CHAPTER

17 BAYESIAN INFERENCE

CHAPTER CONTENTS

In the framework of MLE, parameter $\boldsymbol{\theta}$ in parametric model $q(\boldsymbol{x};\boldsymbol{\theta})$ was treated as a deterministic variable. In this chapter, *Bayesian inference* [15] is introduced which handles parameter $\boldsymbol{\theta}$ as a random variable.

17.1 BAYESIAN PREDICTIVE DISTRIBUTION

In this section, the basic idea of Bayesian inference is explained.

17.1.1 DEFINITION

If $\boldsymbol{\theta}$ is regarded as a random variable, the following probabilities can be determined:

$$p(\boldsymbol{\theta}),\ \ p(\boldsymbol{\theta}|\mathcal{D}),\ \ p(\mathcal{D}|\boldsymbol{\theta}),\ \ \text{and}\ \ p(\mathcal{D},\boldsymbol{\theta}),$$

where

$$\mathcal{D} = \{\boldsymbol{x}_i\}_{i=1}^n.$$

$p(\boldsymbol{\theta}|\mathcal{D})$ is called the *posterior probability* of parameter $\boldsymbol{\theta}$ given training samples $\mathcal{D}$, while $p(\boldsymbol{\theta})$ is called the *prior probability* of $\boldsymbol{\theta}$ before observing training samples $\mathcal{D}$. $p(\mathcal{D}|\boldsymbol{\theta})$ denotes the *likelihood*, which is the same quantity as the one used in MLE (see Section 12.1), but it is regarded as a conditional probability in the Bayesian framework:

$$p(\mathcal{D}|\boldsymbol{\theta}) = \prod_{i=1}^{n} q(\boldsymbol{x}_i|\boldsymbol{\theta}).$$

An Introduction to Statistical Machine Learning. DOI: 10.1016/B978-0-12-802121-7.00028-5

Note that the parametric model $q(\boldsymbol{x}|\boldsymbol{\theta})$ is also represented as a conditional probability in the Bayesian framework. $p(\mathcal{D},\boldsymbol{\theta})$ denotes the *joint probability* of training samples $\mathcal{D}$ and parameter $\boldsymbol{\theta}$.

The joint probability $p(\mathcal{D},\boldsymbol{\theta})$ can be expressed as

$$p(\mathcal{D},\boldsymbol{\theta}) = p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta}),$$

and its marginalization over $\boldsymbol{\theta}$ gives

$$\int p(\mathcal{D},\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} = p(\mathcal{D}).$$

Thus, the marginal probability $p(\mathcal{D})$ can be expressed as

$$p(\mathcal{D}) = \int \left(\prod_{i=1}^{n} q(\boldsymbol{x}_i|\boldsymbol{\theta})\right) p(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}.$$

The solution of Bayesian inference $\widehat{p}_{\mathrm{Bayes}}(\boldsymbol{x})$, called the *Bayesian predictive distribution*, is given as the expectation of model $q(\boldsymbol{x}|\boldsymbol{\theta})$ over the posterior probability $p(\boldsymbol{\theta}|\mathcal{D})$:

$$\widehat{p}_{\mathrm{Bayes}}(\boldsymbol{x}) = \int q(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathcal{D})\mathrm{d}\boldsymbol{\theta}. \tag{17.1}$$

Since the posterior probability $p(\boldsymbol{\theta}|\mathcal{D})$ can be expressed by using *Bayes' theorem* (see Section 5.4) as

$$\begin{aligned} p(\boldsymbol{\theta}|\mathcal{D}) &= \frac{p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathcal{D})} \\ &= \frac{\prod_{i=1}^{n} q(\boldsymbol{x}_i|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int \prod_{i=1}^{n} q(\boldsymbol{x}_i|\boldsymbol{\theta}')p(\boldsymbol{\theta}')\mathrm{d}\boldsymbol{\theta}'}, \end{aligned} \tag{17.2}$$

