Introduction to the Theory of Natural Computing

Appendix: MATH AND STATS

J. Michael Herrmann

16th September 2022

This is a preliminary script[1] for exclusive use in the UoE courses

INFD11007 and INFR11161.

[1] A star next to the title of a chapter or section means that this part could be interesting, but it is not needed for the understanding of other parts or towards a successful completion of the course. In particular, there is usually no point in reading an introduction, at least not before reading the rest of the text.

Appendix A

Appendix: Bits of maths and stats

What is this appendix for?

We try here to give an intuitive understanding of some of the mathematical concepts that are used in Natural Computing. In other words, we present a kind of surrogate mathematics, which is intended to provide a basis for the understanding of the metaheuristic optimisation (MHO) theory, but we have to admit that this will not help much to an understanding of (nor do justice to) the underlying maths itself.

Do we need to know the entire contents of this appendix?

No, but the hope is that you know most of it already. But even if not, just try to follow the main part and come back here only when you got stuck with some term or concept, and you may anyway prefer to check Wikipedia or to ask your fellow students.

A.1 Sets and multisets

A set is a very general concept in mathematics conveying the idea of a union of items. It is not unproblematic to make claims about sets in general. For example, a paper that lists in its bibliography all papers that do not cite themselves [5], is not a clear idea. There are ways to circumvent this difficulty, such as by considering only subsets of a given set. Such a super-set is often called *universe* and can be as simple as the set of non-negative integer numbers from which we can choose the subset of even numbers or the set of all prime number twins[1].

In the context of Genetic Algorithms, a particular problem with sets occurs: A set contains by definition well-distinguished elements, which means that each element can occur only once.

[1]These are natural numbers p and $p+2$ such that both are prime numbers, for example 11 and 13. It is not known whether there are finitely or infinitely may prime number twins.

In a population as a set of individuals, however, two individuals can be identical. In a real population, the two would be physically different. In a numerical realisation of the algorithm the two individuals would correspond to two different memory locations. However, in an abstract algorithm, we cannot take resort to the extrinsic discriminability of the individuals, and have to admit that the population does not fall under the definition of a set.

For example, $\{A, B, B, A\}$ should not be considered as a set, but as a multiset. As a set, it would be identical to the set $\{A, B\}$ as is obvious be the following procedure: Take every element of $\{A, B, B, A\}$ and check whether it is in $\{A, B\}$, and then take every element of $\{A, B\}$ and check whether it is in $\{A, B, B, A\}$. As both tests are affirmative, we would have to concede that $\{A, B, B, A\} = \{A, B\}$, i.e. considered as sets, both are identical.

However, there is life beyond ordinary set theory, and, as usual in such cases in mathematics, the problem is solved by a generalisation. The concept which accounts for this is the *multiset*, simply a set that can have elements that are identical. If there exist some way to discriminate the seemingly identical elements, then we can identify the multiset $\{A, B, B, A\}$ with the set $\{A_1, B_1, B_2, A_2\}$ that has now four discriminable elements. Here, we are not engaging in quantum mechanics, where nondisciminability may be genuine, so for us the distinction between sets and multisets will not be troublesome. Also, we would typically know how many elements we have in the population such that the interpretation of $\{A, B, B, A\}$ as $\{A_1, B_1, B_2, A_2\}$ is possible.

What we actually do in here is considering a list of individuals $\{I(1), \ldots, I(N)\}$, where in the present example $N = 4$ and I is a function (see A.5) that has values in the set of possible individuals (here, this is only A and B), so that $I(1) = A$, $I(2) = B$, $I(3) = B$, $I(4) = A$. In this way $\{A, B, B, A\}$ is just an abbreviated way of writing the list of the $\{I(1), \ldots, I(N)\}$, so we would not have to worry that as a set it is problematic, and we do not have to use the concept of a multiset, which however, occurs in the literature on Genetic Algorithms.

A.2 Permutations

For a set, the number of permutations is the number of possible rearrangement of its elements. For N elements there are $N! = N \cdot (N-1) \cdot \cdots \cdot 4 \cdot 3 \cdot 2 \cdot 1$ possible rearrangements. For a multiset (Sect. A.1), there may be fewer permutations. For example, if in the multiset just one element occurs twice and the other ones are all distinguishable, we only have half as many permutations $N \cdot (N-1) \cdot \cdots \cdot 4 \cdot 3 \cdot 1$. If all elements are identical, there are rearrangements possible, but as there is an arrangement is possible, we would say that there is still one permutation, even so it is a trivial one.

