Networks - Week 1 - Randomness and Matrices

Antonio León Villares

October 2023

Contents

Based on the notes by Renaud Lambiotte and Pete Grindrod, Chapter 1

1 Basic Probability

1.1 Basic Rules with Conditional Probability

1.1.1 Proposition: The Chain Rule

Consider a collection of random variables $X_1, \ldots, X_n$. Then:

$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i \mid X_{i+1}, \ldots, X_n)$$

where:

$$P(X_n \mid X_{n+1}) = P(X_n)$$

Proof. This comes from repeated application of the definition of conditional probability:

$$P(X, Y) = P(X \mid Y)P(Y)$$

□

1.1.2 Proposition: Bayes' Rule

$$P(X \mid Y) = \frac{P(Y \mid X)P(X)}{P(Y)}$$

where:

- $P(X)$ *is the* ***prior***
- $P(X \mid Y)$ *is the* ***posterior***
- $P(Y \mid X)$ *is the* ***likelihood***

Proof. This uses the commutativity in conditional probability:

$$P(X, Y) = P(Y, X) \implies P(X \mid Y)P(Y) = P(Y \mid X)P(X)$$

□

1.1.3 Definition: Odds

*The **odds** of a given random variable X are:*

$$O(X) = \frac{P(X)}{P(\neg X)} = \frac{P(X)}{1 - P(X)}$$

*The **odds** of X given Y are:*

$$O(X \mid Y) = \frac{P(X \mid Y)}{P(\neg X \mid Y)} \in [0, \infty)$$

*Using **Bayes' Rule**, this can be rewritten as:*

$$O(X \mid Y) = \frac{\frac{P(Y \mid X)P(X)}{P(Y)}}{\frac{P(Y \mid \neg X)P(\neg X)}{P(Y)}} = \frac{P(Y \mid X)}{P(Y \mid \neg X)} O(X)$$

1.2 The Binomial Distribution

1.2.1 Definition: Discrete Random Variables

*A **discrete random variable** (DRV) is a variable which takes a number of mutually exclusive, distinct values.*

---

*If these values are **finite**, say $\{\alpha_1, \ldots, \alpha_k\}$, this describes a **multinomial distribution**, with probabilities $p_k, k \in [1, K]$ whose sum is 1.*

1.2.2 Definition: Binomial Distribution

*A **binomial distribution** is a distribution whereby an experiment is repeated n times (independently), and each experiment has 2 possible outcomes (with probability p and $1 - p$).*
If k outcomes are "positive" and $n - k$ are "negative", the probability of such an experiment sequence is:

$$\binom{n}{k} p^k (1-p)^{n-k}$$

1.3 Continuous Random Variables

1.3.1 Definition: Updating Beliefs with Continuous Random Variables

*Let θ be a **continuous random variable** (taking values in some set Ω). The **prior** distribution of θ is given by some non-negative function:*

$$P(\theta) = f(\theta)$$

*If we observe **new data** D, we **update our beliefs** via the chain rule:*

$$P(\theta \mid D) = \frac{P(D \mid \theta) f(\theta)}{P(D)} \propto P(D \mid \theta) f(\theta)$$

*We call $P(D \mid \theta)$ a **model**, since it informs about how a model θ perceives the observed data D.*

Since $P(D)$ is just a constant which normalises the distribution, it plays no role in the actual distribution of θ. If we want to compute a probability distribution, we can just use:

$$P(\theta \mid D) = \frac{P(D \mid \theta) f(\theta)}{\int_\Omega P(D \mid \theta) f(\theta) \, d\theta}$$

