CHAPTER

18 ANALYTIC APPROXIMATION OF MARGINAL LIKELIHOOD

CHAPTER CONTENTS

As discussed in Section 17.2, the use of conjugate priors allows us to avoid the explicit computation of the integration in the marginal likelihood:

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

In this chapter, general choices of prior probabilities are considered and analytic approximation methods of the marginal likelihood are introduced.

18.1 LAPLACE APPROXIMATION

In this section, *Laplace approximation* is introduced, which allows us to analytically approximate the integration of any twice-differentiable non-negative function $f(\boldsymbol{x})$:

$$\int f(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}.$$

18.1.1 APPROXIMATION WITH GAUSSIAN DENSITY

Let $\widehat{\boldsymbol{\theta}}$ be the maximizer of $f(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$:

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\mathrm{argmax}}\, f(\boldsymbol{\theta}).$$

An Introduction to Statistical Machine Learning. DOI: 10.1016/B978-0-12-802121-7.00029-7

Let us apply the *Taylor series expansion* (Fig. 2.8) to $\log f(\boldsymbol{\theta})$ about the maximizer $\widehat{\boldsymbol{\theta}}$:

$$\log f(\boldsymbol{\theta}) = \log f(\widehat{\boldsymbol{\theta}}) + (\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}})^\top \frac{\partial}{\partial \boldsymbol{\theta}} \log f(\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} + \frac{1}{2}(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}})^\top \boldsymbol{H}(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}) + \cdots, \tag{18.1}$$

where $\boldsymbol{H}$ is the *Hessian matrix*:

$$\boldsymbol{H} = \frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} \log f(\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}}.$$

Since $\widehat{\boldsymbol{\theta}}$ is the maximizer of $\log f(\boldsymbol{\theta})$,

$$\frac{\partial}{\partial \boldsymbol{\theta}} \log f(\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} = \mathbf{0}_b$$

holds and thus the first-order term in Eq. (18.1) is zero.

Let $\log \widehat{f}(\boldsymbol{\theta})$ be Eq. (18.1) up to the second-order terms (Fig. 18.1):

$$\log \widehat{f}(\boldsymbol{\theta}) = \log f(\widehat{\boldsymbol{\theta}}) + \frac{1}{2}(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}})^\top \boldsymbol{H}(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}).$$

Exponentiating both sides yields

$$\widehat{f}(\boldsymbol{\theta}) = f(\widehat{\boldsymbol{\theta}}) \exp\left(\frac{1}{2}(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}})^\top \boldsymbol{H}(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}})\right).$$

Recalling that the integration of the normal density is 1,

$$\frac{1}{(2\pi)^{\frac{b}{2}} \det(-\boldsymbol{H})^{\frac{1}{2}}} \int \exp\left(-\frac{1}{2}(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}})^\top (-\boldsymbol{H})(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}})\right) \mathrm{d}\boldsymbol{\theta} = 1,$$

integration of $\widehat{f}(\boldsymbol{\theta})$ yields

$$\int \widehat{f}(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} = f(\widehat{\boldsymbol{\theta}}) \sqrt{\frac{(2\pi)^b}{\det(-\boldsymbol{H})}} \approx \int f(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta},$$

where b denotes the dimensionality of $\boldsymbol{\theta}$. This is Laplace approximation to $\int f(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}$.

Since approximating $\log f(\boldsymbol{\theta})$ by a quadratic function corresponds to approximating $f(\boldsymbol{\theta})$ by an unnormalized Gaussian function, Laplace approximation is quite accurate if $f(\boldsymbol{\theta})$ is close to Gaussian. For this reason, Laplace approximation is also referred to as *Gaussian approximation*.

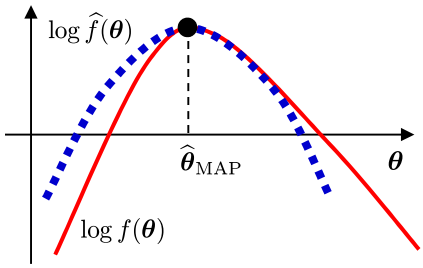

FIGURE 18.1

Laplace approximation.

18.1.2 ILLUSTRATION

Let us illustrate how to compute Laplace approximation to $\int f(\theta)\mathrm{d}\theta$, where

$$f(\theta) = N(\theta; 0, 1^2) + N(\theta; 0, 2^2),$$
$$N(\theta; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\theta-\mu)^2}{2\sigma^2}\right).$$

Note that $N(\theta; \mu, \sigma^2)$ is a normal density and thus the true value is 2.