A permutation can be represented in various ways, just like we would specify an arrangement or order of things. We can say, first place the thing 5, then thing 13, then thing 8 etc., or we can say: Set the 7th item aside, and put the 5th in this place, then put the 9th in the currently free place and finally put 7th where the 9th was before, and then start another cycle, i.e. single permutation would consist of several such cycles. Another possibility to describe permutations is by using only swaps, i.e. 2-cycles. Instead of the cycle mentioned

above, we could use two swaps, namely swapping 7 and 5, and then 7 and 9. We can express this by stating[2] $\left(\underset{\longleftrightarrow}{75}\right) \cdot \left(\underset{\longleftrightarrow}{79}\right) = \left(\underset{\longrightarrow}{759}\right)$, where we may need to be careful in what order the swaps are executed[3]. In fact, the operations with permutations can be understood in quite in the same way as arithmetic with numbers!

A.3 The "curse" of dimensionality

It may seem that the situation in higher dimensions is essentially the same as in one dimension, except that each dimension is to be considered separately. This is not generally true, because the effects in different dimensions interact[4], and in particular for very high dimensions these interaction can be quite complicated, so that a reduction to a collection of one-dimensional cases is not suitable for most purposes.

We know the procedure: Start with two points and connect them, so you'll obtain a line. Take two lines and connect them, so that you get a square. Two connected square can produce a cube. If the procedure became clear already, you will know how to produce a tesseract or a penteract in four or five dimensions. While human imagination can be trained to work in these cases, it is obvious that this becomes increasingly difficult with higher and higher dimensions.

Permutations (A.2) are examples of such high-dimensional objects. In a two-dimensional grid, movements to South, West, North, and East, are possible, but for taking a step away from any given permutation, there are $N\,(N-1)$ options, namely choosing one position (Nchoices) and swapping it with another one ($N-1$ remaining choices). But wait, for the two-dimensional grid, we can go in one direction (East or West) or in the second direction (South or North), but for permutation there is only one direction to go away? Not exactly, as you can start with 7 and swap is with 5 gives the same result as choosing first the 5 and swapping it with 7. This means that the dimension (if this word makes sense in a discrete space) for the set of permutations of a set of N elements is $\frac{1}{2}N\,(N-1)$.

If any physical object is to be optimised that is described by K parameters we will have to search in a K-dimensional parameter space for a good combination. For each preliminary parameter set, we can continue the search in $2K$ directions, which does not include the huge number diagonal directions that appear with increasing dimension.

A particularly weird consequence of high dimensions becomes apparent if we consider a high-dimensional sphere rather than, as above, a cube. A sphere can be obtained by selecting all points in the space that are not further away from a given centre point than a certain radius. If the diameter (twice the radius) of the sphere equals 1, then the area of a circle (the two-dimensional case of a sphere) is just $\pi/4$. This is a bit less than the area of square with a unit side, because we have to cut the corners round to get from a square to the circle. For

[2]Be careful when you permute sets of numbers. Here we are talking only about ordinal for places, and not about what objects (number or any other thing) are kept at these places.

[3]Is the example correct?

[4]The multidimensional normal distribution A.12 specifies these interactions precisely as pairwise correlations, but in general, i.e. in the non-normal case, more complicated relationships occur.

higher dimensions the area or rather volume or rather hyper-volume of the high-dimensional cube will always remain 1, but to get to the sphere we'll have to cut more and more corners: 4 for the square, 8 for the cube, 16 for the tesseract etc. The result is that the volume of the hypersphere becomes smaller and smaller[5].

The study of search in high-dimensional space faces the problem that, if we look within a certain radius of the current point, this will cover only a vanishing fraction of the space, but if we search in a cube around the current point we have to check exponentially many corners. Either way, sampling in high dimensions is not easily successful.

The "curse" of high-dimensional spaces is related to this. If we search the maximum of a one-dimensional function that goes two times up and two times down, we will have to check in two place. If this is similar in N dimensions, we may have to check 2^N locations for a potential maximum, which becomes next to impossible already at $N = 50$. So, a random search may do quite well in two dimensions, but will take exponentially more time in higher dimensions, which should leave us with the idea that increasing dimension often leads to exponential complexity, unless (such as in linear algebra or in Gaussian statistics) a polynomial complexity can be enforce by strong assumption.

There is much else to talk about on this topic, but we conclude with a reference to Ref. [1].

