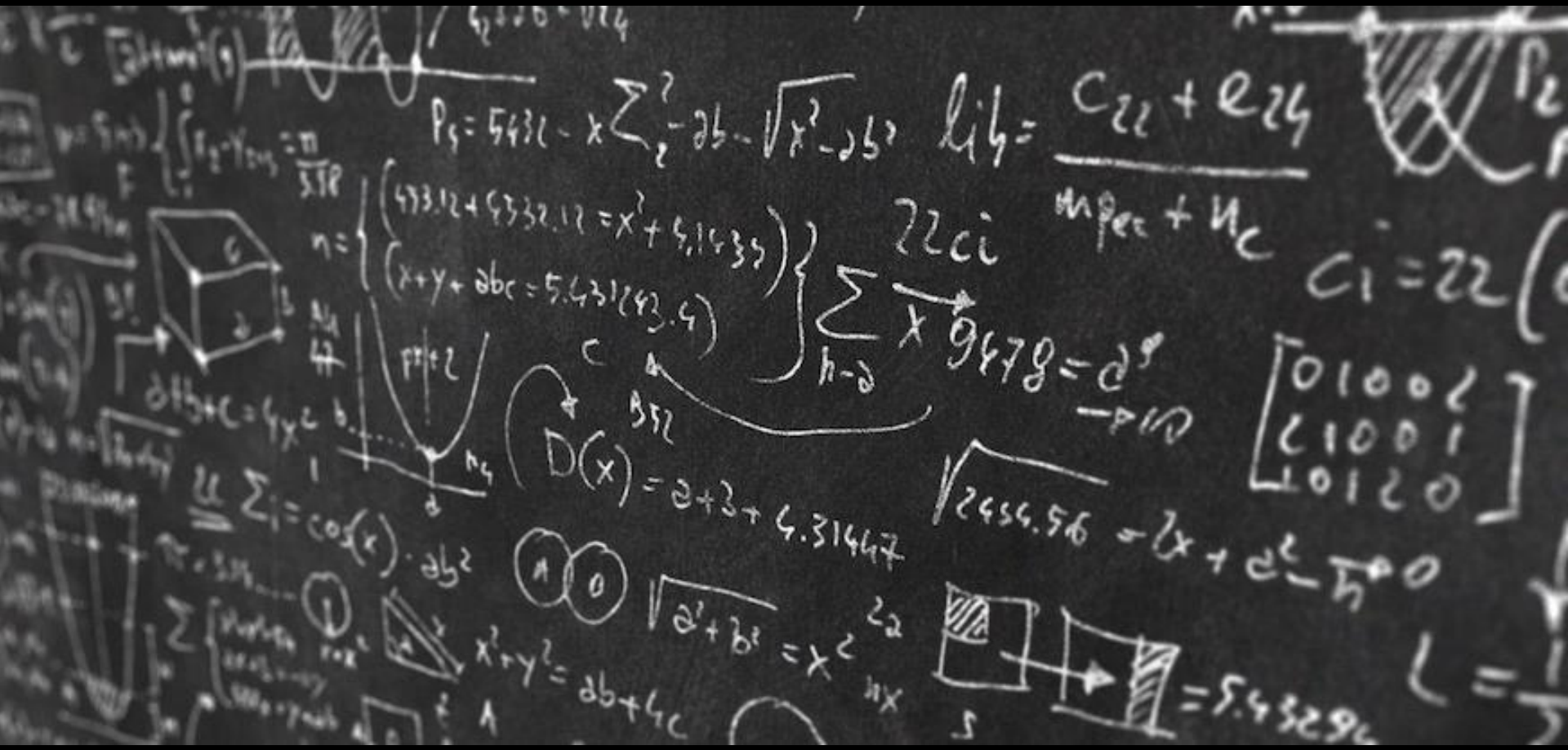


Mathematics Preliminaries for Machine Learning
CS 445/545

Outline

- Linear Algebra Overview
- Statistics/Probability Overview

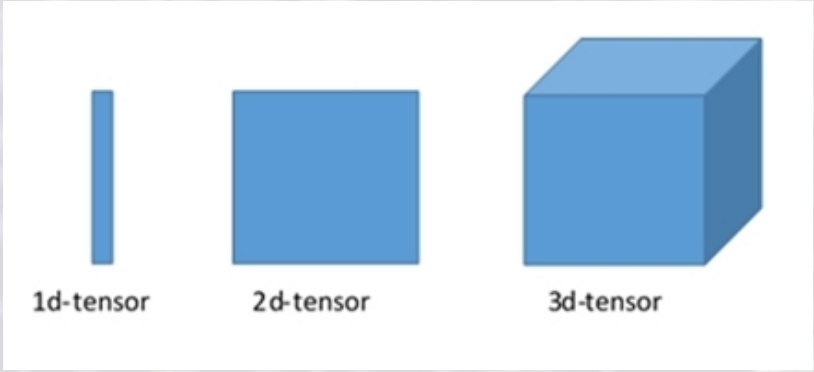
Linear Algebra: Vectors & Matrices

$$\mathbf{x} \in \mathbb{R}^d \rightarrow \mathbf{x} = \langle x_1, \ldots, x_d \rangle$$

vector

$$\mathbf{A} \in \mathbb{R}^{m \times n} \rightarrow \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix}$$

matrix



$$\mathbf{M} \in \mathbb{R}^{m \times n \times p} \rightarrow \begin{bmatrix} a^1_{11} & \cdots & a^1_{1n} \\ \vdots & \ddots & \vdots \\ a^1_{m1} & \cdots & a^1_{mn} \end{bmatrix} \cdots \begin{bmatrix} a^p_{11} & \cdots & a^p_{1n} \\ \vdots & \ddots & \vdots \\ a^p_{m1} & \cdots & a^p_{mn} \end{bmatrix}$$

tensor

Vector & Matrix Operations

dot product: $\mathbf{x} \cdot \mathbf{y} = \sum_{i=1}^{d} x_i y_i \in \mathbb{R}$ results in a scalar; **x**, **y** are **orthogonal** if $\mathbf{x} \cdot \mathbf{y} = 0$

matrix multiplication : $\underset{m \times n}{A} \cdot \underset{n \times p}{B} = \underset{m \times p}{C}$

$$A \cdot B = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \text{row i} & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix} \cdot \begin{bmatrix} b_{11} & \cdots & a_{1p} \\ \vdots & \text{col j} & \vdots \\ b_{n1} & \cdots & a_{np} \end{bmatrix} \rightarrow C_{ij} = \mathbf{a}_{rowi} \cdot \mathbf{b}_{colj}$$

• Matrix multiplication involves a sequence of dot products; element C_{ij} in the resultant matrix is equal to the dot product of *row i* (from the left matrix) and *column j* (from the right matrix).