the Bayesian predictive distribution $\widehat{p}_{\mathrm{Bayes}}(\boldsymbol{x})$ can be expressed as

$$\widehat{p}_{\mathrm{Bayes}}(\boldsymbol{x}) = \int q(\boldsymbol{x}|\boldsymbol{\theta}) \frac{\prod_{i=1}^{n} q(\boldsymbol{x}_i|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int \prod_{i=1}^{n} q(\boldsymbol{x}_i|\boldsymbol{\theta}')p(\boldsymbol{\theta}')\mathrm{d}\boldsymbol{\theta}'}\mathrm{d}\boldsymbol{\theta}. \tag{17.3}$$

This shows that the Bayesian predictive distribution $\widehat{p}_{\mathrm{Bayes}}(\boldsymbol{x})$ can actually be computed *without any learning*, if parametric model $q(\boldsymbol{x}|\boldsymbol{\theta})$ and prior probability $p(\boldsymbol{\theta})$ are specified.

17.1.2 COMPARISON WITH MLE

In MLE, unknown true probability density $p(\boldsymbol{x})$ is approximated by a parametric model $q(\boldsymbol{x};\boldsymbol{\theta})$ with a single parameter estimate $\widehat{\boldsymbol{\theta}}_{\mathrm{ML}}$. On the other hand, in Bayesian inference, infinitely many parameters are simultaneously considered and their

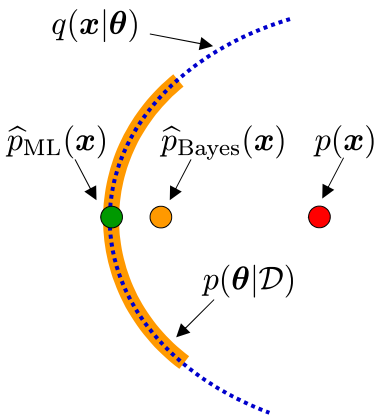


FIGURE 17.1

Bayes vs. MLE. The maximum likelihood solution $\widehat{p}_{\mathrm{ML}}$ is always confined in the parametric model $q(\boldsymbol{x};\boldsymbol{\theta})$, while the Bayesian predictive distribution $\widehat{p}_{\mathrm{Bayes}}(\boldsymbol{x})$ generally pops out from the model.

average over model $q(\boldsymbol{x}|\boldsymbol{\theta})$ weighted according to the posterior probability $p(\boldsymbol{\theta}|\mathcal{D})$ is used as a density estimator. Let us intuitively explain the difference between Bayesian inference and MLE using Fig. 17.1.

A parametric model $q(\boldsymbol{x}|\boldsymbol{\theta})$ is a set of probability density functions and it is denoted by a dotted line in Fig. 17.1. In practice, parametric model $q(\boldsymbol{x}|\boldsymbol{\theta})$ is more or less *misspecified*, meaning that the true probability density $p(\boldsymbol{x})$ is not exactly included in the parametric model $q(\boldsymbol{x}|\boldsymbol{\theta})$. MLE finds the probability density function in the parametric model that maximizes the likelihood, $\widehat{p}_{\mathrm{ML}}(\boldsymbol{x})$, which is equivalent to finding the projection of $p(\boldsymbol{x})$ onto the model under the empirical KL divergence (see Section 14.2). On the other hand, in Bayesian inference, by taking the expectation of parametric model $q(\boldsymbol{x}|\boldsymbol{\theta})$ over the posterior probability $p(\boldsymbol{\theta}|\mathcal{D})$, the solution $\widehat{p}_{\mathrm{Bayes}}(\boldsymbol{x})$ is not generally confined in the model. In the illustration in Fig. 17.1, the solution $\widehat{p}_{\mathrm{Bayes}}(\boldsymbol{x})$ pops out from the model to the right-hand side, and consequently it is closer to the true probability density $p(\boldsymbol{x})$ than $\widehat{p}_{\mathrm{ML}}(\boldsymbol{x})$.