A.4 Eigenvalues

A linear map between two d-dimensional spaces can be characterised by a $(d \times d)$ *matrix*, but not every matrix of this size gives rise to a different map. Eigenvalues are a convenient way to characterise such maps with only d numbers rather than $d \times d$. What can such a map do? Essentially it can stretch or squeeze, rotate or sheer the space, but in different directions it can do different things. The fact that the map is linear leads to a bit more clarity of the deformations caused by the map: There are a few special directions where the maps only extends or compresses and does not sheer at all, and there are rotation axes. In addition, the map can also flip-over, i.e. multiply everything by -1. If there is only stretching and squeezing, then the map essentially only deforms a circle in the one space into an ellipse in the other space, or a sphere into an ellipsoid, and so on for higher dimensions. If there is also flipping-over and rotation, then it is a bit more complicated, but the eigenvalues tell us what happens: If all eigenvalues are positive, then only the elliptic deformation can occur, if some eigenvalues are negative then also flipping over occurs, while for rotations the eigenvalues will not be real numbers and will be complex numbers instead, or, in order words, roots of a negative number such as $\pm\sqrt{-1}$. We cannot go into much more detail here, even so it might be suggestive to ask what happens if the eigenvalue is 0. Then everything collapses into 0 at least what concerns a particular eigen-direction. Or whether there are any exceptions. Yes, there are, but then it starts to get complicated. Apart from this, a few more questions remain:

[5]This is not exactly true, as the volume initially increases more by the dimensional increase than what is cut off by the corners, but from dimension 5 upwards the volume indeed decreases inexorably.

How do we know what the eigenvalues are in a particular case? If the space in one-dimensional ($d = 1$), then the map will be simple $y = ax$, and a is the eigenvalue: if $a < 0$, the y values are flipped over compared to the x values, if $a > 1$ the $y$values are stretched, and $0 < a < 1$ the $y$values are squeezed as compared to the x values, and similarly for $-1 < a < 0$ and $a < -1$, respectively, where we have length changes and flipping-over at the same time.[6] For $d \geq 2$, things get a bit more complicated as we will have to consider d eigenvalues, which can have any combination among larger and smaller, positive and negative, real and imaginary parts. Given a $(d \times d)$-matrix, there are many computer programs that can compute the eigenvalues for us (for example the program *wxmaxima*).

How do we know where the map does what? This question points to eigenvectors: Each eigenvalue is accompanied by an eigenvector. A vector represents a direction in a linear space, and an eigenvector shows the direction where the linear map can be described perfectly by stretching or compressing the eigenvector. If the maps is linear, then the points in between any eigenvectors can be linearly interpolated by the eigenvectors and eigenvalues.

Computer programs can compute both eigenvalues and eigenvectors, but often the direction does not matter. If one eigenvector is larger than 1, then divergence can result, but it may not matter towards what direction precisely the algorithm will blow up. Only, if all eigenvalues are small, the algorithm can be guaranteed to converge. What we do know for a linear map is, if it does something at a certain point, then it does relatively the same also for other points in the same direction. This is is simply a way to express that the map is linear.

How do we know the map is linear? We consider here maps such as describing the change of the population from one generation to the next, so we cannot know whether the map is linear. It may well be the case that the map is non-linear. But then it can still be hoped that there is a linear map that approximates the dynamics of the algorithm well. Perhaps not for a long time, but for one generation to the next. Thus, we may have one linear map for this generation transition, and another one for the next transition. There are methods to deal with non-linear systems, but most of them use the properties of a set of accompanying linear systems.

Even if the map is linear, it may not be given as a matrix. The change that we aim to describe by a matrix may be implied by a number of genetic operators, but it usually does not come as a matrix. Finally, it becomes evident why we are talking about eigenvalues. If we are able to express the algorithm by a matrix (or by a set of matrices), then we can find out what is going to happen from the eigenvalues. As there is not much choice to find this out without this formalism, we have to find a way to express the joint action of the operators as matrices, as we did a few times in this book.

[6] You will naturally ask here: What about $a = 1$? As a well-trained mathematician you know that this is the trivial case, where nothing happens, but from a slightly more practical point of view, you should see that at this point the linear approximation becomes questionable: If the linear dynamic is not dominant in a system, then nonlinear effects will have the opportunity to unfold, and all practical system are more or less non-linear. What about $a = -1$?

A.5 Functions

Function are maps, though not exactly the same as a geographical map, which is supposed to represent a region as accurately as possible. A geographical map is similar to what we call an identity map, although strictly speaking it is a scaled version of an identity map. Now, consider the elevation of a territory: For each point of the map we can state the height above sea level, apart perhaps from a few cliffs or caves where this required some more precise definition. Similarly, we can ask what is the temperature at each point within the globe. This temperature can be measured in some places, while in other places we can only estimate, but there is no doubt that such a function that maps the three-dimensional volume to temperature values. If the temperature changes, we can take the time as a fourth variable and ask at what geological era what point inside the planet had what temperature, or what density, what pressure, what type of mineral etc. So we see, that a function can map any set of variable to any other set of a variable. This mapping may be difficult to pin down in many cases, so if we state that we are having here a fitness function, which for each configuration of our solution template gives us a quality value, that clear state how good the solution is[7], then this is quite a strong assumption. Also, it becomes clear that fitness evaluation can be costly, unreliable, delayed, or dependent on circumstances.

