Linear Algebra Notations

- Vector: $\mathbf{x} = [x_1, \ldots, x_d]^\top = \begin{bmatrix} x_1 \\ \vdots \\ x_d \end{bmatrix}$
  - product: $< \mathbf{x}^{(1)}, \mathbf{x}^{(2)} >= \mathbf{x}^{(1)\top}\mathbf{x}^{(2)} = \sum_{i=1}^{d} x_i^{(1)} x_i^{(2)}$
  - norm: $||\mathbf{x}||_2 = \sqrt{\mathbf{x}^\top \mathbf{x}} = \sqrt{\sum_i x_i^2}$ (Euclidean)
- Matrix: $\mathbf{X} = \begin{bmatrix} X_{1,1} & \ldots & X_{1,m} \\ \vdots & \vdots & \vdots \\ X_{n,1} & \ldots & X_{n,m} \end{bmatrix}$
  - product: $(\mathbf{X}^{(1)}\mathbf{X}^{(2)})_{i,j} = \mathbf{X}^{(1)}_{i,\cdot}\mathbf{X}^{(2)}_{\cdot,j} = \sum_k X^{(1)}_{i,k} X^{(2)}_{k,j}$
  - norm: $||\mathbf{X}||_F = \sqrt{\mathrm{trace}(\mathbf{X}^\top \mathbf{X})} = \sqrt{\sum_i \sum_j X_{i,j}^2}$ (Frobenius)

- Trace of matrix: $\mathrm{trace}(\mathbf{X}) = \sum_i X_{i,i}$
  - trace of products:

$$\mathrm{trace}(\mathbf{X}^{(1)}\mathbf{X}^{(2)}\mathbf{X}^{(3)}) = \mathrm{trace}(\mathbf{X}^{(3)}\mathbf{X}^{(1)}\mathbf{X}^{(2)}) = \mathrm{trace}(\mathbf{X}^{(2)}\mathbf{X}^{(3)}\mathbf{X}^{(1)})$$

- Determinant
  - of triangular matrix: $\det(\mathbf{X}) = \prod_i \mathbf{X}_{i,i}$
  - of transpose of matrix: $\det(\mathbf{X}^\top) = \det(\mathbf{X})$
  - of inverse of matrix: $\det(\mathbf{X}^{-1}) = \det(\mathbf{X})^{-1}$
  - of product of matrix: $\det(\mathbf{X}^{(1)}\mathbf{X}^{(2)}) = \det(\mathbf{X}^{(1)})\det(\mathbf{X}^{(2)})$

- Orthogonal matrix: $\mathbf{X}^\top = \mathbf{X}^{-1}$
- Positive definite matrix: $\mathbf{v}^\top \mathbf{X} \mathbf{v} > 0 \quad \forall \mathbf{v} \in \mathbb{R}^d$
  - if « $\geq$ » , then positive semi-definite

Better Explanation for Positive Definite Matrix: https://www.math.utah.edu/~zwick/Classes/Fall2012_2270/Lectures/Lecture33_with_Examples.pdf

- Set of linearly dependent vectors $\{\mathbf{x}^{(t)}\}$:

$$\exists \mathbf{w}, t^* \text{ such that } \mathbf{x}^{(t^*)} = \sum_{t \neq t^*} w_t \mathbf{x}^{(t)}$$

- Rank of matrix: number of linear independent columns
- Range of a matrix:

$$\mathcal{R}(\mathbf{A}) = \{\mathbf{x} \in \mathbb{R}^n \mid \exists \mathbf{w} \text{ such that } \mathbf{x} = \sum_j w_j \mathbf{A}_{\cdot,j}\}$$

- Null space of a matrix:

$$\mathrm{Null}(\mathbf{A}) = \{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{A}\mathbf{x} = \mathbf{0}\}$$

Some Better Explanations on these terms:
Set of Linearly Dependent Vectors: https://www.sciencedirect.com/topics/mathematics/linearly-dependent

Table 4.1. Equivalent conditions for a subset S of a vector space to be linearly independent or linearly dependent

