Cambridge University Press
978-1-107-07615-0 — Variational Bayesian Learning Theory
Shinichi Nakajima , Kazuho Watanabe , Masashi Sugiyama
Excerpt
More Information

Part I

Formulation

1
Bayesian Learning

Bayesian learning is an inference method based on the fundamental law of probability, called the Bayes theorem. In this first chapter, we introduce the framework of Bayesian learning with simple examples where Bayesian learning can be performed analytically.

1.1 Framework

Bayesian learning considers the following situation. We have observed a set $\mathcal{D}$ of data, which are subject to a *conditional distribution* $p(\mathcal{D}|\boldsymbol{w})$, called the *model distribution*, of the data given unknown *model parameter* $\boldsymbol{w}$. Although the value of $\boldsymbol{w}$ is unknown, vague information on $\boldsymbol{w}$ is provided as a *prior distribution* $p(\boldsymbol{w})$. The conditional distribution $p(\mathcal{D}|\boldsymbol{w})$ is also called the *model likelihood* when it is seen as a function of the unknown parameter $\boldsymbol{w}$.

1.1.1 Bayes Theorem and Bayes Posterior

Bayesian learning is based on the following basic factorization property of the *joint distribution* $p(\mathcal{D}, \boldsymbol{w})$:

$$\underbrace{p(\boldsymbol{w}|\mathcal{D})}_{\text{posterior}} \underbrace{p(\mathcal{D})}_{\text{marginal}} = \underbrace{p(\mathcal{D}, \boldsymbol{w})}_{\text{joint}} = \underbrace{p(\mathcal{D}|\boldsymbol{w})}_{\text{likelihood}} \underbrace{p(\boldsymbol{w})}_{\text{prior}}, \tag{1.1}$$

where the marginal distribution is given by

$$p(\mathcal{D}) = \int_{\mathcal{W}} p(\mathcal{D}, \boldsymbol{w}) d\boldsymbol{w} = \int_{\mathcal{W}} p(\mathcal{D}|\boldsymbol{w}) p(\boldsymbol{w}) d\boldsymbol{w}. \tag{1.2}$$

Here, the integration is performed in the domain $\mathcal{W}$ of the parameter $\boldsymbol{w}$. Note that, if the domain $\mathcal{W}$ is discrete, integration should be replaced with summation, i.e., for any function $f(\boldsymbol{w})$,

$$\int_{\mathcal{W}} f(\boldsymbol{w})d\boldsymbol{w} \to \sum_{\boldsymbol{w}' \in \mathcal{W}} f(\boldsymbol{w}').$$

The *posterior distribution*, the distribution of the unknown parameter $\boldsymbol{w}$ given the observed data set $\mathcal{D}$, is derived by dividing both sides of Eq. (1.1) by the marginal distribution $p(\mathcal{D})$:

$$p(\boldsymbol{w}|\mathcal{D}) = \frac{p(\mathcal{D}, \boldsymbol{w})}{p(\mathcal{D})} \propto p(\mathcal{D}, \boldsymbol{w}). \tag{1.3}$$

Here, we emphasized that the posterior distribution is proportional to the joint distribution $p(\mathcal{D}, \boldsymbol{w})$ because the marginal distribution $p(\mathcal{D})$ is a constant (as a function of $\boldsymbol{w}$). In other words, the joint distribution is an *unnormalized posterior distribution*. Eq. (1.3) is called the *Bayes theorem*, and the posterior distribution computed exactly by Eq. (1.3) is called the *Bayes posterior* when we distinguish it from its approximations.

Example 1.1 (Parametric density estimation) Assume that the observed data $\mathcal{D} = \{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(N)}\}$ consist of N *independent and identically distributed (i.i.d.)* samples from the model distribution $p(\boldsymbol{x}|\boldsymbol{w})$. Then, the model likelihood is given by $p(\mathcal{D}|\boldsymbol{w}) = \prod_{n=1}^{N} p(\boldsymbol{x}^{(n)}|\boldsymbol{w})$, and therefore, the posterior distribution is given by

$$p(\boldsymbol{w}|\mathcal{D}) = \frac{\prod_{n=1}^{N} p(\boldsymbol{x}^{(n)}|\boldsymbol{w})p(\boldsymbol{w})}{\int \prod_{n=1}^{N} p(\boldsymbol{x}^{(n)}|\boldsymbol{w})p(\boldsymbol{w})d\boldsymbol{w}} \propto \prod_{n=1}^{N} p(\boldsymbol{x}^{(n)}|\boldsymbol{w})p(\boldsymbol{w}).$$

