Mathematical preliminaries

Lecture notes for NNDL2025 course

Aapo Hyvärinen
University of Helsinki

March 5, 2025

1 Linear algebra

1.1 Matrices

In matrix algebra, linear transformations and linear systems of equations can be succinctly expressed by matrices. A matrix $\mathbf{M}$ of size $n_1 \times n_2$ is a collection of real numbers arranged into n_1 rows and n_2 columns. The single entries are denoted by m_{ij} where i is the row and j is the column. Thus,

$$\mathbf{M} = \begin{bmatrix} m_{11} & m_{12} & \dots & m_{1n_2} \\ \vdots & & & \vdots \\ m_{n_1 1} & m_{n_1 2} & \dots & m_{n_1 n_2} \end{bmatrix} \tag{1}$$

Consider a vector $\mathbf{z}$ in an n_2 dimensional real space. A linear transformation of the vector $\mathbf{z}$, as defined by the matrix $\mathbf{M}$, is then denoted by

$$\mathbf{y} = \mathbf{M}\mathbf{z} \tag{2}$$

which is basically a short-cut notation for

$$y_i = \sum_{j=1}^{n_2} m_{ij} z_j, \text{ for all } i \tag{3}$$

This operation is also the definition of the product of a matrix and a vector.

If we concatenate two linear transformations, defining

$$\mathbf{s} = \mathbf{N}\mathbf{y} \tag{4}$$

we get another linear transformation. The matrix $\mathbf{P}$ that expresses this linear transformation is obtained by

$$p_{ij} = \sum_{k=1}^{n_1} n_{ik} m_{kj} \tag{5}$$

This is the definition of the product of two matrices: the new matrix $\mathbf{P}$ is denoted by

$$\mathbf{P} = \mathbf{N}\mathbf{M} \tag{6}$$

The definition is quite useful, because it means we can multiply matrices and vectors in any order when we compute $\mathbf{s}$. In fact, we have

$$\mathbf{s} = \mathbf{N}\mathbf{y} = \mathbf{N}(\mathbf{M}\mathbf{z}) = (\mathbf{N}\mathbf{M})\mathbf{z} \tag{7}$$

Another important operation with matrices is the transpose. The transpose $\mathbf{M}^T$ of a matrix $\mathbf{M}$ is the matrix where the indices are exchanged: the i, j-th entry of $\mathbf{M}^T$ is m_{ji}. A matrix $\mathbf{M}$ is called symmetric if $m_{ij} = m_{ji}$, i.e., if $\mathbf{M}$ equals its transpose.

1.2 Determinant

The determinant answers the question: how are volumes changed when the data space is transformed by the linear transformation $\mathbf{M}$? That is, if $\mathbf{z}$ takes values in a cube whose edges are all of length one, what is the volume of the set of the values $\mathbf{y}$ in Equation (2)?. The answer is given by the absolute value of the determinant, denoted by $|\det(\mathbf{M})|$, or sometimes simply as $|\mathbf{M}|$.

Two basic properties of the determinant are very useful.

1. The determinant of a product is the product of the determinants: $\det(\mathbf{MN}) = \det(\mathbf{M})\det(\mathbf{N})$. If you think that the first transformation changes the volume by a factor or 2 and the second by a factor of 3, it is obvious that when you do both transformation, the change in volume is by a factor of $2 \times 3 = 6$.
2. The determinant of a diagonal matrix equals the product of the diagonal elements. If you think in two dimensions, a diagonal matrix simply stretches one coordinate by a factor of, say 2, and the other coordinate by a factor of, say 3, so the volume of a square of area equal to 1 then becomes $2 \times 3 = 6$.

(In Section 1.4 we will see a further important result on the determinant of an orthogonal matrix).

1.3 Inverse

If a linear transformation in Equation (2) does not change the dimension of the data the transformation can usually be inverted. That is, Equation (2) can usually be solved for $\mathbf{z}$: if we know $\mathbf{M}$ and $\mathbf{y}$, we can compute what was the original $\mathbf{z}$. This is the case if the linear transformation is invertible — a technical condition that is almost always true.

In matrix algebra, the coefficients needed to solve an equation can be obtained by computing the inverse of the matrix $\mathbf{M}$, denoted by $\mathbf{M}^{-1}$. So, solving for $\mathbf{z}$ in (2) we have

$$\mathbf{z} = \mathbf{M}^{-1}\mathbf{y} \tag{8}$$

A multitude of numerical methods for computing the inverse of the matrix exist.

Note that the determinant of the inverse matrix is simply the inverse of the determinant: $\det(\mathbf{M}^{-1}) = 1/\det(\mathbf{M})$. Logically, if the transformation changes the volume by a factor of 5 (say), then the inverse must change the volume by a factor of 1/5.

The product of a matrix with its inverse equals the *identity matrix* $\mathbf{I}$:

$$\mathbf{M}\mathbf{M}^{-1} = \mathbf{M}^{-1}\mathbf{M} = \mathbf{I} \tag{9}$$

The identity matrix is a matrix whose diagonal elements are all ones and the off-diagonal elements are all zero. It corresponds to the identity transformation, i.e., a transformation which does not change the vector. This means we have

$$\mathbf{I}\mathbf{z} = \mathbf{z} \tag{10}$$

for any $\mathbf{z}$.

