Home Reading: Mathematics for Machine Learning

Shan Wang

Lingnan College, Sun Yat-sen University

2020 Data Mining and Machine Learning LN3119

https://wangshan731.github.io/DM-ML/

Reference: 2019 CS420 Machine Learning by Weinan Zhang (SJTU)

Areas of Mathematics Essential to Machine Learning

- Machine learning is part of both **statistics** and computer science
  - Probability
  - Statistical inference
  - Validation
  - Estimates of error, confidence intervals
- **Linear Algebra**
  - Hugely useful for compact representation of linear transformations on data
  - Dimensionally reduction techniques
- **Optimization** theory

Notations

- $a \in A$ set membership: *a* is member of set *A*
- $|B|$ cardinality: number of items in set *B*
- $\|\mathbf{v}\|$ norm: length of vector v
- $\sum$ summation
- $\int$ integral
- $\mathbf{x}, \mathbf{y}, \mathbf{z}$ vector (bold, lower case)
- $\mathbf{A}, \mathbf{B}$ matrix (bold, upper case)
- $y = f(x)$ function: assigns unique value in range of *y* to each value in domain of *x*
- $y = f(\mathbf{x})$ function on multiple variables

Probability Spaces

- A probability space models a *random process or experiment* with three components:
  - Ω, the set of possible outcomes O
    - number of possible outcomes = $|\Omega|$
    - Discrete space $|\Omega|$ is finite
    - Continuous space $|\Omega|$ is infinite
  - F, the set of possible events E
    - number of possible events = $|F|$
  - P, the probability distribution
    - function mapping each outcome and event to real number between 0 and 1 (the probability of O or E)
    - probability of an event is sum of probabilities of possible outcomes in event

Axioms of Probability

- Non-negativity:
  - for any event $E \in F,\ p(E) \geq 0$
- All possible outcomes:
  - $p(\Omega) = 1$
- Additivity of disjoint events:
  - For all events $E, E' \in F$ where $E \cap E' = \emptyset$,

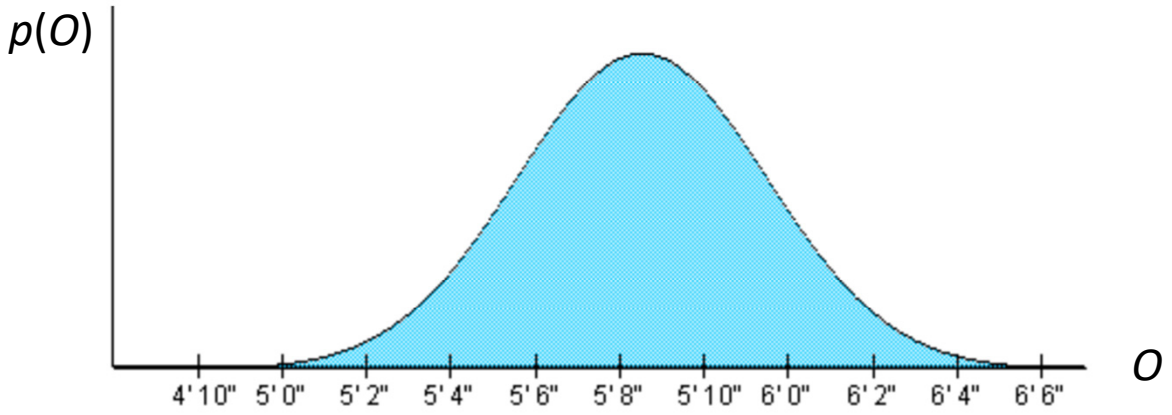
$$p(E \cup E') = p(E) + p(E')$$

Example of Discrete Probability Space

- Three consecutive flips of a coin
  - 8 possible outcomes: O = HHH, HHT, HTH, HTT, THH, THT, TTH, TTT
- 2^8=256 possible events
  - example: E = ($O \in$ { HHT, HTH, THH }), i.e. exactly two flips are heads
  - example: E = ($O \in$ { THT, TTT }), i.e. the first and third flips are tails
- If coin is fair, then probabilities of outcomes are equal
  - p(HHH) = p(HHT) = p(HTH) = p(HTT) = p(THH) = p(THT) = p(TTH) = p(TTT) = 1/8
  - example: probability of event E = (exactly two heads) is p(HHT) + p(HTH) + p(THH) = 3/8

Example of Continuous Probability Space

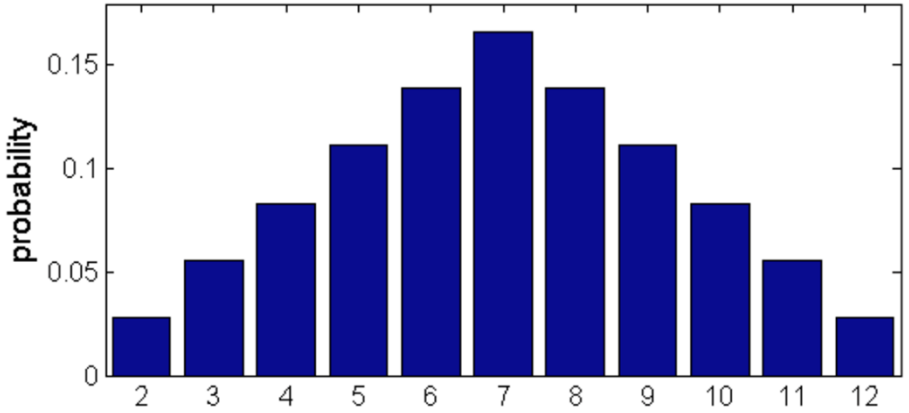
- Height of a randomly chosen American male
  - Infinite number of possible outcomes: *O* has some has some single value in range 2 feet to 8 feet
    - example: $E = (O \mid O < 5.5 \text{ feet})$, i.e. individual chosen is less than 5.5 feet tall
  - Infinite number of possible events
  - Probabilities of outcomes are not equal, and are described by a continuous function, $p(O)$



Probability Distributions

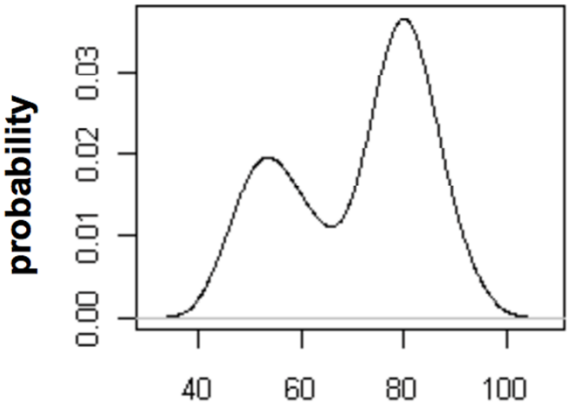
- Discrete: probability mass function (pmf)

example:
sum of two fair dice



- Continuous: probability density function (pdf)

example: waiting time
between eruptions of
Old Faithful (minutes)



Random Variables

- A random variable X is a function that associates a number x with each outcome O of a process
  - Common notation: $X(O) = x$, or just $X = x$
- Basically a way to redefine a probability space to a new probability space
  - X must obey axioms of probability
  - X can be discrete or continuous
- Example: X = number of heads in three flips of a coin
  - Possible values of X are 0, 1, 2, 3
  - $p(X = 0) = p(X = 3) = 1 / 8$, $p(X = 1) = p(X = 2) = 3 / 8$
  - Size of space (number of “outcomes”) reduced from 8 to 4
- Example: X = average height of five randomly chosen American men
  - Size of space unchanged, but pdf of X different than that for single man

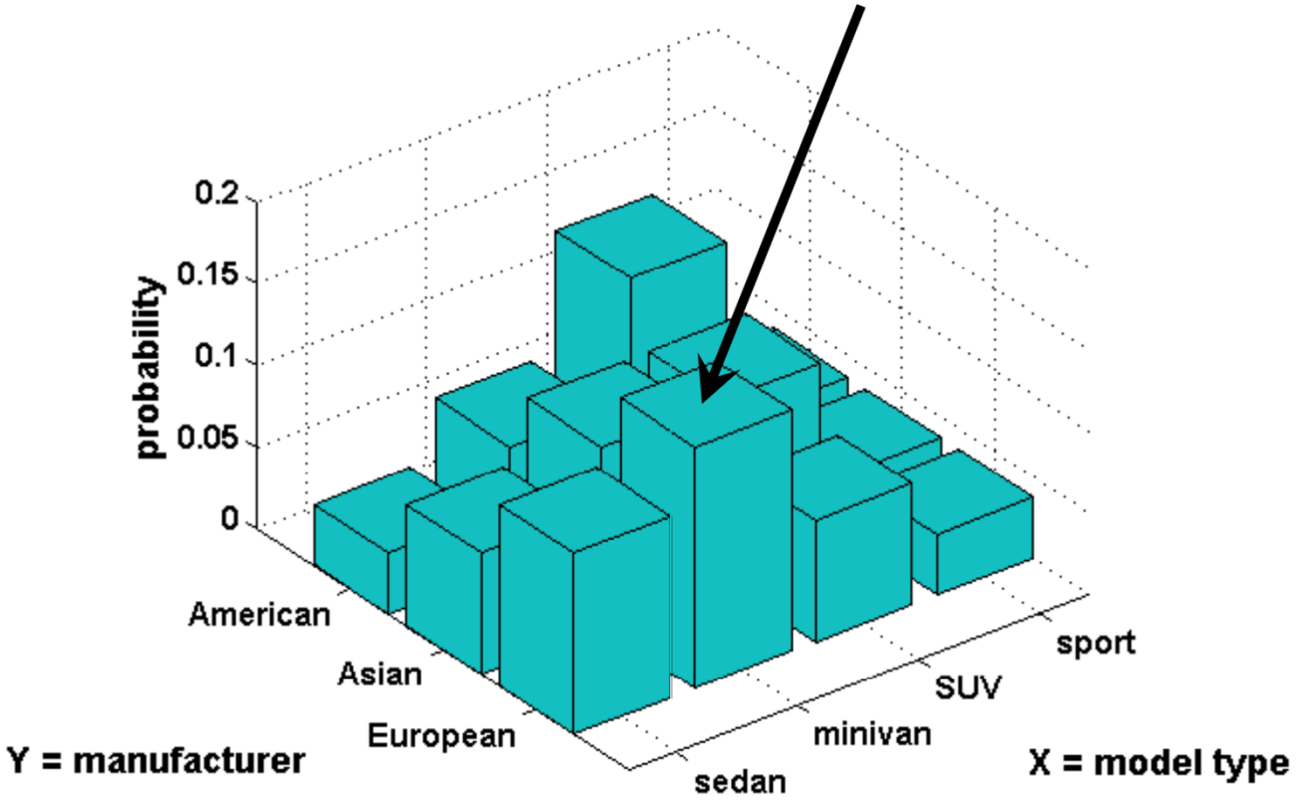
Multivariate Probability Distributions

- Scenario
  - Several random processes occur (doesn't matter whether in parallel or in sequence)
  - Want to know probabilities for each possible combination of outcomes
- Can describe as joint probability of several random variables
  - Example: two processes whose outcomes are represented by random variables *X* and *Y*. Probability that process *X* has outcome *x* and process *Y* has outcome *y* is denoted as

$$p(X = x, Y = y)$$

Example of Multivariate Distribution

joint probability: $p(X = \text{minivan}, Y = \text{European}) = 0.1481$



Multivariate Probability Distributions

- Marginal probability
  - Probability distribution of a single variable in a joint distribution
  - Example: two random variables *X* and *Y*:

