Course 67564 2022/23

Bayesian Machine Learning
Recitation 1: Linear Algebra and Probability

Prof. Yair Weiss *TA: Roy Friedman*

1 Linear Algebra

In the course we will use linear algebra rather intensively, so the main purpose of this document is to go over core concepts we'll see throughout the course and go over all of the relevant definitions. That said, this section is a refresher of linear algebra, not a course in linear algebra, so we won't go around proving anything.

2 Vectors

As a reminder, a vector is a list of numbers which define a point in space:

$$x \in \mathbb{R}^d \qquad x = [x_1, x_2, ..., x_d]^T$$

Here the $\mathbb{R}$ indicates that all of the elements in the vector are real and the superscript $\mathbb{R}^d$ tells us that there are d numbers in the vector. The inner product of two vectors is defined as:

$$\langle x, y\rangle \equiv x^T y \stackrel{\Delta}{=} \sum_{i=1}^{d} x_i y_i \tag{2.1}$$

(the sign "$\equiv$" is to show that the first way of writing the inner product and the second one mean the same thing in this context). For any three (real) vectors $x, y, z \in \mathbb{R}^d$ and a scalar $a \in \mathbb{R}$, the inner product has the following properties:

- Linearity: $\langle ax + z, y\rangle \equiv (ax + z)^T y = a\left(x^T y\right) + z^T y$
- Symmetry: $x^T y = y^T x$
- Positive-definite: $x^T x \geq 0$ and $x^T x = 0 \Leftrightarrow x = 0$

Where $x = 0$ means that $\forall i \in [d]\ x_i = 0$. Geometrically, the inner product is the projection of one vector onto another, which will be a very useful intuition to keep in mind and raises another important definition. Two vectors $x, y \in \mathbb{R}^d$ such that

$$x^T y = 0 \tag{2.2}$$

are said to be orthogonal.

The inner product is just one way of multiplying the vectors and in this course we will also use the *outer product* of 2 vectors:

$$x \in \mathbb{R}^d, y \in \mathbb{R}^m \quad xy^T = \begin{bmatrix} x_1y_1 & x_1y_2 & \cdots & x_1y_m \\ x_2y_1 & x_2y_2 & \cdots & x_2y_m \\ \vdots & & & \vdots \\ x_dy_1 & x_dy_2 & \cdots & x_dy_m \end{bmatrix} \in \mathbb{R}^{d\times m} \tag{2.3}$$

As you can see, the outcome of an outer product is a matrix, instead of a scalar.

Another important quality of vectors are *norms*, which are metrics for their distance from the origin. The most commonly used norm is the Euclidean norm, also called the ℓ_2 norm defined as:

$$\|x\|_2 \stackrel{\Delta}{=} \sqrt{x^T x} = \sqrt{\sum_i x_i^2} \tag{2.4}$$

Sometimes we will write the Euclidean distance simply as $\|\cdot\|$ instead of $\|\cdot\|_2$ - this is because it is by far the most common norm we will see and is usually much simpler to use than other norms. Another norm that is used quite often is the ℓ_1 norm, defined as:

$$\|x\|_1 \triangleq \sum_i |x_i| \tag{2.5}$$

We can also generalize these norms to any $p \geq 1$ - the ℓ_p norm is defined as:

$$\|x\|_p \triangleq \left(\sum_i |x_i|^p\right)^{1/p} \tag{2.6}$$

The last norm that we will talk about for now is the ℓ_∞ norm. As the name suggests, this is the ℓ_p norm when $p \rightarrow \infty$:

$$\|x\|_\infty \triangleq \max_i |x_i| \tag{2.7}$$

Apart from just measuring the distance from the origin, norms also allow us to measure the distance from other vectors. The ℓ_2 distance between x and y is defined as:

$$\|x-y\|_2 = \sqrt{(x-y)^T(x-y)} \tag{2.8}$$

3 Matrices

A matrix is a list of vectors (or a table of numbers) which define a linear transformation of vectors:

$$A \in \mathbb{R}^{n\times m} \quad A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nm} \end{bmatrix} \tag{3.1}$$

and the notation $A \in \mathbb{R}^{n\times m}$ means that the matrix A holds $n \times m$ different real items.

The multiplication of a matrix $A \in \mathbb{R}^{n\times m}$ with a vector $x \in \mathbb{R}^m$ (notice the dimensions) is defined as:

$$[Ax]_j = \sum_i a_{ji}x_i \tag{3.2}$$

where $[Ax]_j$ is the j-th index of the resulting vector from the multiplication. If we defined a_i to be the ith row of the matrix, such that:

$$A = \begin{bmatrix} - & a_1 & - \\ - & a_2 & - \\ & \vdots & \\ - & a_n & - \end{bmatrix} \tag{3.3}$$

then we can write the product of a vector with a matrix more cleanly as:

$$Ax = \begin{bmatrix} a_1^T x \\ a_2^T x \\ \vdots \\ a_n^T x \end{bmatrix} \in \mathbb{R}^n \tag{3.4}$$

The multiplication of a matrix $A \in \mathbb{R}^{n\times m}$ with a matrix $B \in \mathbb{R}^{m\times k}$ has the following elements:

$$C_{ij} = [AB]_{ij} = \sum_\ell a_{i\ell}b_{\ell j} \tag{3.5}$$

where $C \in \mathbb{R}^{n\times k}$.

