Data Fundamentals (H)

Glossary

$$\mathbf{x}\mathbb{R}$$

This document lists the bolded terms for the lecture notes, and provided a brief explanation of each term. The Unit the term is explained in is provided in the section titles.

week_0_intro.ipynb

- **Python** seriously.

week_1_numerical_i.ipynb

- **multidimensional numerical array / nD array** Fundamental data type which holds rectangular, multidimensional arrays of numbers of uniform type in a dense compact form, with a large number of efficient operations that can be applied in a vectorised form.
- **vectorised / vectorised computation** code which does computations on many values simulatenously, without explicit iteration over those values
- **Single Instruction Multiple Data / SIMD** A hardware instruction that applies the same operation (e.g. addition) to multiple values in a single operation.
- **GPU / graphics processor units** The fast number processing unit on your computer.
- **vector** A one dimensional array of numbers.
- **matrix** A two dimensional array of numbers.
- **tensor** an array with >2 dimenions.
- **linear algebra** The algebra of matrices (strictly, of linear maps)
- **linear map** A function which satisfies linearity (see below).
- **vector space** A mathematical space which consists of tuples of numbers (usually reals) and two operations: vector addition and scalar multiplication
- **array functions** functions which are applied to nD arrays
    - **array arithmetic** arithmetic operations on nD arrays, like elemenwise addition
    - **indexing and slicing** Seleting individual elements, or rectangular blocks of elements from an nD array.
    - **array generation** array functions which create new arrays, like np.zeros
    - **array rearranging** array functions which adjust the apparent layout of arrays, like transpose, reshape and fliplr
    - **order operations** array functions which use the rank ordering of values, like np.sort, np.argsort, np.argmin, etc.
    - **aggregate functions** array functions which summarise an array, like np.sum or np.mean
- **Vector operations** array functions which depend on the mathematical structure of vectors, like the cross product
- **Matrix operations** functions which depend on the mathematical structure of matrices, like the determinant or the inverse
- **Signal processing operations** functions which treat arrays as sampled signals, like convolution and the DFT.
- **NumPy** Library providing nD arrays for Python.
- **SciPy** Library providing scientific operations on nD arrays for Python
- **Matplotlib** Library providing plotting facilities for nD arrays for Python.
- **dtype** Datatype of an nD array, e.g. `float32`
- **shape** Size of an nD array, e.g. shape (4,8) is 4 row 8 column 2-dimensional nDarray
- **rank** number of dimensions of an nD array (rank 1->vector, rank 2->matrix, rank n tensor)
- **mutable** being able to be changed after creation
- **slice** a rectangular sub-region of an array (may have "skipped" elements)
- **elementwise arithmetic** arithmetic operation applied each element of an nD array
- **scalar arithmetic** array/scalar operations, like `x+2`
- **array arithemetic** array/array operations, like `x+x`
- **Boolean array** an array of True/False values, the result of a Boolean test on arrays. Can be used directly as an index in NumPy.
- **broadcasting** rules for automatically replicating an array to allow arithmetic between arrays of different shapes, as long as the *last dimensions* match exactly

- **transpose** exchange of rows and columns of a 2D array, or more generally reversal of the order of dimensions
- **tiling** repeating an array in a regular pattern
- **masking** selecting non-rectangular elements of an array, e.g. using a Boolean array as an index
- **cumulative sum** a running sum
- **argsort** the set of (1D) array indices such that indexing the array in that order would result in a sorted array.

week_2_numerical_ii.ipynb

- **striding** The trick used to treat a flat sequence of numbers as a multidimensional array, by storing the offsets to jump to the next element in any dimension
- **stride** the offset (in bytes, typically) to jump to get to the next value in a given dimension
- **C ordering/ row-major** last dimension changes fastest
- **Fortran ordering / column-major** first dimension changes fastest
- **integers** whole numbers
- **floating point numbers** approximations to real numbers with variable absolute precision across their range
- **numerical issues / numerical precision** problems arising from the approximation of real numbers by floating-point numbers
- **scientific notation** numbers written with a single digit before a decimal point, and an explicit exponent, like 5.242e4 = 5.242 x 10^4 = 52420
- **float** a floating point number
- **sign bit** the bit indicated whether a float is negative or positive
- **exponent** the part of a floating point number that specifies the bitshift to be applied (the base 2 exponent)
- **mantissa** the part of the floating point number specifying the fractional part of a (binary) number, following 1.xxxxx...
- **IEE754** standard defining storage formats and standard operations on floating point numbers
- **binary32 / float32 / single precision** IEE754 32 bit floating point numbers
- **binary64 / float64 / double precision** IEEE754 64 bit floating point numbers
- **float128 / quadruple precision** rare format for very precise computations
- **float256 / octuple precision** even rarer format for super precise computations (like astronomical predictions)
- **floating point exception / FPE** an event triggered by a floating point operation which is problematic.
    - **Invalid Operation** operation without defined value, like 0/0 or inf-inf
    - **Division by Zero** division by zero
    - **Overflow** number exceeds maximum number storable for the float type
    - **Underflow** number is closer to zero than the smallest (absolute) number storable for the float type
    - **Inexact** result will not be within machine precision
- **Not A Number/NaN** special "placeholder" number generated by invalid operations, which will propagate to all other operations in which NaN is involved.
- **roundoff error** error due to inexact representation of real numbers as floats
- **relative error** magnitude of difference between real number and float approximation, normalised by the magnitude of the real number
- **endianness** the order in which multi-byte words are stored in memory
    - **little endian** least significant byte first
    - **big endian** most significant byte first
- **row vector** a vector stored as an Nx1 array
- **column vector** a vector stored as a 1xN array
- **singleton dimension** an array dimension of 1
- **promoting** increasing the rank of an array by introducing a new singleton dimension, like (3,3) -> (3,1,3)
- **squeezing** removing all singleton dimensions, like (1,3,1,4) -> (3,4)
- **Einstein summation notation** compact notation to exchange array dimensions, and also compute sums and products
- **First rule of vectorisation:** no `for` loops

week_3_visualisation.ipynb

- **histograms** a plot showing the relative frequencies of values falling into specified bins