A.6 Gradients

At a given position the surface of the earth has an elevation but also an inclination. The mathematical way of describing an inclination is quite simple: Take two nearby points and check their height difference. We then calculate the height difference per distance to get a measure of the inclination. The inclination can be different in different directions. If you are walking on a hillside following a barge there may be no slope, but if you were to go down the brae, then the slope may be steep. Taking the slopes in the coordinate directions gives a vector called gradient.

At the top of any hill, this inclination will be zero, because, if the level at top was rising towards a certain direction, then there would be a higher point in this direction, so by the assumption of the top being the top, we can conclude it is flat. Mathematically, it is possible that the top is a perfect tip that peaks up in a single point that as such cannot be flat. At least directly at the tip the peak is not rising in the sense that nearby points are any higher. Practically, we would expect that the tip is somewhat rounded, or we can assume that either we can identify it by a zero inclination or it is, as in the cases of a pointed peak, not possible to identify an inclination unambiguously.

Natural computing algorithm usually don't use gradients which can be beneficial if the objective function is so complex that any inclination cannot be computed. Some algorithms (such as PSO), however, behave quite similar to a gradient.

[7] at least relative to other solutions

A.7 Optimisation

If we are near a hill, then it is easy to proceed to the top by just following the gradient. A slight problem could arise from doing the gradient following stepwise. Namely, if the step width is too wide, we may jump over the optimum. There are cases when the gradient is so steep that the effect of missing the optimum increases, such that the procedure diverges instead of moving towards the optimum. Mathematically, we can always decrease the step width further, but practically this can lead to dead-slow behaviour, which is likely to occur in cases where different direction have different scales[8]. Assuming, we can find a good step width and represent out problem in a suitable search space, then we may not be anywhere near the hill the top of which is optimum. For example, if we aim at finding the highest elevation on the planet, the problem is more obviously easier if we start near Chomolungma than e.g. in Antarctica. Following the gradient there might lead us to top of Mount Vinson or of any other mountain on the south-polar continent, but does not give us any information that a search elsewhere might be more promising. It seems we need to explore all the surface of the planet in order to be able to solve the task. We could try to start the search at other points or to use many agents to search for us in many different places, so it becomes clear this task of *global optimisation* is a much more demanding problem that the local search that is performed by following the direction of the gradient.

A.8 Approximation

Of we have a good map of the planet, we can search the map for the highest elevation. The coordinates on the map may not be precise, or the heights may have some errors. So, eventually we will need to check in reality, but instead of going everywhere, we could just check a few places which have at least some likelihood to give us the requested answer. How many tests are needed, depends on the accuracy of the map. Instead of hand-drawn maps, any approximation of the problem space will do. And as of today, such approximations are mostly provided by neural networks, which we assume to be known from elsewhere.

A.9 Probability

It is hard to say whether the world is random or whether we just don't know everything. Quantum theory answers this question in favour of true randomness, but this may not be relevant for our daily lives. Probability just states that we don't know some aspects of a thing, and it does not matter, whether we cannot know it, unless we measure it, or whether we could, but just happen not to. The temperature tomorrow can be 20 centigrade, which is a probable value for summer in Scotland, but a improbable value during winter. On the other hand, we don't know whether it will actually be 21 degrees or just 19. It is possible

[8]This is related to the problem of *heteroscedasticity*, which can be solved by what is called the *natural gradient*, a topic beyond the scope of this text.

that an advanced weather prediction system can decide that actually 20 degrees is far more probable than 21 or 19. However, it is also possible that the weather conditions are a bit unstable, so that this cannot be predicted this precisely at the moment. During winter, we may have the same inaccuracy of the prediction that will then fluctuate about other mean values.

Probability is a way to assign numerical values to possibility of event. A probability of near zero means the event is near impossible, and high values tell us that we should expect the event, although we cannot be sure.

A.10 Mean and variance

Average and mean Computing an average is hardly difficult: Just add together all the items, and divide the sum by the total number of items.

$$\text{Mean}\,(A) = \frac{1}{N}\sum_{i=1}^{N} A_i$$

The difference between mean and average is subtle, essentially, a mean is more general than an average, or any average is a mean, but not every mean is an average. It should not be surprising that there are different means: A mean is just any way to define a central or typical point in a data distribution based on a sample. For example, the point half-way between the extrema of the data, can be seen as a mean. In addition to the arithmetic mean, which is the same as the average, there are also geometric and harmonic means.

If the mean value can be predicted from known mathematical analysis, then it is usually referred to as *mathematical expectation*, $\mathbb{E}$, which is a major concept in mathematical statistics. This is so-to-say the presupposed true value that is estimated by the mean.