Linear Independence of S	Linear Dependence of S	Source
If $S = \{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ and $a_1\mathbf{v}_1 + \ldots + a_n\mathbf{v}_n = \mathbf{0}$, then $a_1 = a_2 = \ldots = a_n = 0$. (The zero vector requires zero coefficients.)	If $S = \{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$, then $a_1\mathbf{v}_1 + \ldots + a_n\mathbf{v}_n = \mathbf{0}$ for some scalars $a_1, a_2, \ldots, a_n$, with some $a_i \neq 0$. (The zero vector does not require all coefficients to be zero.)	Definition
No vector in S is a finite linear combination of other vectors in S.	*Some* vector in S is a finite linear combination of other vectors in S.	Theorem 4.8 and Remarks after Example 14
For every $\mathbf{v} \in S$, we have $\mathbf{v} \notin \text{span}(S - \{\mathbf{v}\})$.	There is a $\mathbf{v} \in S$ such that $\mathbf{v} \in \text{span}(S - \{\mathbf{v}\})$.	Alternate characterization
For every $\mathbf{v} \in S$, $\text{span}(S - \{\mathbf{v}\})$ does not contain all the vectors of $\text{span}(S)$.	There is some $\mathbf{v} \in S$ such that $\text{span}(S - \{\mathbf{v}\}) = \text{span}(S)$.	Exercise 12
If $S = \{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$, then for each k $\mathbf{v}_k \notin \text{span}(\{_{v1}, \ldots, \mathbf{v}_{k-1}\})$. (Each $\mathbf{v}_k$ is not a linear combination of the previous vectors in S.)	If $S = \{\mathbf{v}_1, \ldots, _{vn}\}$, some $\mathbf{v}_k$ can be expressed as $\mathbf{v}_k = a_1\mathbf{v}_1 + \ldots + a_{k-1}\mathbf{v}_{k-1}$. (Some $\mathbf{v}_k$ is a linear combination of the previous vectors in S.)	Exercise 22
Every vector in $\text{span}(S)$ can be uniquely expressed as a linear combination of the vectors in S.	*Some* vector in $\text{span}(S)$ can be expressed in more than one way as a linear combination of the vectors in S.	Theorem 4.9 and Theorem 4.10
Every finite subset of S is linearly independent.	*Some* finite subset of S is linearly dependent.	Definition when S is infinite

Rank of Matrix A: https://stattrek.com/matrix-algebra/matrix-rank.aspx
Range and Null Space of a Matrix: https://math.stackexchange.com/questions/2037602/what-is-range-of-a-matrix

- Eigenvalues and eigenvectors

$$\{\lambda_i, \mathbf{u}_i \mid \mathbf{X}\mathbf{u}_i = \lambda_i \mathbf{u}_i \text{ and } \mathbf{u}_i^\top \mathbf{u}_j = 1_{i=j}\}$$

- Properties
  - can write $\mathbf{X} = \sum_i \lambda_i \mathbf{u}_i \mathbf{u}_i^\top$
  - determinant of **any** matrix: $\det(\mathbf{X}) = \prod_i \lambda_i$
  - positive definite if $\lambda_i > 0 \quad \forall i$
  - rank of matrix is the number of non-zero eigenvalues

More info on these at the Matrix Cookbook Pg30.

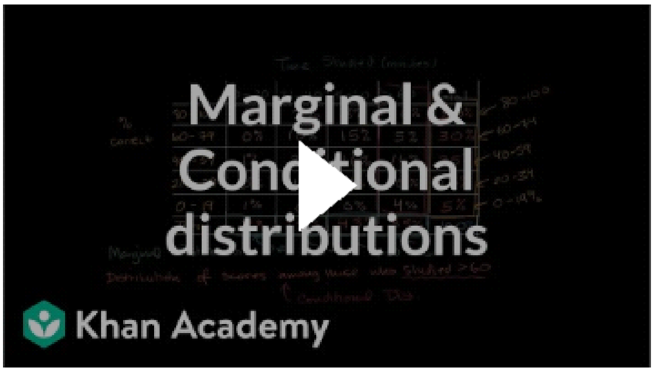
Probability

- Probability space: triplet $(\Omega, \mathcal{F}, P)$
  - Ω is the space of possible outcomes
  - $\mathcal{F}$ is the space of possible events
  - P is a probability measure mapping an ***event*** to its probability [0,1]
  - example: throwing a die
    - $\Omega = \{1,2,3,4,5,6\}$
    - $e = \{1,5\} \in \mathcal{F}$ (i.e. die is either 1 or 5)
    - $P(\{1,5\}) = \frac{2}{6}$
- Properties:

1. $P(\{\omega\}) \geq 0 \quad \forall \omega \in \Omega$ 2. $\sum_{\omega \in \Omega} P(\{\omega\}) = 1$

- Joint distribution: $p(X = x, O = o, S = s)$ $\quad (p(x,s,o)$ for short$)$
  - the probability of a complete assignment of many random variables
  - example: $p(X = 1, O = 1, S = 0) = 0$
- Marginal distribution: $p(o,s) = \sum_x p(x,o,s)$
  - the probability of a partial assignment
  - example: $p(O = 1, S = 0) = \frac{1}{6}$
- Conditional distribution: $p(S = s|O = o)$
  - the probability of some variables, assuming an assignment of other variables
  - example: $p(S = 1|O = 1) = \frac{2}{3}$

