CHAPTER

PROPERTIES OF MAXIMUM LIKELIHOOD ESTIMATION 13

CHAPTER CONTENTS

In the previous chapter, the definition of MLE and its usage in pattern recognition were explained. In this chapter, the asymptotic behavior of MLE is theoretically investigated. Throughout this chapter, the true probability density function $p(\boldsymbol{x})$ is assumed to be included in the parametric model $q(\boldsymbol{x};\boldsymbol{\theta})$, i.e., there exists $\boldsymbol{\theta}^*$ such that $q(\boldsymbol{x};\boldsymbol{\theta}^*) = p(\boldsymbol{x})$.

13.1 CONSISTENCY

First, consistency of MLE is explored. Consistency is the property that the optimal solution $\boldsymbol{\theta}^*$ can be obtained asymptotically (i.e., in the limit that the number of training samples n tends to infinity), which would be a minimum requirement for reasonable estimators.

More precisely, let $\widehat{\boldsymbol{\theta}}_n$ be an estimator obtained from n i.i.d. training samples. Then, an estimator $\widehat{\boldsymbol{\theta}}_n$ is said to be *consistent* if the following property is satisfied for any $\boldsymbol{\theta}^* \in \Theta$ and any $\varepsilon > 0$:

$$\lim_{n\to\infty} \Pr(\|\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*\| \geq \varepsilon) = 0.$$

This means that $\widehat{\boldsymbol{\theta}}_n$ converges in probability to $\boldsymbol{\theta}^*$ (Section 7.3) and is denoted as

$$\widehat{\boldsymbol{\theta}}_n \xrightarrow{\text{p}} \boldsymbol{\theta}^*.$$

MLE was proved to be consistent under mild assumptions. For example, the consistency of the maximum likelihood estimator of the expectation for the one-dimensional Gaussian model, $\widehat{\mu}_{\mathrm{ML}} = \frac{1}{n}\sum_{i=1}^{n} x_i$, was shown in Section 7.3.

An Introduction to Statistical Machine Learning. DOI: 10.1016/B978-0-12-802121-7.00024-8

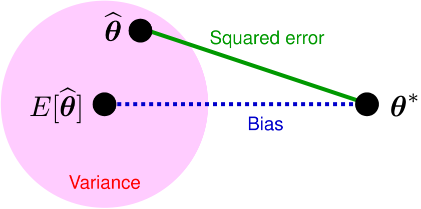
$\widehat{\boldsymbol{\theta}}$
Squared error
$E[\widehat{\boldsymbol{\theta}}]$
$\boldsymbol{\theta}^*$
Bias
Variance

FIGURE 13.1

Bias-variance decomposition of expected squared error.

Next, let us investigate the relation between consistency and the *squared error* $\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|^2$. *Markov's inequality* shown in Section 8.2.1 gives the following upper bound:

$$\Pr(\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\| \geq \varepsilon) = \Pr(\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|^2 \geq \varepsilon^2) \leq \frac{1}{\varepsilon^2}\mathbb{E}\left[\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|^2\right],$$

where $\mathbb{E}\left[\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|^2\right]$ is the *expected squared error* and $\mathbb{E}$ denotes the expectation over all training samples $\{\boldsymbol{x}_i\}_{i=1}^n$ following i.i.d. with $p(\boldsymbol{x})$:

$$\mathbb{E}[\bullet] = \int \cdots \int \bullet\, p(\boldsymbol{x}_1)\cdots p(\boldsymbol{x}_n)\mathrm{d}\boldsymbol{x}_1 \cdots \mathrm{d}\boldsymbol{x}_n. \tag{13.1}$$

This upper bound shows that if the expected squared error of an estimator vanishes asymptotically, the estimator possesses consistency.

