COMP7180: Quantitative Methods for DAAI

(Credits from Prof. Andrew Ng)

(Credits from HKBU)

Course Instructors: Dr. Yang Liu and Dr. Bo Han

Teaching Assistant: Mr. Minghao Li

Course Contents

- Continuous and Discrete Random Variables (Week 7)
- Conditional Probability and Independence (Week 8)
- Maximum Likelihood Estimation (Week 9)
- Mathematical Optimization (Week 10)
- Convex and Non-Convex Optimization (Week 11)
- Quiz and Course Review (Week 12) Our Focus

What is Probability

- A probability can be regard as a function to estimate the value of every event.

As a function, we should have a domain (定义域). What is the domain?

Given a sample space S: set of all possible outcomes of an experiment.

The domain consists of some subsets of S.

An element E in the domain is called event.

What is Probability

Example: Toss a coin (1 time). Then, the outcome is H or T, where H is the head of a coin and T is the tail of a coin.

Then S= { H, T};

The domain is { {H,T}, {H}, {T}, Ø}.

{H,T}, {H}, {T}, Ø are called events.

What is Probability

Example: Toss a coin (1 time). Then, the outcome is H or T, where H is the head of a coin and T is the tail of a coin.

Then S= {H, T}; The domain is { {H,T}, {H}, {T}, Ø}. {H,T}, {H}, {T}, Ø are called events.

- S and Ø should be event;
- S^C = Ø; $\{H\}^C$ = {T}; $\{T\}^C$={H}; $Ø^C$= S;
- S∩ Ø=Ø; S∩{H}={H}; S∩{T}={T}; {H}∩{T}=Ø;
- {H}∪{T}=S; {H}∪ Ø={H}; {T}∪ Ø={T}.

What is Probability

As a function, we should have a range (值域). What is the range?

Given an event E, a probability maps E into [0,1], that is $0 \leq P(E) \leq 1$.

If $P(E)=0$, then this event E will not occur.

If $P(E)=1$, then this event E occurs without uncertainty.

What is Probability

Example. Toss a coin (1 time). There are outcomes: H and T, where H is the head of a coin and T is the tail of a coin.

S= {H, T}; The domain is {{H,T}, {H}, {T}, Ø}.

{H,T}, {H}, {T}, Ø are called events.

P({H,T}) =1; P({H}) = 0.5; P({T}) =0.5; P(Ø)=0.

What is Probability

Probability is a special function, which should satisfy some properities:

- P(S)=1; P(Ø)=0; 0≤P(E)≤ 1;
- If event E belongs to event F, then P(E)≤P(F);
- Given an event E, then P(E^{C}) = 1-P(E);
- Given events E and F, then P(E∪F) = P(E)+P(F)-P(E∩F).

Random Variables

- Generally, it is very complex to represent an event;

- To deal with more complex events, researchers have developed random variables (随机变量).

- Example. Toss a coin (1 time). In the sample space $S=\{H, T\}$, we design a function $X: S \to \{1,-1\}$ such that $X(H)=1$ and $X(T)=-1$. Then X is

  a random variable.

  Moreover, $P(X=1) = P(\{H\}) = 0.5$ and $P(X=-1) = P(\{T\})=0.5$.

What are Random Variables

- A random variable is a variable that can take on different values randomly. We typically denote the random variable itself with an uppercase letter in plain typeface, and the values it can take on with lowercase letters.

- For vector-valued variables, we would write the random variable as **X** and one of its values as **x**.

- Random variables may be discrete or continuous. A discrete random variable is one that has a finite or countably infinite number of states. A continuous random variable is associated with a real value.

Discrete Variables and PMF

- A probability distribution over discrete variables may be described using a probability mass function (PMF, 概率质量函数)
- The probability mass function maps from a state of a random variable to the probability of that random variable taking on that state.
- $0 \leq \mathrm{P}(\mathrm{X} = \mathrm{x}) \leq 1$
- $\sum_{\mathrm{x}} \mathrm{P}(\mathrm{X} = \mathrm{x}) = 1$. We refer to this property as being normalized

Discrete Variables and PMF: Examples

Discrete Random Variable with finite range:

Toss a coin (1 time).

In the sample space $S=\{H, T\}$, we design
a random variable $X: S \rightarrow \{1,-1\}$ such that
$X(H)=1$ and $X(T)=-1$. Then X is a random variable with finite range.

The probability is $P(X=1)=P(X=-1)= 0.5$.

Discrete Variables and PMF: Examples

Discrete Random Variable with infinite range:

Toss a coin (countably infinity times).

We design a random variable X: X = n means that

the first head appears after throwing n times.

Then X is a random variable with countably infinite range.

The probability is P(X=n) = 0.5^{n}.

Discrete Variables and PMF: Examples

- Discrete uniform distribution (均匀分布) is one of the most important discrete distributions
- It is a finite discrete distribution
- Assume that the range is $x_1, x_2 \ldots x_n$, then

- $P(X = x_i) = \frac{1}{n}$; $\sum_i P(X = x_i) = \sum_i \frac{1}{n} = \frac{n}{n} = 1$

Continuous Variables and PDF

- A continuous variable X is a function;
- Range is not discrete and take values in real number;
- There is a probability density function (概率密度函数) $p_{\mathrm{X}}(\mathrm{x})$ such that

1) $p_{\mathrm{X}}(\mathrm{x}) \geq 0$;
2) P(a≤X≤b) = $\int_a^b p_{\mathrm{X}}(\mathrm{x})\, d\mathrm{x}$;
3) $\int_{-\infty}^{+\infty} p_{\mathrm{X}}(\mathrm{x})\, d\mathrm{x}$=1.

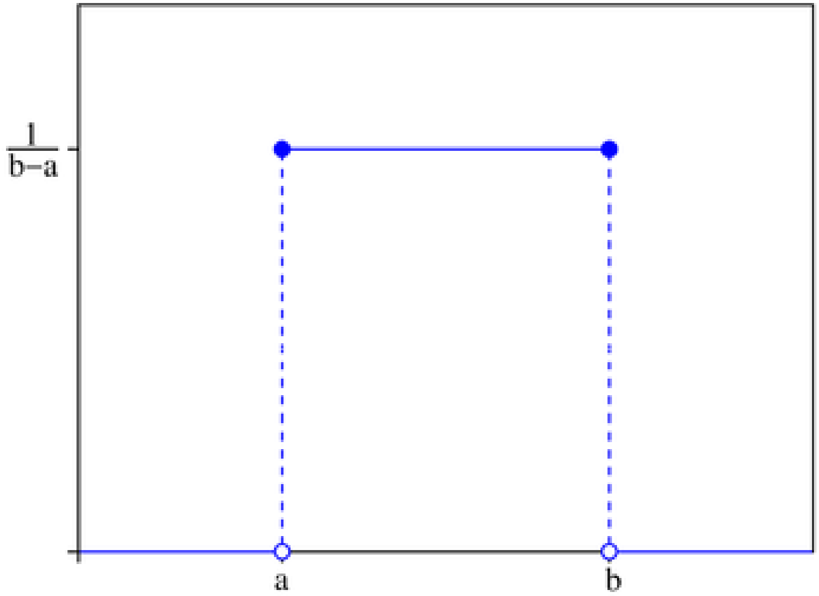
Continuous Variables and PDF

- In principle variables such as height, weight, and temperature are continuous, in practice the limitations of our measuring instruments restrict us to a discrete (though sometimes very finely subdivided) world.

- However, continuous models often approximate real-world situations very well, and continuous mathematics (calculus) is frequently easier to work with than mathematics of discrete variables and distributions.

Continuous Variables and PDF

- Continuous uniform distribution is one of the most important continuous distributions.
- The probability density function of continuous unfiorm distribution can be written as $p(x; a, b) = \frac{1}{b-a}$



Joint Distribution

- In some practice case, we need to consider multiple randon variables

For example, it is clear that the weight is related to the height. So we are interested in knowing the joint distribution related to dog's weight and height.

- If random variables X, Y are discrete random variables, then the joint distribution of X and Y is

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

Joint Distribution

- If random variables X, Y are continuous random variables, then the joint distribution of X and Y is

$$P(a1 \leq X \leq b1,\ a2 \leq Y \leq b2)$$

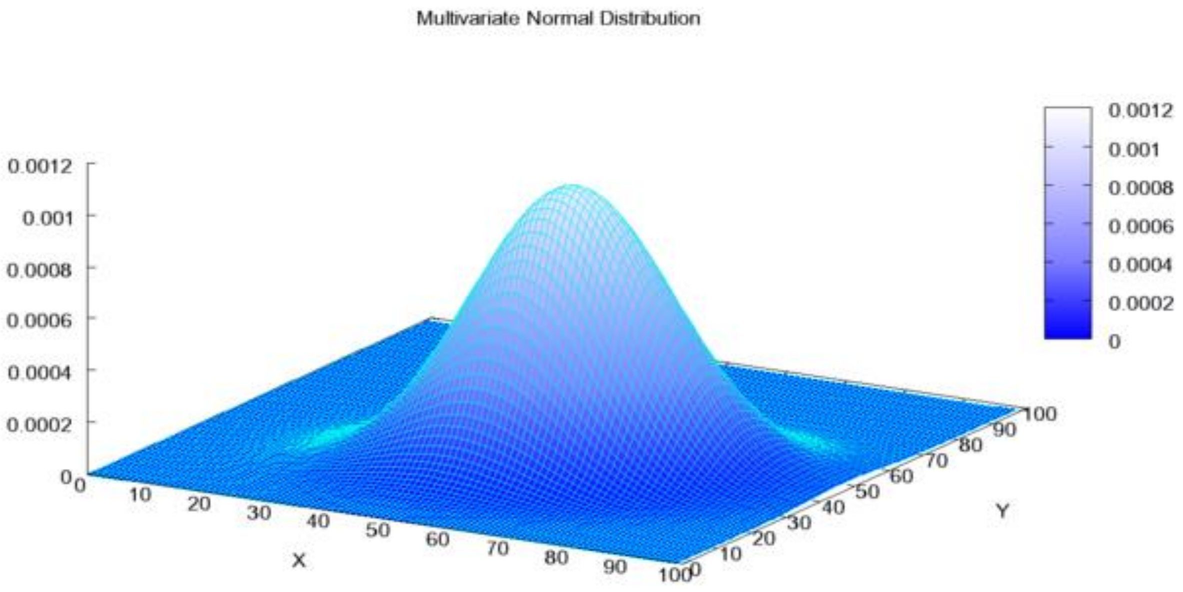
- In fact, when X, Y are continuous random variables, there exists a probability density function for the joint distribution:

$$P(a1 \leq X \leq b1,\ a2 \leq Y \leq b2) = \int_{a1}^{b1} \int_{a2}^{b2} p_{\mathrm{XY}}(x,y)\,dxdy,$$

where $p_{\mathrm{XY}}(x,y)$ is the probability density function.

Joint Distribution

- If X represnets the weight of dog and Y represents the height of dog, then the joint distribuion P(X,Y) is similar to a two-dimensional Gaussian distribution.



Joint Probability

- If random variable X is continuous random variable, and Y is discrete random variable, then the joint distribution of X and Y is

$$P(a1 \leq X \leq b1, Y = y)$$

In fact, when X, is continuous random variable, there exists a probability density function for the joint distribution:

$$P(a1 \leq X \leq b1, Y = y) = \int_{a1}^{b1} p_{XY}(x,y)dx,$$

where $p_{XY}(x,y)$ is continuous with respect to x, but discrete with respect to y.

Marginal Probability

- Using joint distrbution, we can construct marginal distribution:

- If random variables X, Y are discrete random variables, then the marginal distributions are

$$P(X=x) = \sum_{y} P(X = x, Y = y)$$

$$P(Y=y) = \sum_{x} P(X = x, Y = y)$$

Marginal Probability

- If random variables X and Y are continuous random variables, then the marginal distribution with respect to X is

$$P(a\leq X \leq b) = \int_a^b \int_{-\infty}^{+\infty} p_{XY}(x, y)dxdy$$

and the density function of X is

$$p(x) = \int_{-\infty}^{+\infty} p_{XY}(x, y)dy$$

Similarly, we can obtain the marginal distribution with respect to Y and the density function of Y.

Marginal Probability

- If random variable X is continuous random variable, and Y is a discrete random variable, then the marginal distribution with respect to X is

$$\mathrm{P}(\mathrm{a} \leq \mathrm{X} \leq \mathrm{b}) = \sum_{y} \int_{a}^{b} p_{\mathrm{XY}}(\mathrm{x}, \mathrm{y})dx$$

and the density function of X is

$$p(\mathrm{x}) = \sum_{y} p_{\mathrm{XY}}(\mathrm{x}, \mathrm{y})$$

The marginal distribution with respect to Y is

$$\mathrm{P}(\mathrm{Y=y}) = \int_{-\infty}^{+\infty} p_{\mathrm{XY}}(\mathrm{x}, \mathrm{y})dx$$

Independence and Conditional Independence

- If events E and F are independent, then

$$P(E|F) = P(E \cap F)/P(F)=P(E)P(F)/P(F)=P(E).$$

So

$$P(E|F) = P(E)$$

We can also obtain that

$$P(F|E) = P(F)$$

Independence and Conditional Independence

- Two random variables X and Y are independent if their probability distribution can be expressed as a product of two factors:

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

If X and Y are both discrete random variables, then X and Y are independent, if for any x, y

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

We can also obtain that $P(X=x|Y=y) = P(X=x)$ and $P(Y=y|X=x)=P(Y=y)$.

Independence and Conditional Independence

If X and Y are continuous random variables, then X and Y are independent, if for any x, y,

$$p_{\mathrm{XY}}(\mathrm{x,y})=p_{\mathrm{X}}(\mathrm{x})p_{\mathrm{Y}}(\mathrm{y})$$

where p_{XY}(x,y) is the density function of the joint distribution, p_{X}(x) is the density function with respect to random variable X, and p_{Y}(x) is the density function with respect to random variable Y.

Conditional Independence to Random Variables

Two random variables X and Y are conditionally independent given a random variable Z if the conditional probability distribution

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

When X, Y and Z are discrete variable variables, if X and Y are conditionally independent given a random variable Z: for any x, y, z

$$P(X=x,Y=y|Z=z) = P(X=x|Z=z)P(Y=y|Z=z)$$

Conditional Independence to Random Variables

When X, Y and Z are continouos variable variables, if X and Y are conditionally independent given a random variable Z: for any x, y, z,

then the probability density functions satisfy that

$$p_{\mathrm{XY|Z}}(\mathrm{x},\mathrm{y}|\mathrm{z}) = p_{\mathrm{X|Z}}(\mathrm{x}|\mathrm{z})\, p_{\mathrm{Y|Z}}(\mathrm{y}|\mathrm{z})$$

So

$$\frac{p_{\mathrm{XYZ}}(\mathrm{x},\mathrm{y},\mathrm{z})}{p_{\mathrm{Z}}(\mathrm{z})} = \frac{p_{\mathrm{XZ}}(\mathrm{x},\mathrm{z})}{p_{\mathrm{Z}}(\mathrm{z})} \frac{p_{\mathrm{YZ}}(\mathrm{y},\mathrm{z})}{p_{\mathrm{Z}}(\mathrm{z})}$$

$$p_{\mathrm{XYZ}}(\mathrm{x},\mathrm{y},\mathrm{z}) = p_{\mathrm{XZ}}(\mathrm{x},\mathrm{z})\, p_{\mathrm{YZ}}(\mathrm{y},\mathrm{z})$$

Bayes' Rule

- Bayes' theorem is an important tool in statistics and machine learning:

Given events A and B, then $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$.

Proof:

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

Bayes' Rule

- How to use Baye's Rule?

