\boldsymbol{\Phi}$ grows as more data is taken into account. This means (at least) two things:

$\boldsymbol{\Phi}^\top\boldsymbol{\Phi}$ accumulates contributions from the data.

- The posterior precision grows as more and more data is taken into account; therefore, the covariance and the uncertainty about the parameters shrinks.
- The relative influence of the parameter prior vanishes for large N.

Therefore, for $N \to \infty$ the prior plays no role, and the parameter posterior tends to a point estimate, the MAP estimate. ♢

Approach 2: Completing the Squares

Instead of looking at the product of the prior and the likelihood, we can transform the problem into log-space and solve for the mean and covariance of the posterior by completing the squares.

The sum of the log-prior and the log-likelihood is

$$\log\mathcal{N}(\boldsymbol{y} \,|\, \boldsymbol{\Phi}\boldsymbol{\theta}, \sigma^2\boldsymbol{I}) + \log\mathcal{N}(\boldsymbol{\theta} \,|\, \boldsymbol{m}_0, \boldsymbol{S}_0) \tag{9.67}$$

$$= -\frac{1}{2}(\sigma^{-2}(\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{\theta})^\top(\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{\theta}) + (\boldsymbol{\theta} - \boldsymbol{m}_0)^\top\boldsymbol{S}_0^{-1}(\boldsymbol{\theta} - \boldsymbol{m}_0)) + \text{const} \tag{9.68}$$

where the constant contains terms independent of $\boldsymbol{\theta}$. We will ignore the constant in the following. We now factorize (9.68), which yields

$$\begin{aligned} &-\frac{1}{2}(\sigma^{-2}\boldsymbol{y}^\top\boldsymbol{y} - 2\sigma^{-2}\boldsymbol{y}^\top\boldsymbol{\Phi}\boldsymbol{\theta} + \boldsymbol{\theta}^\top\sigma^{-2}\boldsymbol{\Phi}^\top\boldsymbol{\Phi}\boldsymbol{\theta} + \boldsymbol{\theta}^\top\boldsymbol{S}_0^{-1}\boldsymbol{\theta} \\ &- 2\boldsymbol{m}_0^\top\boldsymbol{S}_0^{-1}\boldsymbol{\theta} + \boldsymbol{m}_0^\top\boldsymbol{S}_0^{-1}\boldsymbol{m}_0) \end{aligned} \tag{9.69}$$

$$= -\frac{1}{2}\big(\boldsymbol{\theta}^\top(\sigma^{-2}\boldsymbol{\Phi}^\top\boldsymbol{\Phi} + \boldsymbol{S}_0^{-1})\boldsymbol{\theta} - 2(\sigma^{-2}\boldsymbol{\Phi}^\top\boldsymbol{y} + \boldsymbol{S}_0^{-1}\boldsymbol{m}_0)^\top\boldsymbol{\theta}\big) + \text{const}\,, \tag{9.70}$$

where the constant contains the black terms in (9.69), which are independent of $\boldsymbol{\theta}$. The orange terms are terms that are linear in $\boldsymbol{\theta}$, and the blue terms are the ones that are quadratic in $\boldsymbol{\theta}$. By inspecting (9.70), we find that this equation is quadratic in $\boldsymbol{\theta}$. The fact that the unnormalized log-posterior distribution is a (negative) quadratic form implies that the posterior is Gaussian, i.e.,

$$p(\boldsymbol{\theta}|\boldsymbol{X},\boldsymbol{y}) = \exp(\log p(\boldsymbol{\theta}|\boldsymbol{X},\boldsymbol{y})) \propto \exp(\log p(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{\theta}) + \log p(\boldsymbol{\theta})) \tag{9.71}$$

$$\propto \exp\Big(-\frac{1}{2}\big(\boldsymbol{\theta}^\top(\sigma^{-2}\boldsymbol{\Phi}^\top\boldsymbol{\Phi} + \boldsymbol{S}_0^{-1})\boldsymbol{\theta} - 2(\sigma^{-2}\boldsymbol{\Phi}^\top\boldsymbol{y} + \boldsymbol{S}_0^{-1}\boldsymbol{m}_0)^\top\boldsymbol{\theta}\big)\Big)\,, \tag{9.72}$$

where we used (9.70) in the last expression.

