1

Pretraining

My strongest memory of the class is the very beginning, when he started, not with some deep principle of nature, or some experiment, but with a review of Gaussian integrals. Clearly, there was some calculating to be done.

Joe Polchinski, reminiscing about Richard Feynman's quantum mechanics class [5].

The goal of this book is to develop principles that enable a theoretical understanding of deep learning. Perhaps the most important principle is that wide and deep neural networks are governed by nearly-Gaussian distributions. Thus, to make it through the book, you will need to achieve mastery of Gaussian integration and perturbation theory. Our *pretraining* in this chapter consists of whirlwind introductions to these toolkits as well as a brief overview of some key concepts in statistics that we'll need. The only prerequisite is fluency in linear algebra, multivariable calculus, and rudimentary probability theory.

With that in mind, we begin in §1.1 with an extended discussion of Gaussian integrals. Our emphasis will be on calculational tools for computing averages of monomials against Gaussian distributions, culminating in a derivation of *Wick's theorem*.

Next, in §1.2, we begin by giving a general discussion of *expectation values* and *observables*. Thinking of observables as a way of learning about a probability distribution through repeated experiments, we're led to the statistical concepts of moment and cumulant and the corresponding physicists' concepts of full M-point correlator and connected M-point correlator. A particular emphasis is placed on the connected correlators as they directly characterize a distribution's deviation from Gaussianity.

In §1.3, we introduce the negative log probability or *action* representation of a probability distribution and explain how the action lets us systematically deform Gaussian distributions in order to give a compact representation of non-Gaussian distributions. In particular, we specialize to nearly-Gaussian distributions, for which deviations from Gaussianity are implemented by small *couplings* in the action, and show how perturbation theory can be used to connect the non-Gaussian couplings to observables such as the connected correlators. By treating such couplings perturbatively, we can transform any correlator of a nearly-Gaussian distribution into a sum of Gaussian integrals; each

https://doi.org/10.1017/9781009023405.003 Published online by Cambridge University Press

integral can then be evaluated by the tools we developed in §1.1. This will be one of our most important tricks, as the neural networks we'll study are all governed by nearly-Gaussian distributions, with non-Gaussian couplings that become perturbatively small as the networks become wide.

Since all these manipulations need to be at our fingertips, in this first chapter we've erred on the side of being verbose – in words and equations and examples – with the goal of making these materials as transparent and comprehensible as possible.

1.1 Gaussian Integrals

The goal of this section is to introduce Gaussian integrals and Gaussian probability distributions, and ultimately derive Wick's theorem (1.45). This theorem provides an operational formula for computing any moment of a multivariable Gaussian distribution and will be used throughout the book.

Single-Variable Gaussian Integrals

Let's take it slow and start with the simplest single-variable Gaussian function,

$$e^{-\frac{z^2}{2}}. \tag{1.1}$$

The graph of this function depicts the famous *bell curve*, symmetric around the peak at $z = 0$ and quickly tapering off for large $|z| \gg 1$. By itself, (1.1) cannot serve as a probability distribution since it's not normalized. In order to find out the proper normalization, we need to perform the Gaussian integral

$$I_1 \equiv \int_{-\infty}^{\infty} dz\, e^{-\frac{z^2}{2}}. \tag{1.2}$$

As an ancient object, there exists a neat trick to evaluate such an integral. To begin, consider its square

$$I_1^2 = \left(\int_{-\infty}^{\infty} dz\, e^{-\frac{z^2}{2}}\right)^2 = \int_{-\infty}^{\infty} dx\, e^{-\frac{x^2}{2}} \int_{-\infty}^{\infty} dy\, e^{-\frac{y^2}{2}} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} dxdy\, e^{-\frac{1}{2}(x^2+y^2)}, \tag{1.3}$$

where in the middle we just changed the names of the dummy integration variables. Next, we change variables to polar coordinates $(x, y) = (r\cos\phi, r\sin\phi)$, which transforms the integral measure as $dxdy = rdrd\phi$ and gives us two elementary integrals to compute:

$$\begin{aligned} I_1^2 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} dxdy\, e^{-\frac{1}{2}(x^2+y^2)} &= \int_0^{\infty} rdr \int_0^{2\pi} d\phi\, e^{-\frac{r^2}{2}} \\ &= 2\pi \int_0^{\infty} dr\, re^{-\frac{r^2}{2}} = 2\pi \left| -e^{-\frac{r^2}{2}} \right|_{r=0}^{r=\infty} = 2\pi. \end{aligned} \tag{1.4}$$

https://doi.org/10.1017/9781009023405.003 Published online by Cambridge University Press

Finally, by taking a square root we can evaluate the Gaussian integral (1.2) as

$$I_1 = \int_{-\infty}^{\infty} dz\, e^{-\frac{z^2}{2}} = \sqrt{2\pi}. \tag{1.5}$$

Dividing the Gaussian function with this normalization factor, we define the **Gaussian probability distribution** with unit variance as

$$p(z) \equiv \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}, \tag{1.6}$$

which is now properly normalized, i.e., $\int_{-\infty}^{\infty} dz\, p(z) = 1$. Such a distribution with zero mean and unit variance is sometimes called the *standard normal distribution.*

Extending this result to a Gaussian distribution with **variance** $K > 0$ is super-easy. The corresponding normalization factor is given by

$$I_K \equiv \int_{-\infty}^{\infty} dz\, e^{-\frac{z^2}{2K}} = \sqrt{K} \int_{-\infty}^{\infty} du\, e^{-\frac{u^2}{2}} = \sqrt{2\pi K}, \tag{1.7}$$

where in the middle we rescaled the integration variable as $u = z/\sqrt{K}$. We can then define the Gaussian distribution with variance K as

$$p(z) \equiv \frac{1}{\sqrt{2\pi K}} e^{-\frac{z^2}{2K}}. \tag{1.8}$$

The graph of this distribution again depicts a bell curve symmetric around $z = 0$, but it's now equipped with a scale K characterizing its broadness, tapering off for $|z| \gg \sqrt{K}$. More generally, we can shift the center of the bell curve as

$$p(z) \equiv \frac{1}{\sqrt{2\pi K}} e^{-\frac{(z-s)^2}{2K}}, \tag{1.9}$$

so that it is now symmetric around $z = s$. This center value s is called the **mean** of the distribution, because it is:

$$\begin{aligned}
\mathbb{E}[z] \equiv \int_{-\infty}^{\infty} dz\, p(z)\, z &= \frac{1}{\sqrt{2\pi K}} \int_{-\infty}^{\infty} dz\, e^{-\frac{(z-s)^2}{2K}} z \\
&= \frac{1}{I_K} \int_{-\infty}^{\infty} dw\, e^{-\frac{w^2}{2K}} (s + w) \\
&= \frac{s I_K}{I_K} + \frac{1}{I_K} \int_{-\infty}^{\infty} dw \left(e^{-\frac{w^2}{2K}} w \right) \\
&= s\,,
\end{aligned} \tag{1.10}$$

where in the middle we shifted the variable as $w = z - s$ and in the very last step noticed that the integrand of the second term is odd with respect to the sign flip of the integration variable $w \leftrightarrow -w$ and hence integrates to zero.

Focusing on Gaussian distributions with zero mean, let's consider other **expectation values** for general functions $\mathcal{O}(z)$, i.e.,

$$\mathbb{E}[\mathcal{O}(z)] \equiv \int_{-\infty}^{\infty} dz\, p(z)\, \mathcal{O}(z) = \frac{1}{\sqrt{2\pi K}} \int_{-\infty}^{\infty} dz\, e^{-\frac{z^2}{2K}} \mathcal{O}(z)\,. \tag{1.11}$$

https://doi.org/10.1017/9781009023405.003 Published online by Cambridge University Press

We'll often refer to such functions $\mathcal{O}(z)$ as **observables**, since they can correspond to measurement outcomes of experiments. A special class of expectation values are called **moments** and correspond to the insertion of z^M into the integrand for any integer M:

$$\mathbb{E}\left[z^M\right] = \frac{1}{\sqrt{2\pi K}} \int_{-\infty}^{\infty} dz\ e^{-\frac{z^2}{2K}}\, z^M . \tag{1.12}$$

Note that the integral vanishes for any odd exponent M, because then the integrand is odd with respect to the sign flip $z \leftrightarrow -z$. As for the even number $M = 2m$ of z insertions, we will need to evaluate integrals of the form

$$I_{K,m} \equiv \int_{-\infty}^{\infty} dz\ e^{-\frac{z^2}{2K}}\, z^{2m} . \tag{1.13}$$

As objects almost as ancient as (1.2), again there exists a neat trick to evaluate them:

$$\begin{aligned} I_{K,m} &= \int_{-\infty}^{\infty} dz\ e^{-\frac{z^2}{2K}}\, z^{2m} = \left(2K^2 \frac{d}{dK}\right)^m \int_{-\infty}^{\infty} dz\ e^{-\frac{z^2}{2K}} = \left(2K^2 \frac{d}{dK}\right)^m I_K \\ &= \left(2K^2 \frac{d}{dK}\right)^m \sqrt{2\pi} K^{\frac{1}{2}} = \sqrt{2\pi} K^{\frac{2m+1}{2}} (2m-1)(2m-3)\cdots 1 , \end{aligned} \tag{1.14}$$

where in going to the second line we substituted in our expression (1.7) for I_K. Therefore, we see that the even moments are given by the simple formula[1]

$$\mathbb{E}\left[z^{2m}\right] = \frac{I_{K,m}}{\sqrt{2\pi K}} = K^m\,(2m-1)!!, \tag{1.15}$$

where we have introduced the double factorial

$$(2m-1)!! \equiv (2m-1)(2m-3)\cdots 1 = \frac{(2m)!}{2^m m!} . \tag{1.16}$$

The result (1.15) is Wick's theorem for single-variable Gaussian distributions.

There's actually another nice way to derive (1.15), which can much more naturally be extended to multivariable Gaussian distributions. This derivation starts with the consideration of a Gaussian integral with a **source term** J, which we define as

$$Z_{K,J} \equiv \int_{-\infty}^{\infty} dz\ e^{-\frac{z^2}{2K}+Jz} . \tag{1.17}$$

Note that when setting the source to zero, we recover the normalization of the Gaussian integral, giving the relationship $Z_{K,J=0} = I_K$. In the physics literature $Z_{K,J}$ is sometimes called a **partition function with source** and, as we will soon see, this integral serves as a **generating function** for the moments. We can evaluate $Z_{K,J}$ by completing the square in the exponent,

$$-\frac{z^2}{2K} + Jz = -\frac{(z-JK)^2}{2K} + \frac{KJ^2}{2} , \tag{1.18}$$

[1] This equation with $2m = 2$ makes clear why we called K the variance, since for zero-mean Gaussian distributions with variance K, we have $\text{var}(z) \equiv \mathbb{E}\left[(z - \mathbb{E}[z])^2\right] = \mathbb{E}\left[z^2\right] - \mathbb{E}[z]^2 = \mathbb{E}\left[z^2\right] = K$.

https://doi.org/10.1017/9781009023405.003 Published online by Cambridge University Press

which lets us rewrite the integral (1.17) as

$$Z_{K,J} = e^{\frac{KJ^2}{2}} \int_{-\infty}^{\infty} dz\, e^{-\frac{(z-JK)^2}{2K}} = e^{\frac{KJ^2}{2}} I_K = e^{\frac{KJ^2}{2}} \sqrt{2\pi K}, \tag{1.19}$$

where in the middle equality we noticed that the integrand is just a shifted Gaussian function with variance K.