Vector & Matrix Operations

scalar multiplication: $c\mathbf{x} = c\langle x_1, ..., x_d \rangle = \langle cx_1, ..., cx_d \rangle$

scalar multiplication: $cA = c\begin{bmatrix} a_{11} & \dots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \dots & a_{mn} \end{bmatrix} = \begin{bmatrix} ca_{11} & \dots & ca_{1n} \\ \vdots & \ddots & \vdots \\ ca_{m1} & \dots & ca_{mn} \end{bmatrix}$

• Matrix multiplication is associative: A(BC)= (AB)C (always holds), but not commutative: AB ≠ BA (in general).

vector transpose: $\underset{1\times d}{\mathbf{x}} = [x_1, ..., x_d] \rightarrow \underset{d\times 1}{\mathbf{x}^T} = \begin{bmatrix} x_1 \\ \vdots \\ x_d \end{bmatrix}$

matrix transpose: $A = \begin{bmatrix} a_{11} & \dots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \dots & a_{mn} \end{bmatrix} \rightarrow A^T = \begin{bmatrix} a_{11} & \dots & a_{1m} \\ \vdots & \ddots & \vdots \\ a_{n1} & \dots & a_{nm} \end{bmatrix}$

• A matrix is called "symmetric" if: $A^T = A$.

Vector Norms

• Norms convey the notion of the "magnitude" (i.e. size) of a vector; note that the equivalence (in magnitude) of two vectors is relative to the choice of norm.

• There are many types – even "families" – of norms relevant to ML/data science. Here are several of the most commonly used norms in ML:

(1) **L2 norm** (i.e. "Euclidean norm")

$$\|\mathbf{x}\|_2 = \sqrt{\sum_{i=1}^{d} x_i^2} \qquad \text{Note: } \|\mathbf{x}\|_2^2 = \sum_{i=1}^{d} x_i^2 = \mathbf{x}\cdot\mathbf{x}$$

(2) **L1 norm** (i.e. Manhattan distance)

$$\|\mathbf{x}\|_1 = \sum_{i=1}^{d} |x_i|$$

(3) **∞ norm**

$$\|\mathbf{x}\|_\infty = \max |x_i|$$

*For an ML practitioner, the "choice" of a norm is oftentimes a crucial part of feature engineering and the ML problem formulation process itself; one can think of the different norm choices as striking a balance between "precision" and computational overhead.

*There exist equivalent norms applied to matrices; the above norms are examples from the family of *p-norms*.

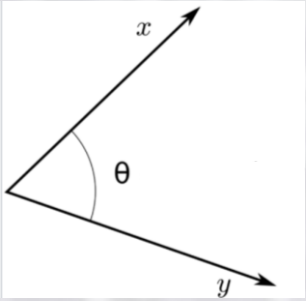
$$\mathbf{x} \cdot \mathbf{y} = \sum_{i=1}^{d} x_i y_i$$

Dot Product

- The dot product has important geometric properties that are useful in ML.

(*) The dot product can be defined equivalently:

$$\mathbf{x} \cdot \mathbf{y} = \|\mathbf{x}\|_2 \|\mathbf{y}\|_2 \cos(\theta)$$



From this equivalent definition of the dot product, we can show that the dot product quantifies the "similarity" between two vectors. Consider (3) cases:

(i) Vectors **x** and **y** are "out of alignment" and meet at a 90 degree angle; in this case:

$$\mathbf{x} \cdot \mathbf{y} = \|\mathbf{x}\|_2 \|\mathbf{y}\|_2 \cos(90°) = 0$$

(ii) Vectors **x** and **y** are "perfectly aligned" (i.e. *parallel* to one another):

$$\mathbf{x} \cdot \mathbf{y} = \|\mathbf{x}\|_2 \|\mathbf{y}\|_2 \cos(0°) = \|\mathbf{x}\|_2 \|\mathbf{y}\|_2$$

(iii) Vectors **x** and **y** are "oppositely aligned (i.e. they are *anti-parallel*):

$$\mathbf{x} \cdot \mathbf{y} = \|\mathbf{x}\|_2 \|\mathbf{y}\|_2 \cos(180°) = -\|\mathbf{x}\|_2 \|\mathbf{y}\|_2$$

Special Matrices

• The **identity matrix** ***I*** is a square (*n*x*n* matrix); the identity matrix multiplied by any matrix *A* (appropriately shaped) results in the matrix *A:*

$$I_n = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & \cdots & 1 \end{bmatrix}_{n \times n} \qquad AI = IA = A$$

• A matrix *A* is said to be **symmetric** if it equals its transpose:

$$A^T = A \qquad e.g., \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix}^T = \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix}$$

$$\text{Note: } (A+B)^T = A^T + B^T, (AB)^T = B^T A^T$$

• For a **diagonal matrix**, all off-diagonal entries are zero (note that diagonal entries are permitted to be zero).

$$D = \begin{bmatrix} d_1 & 0 & \cdots & 0 \\ 0 & d_2 & \cdots & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & \cdots & d_m \end{bmatrix}$$

Special Matrices

• An **upper-triangular matrix** has zero elements *below* the main diagonal. Note that Gaussian Elimination (from elementary linear algebra) yields an upper-triangular matrix.

$$U = \begin{bmatrix} u_{11} & u_{12} & \cdots & u_{1n} \\ 0 & u_{22} & \cdots & u_{2n} \\ 0 & 0 & \ddots & \vdots \\ 0 & 0 & \cdots & u_{mn} \end{bmatrix}$$