1.4 Orthogonality

A linear transformation, or equivalently a matrix, is called *orthogonal* if it does not change the norm of the vector. Likewise, a matrix $\mathbf{A}$ is called orthogonal if the corresponding transformation is orthogonal. An equivalent condition for orthogonality is

$$\mathbf{A}^T\mathbf{A} = \mathbf{I} \tag{11}$$

If you think about the meaning of this equation in detail, you will realize that it says two things: the column vectors of the matrix $\mathbf{A}$ are orthogonal, and all normalized to unit norm. This is because the entries in the matrix $\mathbf{A}^T\mathbf{A}$ are the dot-products $\mathbf{a}_i^T\mathbf{a}_j$ between the column vectors of the matrix $\mathbf{A}$.

Equation (11) shows that the inverse of an orthogonal matrix (or an orthogonal transformation) is trivial to compute: we just need to rearrange the entries by taking the transpose.

The compound transformation of two orthogonal transformation is orthogonal. This is natural since if neither of the transformations changes the norm of the vector, then doing one transformation after the other does not change the norm either.

The determinant of an orthogonal matrix is equal to plus or minus one. This is because because an orthogonal transformation does not change volumes, so the absolute value has to be one. The change in sign is related to reflections. Think of multiplying one-dimensional data by -1: This does not change the "volumes", but "reflects" the data with respect to 0, and corresponds to a determinant of -1.

1.5 Eigenvalues and eigenvectors

If $\mathbf{z}$ fulfills the equation

$$\mathbf{M}\mathbf{z} = \lambda\mathbf{z} \tag{12}$$

for some scalar quantity λ, $\mathbf{z}$ is called an eigenvector of $\mathbf{M}$, and λ is called the corresponding eigenvalue. To each eigenvector corresponds one eigenvalue by definition, but the same eigenvalue can correspond to many eigenvectors.

Typically, an $n \times n$ matrix will have n linearly independent eigenvectors, although some complicated conditions are necessary for this to be exactly true. In general, eigenvectors can be complex-valued, but we will in this course always consider eigenvectors of symmetric matrices, which are guaranteed to be real-valued, and furthermore, orthogonal to each other.

Note that if $\mathbf{z}$ is an eigenvector of $\mathbf{M}$, $\alpha\mathbf{z}$ is also, for any scalar α. Thus, the eigenvectors are defined only up to a scaling. Typically, they are scaled to have unit norm in practical computations, and this is what will always be done in these lecture notes.

One intuitive definition of an eigenvector is that its "direction" is not changed by the linear transformation $\mathbf{M}$, it is only rescaled. However, in this course this interpretation is not very useful. Here, the relevant intuitive interpretation of eigenvalues and vectors is related to the eigenvalue decomposition of a matrix, which will be treated next.

Let us assume that the matrix $\mathbf{M}$ is symmetric. Then, we can decompose the matrix into the following product, called the eigenvalue decomposition:

$$\mathbf{M} = \mathbf{U}\mathbf{D}\mathbf{U}^T \tag{13}$$

where $\mathbf{U}$ is an orthogonal matrix , and $\mathbf{D} = \mathrm{diag}(\lambda_1, \ldots, \lambda_m)$ is diagonal. The columns of $\mathbf{U}$ are the eigenvectors of $\mathbf{M}$, and the λ_i are the corresponding eigenvalues. Very efficient numerical algorithms exist for computing the eigenvalue decomposition of a matrix.

The meaning of the eigenvalue decomposition is that by changing the coordinate frame (by $\mathbf{U}$), or rotating the space, any symmetric matrix can be extremely simplified, making it diagonal.

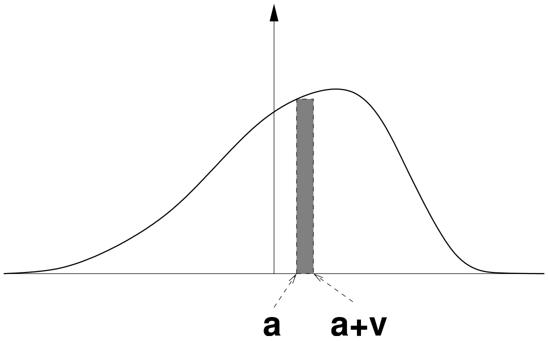

Figure 1: The pdf of a random variable at a point a gives the probability that the random variable takes a value in a small interval $[a, a+v]$, divided by the length of that interval, i.e. v. In other words, the shaded area, equal to $p(a)v$, gives the probability that that the variable takes a value in that interval.

2 Probability theory and statistics

2.1 Multivariate probability distributions

In this chapter, we will denote random variables by $z_1, z_2, \ldots, z_n$ and $s_1, s_2, \ldots, s_n$ for some number n. Taken together, the random variables $z_1, z_2, \ldots, z_n$ form an n-dimensional random vector which we denote by $\mathbf{z}$:

$$\mathbf{z} = \begin{pmatrix} z_1 \\ z_2 \\ \vdots \\ z_n \end{pmatrix} \tag{14}$$

Likewise, the variables $s_1, s_2, \ldots, s_n$ can be collected to a random vector, denoted by $\mathbf{s}$.