Example 1.2 (Parametric regression) Assume that the observed data $\mathcal{D} = \{(\boldsymbol{x}^{(1)}, \boldsymbol{y}^{(1)}), \ldots, (\boldsymbol{x}^{(N)}, \boldsymbol{y}^{(N)})\}$ consist of N i.i.d. input–output pairs from the model distribution $p(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{w}) = p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{w})p(\boldsymbol{x})$. Then, the likelihood function is given by $p(\mathcal{D}|\boldsymbol{w}) = \prod_{n=1}^{N} p(\boldsymbol{y}^{(n)}|\boldsymbol{x}^{(n)}, \boldsymbol{w})p(\boldsymbol{x}^{(n)})$, and therefore, the posterior distribution is given by

$$p(\boldsymbol{w}|\mathcal{D}) = \frac{\prod_{n=1}^{N} p(\boldsymbol{y}^{(n)}|\boldsymbol{x}^{(n)}, \boldsymbol{w})p(\boldsymbol{w})}{\int \prod_{n=1}^{N} p(\boldsymbol{y}^{(n)}|\boldsymbol{x}^{(n)}, \boldsymbol{w})p(\boldsymbol{w})d\boldsymbol{w}} \propto \prod_{n=1}^{N} p(\boldsymbol{y}^{(n)}|\boldsymbol{x}^{(n)}, \boldsymbol{w})p(\boldsymbol{w}).$$

Note that the input distribution $p(\boldsymbol{x})$ does not affect the posterior, and accordingly is often ignored in practice.

1.1.2 Maximum A Posteriori Learning

Since the joint distribution $p(\mathcal{D}, \boldsymbol{w})$ is just the product of the likelihood function and the prior distribution (see Eq. (1.1)), it is usually easy to

CAMBRIDGE

Cambridge University Press
978-1-107-07615-0 — Variational Bayesian Learning Theory
Shinichi Nakajima , Kazuho Watanabe , Masashi Sugiyama
Excerpt
More Information

compute. Therefore, it is relatively easy to perform *maximum a posteriori (MAP) learning*, where the parameters are point-estimated so that the posterior probability is maximized, i.e.,

$$\widehat{\boldsymbol{w}}^{\mathrm{MAP}} = \underset{\boldsymbol{w}}{\mathrm{argmax}}\, p(\boldsymbol{w}|\mathcal{D}) = \underset{\boldsymbol{w}}{\mathrm{argmax}}\, p(\mathcal{D}, \boldsymbol{w}). \tag{1.4}$$

MAP learning includes *maximum likelihood (ML) learning*,

$$\widehat{\boldsymbol{w}}^{\mathrm{ML}} = \underset{\boldsymbol{w}}{\mathrm{argmax}}\, p(\mathcal{D}|\boldsymbol{w}), \tag{1.5}$$

as a special case with the flat prior $p(\boldsymbol{w}) \propto 1$.

1.1.3 Bayesian Learning

On the other hand, *Bayesian learning* requires integration of the joint distribution with respect to the parameter $\boldsymbol{w}$, which is often computationally hard. More specifically, performing Bayesian learning means computing at least one of the following quantities:

***Marginal likelihood* (zeroth moment)**

$$p(\mathcal{D}) = \int p(\mathcal{D}, \boldsymbol{w}) d\boldsymbol{w}. \tag{1.6}$$

This quantity has been already introduced in Eq. (1.2) as the normalization factor of the posterior distribution. As seen in Section 1.1.5 and subsequent sections, the marginal likelihood plays an important role in model selection and hyperparameter estimation.

