4 Gaussian Distribution

The Gaussian distribution plays a prominent role in inference and learning, especially when we deal with the sum of a large number of samples. In this case, a fundamental result in probability theory, known as the *central limit theorem*, states that under conditions often reasonable in applications, the probability density function (pdf) of the sum of independent random variables approaches that of a Gaussian distribution. It is for this reason that the term “Gaussian noise” generally refers to the combined effect of many independent disturbances. In this chapter, we describe the form of the Gaussian distribution for both scalar and vector random variables, and establish several useful properties and integral expressions that will be used throughout our treatment.

4.1 SCALAR GAUSSIAN VARIABLES

We start with the scalar case. Assume $\{\boldsymbol{x}_n, n = 1, 2, \ldots, N\}$ are independent scalar random variables with means $\{\bar{x}_n\}$ and variances $\{\sigma^2_{x,n}\}$ each. Then, as explained in the comments at the end of the chapter, under some weak technical conditions represented by expressions (4.162)–(4.163), the pdf of the normalized variable:

$$\boldsymbol{y} \;\stackrel{\Delta}{=}\; \frac{\sum_{n=1}^{N}(\boldsymbol{x}_n - \bar{x}_n)}{\left(\sum_{n=1}^{N}\sigma^2_{x,n}\right)^{1/2}} \tag{4.1}$$

can be shown to approach that of a Gaussian distribution with zero mean and unit variance, i.e.,

$$f_{\boldsymbol{y}}(y) = \frac{1}{\sqrt{2\pi}}e^{-y^2/2}, \qquad \text{as } N \to \infty \tag{4.2}$$

or, equivalently,

$$\lim_{N\to\infty} \mathbb{P}(\boldsymbol{y} \le a) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{a} e^{-y^2/2}dy \tag{4.3}$$

More generally, we denote a Gaussian distribution with mean $\bar{y}$ and variance σ_y^2 by the notation $\mathcal{N}_{\boldsymbol{y}}(\bar{y}, \sigma_y^2)$ with pdf given by:

$$\boxed{\boldsymbol{y} \sim \mathcal{N}_{\boldsymbol{y}}(\bar{y}, \sigma_y^2) \iff f_{\boldsymbol{y}}(y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left\{ -\frac{1}{2\sigma_y^2}(y-\bar{y})^2 \right\}} \tag{4.4}$$

Figure 4.1 illustrates the form of the Gaussian distribution using $\bar{y} = 2$ and $\sigma_y^2 = 3$. The next example derives three useful integral expressions.

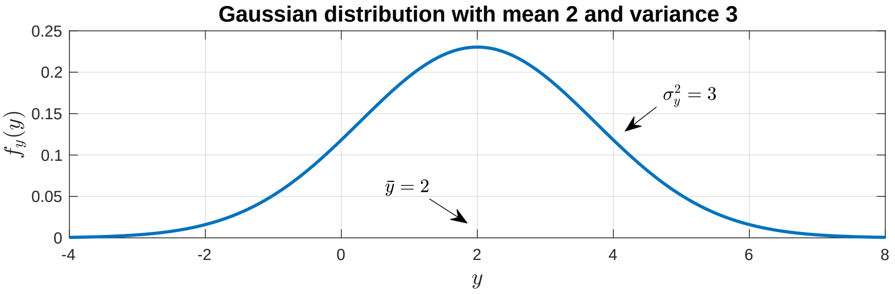

Figure 4.1 Probability density function of a Gaussian distribution with mean $\bar{y} = 2$ and variance $\sigma_y^2 = 3$.

Example 4.1 **(Three useful integral expressions)** There are many results on integrals involving the Gaussian distribution, some of which will appear in our treatment of inference problems. We list some of them here for ease of reference and leave their derivation to the problems.

Let $f_{\boldsymbol{x}}(x)$ denote the standard Gaussian distribution, $\boldsymbol{x} \sim \mathcal{N}_{\boldsymbol{x}}(0,1)$, i.e.,

$$f_{\boldsymbol{x}}(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} = \mathcal{N}_{\boldsymbol{x}}(0,1) \tag{4.5}$$

and introduce its cumulative distribution function (CDF):

$$\Phi(z) \triangleq \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-x^2/2} dx = \int_{-\infty}^{z} \mathcal{N}_{\boldsymbol{x}}(0,1) dx = \mathbb{P}(\boldsymbol{x} \le z) \tag{4.6}$$

This function measures the area under $f_{\boldsymbol{x}}(x)$ from $-\infty$ up to location z. Note that $\Phi(z)$ maps real values z to the interval $[0,1]$. Now, consider a second scalar Gaussian-distributed random variable $\boldsymbol{y}$ with $f_{\boldsymbol{y}}(y) = \mathcal{N}_{\boldsymbol{y}}(\bar{y}, \sigma_y^2)$. One useful integral result is the following identity established in Prob. 4.8 for any a and $\sigma_a > 0$ — a more general result is considered later in Example 4.7:

$$\boxed{Z_0 \triangleq \int_{-\infty}^{\infty} \Phi\left(\frac{y-a}{\sigma_a}\right) \mathcal{N}_{\boldsymbol{y}}(\bar{y}, \sigma_y^2) dy = \Phi\left(\widehat{y}\right)} \tag{4.7}$$

where

$$\widehat{y} \triangleq \frac{\bar{y}-a}{\sqrt{\sigma_y^2+\sigma_a^2}} \tag{4.8}$$

Differentiating (4.7) once and then twice relative to $\bar{y}$ leads to two other useful results involving multiplication of the integrand by y and y^2 — see Prob. 4.9:

$$\boxed{Z_1 \triangleq \int_{-\infty}^{\infty} y\Phi\Big(\frac{y-a}{\sigma_a}\Big)\mathcal{N}_{\boldsymbol{y}}(\bar{y},\sigma_y^2)dy = \bar{y}\,\Phi(\widehat{y}\,) + \frac{\sigma_y^2}{\sqrt{\sigma_y^2+\sigma_a^2}}\,\mathcal{N}_{\widehat{\boldsymbol{y}}}(0,1)} \tag{4.9}$$

and

$$\begin{aligned} Z_2 &\triangleq \int_{-\infty}^{\infty} y^2\Phi\Big(\frac{y-a}{\sigma_a}\Big)\mathcal{N}_{\boldsymbol{y}}(\bar{y},\sigma_y^2)dy \\ &= 2\bar{y}Z_1 + (\sigma_y^2-\bar{y}^2)Z_0 - \frac{\sigma_y^4\,\widehat{y}}{\sigma_y^2+\sigma_a^2}\,\mathcal{N}_{\widehat{\boldsymbol{y}}}(0,1) \end{aligned} \tag{4.10}$$

so that

$$\boxed{Z_2 = (\sigma_y^2+\bar{y}^2)\Phi(\widehat{y}\,) + \frac{\sigma_y^2}{\sqrt{\sigma_y^2+\sigma_a^2}}\left(2\bar{y} - \frac{\sigma_y^2\,\widehat{y}}{\sqrt{\sigma_y^2+\sigma_a^2}}\right)\mathcal{N}_{\widehat{\boldsymbol{y}}}(0,1)} \tag{4.11}$$

All three identities for $\{Z_0, Z_1, Z_2\}$ involve the cumulative function $\Phi(z)$ in the integrand. We will also encounter integrals involving the sigmoid function

$$\sigma(z) \triangleq \frac{1}{1+e^{-z}} \tag{4.12}$$

This function maps real values z to the same interval $[0,1]$. For these integrals, we will employ the useful approximation:

$$\frac{1}{1+e^{-z}} \approx \Phi(bz), \quad \text{where } b^2 = \pi/8 \tag{4.13}$$

4.2 VECTOR GAUSSIAN VARIABLES

Vector Gaussian variables arise frequently in learning and inference problems. We describe next the general form of the pdf for a *vector* Gaussian random variable, and examine several of its properties.

4.2.1 Probability Density Function

We start with a $p\times 1$ random vector $\boldsymbol{x}$ with mean $\bar{x}$ and nonsingular covariance matrix

$$R_x \triangleq \mathbb{E}\,(\boldsymbol{x}-\bar{x})(\boldsymbol{x}-\bar{x})^{\mathsf{T}} > 0 \tag{4.14}$$

We say that $\boldsymbol{x}$ has a Gaussian distribution if its pdf has the form

$$\boxed{f_{\boldsymbol{x}}(x) = \frac{1}{\sqrt{(2\pi)^p}} \frac{1}{\sqrt{\det R_x}} \exp\left\{-\frac{1}{2}(x-\bar{x})^{\mathsf{T}} R_x^{-1}(x-\bar{x})\right\}} \tag{4.15}$$

in terms of the determinant of R_x. Of course, when $p = 1$, the above expression reduces to the pdf considered earlier in (4.4) with R_x replaced by σ_x^2. Figure 4.2 illustrates the form of a two-dimensional Gaussian distribution with

$$\bar{x} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}, \quad R_x = \begin{bmatrix} 1 & 1 \\ 1 & 3 \end{bmatrix} \tag{4.16}$$

The individual entries of $\boldsymbol{x}$ are denoted by $\{\boldsymbol{x}_1, \boldsymbol{x}_2\}$.

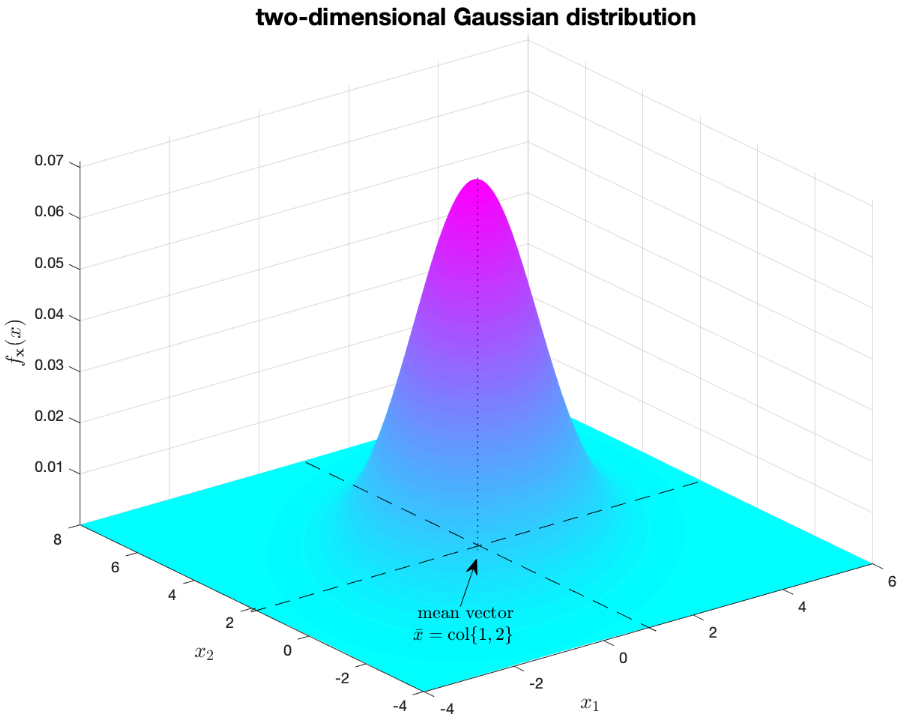

Figure 4.2 Probability density function of a two-dimensional Gaussian distribution with mean vector and covariance matrix given by (4.16).

Example 4.2 **(Noisy measurements of a DC value)** The following example illustrates one possibility by which Gaussian random vectors arise in practice. Consider a collection of N noisy measurements of some unknown constant θ:

$$\boldsymbol{x}(n) = \theta + \boldsymbol{v}(n), \quad n = 1, 2, \ldots, N \tag{4.17}$$

where $\boldsymbol{x}(n)$ is a scalar and $\boldsymbol{v}(n)$ is a zero-mean Gaussian random variable with variance σ_v^2. We assume $\boldsymbol{v}(n)$ and $\boldsymbol{v}(m)$ are independent of each other for all $n \neq m$. Due to the noise, the measurements $\{\boldsymbol{x}(n)\}$ will fluctuate around θ. Each $\boldsymbol{x}(n)$ will be Gaussian distributed with mean θ and variance σ_v^2. We collect the measurements into the vector

$$\boldsymbol{x} \overset{\Delta}{=} \text{col}\Big\{\boldsymbol{x}(1), \boldsymbol{x}(2), \ldots, \boldsymbol{x}(N)\Big\} \tag{4.18}$$

Then, the vector $\boldsymbol{x}$ will have a Gaussian distribution with mean $\theta\mathbb{1}_N$ and covariance matrix $R_x = \sigma_v^2 I_N$:

$$\boldsymbol{x} \sim \mathcal{N}_{\boldsymbol{x}}(\theta\mathbb{1}_N, \sigma_v^2 I_N) \tag{4.19}$$

where the notation $\mathbb{1}$ refers to a vector with all its entries equal to one.

Example 4.3 **(Linear regression model)** Consider next a situation where we observe N noisy scalar measurements $\{\boldsymbol{y}(n)\}$ under Gaussian noise as follows:

$$\boldsymbol{y}(n) = x_n^\mathsf{T}\boldsymbol{w} + \boldsymbol{v}(n), \quad \boldsymbol{v}(n) \sim \mathcal{N}_{\boldsymbol{v}}(0, \sigma_v^2), \quad n = 1, 2, \ldots, N \tag{4.20}$$

where $\{x_n, \boldsymbol{w}\}$ are vectors in $\mathbb{R}^p$ with x_n playing the role of an input vector. The inner product of x_n with $\boldsymbol{w}$ is perturbed by $\boldsymbol{v}(n)$ and results in the measurement $\boldsymbol{y}(n)$. We model the parameter vector $\boldsymbol{w}$ as another Gaussian variable, $\boldsymbol{w} \sim \mathcal{N}_{\boldsymbol{w}}(\bar{w}, R_w)$. For simplicity, we assume the noise samples $\boldsymbol{v}(n)$ and $\boldsymbol{v}(m)$ are independent of each other for $n \neq m$. We also assume $\boldsymbol{v}(n)$ and $\boldsymbol{w}$ are independent of each other for all n. Observe from (4.20) that the inner product $x_n^\mathsf{T}\boldsymbol{w}$ combines the entries of x_n linearly, which explains the designation "linear regression model." We will encounter models of this type frequently in our treatment.

We rewrite the measurement equation in the equivalent form:

$$\boldsymbol{y}(n) = \boldsymbol{g}(x_n) + \boldsymbol{v}(n), \quad \boldsymbol{g}(x_n) \overset{\Delta}{=} x_n^\mathsf{T}\boldsymbol{w} \tag{4.21}$$

where we introduced the scalar-valued function $\boldsymbol{g}(x) = x^\mathsf{T}\boldsymbol{w}$; its values are random in view of the randomness in $\boldsymbol{w}$. We collect the measurements $\{\boldsymbol{y}(n)\}$, the input vectors $\{x_n\}$, and the values of $\boldsymbol{g}(\cdot)$ and $\boldsymbol{v}(\cdot)$ into matrix and vector quantities and write:

$$X \overset{\Delta}{=} \begin{bmatrix} x_1^\mathsf{T} \\ x_2^\mathsf{T} \\ \vdots \\ x_N^\mathsf{T} \end{bmatrix} \tag{4.22a}$$

$$\boldsymbol{g} \overset{\Delta}{=} X\boldsymbol{w}, \quad X \in \mathbb{R}^{N\times p} \tag{4.22b}$$

$$\underbrace{\begin{bmatrix} \boldsymbol{y}(1) \\ \boldsymbol{y}(2) \\ \vdots \\ \boldsymbol{y}(N) \end{bmatrix}}_{\boldsymbol{y}} = \underbrace{\begin{bmatrix} \boldsymbol{g}(x_1) \\ \boldsymbol{g}(x_2) \\ \vdots \\ \boldsymbol{g}(x_N) \end{bmatrix}}_{\boldsymbol{g}} + \underbrace{\begin{bmatrix} \boldsymbol{v}(1) \\ \boldsymbol{v}(2) \\ \vdots \\ \boldsymbol{v}(N) \end{bmatrix}}_{\boldsymbol{v}} \tag{4.22c}$$

$$\boldsymbol{y} = X\boldsymbol{w} + \boldsymbol{v} \tag{4.22d}$$

Note that the input factor $\boldsymbol{g}$ is Gaussian distributed with mean and covariance matrix given by

$$\bar{g} \overset{\Delta}{=} \mathbb{E}\boldsymbol{g} = X\bar{w} \tag{4.23a}$$

and

$$\begin{aligned} R_g &= \mathbb{E}\,(\boldsymbol{g}-\bar{g})(\boldsymbol{g}-\bar{g})^{\mathsf{T}} \\ &= \mathbb{E}\,X(\boldsymbol{w}-\bar{w})(\boldsymbol{w}-\bar{w})^{\mathsf{T}}X^{\mathsf{T}} \\ &= XR_wX^{\mathsf{T}} \end{aligned} \tag{4.23b}$$

where each entry of R_g contains the cross-covariance between individual entries of $\boldsymbol{g}$, namely,

$$[R_g]_{m,n} \;\overset{\Delta}{=}\; \mathbb{E}\Big(\boldsymbol{g}(x_m)-\bar{g}(x_m)\Big)\Big(\boldsymbol{g}(x_n)-\bar{g}(x_n)\Big) \;=\; x_n^{\mathsf{T}}R_wx_m \tag{4.23c}$$