There are a few common families of matrices that we will use, so it will be useful to give them names:

- Matrices with the same number of rows and columns are called *square* matrices
- Matrices where we can change the order of the indices $A_{ij} = A_{ji} \Rightarrow A^T = A$ are called *symmetric* matrices
- Matrices with non-zero values only on the diagonal A_{ii} are called *diagonal matrices* and when the whole diagonal is equal to 1, these matrices are denoted as I. We can think of I as the identity transformation - for any vector x we get $Ix = x$ and for any matrix A we get $IA = A$
- A matrix A that has a corresponding matrix B such that $AB = BA = I$ is called an *invertible* matrix, and B is called the *inverse* of A. If A is invertible, it's inverse is unique - because of this we will write the inverse as A^{-1}. Also, note that from the definition that $AA^{-1} = A^{-1}A = I$ that A must be square in order to be invertible at all
- An orthogonal matrix U is a matrix whose transpose is it's inverse, i.e. $UU^T = U^TU = I$

Example: Inverse of a Diagonal Matrix

Let's say we have a diagonal matrix A such that every value on the diagonal is non-zero:

$$A = \begin{bmatrix} \alpha_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \alpha_n \end{bmatrix}$$

We want to find A^{-1}. In general, it will be hard to find the inverse by simple multiplication, but in this case notice that we can rewrite A as:

$$A = \sum_i e_i e_i^T \alpha_i$$

where e_i are the vectors $e_i = \left[0, ..., 0, \overbrace{1}^{i\text{th index}}, 0, ...0\right]^T$. Equivalently, we see that $I = \sum_i e_i e_i^T$. By simply choosing $A^{-1} = \sum_i e_i e_i^T \frac{1}{\alpha_i}$, we see that:

$$\begin{aligned} AA^{-1} &= \sum_i e_i e_i^T \alpha_i \sum_j e_j e_j^T \frac{1}{\alpha_j} \\ &= \sum_{i,j} e_i e_i^T e_j e_j^T \frac{\alpha_i}{\alpha_j} \end{aligned}$$

however, note that for any $i \neq j$ $e_i^T e_j = 0$, so:

$$AA^{-1} = \sum_i e_i e_i^T \frac{\alpha_i}{\alpha_i} = \sum_i e_i e_i^T = I$$

Finally, we see that:

$$A^{-1} = \begin{bmatrix} \alpha_1^{-1} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \alpha_n^{-1} \end{bmatrix}$$

Example: Blockwise Inversion

Suppose we want to find the inverse of the matrix:

$$M = \begin{bmatrix} A & 0 \\ C & D \end{bmatrix} \tag{3.6}$$

where $A \in \mathbb{R}^{n \times n}$, $C \in \mathbb{R}^{m \times n}$ and $D \in \mathbb{R}^{m \times m}$, so that $M \in \mathbb{R}^{n+m \times n+m}$. Here we have defined M by it's blocks and will find the inverse with respect to these blocks. So we have to find a matrix such that:

$$\begin{bmatrix} A & 0 \\ C & D \end{bmatrix} \begin{bmatrix} T_1 & T_2 \\ T_3 & T_4 \end{bmatrix} = I \tag{3.7}$$

The important thing to notice is that we are still multiplying matrices by their rules, that means that we multiply $[A \quad 0]$ by $[T_1 \quad T_3]$, $[A \quad 0]$ by $[T_2 \quad T_4]$ and so on:

$$\begin{bmatrix} AT_1 + 0T_3 & \cdot \\ \cdot & \cdot \end{bmatrix} = \begin{bmatrix} A & 0 \\ \cdot & \cdot \end{bmatrix} \begin{bmatrix} T_1 & \cdot \\ T_3 & \cdot \end{bmatrix} \tag{3.8}$$

so the result will be:

$$\begin{bmatrix} A & 0 \\ C & D \end{bmatrix} \begin{bmatrix} T_1 & T_2 \\ T_3 & T_4 \end{bmatrix} = \begin{bmatrix} AT_1 + 0T_3 & AT_2 + 0T_4 \\ CT_1 + DT_3 & CT_2 + DT_4 \end{bmatrix} \tag{3.9}$$

The top left corner $(AT_1 + 0T_3)$ must be equal to the identity, which only happens if $T_1 = A^{-1}$. The top right corner must be zero and this only happens if $T_2 = 0$. Let's write this intermediate result:

$$\begin{bmatrix} A & 0 \\ C & D \end{bmatrix} \begin{bmatrix} A^{-1} & 0 \\ T_3 & T_4 \end{bmatrix} = \begin{bmatrix} I & 0 \\ CA^{-1} + DT_3 & DT_4 \end{bmatrix} \tag{3.10}$$