The remaining task is it to bring this (unnormalized) Gaussian into the form that is proportional to $\mathcal{N}(\boldsymbol{\theta}\,|\,\boldsymbol{m}_N,\,\boldsymbol{S}_N)$, i.e., we need to identify the mean $\boldsymbol{m}_N$ and the covariance matrix $\boldsymbol{S}_N$. To do this, we use the concept of *completing the squares*. The desired log-posterior is

completing the squares

$$\log\mathcal{N}(\boldsymbol{\theta}\,|\,\boldsymbol{m}_N,\,\boldsymbol{S}_N) = -\frac{1}{2}\big((\boldsymbol{\theta}-\boldsymbol{m}_N)^\top\boldsymbol{S}_N^{-1}(\boldsymbol{\theta}-\boldsymbol{m}_N)\big) + \text{const} \tag{9.73}$$

$$= -\frac{1}{2}\big(\boldsymbol{\theta}^\top\boldsymbol{S}_N^{-1}\boldsymbol{\theta} - 2\boldsymbol{m}_N^\top\boldsymbol{S}_N^{-1}\boldsymbol{\theta} + \boldsymbol{m}_N^\top\boldsymbol{S}_N^{-1}\boldsymbol{m}_N\big)\,. \tag{9.74}$$

Here, we factorized the quadratic form $(\boldsymbol{\theta}-\boldsymbol{m}_N)^\top\boldsymbol{S}_N^{-1}(\boldsymbol{\theta}-\boldsymbol{m}_N)$ into a term that is quadratic in $\boldsymbol{\theta}$ alone (blue), a term that is linear in $\boldsymbol{\theta}$ (orange), and a constant term (black). This allows us now to find $\boldsymbol{S}_N$ and $\boldsymbol{m}_N$ by matching the colored expressions in (9.70) and (9.74), which yields

$$\boldsymbol{S}_N^{-1} = \boldsymbol{\Phi}^\top\sigma^{-2}\boldsymbol{I}\boldsymbol{\Phi} + \boldsymbol{S}_0^{-1} \iff \boldsymbol{S}_N = (\sigma^{-2}\boldsymbol{\Phi}^\top\boldsymbol{\Phi} + \boldsymbol{S}_0^{-1})^{-1}\,, \tag{9.75}$$

$$\boldsymbol{m}_N^\top\boldsymbol{S}_N^{-1} = (\sigma^{-2}\boldsymbol{\Phi}^\top\boldsymbol{y} + \boldsymbol{S}_0^{-1}\boldsymbol{m}_0)^\top \iff \boldsymbol{m}_N = \boldsymbol{S}_N(\sigma^{-2}\boldsymbol{\Phi}^\top\boldsymbol{y} + \boldsymbol{S}_0^{-1}\boldsymbol{m}_0)\,. \tag{9.76}$$

This is identical to the solution in (9.63)–(9.65), which we obtained by linear transformations of Gaussian random variables.

Remark (Completing the Squares—General Approach). If we are given an equation

$$\boldsymbol{x}^\top\boldsymbol{A}\boldsymbol{x} - 2\boldsymbol{a}^\top\boldsymbol{x} + \text{const}_1\,, \tag{9.77}$$

where $\boldsymbol{A}$ is symmetric and positive definite, which we wish to bring into the form

$$(\boldsymbol{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}(\boldsymbol{x}-\boldsymbol{\mu}) + \text{const}_2\,, \tag{9.78}$$

we can do this by setting

$$\boldsymbol{\Sigma} = \boldsymbol{A} \tag{9.79}$$

$$\boldsymbol{\mu} = \boldsymbol{\Sigma}^{-1}\boldsymbol{a} \tag{9.80}$$

and $\text{const}_2 = \text{const}_1 - \boldsymbol{\mu}^\top \boldsymbol{\Sigma} \boldsymbol{\mu}$. ♢