A probability distribution of a random vector such as $\mathbf{z}$ is usually represented using a *probability density function* (pdf). The pdf at a point in the n-dimensional space is denoted by $p_{\mathbf{z}}$.

The definition of the pdf of a multidimensional random vector is a simple generalization of the definition of the pdf of a random variable in one dimension. Let us first recall that definition. Denote by z a random variable. The idea is that we take a small number v, and look at the probability that z takes a value in the interval $[a, a+v]$ for any given a. Then we divide that probability by v, and that is the value of the probability density function at the point a. That is

$$p_z(a) = \frac{P(z \text{ is in } [a, a+v])}{v} \tag{15}$$

This principle is illustrated in Fig. 1. Rigorously speaking, we should take the limit of an infinitely small v in this definition.

This principle is simple to generalize to the case of an n-dimensional random vector. The value of the pdf function at a point, say $\mathbf{a} = (a_1, a_2, \ldots, a_n)$, gives the probability that an observation of $\mathbf{z}$ belongs to a small neighbourhood of the point $\mathbf{a}$, divided by the volume of the neighbourhood. Computing the probability that the values of each z_i are between the values of a_i and $a_i + v$, we obtain

$$p_{\mathbf{z}}(\mathbf{a}) = \frac{P(z_i \text{ is in } [a_i, a_i + v] \text{ for all } i)}{v^n} \tag{16}$$

where v^n is the volume of the n-dimensional cube whose edges all have length v. Again, rigorously speaking, this equation is true only in the limit of infinitely small v.

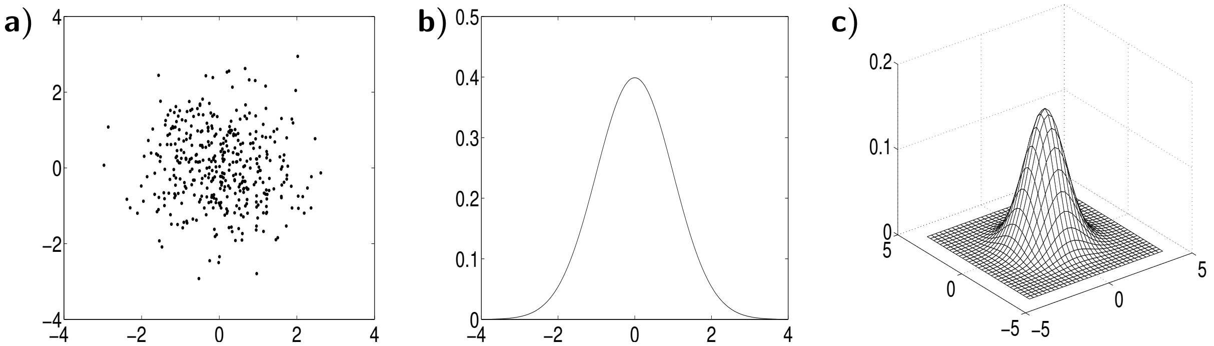

Figure 2: **a)** Scatter plot of the two-dimensional gaussian distribution in Equation (46). **b)** The one-dimensional standardized gaussian pdf. As explained in Section 2.2, it is also the marginal distribution of one of the variables in a), and furthermore, turns out to be equal to the conditional distribution of one variable given the other variable. **c)** The probability density function of the two-dimensional gaussian distribution.

A most important property of a pdf is that it is normalized: its integral is equal to one

$$\int p_{\mathbf{z}}(\mathbf{a})d\mathbf{a} = 1 \tag{17}$$

This constraint means that you cannot just take any non-negative function and say that it is a pdf: you have to normalize the function by dividing it by its integral. (Calculating such an integral can actually be quite difficult and sometimes leads to serious computational problems).

For notational simplicity, we often omit the subscript $\mathbf{z}$. We often also write $p(\mathbf{z})$ which means the value of $p_{\mathbf{z}}$ at the point $\mathbf{z}$. This simplified notation is rather ambiguous because now $\mathbf{z}$ is used as an ordinary vector (like $\mathbf{a}$ above) instead of a random vector. However, often it can be used without any confusion.

Example 1 The most classic probability density function for two variables is the gaussian, or normal, distribution. Let us first recall the one-dimensional gaussian distribution, which in the basic case is given by

$$p(z) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}z^2) \tag{18}$$

It is plotted in Fig. 2 b). This is the "standardized" version (mean is zero and variance is one), as explained below. The most basic case of a two-dimensional gaussian distribution is obtained by taking this one-dimensional pdf separately for each variables, and multiplying them together. (The meaning of such multiplication is that the variables are independent, as will be explained below.) Thus, the pdf is given by

$$p(z_1, z_2) = \frac{1}{2\pi} \exp(-\frac{1}{2}(z_1^2 + z_2^2)) \tag{19}$$

A scatter plot of the distribution is shown in Fig. 2 a). The two-dimensional pdf itself is plotted in Fig. 2 c).