Now, in the same manner as above, we see from the bottom right corner that $T_4 = D^{-1}$, which means we are left with finding T_3 such that:

$$CA^{-1} + DT_3 = 0 \tag{3.11}$$

This will only happen if $T_3 = -D^{-1}CA^{-1}$. So the inverse of M is given by:

$$M = \begin{bmatrix} A & 0 \\ C & D \end{bmatrix} \Rightarrow M^{-1} = \begin{bmatrix} A^{-1} & 0 \\ -D^{-1}CA^{-1} & D^{-1} \end{bmatrix} \tag{3.12}$$

In these derivations, we implicitly assumed that A and D are non-singular. If they are singular, then M will not be invertible.

The above example is a special case of the following rule:

$$M = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \Rightarrow M^{-1} = \begin{bmatrix} A^{-1} + A^{-1}BL^{-1}CA^{-1} & -A^{-1}BL^{-1} \\ -L^{-1}CA^{-1} & L^{-1} \end{bmatrix} \tag{3.13}$$

where $L = D - CA^{-1}B$. In this case, the assumption is that A and L are non-singular. This inversion will become useful in the future, when we talk about the Gaussian distribution.

3.1 Eigenvalues and Eigenvectors

Every matrix has *characteristic directions* (or *characteristic vectors*) - the directions that "matter most" to the matrix. If A is a square matrix, then we call these characteristic directions the *eigenvectors* of A. A vector $u \neq 0$ is an eigenvector of A if:

$$Au = \lambda u \tag{3.14}$$

where λ is a scalar that is called the *eigenvalue* corresponding to the eigenvector u. Notice that if u is an eigenvector of A, then so is $\tilde{u} = \delta u$ for any $\delta \in \mathbb{R}$:

$$A\tilde{u} = A\delta u = \delta Au = \delta\lambda u = \lambda\tilde{u} \tag{3.15}$$

so the eigenvalues u are not unique, while the eigenvalues λ are unique.

The directions of the eigenvectors *are* unique (as long as the rank of A is full, which for the purpose of this course means that A is invertible) and if the matrix is also symmetric, the eigenvectors are always orthogonal to each other. This means that for an $n \times n$ symmetric matrix with full rank, there are exactly n such directions. So as long as A is invertible and symmetric, then the n eigenvectors form a basis and we can rewrite the product of a matrix with a vector as:

$$Ax = A\sum_i \langle u_i, x\rangle \cdot u_i = \sum_i \lambda_i \langle u_i, x\rangle u_i \tag{3.16}$$

(I wrote $\langle u_i, x \rangle$ instead of $u_i^T x$ just to make this form easier to read, but either way of writing is valid).

These eigenvectors also allow us to rewrite the form of A directly using the eigenvectors by noticing the following relationship:

$$\begin{aligned} Ax &= \sum_i \lambda_i u_i^T x u_i \\ &= \sum_i \lambda_i u_i u_i^T x \\ \Leftrightarrow A &= \sum_i \lambda_i u_i u_i^T \end{aligned} \tag{3.17}$$

In vector form, this means any symmetrical matrix A can be decomposed as:

$$A = ULU^T \tag{3.18}$$

where $UU^T = U^T U = I$ and L is a diagonal matrix made up of the eigenvalues of A. Furthermore, the rows in U are the eigenvectors corresponding to the eigenvalues in L. This is called the *eigenvalue decomposition* (EVD) of a symmetrical matrix. Notice that if A is invertible, we can also easily find the decomposition of A^{-1}:

$$I = AA^{-1} \Rightarrow A^{-1} = UL^{-1}U^T \tag{3.19}$$

3.2 Singular Value Decomposition

In a similar way to the eigenvectors and values from above, there is a generalization to all matrices $A \in \mathbb{R}^{m \times n}$. The *singular value decomposition* (SVD) of a matrix A always exists and is defined as:

$$A = U\Sigma V^T \tag{3.20}$$

where $U \in \mathbb{R}^{m \times m}$ is an orthogonal matrix, $\Sigma \in \mathbb{R}^{m \times n}$ is a diagonal matrix (the diagonal Σ_{ii} is non-zero, everything else is a zero) and $V \in \mathbb{R}^{n \times n}$ is also an orthogonal matrix. The terms $\sigma_i = \Sigma_{ii}$ are called the *singular values* of A, are unique to A and are always non-negative. The SVD is directly connected to the EVD in the following manner:

$$AA^T = U\Sigma V^T V \Sigma U^T = U\Sigma^2 U^T \tag{3.21}$$

$$A^T A = V\Sigma U^T U \Sigma V^T = V\Sigma^2 V^T \tag{3.22}$$

and now we can clearly see that U are the eigenvectors of $AA^T$and V are the eigenvectors of $A^T A$.