$$p(X = x) = \sum_{b=\text{all values of } Y} p(X = x, Y = b)$$

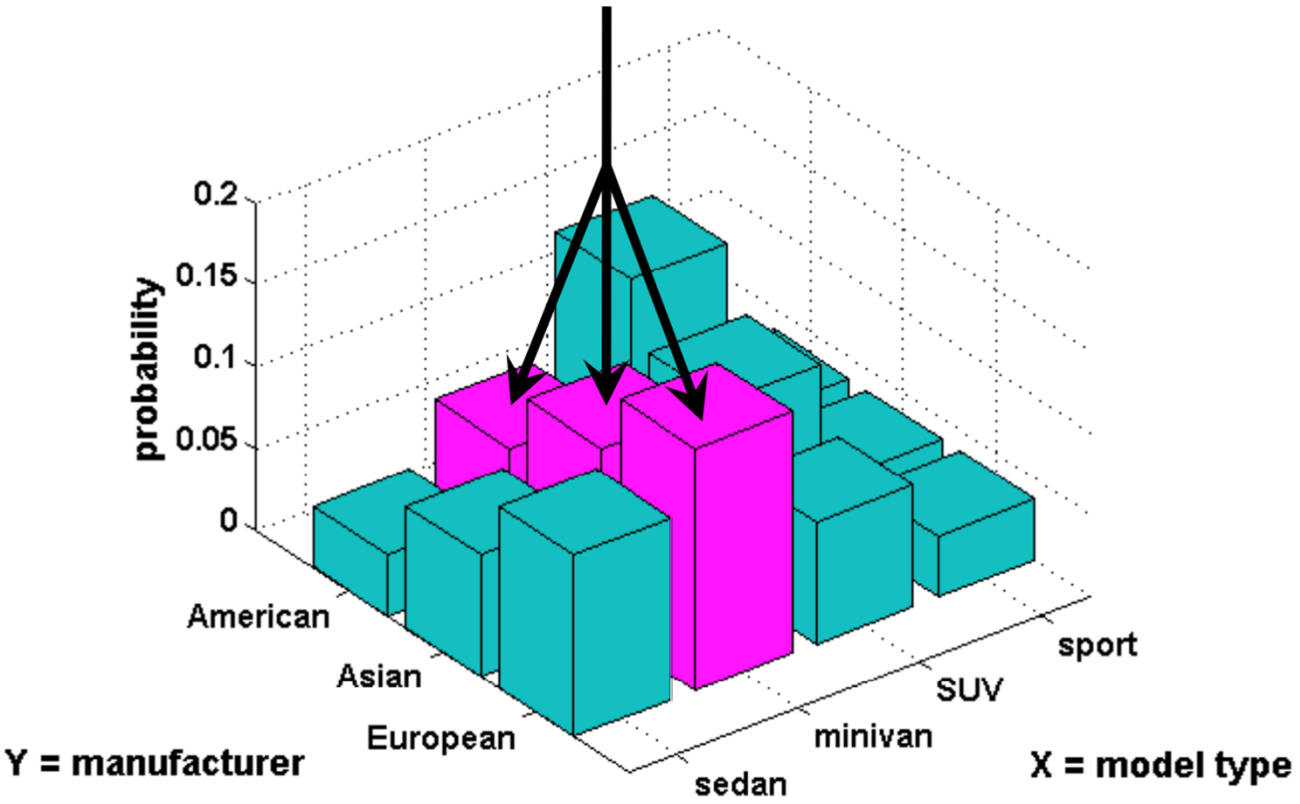
- Conditional probability
  - Probability distribution of one variable given that another variable takes a certain value
  - Example: two random variables X and Y :

$$p(X = x|Y = y) = \frac{p(X=x,Y=y)}{p(Y=y)}$$

Example of Marginal Probability

Marginal probability:

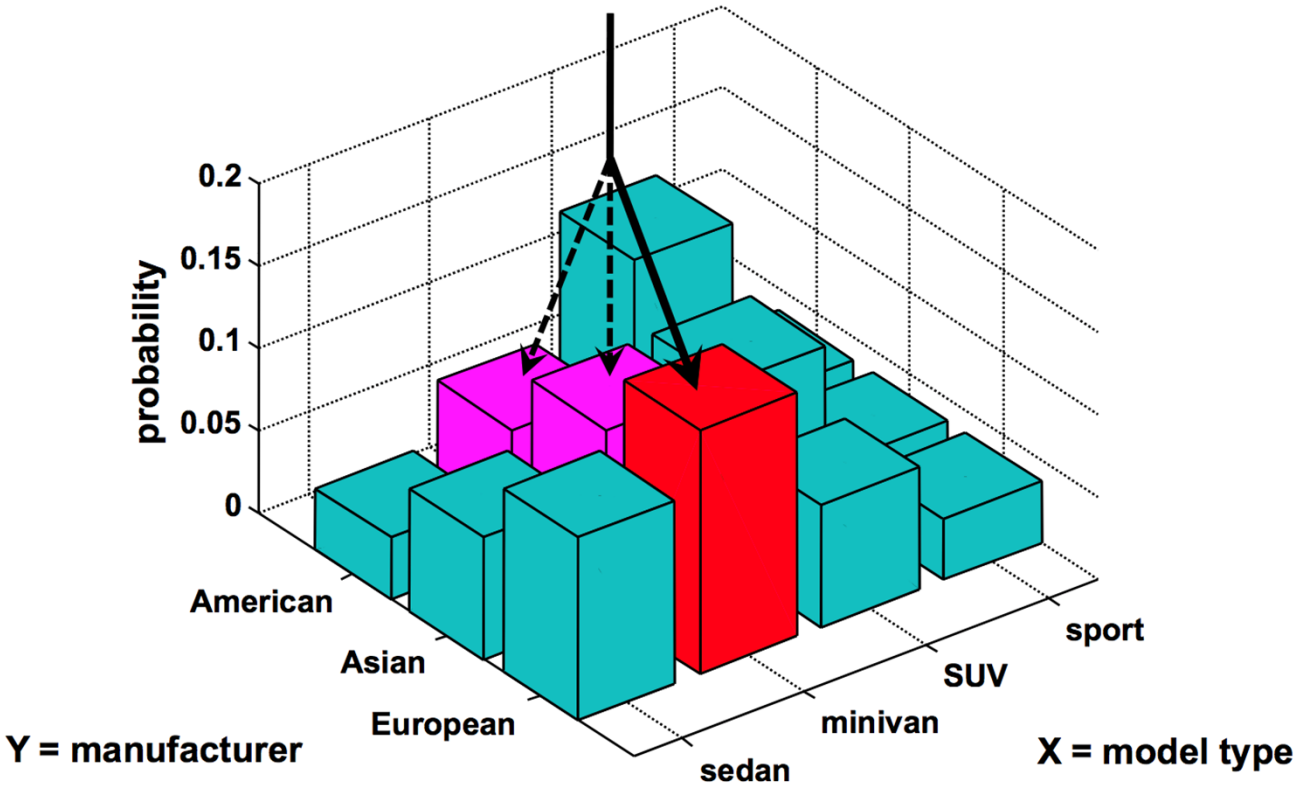
p(X = minivan) = 0.0741 + 0.1111 + 0.1481 = 0.3333



Example of Conditional Probability

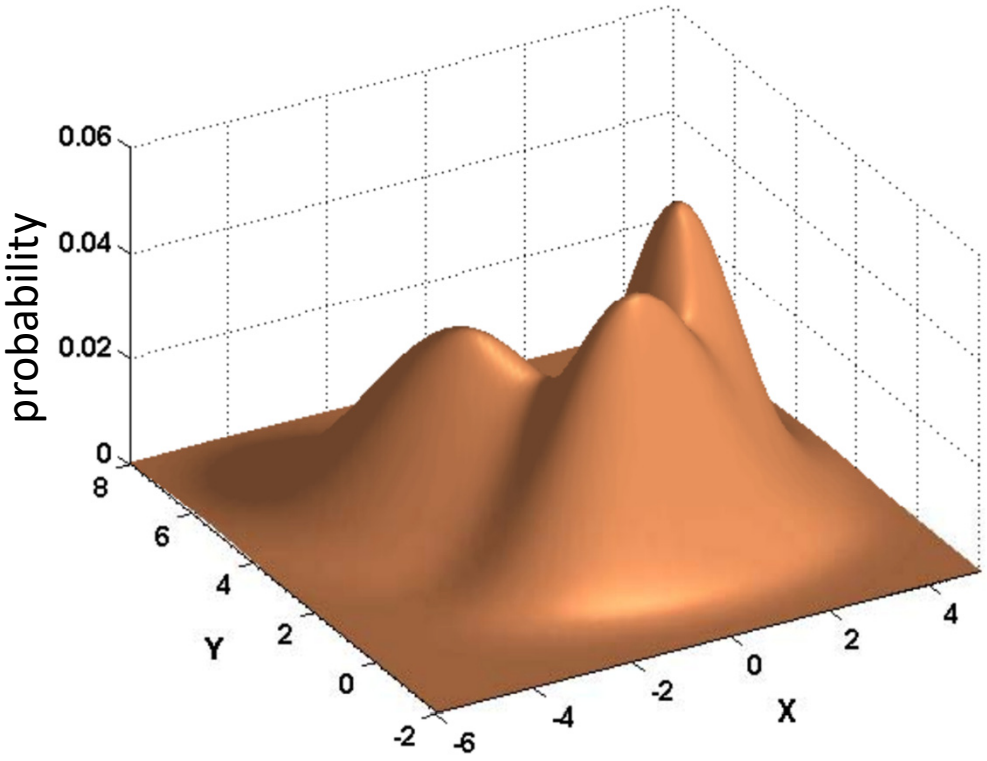
Conditional probability:

p(Y = European | X = minivan) = 0.1481 / (0.0741 + 0.1111 + 0.1481) = 0.4433



Continuous Multivariate Distribution

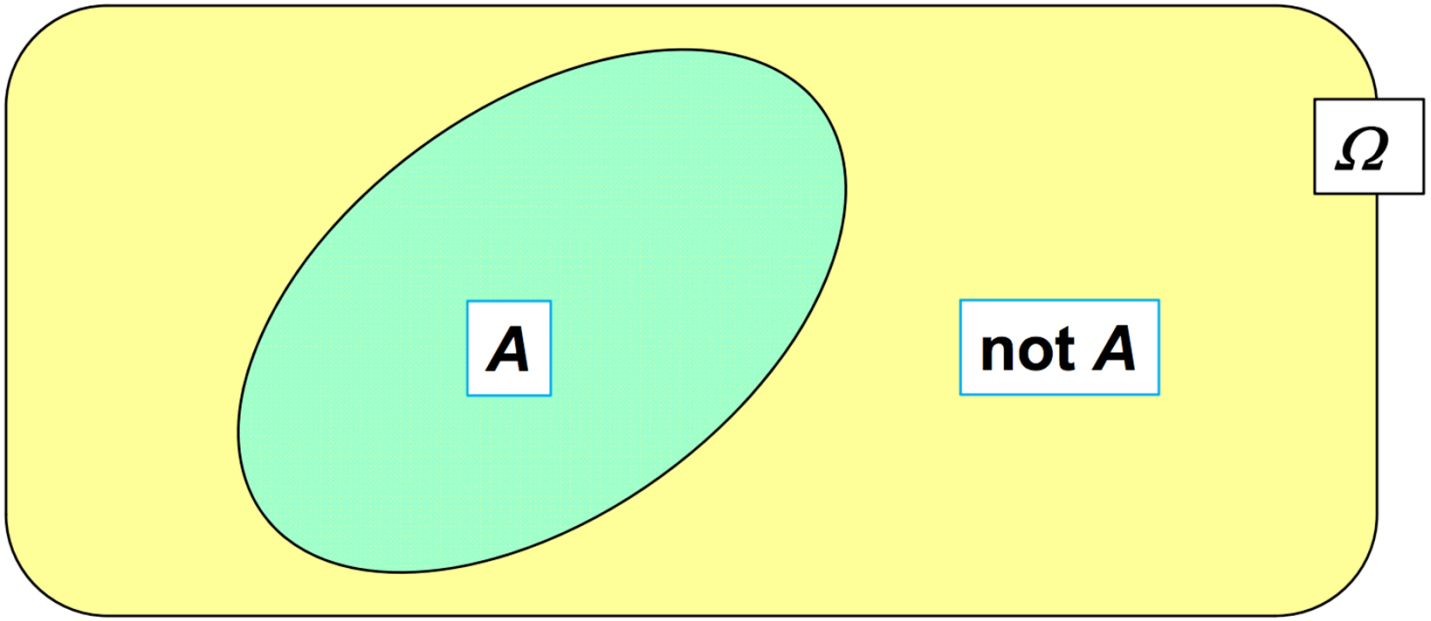
- Example: three-component Gaussian mixture in two dimensions