1.3.2 Definition: Maximum Likelihood Estimation

***Maximum Likelihood Estimation** (MLE) is a technique to find the θ most likely to explain the data, by finding the mode of the distribution:*

$$\theta^* = \text{argmax}_\theta \, P(D \mid \theta)$$

If we have n independent observations $\{x_i\}_{i \in [1,n]}$, the likelihood of the data given θ is given by:

$$\mathcal{L} = \prod_{i=1}^{n} P(x_i \mid \theta)$$

*In practice, we'd opitmise the **log likelihood**:*

$$\log \mathcal{L} = \sum_{i=1}^{n} \log(P(x_i \mid \theta))$$

1.4 Definition: Moments of a Continuous Random Variables

*Let θ be some distribution. Then, the k-th **moment** of the distribution is given by::*

$$\left\langle \theta^k \right\rangle = \frac{\int_\Omega \theta P(D \mid \theta 9 f(\theta)\, d\theta}{\int_\Omega P(D \mid \theta) f(\theta)\, d\theta}$$

In particular:

- *the **moment** with $k = 1$ is the **expected value** of the distribution*
- *the **variance** of the distribution is:*

$$\sigma_\theta^2 = \left\langle (\theta - \langle\theta\rangle)^2 \right\rangle = \left\langle \theta^2 \right\rangle - \langle\theta\rangle^2$$

1.5 Definition: Improper Distributions

*Let $f(\theta)$ be a distribution. Then, $f(\theta)$ is an **improper distribution** if it has **infinite probability mass/density**, and thus can't be summed/integrated to unity.*

Improper distributions will still have maxima and be non-negative, so maximum likelihood methods (like gradient-based methods) can still be applied.

2 Matrices

2.1 Hermitian Matrices

2.1.1 Definition: Hermitian Matrices

*A **Hermitian** (or **self-adjoint**) matrix A is one such that:*

$$A = A^* (= (\overline{A})^T)$$

*where $\overline{A}$ denotes the **complex conjugate** matrix of A.*

2.1.2 Theorem: Spectral Theorem

*Let A be **Hermitian** on the (inner product) vector space $\mathbb{C}^n$. Then, there exists an **orthonormal basis** of $\mathbb{C}^n$ consisting of **eigenvectors** of A. Moreover:*

- *each **eigenvalue** $\lambda_1, \ldots, \lambda_n$ of A is **real***
- *A is **diagonalisable**: in fact, there exists a **unitary** matrix P (that is, a matrix such that $P^*P = \mathbb{I}$), such that:*

$$P^{-1}AP = P^*AP = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$$

.

More details can be found in these notes for Honours Algebra at the University of Edinburgh.

2.1.3 Definition: Normal Matrices

*A matrix is **normal** if it **commutes** with its **adjoint**:*

$$AA^* = A^*A$$

*By definition, all **Hermitian matrices** are **normal**.*

---

*Moreover, a matrix is **normal** if and only if it is **diagonalisable**.*

2.2 The Perron-Frobenius Theorem

2.2.1 Definition: Irreducible Matrix

*Let A be a **non-negative** matrix. Then, A is **irreducible** if:*

$$\forall (i,j),\ \exists k \in \mathbb{N}\ :\ (A^k)_{ij} > 0$$

2.2.2 Definition: Spectral Radius

*Let A be a matrix. The **spectral radius** of A, $\rho(A)$, is the **maximum** of the **absolute values** of its **eigenvalues**.*

2.2.3 Theorem: The Perron-Frobenius Theorem

The Perron-Frobenius Theorem states that real, square matrices with strictly positive entries have a unique largest real eigenvalue, and whose corresponding eigenvector has strictly positive components.

Let A be a $n \times n$ matrix, such that A:

- *is **irreducible***
- ***non-negative***
- *has **spectral radius** $\rho(A) = r > 0$*

Then:

1. *r is an **eigenvalue** of A (called the **Perron-Frobenius** eigenvalue)*
2. *r is **simple**. In particular:*
   - *r has **algebraic multiplicity** 1 (it is not a repeated eigenvalue)*
   - *r has **geometric multiplicity** 1 (both right and left eigenspaces are one-dimensional - this is because geometric multiplicity is bounded by algebraic multiplicity*
3. *A has left/right **eigenvectors** with **eigenvalue** r, and whose components are **all positive***
4. *the only **eigenvectors** whose components are **all positive** are those associated to r*
5. *r is **bounded** above/below by the maximum and minimum **row** sums of A (and also the **column** sums):*

$$\min_{i \in [1,n]} \sum_{j=1}^{n} A_{ij} \leq r \leq \underset{i \in [1,n]}{max} \sum_{j=1}^{n} A_{ij}$$