3.3 Determinant and Trace

The *determinant* of a square matrix A with eigenvalues $\lambda_1, \lambda_2, ..., \lambda_n$ is defined as:

$$\det(A) \equiv |A| \triangleq \prod_i \lambda_i \tag{3.23}$$

(note that the **determinant doesn't have to be positive** even though we write $|A|$!). We can think of the determinant as a measure for how much the space is stretched by the transformation that A implies. If $|A| = 0$, A will be called *singular* and will not be invertible. The term singular originates from the fact that if one of the eigenvalues of the matrix is equal to zero, then there is a direction from which all points are transformed into the origin by the matrix. In turn, there can be no inverse transformation that will move the points from the origin back to their original positions, which is why a singular matrix is not invertible. Two useful properties of determinants are:

- $\left|A^{-1}\right| = \frac{1}{|A|}$
- If A and B are square matrices, then $|AB| = |A|\,|B|$

The *trace* of a square matrix A is defined as:

$$\text{trace}\,[A] \triangleq \sum_i A_{ii} \tag{3.24}$$

if the eigenvalues of A are $\lambda_1, \lambda_2, ..., \lambda_n$ (as before), then:

$$\text{trace}\left[A\right] = \sum_i \lambda_i \tag{3.25}$$

In addition, trace has the following properties:

- $\text{trace}\left[\alpha A + B\right] = \alpha\text{trace}\left[A\right] + \text{trace}\left[B\right]$
- $\text{trace}\left[ABC\right] = \text{trace}\left[CAB\right] = \text{trace}\left[BCA\right]$

3.4 Positive Semi-Definite Matrices

Definition 1. Positive Semi-Definite Matrix
A square, symmetrical, matrix $A \in \mathbb{R}^{n\times n}$ is called positive semi-definite (PSD) if:

$$\forall x \in \mathbb{R}^n \qquad x^T Ax \geq 0$$

and *positive definite* (PD) if:

$$\forall x \neq 0 \in \mathbb{R}^n \qquad x^T Ax > 0$$

There are a few useful characteristics that PD and PSD matrices have, including:

1. A matrix A is PD if and only if it's eigenvalues $\lambda_1, ..., \lambda_n$ are all positive ($\forall i\; \lambda_i > 0$). This also means that a PD matrix is invertible since$|A| = \prod_i \lambda_i > 0$
2. A matrix A is PSD if and only if it's eigenvalues $\lambda_1, ..., \lambda_n$ are all non-negative ($\forall i\; \lambda_i \geq 0$)
3. A matrix A is PD if and only if it can be decomposed as $A = R^T R$ such that R is triangular and invertible. This decomposition is unique, in the sense that there exists only one triangular and invertible matrix R such that $A = R^T R$. This decomposition is called the *Cholesky decomposition*
4. A matrix A is PSD if and only if it can be decomposed as $A = R^T R$

Also, notice that any PD matrix is also PSD, but the opposite isn't true.

Example: Product of a matrix and it's transpose

Suppose we have a matrix $A \in \mathbb{R}^{n\times m}$. We will show that $A^T A$ is PSD for *any* (real) matrix A. We need to show that for any vector x:

$$x^T A^T Ax \geq 0 \tag{3.26}$$

We begin by noticing that $A^T A$ is symmetrical since:

$$\left(A^T A\right)^T = A^T A \tag{3.27}$$

from the definition of transpose (we transpose and change the order). Now, notice we can write the above as an inner product between two vectors:

$$\begin{aligned} x^T A^T Ax &= \left(Ax\right)^T Ax \\ &= \langle Ax, Ax\rangle \\ &= \|Ax\|^2 \end{aligned} \tag{3.28}$$

A norm of a vector is always non-negative, so we see that $x^T A^T Ax \geq 0$, which means that $A^T A$ is a PSD matrix, which is exactly what we wanted to show.
From now on it will be a good idea to remember that for any matrix A, both $A^T A$ and AA^T (you can define $B = A^T$ and then you are looking at the matrix $B^T B$) are PSD matrices.

3.5 Derivatives

Many algorithms include a cost/loss function which we will try to optimize as much as we can. Many times the optimization will be equivalent to finding the minima of the cost function. The simplest (analytical) method to do so when the function is convex/concave, or has a single minima/maxima, is by differentiating the function and equating to 0.

The chain rule for 1D functions is:

$$\frac{\partial f\left(g\left(x\right)\right)}{\partial x}=\frac{\partial f\left(g\left(x\right)\right)}{\partial g\left(x\right)}\frac{\partial g\left(x\right)}{\partial x} \tag{3.29}$$

which you are (hopefully) already comfortable with. However, during this course we will use a lot functions of the form $f:\mathbb{R}^{n}\to\mathbb{R}$, so we will need to first remind ourselves how to treat the derivatives of these functions.