• A **lower-triangular matrix** has zero elements *above* the main diagonal.

$$L = \begin{bmatrix} l_{11} & 0 & \cdots & 0 \\ l_{21} & l_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ l_{m1} & l_{m2} & \cdots & l_{mn} \end{bmatrix}$$

• An **orthogonal matrix** is a matrix with *orthonormal* rows and columns; equivalently, the inverse of an orthogonal matrix is its transpose.

$$Q^T Q = Q Q^T = I$$

Special Matrices

• A square matrix A is **Positive Definite** if:

$$\forall \mathbf{x} \in \mathbb{R}^n \setminus \{\mathbf{0}\}, \mathbf{x}^T A\mathbf{x} > 0$$

• Analogously, a square matrix A is **Positive semi-Definite** if (e.g. covariance matrix):

$$\forall \mathbf{x} \in \mathbb{R}^n, \mathbf{x}^T A\mathbf{x} \geq 0$$

• We say that the matrix A_{mxn} is **invertible** (i.e. *non-singular*) if there exists $A^{-1}{}_{nxm}$, where:

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

Properties:

$$\left(A^T\right)^{-1} = A^{-T} \qquad \left(AB\right)^{-1} = B^{-1}A^{-1} \text{ (if } A,B \text{ non-singular)}$$

Linear Systems

• Commonly we encode, and subsequently solve systems of linear equations :

$$\begin{aligned} a_{11}x_1 + \cdots a_{1n}x_n &= b_1 \\ a_{21}x_1 + \cdots a_{2n}x_n &= b_2 \\ &\vdots \\ a_{m1}x_1 + \cdots a_{mn}x_n &= b_m \end{aligned} \quad \rightarrow \quad \underbrace{A\mathbf{x} = \mathbf{b}}_{\text{Canonical Matrix Form}}$$

• When the *coefficient matrix A* is non-singular, the linear system gives rise to a unique solution:

$$A\mathbf{x} = \mathbf{b} \rightarrow \mathbf{x} = A^{-1}\mathbf{b}$$

* Note that matrix inversion requires roughly on the order of $O(n^3)$ arithmetic operations.

Matrix Factorizations

• Matrix factorizations are immensely useful for identifying an underlying, inherent structure in a matrix (i.e. data).

Here are several important examples:

LU Factorization $A = LU$

• This factorization encodes the result of the Gaussian Elimination (GE) procedure (note that not all matrices admit of an LU factorization). L: denotes a lower-triangular matrix of "multipliers" used in GE. U denotes an upper-triangular (i.e. echelon form) matrix resulting from GE.

PALU Factorization (Permuted LU factorization) $PA = LU$

• This technique is similar to LU Factorization, except that we perform a **pivoting operation first** (i.e. permute the rows of A via a *permutation matrix*, P). LU factorization is subsequently performed; all matrices admit of such a factorization.

Matrix Factorizations

Here are several important examples:

QR Factorization $A = QR$

• Q is an orthogonal matrix and R is upper-triangular -- commonly used for solving both regression problems and linear dynamical systems.

Eigendecomposition $A = V\Sigma V^T$

• This is one of the most useful and commonly-used of all matrix factorizations. The primary use of an eigendecomposition in ML is to perform *dimensionality reduction*; as such, this technique is closely related to **PCA** (*principal component analysis*) and **SVD** (*singular value decomposition* – see below) methods; Σ is a diagonal matrix consisting of the eigenvalues of A, and V is the matrix of corresponding eigenvectors.

Matrix Factorizations

Here are several important examples:

Cholesky Factorization $A = LL^T$

• *L* is lower-triangular; Cholesky can be used to numerically solve linear systems; every positive-definite matrix admits of a Cholesky factorization.

SVD (Singular Value Decomposition) $A = U\Sigma V^T$

• SVD is one of the most essential matrix factorizations for applications of ML. U and V are orthogonal matrices, and Σ is a diagonal matrix containing the "singular values" (i.e. the eigenvalues of A^TA. SVD has many applications (an orthogonal matrix denotes the matrix of eigenvectors of A^TA, including dimensionality reduction and compression. All matrices admit of a singular value decomposition.

Determinants

• Geometrically, the determinant of a square matrix A (written $|A|$) quantifies the unit increase in volume of the *linear transformation* defined by A (note that matrix multiplication defines a linear transformation).

Determinants can be computed through recursion; the general formula for a determinant is:

$$|A| = \sum_{i=1}^{n} a_{ij} (-1)^{i+j} \cdot M_{ij}$$

M_{ij}: the "ij-minor" of A

Note that $|A| = 0$ if and only if A is singular.

Some Properties of Determinants:

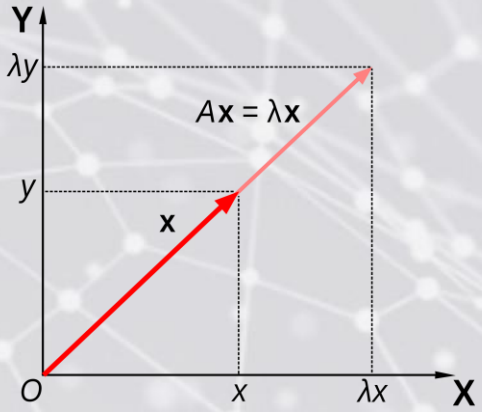
$$|AB| = |A||B| \qquad |A^T| = |A|$$

Eigenvalues

• The eigenvalues λ and eigenvectors **v** of a matrix A satisfy:

$$A\mathbf{v} = \lambda\mathbf{v} \qquad (v \neq 0)$$

• Which means that the eigenvectors of a matrix A are precisely the vectors for which multiplication by A is tantamount to scalar multiplication by λ.



• Determining the exact values of the set of eigenvalues for a matrix A_{nxn} is requires solving the so-called **characteristic equation**: $|A - \lambda I| = 0$, which is an *n*-degree polynomial equation in the variable λ.