Example: A bag I contains 4 white and 6 black balls while another bag II contains 4 white and 3 black balls.

One ball is drawn at random from one of the bags, and it is found to be black.

Find the probability that it was drawn from bag I.

Bayes' Rule

Soultion: Let E1 be the event of choosing bag I, E2 the event of choosing bag II, and A be the event of drawing a black ball.

Then, $P(E1) = P(E2) = 0.5$.

$P(A|E1)$ = P(drawing a black ball from Bag I) = 6/10 = 3/5.

$P(A|E2)$ = P(drawing a black ball from Bag II) = 3/7

$P(A) = P(A|E1)P(E1)+P(A|E2)P(E2) = 18/35$

By using Bayes' theorem, the probability of drawing a black ball from bag I out of two bags,

$$P(E1|A) = P(A|E1)P(E1)/P(A) = 0.6*0.5/(18/35) = 7/12$$

Bayes' Rule

- How to use Baye's Rule?

Example: A man is known to speak the truth 2 out of 3 times. He throws a dice and reports that the number obtained is a four. Find the probability that the number obtained is four.

Bayes' Rule

Soultion: Let A be the event that the man reports that number four is obtained. Let E1 be the event that four is obtained and E2 be its complementary event.

Then, P(E1) = Probability that four occurs = 1/6.

P(E2) = Probability that four does not occur = 1- P(E1) = 1 – (1/6) = 5/6.

P(A|E1)= Probability that man reports four and it is actually a four = 2/3

P(A|E2) = Probability that man reports four and it is not a four = 1/3.

So P(A) = P(A|E1)P(E1)+P(A|E2)P(E2) = 1/9+5/18 = 7/18.

By using Bayes' theorem, probability that number obtained is actually a four, P(E1|A) = P(A|E1)P(E1)/P(A) = 2/18/(7/18) = 2/7.

Expectation (or Mean)

The expectation or mean of a random variable X *is denoted by* E[X] *and defined as:*

- For discrete variable,

$$\mathrm{E[X]} = \sum_{x} \mathrm{P(X = x)x}$$

- For continuous variable,

$$\mathrm{E[X]} = \int_{-\infty}^{+\infty} p_{\mathrm{X}}(\mathrm{x})\mathrm{x}\,\mathrm{dx}$$

In words, we are taking a weighted sum of the values that x can take on, where the weights are the probabilities of those respective values. The expected value has a physical interpretation as the "center of mass" of the distribution.

Expectation of Functions

The expectation or mean of f(X) (a function of random variable X) is denoted by E[f(X)] and defined as:

For discrete variable,

$$E[f(X)] = \sum_{x} P(X = x) f(x)$$

For continuous variable,

$$E[f(X)] = \int_{-\infty}^{+\infty} p_X(x) f(x)\, dx$$

Properties of Expectation

For any two random variables X and Y, functions f and g, and any constants a, b $\in$ R, the following equations hold:

- $E[a] = a$, $\quad E[f(a)] = f(a)$
- $E[X+Y] = E[X]+E[Y]$ $\quad E[f(X)+g(Y)] = E[f(X)]+E[g(X)]$
- $E[aX] = aE[X]$ $\quad E[af(X)] = aE[f(X)]$
- $E[aX+bY]=aE[X]+bE[Y]$ $\quad E[af(X)+bg(Y)]=aE[g(X)]+bE[g(Y)]$

Credits from Dr. Liu Yang

Variance

- Expectation provides measure of the “center” of a distribution, but sometimes we are also interested in what the “spread” is about that center. Therefore, we define the variance Var(X) of a random variable X as follows:

$$Var(X) = E[(X – E[X])^2]$$

- In words, this is the average squared deviation of the values of X from the mean of X.

Properties of Variance

For any random variable X and any a, b ∈ R, the following equations hold:

$$\mathrm{Var}(aX+b) = a^2\mathrm{Var}(X)$$

$$\mathrm{Var}(X) = E[X^2] - E[X]^2$$

Covariance

- The covariance gives some sense of how much two values are linearly related to each other, as well as the scale of these variables

The covariance of two random variables X and Y is denoted by Cov(X, Y) and defined as

$$\mathrm{Cov}(X, Y) = E[(X - E[X])(Y - E[Y])] .$$

Covariance of functions: given functions f and g, then

$$\mathrm{Cov}(f(X), g(Y)) = E[(f(X) - E[f(X)])(g(Y) - E[g(Y)])] .$$

Properties of Covariance

It is clear that

$$\mathrm{Cov}(X,X) = \mathrm{Var}(X);$$

$$\mathrm{Cov}(f(X),f(X)) = \mathrm{Var}(f(X)).$$

Properties of Covariance

If two random variables X and Y are independent, then

$$\mathrm{Cov}(X, Y) = 0 \ .$$

Proof: From the definition of covariance, we have

$\mathrm{Cov}(X, Y) = E[(X - E[X])(Y - E[Y])]$

$= E[XY - E[X]Y - XE[Y] + E[X]E[Y]]$

$= E[XY] - E[E[X]Y] - E[XE[Y]] + E[E[X]E[Y]]$

$= E[XY] - E[X]E[Y] - E[X]E[Y] + E[X]E[Y] = E[XY] - E[X]E[Y] = 0.$

Credits from Dr. Liu Yang

Common Probability Distribution

Distribution	PDF or PMF	Mean	Variance
$Bernoulli(p)$	$\begin{cases} p, & \text{if } x = 1 \\ 1-p, & \text{if } x = 0. \end{cases}$	p	$p(1-p)$
$Binomial(n, p)$	$\binom{n}{k} p^k (1-p)^{n-k}$ for $k = 0, 1, ..., n$	np	$np(1-p)$
$Geometric(p)$	$p(1-p)^{k-1}$ for $k = 1, 2, ...$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
$Poisson(\lambda)$	$\frac{e^{-\lambda}\lambda^k}{k!}$ for $k = 0, 1, ...$	λ	λ
$Uniform(a, b)$	$\frac{1}{b-a}$ for all $x \in (a, b)$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
$Gaussian(\mu, \sigma^2)$	$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ for all $x \in (-\infty, \infty)$	μ	σ^2
$Exponential(\lambda)$	$\lambda e^{-\lambda x}$ for all $x \geq 0, \lambda \geq 0$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$

Credicts from Dr. Griffin Young

Gaussian Distribution (Normal Distribution)

- **Gaussian Distribution (Normal Distribution)**: It is the most widely used model for the distribution of continuous variables. For a single variable X, the Gaussian distribution can be represented as follows:

$$\mathrm{N}\left(\mu, \sigma^2\right).$$

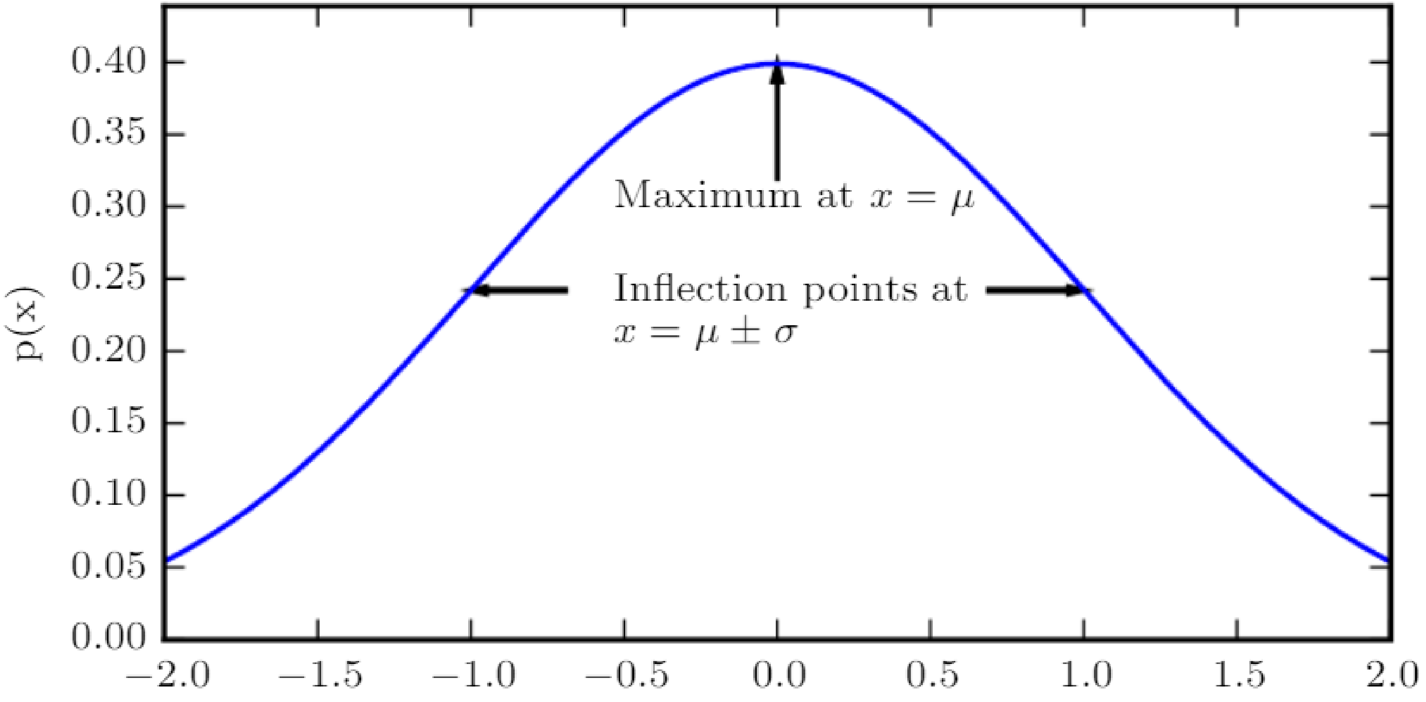
- It is a continuous distribution with the probability density function

$$\sqrt{\frac{1}{2\pi\sigma^2}}\exp\left(-\frac{1}{2\sigma^2}(\mathrm{x}-\mu)^2\right)$$

Gaussian Distribution (Normal Distribution)

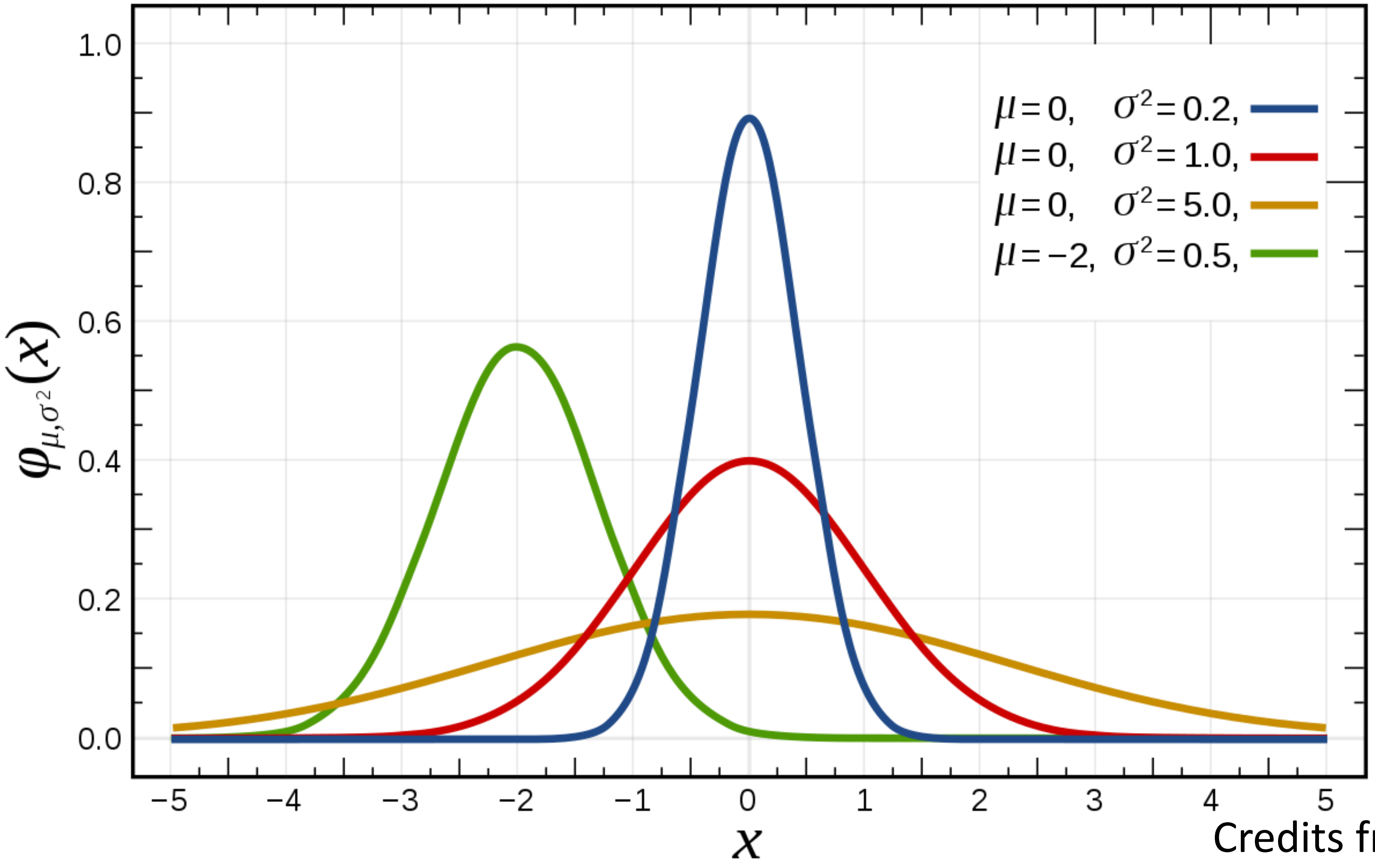
Gaussian distribution depends on two parameters μ, σ.

Then the expectation is μ and the variance is σ^2.



Gaussian Distribution (Normal Distribution)

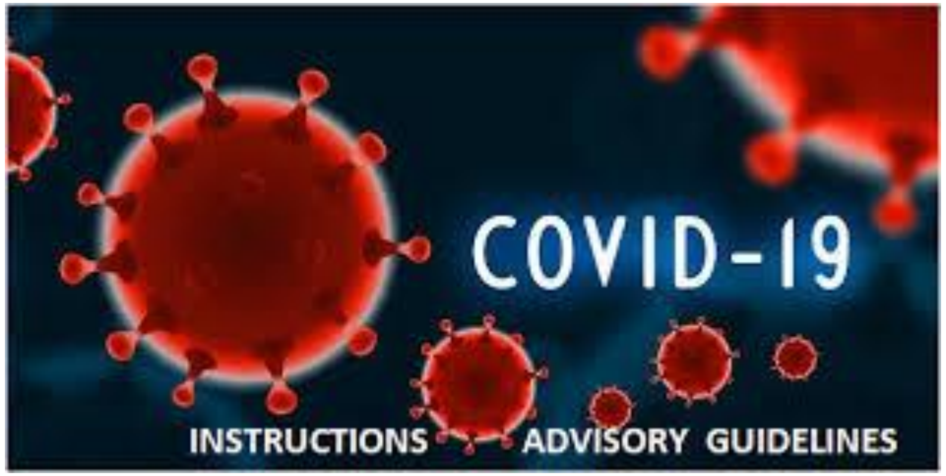
- Gaussian distributions with different expectations and variances



Credits from Wikipedia

Review: Exercise

Consider following table of counts that are obtained from an observed sample of individuals both males and females, who had taken Covid-19. X, Y, Z are random variables. Z represents the gender, X represents whether individuals recover from Covid-19, and Y represents whether individuals have been treated.