***Posterior mean* (first moment)**

$$\widehat{\boldsymbol{w}} = \langle \boldsymbol{w} \rangle_{p(\boldsymbol{w}|\mathcal{D})} = \frac{1}{p(\mathcal{D})} \int \boldsymbol{w} \cdot p(\mathcal{D}, \boldsymbol{w}) d\boldsymbol{w}, \tag{1.7}$$

where $\langle \cdot \rangle_p$ denotes the expectation value over the distribution p, i.e., $\langle \cdot \rangle_{p(\boldsymbol{w})} = \int \cdot p(\boldsymbol{w}) d\boldsymbol{w}$. This quantity is also called the *Bayesian estimator*. The Bayesian estimator or the model distribution with the Bayesian estimator plugged in (see the plug-in predictive distribution (1.10)) can be the final output of Bayesian learning.

***Posterior covariance* (second moment)**

$$\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{w}} = \left\langle (\boldsymbol{w} - \widehat{\boldsymbol{w}})(\boldsymbol{w} - \widehat{\boldsymbol{w}})^{\top} \right\rangle_{p(\boldsymbol{w}|\mathcal{D})} = \frac{1}{p(\mathcal{D})} \int (\boldsymbol{w} - \widehat{\boldsymbol{w}})(\boldsymbol{w} - \widehat{\boldsymbol{w}})^{\top} p(\mathcal{D}, \boldsymbol{w}) d\boldsymbol{w}, \tag{1.8}$$

where $\top$ denotes the transpose of a matrix or vector. This quantity provides the credibility information, and is used to assess the confidence level of the Bayesian estimator.

Predictive distribution **(expectation of model distribution)**

$$p(\mathcal{D}^{\mathrm{new}}|\mathcal{D}) = \left\langle p(\mathcal{D}^{\mathrm{new}}|\boldsymbol{w}) \right\rangle_{p(\boldsymbol{w}|\mathcal{D})} = \frac{1}{p(\mathcal{D})} \int p(\mathcal{D}^{\mathrm{new}}|\boldsymbol{w}) p(\mathcal{D}, \boldsymbol{w}) d\boldsymbol{w}, \qquad (1.9)$$

where $p(\mathcal{D}^{\mathrm{new}}|\boldsymbol{w})$ denotes the model distribution on *unobserved* new data $\mathcal{D}^{\mathrm{new}}$. In the i.i.d. case such as Examples 1.1 and 1.2, it is sufficient to compute the predictive distribution for a single new sample $\mathcal{D}^{\mathrm{new}} = \{\boldsymbol{x}\}$.

Note that each of the four quantities (1.6) through (1.9) requires to compute the expectation of some function $f(\boldsymbol{w})$ over the unnormalized posterior distribution $p(\mathcal{D}, \boldsymbol{w})$ on $\boldsymbol{w}$, i.e., $\int f(\boldsymbol{w}) p(\mathcal{D}, \boldsymbol{w}) d\boldsymbol{w}$. Specifically, the marginal likelihood, the posterior mean, and the posterior covariance are the zeroth, the first, and the second moments of the unnormalized posterior distribution, respectively. The expectation is analytically intractable except for some simple cases, and numerical computation is also hard when the dimensionality of the unknown parameter $\boldsymbol{w}$ is high. This is the main bottleneck of Bayesian learning, with which many approximation methods have been developed to cope.

It hardly happens that the first moment (1.7) or the second moment (1.8) are computationally tractable but the zeroth moment (1.6) is not. Accordingly, we can say that performing Bayesian learning on the parameter $\boldsymbol{w}$ amounts to obtaining the *normalized* posterior distribution $p(\boldsymbol{w}|\mathcal{D})$. It sometimes happens that computing the predictive distribution (1.9) is still intractable even if the zeroth, the first, and the second moments can be computed based on some approximation. In such a case, the model distribution with the Bayesian estimator plugged in, called the *plug-in predictive distribution*,

$$p(\mathcal{D}^{\mathrm{new}}|\widehat{\boldsymbol{w}}), \qquad (1.10)$$

is used for prediction in practice.

1.1.4 Latent Variables

So far, we introduced the observed data set $\mathcal{D}$ as a known variable, and the model parameter $\boldsymbol{w}$ as an unknown variable. In practice, more varieties of known and unknown variables can be involved.