- **scatterplots** a plot showing disconnected markers at x, y locations

- **contour plots** a plot showing the value of a 2D function $f(x, y)$ as a collection of isocontour lines

- **Layered Grammar of Graphics** A model for discussing, implementing and understanding scientific visualisations

- **Stat** a transformation of data, like taking the mean

- **Mapping** a mapping of data attributes onto visual attributes; e.g. price onto x location

- **Scale** a scaling of a data attribute onto the corresponding visual attribute

- **Guide** a visual element indicating the scaling and mapping applied, such that the transformation can be reversed by the reader to return to the data units from the visual units; for example, tick marks.

- **Geom** a geometric object used to visually represent data, like a line or point.

- **Coord** a special scaling that maps data onto 2D locations on the page/screen

- **Layer** visual representation of data with a single type of geom, on a common set of mappings, coords and scales. Multiple layers may be drawn on the same coords, forming a single facet.

- **Facet** Multiple facets represent multiple views of data, on separate coords.

- **Figure** A collection of facets.

- **Caption** Text which describes a figure to help the reader interpret it.

- **Axes** The visual manifestation of coords

- **units** The measurement quantities for a data attribute, like miles-per-hour

- **Ticks** visual guides to indicate equal-spaced steps in the original data units, visible only on outer axes.

- **legend** a guide used to identify distinct layers rendered on a single facet

- **title** a guide used to label a whole facet or figure

- **grid** visual guides to indicate equal-spaced steps in the original data units, visible across the whole facet.

- **annotations** textual labels and accompanying arrows etc. used to highlight important parts of a figure.

- **Markers** point geoms

- **Patches** polygon/area geoms

- **independent variable** a variable which *causes* some relationship. Speed of movement causes stopping distance to vary, for example.

- **dependent variable** a variable which is *caused* by some relationship. Stopping distance is dependent upon speed of movement, for example.

- **line geoms** A geom representing a connection between two data points. May vary in style, thickness, colour and opacity.

- **point geoms** A geom representing a single point in 2D space. May vary in shape, colour, size and opacity

- **ribbon plot** A plot using an area geom which behaves as a line of varying width

- **dimensional quantities** quantities with real-world units, like miles per hour

- **dimensionless** quantities without real-world units, like Mach number

- **aggregate summary statistics** statistics which summarise some larger dataset compactly, like the mean

- **smoothing and regression** finding simple, smooth functions which approximate observed data

- **binning** splitting data sets into discrete "bins" each of which spans a range of a data units; used to form histograms

- **rank plot** a plot of values arranged in rank order
- **Box plot** a plot used to summarise the distribution of a collection of values, showing the median, interquartile range, extrema and outliers
  - **interquartile range** the range between the 25% and 75% percentiles of a dataset. Drawn as a box on Box plot.
  - **median** the value which exactly splits a dataset into one half smaller and one half larger; the 50th percentile. Drawn as a line on Box plot.
  - **extrema** values representing the extreme values of a distribution, for example the 5th and 95th percentiles. Drawn as whiskers on a Box plot
  - **outliers** values outside of the whiskers. Drawn as "fliers", usually crosses or circles on each data point in a Box plot
- **violin plot** a plot similar to a Box plot which shows an approximate, smooth density instead of fixed nonparamateric statistics
- **linear regression** finding a line of best fit to observed data
- **colour map** a mapping from data units to colours
- **colour bar** a guide to indicate the interpretation of a colour map
- **monotonically varying brightness** apparent visual brightness which continuously increases as the data attribute increases
- **perceptually uniform** visual mapping where apparent perceptual change corresponds linearly and uniformly to data unit changes.
- **false contours** apparent contours in data caused by bad colour maps and not by underlying discrete steps in data
- **dash patterns** patterns applied to style line geoms
- **staircase / step** a type of plot where lines are only drawn horizontally and vertically to indicate instantaneous changes of value
- **bar chart** a plot which represents data units as the height of rectangular area geoms.
- **transparency / opacity / alpha** visual appearance of transparency to allow geoms to be partially visible behind other geoms
- **axis limits** range across entire axis, used to define the coord scaling
- **data units** the units of the original data (e.g. miles per hour)
- **visual units** the units displayed (e.g. inches, pixels or RGB values)
- **aspect ratio** the ratio of width of a plot to its height.
- **projection** the model used to map data units to visual units; for example Cartesian or polar
- **logarithmic axes** visual units represent the logarithm of data units
- **symlog** modified logarithmic scale to allow for negative values
- **polar coordinates** visual coordinates in terms of angle and distance
- **double y axes** two different y units overlaid on the same facet
- **standard deviation** the average deviation from the average
- **nonparametric intervals** intervals based on summary statisics of data, like the interquartile range

week_4_matrices_i.ipynb

- **embedding** The representation of objects as points in a vector space (e.g. a word embedding)
- **orthogonal direction** A direction at 90 degrees to (unrelated to) some other direction
- **vector space** A mathematical space which consists of tuples of numbers (usually reals) and two operations: vector addition and scalar multiplication
- **scalar multiplication** multiplying a vector by a scalar value (a single real number)
- **vector addition** adding two vectors, which is just the elementwise sum
- **norm** the length of a vector. Norms must be chosen and there are many choices of norm.
- **inner product** The dot product (sum of elementwise products) of two vectors
- **topological vector space** A vector space equipped with a norm
- **feature vectors** The representation of "features" which define properties as points in a vector space; critical in machine learning.
- **clustered** Partitioned into distinct groups according to spatial affinity
- **vector quantisation** Clustering collections of vectors and replacing the original vectors with cluster centres
- **weighted sums** a sum of values, where each value is scaled by some weight
- **linearly interpolate** find a line connecting two points