We can now relate the Gaussian integral with a source $Z_{K,J}$ to the Gaussian integral with insertions $I_{K,m}$. By differentiating $Z_{K,J}$ with respect to the source J and *then* setting the source to zero, we observe that

$$I_{K,m} = \int_{-\infty}^{\infty} dz\, e^{-\frac{z^2}{2K}} z^{2m} = \left[\left(\frac{d}{dJ}\right)^{2m} \int_{-\infty}^{\infty} dz\, e^{-\frac{z^2}{2K}+Jz}\right]\Bigg|_{J=0} = \left[\left(\frac{d}{dJ}\right)^{2m} Z_{K,J}\right]\Bigg|_{J=0}. \tag{1.20}$$

In other words, the integrals $I_{K,m}$ are simply related to the even Taylor coefficients of the partition function $Z_{K,J}$ around $J=0$. For instance, for $2m=2$ we have

$$\mathbb{E}\left[z^2\right] = \frac{I_{K,1}}{\sqrt{2\pi K}} = \left[\left(\frac{d}{dJ}\right)^2 e^{\frac{KJ^2}{2}}\right]\Bigg|_{J=0} = \left[e^{\frac{KJ^2}{2}}\left(K + K^2J^2\right)\right]\Bigg|_{J=0} = K, \tag{1.21}$$

and for $2m=4$ we have

$$\mathbb{E}\left[z^4\right] = \frac{I_{K,2}}{\sqrt{2\pi K}} = \left[\left(\frac{d}{dJ}\right)^4 e^{\frac{KJ^2}{2}}\right]\Bigg|_{J=0} = \left[e^{\frac{KJ^2}{2}}\left(3K^2 + 6K^3J^2 + K^4J^4\right)\right]\Bigg|_{J=0} = 3K^2. \tag{1.22}$$

Notice that any terms with dangling sources J vanish upon setting $J=0$. This observation gives a simple way to evaluate correlators for general m: Taylor-expand the exponential $Z_{K,J}/I_K = \exp\!\left(\frac{KJ^2}{2}\right)$, and keep the term with the right amount of sources such that the expression doesn't vanish. Doing exactly that, we get

$$\begin{aligned}\mathbb{E}\left[z^{2m}\right] = \frac{I_{K,m}}{\sqrt{2\pi K}} &= \left[\left(\frac{d}{dJ}\right)^{2m} e^{\frac{KJ^2}{2}}\right]\Bigg|_{J=0} = \left\{\left(\frac{d}{dJ}\right)^{2m}\left[\sum_{k=0}^{\infty}\frac{1}{k!}\left(\frac{K}{2}\right)^k J^{2k}\right]\right\}\Bigg|_{J=0} \\ &= \left(\frac{d}{dJ}\right)^{2m}\left[\frac{1}{m!}\left(\frac{K}{2}\right)^m J^{2m}\right] = K^m\frac{(2m)!}{2^m m!} = K^m(2m-1)!!\,,\end{aligned} \tag{1.23}$$

which completes our second derivation of Wick's theorem (1.15) for the single-variable Gaussian distribution. This derivation was much longer than the first neat derivation but can be very naturally extended to the multivariable Gaussian distribution, which we turn to next.

https://doi.org/10.1017/9781009023405.003 Published online by Cambridge University Press

Multivariable Gaussian Integrals

Picking up speed, we are now ready to handle multivariable Gaussian integrals for an N-dimensional variable z_μ with $\mu = 1, \ldots, N$.[2] The multivariable Gaussian function is defined as

$$\exp\left[-\frac{1}{2}\sum_{\mu,\nu=1}^{N} z_\mu (K^{-1})_{\mu\nu}\, z_\nu\right], \tag{1.24}$$

where the variance or **covariance matrix** $K_{\mu\nu}$ is an N-by-N symmetric positive definite matrix, and its inverse $(K^{-1})_{\mu\nu}$ is defined so that their matrix product gives the N-by-N identity matrix

$$\sum_{\rho=1}^{N}(K^{-1})_{\mu\rho}\, K_{\rho\nu} = \delta_{\mu\nu}. \tag{1.25}$$

Here we have also introduced the **Kronecker delta** $\delta_{\mu\nu}$, which satisfies

$$\delta_{\mu\nu} \equiv \begin{cases} 1\,, & \mu = \nu, \\ 0\,, & \mu \neq \nu. \end{cases} \tag{1.26}$$

The Kronecker delta is just a convenient representation of the identity matrix.

Now, to construct a probability distribution from the Gaussian function (1.24), we again need to evaluate the normalization factor

$$\begin{aligned} I_K &\equiv \int d^N z \, \exp\left[-\frac{1}{2}\sum_{\mu,\nu=1}^{N} z_\mu (K^{-1})_{\mu\nu}\, z_\nu\right] \\ &= \int_{-\infty}^{\infty} dz_1 \int_{-\infty}^{\infty} dz_2 \cdots \int_{-\infty}^{\infty} dz_N \, \exp\left[-\frac{1}{2}\sum_{\mu,\nu=1}^{N} z_\mu (K^{-1})_{\mu\nu}\, z_\nu\right]. \end{aligned} \tag{1.27}$$

To compute this integral, first recall from linear algebra that, given an N-by-N symmetric matrix $K_{\mu\nu}$, there is always an orthogonal matrix[3] $O_{\mu\nu}$ that diagonalizes $K_{\mu\nu}$ as $(OKO^T)_{\mu\nu} = \lambda_\mu \delta_{\mu\nu}$ with eigenvalues $\lambda_{\mu=1,\ldots,N}$ and diagonalizes its inverse as $(OK^{-1}O^T)_{\mu\nu} = (1/\lambda_\mu)\,\delta_{\mu\nu}$. With this in mind, after twice inserting the identity matrix as $\delta_{\mu\nu} = (O^T O)_{\mu\nu}$, the sum in the exponent of the integral can be expressed in terms of the eigenvalues as

[2] Throughout this book, we will explicitly write out the component indices of vectors, matrices, and tensors as much as possible, except on some occasions when it is clear enough from context.

[3] An *orthogonal matrix* $O_{\mu\nu}$ is a matrix whose transpose $\left(O^T\right)_{\mu\nu}$ equals its inverse, i.e., $(O^T O)_{\mu\nu} = \delta_{\mu\nu}$.

https://doi.org/10.1017/9781009023405.003 Published online by Cambridge University Press

$$\begin{aligned}\sum_{\mu,\nu=1}^{N} z_\mu (K^{-1})_{\mu\nu} z_\nu &= \sum_{\mu,\rho,\sigma,\nu=1}^{N} z_\mu \, (O^T O)_{\mu\rho} (K^{-1})_{\rho\sigma} (O^T O)_{\sigma\nu} \, z_\nu \\ &= \sum_{\mu,\nu=1}^{N} (Oz)_\mu (OK^{-1}O^T)_{\mu\nu} (Oz)_\nu \\ &= \sum_{\mu=1}^{N} \frac{1}{\lambda_\mu} (Oz)_\mu^2,\end{aligned} \tag{1.28}$$

where to reach the final line we used the diagonalization property of the inverse covariance matrix. Remembering for a positive definite matrix $K_{\mu\nu}$ that the eigenvalues are all positive $\lambda_\mu > 0$, we see that the λ_μ sets the scale of the falloff of the Gaussian function in each of the eigendirections. Next, recall from multivariable calculus that a change of variables $u_\mu \equiv (Oz)_\mu$ with an orthogonal matrix O leaves the integration measure invariant, i.e., $d^N z = d^N u$. All together, this lets us factorize the multivariable Gaussian integral (1.27) into a product of single-variable Gaussian integrals (1.7), yielding

$$\begin{aligned} I_K &= \int_{-\infty}^{\infty} du_1 \int_{-\infty}^{\infty} du_2 \cdots \int_{-\infty}^{\infty} du_N \, \exp\left(-\frac{u_1^2}{2\lambda_1} - \frac{u_2^2}{2\lambda_2} - \cdots - \frac{u_N^2}{2\lambda_N} \right) \\ &= \prod_{\mu=1}^{N} \left[\int_{-\infty}^{\infty} du_\mu \, \exp\left(-\frac{u_\mu^2}{2\lambda_\mu} \right) \right] = \prod_{\mu=1}^{N} \sqrt{2\pi\lambda_\mu} = \sqrt{\prod_{\mu=1}^{N} (2\pi\lambda_\mu)}. \end{aligned} \tag{1.29}$$

Finally, recall one last fact from linear algebra: that the product of the eigenvalues of a matrix is equal to the matrix determinant. Thus, compactly, we can express the value of the multivariable Gaussian integral as

$$I_K = \int d^N z \, \exp\left[-\frac{1}{2} \sum_{\mu,\nu=1}^{N} z_\mu (K^{-1})_{\mu\nu} z_\nu \right] = \sqrt{|2\pi K|}, \tag{1.30}$$

where $|A|$ denotes the determinant of a square matrix A.

Having figured out the normalization factor, we can define the zero-mean **multivariable Gaussian probability distribution** with variance $K_{\mu\nu}$ as

$$p(z) = \frac{1}{\sqrt{|2\pi K|}} \exp\left[-\frac{1}{2} \sum_{\mu,\nu=1}^{N} z_\mu (K^{-1})_{\mu\nu} \, z_\nu \right]. \tag{1.31}$$

While we're at it, let us also introduce the conventions of suppressing the superscript "-1" for the inverse covariance $(K^{-1})_{\mu\nu}$, instead placing the component indices upstairs as

$$K^{\mu\nu} \equiv (K^{-1})_{\mu\nu}. \tag{1.32}$$

This way, we distinguish the covariance $K_{\mu\nu}$ and the inverse covariance $K^{\mu\nu}$ by whether or not component indices are lowered or raised. With this notation, inherited from *general relativity*, the defining equation for the inverse covariance (1.25) is written instead as

https://doi.org/10.1017/9781009023405.003 Published online by Cambridge University Press

$$\sum_{\rho=1}^{N} K^{\mu\rho} K_{\rho\nu} = \delta^{\mu}{}_{\nu}, \tag{1.33}$$

and the multivariable Gaussian distribution (1.31) is written as

$$p(z) = \frac{1}{\sqrt{|2\pi K|}} \exp\left(-\frac{1}{2} \sum_{\mu,\nu=1}^{N} z_\mu K^{\mu\nu} z_\nu \right). \tag{1.34}$$

Although it might take some getting used to, this notation saves us some space and saves you some handwriting pain.[4] Regardless of how it's written, the zero-mean multivariable Gaussian probability distribution (1.34) peaks at $z = 0$, and its falloff is direction-dependent, determined by the covariance matrix $K_{\mu\nu}$. More generally, we can shift the peak of the Gaussian distribution to s_μ:

$$p(z) = \frac{1}{\sqrt{|2\pi K|}} \exp\left[-\frac{1}{2} \sum_{\mu,\nu=1}^{N} (z-s)_\mu \, K^{\mu\nu} \, (z-s)_\nu \right], \tag{1.35}$$

which defines a general multivariable Gaussian distribution with mean $\mathbb{E}[z_\mu] = s_\mu$ and covariance $K_{\mu\nu}$. This is the most general version of the Gaussian distribution.