Some probabilistic models have *latent variables* (or *hidden variables*) $\boldsymbol{z}$, which can be involved in the original model, or additionally introduced for computational reasons. They are typically attributed to each of the observed samples, and therefore have large degrees of freedom. However, they are just additional unknown variables, and there is no reason in inference to distinguish them from the model parameters $\boldsymbol{w}$.[1] The joint posterior over the parameters and the latent variables is given by Eq. (1.3) with $\boldsymbol{w}$ and $p(\boldsymbol{w})$ replaced with $\overline{\boldsymbol{w}} = (\boldsymbol{w}, \boldsymbol{z})$ and $p(\overline{\boldsymbol{w}}) = p(\boldsymbol{z}|\boldsymbol{w})p(\boldsymbol{w})$, respectively.

Example 1.3 (Mixture models) A mixture model is often used for parametric density estimation (Example 1.1). The model distribution is given by

$$p(\boldsymbol{x}|\boldsymbol{w}) = \sum_{k=1}^{K} \alpha_k p(\boldsymbol{x}|\boldsymbol{\tau}_k), \tag{1.11}$$

where $\boldsymbol{w} = \{\alpha_k, \boldsymbol{\tau}_k; \alpha_k \geq 0, \sum_{k=1}^{K} \alpha_k = 1\}_{k=1}^{K}$ is the unknown parameters. The mixture model (1.11) is the weighted sum of K distributions, each of which is parameterized by the component parameter $\boldsymbol{\tau}_k$. The domain of the *mixing weights* $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)^\top$, also called as the *mixture coefficients*, forms the *standard* $(K-1)$*-simplex*, denoted by $\Delta^{K-1} \equiv \{\boldsymbol{\alpha} \in \mathbb{R}_+^K; \sum_{k=1}^{K} \alpha_k = 1\}$ (see Figure 1.1). Figure 1.2 shows an example of the mixture model with three one-dimensional Gaussian components.

The likelihood,

$$\begin{aligned} p(\mathcal{D}|\boldsymbol{w}) &= \prod_{n=1}^{N} p(\boldsymbol{x}^{(n)}|\boldsymbol{w}), \\ &= \prod_{n=1}^{N} \left(\sum_{k=1}^{K} \alpha_k p(\boldsymbol{x}|\boldsymbol{\tau}_k) \right), \end{aligned} \tag{1.12}$$

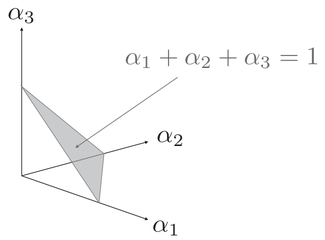

Figure 1.1 $(K-1)$-simplex, Δ^{K-1}, for $K = 3$.

[1] For this reason, the latent variables $\boldsymbol{z}$ and the model parameters $\boldsymbol{w}$ are also called *local latent variables* and *global latent variables*, respectively.

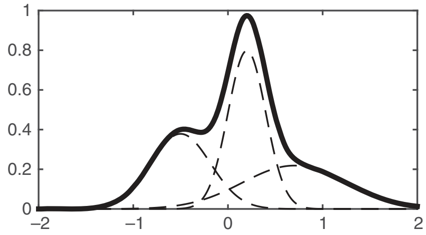

Figure 1.2 Gaussian mixture.

for N observed i.i.d. samples $\mathcal{D} = \{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(N)}\}$ has $O(K^N)$ terms, which makes even ML learning intractable. This intractability arises from the summation inside the multiplication in Eq. (1.12). By introducing latent variables, we can turn this summation into a multiplication, and make Eq. (1.12) tractable.

Assume that each sample $\boldsymbol{x}$ belongs to a single component k, and is drawn from $p(\boldsymbol{x}|\boldsymbol{\tau}_k)$. To describe the assignment, we introduce a latent variable $\boldsymbol{z} \in \mathcal{Z} \equiv \{\boldsymbol{e}_k\}_{k=1}^K$ associated with each observed sample $\boldsymbol{x}$, where $\boldsymbol{e}_k \in \{0, 1\}^K$ is the K-dimensional binary vector, called the *one-of-K representation*, with one at the kth entry and zeros at the other entries:

$$\boldsymbol{e}_k = (\underbrace{0, \ldots, 0, \overbrace{\;1\;}^{k\text{th}}, 0, \ldots, 0}_{K})^\top .$$