Note further that the mean and covariance matrix of the measurement vector are given by

$$\bar{y} \;\overset{\Delta}{=}\; \mathbb{E}\,\boldsymbol{y} \;=\; \mathbb{E}\,X\boldsymbol{w}+\mathbb{E}\,\boldsymbol{v} \;=\; X\bar{w}+0 = X\bar{w} \tag{4.24a}$$

and

$$\begin{aligned} R_y \;&\overset{\Delta}{=}\; \mathbb{E}\,(y-\bar{y})(y-\bar{y})^{\mathsf{T}} \\ &=\; \mathbb{E}\Big(X(\boldsymbol{w}-\bar{w})+\boldsymbol{v}\Big)\Big(X(\boldsymbol{w}-\bar{w})+\boldsymbol{v}\Big)^{\mathsf{T}} \\ &=\; \sigma_v^2 I_N + XR_wX^{\mathsf{T}} \end{aligned} \tag{4.24b}$$

Using future result (4.39) that the sum of two independent Gaussian random variables is another Gaussian random variable, we conclude that $\boldsymbol{y}$ is a Gaussian distributed vector with

$$\boldsymbol{y} \sim \mathcal{N}_{\boldsymbol{y}}\Big(X\bar{w},\; \sigma_v^2 I_N + XR_wX^{\mathsf{T}}\Big) \tag{4.25}$$

Example 4.4 **(Fourth-order moment)** We derive a useful result concerning the fourth-order moment of a Gaussian random vector. Thus, let $\boldsymbol{x}$ denote a real-valued Gaussian random column vector with zero mean and a diagonal covariance matrix, say, $\mathbb{E}\,\boldsymbol{x}\boldsymbol{x}^{\mathsf{T}} = \Lambda$. Then, for any symmetric matrix W of compatible dimensions it holds that:

$$\boxed{\mathbb{E}\Big\{\boldsymbol{x}\boldsymbol{x}^{\mathsf{T}}W\boldsymbol{x}\boldsymbol{x}^{\mathsf{T}}\Big\} = \Lambda\mathrm{Tr}\big(W\Lambda\big) + 2\Lambda W\Lambda} \tag{4.26}$$

Proof of (4.26): The argument is based on the fact that uncorrelated Gaussian random variables are also independent (see Prob. 4.4), so that if $\boldsymbol{x}_n$ is the $n-$th element of $\boldsymbol{x}$, then $\boldsymbol{x}_n$ is independent of $\boldsymbol{x}_m$ for $n \neq m$. Now let S denote the desired matrix, i.e., $S = \mathbb{E}\,\boldsymbol{x}\boldsymbol{x}^{\mathsf{T}}W\boldsymbol{x}\boldsymbol{x}^{\mathsf{T}}$, and let S_{nm} denote its $(n,m)-$th element. Assume also that $\boldsymbol{x}$ is p-dimensional. Then

$$S_{nm} = \mathbb{E}\left\{\boldsymbol{x}_n\boldsymbol{x}_m\left(\sum_{i=0}^{p-1}\sum_{j=0}^{p-1}\boldsymbol{x}_iW_{ij}\boldsymbol{x}_j\right)\right\} \tag{4.27}$$

The right-hand side is nonzero only when there are two pairs of equal indices $\{n = m,\ i = j\}$ or $\{n = i,\ m = j\}$ or $\{n = i,\ m = i\}$. Assume first that $n = m$ (which corresponds to the diagonal elements of S). Then, the expectation is nonzero only for $i = j$, i.e.,

$$S_{nn} = \mathbb{E}\left\{\boldsymbol{x}_n^2\sum_{i=0}^{p-1}W_{ii}\boldsymbol{x}_i^2\right\} \;=\; \sum_{i=0}^{p-1}W_{ii}\mathbb{E}\left\{\boldsymbol{x}_n^2\boldsymbol{x}_i^2\right\} \;=\; \lambda_n\mathrm{Tr}\big(W\Lambda\big) + 2W_{nn}\lambda_n^2 \tag{4.28}$$

where we used the fact that for a zero-mean *real* scalar-valued Gaussian random variable $\boldsymbol{a}$ we have $\mathbb{E}\,\boldsymbol{a}^4 = 3\big(\mathbb{E}\,\boldsymbol{a}^2\big)^2 = 3\sigma_a^4$, where $\sigma_a^2 = \mathbb{E}\,\boldsymbol{a}^2$ — see Prob. 4.13. We are also denoting the diagonal entries of Λ by $\{\lambda_n\}$.

For the off-diagonal elements of S (i.e., for $n \neq m$), we must have either $n = j$, $m = i$, or $n = i$, $m = j$, so that

$$\begin{aligned} S_{nm} &= \mathbb{E}\left\{\boldsymbol{x}_n\boldsymbol{x}_m\big(\boldsymbol{x}_n W_{nm}\boldsymbol{x}_m\big)\right\} + \mathbb{E}\left\{\boldsymbol{x}_n\boldsymbol{x}_m\big(\boldsymbol{x}_m W_{mn}\boldsymbol{x}_n\big)\right\} \\ &= \Big(W_{nm} + W_{mn}\Big)\mathbb{E}\left\{\boldsymbol{x}_n^2\boldsymbol{x}_m^2\right\} \\ &= \Big(W_{nm} + W_{mn}\Big)\lambda_n\lambda_m \end{aligned} \tag{4.29}$$

Using the fact that W is symmetric, so that $W_{nm} = W_{mn}$, and collecting the expressions for S_{nm}, in both cases of $n = m$ and $n \neq m$, into matrix form we arrive at the desired result (4.26).

■

We assumed the covariance matrix of $\boldsymbol{x}$ to be diagonal in expression (4.26) to facilitate the derivation. However, it can be verified that the result holds more generally for arbitrary covariance matrices, $R_x = \mathbb{E}\,\boldsymbol{x}\boldsymbol{x}^\mathsf{T}$, and would take the following form with Λ replaced by R_x — see Prob. 4.21:

$$\boxed{\mathbb{E}\left\{\boldsymbol{x}\boldsymbol{x}^\mathsf{T} W \boldsymbol{x}\boldsymbol{x}^\mathsf{T}\right\} = R_x \mathrm{Tr}\big(W R_x\big) + 2R_x W R_x} \tag{4.30}$$

4.3 USEFUL GAUSSIAN MANIPULATIONS

The fact that the Gaussian pdf is normalized and must integrate to one can be exploited to derive a useful multi-dimensional integration result for quadratic forms, as well as useful expressions for integrals involving the product and division of Gaussian distributions.

Multidimensional integral

Consider a $p \times p$ positive-definite matrix A, a $p \times 1$ vector b, a scalar α, and introduce the quadratic form:

$$J(x) \stackrel{\Delta}{=} -\frac{1}{2}x^\mathsf{T} A x + b^\mathsf{T} x + \alpha \tag{4.31}$$

It is straightforward to verify that

$$J(x) = -\frac{1}{2}(x - A^{-1}b)^\mathsf{T} A (x - A^{-1}b) + \alpha + \frac{1}{2}b^\mathsf{T} A^{-1} b \tag{4.32}$$

so that

$$\begin{aligned}
&\int_{-\infty}^{\infty} e^{J(x)}dx \\
&= \int_{-\infty}^{\infty} \exp\Big\{-\frac{1}{2}(x-A^{-1}b)A(x-A^{-1}b)+\alpha+\frac{1}{2}b^{\mathsf{T}}A^{-1}b\Big\}dx \\
&= \exp\Big\{\alpha+\frac{1}{2}b^{\mathsf{T}}A^{-1}b\Big\}\int_{-\infty}^{\infty} \exp\Big\{-\frac{1}{2}(x-A^{-1}b)A(x-A^{-1}b)\Big\}dx \\
&= \exp\Big\{\alpha+\frac{1}{2}b^{\mathsf{T}}A^{-1}b\Big\}\sqrt{(2\pi)^p}\sqrt{\det A^{-1}}\times \\
&\qquad \int_{-\infty}^{\infty} \underbrace{\frac{1}{\sqrt{(2\pi)^p}}\frac{1}{\sqrt{\det A^{-1}}}\exp\Big\{-\frac{1}{2}(x-A^{-1}b)A(x-A^{-1}b)\Big\}\,dx}_{=\mathcal{N}_{\boldsymbol{x}}(A^{-1}b,A^{-1})}
\end{aligned} \tag{4.33}$$

and, consequently, for $A > 0$:

$$\boxed{\int_{-\infty}^{\infty} \exp\Big\{-\frac{1}{2}x^{\mathsf{T}}Ax+b^{\mathsf{T}}x+\alpha\Big\}dx \;=\; \sqrt{\frac{(2\pi)^p}{\det A}}\times\exp\Big\{\alpha+\frac{1}{2}b^{\mathsf{T}}A^{-1}b\Big\}} \tag{4.34}$$

Sum of Gaussian distributions

The sum of two independent Gaussian distributions is another Gaussian distribution. Specifically, let $\boldsymbol{x} \sim \mathcal{N}_{\boldsymbol{x}}(\bar{x}, R_x)$ and $\boldsymbol{y} \sim \mathcal{N}_{\boldsymbol{y}}(\bar{y}, R_y)$ be two independent Gaussian random variables and introduce their sum $\boldsymbol{z} = \boldsymbol{x} + \boldsymbol{y}$. It is clear that the mean and covariance matrix of $\boldsymbol{z}$ are given by:

$$\bar{z} \overset{\Delta}{=} \mathbb{E}\,\boldsymbol{z} = \bar{x} + \bar{y} \tag{4.35}$$

$$R_z \overset{\Delta}{=} \mathbb{E}\,(\boldsymbol{z}-\bar{z})(\boldsymbol{z}-\bar{z})^{\mathsf{T}} \;=\; R_x + R_y \tag{4.36}$$

The pdf of $\boldsymbol{z}$, on the other hand, is given by the following convolution expression in view of the independence of $\boldsymbol{z}$ and $\boldsymbol{y}$ — recall result (3.160):

$$f_{\boldsymbol{z}}(z) = \int_{-\infty}^{\infty} f_{\boldsymbol{x}}(x)\, f_{\boldsymbol{y}}(z-x)dx \tag{4.37}$$

which involves the integral of the product of two Gaussian distributions. Ignoring the normalization factors we have

$$\begin{aligned}
&f_{\boldsymbol{z}}(z) \propto \\
&\int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2}(x-\bar{x})^{\mathsf{T}}R_x^{-1}(x-\bar{x})\right\}\exp\left\{-\frac{1}{2}(z-x-\bar{y})^{\mathsf{T}}R_y^{-1}(z-x-\bar{y})\right\}dx
\end{aligned} \tag{4.38}$$

The integrand is an exponential function whose exponent is quadratic in x. It can then be verified by using identity (4.34) that the integration leads to a Gaussian pdf with mean $\bar{z} = \bar{x}+\bar{y}$ and covariance matrix $R_z = R_x + R_y$ — see Prob. 4.12:

$$\boxed{\left.\begin{array}{l}\boldsymbol{x}\sim\mathcal{N}_{\boldsymbol{x}}(\bar{x},R_x),\ \boldsymbol{y}\sim\mathcal{N}_{\boldsymbol{y}}(\bar{y},R_y)\\ \boldsymbol{x}\text{ and }\boldsymbol{y}\text{ independent}\end{array}\right\}\Longrightarrow \boldsymbol{z}=\boldsymbol{x}+\boldsymbol{y}\sim\mathcal{N}_{\boldsymbol{z}}(\bar{x}+\bar{y},R_x+R_y)} \tag{4.39}$$

Product of Gaussian distributions

Consider two Gaussian distributions over the *same* random variable $\boldsymbol{x}$, say, $\mathcal{N}_{\boldsymbol{x}}(\bar{x}_a, R_a)$ and $\mathcal{N}_{\boldsymbol{x}}(\bar{x}_b, R_b)$:

$$f_{\boldsymbol{x},a}(x) = \frac{1}{\sqrt{(2\pi)^p}}\,\frac{1}{\sqrt{\det R_a}}\,\exp\left\{-\frac{1}{2}(x-\bar{x}_a)^{\mathsf{T}}R_a^{-1}(x-\bar{x}_a)\right\} \tag{4.40a}$$

$$f_{\boldsymbol{x},b}(x) = \frac{1}{\sqrt{(2\pi)^p}}\,\frac{1}{\sqrt{\det R_b}}\,\exp\left\{-\frac{1}{2}(x-\bar{x}_b)^{\mathsf{T}}R_b^{-1}(x-\bar{x}_b)\right\} \tag{4.40b}$$

Lemma 4.1. (Product of two Gaussians) *Let $g(x) = f_{\boldsymbol{x},a}(x)f_{\boldsymbol{x},b}(x)$ denote the product of two Gaussian distributions over the same variable $\boldsymbol{x}$, where $f_{\boldsymbol{x},a}(x)\sim\mathcal{N}_{\boldsymbol{x}}(\bar{x}_a,R_a)$ and $f_{\boldsymbol{x},b}(x)\sim\mathcal{N}_{\boldsymbol{x}}(\bar{x}_b,R_b)$. The product is an un-normalized Gaussian distribution given by:*

$$g(x) = Z\times\mathcal{N}_{\boldsymbol{x}}(\bar{x}_c,R_c) \tag{4.41}$$

where

$$R_c^{-1} = R_a^{-1}+R_b^{-1} \tag{4.42a}$$

$$\bar{x}_c = R_c\Big(R_a^{-1}\bar{x}_a + R_b^{-1}\bar{x}_b\Big) \tag{4.42b}$$

$$Z = \frac{1}{\sqrt{(2\pi)^p}}\,\frac{1}{\sqrt{\det(R_a+R_b)}}\,\exp\left\{-\frac{1}{2}(\bar{x}_a-\bar{x}_b)^{\mathsf{T}}(R_a+R_b)^{-1}(\bar{x}_a-\bar{x}_b)\right\} \tag{4.42c}$$

Before proving the result, we recall from Prob. 1.11 that we can rewrite the expression for $\bar{x}_c$ in the equivalent forms:

$$\bar{x}_c = \bar{x}_a + R_cR_b^{-1}(\bar{x}_b-\bar{x}_a) \tag{4.43a}$$

$$= \bar{x}_b + R_cR_a^{-1}(\bar{x}_a-\bar{x}_b) \tag{4.43b}$$

Likewise, we can rewrite the expression for R_c as

$$R_c = R_a - R_a(R_a+R_b)^{-1}R_a \tag{4.44a}$$

$$= R_b - R_b(R_a+R_b)^{-1}R_b \tag{4.44b}$$

Observe further that the expression for Z has the form of a Gaussian distribution, say, over the variable $\bar{\boldsymbol{x}}_a$:

$$Z \;=\; \mathcal{N}_{\bar{\boldsymbol{x}}_a}(\bar{x}_b, R_a+R_b) \tag{4.45}$$

Proof of (4.42a)–(4.42c) To begin with, note that

$$\begin{aligned} g(x) = {} & \frac{1}{\sqrt{(2\pi)^p}}\,\frac{1}{\sqrt{\det R_a}} \times \frac{1}{\sqrt{(2\pi)^p}}\,\frac{1}{\sqrt{\det R_b}} \times \\ & \exp\left\{-\frac{1}{2}(x-\bar{x}_a)^{\mathsf{T}} R_a^{-1}(x-\bar{x}_a) - \frac{1}{2}(x-\bar{x}_b)^{\mathsf{T}} R_b^{-1}(x-\bar{x}_b)\right\} \end{aligned} \tag{4.46}$$

The term in the exponent is quadratic in x. Expanding it gives:

$$\begin{aligned} & (x-\bar{x}_a)^{\mathsf{T}} R_a^{-1}(x-\bar{x}_a) + (x-\bar{x}_b)^{\mathsf{T}} R_b^{-1}(x-\bar{x}_b) \\ &= x^{\mathsf{T}} \underbrace{(R_a^{-1}+R_b^{-1})}_{\triangleq\, R_c^{-1}} x - 2\underbrace{(R_a^{-1}\bar{x}_a + R_b^{-1}\bar{x}_b)^{\mathsf{T}}}_{\triangleq\, \bar{x}_c^{\mathsf{T}} R_c^{-1}} x + \underbrace{\bar{x}_a R_a^{-1}\bar{x}_a + \bar{x}_b^{\mathsf{T}} R_b^{-1}\bar{x}_b}_{\triangleq\, \alpha} \\ &= x^{\mathsf{T}} R_c^{-1} x - 2\bar{x}_c^{\mathsf{T}} R_c^{-1} x + \alpha \\ &= x^{\mathsf{T}} R_c^{-1} x - 2\bar{x}_c^{\mathsf{T}} R_c^{-1} x + \bar{x}_c^{\mathsf{T}} R_c^{-1}\bar{x}_c \underbrace{-\bar{x}_c^{\mathsf{T}} R_c^{-1}\bar{x}_c + \alpha}_{\triangleq\, \beta} \\ &= (x-\bar{x}_c)^{\mathsf{T}} R_c^{-1}(x-\bar{x}_c) + \beta \end{aligned} \tag{4.47}$$

where

$$\begin{aligned} \beta &= \alpha - \bar{x}_c^{\mathsf{T}} R_c^{-1}\bar{x}_c \\ &= \bar{x}_a R_a^{-1}\bar{x}_a + \bar{x}_b^{\mathsf{T}} R_b^{-1}\bar{x}_b - (R_a^{-1}\bar{x}_a + R_b^{-1}\bar{x}_b)^{\mathsf{T}} R_c (R_a^{-1}\bar{x}_a + R_b^{-1}\bar{x}_b) \\ &= \bar{x}_a^{\mathsf{T}}(R_a^{-1} - R_a^{-1}R_cR_a^{-1})\bar{x}_a + \bar{x}_b^{\mathsf{T}}(R_b^{-1} - R_b^{-1}R_cR_b^{-1})\bar{x}_b \\ &\quad -\bar{x}_a^{\mathsf{T}}(R_a+R_b)^{-1}x_b - \bar{x}_b^{\mathsf{T}}(R_a+R_b)^{-1}\bar{x}_a \\ &= \bar{x}_a^{\mathsf{T}}(R_a+R_b)^{-1}\bar{x}_a + \bar{x}_b^{\mathsf{T}}(R_a+R_b)^{-1}\bar{x}_b \\ &\quad -\bar{x}_a^{\mathsf{T}}(R_a+R_b)^{-1}x_b - \bar{x}_b^{\mathsf{T}}(R_a+R_b)^{-1}\bar{x}_a \\ &= (\bar{x}_a - \bar{x}_b)^{\mathsf{T}}(R_a+R_b)^{-1}(\bar{x}_a - \bar{x}_b) \end{aligned} \tag{4.48}$$

We therefore conclude that

$$g(x) = Z_1 \times \exp\left\{-\frac{1}{2}(x-\bar{x}_c)^{\mathsf{T}} R_c^{-1}(x-\bar{x}_c)\right\} \tag{4.49}$$

where

$$Z_1 = \frac{1}{(2\pi)^p}\,\frac{1}{\sqrt{\det R_a}}\frac{1}{\sqrt{\det R_b}} \exp\left\{-\frac{1}{2}(\bar{x}_a-\bar{x}_b)^{\mathsf{T}}(R_a+R_b)^{-1}(\bar{x}_a-\bar{x}_b)\right\} \tag{4.50}$$

We can re-normalize $g(x)$ to transform its exponential term into a Gaussian distribution as follows. Introduce the block matrix:

$$X \triangleq \begin{bmatrix} R_a + R_b & R_a \\ R_a & R_a \end{bmatrix} \tag{4.51}$$

We can express the determinant of X in two equivalent ways using the Schur complements relative to $R_a + R_b$ (which is equal to R_c) and the Schur complement relative to R_a (which is equal to R_b):

$$\det X = \det(R_a + R_b) \times \det R_c = \det R_a \times \det R_b \tag{4.52}$$

so that

$$\det R_c = \frac{\det R_a \times \det R_b}{\det(R_a + R_b)} \tag{4.53}$$

Using this expression to replace the terms involving $\det R_a$ and $\det R_b$ in (4.50) we arrive at (4.41).