- **Euclidean norm** the "usual" spatial distance, or L_2 norm

- **Euclidean space** a vector space equipped with the Euclidean norm
- **direction / length** vectors can be considered to have a direction (unit length vector) and a length (a scaling factor applied to the direction)
- **cosine distance** measurement of angle between two vectors, in spaces where inner products are defined
- **unit vectors** a vector normalised to have unit length (norm=1) for some norm
- **document vectors** a way of representing whole text documents as single vectors
- **outer product** the matrix representing every possible product of the elements of two vectors
- **surface normal** the vector pointing "out" from a 3D surface
- **mean vector** the mean of a collection of vectors
- **geometric centroid** the centre of a collection of vectors, which is just the mean vector
- **zero mean** mean at the origin
- **variance** spread of values, defined as the square of the differences from the mean
- **standard deviation** square root of variance, useful as in the same units as the original values
- **correlations** linear relationships between variables
- **ellipse** a squashed circle
- **covariance matrix** a matrix representing the spread of vector valued data sets
- **error ellipse** an ellipse which can be computed from the covariance matrix and will (in general) capture some proportion of the data points.
- **geometric median** the vector which minimises the distance to all vectors in a data set

- **high dimensional vector spaces** a vector space with $D > 3$

- **curse of dimensionality** *very important* the problem that increasing dimensionality exponentially increases the volume of a vector space, making generalisation in high-dimensions very hard.
- **linear maps** a function which satisfies the conditions of linearity

- **linearity** the property that $f(a + b) = f(a) + f(b); f(ca) = cf(a); f(0) = 0$

- **linear functions** linear maps

- **parallelotope** the generalisation of a parallelogram to higher dimensions. A geometric shape with parallel sides, but not necessarily orthogonal sides.
- **linear transform** A linear map from a vector space onto itself, e.g. $\mathbb{R}^N \to \mathbb{R}^N$.
- **projection** A function which is idempotent, such that $f(f(x)) = f(x)$
- **real matrices** A matrix (2D array) consisting of real valued entries
- **linear algebra** The algebra of matrices
- **special forms** Special kinds of matrices, like diagonal, orthogonal or identity matrices
- **left multiplication** Multiplying a matrix A on the left of an expression AB
- **right multiplication** Multiplying a matrix A on the right of an expression BA
- **powers** (of a matrix) The effect of repeated multiplication $A^4 = AAAA$, for example.
- **diagonal** (of a matrix) The elements $a_i i$ of a matrix A; runs down a diagonal top-left to bottom right.
- **anti-diagonal** (of a matrix) The horizontally flipped diagonal.
- **real square matrices** A matrix that is real, and has equal number of rows and columns
- **upper triangular** A matrix with zeros below the diagonal
- **lower triangular** A matrix with zeros above the diagonal
- **symmetric** A matrix with elements mirrored around the diagonal
- **skew-symmetric matrix** A matrix with elements mirrored around the diagonal, but negated on one side
- **sparse** A matrix that mainly consists of zeros, with a few sparse nonzero elements

week_5_matrices_ii.ipynb

- **discrete-continuous interchange** representing apparently discrete problems as continuous problems, and vice versa; e.g. machine translation of strings (discrete) via vector spaces (continuous)
- **directed graph** a graph where edges are directional
- **vertices / nodes** elements on graph which are connected via edges
- **edges** connections between vertices
- **adjacency matrix** a NxN matrix representing a graph of N nodes, where a non-zero entry indicates a connection between two nodes
- **edge weighted graphs** a graph where each edge has a corresponding weight, usually a real value
- **source** a node on a grap that "produces" outgoing flow
- **sink** a node on a graph that "absorbs" incoming flow
- **stochastic matrix** a matrix whose columns (or rows) all sum to 1
- **Markov chains** a stochastic process where the distribution over the next state depends only on the previous state
- **decomposition** reduction to simpler forms (e.g. factorisation)
- **fixed point** a fixed point of a function is a point $f(x) = x$; the function does not alter the point
- **eigenvalues**
- **eigenvectors**
- **eigen**
- **eigenproblems** problems related to eigendecompositions
- **power iteration** a simple iterative algorithm for finding the leading eigenvector
- **leading eigenvector** the eigenvector associated with the largest eigenvalue
- **leading eigenvalue** the largest eigenvalue
- **deflation** the process of "removing" leading eigenvectors from a matrix to reveal the remaining eigenvectors
- **eigenspectrum** the ordered list of eigenvalues of a matrix
- **principal component analysis** the eigendecomposition of the covariance matrix, which indicates the "most important" linear components of a data set. Can be thought of a finding a rotation that would project the data losing the least amount of information
- **perimeter** the sum of the lengths of a bounding volume like a parallelotope
- **volume** the volume of space bounded by a bounding volume
- **positive definite** all eigenvalues are positive
- **positive semi-definite** all eigenvalues are zero or positive
- **negative (semi)-definite** all eigenvalues are (zero) or negative
- **inversion** The inverse of a matrix A is defined such that $A^{-1}A = I$. The inverse (if defined) "undoes" a linear map A.
- **singular** Not invertible (inverse does not exist)
- **non-singular** Invertible
- **singular value decomposition** A very powerful algorithm that factors any matrix A into $U\Sigma V$, where U and V are square orthonormal matrices (orthogonal matrices with each row and column of unit norm), and Σ is diagonal.
- **elimination algorithms** methods for solving linear systems by substitution
- **orthogonal matrix** a matrix where each column is an orthogonal vector. This means the transpose and the inverse are equivalent for orthogonal matrices