The fundamental difference between Bayesian inference and MLE lies in whether parameter $\boldsymbol{\theta}$ is handled as a deterministic or random variable. However, in reality, more significant philosophical difference is involved. More specifically, prior probability $p(\boldsymbol{\theta})$ can contain subjective knowledge in Bayesian inference, which can arbitrarily change the solution. On the other hand, MLE is objective and its solution is purely computed from data. When non-Bayesian inference is contrasted to Bayesian inference, it is sometimes referred to as *frequentist inference*.

17.1.3 COMPUTATIONAL ISSUES

As explained in Section 17.1, the Bayesian predictive distribution $\widehat{p}_{\mathrm{Bayes}}(\boldsymbol{x})$ can be computed without any learning in principle, if parametric model $q(\boldsymbol{x}|\boldsymbol{\theta})$ and prior probability $p(\boldsymbol{\theta})$ are specified:

$$\widehat{p}_{\mathrm{Bayes}}(\boldsymbol{x}) = \int q(\boldsymbol{x}|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathcal{D}) \mathrm{d}\boldsymbol{\theta}, \tag{17.4}$$

where

$$p(\boldsymbol{\theta}|\mathcal{D}) = \frac{\prod_{i=1}^{n} q(\boldsymbol{x}_i|\boldsymbol{\theta}) p(\boldsymbol{\theta})}{\int \prod_{i=1}^{n} q(\boldsymbol{x}_i|\boldsymbol{\theta}') p(\boldsymbol{\theta}') \mathrm{d}\boldsymbol{\theta}'}. \tag{17.5}$$

However, computation of the two integrations above is not straightforward if the dimension of $\boldsymbol{\theta}$ is high. Thus, a main technical challenge in Bayesian inference is how to efficiently compute high-dimensional integrations.

To easily handle the integration in Eq.(17.4), it is preferable to obtain the posterior probability $p(\boldsymbol{\theta}|\mathcal{D})$ *analytically*. One possibility is to choose the prior probability $p(\boldsymbol{\theta})$ so that the parametric form of the posterior probability $p(\boldsymbol{\theta}|\mathcal{D})$ can be explicitly obtained. Such a prior choice will be explained in Section 17.2. For nonconjugate choice of prior probabilities, analytic approximation techniques of the integration in Eq. (17.5) will be discussed in Chapter 18.

For handling the integration in Eq. (17.4), i.e., the expectation of parametric model $q(\boldsymbol{x}|\boldsymbol{\theta})$ over posterior probability $p(\boldsymbol{\theta}|\mathcal{D})$, the simplest approximation scheme would be to use a single point $\widehat{\boldsymbol{\theta}}$ taken from the posterior probability. Such a single-point approximation will be introduced in Section 17.3. Techniques for numerically approximating the posterior expectation will be discussed in Chapter 19.

17.2 CONJUGATE PRIOR

As discussed above, it is convenient to analytically handle the integration in Eq. (17.5). If the prior probability $p(\boldsymbol{\theta})$ is chosen so that the posterior probability $p(\boldsymbol{\theta}|\mathcal{D})$ takes the same parametric form as the prior probability $p(\boldsymbol{\theta})$, the posterior probability $p(\boldsymbol{\theta}|\mathcal{D})$ can be analytically obtained just by specifying its parameters. Such a prior choice is called a *conjugate prior* for the likelihood $p(\mathcal{D}|\boldsymbol{\theta})$.