Z = 1 Male; Z = 0 Female;
X = 1 Recovery (康复）; X = 0 No Recovery ;
Y = 1 Treatment（治疗）; Y = 0 No Treatment;

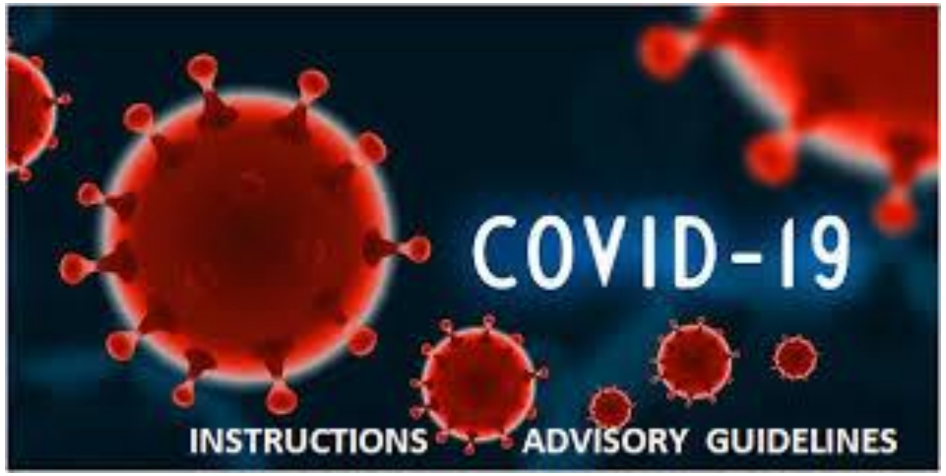
	Z=1, X=1	Z=1, X=0	Z=0, X=1	Z=0, X=0
Y=1	245	105	315	735
Y=0	630	420	70	280

245, 105, 315, 735, 630,420,70, 280 represent the number of individuals corresponding to different values of X, Y and Z.

Review: Exercise

Please compute that

$P(X=1|Y=1,Z=1)$ $P(X=1|Y=0,Z=1)$
$P(X=1|Y=1,Z=0)$ $P(X=1|Y=0,Z=0)$
$P(X=1|Y=1)$ $P(X=1|Y=0)$



	Z=1, X=1	Z=1, X=0	Z=0, X=1	Z=0, X=0
Y=1	245	105	315	735
Y=0	630	420	70	280

Review: Exercise

Solution:

$P(X=1|Y=1,Z=1)=P(X=1,Y=1,Z=1)/P(Y=1,Z=1)=P(X=1,Y=1,Z=1)/(P(X=1,Y=1,Z=1)+P(X=0,Y=1,Z=1))=245/(245+105) = 0.7$

$P(X=1|Y=0,Z=1)=P(X=1,Y=0,Z=1)/P(Y=0,Z=1)=P(X=1,Y=0,Z=1)/(P(X=1,Y=0,Z=1)+P(X=0,Y=0,Z=1))= 630/(630+420) = 0.6$

$P(X=1|Y=1,Z=1)$ means the the recovery probability for individuals who are male and have been treated.

$P(X=1|Y=0,Z=1)$ means the the recovery probability for individuals who are male and have not been treated.

Review: Exercise

Solution:

$P(X=1|Y=1,Z=0)=P(X=1,Y=1,Z=0)/P(Y=1,Z=0)=P(X=1,Y=1,Z=0)/(P(X=1,Y=1,Z=0)+P(X=0,Y=1,Z=0))=315/(315+735) = 0.3$

$P(X=1|Y=0,Z=0)=P(X=1,Y=0,Z=0)/P(Y=0,Z=0)=P(X=1,Y=0,Z=0)/(P(X=1,Y=0,Z=0)+P(X=0,Y=0,Z=0))=70/(70+280) = 0.2$

$P(X=1|Y=1,Z=0)$ means the the recovery probability for individuals who are female and have been treated.

$P(X=1|Y=0,Z=0)$ means the the recovery probability for individuals who are female and have not been treated.

Review: Exercise

Solution:

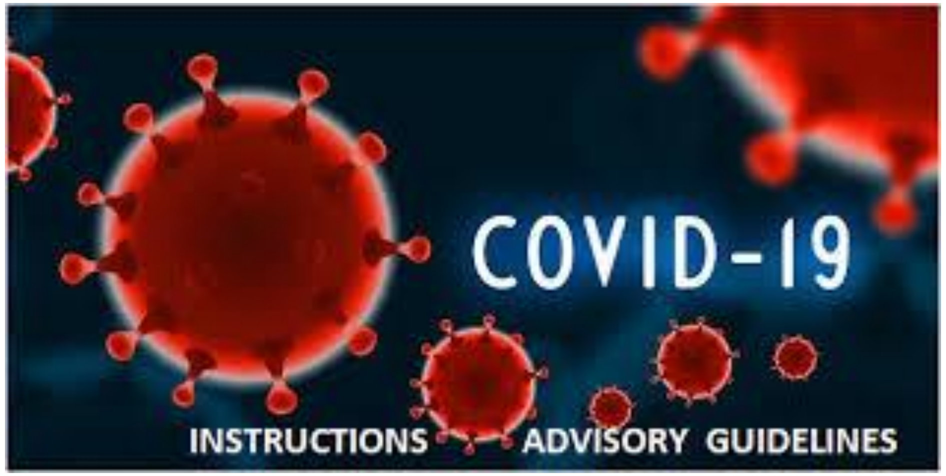
$P(X=1|Y=1)=P(X=1,Y=1)/P(Y=1)=(245+315)/(245+105+315+735) = 0.4$

$P(X=1|Y=0)=P(X=1,Y=0)/P(Y=0)=(630+70)/(630+420+70+280)=0.5$

$P(X=1|Y=1)$ means the the recovery probability for individuals who have been treated.

$P(X=1|Y=0)$ means the the recovery probability for individuals who have not been treated.

Review: Exercise



$P(X=1|Y=1,Z=1)=0.7>P(X=1|Y=0,Z=1)=0.6,$
$P(X=1|Y=1,Z=0)=0.3>P(X=1|Y=0,Z=0)=0.2.$

- The recovery probability for individuals who are male and have been treated is larger than the recovery probability for individuals who are male and have not been treated.

- The recovery probability for individuals who are female and have been treated is larger than the recovery probability for individuals who are female and have not been treated.

Review: Exercise

- The recovery probability for individuals who are male and have been treated is larger than the recovery probability for individuals who are male and have not been treated.

- The recovery probability for individuals who are female and have been treated is larger than the recovery probability for individuals who are female and have not been treated.

- Does this mean that the treatment can make postive affects?

That is: is the recovery probability for individuals who have been treated larger than the recovery probability for individuals who have not been treated?

Review: Exercise

To answer this question, we need to compute

$$P(X=1|Y=1) \text{ and } P(X=1|Y=0)$$

Then, we need to compare them.

We discover that $P(X=1|Y=1)=0.4 < P(X=1|Y=0)=0.5$.

- That is: the recovery probability for individuals who have been treated is smaller than the recovery probability for individuals who have not been treated.

Review: Exercise

It seems that

$P(X=1|Y=1,Z=1)>P(X=1|Y=0,Z=1)$,
$P(X=1|Y=1,Z=0)>P(X=1|Y=0,Z=0)$
and

$P(X=1|Y=1)<P(X=1|Y=0)$ are conflict.

Why does this happen?

It is called Simpson's paradox.

- The recovery probability for individuals who are male and have been treated is larger than the recovery probability for individuals who are male and have not been treated;

  and
- the recovery probability for individuals who are female and have been treated is larger than the recovery probability for individuals who are female and have not been treated.

But

- the recovery probability for individuals who have been treated is smaller than the recovery probability for individuals who have not been treated

Review: Exercise

Note that

$P(Z=1|Y=1)P(X=1|Y=1,Z=1)+P(Z=0|Y=1)P(X=1|Y=1,Z=0)=P(X=1|Y=1)$

$P(Z=1|Y=0)P(X=1|Y=0,Z=1)+P(Z=0|Y=0)P(X=1|Y=0,Z=0)=P(X=1|Y=0)$

Although $P(X=1|Y=1,Z=1)>P(X=1|Y=0,Z=1)$, $P(X=1|Y=1,Z=0)>P(X=1|Y=0,Z=0)$, the conditional probabilities $P(Z=1|Y=1)$, $P(Z=0|Y=1)$, $P(Z=1|Y=0)$ and $P(Z=0|Y=0)$ can affect the values of $P(X=1|Y=1)$ and $P(X=1|Y=0)$.

That is the basic reason why the Simpson's paradox happens.

Review: Exercise

Detailly,

let $u = P(Z=1|Y=1)$ and $v = P(Z = 1|Y = 0)$, then if we hope that

$$P(X=1|Y=1)>P(X=1|Y=0)$$

We need the following inequality:

$$0.7u + 0.3(1 - u) > 0.6v + 0.2(1 - v).$$

Whether the inequality can success depends on the values of u and v.

$$0.7u+0.3(1-u)>0.6v+0.2(1-v) \text{ if and only if } v-u<0.25.$$

But, $u=P(Z=1|Y=1)=0.25$ and $v=P(Z=1|Y=0)=0.75$. So $v-u=0.5>0.25$. That is the reason why $P(X=1|Y=1)<P(X=1|Y=0)$.

Maximum Likelihood (ML) Estimation

The selected distribution $P(X;\alpha^*)$ is the most possible distribution sampling data S=(x1,x2,…,xn), i.i.d..

Understanding above sentence, we can formulate it as follows:

$$\underset{\alpha \in \Delta}{\operatorname{argmax}} \mathrm{P}(x1, x2, \ldots, xn; \alpha)$$

here we assume $\mathrm{P}(X; \alpha)$ is a discrete distribution.

- $\max_{\alpha \in \Delta} \mathrm{P}(x1, x2, \ldots, xn; \alpha)$

  means the largest probability for $\mathrm{P}(X; \alpha)$ that S is observed.

Maximum Likelihood (ML) Estimation

Because (x1,...,xn), are Independent and identically distributed,

$$\underset{\alpha \in \Delta}{\operatorname{argmax}}\, P(x1, x2, \ldots, xn; \alpha)$$

is equal to

$$\underset{\alpha \in \Delta}{\operatorname{argmax}} \prod_{i=1}^{n} P(X = xi; \alpha)$$

Maximum Likelihood (ML) Estimation

To reduce the affects of Multiplier operator $\prod_{i=1}^{n}$, we take a small trick (we use the property of log function to help us):

$$\log \prod_{i=1}^{n} ai = \sum_{i=1}^{n} \log ai$$

Step 1. We take log function.

$$\underset{\alpha \in \Delta}{\operatorname{argmax}} \prod_{i=1}^{n} P(X = xi; \alpha) = \underset{\alpha \in \Delta}{\operatorname{argmax}} \log \prod_{i=1}^{n} P(X = xi; \alpha)$$

Maximum Likelihood (ML) Estimation

$$\underset{\alpha \in \Delta}{\operatorname{argmax}} \prod_{i=1}^{n} P(X = xi; \alpha) = \underset{\alpha \in \Delta}{\operatorname{argmax}} \ \log \prod_{i=1}^{n} P(X = xi; \alpha)$$

Step 2. Using the property of log function:

$$\log \prod_{i=1}^{n} P(X = xi; \alpha) = \sum_{i=1}^{n} \log P(X = xi; \alpha)$$

Maximum Likelihood (ML) Estimation

Therefore,

$$\underset{\alpha \in \Delta}{\operatorname{argmax}} \prod_{i=1}^{n} P(X = xi; \alpha) = \underset{\alpha \in \Delta}{\operatorname{argmax}} \sum_{i=1}^{n} \log P(X = xi; \alpha)$$

Step 3. We need to optimize

$$\underset{\alpha \in \Delta}{\operatorname{argmax}} \sum_{i=1}^{n} \log P(X = xi; \alpha) \quad (1)$$

and obtain the optimal solution.

The solution of Eq. 1 is called Maximum Likelihood Estimation.

Maximum Likelihood (ML) Estimation

If the distribution class consists of continuous distributions, that is P(X;α) is a continuous distribution with respect to all α∈ Δ.

Then the Maximum Likelihood Estimation is

$$\underset{\alpha\in\Delta}{\operatorname{argmax}} \sum_{i=1}^{n} \log p_X(xi;\alpha) \quad (2)$$

where $p_X(x;\alpha)$ is the density function of P(X;α).

Maximum Likelihood (ML) Estimation

How to obtain the solution of

$$\underset{\alpha \in \Delta}{\operatorname{argmax}} \sum_{i=1}^{n} \log P(X = xi; \alpha) \ ?$$

- This is related to optimization problem.

- Generally, there are no unviersal approaches to give soultions to all Maximum Likelihood (ML) Estimation.

- The approaches are case by case.

Maximum Likelihood (ML) Estimation

In this class, we introduce a common used approach.

This approach is based on a simple theorem:

If 1) a function f(x1,x2,...,xd) is differentiable,
2) $\mathrm{x}^* = (\mathrm{x1}^*, \mathrm{x2}^*, \ldots, \mathrm{xd}^*)$ is the maximum point of f, then

$$\frac{\partial f}{\partial \mathrm{xi}}(\mathrm{x1}^*, \mathrm{x2}^*, \ldots, \mathrm{xd}^*) = 0.$$

Maximum Likelihood (ML) Estimation

Using this theorem, if $\sum_{i=1}^{n} \log P(X = xi; \alpha)$ is differentiable, then

Let α=(α1,α2,...,αd),

$$\frac{\partial \sum_{i=1}^{n} \log P(X=xi;\alpha)}{\partial \alpha j} = 0, \ \text{for } j=1,\ldots,d$$

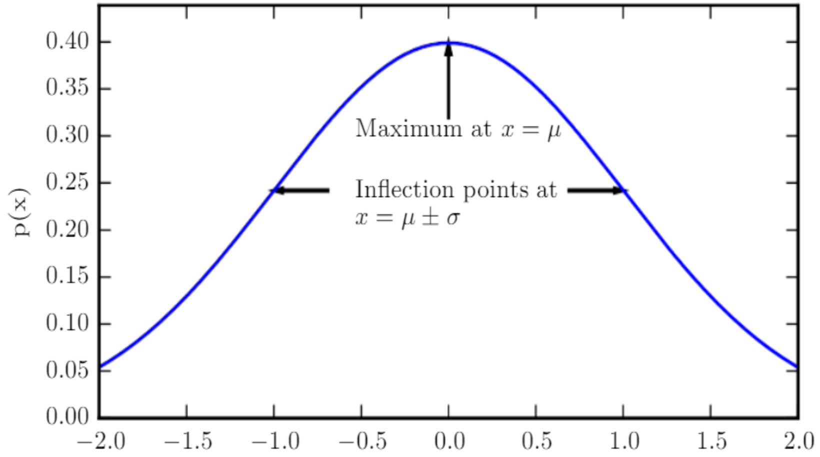
Then, addressing above equations.

Check that you've found a maximum rather than a minimum or saddle-point, and be careful if α belongs to *Δ*.

Exercises: MLE for Gaussian Distribution

Suppose you have x1,x2,…,xn (i.i.d) $N(\mu,\sigma^2)$

$$\sqrt{\frac{1}{2\pi\sigma^2}}\exp(-\frac{1}{2\sigma^2}(\mathrm{x}-\mu)^2)$$



- Assume that you know σ^2
- But you don't know μ

MLE: For which μ is x1, x2, ..., xn most likely?