- **linear system / linear system of equations** An equation of the form $A\mathbf{x} = \mathbf{y}$ or

  $A\mathbf{x} - \mathbf{y} = 0$

- **left singular vectors** The rows of U

- **right singular vectors** The columns of V

- **singular values** The diagonal elements of Σ

- **pseudo-inverse / Moore-Penrose pseudo-inverse** The approximation to the inverse for non-square matrices. Can be used to find nearest solution to $A\mathbf{x} = \mathbf{y}$

  if A is non-square.

- **rank** Number of nonzero singular values of a matrix.
    - **full rank** Rank is equal to number of rows/columns. Only full rank matrices are invertible.
    - **deficient rank** Not full rank
- **condition number** Ratio of largest and smallest singular value. Critical in estimating numerical stability of algorithms like matrix inversion.
    - **well-conditioned** small condition number; numerical problems unlikely
    - **ill-conditioned** large condition number; numerical problems likely
- **whitening** Transformation of a vector data set so that it has zero mean and unit variance
    - **zero mean** a data set where the mean vector is the origin; can always be formed by subtracting the mean vector from any data set
    - **unit covariance** data set where the covariance matrix is the identity matrix ("spherical").
- **low-rank approximation** A simplified representation of a matrix with a small number of vectors
    - **low rank** Rank is (much) less than the number of rows/columns
- **dimensional reduction** A way of mapping high dimensional vector spaces to low dimensional vector spaces, like principal components analysis
- **matrix decompositions** algorithms that can break up matrices into simpler components, like eigendecompositions and the SVD.

week_6_optimisation_i.ipynb

- **optimisation** the process of making things better
- **parameters** values that can vary which affect an objective function
- **parameter space** the space of all parameter settings
- **parameter vector** all of the parameters of an optimisation problem collected into a single vector θ
- **the objective function** a function which maps parameters onto a measure of "goodness" or "badness", typically so that minimising this function will solve an optimisation problem
- **feasible region** / **feasible set** valid parameter settings (i.e. ones that don't violate constraints)
- **cost** / **loss** The value of the objective function for a given set of parameters (the "cost" of a given parameter setting).
- **approximation problems** finding a function which approximates an unknown function from a set of observations $\mathbf{x}$ and outputs $\mathbf{y}$
- **continuous optimization** optimisation with continuously variable parameters
- **discrete optimization** optimisation with parameters that take on discrete values
- **iterative** (optimisation) algorithms which successively approximate the minimisation of the objective functions
- **direct** (optimisation) algorithms which compute the minimum of an objective function in a single step
- **constrained optimisation** optimisation which respects some set of constraints that bound the possible parameter space
- **equality constraint** a constraint that forces some function of the parameters to be constant; typically this represents a tradeoff
- **inequality constraint** a constraint that forces the parameters to lie within some volume
- **box constraint** an inequality constraint that forces the parameters to lie within an axis-aligned box. Commonly supported by constrained optimisers.
- **orthant** the generalisation of a quadrant (2D) or octant (3D) to any dimension
- **convex constraint** an inequality constraint that forces the parameters to lie inside a convex region (i.e. one in which a line between any point on the boundary of the constraints will never cross the constraint volume)
- **unconstrained optimization** optimisation without any constraints at all
- **convex region** the region or volume defined by the convex volume
- **local minima** a parameter settings where any infinitesimal adjustment would make the objective function worse, but which may not be the best possible solution
- **global minimum** the best possible parameter setting, where the absolute smallest value of the objective function can be found
- **convex optimisation** techniques for solving optimisation problems where the objective function *and* the constraints are convex, which includes: linear programming, quadratic programming, semi-quadratic programming and quadratically constrained quadratic programming
- **continuous** a function where an infinitesimal change in the input results in an infinitesimal change in the output
- **differentiable** a function which has a derivative
- **perceptual model** an approximate model of human perception, often used as a proxy when optimising for displays for humans
- **root finding** finding vectors $\mathbf{x}$ where $f(\mathbf{x}) = 0$
- **inverse problem** problems which involve finding the inverse of a function $f^{-1}(x)$