Let us illustrate an example of the conjugate prior for the Gaussian model with expectation 0 and variance σ^2, where the inverse of the variance $\tau = \sigma^{-2}$, called the *precision*, is regarded as a parameter (i.e., the parameter is $\theta = \tau$). Then the parametric model is expressed as

$$q(x|\tau) = \sqrt{\frac{\tau}{2\pi}} \exp\left(-\frac{\tau x^2}{2}\right). \tag{17.6}$$

For this model, let us employ the *gamma distribution* (see Section 4.3) as the prior probability for precision τ:

$$p(\tau; \alpha, \beta) \propto \tau^{\alpha-1} e^{-\beta\tau}. \tag{17.7}$$

For parametric model (17.6) combined with prior probability (17.7), the posterior probability is again the gamma distribution:

$$\begin{aligned} p(\tau|\mathcal{D}) &\propto \prod_{i=1}^{n} q(x_i|\tau)p(\tau;\alpha,\beta) \\ &\propto \tau^{n/2}\exp\left(-\frac{\tau}{2}\sum_{i=1}^{n}x_i^2\right)\tau^{\alpha-1}e^{-\beta\tau} \\ &= \tau^{\widetilde{\alpha}-1}e^{-\widetilde{\beta}\tau}, \end{aligned}$$

where the posterior parameters $\widetilde{\alpha}$ and $\widetilde{\beta}$ are given as

$$\widetilde{\alpha} = \alpha + \frac{n}{2} \quad \text{and} \quad \widetilde{\beta} = \beta + \frac{\sum_{i=1}^{n} x_i^2}{2}.$$

Thus, the posterior probability can be obtained analytically just by computing the posterior parameters $\widetilde{\alpha}$ and $\widetilde{\beta}$.

As shown above, conjugate priors are extremely useful from the viewpoint of computation. However, conjugate priors depend on the parametric form of likelihood $p(\mathcal{D}|\boldsymbol{\theta})$, and they may not be available depending on $p(\mathcal{D}|\boldsymbol{\theta})$. Furthermore, the meaning of choosing conjugate priors is not clear from the viewpoint of statistical inference. For general nonconjugate priors, analytic approximation techniques of the posterior probability will be introduced in Chapter 18.

17.3 MAP ESTIMATION

Given the posterior probability $p(\boldsymbol{\theta}|\mathcal{D})$ analytically, the next step is to compute the posterior expectation:

$$\widehat{p}_{\mathrm{Bayes}}(\boldsymbol{x}) = \int q(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathcal{D})\mathrm{d}\boldsymbol{\theta}.$$

In this section, MAP estimation is introduced, which approximates the above integration by a single point $\widehat{\boldsymbol{\theta}}_{\mathrm{MAP}}$:

$$\widehat{p}_{\mathrm{MAP}}(\boldsymbol{x}) = q(\boldsymbol{x}|\widehat{\boldsymbol{\theta}}_{\mathrm{MAP}}),$$

where $\widehat{\boldsymbol{\theta}}_{\mathrm{MAP}}$ is the maximizer of the posterior probability $p(\boldsymbol{\theta}|\mathcal{D})$ (see Fig. 17.2):

$$\widehat{\boldsymbol{\theta}}_{\mathrm{MAP}} = \operatorname*{argmax}_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|\mathcal{D}).$$

Since MAP estimation approximates the target density by a single point $\widehat{\boldsymbol{\theta}}_{\mathrm{MAP}}$, its property is actually close to MLE. Indeed, the MAP solution $\widehat{\boldsymbol{\theta}}_{\mathrm{MAP}}$ can be expressed as

$$\widehat{\boldsymbol{\theta}}_{\mathrm{MAP}} = \operatorname*{argmax}_{\boldsymbol{\theta}} \left(\sum_{i=1}^{n} \log q(\boldsymbol{x}_i|\boldsymbol{\theta}) + \log p(\boldsymbol{\theta})\right), \tag{17.8}$$