We can see that the terms inside the exponential in (9.72) are of the form (9.77) with

$$\boldsymbol{A} = \sigma^{-2}\boldsymbol{\Phi}^\top \boldsymbol{\Phi} + \boldsymbol{S}_0^{-1}\,, \tag{9.81}$$

$$\boldsymbol{a} = \sigma^{-2}\boldsymbol{\Phi}^\top \boldsymbol{y} + \boldsymbol{S}_0^{-1}\boldsymbol{m}_0\,. \tag{9.82}$$

Since $\boldsymbol{A}, \boldsymbol{a}$ can be difficult to identify in equations like (9.69), it is often helpful to bring these equations into the form (9.77) that decouples quadratic term, linear terms and constants, which simplifies finding the desired solution.

9.3.4 Posterior Predictions

In (9.47), we computed the predictive distribution of y_* at a test input $\boldsymbol{x}_*$ using the parameter prior $p(\boldsymbol{\theta})$. In principle, predictions with the parameter posterior $p(\boldsymbol{\theta}|\boldsymbol{X}, \boldsymbol{y})$ is not fundamentally different given that in our conjugate model the prior and posterior are both Gaussian (with different parameters). Therefore, by following the same reasoning as before, we obtain

$$p(y_*|\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{x}_*) = \int p(y_*|\boldsymbol{x}_*, \boldsymbol{\theta}) p(\boldsymbol{\theta}|\boldsymbol{X}, \boldsymbol{y}) d\boldsymbol{\theta} \tag{9.83}$$

$$= \int \mathcal{N}\big(y_* \,|\, \boldsymbol{\phi}^\top(\boldsymbol{x}_*)\boldsymbol{\theta},\, \sigma^2\big)\mathcal{N}\big(\boldsymbol{\theta} \,|\, \boldsymbol{m}_N,\, \boldsymbol{S}_N\big) d\boldsymbol{\theta} \tag{9.84}$$

$$= \mathcal{N}\big(y_* \,|\, \boldsymbol{\phi}^\top(\boldsymbol{x}_*)\boldsymbol{m}_N,\, \boldsymbol{\phi}^\top(\boldsymbol{x}_*)\boldsymbol{S}_N\boldsymbol{\phi}(\boldsymbol{x}_*) + \sigma^2\big) \tag{9.85}$$

The term $\boldsymbol{\phi}^\top(\boldsymbol{x}_*)\boldsymbol{S}_N\boldsymbol{\phi}(\boldsymbol{x}_*)$ reflects the posterior uncertainty associated with the parameters $\boldsymbol{\theta}$. Note that $\boldsymbol{S}_N$ depends on the training inputs $\boldsymbol{X}$, see (9.63). The predictive mean coincides with the MAP estimate.

Remark (Mean and Variance of Noise-Free Function Values). In many cases, we are not interested in the predictive distribution $p(y_*|\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{x}_*)$ of a (noisy) observation. Instead, we would like to obtain the distribution of the (noise-free) latent function values $f(\boldsymbol{x}_*) = \boldsymbol{\phi}^\top(\boldsymbol{x}_*)\boldsymbol{\theta}$. We determine the corresponding moments by exploiting the properties of means and variances, which yields

$$\begin{aligned}\mathbb{E}[f(\boldsymbol{x}_*)|\boldsymbol{X}, \boldsymbol{y}] &= \mathbb{E}_{\boldsymbol{\theta}}[\boldsymbol{\phi}^\top(\boldsymbol{x}_*)\boldsymbol{\theta}|\boldsymbol{X}, \boldsymbol{y}] = \boldsymbol{\phi}^\top(\boldsymbol{x}_*)\mathbb{E}_{\boldsymbol{\theta}}[\boldsymbol{\theta}|\boldsymbol{X}, \boldsymbol{y}] \\ &= \boldsymbol{\phi}^\top(\boldsymbol{x}_*)\boldsymbol{m}_N\,,\end{aligned} \tag{9.86}$$