The maximizer of $f(\theta)$ is given by

$$\widehat{\theta} = \underset{\theta}{\operatorname{argmax}}\, f(\theta) = 0.$$

Then

$$f(\widehat{\theta}) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\widehat{\theta}^2}{2}\right) + \frac{1}{\sqrt{8\pi}} \exp\left(-\frac{\widehat{\theta}^2}{8}\right) = \frac{3}{2\sqrt{2\pi}},$$
$$f'(\widehat{\theta}) = \frac{\partial}{\partial\theta} \log f(\theta)\Big|_{\theta=\widehat{\theta}} = -\widehat{\theta} N(\widehat{\theta}; 0, 1^2) - \frac{\widehat{\theta}}{4} N(\widehat{\theta}; 0, 2^2) = 0,$$
$$f''(\widehat{\theta}) = \frac{\partial^2}{\partial\theta^2} \log f(\theta)\Big|_{\theta=\widehat{\theta}} = (\widehat{\theta}^2 - 1) N(\widehat{\theta}; 0, 1^2) + \left(\frac{\widehat{\theta}^2}{16} - \frac{1}{4}\right) N(\widehat{\theta}; 0, 2^2)$$
$$= -\frac{9}{8\sqrt{2\pi}},$$

which yield

$$H = \frac{\partial^2}{\partial\theta^2} \log f(\theta)\Big|_{\theta=\widehat{\theta}} = \frac{f''(0) f(0) - f'(0)^2}{f(0)^2} = -\frac{3}{4}.$$

Thus, the Laplace approximation is given by

$$f(0)\sqrt{\frac{2\pi}{-H}} = \sqrt{3} \approx 1.732.$$

18.1.3 APPLICATION TO MARGINAL LIKELIHOOD APPROXIMATION

For the marginal likelihood,

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

Laplace approximation with respect to $\boldsymbol{\theta}$ yields

$$\mathrm{ML}(\boldsymbol{\beta}) \approx \prod_{i=1}^{n} q(\boldsymbol{x}_i|\widehat{\boldsymbol{\theta}}_{\mathrm{MAP}}) p(\widehat{\boldsymbol{\theta}}_{\mathrm{MAP}};\boldsymbol{\beta}) \sqrt{\frac{(2\pi)^b}{\det(-\boldsymbol{H})}},$$

where b denotes the dimensionality of $\boldsymbol{\theta}$ (i.e., the number of parameters), and

$$\begin{aligned} \widehat{\boldsymbol{\theta}}_{\mathrm{MAP}} &= \operatorname*{argmax}_{\boldsymbol{\theta}} \left[\sum_{i=1}^{n} \log q(\boldsymbol{x}_i|\boldsymbol{\theta}) + \log p(\boldsymbol{\theta};\boldsymbol{\beta}) \right], \\ \boldsymbol{H} &= \frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top} \left(\sum_{i=1}^{n} \log q(\boldsymbol{x}_i|\boldsymbol{\theta}) + \log p(\boldsymbol{\theta};\boldsymbol{\beta}) \right) \Bigg|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}_{\mathrm{MAP}}}. \end{aligned} \tag{18.2}$$

If the number of training samples, n, is large, the *central limit theorem* (see Section 7.4) asserts that the posterior probability $p(\boldsymbol{\theta}|\mathcal{D})$ converges in distribution to the Gaussian distribution. Therefore, the Laplace-approximated marginal likelihood would be accurate when a large number of training samples are available.

18.1.4 BAYESIAN INFORMATION CRITERION (BIC)

The logarithm of the Laplace-approximated marginal likelihood is given by

$$\begin{aligned} \log \mathrm{ML}(\boldsymbol{\beta}) \approx & \sum_{i=1}^{n} \log q(\boldsymbol{x}_i|\widehat{\boldsymbol{\theta}}_{\mathrm{MAP}}) + \log p(\widehat{\boldsymbol{\theta}}_{\mathrm{MAP}};\boldsymbol{\beta}) \\ & + \frac{b}{2}\log(2\pi) - \frac{1}{2}\log\left(\det(-\boldsymbol{H})\right). \end{aligned} \tag{18.3}$$

Let us further approximate this under the assumption that the number of training samples, n, is large.