- **linear least squares** an optimisation problem of the form
  $$L(\theta) = \|Ax - y\|_2^2, \; \theta = [a_{11}, a_{12}, \ldots]$$
- **termination criteria** the rule which defines when an iterative optimisation should stop
- **bisection** the process of binary search on a continuous space
- **grid search** searching a parameter space using a grid of evenly spaced points
  - **adaptive grid search** a modified grid search which recursively focuses on points that are likely to contain minima
  - **range** the span of the grid searched
  - **divisions** the number of subdivisions of the range
- **hyperparmeters** values which affect the *way* in which an optimisation algorithm works, but are not part of the objective function
- **random search** an optimisation technique which randomly samples from the parameter space and takes the best result
  - **bogosort** a random search algorithm for sorting which is unbelievably inefficient
- **meta-heuristics** a rule used to enhance a search algorithm, which may improve performance but is not guaranteed to do so
  - **locality** a heuristic that takes advantage of continuous objective functions and assumes that nearby parameters have similar losses
  - **temperature** a heuristic that will take occasionally steps that make the current state worse, but gradually reduce this chance over time
  - **population** a heuristic that maintains a collection of parameter vectors and uses this set to perform the next iteration
  - **memory** a heuristic that remembers the path of parameter vectors explored and uses this to avoid bad previous solutions, or return to good ones
- **local search** optimisation which makes small steps in the neighbourhood of a solution
  - **initial conditions** the parameter setting at which local search begins
  - **trajectory** the history of parameter vectors explored during local search
- **hill climbing** a local search algorithm that searches the neighbourhood and accepts steps that improve the solution
  - **neighbourhood** the local area searched
  - **stochastic hill climbing** hill climbing which uses random search to explore the local neighbourhood
- **multiple restarts** a metaheuristic which repeatedly restarts local search with different (random) initial conditions to mitigate local minima
- **simulated annealing** an algorithm which modifies hill climbing to include a temperature component, allowing occasional "bad" steps
- **temperature schedule** a function which defines how the probability of accepting "bad" steps changes over iterations
- **acceptance probability** the probability of accepting a new solution in simulated annealing
- **genetic algorithms** optimisation algorithms which use a population heuristic and maintain a collection of possible solutions at each iteration
  - **evolution** random search with a "survial of the fittest" rule
  - **mutation** random modification of a parameter vector
  - **natural selection** "survial of the fittest"
  - **breeding / crossover** heuristics which combine elements of multiple solution into a new "offspring"
- **memoryless** optimisation which stores nothing about the solutions it has explored
- **memory** optimisation which does store information about solution explored
- **Tabu search** memory based optimisation which remembers bad solutions (or bad transitions from solutions) and avoids them

- **Ant colony optimisation** heuristics which combine population and memory to find "trails" between potential good solutions and explore them efficiently
- **stigmergy** the process of exchanging information with the environment (like ants' pheromone trails) rather than in the mind. This reduces mental requirements

week_7_optimisation_ii.ipynb

- **weight matrix** the parameterisation of one "layer" of a neural network, which is just an ordinary matrix
- **derivative** the rate of change of a function with respect to an input
- **automatic differentiation** computing derivatives of functions algorithmically (and exactly) without specifying them directly
- **first-order algorithms** optimisation which requires the first derivative of the objective function
- **gradient vector** vector of partial derivatives with respect to each element of the parameter vector. Points in the direction of steepest descent.
- **zeroth order** optimisation which requires no derivatives
- **gradient descent** first-order optimisation algorithm which makes steps along the gradient vector
- **second order** optimisation which requires first and second derivatives
- **slope** gradient
- **analytical derivatives** derivatives in closed form
- **turning points** points at which the derivative of a function is zero
- **optima** turning points which are maxima or minima
- **exact pointwise derivatives** derivatives which can be computed exactly at any given point, but not in closed form
- C^n **continuous** a function which has at least n continuous derivatives
- **step size** the scaling factor of the gradient vector in gradient descent
- **finite differences** a way of approximating the derivative by taking differences of the objective function at closely spaced points
- **momentum** an acceleration heuristic for gradient descent that uses memory of the direction that has been successful to overcome saddle points and plateaus
- **stochastic relaxation** averaging over random instances to make apparently steep or discontinuous function approximately smooth
- **stochastic gradient descent** efficient gradient descent algorithm which uses stochastic relaxation
- **differentiable programming** programming where automatic differentiation is part of the basic language/library
- **computational graph** representation of programming language expressions in terms of a graph relating values (vertices) and operations (edges)
- **minibatch** a subsample of the data for which the average gradient is computed in stochastic gradient descent for approximation problems
- **epoch** a complete run through the entire dataset, such that every data point has been part of a minibatch
- **heuristic search** optimisation algorithms based on rules-of-thumb (heuristics) that are not guaranteed to converge
- **line search** a method for automatically tuning step size in gradient descent
- **attractor basin** The area around a local minimum where gradient descent will "fall into" the minimum
- **critical points** a point on the objective function where the gradient is zero
- **Hessian / Hessian matrix**
- **Second order methods** optimisation which requires the use of the second derivative of the objective function w.r.t the parameters; i.e. uses the Hessian matrix or an approximation to it.
- **type of critical point**
  - **local minima** a critical point where a step in any direction will increase the objective function
  - **local maxima** a critical point where a step in any direction will decrease the objective function

- **plateaus** a critical point where the objective function is completely flat in all directions
- **saddle points** a critical point where the objective function increases in at least one direction and decrease in at least one other direction
- **ridge** a critical point where the objective function is completely flat in at least one but not all dimensions.
- **sub-objective functions** objective functions which contribute to a an overall compromise objective function
- **convex linear sum** a weighted sum of elements where all weights are positive
- **multiplicative interaction** a nonlinear interaction between elements (sub-objective function) where one sub-objective function has a multiplicative effect on another
- **dominance** a solution θ_1 dominates a solution θ_2 if it is better or as good as θ_2 as measured by every sub-objective function
- **strictly dominates** as dominates, but strictly better then
- **Pareto optimal** A Pareto optimal solution is a configuraion θ where any change to θ that would improve a sub-objective function would make at least one other sub-objective function worse.
- **Pareto front** The set of Parteo optimal solutions for a multi-objective optimisation problem.

