Linear Algebra Review

**Vectors** To begin, let us describe an element of the state space as a point with numerical coordinates, that is

\[
x = \begin{pmatrix}
x_1 \\
x_2 \\
\vdots \\
x_n
\end{pmatrix}
\]

Vectors of up to three dimensions are easy to diagram. For example,

\[
x = \begin{pmatrix}
3 \\
2 \\
5
\end{pmatrix}
\]

can be drawn as follows.

![Figure 1: An example of a vector.](image)
**Addition** To add two vectors together simply add their components. To multiply a vector by a scalar (number), multiply each of the components by the scalar. For example, if \( z = x + y \), then if

\[
x = \begin{pmatrix} 3 \\ 5 \\ 2 \end{pmatrix} \quad \text{and} \quad y = \begin{pmatrix} -4 \\ 1 \\ 3 \end{pmatrix}
\]

then \( z = \begin{pmatrix} -1 \\ 6 \\ 5 \end{pmatrix} \).

And if \( z = \alpha x \) for a scalar \( \alpha \), then

\[
z = \alpha \begin{pmatrix} 3 \\ 5 \\ 2 \end{pmatrix} = \begin{pmatrix} 3\alpha \\ 5\alpha \\ 2\alpha \end{pmatrix}
\]

**Dot Product** The *dot product* of two vectors, denoted \( x \cdot y \), is defined as the sum of the product of their pairwise components; that is,

\[
x \cdot y = \sum_{i=1}^{n} x_i y_i
\]

For our example, \( x \cdot y = (3)(-4) + (5)(1) + (2)(3) = -1 \).

Two vectors are said to be *orthogonal* if their dot product is zero; that is, \( x \cdot y = 0 \).

The *length* of a vector, denoted \( ||x|| \), is simply \( \sqrt{x \cdot x} \). The angle \( \theta \) between two vectors \( x \) and \( y \) is given by

\[
\cos \theta = \frac{x \cdot y}{||x|| \ ||y||}
\]

From this equation it is seen that if vectors are orthogonal, then \( \cos \theta = 0 \) or \( \theta = 90^\circ \).

The vector \( (x \cdot y)\frac{y}{||y||} \) is termed the *projection* of \( x \) onto \( y \).
The transpose of a matrix, denoted $A^T$, is simply described using the element notation as $\{a_{ji}\}$. In other words, the elements are “flipped” about the diagonal. A square $n \times n$ matrix is symmetric if $A^T = A$.

**Linear Transformations** For any function $f(x)$, a linear transformation is such that

$$f(ax + by) = af(x) + bf(y)$$

An important linear transformation is *matrix multiplication*. Matrix multiplication $A = BC$ is defined by

$$a_{ij} = \sum_{k=1}^{N} b_{ik}c_{kj}, i = 1, \ldots, P, j = 1, \ldots, Q$$

From this formula it is seen that the number of columns of $B$ has to be the same as the number of rows of $C$ for multiplication to be defined.

**Determinant** To define the *determinant* of a matrix first requires defining the number of inversions in a number sequence. Consider the sequence $\{1, 3, 4, 2\}$. The number of inversions in this sequence is 2 because 3 and 4 come after 2. Similarly the number of inversions in $\{4, 2, 1, 3\}$ is 3. Denote the number of inversions of a sequence as $n$. The determinant of a matrix $A$, denoted $|A|$, is the sum of all $n!$ possible different products that compose elements from columns of the matrix with a term that depends on the number of inversions in the row indices; that is,

$$|A| = \sum (-1)^{n(i_1,i_2,\ldots,i_n)} a_{i_1,1}a_{i_2,2}\cdots a_{i_n,n}$$

Like the inverse of a matrix, the determinant is expensive to calculate for large matrices, and a standard text
should be referred to for an algorithm. For practice calculations, however, it is useful to remember that the determinant of the $2 \times 2$ matrix

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

is given by

$$|A| = ad - bc$$

**Inverse** For square matrices where $N = M$, an important matrix is the *inverse* matrix $A^{-1}$, which is defined by

$$AA^{-1} = I$$

where $I$ is the *identity matrix*

$$I = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \vdots \\ \vdots & \ddots & 0 \\ 0 & \cdots & 1 \end{bmatrix}$$

In general, like the determinant, inverses take some work to calculate, and you should find a numerical routine. For practice, however, it is useful to remember that the inverse of the $2 \times 2$ matrix