2.2.4 Proposition: Singular Value Decomposition of a Matrix

Let M be an $m \times n$ matrix with complex entries. The ***singular value decomposition*** *of M is a* ***factorisation*** *of the form:*

$$M = U\Sigma V^*$$

where:

- *U is an $m \times m$* ***unitary*** *matrix (whose* ***columns*** *are the* ***eigenvectors*** *of M^*M, called the* ***left singular vectors*** *of M).*
- *Σ is an $m \times n$* ***diagonal*** *matrix, with* ***non-negative, real*** *diagonal elements (whose elements are the* ***square root*** *of the non-zero eigenvalues of MM^* or M^*M, called the* ***singular values*** *of M).*
- *V^* is the adjoint of the $n \times n$* ***unitary*** *matrix V (the* ***columns*** *of V are the* ***eigenvectors*** *of MM^*, called the* ***right singular vectors*** *of M).*

- **Is SVD unique?**
  - the **singular values** are unique
  - however, U, V needn't be unique
- **How can SVD be derived?**
  - we exploit the fact that MM^* and M^*M will be **real**, **symmetric** matrices, which are diagonalisable
  - see these notes for extra details
- **How is SVD related to eigenvalue decomposition?**
  - if M is a normal matrix, it is diagonalisable
  - this diagonalisation can be doen through the **eigenvalue decomposition**:

    $$M = UDU^*$$

    where U is a **unitary** matrix whose columns are the **eigenvectors** of M, and D is a diagonal matrix containing the **eigenvalues** of M
  - in this case, the SVD will coincide with the eigenvalue decomposition

---

2.2.5 Proposition: Pseudo-Inverse from SVD

Let M be a matrix. Then, its ***pseudo-inverse*** *is:*

$$M^+ = (A^*A)^{-1}A^*$$

IOf we know the SVD of M, then:

$$M^+ = V\Sigma^+U^*$$

where Σ^+ is the ***pseudo-inverse*** *of Σ (which can be obtained by replacing every non-zero diagonal entry by its reciprocal).*

2.2.6 Exercises

1. **Suppose an $n \times n$ matrix A is non-negative and the spectral radius of A is given by the Perron-Frobenius eigenvalue, r. Let $\alpha \in (0, r)$. Then consider:**

   $$(\mathbb{I} - \alpha A)^{-1}$$

   Show that if this matrix is strictly positive then A is irreducible. Is the converse true? Show that if the matrix:

   $$\exp(A) = \sum_{i=0}^{\infty} \frac{A^i}{i!}$$

   is strictly positive, then A is irreducible. Is the converse true?

2. **Suppose A is normal and invertible. Then there is a unitary U such that $A = U\Lambda U^T$ and Λ is diagonal containing the eigenvalues of A. Let $f : R \to R$ be any function that is well-defined at all of the eigenvalues of A. Define**

   $$f(A) = Uf(\Lambda)U^T$$

   where $f(\Lambda)$ is diagonal; with f applied to each corresponding element of Λ.

   (a) **Show that if Q is any polynomial:**

   $$Q(x) = \sum q_i x^i$$

   then:

   $$Q(A) = \sum q_i A^i$$

   (b) **Similarly, show that:**

   $$Q(A - \mathbb{I}) = \sum q_i (A - \mathbb{I})^i = UQ(\Lambda - I)U^T$$

   (c) **Finally, show that:**

   $$Q(A)^{-1} = UQ(\Lambda)^{-1}U^T$$

3. **Suppose A is normal and its spectral radius is $\rho(A) < \frac{1}{\alpha}$ for some $\alpha > 0$. Then, consider:**

   $$(\mathbb{I} - \alpha)^{-1} = U(\mathbb{I} - \alpha\Lambda)^{-1}U^T$$

   Show that this is the geometric series:

   $$S = \sum \alpha^i A^i$$

2.3 Laplacians of Matrices

2.3.1 Definition: The Laplacian of a Matrix

Let A be a $n \times n$ matrix which is:

- ***real***
- ***non-negative***
- ***normal***

*If $\underline{s} = \underline{1} \in \mathbb{R}^n$, then $A\underline{s} = (d_1, \ldots, d_n)^T$ contains the **row sums** of A.*

If we define:

$$D = \text{diag}(d_1, \ldots, d_n)$$

*the combinatorial **Laplacian** of A is the **symmetric** matrix:*

$$L = D - A$$

Notice, the fact that A is real and normal implies that A is symmetric. In particular, since A is normal, it is diagonalisable, so:

$$A = UDU^T$$

for some orthogonal matrix U. Then:

$$A^T = UD^TU^T = UD^TU = A$$

so A is symmetric.

2.3.2 Proposition: Properties of the Laplacian Matrix

*Let L be the **Laplacian** matrix of some $n \times n$ matrix A. Then:*

1. $$L\underline{s} = \underline{0}$$

2. *For any $\underline{w} \in \mathbb{R}^n$, we have a **quadratic form**:*

$$\underline{w}^T L \underline{w} = \frac{1}{2} \sum_{i,j=1}^{n} (w_i - w_j)^2 A_{ij}$$

*In other words, L is always **positive-semidefinite** and if $\underline{w}$ is an **eigenvector** of L corresponding to the 0 **eigenvalue**, then the components of $\underline{w}$ must all be equal (so $w_i = w_j$ for any i, j).*

Proof.

(1)

$$L\underline{s} = D\underline{s} - A\underline{s} = \underline{0}$$

(2)

We compute directly:

$$\begin{aligned}
\underline{w}^T L \underline{w} &= \underline{w}^T (D - A) \underline{w} \\
&= \underline{w}^T D \underline{w} - \underline{w}^T A \underline{w} \\
&= \sum_{i=1}^{n} d_i w_i^2 - \sum_{i,j=1}^{n} A_{ij} w_i w_j \\
&= \sum_{i,j=1}^{n} A_{ij} w_i^2 - \sum_{i,j=1}^{n} A_{ij} w_i w_j
\end{aligned}$$

But now, notice that since A is symmetric:

$$\sum_{i,j=1}^{n} A_{ij} w_i^2 = \sum_{i,j=1}^{n} A_{ij} w_j^2 \implies \sum_{i,j=1}^{n} A_{ij} w_i^2 = \frac{1}{2} \sum_{i,j=1}^{n} A_{ij} (w_i^2 + w_j^2)$$

Hence:

$$\underline{w}^T L \underline{w} = \frac{1}{2} \sum_{i,j=1}^{n} (w_i - w_j)^2 A_{ij}$$

□

3 Markov Chains

3.1 Definition: Markov Chain

Consider some structure consisting of n ***states*** *in* ***discrete time****. A* ***Markov chain*** *is a* ***stochastic process****, whereby the probability of observing a state at time $t+1$, X_{t+1} solely depends on the previous state X_t.*

3.2 Definition: Stationary Markov Chains

A ***stationary Markov Chain*** *is a* ***Markov Chain*** *where the* ***transition probability*** *doesn't depend on t:*

$$P(X_{t+1} = j \mid X_t)$$

These ***stationary transition probabilities*** *can be stored as a* ***transition matrix*** *with entries:*

$$T_{ij} = P(X_{t+1} = j \mid X_t)$$

Moreover, we require that:

$$\sum_{j=1}^{n} T_{ij} = 1$$

(from a given state, we must always go to some state, including the same one)

3.3 Types of States

3.3.1 Definition: Ergodic Set

Let S be a set of states. S is an ***ergodic set*** *if:*

- *for any $i, j \in S$, one can reach j from i solely through elements of S*
- *once an element of S is reached, all subsequent transitions happen within S*