Variance and standard deviation The variance is the mean square deviation from the mean. If the mean is known, then the squared distance of each value from the mean is calculated and averaged. For example, if one value is 5 and the second one is 7, the mean is 6, and the squared deviations ($(5-6)^2 = (7-6)^2 = 1$) are 1 on both cases, so the variance is 1. It is legitimate to ask why in particular squares are used (rather than absolute differences), but the answer requires a bit more maths. With some simplification, we can refer to the Pythagorean theorem: If we add together two independent random variables A and B to $A + B$, then they will sometimes fluctuate in opposite directions, so that the combined deviations of A and B are larger than the deviations of $A + B$. However, the squared deviations add up perfectly, i.e.

$$\text{Variance}\,(A+B) = \text{Variance}\,(A) + \text{Variance}\,(B)$$

where we advise the reader to check any resources on probability what the keyword independence means in this context.

There is another fine point that is often left unclear, and which we can only mention here without a precise explanation. The formula for the variance is

$$\text{Variance}\,(A) = \frac{1}{N}\sum_{i=1}^{N}\left(A_i - \mathbb{E}\,(A)\right)^2$$

while often the variant

$$\text{Variance}\,(A) = \frac{1}{N-1}\sum_{i=1}^{N}\left(A_i - \text{Mean}\,(A)\right)^2$$

is used. The difference is not very big, as $\frac{1}{N}$ and $\frac{1}{N-1}$ are similar for large N, and also the mathematical expectation has usually a similar value as the mean. The idea is that the two differences in the above formulae are compensating each other. Namely, if the mean is calculated from the same data as the variance, it can happen that the particular sample is bias a bit to one side, which leads to a biased mean. The variance with respect to the biased mean is a bit smaller, and, interestingly, dividing by the smaller number $N-1$ rather than by N this bias is compensated. For the ideal value provided by the mathematical expectation in the first formula, this compensation is not necessary. The fine points between mean, average, and expectation do not have direct counterparts for the variance, but are usually clear from the context or by explicit statement (*sample* variance).

Calculating mean and variance. In order to get mean and variance, we can go through all data and calculate first the mean, and then, in a second pass, also the variance. It is also possible to calculate the variance in a single pass. For this purpose we expand the above formula

$$\begin{aligned}
\text{Variance}\,(A) &= \frac{1}{N}\sum_{i=1}^{N}\left(A_i - \text{Mean}\,(A)\right)^2 \\
&= \frac{1}{N}\sum_{i=1}^{N} A_i^2 - 2A_i\text{Mean}\,(A) + \text{Mean}\,(A)^2 \\
&= \frac{1}{N}\sum_{i=1}^{N} A_i^2 - \left(2\frac{1}{N}\sum_{i=1}^{N} A_i\right)\text{Mean}\,(A) + \text{Mean}\,(A)^2 \\
&= \frac{1}{N}\sum_{i=1}^{N} A_i^2 - 2\text{Mean}\,(A)^2 + \text{Mean}\,(A)^2 \\
&= \frac{1}{N}\sum_{i=1}^{N} A_i^2 - \text{Mean}\,(A)^2
\end{aligned}$$

where we have assumed for simplicity that $\text{Mean}\,(A) = \mathbb{E}\,(A)$. If we had instead used the formula with the $\frac{1}{N-1}$ factor, we would have obtained the similar expression

$$\text{Variance}\,(A) = \left(\frac{1}{N-1}\sum_{i=1}^{N} A_i^2\right) - \frac{N+1}{N-1}\text{Mean}\,(A)^2$$

and, in either case, we can average both $A_i$and A_i^2 in a single run and use the above expression to calculate the variance.

Representing data For the graphical representation of any data, the variance is of little help. Just consider that a quantity in kilometres, where the variance would have the unit km^2. Taking the square root of the variance yields the standard deviation, which is more suitable to illustrate the range over which the considered quantity fluctuates. The quantity would be most of the time within its standard deviation from its mean, but it would also be frequently outside. If we require merely that the quantity stays within three times the standard deviation, then this would be violated less than a percent of the time[9].

Is there always a mean, an expectation, and a variance? Given N numbers, we can always compute the sum and thus the average or the mean value. There is, nevertheless the problem, that it is not always clear how representative the value is for the all data that can be obtained in the same way. A classical example is the "mean energy released during an earthquake". Taking data from the past N earthquakes, we can calculate the mean, but we cannot hope that more observations give us a more and more precise estimate of the mean energy. Instead, we need to be aware, that a single huge earthquake can change the mean value considerably. This is due to the distribution of earthquake powers, which essentially says, that all earthquakes of a certain magnitude have the same power. The huge number of very small earthquakes has the same power as the few earthquakes of magnitude 9, and if a single earthquake of magnitude 10 occurs, it will change the empirical mean considerably as compared to the lucky case where it has not yet occurred. Prominent examples of distribution with the strange property that mean and variance do not exist are power laws. These are typical in situations where any events lead to more of the same, as we have discussed in the context of PSO [3], see also cuckoo search [8], and below in Sect. A.16.

