2

Gaussian Distributions and the Heat Equation

In this chapter the Gaussian distribution is defined and its properties are explored. The chapter starts with the definition of a Gaussian distribution on the real line. In the process of exploring the properties of the Gaussian on the line, the Fourier transform and heat equation are introduced, and their relationship to the Gaussian is developed. The Gaussian distribution in multiple dimensions is defined, as are clipped and folded versions of this distribution. Some concepts from probability and statistics such as mean, variance, marginalization, and conditioning of probability densities are introduced in a concrete way using the Gaussian as the primary example. The properties of the Gaussian distribution are fundamental to understanding the concept of white noise, which is the driving process for all of the stochastic processes studied in this book.

The main things to take away from this chapter are:

- To become familiar with the Gaussian distribution and its properties, and to be comfortable in performing integrals involving multi-dimensional Gaussians;
- To become acquainted with the concepts of mean, covariance, and information-theoretic entropy;
- To understand how to marginalize and convolve probability densities, to compute conditional densities, and to fold and clip Gaussians;
- To observe that there is a relationship between Gaussian distributions and the heat equation;
- To know where to begin if presented with a diffusion equation, the symmetries of which are desired.

2.1 The Gaussian Distribution on the Real Line

2.1.1 Defining Parameters

The Gaussian distribution on the real line is any function of the form $\rho(x - x_0)$ where

$$\rho(x) = ce^{-ax^2} \tag{2.1}$$

and $c \in \mathbb{R}_{>0}$ is related to $a \in \mathbb{R}_{>0}$ by the constraint that

$$I \doteq \int_{-\infty}^{\infty} \rho(x)dx = 1. \tag{2.2}$$

G.S. Chirikjian, *Stochastic Models, Information Theory, and Lie Groups, Volume 1: Classical Results and Geometric Methods*, Applied and Numerical Harmonic Analysis, DOI 10.1007/978-0-8176-4803-9_2,

This constraint, together with the fact that $\rho(x) \geq 0$ makes it a *probability density function* (or *pdf* for short). That is, any non-negative function satisfying (2.2) (not only those of the form in (2.1)) is a pdf.

The Gaussian distribution is the "bell curve" so often referred to when discussing statistical quantities. It is an infinitely differentiable function. Taking the first derivative gives

$$\frac{d\rho}{dx} = -2acxe^{-ax^2}.$$

From this it is clear that $\rho(x)$ has a critical point at $x = 0$, and this is its only critical point. The second derivative of $\rho(x)$ evaluated at $x = 0$ is

$$\frac{d^2\rho}{dx^2}|_{x=0} = -2ac,$$

which is always negative, indicating that $x = 0$ is a maximum, and the maximal value that $\rho(x)$ can attain is c. Furthermore, due to the negative sign in the exponential, the function $\rho(x)$ decays to zero very rapidly as $|x|$ increases. The Gaussian distribution is called *unimodal* because it has only one local maximum, or mode.

To determine the functional relationship between c and a that ensures that $I = 1$, the following trick can be used. First evaluate

$$I^2 = c^2 \left(\int_{-\infty}^{\infty} e^{-ax^2} dx \right)^2 = c^2 \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-a(x^2+y^2)} dxdy.$$

Then, changing to the polar coordinates $x = r\cos\theta$ and $y = r\sin\theta$ it becomes clear that

$$I^2 = c^2 \int_0^{2\pi} \int_0^{\infty} e^{-ar^2} rdrd\theta.$$

The integral over θ reduces to 2π and the integral over r can be performed in closed form. The resulting relationship between c and a is then $I^2 = c^2\pi/a = 1$, or

$$c = \sqrt{\frac{a}{\pi}}. \tag{2.3}$$

The Gaussian distribution is an *even function*, and for any finite positive value of a it is also a "nice" function. An even function is one for which $f_e(x) = f_e(-x)$ and an *odd function* is one for which $f_o(x) = -f_o(-x)$. Any function can be decomposed into a sum of even and odd functions as $f(x) = f_e(x) + f_o(x)$ where

$$f_e(x) = \frac{1}{2}[f(x) + f(-x)] \quad \text{and} \quad f_o(x) = \frac{1}{2}[f(x) - f(-x)].$$

Furthermore, the product of two even functions is even, the product of two odd functions is even, and the product of one even and one odd function is odd. The integral of any well-behaved odd function over any finite interval that is symmetric around the origin is always zero. This can be seen as follows:

$$\int_{-b}^{b} f_o(x)dx = \int_{-b}^{0} f_o(x)dx + \int_{0}^{b} f_o(x)dx,$$

but from the definition of an odd function,

$$\int_{-b}^{0} f_o(x)dx = -\int_{-b}^{0} f_o(-x)dx = -\int_{0}^{b} f_o(y)dy,$$

and so

$$\int_{-b}^{b} f_o(x)dx = 0.$$

For an even function

$$\int_{-b}^{b} f_e(x)dx = 2\int_{0}^{b} f_e(x).$$

For an even function, the product $x \cdot f_e(x)$ must be an odd function, and since odd functions integrate to zero over any interval $[-b, b]$, it follows that

$$\int_{-\infty}^{\infty} x f_e(x)dx = \lim_{b\to\infty}\int_{-b}^{b} x f_e(x)dx = 0.$$

This limit would exist even if the upper and lower integrands go to ±∞ at different rates because $f_e(x)$, like the other functions in this book, is restricted to be a "nice" function in the sense defined in (1.19), and hence it must decay to zero faster than $1/x$ as $x \to \pm\infty$. More generally, the quantity μ defined by the integral

$$\mu \doteq \int_{-\infty}^{\infty} x f(x)dx$$

for any probability density function, $f(x)$, is called the *mean.*

From the shift-invariance property of integration of an arbitrary integrable function on the real line,[1]

$$\int_{-\infty}^{\infty} f(x - x_0)dx = \int_{-\infty}^{\infty} f(x)dx,$$

it follows that for the special case of a Gaussian distribution shifted by μ, $\rho(x-\mu)$,

$$\int_{-\infty}^{\infty} x\rho(x-\mu)dx = \int_{-\infty}^{\infty} (y+\mu)\rho(y)dy = 0 + \mu \cdot I = \mu.$$

The *median* of the Gaussian distribution is the point m for which

$$\int_{-\infty}^{m} \rho(x)dx = \int_{m}^{\infty} \rho(x)dx.$$

Due to the fact that the Gaussian distribution is an even function, $m = 0$.

In statistics it is useful to have indicators that describe how concentrated or how spread out a distribution is. One such indicator is the *variance*, defined as

$$\sigma^2 \doteq \int_{-\infty}^{\infty} x^2 f(x-\mu)dx. \tag{2.4}$$

[1] Another often-glossed-over property of integration of functions on the real line that will be useful later is invariance under inversion of the argument:

$$\int_{-\infty}^{\infty} f(-x)dx = \int_{-\infty}^{\infty} f(x)dx.$$

The square root of the variance is called the *standard deviation.* Note that this is different from

$$s \doteq \int_{-\infty}^{\infty} |x| f(x-\mu)dx, \tag{2.5}$$

which is called the *spread.* Of course, the concepts of mean, mode, variance, and spread are not limited to the study of Gaussian distributions. They can be calculated for any pdf.

For the Gaussian distribution in (2.1) with normalization (2.3), the mean, median, variance, and spread can be calculated in the following closed form:

$$\mu = m = 0\,, \qquad \sigma^2 = \frac{1}{2a}\,, \qquad \text{and} \qquad s = \frac{1}{\sqrt{\pi a}}. \tag{2.6}$$

In general, non-Gaussian pdfs can have multiple modes, the mean and median need not be at the same point, and the relationship between spread and variance need not be so simple.

Since for a Gaussian these quantities are directly related to a, the Gaussian distribution can be redefined with σ^2 or s incorporated into the definition. The most common choice is to use σ^2, in which case the Gaussian distribution with mean at μ and standard deviation σ is denoted[2]

$$\boxed{\rho(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}.} \tag{2.7}$$

In some instances, such as in the following subsections, it will be more convenient to write this as $\rho_{(\mu,\sigma^2)}(x)$. Note: another common name for the Gaussian distribution is the *normal distribution.* Figure 2.1 shows a plot of the Gaussian distribution with $\mu = 0$ and $\sigma = 1$ plotted over the range $[-3, 3]$. Most (approximately 97 percent) of the probability density falls on this finite interval. Changing the value of μ or σ would only shift or uniformly stretch this plot.

The integral

$$F(x;\mu,\sigma^2) = \int_{-\infty}^{x} \rho(\xi;\mu,\sigma^2)d\xi$$

is called the *cumulative distribution function.* This function is known to have a “closed-form” solution in terms of error integrals. In the limit as $\sigma \to 0$, $F(x;\mu,\sigma^2)$ exhibits a sharp transition from a value of 0 for $x < \mu$ to a value of 1 for $x > \mu$. When $\mu = 0$ this is idealized with the *Heaviside step function*

$$H(x) \doteq \begin{cases} 1 \text{ for } x > 0 \\ 0 \text{ for } x \leq 0. \end{cases} \tag{2.8}$$

2.1.2 The Maximum Entropy Property

The entropy of a pdf $f(x)$ is defined by the integral [23]

$$\boxed{S(f) = -\int_{-\infty}^{\infty} f(x)\log f(x)dx} \tag{2.9}$$

[2] The symbols $f(x)$ and $\rho(x)$ often will be used to denote generic pdfs, but when appended as $\rho(x;\mu,\sigma^2)$, this will always denote a Gaussian.

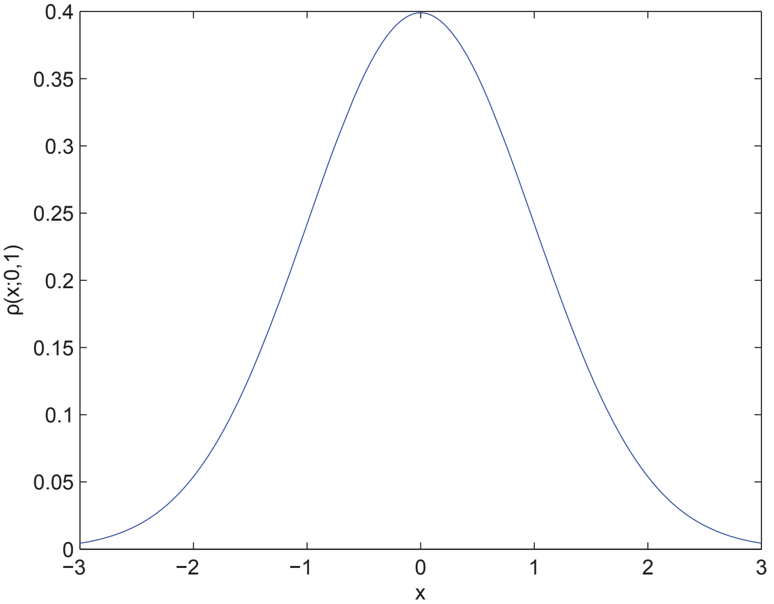

Fig. 2.1. The Gaussian Distribution $\rho(x; 0, 1)$ Plotted over $[-3, 3]$

where here $\log = \log_e = \ln$. This entropy is written as $S(f)$ rather than $S(f(x))$ because it is not a function of x, but rather it is a "functional" of f, since all dependence on x has been integrated out.

S is computed in closed form for the Gaussian distribution as

$$S(\rho_{(\mu,\sigma^2)}) = \log(\sqrt{2\pi e}\,\sigma). \tag{2.10}$$

Interestingly, for any given value of variance, the Gaussian distribution is the pdf with maximal entropy. This can be shown by performing the following optimization:

$$\max_f S(f) \quad \text{subject to} \quad f(x) \geq 0$$

and

$$\int_{-\infty}^{\infty} f(x)dx = 1\,, \quad \int_{-\infty}^{\infty} xf(x)dx = \mu\,, \quad \int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx = \sigma^2. \tag{2.11}$$

To find the distribution that satisfies these conditions, Lagrange multipliers[3] are introduced to enforce constraints, and the following necessary conditions are calculated:

$$\frac{\partial C}{\partial f} = 0 \quad \text{where} \quad C = -f\log f + \lambda_1 f + \lambda_2 xf + \lambda_3 (x-\mu)^2 f.$$

Performing the above calculation and solving for f and the λ_i that satisfy (2.11) gives $f(x) = \rho_{(\mu,\sigma^2)}(x)$. Note that the constraint $f(x) \geq 0$ was not actively enforced in the above derivation, but the result satisfies this condition anyway.