Joint Distribution: https://www.statisticshowto.com/joint-probability-distribution/
Marginal Distribution: https://www.statisticshowto.com/probability-and-statistics/statistics-definitions/marginal-distribution/
Conditional Distribution: https://www.statisticshowto.com/conditional-distribution/, Marginal distribution and conditional distribution | AP Statistics | Khan Academy



Statistics

- Sample mean:

$$\widehat{\boldsymbol{\mu}} = \frac{1}{T} \sum_t \mathbf{x}^{(t)}$$

- Sample variance:

$$\widehat{\boldsymbol{\sigma}}^2 = \frac{1}{T-1} \sum_t (\mathbf{x}^{(t)} - \widehat{\boldsymbol{\mu}})^2$$

- Sample covariance matrix:

$$\widehat{\Sigma} = \frac{1}{T-1} \sum_t (\mathbf{x}^{(t)} - \widehat{\boldsymbol{\mu}})(\mathbf{x}^{(t)} - \widehat{\boldsymbol{\mu}})^\top$$

- These estimators are unbiased, i.e.:

$$\mathrm{E}[\widehat{\boldsymbol{\mu}}] = \boldsymbol{\mu} \quad \mathrm{E}[\widehat{\sigma}^2] = \sigma^2 \quad \mathrm{E}\left[\widehat{\Sigma}\right] = \Sigma$$

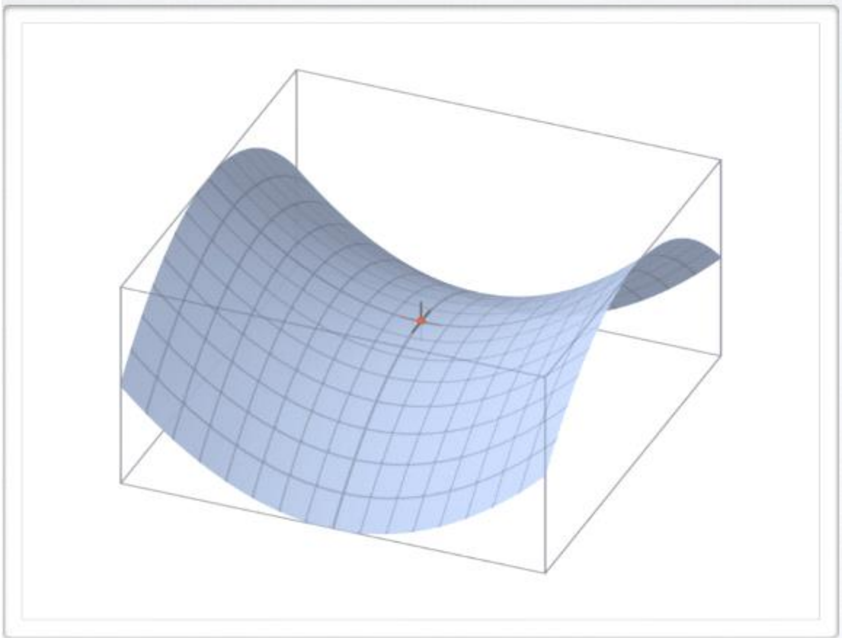
Machine Learning

- Critical points: $\{\mathbf{x} \in \mathbb{R}^d \mid \nabla_{\mathbf{x}} f(\mathbf{x}) = 0\}$

- Curvature in direction $\mathbf{v}$: $\mathbf{v}^\top \nabla^2_{\mathbf{x}} f(\mathbf{x}) \mathbf{v}$

- Types of critical points:
  - local minima: $\mathbf{v}^\top \nabla^2_{\mathbf{x}} f(\mathbf{x}) \mathbf{v} > 0 \ \ \forall \mathbf{v}$ (i.e. $\nabla^2_{\mathbf{x}} f(\mathbf{x})$ positive definite)
  - local maxima: $\mathbf{v}^\top \nabla^2_{\mathbf{x}} f(\mathbf{x}) \mathbf{v} < 0 \ \ \forall \mathbf{v}$ (i.e. $\nabla^2_{\mathbf{x}} f(\mathbf{x})$ negative definite)
  - saddle point: curvature is positive in certain directions and negative in others

saddle point



- Parametric model: its capacity is fixed and does not increase with the amount of training data
  - examples: linear classifier, neural network with fixed number of hidden units, etc.
- Non-parametric model: the capacity increases with the amount of training data
  - examples: k nearest neighbors classifier, neural network with adaptable hidden layer size, etc.