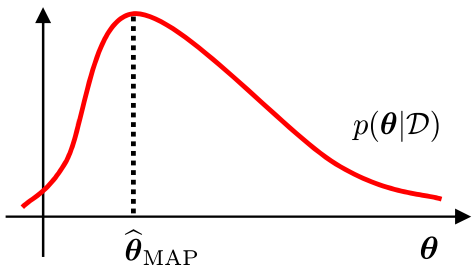
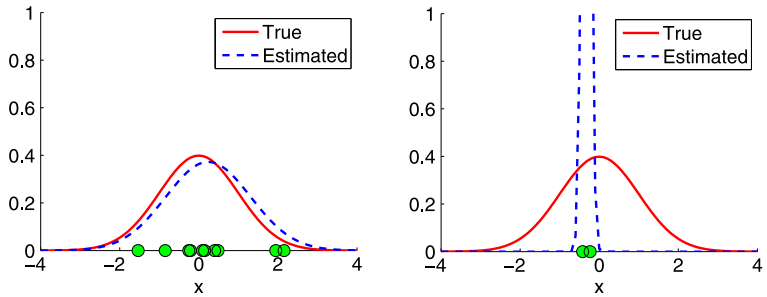


FIGURE 17.2

MAP estimation.



(a) When $n = 10$ (b) When $n = 2$

FIGURE 17.3

Example of MLE for Gaussian model. When the number of training samples, n, is small, MLE tends to overfit the samples.

and thus it actually minimizes the sum of the log-likelihood and an additional term $\log p(\boldsymbol{\theta})$.

As explained in Chapter 13, MLE tends to *overfit* the training samples if the sample size is small (Fig. 17.3). The additional term $\log p(\boldsymbol{\theta})$ in Eq. (17.8) can work as a penalty to mitigate overfitting. For this reason, MAP estimation is also referred to as *penalized MLE*. MAP estimation tries to increase not only the likelihood but also the prior probability, and therefore the solution tends to be biased toward the parameter having a larger prior probability. Penalizing the objective function in this way is also called *regularization* (see Chapter 23 for details).

Let us specifically compute the MAP solution for the Gaussian model with expectation $\boldsymbol{\mu}$ and variance-covariance matrix $\boldsymbol{I}_d$ (i.e., the parameter is $\boldsymbol{\theta} = \boldsymbol{\mu}$):

$$q(\boldsymbol{x}|\boldsymbol{\mu}) = \frac{1}{(2\pi)^{\frac{d}{2}}} \exp\left(-\frac{(\boldsymbol{x}-\boldsymbol{\mu})^\top(\boldsymbol{x}-\boldsymbol{\mu})}{2}\right).$$

Let us consider the following Gaussian prior:

$$p(\boldsymbol{\mu};\beta) = \frac{1}{(2\pi\beta^2)^{\frac{d}{2}}} \exp\left(-\frac{\boldsymbol{\mu}^\top \boldsymbol{\mu}}{2\beta^2}\right), \tag{17.9}$$

which prefers $\boldsymbol{\mu}$ closer to the origin. For this setup, the penalized log-likelihood is given by

$$\begin{aligned}\mathrm{PL}(\boldsymbol{\mu}) &= \sum_{i=1}^{n} \log q(\boldsymbol{x}_i|\boldsymbol{\mu}) + \log p(\boldsymbol{\mu}) \\ &= -\frac{nd}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{n}\|\boldsymbol{x}_i - \boldsymbol{\mu}\|^2 - \frac{d}{2}\log(2\pi\beta^2) - \frac{1}{2\beta^2}\|\boldsymbol{\mu}\|^2.\end{aligned}$$

Taking the derivative of $\mathrm{PL}(\boldsymbol{\mu})$ and setting it to zero yield

$$\frac{\partial}{\partial \boldsymbol{\mu}}\mathrm{PL}(\boldsymbol{\mu}) = \sum_{i=1}^{n}(\boldsymbol{x}_i - \boldsymbol{\mu}) - \frac{1}{\beta^2}\boldsymbol{\mu} = \mathbf{0}_d,$$

from which the MAP solution $\widehat{\boldsymbol{\mu}}_{\mathrm{MAP}}$ is obtained as

$$\widehat{\boldsymbol{\mu}}_{\mathrm{MAP}} = \frac{1}{n+\beta^{-2}}\sum_{i=1}^{n}\boldsymbol{x}_i.$$