13.2 ASYMPTOTIC UNBIASEDNESS

The expected squared error can be decomposed as

$$\begin{aligned}
&\mathbb{E}\left[\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|^2\right] \\
&= \mathbb{E}\left[\|\widehat{\boldsymbol{\theta}} - \mathbb{E}[\widehat{\boldsymbol{\theta}}] + \mathbb{E}[\widehat{\boldsymbol{\theta}}] - \boldsymbol{\theta}^*\|^2\right] \\
&= \mathbb{E}\left[\|\widehat{\boldsymbol{\theta}} - \mathbb{E}[\widehat{\boldsymbol{\theta}}]\|^2\right] + \|\mathbb{E}[\widehat{\boldsymbol{\theta}}] - \boldsymbol{\theta}^*\|^2 + 2\mathbb{E}\left[(\widehat{\boldsymbol{\theta}} - \mathbb{E}[\widehat{\boldsymbol{\theta}}])^\top(\mathbb{E}[\widehat{\boldsymbol{\theta}}] - \boldsymbol{\theta}^*)\right] \\
&= \mathbb{E}\left[\|\widehat{\boldsymbol{\theta}} - \mathbb{E}[\widehat{\boldsymbol{\theta}}]\|^2\right] + \|\mathbb{E}[\widehat{\boldsymbol{\theta}}] - \boldsymbol{\theta}^*\|^2 + 2[(\mathbb{E}[\widehat{\boldsymbol{\theta}}] - \mathbb{E}[\widehat{\boldsymbol{\theta}}])^\top(\mathbb{E}[\widehat{\boldsymbol{\theta}}] - \boldsymbol{\theta}^*)] \\
&= \mathbb{E}\left[\|\widehat{\boldsymbol{\theta}} - \mathbb{E}[\widehat{\boldsymbol{\theta}}]\|^2\right] + \|\mathbb{E}[\widehat{\boldsymbol{\theta}}] - \boldsymbol{\theta}^*\|^2,
\end{aligned}$$

where the first term and second term are called the *variance term* and the *bias term*, respectively (Fig. 13.1).

An estimator $\widehat{\boldsymbol{\theta}}$ is said to be *unbiased* if

$$\mathbb{E}[\widehat{\boldsymbol{\theta}}] = \boldsymbol{\theta}^*,$$

and an estimator $\widehat{\boldsymbol{\theta}}_n$ is said to be *asymptotically unbiased* if

$$\mathbb{E}[\widehat{\boldsymbol{\theta}}_n] \xrightarrow{\text{p}} \boldsymbol{\theta}^*.$$

MLE was shown to be asymptotically unbiased under mild assumptions.

13.3 ASYMPTOTIC EFFICIENCY

Consistency and asymptotic unbiasedness are properties for infinitely many training samples. However, in practice, the number of training samples cannot be infinity, and therefore an estimator with consistency and asymptotic unbiasedness is not necessarily superior. Indeed, as explained in Section 13.2, not only the bias but also the variance has to be taken into account to reduce the expected squared error.

13.3.1 ONE-DIMENSIONAL CASE

Efficiency concerns the variance of an estimator. To explain the concept of efficiency more precisely, let us first consider a one-dimensional case where the parameter to be estimated, θ, is a scalar. Let $\mathbb{V}$ denote the variance operator over all training samples $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ following i.i.d. with $p(\boldsymbol{x})$:

$$\mathbb{V}(\bullet) = \mathbb{E}[(\bullet - \mathbb{E}[\bullet])^2], \tag{13.2}$$

where $\mathbb{E}$ is defined in Eq. (13.1). Then the variance of any unbiased estimator $\widehat{\theta}$ is lower-bounded as

$$\mathbb{V}(\widehat{\theta}) = \mathbb{E}[(\widehat{\theta} - \theta^*)^2] \geq \frac{1}{nF(\theta^*)},$$

which is called the *Cramér-Rao inequality* [31, 82]. Here, $F(\theta)$ is called *Fisher information*:

$$F(\theta) = \int \left(\frac{\partial}{\partial \theta} \log q(\boldsymbol{x}; \theta) \right)^2 q(\boldsymbol{x}; \theta) \mathrm{d}\boldsymbol{x}.$$