■

We conclude from (4.41) that the product of two Gaussian distributions is an *unnormalized* Gaussian. Equivalently, we obtain a properly normalized Gaussian through scaling by Z as follows:

$$\boxed{\frac{1}{Z} \times \mathcal{N}_{\boldsymbol{x}}(\bar{x}_a, R_a) \times \mathcal{N}_{\boldsymbol{x}}(\bar{x}_b, R_b) \;=\; \mathcal{N}_{\boldsymbol{x}}(\bar{x}_c, R_c)} \tag{4.54}$$

Obviously, the scaling factor Z has the interpretation

$$Z = \int_{-\infty}^{\infty} \mathcal{N}_{\boldsymbol{x}}(\bar{x}_a, R_a) \times \mathcal{N}_{\boldsymbol{x}}(\bar{x}_b, R_b) dx \tag{4.55}$$

Division of Gaussian distributions

Consider the same Gaussian distributions over the random variable $\boldsymbol{x}$. Let now $g(x) = f_{\boldsymbol{x},a}(x)/f_{\boldsymbol{x},b}(x)$ denote their ratio. Repeating the previous arguments we find that

$$\boxed{g(x) = Z_1 \times \exp\left\{-\frac{1}{2}(x - \bar{x}_c)^{\mathsf{T}} R_c^{-1}(x - \bar{x}_c)\right\}} \tag{4.56}$$

where now

$$R_c^{-1} = R_a^{-1} - R_b^{-1} \tag{4.57a}$$

$$\bar{x}_c = R_c(R_a^{-1}\bar{x}_a - R_b^{-1}\bar{x}_b) \tag{4.57b}$$

$$Z_1 = \frac{\sqrt{\det R_b}}{\sqrt{\det R_a}} \exp\left\{-\frac{1}{2}(\bar{x}_a - \bar{x}_b)^{\mathsf{T}}(R_a - R_b)^{-1}(\bar{x}_a - \bar{x}_b)\right\} \tag{4.57c}$$

Observe that in this case, the matrix R_c is not guaranteed to be positive-definite; it can become indefinite. When R_c is positive-definite (which happens when $R_a < R_b$), we observe that $g(x)$ will have the form of an *unnormalized* Gaussian distribution and it can be normalized by noting that

$$\frac{1}{Z_1}\frac{1}{\sqrt{(2\pi)^p}}\frac{1}{\sqrt{\det R_c}}\, g(x) \;=\; \mathcal{N}_{\boldsymbol{x}}(\bar{x}_c, R_c) \tag{4.58}$$

and, consequently,

$$\boxed{\frac{1}{Z}\frac{\mathcal{N}_{\boldsymbol{x}}(\bar{x}_a, R_a)}{\mathcal{N}_{\boldsymbol{x}}(\bar{x}_b, R_b)} \;=\; \mathcal{N}_{\boldsymbol{x}}(\bar{x}_c, R_c)} \tag{4.59}$$

where

$$Z \;\overset{\Delta}{=}\; \sqrt{(2\pi)^p}\,\sqrt{\det R_c}\; Z_1 \tag{4.60}$$

Let us introduce the block matrix

$$X \;\overset{\Delta}{=}\; \begin{bmatrix} R_b - R_a & R_b \\ R_b & R_b \end{bmatrix} \tag{4.61}$$

Its Schur complement relative to $(R_b - R_a)$ is equal to $-R_c$, whereas its Schur complement relative to R_b is $-R_a$. It follows that

$$\det X = \det(R_b - R_a) \times \det(-R_c) = \det R_b \times \det(-R_a) \tag{4.62}$$

Using the fact that $\det(-A) = (-1)^p \det(A)$ for $p \times p$ matrices, we conclude that

$$\det R_c = \frac{\det R_b \det R_a}{\det(R_b - R_a)} \tag{4.63}$$

so that Z admits the following expression as well:

$$\boxed{Z = \sqrt{(2\pi)^p}\,\frac{\det R_b}{\sqrt{\det(R_b - R_a)}}\exp\left\{-\frac{1}{2}(\bar{x}_a - \bar{x}_b)^{\mathsf{T}}(R_a - R_b)^{-1}(\bar{x}_a - \bar{x}_b)\right\}} \tag{4.64}$$

Stein lemma

A useful result pertaining to the evaluation of expectations involving transformations of Gaussian variables is Stein Lemma. We state the result for vector random variables. Let $\boldsymbol{x} \in \mathbb{R}^p$ denote a Gaussian-distributed random variable with mean $\bar{x}$ and covariance matrix R_x:

$$f_{\boldsymbol{x}}(x) = \frac{1}{\sqrt{(2\pi)^p}}\,\frac{1}{\sqrt{\det R_x}}\,\exp\left\{-\frac{1}{2}(x-\bar{x})^{\mathsf{T}}R_x^{-1}(x-\bar{x})\right\} \tag{4.65}$$

We will often encounter situations where it is necessary to compute expectations of terms of the form $\boldsymbol{x}g(\boldsymbol{x})$ for some scalar-valued function $g(x)$. This computation is equivalent to evaluating integral expressions of the form:

$$\mathbb{E}\,\boldsymbol{x}g(\boldsymbol{x}) = \frac{1}{\sqrt{(2\pi)^p}}\,\frac{1}{\sqrt{\det R_x}}\int_{-\infty}^{\infty} xg(x)\exp\left\{-\frac{1}{2}(x-\bar{x})^{\mathsf{T}}R_x^{-1}(x-\bar{x})\right\}dx \tag{4.66}$$

Lemma 4.2. (Stein lemma) *Assume the function $g(\boldsymbol{x})$ satisfies the finite expectation conditions $\mathbb{E}\,|\partial g(\boldsymbol{x})/\partial x_m| < \infty$, relative to the individual entries of x. Then, it holds that*

$$\boxed{\mathbb{E}\,(\boldsymbol{x}-\bar{x})g(\boldsymbol{x}) \;=\; R_x\,\mathbb{E}\,\nabla_{\boldsymbol{x}^{\mathsf{T}}}\,g(\boldsymbol{x})} \tag{4.67}$$

For scalar Gaussian random variables $\boldsymbol{x} \sim \mathcal{N}_{\boldsymbol{x}}(\bar{x}, \sigma_x^2)$, the lemma reduces to

$$\mathbb{E}\,(\boldsymbol{x}-\bar{x})g(\boldsymbol{x}) = \sigma_x^2\,\mathbb{E}\,g'(\boldsymbol{x}) \tag{4.68}$$

in terms of the derivative of $g(x)$. Later, in Example 5.2 we extend Stein lemma to the exponential family of distributions.

Proof: We establish the result for scalar $\boldsymbol{x}$ and defer the vector case to Prob. 4.33 — see also future Example 5.2. Thus, note that in the scalar case:

$$\mathbb{E}\,(\boldsymbol{x}-\bar{x})g(\boldsymbol{x}) = \frac{1}{\sqrt{2\pi\sigma_x^2}}\int_{-\infty}^{\infty}(x-\bar{x})g(x)\exp\left\{-\frac{1}{2\sigma_x^2}(x-\bar{x})^2\right\}dx \tag{4.69}$$

We carry out the integration by parts. Let

$$u = g(x) \implies du = g'(x)dx \tag{4.70}$$

and

$$dv = (x-\bar{x})\exp\left\{-\frac{1}{2\sigma_x^2}(x-\bar{x})^2\right\}dx \implies v = -\sigma_x^2\exp\left\{-\frac{1}{2\sigma_x^2}(x-\bar{x})^2\right\} \tag{4.71}$$

It follows that

$$\begin{aligned}
\mathbb{E}\,(\boldsymbol{x}-\bar{x})g(\boldsymbol{x}) &= \frac{1}{\sqrt{2\pi\sigma_x^2}}\left\{uv\Big|_{-\infty}^{\infty} - \int_u vdu\right\} \\
&= -\frac{1}{\sqrt{2\pi\sigma_x^2}}g(x)\sigma_x^2\exp\left\{-\frac{1}{2\sigma_x^2}(x-\bar{x})^2\right\}\Bigg|_{-\infty}^{\infty} + \\
&\quad \sigma_x^2\left(\frac{1}{\sqrt{2\pi\sigma_x^2}}\int_{-\infty}^{\infty} g'(x)\exp\left\{-\frac{1}{2\sigma_x^2}(x-\bar{x})^2\right\}dx\right) \\
&= 0 + \sigma_x^2\,\mathbb{E}\,g'(\boldsymbol{x})
\end{aligned} \tag{4.72}$$

as claimed. The mean of $g'(\boldsymbol{x})$ exists in view of the condition $\mathbb{E}\,|g'(\boldsymbol{x})| < \infty$.

■

Example 4.5 (**Fifth-order moment of Gaussian**) Let us apply Stein lemma to evaluate the 5th-order moment of a Gaussian distribution, $\boldsymbol{x} \sim \mathcal{N}_{\boldsymbol{x}}(\bar{x}, \sigma_x^2)$. Thus, note the following sequence of calculations using (4.68):

$$\begin{aligned}
\mathbb{E}\,\boldsymbol{x}^5 &= \mathbb{E}\,(\boldsymbol{x}-\bar{x}+\bar{x})\boldsymbol{x}^4 \\
&= \mathbb{E}\,(\boldsymbol{x}-\bar{x})\boldsymbol{x}^4 \;+\; \bar{x}\mathbb{E}\,\boldsymbol{x}\boldsymbol{x}^3 \\
&= 4\sigma_x^2\mathbb{E}\,\boldsymbol{x}^3 \;+\; \bar{x}\mathbb{E}\,(\boldsymbol{x}-\bar{x}+\bar{x})\boldsymbol{x}^3 \\
&= 4\sigma_x^2\mathbb{E}\,\boldsymbol{x}^3 \;+\; \bar{x}\mathbb{E}\,(\boldsymbol{x}-\bar{x})\boldsymbol{x}^3 + \bar{x}^2\mathbb{E}\,\boldsymbol{x}^3 \\
&= (4\sigma_x^2+\bar{x}^2)\mathbb{E}\,\boldsymbol{x}^3 \;+\; 3\bar{x}\sigma_x^2\mathbb{E}\,\boldsymbol{x}^2 \\
&= (4\sigma_x^2+\bar{x}^2)\mathbb{E}\,(\boldsymbol{x}-\bar{x}+\bar{x})\boldsymbol{x}^2 \;+\; 3\bar{x}\sigma_x^2\mathbb{E}\,\boldsymbol{x}^2 \\
&= 2(4\sigma_x^2+\bar{x}^2)\sigma_x^2\mathbb{E}\,\boldsymbol{x} + (7\sigma_x^2+\bar{x}^2)\bar{x}\mathbb{E}\,\boldsymbol{x}^2 \\
&= 2(4\sigma_x^2+\bar{x}^2)\sigma_x^2\bar{x} + (7\sigma_x^2+\bar{x}^2)\bar{x}(\sigma_x^2+\bar{x}^2) \\
&= 15\bar{x}\sigma_x^4 + 10\bar{x}^3\sigma_x^2 + \bar{x}^5
\end{aligned} \tag{4.73}$$

4.4 JOINTLY-DISTRIBUTED GAUSSIAN VARIABLES

Consider two random vectors $\boldsymbol{x}$ of size $p \times 1$ and $\boldsymbol{y}$ of size $q \times 1$. We denote their respective means by $\{\bar{x}, \bar{y}\}$ and their respective covariance matrices by:

$$R_x \;\stackrel{\Delta}{=}\; \mathbb{E}\,(\boldsymbol{x}-\bar{x})(\boldsymbol{x}-\bar{x})^{\mathsf{T}} \tag{4.74a}$$

$$R_y \;\stackrel{\Delta}{=}\; \mathbb{E}\,(\boldsymbol{y}-\bar{y})(\boldsymbol{y}-\bar{y})^{\mathsf{T}} \tag{4.74b}$$

We further let R_{xy} denote the cross-covariance matrix between $\boldsymbol{x}$ and $\boldsymbol{y}$, i.e.,

$$R_{xy} \;\stackrel{\Delta}{=}\; \mathbb{E}\,(\boldsymbol{x}-\bar{x})(\boldsymbol{y}-\bar{y})^{\mathsf{T}} \;=\; R_{yx}^{\mathsf{T}} \tag{4.75}$$

and introduce the covariance matrix, R, of the aggregate vector $\text{col}\{\boldsymbol{x},\boldsymbol{y}\}$:

$$R \stackrel{\Delta}{=} \mathbb{E}\left(\begin{bmatrix}\boldsymbol{x}\\ \boldsymbol{y}\end{bmatrix}-\begin{bmatrix}\bar{x}\\ \bar{y}\end{bmatrix}\right)\left(\begin{bmatrix}\boldsymbol{x}\\ \boldsymbol{y}\end{bmatrix}-\begin{bmatrix}\bar{x}\\ \bar{y}\end{bmatrix}\right)^{\mathsf{T}} = \begin{bmatrix}R_x & R_{xy}\\ R_{xy}^{\mathsf{T}} & R_y\end{bmatrix} \tag{4.76}$$

We then say that the random variables $\{\boldsymbol{x},\boldsymbol{y}\}$ have a *joint* Gaussian distribution if their joint pdf has the form:

$$\begin{aligned} &f_{\boldsymbol{x},\boldsymbol{y}}(x,y) \qquad\qquad (4.77)\\ &= \frac{1}{\sqrt{(2\pi)^{p+q}}}\frac{1}{\sqrt{\det R}}\exp\left\{-\frac{1}{2}\begin{bmatrix}(x-\bar{x})^{\mathsf{T}} & (y-\bar{y})^{\mathsf{T}}\end{bmatrix}R^{-1}\begin{bmatrix}x-\bar{x}\\ y-\bar{y}\end{bmatrix}\right\}\end{aligned}$$

It can be seen that the joint pdf of $\{\boldsymbol{x},\boldsymbol{y}\}$ is completely determined by the mean, covariances, and cross-covariance of $\{\boldsymbol{x},\boldsymbol{y}\}$, i.e., by the first and second-order moments $\{\bar{x},\bar{y},R_x,R_y,R_{xy}\}$. It is also straightforward to conclude from (4.77) that uncorrelated Gaussian random vectors are independent — see Prob. 4.4. It takes more effort though to show that if $\{\boldsymbol{x},\boldsymbol{y}\}$ are jointly Gaussian-distributed as above, then each of the variables is individually Gaussian-distributed as well, namely, it holds:

$$\boldsymbol{x}\sim\mathcal{N}_{\boldsymbol{x}}(\bar{x},R_x),\qquad \boldsymbol{y}\sim\mathcal{N}_{\boldsymbol{y}}(\bar{y},R_y) \tag{4.78}$$

so that

$$f_{\boldsymbol{x}}(x) = \frac{1}{\sqrt{(2\pi)^p}}\,\frac{1}{\sqrt{\det R_x}}\;\exp\left\{-\frac{1}{2}(x-\bar{x})^{\mathsf{T}}R_x^{-1}(x-\bar{x})\right\} \tag{4.79a}$$

$$f_{\boldsymbol{y}}(y) = \frac{1}{\sqrt{(2\pi)^q}}\,\frac{1}{\sqrt{\det R_y}}\;\exp\left\{-\frac{1}{2}(y-\bar{y})^{\mathsf{T}}R_y^{-1}(y-\bar{y})\right\} \tag{4.79b}$$

LEMMA 4.3. (Marginal and conditional pdfs) *Consider two random vectors $\{\boldsymbol{x},\boldsymbol{y}\}$ that are jointly Gaussian distributed as in (4.77), namely,*

$$\begin{bmatrix}\boldsymbol{x}\\ \boldsymbol{y}\end{bmatrix}\sim\mathcal{N}_{\boldsymbol{x},\boldsymbol{y}}\left(\begin{bmatrix}\bar{x}\\ \bar{y}\end{bmatrix},\begin{bmatrix}R_x & R_{xy}\\ R_{xy}^{\mathsf{T}} & R_y\end{bmatrix}\right) \tag{4.80}$$

It follows that the individual marginal distributions are Gaussian, $\boldsymbol{x}\sim\mathcal{N}_{\boldsymbol{x}}(\bar{x},R_x)$ and $\boldsymbol{y}\sim\mathcal{N}_{\boldsymbol{y}}(\bar{y},R_y)$. Moreover, by marginalizing over $\boldsymbol{y}$ and $\boldsymbol{x}$ separately, the resulting conditional pdfs turn out to be Gaussian as well, as listed in Table 4.1.