The first term $\sum_{i=1}^{n} \log q(\boldsymbol{x}_i|\widehat{\boldsymbol{\theta}}_{\mathrm{MAP}})$ in Eq. (18.3) has asymptotic order n (see Fig. 14.4):

$$\sum_{i=1}^{n} \log q(\boldsymbol{x}_i|\widehat{\boldsymbol{\theta}}_{\mathrm{MAP}}) = O(n).$$

On the other hand, the second term $\log p(\widehat{\boldsymbol{\theta}}_{\mathrm{MAP}}; \boldsymbol{\beta})$ and the third term $\frac{b}{2}\log(2\pi)$ are independent of n:

$$\log p(\widehat{\boldsymbol{\theta}}_{\mathrm{MAP}}; \boldsymbol{\beta}) = O(1) \quad \text{and} \quad \frac{b}{2}\log(2\pi) = O(1).$$

According to the law of large numbers (see Section 7.3), the Hessian matrix $\boldsymbol{H}$ defined by Eq. (18.2) divided by n converges in probability to $\widetilde{\boldsymbol{H}}$:

$$\frac{1}{n}\boldsymbol{H} = \frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}\left(\frac{1}{n}\sum_{i=1}^{n}\log q(\boldsymbol{x}_i|\boldsymbol{\theta}) + \frac{1}{n}\log p(\boldsymbol{\theta};\boldsymbol{\beta})\right)\Bigg|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}_{\mathrm{MAP}}} \xrightarrow{\mathrm{p}} \widetilde{\boldsymbol{H}},$$

where

$$\widetilde{\boldsymbol{H}} = \frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}\Big(E\left[\log q(\boldsymbol{x}|\boldsymbol{\theta})\right]\Big)\Big|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}_{\mathrm{MAP}}}.$$

Since $\widetilde{\boldsymbol{H}}$ is a $b \times b$ matrix,

$$\det(n\widetilde{\boldsymbol{H}}) = n^b \det(\widetilde{\boldsymbol{H}})$$

holds and therefore

$$\frac{1}{2}\log\left(\det(-\boldsymbol{H})\right) \xrightarrow{\mathrm{p}} \frac{b}{2}\log n + \frac{1}{2}\log\Big(\det(-\widetilde{\boldsymbol{H}})\Big)$$

is obtained. The first term $\frac{b}{2}\log n$ is proportional to $\log n$, while the second term $\frac{1}{2}\log\Big(\det(-\widetilde{\boldsymbol{H}})\Big)$ is independent of n:

$$\frac{b}{2}\log n = O(\log n) \quad \text{and} \quad \frac{1}{2}\log\Big(\det(-\widetilde{\boldsymbol{H}})\Big) = O(1).$$

Here, suppose that n is large enough so that terms with $O(1)$ can be ignored. Then only $\sum_{i=1}^{n}\log q(\boldsymbol{x}_i|\widehat{\boldsymbol{\theta}}_{\mathrm{MAP}})$ and $\frac{b}{2}\log n$ remain. Furthermore, the difference between the MAP solution $\widehat{\boldsymbol{\theta}}_{\mathrm{MAP}}$ and the maximum likelihood solution $\widehat{\boldsymbol{\theta}}_{\mathrm{MLE}}$ is known to be $O(n^{-1})$:

$$\sum_{i=1}^{n}\log q(\boldsymbol{x}_i|\widehat{\boldsymbol{\theta}}_{\mathrm{MAP}}) = \sum_{i=1}^{n}\log q(\boldsymbol{x}_i|\widehat{\boldsymbol{\theta}}_{\mathrm{MLE}}) + O(n^{-1}).$$

Consequently, Laplace approximation of the log marginal likelihood given by Eq. (18.3) can be further approximated as

$$\log \mathrm{ML}(\boldsymbol{\beta}) \approx \sum_{i=1}^{n}\log q(\boldsymbol{x}_i|\widehat{\boldsymbol{\theta}}_{\mathrm{MLE}}) - \frac{b}{2}\log n.$$

The negative of the right-hand side is called the BIC:

$$\mathrm{BIC} = -\sum_{i=1}^{n} \log q(\boldsymbol{x}_i|\widehat{\boldsymbol{\theta}}_{\mathrm{MLE}}) + \frac{b}{2}\log n.$$

BIC is quite simple and thus is popularly used in model selection. However, BIC is no longer dependent on prior probabilities and thus cannot be used for setting the prior probability. It is also known that BIC is equivalent to the *minimum description length* (MDL) criterion [84], which was derived in a completely different framework.