Example 2 Let us next consider the following two-dimensional pdf:

$$p(z_1, z_2) = \begin{cases} 1, \text{ if } |z_1| + |z_2| < 1 \\ 0, \text{ otherwise} \end{cases} \tag{20}$$

This means that the data is uniformly distributed inside a square which has been rotated 45 degrees. A scatter plot of data from this distribution is shown in Figure 3 a).

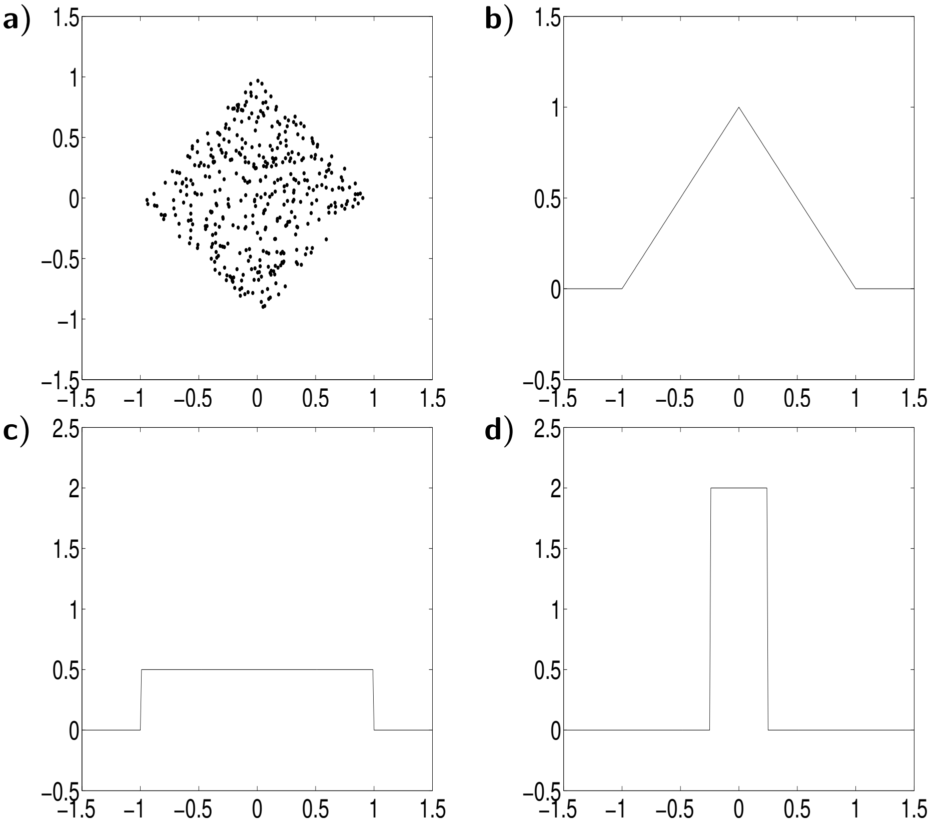

Figure 3: **a)** scatter plot of data obtained from the pdf in Eq. (20). **b)** marginal pdf of one of the variables in a). **c)** conditional pdf of z_2 given $z_1 = 0$. **d)** conditional pdf of z_2 given $z_1 = .75$

2.2 Marginal and joint probabilities

Consider the random vector $\mathbf{z}$ whose pdf is denoted by $p_{\mathbf{z}}$. It is important to make a clear distinction between the *joint* pdf and the *marginal* pdf's. The joint pdf is just what we called pdf above. The marginal pdf's are what you might call the "individual" pdf's of z_i, i.e. the pdf's of those variables, $p_{z_1}(z_1), p_{z_2}(z_2), \ldots$ when we just consider one of the variables and ignore the existence of the other variables.

There is actually a simple connection between marginal and joint pdf's. We can obtain a marginal pdf by integrating the joint pdf over one of the variables. This is sometimes called "integrating out". Consider for simplicity the case where we only have two variables, z_1 and z_2. Then, the marginal pdf of z_1 is obtained by

$$p_{z_1}(z_1) = \int p_{\mathbf{z}}(z_1, z_2) dz_2 \tag{21}$$

This is a continuous-space version of the intuitive idea that for a given value of z_1, we "count" how many observations we have with that value, going through all the possible values of z_2.[1] (In this continuous-valued case, no observed values of z_1 are likely to be exactly equal to the specified value, but we can use the idea of a small interval centred around that value as in the definition of the pdf above.)

Example 3 In the case of the gaussian distribution in Equation (46), we have

$$p(z_1) = \int p(z_1, z_2) dz_2 = \int \frac{1}{2\pi} \exp(-\frac{1}{2}(z_1^2 + z_2^2)) dz_2 = \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2} z_1^2) \int \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2} z_2^2) dz_2 \tag{23}$$

Here, we used the fact that the pdf is factorizable since $\exp(a + b) = \exp(a)\exp(b)$. In the last integral, we recognize the pdf of the one-dimensional gaussian distribution of zero mean and unit variance given in Equation (18). Thus, that integral is one, because the integral of any pdf is equal to one. This means that the marginal distribution $p(z_1)$ is just the the classic one-dimensional standardized gaussian pdf.