Table 4.1 Conditional Gaussian distributions.

$f_{\boldsymbol{x}\vert\boldsymbol{y}}(x\vert y)\;\sim\;\mathcal{N}_{\boldsymbol{x}}(\widehat{x},\,\Sigma_x)$	$f_{\boldsymbol{y}\vert\boldsymbol{x}}(y\vert x)\;\sim\;\mathcal{N}_{\boldsymbol{y}}(\widehat{y},\,\Sigma_y)$
$\widehat{x}=\bar{x}+R_{xy}R_y^{-1}(y-\bar{y})$	$\widehat{y}=\bar{y}+R_{yx}R_x^{-1}(x-\bar{x})$
$\Sigma_x=R_x-R_{xy}R_y^{-1}R_{yx}$	$\Sigma_y=R_y-R_{yx}R_x^{-1}R_{xy}$

Proof: We start by noting that the block covariance matrix R in (4.76) can be factored into a product of three upper-triangular, diagonal, and lower-triangular matrices, as follows (this can be checked by straightforward algebra or see (1.63)):

$$R = \left[\begin{array}{cc} I_p & R_{xy}R_y^{-1} \\ 0 & I_q \end{array}\right] \left[\begin{array}{cc} \Sigma_x & 0 \\ 0 & R_y \end{array}\right] \left[\begin{array}{cc} I_p & 0 \\ R_y^{-1}R_{yx} & I_q \end{array}\right] \tag{4.81}$$

where we introduced the Schur complement $\Sigma_x = R_x - R_{xy}R_y^{-1}R_{yx}$. The matrix Σ_x is guaranteed to be positive-define in view of the assumed positive-definiteness of R itself — recall Example 1.5. It follows that the determinant of R factors into the product

$$\det R = \det \Sigma_x \; \times \; \det R_y \tag{4.82}$$

Inverting both sides of (4.81), we find that the inverse of R can be factored as

$$R^{-1} = \left[\begin{array}{cc} I_p & 0 \\ -R_y^{-1}R_{yx} & I_q \end{array}\right] \left[\begin{array}{cc} \Sigma_x^{-1} & 0 \\ 0 & R_y^{-1} \end{array}\right] \left[\begin{array}{cc} I_p & -R_{xy}R_y^{-1} \\ 0 & I_q \end{array}\right] \tag{4.83}$$

where we used the fact that for any matrix A of appropriate dimensions,

$$\left[\begin{array}{cc} I_p & 0 \\ A & I_q \end{array}\right]^{-1} = \left[\begin{array}{cc} I_p & 0 \\ -A & I_q \end{array}\right], \qquad \left[\begin{array}{cc} I_p & A \\ 0 & I_q \end{array}\right]^{-1} = \left[\begin{array}{cc} I_p & -A \\ 0 & I_q \end{array}\right] \tag{4.84}$$

Then, substituting into the exponent of the joint distribution (4.77) we can rewrite it in the equivalent form:

$$\begin{aligned} &\exp\left\{-\frac{1}{2}\left[\begin{array}{cc} (x-\bar{x})^{\mathsf{T}} & (y-\bar{y})^{\mathsf{T}} \end{array}\right] R^{-1} \left[\begin{array}{c} x-\bar{x} \\ y-\bar{y} \end{array}\right]\right\} \\ &= \exp\Big\{-\frac{1}{2}(x-\widehat{x})^{\mathsf{T}}\Sigma_x^{-1}(x-\widehat{x})\Big\} \; \exp\Big\{-\frac{1}{2}(y-\bar{y})^{\mathsf{T}}R_y^{-1}(y-\bar{y})\Big\} \end{aligned} \tag{4.85}$$

where we introduced $\widehat{x} = \bar{x} + R_{xy}R_y^{-1}(y-\bar{y})$. Substituting (4.85) into (4.77) and using (4.82) we find that the joint pdf of $\{\boldsymbol{x}, \boldsymbol{y}\}$ factorizes into the form

$$\begin{aligned} f_{\boldsymbol{x},\boldsymbol{y}}(x,y) = & \frac{1}{\sqrt{(2\pi)^p}} \frac{1}{\sqrt{\det \Sigma_x}} \exp\left\{-\frac{1}{2}(x-\widehat{x})^{\mathsf{T}}\Sigma_x^{-1}(x-\widehat{x})\right\} \times \\ & \frac{1}{\sqrt{(2\pi)^q}} \frac{1}{\sqrt{\det R_y}} \exp\left\{-\frac{1}{2}(y-\bar{y})^{\mathsf{T}}R_y^{-1}(y-\bar{y})\right\} \end{aligned} \tag{4.86}$$

We conclude from Bayes rule (3.39) that the marginal pdf of $\boldsymbol{y}$ is Gaussian with mean $\bar{y}$ and covariance matrix R_y:

$$\boldsymbol{y} \;\sim\; \mathcal{N}_{\boldsymbol{y}}(\bar{y}, R_y) \tag{4.87}$$

and, moreover, the conditional pdf $f_{\boldsymbol{x}|\boldsymbol{y}}(x|y)$ is Gaussian with mean $\widehat{x}$ and covariance matrix Σ_x:

$$f_{\boldsymbol{x}|\boldsymbol{y}}(x|y) \;\sim\; \mathcal{N}_{\boldsymbol{x}}(\widehat{x}, \Sigma_x) \tag{4.88a}$$

$$\widehat{x} = \bar{x} + R_{xy}R_y^{-1}(y-\bar{y}) \tag{4.88b}$$

$$\Sigma_x = R_x - R_{xy}R_y^{-1}R_{yx} \tag{4.88c}$$

If we repeat the same argument using instead the following alternative factorization for R (which can again be checked by straightforward algebra or using (1.63)):

$$R = \left[\begin{array}{cc} I_q & 0 \\ R_{yx}R_x^{-1} & I_p \end{array}\right] \left[\begin{array}{cc} R_x & 0 \\ 0 & \Sigma_y \end{array}\right] \left[\begin{array}{cc} I_q & R_x^{-1}R_{xy} \\ 0 & I_p \end{array}\right] \tag{4.89}$$

where $\Sigma_y = R_y - R_{yx}R_x^{-1}R_{xy}$, then we can similarly conclude that the marginal pdf of $\boldsymbol{x}$ is Gaussian with mean $\bar{x}$ and covariance matrix R_x:

$$\boldsymbol{x} \;\sim\; \mathcal{N}_{\boldsymbol{x}}(\bar{x}, R_x) \tag{4.90}$$

Moreover, the reverse conditional pdf $f_{\boldsymbol{y}|\boldsymbol{x}}(y|x)$ is also Gaussian with mean $\widehat{y}$ and covariance matrix Σ_y:

$$f_{\boldsymbol{y}|\boldsymbol{x}}(y|x) \;\sim\; \mathcal{N}_{\boldsymbol{y}}(\widehat{y},\, \Sigma_y) \tag{4.91a}$$

$$\widehat{y} = \bar{y} + R_{yx}R_x^{-1}(x - \bar{x}) \tag{4.91b}$$

$$\Sigma_y = R_y - R_{yx}R_x^{-1}R_{xy} \tag{4.91c}$$

■

Example 4.6 (**Joint pdf from marginal and conditional pdfs**) Consider two random variables $\boldsymbol{x}$ and $\boldsymbol{y}$ with marginal and conditional pdfs given by

$$f_{\boldsymbol{x}}(x) \sim \mathcal{N}_{\boldsymbol{x}}(\bar{x}, R_x), \quad f_{\boldsymbol{y}|\boldsymbol{x}}(x|y) \sim \mathcal{N}_{\boldsymbol{y}}(Fx, P) \tag{4.92}$$

for some matrices F and $P > 0$. The resulting joint pdf is Gaussian and given by

$$f_{\boldsymbol{x},\boldsymbol{y}}(x, y) \sim \mathcal{N}_{\boldsymbol{x},\boldsymbol{y}}\left(\begin{bmatrix} \bar{x} \\ F\bar{x} \end{bmatrix}, \begin{bmatrix} R_x & R_xF^{\mathsf{T}} \\ FR_x & FR_xF^{\mathsf{T}} + P \end{bmatrix} \right) \tag{4.93}$$

Proof: We rewrite the Gaussian distribution for $\boldsymbol{y}$ conditioned on $\boldsymbol{x}$ in the following form by adding and subtracting $F\bar{x}$ in the second line:

$$\begin{aligned} f_{\boldsymbol{y}|\boldsymbol{x}}(x|y) &\propto \exp\Big\{-\frac{1}{2}(y - Fx)^{\mathsf{T}}P^{-1}(y - Fx)\Big\} \\ &= \exp\Big\{-\frac{1}{2}\Big((y - F\bar{x}) - F(x - \bar{x})\Big)^{\mathsf{T}}P^{-1}\Big((y - F\bar{x}) - F(x - \bar{x})\Big)\Big\} \end{aligned} \tag{4.94}$$

From Bayes rule (3.39), the joint pdf is given by:

$$\begin{aligned} & f_{\boldsymbol{x},\boldsymbol{y}}(x, y) = f_{\boldsymbol{x}}(x)f_{\boldsymbol{y}|\boldsymbol{x}}(x|y) \\ & \propto \exp\Big\{-\frac{1}{2}(x - \bar{x})^{\mathsf{T}}R_x^{-1}(x - \bar{x})\Big\} \times \\ & \quad \exp\Big\{-\frac{1}{2}\Big((y - F\bar{x}) - F(x - \bar{x})\Big)^{\mathsf{T}}P^{-1}\Big((y - F\bar{x}) - F(x - \bar{x})\Big)\Big\} \\ & = \exp\left\{ -\frac{1}{2}\begin{bmatrix} (x - \bar{x})^{\mathsf{T}} & (y - F\bar{x})^{\mathsf{T}} \end{bmatrix} \begin{bmatrix} R_x^{-1} + F^{\mathsf{T}}P^{-1}F & -F^{\mathsf{T}}P^{-1} \\ -P^{-1}F & P^{-1} \end{bmatrix} \begin{bmatrix} x - \bar{x} \\ y - F\bar{x} \end{bmatrix} \right\} \\ & \overset{(1.67)}{=} \exp\left\{ -\frac{1}{2}\begin{bmatrix} (x - \bar{x})^{\mathsf{T}} & (y - F\bar{x})^{\mathsf{T}} \end{bmatrix} \begin{bmatrix} R_x & R_xF^{\mathsf{T}} \\ FR_x & FR_xF^{\mathsf{T}} + P \end{bmatrix}^{-1} \begin{bmatrix} x - \bar{x} \\ y - F\bar{x} \end{bmatrix} \right\} \end{aligned} \tag{4.95}$$

from which we conclude that (4.93) holds.

■

Example 4.7 (**Useful integral expressions**) We generalize Example 4.1 to vector Gaussian random variables. Let $f_{\boldsymbol{x}}(x)$ denote the standard $N-$dimensional Gaussian distribution, $\boldsymbol{x} \sim \mathcal{N}_{\boldsymbol{x}}(0, I_N)$, and introduce its cumulative distribution function (CDF):

$$\Phi(z) \;\triangleq\; \int_{-\infty}^{z} \frac{1}{\sqrt{(2\pi)^N}} \exp\Big\{-\frac{1}{2}\|x\|^2\Big\}dx \;=\; \int_{-\infty}^{z} \mathcal{N}_{\boldsymbol{x}}(0, I_N)dx \tag{4.96}$$

which is now a multi-dimensional integral since x is a vector. Let $\boldsymbol{y} \sim \mathcal{N}_{\boldsymbol{y}}(\bar{y}, R_y)$ denote an $M-$dimensional Gaussian distribution. We wish to evaluate an integral expression of the following form involving the product of a Gaussian distribution and the cumulative distribution:

$$\begin{aligned} Z_0 &\overset{\Delta}{=} \int_{-\infty}^{\infty} \Phi(Ay+b)\,\mathcal{N}_{\boldsymbol{y}}(\bar{y}, R_y)dy, \quad \text{for some given } A \in \mathbb{R}^{N\times M},\, b \in \mathbb{R}^N \\ &= \frac{1}{\sqrt{(2\pi)^{M+N}}}\frac{1}{\sqrt{\det R_y}} \times \\ &\quad \int_{-\infty}^{\infty}\int_{-\infty}^{Ay+b} \exp\Big\{-\frac{1}{2}\|x\|^2\Big\} \exp\Big\{-\frac{1}{2}(y-\bar{y})R_y^{-1}(y-\bar{y})\Big\}dxdy \end{aligned} \tag{4.97}$$

The evaluation takes some effort. We start with the change of variables

$$w \overset{\Delta}{=} y - \bar{y} \in \mathbb{R}^M, \quad z \overset{\Delta}{=} x - Aw \in \mathbb{R}^N \tag{4.98}$$

and replace the integration over x and y by an integration over z and w:

$$\begin{aligned} Z_0 = &\frac{1}{\sqrt{(2\pi)^{M+N}}}\frac{1}{\sqrt{\det R_y}} \times \\ &\int_{-\infty}^{\infty}\int_{-\infty}^{A\bar{y}+b} \exp\Big\{-\frac{1}{2}(z+Aw)^{\mathsf{T}}(z+Aw)\Big\} \exp\Big\{-\frac{1}{2}w^{\mathsf{T}}R_y^{-1}w\Big\}dzdw \end{aligned} \tag{4.99}$$

The advantage of the change of variables is that the limit of the inner integral, $A\bar{y}+b$, is now independent of the variables $\{z, w\}$ over which the integration is performed. The exponent in (4.99) is quadratic in $\{z, w\}$ since

$$(z+Aw)^{\mathsf{T}}(z+Aw) + w^{\mathsf{T}}R_y^{-1}w = \begin{bmatrix} z^{\mathsf{T}} & w^{\mathsf{T}} \end{bmatrix} \begin{bmatrix} I_N & A \\ A^{\mathsf{T}} & R_y^{-1} + A^{\mathsf{T}}A \end{bmatrix} \begin{bmatrix} z \\ w \end{bmatrix} \tag{4.100}$$

This means that the integrand in (4.99) can be written as a joint Gaussian distribution over the extended variable col$\{z, w\}$ with zero mean and covariance matrix

$$R \overset{\Delta}{=} \begin{bmatrix} I_N & A \\ A^{\mathsf{T}} & R_y^{-1} + A^{\mathsf{T}}A \end{bmatrix}^{-1} \overset{(1.67)}{=} \begin{bmatrix} I_N + AR_yA^{\mathsf{T}} & -AR_y \\ -R_yA^{\mathsf{T}} & R_y \end{bmatrix} \tag{4.101}$$

This shows that the covariance matrix of $\boldsymbol{z}$ is

$$R_z \overset{\Delta}{=} I_N + AR_yA^{\mathsf{T}} \tag{4.102}$$

Note further that $\det R = \det R_y$ since the Schur complement of R relative to R_y is the identity matrix. It follows that

$$Z_0 = \int_{-\infty}^{A\bar{y}+b} \underbrace{\left(\int_{-\infty}^{\infty} \mathcal{N}_{\boldsymbol{z},\boldsymbol{w}}(0, R)dw\right)}_{\textbf{marginalization}} dz \tag{4.103}$$

The inner integral amounts to marginalizing the joint distribution of $\{\boldsymbol{z}, \boldsymbol{w}\}$ over w so that the result is the marginal distribution for $\boldsymbol{z}$, namely, $f_{\boldsymbol{z}}(z) = \mathcal{N}_{\boldsymbol{z}}(0, R_z)$ and, consequently,

$$\begin{aligned} Z_0 &= \int_{-\infty}^{A\bar{y}+b} \mathcal{N}_{\boldsymbol{z}}(0, R_z)dz \\ &= \frac{1}{\sqrt{(2\pi)^N}}\frac{1}{\sqrt{\det R_z}} \int_{-\infty}^{A\bar{y}+b} \exp\Big\{-\frac{1}{2}z^{\mathsf{T}}R_z^{-1}z\Big\}dz \end{aligned} \tag{4.104}$$

This expression almost has the form of a cumulative distribution calculation except that the Gaussian distribution is not standard (it has covariance matrix R_z rather than the identity matrix). Let X denote a square-root factor for R_z (recall the definition in Sec. 1.8), namely, X is any invertible square matrix that satisfies $R_z = XX^{\mathsf{T}}$. We can also use the more explicit notation $R_z^{1/2}$ to refer to X. One choice for X arises from the eigendecomposition $R_z = U\Lambda U^{\mathsf{T}}$, where U is orthogonal and Λ is diagonal with positive entries. Using U and Λ, we can select $X = U\Lambda^{1/2}$. Note that

$$R_z = XX^{\mathsf{T}} \implies \det R_z = (\det X)^2 \tag{4.105}$$

Next, we introduce the change of variables

$$s = X^{-1}z \implies z^{\mathsf{T}}R_z^{-1}z = s^{\mathsf{T}}s \tag{4.106}$$

and the $N \times N$ Jacobian matrix J whose entries consist of the partial derivatives:

$$\big[J\big]_{m,n} = \frac{\partial z_m}{\partial s_n} = X \tag{4.107}$$

where z_m is the $m-$th entry of z and s_n is the $n-$th entry of s. We know from the study of multi-dimensional integrals that when a change of variables is used, we need to account for the (absolute value of the) determinant of the Jacobian matrix so that expression (4.104) is replaced by

$$\begin{aligned} Z_0 &= \frac{1}{\sqrt{(2\pi)^N}}\frac{1}{\sqrt{\det R_z}}\int_{-\infty}^{X^{-1}(A\bar{y}+b)} \exp\Big\{-\frac{1}{2}\|s\|^2\Big\}\,|\det X|\,ds \\ &= \underbrace{\frac{|\det X|}{\sqrt{\det R_z}}}_{=1}\int_{-\infty}^{X^{-1}(A\bar{y}+b)} \frac{1}{\sqrt{(2\pi)^N}}\exp\Big\{-\frac{1}{2}\|s\|^2\Big\}ds \\ &\overset{(4.105)}{=} \Phi\Big(X^{-1}(A\bar{y}+b)\Big) \end{aligned} \tag{4.108}$$