What the above shows is that $\rho_{(\mu,\sigma^2)}(x)$ extremizes the entropy subject to the given constraints. In other words, $\rho_{(\mu,\sigma^2)}$ is a critical point of the functional $S(f)$ subject

[3] See Section A.11.1 for a definition.

to the constraints (2.11). However, this could be a minimum, maximum, or point of inflection. To show that it actually maximizes the entropy (at least in a local sense), it is possible to define a perturbed version of this pdf as

$$f(x) = \rho_{(\mu,\sigma^2)}(x) \cdot [1 + \epsilon(x)] \tag{2.12}$$

where $\epsilon(x)$ is arbitrary except for the fact that[4] $|\epsilon(x)| << 1$ and it is defined such that $f(x)$ satisfies (2.11). In other words,

$$\int_{-\infty}^{\infty} \rho_{(\mu,\sigma^2)}(x)\epsilon(x)dx = \int_{-\infty}^{\infty} x\rho_{(\mu,\sigma^2)}(x)\epsilon(x)dx = \int_{-\infty}^{\infty} (x-\mu)^2\rho_{(\mu,\sigma^2)}(x)\epsilon(x)dx = 0.$$

Substituting (2.12) into (2.9) and using the Taylor series approximation $\log(1+\epsilon) \approx \epsilon - \epsilon^2/2$,

$$\begin{aligned} S(f) &= -\int_{-\infty}^{\infty} \rho_{(\mu,\sigma^2)}(x) \cdot [1+\epsilon(x)] \log(\rho_{(\mu,\sigma^2)}(x) \cdot [1+\epsilon(x)])dx \\ &= -\int_{-\infty}^{\infty} \rho_{(\mu,\sigma^2)}(x) \cdot [1+\epsilon(x)] \cdot [\log(\rho_{(\mu,\sigma^2)}(x)) + \log(1+\epsilon(x))]dx \\ &= S(\rho_{(\mu,\sigma^2)}) - F(\epsilon^2) + O(\epsilon^3) \end{aligned}$$

where the functional F is always positive and the cross terms that are linear in ϵ all vanish due to the integral constraints on ϵ. This means that at least locally a Gaussian maximizes entropy. Determining the exact form of the functional F is left as an exercise.

2.1.3 The Convolution of Gaussians

The *convolution* of two pdfs on the real line is defined as

$$(f_1 * f_2)(x) \doteq \int_{-\infty}^{\infty} f_1(\xi)f_2(x-\xi)d\xi. \tag{2.13}$$

Sometimes this is written as $f_1(x)*f_2(x)$. Note that convolution on the real line is commutative: $(f_1 * f_2)(x) = (f_2 * f_1)(x)$. This is a direct consequence of the commutativity of addition: $x + y = y + x$.

In order for the convolution integral to exist, $f_1(x)$ and $f_2(x)$ must both decay to zero sufficiently fast as $x \to \pm\infty$. In addition, the scope here is restricted to "nice" functions in the sense of (1.19) with $D = \mathbb{R}$. Therefore these functions are infinitely differentiable and have integrals of their square and absolute values that are finite. It can be shown that the convolution integral will always exist for such "nice" functions, and furthermore

$$f_i \in \mathcal{N}(\mathbb{R}) \implies f_1 * f_2 \in \mathcal{N}(\mathbb{R}).$$

In (2.13) ξ is a dummy variable of integration, the name of which is unimportant. A geometric interpretation of (2.13) is as follows. First, the function $f_2(x)$ is shifted along the real line in the positive direction by an amount ξ, resulting in $f_2(x-\xi)$. Then, the function f_1 evaluated at the amount of shift, $f_1(\xi)$, is used to weight $f_2(x-\xi)$. Finally, all copies of the product $f_1(\xi)f_2(x-\xi)$ are "added up" by integrating over all values of the shift. This has the effect of "smearing" f_2 over f_1.

[4] To be concrete, $\epsilon = 0.01 << 1$. Then $\epsilon^3 = 10^{-6}$ is certainly negligible in comparison to quantities that are on the order of 1.

In the case when $f_1(x) = \delta(x)$, i.e., the Dirac delta function, which is the probability density function with all of its mass concentrated at $x = 0$, $(\delta * f)(x) = f(x)$. This is because the only shift that the delta function allows is $\xi = 0$. All other shifts are weighted by a value of zero, and therefore do not contribute. While $\delta(x)$ is not a "nice" function, it is possible to approximate it with a Gaussian distribution with very small variance, ϵ, which is a "nice" function. The approximation of the Dirac delta function as $\delta(x) \approx \rho(x; 0, \epsilon)$ is deemed to be "good enough" if the integral of $|\rho(x; 0, \epsilon) * f(x) - f(x)|$ and the integral of the square of this are both "small enough" when $f(x)$ is a nice function.

The Gaussian distribution has the property that the convolution of two Gaussians is a Gaussian:

$$\rho(x; \mu_1, \sigma_1^2) * \rho(x; \mu_2, \sigma_2^2) = \rho(x; \mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2). \tag{2.14}$$

The Dirac δ-function can be viewed as the limit

$$\delta(x) = \lim_{\sigma \to 0} \rho(x; 0, \sigma^2). \tag{2.15}$$

It then follows from (2.14) that

$$\rho(x; \mu_1, \sigma_1^2) * \delta(x) = \rho(x; \mu_1, \sigma_1^2).$$

2.1.4 The Fourier Transform of the Gaussian Distribution

The Fourier transform of a "nice" function $f \in \mathcal{N}(\mathbb{R})$ is defined as

$$[\mathcal{F}(f)](\omega) \doteq \int_{-\infty}^{\infty} f(x) e^{-i\omega x} dx. \tag{2.16}$$

The shorthand $\hat{f}(\omega) \doteq [\mathcal{F}(f)](\omega)$ will be used frequently.

The conditions for existence and properties of the Fourier transform of functions on the real line are described in detail in [6, 11, 15]. Tools for the computation of fast sampled versions of the Fourier transform of periodic functions can be found in many books such as [7, 10, 24]. From the definition of the Fourier transform, it can be shown that

$$(\widehat{f_1 * f_2})(\omega) = \hat{f}_1(\omega) \hat{f}_2(\omega) \tag{2.17}$$

(i.e., the Fourier transform of the convolution is the product of Fourier transforms) and

$$f(x) = [\mathcal{F}^{-1}(\hat{f})](x) \doteq \frac{1}{2\pi} \int_{-\infty}^{\infty} \hat{f}(\omega) e^{i\omega x} d\omega. \tag{2.18}$$

This is called the *inverse Fourier transform* or *Fourier reconstruction formula.*

The proof of the property (2.17) is left as an exercise, whereas (2.18) is proven below. For more details about classical Fourier analysis and its extensions, see [8] and references therein.

The fact that a function is recovered from its Fourier transform is found by first observing that it is true for the special case of $g(x) = e^{-ax^2}$ for $a > 0$. One way to calculate

$$\hat{g}(\omega) = \int_{-\infty}^{\infty} e^{-ax^2} e^{-i\omega x} dx$$

is to differentiate both sides with respect to ω, which yields

$$\frac{d\hat{g}}{d\omega} = -i\int_{-\infty}^{\infty} xe^{-ax^2} e^{-i\omega x} dx = \frac{i}{2a}\int_{-\infty}^{\infty} \frac{dg}{dx} e^{-i\omega x} dx.$$

Integrating by parts, and observing that $e^{-i\omega x} g(x)$ vanishes at the limits of integration yields

$$\frac{d\hat{g}}{d\omega} = -\frac{\omega}{2a}\hat{g}.$$

The solution of this first-order ordinary differential equation is of the form

$$\hat{g}(\omega) = \hat{g}(0) e^{-\frac{\omega^2}{4a}}$$

where

$$\hat{g}(0) = \int_{-\infty}^{\infty} e^{-ax^2} dx = \sqrt{\frac{\pi}{a}}.$$

Having found the form of $\hat{g}(\omega)$, it is easy to see that $g(x)$ is reconstructed from $\hat{g}(\omega)$ using the inversion formula (2.18) (the calculation is essentially the same as for the forward Fourier transform). Likewise, the Gaussian function

$$\rho_{(0,\sigma^2)}(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}}$$

has Fourier transform

$$\hat{\rho}_{(0,\sigma^2)}(\omega) = e^{-\frac{\sigma^2}{2}\omega^2}$$

and the reconstruction formula holds. As σ becomes small, $\rho_{(0,\sigma^2)}(x)$ becomes like $\delta(x)$. From the property that $(\delta * f)(x) = f(x)$, the convolution theorem, and the above properties of Gaussian approximations to the Dirac δ-function, (2.18) immediately follows.

2.1.5 Diffusion Equations

A one-dimensional linear diffusion equation with constant coefficients has the form

$$\frac{\partial u}{\partial t} = a\frac{\partial u}{\partial x} + b\frac{\partial^2 u}{\partial x^2} \tag{2.19}$$

where $a \in \mathbb{R}$ is called the *drift coefficient* and $b \in \mathbb{R}_{>0}$ is called the *diffusion coefficient.* When modeling diffusion phenomena in an infinite medium, the above diffusion equation for $u(x,t)$ has initial conditions of the form $u(x,0) = f(x)$. The boundary conditions

$$u(\pm\infty, 0) = \frac{\partial u}{\partial x}(\pm\infty, 0) = 0$$

are implicit in this problem, because otherwise the solutions will not be pdfs, or in the class $\mathcal{N}(\mathbb{R})$.

Note that (2.19) is a special case of the *Fokker–Planck equation*[5] which will be examined in great detail in Chapter 4. When the drift coefficient is zero, the diffusion equation is called the *heat equation.*

Taking the Fourier transform of $u(x,t)$ for each value of t (i.e., treating time as a constant for the moment and x as the independent variable) produces $\hat{u}(\omega, t)$. Then applying the Fourier transform to both sides of (2.19) and the initial conditions results

[5] Also known as *Kolmogorov's forward equation.*

in a linear first-order ordinary differential equation with t as the independent variable, together with initial conditions, for each fixed frequency ω:

$$\frac{d\hat{u}}{dt} = (ia\omega - b\omega^2)\hat{u} \quad \text{with} \quad \hat{u}(\omega, 0) = \hat{f}(\omega).$$

The solution to this initial value problem is of the form

$$\hat{u}(\omega, t) = \hat{f}(\omega) e^{(ia\omega - b\omega^2)t}.$$

Application of the inverse Fourier transform yields a solution. The above expression for $\hat{u}(\omega, t)$ is a Gaussian with phase factor, and on inversion this becomes a shifted Gaussian:

$$[\mathcal{F}^{-1}(e^{iat\omega} e^{-b\omega^2 t})](x) = \frac{1}{\sqrt{4\pi bt}} \exp\left(-\frac{(x+at)^2}{4bt}\right).$$

Using the convolution theorem in reverse then gives

$$u(x,t) = \frac{1}{\sqrt{4\pi bt}} \int_{-\infty}^{\infty} f(\xi) \exp\left(-\frac{(x+at-\xi)^2}{4bt}\right) d\xi. \tag{2.20}$$

2.1.6 Stirling's Formula

In probability theory for discrete variables, the *binomial distribution* is defined as

$$f(k; n, p) \doteq \binom{n}{k} p^k (1-p)^{n-k} \quad \text{where} \quad \binom{n}{k} \doteq \frac{n!}{k!(n-k)!} \qquad 0 \leq p \leq 1 \tag{2.21}$$

and $k = 0, 1, 2, ..., n$, and the values $\binom{n}{k}$ are called *binomial coefficients*. From the *binomial theorem*,

$$(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k},$$

it follows that

$$\sum_{k=0}^{n} f(k; n, p) = (1 - p + p)^n = 1,$$

and from the definition in (2.21)

$$\sum_{k=0}^{n} k \cdot f(k; n, p) = np \cdot \sum_{k'=0}^{n-1} f(k'; n-1, p) = np \quad \text{where} \quad k' = k - 1.$$

The factorial $n!$ can be approximated using the *Stirling series*:

$$n! = \sqrt{2\pi n} \left(\frac{n}{e}\right)^n \left(1 + \frac{1}{12n} + \frac{1}{288n^2} + \cdots\right).$$