A.11 The normal distribution

If we add many random quantities, we often find that the resulting distribution of the values of the sum is of bell shape. As an observation this was known empirically and Abraham de Moivre had found such a shape when studying the binomial expansion

$$(a+b)^n = \sum_{k=0}^{n} \left(\begin{array}{c} n \\ k \end{array} \right) a^k b^{n-k}$$

which explains that in a sum of binary items the deviations from the mean tend to average out, although not quite, but large deviations become rarer and rarer the further we move away

[9]This holds for normally distributed random quantities. There are other weirder distributions, Chebyshev's inequality assures us that violations of the three standard deviations range occur at most in 11.11% of the cases. If we want to have in general a confidence of 1% for our result, then we need to grant ten standard deviations. There are even weirder distributions, where the standard deviation cannot even be computed as usual, see next paragraph and Sect. A.16.

from the mean. Interestingly, it does not matter much, what contributions are combined as long as there are many, and they do not show avalanche-like dependencies among each other (see A.16). This means that the occurrence of one elementary event makes another on more likely, and this one in turn another one and so forth. If this is the case, then we will arrive at distributions discussed below in Sect. A.16, and if not a normal distribution will appear as the central limit theorem[10] guarantees.

Carl Friedrich Gauss nearly 200 years ago described the beautiful shape of the normal distribution by the beautiful formula[11]

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{A.1}$$

which tells that the probability of a measurement in a small interval Δx around x is approximately $x\Delta x$. This is called the *normal* distribution, often also called *Gaussian*[12], with μ is the mean and σ the standard deviation.

Note, that we cannot simply say: *The probability is* $p(x)$, because the probability that a random draw of a normally distributed random variable has a particular precise value is actually zero. But, if we ask how probable it is get a value near x then we get a reasonable answer, but we need to say what we mean by *near*, and we just said this, namely Δx, or, to be precise, in an interval from $x - \frac{\Delta x}{2}$ to $x + \frac{\Delta x}{2}$. Now, if we assume that $\Delta x = 0$, i.e. if we want to be precise, then we see that the respective probability become zero, as we also just have said.

Further note, that not all probabilities are normal. There are many other distributions which say something about the nature of the underlying process. In particular, if the process is similar to the *normal* case an addition of many terms, but these depend in certain ways on each other[13], then the result may have higher fluctuations than implied by the normal distribution for any width σ.

A.12 The multidimensional normal distribution

Now we are using what we said above in Sect. A.3. Changing σ^2 to Σ, we can write instead of Eq. A.1

$$p(\boldsymbol{x}) = \frac{1}{\sqrt{(2\pi)^d \det \Sigma}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})\,\Sigma^{-1}\,(\boldsymbol{x}-\boldsymbol{\mu})\right). \tag{A.2}$$

where now Σ is a $d \times d$ matrix[14] and the variable and its mean are vectors $\boldsymbol{x}$, $\boldsymbol{\mu} \in \mathbb{R}^d$. The expression $\det \Sigma$ refers to the determinant of this matrix which is a complex operation that

[10] This belongs to higher maths and can be easily found in Wikipedia.

[11] You can ignore the formula, if you believe that maths from 200 years ago (not to speak of more recent maths) is unimportant in real life.

[12] pronounced *Gouzie-ahn*, with the *ou* as in *house.*

[13] like e.g. decisions of stock market traders

[14] Σ is the Greek capital letter Sigma, it is also used to denote sums, but this is not the use in which we use it here.

calculates a single number from the $d \times d$ array the constitutes Σ.

For $d = 1$, we have luckily just one number in the array and this number is then also equal to the determinant, and we are immediately back to Eq. A.1, while for $d > 1$ we need to consider the determinant $|\Sigma|$ and the inverse Σ^{-1} of the *covariance matrix* Σ. Try not to be confused by the missing square in Σ, it is rather σ that is oddly named: The expression σ^2 denotes the variance of the random variable x, while σ is used for its standard deviation, a mentioned above. The standard deviation is more easily understood: It is the typical deviation of a random quantity from its mean value. The variance, on the other hand, is easier to compute with: If there are more than one dimensions and if these dimensions are independent, then the variance can be added up very much like in the Pythagorean theorem. If the dimensions are not independent, things can get complicated, but it is often a useful attempt to capture the dependencies by the covariance matrix, the (k, l)-element of which simply says how the k-th and the l-th dimensions correlate. If the dependencies are fully described in this way, then the vector variable is normally distributed. Note, however, that it's just in textbook example that Σ is somehow given together with the fact that the normal distribution is the correct choice, whereas in practice a normal distribution may be a good start, but is usually just an assumption or approximation.