Example 4 Going back to our example in Eq. (20), we can calculate the marginal pdf of z_1 to equal

$$p_{z_1}(z_1) = \begin{cases} 1 - |z_1|, \text{ if } |z_1| < 1 \\ 0, \text{ otherwise} \end{cases} \tag{24}$$

which is plotted in Fig. 3 b), and shows the fact that there is more "stuff" near the origin, and no observation can have an absolute value larger than one. Due to symmetry, the marginal pdf of z_2 has exactly the same form.

2.3 Conditional probabilities

Another important concept is the *conditional* pdf of z_2 given z_1. This means the pdf of z_2 when we have observed the value of z_1. Let us denote the observed value of z_1 by a. The conditional pdf is basically obtained by just fixing the value of z_1 to a in the pdf, which gives $p_{\mathbf{z}}(a, z_2)$. However, this is not enough because a pdf must have an integral equal to one. Therefore, we must normalize $p_{\mathbf{z}}(a, z_2)$ by dividing it by its integral. Thus, we obtain the conditional pdf, denoted by $p(z_2 \,|\, z_1 = a)$ as

$$p(z_2 \,|\, z_1 = a) = \frac{p_{\mathbf{z}}(a, z_2)}{\int p_{\mathbf{z}}(a, z_2) dz_2} \tag{25}$$

[1] Note again that the notation in Eq. (21) is sloppy, because now z_1 in the parentheses, both on the left and the right-hand side, stands for any value z_1 might obtain, although the same notation is used for the random quantity itself. A more rigorous notation would be something like:

$$p_{z_1}(v_1) = \int p_{\mathbf{z}}(v_1, v_2) dv_2 \tag{22}$$

where we have used two new variables, v_1 to denote the point where we want to evaluate the marginal density, and v_2 which is the integration variable. However, in practice we often do not want to introduce new variable names in order to keep things simple, so we use the notation in Eq. (21).

Note that the integral in the denominator equals the marginal pdf of z_1 at point a, so we can also write

$$p(z_2 \,|\, z_1 = a) = \frac{p_{\mathbf{z}}(a, z_2)}{p_{z_1}(a)} \tag{26}$$

Again, for notational simplicity, we can omit the subscripts and just write

$$p(z_2 \,|\, z_1 = a) = \frac{p(a, z_2)}{p(a)} \tag{27}$$

or, we can even avoid introducing the new quantity a and write

$$p(z_2 \,|\, z_1) = \frac{p(z_1, z_2)}{p(z_1)} \tag{28}$$

Example 5 For the gaussian density in Equation (46), the computation of the conditional pdf is quite simple, if we use the same factorization as in Equation (23):

$$p(z_2|z_1) = \frac{p(z_1, z_2)}{p(z_1)} = \frac{\frac{1}{\sqrt{2\pi}}\exp(-\frac{1}{2}z_1^2)\frac{1}{\sqrt{2\pi}}\exp(-\frac{1}{2}z_2^2)}{\frac{1}{\sqrt{2\pi}}\exp(-\frac{1}{2}z_1^2)} = \frac{1}{\sqrt{2\pi}}\exp(-\frac{1}{2}z_2^2) \tag{29}$$

which turns out to be the same as the marginal distribution of z_2. (This kind of situation where $p(z_2|z_1) = p(z_2)$ is related to independence as discussed in Section 2.4 below.)

Example 6 In our example pdf in Eq. (20), the conditional pdf changes quite a lot as a function of the value a of z_1. If z_1 is zero (i.e. $a = 0$), the conditional pdf of z_2 is the uniform density in the interval $[-1, 1]$. In contrast, if z_1 is close to 1 (or -1), the values that can be taken by z_2 are quite small. Simply fixing $z_1 = a$ in the pdf, we have

$$p(a, z_2) = \begin{cases} 1, \text{ if } |z_2| < 1 - |a| \\ 0 \text{ otherwise} \end{cases} \tag{30}$$

which can be easily integrated:

$$\int p(a, z_2) dz_2 = 2(1 - |a|) \tag{31}$$

(This is just the length of the segment in which z_2 is allowed to take values.) So, we get

$$p(z_2|z_1) = \begin{cases} \frac{1}{2-2|z_1|}, \text{ if } |z_2| < 1 - |z_1| \\ 0 \text{ otherwise} \end{cases} \tag{32}$$

where we have replaced a by z_1. This pdf is plotted for $z_1 = 0$ and $z_1 = 0.75$ in Fig. 3 a) and b), respectively.

Generalization to many dimensions The concepts of marginal and conditional pdf's extend naturally to the case where we have n random variables instead of just two. The point is that instead of two random variables, z_1 and z_2, we can have two random vectors, say $\mathbf{z}_1$ and $\mathbf{z}_2$, and use exactly the same formulas as for the two random variables. So, starting with a random vector $\mathbf{z}$, we take some of its variables and put them in the vector $\mathbf{z}_1$, and leave the rest in the vector $\mathbf{z}_2$

$$\mathbf{z} = \begin{pmatrix} \mathbf{z}_1 \\ \mathbf{z}_2 \end{pmatrix} \tag{33}$$

Now, the marginal pdf of $\mathbf{z}_1$ is obtained by the same integral formula as above:

$$p_{\mathbf{z}_1}(\mathbf{z}_1) = \int p_{\mathbf{z}}(\mathbf{z}_1, \mathbf{z}_2) d\mathbf{z}_2 \tag{34}$$

and, likewise, the conditional pdf of $\mathbf{z}_2$ given $\mathbf{z}_1$ is given by:

$$p(\mathbf{z}_2 \,|\, \mathbf{z}_1) = \frac{p(\mathbf{z}_1, \mathbf{z}_2)}{p(\mathbf{z}_1)} \tag{35}$$

Both of these are, naturally, multidimensional pdf's.