Then, we have the following model:

$$p(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{w}) = p(\boldsymbol{x}|\boldsymbol{z}, \boldsymbol{w})p(\boldsymbol{z}|\boldsymbol{w}), \tag{1.13}$$

$$\text{where} \qquad p(\boldsymbol{x}|\boldsymbol{z}, \boldsymbol{w}) = \prod_{k=1}^K \{p(\boldsymbol{x}|\boldsymbol{\tau}_k)\}^{z_k}, \qquad p(\boldsymbol{z}|\boldsymbol{w}) = \prod_{k=1}^K \alpha_k^{z_k}.$$

The conditional distribution (1.13) on the observed variable $\boldsymbol{x}$ and the latent variable $\boldsymbol{z}$ given the parameter $\boldsymbol{w}$ is called the *complete likelihood*.

Note that marginalizing the complete likelihood over the latent variable recovers the original mixture model:

$$p(\boldsymbol{x}|\boldsymbol{w}) = \int_{\mathcal{Z}} p(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{w}) d\boldsymbol{z} = \sum_{\boldsymbol{z} \in \{\boldsymbol{e}_k\}_{k=1}^K} \prod_{k=1}^K \{\alpha_k p(\boldsymbol{x}|\boldsymbol{\tau}_k)\}^{z_k} = \sum_{k=1}^K \alpha_k p(\boldsymbol{x}|\boldsymbol{\tau}_k).$$

This means that, if samples are generated from the model distribution (1.13), and only $\boldsymbol{x}$ is recorded, the observed data follow the original mixture model (1.11).

In the literature, latent variables tend to be marginalized out even in MAP learning. For example, the *expectation-maximization (EM) algorithm* (Dempster et al., 1977), a popular MAP solver for latent variable models, seeks a (local) maximizer of the posterior distribution with the latent variables marginalized out, i.e.,

$$\widehat{\boldsymbol{w}}^{\mathrm{EM}} = \mathop{\mathrm{argmax}}_{\boldsymbol{w}} p(\boldsymbol{w}|\mathcal{D}) = \mathop{\mathrm{argmax}}_{\boldsymbol{w}} \int_{\mathcal{Z}} p(\mathcal{D}, \boldsymbol{w}, \boldsymbol{z}) d\boldsymbol{z}. \qquad (1.14)$$

However, we can also maximize the posterior jointly over the parameters and the latent variables, i.e.,

$$(\widehat{\boldsymbol{w}}^{\mathrm{MAP-hard}}, \widehat{\boldsymbol{z}}^{\mathrm{MAP-hard}}) = \mathop{\mathrm{argmax}}_{\boldsymbol{w},\boldsymbol{z}} p(\boldsymbol{w}, \boldsymbol{z}|\mathcal{D}) = \mathop{\mathrm{argmax}}_{\boldsymbol{w},\boldsymbol{z}} p(\mathcal{D}, \boldsymbol{w}, \boldsymbol{z}). \qquad (1.15)$$

For clustering based on the mixture model in Example 1.3, the EM algorithm (1.14) gives a *soft assignment*, where the expectation value $\widehat{\boldsymbol{z}}^{\mathrm{EM}} \in \Delta^{K-1} \subset [0, 1]^K$ is substituted into the joint distribution $p(\mathcal{D}, \boldsymbol{w}, \boldsymbol{z})$, while the joint maximization (1.15) gives the *hard assignment*, where the optimal assignment $\widehat{\boldsymbol{z}}^{\mathrm{MAP-hard}} \in \{\boldsymbol{e}_k\}_{k=1}^K \subset \{0, 1\}^K$ is looked for in the binary domain.

1.1.5 Empirical Bayesian Learning

In many practical cases, it is reasonable to use a prior distribution parameterized by *hyperparameters* $\boldsymbol{\kappa}$. The hyperparameters can be tuned by hand or based on some criterion outside the Bayesian framework. A popular method of the latter is the *cross validation*, where the hyperparameters are tuned so that an (preferably unbiased) estimator of the performance criterion is optimized. In such cases, the hyperparameters should be treated as *known* variables when Bayesian learning is performed.