Week_8_probability_I.ipynb

- **experiment** (or **trial**) An occurrence with an uncertain outcome. For example, losing a submarine -- the location of the submarine is now unknown.
- **outcome** The result of an experiment; one particular state of the world. For example: the submarine is in ocean grid square [2,3].
- **sample Space** The set of *all possible* outcomes for the experiment. For example, ocean grid squares {[0,0], [0,1], [0,2], [0,3], ..., [8,7], [8,8], [8,9]}.
- **event** A *subset* of possible outcomes with some common property. For example, the grid squares which are south of the equator.
- **probability** The probability of an event *with respect to a sample space* is the number of outcomes from the sample space that are in the event, divided by the total number of outcomes in the sample space. Since it is a ratio, probability will always be a number between 0 (representing an impossible event) and 1 (representing a certain event). For example, the probability of the submarine being below the equator.
- **frequency** a number describing how often an outcome occurs in a sequence of experiments; which could either be a count of occurrences (the last 10 times, we found submarines below the equator after losing them) or a proportion of the total experiments (the 0.512 of the time, lost submarines are found below the equator).
- **distribution** A mapping from outcome to frequency for each outcome in a sample space.
- **probability distribution** A distribution that has been *normalized* so that the sum of the frequencies is 1. This is because an outcome must happen (with probability 1) so the sum of all possible outcomes together will be 1.
- **calculus of belief** a set of rules for updating beliefs
- **Bayesians** those who use probability as a calculus of belief (incorporating priors)
- **frequentist** those who use probability as the long-term average of an infinite series of experiments
- **subjective probability** probability which relies of priors
- **objective probability** probability which does involve priors
- **degree of belief** how much you believe in something; Bayesians use probability as a degree of belief.
- **urn problem** a problem related to balls being distributed among and drawn from urns. Often used as a simple model of a random system.
- **forward probability** probabilities related to observations ("What is the probability that the next ball that is drawn will be white?")
- **inverse probability** probabilities related to unseen processes generating observations ("What is the distribution of white and black balls in the urn?")
- **boundedness** all probabilities are in the range [0,1]
- **unitarity** for a given experiment, the probability of every outcome must sum to 1
- **sum rule** the probability of either of two events is the sum of the probabilities of both minus the probability of both at the same time.
- **conditional probability** the probability of an event *given that we know* that another event has occured
- **random variable (rv)** a variable whose value is unknown (random) but whose distribution is known
- **discrete random variables** random variable taking on a discrete set of possible states
- **probability mass function (PMF)** a function mapping each outcome of a discrete random variable to its probability
- **continuous random variables** a random variable which can take on an uncountably infinite number of states

- **probability density function (PDF)** a function which describes the probability distribution of a continuous random variable
- **samples / observations** known, observed values drawn from a population
- **uniformly distributed** distribution such that every value in a fixed range (e.g. [0,1]) has equal probability
- **pseudo-random numbers** numbers generated algorithmically which are statistically similar to true random numbers
- **empirical distribution** the distribution of a discrete random variable estimated from counts of observations
- **joint probability** the probability distribution of two or more random variables

  $P(X, Y)$

- **marginal probability** the probability distribution of one random variable, computed from the joint
- **marginalisation** computing marginal distributions by integrating/summing joint probabilities
- **independent** not dependent. Independent random variables have

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

- **conditional probability** $P(X|Y)$; the probability of $X = x$ *given that we know*

  $Y = y$

- **bigram** a pair of symbols, e.g. a pair of characters or words
- **n-gram** generalisation of bigram to n-tuples of symbols
- **entropy** a measure of surprise, which applies to random variables. The entropy indicates how surprising the outcome of an experiment would be.
- **information theory** The theory of communication defined in terms of entropy.
- **Bayes' rule** The rule that defines how likelihood, prior and evidence can be combined to estimate the posterior. Used to invert conditional probability

  distributions $P(B|A)$ to get $P(A|B)$

- **posterior** the distribution obtained as a result of applying Bayes Rule
- **likelihood** how likely the observed data is given a particular model
- **prior** how likely a particular model is, before we observe data
- **evidence** how likely the data is; equal to the sum of the product of the likelihood and prior for all possible models
- **integration over the evidence** summing over possible models to compute the

  evidence $P(B)$ in $P(B|A)P(A)/P(B)$

- **hypothesis** a possible model of the world
- **data** observed samples
- **false positive rate** the rate at which a test is positive if the condition is not present
- **false negative rate** the rate at which a test is negative if the condition is present
- **natural frequency** a way of visualising probability problems by imagining objects (typically people in a room), that helps avoid logical fallacies about probability
- **odds** a way of expressing probabilities that is easy for humans to understand.

  $p = 0.01$ is the same as 99:1

- **log-odds / logit** the logarithm of (the reciprocal of) the odds. Useful for very rare events.
- **log-probabilities** logarithm of probabilities. Widely used in computations to avoid floating point underflow
- **log-likelihood** logarithm of the likelihood.
- **cumulative distribution function** The integral of a probability density function up to some point. Often easier to work with than the PDF itself.

