the estimation problem can be reduced to a nonlinear function minimization problem, as in the regression case 16.3.4. We give examples of such techniques in Section 16.4.

### 16.3.6 Adjusting Hyperparameters for Gaussian Processes

More often than not, we will not know beforehand the exact amount of additive noise, or the specific form of the covariance kernel. To address this problem, the hyperparameter formalism of Section 16.1.4 is needed. To avoid technicalities, we only discuss the application of the MAP2 estimate (16.24) for the special case of regression with additive Gaussian noise, and refer the reader to [600, 193, 151, 426] and the references therein for integration methods based on Markov Chain Monte Carlo approximations (see also [486] for a more recent overview).

We denote by $\omega$ the set of hyperparameters we would like to adjust. In more compact notation, (16.48) becomes (now conditioned on $\omega$)

$$p(\mathbf{y}|\omega) = \frac{1}{\sqrt{(2\pi)^m \det(K + \sigma^2 \mathbf{1})}} \exp\left(-\frac{1}{2}\mathbf{y}^\top (K + \sigma^2 \mathbf{1})^{-1}\mathbf{y}\right) \tag{16.55}$$

where $K, \sigma$ are functions of $\omega$. In other words, (16.55) tells us how likely it is that we observe $\mathbf{y}$, if we know $\omega$.

Recall that the basic idea of the MAP2 estimate (16.24) is to maximize $p(\omega|\mathbf{y})$ by maximizing $p(\mathbf{y}|\omega)p(\omega)$. In practice, this is achieved by gradient ascent (see Section 6.2.2) or second order methods (see Section 6.2.1 for Newton's method) on $p(\mathbf{y}|\omega)p(\omega)$. Both cases require information about the gradient of (16.55) with respect to $\omega$. We give an explicit expression for the gradient below.

Gradient wrt. Hyperparameters

Since the logarithm is monotonic, we can equivalently minimize the negative log posterior, $\ln p(\mathbf{y}|\omega)p(\omega)$. With the shorthand $Q := K + \sigma^2\mathbf{1}$, we obtain

$$\begin{aligned} &\partial_\omega \left[-\ln p(\mathbf{y}|\omega)p(\omega)\right] \\ &= \frac{1}{2}\partial_\omega(\ln\det Q) - \frac{1}{2}\partial_\omega\left[\mathbf{y}^\top Q^{-1}\mathbf{y}\right] - \partial_\omega \ln p(\omega) \qquad (16.56) \\ &= -\frac{1}{2}\operatorname{tr}\left(Q^{-1}\partial_\omega Q\right) + \frac{1}{2}\mathbf{y}^\top Q^{-1}(\partial_\omega Q)\,Q^{-1}\mathbf{y} - \partial_\omega \ln p(\omega). \qquad (16.57) \end{aligned}$$

Here (16.57) follows from (16.56) via standard matrix algebra [337]. Likewise, we could compute the Hessian of $\ln p(\mathbf{y}|\omega)p(\omega)$ with respect to $\omega$ and use a second order optimization method.[6]

Flat Hyperprior

If we assume a flat hyperprior ($p(\omega) = \text{const.}$), optimization over $\omega$ simply becomes gradient descent in $-\ln p(\mathbf{y}|\omega)$; in other words, the term depending on $p(\omega)$ vanishes. Computing (16.57) is still very expensive numerically since it involves the inversion of $Q$, which is an $m \times m$ matrix.

There exist numerous techniques, such as sparse greedy approximation meth-

6. This is rather technical, and the reader is encouraged to consult the literature for further detail [339, 426, 383, 197].