$$\begin{aligned}\mathbb{V}_{\boldsymbol{\theta}}[f(\boldsymbol{x}_*)|\boldsymbol{X}, \boldsymbol{y}] &= \mathbb{V}_{\boldsymbol{\theta}}[\boldsymbol{\phi}^\top(\boldsymbol{x}_*)\boldsymbol{\theta}|\boldsymbol{X}, \boldsymbol{y}] \\ &= \boldsymbol{\phi}^\top(\boldsymbol{x}_*)\mathbb{V}_{\boldsymbol{\theta}}[\boldsymbol{\theta}|\boldsymbol{X}, \boldsymbol{y}]\boldsymbol{\phi}(\boldsymbol{x}_*) = \boldsymbol{\phi}^\top(\boldsymbol{x}_*)\boldsymbol{S}_N\boldsymbol{\phi}(\boldsymbol{x}_*)\end{aligned} \tag{9.87}$$

We see that the predictive mean is the same as the predictive mean for noisy observations as the noise has mean 0, and the predictive variance only differs by σ^2, which is the variance of the measurement noise: When we predict noisy function values, we need to include σ^2 as a source of uncertainty, but this term is not needed for noise-free predictions. Here, the only remaining uncertainty stems from the parameter posterior. ♢

Integrating out parameters induces a distribution over functions.

Remark (Distribution over Functions). The fact that we integrate out the parameters $\boldsymbol{\theta}$ induces a distribution over functions: If we sample $\boldsymbol{\theta}_i \sim p(\boldsymbol{\theta}|\boldsymbol{X}, \boldsymbol{y})$ from the parameter posterior, we obtain a single function realization $\boldsymbol{\theta}_i^\top \boldsymbol{\phi}(\cdot)$. The *mean function*, i.e., the set of all expected function values $\mathbb{E}_{\boldsymbol{\theta}}[f(\cdot)|\boldsymbol{\theta}, \boldsymbol{X}, \boldsymbol{y}]$, of this distribution over functions is $\boldsymbol{m}_N^\top \boldsymbol{\phi}(\cdot)$. The (marginal) variances, i.e., the variance of the function $f(\cdot)$, are given by $\boldsymbol{\phi}^\top(\cdot)\boldsymbol{S}_N\boldsymbol{\phi}(\cdot)$. ♢

mean function

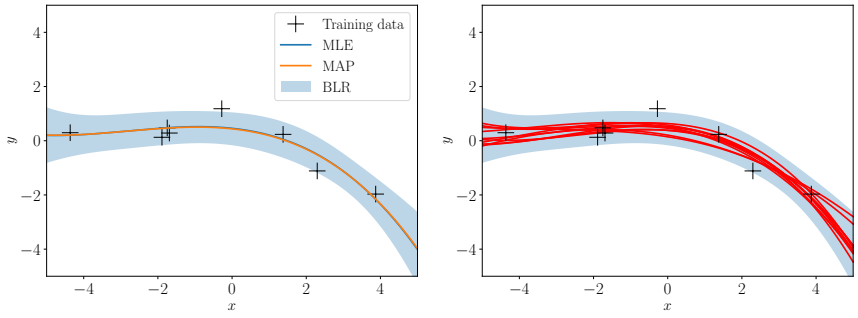
Example (Posterior over Functions)

Let us revisit the Bayesian linear regression problem with polynomials of degree 5. We choose a parameter prior $p(\boldsymbol{\theta}) = \mathcal{N}\left(\mathbf{0}, \frac{1}{4}\boldsymbol{I}\right)$, and. Figure 9.9 visualizes the prior over functions induced by the parameter prior.

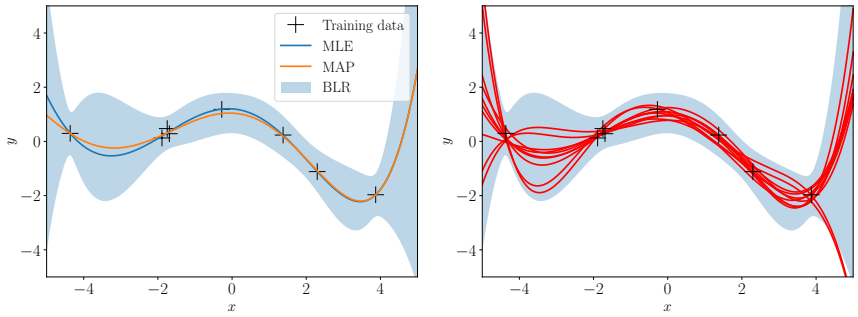
Figure 9.11 shows the posterior we obtain via Bayesian linear regression. When observing 8 noisy function values, see Figure 9.11(a), the posterior distribution over functions is shown in Figure 9.11(b), including the functions we would obtain via maximum likelihood and MAP estimation. The function we obtain using the MAP estimate also corresponds to the posterior mean function in the Bayesian linear regression setting. Figure 9.11(c) shows some plausible realizations (samples) of functions under that posterior over functions.