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

is given by

$$A^{-1} = \frac{1}{|A|} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$
**Trace**  The *trace* of a matrix $A$ is the sum of its diagonal elements; that is,

$$Tr(A) = \sum_{i=1}^{N} a_{ii}$$

**Positive Definite**  A matrix $A$ is *positive definite* if for every $x$,

$$x^T A x > 0$$

and *positive semidefinite* if

$$x^T A x \geq 0$$

**Orthonormal Transformation**  A transformation matrix $A$ is orthonormal when

$$A^{-1} = A^T$$

As a consequence

$$AA^T = I$$
Conditional probability

The centerpiece of probabilistic reasoning is the calculation of the probability of a conditional event, \( P(A|B) \), read as “the probability of \( A \) given that \( B \) happened.” The figure motivates this rule using Venn diagrams. Given that \( B \) occurred, the right thing to do is to renormalize the universe so that the probability of \( A \) occurring is expressed as a fraction of \( B \)’s area.

Figure 2: (a) Using the Venn diagram, the probability of an event \( B \) can be thought of as the ratio of the area allocated to \( B \) scaled by the area allocated to the universal event \( U \). (b) Bayes’ rule works similarly. Once it is known that \( B \) occurred, the appropriate area for the event \( A \) is given by \( A \cap B \) and the appropriate area of the new universe is now that of \( B \).
Bayes’ Rule

\[ P(A|B) = \frac{P(A, B)}{P(B)} \]

An extremely important setting for Bayes’ rule is that of maximum likelihood. In this setting the goal is to calculate the probability of a hypothesis or model \( M \), given data \( D \), or

\[ \max_M P(M|D) \]

This rule is very natural and intuitive, but the probability \( P(M|D) \) is typically difficult to measure. It is much easier to obtain a priori distributions \( P(D|M) \). These can be related to a posteriori distributions \( P(M|D) \) by Bayes’ rule: \( P(M|D) = \frac{P(D|M)P(M)}{P(D)} \). It is assumed that all the terms on the right-hand side (RHS) can be measured ahead of time or from on-line exemplars. The term \( P(D|M)P(M) \) is the probability or likelihood that the data was produced by model \( M \). Assigning it so is using the maximum likelihood rule.
**Example: Vision Test**  A robot’s job is to pick up boxes and take them to a designated location. However, the robot’s picking-up device is faulty and only works with probability 0.8. If the robot turns up without a box, it is sent back for another try. You decide to install a camera to test whether a box has been picked up. Let $'B'$ represent the datum that the camera test reports the box picked up and $B$ be the hypothesis that it is actually picked up. The specifications for the camera are in terms of its conditional *a posteriori* probability distribution:

<table>
<thead>
<tr>
<th></th>
<th>$'B'$</th>
<th>$\neg'B'$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$B$</td>
<td>$P('B'</td>
<td>B) = 0.9$</td>
</tr>
<tr>
<td>$\neg B$</td>
<td>$P('B'</td>
<td>\neg B) = 0.2$</td>
</tr>
</tbody>
</table>

Presumably these data are obtained through many tests.
If the camera test says the box was picked up, what is the probability that the box was actually picked up? To solve this problem you would like to know \( P(B'|B') \). This can be obtained from the Bayes inversion formula:

\[
P(B'|B') = \frac{P(B'|B)P(B)}{P(B')}
\]

where \( P(B') \) can be computed from

\[
P(B') = P(B', B) + P(B', \neg B)
\]

So that

\[
P(B'|B') = \frac{P(B'|B)P(B)}{P(B', B) + P(B', \neg B)}
\]

\[
= \frac{P(B'|B)P(B)}{P(B'|B)P(B) + P(B'|\neg B)P(\neg B)}
\]

\[
= \frac{1}{1 + \frac{P(B'|\neg B)P(\neg B)}{P(B'|B)P(B)}}
\]

Combining and substituting,

\[
P(B'|B') = \frac{1}{1 + \frac{(0.2)(0.2)}{(0.9)(0.8)}} = \frac{1}{1.056} \approx 0.95
\]

In other words, the camera test has improved the situation: With the test, the robot is sent back only 5% of the time, versus 20% without it.
Probability Distributions

In the foregoing example, the event space is discrete and small scale. But dealing with large systems of neurons requires analyzing large numbers of discrete events and also continuous events. To accomplish this purpose requires the notion of a probability distribution.