The determinant describes the scalar factor by how much Σ stretches or squeezes a volume. For example, if a two-dimensional matrix stretches the space in each direction by a factor of 3, then the determinant would be $3 \times 3 = 9$. Likewise, if a two-dimensional variable has a variance of 3 in each direction. Note, that $|\Sigma|$ can be negative in principle, which would mean that the matrix's action flips the volume over, but this is here irrelevant as covariance matrices don't do this. It is of course possible that two of the dimensions are negatively correlated, i.e. the large one dimension get, the smaller will be the other one. But because a one-dimensional variable cannot be negatively correlated to itself (which refers to the positive diagonal elements (k, k) of the matrix) and because not all dimensions can be negatively correlated among each other, the general effect of the matrix is more positive than negative, i.e. more stretching or squeezing than flipping over, causing the determinant to be positive[15].

The inverse of the matrix arises, because we divide by σ^2 in Eq. A.1. Writing Σ^{-1} expresses the same idea. It is necessary, because we cannot simply divide by a matrix, so we define a matrix Σ^{-1} for the given or somehow fitted matrix Σ, and we can multiply by Σ^{-1} to get the same effect as by a division. Computing systems like MATLAB have usually no problems to calculate determinants or inverses of matrices, so there is no need go into more detail here.

Finally, the construction $(\boldsymbol{x} - \boldsymbol{\mu})\,\Sigma^{-1}\,(\boldsymbol{x} - \boldsymbol{\mu})$ in Eq. A.2 makes sure that the dimensions meet in a correct way: From right to left, a vector is multiplied to a matrix giving a vector, which is then multiplied with a vector to give a number. This is very useful as function like the exponential cannot easily take matrix or vector arguments, but the can take numbers to give us eventually an answer on the probability density near a certain vector $\boldsymbol{x}$.

[15] There is an interesting relation between covariance matrices and the volume of a generalised tetrahedron which described by the Cayley–Menger determinant [6].

A.13 Principal component analysis

PCA is very handy when it comes to finding important directions in a data set. It determines the eigenvectors of the covariance matrix Z (Sect. A.12) or, if the data are not Gaussian, then of the best fit of a covariance matrix to the data).

PCA is sensitive to the dimensional units of the data, i.e. it works best for numerical data where all columns and rows have the same (or no) unit.

If the data are not (more or less) a nicely ellipsoid-shaped Gaussian, then in different point different ellipsoids can be fitted. This plays a role in the covariance matrix adaptation (CMA) variants of evolutionary algorithms.

A.14 Bayes' rule

The Bayesian rule can be easily memorised or written down, but it is actually quite hard to understand. For example, if tests of any kind are evaluated, then the rule gives an idea how much we can trust the tests, and how many false alarms or missed results are to be expected. So it is potentially very useful, but it should be remembered that Bayes' formula is not merely a triviality implied by the definition of a conditional probability. Instead, it describes how we should change out beliefs if we obtain new evidence that is not totally unexpected.

The Bayesian rule as such can not be questioned as it is mathematically trivial[16]. Even the typical problem that mathematical concepts do not typically match the complexities of the real world, is irrelevant in Bayes rule[17], because it is about beliefs: If the strength of your belief in H can be measured on a scale between 0 and 1, and it has the strength $P(H)$ and you get data D, then you are free to maintain a certain belief or to overturn if completely, but it is usually optimal if you change your belief exactly according to this rule:

$$P(H|D) = \frac{P(D|H)\,P(H)}{P(D)}. \tag{A.3}$$

A.15 Stochastic processes

Many metaheuristic optimisation algorithms can be seen as a stochastic process that adds noise to each particle or individual in the population and then apply a potential to guide the

[16]Eq. A.3 makes sense only if $P(D) \neq 0$. In this case, we can multiply the equation by $P(D)$ leaving us with $P(H|D)\,P(D) = P(D|H)\,P(H)$, and both sides of this equation are *by definition* the same as the joint probability $P(H, D)$, which completes the proof of Bayes' rule as a mathematical theorem.

[17]This is not fully true, but was deliberately formulated this way, in order to distract from the need to further discuss the rule, and in order to rather focus on its implications. If you are interested in formalities, you should make sure you have an idea how much any new data depend on old data, what processes underlie the data generation in order to rate the probability of the data, whether the data are likely at all, i.e. $P(D) > 0$, how data can support a hypothesis or be implied by a hypothesis in the particular case, and whether the hypothesis you believe in, is consistent, meaningful, and worthwhile.

particle towards good positions. In some algorithms particles are annihilated (if the fall short of a threshold) or created based on information from other particles, which is not considered in the standard models of stochastic processes. It is nevertheless very useful to recapitulate a few general insights from the theory of stochastic processes.

A.16 Power laws

The art of search, just like the ability to learn, to sense, and to act, requires a balance between the exploitation of available information and the acquisition of new information. There have been attempts to use the concept of criticality to make the idea of balance more precise.