Example

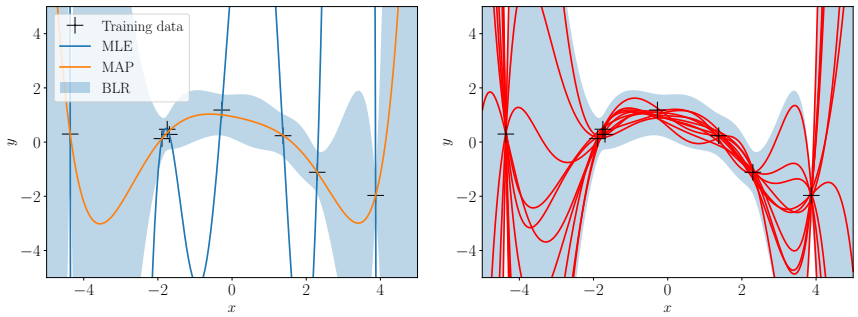
Figure 9.10 shows some examples of the posterior distribution over functions, induced by the parameter posterior. The left panels show the maximum likelihood estimate, the MAP estimate (which is identical to the posterior mean function) and the 95% predictive confidence bounds, represented by the shaded area. The right panels show samples from the posterior over functions: Here, we sampled parameters $\boldsymbol{\theta}_i$ from the parameter posterior and computed the function $\boldsymbol{\phi}^\top(\boldsymbol{x}_*)\boldsymbol{\theta}_i$, which is a single realization of a function under the posterior distribution over functions. For low-order polynomials, the parameter posterior does not allow the parameters to vary much: The sampled functions are nearly identical. When we make the model more flexible by adding more parameters (i.e., we end up with a higher-order polynomial), these parameters are not sufficiently constrained by the posterior, and the sampled functions can be easily visually separated. We also see in the corresponding panels on the left how



(a) Posterior distribution for polynomials of degree $M = 3$ (left) and samples from the posterior over functions (right).



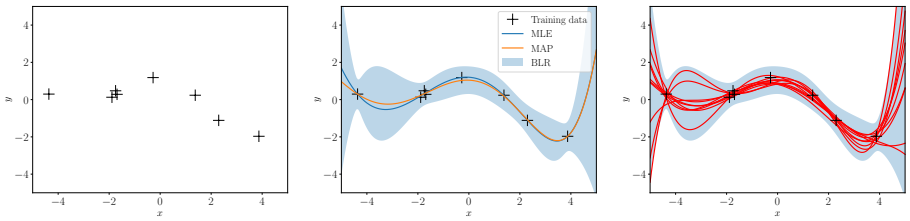
(b) Posterior distribution for polynomials of degree $M = 5$ (left) and samples from the posterior over functions (right).



(c) Posterior distribution for polynomials of degree $M = 7$ (left) and samples from the posterior over functions (right).

Figure 9.10 Bayesian linear regression. Left panels: The shaded area indicates the 95% predictive confidence bounds. The mean of the Bayesian linear regression model coincides with the MAP estimate. The predictive uncertainty is the sum of the noise term and the posterior parameter uncertainty, which depends on the location of the test input. Right panels: Sampled functions from the posterior distribution.

the uncertainty increases, especially at the boundaries. Although for a 7th-order polynomial the MAP estimate yields a reasonable fit, the Bayesian linear regression model additionally tells us that the posterior uncertainty is huge. This information can be critical, if we use these predictions in a decision-making system, where bad decisions can have significant consequences (e.g., in reinforcement learning or robotics).