Complement Rule

- Given: event A, which can occur or not

$$p(\text{not } A) = 1 - p(A)$$

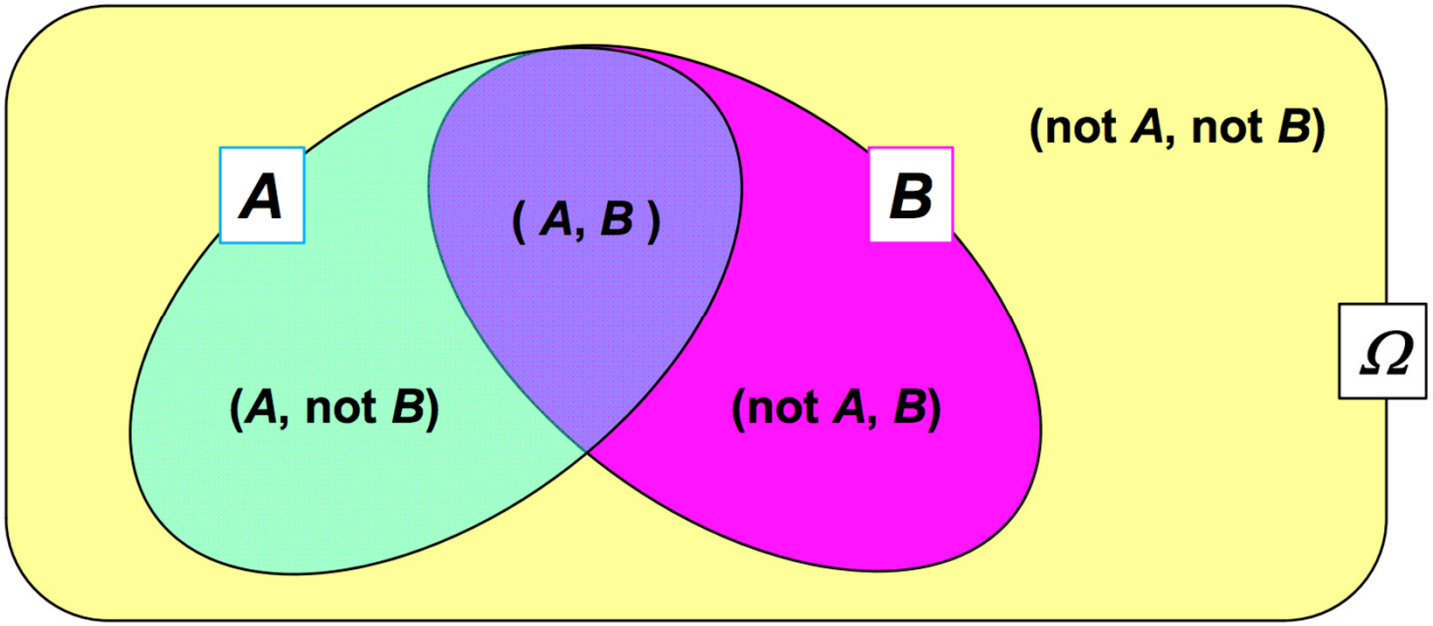


areas represent relative probabilities

Product Rule

- Given: events A and B, which can co-occur (or not)

$$p(A, B) = p(A|B) \cdot p(B)$$

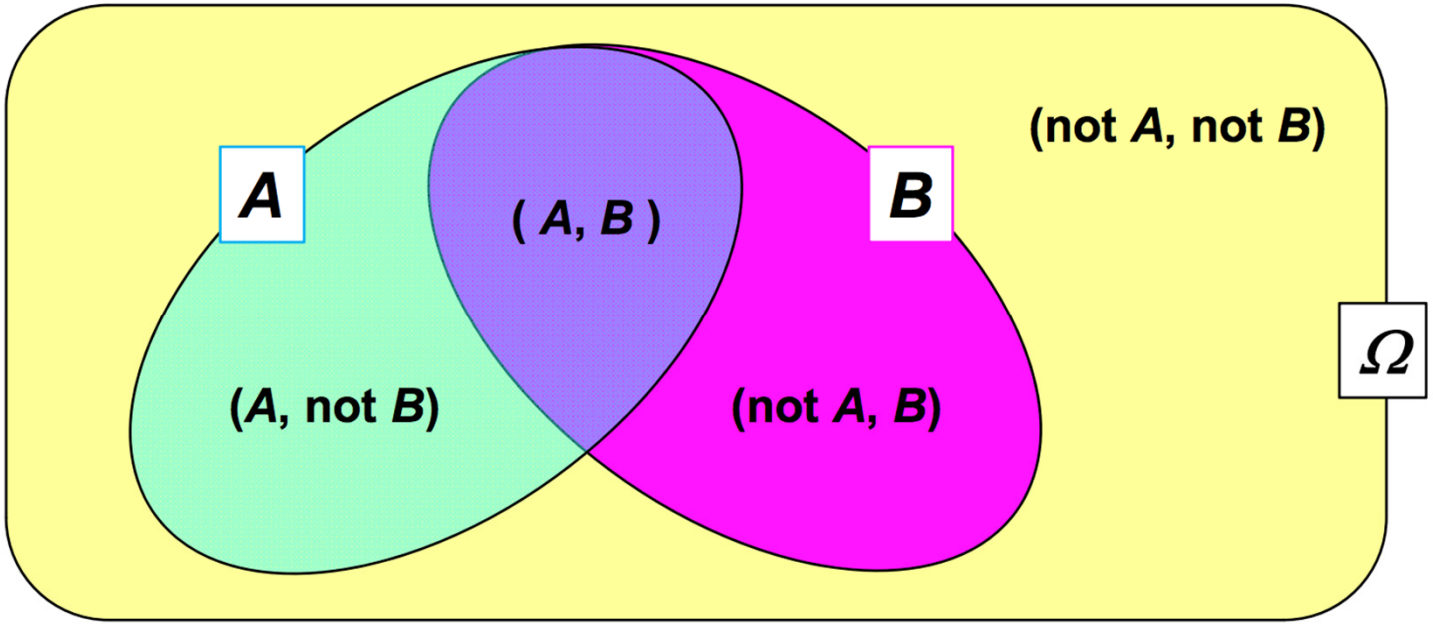


areas represent relative probabilities

Rule of Total Probability

- Given: events A and B, which can co-occur (or not)

$$p(A) = p(A, B) + p(A, \text{not } B)$$
$$= p(A|B) \cdot p(B) + p(A|\text{not } B) \cdot p(\text{not } B)$$

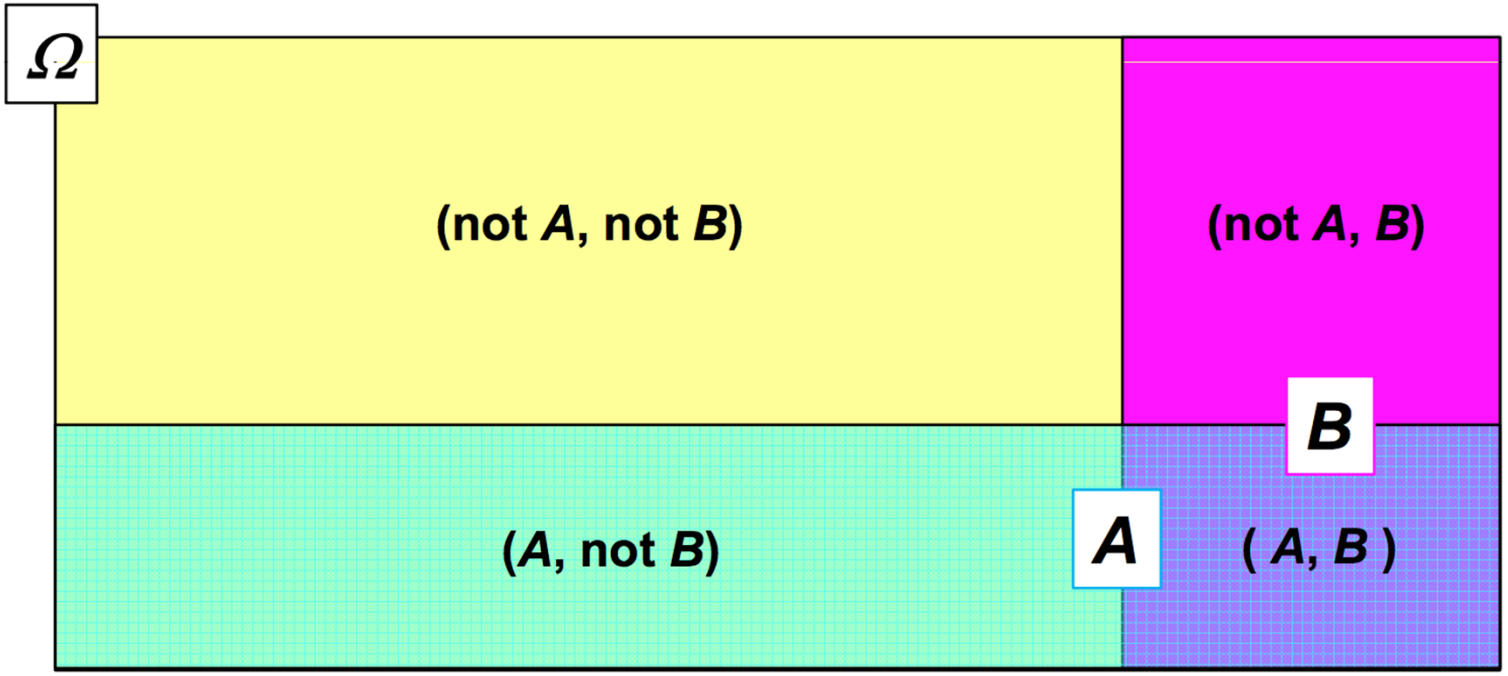


areas represent relative probabilities

Independence

- Given: events A and B, which can co-occur (or not)

$$p(A|B) = p(A) \quad \text{or} \quad p(A, B) = p(A) \cdot p(B)$$



areas represent relative probabilities

Example of Independence/Dependence

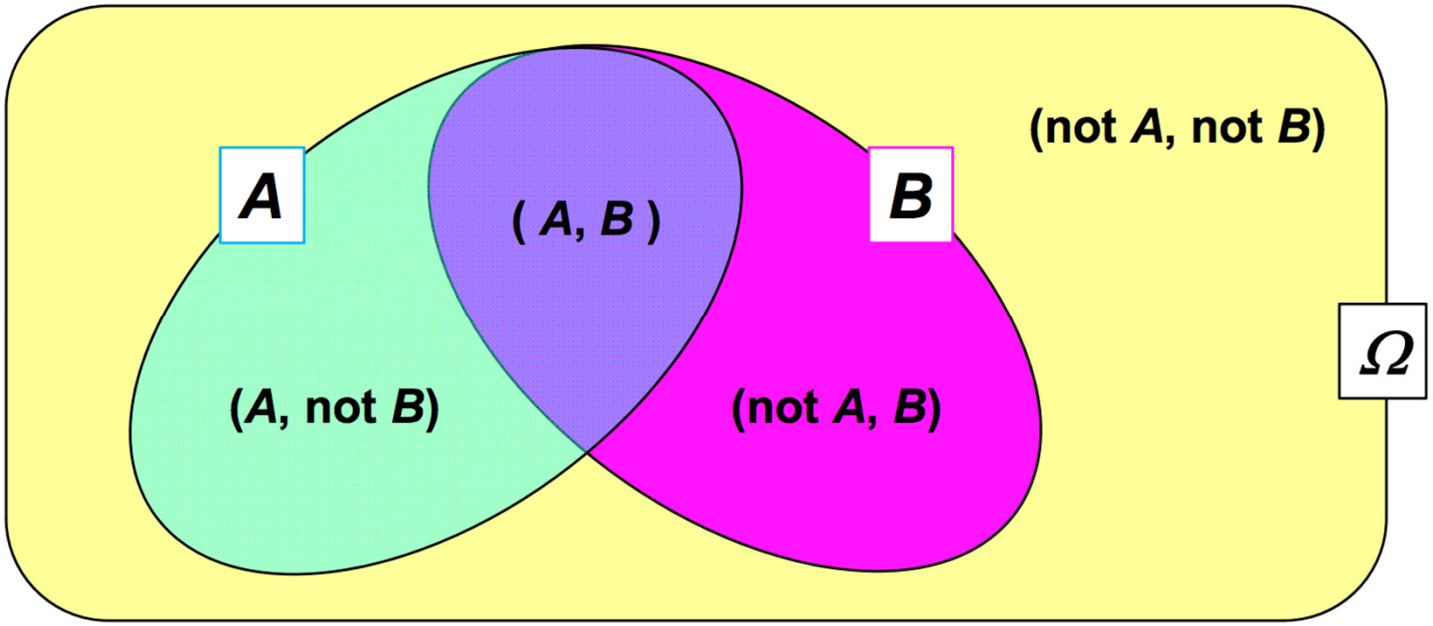
- Independence:
  - Outcomes on multiple flips of a coin
  - Height of two unrelated individuals
  - Probability of getting a king on successive draws from a deck, if card from each draw is ***replaced***

- Dependence:
  - Height of two related individuals
  - Probability of getting a king on successive draws from a deck, if card from each draw is ***not replaced***

Bayes Rule

- A way to find conditional probabilities for one variable when conditional probabilities for another variable are known.