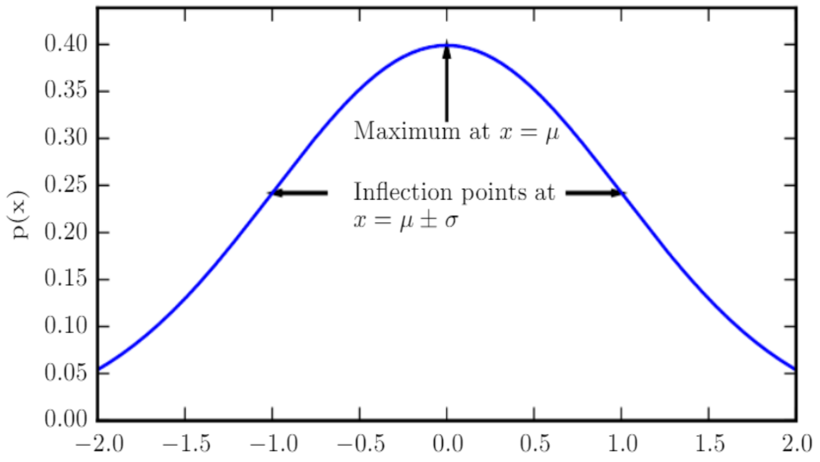
Credicts from Dr. Liu Yang

Exercises: MLE for Gaussian Distribution

Compute the MLE $\underset{\mu \in R}{\operatorname{argmax}} \sum_{i=1}^{n} \log p_X(xi; \mu)$

$$\underset{\mu \in R}{\arg\max} \frac{1}{\sqrt{2\pi}\ \sigma} \sum_{i=1}^{n} -\frac{(xi-\mu)^2}{2\sigma^2}$$

$$= \underset{\mu \in R}{\arg\min} \sum_{i=1}^{n} (xi-\mu)^2$$

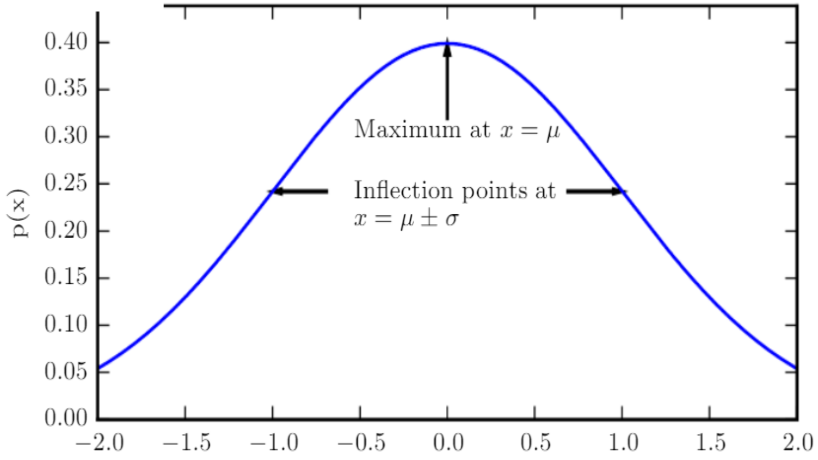


Noticed that σ is known, we can ignore it when finding the proper μ

Exercises: MLE for Gaussian Distribution

Derivation the equation $\underset{\mu \in R}{\arg\min} \sum_{i=1}^{n} (\mathrm{xi} - \mu)^2$

$$\frac{\mathrm{d} \sum_{\mathrm{i}=1}^{n} (\mathrm{xi} - \mu)^2}{\mathrm{d}\mu} = 2 \sum_{\mathrm{i}=1}^{n} (\mathrm{xi} - \mu) = 0$$



So the solution is

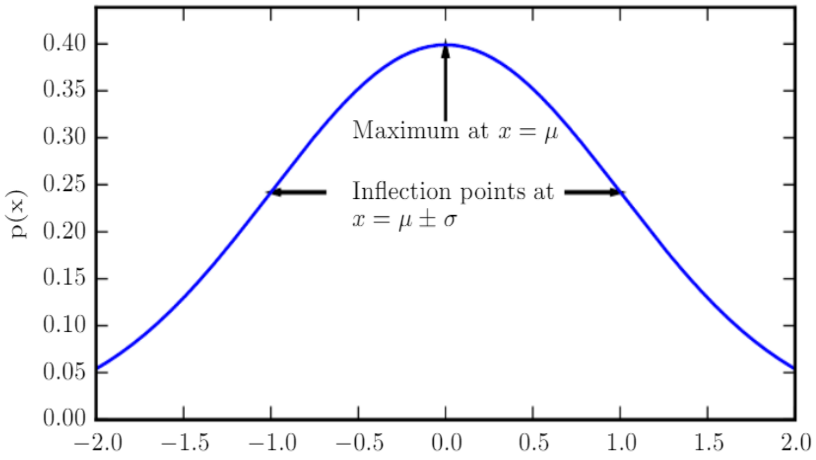
$$\mu = \frac{\sum_{i=1}^{n} \mathrm{xi}}{n}$$

Credicts from Dr. Liu Yang

Exercises: MLE for Gaussian Distribution

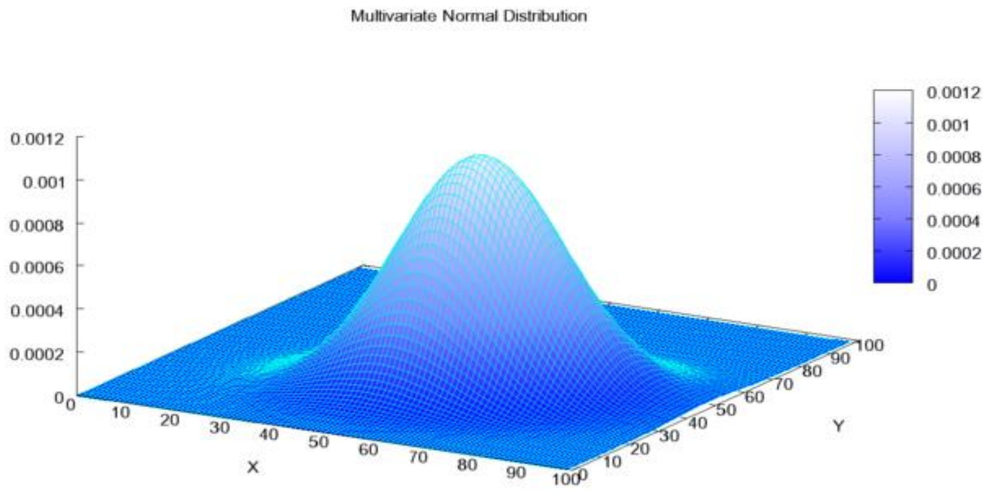
- In conclusion, the best estimate of the mean of a gaussian distribution is the mean of the sample!

$$\mu = \frac{\sum_{i=1}^{n} \mathrm{xi}}{n}$$



Exercises: MLE for high-dimensional Gaussian Distribution

- Given a 2 $\times$ 2 positive semi-definite matrix (半正定) Σ and a 2 $\times$ 1 vector μ, a three dimensional normal distribution $N(\mu, \Sigma)$ can be represented as follows: the density function of this distribution is



$$p_{XY}(x, y; \mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^2|\Sigma|}} e^{-\frac{1}{2}(\mathbf{x}-\mu)^{\mathrm{T}}\Sigma^{-1}(\mathbf{x}-\mu)}$$

where $|\Sigma|$ is the determinant (行列式) of Σ and $\mathbf{x}=(x, y)^{\mathbf{T}}$.

Definition: Positive Semi-Definite Matrix

- Definite Matrix:
  - An $n \times n$ symmetric real matrix A is said to be positive definite if $x^T Ax > 0$ **for all non-zero** $x \in \mathrm{R}^n$
  - An $n \times n$ symmetric real matrix A is said to be negative definite if $x^T Ax < 0$ **for all non-zero** $x \in \mathrm{R}^n$
- Semi-Definite Matrix:
  - An $n \times n$ symmetric real matrix A is said to be positive semi-definite if $x^T Ax \geq 0$ **for all non-zero** $x \in \mathrm{R}^n$
  - An $n \times n$ symmetric real matrix A is said to be negative semi-definite if $x^T Ax \leq 0$ **for all non-zero** $x \in \mathrm{R}^n$

Useful Properties

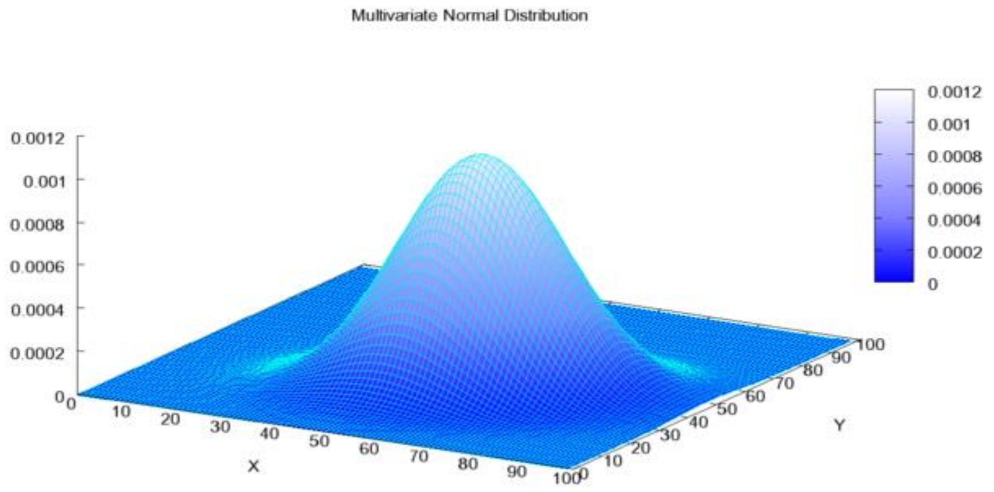
- For Positive Definite Matrix, all the eigenvalues (特征根) $\lambda > 0$.
- For Negative Definite Matrix, all the eigenvalues $\lambda < 0$.
- For Positive Semi-Definite Matrix, all the eigenvalues $\lambda \geq 0$.
- For Negative Semi-Definite Matrix, all the eigenvalues $\lambda \leq 0$.

- These properties are useful to check whether the matrix is definite, semi-definite or not.

Exercises: MLE for high-dimensional Gaussian Distribution

- If μ=(a1,a2) and Σ is a diagonal matrix with eigenvalues λ1, λ2 (λ1 > 0, λ2 > 0),

$$\Sigma = \begin{bmatrix} \lambda 1 & 0 \\ 0 & \lambda 2 \end{bmatrix}$$



then $f(x, y)$ can be writeen as:

$$p_{XY}(x, y; \mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^2 \lambda 1 * \lambda 2}} e^{-\frac{1}{2\lambda 1}(x-a1)^2 - \frac{1}{2\lambda 2}(y-\mathrm{a}2)^2}$$

Two Dimension Gaussian Distribution

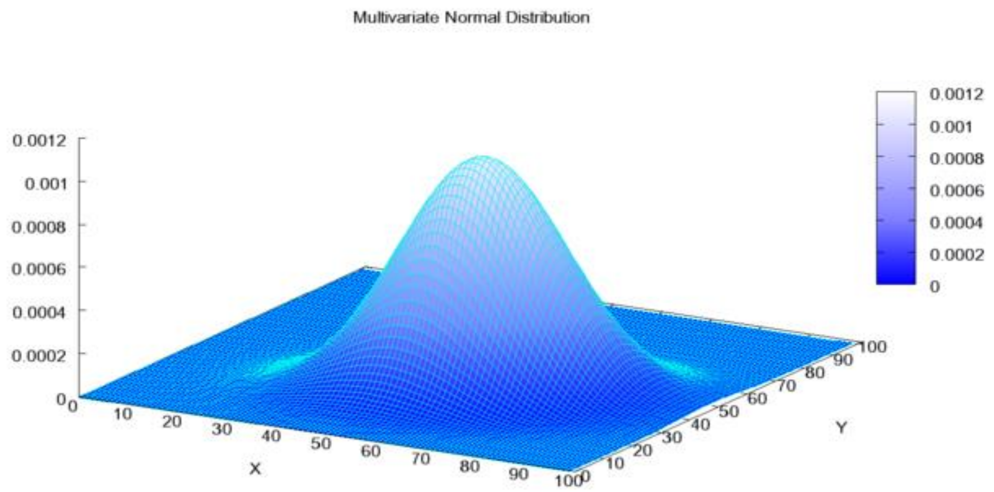
- How to get $p_{XY}(x, y; \mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^2 \lambda 1 * \lambda 2}} e^{-\frac{1}{2\lambda 1}(x-a1)^2 - \frac{1}{2\lambda 2}(y-\mathrm{a}2)^2}$?
- Consider one dimension version:
- $p_X(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi\lambda 1}} e^{-\frac{1}{2\lambda 1}(x-a1)^2}$ (we have $\sigma = \lambda 1, \mu = a1$)
- Σ =$\begin{bmatrix} \lambda 1 & 0 \\ 0 & \lambda 2 \end{bmatrix}$ tells us X and Y are independent.
- Then we have:
- $p_{XY}(x, y; \mu, \Sigma) = \frac{1}{\sqrt{2\pi\lambda 1}} e^{-\frac{1}{2\lambda 1}(x-a1)^2} \times \frac{1}{\sqrt{2\pi\lambda 2}} e^{-\frac{1}{2\lambda 2}(x-a2)^2}$

$$= \frac{1}{\sqrt{(2\pi)^2 \lambda 1 * \lambda 2}} e^{-\frac{1}{2\lambda 1}(x-a1)^2 - \frac{1}{2\lambda 2}(y-\mathrm{a}2)^2}$$

Exercises: MLE for high-dimensional Gaussian Distribution

- If $\Sigma = \begin{bmatrix} \lambda 1 & 0 \\ 0 & \lambda 2 \end{bmatrix}$

and we have n data (x1,y1),...,(xn,yn) sampled from a two-dimensional Gaussian Distribution $N(\mu, \Sigma)$, i.i.d., calculate μ by the maximum likelihood estimation method.



Exercises: MLE for high-dimensional Gaussian Distribution

- Maximum Likelihood (ML) Estimation：

$$\underset{\mu}{\operatorname{argmax}} \ \sum_{\mathrm{i=1}}^{\mathrm{n}} \log p_{XY}(\mathrm{xi}, \mathrm{yi}; \mu, \Sigma)$$

It is equal to

$$\underset{\mathrm{a1,a2}}{\operatorname{argmin}} \sum_{i=1}^{n} \left(\frac{1}{2\lambda 1}(\mathrm{xi} - \mathrm{a1})^2 + \frac{1}{2\lambda 2}(\mathrm{yi} - \mathrm{a2})^2\right)$$

Exercises: MLE for high-dimensional Gaussian Distribution

Derivation the equation $\mathrm{G(a1, a2)} = \sum_{i=1}^{n} (\frac{1}{2\lambda 1}(\mathrm{xi} - \mathrm{a1})^2 + \frac{1}{2\lambda 2}(\mathrm{yi} - \mathrm{a2})^2)$

$$\frac{\partial \mathrm{G}}{\partial \mathrm{a1}} = \sum_{i=1}^{n} \frac{\mathrm{a1} - \mathrm{xi}}{\lambda 1} = 0, \qquad \frac{\partial \mathrm{G}}{\partial \mathrm{a2}} = \sum_{i=1}^{n} \frac{\mathrm{a2} - \mathrm{yi}}{\lambda 2} = 0$$

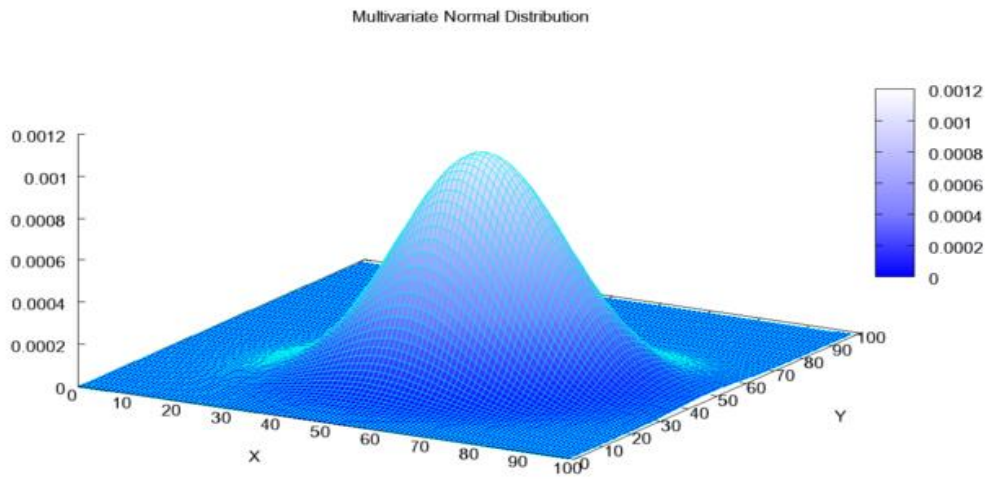
So a1 = $\frac{1}{\mathrm{n}} \sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{xi}$, a2 = $\frac{1}{\mathrm{n}} \sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{yi}$

(As we only need to estimate μ, so the λ will not influence the estimation.)