(a) Training data. (b) Posterior over functions. (c) Samples from the posterior over functions.

Figure 9.11 Bayesian linear regression and posterior over functions. (a) Training data; (b) posterior distribution over functions represented by the marginal uncertainties (shaded) showing the 95% predictive confidence bounds, the maximum likelihood estimate (MLE) and the MAP estimate (MAP), which is identical to the posterior mean function; (c) Samples from the posterior over functions, which are induced by the samples from the parameter posterior.

9.4 Maximum Likelihood as Orthogonal Projection

In the following, we will provide a geometric interpretation of maximum likelihood estimation.

Let us look at a simple linear regression setting

$$y = x\theta + \epsilon, \quad \epsilon \sim \mathcal{N}\left(0, \sigma^2\right), \tag{9.88}$$

in which we consider linear functions $f : \mathbb{R} \to \mathbb{R}$ that go through the origin (we omit features here for clarity). The parameter θ determines the slope of the line. Figure 9.12 illustrates a one dimensional dataset.

With a training data set $\boldsymbol{X} = [x_1, \ldots, x_N]^\top \in \mathbb{R}^N$, $\boldsymbol{y} = [y_1, \ldots, y_N]^\top \in \mathbb{R}^N$, we recall the results from Section 9.2.1 and obtain the maximum likelihood estimator for the slope parameter as

$$\theta_{\text{ML}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y} = \frac{\boldsymbol{X}^\top \boldsymbol{y}}{\boldsymbol{X}^\top \boldsymbol{X}} \in \mathbb{R}\,. \tag{9.89}$$

This means for the training inputs $\boldsymbol{X}$ we obtain the optimal (maximum likelihood) reconstruction of the training data, i.e., the approximation with the minimum least squares error

$$\boldsymbol{X}\theta_{\text{ML}} = \boldsymbol{X}\frac{\boldsymbol{X}^\top \boldsymbol{y}}{\boldsymbol{X}^\top \boldsymbol{X}} = \frac{\boldsymbol{X}\boldsymbol{X}^\top}{\boldsymbol{X}^\top \boldsymbol{X}}\boldsymbol{y}\,. \tag{9.90}$$

As we are basically looking for a solution of the linear equation system $\boldsymbol{y} = \boldsymbol{X}\theta$, we can think of linear regression as a problem for solving linear equation systems. Therefore, we can relate to concepts from linear algebra and analytic geometry that we discussed in Chapters 2 and 3. In particular, looking carefully at (9.90) we see that the maximum likelihood estimator θ_{ML} in our example from (9.88) effectively does an orthogonal projection of $\boldsymbol{y}$ onto the one-dimensional subspace spanned by $\boldsymbol{X}$. Recalling the results on orthogonal projections from Section 3.6, we identify $\frac{\boldsymbol{X}\boldsymbol{X}^\top}{\boldsymbol{X}^\top \boldsymbol{X}}$ as the

Linear regression with maximum likelihood parameters performs an orthogonal projection.

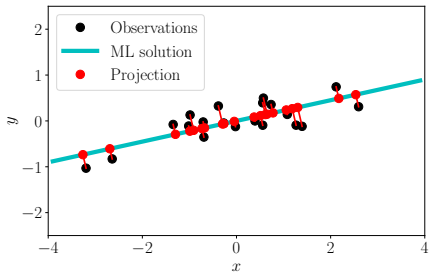


Figure 9.12 Geometric interpretation of least squares. The red dots are the projections of the noisy observations (black dots) onto the line $\theta_{\text{ML}}x$. The maximum likelihood solution to a linear regression problem finds a subspace (line) onto which the projection error (red lines) of the observations is minimized.

projection matrix, θ_{ML} as the coordinates of the projection onto the one-dimensional subspace of $\mathbb{R}^N$ spanned by $\boldsymbol{X}$ and $\boldsymbol{X}\theta_{\text{ML}}$ as the orthogonal projection of $\boldsymbol{y}$ onto this subspace.