Criticality is known from the phenomenon of phase transitions: At the very moment when water freezes into ice, liquid water coexists with chunks of solid ice that have a wide distribution of sizes. This situation can be characterised by a length-scale parameter. In liquid water neighbouring molecules do not have much to do with each other. Although the interaction of water molecules in the liquid phase is already very complex[18], we can assume here that the interaction between two molecules decays exponentially quickly with the respective distance. On the other hand, in solid water, the length of interaction is very large, as you can literally hear if you throw a stone on the surface of a frozen pond[19]: The propagation of sound will cover the whole pond.

How does the exponentially decaying range changes into a extended range? At criticality the coordination of molecules covers a wider range than exponential. Although the interaction does still decay, it follows a power-law.

A power-law distribution is defined by a function

$$f(x) = x^{\alpha} \tag{A.4}$$

for some α. These are a few technical difficulties here [20], but we should first ask: Why would we need such distribution? Aren't normal distribution good enough in most cases?

Not everything is Gaussian, although the central limit theorem seems to imply this. This theorem states that if many independent small random quantities are added up and divided by the standard deviation, then they will be approximately normally distributed. This is demonstrated nicely by the *Galton board*.

If the quantities are not all small, then the sum may be dominated but one or a few of them, so whatever the big contributions are will determine the result. If they are not independent, then they may not average out in a way that produces the known Gauss bell shape. This happens, in particular, if the occurrence of some values in some variable makes similar values

[18] You may not know that pure water freezes only at about -40°C.

[19] The are reasons that you may not want to actually do it.

[20] In order to be useful as a probability distribution, $f(x)$ integrated over all x should give the result 1. For $\alpha \geq 0$ this is hopeless, but for negative α it can work provided that we exclude $x = 0$. Also, if α is small we may run into problems when integrating all the way to positive infinity.

in other variables more likely. For example, the number of children born in a family is not random, but it determined to some extent by the genes in this family, so if a family has many children, also the children may have a higher probability to have many children, at least this was the conclusion of a 1875 paper [7] that aimed at explaining the frequency distribution of family names, which is clearly non-Gaussian.

All these phenomena have in common that certain events make some other events more likely. For example, a small shock of the earth surface propagates along the surface of the earth and can trigger another shock nearby. Then the two shocks can cause a third one, so that a substantial earthquake results. While different earthquakes will usually be statistically independent, each of them consists of a sequence of dependent events. The statistics of these events is in most cases not Gaussian, and it is known for some time that the frequency and the sizes of earthquakes [4] follow a power-law distributions. Meanwhile, many other phenomena have been identified to have a critical dynamics [2], including the sizes of forest fires, the intensity of brain activity, stock market crashes, the palaeontological extinction events, the amount of rainfall, or the sizes of cities in a country. All carry the hallmark that if something starts to happen, then more of the same is likely to happen.

In search and optimisation, this principle is used in order to intensify search if there is some benefit, i.e. to search more where the search had already some preliminary success. The power-law distribution implies not only many small steps that are useful for a local search, but also relatively many larger jumps (provided the exponent in Eq. A.4 is small and negative) to escape from local minima. In some algorithms (e.g. cuckoo search [8]), this is achieved by adding a noise terms with a power-law distribution, while in other algorithms (e.g. PSO [3]) parameter values are known for which the standard algorithm will produce power-law distributed search steps.

This concludes appendix A, but if you miss any topic here or find any errors, please contact the author.

Bibliography

[1] Philip W. Anderson. More is different: Broken symmetry and the nature of the hierarchical structure of science. *Science*, 177(4047):393–396, 1972. 5

[2] Per Bak. *How nature works: The science of self-organized criticality*. Copernicus, New York, 1996. 16

[3] Adam Erskine, Thomas Joyce, and J. Michael Herrmann. Stochastic stability of particle swarm optimisation. *Swarm Intelligence*, 11(3-4):295–315, 2017. 11, 16

[4] Beno Gutenberg and Carl F. Richter. Earthquake magnitude, intensity, energy, and acceleration. *Bulletin of the Seismological Society of America*, 46(2):105–145, 1956. 16

[5] J. Michael Herrmann. *Natural Computing (Lecture Notes)*. unpublished, 2022. 2

[6] Karl Menger. Untersuchungen über allgemeine Metrik. *Mathematische Annalen*, 100(1):75–163, 1928. 13

[7] Henry William Watson and Francis Galton. On the probability of the extinction of families. *The Journal of the Anthropological Institute of Great Britain and Ireland*, 4:138–144, 1875. 16

[8] Xin-She Yang and Suash Deb. Cuckoo search via Lévy flights. In *World congress on nature & biologically inspired computing (NaBIC)*, pages 210–214. IEEE, 2009. 11, 16