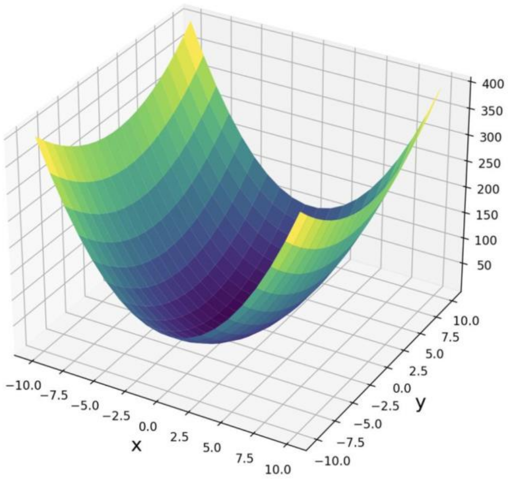
Exercises: MLE for high-dimensional Gaussian Distribution

- In conclusion, the best estimate of the mean of a two-dimensional gaussian distribution is the mean of the sample!

In fact, it also holds for high-dimensional gaussian distribution.



Convex Optimization



Definition of Convex Optimization Problem:

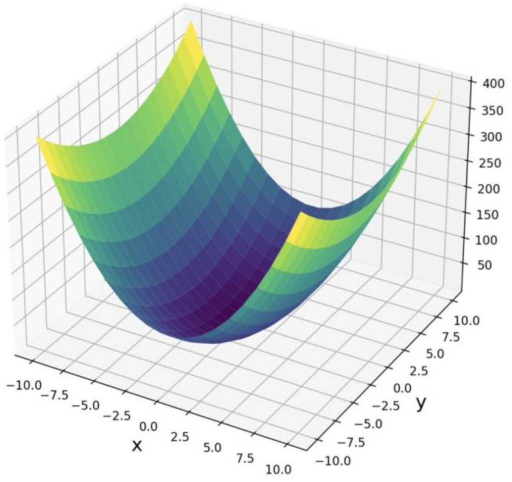
Minimize $f(\mathbf{x})$

Subject to $g_i(\mathbf{x}) \leq 0$, $i=1,\ldots,m$,

$h_j(\mathbf{x}) = 0$, $j=1,\ldots,n$.

- $g_i(\mathbf{x})$ is convex function, $i=1,\ldots,m$
- $h_j(\mathbf{x})$ is linear function $\mathbf{A}_j\mathbf{x}+\mathbf{b}_j$, $j=1,\ldots,n$
- $f(\mathbf{x})$ is a convex function

Convex Optimization



Examples of Convex Optimization:

- Linear optimization belongs to Convex Optimization. Because linear functions are also convex function.

- Minimize $f(\mathbf{x}) = \mathbf{x}^{T}\mathbf{M}\mathbf{x}+\mathbf{b}\mathbf{x}$ is a Convex Optimization problem without constraints, if **M** is a positive semi-definite matrix.

Convex Optimization without Constraints

We first introduce how to address convex optimization without constraints: that is

Minimize f(**x**), where f(**x**) is convex

We want to ask some issues:

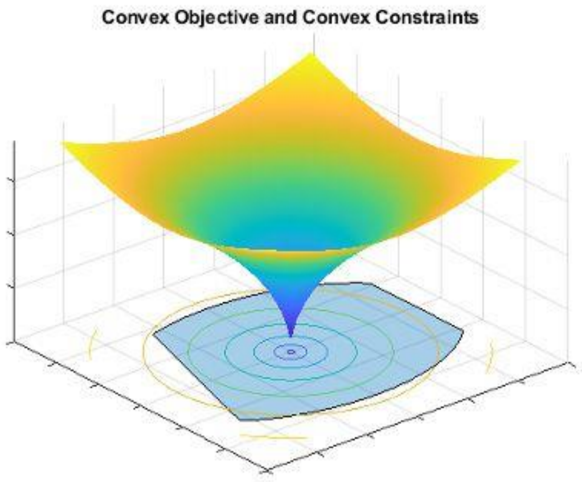
- Issue 1. Whether we can find a solution to this issue?

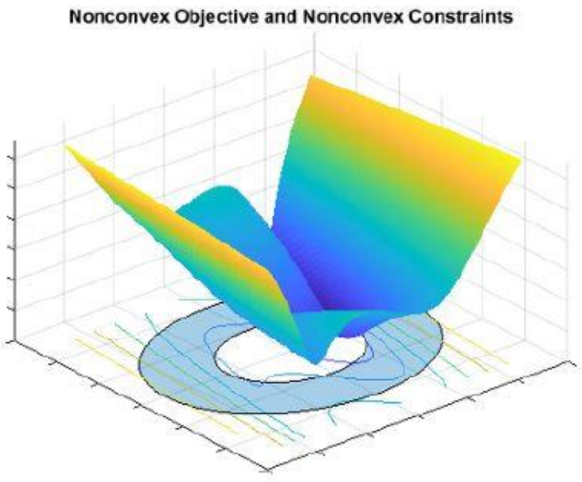
- Issue 2. Whether the solution is unique.

Convex Optimization without Constraints

Exercise: $\mathbf{M} = \begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{pmatrix}, \mathbf{b} = (1,0,0)$

what is the solution of Minimize f(**x**) =$\mathbf{x}^{\mathrm{T}}\mathbf{Mx+bx}$**?**





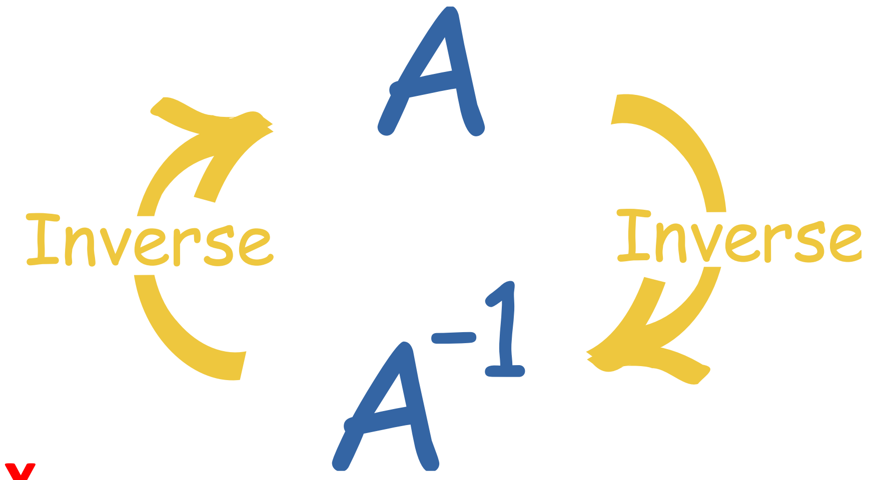
Convex Optimization without Constraints

Solution: Firstly, we note that $\mathbf{M}$ is an inverse matrix.

What is the inverse matrix of $\mathbf{M}$?

The inverse matrix $\mathbf{M}^{-1}$ of $\mathbf{M}$ satisfies that

$$\mathbf{M}^{-1}\mathbf{M} = \mathbf{I}$$, where **I** is the **identity matrix**.



How to compute the **inverse matrix**?

Convex Optimization without Constraints

In matlab, we can use code **inv(M)** to compute the inverse matrix of $\mathbf{M}$.

To understand how to compute the inverse matrix, we need to know the determinant （行列式） of $\mathbf{M}$.

$$\det(\mathbf{M}) = \begin{vmatrix} m_{11} & m_{12} & \cdots & m_{1d} \\ m_{21} & m_{22} & \cdots & m_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ m_{d1} & m_{d2} & \cdots & m_{dd} \end{vmatrix} = \sum_{i=1}^{d} (-1)^{i+j} m_{ij} \det(\mathbf{M}_{\mathrm{ij}})$$

where $\mathbf{M}_{\mathrm{ij}}$ is the (d-1)×(d-1) submatrix obtained by deleting row i and column j from $\mathbf{M}$.

Convex Optimization without Constraints

$$\det(\mathbf{M}) = \begin{vmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{vmatrix} = \sum_{i=1}^{d} (-1)^{i+j}\, m_{ij} \det(\mathbf{M}_{\mathrm{ij}})$$

$$= \ 2\begin{vmatrix} 2 & -1 \\ -1 & 2 \end{vmatrix} + 1\begin{vmatrix} -1 & -1 \\ 0 & 2 \end{vmatrix} + 0\begin{vmatrix} -1 & 2 \\ 0 & -1 \end{vmatrix} + 1\begin{vmatrix} -1 & 0 \\ -1 & 2 \end{vmatrix}$$

$$+ \ 2\begin{vmatrix} 2 & 0 \\ 0 & 2 \end{vmatrix} + 1\begin{vmatrix} 2 & -1 \\ 0 & -1 \end{vmatrix} + 0\begin{vmatrix} -1 & 0 \\ 2 & -1 \end{vmatrix} + 1\begin{vmatrix} 2 & 0 \\ -1 & -1 \end{vmatrix} + 2\begin{vmatrix} 2 & -1 \\ -1 & 2 \end{vmatrix}$$

Now we need to compute the determinant of 2×2 matrix using folloing equations

$$\begin{vmatrix} a & b \\ c & d \end{vmatrix} = ad - bc$$

Convex Optimization without Constraints

$$\begin{vmatrix} 2 & -1 \\ -1 & 2 \end{vmatrix} = 2*2- \ (-1)* \ (-1) \ = 3, \begin{vmatrix} -1 & -1 \\ 0 & 2 \end{vmatrix} = (-1)*2 - 0*(-1) = -2$$

We omit the computing process of other 7 matrics.

$$\det(\mathbf{M}) = \begin{vmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{vmatrix} = 4$$

Before computing the inverse matrix, we need to know how to compute the adjoint matrix (伴随矩阵).

Convex Optimization without Constraints

What is adjoint matrix?

$$\textbf{adj}(\mathbf{M}) = \begin{pmatrix} A_{11} & A_{12} & \cdots & A_{1d} \\ A_{21} & A_{22} & \cdots & A_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ A_{d1} & A_{d2} & \cdots & A_{dd} \end{pmatrix}^{\mathbf{T}}$$

where $A_{ij} = (-1)^{i+j}$

$\mathbf{M}_{ij}$,

here $\mathbf{M}_{ij}$ is the (d-1)×(d-1) submatrix obtained by deleting row i and column j from $\mathbf{M}$.

Convex Optimization without Constraints

Because $\mathbf{M}$ = $\begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{pmatrix}$ **adj(M)** = $\begin{pmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{33} \end{pmatrix}^{\mathbf{T}}$

$$A_{11} = (-1)^2 \begin{vmatrix} 2 & -1 \\ -1 & 2 \end{vmatrix} = 3, \quad A_{12} = (-1)^3 \begin{vmatrix} -1 & -1 \\ 0 & 2 \end{vmatrix} = 2, A_{13} = (-1)^4 \begin{vmatrix} -1 & 2 \\ 0 & -1 \end{vmatrix} = 1,$$

$$A_{21} = (-1)^3 \begin{vmatrix} -1 & 0 \\ -1 & 2 \end{vmatrix} = 2, \quad A_{22} = (-1)^4 \begin{vmatrix} 2 & 0 \\ 0 & 2 \end{vmatrix} = 4, A_{23} = (-1)^5 \begin{vmatrix} 2 & -1 \\ 0 & -1 \end{vmatrix} = 2,$$

$$A_{31} = (-1)^4 \begin{vmatrix} -1 & 0 \\ 2 & -1 \end{vmatrix} = 1, \quad A_{32} = (-1)^5 \begin{vmatrix} 2 & 0 \\ -1 & -1 \end{vmatrix} = 2, A_{33} = (-1)^6 \begin{vmatrix} 2 & -1 \\ -1 & 2 \end{vmatrix} = 3$$

Convex Optimization without Constraints

Then, the adjoint matrix is $\mathbf{adj(M)} = \begin{pmatrix} 3 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 3 \end{pmatrix}$

Finally, the inverse matrix is

$$\mathbf{M}^{-1} = \mathbf{adj(M)}/\det(\mathbf{M})$$

Therefore, the inverse matrix

$$\mathbf{M}^{-1} = \begin{pmatrix} 0.75 & 0.5 & 0.25 \\ 0.5 & 1 & 0.5 \\ 0.25 & 0.5 & 0.75 \end{pmatrix}$$

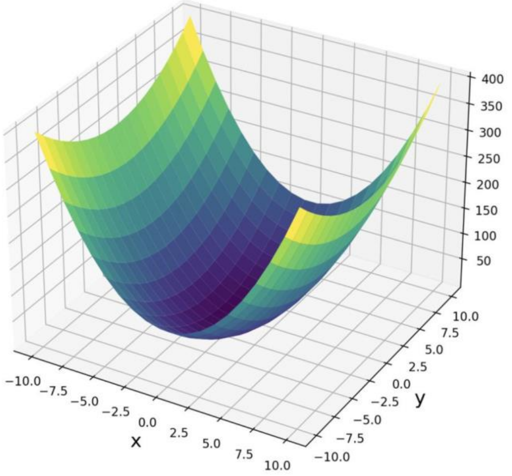
More details about the inverse matrix can be found in ***Lecture 1: Introduction to Linear Algebra: Vectors and Matrices*** and ***Lecture 2: Linear Independence, Rank, and Orthogonality***

Convex Optimization without Constraints

Solution: Firstly, we note that **M** is an inverse matrix, and the inverse is

$$\mathbf{M}^{-1} = \begin{pmatrix} 0.75 & 0.5 & 0.25 \\ 0.5 & 1 & 0.5 \\ 0.25 & 0.5 & 0.75 \end{pmatrix}$$

Secondly, we need to compute the gradient $\nabla f(\mathbf{x})$

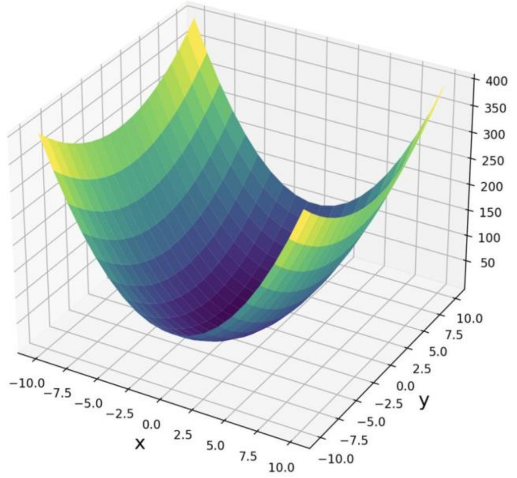


How to compute the gradient?