Next, let's consider the moments of the mean-zero multivariable Gaussian distribution

$$\begin{aligned} \mathbb{E}\left[z_{\mu_1} \cdots z_{\mu_M}\right] &\equiv \int d^N z \; p(z)\, z_{\mu_1} \cdots z_{\mu_M} \\ &= \frac{1}{\sqrt{|2\pi K|}} \int d^N z \; \exp\left(-\frac{1}{2} \sum_{\mu,\nu=1}^{N} z_\mu K^{\mu\nu} z_\nu \right) z_{\mu_1} \cdots z_{\mu_M} = \frac{I_{K,(\mu_1,\ldots,\mu_M)}}{I_K}, \end{aligned} \tag{1.36}$$

where we introduced multivariable Gaussian integrals with insertions

$$I_{K,(\mu_1,\ldots,\mu_M)} \equiv \int d^N z \; \exp\left(-\frac{1}{2} \sum_{\mu,\nu=1}^{N} z_\mu K^{\mu\nu} z_\nu \right) z_{\mu_1} \cdots z_{\mu_M}. \tag{1.37}$$

Following our approach in the single-variable case, let's construct the generating function for the integrals $I_{K,(\mu_1,\ldots,\mu_M)}$ by including a source term J^μ as

$$Z_{K,J} \equiv \int d^N z \; \exp\left(-\frac{1}{2} \sum_{\mu,\nu=1}^{N} z_\mu K^{\mu\nu} z_\nu + \sum_{\mu=1}^{N} J^\mu z_\mu \right). \tag{1.38}$$

[4] If you like, in your notes you can also go full general-relativistic mode and adopt *Einstein summation convention*, suppressing the summation symbol any time indices are repeated in upstairs-downstairs pairs. For instance, if we adopted this convention, we would write the defining equation for the inverse simply as $K^{\mu\rho} K_{\rho\nu} = \delta^{\mu}{}_{\nu}$ and the Gaussian function as $\exp\left(-\frac{1}{2} z_\mu K^{\mu\nu} z_\nu\right)$.

Specifically for neural networks, you might find the Einstein summation convention helpful for *sample* indices but sometimes confusing for *neural* indices. For extra clarity, we won't adopt this convention in the text of the book, but we mention it now since we do often use such a convention to simplify our own calculations in private.

https://doi.org/10.1017/9781009023405.003 Published online by Cambridge University Press

As the name suggests, differentiating the generating function $Z_{K,J}$ with respect to the source J^μ brings down a power of z_μ such that after M such differentiations, we have

$$\left[\frac{d}{dJ^{\mu_1}}\frac{d}{dJ^{\mu_2}}\cdots\frac{d}{dJ^{\mu_M}}Z_{K,J}\right]\Bigg|_{J=0} \tag{1.39}$$
$$= \int d^N z\ \exp\left(-\frac{1}{2}\sum_{\mu,\nu=1}^{N} z_\mu K^{\mu\nu} z_\nu\right)\ z_{\mu_1}\cdots z_{\mu_M} = I_{K,(\mu_1,\ldots,\mu_M)}.$$

So, as in the single-variable case, the Taylor coefficients of the partition function $Z_{K,J}$ expanded around $J^\mu = 0$ are simply related to the integrals with insertions $I_{K,(\mu_1,\ldots,\mu_M)}$. Therefore, if we knew a closed-form expression for $Z_{K,J}$, we could easily compute the values of the integrals $I_{K,(\mu_1,\ldots,\mu_M)}$.

To evaluate the generating function $Z_{K,J}$ in a closed form, again we follow the lead of the single-variable case and complete the square in the exponent of the integrand in (1.38) as

$$-\frac{1}{2}\sum_{\mu,\nu=1}^{N} z_\mu K^{\mu\nu} z_\nu + \sum_{\mu=1}^{N} J^\mu z_\mu \tag{1.40}$$
$$= -\frac{1}{2}\sum_{\mu,\nu=1}^{N}\left(z_\mu - \sum_{\rho=1}^{N} K_{\mu\rho}J^\rho\right)K^{\mu\nu}\left(z_\nu - \sum_{\lambda=1}^{N} K_{\nu\lambda}J^\lambda\right) + \frac{1}{2}\sum_{\mu,\nu=1}^{N} J^\mu K_{\mu\nu} J^\nu$$
$$= -\frac{1}{2}\sum_{\mu,\nu=1}^{N} w_\mu K^{\mu\nu} w_\nu + \frac{1}{2}\sum_{\mu,\nu=1}^{N} J^\mu K_{\mu\nu} J^\nu\,,$$

where we have introduced the shifted variable $w_\mu \equiv z_\mu - \sum_{\rho=1}^{N} K_{\mu\rho}J^\rho$. Using this substitution, the generating function can be evaluated explicitly:

$$Z_{K,J} = \exp\left(\frac{1}{2}\sum_{\mu,\nu=1}^{N} J^\mu K_{\mu\nu} J^\nu\right)\int d^N w\ \exp\left[-\frac{1}{2}\sum_{\mu,\nu=1}^{N} w_\mu K^{\mu\nu} w_\nu\right] \tag{1.41}$$
$$= \sqrt{|2\pi K|}\exp\left(\frac{1}{2}\sum_{\mu,\nu=1}^{N} J^\mu K_{\mu\nu} J^\nu\right)\,,$$

where at the end we used our formula for the multivariable integral I_K, (1.30). With our closed-form expression (1.41) for the generating function $Z_{K,J}$, we can compute the Gaussian integrals with insertions $I_{K,(\mu_1,\ldots,\mu_M)}$ by differentiating it, using (1.39). For an even number $M = 2m$ of insertions, we find a really nice formula

$$\mathbb{E}\left[z_{\mu_1}\cdots z_{\mu_{2m}}\right] = \frac{I_{K,(\mu_1,\ldots,\mu_{2m})}}{I_K} = \frac{1}{I_K}\left[\frac{d}{dJ^{\mu_1}}\cdots\frac{d}{dJ^{\mu_{2m}}}Z_{K,J}\right]\Bigg|_{J=0} \tag{1.42}$$
$$= \frac{1}{2^m m!}\frac{d}{dJ^{\mu_1}}\frac{d}{dJ^{\mu_2}}\cdots\frac{d}{dJ^{\mu_{2m}}}\left(\sum_{\mu,\nu=1}^{N} J^\mu K_{\mu\nu} J^\nu\right)^m .$$

https://doi.org/10.1017/9781009023405.003 Published online by Cambridge University Press

For an odd number $M = 2m + 1$ of insertions, there is dangling source upon setting $J = 0$, and so those integrals vanish. You can also see this by looking at the integrand for any odd moment and noticing that it is odd with respect to the sign flip of the integration variables $z_\mu \leftrightarrow -z_\mu$.

Now, let's take a few moments to evaluate a few moments using this formula. For $2m = 2$, we have

$$\mathbb{E}\left[z_{\mu_1} z_{\mu_2}\right] = \frac{1}{2} \frac{d}{dJ^{\mu_1}} \frac{d}{dJ^{\mu_2}} \left(\sum_{\mu,\nu=1}^{N} J^\mu K_{\mu\nu} J^\nu \right) = K_{\mu_1\mu_2}. \tag{1.43}$$

Here, there are $2! = 2$ ways to apply the product rule for derivatives and differentiate the two J's, both of which evaluate to the same expression due to the symmetry of the covariance, $K_{\mu_1\mu_2} = K_{\mu_2\mu_1}$. This expression (1.43) validates in the multivariable setting why we have been calling $K_{\mu\nu}$ the covariance, because we see explicitly that it is the covariance.

Next, for $2m = 4$ we get a more complicated expression:

$$\begin{aligned}\mathbb{E}\left[z_{\mu_1} z_{\mu_2} z_{\mu_3} z_{\mu_4}\right] &= \frac{1}{2^2 2!} \frac{d}{dJ^{\mu_1}} \frac{d}{dJ^{\mu_2}} \frac{d}{dJ^{\mu_3}} \frac{d}{dJ^{\mu_4}} \left(\sum_{\mu,\nu=1}^{N} J^\mu K_{\mu\nu} J^\nu \right) \left(\sum_{\rho,\lambda=1}^{N} J^\rho K_{\rho\lambda} J^\lambda \right) \\ &= K_{\mu_1\mu_2} K_{\mu_3\mu_4} + K_{\mu_1\mu_3} K_{\mu_2\mu_4} + K_{\mu_1\mu_4} K_{\mu_2\mu_3}. \end{aligned} \tag{1.44}$$

Here we note that there are now $4! = 24$ ways to differentiate the four J's, though only three distinct ways to pair the four auxiliary indices $1, 2, 3, 4$ that sit under μ. This gives $24/3 = 8 = 2^2 2!$ equivalent terms for each of the three pairings, which cancels against the overall factor $1/(2^2 2!)$.