Linear Independence, Span and Basis

- A set of vectors is called **linearly independent** if the set contains "no redundancy"; formally:

Def. A set of vectors $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ is **linearly independent** if:

$\alpha_1 \mathbf{v}_1 + \cdots \alpha_n \mathbf{v}_n = \mathbf{0}$ implies $\alpha_i = 0 \ \forall \ 1 \leq i \leq n$

- The **span** of a set is defined as the set of all linear combinations of the set of vectors.

$$\textbf{Def}.\ \text{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_n\} = \{\alpha_{i1}\mathbf{v}_1 + \cdots \alpha_{in}\mathbf{v}_n \mid i \in \mathbb{R}\}$$

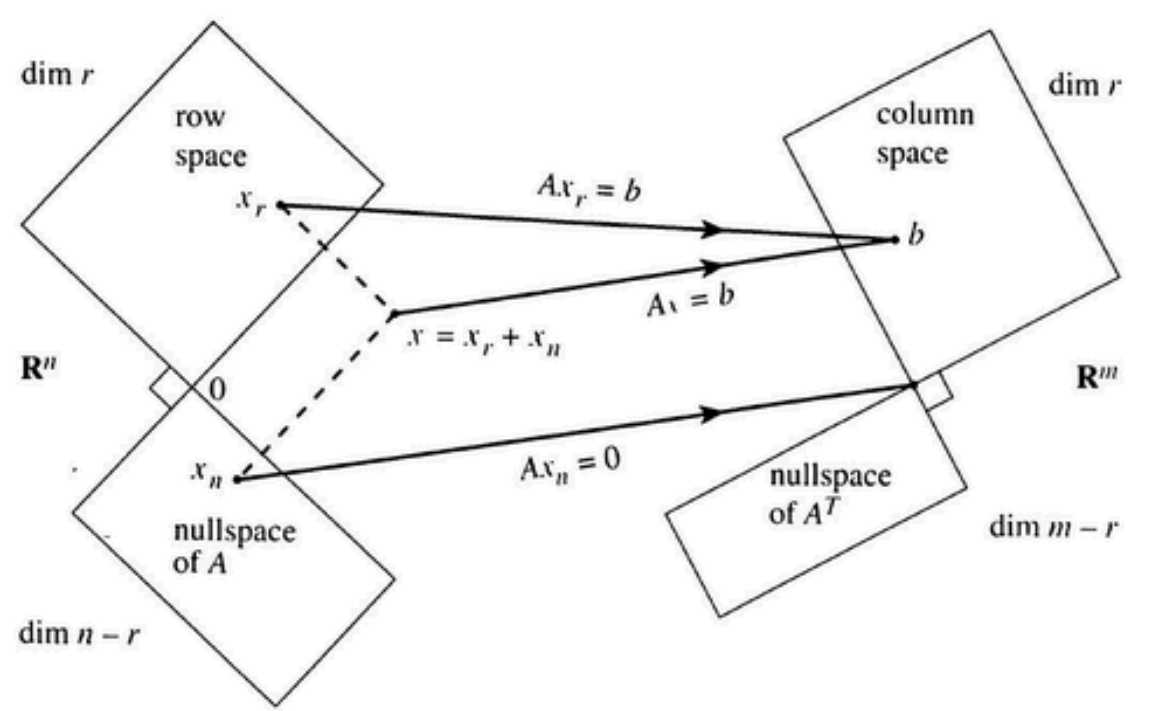
- A **basis** is a set of linearly independent vectors that spans the parent vector space.

$$e.g., \left\{ \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \right\} \text{ is the "standard" basis set for } \mathbb{R}^3$$

The Four Fundamental Subspaces

- The Four Fundamental subspaces of a matrix A_{mxn}:

1. (Column Space) **Col(A)**: the span of the column vectors of A.
2. (Null Space) **Null(A)**: the set of all vectors that satisfy A**x**=**0**.
3. (Row Space) **Row(A)**: the span of the row vectors of A.
4. (Null Space of A^T) **Null(A^T)**: the set of all vectors that satisfy A^T**x**=**0**.



Overview of Statistics/Probability

- We use statistics and probability to quantify and summarize our beliefs about a "state of the world" in the face of incomplete or partial knowledge.

- Denote a random event E; the **sample space S** consists of the set of all possible outcomes associated with E (e.g. if E="coin flip", S={H,T}).

- A **random variable** (e.g. X, Y) is a variable that is assigned a number based on the outcome of the random event E.

- Random variables are either **Discrete** (e.g., 0/1) or **Continuous** (e.g., height, time).

Overview of Statistics/Probability

Probability Distributions

• A **probability distribution** summarizes our total knowledge about the random event E, via the random variable X.

• For a discrete random variable, the probability distribution of X is called a **probability mass function** (pmf); a pmf satisfies the following properties, with $|S|=k$:

$$1.\ 0 \le p(X_i) \le 1 \quad 1 \le i \le k$$

$$2.\ \sum_{i=1}^{k} p(X_i) = 1$$

• Similarly, for a continuous random variable, the probability distribution of X is called a **probability density function** (pdf); a pdf satisfies the following properties, with $|S|=\infty$:

$$1.\ 0 \le p(X_i) \le 1$$

$$2.\ \int_S p(X)\,dx = 1$$

Overview of Statistics/Probability

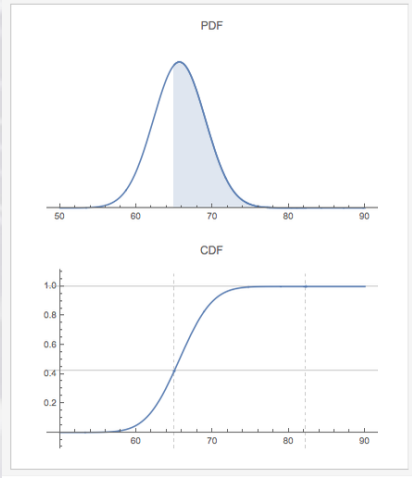
Probability Distributions

• A **cumulative density function** (cdf) is defined as the cumulative probability up to a given value of a random variable:

$$F_X(x) = p(X \le x) = \int_{-\infty}^{x} p(u)\, du$$

• Percentiles and quartiles can be defined in a natural way with respect to a cdf:

$$\underbrace{F_X(x) = 0.25}_{x=Q_1} \qquad \underbrace{F_X(x) = 0.5}_{x=Q_2 \text{ (median)}} \qquad \underbrace{F_X(x) = 0.75}_{x=Q_3}$$



• Note that due to the *Fundamental Theorem of Calculus*, it follows that:

$$\frac{d}{dx}\underbrace{F_X(x)}_{cdf} = \underbrace{p(x)}_{pdf}$$

Overview of Statistics/Probability

Properties of Probability Distributions

- Two random events E_1 and E_2 are disjoint if: $S_1 \cap S_2 = \emptyset$.