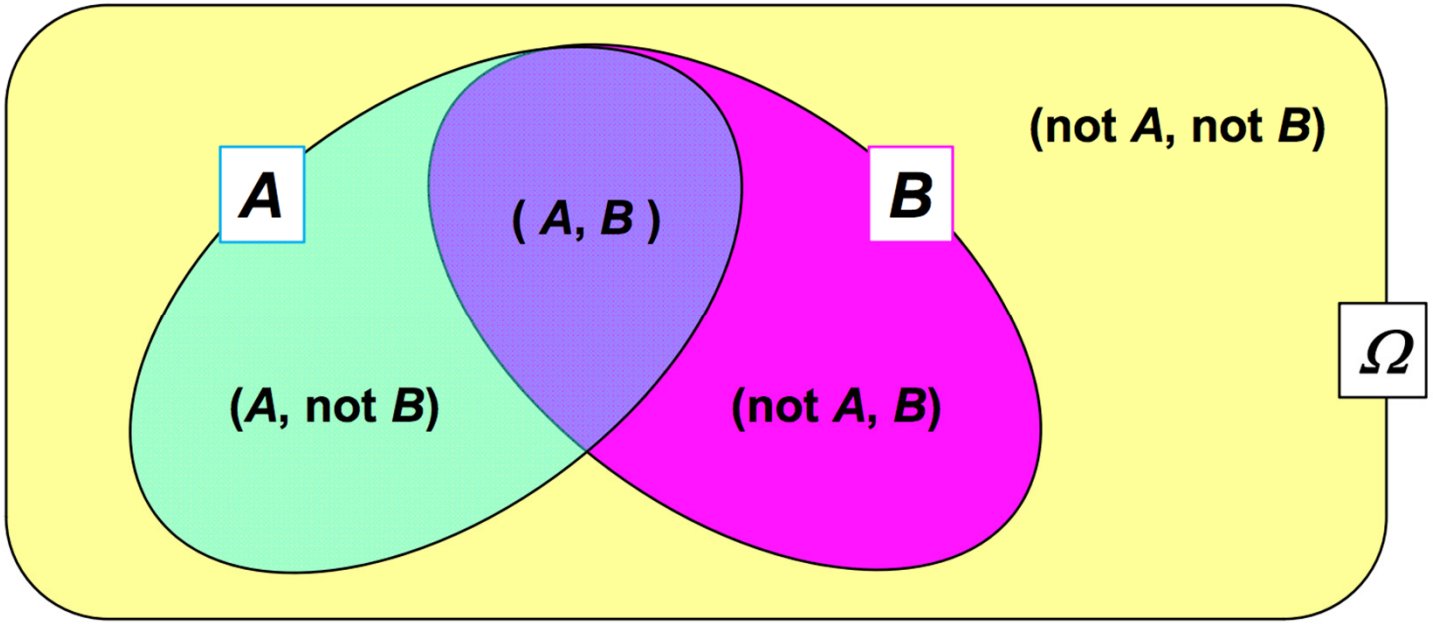
$$p(B|A) = \frac{p(A|B) \cdot p(B)}{p(A)}$$



Bayes Rule

$$p(B|A) \propto p(A|B) \cdot p(B)$$

$$\text{posterior probability} \propto \text{likelihood} \times \text{prior probability}$$



Example of Bayes Rule

- In recent years, it has rained only 5 days each year in a desert. The weatherman is forecasting rain for tomorrow. When it actually rains, the weatherman has forecast rain 90% of the time. When it doesn't rain, he has forecast rain 10% of the time. What is the probability it will rain tomorrow?
- Event A: The weatherman has forecast rain.
- Event B: It rains.
- We know:
  - $P(B) = 5/365 = 0.0137$ [It rains 5 days out of the year.]
  - $P(\text{not } B) = 1-0.0137 = 0.9863$
  - $P(A|B) = 0.9$ [When it rains, the weatherman has forecast rain 90% of the time.
  - $P(A|\text{not } B)=0.1$ [When it does not rain the weatherman has forecast rain 10% of the time.]

Example of Bayes Rule, cont'd

- We want to know *P*(B|A), the probability it will rain tomorrow, given a forecast for rain by the weatherman. The answer can be determined from Bayes rule:

$$
\begin{aligned}
p(B|A) &= p(A|B) \cdot p(B)/p(A) \\
p(A) &= p(A|B) \cdot p(B) + p(A|\text{not } B) \cdot p(\text{not } B) \\
&= 0.9 \times 0.0137 + 0.1 \times 0.9863 = 0.1110 \\
p(B|A) &= 0.9 \times 0.0137/0.1110 = 0.1111
\end{aligned}
$$

- The result seems unintuitive but is correct. Even when the weatherman predicts rain, it only rains only about 11% of the time, which is much higher than average.

Expected Value

- Given:
  - A discrete random variable *X*, with possible values $X = x_1, x_2, \ldots, x_n$
  - Probabilities $p(X = x_i)$ that *X* takes on the takes on the various values of x_i
  - A function $y_i = f(x_i)$ defined on *X*

- The expected value of *f* is the probability-weighted "average" value of $f(x_i)$:

$$\mathbb{E}(f) = \sum_i p(x_i) f(x_i)$$

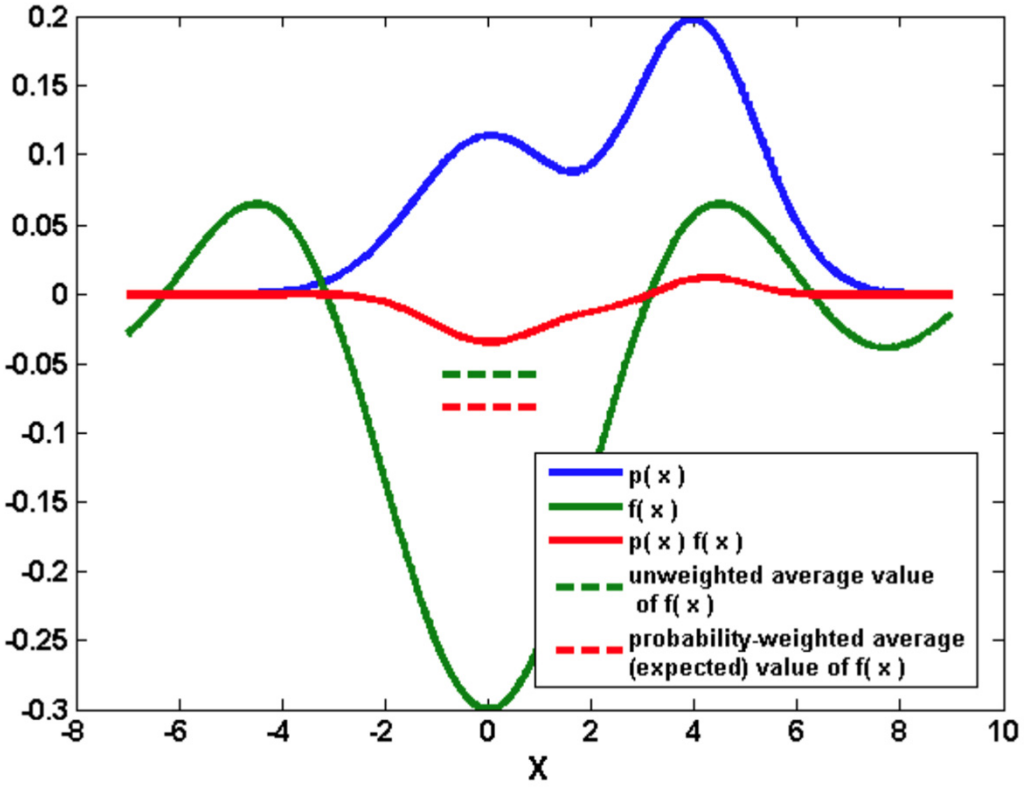
Example of Expected Value

- Process: game where one card is drawn from the deck
  - If face card, the dealer pays you $10
  - If not a face card, you pay dealer $4
- Random variable X = {face card, not face card}
  - *P*(face card) = 3/13
  - *P*(not face card) = 10/13
- Function *f*(*X*) is payout to you
  - *f*(face card) = 10
  - *f* (not face card) = -4
- Expected value of payout is

$$\mathbb{E}(f) = \sum_i p(x_i) f(x_i) = 3/13 \cdot 10 + 10/13 \cdot -4 = -0.77$$

Expected Value in Continuous Spaces

$$\mathbb{E}(f) = \int_{x=a \to b} p(x) f(x)$$



Common Forms of Expected Value (1)

- Mean μ

$$f(x_i) = x_i \quad \Longrightarrow \quad \mu = \mathbb{E}(f) = \sum_i p(x_i) x_i$$

  - Average value of $X = x_i$, taking into account probability of the various x_i
  - Most common measure of "center" of a distribution

- Estimate mean from actual samples

$$\mu = \frac{1}{N} \sum_{i=1}^{N} x_i$$

Common Forms of Expected Value (2)

- Variance σ^2

$$f(x_i) = (x_i - \mu) \quad \Longrightarrow \quad \sigma^2 = \sum_i p(x_i) \cdot (x_i - \mu)^2$$

  - Average value of squared deviation of $X = x_i$ from mean μ, taking into account probability of the various x_i
  - Most common measure of "spread" of a distribution
  - σ is the standard deviation

- Estimate variance from actual samples:

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

https://www.zhihu.com/question/20099757

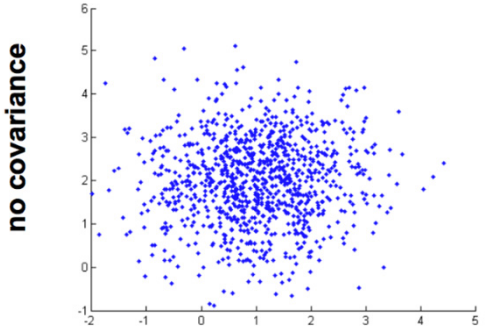
Common Forms of Expected Value (3)

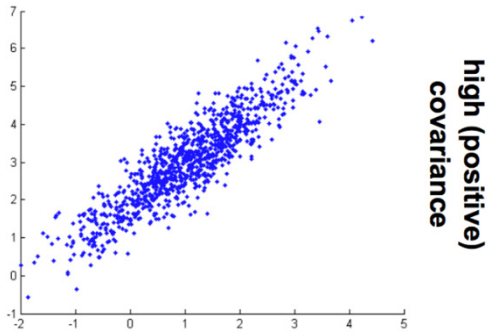
- Covariance

$$f(x_i) = (x_i - \mu_x), \quad g(y_i) = (y_i - \mu_y)$$

$$cov(x, y) = \sum_i p(x_i, y_i) \cdot (x_i - \mu_x) \cdot (y_i - \mu_y)$$

  - Measures tendency for *x* and *y* to deviate from their means in same (or opposite) directions at same time





- Estimate covariance from actual samples

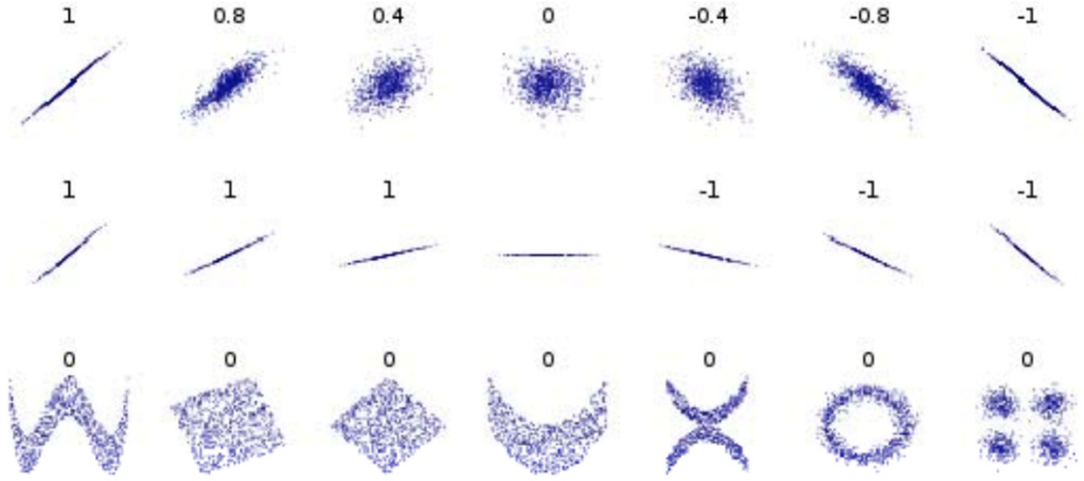
$$cov(x, y) = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \mu_x)(y_i - \mu_y)$$

Correlation

- Pearson's correlation coefficient is covariance normalized by the standard deviations of the two variables

$$corr(x, y) = \frac{cov(x, y)}{\sigma_x \sigma_y}$$

  - Always lies in range -1 to 1
  - Only reflects *linear dependence* between variables



Linear dependence with noise

Linear dependence without noise

Various nonlinear dependencies

Estimation of Parameters

- Suppose we have random variables $X_1, \ldots, X_n$ and corresponding observations $x_1, \ldots, x_n$.