The partial derivative of $\log q(\boldsymbol{x}; \theta)$,

$$\frac{\partial}{\partial \theta} \log q(\boldsymbol{x}; \theta), \tag{13.3}$$

is often called the *Fisher score*. An unbiased estimator $\widehat{\theta}$ is said to be efficient if the Cramér-Rao lower bound is attained with strict equality:

$$\mathbb{V}(\widehat{\theta}) = \frac{1}{nF(\theta^*)}.$$

13.3.2 MULTIDIMENSIONAL CASES

To extend the above definition to multidimensional cases, let us define the *Fisher information matrix* $\boldsymbol{F}(\boldsymbol{\theta})$ as

$$\boldsymbol{F}(\boldsymbol{\theta}) = \int \left(\frac{\partial}{\partial \boldsymbol{\theta}} \log q(\boldsymbol{x};\boldsymbol{\theta})\right)\left(\frac{\partial}{\partial \boldsymbol{\theta}^\top} \log q(\boldsymbol{x};\boldsymbol{\theta})\right) q(\boldsymbol{x};\boldsymbol{\theta}) \mathrm{d}\boldsymbol{x}. \tag{13.4}$$

Here, $\frac{\partial}{\partial \boldsymbol{\theta}}$ and $\frac{\partial}{\partial \boldsymbol{\theta}^\top}$ denote the vertical and horizontal vectors of partial derivatives, respectively. That is, the (j, j')th element of $\boldsymbol{F}(\boldsymbol{\theta})$ for $\boldsymbol{\theta} = (\theta^{(1)}, \ldots, \theta^{(b)})^\top$ is given by

$$F_{j,j'}(\boldsymbol{\theta}) = \int \left(\frac{\partial}{\partial \theta^{(j)}} \log q(\boldsymbol{x};\boldsymbol{\theta})\right)\left(\frac{\partial}{\partial \theta^{(j')}} \log q(\boldsymbol{x};\boldsymbol{\theta})\right) q(\boldsymbol{x};\boldsymbol{\theta}) \mathrm{d}\boldsymbol{x}.$$

Then a multidimensional version of the Cramér-Rao inequality is given for any unbiased estimator $\widehat{\boldsymbol{\theta}}$ as

$$\mathbb{V}(\widehat{\boldsymbol{\theta}}) = \mathbb{E}[(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^\top] \geq \frac{1}{n} \boldsymbol{F}(\boldsymbol{\theta}^*)^{-1}.$$

Here, the inequality $\boldsymbol{A} \geq \boldsymbol{B}$ for square matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ means that $\boldsymbol{A} - \boldsymbol{B}$ is a *positive semidefinite matrix*: a matrix $\boldsymbol{C}$ is said to be positive semidefinite if it satisfies $\boldsymbol{\varphi}^\top \boldsymbol{C} \boldsymbol{\varphi} \geq 0$ for any vector $\boldsymbol{\varphi}$. An unbiased estimator $\widehat{\boldsymbol{\theta}}$ is said to be efficient if the above multidimensional version of the Cramér-Rao lower bound is attained with strict equality:

$$\mathbb{V}(\widehat{\boldsymbol{\theta}}) = \frac{1}{n} \boldsymbol{F}(\boldsymbol{\theta}^*)^{-1}.$$

The concept of efficiency is defined for unbiased estimators. Therefore, since MLE is asymptotically unbiased, but not generally unbiased with finite samples, MLE is not efficient. An estimator is said to be *asymptotic efficient* if the Cramér-Rao lower bound is attained asymptotically:

$$n\mathbb{V}(\widehat{\boldsymbol{\theta}}) \xrightarrow{\mathrm{p}} \boldsymbol{F}(\boldsymbol{\theta}^*)^{-1}.$$

MLE was shown to be asymptotically efficient under mild assumptions.

Suppose that $q(\boldsymbol{x};\boldsymbol{\theta})$ is twice differentiable. Then