Jacobian

The *Jacobian* of a differentiable function $f:\mathbb{R}^{n}\to\mathbb{R}^{m}$ is a matrix with dimensions $m\times n$ and we define it as:

$$\left[J_{x}\left[f\left(x\right)\right]\right]_{ij}\triangleq\frac{\partial\left[f\left(x\right)\right]_{i}}{\partial x_{j}} \tag{3.30}$$

In this sense, the Jacobian is a sort of generalization of the derivative in higher dimensions. If the function is of the form $g:\mathbb{R}^{n}\to\mathbb{R}$, the transpose of the Jacobian will be a vector that is called the *gradient*:

$$J_{x}\left[g\left(x\right)\right]^{T}\equiv\nabla g\left(x\right)\triangleq\left[\frac{\partial g\left(x\right)}{\partial x_{1}},\ \frac{\partial g\left(x\right)}{\partial x_{2}},\ ...,\ \frac{\partial g\left(x\right)}{\partial x_{n}}\right]^{T}$$

We will often use a different notation for high-order derivatives, that is closer to the 1D definition of derivatives. In our notation, we will use the transpose of the Jacobian:

$$\frac{\partial f\left(x\right)}{\partial x}\triangleq J_{x}\left[f\left(x\right)\right]^{T} \tag{3.31}$$

In other words, if $f:\mathbb{R}^{n}\to\mathbb{R}^{m}$, then $\frac{\partial f(x)}{\partial x}$ is an $n\times m$ matrix. This definition aligns with the definition of the gradient such that:

$$\frac{\partial g\left(x\right)}{\partial x}\equiv\nabla g\left(x\right) \tag{3.32}$$

and will make differentiating a bit easier to understand later on.

Example: Gradient of the squared norm of a vector

Let's look at the function $g\left(x\right)=\left\Vert x\right\Vert ^{2}=\sum_{i}x_{i}^{2}$. The elements of the gradient of this function will be:

$$\left[\nabla g\left(x\right)\right]_{i}=\frac{\partial\sum_{i}x_{i}^{2}}{\partial x_{i}}=\frac{\partial x_{i}^{2}}{\partial x_{i}}=2x_{i}$$

so the whole gradient will be:

$$\frac{\partial g\left(x\right)}{\partial x}=2x$$

(I'm switching notations constantly on purpose - this is to get you accustomed with the fact that both ways to write the gradient mean the same thing, one of them just reminds us that the gradient is a vector and not a number).

Chain Rule

Many times we want to find the gradient of $g\left(f\left(x\right)\right)$ where $f:\mathbb{R}^{n}\to\mathbb{R}^{m}$ and $g:\mathbb{R}^{m}\to\mathbb{R}$ (in this case we are deriving a *scalar* $g\left(f\left(x\right)\right)$ by the *vector* x). In this case, the chain rule is:

$$\frac{\partial g\left(f\left(x\right)\right)}{\partial x}=\underbrace{J_{x}\left[f\left(x\right)\right]^{T}}_{n\times m}\underbrace{\nabla_{f\left(x\right)}g\left(f\left(x\right)\right)}_{m\times1} \tag{3.33}$$

As you can see, the derivative $\frac{\partial g(f(x))}{\partial x}$ will be a vector in $\mathbb{R}^n$, which makes sense since the vector we are differentiating by, x, is in $\mathbb{R}^n$.

This notation makes the chain rule look more intimidating than it is, by using the notation of normal derivatives we get:

$$\frac{\partial g\left(f\left(x\right)\right)}{\partial x}=\frac{\partial f\left(x\right)}{\partial x}\frac{\partial g\left(f\left(x\right)\right)}{\partial f\left(x\right)} \tag{3.34}$$

which looks exactly like the normal chain rule. However, the distinction that is easy to see in the more "formal" notation in 3.33 is that $\nabla_{f(x)}g\left(f\left(x\right)\right)$ is a *vector* and $J_x\left[f\left(x\right)\right]$ is a *matrix*; this is important to remember, as in this case the order of multiplication *is* important, unlike when the function is 1 dimensional.

Example: Gradient of the squared norm of a product

Let's build on the previous example by looking at the function $g\left(x\right)=\|Ax\|^2$, where $A\in\mathbb{R}^{m\times n}$ and $x\in\mathbb{R}^n$. In this case, $g\left(y\right)=\|y\|^2$ and $f\left(x\right)=Ax$. Using the chain rule, we have:

$$\frac{\partial g\left(f\left(x\right)\right)}{\partial x}=J_x\left[f\left(x\right)\right]^T\nabla g\left(y\right)$$

We already know that $\nabla g\left(y\right)=2y$, so all that remains is to find $J_x\left[f\left(x\right)\right]$:

$$\frac{\partial\left(Ax\right)_j}{\partial x_i}=\frac{\partial\sum_k A_{jk}x_k}{\partial x_i}=A_{ji}$$

so we see that $\left[J_x\left[Ax\right]\right]_{ij}=A_{ij}$, i.e.:

$$\frac{\partial Ax}{\partial x}=A^T$$

Using the chain rule, we get:

$$\frac{\partial\|Ax\|^2}{\partial x}=2A^TAx$$

4 Probability

Almost all of the material we will see in this course will be probabilistic in nature; we will almost always want to model the variance of the solution and not just find the most optimal solution. To really understand everything that happens, we must have a good understanding of probability. The following section will be a refresher for some of the key concepts in probability[1] which we will use throughout the course.