For general $2m$, there are $(2m)!$ ways to differentiate the sources, of which $2^m m!$ of those ways are equivalent. This gives $(2m)!/(2^m m!) = (2m-1)!!$ distinct terms, corresponding to the $(2m-1)!!$ distinct pairings of $2m$ auxiliary indices $1, \ldots, 2m$ that sit under μ. The factor of $1/(2^m m!)$ in the denominator of (1.42) ensures that the coefficient of each of these terms is normalized to unity. Thus, most generally, we can express the moments of the multivariable Gaussian with the following formula:

$$\mathbb{E}\left[z_{\mu_1} \cdots z_{\mu_{2m}}\right] = \sum_{\text{all pairing}} K_{\mu_{k_1}\mu_{k_2}} \cdots K_{\mu_{k_{2m-1}}\mu_{k_{2m}}}, \tag{1.45}$$

where, to reiterate, the sum is over all the possible distinct pairings of the $2m$ auxiliary indices under μ such that the result has the $(2m-1)!!$ terms that we described above. Each factor of the covariance $K_{\mu\nu}$ in a term in sum is called a **Wick contraction**, corresponding to a particular pairing of auxiliary indices. Each term then is composed of m different Wick contractions, representing a distinct way of pairing up all the auxiliary indices. To make sure you understand how this pairing works, look back at the $2m = 2$ case (1.43) – with a single Wick contraction – and the $2m = 4$ case (1.44) – with three distinct ways of making two Wick contractions – and try to work out the $2m = 6$ case, which yields $(6-1)!! = 15$ distinct ways of making three Wick contractions:

https://doi.org/10.1017/9781009023405.003 Published online by Cambridge University Press

$$
\begin{aligned}
\mathbb{E}\left[z_{\mu_1} z_{\mu_2} z_{\mu_3} z_{\mu_4} z_{\mu_5} z_{\mu_6}\right] = &\, K_{\mu_1\mu_2} K_{\mu_3\mu_4} K_{\mu_5\mu_6} + K_{\mu_1\mu_3} K_{\mu_2\mu_4} K_{\mu_5\mu_6} + K_{\mu_1\mu_4} K_{\mu_2\mu_3} K_{\mu_5\mu_6} \\
&+ K_{\mu_1\mu_2} K_{\mu_3\mu_5} K_{\mu_4\mu_6} + K_{\mu_1\mu_3} K_{\mu_2\mu_5} K_{\mu_4\mu_6} + K_{\mu_1\mu_5} K_{\mu_2\mu_3} K_{\mu_4\mu_6} \\
&+ K_{\mu_1\mu_2} K_{\mu_5\mu_4} K_{\mu_3\mu_6} + K_{\mu_1\mu_5} K_{\mu_2\mu_4} K_{\mu_3\mu_6} + K_{\mu_1\mu_4} K_{\mu_2\mu_5} K_{\mu_3\mu_6} \\
&+ K_{\mu_1\mu_5} K_{\mu_3\mu_4} K_{\mu_2\mu_6} + K_{\mu_1\mu_3} K_{\mu_5\mu_4} K_{\mu_2\mu_6} + K_{\mu_1\mu_4} K_{\mu_5\mu_3} K_{\mu_2\mu_6} \\
&+ K_{\mu_5\mu_2} K_{\mu_3\mu_4} K_{\mu_1\mu_6} + K_{\mu_5\mu_3} K_{\mu_2\mu_4} K_{\mu_1\mu_6} + K_{\mu_5\mu_4} K_{\mu_2\mu_3} K_{\mu_1\mu_6} .
\end{aligned}
$$

The formula (1.45) is **Wick's theorem**. Put a box around it. Take a few moments for reflection.

. . .

. . .

. . .

Good. You are now a Gaussian sensei. Exhale, and then say, as Neo would say, "I know Gaussian integrals."

Now that the moments have passed, it is an appropriate time to transition to the next section, where you will learn about more general probability distributions.

1.2 Probability, Correlation and Statistics, and All That

In introducing the Gaussian distribution in the last section, we briefly touched upon the concepts of expectation and moments. These are defined for non-Gaussian probability distributions too, so now let us reintroduce these concepts and expand on their definitions, with an eye toward understanding the nearly-Gaussian distributions that describe wide neural networks.

Given a **probability distribution** $p(z)$ of an N-dimensional random variable z_μ, we can learn about its statistics by measuring functions of z_μ. We'll refer to such measurable functions in a generic sense as **observables** and denote them as $\mathcal{O}(z)$. The **expectation value** of an observable

$$
\mathbb{E}\left[\mathcal{O}(z)\right] \equiv \int d^N z \; p(z)\, \mathcal{O}(z) \tag{1.46}
$$

characterizes the mean value of the random function $\mathcal{O}(z)$. Note that the observable $\mathcal{O}(z)$ need not be a scalar-valued function, e.g., the second moment of a distribution is a matrix-valued observable given by $\mathcal{O}(z) = z_\mu z_\nu$.

Operationally, an observable is a quantity that we measure by conducting experiments in order to connect to a theoretical model for the underlying probability distribution describing z_μ. In particular, we repeatedly measure the observables that are naturally accessible to us as experimenters, collect their statistics, and then compare them with predictions for the expectation values of those observables computed from some theoretical model of $p(z)$.

With that in mind, it's very natural to ask: what kind of information can we learn about an underlying distribution $p(z)$ by measuring an observable $\mathcal{O}(z)$? For an a priori unknown distribution, is there a set of observables that can serve as a sufficient probe

https://doi.org/10.1017/9781009023405.003 Published online by Cambridge University Press

of $p(z)$ such that we could use that information to predict the result of all future experiments involving z_μ?

Consider a class of observables that we've already encountered, the **moments** or ***M*-point correlators** of z_μ, given by the expectation[5]

$$\mathbb{E}\left[z_{\mu_1} z_{\mu_2} \cdots z_{\mu_M}\right] = \int d^N z \; p(z)\, z_{\mu_1} z_{\mu_2} \cdots z_{\mu_M} . \tag{1.47}$$

In principle, knowing the M-point correlators of a distribution lets us compute the expectation value of any analytic observable $\mathcal{O}(z)$ via Taylor expansion:

$$\begin{aligned}\mathbb{E}\left[\mathcal{O}(z)\right] &= \mathbb{E}\left[\sum_{M=0}^{\infty} \frac{1}{M!} \sum_{\mu_1,\ldots,\mu_M=1}^{N} \frac{\partial^M \mathcal{O}}{\partial z_{\mu_1} \cdots \partial z_{\mu_M}}\bigg|_{z=0} z_{\mu_1} z_{\mu_2} \cdots z_{\mu_M}\right] \\ &= \sum_{M=0}^{\infty} \frac{1}{M!} \sum_{\mu_1,\ldots,\mu_M=1}^{N} \frac{\partial^M \mathcal{O}}{\partial z_{\mu_1} \cdots \partial z_{\mu_M}}\bigg|_{z=0} \mathbb{E}\left[z_{\mu_1} z_{\mu_2} \cdots z_{\mu_M}\right] ,\end{aligned} \tag{1.48}$$

where on the last line we took the Taylor coefficients out of the expectation by using the linearity property of the expectation, inherited from the linearity property of the integral in (1.46). As such, it's clear that the collection of all the M-point correlators completely characterizes a probability distribution for all intents and purposes.[6]

However, this description in terms of all the correlators is somewhat cumbersome and operationally infeasible. To get a reliable estimate of the M-point correlator, we must simultaneously measure M components of a random variable for each draw and repeat such measurements many times. As M grows, this task quickly becomes impractical. In fact, if we could easily perform such measurements for all M, then our theoretical model of $p(z)$ would no longer be a useful abstraction: from (1.48) we would already know the outcome of all possible experiments that we could perform, leaving nothing for us to predict.

To that point, essentially all useful distributions can be effectively described in terms of a finite number of quantities, giving them a parsimonious representation. For instance,

[5] In the rest of this book, we'll often use the physics term *M-point correlator* rather than the statistics term *moment*, though they mean the same thing and can be used interchangeably.

[6] In fact, the moments offer a dual description of the probability distribution through either the Laplace transform or the Fourier transform. For instance, the Laplace transform of the probability distribution $p(z)$ is given by

$$Z_J \equiv \mathbb{E}\left[\exp\left(\sum_\mu J^\mu z_\mu\right)\right] = \int \left[\prod_\mu dz_\mu\right] p(z) \exp\left(\sum_\mu J^\mu z_\mu\right) . \tag{1.49}$$

As in the Gaussian case, this integral gives a generating function for the M-point correlators of $p(z)$, which means that Z_J can be reconstructed from these correlators. The probability distribution can then be obtained through the inverse Laplace transform.

https://doi.org/10.1017/9781009023405.003 Published online by Cambridge University Press

consider the zero-mean n-dimensional Gaussian distribution with the variance $K_{\mu\nu}$. The nonzero $2m$-point correlators are given by Wick's theorem (1.45) as

$$\mathbb{E}\left[z_{\mu_1} z_{\mu_2} \cdots z_{\mu_{2m}}\right] = \sum_{\text{all pairing}} K_{\mu_{k_1}\mu_{k_2}} \cdots K_{\mu_{k_{2m-1}}\mu_{k_{2m}}} \tag{1.50}$$

and are determined entirely by the $N(N+1)/2$ independent components of the variance $K_{\mu\nu}$. The variance itself can be estimated by measuring the two-point correlator

$$\mathbb{E}\left[z_\mu z_\nu\right] = K_{\mu\nu}. \tag{1.51}$$

This is consistent with our description of the distribution itself as "the zero-mean N-dimensional Gaussian distribution with the variance $K_{\mu\nu}$" in which we only had to specify these same set of numbers, $K_{\mu\nu}$, to pick out the particular distribution we had in mind. For zero-mean Gaussian distributions, there's no reason to measure or keep track of any of the higher-point correlators as they are completely constrained by the variance through (1.50).

More generally, it would be nice if there were a systematic way to learn about non-Gaussian probability distributions without performing an infinite number of experiments. For nearly-Gaussian distributions, a useful set of observables is given by what statisticians call **cumulants** and physicists call **connected correlators**.[7] As the formal definition of these quantities is somewhat cumbersome and unintuitive, let's start with a few simple examples.

The first cumulant or the connected one-point correlator is the same as the full one-point correlator:

$$\mathbb{E}\left[z_\mu\right]\Big|_{\text{connected}} \equiv \mathbb{E}\left[z_\mu\right]. \tag{1.52}$$

This is just the *mean* of the distribution. The second cumulant or the connected two-point correlator is given by

$$\begin{aligned}\mathbb{E}\left[z_\mu z_\nu\right]\Big|_{\text{connected}} &\equiv \mathbb{E}\left[z_\mu z_\nu\right] - \mathbb{E}\left[z_\mu\right]\mathbb{E}\left[z_\nu\right] \\ &= \mathbb{E}\left[\left(z_\mu - \mathbb{E}\left[z_\mu\right]\right)\left(z_\nu - \mathbb{E}\left[z_\nu\right]\right)\right] \equiv \mathrm{Cov}[z_\mu,\, z_\nu]\,,\end{aligned} \tag{1.53}$$

which is also known as the *covariance* of the distribution. Note how the mean is subtracted from the random variable z_μ before taking the square in the connected version. The quantity $\widehat{\Delta z}_\mu \equiv z_\mu - \mathbb{E}\left[z_\mu\right]$ represents a **fluctuation** of the random variable around its mean. Intuitively, such fluctuations are equally likely to contribute positively as they are likely to contribute negatively, $\mathbb{E}\left[\widehat{\Delta z}_\mu\right] = \mathbb{E}\left[z_\mu\right] - \mathbb{E}\left[z_\mu\right] = 0$, so it's necessary to take the square in order to get an estimate of the magnitude of such fluctuations.

[7] Outside of this chapter, just as we'll often use the term M-point correlator rather than the term moment, we'll use the term M-point connected correlator rather than the term cumulant. When we want to refer to the moment and not the cumulant, we might sometimes say *full correlator* to contrast with *connected correlator*.

https://doi.org/10.1017/9781009023405.003 Published online by Cambridge University Press

At this point, let us restrict our focus to distributions that are invariant under a sign-flip symmetry $z_\mu \to -z_\mu$, which holds for the zero-mean Gaussian distribution (1.34). Importantly, this *parity symmetry* will also hold for the nearly-Gaussian distributions that we will study in order to describe neural networks. For all such even distributions with this symmetry, all odd moments and all odd-point connected correlators vanish.