In summary, we arrive at the result:

$$\boxed{Z_0 = \int_{-\infty}^{\infty}\Phi(Ay+b)\,\mathcal{N}_{\boldsymbol{y}}(\bar{y},R_y)dy \;=\; \Phi(\widehat{y}\,)} \tag{4.109}$$

where

$$\widehat{y} \;\stackrel{\Delta}{=}\; R_z^{-1/2}(A\bar{y}+b) \tag{4.110}$$

It is easy to see that the above result reduces to (4.7) with the identifications $A \leftarrow 1/\sigma_a$, $b \leftarrow -a/\sigma_a$, and $R_z \leftarrow 1+\sigma_y^2/\sigma_a^2$. Another useful special case is when A is a row vector and b is a scalar, say, $A = h^{\mathsf{T}}$ and $b = \alpha$, in which case we get

$$\boxed{Z_0 = \int_{-\infty}^{\infty}\Phi(h^{\mathsf{T}}y+\alpha)\,\mathcal{N}_{\boldsymbol{y}}(\bar{y},R_y)dy \;=\; \Phi(\widehat{y}\,)} \tag{4.111}$$

where now

$$\widehat{y} \;\stackrel{\Delta}{=}\; \frac{h^{\mathsf{T}}\bar{y}+\alpha}{\sqrt{1+h^{\mathsf{T}}R_y h}} \tag{4.112}$$

If we differentiate (4.111) relative to $\bar{y}$ once and then twice, we arrive at two additional relations where the integrands are further multiplied by y and yy^{T} — see Prob. 4.34:

$$Z_1 = \int_{-\infty}^{\infty} y\Phi(h^{\mathsf{T}}y+\alpha)\,\mathcal{N}_{\boldsymbol{y}}(\bar{y},R_y)dy \;=\; \bar{y}\Phi(\widehat{y}\,) + \frac{R_y h}{\sqrt{1+h^{\mathsf{T}}R_y h}}\mathcal{N}_{\widehat{\boldsymbol{y}}}(0,1) \tag{4.113}$$

and

$$\begin{aligned} Z_2 &\overset{\Delta}{=} \int_{-\infty}^{\infty} yy^{\mathsf{T}}\Phi(h^{\mathsf{T}}y+\alpha)\,\mathcal{N}_y(\bar{y},R_y)dy && (4.114) \\ &= (R_y+\bar{y}\bar{y}^{\mathsf{T}})\Phi(\widehat{y}) + \frac{1}{\sqrt{1+h^{\mathsf{T}}R_yh}}\left(2R_yh\bar{y}^{\mathsf{T}} - \frac{\widehat{y}R_yhh^{\mathsf{T}}R_y}{\sqrt{1+h^{\mathsf{T}}R_yh}}\right)\mathcal{N}_{\widehat{y}}(0,1) \end{aligned}$$

4.5 GAUSSIAN PROCESSES

If we consider a $p-$dimensional Gaussian random vector $\boldsymbol{x} \sim \mathcal{N}_{\boldsymbol{x}}(\bar{x}, R_x)$ with entries $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_p\}$, then we know from Lemma 4.3 that any sub-collection of entries of $\boldsymbol{x}$ will be jointly Gaussian as well. In other words, the Gaussianity property is inherited by any sub-grouping within $\boldsymbol{x}$. For example, $\{\boldsymbol{x}_1, \boldsymbol{x}_2\}$ will be jointly Gaussian with mean vector col$\{\bar{x}_1, \bar{x}_2\}$ formed from the two top entries of $\bar{x}$ and with covariance matrix formed from the leading 2×2 submatrix of R_x.

The notion of Gaussian processes (GPs) allows us to extend this property to *sequences* of vectors. The concept will be useful when we study learning and inference problems in the kernel domain later in this text. For now, we define a *Gaussian process* as a sequence of random vectors where any finite sub-collection of entries in each vector is jointly Gaussian-distributed. Moreover, entries across vectors can be correlated with each other.

Remark 4.1. (Terminology) We will discuss "random processes" in Chapter 7. We explain there that random processes consist of *sequences* of random variables (or vectors). We will denote the random process by the notation $\boldsymbol{x}_n$, with a time (or space) index n added. This notation means that a realization for $\boldsymbol{x}$ is selected at each instant n, and the index n evolves sequentially from lower values to higher values. Moreover, there can be correlation between different samples $\boldsymbol{x}_n$ and $\boldsymbol{x}_m$. For the Gaussian processes discussed in this section, the index n need not be time or space (e.g., it can refer to something more abstract, such as repeated experiments involving separate data collections as the next example and the discussion following it illustrate).

■

Example 4.8 (**Nonlinear transformations**) We reconsider the linear regression model from Example 4.3, namely,

$$\boldsymbol{y}(n) = x_n^{\mathsf{T}}\boldsymbol{w} + \boldsymbol{v}(n), \quad \boldsymbol{v}(n) \sim \mathcal{N}_{\boldsymbol{v}}(0, \sigma_v^2), \quad n = 1, 2, \ldots, N \tag{4.115}$$

where $x_n, \boldsymbol{w} \in \mathbb{R}^M$. This expression models the observation $\boldsymbol{y}(n)$ as a noisy measurement of a linear combination of the individual entries $\{x_{n,m}\}$ of x_n, written explicitly as

$$\boldsymbol{y}(n) = \sum_{m=1}^{M} x_{n,m}^{\mathsf{T}}\boldsymbol{w}_m + \boldsymbol{v}(n) \tag{4.116}$$

In many situations, it will be advantageous to consider more elaborate models for the mapping from the input x_n to the output $\boldsymbol{y}(n)$. One possibility is to replace the

$M \times 1$ vector x_n by a longer $M_\phi \times 1$ vector $\phi(x_n)$, where the notation $\phi(\cdot)$ represents some nonlinear transformation applied to the entries of x_n. For example, if x_n is two-dimensional with individual entries $x_n = [a\ b]$, then one possibility is to use $\phi(x_n) = [a\ b\ a^2\ b^2\ ab]$. For this case, $M = 2$ and $M_\phi = 5$. The $M-$dimensional weight vector $\boldsymbol{w}$ would also be extended and replaced by $\boldsymbol{w}^\phi \in \mathbb{R}^{M_\phi}$, in which case the original model would be:

$$\boldsymbol{y}(n) = (\phi(x_n))^\mathsf{T}\boldsymbol{w}^\phi + \boldsymbol{v}(n) \tag{4.117}$$

This representation captures more nonlinear dynamics from x_n to $\boldsymbol{y}(n)$. Many other choices for $\phi(\cdot)$ are of course possible. We rewrite the measurement equation in the equivalent form:

$$\boldsymbol{y}(n) = \boldsymbol{g}(x_n) + \boldsymbol{v}(n), \quad \boldsymbol{g}(x_n) \;\stackrel{\Delta}{=}\; (\phi(x_n))^\mathsf{T}\boldsymbol{w}^\phi \tag{4.118}$$

where the scalar-valued function $\boldsymbol{g}(x)$ now depends on x_n through the transformation $\phi(\cdot)$; it assumes random values because we model $\boldsymbol{w}^\phi$ as Gaussian-distributed:

$$\boldsymbol{w}^\phi \sim \mathcal{N}_{\boldsymbol{w}^\phi}(\bar{w}^\phi, R_w^\phi) \tag{4.119}$$

for some mean vector $\bar{w}^\phi$ and covariance matrix R_w^ϕ. We collect the measurements $\{\boldsymbol{y}(n)\}$, the input vectors $\{x_n\}$, and the values of $\boldsymbol{g}(\cdot)$ and $\boldsymbol{v}(\cdot)$ into matrix and vector quantities and write:

$$\Phi \;\stackrel{\Delta}{=}\; \begin{bmatrix} (\phi(x_1))^\mathsf{T} \\ (\phi(x_2))^\mathsf{T} \\ \vdots \\ (\phi(x_N))^\mathsf{T} \end{bmatrix} \tag{4.120a}$$

$$\boldsymbol{g} \;\stackrel{\Delta}{=}\; \Phi\boldsymbol{w}^\phi, \quad \Phi \in \mathbb{R}^{N\times M_\phi} \tag{4.120b}$$

$$\underbrace{\begin{bmatrix} \boldsymbol{y}(1) \\ \boldsymbol{y}(2) \\ \vdots \\ \boldsymbol{y}(N) \end{bmatrix}}_{\boldsymbol{y}} = \underbrace{\begin{bmatrix} \boldsymbol{g}(x_1) \\ \boldsymbol{g}(x_2) \\ \vdots \\ \boldsymbol{g}(x_N) \end{bmatrix}}_{\boldsymbol{g}} + \underbrace{\begin{bmatrix} \boldsymbol{v}(1) \\ \boldsymbol{v}(2) \\ \vdots \\ \boldsymbol{v}(N) \end{bmatrix}}_{\boldsymbol{v}} \tag{4.120c}$$

$$\boldsymbol{y} \;=\; \Phi\boldsymbol{w}^\phi + \boldsymbol{v} \tag{4.120d}$$

where $\boldsymbol{y}$ continues to be Gaussian-distributed with mean and covariance matrix given by

$$\boldsymbol{y} \sim \mathcal{N}_{\boldsymbol{y}}\Big(\Phi\bar{w}^\phi,\ \sigma_v^2 I_N + \Phi R_w^\phi \Phi^\mathsf{T}\Big) \tag{4.121}$$

Moreover, the factor $\boldsymbol{g}$ is Gaussian-distributed with mean and covariance matrix given by

$$\bar{g} = \Phi\bar{w}^\phi, \qquad R_g \;=\; \Phi R_w^\phi \Phi^\mathsf{T} \tag{4.122a}$$

where each entry of R_g corresponds to the cross-covariance:

$$[R_g]_{m,n} = (\phi(x_n))^\mathsf{T} R_w^\phi\ \phi(x_m) \tag{4.122b}$$

The vector $\boldsymbol{g}$ is an example of a Gaussian process: any sub-collection of entries in $\boldsymbol{g}$ follows a joint Gaussian distribution. The same is true for the vector $\boldsymbol{g}$ defined earlier in (4.22d). However, the addition of the nonlinear mapping $\phi(\cdot)$ to the model enriches the scenario under consideration, as we proceed to explain.

Model (4.120d) involves a finite number of observations in $\boldsymbol{y}$; this observation vector is a perturbed version of the Gaussian process $\boldsymbol{g}$. The mean and covariance matrix of $\boldsymbol{g}$ are described by (4.122a); they both depend on Φ, which in turn is defined in terms of the input vectors $\{x_n\}$. It would appear at first sight that we are dealing with a Gaussian vector with a finite number of elements in it. However, on closer examination, $\boldsymbol{g}(\cdot)$ is a Gaussian process. This is because, in general, we would not know beforehand which input vectors $\{x_n\}$ to expect. For instance, in a second experiment, the observation vector $\boldsymbol{y}$ will be determined by some other collection of input vectors $\{x'_n\}$. In that case, the mean and covariance matrix of this new observation $\boldsymbol{y}'$ would not be given by (4.121) because the matrix Φ will now be different and defined in terms of the $\{x'_n\}$ rather than $\{x_n\}$. Nevertheless, we would still be able to identify the mean and covariance matrix of the new observation vector $\boldsymbol{y}'$ if we define the mean and covariance matrix of the Gaussian process $\boldsymbol{g}$ more broadly, for any possible choice of its arguments $\{x_n\}$. To do so, we proceed as follows.

We let $\boldsymbol{g}(x)$ denote any generic entry of the Gaussian process. Observe that the argument is the $M-$dimensional vector x; it can assume any value in $\mathbb{R}^M$. We associate with the process $\boldsymbol{g}(\cdot)$ a mean function and a covariance function defined as follows:

$$m(x) \overset{\Delta}{=} \mathbb{E}\,\boldsymbol{g}(x) \tag{4.123a}$$

$$K(x,x') \overset{\Delta}{=} \mathbb{E}\left(\boldsymbol{g}(x)-m(x)\right)\left(\boldsymbol{g}(x')-m(x')\right) \tag{4.123b}$$

Using these functions, we can evaluate the mean of $\boldsymbol{g}(x)$ for *any* x, and the cross-covariance between $\boldsymbol{g}(x)$ and $\boldsymbol{g}(x')$ for *any* x, x'. The expectations are over the sources of randomness in $\boldsymbol{g}(x)$. For example, for the case studied above we have

$$m(x) = (\phi(x))^{\mathsf{T}}\bar{w}^{\phi}, \qquad K(x,x') = (\phi(x))^{\mathsf{T}}R_w^{\phi}\,\phi(x') \tag{4.124}$$

so that the mean and covariance matrix that correspond to the particular input vectors $\{x_n\}$ would be constructed as follows (say, for $N=4$ measurements):

$$\bar{g} = \begin{bmatrix} m(x_1)\\ m(x_2)\\ m(x_3)\\ m(x_4)\end{bmatrix}, \quad R_g = \begin{bmatrix} K(x_1,x_1) & K(x_1,x_2) & K(x_1,x_3) & K(x_1,x_4)\\ K(x_2,x_1) & K(x_2,x_2) & K(x_2,x_3) & K(x_2,x_4)\\ K(x_3,x_1) & K(x_3,x_2) & K(x_3,x_3) & K(x_3,x_4)\\ K(x_4,x_1) & K(x_4,x_2) & K(x_4,x_3) & K(x_4,x_4)\end{bmatrix} \tag{4.125}$$

We will denote a Gaussian process by the notation

$$\boldsymbol{g} \sim \mathcal{GP}_{\boldsymbol{g}}\Big(m(x),\, K(x,x')\Big) \tag{4.126}$$

It is important to note that not any function $K(x,x')$ can be selected as the covariance function for a Gaussian process. This is because when $K(x,x')$ is applied to data to construct matrices like R_g above, these matrices will need to behave like covariance matrices (i.e., they will need to be symmetric and nonnegative definite). We will examine conditions on $K(x,x')$ in greater detail later in Chapter 63 when we study kernel methods. Here, we summarize the main requirement and leave the details to that chapter.

The function $K(x, x')$ will generally be selected to be a kernel, which is a function that maps two vector arguments (x, x') into the *inner-product* of similarly-transformed versions of these same vectors, namely, a function that can be written in the form:

$$K(x, x') = (\phi(x))^{\mathsf{T}}\phi(x') \tag{4.127}$$

for some mapping $\phi(\cdot)$. Note that the covariance function used in (4.124) is of this form. It is written in terms of a weighted inner product but can be easily recast in the standard form (4.127). For instance, assume for simplicity that R_w^{ϕ} is positive-definite and introduce the eigendecomposition $R_w^{\phi} = U\Lambda U^{\mathsf{T}}$ (where U is orthogonal and Λ is diagonal with positive entries). Then, we can redefine

$$\phi(x) \leftarrow \Lambda^{1/2}U^{\mathsf{T}}\phi(x) \tag{4.128}$$

and $K(x, x')$ would reduce to the form (4.127). In the above notation, $\Lambda^{1/2}$ is a diagonal matrix with the positive square roots of the entries of Λ.

Note from definition (4.127) that kernels are symmetric functions since

$$K(x, x') = K(x', x) \tag{4.129}$$

We say that kernel functions induce inner-product operations in the transformed domain. Obviously, not every function, $K(x, x')$, can be expressed in the inner-product form (4.127) and, therefore, not every function is a kernel. A fundamental theorem in functional analysis, known as *Mercer theorem*, clarifies which functions $K(x, x')$ can be expressed in the form (4.127) — see future Sec. 63.2. For any integer N, we introduce the following $N \times N$ Gramian matrix, R_N, which is symmetric:

$$[R_N]_{m,n} \;\stackrel{\Delta}{=}\; K(x_m, x_n), \quad m, n = 1, 2, \ldots, N \tag{4.130}$$

(Mercer theorem). *The theorem affirms that a symmetric and square-integrable function $K(x, x')$ is a kernel if, and only if, the Gramian matrix R_N defined by (4.130) is positive semidefinite for any size N and any data $\{x_n\}$.*

There are many popular choices for $K(x, x')$ that satisfy Mercer condition. One choice is the Gaussian kernel (also called the *radial basis function* or the squared exponential kernel):

$$K(x, x') \;\stackrel{\Delta}{=}\; \exp\left\{-\frac{1}{2\sigma^2}\|x - x'\|^2\right\} \tag{4.131}$$

for some parameter $\sigma^2 > 0$. This parameter controls the width of the Gaussian pulse. One could also replace the exponent by a weighted squared norm such as

$$K(x, x') \;\stackrel{\Delta}{=}\; \exp\left\{-\frac{1}{2}(x - x')^{\mathsf{T}}W(x - x')\right\} \tag{4.132}$$

with different choices for the positive-definite matrix W, such as

$$W = \frac{1}{\sigma^2} I_N \quad \text{or} \quad W = \text{a diagonal matrix} \tag{4.133}$$

Another kernel choice is the Ornstein-Uhlenbeck kernel:

$$K(x, x') \overset{\Delta}{=} \exp\left\{-\frac{1}{\sigma}\|x - x'\|\right\}, \quad \sigma > 0 \tag{4.134}$$

defined in terms of the distance between x and x' rather than their squared distance. The parameter σ is referred to as the *length-scale* of the process; it determines how close points x and x' will need to be to each other for a meaningful correlation between them. We can interpret these kernel functions as providing measures of "similarity" between points in space.