If the first term is kept, the result is *Stirling's formula*:

$$n! \approx \sqrt{2\pi n} \left(\frac{n}{e}\right)^n. \tag{2.22}$$

Stirling's formula is used extensively in probability theory to establish limiting behaviors. In the current context, it can be used to show that the Gaussian distribution is the limiting distribution of the binomial distribution in the sense that [22]

$$\lim_{n \to \infty} \frac{f(k; n, p)}{\rho(k; np, np(1-p))} = 1 \quad \text{for finite} \quad |k - np| / \sqrt{np(1-p)}. \tag{2.23}$$

2.2 The Multivariate Gaussian Distribution

The multivariate Gaussian distribution on $\mathbb{R}^n$ is defined as[6]

$$\boxed{\rho(\mathbf{x};\boldsymbol{\mu},\Sigma) \doteq \frac{1}{(2\pi)^{n/2}|\det\Sigma|^{\frac{1}{2}}}\exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}.} \tag{2.24}$$

This is the maximum entropy distribution subject to the constraints[7]

$$\int_{\mathbb{R}^n}\rho(\mathbf{x};\boldsymbol{\mu},\Sigma)\,d\mathbf{x}=1;\ \int_{\mathbb{R}^n}\mathbf{x}\,\rho(\mathbf{x};\boldsymbol{\mu},\Sigma)\,d\mathbf{x}=\boldsymbol{\mu};\ \int_{\mathbb{R}^n}(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T\rho(\mathbf{x};\boldsymbol{\mu},\Sigma)\,d\mathbf{x}=\Sigma. \tag{2.25}$$

The integral is calculated with respect to the differential volume element for $\mathbb{R}^n$, denoted above as $d\mathbf{x}=dx_1dx_2\cdots dx_n$. The above properties can be proved by changing coordinates as $\mathbf{y}=\Sigma^{-\frac{1}{2}}(\mathbf{x}-\boldsymbol{\mu})$, which reduces the problem to many one-dimensional integrals. The meaning of a fractional power of a matrix is reviewed in the appendix. Given a multi-dimensional coordinate transformation $\mathbf{y}=\mathbf{y}(\mathbf{x})$ (which is written in components as $y_i=y_i(x_1,...,x_n)$ for $i=1,...,n$), the following well-known integration rule (which is a restatement of (1.38) in different notation) holds:

$$\int_{\mathbf{y}(D)}F(\mathbf{y})d\mathbf{y}=\int_D F(\mathbf{y}(\mathbf{x}))|\det J|d\mathbf{x} \tag{2.26}$$

where $d\mathbf{x}=dx_1dx_2\cdots dx_n$, $d\mathbf{y}=dy_1dy_2\cdots dy_n$, and

$$J=\left[\frac{\partial\mathbf{y}}{\partial x_1},...,\frac{\partial\mathbf{y}}{\partial x_n}\right]$$

is the Jacobian matrix of the transformation and $|\det J|$ gives a measure of *local volume change*. D is the domain of integration in terms of the coordinates $\mathbf{x}$, and $\mathbf{y}(D)$ is the new domain to which each point in D is mapped under the transformation $\mathbf{y}(\mathbf{x})$. In the current context, the range of integrals over $\mathbf{x}$ and $\mathbf{y}$ are both copies of $\mathbb{R}^n$, i.e., $D=\mathbf{y}(D)=\mathbb{R}^n$.

2.2.1 Conditional and Marginal Densities

A vector $\mathbf{x}\in\mathbb{R}^n$ can be partitioned as

$$\mathbf{x}=\begin{pmatrix}\mathbf{x}_1\\ \mathbf{x}_2\end{pmatrix}=[\mathbf{x}_1^T,\mathbf{x}_2^T]^T\in\mathbb{R}^{n_1+n_2}$$

where $\mathbf{x}_1\in\mathbb{R}^{n_1}$ and $\mathbf{x}_2\in\mathbb{R}^{n_2}$. The notation $[\mathbf{x}_1^T,\mathbf{x}_2^T]^T$, which takes advantage of the fact that the “transpose of a transpose is the original,” has the benefit that it can be

[6] It is unfortunate that the notation for the one-dimensional case, $\rho(x;\mu,\sigma^2)$, is inconsistent with the multivariate case since σ^2 becomes Σ (rather than Σ^2), but this is the notation that is standard in the field.

[7] In Chapter 1 the notation $d(\mathbf{x})$ was used to denote the volume element $dx_1dx_2\cdots dx_n$. In the expressions in this chapter, the parentheses will be dropped to reduce the amount of clutter, and $d\mathbf{x}$ will be used as shorthand for $d(\mathbf{x})$. This will not cause trouble because $\mathbf{x}(t+dt)-\mathbf{x}(t)$ does not appear in any of these calculations.

written on one line and included in a sentence, whereas it is difficult to do so for a column vector.

If $f(\mathbf{x}) = f([\mathbf{x}_1^T, \mathbf{x}_2^T]^T)$ (which also will be referred to as $f(\mathbf{x}_1, \mathbf{x}_2)$) is any pdf on $\mathbb{R}^{n_1+n_2}$, then the marginal density $f_1(\mathbf{x}_1)$ is defined by integrating over all values of $\mathbf{x}_2$:

$$f_1(\mathbf{x}_1) = \int_{\mathbb{R}^{n_2}} f(\mathbf{x}_1, \mathbf{x}_2)\, d\mathbf{x}_2.$$

$f_2(\mathbf{x}_2)$ is obtained from $f(\mathbf{x}_1, \mathbf{x}_2)$ in a similar way by integrating over all values of $\mathbf{x}_1$.

The mean and variance of $f_1(\mathbf{x}_1)$ are obtained from the mean and variance of $f(\mathbf{x})$ by observing that

$$\begin{aligned}\boldsymbol{\mu}_1 &= \int_{\mathbb{R}^{n_1}} \mathbf{x}_1 f_1(\mathbf{x}_1)\, d\mathbf{x}_1 \\ &= \int_{\mathbb{R}^{n_1}} \mathbf{x}_1 \left(\int_{\mathbb{R}^{n_2}} f(\mathbf{x}_1, \mathbf{x}_2)\, d\mathbf{x}_2 \right) d\mathbf{x}_1 \\ &= \int_{\mathbb{R}^{n_1}} \int_{\mathbb{R}^{n_2}} \mathbf{x}_1 f(\mathbf{x}_1, \mathbf{x}_2)\, d\mathbf{x}_2\, d\mathbf{x}_1\end{aligned}$$

and

$$\begin{aligned}\Sigma_{11} &= \int_{\mathbb{R}^{n_1}} (\mathbf{x}_1 - \boldsymbol{\mu}_1)(\mathbf{x}_1 - \boldsymbol{\mu}_1)^T f_1(\mathbf{x}_1)\, d\mathbf{x}_1 \\ &= \int_{\mathbb{R}^{n_1}} (\mathbf{x}_1 - \boldsymbol{\mu}_1)(\mathbf{x}_1 - \boldsymbol{\mu}_1)^T \left(\int_{\mathbb{R}^{n_2}} f(\mathbf{x}_1, \mathbf{x}_2)\, d\mathbf{x}_2 \right) d\mathbf{x}_1 \\ &= \int_{\mathbb{R}^{n_1}} \int_{\mathbb{R}^{n_2}} (\mathbf{x}_1 - \boldsymbol{\mu}_1)(\mathbf{x}_1 - \boldsymbol{\mu}_1)^T f(\mathbf{x}_1, \mathbf{x}_2)\, d\mathbf{x}_2\, d\mathbf{x}_1.\end{aligned}$$

In other words, the mean vector and covariance matrix for the marginal density are obtained directly from those of the full density. For example, $\boldsymbol{\mu} = [\boldsymbol{\mu}_1^T, \boldsymbol{\mu}_2^T]^T$.

Given a (multivariate) Gaussian distribution $\rho(\mathbf{x}; \boldsymbol{\mu}, \Sigma)$, the associated covariance matrix can be written in terms of blocks as

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

where $\Sigma_{11} = \Sigma_{11}^T$, $\Sigma_{22} = \Sigma_{22}^T$, and $\Sigma_{21} = \Sigma_{12}^T$. The block Σ_{ij} has dimensions $n_i \times n_j$. In other words, $\Sigma_{ij} \in \mathbb{R}^{n_i \times n_j}$ where i and j can either be 1 or 2.

The *marginal density* that results from integrating the Gaussian distribution $\rho(\mathbf{x}, \boldsymbol{\mu}, \Sigma)$ over all values of $\mathbf{x}_2$ is

$$\int_{\mathbb{R}^{n_2}} \rho([\mathbf{x}_1^T, \mathbf{x}_2^T]^T; \boldsymbol{\mu}, \Sigma) d\mathbf{x}_2 = \rho(\mathbf{x}_1; \boldsymbol{\mu}_1, \Sigma_{11}). \tag{2.27}$$

This should not come as a surprise, since a Gaussian is defined completely by the values of its mean and covariance.

Another operation that is important in probability and statistics is that of conditioning. Given $f(\mathbf{x}_1, \mathbf{x}_2)$, the conditional density of $\mathbf{x}_1$ given $\mathbf{x}_2$ is

$$f(\mathbf{x}_1 | \mathbf{x}_2) \doteq f(\mathbf{x}_1, \mathbf{x}_2)/f_2(\mathbf{x}_2). \tag{2.28}$$

Evaluating this expression using a Gaussian gives

$$\rho([\mathbf{x}_1^T,\mathbf{x}_2^T]^T;\boldsymbol{\mu},\Sigma)/\rho(\mathbf{x}_2;\boldsymbol{\mu}_2,\Sigma_2)=\rho(\mathbf{x}_1;\boldsymbol{\mu}_1+\Sigma_{12}\Sigma_{22}^{-1}(\mathbf{x}_2-\boldsymbol{\mu}_2),\Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}). \tag{2.29}$$

The above formulas follow from decomposing Σ into a product of block lower triangular, block diagonal, and block upper triangular matrices as in Appendix A.4.3. Each of these can then be inverted in closed form resulting in explicit expressions for Σ^{-1} in terms of the blocks of Σ.

In summary, the set of Gaussian distributions has the remarkable property that it is closed under marginalization and conditioning, and as was demonstrated previously in the 1D case, it is also closed under convolution.

2.2.2 Multi-Dimensional Integrals Involving Gaussians

Several integral identities involving Gaussian distributions are used throughout this book. These are stated here and proved in the following subsections.

First, it is well known that

$$\int_{-\infty}^{\infty} e^{-\frac{1}{2}x^2}dx=\sqrt{2\pi} \quad\Longrightarrow\quad \int_{\mathbb{R}^n}\exp\left(-\frac{1}{2}\mathbf{x}^T\mathbf{x}\right)d\mathbf{x}=(2\pi)^{\frac{n}{2}}. \tag{2.30}$$

Here $\mathbf{x}\in\mathbb{R}^n$ and $d\mathbf{x}=dx_1dx_2\cdots dx_n$. Note also that

$$\int_{-\infty}^{\infty} x^2 e^{-\frac{1}{2}x^2}dx=\sqrt{2\pi}. \tag{2.31}$$

These identities are used below to prove

$$\int_{\mathbb{R}^n}\exp(-\frac{1}{2}\mathbf{x}^TM\mathbf{x}-\mathbf{m}^T\mathbf{x})d\mathbf{x}=(2\pi)^{n/2}|\mathrm{det}M|^{-\frac{1}{2}}\exp\left(\frac{1}{2}\mathbf{m}^TM^{-1}\mathbf{m}\right) \tag{2.32}$$

and

$$\int_{\mathbb{R}^n}\mathbf{x}^TG\mathbf{x}\exp\left(-\frac{1}{2}\mathbf{x}^TA\mathbf{x}\right)d\mathbf{x}=(2\pi)^{n/2}\,\frac{\mathrm{tr}(GA^{-1})}{|\det A|^{\frac{1}{2}}}. \tag{2.33}$$

These integrals have applications in the analysis of elastic network models of proteins [9].