- We prescribe a parametric model and fit the parameters of the model to the data.

- How do we choose the values of the parameters?

Maximum Likelihood Estimation(MLE)

- The basic idea of MLE is to maximize the probability of the data we have seen.

$$\hat{\theta}_{MLE} = \arg\max_{\theta} \mathcal{L}(\theta)$$

- where L is the likelihood function

$$\mathcal{L}(\theta) = p(x_1, \ldots, x_n; \theta)$$

- Assume that $X_1, \ldots, X_n$ are i.i.d, then we have

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

- Take log on both sides, we get log-likelihood

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

Example

- X_i are independent Bernoulli random variables with unknown parameter ϑ.

$$f(x_i;\theta) = \theta^{x_i}(1-\theta)^{1-x_i}$$

$$\mathcal{L}(\theta) = \prod_{i=1}^{n} f(x_i;\theta) = \theta^{\sum x_i}(1-\theta)^{n-\sum x_i}$$

$$\log \mathcal{L}(\theta) = (\sum x_i)\log\theta + (n-\sum x_i)\log(1-\theta)$$

$$\frac{\partial \log \mathcal{L}(\theta)}{\partial\theta} = 0 \Rightarrow \hat{\theta} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Maximum A Posteriori Estimation (MAP)

- We assume that the parameters are a random variable, and we specify a prior distribution p(θ).

- Employ Bayes' rule to compute the posterior distribution

$$p(\theta|x_1,\ldots,x_n) \propto p(\theta)p(x_1,\ldots,x_n|\theta)$$

- Estimate parameter θ by maximizing the posterior

$$\hat{\theta}_{MAP} = \arg\max_{\theta} p(\theta)p(x_1,\ldots,x_n|\theta)$$

$$\hat{\theta}_{MAP} = \arg\max_{\theta} \log p(\theta) + \sum_{i=1}^{n} \log(x_i|\theta)$$

Example

- X_i are independent Bernoulli random variables with unknown parameter ϑ. Assume that ϑ satisfies normal distribution.
- Normal distribution:

$$\mathcal{N}(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- Maximize:

$$\arg\max_{\theta} -\frac{(\theta-\mu)^2}{2\sigma^2} + (\sum x_i)\log\theta + (n - \sum x_i)\log(1-\theta)$$

Comparison between MLE and MAP

- MLE: For which ϑ is $X_1, \ldots, X_n$ most likely?
- MAP: Which ϑ maximizes $p(\vartheta | X_1, \ldots, X_n)$ with prior $p(\vartheta)$?
- The prior can be regard as regularization - to reduce the overfitting.

Example

- Flip a unfair coin 10 times. The result is HHTTHHHHHT

$$f(x_i;\theta) = \theta^{x_i}(1-\theta)^{1-x_i}$$

- x_i = 1 if the result is head.

- MLE estimates ϑ = 0.7

- Assume the prior of ϑ is N(0.5,0.01), MAP estimates ϑ=0.558

What happens if we have more data?

- Flip the unfair coins 100 times, the result is 70 heads and 30 tails.
  - The result of MLE does not change, $\vartheta = 0.7$
  - The estimation of MAP becomes $\vartheta = 0.663$

- Flip the unfair coins 1000 times, the result is 700 heads and 300 tails.
  - The result of MLE does not change, $\vartheta = 0.7$
  - The estimation of MAP becomes $\vartheta = 0.696$

Unbiased Estimators

- An estimator of a parameter is unbiased if the expected value of the estimate is the same as the true value of the parameters.
- Assume X_i is a random variable with mean μ and variance σ^2

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$\mathbb{E}(\bar{X}) = \mathbb{E}(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(X_i) = \frac{1}{n}n\mu = \mu$$

- $\bar{X}$ is unbiased estimation

Estimator of Variance

- Assume X_i is a random variable with mean μ and variance σ^2
- Is $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$ unbiased?

$$\begin{aligned}
\hat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(X_i^2 - 2X_i\bar{X} + \bar{X}^2) \\
&= \frac{1}{n}\sum_{i=1}^{n}X_i^2 - 2\bar{X}\frac{1}{n}\sum_{i=1}^{n}X_i + \frac{1}{n}n\bar{X}^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}X_i^2 - \bar{X}^2
\end{aligned}$$

Estimator of Variance

$$
\begin{aligned}
\mathbb{E}(\hat{\sigma}^2) =& \mathbb{E}[\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2] \\
=& \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}(X_i^2) - \mathbb{E}(\bar{X}^2) \\
=& \frac{1}{n}\sum_{i=1}^{n} (\sigma^2 + \mu^2) - (\sigma^2/n + \mu^2) \\
=& \sigma^2 - \sigma^2/n \\
=& \frac{(n-1)\sigma^2}{n} \neq \sigma^2
\end{aligned}
$$

- where we use

$$var(X) = \sigma^2 = E(X^2) - \mu^2 \text{ , } \quad var(\bar{X}) = \sigma^2/n = E(\bar{X}^2) - \mu^2$$

Estimator of Variance

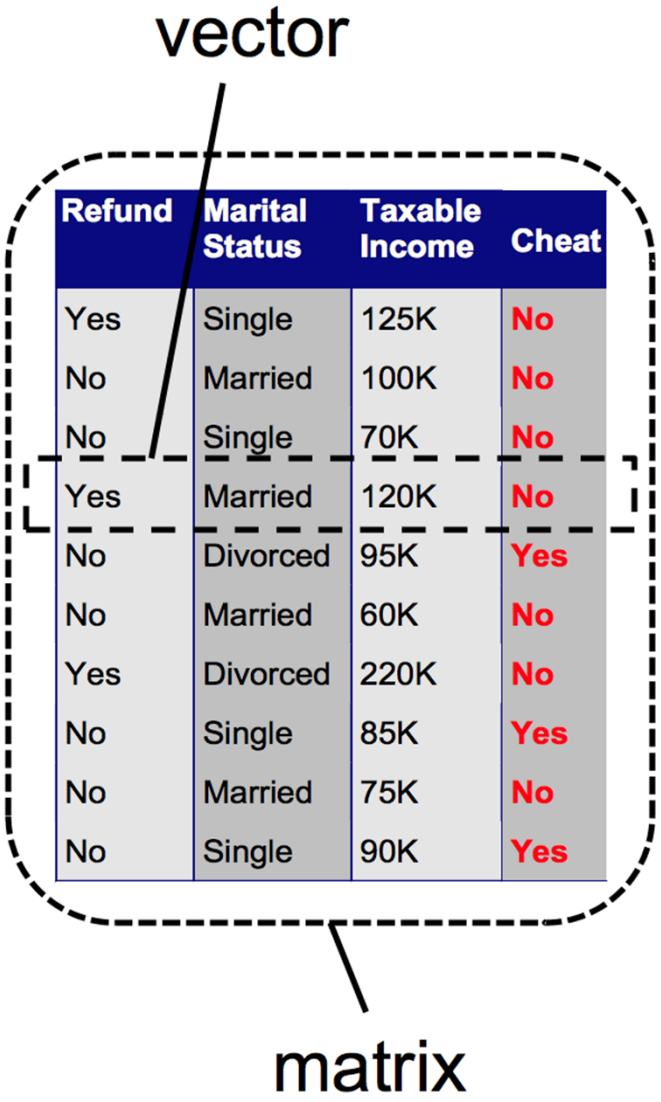
$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

$$\mathbb{E}(\hat{\sigma}^2) = \sigma^2$$

- $\hat{\sigma}^2$ is a unbiased estimation

Linear Algebra Applications

- Why vectors and matrices?
  - Most common form of data organization for machine vector organization for machine learning is a 2D array, where
    - rows represent samples
    - columns represent attributes
  - Natural to think of each sample as a vector of attributes, and whole array as a matrix



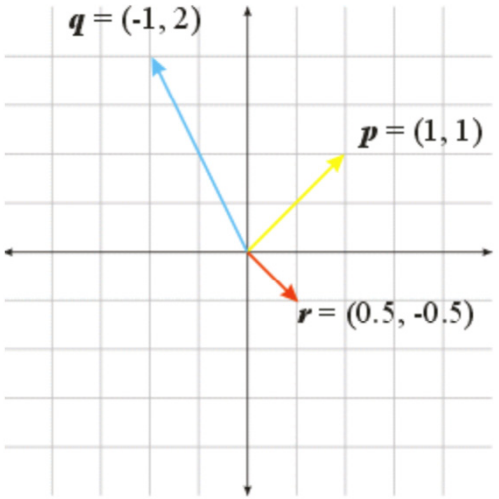
Vectors

- Definition: an *n*-tuple of values
  - *n* referred to as the *dimension* of the vector
- Can be written in column form or row form

$$x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \qquad x^\top = \begin{pmatrix} x_1 & \cdots & x_n \end{pmatrix}$$

$\top$ means "transpose"

- Can think of a vector as
  - a point in space *or*
  - a directed line segment with a magnitude and direction



Vector Arithmetic

- Addition of two vectors
  - add corresponding elements

$$\mathbf{z} = \mathbf{x} + \mathbf{y} = \begin{pmatrix} x_1 + y_1 & \cdots & x_n + y_n \end{pmatrix}^\top$$

- Scalar multiplication of a vector
  - multiply each element by scalar

$$\mathbf{y} = a\mathbf{x} = \begin{pmatrix} ax_1 & \cdots & ax_n \end{pmatrix}^\top$$

- Dot product of two vectors
  - Multiply corresponding elements, then add products

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

  - Result is a scalar

Vector Norms

- A norm is a function $\|\cdot\|$ that satisfies:
  - $\|\mathbf{x}\| \geq 0$ with equality if and only if $\mathbf{x} = 0$
  - $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$
  - $\|a\mathbf{x}\| = |a|\|\mathbf{x}\|$
- 2-norm of vectors

$$\|\mathbf{x}\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2}$$

- Cauchy-Schwarz inequality

$$\mathbf{x} \cdot \mathbf{y} \leq \|\mathbf{x}\|_2 \|\mathbf{y}\|_2$$

Matrices

- Definition: an $m \times n$ two-dimensional array of values
  - m rows
  - n columns
- Matrix referenced by two-element subscript
  - first element in subscript is row
  - Second element in subscript is column
  - example: $\mathbf{A}_{24}$ or a_{24} is element in second row, fourth column of ***A***

$$\mathbf{A} = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{pmatrix}$$

Matrices

- A vector can be regarded as special case of a matrix, where one of matrix dimensions is 1.
- Matrix transpose (denoted $^\top$)
  - swap columns and rows
  - $m \times n$ matrix becomes n x m matrix
  - example:

$$\mathbf{A} = \begin{pmatrix} 2 & 7 & -1 & 0 & 3 \\ 4 & 6 & -3 & 1 & 8 \end{pmatrix} \qquad \mathbf{A}^\top = \begin{pmatrix} 2 & 4 \\ 7 & 6 \\ -1 & -3 \\ 0 & 1 \\ 3 & 8 \end{pmatrix}$$

Matrix Arithmetic

- Addition of two matrices
  - matrices must be same size
  - add corresponding elements:

$$c_{ij} = a_{ij} + b_{ij}$$

  - result is a matrix of same size

$$\mathbf{C} = \mathbf{A} + \mathbf{B} = \begin{pmatrix} a_{11} + b_{11} & \cdots & a_{1n} + b_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} + b_{m1} & \cdots & a_{mn} + b_{mn} \end{pmatrix}$$

- Scalar multiplication of a matrix
  - multiply each element by scalar:

$$b_{ij} = d \cdot a_{ij}$$

  - result is a matrix of same size

$$\mathbf{B} = d \cdot \mathbf{A} = \begin{pmatrix} d \cdot a_{11} & \cdots & d \cdot a_{1n} \\ \vdots & \ddots & \vdots \\ d \cdot a_{m1} & \cdots & d \cdot a_{mn} \end{pmatrix}$$

Matrix Arithmetic

- Matrix-matrix multiplication
  - the column dimension of the previous matrix must match the row dimension of the following matrix

$$\mathbf{C}_{p\times n} = \mathbf{A}_{p\times m}\mathbf{B}_{m\times n} \qquad c_{ij} = \sum_{k=1}^{m} a_{ik}b_{kj}$$

- Multiplication is associative

$$\mathbf{A}\cdot(\mathbf{B}\cdot\mathbf{C}) = (\mathbf{A}\cdot\mathbf{B})\cdot\mathbf{C}$$