$$\begin{aligned}
&\int \left(\frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} \log q(\boldsymbol{x};\boldsymbol{\theta})\right) q(\boldsymbol{x};\boldsymbol{\theta}) \mathrm{d}\boldsymbol{x} \\
&= \int \frac{\frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} q(\boldsymbol{x};\boldsymbol{\theta})}{q(\boldsymbol{x};\boldsymbol{\theta})} q(\boldsymbol{x};\boldsymbol{\theta}) \mathrm{d}\boldsymbol{x} - \int \frac{\frac{\partial}{\partial \boldsymbol{\theta}} q(\boldsymbol{x};\boldsymbol{\theta}) \frac{\partial}{\partial \boldsymbol{\theta}^\top} q(\boldsymbol{x};\boldsymbol{\theta})}{q(\boldsymbol{x};\boldsymbol{\theta})^2} q(\boldsymbol{x};\boldsymbol{\theta}) \mathrm{d}\boldsymbol{x} \\
&= \int \frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} q(\boldsymbol{x};\boldsymbol{\theta}) \mathrm{d}\boldsymbol{x} - \int \left(\frac{\partial}{\partial \boldsymbol{\theta}} \log q(\boldsymbol{x};\boldsymbol{\theta})\right)\left(\frac{\partial}{\partial \boldsymbol{\theta}^\top} \log q(\boldsymbol{x};\boldsymbol{\theta})\right) q(\boldsymbol{x};\boldsymbol{\theta}) \mathrm{d}\boldsymbol{x} \\
&= \frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} \int q(\boldsymbol{x};\boldsymbol{\theta}) \mathrm{d}\boldsymbol{x} - \boldsymbol{F}(\boldsymbol{\theta}) = \frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} 1 - \boldsymbol{F}(\boldsymbol{\theta}) = -\boldsymbol{F}(\boldsymbol{\theta}).
\end{aligned}$$

Therefore, the Fisher information matrix defined in Eq. (13.4) can be expressed as

$$\boldsymbol{F}(\boldsymbol{\theta}) = -\int \left(\frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} \log q(\boldsymbol{x};\boldsymbol{\theta})\right) q(\boldsymbol{x};\boldsymbol{\theta}) \mathrm{d}\boldsymbol{x}. \tag{13.5}$$

The matrix

$$\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}\log q(\boldsymbol{x};\boldsymbol{\theta})$$

is called the *Hessian matrix* of $\log q(\boldsymbol{x};\boldsymbol{\theta})$, which plays an important role in optimization theory. Eq. (13.5) shows that the Fisher information matrix agrees with the expected negative Hessian matrix of $\log q(\boldsymbol{x};\boldsymbol{\theta})$.

13.4 ASYMPTOTIC NORMALITY

If an estimator approximately follows the normal distribution when the number of training samples n is large, it is said to possess *asymptotic normality*. In this section, asymptotic normality of the maximum likelihood estimator is explained.

As explained in Section 7.4, the *central limit theorem* asserts that, for n one-dimensional i.i.d. samples $\{x_i\}_{i=1}^n$ having expectation 0 and variance 1, their mean $\overline{x}_n = \frac{1}{n}\sum_{i=1}^n x_i$ satisfies

$$\lim_{n\to\infty}\Pr(a \le \sqrt{n}\overline{x}_n \le b) = \frac{1}{\sqrt{2\pi}}\int_a^b \exp\left(-\frac{x^2}{2}\right)\mathrm{d}x.$$

This means that $\sqrt{n}\overline{x}_n$ asymptotically follows the standard normal distribution.

The central limit theorem can be extended to multidimensions as follows. Let $\{\boldsymbol{x}_i\}_{i=1}^n$ be the i.i.d. samples having expectation $\boldsymbol{0}_d$ and variance-covariance matrix $\boldsymbol{\Sigma}$, where d denotes the dimensionality of $\boldsymbol{x}_i$. Let $\overline{\boldsymbol{x}}_n$ be the sample average:

$$\overline{\boldsymbol{x}}_n = \frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i.$$

Then $\sqrt{n}\overline{\boldsymbol{x}}_n$ approximately follows the d-dimensional normal distribution with expectation $\boldsymbol{0}_d$ and variance-covariance matrix $\boldsymbol{\Sigma}$ (Section 6.2):

$$\sqrt{n}\overline{\boldsymbol{x}}_n \xrightarrow{\mathrm{d}} N(\boldsymbol{0}_d,\boldsymbol{\Sigma}),$$

where “d” denotes the convergence in distribution (see Section 7.4).