Proof of Equation (2.32)

Consider the integral

$$I=\int_{\mathbb{R}^n}\exp(-\frac{1}{2}\mathbf{x}^TM\mathbf{x}-\mathbf{m}^T\mathbf{x})d\mathbf{x}.$$

Using the change of variables $\mathbf{z}=M^{\frac{1}{2}}\mathbf{x}-M^{-\frac{1}{2}}\mathbf{m}$ implies that $d\mathbf{z}=|\det M|^{\frac{1}{2}}d\mathbf{x}$ and $\mathbf{x}=M^{-\frac{1}{2}}(\mathbf{z}+M^{-\frac{1}{2}}\mathbf{m})$. Therefore

$$\begin{aligned} I&=\frac{1}{|\det M|^{\frac{1}{2}}}\int_{\mathbb{R}^n}\exp(-\frac{1}{2}\mathbf{z}^T\mathbf{z}+\frac{1}{2}\mathbf{m}^TM^{-1}\mathbf{m})d\mathbf{z}\\ &=\frac{\exp\left(\frac{1}{2}\mathbf{m}^TM^{-1}\mathbf{m}\right)}{|\det M|^{\frac{1}{2}}}\int_{\mathbb{R}^n}\exp(-\frac{1}{2}\mathbf{z}^T\mathbf{z})d\mathbf{z}. \end{aligned}$$

And so, (2.32) follows from (2.30).

Proof of Equation (2.33)

It is also convenient to have closed-form solutions for integrals of the form

$$J = \int_{\mathbb{R}^n} \mathbf{x}^T G \mathbf{x} \exp\left(-\frac{1}{2}\mathbf{x}^T A \mathbf{x}\right) d\mathbf{x}.$$

Let $\mathbf{z} = A^{\frac{1}{2}}\mathbf{x}$. Then

$$J = \frac{1}{|\det A|^{\frac{1}{2}}} \int_{\mathbb{R}^n} \mathbf{z}^T A^{-\frac{1}{2}} G A^{-\frac{1}{2}} \mathbf{z} \exp\left(-\frac{1}{2}\mathbf{z}^T \mathbf{z}\right) d\mathbf{z}.$$

Now let $G' = A^{-\frac{1}{2}} G A^{-\frac{1}{2}}$. Then it is clear that off-diagonal terms of G' do not contribute to this integral since odd moments of Gaussians are zero. Therefore,

$$\begin{aligned} J &= \frac{1}{|\det A|^{\frac{1}{2}}} \int_{\mathbb{R}^n} \sum_{i=1}^{n} g_{ii}^{'} z_i^2 \exp\left(-\frac{1}{2}\mathbf{z}^T \mathbf{z}\right) d\mathbf{z} \\ &= \frac{1}{|\det A|^{\frac{1}{2}}} \sum_{i=1}^{n} g_{ii}^{'} \int_{-\infty}^{\infty} z_i^2 e^{-\frac{1}{2} z_i^2} dz_i \int_{\mathbb{R}^{n-1}} \exp\left(-\frac{1}{2}\mathbf{y}_i^T \mathbf{y}_i\right) d\mathbf{y}_i \end{aligned}$$

where $\mathbf{y}_i \in \mathbb{R}^{n-1}$ is the part of $\mathbf{z} \in \mathbb{R}^n$ with the z_i component removed. The value of the integrals are independent of i, and

$$\sum_{i=1}^{n} g_{ii}^{'} = \operatorname{tr}(G') = \operatorname{tr}(A^{-\frac{1}{2}} G A^{-\frac{1}{2}}) = \operatorname{tr}(G A^{-1}),$$

and so, (2.33) follows.

2.3 The Volume of Spheres and Balls in $\mathbb{R}^n$

The volume of the $(n-1)$-dimensional hyper-sphere with unit radius, $S^{n-1} \subset \mathbb{R}^n$, and of the open ball $B^n \subset \mathbb{R}^n$ enclosed by S^{n-1} appear in a number of geometric and statistical applications. The argument used here for computing these volumes follows that given in [12]. Before proceeding, a note is in order regarding the use of the word "volume." In the case of $n = 3$, the "volume" of the sphere S^2 is its surface area, and in the case of $n = 2$, the "volume" of the circle S^1 is its perimeter. In contrast, the "volume" of the ball B^2 is the area on the interior of a circle, and the "volume" of B^3 is the classical volume in $\mathbb{R}^3$ bounded by the sphere S^2. In general, the volume of an n-dimensional manifold will be an n-dimensional measurement.

Consider the isotropic Gaussian distribution on $\mathbb{R}^n$ with zero mean written as

$$\rho(\mathbf{x}; \boldsymbol{\mu} = \mathbf{0}, \Sigma = \sigma^2 I) = \frac{1}{(2\pi)^{n/2}\sigma^n} \exp(-\frac{1}{2}\|\mathbf{x}\|^2/\sigma^2).$$

If $\mathbf{x} = r\mathbf{u}$ where r and $\mathbf{u} = \mathbf{u}(\phi_1, \phi_2, ..., \phi_{n-1})$ represent "hyper-spherical" coordinates, then the Jacobian determinant relates the change from Cartesian coordinates as

$$d\mathbf{x} = \left|\det\left[\frac{\partial \mathbf{x}}{\partial r}, \frac{\partial \mathbf{x}}{\partial \phi_1}, \ldots, \frac{\partial \mathbf{x}}{\partial \phi_{n-1}}\right]\right| dr d\phi_1 \cdots d\phi_{n-1} = dV(\phi) r^{n-1} dr$$

where $dV(\phi)$ is the volume element for the sphere S^{n-1}. The volume of S^{n-1} is then

$$Vol(S^{n-1}) = \int_{S^{n-1}} dV(\phi).$$

This can be computed directly by extending the usual spherical coordinates to higher dimensions in the natural way as

$$\mathbf{u}^{(2)} = \begin{pmatrix} \cos\phi_1 \\ \sin\phi_1 \end{pmatrix}; \quad \mathbf{u}^{(3)} = \begin{pmatrix} \cos\phi_1 \sin\phi_2 \\ \sin\phi_1 \sin\phi_2 \\ \cos\phi_2 \end{pmatrix}; \quad \mathbf{u}^{(3)} = \begin{pmatrix} \cos\phi_1 \sin\phi_2 \sin\phi_3 \\ \sin\phi_1 \sin\phi_2 \sin\phi_3 \\ \cos\phi_2 \sin\phi_3 \\ \cos\phi_3 \end{pmatrix}; \quad \text{etc.},$$

computing Jacobian determinants for each case, and then integrating over the appropriate range of angles, $0 \leq \phi_1 < 2\pi$ and $0 \leq \phi_i < \pi$ for $1 < i \leq n-1$. Or, the volume of the unit sphere can be calculated indirectly, as it is done below.

From the fact that ρ is a pdf,

$$\begin{aligned} 1 &= \int_{\mathbb{R}^n} \rho(\mathbf{x}; \mathbf{0}, \sigma^2 I) d\mathbf{x} \\ &= \int_0^\infty \int_{S^{n-1}} \rho(r\mathbf{u}; \mathbf{0}, \sigma^2 I) dV(u) r^{n-1} dr \\ &= \frac{1}{(2\pi)^{n/2}\sigma^n} \left(\int_0^\infty \exp(-r^2/(2\sigma^2)) r^{n-1} dr \right) Vol(S^{n-1}). \end{aligned}$$

Therefore, it must be that

$$\frac{1}{(2\pi)^{n/2}\sigma^n} \int_0^\infty \exp(-r^2/(2\sigma^2)) r^{n-1} dr = 1/Vol(S^{n-1})$$

for any value of σ. Letting $s = r/(\sqrt{2}\sigma)$, the integral on the left becomes

$$\int_0^\infty \exp(-r^2/(2\sigma^2)) r^{n-1} dr = 2^{n/2}\sigma^n \int_0^\infty \exp(-s^2) s^{n-1} ds = \frac{1}{2} 2^{n/2} \sigma^n \Gamma(n/2).$$

This can be taken as the definition of the *Gamma function*, or it can be viewed as the result of the change of coordinates $t = s^2$ from the more standard definition

$$\Gamma(\alpha) = \int_0^\infty e^{-t} t^{\alpha-1} dt \tag{2.34}$$

with $\alpha = n/2$.

In any case, since the Gaussian pdf integrates to unity, the factors of $2^{n/2}\sigma^n$ cancel, and it must be that $\frac{1}{2}\Gamma(n/2) Vol(S^{n-1}) = (\pi)^{n/2}$, or

$$Vol(S^{n-1}) = \frac{2(\pi)^{n/2}}{\Gamma\left(\frac{n}{2}\right)}. \tag{2.35}$$

This is the volume of a unit hyper-sphere $S^{n-1} \subset \mathbb{R}^n$. The volume of a hyper-sphere of radius r would be r^{n-1} times this quantity. The volume of the unit ball $B^n \subset \mathbb{R}^n$ is then obtained by integrating over all of these spherical shells as

$$Vol(B^n) = \int_0^1 Vol(S^{n-1}) r^{n-1} dr = \frac{2(\pi)^{n/2}}{\Gamma\left(\frac{n}{2}\right)} \int_0^1 r^{n-1} dr.$$

In other words,

$$Vol(B^n) = \frac{2(\pi)^{n/2}}{n \cdot \Gamma\left(\frac{n}{2}\right)} = \frac{(\pi)^{n/2}}{\Gamma\left(\frac{n}{2}+1\right)}. \tag{2.36}$$

The first few values of $\Gamma(n/2)$ are given in the following table:

Table 2.1. The First Few Half-Integer Values of the Γ-Function

n	$\Gamma(n/2)$
1	$\sqrt{\pi}$
2	1
3	$\sqrt{\pi}/2$
4	1
5	$3\sqrt{\pi}/4$
6	2

Note that for integer arguments, $\Gamma(m) = (m-1)!$.

The shorthand notation

$$Vol(S^{n-1}) = \mathcal{O}_n \quad \text{and} \quad Vol(B^n) = \frac{\mathcal{O}_n}{n} \tag{2.37}$$

will be useful.

2.4 Clipped Gaussian Distributions

The Gaussian distribution has many interesting and useful properties. For example, it is the maximum entropy distribution of given mean and covariance, it satisfies a diffusion equation, as a family of parametric distributions it is closed under the operations of convolution and conditioning. In addition, its higher moments can be computed as closed-form integrals. It would be useful to take advantage of these properties when fitting a density to measured data on other domains such as spheres. However, a problem that immediately arises is that for compact domains, something must be done with the infinite tails of the Gaussian distribution. Two options are to wrap the tails around (resulting in a "folded" Gaussian), or to clip the tails. The folded Gaussian for the circle is discussed in Section 2.5. While this is a viable option in some cases, a more general procedure that can be used for other finite domains is clipping. In the subsections that follow, the properties of the univariate clipped Gaussian are obtained, and extended to the multi-dimensional case.

2.4.1 One-Dimensional Clipped Gaussian Distributions

Suppose that we want to clip the Gaussian distribution with mean at $x = 0$ defined by

$$\rho(x; 0, \sigma_0) = \frac{1}{\sqrt{2\pi}\sigma_0} e^{-x^2/2\sigma_0^2}.$$

This is defined on the real line. By restricting it to the unit circle, which we identify with the interval $[-\pi, \pi]$, the mass is reduced from unity to

$$r(\sigma_0) \doteq \int_{-\pi}^{\pi} \rho(x;0,\sigma_0)dx < 1. \tag{2.38}$$

An exact expression for $r(\sigma_0)$ can be found in terms of the *error function*

$$\operatorname{erf}(x) \doteq \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt. \tag{2.39}$$

However, if $k\sigma_0 < \pi$ for $k \geq 3$, then $r(\sigma_0) \approx 1$ is a good approximation.

The variance of a clipped Gaussian is then

$$\sigma^2 = \frac{1}{\sqrt{2\pi}\sigma_0 r(\sigma_0)} \int_{-\pi}^{\pi} x^2 e^{-x^2/2\sigma_0^2} dx = \frac{\sigma_0^2}{\sqrt{2\pi} r(\sigma_0)} \int_{-\pi/\sigma_0}^{\pi/\sigma_0} y^2 e^{-y^2/2} dy.$$

This can be written as

$$\sigma^2 = \frac{\sigma_0^2}{\sqrt{2\pi} r(\sigma_0)} \left[\sqrt{2\pi} - \frac{2\pi}{\sigma_0} e^{-\pi^2/(2\sigma_0^2)} \right]$$

by using integration by parts. As $\sigma_0 \to 0$, then $\sigma \to \sigma_0$.