We will explain later in Example 63.4 that the Gaussian kernel (4.131) can be written in the inner-product form (4.127) for some function $\phi(\cdot)$; similarly, for the Ornstein-Uhlenbeck kernel. Fortunately, explicit knowledge of $\phi(\cdot)$ is unnecessary (and this important fact is what makes kernel methods powerful; as explained later in Chapter 63). Observe that the kernel functions in (4.131)–(4.134) are written directly in terms of the input vectors (x, x') and not their transformed versions $\phi(x)$ or $\phi(x')$. Usually, the mean function of a Gaussian process is taken to be zero. In this way, characterization of the first and second-order moments of $\boldsymbol{g} \sim \mathcal{GP}_{\boldsymbol{g}}(0, K(x, x'))$ would not require knowledge of the nonlinear mapping $\phi(\cdot)$. Once a kernel function is specified, we are implicitly assuming some nonlinear mapping is applied to the input vectors $\{x_n\}$.

Example 4.9 **(Polynomial kernel)** Let us illustrate the last point by considering a simplified example. Assume $x \in \mathbb{R}^2$ with entries $x = [a\ b]$. We select the polynomial kernel

$$K(x, x') \overset{\Delta}{=} (1 + x^{\mathsf{T}} x')^2 \tag{4.135}$$

and verify that it is a kernel function. To do so, and according to definition (4.127), we need to identify a transformation $\phi(x)$ that allows us to express $K(x, x')$ as the inner product $(\phi(x))^{\mathsf{T}} \phi(x')$. Indeed, note that

$$\begin{aligned} K(x, x') &= (1 + aa' + bb')^2 \\ &= (1 + aa')^2 \; + \; b^2 b'^2 \; + \; 2(1 + aa')bb' \\ &= 1 + a^2 a'^2 \; + \; 2aa' \; + \; b^2 b'^2 \; + \; 2bb' \; + \; 2aa'bb' \end{aligned} \tag{4.136}$$

which we can express more compactly as follows. We introduce the transformed vector:

$$\phi(x) = \text{col}\left\{1,\, \sqrt{2}a,\, \sqrt{2}b,\, \sqrt{2}ab,\, a^2,\, b^2\right\} \tag{4.137}$$

and note from (4.136) that $K(x, x') = (\phi(x))^{\mathsf{T}} \phi(x')$. In other words, the function (4.135) can be expressed as an inner product between the two transformed vectors $(\phi(x), \phi(x'))$, both of dimension 6×1. Observe further the important fact that the vectors, x and x', have both been transformed in an identical manner.

4.6 CIRCULAR GAUSSIAN DISTRIBUTION[1]

The Gaussian distribution can be extended to complex variables as well. Thus, consider a complex random vector $\boldsymbol{z} = \boldsymbol{x} + j\boldsymbol{y} \in \mathbb{C}^p$. We say $\boldsymbol{z}$ is Gaussian-distributed if its real and imaginary parts are jointly Gaussian (cf. (4.77)), namely, their joint pdf is of the form:

$$f_{\boldsymbol{x},\boldsymbol{y}}(x,y) = \frac{1}{(2\pi)^p}\frac{1}{\sqrt{\det R}}\exp\left\{-\frac{1}{2}\left[\begin{array}{cc}(x-\bar{x})^{\mathsf{T}} & (y-\bar{y})^{\mathsf{T}}\end{array}\right]R^{-1}\left[\begin{array}{c}x-\bar{x}\\ y-\bar{y}\end{array}\right]\right\} \tag{4.138}$$

The mean of $\boldsymbol{z}$ is clearly

$$\bar{z} = \mathbb{E}\,\boldsymbol{z} = \bar{x} + j\bar{y} \tag{4.139}$$

while its covariance matrix is

$$R_z \stackrel{\Delta}{=} \mathbb{E}\,(\boldsymbol{z}-\bar{z})(\boldsymbol{z}-\bar{z})^* \;=\; (R_x + R_y) + j(R_{yx} - R_{xy}) \tag{4.140}$$

which is expressed in terms of *both* the covariances and cross-covariance of $\{\boldsymbol{x},\boldsymbol{y}\}$. Note that the variable $\boldsymbol{z}$ can be regarded as a function of the two variables $\{\boldsymbol{x},\boldsymbol{y}\}$ and, therefore, its probabilistic nature is fully characterized by the joint pdf of $\{\boldsymbol{x},\boldsymbol{y}\}$. This joint pdf is defined in terms of the first and second-order moments of $\{\boldsymbol{x},\boldsymbol{y}\}$, i.e., in terms of $\{\bar{x},\bar{y},R_x,R_y,R_{xy}\}$.

It is useful to verify whether it is possible to express the pdf of $\boldsymbol{z}$ *directly* in terms of its own first and second-order moments, i.e., in terms of $\{\bar{z},R_z\}$. It turns out that this is *not* always possible. This is because knowledge of $\{\bar{z},R_z\}$ alone is not enough to recover the moments $\{\bar{x},\bar{y},R_x,R_y,R_{xy}\}$. More information is needed in the form of a *circularity* condition. To see this, assume we only know $\{\bar{z},R_z\}$. Then, this information is enough to recover $\{\bar{x},\bar{y}\}$ since $\bar{z} = \bar{x} + j\bar{y}$. However, the information is not enough to recover $\{R_x,R_y,R_{xy}\}$. This is because, as we see from (4.140), knowledge of R_z allows us to recover $(R_x + R_y)$ and $(R_{yx} - R_{xy})$ via

$$R_x + R_y \;=\; \mathrm{Re}(R_z), \qquad R_{yx} - R_{xy} \;=\; \mathrm{Im}(R_z) \tag{4.141}$$

in terms of the real and imaginary parts of R_z. This information is not sufficient to determine the individual covariances $\{R_x,R_y,R_{xy}\}$.

In order to be able to uniquely recover $\{R_x,R_y,R_{xy}\}$ from R_z, it will be further assumed that the random variable $\boldsymbol{z}$ satisfies a *circularity* condition, namely, that

$$\boxed{\mathbb{E}\,(\boldsymbol{z}-\bar{z})(\boldsymbol{z}-\bar{z})^{\mathsf{T}} \;=\; 0} \qquad (\textbf{circularity condition}) \tag{4.142}$$

with the transposition symbol T used instead of Hermitian conjugation. Knowledge of R_z, along with circularity, are enough to recover $\{R_x,R_y,R_{xy}\}$ from R_z.

[1] This section can be skipped on a first reading.

Indeed, using the fact that

$$\mathbb{E}\,(\boldsymbol{z}-\bar{z})(\boldsymbol{z}-\bar{z})^{\mathsf{T}} \;=\; (R_x - R_y) + j(R_{yx} + R_{xy}) \tag{4.143}$$

we find that, in view of the circularity assumption (4.142), it must now hold that $R_x = R_y$ and $R_{xy} = -R_{yx}$. Consequently, combining with (4.141), we can solve for $\{R_x, R_y, R_{xy}\}$ to get

$$R_x \;=\; R_y \;=\; \frac{1}{2}\,\text{Re}(R_z) \qquad \text{and} \qquad R_{xy} = -R_{yx} = -\frac{1}{2}\,\text{Im}(R_z) \tag{4.144}$$

It follows that the covariance matrix of col$\{\boldsymbol{x},\boldsymbol{y}\}$ can be expressed in terms of R_z as

$$R = \frac{1}{2}\begin{bmatrix} \text{Re}(R_z) & -\text{Im}(R_z) \\ \text{Im}(R_z) & \text{Re}(R_z) \end{bmatrix} \tag{4.145}$$

Actually, it also follows that R should have the symmetry structure:

$$R \;=\; \begin{bmatrix} R_x & R_{xy} \\ -R_{xy} & R_x \end{bmatrix} \tag{4.146}$$

with the same matrix R_x appearing on the diagonal, and with R_{xy} and its negative appearing at the off-diagonal locations. Observe further that when $\boldsymbol{z}$ happens to be scalar-valued, then R_{xy} becomes a scalar, say, σ_{xy}, and the condition $R_{xy} = -R_{yx}$ can only hold if $\sigma_{xy} = 0$. That is, the real and imaginary parts of $\boldsymbol{z}$ will need to be independent in the scalar case.

Using result (4.146), we can now verify that the joint pdf of $\{\boldsymbol{x},\boldsymbol{y}\}$ in (4.138) can be rewritten in terms of $\{\bar{z}, R_z\}$ as shown below — compare with (4.79a) in the real case. Observe in particular that the factors of 2, as well as the square-roots, disappear from the pdf expression in the complex case:

$$\boxed{f_{\boldsymbol{z}}(z) = \frac{1}{\pi^p}\frac{1}{\det R_z}\;\exp\Big\{-(z-\bar{z})^* R_z^{-1}(z-\bar{z})\Big\}} \tag{4.147}$$

We say that $\boldsymbol{z} \in \mathbb{C}^p$ is a circular or spherically-invariant Gaussian random variable. When (4.147) holds, we can check that uncorrelated jointly Gaussian random variables will also be independent; this is one of the main reasons for the assumption of circularity — see Prob. 4.22.

Proof of (4.147): Using (4.146) and the determinantal formula (1.64a), we have

$$\det R = \det(R_x)\,\det(R_x + R_{xy}R_x^{-1}R_{xy}) \tag{4.148}$$

Likewise, using the expression $R_z = 2(R_x - jR_{xy})$, we obtain

$$\begin{aligned}(\det R_z)^2 &= \det(R_z)\,\det(R_z^{\mathsf{T}}) \\ &= 2^{2p}\det(R_x(I - jR_x^{-1}R_{xy}))\,\det(R_x - jR_{xy}^{\mathsf{T}})\end{aligned} \tag{4.149}$$

Noting that

$$R_{xy}^{\mathsf{T}} = R_{yx} \overset{(4.144)}{=} -R_{xy} \tag{4.150}$$

and, for matrices A and B of compatible dimensions, $\det(AB) = \det(BA)$, we get

$$\begin{aligned}(\det R_z)^2 &= 2^{2p} \det R_x \det[(R_x + jR_{xy})(I - jR_x^{-1}R_{xy})] \\ &= 2^{2p} \det(R_x) \det(R_x + R_{xy}R_x^{-1}R_{xy}) \\ &\overset{(4.148)}{=} 2^{2p} \det R \qquad (4.151)\end{aligned}$$

so that

$$\frac{1}{(2\pi)^p} \frac{1}{\sqrt{\det R}} = \frac{1}{\pi^p} \frac{1}{\det R_z} \qquad (4.152)$$

To conclude the argument, some algebra will show that the exponents in (4.138) and (4.147) are identical — see Prob. 4.23.

■

We can also determine an expression for the fourth-order moment of a circular Gaussian random variable. Following the same argument that led to (4.26), we can similarly verify that if $\boldsymbol{z}$ is a circular Gaussian vector with zero mean and covariance matrix $\mathbb{E}\,\boldsymbol{z}\boldsymbol{z}^* = R_z$ then, for any Hermitian matrix W of compatible dimensions:

$$\boxed{\mathbb{E}\left\{\boldsymbol{z}\boldsymbol{z}^*W\boldsymbol{z}\boldsymbol{z}^*\right\} = R_z\text{Tr}\big(WR_z\big) + R_zWR_z} \qquad (4.153)$$

4.7 COMMENTARIES AND DISCUSSION

Gaussian distribution. The origin of the Gaussian distribution is attributed to the German mathematician **Carl Friedrich Gauss (1777–1855)** who published it in Gauss (1809) while working on two other original ideas, namely, the formulation of the least-squares criterion and the formulation of an early version of the maximum likelihood criterion. He started from a collection of N independent noisy measurements, $\boldsymbol{y}(n) = \theta + \boldsymbol{v}(n)$, of some unknown parameter θ where the perturbation error was assumed to arise from some unknown probability density function, $f_{\boldsymbol{v}}(v)$. Gauss (1809) formulated the problem of estimating the parameter by maximizing the product of the individual probabilities:

$$\widehat{\theta} = \underset{\theta}{\text{argmax}} \left\{ \prod_{n=1}^{N} f_{\boldsymbol{v}}\Big(y(n) - \theta\Big) \right\} \qquad (4.154)$$

He actually worked on a "reverse" problem. He posed the question of determining the form of the noise pdf that would result in an estimate for θ that is equal to the sample mean of the observations, namely, he wanted to arrive at a solution of the form:

$$\widehat{\theta} = \frac{1}{N}\sum_{n=1}^{N} y(n) \qquad (4.155)$$

He argued that what we refer to today as the Gaussian distribution is the answer to his inquiry, i.e.,

$$f_{\boldsymbol{v}}(v) = \frac{1}{\sqrt{2\pi\sigma_v^2}} e^{-\frac{v^2}{2\sigma_v^2}} \qquad (4.156)$$

For this choice of pdf, the maximum-likelihood formulation (4.154) reduces to the least-squares problem

$$\widehat{\theta} \;=\; \underset{\theta}{\operatorname{argmin}} \;\left\{ \sum_{n=1}^{N} (y(n) - \theta)^2 \right\} \tag{4.157}$$

whose solution is (4.155).

Independently of Gauss, the Irish-American mathematician **Robert Adrain (1775–1843)** also arrived at the Gaussian distribution (as well as the least-squares formulation) in the work by Adrain (1808). He considered a similar estimation problem involving a collection of noisy measurements and postulated that the size of the error in each measurement should be proportional to the size of the measurement itself (namely, larger measurements should contain larger errors). He also postulated that the errors across measurements are independent of each other and moved on to derive the form of the error probability measure that would satisfy these properties, arriving again at the Gaussian distribution — he arrived at the exponential curve e^{-v^2} and refers to it as "*the simplest form of the equation expressing the nature of the curve of probability.*" He used this conclusion to determine the most probable value for the unknown parameter from the noisy observations and arrived again at the sample mean estimate (4.155). In the solution to this problem, he writes: "*Hence the following rule: Divide the sum of all the observed values by their number, and the quotient will be the most probable value required.*" For further details on the early developments of this branch of statistical analysis, the reader may refer to Stigler (1986).

We will encounter the maximum likelihood formulation more generally in later chapters. It has become a formidable tool in modern statistical signal analysis, pushed largely by the foundational work of the English statistician **Ronald Fisher (1890–1962)**, who formulated and studied the maximum likelihood approach in its generality in Fisher (1912,1922,1925).

Central limit theorem. We explained in the introductory section of this chapter that the Gaussian distribution, also called *normal distribution*, derives its eminence from the central limit theorem. According to Feller (1945), the name "central limit theorem" is due to Pólya (1920). The earliest formulation of the central limit theorem, and its recognition as a powerful universal approximation law for sums of independent random variables, is due to the French mathematician **Pierre Laplace (1749–1827)** in the treatise by Laplace (1812). He considered a collection of N independent and identically distributed scalar random variables $\{\boldsymbol{x}_n\}$ with mean $\bar{x}$ and finite variance σ_x^2 and showed that the normalized variable:

$$\boldsymbol{y} \;\stackrel{\Delta}{=}\; \sqrt{N}\left(\frac{1}{N}\sum_{n=1}^{N}(\boldsymbol{x}_n - \bar{x})\right) \tag{4.158}$$

converges in distribution to $\mathcal{N}_{\boldsymbol{y}}(0, \sigma_x^2)$ as $N \to \infty$, written as

$$\boldsymbol{y} \xrightarrow{d} \mathcal{N}_{\boldsymbol{y}}(0, \sigma_x^2) \tag{4.159}$$

It was not, however, until almost a century later in the works by the Russian and Finnish mathematicians **Aleksandr Lyapunov (1857–1918)** and **Jarl Lindeberg (1876–1932)**, respectively, that the central limit theorem was generalized and placed on a more solid and formal footing — see the treatments by Billingsley (1986) and Fischer (2011). Weaker versions of the theorem were developed where, for example, the requirement of identically distributed random variables was dropped. Both Lyapunov (1901) and Lindeberg (1922) derived sufficient conditions under which the theorem would continue to hold, with Lindeberg condition being one of the weakest sufficient (and almost necessary) condition available.

More specifically, let $\{\boldsymbol{x}_n, n = 1, 2, \ldots, N\}$ denote a collection of independent scalar random variables, with possibly different means and variances denoted by $\{\bar{x}_n, \sigma_{x,n}^2 < \infty\}$. We introduce the sum of variances factor

$$\sigma_N^2 \triangleq \sum_{n=1}^{N} \sigma_{x,n}^2 \tag{4.160}$$

and consider the normalized variable

$$\boldsymbol{y} \triangleq \frac{1}{\sigma_N} \sum_{n=1}^{N} (\boldsymbol{x}_n - \bar{x}_n) \tag{4.161}$$

Lyapunov condition guarantees the convergence of the distribution of $\boldsymbol{y}$ to $\mathcal{N}_{\boldsymbol{y}}(0,1)$ if there exists some $\lambda > 0$ for which

$$\lim_{N\to\infty} \left\{ \frac{1}{\sigma_N^{2+\lambda}} \sum_{n=1}^{N} \mathbb{E}\,(\boldsymbol{x}_n - \bar{x}_n)^{2+\lambda} \right\} = 0 \tag{4.162}$$

A weaker condition is Lindeberg requirement: it guarantees the convergence of the distribution of $\boldsymbol{y}$ to $\mathcal{N}_{\boldsymbol{y}}(0,1)$ if for every $\epsilon > 0$ it holds that

$$\lim_{N\to\infty} \left\{ \frac{1}{\sigma_N^2} \sum_{n=1}^{N} \mathbb{E}\,(\boldsymbol{x}_n - \bar{x}_n)^2\, \mathbb{I}\Big[|\boldsymbol{x}_n - \bar{x}_n| > \epsilon\sigma_N\Big] \right\} = 0 \tag{4.163}$$

where the notation $\mathbb{I}[x]$ denotes the indicator function and is defined as follows:

$$\mathbb{I}[x] \triangleq \begin{cases} 1, & \text{when argument } x \text{ is true} \\ 0, & \text{otherwise} \end{cases} \tag{4.164}$$

It can be verified that if condition (4.162) holds then so does (4.163) so that Lindeberg condition is weaker than Lyapunov condition. Both conditions essentially amount to requiring the summands that appear in (4.162) and (4.163) to assume small values with high probability.