3.3.2 Definition: Absorbing State

*State i is **absorbing** if it can't be escaped once reached:*

$$T_{ii} = 1$$

***Absorbing states** form a 1-element **ergodic set**.*

3.3.3 Definition: Transient State

*A state i is **transient** if it isn't part of **any erogdic set**.*

3.4 Evolution of Markov Chain Process

- **At some time $t+1$, how can we compute the probability of reaching some state j from the previous state?**
  - let $p_j(t)$ denote the probability of reaching state j at time t
  - then, since we assume that **transitions** and **states** are independent:

$$p_j(t+1) = \sum_{i=1}^{n} p_i(t) T_{ij}$$

  - if we want to compute probabilities for **all** states, we can use **matrix multiplication**:

$$\underline{p}(t+1) = \underline{p}(t) T$$

    where $\underline{p}$ is a row vector $(p_1(t), \ldots, p_n(t))$
  - even more succintly (depending solely on the **initial state**):

$$\underline{p}(t) = \underline{p}(0) T^t$$

3.4.1 Definition: Stationary Density

*The **non-negative stationary density** is a vector:*

$$\underline{p}^* = (p_1^*, \ldots, p_n^*)$$

where:

$$p_i^* = \lim_{t \to \infty} p_i(t)$$

and:

$$\underline{p}^* = \underline{p}^* T$$

- **How is the stationary density related to T?**
  - $\underline{p}^*$ is the **left eigenvector** of T, with **eigenvalue** 1
- **Under what conditions does an eigenvalue of unity exist for T?**
  - if the set of n states is **ergodic**, then T will have an eigenvalue 1
- **What special type of eigenvalue is 1?**
  - for a **transition matrix** T with an **erogdic set** of states, the **eigenvalue** 1 will be the **Perron-Frobenius Eigenvalue**
  - the **stationary density** is the **Perron-Frobenius Eigenvector** (which we know has all positive components, as expected)
- **How does the difference between $\underline{p}^*$ and $\underline{p}(t)$ vary as $t \to \infty$?**
  - the discrepancy decays **exponentially**
  - it depends on the second eigenvalue with the largest modulo:

$$\propto |\lambda_2|^t$$

  - in general, speed of convergence depends on the difference or ratio of λ_2 and r (the Perron-Frobenius eigenvalue, which won't always be 1)
- **What is the spectral gap?**
  - the value $1 - \lambda_2$
  - if the spectral gap is **large**, the Markov chain converges rapidly

4 Poisson Processes

4.1 Definition: Poisson Processes

*A **Poisson Process** is a model for events which occur discretely, and in apparent random fashion.*

---

*In particular, consider a window of time Δt, with probability of an event happening during the window of q. Then, the **event rate** is given by:*

$$\lambda = \frac{q}{\Delta t}$$

For λ to be well-defined, we require that:

- $q \to 0$ *as* $\Delta t \to 0$
- *as* $\Delta t \to 0$ *we don't allow multiple events to happen in a single time window*

4.2 Properties of Poisson Processes

4.2.1 Proposition: Distribution of Inter-Event Times

4.2.2 Proposition: Distribution of Number of Events

5 Random Walks

Random walks are useful in modelling trajectories in space, which can, for example, extract information from the structure of networks

5.1 Definition: Random Walks

Consider a one-dimensional space (i.e the real line). A ***random walker*** *performs a jump whose* ***length*** *and* ***direction*** *are* ***random variables****.*

In particular, the ***probability density*** *of transition is denoted $f(r)$, such that the probability that a walker at x arrives in*

$$[x + r, x + r + \Delta r$$

in 1 jump is:

$$f(r)\Delta r$$

Moreover, we must have that:

$$\int_{-\infty}^{\infty} f(r)\ dr = 1$$

5.2 Proposition: Solution to Random Walks

Let $p(x, t)$ denote the probability of a ***random walker*** *being at x after t steps. Then, if $f(r)$ has* ***finite*** *mean and variance:*

$$p(x; t) = \frac{1}{(2\pi Dt)^{1/2}} e^{-\frac{(x-vt)^2}{4Dt}}$$

where D, v are constants.