2.4.2 Multi-Dimensional Clipped Gaussian Distributions

The integral of a multi-dimensional Gaussian distribution over the interior of an ellipsoid defined by

$$\mathbf{x}^T \Sigma_0^{-1} \mathbf{x} = a^2$$

can be computed in closed form (using error integrals). We can therefore clip a multi-dimensional Gaussian distribution along the boundary of such an ellipsoid and renormalize the resulting distribution so as to be a pdf. In other words, a clipped Gaussian is defined relative to a Gaussian as

$$\rho_c(\mathbf{x}, \Sigma_0, a) \doteq \begin{cases} \rho(\mathbf{x}, \Sigma_0)/r(\Sigma_0, a) & \text{for } \mathbf{x}^T \Sigma_0^{-1} \mathbf{x} < a^2 \\ 0 & \text{otherwise} \end{cases} \tag{2.40}$$

where

$$r(\Sigma_0, a) \doteq \int_{\mathbf{x}^T \Sigma_0^{-1} \mathbf{x} < a^2} \rho(\mathbf{x}, \Sigma_0)\, d\mathbf{x}.$$

The covariance of a clipped Gaussian is then

$$\Sigma = \int_{\mathbf{x}^T \Sigma_0^{-1} \mathbf{x} < a^2} \mathbf{x}\mathbf{x}^T \rho_c(\mathbf{x}, \Sigma_0, a)\, d\mathbf{x}. \tag{2.41}$$

By making the change of variables $\mathbf{y} = \Sigma_0^{-\frac{1}{2}} \mathbf{x}$, it follows that

$$\Sigma = \Sigma_0^{\frac{1}{2}} \left[\int_{\mathbf{y}^T \mathbf{y} < a^2} \mathbf{y}\mathbf{y}^T \rho_c(\mathbf{y}, I, a)\, d\mathbf{y} \right] \Sigma_0^{\frac{1}{2}}. \tag{2.42}$$

The above integral can be computed in closed form. This is done below for the three-dimensional case. The two-dimensional case is left as an exercise.

It will be convenient to define

$$f_0(a) \doteq \int_0^a e^{-r^2/2} dr = \sqrt{\frac{\pi}{2}} \mathrm{erf}(a/\sqrt{2})$$

$$f_1(a) \doteq \int_0^a r^2 e^{-r^2/2} dr = -ae^{-a^2/2} + f_0(a)$$

and

$$f_2(a) \doteq \int_0^a r^4 e^{-r^2/2} dr = 3f_1(a) - a^3 e^{-a^2/2}.$$

Then

$$m(\Sigma_0, a) \doteq \int_{\mathbf{x}^T \Sigma_0^{-1} \mathbf{x} < a^2} \exp\{-\mathbf{x}^T \Sigma_0^{-1} \mathbf{x}\} d\mathbf{x} = 4\pi f_1(a) \cdot |\Sigma_0|^{\frac{1}{2}}$$

and

$$r(\Sigma_0, a) = m(\Sigma_0, a)/(2\pi)^{\frac{3}{2}} |\Sigma_0|^{\frac{1}{2}} = \sqrt{\frac{2}{\pi}} f_1(a).$$

Using spherical coordinates,

$$\mathbf{y} = \begin{pmatrix} r \sin\theta \cos\phi \\ r \sin\theta \sin\phi \\ r \cos\theta \end{pmatrix},$$

$$\int_{\mathbf{y}^T \mathbf{y} < a^2} \mathbf{y}\mathbf{y}^T \rho_c(\mathbf{y}, I, a)\, d\mathbf{y} = \int_{r=0}^{a} \int_{\phi 0}^{2\pi} \int_{\theta=0}^{\pi} \mathbf{y}\mathbf{y}^T \rho_c(\mathbf{y}, I, a)\, r^2 dr d\phi d\theta = \sqrt{\frac{2}{\pi}} \frac{f_2(a)}{3} I$$

where

$$f_2(a) = \int_0^a r^4 e^{-r^2/2} dr.$$

This can be computed in closed form using integration by parts. Therefore (2.42) reduces to

$$\Sigma = \frac{f_2(a)}{3 \cdot f_1(a)} \Sigma_0. \tag{2.43}$$

As $a \to \infty$, $\Sigma \to \Sigma_0$.

2.5 Folded, or Wrapped, Gaussians

In some applications, data on the circle is given, and a corresponding concept of Gaussian distribution is needed. One approach that was discussed in the previous section that could be applied to this end is to "clip the tails" of a Gaussian outside of the range of values $\theta \in [-\pi, \pi]$ and renormalize the result in order to make it a valid pdf. In contrast, the tails can be "wrapped around" the circle as

$$\rho_W(\theta; \mu, \sigma) \doteq \sum_{k=-\infty}^{\infty} \rho(\theta - 2\pi k; \mu, \sigma), \tag{2.44}$$

where if μ is outside of the range $[-\pi, \pi]$, it can be "put back in the range" by subtracting $2\pi N$ from it for some $N \in \mathbb{Z}$ until it is in range.

If σ is very small and $\mu = 0$, only the $k = 0$ term in the above sum needs to be retained, and there is no distinction between the original Gaussian restricted to the range $\theta \in [-\pi, \pi]$, the Gaussian clipped to this range, and the folded Gaussian. But as

σ increases, so too do the values of $|k|$ that need to be retained. As σ becomes very large, it becomes impractical to compute (2.44).

However, there is an alternative representation of the folded Gaussian that uses the fact that it is a periodic function. Recall that any 2π-periodic function, i.e., a "function on the unit circle," can be expanded in a Fourier series:

$$f(\theta) = \frac{1}{2\pi}\sum_{n=-\infty}^{\infty} \hat{f}(n)e^{in\theta} \quad \text{where} \quad \hat{f}(n) = \int_0^{2\pi} f(\theta)e^{-in\theta}d\theta, \tag{2.45}$$

where $e^{in\theta} = \cos n\theta + i\sin n\theta$ and $i = \sqrt{-1}$. Here $\hat{f}(n)$ are called the *Fourier coefficients*, or circular Fourier transform. These coefficients can be computed in closed form for (2.44). This leads to the Fourier series representation of the folded Gaussian distribution:

$$\boxed{\rho_W(\theta;\mu,\sigma) = \frac{1}{2\pi} + \frac{1}{\pi}\sum_{n=1}^{\infty} e^{-\frac{\sigma^2}{2}n^2}\cos\left(n(\theta-\mu)\right).} \tag{2.46}$$

As σ becomes large, very close approximations can be achieved with the first couple of terms in the summation in (2.46). In contrast, as σ becomes very small, using very few of the terms in the series (2.44) will produce a very good approximation when $\mu = 0$.

The general theme that a Gaussian on a space other than the real line can be approximated well as a Gaussian restricted to a smaller domain when σ is small, or as a generalized Fourier series expansion when σ is large, will recur many times throughout this book.

Note that the above "folding" process is not restricted to Gaussian distributions; any well-behaved function, $f(x)$, defined on the line can be wrapped around the circle. The resulting folded function, which is 2π-periodic, is related to the Fourier transform of the original non-periodic function on the real line through the *Poisson summation formula* [1]:

$$\sum_{n=-\infty}^{\infty} f(\theta + 2\pi n) = \frac{1}{2\pi}\sum_{k=-\infty}^{\infty} [\mathcal{F}(f)](k)e^{ik\theta}. \tag{2.47}$$

In other words, the Fourier coefficients of the folded function are related to the Fourier transform of the original function as

$$\hat{f}(k) = [\mathcal{F}(f)](k).$$

2.6 The Heat Equation

In this section, the relationship between the Gaussian distribution and the heat equation (also called the diffusion equation) is developed.

Sometimes the exact solution of an equation is not as critical as knowing how its mean and covariance behave as a function of time. This is illustrated both in the one-dimensional and multi-dimensional settings in the following subsections.

2.6.1 The One-Dimensional Case

Consider the diffusion equation on the real line with time-varying diffusion and drift coefficients, $k(t)$ and $a(t)$:

$$\frac{\partial f}{\partial t} = \frac{1}{2}k(t)\frac{\partial^2 f}{\partial x^2} - a(t)\frac{\partial f}{\partial x}. \tag{2.48}$$

The initial condition is $f(x, 0) = \delta(x)$. The solution $f(x, t)$ can be obtained in closed form, following essentially the same procedure as in Section 2.1.5, and then the mean and variance can be computed from this solution as

$$\mu(t) = \int_{-\infty}^{\infty} x f(x, t) dx \quad \text{and} \quad \sigma^2(t) = \int_{-\infty}^{\infty} [x - \mu(t)]^2 f(x, t) dx. \tag{2.49}$$

Alternatively, the mean and variance of $f(x, t)$ can be computed directly from (2.48) without actually knowing the solution $f(x, t)$. In fact, many properties of $f(x, t)$ can be determined from (2.48) and the corresponding initial conditions without knowing $f(x, t)$. For example, integrating both sides of (2.48) with respect to x yields

$$\frac{d}{dt}\int_{-\infty}^{\infty} f(x, t) dx = 0.$$

This follows because

$$\int_{-\infty}^{\infty} \frac{\partial f}{\partial x} dx = f(x, t)\Big|_{x=-\infty}^{\infty} \quad \text{and} \quad \int_{-\infty}^{\infty} \frac{\partial^2 f}{\partial x^2} dx = \frac{\partial f}{\partial x}\Big|_{x=-\infty}^{\infty}$$

and under the boundary conditions that $f(x, t)$ and $\partial f/\partial x$ decay rapidly to zero as $x \to \pm\infty$, these terms become zero. Since the initial conditions are a delta function in x, it follows that

$$\int_{-\infty}^{\infty} f(x, t) dx = 1.$$

In other words, (2.48) preserves the initial mass of the distribution over all values of time after $t = 0$.

To compute $\mu(t)$, multiply both sides of (2.48) by x and integrate. On the one hand,

$$\int_{-\infty}^{\infty} x \frac{\partial f}{\partial t} dx = \frac{d}{dt}\int_{-\infty}^{\infty} x f(x, t) dx = \frac{d\mu}{dt}.$$

On the other hand,

$$\int_{-\infty}^{\infty} x \frac{\partial f}{\partial t} dx = \frac{1}{2}k(t)\int_{-\infty}^{\infty} x \frac{\partial^2 f}{\partial x^2} dx - a(t)\int_{-\infty}^{\infty} x \frac{\partial f}{\partial x} dx.$$

Evaluating both integrals on the right side by integrating by parts and using the conditions that both $f(x, t)$ and $\partial f/\partial x$ decay rapidly to zero as $x \to \pm\infty$, it becomes clear that

$$\frac{d\mu}{dt} = a(t) \quad \text{or} \quad \mu(t) = \int_0^t a(s) ds. \tag{2.50}$$

A similar argument shows that

$$\frac{d}{dt}(\sigma^2) = k(t) \quad \text{or} \quad \sigma^2(t) = \int_0^t k(s) ds. \tag{2.51}$$

2.6.2 The Multi-Dimensional Case

Consider the following time-varying diffusion equation without drift:

$$\frac{\partial f}{\partial t} = \frac{1}{2} \sum_{i,j=1}^{n} D_{ij}(t) \frac{\partial^2 f}{\partial x_i \partial x_j}, \tag{2.52}$$

where $D_{ij}(t) = D_{ji}(t)$ are the *time-varying diffusion constants*. If $f(\mathbf{x}, 0) = \delta(\mathbf{x})$, then integrating (2.52) both sides over $\mathbb{R}^n$ and using integration by parts in $\mathbf{x}$ shows that the unit volume under the curve is preserved.

Multiplying both sides by $x_k x_l$ and integrating over $\mathbf{x} \in \mathbb{R}^n$ gives

$$\frac{d}{dt}(\sigma_{kl}) = \frac{1}{2} \sum_{i,j=1}^{n} D_{ij}(t) \int_{\mathbb{R}^n} x_k x_l \frac{\partial^2 f}{\partial x_i \partial x_j} d\mathbf{x}. \tag{2.53}$$

Let the integral over $\mathbb{R}^{n-1}$ resulting from the exclusion of the integral over x_i be denoted as

$$\int_{\mathbf{x}-x_i} f(\mathbf{x}) d\mathbf{x}/dx_i =$$

$$\int_{x_1=-\infty}^{\infty} \cdots \int_{x_{i-1}=-\infty}^{\infty} \int_{x_{i+1}=-\infty}^{\infty} \cdots \int_{x_n=-\infty}^{\infty} f(x_1, ..., x_n) dx_1 \cdots dx_{i-1} dx_{i+1} \cdots dx_n$$

so that

$$\int_{\mathbb{R}^n} f(\mathbf{x}) d\mathbf{x} = \int_{-\infty}^{\infty} \left(\int_{\mathbf{x}-x_i} f(\mathbf{x}) d\mathbf{x}/dx_i \right) dx_i = \int_{\mathbf{x}-x_i} \left(\int_{-\infty}^{\infty} f(\mathbf{x}) dx_i \right) d\mathbf{x}/dx_i.$$

An integral over $n-2$ degrees of freedom denoted by the integral with subscript $\mathbf{x}-x_i-x_j$ follows in a similar way.