On the other hand, the maximum likelihood solution $\widehat{\boldsymbol{\mu}}_{\mathrm{MLE}}$ for this model is given by

$$\widehat{\boldsymbol{\mu}}_{\mathrm{MLE}} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i.$$

$\beta > 0$ implies

$$\left|\widehat{\mu}^{(j)}_{\mathrm{MAP}}\right| = \frac{1}{n+\beta^{-2}}\left|\sum_{i=1}^{n} x_i^{(j)}\right| < \frac{1}{n}\left|\sum_{i=1}^{n} x_i^{(j)}\right| = \left|\widehat{\mu}^{(j)}_{\mathrm{MLE}}\right|,$$

where $x_i^{(j)}$, $\widehat{\mu}^{(j)}_{\mathrm{MAP}}$, and $\widehat{\mu}^{(j)}_{\mathrm{MLE}}$ denote the jth elements of vectors $\boldsymbol{x}_i$, $\widehat{\boldsymbol{\mu}}_{\mathrm{MAP}}$, and $\widehat{\boldsymbol{\mu}}_{\mathrm{MLE}}$, respectively. Thus, the MAP solution $\widehat{\boldsymbol{\mu}}_{\mathrm{MAP}}$ is always closer to the origin than the maximum likelihood solution $\widehat{\boldsymbol{\mu}}_{\mathrm{MLE}}$, which the Gaussian prior (17.9) favors.

A MATLAB code for penalized MLE with one-dimensional Gaussian model is given in Fig. 17.4, and its behavior is illustrated in Fig. 17.5. This demonstrates that, if the prior probability is chosen properly (i.e., $\beta \approx 1$), MAP estimation can give a better solution than MLE.

In MAP estimation, the maximizer of the posterior probability (i.e., the *mode*) was used. An alternative idea is to use the mean of the posterior probability:

$$\widehat{p}(\boldsymbol{x}) = q(\boldsymbol{x}|\overline{\boldsymbol{\theta}}),$$

```
n=12; mu=0.5; x=randn(n,1)+mu;
bs=[0.01:0.01:3]; bl=length(bs);
MLE=mean(x);
for i=1:bl
  MAP(i)=sum(x)/(n+bs(i).^(-2));
end

figure(1); clf; hold on;
plot(bs,mu*ones(1,bl),'k:');
plot(bs,MAP,'r-');
plot(bs,MLE*ones(1,bl),'b-.');
xlabel('\beta'); ylabel('\mu');
legend('True','MAP','MLE',4);
```

FIGURE 17.4

MATLAB code for penalized MLE with one-dimensional Gaussian model.

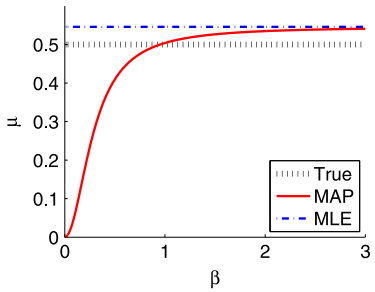


FIGURE 17.5

Example of MAP estimation with one-dimensional Gaussian model.

where

$$\overline{\boldsymbol{\theta}} = \int \boldsymbol{\theta} p(\boldsymbol{\theta}|\mathcal{D}) \mathrm{d}\boldsymbol{\theta}.$$

If the posterior probability $p(\boldsymbol{\theta}|\mathcal{D})$ is a popular probability distribution, its expectation may be known analytically and then the posterior expectation $\overline{\boldsymbol{\theta}}$ can also be obtained analytically.

However, single-point approximation of the Bayesian predictive distribution $\widehat{p}_{\text{Bayes}}(\boldsymbol{x})$ loses the distinctive feature of Bayesian inference that the solution $\widehat{p}_{\text{Bayes}}(\boldsymbol{x})$

can pop out from the parametric model $q(\boldsymbol{x}|\boldsymbol{\theta})$ (see Fig. 17.1). To enjoy Bayesianity, it is essential to compute the integration at least approximately. In Chapter 19, techniques for numerically approximating the posterior expectation will be discussed.