- **Gaussian / normal distribution** A very important distribution which has many nice mathematical properties. Widely used as an approximation.
- **central limit theorem** a result which indicates that *any* set of observations which is really the sum of many independent random factors will tend to a normal distribution.
- **properties of distributions**
  - **location** the "shifting" of a distribution; the **location** of the normal distribution is the mean μ
  - **scale** the "spread" of a distribution; the **spread** of the normal distribution is σ
  - **support** the values for which the distribution has non-zero probability
  - **compact support** a distribution which has a fixed range of possible values, like the uniform distribution over the range $[a, b]$; values less than a or more than b are impossible (probability zero)
  - **infinite support** a distribution that has an infinite range of possible values. The normal distribution has infinite support
- **statistics** functions of data that summarise the data in some way, like the mean
- **estimators** statistics which estimate some unknown parameter of a generating process. The sample mean is an estimator of the parameter μ of a normal distribution
- **maximum likelihood estimation** the process of finding a set of parameters that "best explains" observations by optimisation
- **multivariate distributions** probability distributions over multiple dimensions, like $\mathbb{R}^N$
- **univariate** probability distributions over a single dimension, like $\mathbb{R}$
- **multivariate normal** the normal distribution generalised to $\mathbb{R}^N$. It is parameterised by:
  - **mean vector** μ, which specifies a point where the distribution is centered
  - **covariance matrix** Σ which specifies how the distribution is shaped, effectively as an ellipsoidal shape
- **joint probability density** the probability density function for a multivariate distribution, where each dimension can be seen as a different random variable
- **marginal probability density** the probability density function for a specific variable in a a multivariate distribution, which can be computed by integrating over all of the other dimensions
- **recursive Bayesian estimation** the process of updating beliefs over time, using the posterior from one step as the prior for the next.

lab_8.ipynb

- (stochastic) **simulation** simulating the behaviour of a stochastic process using known parameters and random number generation
- (stochastic) **estimation** inferring the parameters of a stochastic process from a sequence of observations; e.g. the biasedness of a coin from a series of observed tosses
- **random process / stochastic process** a process (that is, something that unfolds over time) which has an element of randomness. A random process depends on previous states
- **Markov process** a stochastic process whose behaviour depends on the *immediately* previous state, and no other history.
- **Markov property** The property of being independent of all previous states, except for the immediately prior state
- **transition distribution** The distribution that specifies how a random process will be distributed *given a previous known state*
- **discrete states** states with a fixed number of options, like a coin toss can have heads or tails
- **continuous states** states with a continuum of values, like the value of the stock market at a given time
- **Markov chains** a Markov process
- **Gilbert model** a binary Markov chain which simulates "sticky" processes like flickering candles or network packet traffic
- **Brownian motion** a continuous random process where the next state is a small random perturbation, normally distributed around the current state
- **random walk** a Markov process; so called because the result is slow "wandering" about some space
- **t-distributed** distributed according to the t-distribution, which has a much higher chance of generating extreme values than the normal distribution
- **maximise likelihood** to find parameters that maximise the likelihood function for some fixed, known observations

week_9_probability_ii.ipynb

- **expectation / expected value** the "average" expected outcome of a random variable with a numerical value
- **population mean** the mean of the whole (usually unobserved, and often infinite) population
- **central tendency**
- **rational decisions**
- **decision theory** the theory of making rational decisions
- **utility** the value of an outcome (e.g. the utility of winning a betting game would be the monetary prize)
- **summary statistics** summarising observations or samples
- **inferential statistics** inferring unseen populations from samples
- **population** the complete set of unseen and potentially infinite possibilities from which samples are drawn (e.g. the height of all adult men)
- **(population) parameter** a parameter which defines the process by which it is assumed that the population is being generated (e.g. the standard deviation of adult men)
- **sample** an observation drawn from a population; for example, the height of one specific man measured
- **statistic** a function of observations or samples which summarises them
- **infer** to determine an unobserved state from an observed one
- **sample mean** the mean of definite, known samples; a function of data
- **descriptive statistics** statistics which describe data, but are not directly used to infer population parameters
- **robust statistics** statistics which are insensitive to extreme values or outliers. The median is robust; the mean is not
- **interquartile range** range between the 25th and 75th percentile
- **Monte Carlo** a broad and important class of algorithms which use random simulation to approximate problems
- **bootstrap** a simple resampling algorithm which can provide estimates of the certainty of sample statistics like the mean by repeatedly randomly resampling a set of observations with replacement and computing a statistic for each one
- **generative models** a statistical model which describes how samples are being generated (by an unseen process) rather than simply describing the samples
- **probabilistic programming** a programming language library that supports random variables as first-class values
- **Markov Chain Monte Carlo** a class of algorithms for performing (relatively) efficient inference by creating a stochastic process (Markov Chain) which will approximately sample from hard to sample from distribution. This is typically used to draw samples from an intractable posterior in Bayesian inference.
- **probabilistic graphical model** a formalism for representing the relation between variables in the description of a generative model. It has nodes of various types
  - **Stochastic** stochastic nodes represent unknowns, whose value is not known but for which a prior distribution is available
    - **observed** an observed stochastic node has known, observed samples associated with it. Its distribution can be updated from samples.
    - **unobserved** an unobserved stochastic node does not have any known observed samples. Its distribution must be inferred from its relation to other nodes
  - **deterministic** a deterministic node is known if its parents (the values it depends on) are known.

- **linear regression** the estimation of a best-fitting line (or generally a hyperplane) from a set of observed $\mathbf{x}$ and $\mathbf{y}$

- **latent structure** assumed structure which is a simple explanation of the observations made

- **maximum likelihood linear regression** linear regression by maximising the likelihood (minimising log-likelihood) of the observations given a set of parameters. Equivalent to ordinary least squares.