With this restriction, the next simplest observable is the fourth cumulant or the connected four-point correlator, given by the formula

$$\begin{aligned}
&\mathbb{E}\left[z_{\mu_1} z_{\mu_2} z_{\mu_3} z_{\mu_4}\right]\big|_{\text{connected}} \\
&\quad= \mathbb{E}\left[z_{\mu_1} z_{\mu_2} z_{\mu_3} z_{\mu_4}\right] \\
&\qquad - \mathbb{E}\left[z_{\mu_1} z_{\mu_2}\right]\mathbb{E}\left[z_{\mu_3} z_{\mu_4}\right] - \mathbb{E}\left[z_{\mu_1} z_{\mu_3}\right]\mathbb{E}\left[z_{\mu_2} z_{\mu_4}\right] - \mathbb{E}\left[z_{\mu_1} z_{\mu_4}\right]\mathbb{E}\left[z_{\mu_2} z_{\mu_3}\right].
\end{aligned} \tag{1.54}$$

For the Gaussian distribution, recalling Wick's theorem (1.50), the last three terms precisely subtract off the three pairs of Wick contractions used to evaluate the first term, meaning

$$\mathbb{E}\left[z_{\mu_1} z_{\mu_2} z_{\mu_3} z_{\mu_4}\right]\big|_{\text{connected}} = 0. \tag{1.55}$$

Essentially by design, the connected four-point correlator vanishes for the Gaussian distribution, and a nonzero value signifies a deviation from Gaussian statistics.[8] In fact, the connected four-point correlator is perhaps the simplest measure of non-Gaussianity.

Now that we have a little intuition, we are as ready as we'll ever be to discuss the definition for the M-th cumulant or the M-point connected correlator. For completeness, we'll give the general definition before restricting again to distributions that are symmetric under parity $z_\mu \to -z_\mu$. The definition is *inductive* and somewhat counterintuitive, expressing the M-th moment in terms of connected correlators from degrees 1 to M:

$$\begin{aligned}
&\mathbb{E}\left[z_{\mu_1} z_{\mu_2} \cdots z_{\mu_M}\right] \\
&\quad\equiv \mathbb{E}\left[z_{\mu_1} z_{\mu_2} \cdots z_{\mu_M}\right]\big|_{\text{connected}} \\
&\qquad + \sum_{\text{all subdivisions}} \mathbb{E}\left[z_{\mu_{k_1^{[1]}}} \cdots z_{\mu_{k_{\nu_1}^{[1]}}}\right]\Bigg|_{\text{connected}} \cdots \mathbb{E}\left[z_{\mu_{k_1^{[s]}}} \cdots z_{\mu_{k_{\nu_s}^{[s]}}}\right]\Bigg|_{\text{connected}},
\end{aligned} \tag{1.56}$$

where the sum is over all the possible subdivisions of M variables into $s > 1$ clusters of sizes $(\nu_1, \ldots, \nu_s)$ as $(k_1^{[1]}, \ldots, k_{\nu_1}^{[1]}), \ldots, (k_1^{[s]}, \ldots, k_{\nu_s}^{[s]})$. By decomposing the M-th moment into a sum of products of connected correlators of degree M and lower, we see that the connected M-point correlator corresponds to a *new* type of correlation that cannot be expressed by the connected correlators of a lower degree. We saw an example of this above when discussing the connected four-point correlator as a simple measure of non-Gaussianity.

[8] In statistics, the connected four-point correlator for a single random variable z is called the *excess kurtosis* when normalized by the square of the variance. It is a natural measure of the tails of the distribution, as compared to a Gaussian distribution, and also serves as a measure of the potential for outliers. In particular, a positive value indicates fatter tails while a negative value indicates thinner tails.

https://doi.org/10.1017/9781009023405.003 Published online by Cambridge University Press

To see how this abstract definition actually works, let's revisit the examples. First, we trivially recover the relation between the mean and the one-point connected correlator:

$$\mathbb{E}\left[z_{\mu}\right]\big|_{\text{connected}} = \mathbb{E}\left[z_{\mu}\right], \tag{1.57}$$

as there is no subdivision of an $M = 1$ variable into any smaller pieces. For $M = 2$, the definition (1.56) gives

$$\begin{aligned}\mathbb{E}\left[z_{\mu_1} z_{\mu_2}\right] &= \mathbb{E}\left[z_{\mu_1} z_{\mu_2}\right]\big|_{\text{connected}} + \mathbb{E}\left[z_{\mu_1}\right]\big|_{\text{connected}}\mathbb{E}\left[z_{\mu_2}\right]\big|_{\text{connected}} \\ &= \mathbb{E}\left[z_{\mu_1} z_{\mu_2}\right]\big|_{\text{connected}} + \mathbb{E}\left[z_{\mu_1}\right]\mathbb{E}\left[z_{\mu_2}\right].\end{aligned} \tag{1.58}$$

Rearranging to solve for the connected two-point function in terms of the moments, we see that this is equivalent to our previous definition for the covariance (1.53).

At this point, let us again restrict to parity-symmetric distributions invariant under $z_{\mu} \to -z_{\mu}$, remembering that this means that all the odd-point connected correlators will vanish. For such distributions, evaluating the definition (1.56) for $M = 4$ gives

$$\begin{aligned}\mathbb{E}\left[z_{\mu_1} z_{\mu_2} z_{\mu_3} z_{\mu_4}\right] = {}& \mathbb{E}\left[z_{\mu_1} z_{\mu_2} z_{\mu_3} z_{\mu_4}\right]\big|_{\text{connected}} \\ &+ \mathbb{E}\left[z_{\mu_1} z_{\mu_2}\right]\big|_{\text{connected}}\mathbb{E}\left[z_{\mu_3} z_{\mu_4}\right]\big|_{\text{connected}} \\ &+ \mathbb{E}\left[z_{\mu_1} z_{\mu_3}\right]\big|_{\text{connected}}\mathbb{E}\left[z_{\mu_2} z_{\mu_4}\right]\big|_{\text{connected}} \\ &+ \mathbb{E}\left[z_{\mu_1} z_{\mu_4}\right]\big|_{\text{connected}}\mathbb{E}\left[z_{\mu_2} z_{\mu_3}\right]\big|_{\text{connected}}.\end{aligned} \tag{1.59}$$

Since $\mathbb{E}\left[z_{\mu_1} z_{\mu_2}\right] = \mathbb{E}\left[z_{\mu_1} z_{\mu_2}\right]\big|_{\text{connected}}$ when the mean vanishes, this is also just a rearrangement of our previous expression (1.54) for the connected four-point correlator for such zero-mean distributions.

In order to see something new, let us carry on for $M = 6$:

$$\begin{aligned}\mathbb{E}\left[z_{\mu_1} z_{\mu_2} z_{\mu_3} z_{\mu_4} z_{\mu_5} z_{\mu_6}\right] = {}& \mathbb{E}\left[z_{\mu_1} z_{\mu_2} z_{\mu_3} z_{\mu_4} z_{\mu_5} z_{\mu_6}\right]\big|_{\text{connected}} \\ &+ \mathbb{E}\left[z_{\mu_1} z_{\mu_2}\right]\big|_{\text{connected}}\mathbb{E}\left[z_{\mu_3} z_{\mu_4}\right]\big|_{\text{connected}}\mathbb{E}\left[z_{\mu_5} z_{\mu_6}\right]\big|_{\text{connected}} \\ &+ \left[14 \text{ other } (2,2,2) \text{ subdivisions}\right] \\ &+ \mathbb{E}\left[z_{\mu_1} z_{\mu_2} z_{\mu_3} z_{\mu_4}\right]\big|_{\text{connected}}\mathbb{E}\left[z_{\mu_5} z_{\mu_6}\right]\big|_{\text{connected}} \\ &+ \left[14 \text{ other } (4,2) \text{ subdivisions}\right],\end{aligned} \tag{1.60}$$

in which we have expressed the full six-point correlator in terms of a sum of products of connected two-point, four-point, and six-point correlators. Rearranging the above expression and expressing the two-point and four-point connected correlators in terms of their definitions, (1.53) and (1.54), we obtain the following expression for the connected six-point correlator:

$$\begin{aligned}&\mathbb{E}\left[z_{\mu_1} z_{\mu_2} z_{\mu_3} z_{\mu_4} z_{\mu_5} z_{\mu_6}\right]\big|_{\text{connected}} \\ &\quad = \mathbb{E}\left[z_{\mu_1} z_{\mu_2} z_{\mu_3} z_{\mu_4} z_{\mu_5} z_{\mu_6}\right] \\ &\quad\quad - \left\{\mathbb{E}\left[z_{\mu_1} z_{\mu_2} z_{\mu_3} z_{\mu_4}\right]\mathbb{E}\left[z_{\mu_5} z_{\mu_6}\right] + \left[14 \text{ other } (4,2) \text{ subdivisions}\right]\right\} \\ &\quad\quad + 2\left\{\mathbb{E}\left[z_{\mu_1} z_{\mu_2}\right]\mathbb{E}\left[z_{\mu_3} z_{\mu_4}\right]\mathbb{E}\left[z_{\mu_5} z_{\mu_6}\right] + \left[14 \text{ other } (2,2,2) \text{ subdivisions}\right]\right\}.\end{aligned} \tag{1.61}$$

https://doi.org/10.1017/9781009023405.003 Published online by Cambridge University Press

The rearrangement is useful for computational purposes, in that it's simple to first compute the moments of a distribution and then organize the resulting expressions in order to evaluate the connected correlators.

Focusing back on (1.60), it's easy to see that the connected six-point correlator vanishes for Gaussian distributions. Remembering that the connected four-point correlator also vanishes for Gaussian distributions, we see that the fifteen $(2, 2, 2)$ subdivision terms are exactly equal to the fifteen terms generated by the Wick contractions resulting from evaluating the full correlator on the left-hand side of the equation. In fact, applying the general definition of connected correlators (1.56) to the zero-mean Gaussian distribution, we see inductively that all M-point connected correlators for $M > 2$ will vanish.[9] Thus, the connected correlators are a very natural measure of how a distribution deviates from Gaussianity.

With this in mind, we can finally define a **nearly-Gaussian distribution** as a distribution for which all the connected correlators for $M > 2$ are *small*.[10] In fact, the non-Gaussian distributions that describe neural networks generally have the property that, as the network becomes wide, the connected four-point correlator becomes small and the higher-point connected correlators become even smaller. For these nearly-Gaussian distributions, a few leading connected correlators give a concise and accurate description of the distribution, just as a few leading Taylor coefficients can give a good description of a function near the point of expansion.

1.3 Nearly-Gaussian Distributions

Now that we have defined nearly-Gaussian distributions in terms of measurable deviations from Gaussian statistics, i.e., via small but nonzero connected correlators, it's natural to ask how we can link these observables to the actual functional form of the distribution, $p(z)$. We can make this connection through the action.