Convex Optimization without Constraints

- How to compute the gredient?

We introduce the Matrix derivatives:



$$\frac{\partial \mathbf{x}^{\mathrm{T}}\mathbf{M}\mathbf{x}}{\partial \mathbf{x}} = (\mathbf{M} + \mathbf{M}^{\mathrm{T}})\mathbf{x},$$

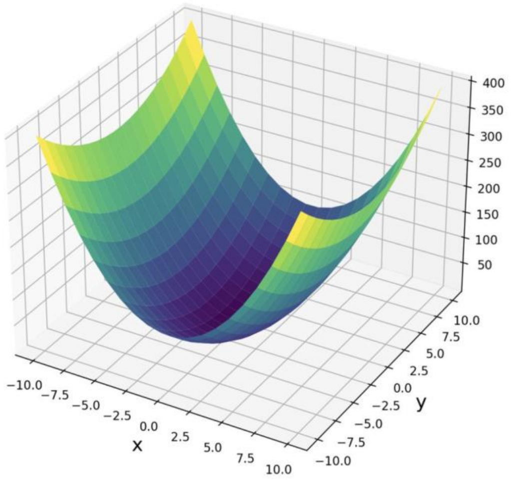
$$\frac{\partial \mathbf{b}\mathbf{x}}{\partial \mathbf{x}} = \mathbf{b}^{\mathrm{T}},$$

More details can be found in https://cloud.tencent.com/developer/article/1551901

Convex Optimization without Constraints

Because **M** is sysmetric, then

$$\frac{\partial \mathbf{x}^{\mathrm{T}}\mathbf{M}\mathbf{x}}{\partial \mathbf{x}} = (\mathbf{M} + \mathbf{M}^{\mathrm{T}})\mathbf{x} \ = \ \mathbf{2Mx}.$$



So we obtain that

$$\nabla \mathrm{f}(\mathbf{x}) = \mathbf{2Mx} + \mathbf{b}^{\mathrm{T}}$$

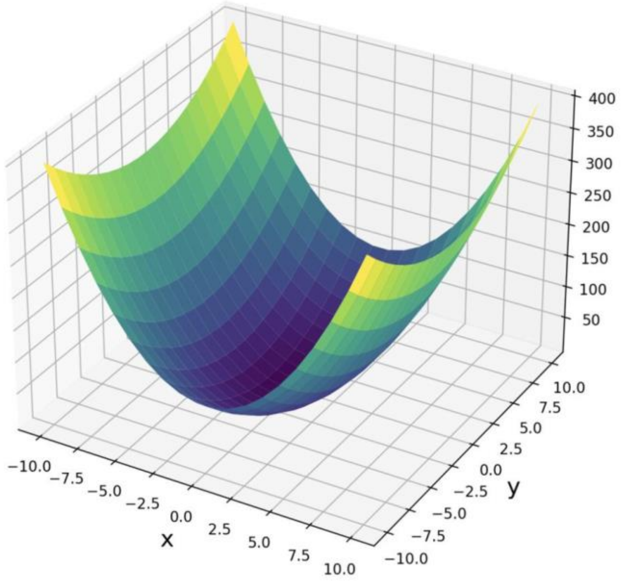
Then according to Theorem 4, the optimal solution sholud satisfy that $\mathbf{2Mx_0} + \mathbf{b}^{\mathrm{T}} \ = \mathbf{0}$

Convex Optimization without Constraints

Then according to Theorem 4, the optimal solution sholud satisfy that $\mathbf{2Mx_0}+\mathbf{b}^{\mathrm{T}} = \mathbf{0}$

The solution of $\mathbf{2Mx_0}+\mathbf{b}^{\mathrm{T}} = \mathbf{0}$ is $\frac{-\mathbf{M}^{-1}\mathbf{b}^{\mathrm{T}}}{2} = \frac{-1}{2}\begin{pmatrix} 0.75 & 0.5 & 0.25 \\ 0.5 & 1 & 0.5 \\ 0.25 & 0.5 & 0.75 \end{pmatrix}\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$

So the optimal solution is $\begin{pmatrix} -3/8 \\ -1/4 \\ -1/8 \end{pmatrix}$



Convex Optimization with Constraints

We next introduce convex optimization with constraints: that is

Minimize $f(\mathbf{x})$,

Subject to $g_i(\mathbf{x}) \leq 0$, $i=1,\ldots,m$,

$h_j(\mathbf{x}) = 0$, $j=1,\ldots,n$.

We want to ask an issue:

- Issue. Whether we can find a solution to this issue? (解的存在性)

Convex Optimization with Constraints

- Issue. Whether we can find a solution to this issue? (解的存在性)

Following theorem gives the answer:

Theorem 8. Assume that f(**x**) is differential, then $\mathbf{x}_0$ is the optimal solution of the Convex optimization problem with constraints if and only if

$$\nabla f(\mathbf{x}_0)^T(\mathbf{y}-\mathbf{x}_0) \geq 0,$$

for all **y** satisfy the constraints.

The proof can be found in Section 4.2.3 in *https://web.stanford.edu/~boyd/cvxbook/bv_cvxbook.pdf*

Convex Optimization with Constraints

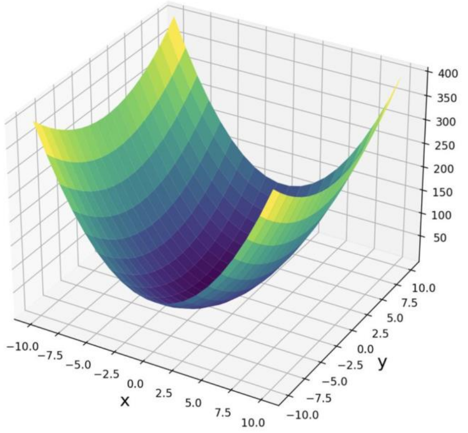
- We consider a conic form problem

Minimize **bx**

Subject to $\mathbf{x}^{\mathrm{T}}\mathbf{M}\mathbf{x} - \mathrm{c} \leq 0,$

where

$$\mathbf{M} = \begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{pmatrix}, \mathbf{b} = (1,0,0), \mathrm{c} = 1$$



Convex Optimization with Constraints

- We firstly transform this problem to a simple form

Note that **M** is positive definite. According to the property of positive definite matrix, **M** can be decomposed as folloing form:

$$\mathbf{M} = \mathrm{U}^{\mathrm{T}}\mathrm{D}\mathrm{U},$$

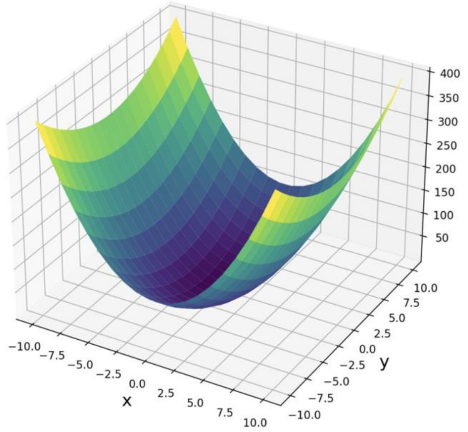
where U is the orthogonal matrix and D is the diagonal matrix whose elements in the diagonal elements are M's eigenvalues.

This decomposition is called singular value decomposition, see *https://en.wikipedia.org/wiki/Singular_value_decomposition*

Convex Optimization with Constraints

- D can be written as follows:

$$\begin{pmatrix} a_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & a_d \end{pmatrix},$$



where a_i >0 (i=1,...,d), because **M** is positive definite.

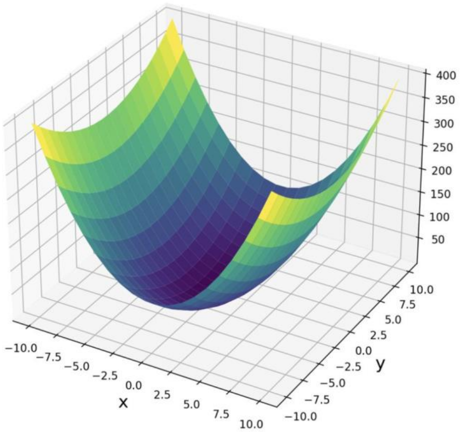
Convex Optimization with Constraints

- D can be written as follows:

$$\sqrt{\mathrm{D}}^{\mathrm{T}}\sqrt{\mathrm{D}},$$

where $\sqrt{\mathrm{D}}$ is the diagonal matrix whose elements in the square roots of diagonal elements

$$\sqrt{\mathrm{D}} = \begin{pmatrix} \sqrt{a_1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sqrt{a_d} \end{pmatrix}$$

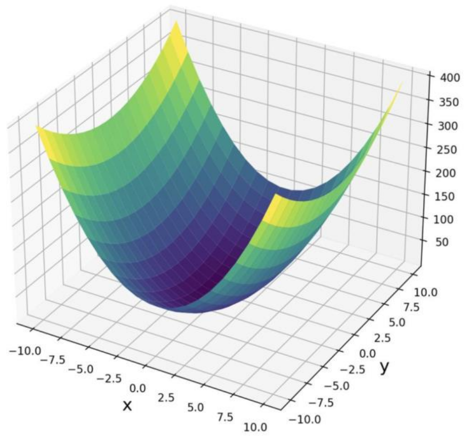


Convex Optimization with Constraints

Then if we set (We transform this problem to a simple form as follows)

$$\mathbf{y} = \sqrt{\mathrm{D}}\,\mathrm{U}\,\mathbf{x}, \text{ so } \mathrm{U}^{\mathrm{T}}\sqrt{\mathrm{D}}^{-1}\,\mathbf{y} = \mathbf{x} \text{ and } \mathbf{x}^{\mathbf{T}}\mathbf{M}\mathbf{x} = \mathbf{y}^{\mathbf{T}}\mathbf{y}$$

Then, the problem will be transformed into

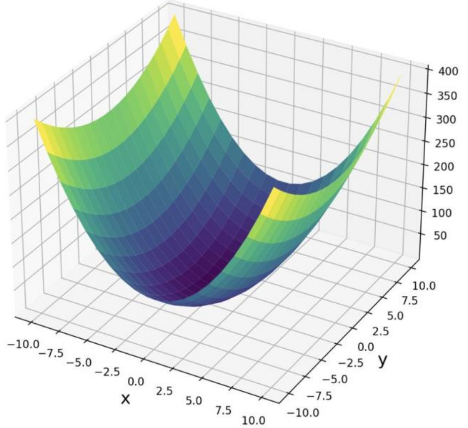


$$\text{Minimize } \mathbf{b}\,\mathrm{U}^{\mathrm{T}}\sqrt{\mathrm{D}}^{-1}\,\mathbf{y}$$

$$\text{Subject to } \mathbf{y}^{\mathbf{T}}\mathbf{y} - \mathrm{c} \leq 0.$$

Convex Optimization with Constraints

Now we address this simpler issue



$$\text{Minimize } \mathbf{b}\,\mathrm{U}^{\mathrm{T}}\sqrt{\mathrm{D}}^{-1}\,\mathbf{y}$$

$$\text{Subject to } \mathbf{y}^{\mathbf{T}}\mathbf{y} - \mathrm{c} \leq 0.$$

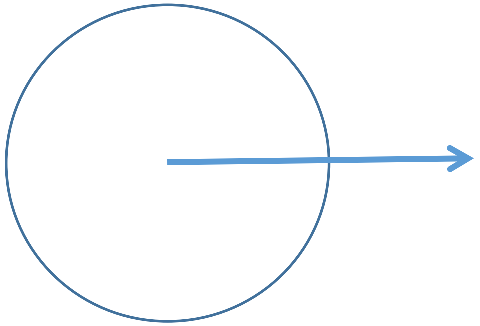
Using **Theorem 8,** we obtain that: if $\mathbf{y_0}$ is the optimal solution, then

$$\nabla \mathrm{f}(\mathbf{y}_0)^{\mathrm{T}}(\mathbf{y} - \mathbf{y}_0) \geq 0, \text{ for all } \mathbf{y} \text{ satisfy } \mathbf{y}^{\mathbf{T}}\mathbf{y} - \mathrm{c} \leq 0.$$

which is equal to $\mathbf{b}\,\mathrm{U}^{\mathrm{T}}\sqrt{\mathrm{D}}^{-1}\,\mathbf{y} \geq \mathbf{b}\,\mathrm{U}^{\mathrm{T}}\sqrt{\mathrm{D}}^{-1}\,\mathbf{y}_0$

Convex Optimization with Constraints

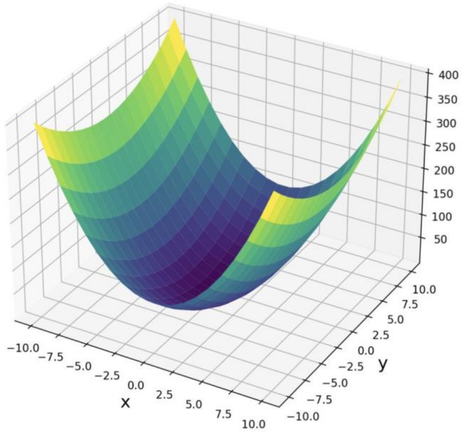
Note that $\mathbf{y}^{\mathbf{T}}\mathbf{y} - \mathrm{c} \leq 0$ means a ball with radius $\sqrt{\mathrm{c}}$.



$\sqrt{\mathrm{D}}^{-1}\mathrm{U}\mathbf{b}^{\mathbf{T}}$

$$\mathbf{b}\,\mathrm{U}^{\mathrm{T}}\sqrt{\mathrm{D}}^{-1}\mathbf{y} \geq \mathbf{b}\,\mathrm{U}^{\mathrm{T}}\sqrt{\mathrm{D}}^{-1}\mathbf{y}_0$$

means that inner product between the optimal solution $\mathbf{y}_0$ and $\mathbf{b}\,\mathrm{U}^{\mathrm{T}}\sqrt{\mathrm{D}}^{-1}$ should be smallest in the ball.

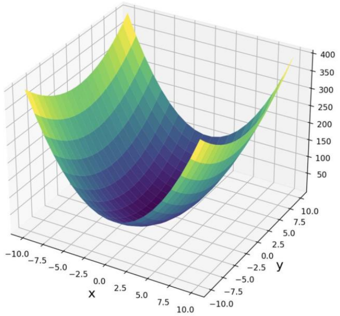


Convex Optimization with Constraints

Cauchy inequality: *https://en.wikipedia.org/wiki/Cauchy%E2%80%93Schwarz_inequality*

$$\mathbf{x}^{\mathrm{T}}\,\mathbf{y} \geq -\,\|\mathbf{x}\|\|\mathbf{y}\| \quad \textbf{(}\|*\|\textbf{ is L2 norm)}$$

and $\mathbf{x}^{\mathrm{T}}\mathbf{y} = -\|\mathbf{x}\|\|\mathbf{y}\|$ if and only if **x = -**k**y,** where k is any positive constant.