17.4 BAYESIAN MODEL SELECTION

In Bayesian inference, the prior probability of parameters is utilized. If such prior knowledge is not available, the prior probability has to be determined by a user. Since the solution of Bayesian inference depends on the choice of the prior probability (see Fig. 17.5), it must be determined in an objective and appropriate way. In this section, choice of prior probabilities and models in the Bayesian framework is addressed.

Suppose that the prior probability is parameterized by $\boldsymbol{\beta}$:

$$p(\boldsymbol{\theta};\boldsymbol{\beta}),$$

where $\boldsymbol{\beta}$ is called the *hyperparameter* and this should be distinguished from ordinary parameter $\boldsymbol{\theta}$ in the parametric model $q(\boldsymbol{x}|\boldsymbol{\theta})$. As explained in Section 14.3, model selection of MLE is possible by *cross validation*, and it can also be applied to Bayesian inference. Below, an alternative approach that is specific to Bayesian inference is introduced.

MLE is aimed at setting parameter $\boldsymbol{\theta}$ so that training samples $\{\boldsymbol{x}\}_{i=1}^n$ at hand are most typically generated. The fundamental idea of Bayesian model selection is to apply MLE to hyperparameter $\boldsymbol{\beta}$. More specifically, the probability that training samples $\mathcal{D} = \{\boldsymbol{x}\}_{i=1}^n$ at hand are generated is expressed as

$$p(\mathcal{D};\boldsymbol{\beta}) = \int \prod_{i=1}^{n} q(\boldsymbol{x}_i|\boldsymbol{\theta})p(\boldsymbol{\theta};\boldsymbol{\beta})\mathrm{d}\boldsymbol{\theta} = \mathrm{ML}(\boldsymbol{\beta}). \tag{17.10}$$

Eq. (17.10) viewed as a function of $\boldsymbol{\beta}$ is called the *marginal likelihood*, which is also referred to as the *evidence* and its negative log is called the *free energy*.

The method of determining hyperparameter $\boldsymbol{\beta}$ so that the marginal likelihood is maximized is called *empirical Bayes*, *type-II MLE*, or *evidence maximization*:

$$\boldsymbol{\beta}_{EB} = \operatorname*{argmax}_{\boldsymbol{\beta}} \mathrm{ML}(\boldsymbol{\beta}).$$

In addition to the hyperparameter, parametric model $q(\boldsymbol{x}|\boldsymbol{\theta})$ may also be selected by empirical Bayes. That is, among a set of model candidates, the one that maximizes the marginal likelihood may be selected as the most promising one.

Let us specifically compute the marginal likelihood for one-dimensional Gaussian parametric model with expectation μ and variance 1,

$$q(x|\mu) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2}\right), \tag{17.11}$$

and a Gaussian prior probability with expectation 0 and variance β^2:

$$p(\mu;\beta) = \frac{1}{\sqrt{2\pi\beta^2}} \exp\left(-\frac{\mu^2}{2\beta^2}\right). \tag{17.12}$$

The marginal likelihood can be expressed as

$$\begin{aligned}\mathrm{ML}(\beta) &= \int \prod_{i=1}^{n} q(x_i|\mu)p(\mu;\beta)\mathrm{d}\mu \\ &= (2\pi)^{-\frac{n}{2}}(2\pi\beta^2)^{-\frac{1}{2}} \int \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2 - \frac{\mu^2}{2\beta^2}\right)\mathrm{d}\mu.\end{aligned}$$