2.4 Independence

Let us consider two random variables, z_1 and z_2. Basically, the variables z_1 and z_2 are said to be statistically independent if information on the value taken by z_1 does not give any information on the value of z_2, and vice versa.

The idea that z_1 gives no information on z_2 can be intuitively expressed using conditional probabilities: the conditional probability $p(z_2 \,|\, z_1)$ should be just the same as $p(z_2)$:

$$p(z_2 \,|\, z_1) = p(z_2) \tag{36}$$

for any observed value a of z_1. This implies

$$\frac{p(z_1, z_2)}{p(z_1)} = p(z_2) \tag{37}$$

or

$$p(z_1, z_2) = p(z_1)p(z_2) \tag{38}$$

for any values of z_1 and z_2. Equation (38) is usually taken as the definition of independence because it is mathematically so simple. It simply says that the joint pdf must be a product of the marginal pdf's. The joint pdf is then called factorizable.

The definition is easily generalized to n variables $z_1, z_2, \ldots, z_n$, in which case it is

$$p(z_1, z_2, \ldots, z_n) = p(z_1)p(z_2)\ldots p(z_n) \tag{39}$$

Dependence of two variables is often measured by the *covariance*:

$$\text{cov}(z_1, z_2) = E\{z_1 z_2\} - E\{z_1\}E\{z_2\} \tag{40}$$

which is zero for independent variables, but not the other way round: so this is an incomplete measure of dependence.

Example 7 For the gaussian distribution in Equation (46) and Fig. 2, we have

$$p(z_1, z_2) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}z_1^2) \times \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}z_2^2) \tag{41}$$

So, we have factorized the joint pdf as the product of two pdf's, each of which depends on only one of the variables. Thus, z_1 and z_2 are independent. This can also be seen in the form of the conditional pdf in Equation (29), which does not depend on the conditioning variable at all.

Example 8 For our second pdf in Eq. (20), we computed the conditional pdf $p(z_2|z_1)$ in Eq. (32). This is clearly not the same as the marginal pdf in Eq. (24); it depends on z_1. So the variables are not independent. (See the discussion just before Eq. (30) for an intuitive explanation of the dependencies.)

Example 9 Consider the uniform distribution on a square:

$$p(z_1, z_2) = \begin{cases} \frac{1}{12}, & \text{if } |z_1| \leq \sqrt{3} \text{ and } |z_2| \leq \sqrt{3} \\ 0, & \text{otherwise} \end{cases} \tag{42}$$

A scatter plot from this distribution is shown in Fig. 4. Now, z_1 and z_2 are independent because the pdf can be expressed as the product of the marginal distributions, which are

$$p(z_1) = \begin{cases} \frac{1}{2\sqrt{3}}, & \text{if } |z_1| \leq \sqrt{3} \\ 0, & \text{otherwise} \end{cases} \tag{43}$$

and the same for z_2.

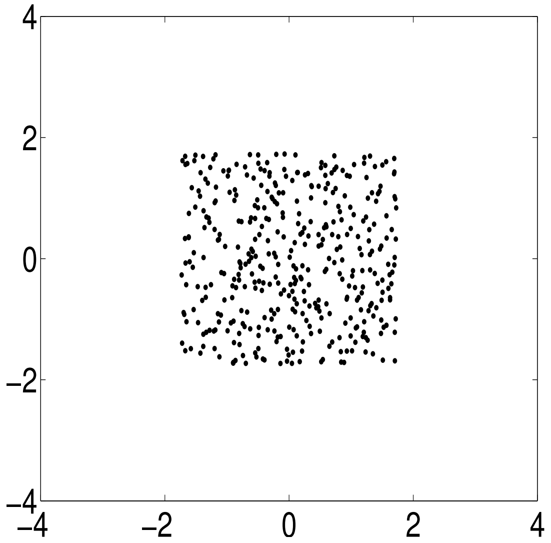

Figure 4: A scatter plot of the two-dimensional uniform distribution in Equation (42)

2.5 Multivariate Gaussian distribution

The general Gaussian distribution in n dimensions is defined as $\mathbf{z} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ which means

$$p(\mathbf{z}) = \frac{1}{(2\pi)^{n/2} |\det \boldsymbol{\Sigma}|^{1/2}} \exp(-\frac{1}{2}(\mathbf{z} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{z} - \boldsymbol{\mu})) \tag{44}$$

where $\boldsymbol{\mu}$ is the mean vector, $\boldsymbol{\Sigma}$ is the matrix of covariances:

$$\boldsymbol{\Sigma}_{ij} = \text{cov}(z_i, z_j) = E\{z_i z_j\} - E\{z_i\}E\{z_j\}. \tag{45}$$

The Gaussian distribution has some exceptional properties. Let us consider, for simplicity, the 2D case with zero means and general covariance. The pdf is given by

$$p(z_1, z_2) = \frac{1}{2\pi} \exp(-\frac{1}{2}(\alpha z_1^2 + \beta z_2^2 + \gamma z_1 z_2)) \tag{46}$$

where γ is negative covariance multiplied by a positive function of variances (it is the off-diagonal entry in the inverse covariance matrix).