- If two events E_1 and E_2 are disjoint, then: $P(E_1 \cup E_2) = P(E_1) + P(E_2)$

- More generally, the **addition rule of probability** states, that for *any* two events E_1 and E_2:

$$P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2)$$

Overview of Statistics/Probability

Conditional Probability

- Def. **Conditional Probability**:

$$P\underbrace{(A|B)}_{\substack{\text{probability of} \\ \text{"A given B"}}} = \frac{P(A \cap B)}{P(B)}$$

- From this definition, we can derive the **multiplication rule of probability**:

$$P(A \cap B) = P(A|B)P(B)$$

- Equivalently,

$$P(A \cap B) = P(B|A)P(A)$$

Overview of Statistics/Probability

Independence

• We say that events A & B are **independent** if the outcome of A has no bearing on B (and vice versa); more formally the joint probability distribution p(A,B) factors.

• **Def**. A & B are independent if: $P(A \cap B) = P(A)P(B)$

• Equivalently, if A & B are independent, it also follows that:

$$P(A|B) = P(A) \qquad P(B|A) = P(B)$$

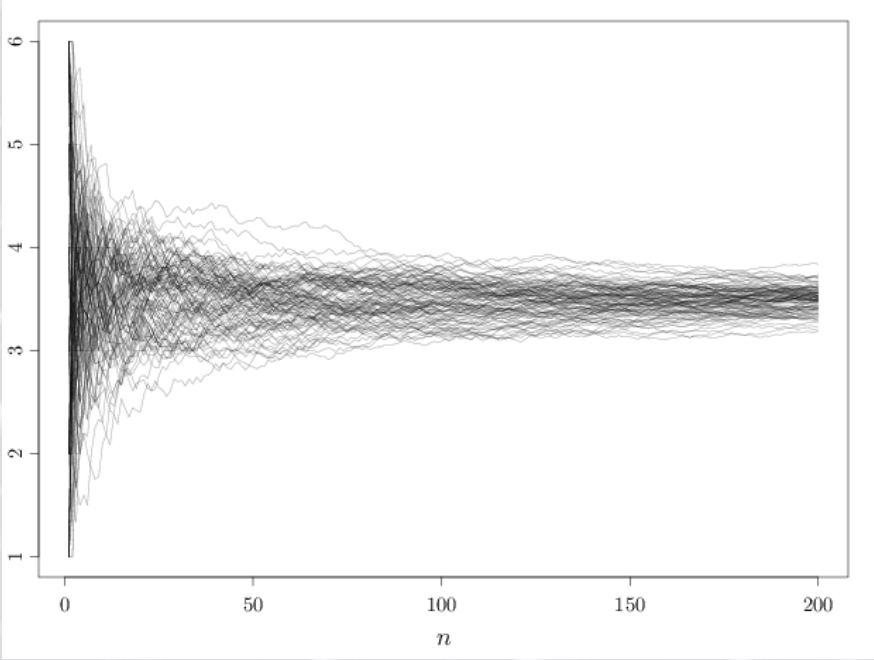
Thus, in summary, if A & B are independent: $P(A \cap B) = (A|B)P(B) = P(A)P(B)$

*Independence is commonly denoted: $A \perp B$

Overview of Statistics/Probability

• Two major theorems in elementary statistics: (1) the **Law of Large Numbers** and (2) the **Central Limit Theorem**.

• The *Law of Large Numbers* states (paraphrasing): Experimental (i.e. empirical) probabilities converge to their associate theoretical probability as the number of trials tends to infinity.

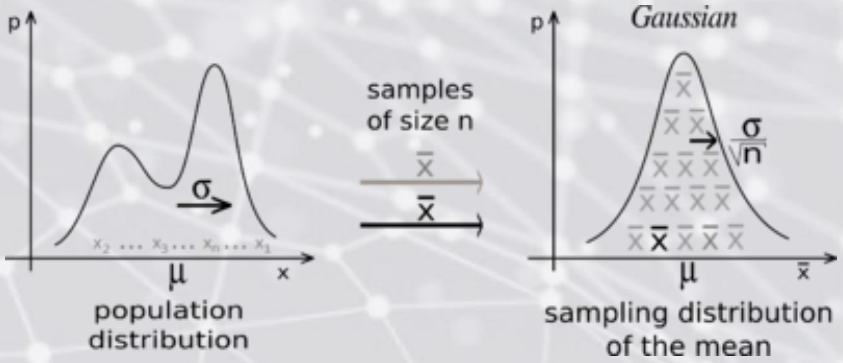


Overview of Statistics/Probability

The *Central Limit Theorem* (a conceptual pillar of statistics)

In words: given a sufficiently large sample size from a population (with a finite level of variance), the mean of all samples from the same population will be approximately equal to the mean of the population. Furthermore, all of the samples will follow an approximate normal distribution pattern, with all variances being approximately equal to the variance of the population divided by each sample's size.

In a picture:



Whatever the form of the population distribution, the sampling distribution tends to a Gaussian, and its dispersion is given by the Central Limit Theorem.