- Multiplication is not commutative

$$\mathbf{A}\cdot\mathbf{B} \neq \mathbf{B}\cdot\mathbf{A}$$

- Transposition rule

$$(\mathbf{A}\cdot\mathbf{B})^{\top} = \mathbf{B}^{\top}\cdot\mathbf{A}^{\top}$$

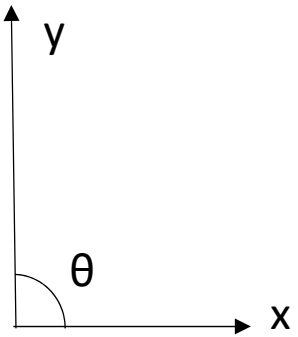
Orthogonal Vectors

- Alternative form of dot product:

$$\mathbf{x}^\top \mathbf{y} = \|\mathbf{x}\| \|\mathbf{y}\| \cos\theta$$

- A pair of vector x and y are *orthogonal* if

$$\mathbf{x}^\top \mathbf{y} = 0$$



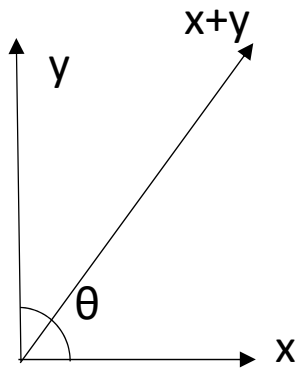
- A set of vectors S is orthogonal if its elements are pairwise orthogonal
  - for $\mathbf{x}, \mathbf{y} \in S, \mathbf{x} \neq \mathbf{y} \Rightarrow \mathbf{x}^\top \mathbf{y} = 0$
- A set of vectors S is orthonormal if it is orthogonal and, every $\mathbf{x} \in S$ has $\|\mathbf{x}\| = 1$

Orthogonal Vectors

- Pythagorean theorem:
  - If x and y are orthogonal, then

    $\|\mathbf{x}+\mathbf{y}\|^2 = \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2$
  - Proof: we know $\mathbf{x}^\top \mathbf{y} = 0$, then

$$\begin{aligned}\|\mathbf{x}+\mathbf{y}\|^2 =& (\mathbf{x}+\mathbf{y})^\top(\mathbf{x}+\mathbf{y}) \\ =& \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 + \mathbf{x}^\top\mathbf{y} + \mathbf{y}^\top\mathbf{x} \\ =& \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2\end{aligned}$$



- General case: a set of vectors is orthogonal

$$\|\sum_{i=1}^{n}\mathbf{x}_i\|^2 = \sum_{i=1}^{n}\|\mathbf{x}_i\|^2$$

Orthogonal Matrices

- A square matrix $\mathbf{Q} \in \mathbb{R}^{n \times n}$ is orthogonal if

$$\mathbf{Q}^\top \mathbf{Q} = \mathbf{I} \quad \text{i.e., } \mathbf{Q}^\top = \mathbf{Q}^{-1}$$

- In terms of the columns of Q, the product can be written as

$$\begin{pmatrix} \mathbf{q}_1^\top \\ \mathbf{q}_2^\top \\ \vdots \\ \mathbf{q}_n^\top \end{pmatrix} \begin{pmatrix} \mathbf{q}_1 & \mathbf{q}_2 & \cdots \mathbf{q}_n \end{pmatrix} = \begin{pmatrix} 1 & & & \\ & 1 & & \\ & & \ddots & \\ & & & 1 \end{pmatrix}$$

Orthogonal Matrices

$$\begin{pmatrix} \mathbf{q}_1^\top \\ \mathbf{q}_2^\top \\ \vdots \\ \mathbf{q}_n^\top \end{pmatrix} \begin{pmatrix} \mathbf{q}_1 & \mathbf{q}_2 & \cdots \mathbf{q}_n \end{pmatrix} = \begin{pmatrix} 1 & & & \\ & 1 & & \\ & & \ddots & \\ & & & 1 \end{pmatrix}$$

$$\mathbf{q}_i^\top \mathbf{q}_j = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$$

- The columns of orthogonal matrix Q form an orthonormal basis

Orthogonal matrices

- The processes of multiplication by an orthogonal matrices preserves geometric structure
  - Dot products are preserved

    $$(\mathbf{Q}\mathbf{x}) \cdot (\mathbf{Q}\mathbf{y}) = \mathbf{x} \cdot \mathbf{y}$$

    $$(\mathbf{Q}\mathbf{x})^\top (\mathbf{Q}\mathbf{y}) = \mathbf{x}^\top \mathbf{Q}^\top \mathbf{Q}\mathbf{y} = \mathbf{x}^\top \mathbf{y}$$

  - Lengths of vectors are preserved

    $$\|\mathbf{Q}\mathbf{x}\| = \|\mathbf{x}\|$$

  - Angles between vectors are preserved

    $$\cos\theta = \frac{(\mathbf{Q}\mathbf{x})^\top (\mathbf{Q}\mathbf{y})}{\|\mathbf{Q}\mathbf{x}\|\|\mathbf{Q}\mathbf{y}\|} = \frac{\mathbf{x}^\top \mathbf{y}}{\|\mathbf{x}\|\|\mathbf{y}\|}$$

Tall Matrices with Orthonormal Columns

- Suppose matrix $\mathbf{Q} \in \mathbb{R}^{m \times n}$ is tall (m>n) and has orthogonal columns

- Properties:

$$\mathbf{Q}^\top \mathbf{Q} = \mathbf{I}$$

$$\mathbf{Q}\mathbf{Q}^\top \neq \mathbf{I}$$

Matrix Norms

- Vector p-norms:

$$\|\mathbf{x}\|_p = \left(|x_1|^p + |x_2|^p + \cdots + |x_n|^p\right)^{1/p}$$

- Matrix p-norms:

$$\|\mathbf{A}\|_p = \max_{\mathbf{x}\neq 0} \frac{\|\mathbf{A}\mathbf{x}\|_p}{\|\mathbf{x}\|_p}$$

- Example: 1-norm $\|\mathbf{A}\|_1 = \max_j \sum_{i=1}^{n} |a_{ij}|$

- Matrix norms which induced by vector norm are called operator norm.

General Matrix Norms

- A norm is a function $\|\cdot\|$ that satisfies:
  - $\|\mathbf{A}\| \geq 0$ with equality if and only if $\mathbf{A} = 0$
  - $\|\mathbf{A} + \mathbf{B}\| \leq \|\mathbf{A}\| + \|\mathbf{B}\|$
  - $\|\alpha\mathbf{A}\| = |\alpha|\|\mathbf{A}\|$

- Frobenius norm
  - The Frobenius norm of $\mathbf{A} \in \mathbb{R}^{m \times n}$ is:

$$\|\mathbf{A}\|_F = (\sum_{i=1}^{m}\sum_{j=1}^{n}|a_{ij}|^2)^{1/2}$$

Some Properties

- $\|\mathbf{A}\|_F^2 = trace(\mathbf{A}^\top \mathbf{A})$
- $\|\mathbf{AB}\|_F = \|\mathbf{A}\|_F \|\mathbf{B}\|_F$
- Invariance under orthogonal Multiplication

  $$\|\mathbf{QA}\|_2 = \|\mathbf{A}\|_2 \qquad \|\mathbf{QA}\|_F = \|\mathbf{A}\|_F$$

  - Q is an orthogonal matrix

Eigenvalue Decomposition

- For a square matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$, we say that a nonzero vector $\mathbf{x} \in \mathbb{R}^n$ is an eigenvector of A corresponding to eigenvalue λ if

$$\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$$

- An eigenvalue decomposition of a square matrix A is

$$\mathbf{A} = \mathbf{X}\mathbf{\Lambda}\mathbf{X}^{-1}$$

- X is nonsingular and consists of eigenvectors of A
- $\mathbf{\Lambda}$ is a diagonal matrix with the eigenvalues of A on its diagonal.

Eigenvalue Decomposition

- Not all matrix has eigenvalue decomposition.
  - A matrix has eigenvalue decomposition if and only if it is diagonalizable.

- Real symmetric matrix has real eigenvalues.
- It's eigenvalue decomposition is the following form:

$$\mathbf{A} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^{\top}$$

- Q is orthogonal matrix.

Singular Value Decomposition(SVD)

- every matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ has an SVD as follows:

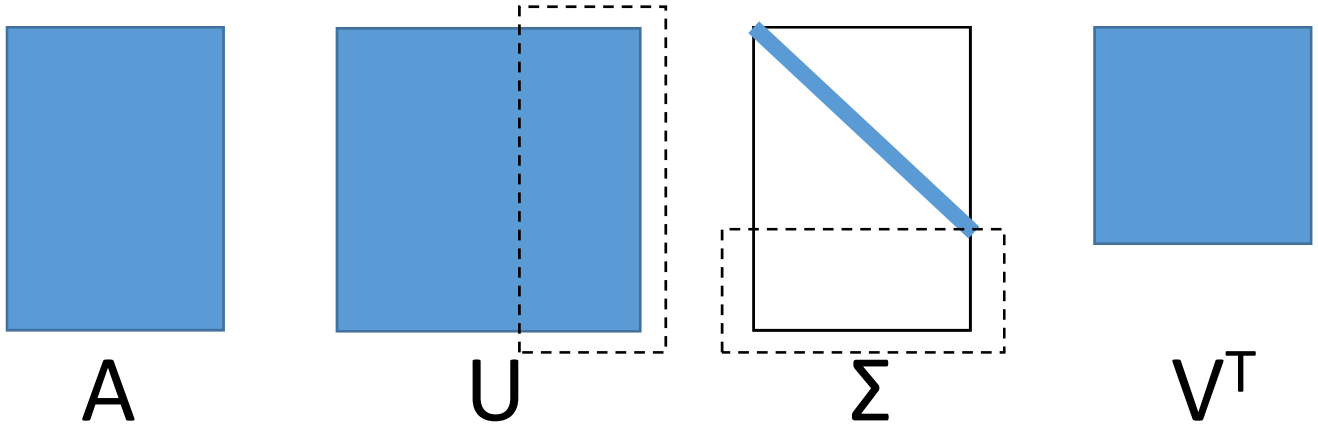
$$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\top}$$

- $\mathbf{U} \in \mathbb{R}^{m \times m}$ and $\mathbf{V} \in \mathbb{R}^{n \times n}$ are orthogonal matrices
- $\mathbf{\Sigma} \in \mathbb{R}^{m \times n}$ is a diagonal matrix with the singular values of A on its diagonal.
- Suppose the rank of A is r, the singular values of A is

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r \geq \sigma_{r+1} = \ldots \sigma_{min(m,n)} = 0$$

Full SVD and Reduced SVD

- Assume that $m \geq n$
  - Full SVD: *U* is $m \times m$ matrix, *Σ* is $m \times n$ matrix.
  - Reduced SVD: *U* is $m \times n$ matrix, *Σ* is $n \times n$ matrix.



Properties via the SVD

- The nonzero singular values of *A* are the square roots of the nonzero eigenvalues of A^TA.

$$\mathbf{A}^\top\mathbf{A} = (\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top)^\top(\mathbf{U}\mathbf{\Sigma}\mathbf{V}) = \mathbf{V}\mathbf{\Sigma}^\top\mathbf{U}^\top\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top = \mathbf{V}(\mathbf{\Sigma}^\top\mathbf{\Sigma})\mathbf{V}^\top$$

- If $A=A^T$, then the singular values of *A* are the absolute values of the eigenvalues of *A*.

$$\mathbf{A} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^\top = \mathbf{Q}|\mathbf{\Lambda}|sign(\mathbf{\Lambda})\mathbf{Q}^\top$$

Properties via the SVD

- $\|\mathbf{A}\|_2 = \sigma_1$

$$\|\mathbf{A}\|_F = \sqrt{\sigma_1^2 + \sigma_2^2 + \cdots + \sigma_r^2}$$

- Denote $\mathbf{U} = \begin{pmatrix} \mathbf{u}_1 & \mathbf{u}_2 & \dots & \mathbf{u}_m \end{pmatrix}$

$$\mathbf{V} = \begin{pmatrix} \mathbf{v}_1 & \mathbf{v}_2 & \dots & \mathbf{v}_n \end{pmatrix}$$

$$\mathbf{A} = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top$$

Low-rank Approximation

- $\mathbf{A} = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top$
- For any $0 < k < r$, define $\mathbf{A}_k = \sum_{i=1}^{k} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top$
- Eckart-Young Theorem:

$$\min_{rank(\mathbf{B}) \leq k} \|\mathbf{A} - \mathbf{B}\|_2 = \|\mathbf{A} - \mathbf{A}_k\|_2 = \sigma_{k+1}$$

$$\min_{rank(\mathbf{B}) \leq k} \|\mathbf{A} - \mathbf{B}\|_F = \|\mathbf{A} - \mathbf{A}_k\|_F = \sqrt{\sigma_{k+1}^2 + \cdots + \sigma_r^2}$$

- A_k is the best rank-k approximation of A.