The **action** $S(z)$ is a function that defines a probability distribution $p(z)$ through the relation

$$p(z) \propto e^{-S(z)}. \tag{1.62}$$

In the statistics literature, the action $S(z)$ is sometimes called the *negative log probability*, but we will again follow the physics literature and call it the action. In order for (1.62) to make sense as a probability distribution, $p(z)$ needs be normalizable so that we can satisfy

$$\int d^N z \; p(z) = 1. \tag{1.63}$$

[9] To see this, note that if all the higher-point connected correlators vanish, then the definition (1.56) is equivalent to Wick's theorem (1.50), with nonzero terms in (1.56) – the subdivisions into clusters of sizes $(2, \ldots, 2)$ – corresponding exactly to the different pairings in (1.50).

[10] As we discussed in §1.1, the variance sets the scale of the Gaussian distribution. For nearly-Gaussian distributions, we require that all $2m$-point connected correlators be parametrically small when compared to an appropriate power of the variance, i.e., $|\mathbb{E}\left[z_{\mu_1} \cdots z_{\mu_{2m}}\right]|_{\text{connected}}| \ll |K_{\mu\nu}|^m$, schematically.

https://doi.org/10.1017/9781009023405.003 Published online by Cambridge University Press

That's where the *normalization factor* or **partition function**

$$Z \equiv \int d^N z \; e^{-S(z)} \tag{1.64}$$

comes in. After computing the partition function, we can define a probability distribution for a particular action $S(z)$ as

$$p(z) \equiv \frac{e^{-S(z)}}{Z}. \tag{1.65}$$

Conversely, given a probability distribution, we can associate an action, $S(z) = -\log\left[p(z)\right]$, up to an additive ambiguity: the ambiguity arises because a constant shift in the action can be offset by the multiplicative factor in the partition function.[11]

The action is a very convenient way to approximate certain types of statistical processes, particularly those with nearly-Gaussian statistics. To demonstrate this, we'll first start with the simplest action, which describes the Gaussian distribution, and then we'll show how to systematically perturb it in order to include various non-Gaussianities.

Quadratic Action and the Gaussian Distribution

Since we already know the functional form of the Gaussian distribution, it's simple to identify the action by reading it off from the exponent in (1.34),

$$S(z) = \frac{1}{2} \sum_{\mu,\nu=1}^{N} K^{\mu\nu} z_\mu z_\nu, \tag{1.66}$$

where, as a reminder, the matrix $K^{\mu\nu}$ is the inverse of the variance matrix $K_{\mu\nu}$. The partition function is given by the normalization integral (1.30) that we computed in §1.1:

$$Z = \int d^N z \; e^{-S(z)} = I_K = \sqrt{|2\pi K|}. \tag{1.67}$$

This **quadratic action** is the simplest normalizable action and serves as a starting point for defining other distributions.

As we will show next, integrals against the Gaussian distribution are a primitive for evaluating expectations against nearly-Gaussian distributions. Therefore, in order to differentiate between a general expectation and an integral against the Gaussian distribution, let us introduce a special *bra-ket*, or $\langle\cdot\rangle$ notation for computing *Gaussian* expectation values. For an observable $\mathcal{O}(z)$, define a **Gaussian expectation** as

$$\langle \mathcal{O}(z) \rangle_K \equiv \frac{1}{\sqrt{|2\pi K|}} \int \left[\prod_{\mu=1}^{N} dz_\mu\right] \exp\left(-\frac{1}{2}\sum_{\mu,\nu=1}^{N} K^{\mu\nu} z_\mu z_\nu\right) \mathcal{O}(z). \tag{1.68}$$

[11] One convention is to pick the constant such that the action vanishes when evaluated at its global minimum.

https://doi.org/10.1017/9781009023405.003 Published online by Cambridge University Press

In particular, with this notation we can write Wick's theorem as

$$\langle z_{\mu_1} z_{\mu_2} \cdots z_{\mu_{2m}} \rangle_K = \sum_{\text{all pairing}} K_{\mu_{k_1}\mu_{k_2}} \cdots K_{\mu_{k_{2m-1}}\mu_{k_{2m}}} . \tag{1.69}$$

If we're talking about a Gaussian distribution with variance $K_{\mu\nu}$, then we can use the notations $\mathbb{E}\left[\,\cdot\,\right]$ and $\langle\cdot\rangle_K$ interchangeably. If instead we're talking about a nearly-Gaussian distribution $p(z)$, then $\mathbb{E}\left[\,\cdot\,\right]$ indicates expectation with respect to $p(z)$, (1.46). However, in the evaluation of such an expectation, we'll often encounter Gaussian integrals, for which we'll use this bra-ket notation $\langle\cdot\rangle_K$ to simplify expressions.

Quartic Action and Perturbation Theory

Now, let's find an action that represents a nearly-Gaussian distribution with a connected four-point correlator that is *small* but nonvanishing,

$$\mathbb{E}\left[z_{\mu_1} z_{\mu_2} z_{\mu_3} z_{\mu_4}\right]\big|_{\text{connected}} = O(\epsilon) . \tag{1.70}$$

Here we have introduced a small parameter $\epsilon \ll 1$ and indicated that the correlator should be of order ϵ. For neural networks, we will later find that the role of the small parameter ϵ is played by $1/\text{width}$.

We should be able to generate a small connected four-point correlator by *deforming* the Gaussian distribution through the addition of a small quartic term to the quadratic action (1.66), giving us a **quartic action**

$$S(z) = \frac{1}{2} \sum_{\mu,\nu=1}^{N} K^{\mu\nu} z_\mu z_\nu + \frac{\epsilon}{4!} \sum_{\mu,\nu,\rho,\lambda=1}^{N} V^{\mu\nu\rho\lambda} z_\mu z_\nu z_\rho z_\lambda , \tag{1.71}$$

where the **quartic coupling** $\epsilon V^{\mu\nu\rho\lambda}$ is an $(N \times N \times N \times N)$-dimensional **tensor** that is completely symmetric in all of its four indices. The factor of $1/4!$ is conventional in order to compensate for the overcounting in the sum due to the symmetry of the indices. While it's not a proof of the connection, note that the coupling $\epsilon V^{\mu\nu\rho\lambda}$ has the right number of components to faithfully reproduce the four-point connected correlator (1.70), which is also an $(N \times N \times N \times N)$-dimensional symmetric tensor. At least from this perspective we're off to a good start.

Let us now establish this correspondence between the quartic coupling and connected four-point correlator. Note that in general it is impossible to compute any expectation value in closed form with a non-Gaussian action – this includes even the partition function. Instead, in order to compute the connected four-point correlator, we'll need to employ **perturbation theory** to expand everything to first order in the small parameter ϵ, each term of which can then be evaluated in a closed form. As this is easier done than said, let's get to the computations.

https://doi.org/10.1017/9781009023405.003 Published online by Cambridge University Press

To start, let's evaluate the partition function:

$$\begin{aligned}
Z &= \int \left[\prod_{\mu} dz_{\mu}\right] \, e^{-S(z)} \qquad (1.72)\\
&= \int \left[\prod_{\mu} dz_{\mu}\right] \exp\left(-\frac{1}{2}\sum_{\mu,\nu} K^{\mu\nu} z_{\mu} z_{\nu} - \frac{\epsilon}{24}\sum_{\rho_1,\dots,\rho_4} V^{\rho_1\rho_2\rho_3\rho_4} z_{\rho_1} z_{\rho_2} z_{\rho_3} z_{\rho_4}\right)\\
&= \sqrt{|2\pi K|}\left\langle \exp\left(-\frac{\epsilon}{24}\sum_{\rho_1,\dots,\rho_4} V^{\rho_1\rho_2\rho_3\rho_4} z_{\rho_1} z_{\rho_2} z_{\rho_3} z_{\rho_4}\right)\right\rangle_K .
\end{aligned}$$

In the second line we inserted our expression for the quartic action (1.71), and in the last line we used our bra-ket notation (1.68) for a Gaussian expectation with variance $K_{\mu\nu}$. As advertised, the Gaussian expectation in the final line cannot be evaluated in closed form. However, since our parameter ϵ is small, we can Taylor-expand the exponential to express the partition function as a sum of simple Gaussian expectations that can be evaluated using Wick's theorem (1.69):

$$\begin{aligned}
Z &= \sqrt{|2\pi K|}\left\langle 1 - \frac{\epsilon}{24}\sum_{\rho_1,\dots,\rho_4} V^{\rho_1\rho_2\rho_3\rho_4} z_{\rho_1} z_{\rho_2} z_{\rho_3} z_{\rho_4} + O\!\left(\epsilon^2\right)\right\rangle_K \qquad (1.73)\\
&= \sqrt{|2\pi K|}\left[1 - \frac{\epsilon}{24}\sum_{\rho_1,\dots,\rho_4} V^{\rho_1\rho_2\rho_3\rho_4} \left\langle z_{\rho_1} z_{\rho_2} z_{\rho_3} z_{\rho_4}\right\rangle_K + O\!\left(\epsilon^2\right)\right]\\
&= \sqrt{|2\pi K|}\left[1 - \frac{\epsilon}{24}\sum_{\rho_1,\dots,\rho_4} V^{\rho_1\rho_2\rho_3\rho_4} \left(K_{\rho_1\rho_2}K_{\rho_3\rho_4} + K_{\rho_1\rho_3}K_{\rho_2\rho_4} + K_{\rho_1\rho_4}K_{\rho_2\rho_3}\right) + O\!\left(\epsilon^2\right)\right]\\
&= \sqrt{|2\pi K|}\left[1 - \frac{1}{8}\epsilon\sum_{\rho_1,\dots,\rho_4} V^{\rho_1\rho_2\rho_3\rho_4} K_{\rho_1\rho_2}K_{\rho_3\rho_4} + O\!\left(\epsilon^2\right)\right].
\end{aligned}$$

In the final line, we were able to combine the three K^2 terms together by using the total symmetry of the quartic coupling and then relabeling some of the summed-over dummy indices.