In a theorem: Suppose $\{X_1, X_2, \ldots,\}$ is a sequence of I.I.D. random variables with $E[X_i]=\mu$ and $Var[X_i]=\sigma^2<\infty$ Then as n approaches infinity, the random variable $(1/n)(X_1+\ldots+X_n)$ converges to a normal $N(0, \sigma^2/n)$:

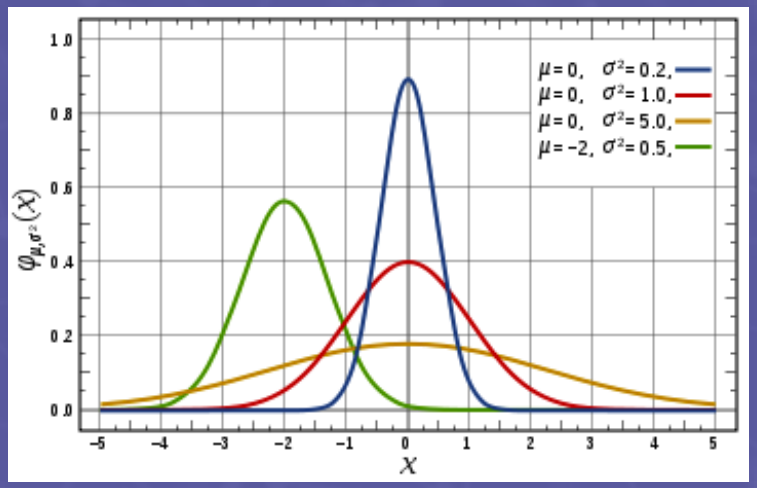
$$\left(\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)-\mu\right)\xrightarrow{d} N\left(0,\frac{\sigma^2}{n}\right)$$

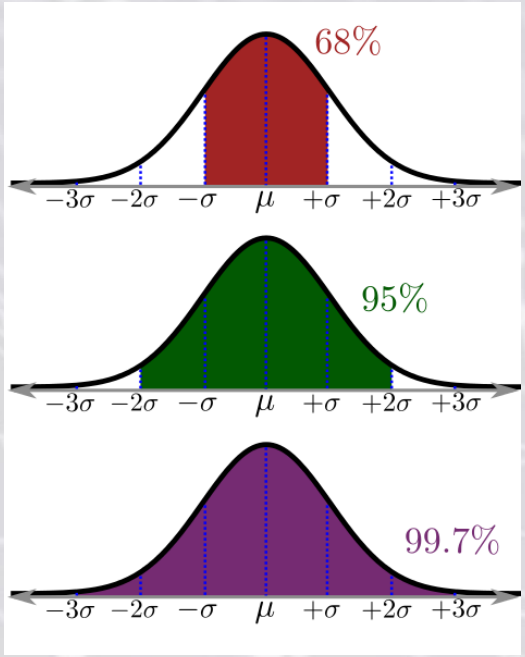
Probability Distributions

• Here are some (but certainly not all) of the essential probability distributions for ML and applied statistics:

1-D Gaussian (i.e. Normal)

$$N(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$



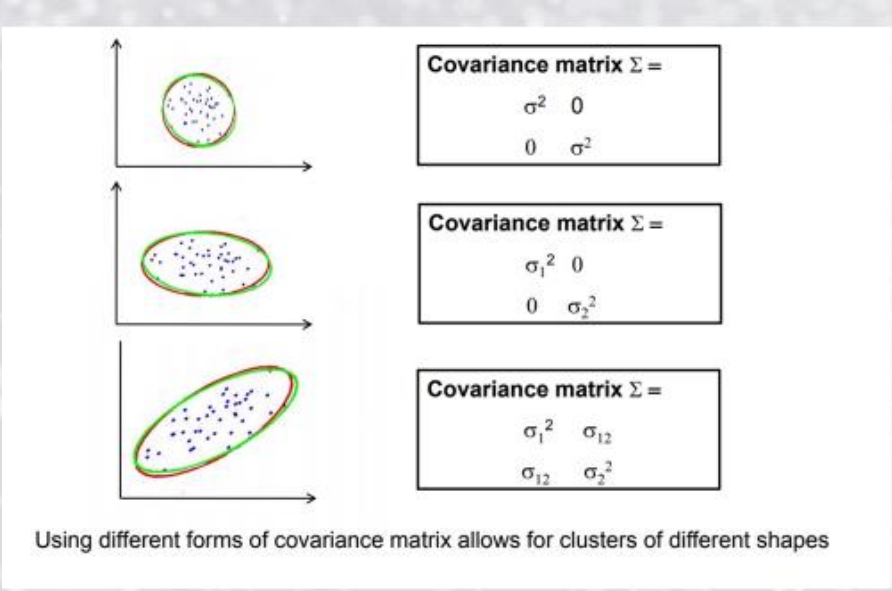


• When μ=0 and σ=1 (i.e. N(0,1)) we call this the **standard Normal** model.

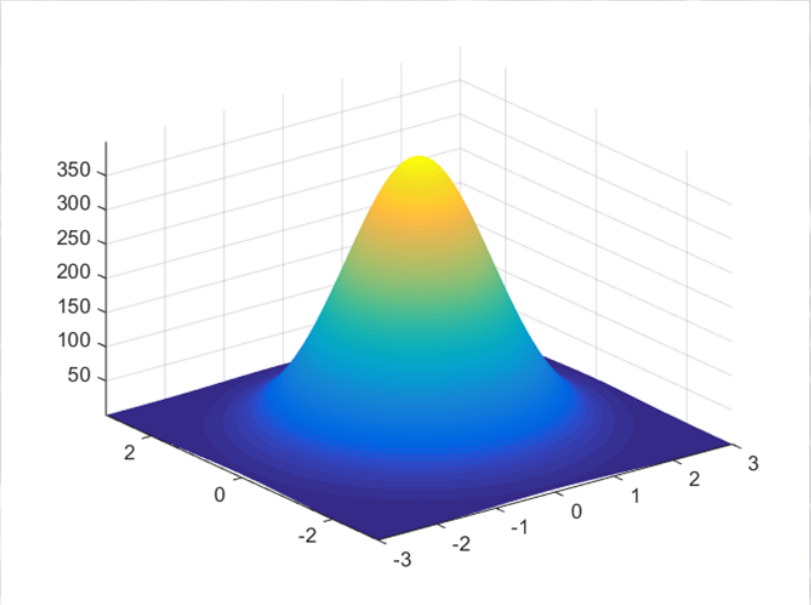
Probability Distributions

Multivariate Gaussian (i.e. MVN)

$$N(x;\mu,\Sigma) = \frac{1}{(2\pi)^{D/2}|\Sigma|^{1/2}} exp\left\{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right\}$$



Using different forms of covariance matrix allows for clusters of different shapes



Probability Distributions

Bernoulli & Binomial Distributions

• The **Bernoulli distribution** is a single variable, discrete distribution, describing a random variable with two discrete states (e.g. heads/tails for a single coin flip). The Bernoulli distribution forms the basis of the **Binomial distribution**, which models repetitions of independent Bernoulli trials (N total).