4.1 Discrete Probabilities

In general, we define a discrete probability to be a function $P:\Omega\rightarrow[0,1]$ such that $\sum_{\omega\in\Omega}P\left(\omega\right)=1$. A random variable is a variable that can take any value from Ω. From now on, we will denote the probability that a random variable X takes on the value x as:

$$P\left(x\right)\triangleq P\left(X=x\right) \tag{4.1}$$

This will greatly shorten the amount we need to write in the future.

Usually there won't be only one variable of interest, so we need to find a way to introduce a probability over the interaction of several random variables. The probability that two random variables X and Y will take on specific values is called the *joint probability* and is notated by:

$$P\left(x,y\right)\triangleq P\left(X=x,Y=y\right) \tag{4.2}$$

Because of the structure of the probability function, we can always move from the joint probability to one of the *marginal probabilities* by summing out the other variable:

$$P\left(x\right)=\sum_y P\left(x,y\right) \tag{4.3}$$

[1] See Bishop 1.2 and Murphy 2.2, although they also assume that most of the content is known ahead of time

Of course, if Y takes a specific value, then this may effect X in some manner. We notate this *conditional probability* as $P(x|y)$; this new function is also a probability function, i.e.

$$\sum_{x} P(x|y) = 1 \tag{4.4}$$

We can think of this behavior on the side of y as a "re-weighting" of specific values that X may take. The joint probability can be written in terms of this re-weighting as:

$$P(x,y) = P(x|y)\,P(y) \tag{4.5}$$

If the value of X doesn't depend on the value of Y at all (and vice-versa) we will say that the two variables are *independent*. In this case the joint probability takes the form of:

$$P(x,y) = P(x)\,P(y) \tag{4.6}$$

which is a direct result from the rule given by 4.5. We define the conditional probability according to *Bayes' law*:

$$P(y|x) = \frac{P(x,y)}{P(x)} = \frac{P(x|y)\,P(y)}{P(x)} \tag{4.7}$$

Finally, if we have a random variable X and a random variable Y that partitions the space, then we can rewrite the probability for X under the conditionals of Y - this is also called the *law of total probability*:

$$P(x) = \sum_{y} P(x|y)\,P(y) \tag{4.8}$$

4.2 Continuous Probabilities

While it is of course important to understand the rules of probability in the discrete case, most of the course we will be dealing with *continuous random variables*; these random variables can take any value in $\mathbb{R}$ or a section of it. If we simply try to scale the definition we saw before to the continuous case, then for any non-zero probability over any bounded section Γ of $\mathbb{R}$, we have:

$$\sum_{x\in\Gamma} P(x) \geq \sum_{x\in\Gamma} \min_{y\in\Gamma} P(y) \to \infty$$

since there are an infinite number of points in the section Γ. Clearly, we can't use the same reasoning to describe probabilities over continuous variables. Instead, for any section $[a,b] \subseteq \mathbb{R}$, we will define:

$$P(a \leq X \leq b) \triangleq \int_{a}^{b} p(x)\,dx \tag{4.9}$$

where $p(x)$ is called the *probability density function* (PDF) of the variable X. Under this logic, the only restrictions on $p(\cdot)$ are that for any $x \in \mathbb{R}$ $p(x) \geq 0$ and:

$$\int_{-\infty}^{\infty} p(x)\,dx = 1$$

After defining the PDF, all of the rules we have defined earlier apply, only using integrals instead of sums.

4.3 Expectation

One of the most useful statistics involving probabilities we will need in this course is that of finding the weighted average of functions. The average of some function $f(x)$ under a probability function $p(x)$ is called the *expectation* of $f(x)$ and is denoted as $\mathbb{E}[f]$. For a discrete distribution it is given by:

$$\mathbb{E}[f(x)] \triangleq \sum_{x} p(x)\,f(x) \tag{4.10}$$

This has a very clear interpretation, since $p(x)$ sums up to 1: it is the averaging of f, weighted by the relative probabilities of the variable x. For continuous variables, we exchange the sum with an integral to get:

$$\mathbb{E}\left[f(x)\right] \triangleq \int\limits_{-\infty}^{\infty} p(x) f(x)\, dx$$

By definition, the expectation is a *linear operator*, i.e.:

$$\mathbb{E}\left[ax + y\right] = a\mathbb{E}\left[x\right] + \mathbb{E}\left[y\right] \tag{4.11}$$

In either case, if we are given a finite number of points, N, sampled independently and identically from the distribution, then the expectation can be approximated as:

$$\mathbb{E}\left[f(x)\right] \approx \frac{1}{N}\sum_i f(x_i)$$

At the limit $N \to \infty$, this approximation is exact.