Discrete Distributions

Discrete events can be described with a probability mass function $P(x)$ that keeps track of the probability associated with each event. It is also useful to know certain features of this function, such as its average value, or mean, and the spread, or variance, about that value. The standard deviation $\sigma$, which is just the square root of the variance, is also useful. The mean may be estimated from $n$ samples of $x_i$ that occur with $P(x)$ as

$$m = \frac{1}{n} \sum_{i=1}^{n} x_i$$

and the variance as

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - m)^2$$

Two extremely important probability mass functions have an analytic form. These are the binomial distribution and its approximation for special cases, the Poisson distribution.
**Binomial Distribution**  The starting point for the analysis is a set of $n$ independent experiments, the best example of which is tossing a coin. Suppose that you toss a coin three times. The possible outcomes are

$$HHH, HHT, HTH, THH, HTT, THT, TTH, TTT$$

Suppose the problem is to find the probability of exactly 0, 1, 2, or 3 heads. In this example it can be done by counting. Thus

$$P(0) = \frac{1}{8}, \ P(1) = \frac{3}{8}, \ P(2) = \frac{3}{8}, \ P(3) = \frac{1}{8}$$

For $P(1)$ the probability of each individual outcome is $\frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{8}$, and there are three ways it can happen. In general, if the probability of success is $p$ and that of failure $q = 1 - p$, then the general formula is the binomial distribution given by

$$P(k) = \binom{n}{k} p^k q^{n-k} \quad (1)$$

where $\binom{n}{k}$ is the term that keeps track of all the different ways of getting $k$ “heads” in $n$ tosses of the coin, and is given by

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

The binomial distribution has a mean given by

$$m = np$$

and a variance given by

$$\sigma^2 = npq$$
Poisson Distribution  The binomial distribution is extremely useful but difficult to calculate, so there is a premium on approximations to it that are easier to compute. The most useful of these is the Poisson distribution, which works when \( p \) is small. In this case,

\[
\binom{n}{k} p^k q^{n-k} \approx \frac{(np)^k}{k!} e^{-np}
\]

The Poisson distribution has a mean given by

\[ m = np \]

and a variance given by

\[ \sigma^2 = np \]

The Poisson distribution is usually used when dealing with events that occur with respect to a continuous quantity, such as length or time. Thus the probability is usually measured as the number of occurrences in a given interval, which in turn is calculated by a rate of occurrence \( \times \) a reference interval. Letting \( I \) be one over the reference interval and \( N \) the size of the interval of interest, the formula becomes

\[
P(k) = \frac{(NI)^k}{k!} e^{-NI}
\]
**Example** if a neuron produces an average of one spike per 100 ms, what is the probability of seeing three spikes in $\frac{1}{4}$ second? Here $I = \frac{1}{100}$ and $N = 250$, so that

$$P(3) = \frac{(2.5)^3}{3!}e^{-2.5} \approx 0.15$$

For the Poisson approximation to work, the rate of occurrence must obey certain properties:

1. For an interval of given length, the frequency with which an event will occur is proportional to the length of the interval.
2. The occurrence of more than one of the events in the interval is rare.
3. The occurrence of an event in one interval is independent of the occurrence of an event in another interval.
Standard Normal Distribution  One of the most useful parametric distributions is the normal distribution or Gaussian of a continuous random variable $X$. It is useful because most observed random variables tend to be the sum of several random components, and the sum of random components tends to be normally distributed.

The normal distribution is given by

$$p(x) = \frac{e^{-\frac{1}{2}d^2(x,m,\sigma)}}{\sqrt{2\pi\sigma}}$$

where $d^2$ is given by

$$d^2(x, m, \sigma) = \frac{(x - m)^2}{\sigma}$$

Central Limit Theorem

The distribution of the sum of $N$ i.i.d. random variables becomes increasingly Gaussian as $N$ grows.

Example: $N$ uniform $[0,1]$ random variables.
Cumulative Distribution

Especially useful is the cumulative distribution, which is given by

\[ \Phi(x) = P(X < x) = \int_{-\infty}^{x} p(u) du \]

This distribution can be tabulated for \( m = 0 \) and \( \sigma = 1 \), as is done in Table 1. When the mean and variance are not zero and one, respectively, the table can still be used, but instead of \( x \) you use the argument \( (x - m)/\sigma \).

Table 1: The integral under the Gaussian probability density function.

<table>
<thead>
<tr>
<th>( x )</th>
<th>( \Phi(x) )</th>
<th>( x )</th>
<th>( \Phi(x) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>0.500</td>
<td>1.4</td>
<td>0.919</td>
</tr>
<tr>
<td>0.2</td>
<td>0.579</td>
<td>1.6</td>
<td>0.945</td>
</tr>
<tr>
<td>0.4</td>
<td>0.655</td>
<td>1.8</td>
<td>0.964</td>
</tr>
<tr>
<td>0.6</td>
<td>0.726</td>
<td>2.0</td>
<td>0.977</td>
</tr>
<tr>
<td>0.8</td>
<td>0.788</td>
<td>2.2</td>
<td>0.986</td>
</tr>
<tr>
<td>1.0</td>
<td>0.841</td>
<td>2.4</td>
<td>0.992</td>
</tr>
<tr>
<td>1.2</td>
<td>0.885</td>
<td>2.6</td>
<td>0.995</td>
</tr>
</tbody>
</table>
Gaussian Approximation to a Binomial Distribution  To see how the cumulative distribution is useful, consider the problem of calculating the probabilities of the binomial distribution when $n$ is large. Using Equation 1 involves a lot of calculation. It turns out, however, that the number of successes in a binomial distribution can be approximated with a normal distribution for large $n$. Let us phrase the question as, Given a set of $n$ trials, what is the probability that the frequency of successes falls between the interval specified by $f_1$ and $f_2$? The normal distribution allows the following approximation:

$$\Phi(x_2) - \Phi(x_1) = P(f_1 < X \leq f_2) = \frac{1}{\sqrt{2\pi}} \int_{x_1}^{x_2} e^{-\frac{u^2}{2}} du$$

where

$$x_1 = \sqrt{\frac{n}{pq}} (f_1 - p)$$

and

$$x_2 = \sqrt{\frac{n}{pq}} (f_2 - p)$$
Example  Use the Gaussian approximation to the binomial distribution to estimate $P(k > 27)$ for $n = 50$, $p = \frac{2}{3}$.

To solve this first use the fact that

$$P(k > 27) = 1 - P(k \leq 27)$$

Next, since $f_2 = \frac{k}{n}$, $x_2$ is given by

$$x_2 = \sqrt{\frac{n}{pq}}(f_2 - p) = \sqrt{\frac{50}{\frac{4}{9}} \left(\frac{27}{50} - \frac{2}{3}\right)} = -1.344$$

So that

$$\Phi(x_2) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x_2} e^{-\frac{u^2}{2}} du = 1 - \Phi(x_2)$$

Interpolating from Table 1, $\Phi(1.344) = 0.90$, so that, putting all this together, the probability of $k$ exceeding 27 is 0.90.

Does this jibe with your intuition? From Table 1, $\Phi(0) = 0.5$. This is the result when $f = p$ as you might expect. With an increasing number of samples, it becomes more and more likely that the measured frequency will be close to the mean.
Random Vectors

In this case vectors are drawn from some random distribution that captures the natural variations in the world. A random vector $\mathbf{X}$ is specified by a probability density function $p(\mathbf{X})$, where formally

$$p(\mathbf{X}) = \lim_{\Delta x_i \to 0} \frac{P(\mathbf{X} \in I)}{\prod \Delta x_i}$$

where

$$I = \{ \mathbf{X} : x_i < X_i \leq x_i + \Delta x_i, \forall i \}$$

Although a random vector is fully characterized by its density function, such functions are often difficult to determine or mathematically complex to use. These limitations motivate modeling distributions with functions that can be described by a low number of parameters. The most important of such parameters are the mean vector and covariance matrix, which are just generalizations of the mean and variance of scalar random variables to vector random variables.

The mean vector is defined by

$$\mathbf{M} = E\{\mathbf{X}\} = \int \mathbf{X} p(\mathbf{X}) d\mathbf{X}$$

and the covariance matrix by

$$\Sigma = E\{(\mathbf{X} - \mathbf{M})(\mathbf{X} - \mathbf{M})^T\}$$
In practice, with real data you will use the sample mean vector and sample covariance matrix. Where $X^k, k = 1, N$ are the samples,

$$M = \frac{1}{N} \sum_{k=1}^{N} X^k$$

$$\Sigma = \frac{1}{N} \sum_{k=1}^{N} (X^k - M)(X^k - M)^T$$

You will need more than three samples in practice, but to illustrate the calculations, suppose

$$X^1 = \begin{pmatrix} -1 \\ 3 \\ 1 \end{pmatrix}, \quad X^2 = \begin{pmatrix} 2 \\ 1 \\ -1 \end{pmatrix}, \quad X^3 = \begin{pmatrix} 2 \\ 2 \end{pmatrix}$$

Then the mean value is

$$M = \frac{1}{3} \begin{pmatrix} 3 \\ 6 \\ 3 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix}$$