BIC is similar to AIC explained in Section 14.3, but the second term is different:

$$\mathrm{AIC} = -\sum_{i=1}^{n} \log q(\boldsymbol{x}_i;\widehat{\boldsymbol{\theta}}_{\mathrm{MLE}}) + b.$$

When $n > e^2 \approx 7.39$, BIC has a stronger penalty than AIC and thus a simpler model would be chosen. However, since AIC and BIC are derived in completely different frameworks (KL divergence approximation and the marginal likelihood approximation), it cannot be simply concluded which one is more superior than the other.

18.2 VARIATIONAL APPROXIMATION

When the integrand $\prod_{i=1}^{n} q(\boldsymbol{x}_i|\boldsymbol{\theta})p(\boldsymbol{\theta};\boldsymbol{\beta})$ is not close to Gaussian, Laplace-approximated marginal likelihood may not be accurate. In this section, *variational approximation* is introduced, which finds the best approximation to the marginal likelihood in a limited function class that is easier to compute [71].

18.2.1 VARIATIONAL BAYESIAN EM (VBEM) ALGORITHM

The marginal likelihood can be expressed as

$$\mathrm{ML}(\boldsymbol{\beta}) = p(\mathcal{D};\boldsymbol{\beta}) = \iint p(\mathcal{D},\boldsymbol{\eta},\boldsymbol{\theta};\boldsymbol{\beta})\mathrm{d}\boldsymbol{\eta}\mathrm{d}\boldsymbol{\theta},$$

where $\boldsymbol{\eta}$ is called a *latent variable*. Let us consider probability density functions for $\boldsymbol{\eta}$ and $\boldsymbol{\theta}$, denoted by $q(\boldsymbol{\eta})$ and $r(\boldsymbol{\theta})$, called the *trial distributions*. Then Jensen's inequality (see Section 8.3.1) gives the following lower bound:

$$\begin{aligned}\log \mathrm{ML}(\boldsymbol{\beta}) &= \log \iint p(\mathcal{D},\boldsymbol{\eta},\boldsymbol{\theta};\boldsymbol{\beta})\mathrm{d}\boldsymbol{\eta}\mathrm{d}\boldsymbol{\theta} \\ &= \log \iint q(\boldsymbol{\eta})r(\boldsymbol{\theta})\frac{p(\mathcal{D},\boldsymbol{\eta},\boldsymbol{\theta};\boldsymbol{\beta})}{q(\boldsymbol{\eta})r(\boldsymbol{\theta})}\mathrm{d}\boldsymbol{\eta}\mathrm{d}\boldsymbol{\theta} \geq -F(q,r),\end{aligned}$$

where

$$F(q,r) = \iint q(\boldsymbol{\eta})r(\boldsymbol{\theta})\log\frac{q(\boldsymbol{\eta})r(\boldsymbol{\theta})}{p(\mathcal{D},\boldsymbol{\eta},\boldsymbol{\theta};\boldsymbol{\beta})}\mathrm{d}\boldsymbol{\eta}\mathrm{d}\boldsymbol{\theta}$$

is a functional of probability density functions q and r called the *variational free energy* (while $-\log \mathrm{ML}(\boldsymbol{\beta})$ is called the *free energy*). If q and r are chosen to minimize the variational free energy, a good approximator to the log marginal likelihood may be obtained.

Setting the partial derivatives of the variational free energy at zero,

$$\frac{\partial}{\partial q} F(q,r) = 0 \quad \text{and} \quad \frac{\partial}{\partial r} F(q,r) = 0,$$

yields that the solutions should satisfy

$$q(\boldsymbol{\eta}) \propto \exp\left(\int r(\boldsymbol{\theta}) \log p(\mathcal{D}, \boldsymbol{\eta} | \boldsymbol{\theta}; \boldsymbol{\beta}) \mathrm{d}\boldsymbol{\theta} \right), \tag{18.4}$$

$$r(\boldsymbol{\theta}) \propto p(\boldsymbol{\theta}) \exp\left(\int q(\boldsymbol{\eta}) \log p(\mathcal{D}, \boldsymbol{\eta} | \boldsymbol{\theta}; \boldsymbol{\beta}) \mathrm{d}\boldsymbol{\eta} \right). \tag{18.5}$$

Since no method is known to analytically solve Eq. (18.4) and Eq. (18.5), these equations are used as updating formulas like the EM algorithm (see Section 15.4). Such an EM-like algorithm is called the VBEM algorithm, and Eq. (18.4) and Eq. (18.5) are, respectively, called the VB-E step and VB-M step.