Proof. Assuming that jumps are independent events, the probability of reaching x at time t from any other x' is:

$$p(x; t) = \int_{-\infty}^{\infty} f(x - x')p(x'; t - 1)\ dx'$$

Notice, this looks exactly like a convolution between f, p. If we apply the Fourier transform, we can convert this into a product:

$$\hat{p}(k; t) = \hat{f}(k)\hat{p}(k; t - 1)$$

where:

$$\hat{g}(k) = \int_{-\infty}^{\infty} g(x)e^{-ikx}\ dx$$

Now, at the start of the walk ($t = 0$) we know for certain where the random walker is, so we can model:

$$p(x; 0) = \delta(x)$$

where δ is the Dirac distribution. But the Fourier Transform of δ is:

$$\hat{p}(k;0) = 1$$

so it follows that:

$$\hat{p}(k;t) = \hat{f}(k)\hat{p}(k;t-1) \implies \hat{p}(k;t) = [\hat{f}(k)]^t$$

Now, if we take the Inverse Fourier Transform:

$$p(x;t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} [\hat{f}(k)]^t e^{ikx}\ dk$$

Whilst the function depends on $\hat{f}$, the behaviour of the random walk as t grows only depends on some of its properties.

In particular, if the mean and variance of f are finite, the solution converges to:

$$p(x;t) = \frac{1}{(2\pi Dt)^{1/2}} e^{-\frac{(x-vt)^2}{4Dt}}$$

□

Notice, it is expected that a Gaussian profile appears: after all, a random walk is nothing but a sum of independent steps, drawn from a smooth distribution f with finite mean and variance. That is, the Central Limit Theorem applies!

5.3 Definition: Lévy Flight

> *A **Lévy Flight** is a **non-diffusive** spatial process: f doesn't have finite variance, so large jumps are possible.*

6 Power Law Distributions

Power Law distributions are defined by properties whose probability density changes as powers.

6.1 Definition: Pareto Distribution

*The **Pareto Distribution** is a power-law distribution defined by:*

$$p(x) = Cx^{-\alpha}$$

where:

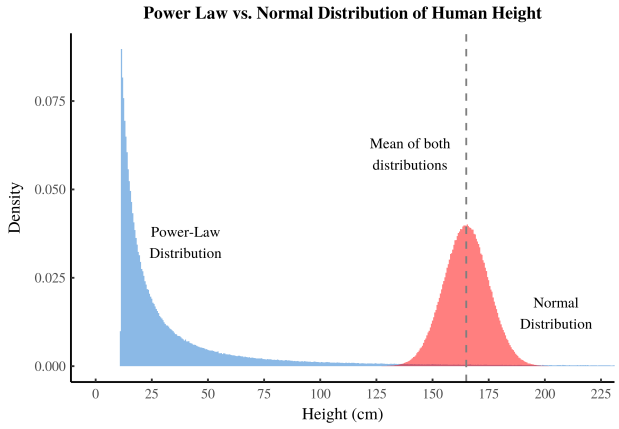
- $x > x_{min}$, *and* x_{min} *is the minimum value taken by the random variable*
- $\alpha > 1$
- C *is a* ***normalisation constant***

$$C = (\alpha - 1)x^{\alpha - 1}$$

such that:

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

- **How do power law distributions differ from Gaussian distributions?**
  - **Gaussian distributions** are more "balanced", with very little probability density assigned to its tails
  - on the other hand, power law distributions have:
    - a vast majority of instances with small values
    - few (but not negligible) very large values
  - **power-law distributions** are said to have a "fat tail", as it is more populated than other distributions (like the exponential distribution)