Stein identity. Stein lemma (4.68), also known as Stein identity, is a useful tool in the study of Gaussian-distributed random variables — see Prob. 4.32. The identity is due to Stein (1973,1981) and was generalized by Hudson (1978) to the family of exponential distributions, as shown later in Example 5.2. It was also extended to the vector case by Arnold, Castillo, and Sarabia (2001). The identity is useful in computing moments of transformations of Gaussian random variables. It arises, for example, in the context of the expectation propagation algorithm, which we study in a later chapter. It has also found applications in many domains, including in finance and asset pricing — see, e.g., Ingersoll (1987) and Cochrane (2001). Several of the other integral identities involving Gaussian distributions derived in Examples 4.1 and 4.7 are motivated by arguments and derivations from Owen (1980), Patel and Read (1996), and Rasmussen and Williams (2006, Sec. 3.9). The proofs of Lemma 4.3 and Example 4.4 are motivated by the derivations from Sayed (2003,2008).

Gaussian processes. We introduced in Sec. 4.5 the notion of Gaussian processes and commented on their relation to kernel methods in learning and inference; we will discuss these latter methods in greater detail in Chapter 63. Gaussian processes are a useful modeling tool in the study of learning algorithms, as detailed, for example, in the text by Rasmussen and Williams (2006). We observe from expressions (4.131) and (4.134) for the Gaussian and Ornstein-Uhlenbeck kernels that the "correlation" between two points (x, x') in space decreases as the points move further apart from each other. For this reason, when Gaussian processes are used to model and solve learning and inference problems, it is noted that this property of their kernel translates into the inference decisions being based largely on the closest points in the training data (this behavior is similar to the nearest-neighbor rule discussed later in Chapter 52). One early reference on the application of Gaussian processes to statistical inference is the work by O'Hagan (1978). Other notable references on the use of Gaussian processes and kernels for regression and learning applications include Blight and Ott (1975), Poggio and Girosi (1990), Neal (1995,1996), and Williams and Rasmussen (1995). Similar techniques have also been used in geostatistics under the name of "krigging" — see, e.g., Journel and Huijbregts (1978), Ripley (1981), and Fedorov (1987).

Circular Gaussian distribution. The presentation in Sec. 4.6 is adapted from Kailath, Sayed, and Hassibi (2000). Expression (4.147) shows the form of a complex Gaussian distribution under the circularity assumption. The original derivation of this form is due to Wooding (1956) — see also Goodman (1963) and the texts by Miller (1974) and Picinbono (1993). This distribution was derived under the circularity condition (4.142), which enables the (pdf) to be completely characterized by the first and second-order moments of the complex variable. Under this same condition, uncorrelated Gaussian variables continue to be independent.

PROBLEMS[2]

4.1 Consider two independent and zero-mean real random variables $\{\boldsymbol{u}, \boldsymbol{w}\}$, where $\boldsymbol{u}$ and $\boldsymbol{w}$ are column vectors; both are M-dimensional. The covariance matrices of $\boldsymbol{u}$ and $\boldsymbol{w}$ are defined by $\mathbb{E}\,\boldsymbol{u}\boldsymbol{u}^{\mathsf{T}} = \sigma_u^2 I$ and $\mathbb{E}\,\boldsymbol{w}\boldsymbol{w}^{\mathsf{T}} = C$. Let $\boldsymbol{e}_a = \boldsymbol{u}^{\mathsf{T}}\boldsymbol{w}$.

(a) Show that $\mathbb{E}\,\boldsymbol{e}_a^2 = \sigma_u^2 \text{Tr}(C)$.

(b) Assume $\boldsymbol{u}$ is Gaussian-distributed. Show that $\mathbb{E}\,\boldsymbol{e}_a^2\|\boldsymbol{u}\|^2 = (M+2)\sigma_u^4\text{Tr}(C)$, where the notation $\|\cdot\|$ denotes the Euclidean norm of its argument.

4.2 Consider K Gaussian distributions with mean μ_k and variance σ_k^2 each. We index these components by $k = 1, 2, \ldots, K$. We select one component k at random with probability π_k. Subsequently, we generate a random variable $\boldsymbol{y}$ according to the selected Gaussian distribution $\mathcal{N}_{\boldsymbol{y}}(\mu_k, \sigma_k^2)$.

(a) What is the pdf of $\boldsymbol{y}$?

(b) What is the mean of $\boldsymbol{y}$?

(c) What is the variance of $\boldsymbol{y}$?

4.3 Let $\boldsymbol{x}$ be a real-valued random variable with pdf $f_{\boldsymbol{x}}(x)$. Define $\boldsymbol{y} = \boldsymbol{x}^2$.

(a) Use the fact that for any nonnegative $\boldsymbol{y}$, the event $\{\boldsymbol{y} \leq y\}$ occurs whenever $\{-\sqrt{y} \leq \boldsymbol{x} \leq \sqrt{y}\}$ to conclude that the pdf of $\boldsymbol{y}$ is given by

$$f_{\boldsymbol{y}}(y) = \frac{1}{2}\frac{f_{\boldsymbol{x}}(\sqrt{y})}{\sqrt{y}} + \frac{1}{2}\frac{f_{\boldsymbol{x}}(-\sqrt{y})}{\sqrt{y}}, \quad y > 0$$

(b) Assume $\boldsymbol{x}$ is Gaussian with zero mean and unit variance. Conclude that $f_{\boldsymbol{y}}(y) = \frac{1}{\sqrt{2\pi y}}e^{-y/2}$ for $y > 0$. *Remark.* The above pdf is known as the Chi-square distribution with one degree of freedom. A Chi-square distribution with k degrees of freedom is characterized by the pdf:

$$f_{\boldsymbol{y}}(y) = \frac{1}{2^{k/2}\Gamma(k/2)}\, y^{(k-2)/2}e^{-y/2}, \quad y > 0$$

where $\Gamma(\cdot)$ is the so-called Gamma function; it is defined by the integral $\Gamma(z) = \int_0^\infty s^{z-1}e^{-s}ds$ for $z > 0$. The function $\Gamma(\cdot)$ has the following useful properties: $\Gamma(1/2) = \sqrt{\pi}$, $\Gamma(z+1) = z\Gamma(z)$ for any $z > 0$, and $\Gamma(n+1) = n!$ for any integer

[2] Some problems in this section are adapted from exercises in Sayed (2003,2008).

$n \geq 0$. The Chi-square distribution with $k-$degrees of freedom is a special case of the Gamma distribution considered later in Prob. 5.2 using the parameters $\alpha = k/2$ and $\beta = 1/2$. The mean and variance of $\boldsymbol{y}$ are $\mathbb{E}\,\boldsymbol{y} = k$ and $\sigma_y^2 = 2k$.

(c) Let $\boldsymbol{y} = \sum_{j=1}^{k} \boldsymbol{x}_j^2$ denote the sum of the squares of k independent Gaussian random variables $\boldsymbol{x}_j$, each with zero mean and unit variance. Show that $\boldsymbol{y}$ is chi-square distributed with k degrees of freedom.

4.4 Refer to the pdf expression (4.77) for jointly-distributed Gaussian random vectors. Show that if $\boldsymbol{x}$ and $\boldsymbol{y}$ are uncorrelated, then they are also independent.

4.5 Consider the product of three Gaussian distributions over the same random variable $\boldsymbol{x} \in \mathbb{R}^M$:

$$g(x) = \mathcal{N}_{\boldsymbol{x}}(\bar{x}_a, R_a) \times \mathcal{N}_{\boldsymbol{x}}(\bar{x}_b, R_b) \times \mathcal{N}_{\boldsymbol{x}}(\bar{x}_c, R_c)$$

Find an expression for $g(x)$. How should the product be normalized so that $g(x)/Z$ is a Gaussian distribution over x? Find Z.

4.6 Consider a Gaussian distribution over $\boldsymbol{\theta} \sim \mathcal{N}_{\boldsymbol{\theta}}(\bar{\theta}, R_\theta)$, and a second Gaussian distribution over $\boldsymbol{y} \sim \mathcal{N}_{\boldsymbol{y}}(\theta, R_y)$ where the mean is defined in terms of a realization for $\theta \in \mathbb{R}^M$. Find closed-form expressions for the following integrals:

$$I_1 = \int_\theta \theta\, \mathcal{N}_{\boldsymbol{y}}(\theta, R_y)\, \mathcal{N}_\theta(\bar{\theta}, R_\theta) d\theta, \quad I_2 = \int_\theta \theta\theta^{\mathsf{T}}\, \mathcal{N}_{\boldsymbol{y}}(\boldsymbol{\theta}, R_y)\, \mathcal{N}_{\boldsymbol{\theta}}(\bar{\theta}, R_\theta) d\theta$$

4.7 Consider two distributions over the random variables $\boldsymbol{\theta}, \boldsymbol{y} \in \mathbb{R}^M$ of the form:

$$\begin{aligned} &f_{\boldsymbol{\theta}}(\theta) = \mathcal{N}_{\boldsymbol{\theta}}(\bar{\theta}, R_\theta) \\ &f_{\boldsymbol{y}|\boldsymbol{\theta}}(y|\theta) = (1-\alpha)\mathcal{N}_{\boldsymbol{y}}(\theta, R_1) + \alpha \mathcal{N}_{\boldsymbol{y}}(0, R_2), \quad \alpha \in (0,1) \end{aligned}$$

In other words, the conditional pdf of $\boldsymbol{y}$ is parameterized by θ, which appears as the mean of the first Gaussian term. Determine a closed-form expression for the marginal of $\boldsymbol{y}$.

4.8 Let $\boldsymbol{y}$ be a scalar Gaussian-distributed random variable, $\boldsymbol{y} \sim \mathcal{N}_{\boldsymbol{y}}(\bar{y}, \sigma_y^2)$. Establish (4.7). *Remark.* The reader can refer to Owen (1980) and Patel and Read (1996) for a list of similar integral expressions involving Gaussian distributions. Related discussions also appear in Rasmussen and Williams (2006, Ch. 3).

4.9 Differentiate both sides of the identity (4.7) once and twice relative to $\bar{y}$ and establish identities (4.9) and (4.11).

4.10 Consider a standard Gaussian distribution with zero mean and unit variance. The error function that is associated with it is denoted by $\text{erf}(z)$ and is defined as the integral

$$\text{erf}(z) \;\triangleq\; \frac{2}{\sqrt{\pi}} \int_0^z e^{-x^2} dx$$

The complementary error function is defined by $\text{erfc}(z) = 1 - \text{erf}(z)$.

(a) Verify that $\text{erf}(0) = 0$ and that the function tends to ± 1 as $z \to \pm\infty$.

(b) Comparing with (4.6), verify that $\text{erf}(z) = 2\Phi(\sqrt{2}z) - 1$.

4.11 Consider a Gaussian random vector $\boldsymbol{w} \sim \mathcal{N}_{\boldsymbol{w}}(0, R_w)$ where $\boldsymbol{w} \in \mathbb{R}^M$. Show that

$$\int_{-\infty}^{\infty} \text{erf}\Big(h_a^{\mathsf{T}} w\Big)\, \text{erf}\Big(h_b^{\mathsf{T}} w\Big)\, \mathcal{N}_{\boldsymbol{w}}(0, R_w) dw = \frac{2}{\pi} \arcsin\left(\frac{2h_a^{\mathsf{T}} R_w h_b}{\sqrt{(1+2h_a^{\mathsf{T}} R_w h_a)(1+2h_b^{\mathsf{T}} R_w h_b)}}\right)$$

Remark. See Williams (1996) for a related discussion.

4.12 Use identity (4.34) to show that the calculation in (4.38) leads to a Gaussian pdf with mean $\bar{z} = \bar{x} + \bar{y}$ and covariance matrix $R_z = R_x + R_y$.

4.13 Let $\boldsymbol{a}$ denote a real scalar-valued Gaussian random variable with zero mean and variance σ_a^2. Show that $\mathbb{E}\,\boldsymbol{a}^4 = 3\sigma_a^4$.

4.14 Let $\boldsymbol{a}$ denote a complex circular Gaussian random variable with zero mean and variance σ_a^2. Show that $\mathbb{E}\,|\boldsymbol{a}|^4 = 2\sigma_a^4$.

4.15 Assume $\boldsymbol{u}$ is a real Gaussian random column vector with a diagonal covariance matrix Λ. Define $\boldsymbol{z} = \|\boldsymbol{u}\|^2$. What is the variance of $\boldsymbol{z}$?

4.16 Consider two column vectors $\{\boldsymbol{w}, \boldsymbol{z}\}$ that are related via $\boldsymbol{z} = \boldsymbol{w} + \mu \boldsymbol{u}(\boldsymbol{d} - \boldsymbol{u}^{\mathsf{T}}\boldsymbol{w})$, where $\boldsymbol{u}$ is a real Gaussian column vector with a diagonal covariance matrix, $\mathbb{E}\,\boldsymbol{u}\boldsymbol{u}^{\mathsf{T}} = \Lambda$. Moreover, μ is a positive constant and $\boldsymbol{d} = \boldsymbol{u}^{\mathsf{T}} w^o + \boldsymbol{v}$, for some constant vector w^o and random scalar $\boldsymbol{v}$ with variance σ_v^2. The variables $\{\boldsymbol{v}, \boldsymbol{u}, \boldsymbol{w}\}$ are independent of each other. Define $\boldsymbol{e}_a = \boldsymbol{u}^{\mathsf{T}}(w^o - \boldsymbol{w})$, as well as the error vectors $\widetilde{\boldsymbol{z}} = w^o - \boldsymbol{z}$ and $\widetilde{\boldsymbol{w}} = w^o - \boldsymbol{w}$, and denote their covariance matrices by $\{R_{\tilde{z}}, R_{\tilde{w}}\}$. Assume $\mathbb{E}\,\boldsymbol{z} = \mathbb{E}\,\boldsymbol{w} = w^o$, while all other random variables are zero-mean.

(a) Verify that $\widetilde{\boldsymbol{z}} = \widetilde{\boldsymbol{w}} - \mu \boldsymbol{u}(\boldsymbol{e}_a + \boldsymbol{v})$.

(b) Show that $R_{\tilde{z}} = R_{\tilde{w}} - \mu R_{\tilde{w}}\Lambda - \mu \Lambda R_{\tilde{w}} + \mu^2\left(\Lambda \mathrm{Tr}\big(R_{\tilde{w}}\Lambda\big) + 2\Lambda R_{\tilde{w}}\Lambda\right) \;+\; \mu^2\sigma_v^2\Lambda$.

4.17 Consider a collection of N measurements $\{\boldsymbol{\gamma}(n), \boldsymbol{h}_n\}$ where each scalar $\boldsymbol{\gamma}(n)$ is modeled as a noisy perturbation of some Gaussian process $\boldsymbol{g}(h_n)$ defined over the $M-$dimensional vectors $\{h_n\}$:

$$\boldsymbol{\gamma}(n) = \boldsymbol{g}(h_n) + \boldsymbol{v}(n), \qquad \boldsymbol{g} \sim \mathcal{GP}_{\boldsymbol{g}}\Big(0, K(h, h')\Big)$$

Assume the mean function for the Gaussian process $\boldsymbol{g}(\cdot)$ is zero and denote its covariance function by the kernel $K(h, h')$. Assume further that the noise $\boldsymbol{v}(n) \sim \mathcal{N}_{\boldsymbol{v}}(0, \sigma_v^2)$ is white Gaussian with variance σ_v^2 and independent of $\boldsymbol{g}(\cdot)$. Collect the measurements $\{\boldsymbol{\gamma}(n)\}$, the Gaussian process values of $\boldsymbol{g}(\cdot)$, and the perturbations $\{\boldsymbol{v}(n)\}$ into vector quantities:

$$\underbrace{\begin{bmatrix} \boldsymbol{\gamma}(1) \\ \boldsymbol{\gamma}(2) \\ \vdots \\ \boldsymbol{\gamma}(N) \end{bmatrix}}_{\boldsymbol{\gamma}_N} = \underbrace{\begin{bmatrix} \boldsymbol{g}(h_1) \\ \boldsymbol{g}(h_2) \\ \vdots \\ \boldsymbol{g}(h_N) \end{bmatrix}}_{\boldsymbol{g}_N} + \underbrace{\begin{bmatrix} \boldsymbol{v}(1) \\ \boldsymbol{v}(2) \\ \vdots \\ \boldsymbol{v}(N) \end{bmatrix}}_{\boldsymbol{v}_N}$$

so that $\boldsymbol{\gamma}_N = \boldsymbol{g}_N + \boldsymbol{v}_N$. Let R_N denote the covariance matrix of the vector $\boldsymbol{g}_N$ evaluated at the given feature data, $R_N = [K(h_n, h_m)]_{n,m=0}^{N-1}$.

(a) Argue that $\boldsymbol{\gamma}_N$ has a Gaussian distribution. What is its mean and covariance matrix?

(b) Consider a new vector h and its label $\boldsymbol{\gamma}$. What is the conditional pdf of $\boldsymbol{\gamma}$ given the past data $\{\gamma(n), h_n\}_{n=1}^{N}$?