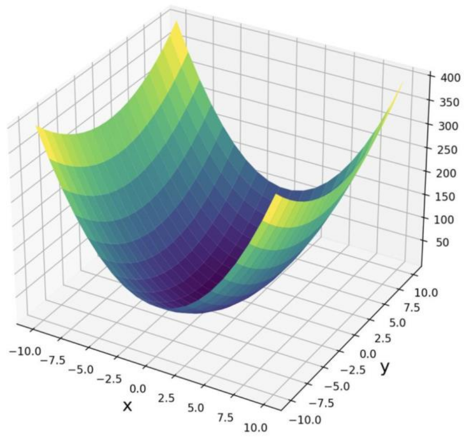
Using Cauchy inequality

$$\mathbf{b}\,\mathrm{U}^{\mathrm{T}}\sqrt{\mathrm{D}}^{-1}\mathbf{y} \geq \mathbf{b}\,\mathrm{U}^{\mathrm{T}}\sqrt{\mathrm{D}}^{-1}\mathbf{y}_0 \geq -\left\|\mathbf{b}\,\mathrm{U}^{\mathrm{T}}\sqrt{\mathrm{D}}^{-1}\right\|\|\mathbf{y}_0\| \geq -\sqrt{c}\left\|\mathbf{b}\,\mathrm{U}^{\mathrm{T}}\sqrt{\mathrm{D}}^{-1}\right\|$$

Convex Optimization with Constraints

So

$$\mathbf{b}\,\mathrm{U}^{\mathrm{T}}\sqrt{\mathrm{D}}^{-1}\mathbf{y} \geq \; -\sqrt{\mathrm{c}}\;\left\|\mathbf{b}\,\mathrm{U}^{\mathrm{T}}\sqrt{\mathrm{D}}^{-1}\right\|$$



If we take $\mathbf{y}_0 = -\sqrt{\mathrm{c}}\sqrt{\mathrm{D}}^{-1}\mathrm{U}\mathbf{b}^{\mathbf{T}} / \left\|\mathbf{b}\,\mathrm{U}^{\mathrm{T}}\sqrt{\mathrm{D}}^{-1}\right\|$,

Then, it is clear that $\mathbf{b}\,\mathrm{U}^{\mathrm{T}}\sqrt{\mathrm{D}}^{-1}\mathbf{y}_0 = -\sqrt{\mathrm{c}}\;\left\|\mathbf{b}\,\mathrm{U}^{\mathrm{T}}\sqrt{\mathrm{D}}^{-1}\right\|$.

So $\mathbf{y}_0 = \; -\sqrt{\mathrm{c}}\sqrt{\mathrm{D}}^{-1}U\mathbf{b}^{\mathbf{T}} / \left\|\mathbf{b}\,\mathrm{U}^{\mathrm{T}}\sqrt{\mathrm{D}}^{-1}\right\|$ is the optimal solution.

Convex Optimization with Constraints

So $\mathbf{x}_0 = \mathrm{U}^{\mathrm{T}}\sqrt{\mathrm{D}}^{-1}\,\mathbf{y}_0 = -\sqrt{\mathrm{c}}\,\mathrm{U}^{\mathrm{T}}\sqrt{\mathrm{D}}^{-1}\sqrt{\mathrm{D}}^{-1}\mathrm{U}\mathbf{b}^{\mathbf{T}} / \left\|\mathbf{b}\,\mathrm{U}^{\mathrm{T}}\sqrt{\mathrm{D}}^{-1}\right\|$

$$= -\sqrt{\mathrm{c}}\,\mathbf{M}^{-1}\mathbf{b}^{\mathbf{T}} / \left\|\mathbf{b}\,\mathrm{U}^{\mathrm{T}}\sqrt{\mathrm{D}}^{-1}\right\|$$

$$\left\|\mathbf{b}\,\sqrt{\mathrm{D}}^{-1}\mathrm{U}^{\mathrm{T}}\right\| = \sqrt{\mathbf{b}\,\mathrm{U}^{\mathrm{T}}\sqrt{\mathrm{D}}^{-1}\,\sqrt{\mathrm{D}}^{-1}\mathrm{U}\mathbf{b}^{\mathrm{T}}} = \sqrt{\mathbf{b}\mathbf{M}^{-1}\mathbf{b}^{\mathbf{T}}}$$

So, the optimal solution should be $-\sqrt{\mathrm{c}}\,\mathbf{M}^{-1}\mathbf{b}^{\mathbf{T}} / \sqrt{\mathbf{b}\mathbf{M}^{-1}\mathbf{b}^{\mathbf{T}}}$

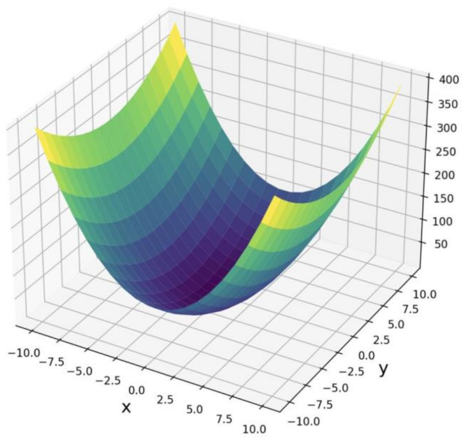
Convex Optimization with Constraints

$$\mathbf{M^{-1}} = \begin{pmatrix} 0.75 & 0.5 & 0.25 \\ 0.5 & 1 & 0.5 \\ 0.25 & 0.5 & 0.75 \end{pmatrix}, \mathbf{b} = (1,0,0), \mathrm{c} = 1$$

So the solution

$$-\sqrt{\mathrm{c}}\,\mathbf{M}^{-1}\mathbf{b}^{\mathbf{T}}/\sqrt{\mathbf{b}\mathbf{M}^{-1}\mathbf{b}^{\mathbf{T}}}$$

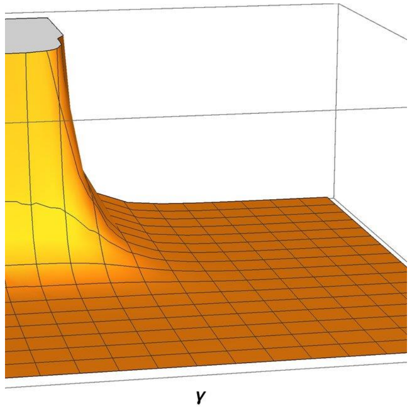
is equal to $\begin{pmatrix} -\frac{\sqrt{3}}{2} \\ -\frac{\sqrt{3}}{3} \\ -\frac{\sqrt{3}}{6} \end{pmatrix}$

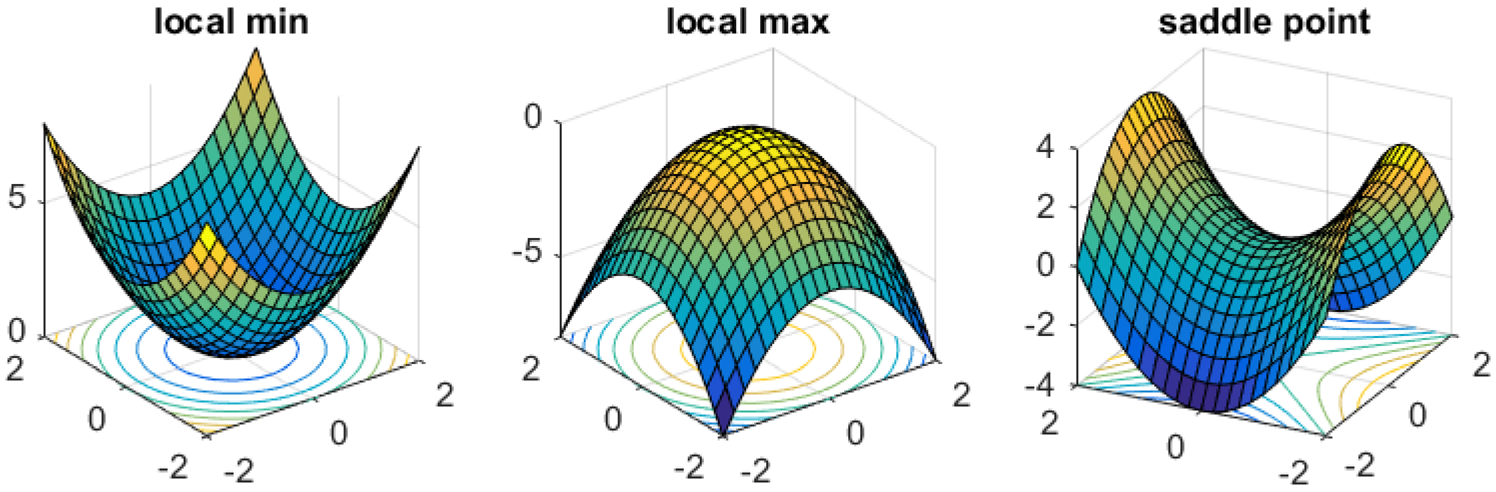


Non-Convex Optimization

Why is non-convex optimization hard?

- Potentially many local minimal points
- Saddle points
- Very flat regions





Credits from Dr. Christopher De Sa

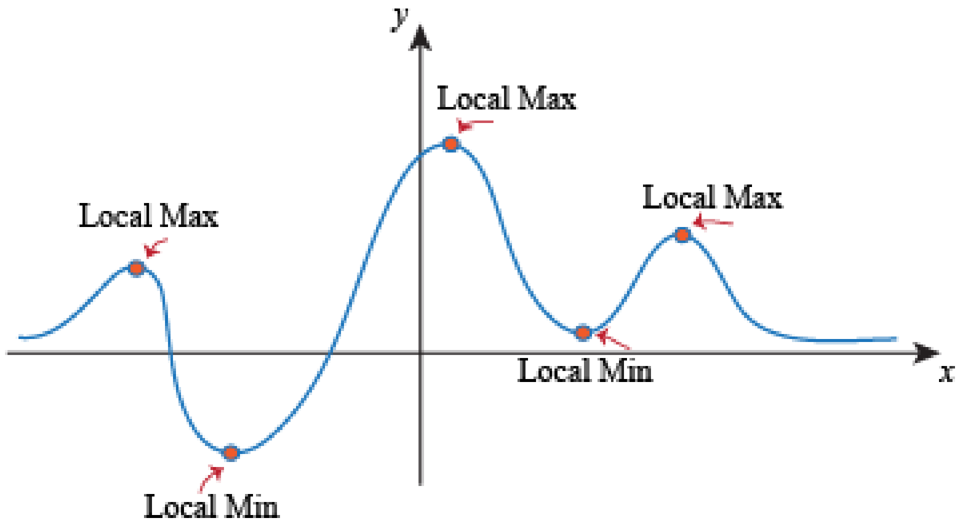
Non-Convex Optimization

- **Global minimum point:**

A real-valued function f defined on a domain Ω has a global maximum point at x∗, if f(x∗) ≤ f(x) for all x in Ω.

- **Local minimum point:**

A real-valued function f defined on a domain Ω has a local maximum point at x∗, if there exists some ε > 0 such that f(x∗) ≤ f(x) for all x in Ω within distance ε of x∗



Credits from Wikipedia

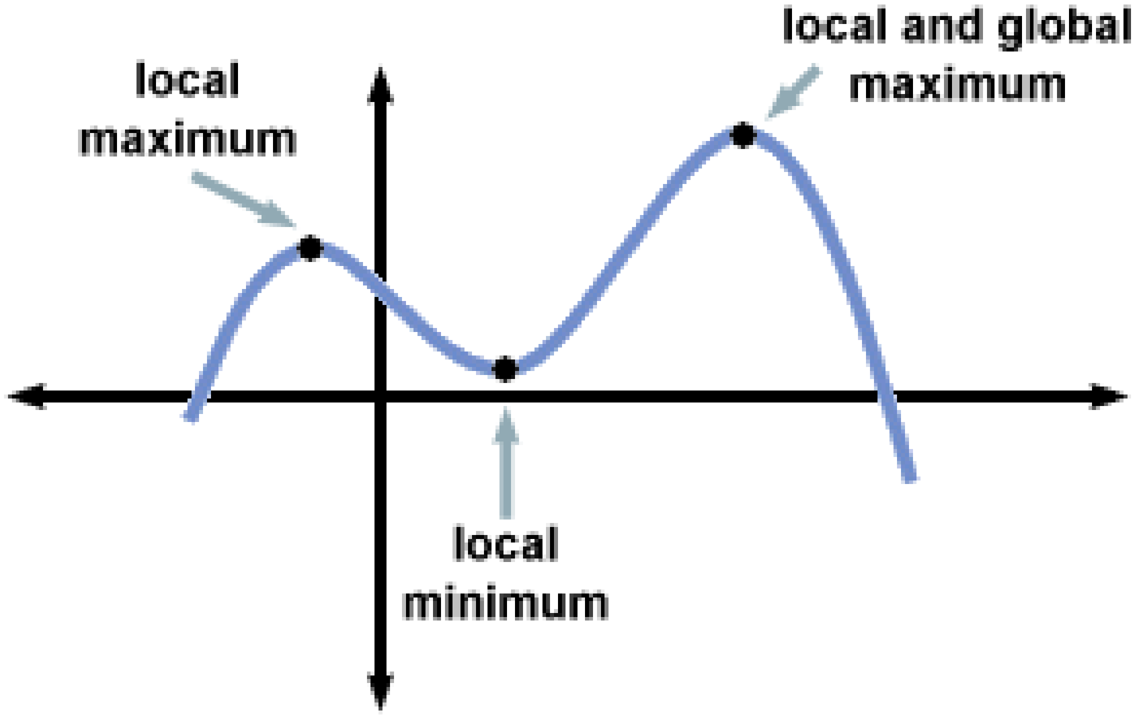
Non-Convex Optimization

- **Global maximum point:**

A real-valued function f defined on a domain Ω has a global maximum point at $x*$, if $f(x*) \geq f(x)$ for all x in Ω.

- **Local maximum point:**

A real-valued function f defined on a domain Ω has a local maximum point at $x*$, if there exists some $\varepsilon > 0$ such that $f(x*) \geq f(x)$ for all x in Ω within distance ε of $x*$

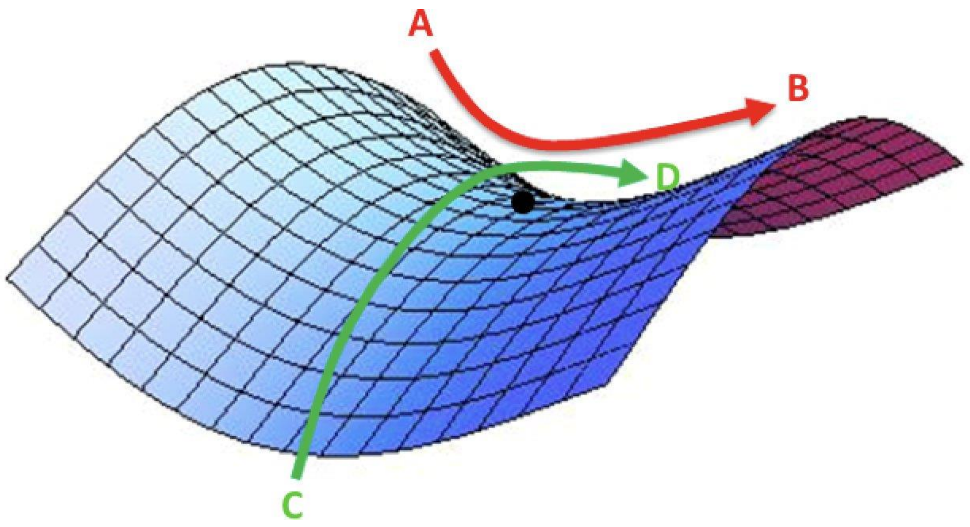


Credits from Wikipedia

Non-Convex Optimization

- **Saddle Point:**

A saddle point or minimax point is a point on the surface of the graph of a function where the slopes (derivatives) in orthogonal directions are all zero (a critical point), but which is not a local extremum (local minimal point or local maximal point) of the function.