Based on the above central limit theorem, asymptotic normality of the maximum likelihood estimator is explained. Let $\boldsymbol{\xi}(\boldsymbol{\theta})$ be the sample average of the *Fisher score* (see Eq. (13.3)):

$$\boldsymbol{\xi}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^n \frac{\partial}{\partial\boldsymbol{\theta}}\log q(\boldsymbol{x}_i;\boldsymbol{\theta}).$$

Since the maximum likelihood estimator $\widehat{\boldsymbol{\theta}}_{\mathrm{ML}}$ maximizes the log-likelihood, it satisfies $\boldsymbol{\xi}(\widehat{\boldsymbol{\theta}}_{\mathrm{ML}}) = \boldsymbol{0}_b$. Then the *Taylor series expansion* (Fig. 2.8) of the left-hand side about the true parameter $\boldsymbol{\theta}^*$ yields

$$\boldsymbol{\xi}(\boldsymbol{\theta}^*) - \widehat{\boldsymbol{F}}(\boldsymbol{\theta}^*)(\widehat{\boldsymbol{\theta}}_{\mathrm{ML}} - \boldsymbol{\theta}^*) + \boldsymbol{r} = \boldsymbol{0}_b,$$

where $\boldsymbol{r}$ is the residual vector that contains higher-order terms and

$$\widehat{\boldsymbol{F}}(\boldsymbol{\theta}) = -\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}\log q(\boldsymbol{x}_i;\boldsymbol{\theta}).$$

Since $\widehat{\boldsymbol{F}}(\boldsymbol{\theta})$ is a sample approximation of Fisher information matrix (13.5), it converges in probability to the true Fisher information matrix in the limit $n \to \infty$:

$$\widehat{\boldsymbol{F}}(\boldsymbol{\theta}) \xrightarrow{\mathrm{p}} \boldsymbol{F}(\boldsymbol{\theta}).$$

Also, consistency of MLE implies that the residual $\boldsymbol{r}$ can be ignored in the limit $n \to \infty$. Thus,

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_{\mathrm{ML}} - \boldsymbol{\theta}^*) \xrightarrow{\mathrm{p}} \boldsymbol{F}(\boldsymbol{\theta}^*)^{-1}\sqrt{n}\boldsymbol{\xi}(\boldsymbol{\theta}^*). \tag{13.6}$$

The true parameter $\boldsymbol{\theta}^*$ maximizes the expected log-likelihood $E[\log q(\boldsymbol{x};\boldsymbol{\theta})]$ with respect to $\boldsymbol{\theta}$, where E is the expectation operator over $p(\boldsymbol{x})$:

$$E[\bullet] = \int \bullet\, p(\boldsymbol{x})\mathrm{d}\boldsymbol{x}.$$

Then the first-order optimality (Fig. 12.1) yields

$$E\left[\frac{\partial}{\partial\boldsymbol{\theta}}\log q(\boldsymbol{x};\boldsymbol{\theta})\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}\right] = \boldsymbol{0}_b.$$

Furthermore, Eq. (13.4) implies

$$V\left[\frac{\partial}{\partial\boldsymbol{\theta}}\log q(\boldsymbol{x};\boldsymbol{\theta})\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}\right] = \boldsymbol{F}(\boldsymbol{\theta}^*),$$

where V denotes the variance operator over $p(\boldsymbol{x})$:

$$V[\bullet] = \int (\bullet - E[\bullet])(\bullet - E[\bullet])^\top p(\boldsymbol{x})\mathrm{d}\boldsymbol{x}.$$

Therefore, the central limit theorem asserts that $\sqrt{n}\boldsymbol{\xi}(\boldsymbol{\theta}^*)$ asymptotically follows the normal distribution with expectation $\boldsymbol{0}_b$ and variance-covariance matrix $\boldsymbol{F}(\boldsymbol{\theta}^*)$:

$$\sqrt{n}\boldsymbol{\xi}(\boldsymbol{\theta}^*) \xrightarrow{\mathrm{d}} N(\boldsymbol{0}_b, \boldsymbol{F}(\boldsymbol{\theta}^*)).$$