4.18 Let $\boldsymbol{a}$ and $\boldsymbol{b}$ be scalar real-valued zero-mean jointly Gaussian random variables and denote their correlation by $\rho = \mathbb{E}\,\boldsymbol{a}\boldsymbol{b}$. Price theorem states that for any function $f(\boldsymbol{a}, \boldsymbol{b})$, for which the required derivatives and integrals exist, the following equality due to Price (1958) holds

$$\frac{\partial^n \mathbb{E}\, f(\boldsymbol{a}, \boldsymbol{b})}{\partial \rho^n} \;=\; \mathbb{E}\left(\frac{\partial^{2n} f(\boldsymbol{a}, \boldsymbol{b})}{\partial \boldsymbol{a}^n \partial \boldsymbol{b}^n}\right)$$

in terms of the n-th and $2n-$th order partial derivatives. In simple terms, Price theorem allows us to move the expectation on the left-hand side outside of the differentiation operation.

(a) Choose $n = 1$ and assume $f(\boldsymbol{a}, \boldsymbol{b})$ has the form $f(\boldsymbol{a}, \boldsymbol{b}) = \boldsymbol{a}g(\boldsymbol{b})$. Verify from Price theorem that

$$\frac{\partial \mathbb{E}\,\boldsymbol{a}g(\boldsymbol{b})}{\partial \rho} = \mathbb{E}\left(\frac{dg}{dx}\boldsymbol{b}\right)$$

in terms of the derivative of $g(\cdot)$. Integrate both sides over ρ to establish that $\mathbb{E}\,\boldsymbol{a}g(\boldsymbol{b}) = (\mathbb{E}\,\boldsymbol{a}\boldsymbol{b})\;\mathbb{E}\,(dg/dx)\boldsymbol{b}$.

(b) Show further that $\mathbb{E}\,\boldsymbol{b}g(\boldsymbol{b}) = \sigma_b^2\,\mathbb{E}\,(dg/dx)\boldsymbol{b}$ and conclude that the following relation also holds:

$$\mathbb{E}\,\boldsymbol{a}g(\boldsymbol{b}) \;=\; \frac{\mathbb{E}\,\boldsymbol{ab}}{\sigma_b^2}\,\mathbb{E}\,\boldsymbol{b}g(\boldsymbol{b})$$

(c) Assume $g(\boldsymbol{b}) = \text{sign}(\boldsymbol{b})$. Conclude from part (b) that

$$\mathbb{E}\,\boldsymbol{a}\;\text{sign}(\boldsymbol{b}) \;=\; \sqrt{\frac{2}{\pi}}\,\frac{1}{\sigma_b}\,\mathbb{E}\,\boldsymbol{ab}$$

4.19 Bussgang theorem is a special case of Price theorem and is due to Bussgang (1952). Let $\{\boldsymbol{a}, \boldsymbol{b}\}$ be two real zero-mean Gaussian random variables and define the function

$$g(\boldsymbol{b}) \;\triangleq\; \int_0^{\boldsymbol{b}} e^{-z^2/\sigma^2}\,dz$$

for some $\sigma > 0$. Bussgang theorem states that

$$\mathbb{E}\,\boldsymbol{a}g(\boldsymbol{b}) \;=\; \frac{1}{\sqrt{\frac{\sigma_b^2}{\sigma^2}+1}}\,\mathbb{E}\,\boldsymbol{ab}$$

The proof of the theorem is as follows. Let $\rho = \mathbb{E}\,\boldsymbol{ab}$. Use Price general statement from Prob. 4.18 to verify that

$$\frac{\partial\mathbb{E}\,\boldsymbol{a}g(\boldsymbol{b})}{\partial\rho} \;=\; \mathbb{E}\left(\frac{\partial^2\boldsymbol{a}g(\boldsymbol{b})}{\partial\boldsymbol{a}\partial b}\right) \;=\; \mathbb{E}\left(e^{-\boldsymbol{b}^2/\sigma_z^2}\right)$$

Integrate both sides over ρ to establish Bussgang theorem.

4.20 Consider two real-valued zero-mean jointly Gaussian random variables $\{\boldsymbol{x}, \boldsymbol{y}\}$ with covariance matrix

$$\mathbb{E}\begin{bmatrix}\boldsymbol{x}\\ \boldsymbol{y}\end{bmatrix}\begin{bmatrix}\boldsymbol{x} & \boldsymbol{y}\end{bmatrix} \;=\; \begin{bmatrix}1 & \rho\\ \rho & 1\end{bmatrix}$$

That is, $\{\boldsymbol{x}, \boldsymbol{y}\}$ have unit variances and correlation ρ. Define the function

$$g(\boldsymbol{x}, \boldsymbol{y}) = \frac{2}{\pi\sigma^2}\int_0^{\boldsymbol{x}}\int_0^{\boldsymbol{y}} e^{-\alpha^2/2\sigma^2}e^{-\beta^2/2\sigma^2}\,d\alpha d\beta$$

for some $\sigma > 0$.

(a) Verify that $\partial^2 g(\boldsymbol{x}, \boldsymbol{y})/\partial\boldsymbol{x}\partial\boldsymbol{y} = \frac{2}{\pi\sigma^2}e^{-\boldsymbol{x}^2/2\sigma^2}e^{-\boldsymbol{y}^2/2\sigma^2}$, and show that

$$\mathbb{E}\,\frac{\partial^2 g(\boldsymbol{x}, \boldsymbol{y})}{\partial\boldsymbol{x}\partial\boldsymbol{y}} \;=\; \frac{2}{\pi}\,\frac{1}{\sqrt{(\sigma^2+1)^2-\rho^2}}$$

(b) Integrate the equality of part (a) over $\rho \in (0,1)$ and conclude that

$$\int_0^1 \mathbb{E}\left(\frac{\partial^2 g(\boldsymbol{x}, \boldsymbol{y})}{\partial\boldsymbol{x}\partial\boldsymbol{y}}\right)d\rho \;=\; \frac{2}{\pi}\arcsin\left(\frac{1}{1+\sigma^2}\right)$$

(c) Use Price identity (cf. Prob. 4.18) to conclude that

$$\mathbb{E}\,g(\boldsymbol{x}, \boldsymbol{y}) \;=\; \frac{2}{\pi}\arcsin\left(\frac{1}{1+\sigma^2}\right)$$

4.21 Start from (4.26) and show that result (4.30) holds.

4.22 Refer to the general form (4.147) of a circular Gaussian random vector. Show that uncorrelated Gaussian vectors are also independent.

4.23 Show that the exponents in (4.138) and (4.147) coincide.

4.24 Prove that if condition (4.162) holds then so does condition (4.163).

4.25 Let $\boldsymbol{x}$ denote a Bernoulli random variable, assuming the value one with probability p and the value zero with probability $1-p$. Let $\boldsymbol{S}_N$ denote the sum of N independent Bernoulli experiments. What is the asymptotic distribution of $\boldsymbol{S}_N/N$?

4.26 Let $\boldsymbol{x}_n$ denote a Bernoulli random variable, assuming the value one with probability p_n and the value zero with probability $1-p_n$. Note that we are allowing the probability of success to vary across experiments. Set $\lambda = 1$ and show that Lyapunov condition (4.162) is satisfied if

$$\lim_{N\to\infty} \sum_{n=1}^{N} p_n(1-p_n) = \infty$$

4.27 Let $\boldsymbol{x}_n$ denote a sequence of independent and identically distributed random variables with mean μ and variance $\sigma_x^2 < \infty$. Show that Lindeberg condition (4.163) is satisfied.

4.28 Let $\boldsymbol{x}_n$ denote a sequence of independent and identically distributed scalar random variables with mean $\mathbb{E}\,\boldsymbol{x}_n = \mu$ and finite variance, $\sigma_x^2 = \mathbb{E}\,(\boldsymbol{x}(n) - \mu)^2 < \infty$. Introduce the sample average estimator $\widehat{\boldsymbol{\mu}} = \frac{1}{N}\sum_{n=1}^{N}\boldsymbol{x}(n)$ and let $\sigma_{\widehat{\mu}}^2$ denote its variance.

(a) Verify that $\sigma_{\widehat{\mu}}^2 = \sigma_x^2/N$.

(b) Use Chebyshev inequality (3.28) to conclude the validity of the *weak law of large numbers*, namely, the fact that the sample average converges in probability to the actual mean as $N \to \infty$ — see also future Prob. 3.54:

$$\lim_{N\to\infty} \mathbb{P}\left(|\widehat{\boldsymbol{\mu}} - \mu| \geq \epsilon\right) = 0, \quad \text{for any } \epsilon \geq 0$$

4.29 Let $\{\boldsymbol{x}(n),\ n = 1, \ldots, N\}$ denote N independent realizations with mean μ and and finite variance, $\sigma_x^2 = \mathbb{E}\,(\boldsymbol{x}(n) - \mu)^2 < \infty$. Introduce the *weighted* sample average $\widehat{\boldsymbol{\mu}} = \sum_{n=1}^{N}\alpha(n)\boldsymbol{x}(n)$, where the scalars $\{\alpha(n)\}$ satisfy

$$\alpha(n) \geq 0, \quad \sum_{n=1}^{N}\alpha(n) = 1, \quad \lim_{N\to\infty}\sum_{n=1}^{N}\alpha^2(n) = 0$$

(a) Verify that $\mathbb{E}\,\widehat{\boldsymbol{\mu}} = \mu$ and $\sigma_{\widehat{\mu}}^2 = \sigma_x^2\left(\sum_{n=1}^{N}\alpha^2(n)\right)$.

(b) Conclude that $\sigma_{\widehat{\mu}}^2 \to 0$ as $N \to \infty$.

4.30 A sequence of $M \times 1$ random vectors $\boldsymbol{x}_n$ converges in distribution to a Gaussian random vector $\boldsymbol{x}$ with zero mean and covariance matrix R_x. A sequence of $M \times M$ random matrices $\boldsymbol{A}_n$ converges in probability to a constant matrix A. What is the asymptotic distribution of the random sequence $\boldsymbol{A}_n\boldsymbol{x}_n$?

4.31 Consider a collection of N independent and identically distributed random variables $\{\boldsymbol{x}_n\}$, each with mean μ and variance σ_x^2. Introduce the sample mean estimator $\widehat{\boldsymbol{\mu}}_N = (1/N)\sum_{n=1}^{N}\boldsymbol{x}_n$, whose mean and variance are given by $\mathbb{E}\,\widehat{\boldsymbol{\mu}} = \mu$ and $\sigma_{\widehat{\mu}}^2 = \sigma_x^2/N$.

(a) Let a and δ be small positive numbers. Use Chebyshev inequality (3.28) to conclude that at least $N \geq 1/a^2\delta$ samples are needed to ensure that the sample mean lies within the interval $\mu \pm a\sigma_x$ with high likelihood of at least $1-\delta$, namely, $\mathbb{P}(|\widehat{\boldsymbol{\mu}}_N - \mu| < a\sigma_x) \geq 1 - \delta$.

(b) Use the central limit theorem (4.158) to find that the conclusion holds by selecting N to satisfy $\delta \leq 2Q(a\sqrt{N})$, where $Q(\cdot)$ denotes the standard Gaussian cumulative distribution function (i.e., the area under the standard Gaussian distribution $\mathcal{N}(0,1)$):

$$Q(x) \triangleq \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-\frac{1}{2}t^2}\,dt$$

The $Q-$function is usually tabulated in books on statistics.

(c) Compare the results of parts (a) and (b) to ensure that $\widehat{\boldsymbol{\mu}}$ lies within the interval $\mu \pm 0.05\sigma_x$ with probability larger than or equal to 0.995.

4.32 Consider the same setting of Stein lemma stated in (4.68). Consider two jointly Gaussian distributed scalar random variables $\boldsymbol{x}$ and $\boldsymbol{y}$. Show that it also holds

$$\mathbb{E}\left(g(\boldsymbol{x}) - \mathbb{E}\, g(\boldsymbol{x})\right)(\boldsymbol{y} - \bar{y}) \;=\; \mathbb{E}\,(\boldsymbol{x} - \bar{x})(\boldsymbol{y} - \bar{y})\mathbb{E}\, g'(\boldsymbol{x})$$

4.33 Repeat the proof of Stein Lemma and establish the validity of (4.67) for vector random variables $\boldsymbol{x} \in \mathbb{R}^M$.

4.34 Refer to expression (4.111) for Z_0. Compute the gradient vector and Hessian matrix of both sides of the equality and establish the validity of results (4.113) and (4.114).

REFERENCES

Adrain, R. (1808), "Research concerning the probabilities of the errors which happen in making observations," *The Analyst*, vol. I, no. 4, pp. 93–109.

Arnold, B. C., E. Castillo, and J. M. Sarabia (2001), "A multivariate version of Stein?s identity with applications to moment calculations and estimation of conditionally specified distributions," *Comm. Statist. Theory Methods*, vol. 30, pp. 2517–2542.

Billingsley, P. (1986), *Probability and Measure*, 2nd edition, Wiley, NY.

Blight, B. J. N. and L. Ott (1975), "A Bayesian approach to model inadequacy for polynomial regression," *Biometrika*, vol. 62, no. 1, pp. 79–88.

Bussgang, J. J. (1952), *Cross-Correlation Functions of Amplitude Distorted Gaussian Signals*, Tech. Report 216, MIT Research Laboratory of Electronics, Cambridge, MA.

Cochrane, J. H. (2001), *Asset Pricing*, Princeton University Press.

Fedorov, V. (1987), "Kriging and other estimators of spatial field characteristics," working paper WP-87-99, *International Institute for Applied Systems Analysis* (IIASA), Austria.

Feller, W. (1945), "The fundamental limit theorems in probability," *Bull. Amer. Math. Soc.*, vol. 51, pp. 800–832.

Fischer, H. (2011), *A History of the Central Limit Theorem*, Springer, NY.

Fisher, R. A. (1912), "On an absolute criterion for fitting frequency curves," *Messeg. Math.*, vol. 41, pp. 155–160.

Fisher, R. A. (1922), "On the mathematical foundations of theoretical statistics," *Philos. Trans. Roy. Soc. London Ser. A.*, vol. 222, pp. 309–368.

Fisher, R. A. (1925), "Theory of statistical estimation," *Proc. Cambridge Philos. Soc.*, vol. 22, pp. 700–725.

Gauss, C. F. (1809), *Theory of the Motion of the Heavenly Bodies Moving about the Sun in Conic Sections*, English translation by C. H. Davis, 1857, Little, Brown, and Company, Boston, MA.

Goodman, N. (1963), "Statistical analysis based on a certain multivariate complex Gaussian distribution," *Ann. Math. Stat.*, vol. 34, pp. 152–177.

Hudson, H. M. (1978), "A natural identity for exponential families with application in a multiparameter estimation," *Ann. Statist.*, vol. 6, pp. 473–484.

Ingersoll, J. (1987), *Theory of Financial Decision Making*, Rowman and Littlefield.

Journel, A. G. and C. J. Huijbregts (1978), *Mining Geostatistics*, Academic Press.

Kailath, T., A. H. Sayed, and B. Hassibi (2000), *Linear Estimation*, Prentice Hall, NJ.

Laplace, P. S. (1812), *Théorie Analytique des Probabilités*, Paris.

Lindeberg, J. W. (1922), "Eine neue Herleitung des Exponentialgesetzes in der Wahrscheinlichkeitsrechnung," *Mathematische Zeitschrift*, vol. 15, no.1, pp. 211–225.

Lyapunov, A. M. (1901), "Nouvelle forme du théoreme sur la limite de probabilité," *Mémoires de l'Académie de Saint-Petersbourg*, vol. 12, no. 8, pp. 1–24.

Miller, K. (1974), *Complex Stochastic Processes*, Addison-Wesley, Reading, MA.

Neal, R. (1995), *Bayesian Learning for Neural Networks*, PhD dissertation, Dept. of Computer Science, University of Toronto, Canada.

Neal, R. (1996), *Bayesian Learning for Neural Networks*, Springer, NY.

O'Hagan, A. (1978), "Curve fitting and optimal design for prediction," *J. Royal Statistical Society*, Series B (Methodological), vol. 40, no. 1, pp. 1–42.

Owen, D. (1980), "A table of normal integrals" *Communications in Statistics: Simulation and Computation*, vol. B9, pp. 389–419.

Patel, J. K. and C. B. Read (1996), *Handbook of the Normal Distribution*, 2nd edition, CRC Press.

Picinbono, B. (1993), *Random Signals and Systems*, Prentice-Hall, NJ.

Poggio, T. and F. Girosi (1990), "Networks for approximation and learning," *Proc. IEEE*, vol. 78, no. 9, pp. 1481–1497.

Polya, G. (1920), "Ueber den zentralen Grenzwertsatz der Wahrscheinlichkeitsrechnung und das Momentenproblem," *Math. Zeit.* vol. 8, pp. 171–181.

Price, R. (1958), "A useful theorem for nonlinear devices having Gaussian inputs," *IRE Trans. Inform. Theory*, vol. 4, pp. 69–72.

Rasmussen, C. E. and C. K. I. Williams (2006), *Gaussian Processes for Machine Learning*, MIT Press.

Ripley, B. D. (1981), *Spatial Statistics*, Wiley, NY.

Sayed, A. H. (2003), *Fundamentals of Adaptive Filtering*, Wiley, NJ.

Sayed, A. H. (2008), *Adaptive Filters*, Wiley, NJ.

Stein, C. M (1973), "Estimation of the mean of a multivariate normal distribution," *Proc. Symposium on Asymptotic Statistics*, pp. 345–381, Prague.

Stein, C. M. (1981), "Estimation of the mean of a multivariate normal distribution," *Ann. Statist.*, vol. 9, no. 6, pp. 1135–1151.

Stigler, S. M. (1986), *The History of Statistics: The Measurement of Uncertainty Before 1900*, Harvard University Press.

Williams, C. K. I. and C. E. Rasmussen (1995), "Gaussian processes for regression," in *Proc. Neural Information Processing Systems* (NIPS), pp. 514–520, Denver, CO.

Wooding, R. (1956), "The multivariate distribution of complex normal variables," *Biometrika*, vol. 43, pp. 212–215.