The mean of the distribution $p(x)$ is simply the expected value of x itself, i.e.:

$$\mathbb{E}\left[x\right] = \int\limits_{-\infty}^{\infty} xp(x)\, dx \tag{4.12}$$

We can of course give the same treatment to joint probabilities:

$$\mathbb{E}_{x,y}\left[f(x,y)\right] = \int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{\infty} f(x,y)\, p(x,y)\, dxdy$$

Moreover, we can look at the averages according to only one of the marginals of the distribution, i.e.:

$$\mathbb{E}_x\left[f(x,y)\right] = \int\limits_{-\infty}^{\infty} f(x,y)\, p(x)\, dx \tag{4.13}$$

here we have added the subscript x to denote that we are averaging over x and not y. In this case, the expectation will be a function of y, as it is still a free variable. We can also consider *conditional expectations*, that is the weighted average of function over the conditional expectations:

$$\mathbb{E}\left[f(x)\,|y\right] = \int\limits_{-\infty}^{\infty} f(x)\, p(x|y)\, dx \tag{4.14}$$

4.4 Variance and Covariance

Many times we would also like to measure how much variability there is to the values of the function $f(\cdot)$. The *variance* of $f(\cdot)$, defined as:

$$\mathrm{var}\left[f(x)\right] \triangleq \mathbb{E}\left[\left(f(x) - \mathbb{E}\left[f(x)\right]\right)^2\right] = \mathbb{E}\left[f(x)^2\right] - \mathbb{E}\left[f(x)\right]^2 \tag{4.15}$$

measures exactly that. Of course, we can also consider the variance of the variable itself:

$$\mathrm{var}\left[x\right] = \mathbb{E}\left[x^2\right] - \mathbb{E}\left[x\right]^2$$

Another measure that we will see during the course is the *standard deviation*. The standard deviation of a random variable is defined as:

$$\sigma_x \triangleq \sqrt{\mathrm{var}\left[x\right]}$$

When we have many dependent variables, we may also want to see how much each random variable is effected by the other variables. The *covariance* measures this and is defined by:

$$\mathrm{cov}\left[x,y\right] \triangleq \mathbb{E}\left[\left(x - \mathbb{E}\left[x\right]\right)\left(y - \mathbb{E}\left[y\right]\right)\right] = \mathbb{E}\left[xy\right] - \mathbb{E}\left[x\right]\mathbb{E}\left[y\right] \tag{4.16}$$

Directly from the definition, we can see that the covariance of a random variable with itself is simply its variance:

$$\mathrm{cov}\left[x,x\right] = \mathbb{E}\left[x^2\right] - \mathbb{E}\left[x\right]^2 = \mathrm{var}\left[x\right]$$

4.5 Random Vectors

In many applications we will have many random variables that somehow depend on each other - $x_1, ..., x_n$. Usually, it will by much easier to group them together into a vector $\boldsymbol{x} = (x_1, ..., x_n)^T$ than to consider them individually. In this case, we will also write out the PDF as a function of a vector, so that:

$$p : \mathbb{R}^n \rightarrow \mathbb{R}_+$$

and we will simply write $p(\boldsymbol{x})$ instead of $p(x_1, x_2, ..., x_n)$. Of course, all of the attributes we have introduced above are also available to random vectors. The only real difference from before is the notation. The expectation of a random vector is defined to also be a vector, where each coordinate is the expectation of the random variable from the same coordinate:

$$\mathbb{E}[\boldsymbol{x}] \in \mathbb{R}^n \mathbb{E}[\boldsymbol{x}]_i = \mathbb{E}[x_i] \tag{4.17}$$

This definition of random vectors allows us to define the *covariance matrix*. For two random vectors $\boldsymbol{x} = (x_1, ..., x_n)^T$ and $\boldsymbol{y} = (y_1, ..., y_m)^T$, we define the covariance matrix as:

$$\text{cov}[\boldsymbol{x}, \boldsymbol{y}] = \mathbb{E}\left[(\boldsymbol{x} - \mathbb{E}[\boldsymbol{x}])(\boldsymbol{y} - \mathbb{E}[\boldsymbol{y}])^T\right] = \mathbb{E}\left[\boldsymbol{x}\boldsymbol{y}^T\right] - \mathbb{E}[\boldsymbol{x}]\,\mathbb{E}\left[\boldsymbol{y}^T\right] \tag{4.18}$$

the result is a matrix of dimension $n \times m$ (yes, $\boldsymbol{x}$ and $\boldsymbol{y}$ don't have to have the same dimension), with the elements:

$$\text{cov}[\boldsymbol{x}, \boldsymbol{y}]_{ij} = \text{cov}[x_i, y_j]$$

as expected. For notational convenience, we may use the definition:

$$\text{cov}[\boldsymbol{x}] \triangleq \text{cov}[\boldsymbol{x}, \boldsymbol{x}]$$

which is the matrix of the covariances between different variables in the random vector $\boldsymbol{x}$.