The variational free energy can be expressed by using the KL divergence (see Section 14.2) as

$$F(q,r) = \mathrm{KL}\big(q(\boldsymbol{\eta}) r(\boldsymbol{\theta}) \big\| p(\boldsymbol{\eta}, \boldsymbol{\theta} | \mathcal{D}; \boldsymbol{\beta})\big) - \log \mathrm{ML}(\boldsymbol{\beta}).$$

This shows that reducing the variational free energy $F(q,r)$ corresponds to reducing $\mathrm{KL}\big(q(\boldsymbol{\eta}) r(\boldsymbol{\theta}) \big\| p(\boldsymbol{\eta}, \boldsymbol{\theta} | \mathcal{D}; \boldsymbol{\beta})\big)$, and therefore the VBEM algorithm can be regarded as approximating $p(\boldsymbol{\eta}, \boldsymbol{\theta} | \mathcal{D}; \boldsymbol{\beta})$ by $q(\boldsymbol{\eta}) r(\boldsymbol{\theta})$.

For this reason, $r(\boldsymbol{\theta})$ obtained by the VBEM algorithm may be a good approximation to the posterior probability $p(\boldsymbol{\theta} | \mathcal{D}; \boldsymbol{\beta})$.

18.2.2 RELATION TO ORDINARY EM ALGORITHM

As explained in Section 15.4, the ordinary EM algorithm maximized a lower bound of the likelihood. Here, it is shown that the ordinary EM algorithm can also be interpreted as a variational approximation. *Jensen's inequality* (Section 8.3.1) yields the following lower bound of the log-likelihood:

$$\begin{aligned} \log p(\mathcal{D} | \boldsymbol{\theta}) &= \log \int p(\mathcal{D}, \boldsymbol{\eta} | \boldsymbol{\theta}) \mathrm{d}\boldsymbol{\eta} \\ &= \log \int q(\boldsymbol{\eta} | \mathcal{D}, \boldsymbol{\theta}') \frac{p(\mathcal{D}, \boldsymbol{\eta} | \boldsymbol{\theta})}{q(\boldsymbol{\eta} | \mathcal{D}, \boldsymbol{\theta}')} \mathrm{d}\boldsymbol{\eta} \geq b(q, \boldsymbol{\theta}), \end{aligned}$$

where

$$b(q, \boldsymbol{\theta}) = \int q(\boldsymbol{\eta} | \mathcal{D}, \boldsymbol{\theta}') \log \frac{p(\mathcal{D}, \boldsymbol{\eta} | \boldsymbol{\theta})}{q(\boldsymbol{\eta} | \mathcal{D}, \boldsymbol{\theta}')} \mathrm{d}\boldsymbol{\eta}.$$

$\frac{\partial}{\partial q} b(q, \boldsymbol{\theta}) = 0$ yields $q(\boldsymbol{\eta}|\mathcal{D}, \boldsymbol{\theta}') = p(\boldsymbol{\eta}|\mathcal{D}, \boldsymbol{\theta})$, which is the E-step of the ordinary EM algorithm. Then finding $\boldsymbol{\theta}'$ that satisfies $\frac{\partial}{\partial \boldsymbol{\theta}} b(q, \boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}'} = \mathbf{0}$ is the M-step of the ordinary EM algorithm.

The relation between the ordinary EM algorithm and the VBEM algorithm can be further elucidated by the use of *Dirac's delta function* $\delta(\cdot)$, which satisfies for any function $g : \mathbb{R} \to \mathbb{R}$ and any real number κ,

$$\int_{-\infty}^{\infty} g(\tau)\delta(\kappa - \tau)\mathrm{d}\tau = g(\kappa).$$

Thus, *convolution* with Dirac's delta function allows us to extract the value of an arbitrary function $g(\cdot)$ at an arbitrary point κ. Dirac's delta function $\delta(\cdot)$ can be expressed as the limit of the normal density:

$$\delta(\tau) = \lim_{\sigma \to 0} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\tau^2}{2\sigma^2}\right).$$

For multidimensional $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_d)^\top$, Dirac's delta function is defined in an elementwise manner as

$$\delta(\boldsymbol{\tau}) = \delta(\tau_1) \times \cdots \times \delta(\tau_d).$$

Dirac's delta function, setting

$$r(\boldsymbol{\theta}') = \delta(\boldsymbol{\theta}' - \boldsymbol{\theta})$$

in the VB-E step yields

$$q(\boldsymbol{\eta}) \propto p(\mathcal{D}, \boldsymbol{\eta}|\boldsymbol{\theta}) \propto p(\boldsymbol{\eta}|\mathcal{D}, \boldsymbol{\theta}),$$

which agrees with the ordinary E-step.