Therefore, the maximum likelihood solution provides also a geometrically optimal solution by finding the vectors in the subspace spanned by $\boldsymbol{X}$ that are "closest" to the corresponding observations $\boldsymbol{y}$, where "closest" means the smallest (squared) distance of the function values y_n to $x_n\theta$. This is achieved by orthogonal projections.

In the general linear regression case where

$$y = \boldsymbol{\phi}^\top(\boldsymbol{x})\boldsymbol{\theta} + \epsilon, \quad \epsilon \sim \mathcal{N}\left(0, \sigma^2\right) \tag{9.91}$$

with vector-valued features $\boldsymbol{\phi}(\boldsymbol{x}) \in \mathbb{R}^K$, we again can interpret the maximum likelihood result

$$\boldsymbol{y} \approx \boldsymbol{\Phi}\boldsymbol{\theta}_{\text{ML}}\,, \tag{9.92}$$

$$\boldsymbol{\theta}_{\text{ML}} = \boldsymbol{\Phi}(\boldsymbol{\Phi}^\top\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^\top\boldsymbol{y} \tag{9.93}$$

as a projection onto a K-dimensional subspace of $\mathbb{R}^N$, which is spanned by the columns of the feature matrix $\boldsymbol{\Phi}$, see Section 3.6.2.

If the feature functions ϕ_k that we use to construct the feature matrix $\boldsymbol{\Phi}$ are orthonormal (see Section 3.5), we obtain a special case where the columns of $\boldsymbol{\Phi}$ form an orthonormal basis (see Section 3.7), such that $\boldsymbol{\Phi}^\top\boldsymbol{\Phi} = \boldsymbol{I}$. This will then lead to the projection

$$\boldsymbol{\Phi}(\boldsymbol{\Phi}^\top\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}\boldsymbol{y} = \boldsymbol{\Phi}\boldsymbol{\Phi}^\top\boldsymbol{y} = \left(\sum_{k=1}^{K}\boldsymbol{\phi}_k\boldsymbol{\phi}_k^\top\right)\boldsymbol{y} \tag{9.94}$$

so that the coupling between different features has disappeared and the maximum likelihood projection is simply the sum of projections of $\boldsymbol{y}$ onto the individual basis vectors $\boldsymbol{\phi}_k$, i.e., the columns of $\boldsymbol{\Phi}$. Many popular basis functions in signal processing, such as wavelets and Fourier bases, are orthogonal basis functions. When the basis is not orthogonal, one can convert a set of linearly independent basis functions to an orthogonal basis by using the Gram- Schmidt process.

9.5 Further Reading

In this chapter, we discussed linear regression for Gaussian likelihoods and conjugate Gaussian priors on the parameters of the model. This allowed for closed-form Bayesian inference. However, in some applications we may want to choose a different likelihood function. For example, in a binary *classification* setting, we observe only two possible (categorical) outcomes, and a Gaussian likelihood is inappropriate in this setting. Instead, we can choose a Bernoulli likelihood that will return a probability of the predicted label to be 1 (or 0). We refer to the books by Bishop (2006); Murphy (2012); Barber (2012) for an in-depth introduction to classification problems. A different example where non-Gaussian likelihoods are important is count data. Counts are non-negative integers, and in this case a Binomial or Poisson likelihood would be a better choice than a Gaussian. All these examples fall into the category of *generalized linear models*, a flexible generalization of linear regression that allows for response variables that have error distribution models other than a Gaussian distribution. The GLM generalizes linear regression by allowing the linear model to be related to the observed values via a smooth and invertible function $\sigma(\cdot)$ that may be nonlinear so that $y = \sigma(f)$, where $f = \boldsymbol{\theta}^\top \boldsymbol{\phi}(\boldsymbol{x})$ is the linear regression model from (9.19). We can therefore think of a generalized linear model in terms of function composition $y = \sigma \circ f$ where f is a linear regression model and σ the activation function. Note, that although we are talking about "generalized linear models" the outputs y are no longer linear in the parameters $\boldsymbol{\theta}$. In *logistic regression*, we choose the *logistic sigmoid* $\sigma(f) = \frac{1}{1+\exp(-f)} \in [0, 1]$, which can be interpreted as the probability of observing a binary output $y = 1$ of a Bernoulli random variable. The function $\sigma(\cdot)$ is called *transfer function* or *activation function*, its inverse is called the *canonical link function*. From this perspective, it is also clear that generalized linear models are the building blocks of (deep) feedforward neural networks: If we consider a generalized linear model $\boldsymbol{y} = \sigma(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b})$, where $\boldsymbol{A}$ is a weight matrix and $\boldsymbol{b}$ a bias vector, we identify this generalized linear model as a single-layer neural network with activation function $\sigma(\cdot)$. We can now recursively compose these functions via