So that

$$X^1 - M = \begin{pmatrix} -2 \\ 1 \\ 0 \end{pmatrix}, \quad X^2 - M = \begin{pmatrix} 1 \\ -1 \\ -2 \end{pmatrix}, \quad X^3 - M = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

and the covariance matrix is given by

$$\Sigma = \frac{1}{3} \left\{ \begin{bmatrix} 4 & -2 & 0 \\ -2 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} + \begin{bmatrix} 1 & -1 & -2 \\ -1 & 1 & 2 \\ -2 & 2 & 4 \end{bmatrix} + \begin{bmatrix} 1 & 0 & 2 \\ 0 & 0 & 0 \\ 2 & 0 & 4 \end{bmatrix} \right\}$$

$$= \frac{1}{3} \begin{bmatrix} 6 & -3 & 0 \\ -3 & 2 & 2 \\ -2 & 2 & 8 \end{bmatrix}$$
The Gaussian Distribution

\[
N(x|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp \left\{ -\frac{1}{2\sigma^2} (x - \mu)^2 \right\}
\]

\[
N(x|\mu, \Sigma) = \frac{1}{(2\pi)^{D/2} |\Sigma|^{1/2}} \exp \left\{ -\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right\}
\]
Maximum Likelihood for the Gaussian (1)

Given i.i.d. data $X = (x_1, \ldots, x_N)^T$, the log likelihood function is given by

$$\ln p(X | \mu, \Sigma) = -\frac{ND}{2} \ln(2\pi) - \frac{N}{2} \ln |\Sigma| - \frac{1}{2} \sum_{n=1}^{N} (x_n - \mu)^T \Sigma^{-1} (x_n - \mu)$$

Sufficient statistics

$$\sum_{n=1}^{N} x_n \quad \sum_{n=1}^{N} x_n x_n^T$$
Maximum Likelihood for the Gaussian (3)

Under the true distribution

$$\mathbb{E}[\mu_{\text{ML}}] = \mu$$

$$\mathbb{E}[\Sigma_{\text{ML}}] = \frac{N-1}{N} \Sigma.$$ 

Hence define

$$\tilde{\Sigma} = \frac{1}{N-1} \sum_{n=1}^{N} (x_n - \mu_{\text{ML}})(x_n - \mu_{\text{ML}})^T.$$
Mixtures of Gaussians (1)

Old Faithful data set

Single Gaussian

Mixture of two Gaussians
Mixtures of Gaussians

Suppose that the observed data can now be a mixture of two states, \( k = 1, 2 \), and that \( P(k) \) is the probability of the \( k \)th state.

Let’s assume that each of these can be described by a Gaussian; that is,

\[
p(x_i|k) = \frac{1}{\sqrt{2\pi\sigma_k}} e^{-\frac{(x_i-m_k)^2}{2\sigma_k^2}} \text{ for } k = 1, 2; i = 1, \ldots, n
\]

Now the probability of seeing the data sample can be written in terms of the internal state using Bayes’ rule:

\[
p(x) = \sum_k p(x|k)P(k)
\]

so that if you are given the data and want to estimate the probability that it came from a particular state you can use Bayes’ rule:

\[
P(k|x_i) = \frac{p(x_i|k)P(k)}{\sum_j p(x_i|j)P(j)}
\]  \(2\)

This equation in turn allows the \( a \ priori \) state probabilities to be estimated as an average

\[
P(k) = \frac{1}{n} \sum_i P(k|x_i)
\]  \(3\)

Now the probability of seeing all the data, given that the samples are generated independently, is

\[
p(x_1, \ldots, x_n) = \prod_i \sum_k p(x_i|k)P(k)
\]
**Maximum Likelihood Estimation** To choose the parameters for each of the internal states, use maximum likelihood estimation by maximizing the log likelihood function given by

$$\log L = \log p(x_1, \ldots, x_n) = \sum_i \log \sum_k p(x_i|k)P(k)$$

From calculus you know that at the maximum, the partial derivatives with respect to the parameters will be zero; that is,

$$\frac{\partial \log L}{\partial m_k} = \sum_i p(x_i|k)\frac{(x_i - m_k)}{\sigma^2_k} = 0 \quad \text{(4)}$$

$$\frac{\partial \log L}{\partial \sigma^2_k} = \sum_i p(x_i|k)\frac{(x_i - m_k)^2}{2\sigma^4_k} = 0 \quad \text{(5)}$$

These equations can be solved for expressions for $m_k$ and $\sigma^2_k$ in terms of the data points $x_i$. 
Algorithm: Estimating States with Expectation Maximization

Initialize the model by guessing values for $m_k$ and $\sigma_k$. Until the model parameters converge, do the following:

1. Choose the internal state probabilistically according to Equation 3, which in turn uses Equation 2.
2. Now that the state is estimated, update its parameters by using Equation 4

$$m_k = \frac{\sum_i P(k|x_i)x_i}{\sum_i P(k|x_i)}$$

and Equation 5

$$\sigma_k^2 = \frac{\sum_i P(k|x_i)(x_i - m_k)^2}{\sum_i P(k|x_i)}$$