Example

- Image Compression

original
(390*390)

k=10 k=20 k=50

Positive (semi-)definite matrices

- A symmetric matrix A is positive semi-definite(PSD) if for all $\mathbf{x} \in \mathbb{R}^n, \mathbf{x}^\top \mathbf{A}\mathbf{x} \geq 0$

- A symmetric matrix A is positive definite(PD) if for all nonzero $\mathbf{x} \in \mathbb{R}^n, \mathbf{x}^\top \mathbf{A}\mathbf{x} > 0$

- Positive definiteness is a strictly stronger property than positive semi-definiteness.

- Notation: $\mathbf{A} \succeq 0$ if A is PSD, $\mathbf{A} \succ 0$ if A is PD

Properties of PSD matrices

- A symmetric matrix is PSD if and only if all of its eigenvalues are nonnegative.
  - Proof: let *x* be an eigenvector of *A* with eigenvalue *λ*.

$$0 \leq \mathbf{x}^\top \mathbf{A}\mathbf{x} = \mathbf{x}^\top (\lambda \mathbf{x}) = \lambda \mathbf{x}^\top \mathbf{x} = \lambda \|\mathbf{x}\|_2^2$$

- The eigenvalue decomposition of a symmetric PSD matrix is equivalent to its singular value decomposition.

Properties of PSD matrices

- For a symmetric PSD matrix *A*, there exists a unique symmetric PSD matrix *B* such that

$$\mathbf{B}^2 = \mathbf{A}$$

- Proof: We only show the existence of *B*
  - Suppose the eigenvalue decomposition is

$$\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top$$

  - Then, we can get *B*:

$$\mathbf{B} = \mathbf{U}\mathbf{\Lambda}^{\frac{1}{2}}\mathbf{U}^\top$$

$$\mathbf{B}^2 = \mathbf{U}\mathbf{\Lambda}^{\frac{1}{2}}\mathbf{U}^\top\mathbf{U}\mathbf{\Lambda}^{\frac{1}{2}}\mathbf{U}^\top = \mathbf{U}\mathbf{\Lambda}^{\frac{1}{2}}\mathbf{\Lambda}^{\frac{1}{2}}\mathbf{U}^\top = \mathbf{A}$$

Convex Optimization

Gradient and Hessian

- The gradient of $f : \mathbb{R}^d \to \mathbb{R}$ is

$$\nabla f = \begin{pmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \vdots \\ \frac{\partial f}{\partial x_d} \end{pmatrix}$$

- The Hessian of $f : \mathbb{R}^d \to \mathbb{R}$ is

$$\nabla^2 f = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1^2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_d} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_d \partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_d^2} \end{pmatrix}$$

What is Optimization?

- Finding the minimizer of a function subject to constraints:

$$\begin{aligned} \min_{x} \; & f(x) \\ s.t. \; & g_i(x) \leq 0, i = 1, 2, \ldots, m \\ & h_j(x) = 0, j = 1, 2, \ldots, n \end{aligned}$$

Why optimization?

- Optimization is the key of many machine learning algorithms
  - Linear regression:

$$\min_w \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2$$

  - Logistic regression:

$$\min_w \sum_{i=1}^{n} \log(1 + \exp(-\mathbf{y}_i \mathbf{x}_i^\top \mathbf{w}))$$

  - Support vector machine:

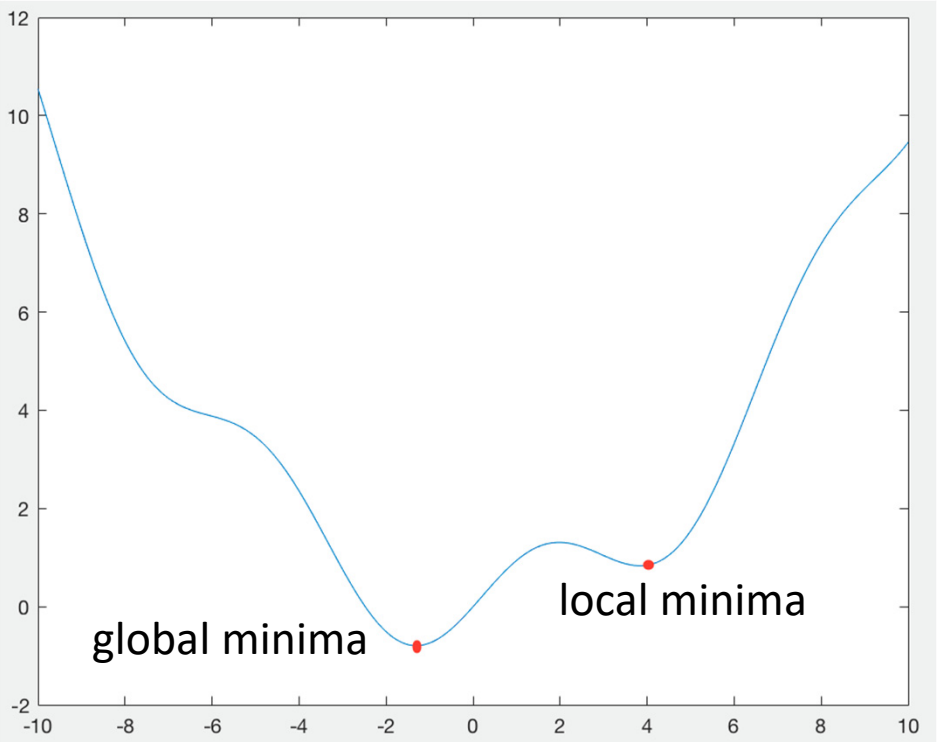
$$\min \|\mathbf{w}\|^2 + C \sum_{i=1}^{n} \xi_i$$

$$s.t.\ \xi_i \geq 1 - \mathbf{y}_i \mathbf{x}_i^\top \mathbf{w}$$

$$\xi_i \geq 0$$

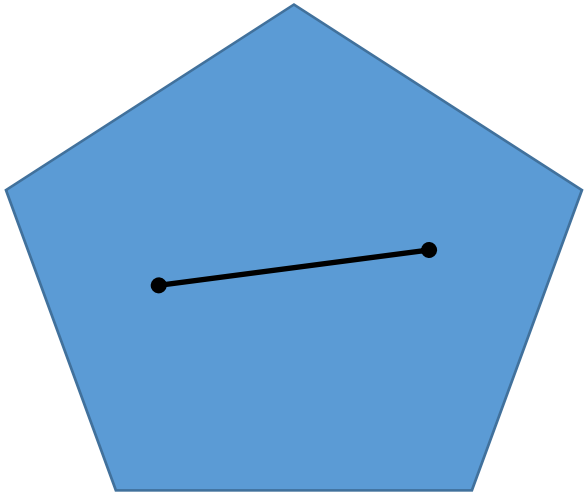
Local Minima and Global Minima

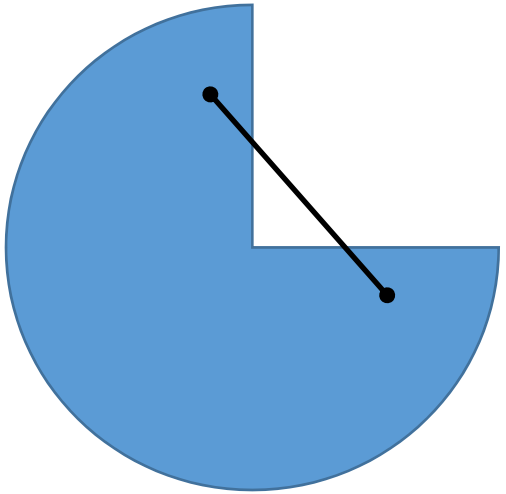
- Local minima
  - a solution that is optimal within a neighboring set
- Global minima
  - the optimal solution among all possible solutions



Convex Set

- A set $C \subseteq \mathbb{R}^n$ is convex if for any $x, y \in C$,
$$tx + (1-t)y \in C \text{ for all } t \in [0,1]$$





Example of Convex Sets

- Trivial: empty set, line, point, etc.
- Norm ball: $\{x : \|x\| \leq r\}$, for given radius *r*
- Affine space: $\{x : Ax = b\}$, given *A*, *b*
- Polyhedron: $\{x : Ax \leq b\}$, where inequality ≤ is interpreted component-wise.

Operations preserving convexity

- Intersection: the intersection of convex sets is convex

- Affine images: if $f(x) = Ax + b$ and C is convex, then

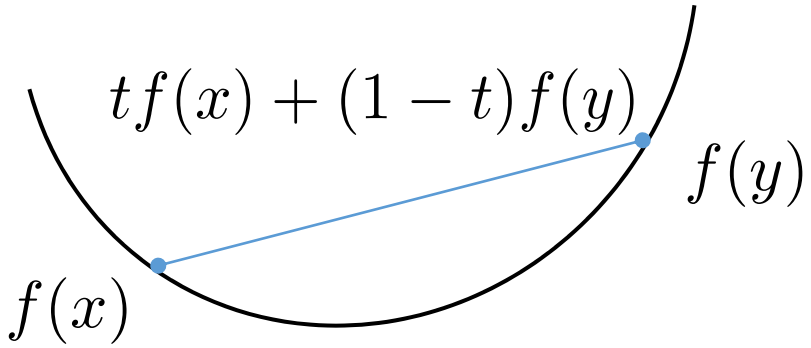
$$f(C) = \{Ax + b : x \in C\}$$

  is convex

Convex functions

- A function $f : \mathbb{R}^n \to \mathbb{R}$ is convex if for $x, y \in dom f$,

$$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y), \text{ for all } t \in [0,1]$$



Strictly Convex and Strongly Convex

- Strictly convex:
  - $$f(tx + (1-t)y) < tf(x) + (1-t)f(y)$$
    $$\text{for } x \neq y \text{ and } 0 < t < 1$$
  - Linear function is not strictly convex.

- Strongly convex:
  - For $m > 0 : f(x) - \frac{m}{2}\|x\|^2$ is convex

- Strong convexity ⇒ strict convexity ⇒ convexity

Example of Convex Functions

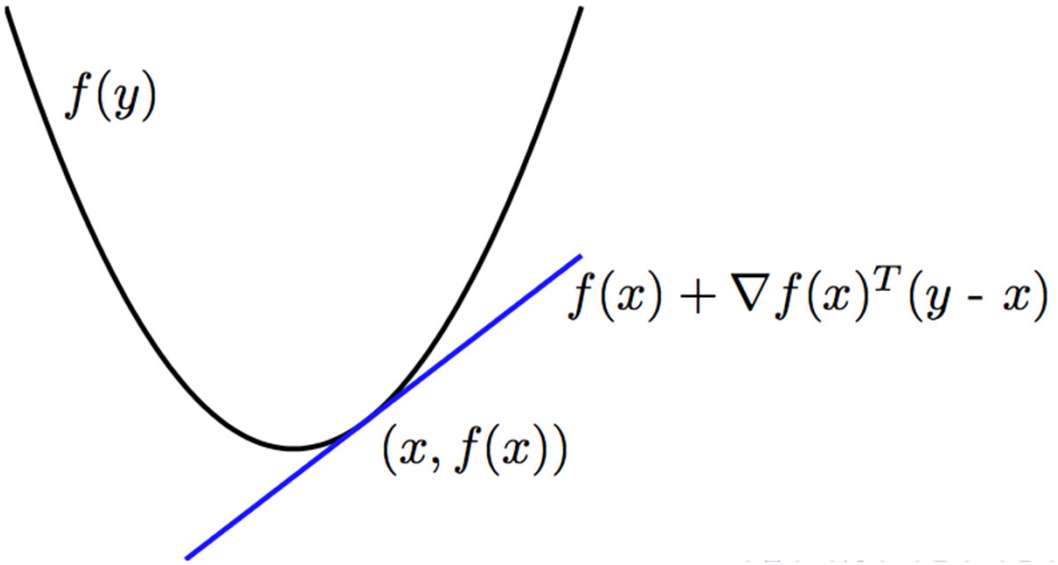
- Exponential function: e^{ax}
- logarithmic function log(*x*) is concave
- Affine function: $a^\top x + b$
- Quadratic function: $x^\top Q x + b^\top x + c$ is convex if Q is positive semidefinite (PSD)
- Least squares loss: $\|y - Ax\|_2^2$
- Norm: $\|x\|$ is convex for any norm

$$\|x\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2} \qquad \|x\|_1 = \sum_{i=1}^{n} |x_i|$$

First order convexity conditions

- Theorem:
- Suppose f is differentiable. Then f is convex if and only if for all $x, y \in dom f$

$$f(y) \geq f(x) + \nabla f(x)^\top (y - x)$$



Second order convexity conditions

- Suppose f is twice differentiable. Then f is convex if and only if for all $x \in dom f$

$$\nabla^2 f(x) \succeq 0$$

Properties of convex functions

- If *x* is a local minimizer of a convex function, it is a global minimizer.