The covariance matrix of a single variable $\boldsymbol{x}$ is a PSD matrix, as we can see by writing the covariance explicitly:

$$\text{cov}[\boldsymbol{x}] = \mathbb{E}\left[\boldsymbol{x}\boldsymbol{x}^T\right] - \mathbb{E}[\boldsymbol{x}]\,\mathbb{E}\left[\boldsymbol{x}^T\right] = \mathbb{E}\left[(\boldsymbol{x} - \mathbb{E}[\boldsymbol{x}])(\boldsymbol{x} - \mathbb{E}[\boldsymbol{x}])^T\right] \tag{4.19}$$

That is, the covariance is the result of a matrix times it's transpose, which is a family of matrices we have shown to be PSD.

4.6 Change of Variable

As defined so far, every random variable X comes with it's own PDF, $p_x(\cdot)$. When we have two different random variables X and Y, we will have two *different* pdfs $p_x(\cdot)$ and $p_y(\cdot)$. Sometimes it will be helpful to move from one PDF to the other, if the density of one variable depends on the other. However, it isn't clear how we can do this without violating the elementary conditions that PDFs must satisfy. We can bypass this by working with the *cumulative distribution function* (CDF) of the random variable, instead of the PDF, defined as:

$$P_y(y) \triangleq P(Y \leq y) = \int_{-\infty}^{y} p_y(\tilde{y})\, d\tilde{y} \tag{4.20}$$

As is maybe obvious, we can get back to the PDF by deriving the CDF. If we have a function $f : X \rightarrow Y$ that maps between the random variables, then[2]:

$$P(Y \leq y) = P(f(X) \leq y) = P(X \in \{x \,|\, f(x) \leq y\}) \tag{4.21}$$

If $f(\cdot)$ is invertible, we can further simplify this:

$$P(f(x) \leq y) = P\left(X \leq f^{-1}(y)\right) = P_x\left(f^{-1}(y)\right) \tag{4.22}$$

Differentiating, we move back to the PDF:

$$p_y(y) \triangleq \frac{\partial}{\partial y} P_y(y) = \frac{\partial}{\partial y} P_x\left(f^{-1}(y)\right) = \frac{\partial f^{-1}(y)}{\partial y} \frac{\partial}{\partial f^{-1}(y)} P_x\left(f^{-1}(y)\right) = \frac{\partial f^{-1}(y)}{\partial y} p_x\left(f^{-1}(y)\right) \tag{4.23}$$

[2] See the Wikipedia page for the change of variables for a slightly more organized explanation of what's happening here

In general, since the PDF is non-negative, we will take the absolute value of the derivative to be sure of the result:

$$p_y(y) = \left| \frac{\partial f^{-1}(y)}{\partial y} \right| p_x\left(f^{-1}(y)\right) \tag{4.24}$$

The derivative term is a re-normalization of the PDF from the Y space to the X space, where $\partial x = \partial f^{-1}(y)$ and ∂y are measures of the volume of the X and Y spaces respectively - the term in the derivative can then generally be thought of as re-normalizing the PDF so that it is measured in units of volume of the X space instead of units of volume in the Y space.

The same story unfolds in the multivariate case, with the only catch that now we have to use the Jacobian of the transformation and not a simple derivative. Using the same analogy as before (I know this isn't a proof, but that will be harder and not particularly useful), while the Jacobian measures the *change* of the function in each of it's directions, the determinant of the Jacobian measures the *change in volume* before and after the function (the same as above). With this knowledge, given a function $f : \boldsymbol{x} \rightarrow \boldsymbol{y}$, the change of variable will be:

$$p_y(\boldsymbol{y}) = p_x\left(f^{-1}(\boldsymbol{y})\right) \left| J_y\left[f^{-1}(\boldsymbol{y})\right] \right|$$

Example: Change of variable

Let's build an example for the change of variable rule. Let x be a uniform random variable in the range $[0,\ \alpha]$; in other words:

$$p_x(x) = \begin{cases} \frac{1}{\alpha} & x \in [0,\ \alpha] \\ 0 & \text{otherwise} \end{cases} \tag{4.25}$$

Also, we will define $z = x^2$ and want to find $p_z(\cdot)$ in terms of $p_x(\cdot)$.
Notice that in the range $0 \leq x \leq \alpha$ - the range where $p_x(\cdot)$ is non-zero (also called the *support* of $p_x(\cdot)$)- the function:

$$z = f(x) = x^2 \tag{4.26}$$

is invertible, which means that we can use the change of variable rule. The inverse function in the same range is given by:

$$x = f^{-1}(z) = \sqrt{z} \tag{4.27}$$

The derivative of the inverse function is the following:

$$\frac{\partial f^{-1}(z)}{\partial z} = \frac{1}{2} \cdot \frac{1}{\sqrt{z}} \tag{4.28}$$

so the PDF of z can be rewritten in terms of the PDF of x as follows:

$$p_z(z) = \frac{1}{2} \cdot \frac{1}{\sqrt{z}} p_x\left(\sqrt{z}\right) = \begin{cases} \frac{1}{2\alpha} \cdot \frac{1}{\sqrt{z}} & 0 \leq \sqrt{z} \leq \alpha \\ 0 & \text{otherwise} \end{cases} \tag{4.29}$$