Moreover, the variance-covariance matrix of $\boldsymbol{F}(\boldsymbol{\theta}^*)^{-1}\sqrt{n}\boldsymbol{\xi}(\boldsymbol{\theta}^*)$ is given by

$$\begin{aligned}\mathbb{V}\left[\boldsymbol{F}(\boldsymbol{\theta}^*)^{-1}\sqrt{n}\boldsymbol{\xi}(\boldsymbol{\theta}^*)\right] &= \boldsymbol{F}(\boldsymbol{\theta}^*)^{-1}V\left[\sqrt{n}\boldsymbol{\xi}(\boldsymbol{\theta}^*)\right]\boldsymbol{F}(\boldsymbol{\theta}^*)^{-1}\\ &= \boldsymbol{F}(\boldsymbol{\theta}^*)^{-1}\boldsymbol{F}(\boldsymbol{\theta}^*)\boldsymbol{F}(\boldsymbol{\theta}^*)^{-1} = \boldsymbol{F}(\boldsymbol{\theta}^*)^{-1},\end{aligned}$$

```
n=10; t=10000; s=1/12/n;
x=linspace(-0.4,0.4,100);
y=1/sqrt(2*pi*s)*exp(-x.^2/(2*s));
z=mean(rand(t,n)-0.5,2);

figure(1); clf; hold on
b=20; hist(z,b); c=max(z)-min(z);
h=plot(x,y*t/b*c,’r-’);
```

FIGURE 13.2

MATLAB code for illustrating asymptotic normality of MLE.

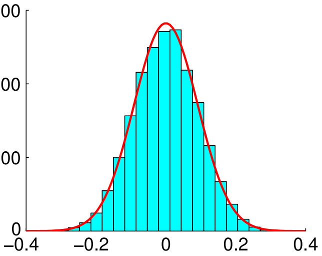

FIGURE 13.3

Example of asymptotic normality of MLE.

where $\mathbb{V}$ denotes the variance with respect to all training samples $\{\boldsymbol{x}_i\}_{i=1}^{n}$ drawn i.i.d. from $p(\boldsymbol{x})$ (see Eq. (13.2)). Therefore, Eq. (13.6) yields

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_{\mathrm{ML}} - \boldsymbol{\theta}^*) \xrightarrow{\mathrm{d}} N(\boldsymbol{0}_b, \boldsymbol{F}(\boldsymbol{\theta}^*)^{-1}).$$

This means that the maximum likelihood estimator $\widehat{\boldsymbol{\theta}}_{\mathrm{ML}}$ asymptotically follows the normal distribution with expectation $\boldsymbol{\theta}^*$ and variance-covariance matrix $\frac{1}{n}\boldsymbol{F}(\boldsymbol{\theta}^*)^{-1}$.

The above asymptotic normality of MLE implies that MLE is asymptotically unbiased. Furthermore, the variance-covariance matrix $\frac{1}{n}\boldsymbol{F}(\boldsymbol{\theta}^*)^{-1}$ vanishes asymptotically, meaning that the bias and variance terms explained in Section 13.2 also vanish asymptotically. Therefore, the expected squared error, which is the sum of the bias and variance terms, also vanishes asymptotically. Moreover, the above results show that the Cramér-Rao lower bound is attained asymptotically, meaning that MLE is asymptotically efficient.

A MATLAB code for illustrating the asymptotic normality of

$$\widehat{\mu}_{\mathrm{ML}} = \frac{1}{n}\sum_{i=1}^{n} x_i,$$

when $p(x)$ is the uniform distribution on $[-0.5, 0.5]$, is given in Fig. 13.2, and its behavior is illustrated in Fig. 13.3.

13.5 SUMMARY

MLE is consistent and asymptotically unbiased, and therefore its validity is theoretically guaranteed when the number of training samples is infinite. Furthermore, since MLE is asymptotically efficient, its high reliability is guaranteed even when the number of training samples is not infinite, but large.

However, when the number of training samples is not large, MLE is not necessarily a good method. Also, efficiency just guarantees that the variance is the smallest among unbiased estimators; the expected squared error, which is the sum of the bias and variance terms, is not guaranteed to be small. Indeed, a slightly biased estimator can have significantly smaller variance than the efficient estimator. In such a case, a biased estimator can have much smaller expected squared error than the efficient estimator. Such a biased estimator will be discussed in Chapter 17 and Chapter 23.