- **proposal distribution** a distribution used in MCMC to propose possible new steps from the current state. Typically a simple distribution like a normal distribution.

- **acceptance probability** The probability of accepting a new step. This is designed such that the samples accepted will (asymptotically) be samples from the distribution the MCMC process is trying to sample from.

- **trace** the sequence of samples draw in an MCMC process. The trace is the collection of posterior samples.

- **posterior** estimated distribution of the model parameters after observing data (via Bayesian inference)

- **predictive posterior** the estimated distribution over observations given a

week_10_signals.ipynb

- **sample** a measurement of a continuous signal at a precise instant or point
- **quantisation** the reduction of a continuous signal to discrete steps
- **sampled signal** the discrete representation of a continuous signal as a set of evenly spaced samples
- **sampling rate** the rate at which samples are taken; $1/\Delta T$
- **digital signal processing** the name for techniques applied to process sampled signals computationally
- **Nyquist limit** the highest frequency representable for a given sample rate; equal to half the sample rate
- **aliasing** the distortion introduced when a signal with frequencies greater than the Nyquist rate are present when a signal is sampled
- **wagon wheel effect** temporal aliasing in the video domain
- **interpolation** estimating the value of a function in between known samples
- **gridding** the combination of interpolation with regular sampling to reinterpolate an irregularly sampled signal into a regularly sampled signal
- **interpolation function** a function used to interpolate between measurements
    - **constant / nearest-neighbour** assumes function constant between samples
    - **linear** assumes a linear slope between samples
    - **polynomial** fits a polynomial (e.g. cubic) to successive groups of samples
- **piecewise interpolation** interpolation applied separately to different regions of measurements (e.g. every pair of samples), rather than regression over the entire sequecne of measurements
- **amplitude quantisation** reduction of amplitude to a fixed number of distinct levels
- **noise** components of a signal unrelated to the signal of interest
- **signal to noise ratio** the ratio of the signal of interest to noise
- **decibels** a logarithmic unit used for ratios, equal to $10\log(S/N)$ for signal to noise ratios.
- **moving average** a smoothing operation which applies the mean statistics to a sliding window
- **sliding window** an operation which takes successive "windows" of fixed length groups of samples from a sampled signal. Critical in implementing DSP algorithms, as it turns unbounded length signals into a collection of fixed length vectors.
- **feedback filtering** filtering which uses the previous output of the filtering operation as one of the inputs at the next step
- **exponential smooth** a very simple but effective linear feedback filter, with one step history
- **linear filters** filters which are just linear sums (weighted sums) of neighbouring samples
- **nonlinear filter** any filter which is not a linear filter
- **median filter** a nonlinear filter which applies the median filter to a sliding window
- **Order filters** a nonlinear filter which applies any order statistic (median, max, min, percentile) to a sliding window
- **multiply and accumulate** a type of hardware instruction that can efficiently accumulate a weighted sum
- **convolution** the key operation in signal processing, which computes weighted sums of a sliding window
- **convolution kernel** the array of values which define the convolution apply
- **algebraic properties of convolution** key properties of convolution which make it efficient and useful:
    - **commutes** $A * B = B * A$

- **associates** $A * (B * C) = (A * B) * C$

- **distributes** $A * (B + C) = A * B + A * C$

- **delta function** a (pseudo) function which is zero everywhere, except for a value of 1 at the origin
- **impulse** a delta function
- **impulse response recovery** a way of extracting the convolution kernel from a system by feeding an impulse in and measuring the response
- **reverberation** the acoustic effect of multiple reflections, as in the characteristic sound of a room
- **the time domain** the representation of signals as amplitudes varying over time (or over space, in an imaging context)
- **the frequency domain** the representation of signals as the superposition of oscillations of different frequencies

- **sine wave** a pure oscillation defined by $A\sin(wt + \theta)$

  - **amplitude** A, the intensity or strength of the oscillation

  - **phase** θ, a shift of the oscillation in time

  - **frequency** ω, the rate of oscillation

- **Fourier theorem** every periodic signal can be represented precisely as a sum of sine waves
- **correlation** the (linear) similarity between two signals
- **magnitude** amplitude
- **sinusoids** sine waves
- **real signals** signals without imaginary values
- **Fourier transform** function which can transform the time domain (measurements over time) to the frequency domain (sum of oscillations)
- **inverse Fourier transform** transforms frequency domain (exactly) back to the time domain
- **discrete Fourier transform (DFT)** Fourier transform for sampled signals
- **complex numbers** numbers with a real and imaginary component
- **Argand diagram** a way of drawing complex numbers as 2D points on the plane. This can be used to see a **polar** representation of complex numbers, in terms of phase and magnitude.
  - **phase** the angle of a complex number to the x-axis
  - **magnitude** the distance of a complex number from the origin
- **fast Fourier transform (FFT)** a very efficient algorithm for computing the DFT in $O(NlogN)$ time (in some cases), instead of $O(N^2)$ time

- **convolution theorem** convolution in the time domain is identical to elementwise multiplication in the frequency domain, and vice versa
- **smoothing filter** a filter which reduces high frequency components ("smooths them out")
- **lowpass filter** a filter which reduces high frequency components, like noise
- **highpass filter** a filter which reduces low frequency components, like slow trends
- **bandpass filter** a filter which selects a band of frequencies of interest (like tuning into a radio)
- **notch filter / bandstop filter** a filter which removes a band of frequencies, e.g. removing 50Hz mains hum
- **filter design** the process of designing a filter in the time domain given frequency domain specifications