- Suppose *f* is differentiable and convex. Then, *x* is a global minimizer of *f*(*x*) if and only if

$$\nabla f(x) = 0$$

- Proof:
  - $\nabla f(x) = 0$. We have

$$f(y) \geq f(x) + \nabla f(x)^\top (y - x) = f(x)$$

  - $\nabla f(x) \neq 0$. There is a direction of descent.

Gradient Descent

- The simplest optimization method.

- Goal:

$$\min_x f(x)$$

- Iteration:

$$x_{t+1} = x_t - \eta_t \nabla f(x_t)$$

- η_t is step size.

How to choose step size

- If step size is too big, the value of function can diverge.
- If step size is too small, the convergence is very slow.
- Exact line search:

$$\eta_t = \arg\min_{\eta} \; f(x - \eta \nabla f(x))$$

- Usually impractical.

Backtracking Line Search

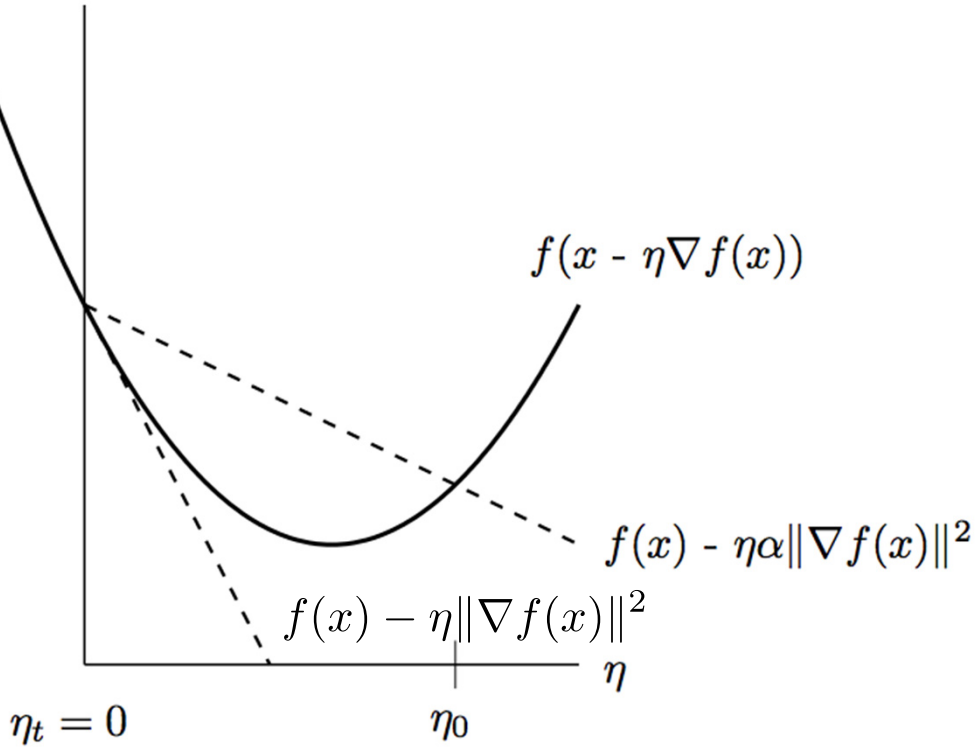
- Let $\alpha \in (0, 1/2], \beta \in (0, 1)$. Start with $\eta = 1$ and multiply $\eta = \beta\eta$ until

$$f(x - \eta \nabla f(x)) \leq f(x) - \alpha\eta \|\nabla f(x)\|^2$$

- Work well in practice.

Backtracking Line Search

- Understanding backtracking Line Search



$$f(x - \eta\nabla f(x)) \leq f(x) - \alpha\eta\|\nabla f(x)\|^2$$

Convergence Analysis

- Assume that f convex and differentiable.
- Lipschitz continuous:

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2$$

- Theorem:
  - Gradient descent with fixed step size η ≤ 1/L satisfies

$$f(x_t) - f^* \leq \frac{\|x_0 - x^*\|_2^2}{2t\eta}$$

  - To get $f(x_t) - f^* \leq \epsilon$, we need *O*(1/ϵ) iterations.
  - Gradient descent with backtracking line search have the same order convergence rate.

Convergence Analysis under Strong Convexity

- Assume f is strongly convex with constant m.
- Theorem:
  - Gradient descent with fixed step size $t \leq 2/(m + L)$ or with backtracking line search satisfies

$$f(x_t) - f^* \leq c^t \frac{L}{2} \|x_0 - x^*\|_2^2$$

  - where $0 < c < 1$.
  - To get $f(x_t) - f^* \leq \epsilon$, we need $O(\log(1/\epsilon))$ iterations.
  - Called linear convergence.

Newton's Method

- Idea: minimize a second-order approximation

$$f(x+v) \approx f(x) + \nabla f(x)^\top v + \frac{1}{2} v^\top \nabla^2 f(x) v$$

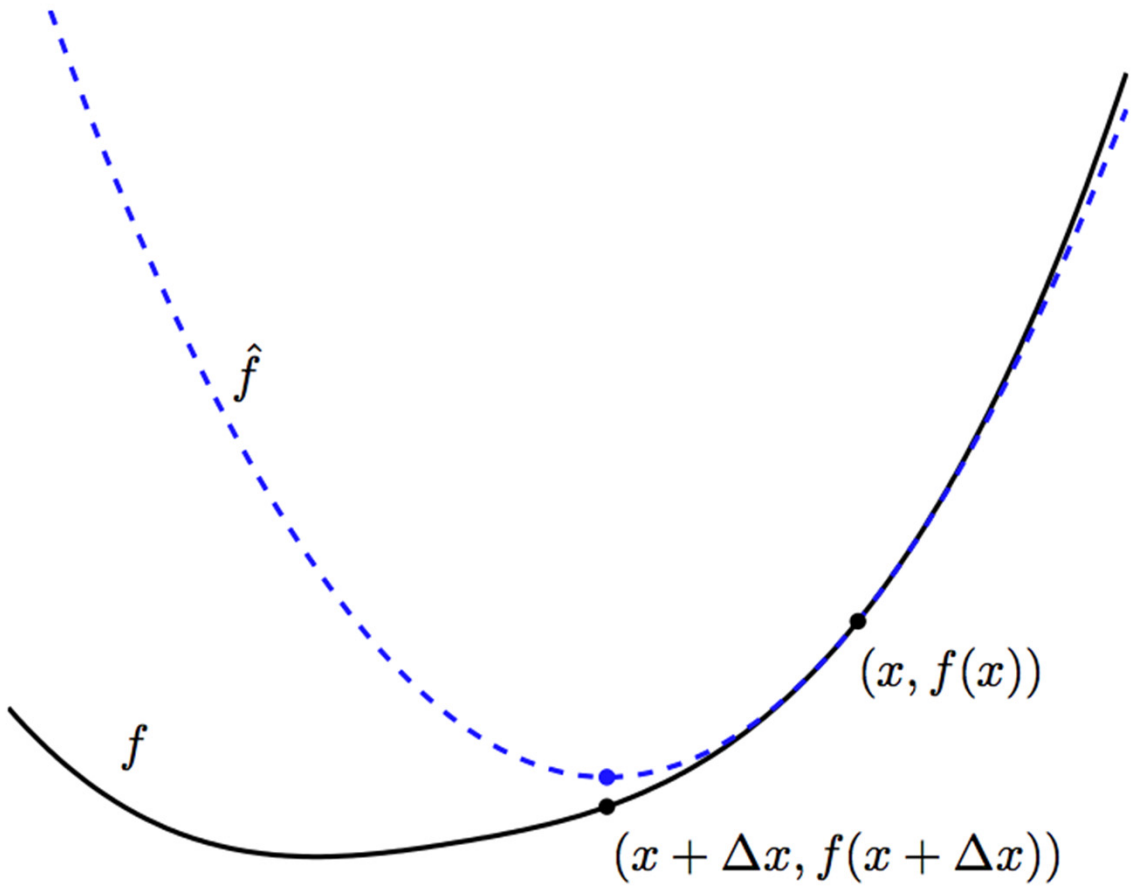
- Choose v to minimize above

$$v = -[\nabla^2 f(x)]^{-1} \nabla f(x)$$

- Newton step:

$$x_{t+1} = x_t - [\nabla^2 f(x_t)]^{-1} \nabla f(x_t)$$

Newton step



Newton's Method

- f is strongly convex
- $\nabla f(x), \nabla^2 f(x)$ are Lipschitz continuous
- Quadratic convergence:
  - convergence rate is O(log log(1/ϵ))
- Locally quadratic convergence: we are only guaranteed quadratic convergence after some number of steps k.

- Drawback: computing the inverse of Hessian is usually very expensive.
- Quasi-Newton, Approximate Newton...

Lagrangian

- Start with optimization problem:

$$\min_x \ f(x)$$
$$s.t. \ g_i(x) \leq 0, i = 1, 2, \ldots, m$$
$$h_j(x) = 0, j = 1, 2, \ldots, n$$

- We define Lagrangian as

$$L(x, u, v) = f(x) + \sum_{i=1}^{m} u_i g_i(x) + \sum_{j=1}^{n} v_j h_j(x)$$

- where $u_i \geq 0$

Property

- Lagrangian

$$L(x, u, v) = f(x) + \sum_{i=1}^{m} u_i g_i(x) + \sum_{j=1}^{n} v_j h_j(x)$$

- For any $u \geq 0$ and v, any feasible x,

$$L(x, u, v) \leq f(x)$$

Lagrange Dual Function

- Let *C* denote primal feasible set, *f** denote primal optimal value. Minimizing *L*(*x*, *u*, *v*) over all *x* gives a lower bound on *f** for any *u* ≥ 0 and *v*.

$$f^* \geq \min_{x \in C} L(x, u, v) \geq \min_{x} L(x, u, v) = g(u, v)$$

- Form dual function:

$$g(u, v) = \min_{x} L(x, u, v)$$

Lagrange Dual Problem

- Given primal problem

$$\min_x \ f(x)$$
$$s.t. \ g_i(x) \leq 0, i = 1, 2, \ldots, m$$
$$h_j(x) = 0, j = 1, 2, \ldots, n$$

- The Lagrange dual problem is:

$$\max_{u,v} \ g(u, v)$$
$$s.t. \ u \geq 0$$

Property

- Weak duality:

$$f^* \geq g^*$$

- The dual problem is a convex optimization problem (even when primal problem is not convex)

$$g(u,v) = \min_x \{ f(x) + \sum_{i=1}^{m} u_i g_i(x) + \sum_{j=1}^{n} v_j h_j(x) \}$$

- *g*(*u*,*v*) is concave.

Strong duality

- In some problems we have observed that actually

$$f^* = g^*$$

  which is called strong duality.

- Slater's condition: if the primal is a convex problem, and there exists at least one strictly feasible *x*, i.e,

$$g_1(x) < 0, \ldots, g_m(x) < 0 \text{ and } h_1(x) = \ldots h_n(x) = 0$$

  then strong duality holds

Example

- Primal problem

$$\min_x \ x^\top x$$

$$s.t. \ Ax \leq b$$

- Dual function

$$g(u) = \min_x \{x^\top x + u^\top (Ax - b)\}$$

$$= -\frac{1}{4} u^\top A A^\top u - b^\top u$$

- Dual problem

$$\max_u \ -\frac{1}{4} u^\top A A^\top u - b^\top u, s.t. \ u \geq 0$$

- Slater's condition always holds.

References

- A majority part of this lecture is based on CSS 490 / 590 - Introduction to Machine Learning
  - https://courses.washington.edu/css490/2012.Winter/lecture_slides/02_math_essentials.pdf
- The Matrix Cookbook – Mathematics
  - https://www.math.uwaterloo.ca/~hwolkowi/matrixcookbook.pdf
- E-book of Mathematics for Machine Learning
  - https://mml-book.github.io/

References

- Optimization for machine learning
  - https://people.eecs.berkeley.edu/~jordan/courses/294-fall09/lectures/optimization/slides.pdf

- A convex optimization course
  - http://www.stat.cmu.edu/~ryantibs/convexopt-F15/