$$\begin{aligned} \boldsymbol{x}_{k+1} &= \boldsymbol{f}_k(\boldsymbol{x}_k) \\ \boldsymbol{f}_k(\boldsymbol{x}_k) &= \sigma_k(\boldsymbol{A}_k \boldsymbol{x}_k + \boldsymbol{b}_k) \end{aligned} \tag{9.95}$$

for $k = 0, \ldots, K-1$ where $\boldsymbol{x}_0$ are the input features and $\boldsymbol{x}_K = \boldsymbol{y}$ are the observed outputs, such that $\boldsymbol{f}_{K-1} \circ \cdots \circ \boldsymbol{f}_0$ is a K-layer deep neural network. Therefore, the building blocks of this deep neural network are the generalized linear models defined in (9.95). A great post on the relation between GLMs and deep networks is available at `https://tinyurl.com/glm-dnn`. Neural networks (Bishop, 1995; Goodfellow et al., 2016) are significantly more expressive and flexible than linear re-

classification

generalized linear models

logistic regression

logistic sigmoid

transfer function

activation function

canonical link function

For ordinary linear regression the activation function would simply be the identity.

Generalized linear models are the building blocks of deep neural networks.

gression models. However, maximum likelihood parameter estimation is a non-convex optimization problem, and marginalization of the parameters in a fully Bayesian setting is analytically intractable.

We briefly hinted at the fact that a distribution over parameters induces a distribution over regression functions. *Gaussian processes* (Rasmussen and Williams, 2006) are regression models where the concept of a distribution over function is central. Instead of placing a distribution over parameters a Gaussian process places a distribution directly on the space of functions without the "detour" via the parameters. To do so, the Gaussian process exploits the *kernel trick* (Schölkopf and Smola, 2002), which allows us to compute inner products between two function values $f(\boldsymbol{x}_i), f(\boldsymbol{x}_j)$ only by looking at the corresponding input $\boldsymbol{x}_i, \boldsymbol{x}_j$. A Gaussian process is closely related to both Bayesian linear regression and support vector regression but can also be interpreted as a Bayesian neural network with a single hidden layer where the number of units tends to infinity (Neal, 1996; Williams, 1997). An excellent introduction to Gaussian processes can be found in (MacKay, 1998; Rasmussen and Williams, 2006).

Gaussian processes

kernel trick

We focused on Gaussian parameter priors in the discussions in this chapters because they allow for closed-form inference in linear regression models. However, even in a regression setting with Gaussian likelihoods we may choose a non-Gaussian prior. Consider a setting where the inputs are $\boldsymbol{x} \in \mathbb{R}^D$ and our training set is small and of size $N \ll D$. This means that the regression problem is under-determined. In this case, we can choose a parameter prior that enforces sparsity, i.e., a prior that tries to set as many parameters to 0 as possible (*variable selection*). This prior provides a stronger regularizer than the Gaussian prior, which often leads to an increased prediction accuracy and interpretability of the model. The Laplace prior is one example that is frequently used for this purpose. A linear regression model with the Laplace prior on the parameters is equivalent to linear regression with L1 regularization (*LASSO*) (Tibshirani, 1996). The Laplace distribution is sharply peaked at zero (its first derivative is discontinuous) and it concentrates its probability mass closer to zero than the Gaussian distribution, which encourages parameters to be 0. Therefore, the non-zero parameters are relevant for the regression problem, which is the reason why we also speak of "variable selection".

variable selection

LASSO