From integration by parts

$$\int_{\mathbb{R}^n} x_k x_l \frac{\partial^2 f}{\partial x_i \partial x_j} d\mathbf{x} = \int_{\mathbf{x}-x_i} \left[x_k x_l \left. \frac{\partial f}{\partial x_j} \right|_{x_i=-\infty}^{\infty} - \int_{-\infty}^{\infty} \frac{\partial}{\partial x_i}(x_k x_l) \frac{\partial f}{\partial x_j} dx_i \right] d\mathbf{x}/dx_i.$$

The assumption that $f(\mathbf{x}, t)$ decays rapidly as $\|\mathbf{x}\| \to \infty$ for all values of t makes the first term in the brackets disappear. Using the fact that $\partial x_i / \partial x_j = \delta_{ij}$, and integrating by parts again (over x_j) reduces the above integral to

$$\int_{\mathbb{R}^n} x_k x_l \frac{\partial^2 f}{\partial x_i \partial x_j} d\mathbf{x} = \delta_{kj}\delta_{il} + \delta_{ik}\delta_{lj}.$$

Substituting this into (2.53) results in

$$\frac{d}{dt}(\sigma_{kl}) = D_{kl}(t) \quad \text{or} \quad \sigma_{kl}(t) = \int_0^t D_{kl}(s) ds. \tag{2.54}$$

Therefore, even without knowing the form of the time-varying pdf that solves (2.52) it is possible to obtain an exact expression for the covariance of the solution.

2.6.3 The Heat Equation on the Unit Circle

The heat equation on the circle is exactly the same as the heat equation on the real line (with θ replacing x as the spatial variable). However, the topological constraint that $\theta = \pm\pi$ represents the same point means that the long-time solution will be completely different than in the unconstrained case on the real line. Whereas the Fourier transform can be used to solve the heat equation on the line, the Fourier series expansion is used on the circle.

The result is that the solution on the line can be folded around the circle. In other words, the solution to the *heat equation on the circle* for constant diffusion coefficient k,

$$\frac{\partial f}{\partial t} = \frac{1}{2}k\frac{\partial^2 f}{\partial \theta^2} \quad \text{subject} \quad \text{to} \quad f(\theta, 0) = \delta(\theta),$$

is

$$f(\theta,t) = \sum_{k=-\infty}^{\infty} \rho(\theta - 2\pi k; 0, (kt)^{\frac{1}{2}}) = \frac{1}{2\pi} + \frac{1}{\pi}\sum_{n=1}^{\infty} e^{-ktn^2/2}\cos n\theta. \tag{2.55}$$

This is the folded Gaussian in (2.46) with $\sigma^2 = kt$ and $\mu = 0$.

2.7 Gaussians and Multi-Dimensional Diffusions

In the previous section, the evolution of the mean and covariance of a diffusion equation was obtained without knowing the time-varying pdf. Here, the pdf is sought.

2.7.1 The Constant Diffusion Case

Consider the diffusion equation

$$\frac{\partial f}{\partial t} = \frac{1}{2}\sum_{i,j=1}^{n} D_{ij}\frac{\partial f^2}{\partial x_i \partial x_j} \tag{2.56}$$

subject to the initial conditions $f(\mathbf{x}, t) = \delta(\mathbf{x})$, where $D = [D_{ij}] = D^T$ is a constant matrix of diffusion constants.

Since diffusion equations preserve mass (see Section 2.6.2), it follows that

$$\int_{\mathbb{R}^n} f(\mathbf{x}, t)d\mathbf{x} = 1 \tag{2.57}$$

for all values of time, $t \in \mathbb{R}_{>0}$.

Try a solution of the form

$$f(\mathbf{x}, t) = c(t)\exp(-\frac{1}{2}\mathbf{x}^T A(t)\mathbf{x}) \tag{2.58}$$

where $A(t) = \phi(t)A_0$ and $A_0 = [\alpha_{ij}] = A_0^T$. Then, from (2.57) and the formula (2.83) derived in the exercises, it follows that

$$c(t) = \left(\frac{\phi(t)}{2\pi}\right)^{n/2} |\det A_0|^{\frac{1}{2}}.$$

With this constraint in mind, substituting $f(\mathbf{x},t)$ into (2.56) produces the following conditions on $\phi(t)$ and A_0:

$$n\phi' = -\phi^2 \sum_{i,j=1}^{n} D_{ij}\alpha_{ij}$$

$$\phi'\mathbf{x}^T A_0 \mathbf{x} = -\phi^2 \sum_{i,j=1}^{n} D_{ij} \left(\sum_{k=1}^{n} \alpha_{ik}x_k\right)\left(\sum_{l=1}^{n} \alpha_{jl}x_l\right)$$

where $\phi' = d\phi/dt$.

Both of the conditions (2.59) are satisfied if $A_0 = \alpha_0 D^{-1}$ and $\phi(t) = (\alpha_0 t)^{-1}$ for some arbitrary constant $\alpha_0 \in \mathbb{R}_{>0}$. But since $A(t) = \phi(t)A_0 = t^{-1}D^{-1}$, this constant does not matter.

Putting all of this together,

$$\boxed{f(\mathbf{x},t) = \frac{1}{(2\pi t)^{n/2}|\det D|^{\frac{1}{2}}} \exp(-\frac{1}{2t}\mathbf{x}^T D^{-1}\mathbf{x}).} \tag{2.59}$$

Stated in another way, the solution to (2.56) is a time-varying Gaussian distribution with $\Sigma(t) = tD$ when D is symmetric.

2.7.2 The Time-Varying Case

Consider again (2.56), but now let $D = D(t)$. Try a solution of the form

$$f(\mathbf{x},t) = (2\pi)^{-n/2}|\det \Sigma(t)|^{-\frac{1}{2}} \exp(-\frac{1}{2}\mathbf{x}^T\Sigma^{-1}(t)\mathbf{x}) \tag{2.60}$$

where $\Sigma(t)$ is a time-varying covariance matrix, the form of which is as yet undetermined. This guess is simply $f(\mathbf{x},t) = \rho(\mathbf{x};\mathbf{0},\Sigma(t))$.

The derivatives with respect to x_i are evaluated as before, using the chain rule. The time derivative is evaluated as follows:

$$\begin{aligned}\frac{\partial f}{\partial t} &= (2\pi)^{-n/2}\frac{d(|\det \Sigma|^{-\frac{1}{2}})}{dt}\exp(-\frac{1}{2}\mathbf{x}^T\Sigma^{-1}\mathbf{x}) \\ &\quad +(2\pi)^{-n/2}|\det \Sigma|^{-\frac{1}{2}}\frac{d}{dt}\left[\exp(-\frac{1}{2}\mathbf{x}^T\Sigma^{-1}\mathbf{x})\right] \\ &= -\frac{1}{2}(2\pi)^{-n/2}|\det \Sigma|^{-\frac{3}{2}}\frac{d(\det \Sigma)}{dt}\exp(-\frac{1}{2}\mathbf{x}^T\Sigma^{-1}\mathbf{x}) \\ &\quad -\frac{1}{2}(2\pi)^{-n/2}|\det \Sigma|^{-\frac{1}{2}}\left(\mathbf{x}^T\frac{d}{dt}[\Sigma^{-1}]\mathbf{x}\right)\exp\left(-\frac{1}{2}\mathbf{x}^T\Sigma^{-1}\mathbf{x}\right).\end{aligned}$$

On the other hand,

$$\frac{1}{2}\sum_{i,j=1}^{n} D_{ij}\frac{\partial^2 f}{\partial x_i \partial x_j} = \frac{1}{2}\left\{-\mathrm{tr}(D\Sigma^{-1}) + \mathbf{x}^T(\Sigma^{-1}D\Sigma^{-1})\mathbf{x}\right\} f(\mathbf{x},t).$$

Therefore, if

$$|\det \Sigma|^{-1}\frac{d(\det \Sigma)}{dt} = \mathrm{tr}(D\Sigma^{-1}) \quad \text{and} \quad \frac{d}{dt}[\Sigma^{-1}] = -\Sigma^{-1}D\Sigma^{-1}, \tag{2.61}$$

then (2.56) with variable diffusion coefficients will be satisfied. Since[8]

$$\frac{d}{dt}(\Sigma\Sigma^{-1}) = \mathbb{O} \quad \Longrightarrow \quad \frac{d}{dt}[\Sigma^{-1}] = -\Sigma^{-1}\dot{\Sigma}\Sigma^{-1},$$

the second equality in (2.61) will be satisfied if $D = \dot{\Sigma}$. In this case the first equality in (2.61) becomes

$$\frac{d}{dt}\log(\det\Sigma) = \mathrm{tr}(\dot{\Sigma}\Sigma^{-1}). \tag{2.62}$$

Under what conditions will this be true?

Case 1:

From Systems Theory (as reviewed in the appendix), if $\Sigma = \exp(tS_0)$ where $S_0 = S_0^T$ is constant, then

$$\det\Sigma = e^{\mathrm{tr}(tS_0)} = e^{t(\mathrm{tr}S_0)}.$$

Therefore, in this special case

$$\frac{d}{dt}\log(\det\Sigma) = \mathrm{tr}(S_0).$$

Likewise, if $\Sigma = \exp(tS_0)$, then $\mathrm{tr}(\dot{\Sigma}\Sigma^{-1}) = \mathrm{tr}(S_0)$. Therefore, it can be concluded that a sufficient condition for the Gaussian in (2.60) to be a solution to (2.56) is if a constant symmetric matrix S_0 can be found such that $D(t) = S_0\exp(tS_0)$.

Case 2:

The condition in (2.62) will be satisfied if $\Sigma = \sigma(t)\Sigma_0$ where $\sigma(t)$ is a differentiable scalar function of time and $\Sigma_0 = \Sigma_0^T$. Substitution into (2.62) yields the condition

$$\frac{d}{dt}\log(\sigma^n \det\Sigma_0) = \dot{\sigma}\sigma^{-1}\mathrm{tr}(\mathbb{I}).$$

Since $\log(a\cdot b) = \log a + \log b$, and $\frac{d}{dt}\log a(t) = \dot{a}/a$, the above condition becomes

$$\frac{1}{\sigma^n}n\sigma^{n-1}\dot{\sigma} = n\dot{\sigma}\sigma^{-1},$$

which is always true. Therefore any $\sigma(t)$ will work.

A broader condition that encompasses both Case 1 and Case 2 is $D(t) = \dot{S}(t)\exp S(t)$ where $S = S^T$ and $[\dot{S}, S] \doteq \dot{S}S - S\dot{S} = \mathbb{O}$.

Under this condition,

$$\Sigma(t) = \int_0^t D(s)ds. \tag{2.63}$$

2.8 Symmetry Analysis of Evolution Equations

The concept of symmetry can have several meanings when applied to *evolution equations*.[9] For example, the diffusion matrix in the multi-dimensional heat equation might have symmetries in it other than the primary symmetry $D = D^T$. That kind of symmetry is reflected in the solution of the equation. Another kind of symmetry is that the equation itself can be solved when the independent variables undergo a non-linear change of coordinates. Both of these concepts of symmetry are addressed in this section.

[8] Here $\mathbb{O} = \frac{d}{dt}(\mathbb{I})$ is the zero matrix.

[9] These are equations with a single partial derivative in time, and multiple partial derivatives in space. They include, but are not limited to, diffusion equations.

2.8.1 Symmetries in Parameters

Consider a drift-free diffusion in $\mathbb{R}^n$ with constant diffusion matrix $D = D^T$, and let the solution be denoted as $f(\mathbf{x}, t; D)$. Since the dependence on D and t always appears as their product, the solution has a continuous scale symmetry of the form

$$f(\mathbf{x}, t; D) = f(\mathbf{x}, t/\alpha; \alpha D)$$

for any $\alpha \in \mathbb{R}_{>0}$.