Similarly, let's evaluate the two-point correlator:

$$\begin{aligned}
\mathbb{E}\left[z_{\mu_1} z_{\mu_2}\right] &= \frac{1}{Z}\int \left[\prod_{\mu} dz_{\mu}\right] \, e^{-S(z)} \, z_{\mu_1} z_{\mu_2} \qquad (1.74)\\
&= \frac{\sqrt{|2\pi K|}}{Z}\left\langle z_{\mu_1} z_{\mu_2} \exp\left(-\frac{\epsilon}{24}\sum_{\rho_1,\dots,\rho_4} V^{\rho_1\rho_2\rho_3\rho_4} z_{\rho_1} z_{\rho_2} z_{\rho_3} z_{\rho_4}\right)\right\rangle_K\\
&= \frac{\sqrt{|2\pi K|}}{Z}\left[\left\langle z_{\mu_1} z_{\mu_2}\right\rangle_K - \frac{\epsilon}{24}\sum_{\rho_1,\dots,\rho_4} V^{\rho_1\rho_2\rho_3\rho_4} \left\langle z_{\mu_1} z_{\mu_2} z_{\rho_1} z_{\rho_2} z_{\rho_3} z_{\rho_4}\right\rangle_K + O\!\left(\epsilon^2\right)\right]
\end{aligned}$$

https://doi.org/10.1017/9781009023405.003 Published online by Cambridge University Press

$$
\begin{aligned}
&= \left[1 + \frac{1}{8}\epsilon \sum_{\rho_1,\ldots,\rho_4} V^{\rho_1\rho_2\rho_3\rho_4} K_{\rho_1\rho_2} K_{\rho_3\rho_4}\right] K_{\mu_1\mu_2} \\
&\quad - \frac{\epsilon}{24} \sum_{\rho_1,\ldots,\rho_4} V^{\rho_1\rho_2\rho_3\rho_4} \left(3K_{\mu_1\mu_2}K_{\rho_1\rho_2}K_{\rho_3\rho_4} + 12K_{\mu_1\rho_1}K_{\mu_2\rho_2}K_{\rho_3\rho_4}\right) + O\!\left(\epsilon^2\right) \\
&= K_{\mu_1\mu_2} - \frac{\epsilon}{2} \sum_{\rho_1,\ldots,\rho_4} V^{\rho_1\rho_2\rho_3\rho_4} K_{\mu_1\rho_1}K_{\mu_2\rho_2}K_{\rho_3\rho_4} + O\!\left(\epsilon^2\right).
\end{aligned}
$$

Here, to go from the first line to the second line we inserted our expression for the quartic action (1.71) and rewrote the integral as a Gaussian expectation. Then, after expanding in ϵ to first order, in the next step we substituted (1.73) for the partition function Z in the denominator and expanded $1/Z$ to the first order in ϵ using the expansion $1/(1-x) = 1 + x + O(x^2)$. In that same step, we also noted that, of the fifteen terms coming from the Gaussian expectation $\langle z_{\mu_1} z_{\mu_2} z_{\rho_1} z_{\rho_2} z_{\rho_3} z_{\rho_4}\rangle_K$, there are three ways in which z_{μ_1} and z_{μ_2} contract with each other but twelve ways in which they don't. Given again the symmetry of $V^{\rho_1\rho_2\rho_3\rho_4}$, this is the only distinction that matters.

Finally, let's compute the full four-point correlator:

$$
\begin{aligned}
&\mathbb{E}\left[z_{\mu_1} z_{\mu_2} z_{\mu_3} z_{\mu_4}\right] = \frac{1}{Z} \int \left[\prod_{\mu} dz_{\mu}\right] e^{-S(z)}\, z_{\mu_1} z_{\mu_2} z_{\mu_3} z_{\mu_4} \qquad (1.75) \\
&= \frac{\sqrt{|2\pi K|}}{Z} \left[\langle z_{\mu_1} z_{\mu_2} z_{\mu_3} z_{\mu_4}\rangle_K - \frac{\epsilon}{24} \sum_{\rho_1,\ldots,\rho_4} V^{\rho_1\rho_2\rho_3\rho_4} \langle z_{\mu_1} z_{\mu_2} z_{\mu_3} z_{\mu_4} z_{\rho_1} z_{\rho_2} z_{\rho_3} z_{\rho_4}\rangle_K + O\!\left(\epsilon^2\right)\right] \\
&= \left[1 + \frac{1}{8}\epsilon \sum_{\rho_1,\ldots,\rho_4} V^{\rho_1\rho_2\rho_3\rho_4} K_{\rho_1\rho_2} K_{\rho_3\rho_4}\right] \left[K_{\mu_1\mu_2}K_{\mu_3\mu_4} + K_{\mu_1\mu_3}K_{\mu_2\mu_4} + K_{\mu_1\mu_4}K_{\mu_2\mu_3}\right] \\
&\quad - \frac{\epsilon}{24} \sum_{\rho_1,\ldots,\rho_4} V^{\rho_1\rho_2\rho_3\rho_4} \\
&\quad \times \Big(3K_{\mu_1\mu_2}K_{\mu_3\mu_4}K_{\rho_1\rho_2}K_{\rho_3\rho_4} + 12K_{\mu_1\rho_1}K_{\mu_2\rho_2}K_{\mu_3\mu_4}K_{\rho_3\rho_4} + 12K_{\mu_3\rho_1}K_{\mu_4\rho_2}K_{\mu_1\mu_2}K_{\rho_3\rho_4} \\
&\quad + 3K_{\mu_1\mu_3}K_{\mu_2\mu_4}K_{\rho_1\rho_2}K_{\rho_3\rho_4} + 12K_{\mu_1\rho_1}K_{\mu_3\rho_2}K_{\mu_2\mu_4}K_{\rho_3\rho_4} + 12K_{\mu_2\rho_1}K_{\mu_4\rho_2}K_{\mu_1\mu_3}K_{\rho_3\rho_4} \\
&\quad + 3K_{\mu_1\mu_4}K_{\mu_2\mu_3}K_{\rho_1\rho_2}K_{\rho_3\rho_4} + 12K_{\mu_1\rho_1}K_{\mu_4\rho_2}K_{\mu_2\mu_3}K_{\rho_3\rho_4} + 12K_{\mu_2\rho_1}K_{\mu_3\rho_2}K_{\mu_1\mu_4}K_{\rho_3\rho_4} \\
&\quad + 24K_{\mu_1\rho_1}K_{\mu_2\rho_2}K_{\mu_3\rho_3}K_{\mu_4\rho_4}\Big) + O\!\left(\epsilon^2\right).
\end{aligned}
$$

To go from the first line to the second line, we inserted our expression for the quartic action (1.71), expanded to first order in ϵ, and rewrote in the bra-ket notation (1.68). On the third line, we again substituted in the expression (1.73) for the partition function Z, expanded $1/Z$ to first order in ϵ, and then used Wick's theorem (1.69) to evaluate the fourth and eighth Gaussian moments. (Yes, we know that the evaluation of $\langle z_{\mu_1} z_{\mu_2} z_{\mu_3} z_{\mu_4} z_{\rho_1} z_{\rho_2} z_{\rho_3} z_{\rho_4}\rangle_K$ is not fun. The breakdown of the terms depends again on whether or not the μ-type indices are contracted with the ρ-type indices or not.) We can simplify this expression by noticing that some terms cancel due to $\frac{1}{8} - \frac{3}{24} = 0$, and some other terms can be nicely regrouped once we notice through the expression for the two-point correlator (1.74) that

https://doi.org/10.1017/9781009023405.003 Published online by Cambridge University Press

$$K_{\mu_1\mu_2}K_{\mu_3\mu_4} - \frac{\epsilon}{24}\sum_{\rho_1,\ldots,\rho_4} V^{\rho_1\rho_2\rho_3\rho_4}\left(12K_{\mu_1\rho_1}K_{\mu_2\rho_2}K_{\mu_3\mu_4}K_{\rho_3\rho_4} + 12K_{\mu_3\rho_1}K_{\mu_4\rho_2}K_{\mu_1\mu_2}K_{\rho_3\rho_4}\right)$$
$$= \mathbb{E}\left[z_{\mu_1}z_{\mu_2}\right]\mathbb{E}\left[z_{\mu_3}z_{\mu_4}\right] + O\left(\epsilon^2\right), \tag{1.76}$$

yielding in the end

$$\begin{aligned}&\mathbb{E}\left[z_{\mu_1}z_{\mu_2}z_{\mu_3}z_{\mu_4}\right] \\ &\quad= \mathbb{E}\left[z_{\mu_1}z_{\mu_2}\right]\mathbb{E}\left[z_{\mu_3}z_{\mu_4}\right] + \mathbb{E}\left[z_{\mu_1}z_{\mu_3}\right]\mathbb{E}\left[z_{\mu_2}z_{\mu_4}\right] + \mathbb{E}\left[z_{\mu_1}z_{\mu_4}\right]\mathbb{E}\left[z_{\mu_2}z_{\mu_3}\right] \\ &\qquad - \epsilon\sum_{\rho_1,\ldots,\rho_4} V^{\rho_1\rho_2\rho_3\rho_4}K_{\mu_1\rho_1}K_{\mu_2\rho_2}K_{\mu_3\rho_3}K_{\mu_4\rho_4} + O\left(\epsilon^2\right).\end{aligned} \tag{1.77}$$

Given the full four-point correlator (1.75) and the two-point correlator (1.74), we can finally evaluate the connected four-point correlator (1.54) as

$$\mathbb{E}\left[z_{\mu_1}z_{\mu_2}z_{\mu_3}z_{\mu_4}\right]\Big|_{\text{connected}} = -\epsilon\sum_{\rho_1,\ldots,\rho_4} V^{\rho_1\rho_2\rho_3\rho_4}K_{\mu_1\rho_1}K_{\mu_2\rho_2}K_{\mu_3\rho_3}K_{\mu_4\rho_4} + O\left(\epsilon^2\right). \tag{1.78}$$

This makes explicit the relationship between the connected four-point correlator and the quartic coupling in the action, when both are small. We see that for the nearly-Gaussian distribution realized by the quartic action (1.71), the distribution is – as promised – *nearly* Gaussian: the strength of the coupling $\epsilon V^{\rho_1\rho_2\rho_3\rho_4}$ directly controls the distribution's deviation from Gaussian statistics, as measured by the connected four-point correlator. This also shows that the four-index tensor $V^{\rho_1\rho_2\rho_3\rho_4}$ creates nontrivial correlations between the components $z_{\rho_1}z_{\rho_2}z_{\rho_3}z_{\rho_4}$ that cannot otherwise be built up by the correlation $K_{\mu\nu}$ in any pair of random variables $z_\mu z_\nu$.

Finally, note that the connected two-point correlator (1.74) – i.e., the covariance of this nearly-Gaussian distribution – is also shifted from its Gaussian value of $K_{\mu_1\mu_2}$ by the quartic coupling $\epsilon V^{\rho_1\rho_2\rho_3\rho_4}$. Thus, the nearly-Gaussian deformation not only creates complicated patterns of four-point correlation as measured by the connected four-point correlator (1.78), but it also can modify the details of the Gaussian two-point correlation.

Now that we see how to compute the statistics of a nearly-Gaussian distribution, let's take a step back and think about what made this possible. We can perform these perturbative calculations any time there exists in the problem a dimensionless parameter ϵ that is small $\epsilon \ll 1$, but nonzero $\epsilon > 0$. This makes perturbation theory an extremely powerful tool for theoretical analysis any time a problem has any extreme scales, small *or* large.

Importantly, this is directly relevant to theoretically understanding neural networks in practice. As we will explain in the following chapters, real networks have a parameter n – the number of neurons in a layer – that is typically large $n \gg 1$, but certainly not infinite $n < \infty$. This means that we can expand the distributions that describe such networks in the inverse of the large parameter as $\epsilon = 1/n$. Indeed, when the parameter n is large – as is typical in practice – the distributions that describe neural networks become nearly-Gaussian and thus theoretically tractable. This type of expansion is known as the **$1/n$ expansion** or **large-n expansion** and will be one of our main tools for learning the principles of deep learning theory.

https://doi.org/10.1017/9781009023405.003 Published online by Cambridge University Press

Aside: Statistical Independence and Interactions

The quartic action (1.71) is one of the simplest models of an **interacting theory**. We showed this explicitly by connecting the quartic coupling to the non-Gaussian statistics of the nonvanishing connected four-point correlator. Here, let us try to offer an intuitive meaning of *interaction* by appealing to the notion of *statistical independence*.