Now, we can ask: When are these independent? That is the case if and only if $\gamma = 0$. Detailed maths are omitted here, but γ is zero if and only if the covariance is zero. Thus, zero covariance ("uncorrelatedness") implies independence. This is a remarkable property which holds *only* with the Gaussian distribution!

Likewise, let as consider the conditional distribution: $p(z_1|z_2) = p(z_1, z_2)/p(z_2)$. Without going into the detailed derivation, just consider the following heuristics. Consider z_2 fixed in the pdf, and normalize the remaining by some function of z_2. Now, the conditional pdf depends of z_1 as $\exp(q(z_1))$ where q is a second-order polynomial. Thus, conditional distribution is Gaussian as well! Its mean is not zero but a function of z_2 —in fact it is a linear function; again, this is a special property of the Gaussian distribution!

2.6 Expectation

The expectation of a random vector, or its "mean" value, is, in theory, obtained by the same kind of integral as for a single random variable

$$E\{\mathbf{z}\} = \int p_{\mathbf{z}}(\mathbf{z})\, \mathbf{z}\, d\mathbf{z} \tag{47}$$

The expectation can be computed by taking the expectation of each variable separately, completely ignoring the existence of the other variables

$$E\{\mathbf{z}\} = \begin{pmatrix} E\{z_1\} \\ E\{z_2\} \\ \vdots \\ E\{z_n\} \end{pmatrix} = \begin{pmatrix} \int p_{z_1}(z_1) z_1 \, dz_1 \\ \int p_{z_2}(z_2) z_2 \, dz_2 \\ \vdots \\ \int p_{z_n}(z_n) z_n \, dz_n \end{pmatrix} \tag{48}$$

The expectation of any transformation $\mathbf{g}$, whether one- or multidimensional, can be computed as:

$$E\{\mathbf{g}(\mathbf{z})\} = \int p_{\mathbf{z}}(\mathbf{z}) \mathbf{g}(\mathbf{z}) \, d\mathbf{z} \tag{49}$$

The expectation is a linear operation, which means

$$E\{a\mathbf{z} + b\mathbf{s}\} = aE\{\mathbf{z}\} + bE\{\mathbf{s}\} \tag{50}$$

for any constants a and b. In fact, this generalizes to any multiplication by a matrix $\mathbf{M}$:

$$E\{\mathbf{M}\mathbf{z}\} = \mathbf{M}E\{\mathbf{z}\} \tag{51}$$

The *conditional expectation* is the expectation of a conditional distribution:

$$E\{z_1|z_2\} = \int p(z_1|z_2) z_1 dz_1 \tag{52}$$

and it is a function of the conditioning variable (z_2 here).

The expectation is closely related to the average over a sample. Expectations are theoretical quantities which usually cannot be computed from an observed sample $\mathbf{z}_1, ...\mathbf{z}_N$. The expectation can be approximated, however, by the sample average

$$E\{\mathbf{g}(\mathbf{z})\} \frac{1}{N} \approx \sum_{i=1}^{n} \mathbf{g}(\mathbf{z}_i) \tag{53}$$

where the approximation becomes exact in the limit of an infinite sample, i.e. an infinite number of observations.

2.7 Parameter estimation and likelihood

A *statistical model* describes the pdf of the observed random vector using a number of parameters. The parameters typically have an intuitive interpretation. A model is basically a conditional density of the observed data variable, $p(z \,|\, \alpha)$, where α is the parameter. The parameter could be a multidimensional vector as well. Different values of the parameter imply different distributions for the data, which is why this can also be thought of as a conditional density in Bayesian theory.

For example, consider the one-dimensional gaussian pdf

$$p(z \,|\, \alpha) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}(z - \alpha)^2) \tag{54}$$

Here, the parameter α has an intuitive interpretation as the mean of the distribution. Given α, the observed data variable z then takes values around α, with variance equal to one.

Typically, we have a large number of observations of the random variable z, which might come from measuring some phenomenon n times, and these observations are independent. The set of observations is called a *sample* in statistics.[2]

[2] In signal processing, sampling refers the process of reducing a continuous signal to a discrete signal. For example, an image $I(x, y)$ with continuous-valued coordinates x and y is reduced to a finite-dimensional vector in which the coordinates x and y take only a limited number of values (e.g. as on a rectangular grid). These two meanings of the word “sample” need to be distinguished.

So, we want to use all the observations to better estimate the parameters. For example, in the model in (54), it is obviously not a very good idea to estimate the mean of the distribution based on just a single observation.