On the other hand, the hyperparameters can be estimated within the Bayesian framework. In this case, there is again no reason to distinguish the hyperparameters from the other unknown variables $(\boldsymbol{w}, \boldsymbol{z})$. The joint posterior over all unknown variables is given by Eq. (1.3) with $\boldsymbol{w}$ and $p(\boldsymbol{w})$ replaced with $\overline{\boldsymbol{w}} = (\boldsymbol{w}, \boldsymbol{\kappa}, \boldsymbol{z})$ and $p(\overline{\boldsymbol{w}}) = p(\boldsymbol{z}|\boldsymbol{w})p(\boldsymbol{w}|\boldsymbol{\kappa})p(\boldsymbol{\kappa})$, respectively, where $p(\boldsymbol{\kappa})$ is called a *hyperprior*. A popular approach, called *empirical Bayesian (EBayes) learning* (Efron and Morris, 1973), applies Bayesian learning on $\boldsymbol{w}$ (and $\boldsymbol{z}$) and point-estimate $\boldsymbol{\kappa}$, i.e.,

$$\widehat{\boldsymbol{\kappa}}^{\mathrm{EBayes}} = \mathop{\mathrm{argmax}}_{\boldsymbol{\kappa}} p(\mathcal{D}, \boldsymbol{\kappa}) = \mathop{\mathrm{argmax}}_{\boldsymbol{\kappa}} p(\mathcal{D}|\boldsymbol{\kappa}) p(\boldsymbol{\kappa}),$$

$$\text{where} \quad p(\mathcal{D}|\boldsymbol{\kappa}) = \int p(\mathcal{D}, \boldsymbol{w}, \boldsymbol{z}|\boldsymbol{\kappa}) d\boldsymbol{w} d\boldsymbol{z}.$$

Here the marginal likelihood $p(\mathcal{D}|\boldsymbol{\kappa})$ is seen as the likelihood of the hyperparameter $\boldsymbol{\kappa}$, and MAP learning is performed by maximizing the joint distribution $p(\mathcal{D}, \boldsymbol{\kappa})$ of the observed data $\mathcal{D}$ and the hyperparameter $\boldsymbol{\kappa}$, which can be seen as an *unnormalized posterior distribution* of the hyperparameter. The hyperprior is often assumed to be flat: $p(\boldsymbol{\kappa}) \propto 1$.

With an appropriate design of priors, empirical Bayesian learning combined with approximate Bayesian learning is often used for *automatic relevance determination (ARD)*, where irrelevant degrees of freedom of the statistical model are automatically pruned out. Explaining the ARD property of approximate Bayesian learning is one of the main topics of theoretical analysis in Parts III and IV.

1.2 Computation

Now, let us explain how Bayesian learning is performed in simple cases. We start from introducing *conjugacy*, an important notion in performing Bayesian learning.

1.2.1 Popular Distributions

Table 1.1 summarizes several distributions that are frequently used as a model distribution (or likelihood function) $p(\mathcal{D}|\boldsymbol{w})$ or a prior distribution $p(\boldsymbol{w})$ in Bayesian learning. The domain $\mathcal{X}$ of the random variable $\boldsymbol{x}$ and the domain $\mathcal{W}$ of the parameters $\boldsymbol{w}$ are shown in the table.

Some of the distributions in Table 1.1 have complicated function forms, involving Beta or Gamma functions. However, such complications are mostly in the *normalization constant*, and can often be ignored when it is sufficient to find the *shape* of a function. In Table 1.1, the normalization constant is separated by a dot, so that one can find the simple main part. As will be seen shortly, we often refer to the normalization constant when we need to perform integration of a function, which is in the same form as the main part of a popular distribution.

Below we summarize abbreviations of distributions:

$$\mathrm{Gauss}_M(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \equiv \frac{1}{(2\pi)^{M/2} \det(\boldsymbol{\Sigma})^{1/2}} \cdot \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right), \tag{1.16}$$

$$\mathrm{Gamma}(x; \alpha, \beta) \equiv \frac{\beta^\alpha}{\Gamma(\alpha)} \cdot x^{\alpha-1} \exp(-\beta x), \tag{1.17}$$