For

$$\widehat{\mu}_{\mathrm{MAP}} = \frac{1}{n+\beta^{-2}}\sum_{i=1}^{n} x_i,$$

completing the square (see Eq. (4.2)) yields

$$\begin{aligned}\mathrm{ML}(\beta) &= (2\pi)^{-\frac{n}{2}}(2\pi\beta^2)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(x_i-\widehat{\mu}_{\mathrm{MAP}})^2 - \frac{\widehat{\mu}_{\mathrm{MAP}}^2}{2\beta^2}\right) \\ &\quad \times \int \exp\left(-\frac{(\mu-\widehat{\mu}_{\mathrm{MAP}})^2}{2(n+\beta^{-2})^{-1}}\right)\mathrm{d}\mu \\ &= (2\pi)^{-\frac{n}{2}}(n\beta^2+1)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(x_i-\widehat{\mu}_{\mathrm{MAP}})^2 - \frac{1}{2\beta^2}\widehat{\mu}_{\mathrm{MAP}}^2\right),\end{aligned}$$

where *Gaussian integral* shown in Fig. 4.1,

$$\int \exp\left(-\frac{(\mu-\widehat{\mu}_{\mathrm{MAP}})^2}{2(n+\beta^{-2})^{-1}}\right)\mathrm{d}\mu = \left(\frac{2\pi}{n+\beta^{-2}}\right)^{\frac{1}{2}},$$

is used. Consequently, the log marginal likelihood is expressed as

$$\begin{aligned}\log \mathrm{ML}(\beta) = &-\frac{n}{2}\log 2\pi - \frac{1}{2}\log(n\beta^2+1) \\ &- \frac{1}{2}\sum_{i=1}^{n}(x_i-\widehat{\mu}_{\mathrm{MAP}})^2 - \frac{\widehat{\mu}_{\mathrm{MAP}}^2}{2\beta^2}.\end{aligned}$$

A MATLAB code of empirical Bayes for parametric model (17.11) and prior probability (17.12) is given in Fig. 17.6, and its behavior is illustrated in Fig. 17.7. In this example, $\beta_{EB} = 2.19$ is chosen as the empirical Bayes solution, and true $\mu = 0.5$ was estimated by the MAP method as $\widehat{\mu}_{\mathrm{MAP}} = 0.537$, which is slightly better than MLE $\widehat{\mu}_{\mathrm{MLE}} = 0.546$.

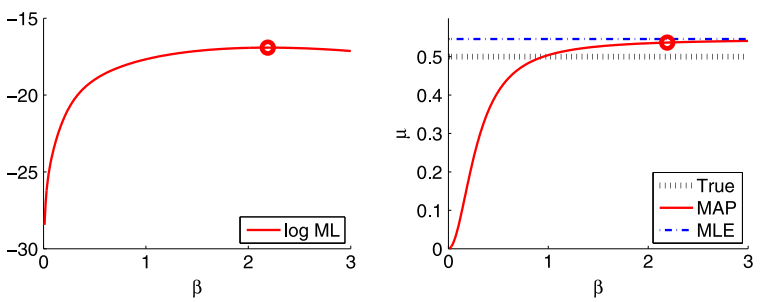
```
n=12; mu=0.5; x=randn(n,1)+mu;
bs=[0.01:0.01:3]; bl=length(bs);
MLE=mean(x);
for i=1:bl
  bb=bs(i)^(-2); MAP(i)=sum(x)/(n+bb);
  logML(i)=-n/2*log(2*pi)-sum((x-MAP(i)).^2)/2 ...
           -MAP(i)^2/(2*bb)-log(n*bb+1);
end
[dummy,c]=max(logML);

figure(1); clf; hold on;
plot(bs,logML,'r-');
plot(bs(c),logML(c),'ro');
xlabel('\beta'); legend('log ML',4);

figure(2); clf; hold on;
plot(bs,mu*ones(1,bl),'k:','LineWidth',5);
plot(bs,MAP,'r-','LineWidth',2);
plot(bs,MLE*ones(1,bl),'b-.','LineWidth',2);
plot(bs(c),MAP(c),'ro','LineWidth',4,'MarkerSize',10);
xlabel('\beta'); ylabel('\mu');
legend('True','MAP','MLE',4);
```

FIGURE 17.6

MATLAB code for empirical Bayes.



(a) Log marginal likelihood (b) MLE vs. MAP

FIGURE 17.7

Example of empirical Bayes.