Bernoulli pmf

$$p(X=1)=\theta \quad p(X=0)=1-\theta \qquad 0\le\theta\le 1$$

Binomial pmf

$$\underbrace{p(X=k)}_{\substack{\text{probability of}\\ \text{"k successes in n trials"}}} = \binom{n}{k}\theta^k(1-\theta)^{n-k}$$

• For example, with a biased coin ($\theta = 0.6$), we have:

$$p(\text{exactly 7 H in 10 flips}) = \binom{10}{7}0.6^7(1-0.6)^{10-7}$$

Summary Statistics for Random Variables

Expectation and Variance of a Random Variable

• The **Expected Value** of a random variable X summarizes the outcome: "if the trial were executed once, on average, this is the numerical value we would expect for X"; E[X] accordingly computes the arithmetic mean of a random variable, i.e. E[X]=μ.

$$E[X]=\sum_{i=1}^{k} x_i P(X=x_i) \quad \text{(Discrete RV)} \qquad E[X]=\int_S xp(x)dx \quad \text{(Continuous RV)}$$

For example, to compute the expected number of heads X in 10 flips of a fair coin (X~Binomial) we have:

$$E[X]=\sum_{i=0}^{10} k\binom{10}{k}\left(\frac{1}{2}\right)^k\left(\frac{1}{2}\right)^{n-k}=5$$

Summary Statistics for Random Variables

Properties of Expected Value and Variance

• Expected Value is a **linear operator** (as are matrix multiplication, limits, differentiation and integration, among other common mathematical operators) -- meaning that it obeys the following two *linearity properties:*

1.For any two random variables X, Y: $E[X+Y]=E[X]+E[Y]$

2.For any $c \in \mathbb{R}$: $E[cX]=cE[X]$

The following corollary is also useful: $Var[X]=E[X^2]-E[X]^2$

Proof.

$$Var[X]=E\left[(X-\mu)^2\right]=E\left[X^2-2\mu X+\mu^2\right]$$

$$\underset{\text{by linearity of } E[\cdot]}{=} E[X^2]-E[2\mu X]+E[\mu^2]$$

$$=E[X^2]-\underset{\text{by linearity of } E[\cdot]}{2\mu}E[X]+\underset{E[const]=const}{\mu^2}$$

$$=E[X^2]-2\mu^2+\mu^2$$

Covariance

• **Covariance** is a measure of the linear relationship between two random variables, X and Y. If Cov(X,Y) > 0, this indicates a positive linear relationship between the random variables (i.e. as X increases, Y increases; as X decreases, Y decreases); when Cov(X,Y) < 0 the variables share a negative linear relationship; Cov(X,Y) = 0 indicates the absence of a linear relationship.

Def. $$Cov(X,Y) = E\left[\left(X - E[Y]\right)\left(X - E[Y]\right)\right]$$

Lemma

$$\text{If } X \perp Y (i.e. \text{ if } X \text{ and } Y \text{ are independent}), \text{ then } Cov(X,Y) = 0$$

* Note that the *converse* of the lemma above fails; in other words Cov(X,Y)=0 need not imply that X and Y are independent.

Covariance

• The **Covariance Matrix** (Σ) for a set of random variables $\{X_1,\ldots,X_N\}$ is defined as the matrix of pairwise covariances:

Def.

$$\text{Let } \mathbf{X} = \{X_1,\ldots,X_N\}, \quad \underbrace{\Sigma_{ij}}_{\text{matrix of Covariances}} = Cov(X_i, X_j) = E\left[\left(X_i - E[X_i]\right)\left(X_j - E[X_j]\right)\right]$$

$$\Sigma = \begin{bmatrix} Var[X_1] & Cov(X_1, X_2) & \cdots & Cov(X_1, X_N) \\ Cov(X_2, X_1) & Var[X_2] & \cdots & Cov(X_2, X_N) \\ \vdots & \vdots & \ddots & \vdots \\ Cov(X_N, X_1) & Cov(X_N, X_2) & \cdots & Var[X_N] \end{bmatrix}$$

Note that Σ is symmetric and positive semi-definite. The covariance matrix is used to parameterize the MVN (multivariate normal distribution); the covariance matrix can likewise be computed for a dataset.

Bayes' Theorem

• **Bayes' Theorem** is a vital (yet simple) conditional probability formula; today its use is omnipresent across ML.

Def. $$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Derivation: $$P(A|B) = \underbrace{\frac{P(A \cap B)}{P(B)}}_{\text{by definition of conditional probability}} = \underbrace{\frac{P(B|A)P(A)}{P(B)}}_{\text{by multiplication rule}}$$

• More importantly, Bayes' Theorem can be generalized to encapsulate the whole of the inductive element of the scientific method. To this end, consider H (hypothesis) and D (data):

$$P(H|D) = \frac{P(D|H)P(H)}{P(D)}$$

• In this case, Bayes' Theorem yields a natural mechanism for updating our belief about the world/the plausibility of a hypothesis (H) given an observation (D). P(H|D) is referred to as the **posterior probability** of H, P(H) is called the **prior probability** of H, P(D|H) defines the **likelihood of the data**, and P(D) is the **data prior**.

Bayesian and Frequentist Statistics

• There exist two general paradigms for modern statistics: the **frequentist** and **Bayesian** approaches.

Frequentists: Generally consider model parameters (θ) as fixed; data are drawn from some objective distribution, defined by θ. There exists various well-known pathologies associated with frequentism, including the "problem of induction" (Hume), the Black Swan Paradox, limited exact solutions and a heavy reliance upon long-term frequencies.