In addition, since the solution is the Gaussian distribution in (2.59), it can be verified that

$$f(\mathbf{x}, t; D) = \beta^{n/2} f(\sqrt{\beta}\mathbf{x}, t; \beta D).$$

If $D = \sigma^2 I$, then any change of spatial coordinates of the form $\mathbf{y} = Q\mathbf{x}$ where $Q^T Q = I$ will preserve the solution:

$$f(Q\mathbf{x}, t; \sigma^2 I) = f(\mathbf{x}, t; \sigma^2 I).$$

In contrast, if $n = 3$ and $D = \text{diag}[\sigma_1^2, \sigma_1^2, \sigma_3^2]$ is the diagonal matrix with the indicated entries on the diagonal, then

$$R_3(\theta)^T D R_3(\theta) = D \quad \text{where} \quad R_3(\theta) = \begin{pmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

and so

$$f(R_3(\theta)\mathbf{x}, t; D) = f(\mathbf{x}, t; D).$$

These symmetries all involve simple transformations of the coordinates. Less obvious symmetries result by examining operators which, when applied to the equation of interest, leave it invariant in a sense that will be made precise.

2.8.2 Infinitesimal Symmetry Operators of the Heat Equation

Let $Qf = 0$ denote any partial differential equation, where Q is a differential operator in temporal and spatial variables $(t, \mathbf{x}) \in \mathbb{R}_{\geq 0} \times \mathbb{R}^n$. For example, for the heat equation on the real line where there is only one spatial variable $(t, \mathbf{x})$ becomes (t, x) and

$$Q = \frac{\partial}{\partial t} - \frac{\partial^2}{\partial x^2}$$

where the diffusion constant, k, is chosen to be $k = 2$ here for convenience.

A body of literature exists that addresses the question of how to obtain new solutions of $Qf = 0$ from old ones. In particular, if it is possible to find a first-order operator of the form

$$L = T(\mathbf{x}, t)\frac{\partial}{\partial t} + \sum_{i=1}^{n} X_i(\mathbf{x}, t)\frac{\partial}{\partial x_i} + Z(\mathbf{x}, t) \tag{2.64}$$

where $T(\mathbf{x}, t)$, $X_i(\mathbf{x}, t)$, and $Z(\mathbf{x}, t)$ are analytic functions such that

$$[L, Q]f(\mathbf{x}, t) = R(\mathbf{x}, t)Qf \quad \text{where} \quad [L, Q] = LQ - QL, \tag{2.65}$$

then $f' \doteq Lf$ will solve $Qf' = 0$.

At first this might seem surprising, but since the condition in (2.64) reads $LQf - QLf = RQf$, and since $Qf = 0$, it must be that $0 = QLf = Q(Lf) = Qf'$.

Following [2, 3, 4, 17, 19, 21], the infinitesimal operators that transform solutions of the heat equation into new solutions are presented below. In this case there is one spatial dimension and so

$$L = T(x,t)\frac{\partial}{\partial t} + X(x,t)\frac{\partial}{\partial x} + Z(x,t).$$

Some mundane calculus yields

$$\begin{aligned} QLf &= \left(\frac{\partial}{\partial t} - \frac{\partial^2}{\partial x^2}\right)\left(T(x,t)\frac{\partial}{\partial t} + X(x,t)\frac{\partial}{\partial x} + Z(x,t)\right) \\ &= \left(\frac{\partial T}{\partial t}\right)\left(\frac{\partial f}{\partial t}\right) + T\left(\frac{\partial^2 f}{\partial t^2}\right) + \left(\frac{\partial X}{\partial t}\right)\left(\frac{\partial f}{\partial x}\right) + X\left(\frac{\partial^2 f}{\partial t \partial x}\right) + \left(\frac{\partial Z}{\partial t}\right) f \\ &\quad + Z\left(\frac{\partial f}{\partial t}\right) - \left(\frac{\partial^2 T}{\partial x^2}\right)\left(\frac{\partial f}{\partial t}\right) - 2\left(\frac{\partial T}{\partial x}\right)\left(\frac{\partial^2 f}{\partial t \partial x}\right) - T\left(\frac{\partial^3 f}{\partial t \partial x^2}\right) \\ &\quad - \left(\frac{\partial^2 X}{\partial x^2}\right)\left(\frac{\partial f}{\partial x}\right) - 2\left(\frac{\partial X}{\partial x}\right)\left(\frac{\partial^2 f}{\partial x^2}\right) - X\left(\frac{\partial^3 f}{\partial x^3}\right) \\ &\quad - \left(\frac{\partial^2 Z}{\partial x^2}\right) f - 2\left(\frac{\partial Z}{\partial x}\right)\left(\frac{\partial f}{\partial x}\right) - Z\left(\frac{\partial^2 f}{\partial x^2}\right) \end{aligned}$$

and

$$\begin{aligned} LQf &= \left(T(x,t)\frac{\partial}{\partial t} + X(x,t)\frac{\partial}{\partial x} + Z(x,t)\right)\left(\frac{\partial}{\partial t} - \frac{\partial^2}{\partial x^2}\right) f \\ &= T\frac{\partial^2 f}{\partial t^2} - T\frac{\partial^3 f}{\partial x^2 \partial t} + X\frac{\partial^2 f}{\partial x \partial t} - X\frac{\partial^3 f}{\partial x^3} + Z\frac{\partial f}{\partial t} - Z\frac{\partial^2 f}{\partial x^2}. \end{aligned}$$

Note that every term in LQf can also be found in QLf. Subtracting, and reorganizing the terms that result, yields

$$\begin{aligned} [Q,L]f &= \left(\frac{\partial T}{\partial t} - \frac{\partial^2 T}{\partial x^2}\right)\frac{\partial f}{\partial t} + \left(\frac{\partial X}{\partial t} - \frac{\partial^2 X}{\partial x^2} - 2\frac{\partial Z}{\partial x}\right)\frac{\partial f}{\partial x} \\ &\quad + \left(-2\frac{\partial T}{\partial x}\right)\frac{\partial^2 f}{\partial x \partial t} + \left(-2\frac{\partial X}{\partial x}\right)\frac{\partial^2 f}{\partial x^2} + \left(\frac{\partial Z}{\partial t} - \frac{\partial^2 Z}{\partial x^2}\right) f. \end{aligned}$$

Since $[Q,L] = -[L,Q]$, (2.64) is the same as computing $[Q,L]f = -RQf$ where

$$RQf = R\frac{\partial f}{\partial t} - R\frac{\partial^2 f}{\partial x^2}.$$

Then equating the coefficients in front of each term involving f, the following five equations result:

$$\frac{\partial T}{\partial t} - \frac{\partial^2 T}{\partial x^2} = -R \tag{2.66}$$

$$2\frac{\partial X}{\partial x} = -R \tag{2.67}$$

$$\frac{\partial X}{\partial t} - \frac{\partial^2 X}{\partial x^2} - 2\frac{\partial Z}{\partial x} = 0 \tag{2.68}$$

$$\frac{\partial T}{\partial x} = 0 \tag{2.69}$$

$$\frac{\partial Z}{\partial t} - \frac{\partial^2 Z}{\partial x^2} = 0. \tag{2.70}$$

These equations completely determine the structure of the operator L that transforms solutions into solutions.

Starting with (2.69), the restriction $T(x,t) = T(t)$ must be observed. Then, using this result in (2.66) means $-R(x,t) = T'(t)$. This in turn can be substituted into (2.67) to yield

$$X(x,t) = \frac{1}{2}T'(t)x + c_1(t)$$

where $c_1(t)$ is a yet-to-be-determined function resulting from integration over x. Substituting this into (2.68) forces the form of $Z(x,t)$ to be

$$Z(x,t) = \frac{1}{8}T''(t)x^2 + \frac{1}{2}c_1'(t)x + c_2(t).$$

Substituting this into (2.70) forces

$$T'''(t) = 0; \qquad c_1''(t) = 0; \qquad c_2'(t) = \frac{1}{4}T''(t).$$

It follows that

$$T(t) = a_0t^2 + b_0t + c_0; \qquad c_1(t) = \alpha_0 t + \beta_0; \qquad c_2(t) = \frac{1}{2}a_0 t + \gamma_0$$

where $a_0, b_0, c_0, \alpha_0, \beta_0, \gamma_0$ are all free constants.

This means that any L with the following form will map solutions of the heat equation into solutions:

$$\begin{aligned} T(x,t) &= a_0t^2 + b_0t + c_0 \\ X(x,t) &= (a_0t + b_0/2)x + \alpha_0 t + \beta_0 \\ Z(x,t) &= \frac{1}{4}a_0x^2 + \frac{1}{2}\alpha_0 x + \frac{1}{2}a_0 t + \gamma_0. \end{aligned}$$

In fact, the space of all allowable L operators is a vector space with elements of the form

$$L = a_0L_1 + b_0L_2 + c_0L_3 + \alpha_0L_4 + \beta_0L_5 + \gamma_0L_6$$

where the following serves as a basis:

$$L_1 = t^2\frac{\partial}{\partial t} + xt\frac{\partial}{\partial x} + \frac{1}{4}x^2 + \frac{1}{2}t \tag{2.71}$$

$$L_2 = t\frac{\partial}{\partial t} + \frac{1}{2}x\frac{\partial}{\partial x} \tag{2.72}$$

$$L_3 = \frac{\partial}{\partial t} \tag{2.73}$$

$$L_4 = t\frac{\partial}{\partial x} + \frac{1}{2}x \tag{2.74}$$

$$L_5 = \frac{\partial}{\partial x} \tag{2.75}$$

$$L_6 = 1. \tag{2.76}$$

In addition to being a vector space, operators of the form of L given above are also closed under the Lie bracket, $[\cdot,\cdot]$. In other words, $[L_i, L_j]$ for any $i, j \in \{1, ..., 6\}$ will result in a linear combination of these same basis elements. This makes the space of all L operators that map solutions of the heat equation into solutions a *Lie algebra* [16]. This concept will be defined more rigorously in the appendix and in Volume 2.

2.8.3 Non-Linear Transformations of Coordinates

Consider the heat equation

$$\frac{\partial f}{\partial t} = \frac{\partial f^2}{\partial x^2}$$

and assume that an $f(x,t)$ has been obtained that satisfies this equation. For the moment, the initial conditions will be left unspecified.

The following matrices can be defined [17]:

$$B = B(u,v,w) = \begin{pmatrix} 1 & v & 2w + uv/2 \\ 0 & 1 & u \\ 0 & 0 & 1 \end{pmatrix} \quad \text{where} \quad u, v, w \in \mathbb{R} \tag{2.77}$$

and

$$A = A(\alpha,\beta,\gamma,\delta) = \begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix} \quad \text{where} \quad \alpha,\beta,\gamma,\delta \in \mathbb{R} \quad \text{and} \quad \alpha\delta - \beta\gamma = 1. \tag{2.78}$$

It is clear that since $\det A = 1$ by definition, then the product of two such matrices also satisfies this condition: $\det(A_1 A_2) = \det A_1 \det A_2 = 1$. Likewise, the form of the B matrices are preserved under matrix multiplication, and

$$B(u,v,w)B(u',v',w') = B(u+u', v+v', w+w' + (vu' - uv')/4).$$

These are examples of *matrix Lie groups* which, roughly speaking, are groups of continuous transformations, the elements of which are matrices. The group operation is matrix multiplication.