Recall from probability theory that two random variables x and y are statistically independent if their joint distribution factorizes as

$$p(x, y) = p(x)p(y). \tag{1.79}$$

For the Gaussian distribution, if the variance matrix $K_{\mu\nu}$ is diagonal, there is no correlation at all between different components of z_μ; they are manifestly statistically independent from each other.

Even if $K_{\mu\nu}$ is not diagonal, we can still unwind the correlation of a Gaussian distribution by rotating to the right basis. As discussed in §1.1, there always exists an orthogonal matrix O that diagonalizes the covariance as $(OKO^T)_{\mu\nu} = \lambda_\mu \delta_{\mu\nu}$. In terms of the variables $u_\mu \equiv (Oz)_\mu$, the distribution looks like

$$p(z) = \frac{1}{\sqrt{|2\pi K|}} \exp\left(-\sum_{\mu=1}^{N} \frac{u_\mu^2}{2\lambda_\mu} \right) = \prod_{\mu=1}^{N} \left(\frac{e^{-\frac{u_\mu^2}{2\lambda_\mu}}}{\sqrt{2\pi\lambda_\mu}} \right) = p(u_1)\cdots p(u_N). \tag{1.80}$$

Thus, we see that in the u-coordinate basis, the original multivariable Gaussian distribution factorizes into N single-variable Gaussians that are statistically independent.

We also see that in terms of the action, statistical independence is characterized by the action breaking into a sum over separate terms. This unwinding of interaction between variables is generically impossible when there are nonzero non-Gaussian couplings. For instance, there are $\sim N^2$ components of an orthogonal matrix $O_{\mu\nu}$ to change basis, while there are $\sim N^4$ components of the quartic coupling $\epsilon V^{\mu\nu\rho\lambda}$ that correlate random variables, so it is generically impossible to re-express the quartic action as a sum of functions of N different variables. Since the action cannot be put into a sum over N separate terms, the joint distribution cannot factorize, and the components will not be independent from each other. Thus, it is impossible to factor the nearly-Gaussian distribution into the product of N statistically independent distributions. In this sense, what is meant by *interaction* is the breakdown of *statistical independence*.[12]

[12] An astute reader might wonder if there is any interaction when we consider a single-variable distribution with $N = 1$, since there are no other variables to interact with. For nearly-Gaussian distributions, even if $N = 1$, we saw in (1.74) that the variance of the distribution is shifted from its Gaussian value, K, and depends on the quartic coupling ϵV. In physics, we say that this shift is due to the *self-interaction* induced by the quartic coupling ϵV, since it modifies the value of observables from the *free* Gaussian theory that we are comparing to, even though there's no notion of statistical independence to appeal to here.

Said another way, even though the action just involves one term, such a non-Gaussian distribution does not have a closed-form solution for its partition function or correlators; i.e., there's no trick that lets us compute integrals of the form $e^{-S(z)}$ exactly, when $S(z) = \frac{z^2}{2K} + \frac{1}{4!}\epsilon V z^4$. This means that we still have to make use of perturbation theory to analyze the self-interaction in such distributions.

https://doi.org/10.1017/9781009023405.003 Published online by Cambridge University Press

Nearly-Gaussian Actions

Having given a concrete example in which we illustrated how to deform the quadratic action to realize the simplest nearly-Gaussian distribution, we now give a more general perspective on nearly-Gaussian distributions. In what follows, we will continue to require that our distributions are invariant under the parity symmetry that takes $z_\mu \to -z_\mu$. In the action representation, this corresponds to including only terms of even degree.[13]

With that caveat in mind, though otherwise very generally, we can express a **non-Gaussian distribution** by *deforming* the Gaussian action as

$$S(z) = \frac{1}{2}\sum_{\mu,\nu=1}^{N} K^{\mu\nu} z_\mu z_\nu + \sum_{m=2}^{k} \frac{1}{(2m)!} \sum_{\mu_1,\ldots,\mu_{2m}=1}^{N} s^{\mu_1\cdots\mu_{2m}} z_{\mu_1}\cdots z_{\mu_{2m}}\,, \tag{1.81}$$

where the factor of $1/(2m)!$ is conventional in order to compensate for the overcounting in the sum due to the implied symmetry of the indices $\mu_1, \ldots, \mu_{2m}$ in the coefficients $s^{\mu_1\cdots\mu_{2m}}$, given the permutation symmetry of the product of variables $z_{\mu_1}\cdots z_{\mu_{2m}}$. The number of terms in the non-Gaussian part of the action is controlled by the integer k. If k were unbounded, then $S(z)$ would be an arbitrary even function, and $p(z)$ could be any parity-symmetric distribution. The action is most useful when the expanded polynomial $S(z)$ truncated to reasonably small degree k – like $k = 2$ for the quartic action – yields a good representation for the statistical process of interest.

The coefficients $s^{\mu_1\cdots\mu_{2m}}$ are generally known as **non-Gaussian couplings**, and they control the **interactions** of the z_μ.[14] In particular, there is a direct correspondence between the product of the specific components z_μ that appear together in the action and the presence of connected correlation between those variables, with the degree of the term in (1.81) directly contributing to connected correlators of that degree. We saw an example of this in (1.78), which connected the quartic term to the connected four-point correlator. In this way, the couplings give a very direct way of controlling the degree and pattern of non-Gaussian correlation, and the overall degree of the action offers a way of systematically including more and more complicated patterns of such correlations.

If you recall from §1.2, we defined nearly-Gaussian distributions as ones for which all these connected correlators are small. Equivalently, from the action perspective, a nearly-Gaussian distribution is a non-Gaussian distribution with an action of the form (1.81) for which all the couplings $s^{\mu_1\cdots\mu_{2m}}$ are parametrically small for all $1 \le m \le k$:

$$\left|s^{\mu_1\cdots\mu_{2m}}\right| \ll \left|K^{\mu\nu}\right|^m, \tag{1.82}$$

[13] The imposition of such a parity symmetry, and thus the absence of odd-degree terms in the action, means that all of the odd moments and hence all of the odd-point connected correlators will vanish.

[14] In a similar vein, the coefficient $K^{\mu\nu}$ in the action is sometimes called a *quadratic coupling* since the *coupling* of the component z_μ with the component z_ν in the quadratic action leads to a nontrivial *correlation*, i.e., $\mathrm{Cov}[z_\mu, z_\nu] = K_{\mu\nu}$.

https://doi.org/10.1017/9781009023405.003 Published online by Cambridge University Press

where this equation is somewhat schematic given the mismatch of the indices.[15] Importantly, the comparison is with an appropriate power of the inverse variance or quadratic coupling $K^{\mu\nu}$ since, as we already explained, the variance sets the scale of the Gaussian distribution to which we are comparing these nearly-Gaussian distributions.

As we will see in §4, wide neural networks are described by nearly-Gaussian distributions. In particular, we will find that such networks are described by a special type of nearly-Gaussian distribution where the connected correlators are *hierarchically* small, scaling as

$$\mathbb{E}\left[z_{\mu_1}\cdots z_{\mu_{2m}}\right]\Big|_{\text{connected}} = O(\epsilon^{m-1}), \tag{1.83}$$

with the same parameter ϵ controlling the different scalings for each of the $2m$-point connected correlators. Importantly, the non-Gaussianities coming from higher-point connected correlators become parametrically less important as ϵ becomes smaller.

This means that for a nearly-Gaussian distribution with hierarchical scalings (1.83), we can consistently approximate the distribution by *truncating* the action at some fixed order in ϵ. To be concrete, we can use an action of the form (1.81) to faithfully represent all the correlations up to order $O(\epsilon^{k-1})$, neglecting connected correlations of order $O(\epsilon^k)$ and higher. The resulting action offers a useful and effective description for the statistical process of interest, as long as ϵ is small enough and k is high enough that $O(\epsilon^k)$ is negligible.

In practice, a quartic action (1.71) truncated to $k = 2$ will let us model realistic finite-width neural networks. This quartic action captures the important *qualitative* difference between nearly-Gaussian distributions and the Gaussian distribution, incorporating nontrivial interactions between the different components of the random variable. In addition,

[15] This schematic equation is, nonetheless, dimensionally consistent. To support that remark, let us give a brief introduction to *dimensional analysis*: let the random variable z_μ have dimension ζ, which we denote as $[z_\mu] = \zeta^1$. By *dimension*, you should have in mind something like a unit of length, so, e.g., we read the expression $[z_\mu] = \zeta^1$ as "a component of z is measured in units of ζ." The particular units are arbitrary: e.g., for length, we can choose between `meters` or `inches` or `parsecs` as long as we use a unit of length but *not*, say, `meters`^{2}, which instead would be a unit of area. Importantly, we cannot add or equate quantities that have different units: it doesn't make any logical sense to add a length to an area. This is similar to the concept of *type safety* in computer science: e.g., we should not add a type `str` variable to a type `int` variable.

Now, since the action $S(z)$ is the argument of an exponential $p(z) \propto e^{-S(z)}$, it must be *dimensionless*; otherwise, the exponential $e^{-S} = 1 - S + \frac{S^2}{2} + \cdots$ would violate the addition rule that we just described. From this dimensionless requirement for the action, we surmise that the inverse of the covariance matrix has dimension $[K^{\mu\nu}] = \zeta^{-2}$ and that the covariance itself has dimension $[K_{\mu\nu}] = \zeta^2$. Similarly, all the non-Gaussian couplings in (1.81) have dimensions $[s^{\mu_1\cdots\mu_{2m}}] = \zeta^{-2m}$. Thus, both sides of (1.82) have the same dimension, making this equation dimensionally consistent.

Even more concretely, consider the quartic action (1.71). If we let the tensorial part of the quartic coupling have dimensions $[V^{\mu\nu\rho\lambda}] = \zeta^{-4}$, then the parameter ϵ is dimensionless, as claimed. This means that we can consistently compare ϵ to unity, and its parametric smallness $\epsilon \ll 1$ means that the full quartic coupling $\epsilon V^{\mu\nu\rho\lambda}$ is much smaller than the square of the quadratic coupling and that the connected four-point correlator (1.78) is much smaller than the square of the connected two-point correlator (1.74).

https://doi.org/10.1017/9781009023405.003 Published online by Cambridge University Press

the difference between the statistics (1.83) of a nearly-Gaussian distribution truncated to $O(\epsilon)$ versus one truncated to $O(\epsilon^2)$ is mostly *quantitative*: in both cases there are nontrivial non-Gaussian correlations, but the pattern of higher-order correlation differs only in a small way, with the difference suppressed as $O(\epsilon^2)$. In this way, the distribution represented by the quartic action is complex enough to capture the most salient non-Gaussian effects in neural networks while still being simple enough to be analytically tractable.

https://doi.org/10.1017/9781009023405.003 Published online by Cambridge University Press

https://doi.org/10.1017/9781009023405.003 Published online by Cambridge University Press