Estimation has a very boring mathematical definition, but basically it means that we want to find a reasonable approximation of the value of the parameter based on the observations in the sample. A method (a formula or an algorithm) that estimates α is called an estimator. The value given by the estimator for a particular sample is called an estimate. Both are usually denoted by a hat: $\hat{\alpha}$.

Assume we now have a sample of n observations. Let us denote the observed values by $z(1), z(2), \ldots, z(n)$. Because the observations are independent, the joint probability is simply obtained by multiplying the probabilities of the observations, so we have

$$p(z(1), z(2), \ldots, z(n) \mid \alpha) = p(z(1) \mid \alpha) \times p(z(2) \mid \alpha) \times \ldots \times p(z(n) \mid \alpha) \tag{55}$$

This conditional density is called the *likelihood*. It is often simpler to consider the logarithm, which transforms products into sums. If we take the logarithm, we have the log-likelihood as

$$\log p(z(1), z(2), \ldots, z(n) \mid \alpha) = \log p(z(1) \mid \alpha) + \log p(z(2) \mid \alpha) + \ldots + \log p(z(n) \mid \alpha) \tag{56}$$

The question is then, How can we estimate α?

One basic principle is *maximum likelihood estimation*. It means that we take the value of the parameters which gives the largest value for the likelihood, when likelihood is considered as a function of the parameters with the sample being fixed. The maximum likelihood estimator has thus the intuitive interpretation: it gives *the parameter value that gives the highest probability for the observed data*.

Sometimes the maximum likelihood estimator can be computed by a simple algebraic formula, but in most cases, the maximization has to be done numerically. In some cases it is computationally so difficult that other methods are preferred.

In general, estimation can be accomplished by finding some system of equations which the right parameter values fulfill and then solving it. Of course, this is a very general statement but we will see some examples in this course.

Example 10 In the case of the model in Eq. (54), we have

$$\log p(z \mid \alpha) = -\frac{1}{2}(z - \alpha)^2 + \text{const.} \tag{57}$$

where the constant is not important because it does not depend on α. So, we have for a sample

$$\log p(z(1), z(2), \ldots, z(n) \mid \alpha) = -\frac{1}{2}\sum_{i=1}^{n}(z(i) - \alpha)^2 + \text{const.} \tag{58}$$

It can be shown (this is left as an exercise) that this is maximized by

$$\hat{\alpha} = \frac{1}{n}\sum_{i=1}^{n} z(i) \tag{59}$$

Thus, the maximum likelihood estimator is given by the average of the observed values. This is not a trivial result: in some other models, the maximum likelihood estimator of such a location parameter is given by the median.

Example 11 Here's an example of maximum likelihood estimation with a less obvious result. Consider the exponential distribution

$$p(z|\alpha) = \alpha \exp(-\alpha z) \tag{60}$$

where z is constrained to be positive. The parameter α determines how likely large values are and what the mean is. Some examples of this pdf are shown in Fig. 5. The log-pdf is given by

$$\log p(z|\alpha) = \log \alpha - \alpha z \tag{61}$$

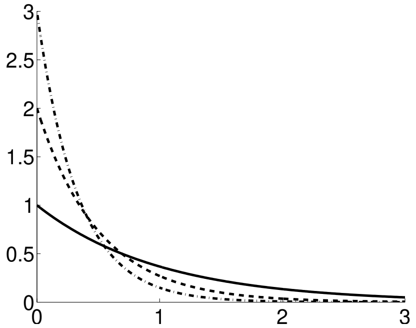

Figure 5: The exponential pdf in Equation (60) plotted for three different values of α, which is equal 1,2, or 3. The value of α is equal to the value of the pdf at zero.

so the log-likelihood for a sample equals

$$\log p(z(1), z(2), \ldots, z(n) \,|\, \alpha) = n \log \alpha - \alpha \sum_{i=1}^{n} z(i) \tag{62}$$

To solve for the α which maximizes the likelihood, we take the derivative of this with respect to α and find the point where it is zero. This gives

$$\frac{n}{\alpha} - \sum_{i=1}^{n} z(i) = 0 \tag{63}$$

from which we obtain

$$\hat{\alpha} = \frac{1}{\frac{1}{n}\sum_{i=1}^{n} z(i)} \tag{64}$$

So, the estimate is the reciprocal of the mean of the z in the sample.

Properties of estimators Consider estimator $\hat{\boldsymbol{\theta}}$ of parameter $\boldsymbol{\theta}$ (possibly vector). The estimator is called *unbiased* if for finite sample of any size:

$$\mathbf{E}\{\hat{\boldsymbol{\theta}}\} = \boldsymbol{\theta} \tag{65}$$

Unbiasedness is a very strong property, and it rarely holds in machine learning, let alone deep learning. A more realistic requirement is that the estimator is *consistent:*

$$\hat{\boldsymbol{\theta}} \xrightarrow[p]{N\to\infty} \boldsymbol{\theta} \tag{66}$$

where N is sample size. (Here, p means convergence in probability, don't worry of you don't know what that means.) In other words, consistency means, roughly speaking, convergence of the estimator to the right value, in the limit of infinite data. (Rather confusingly, in deep learning, consistency may be very loosely called unbiasedness, sometimes "asymptotic unbiasedness".)