It can be shown (see Exercise 2.18) that transformations of the following form convert solutions into solutions [17]:

$$\boxed{(T_1(B)f)(x,t) = \exp\frac{1}{2}\left[b_{13} + b_{23}x + \frac{1}{2}b_{23}^2 t\right] f(x + b_{12} + b_{23}t, t)} \tag{2.79}$$

and

$$\boxed{(T_2(A)f)(x,t) = \exp\left(-\frac{x^2\beta/4}{\delta + t\beta}\right)(\delta + t\beta)^{-\frac{1}{2}} f\left(\frac{x}{\delta + t\beta}, \frac{\gamma + t\alpha}{\delta + t\beta}\right).} \tag{2.80}$$

In other words, if $f(x,t)$ is a solution to the heat equation, then so too are $f_1(x,t) = (T_1(B)f)(x,t)$ and $f_2(x,t) = (T_2(A)f)(x,t)$. This means that applying these transformations twice with different permissible matrices A_i and B_i will also take solutions into solutions:

$$f_1(x,t) = (T_1(B_2)T_1(B_1)f)(x,t) = (T_1(B_2)(T_1(B_1)f))(x,t)$$

and

$$f_2(x,t) = (T_2(A_2)T_2(A_1)f)(x,t) = (T_2(A_2)(T_2(A_1)f))(x,t).$$

This gets really interesting when these definitions are combined with the closure property under multiplication of matrices of the same kind since

$$T_1(B_2)T_1(B_1) = T_1(B_2B_1) \quad \text{and} \quad T_2(A_2)T_2(A_1) = T_2(A_2A_1). \tag{2.81}$$

What this means is that there are two independent sets of three-parameter transformations that can map solutions into solutions. And furthermore, these can be combined since $(T_2(A)(T_1(B)f))(x,t)$ and $(T_1(B)(T_2(A)f))(x,t)$ must also be solutions.

In Volume 2 this example will be revisited as an example of a six-dimensional Lie group, where the A matrices and B matrices each independently form three-dimensional subgroups.

2.9 Chapter Summary

Many aspects of the Gaussian distribution were reviewed. These include the parametrization of multi-dimensional Gaussians by their mean and covariance, the form of marginals and conditionals of Gaussians, the properties of Gaussians under convolution, the maximum entropy property, and the relationship between Gaussians and diffusion/heat equations.[10] Finally, a brief review of the theory of symmetry analysis of partial differential equations, as applied to diffusion equations, was presented. This forms the first of many links between the topic of diffusion equations and Lie groups that will be forged throughout these books.

The connection between Lie group methods and partial differential equations has a long history dating back to the 1950s [25, 26, 27, 28]. In addition to those references cited earlier in this chapter, significant progress on this topic was made through the 1970s and 1980s including [5, 13, 14, 18, 20]. These approaches have been used for very complicated partial differential equations, such as in [21].

The next chapter will serve as a more formal introduction to probability and information theory. With the concrete example of the Gaussian distribution in mind, it should be easier to tackle these problems. Furthermore, the maximum entropy property of Gaussians, as well as their role in the central limit theorem will justify what might appear to be a preoccupation with Gaussians in the current chapter.

2.10 Exercises

2.1. Verify (2.6) by performing the integrals in the definitions of σ^2 and s.

2.2. Verify (2.10).

2.3. Verify (2.14) by: (a) directly computing the convolution integral in (2.13); (b) using the convolution property of the Fourier transform (2.17).

[10] Note that although these equations were written in Cartesian coordinates in this chapter, it is possible to convert to polar, spherical, or other coordinates. For covariance matrices with symmetry, this can be more convenient. See [8] for a detailed discussion of different curvilinear coordinate systems.

2.4. Using the same reasoning as in Section 2.1.2, compute: (a) the maximum entropy distribution on the real line subject to the constraint that it has a specified value of the spread (rather than variance); (b) the maximum entropy distribution on the finite interval $[a, b]$ subject to no constraints.

2.5. What is the exact expression for the functional $F(\epsilon^2)$ in Section 2.1.2?

2.6. Prove that for any suitable $f_1(x)$ and $f_2(x)$, the convolution theorem (2.17) holds. Hint: a change of variables and a change in the order in which integrals are performed will be required.

2.7. Verify (2.27). Hint: Use the block decomposition in (A.70) to obtain an explicit expression for Σ^{-1}.

2.8. Verify (2.29). Hint: Use the property of the exponential function $e^{a+b} = e^a e^b$.

2.9. Calculate the exact form of $r(\sigma_0)$ in (2.38) in terms of the error function in (2.39).

2.10. Work out the covariance matrix for the 2D clipped Gaussian in analogy with the 3D case presented in (2.43).

2.11. Following the steps in Section 2.1.5, derive the closed-form solution $f(x,t)$ that satisfies (2.48) subject to the initial conditions $f(x,0) = \delta(x)$.

2.12. Using (2.49), show that the mean and variance of $f(x,t)$ computed from (2.48) in (2.50) and (2.51) are the same as computed directly from the closed-form solution of $f(x,t)$ obtained in the previous exercise.

2.13. Verify (2.46) analytically by computing the Fourier coefficients $\hat{\rho}_W(n;\mu,\sigma)$ of (2.44).

2.14. Using integration by parts, prove (2.50) and (2.51).

2.15. Show that the matrices in (2.77) and (2.78) are invertible.

2.16. Find the nine basis operators $\{L_i\}$ that take solutions of

$$\frac{\partial f}{\partial t} = \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} \tag{2.82}$$

into other solutions.

2.17. Find the thirteen basis operators $\{L_i\}$ that take solutions of

$$\frac{\partial f}{\partial t} = \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} + \frac{\partial^2 f}{\partial z^2}$$

into other solutions.

2.18. Verify that transformations of the form in (2.79) and (2.80) will transform one solution into another. Hint: Use the chain rule.

2.19. Verify that the two equations in (2.81) hold. That is, first compute two concatenated transformations, and then compute the single transformation resulting from the matrix products, and compare.

2.20. Show that for $A \in \mathbb{R}^{n\times n}$ with $A = A^T > 0$,

$$\int_{\mathbb{R}^n} \exp(-\frac{1}{2}\mathbf{x}^T A\mathbf{x})d\mathbf{x} = (2\pi)^{n/2}|\det A|^{-\frac{1}{2}}. \tag{2.83}$$

Hint: Decompose $A = Q\Lambda Q^T$ where Q is orthogonal and Λ is the diagonal matrix consisting of eigenvalues of A, which are all positive.

2.21. Verify (2.59) by substituting (2.58) into (2.56) and using the chain rule.

2.22. Can a 2×2 matrix $S(t)$ be constructed such that $\Sigma = \exp S$ which does not fall into Case 1 or Case 2? If so, provide an example. If not, explain why not.

2.23. Determine conditions under which the time-dependent diffusion with drift

$$\frac{\partial f}{\partial t} = \frac{1}{2}\sum_{k,l=1}^{n} D_{kl}(t)\frac{\partial f^2}{\partial x_k \partial x_l} - \sum_{k=1}^{n} d_k(t)\frac{\partial f}{\partial x_k} \tag{2.84}$$

will have a solution of the form

$$f(\mathbf{x},t) = c(t)\exp\left[-\frac{1}{2}[\mathbf{x}-\mathbf{a}(t)]^T C(t)[\mathbf{x}-\mathbf{a}(t)]\right]. \tag{2.85}$$

2.24. Show that the following transformations take solutions of (2.82) into solutions [17]:

$$T_1(\mathbf{w},\mathbf{z},\omega)f(\mathbf{x},t) = \exp\left[\frac{1}{2}\mathbf{x}\cdot\mathbf{w} + \frac{1}{4}t\|\mathbf{w}\|^2 + \omega\right] f(\mathbf{x}+t\mathbf{w}+\mathbf{z},t) \tag{2.86}$$

where $\mathbf{w},\mathbf{z} \in \mathbb{R}^2$ and $\omega \in \mathbb{R}$;

$$T_2(A)f(\mathbf{x},t) = \exp\left[-\frac{1}{4}(\delta+t\beta)^{-1}\beta\|\mathbf{x}\|^2\right](\delta+t\beta)^{-1} f\left((\delta+t\beta)^{-1}\mathbf{x}, \frac{\gamma+t\alpha}{\delta+t\beta}\right) \tag{2.87}$$

where $A \in \mathbb{R}^{2\times 2}$ with $\det A = 1$;

$$T_3(\theta)f(\mathbf{x},t) = f(R^T(\theta)\mathbf{x},t) \quad \text{where} \quad R(\theta) = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}. \tag{2.88}$$

References

1. Benedetto, J.J., Zimmermann, G., "Sampling multipliers and the Poisson summation formula," *J. Fourier Anal. Appl.,* 3, pp. 505–523, 1997.
2. Bluman, G., *Construction of Solutions to Partial Differential Equations by the Use of Transformation Groups*, PhD Dissertation, Caltech, 1967.
3. Bluman, G., Cole, J., "The general similarity solution of the heat equation," *J. Math. Mech.,* 18, pp. 1025–1042, 1969.
4. Bluman, G., Cole, J., *Similarity Methods for Differential Equations*, Springer-Verlag, New York, 1974.
5. Boyer, C., "The maximal kinematical invariance group for an arbitrary potential," *Helv. Phys. Acta*, 47, pp. 589–605, 1974.
6. Bracewell, R.N., *The Fourier Transform and Its Applications*, 2nd ed., McGraw-Hill, New York, 1986.
7. Burrus, C.S., Parks, T.W., *DFT/FFT and Convolution Algorithms*, John Wiley and Sons, New York, 1985.

8. Chirikjian, G.S., Kyatkin, A.B., *Engineering Applications of Noncommutative Harmonic Analysis*, CRC Press, Boca Raton, FL, 2001.
9. Chirikjian, G.S., "A methodology for determining mechanical properties of macromolecules from ensemble motion data," *Trends Anal. Chem.,* 22, pp. 549–553, 2003.
10. Cooley, J.W., Tukey, J., "An algorithm for the machine calculation of complex Fourier series," *Math. Comput.,* 19, pp. 297–301, 1965.
11. Fourier, J.B.J., *Théorie Analytique de la Chaleur*, F. Didot, Paris, 1822.
12. Gray, A., *Tubes*, 2nd ed., Birkhäuser, Boston, 2004.
13. Kalnins, E., Miller, W., Jr., "Symmetry and separation of variables for the heat equation," *Proc. Conf. on Symmetry, Similarity and Group-Theoretic Methods in Mechanics*, pp. 246–261, University of Calgary, Calgary, Canada, 1974.
14. Kalnins, E., Miller, W., Jr., "Lie theory and separation of variables 4: The groups $SO(2,1)$ and $SO(3)$," *J. Math. Phys.,* 15, pp. 1263–1274, 1974.
15. Körner, T.W., *Fourier Analysis*, Cambridge University Press, London, 1988 (reprinted 1993).
16. Lie, S., "Über die Integration durch bestimmte Integral von einer Klasse linearer partieller Differentialgleichungen," *Arch. Math.,* 4(3), Kristiana, p. 328, 1881.
17. Miller, W., Jr., *Symmetry and Separation of Variables*, Encyclopedia of Mathematics and Its Applications, Vol. 4, G.-C. Rota, ed., Addison-Wesley, Reading MA, 1974.
18. Olver, P.J., *Applications of Lie Groups to Differential Equations*, Springer-Verlag, New York, 1986.
19. Ovsiannikov, L.V., *Group Analysis of Differential Equations*, Academic Press, New York, 1982.
20. Patera, J., Winternitz, P., "A new basis for the representations of the rotation group: Lamé and Heun polynomials," *J. Math. Phys.,* 14, pp. 1130–1139, 1973.
21. Poluyanov, L.V., Aguilar, A., González, M., *Group Properties of the Acoustic Differential Equation*, Taylor & Francis, London, 1995.
22. Rényi, A., *Probability Theory*, North-Holland, Amsterdam, 1970.
23. Shannon, C.E., "A mathematical theory of communication," *Bell Syst. Tech. J.,* 27, pp. 379–423 and 623–656, July and October, 1948.
24. Van Loan, C., *Computational Frameworks for the Fast Fourier Transform*, SIAM, Philadelphia, 1992.
25. Weisner, L., "Group-theoretic origin of certain generating functions," *Pacific J. Math.,* 5, pp. 1033–1039, 1955.
26. Weisner, L., "Generating functions for Hermite functions," *Can. J. Math.,* 11, pp. 141–147, 1959.
27. Weisner, L., "Generating functions for Bessel functions," *Can. J. Math.,* 11, pp. 148–155, 1959.
28. Winternitz, P., Fris, I., "Invariant expansions of relativistic amplitudes and subgroups of the proper Lorentz group," *Soviet Physics* JNP, 1, pp. 636–643, 1965.

http://www.springer.com/978-0-8176-4802-2

Stochastic Models, Information Theory, and Lie Groups,
Volume 1
Classical Results and Geometric Methods
Chirikjian, G.S.
2009, XXII, 383 p. 13 illus., Hardcover
ISBN: 978-0-8176-4802-2
A product of Birkhäuser Basel