- **How are power laws related to Zipf's Law?**

- **Zipf's Law** gives a relationship between **frequency** and **ranking** of certain phenomena (typically languages - see these notes on NLP)
- turns out that **Zipf's Law** is just a specific instance of a **power-law distribution**
- beyond linguistics, **power-law distributions** can also be used to model individual wealth and city populations (for example)

6.2 Proposition: Moments of the Pareto Distribution

*Let $\beta > \alpha - 1$. Then, the βth moment of the **Pareto distribution** is:*

$$\left\langle x^{\beta} \right\rangle = \int_{x_{min}}^{\infty} x^{betap(x)\ dx = \frac{\alpha-1}{\alpha-1-\beta} x_{min}^{\beta}}$$

Clearly, such moments are undefined when $\beta > \alpha - 1$

6.3 Definition: Cauchy Distribution

*The **Cauchy Distribution** is proportional to:*

$$\frac{1}{1+x^2}$$

*and behaves **asymptotically** like the **Pareto distribution** with $\alpha = 2$.*

- **Does the Cauchy Distribution have a well-defined mean?**
  - notice, when $\alpha = 2$, the mean of the **Pareto Distribution** diverges
  - since the **Cauchy Distribution** behaves asymptotically like the **Pareto Distribution**, it doesn't have a defined mean (or variance)
  - in particular, this menas tha tthe CLT doesn't apply

6.4 Proposition: Properties of Power-Law Distributions

1. ***Scale Invariance**:*

   $$p(c_1 x) = c_2 p(x)$$

   In other words, the properties of the sytem aren't affected by a change in units.

2. ***Log-Log Plot**:*

   $$\log(p(x)) = \log C - \alpha \log(x)$$

7 Information Theory

7.1 Definition: Entropy of Random Variable

*The **entropy** of a **random variable** (denoted H) is a measure of the **uncertainty** we have about the variable, before observing it:*

$$H(X)0 - \sum_x p(x)\log(p(x))$$

- **What is the minimum value of entropy?**
  - when the RV is **deterministic** ($P(X = x_0) = 1$ for some x_0), we get that $H(X) = 0$
  - this corresponds with the notion that there is **no uncertainty**
- **When does entropy achieve its maximum value?**
  - if $p(x)$ is **uniformly distributed** such that:

$$p(x) = \frac{1}{n}$$

then H is maximised, and:

$$H(X) = \log(n)$$

7.2 Definition: Joint Entropy

*Let X, Y be a pair of **discrete random variables** with joint distribution $p(x, y)$. Then, their **joint entropy** is:*

$$H(X, Y) = \sum_x \sum_y p(x, y)\log(p(x, y))$$

7.3 Definition: Conditional Entropy

Let X, Y be a pair of **discrete random variables** with joint distribution $p(x, y)$. Then, their **conditional entropy**:

$$H(X \mid Y) = \sum_y p(y)H(X \mid Y = y) = \sum_x \sum_y p(x, y)\log(p(x \mid y))$$

7.4 Definition: Chain Rule of Entropy

*The **joint entropy** and **conditional entropy** are related by the **chain rule**:*

$$H(X,Y) = H(X) + H(Y \mid X)$$

7.5 Definition: Mutual Information

*Let X, Y be a pair of **discrete random variables** with joint distribution $p(x,y)$. Then, their **mutual information** is the amount of information gained on X by knowing the value of Y:*

$$I(X,Y) = H(X) - H(X \mid Y) = H(X) + H(Y) - H(X,Y)$$

Alternatively:

$$I(X,Y) = \sum_x \sum_y p(x,y) \log \left(\frac{p(x,y)}{p(x)p(y)} \right)$$

- **To what does mutual information get reduced if Y is perfectly informative; that is, it tells us everything about X?**
  - in such a case, $H(X \mid Y) = 0$, and:
    $$I(X,Y) = H(X)$$
- **intuitively, what does mutual information aim to measure?**
  - the **non-linear correlations** between random variables
  - it measures the cost of assuming that 2 variables are independent (when in fact they aren't)