Bayesians: (Observed) data are fixed; data are observed from a realized sample; we encode prior beliefs, and parameters values are described probabilistically.

- Frequentists use the **Maximum Likelihood Estimate** (MLE) for point estimates of parameters θ :

$$\hat{\theta}_{MLE} = \arg\max_{\theta} P(D \mid \theta)$$

- Bayesians instead use the Maximum A Posterior (MAP) for parameter estimates:

$$\hat{\theta}_{MAP} = \arg\max_{\theta} P(\theta \mid D) = \arg\max_{\theta} P(D \mid \theta) P(\theta)$$

(Very Brief) Information Theory

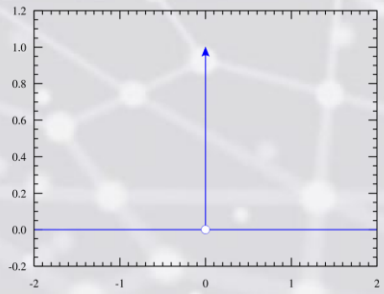
• The **entropy** of a discrete random variable X (equivalently: the entropy of the pmf associated with X) is defined:

$$H(X) = -\sum_i p(X = x_i) \log p(X = x_i)$$

• The **differential entropy** of a continuous random variable is defined analogously:

$$H(X) = -\int_S p(x) \log p(x) dx$$

- Entropy quantifies disorder/"surprise"; the **Principle of Insufficient Reasons** (PIR) states (paraphrasing) that in the absence of compelling evidence, one should adopt a **maximum entropy probability distribution**. The uniform distribution is a maximum entropy distribution; the Gaussian distribution is a likewise a maximum entropy distribution (up to second moments). Entropy is minimized (i.e. zero) for deterministic events, e.g. *Dirac delta function*.



(Very Brief) Information Theory

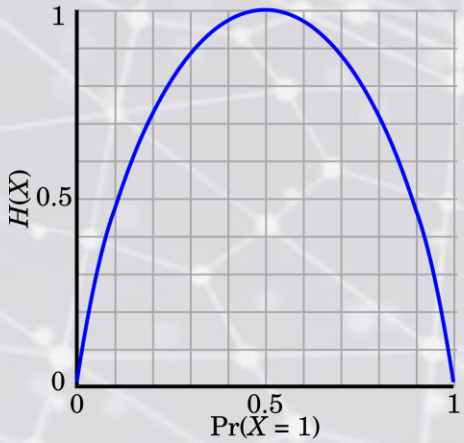
Ex. The entropy of a Bernoulli random variable X is given by:

$$p(X=1)=\theta \quad p(X=0)=1-\theta \qquad 0\le\theta\le 1$$

$$H(X)=-\sum_i p(X=x_i)\log p(X=x_i)$$

$$=-\left(p(X=1)\log p(X=1)+p(X=0)\log p(X=0)\right)$$

$$=-\left(\theta\log\theta+(1-\theta)\log(1-\theta)\right)$$



* Notice that entropy is maximized in this case when $\theta = 0.5$, which corresponds with a binary uniform distribution; conversely, entropy is minimized when either $\theta = 0$ or $\theta = 1$, in which case the even is deterministic.

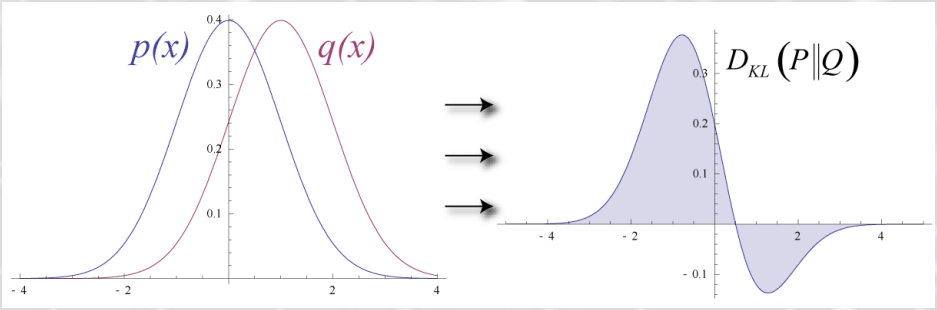
(Very Brief) Information Theory

• The **Kullback-Leibler Divergence** quantifies the difference between two probability distributions, p(x) and q(x).

Def.

$$KL(p \| q) = \sum_i p(X = x_i) \log \frac{p(X = x_i)}{q(X = x_i)}$$

$$KL(p \| q) = \int_S p(x) \log \frac{p(x)}{q(x)} dx$$



The **Information Inequality** states:

$$KL(p \| q) \geq 0 \text{ and } KL(p \| q) = 0 \leftrightarrow p = q$$

(Very Brief) Information Theory

$$KL(p\|q)=\sum_i p(X=x_i)\log\frac{p(X=x_i)}{q(X=x_i)}$$

$$KL(p\|q)=\int_S p(x)\log\frac{p(x)}{q(x)}dx$$

$$KL(p\|q)\geq 0 \text{ and } KL(p\|q)=0\leftrightarrow p=q$$

*Recall that covariance/correlation are inherent measures of the linear relationship between two random variables. Using KL-divergence, we can develop a more general notion of independence, called **mutual information**.

Def.

$$MI(X,Y)=KL(p(X,Y)\|p(X)p(Y))=\sum_X\sum_Y p(X,Y)\log\frac{p(X,Y)}{p(X)p(Y)}$$

$$MI(X,Y)=KL(p(X,Y)\|p(X)p(Y))=\int_{S_X}\int_{S_Y} p(x,y)\log\frac{p(x,y)}{p(x)p(y)}dxdy$$

*From the information inequality, it follows that:

$$MI(X,Y)\geq 0 \text{ and } MI(X,Y)=0\leftrightarrow p(X,Y)\|p(X)p(Y)$$

$$\text{(i.e. } MI(X,Y)=0\leftrightarrow X\perp Y)$$

Thus, MI can be seen as a more general measure of statistical independence than covariance.

Fin