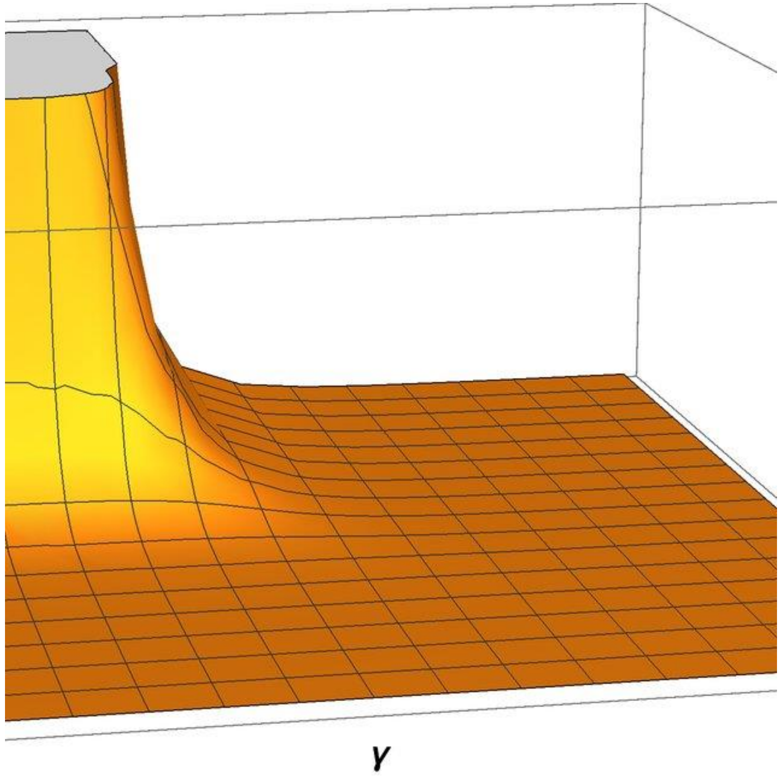
For example, x^2-y^2 (see the figure). (0,0) is a saddle point, because the gradient at (0,0) is zero, but it is not the local extremum

Credits from Wikipedia

Non-Convex Optimization

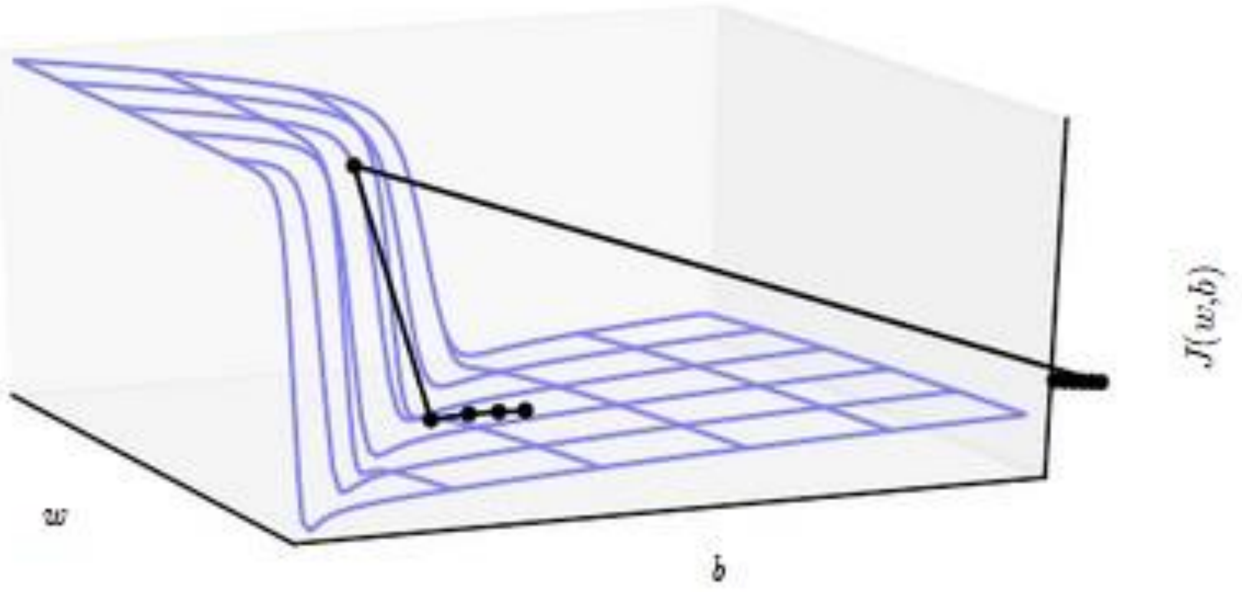
- **Very Flat Regions:**

Very Flat Regions is an area on the surface of the graph of a function where the gredients (derivatives) in orthogonal directions are very close to zero, but which is not a local extremum of the function.



Credits from Wikipedia

Cliffs and Exploding Gradients



Neural networks with many layers will have cliffs and exploding gradients. Therefore, gradient clipping is useful

(Credits from Deep Learning Book)

Non-Convex Optimization

How to solve non-convex problems?

- Gradient descent
- Stochastic gradient descent https://en.wikipedia.org/wiki/Stochastic_gradient_descent
- Adaptive gradient algorithm https://conferences.mpi-inf.mpg.de/adfocs/material/alina/adaptive-L1.pdf
- RMSprop https://optimization.cbe.cornell.edu/index.php?title=RMSProp
- Momentum https://en.wikipedia.org/wiki/Momentum

Credits from Dr. Christopher De Sa

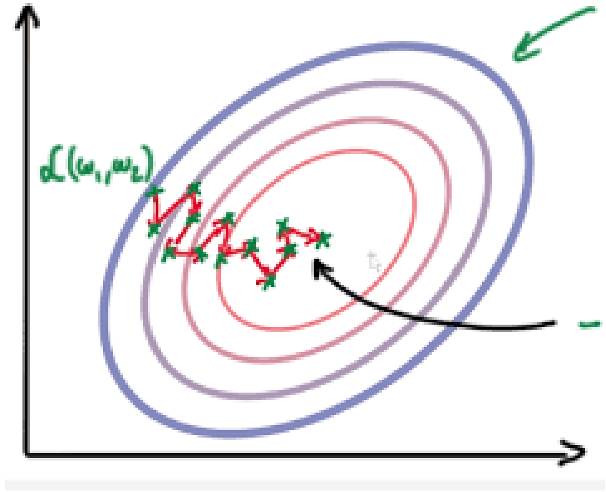
Gradient Descent

- Consider the following non-convex optimization problem:

$$\text{Minimize } f(\mathbf{x}),$$

where f is non-convex and is defined in R^{d}.

Given an inital point $\mathbf{x_0}$ (which is also called initial weight)



Then the next updated point should be

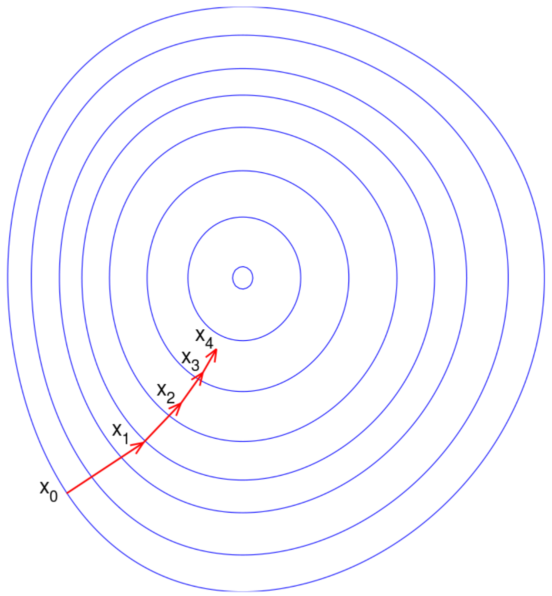
$$\mathbf{x_1} = \mathbf{x_0} - t\nabla f(\mathbf{x_0})$$

where t (t>0) is called learning rate.

Gradient Descent

After we obtain the n-th weight $\mathbf{x_n}$, then the (n+1)-th weight $\mathbf{x_{n+1}}$ is

$$\mathbf{x_{n+1}} = \mathbf{x_n} - t\nabla f(\mathbf{x_n})$$



Motivation 1. We hope the final point $\mathbf{x}$ will get close to the a critical point $\nabla f(\mathbf{x})=0$.

Motivation 2. to take repeated steps in the opposite direction of the gradient (or approximate gradient) of the function at the current point, because this is the direction of steepest descent (下降最快的方向).

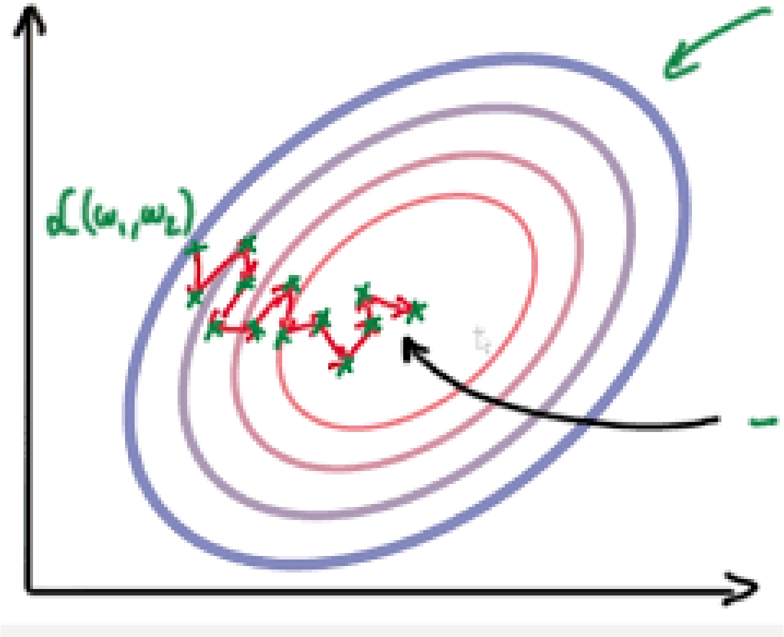
- Doesn't necessarily go towards optimal point.

Gradient Descent

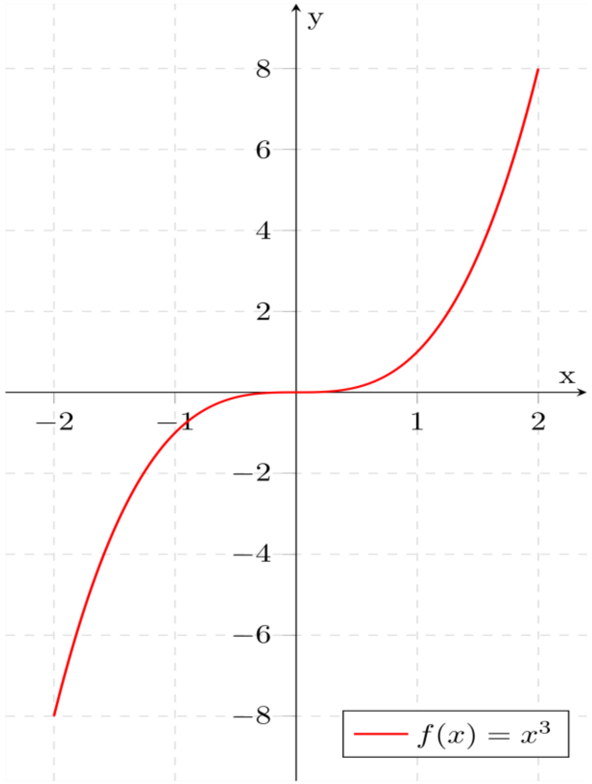
Exercises

- Minimize f(x), where f(x) = x^3.

Initial weight x_0 = 1, learning rate t = 1/3. Then what is the convergent point?



Gradient Descent



Solution

- Minimize f(x), where f(x) = x^3.

Then the grendent descent formula is

$$x_{n+1} = x_n - 3tx_n^{\ 2}$$

Because x_0= 1, t= 1/3, then

$$x_1=0$$

Note that $\nabla f(\mathbf{x_1})$ = 0, so x_2=x_1, which implies $x_{n+1} = x_1 = 0$.

So 0 is the convergent point. But 0 is a saddle point of f(x).

Stochastic Gradient Descent

Algorithm 8.1 Stochastic gradient descent (SGD) update

Require: Learning rate schedule $\epsilon_1, \epsilon_2, \dots$
Require: Initial parameter $\boldsymbol{\theta}$
  $k \leftarrow 1$
  while stopping criterion not met **do**
    Sample a minibatch of m examples from the training set $\{\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(m)}\}$ with corresponding targets $\boldsymbol{y}^{(i)}$.
    Compute gradient estimate: $\hat{\boldsymbol{g}} \leftarrow \frac{1}{m} \nabla_{\boldsymbol{\theta}} \sum_i L(f(\boldsymbol{x}^{(i)}; \boldsymbol{\theta}), \boldsymbol{y}^{(i)})$
    Apply update: $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \epsilon_k \hat{\boldsymbol{g}}$
    $k \leftarrow k + 1$
  end while

Sample i uniformly from $\{1, \cdots, n\}$, and update θ by

$$\theta = \theta - \epsilon \nabla_\theta J^{(i)}(\theta)$$

AdaGrad

Algorithm 8.4 The AdaGrad algorithm

Require: Global learning rate ϵ
Require: Initial parameter $\boldsymbol{\theta}$
Require: Small constant δ, perhaps 10^{-7}, for numerical stability
Initialize gradient accumulation variable $\boldsymbol{r} = \boldsymbol{0}$
while stopping criterion not met **do**
Sample a minibatch of m examples from the training set $\{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(m)}\}$ with corresponding targets $\boldsymbol{y}^{(i)}$.
Compute gradient: $\boldsymbol{g} \leftarrow \frac{1}{m} \nabla_{\boldsymbol{\theta}} \sum_i L(f(\boldsymbol{x}^{(i)}; \boldsymbol{\theta}), \boldsymbol{y}^{(i)})$.
Accumulate squared gradient: $\boldsymbol{r} \leftarrow \boldsymbol{r} + \boldsymbol{g} \odot \boldsymbol{g}$.
Compute update: $\Delta\boldsymbol{\theta} \leftarrow -\frac{\epsilon}{\delta + \sqrt{\boldsymbol{r}}} \odot \boldsymbol{g}$. (Division and square root applied element-wise)
Apply update: $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \Delta\boldsymbol{\theta}$.
end while

$$g_t = \nabla_{\theta_t} J(\theta_t)$$

AdaGrad: $\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{G_t + \epsilon}} \odot g_t$

(Credits from Deep Learning Book)

RMSProp

Algorithm 8.5 The RMSProp algorithm

Require: Global learning rate ϵ, decay rate ρ
Require: Initial parameter $\boldsymbol{\theta}$
Require: Small constant δ, usually 10^{-6}, used to stabilize division by small numbers
  Initialize accumulation variables $\boldsymbol{r} = 0$
  while stopping criterion not met **do**
    Sample a minibatch of m examples from the training set $\{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(m)}\}$ with corresponding targets $\boldsymbol{y}^{(i)}$.
    Compute gradient: $\boldsymbol{g} \leftarrow \frac{1}{m} \nabla_{\boldsymbol{\theta}} \sum_i L(f(\boldsymbol{x}^{(i)}; \boldsymbol{\theta}), \boldsymbol{y}^{(i)})$.
    Accumulate squared gradient: $\boldsymbol{r} \leftarrow \rho \boldsymbol{r} + (1 - \rho) \boldsymbol{g} \odot \boldsymbol{g}$.
    Compute parameter update: $\Delta \boldsymbol{\theta} = -\frac{\epsilon}{\sqrt{\delta + \boldsymbol{r}}} \odot \boldsymbol{g}$. ($\frac{1}{\sqrt{\delta + \boldsymbol{r}}}$ applied element-wise)
    Apply update: $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \Delta \boldsymbol{\theta}$.
  end while

Changing the gradient accumulation into an exponentially weighted moving average

(Credits from Deep Learning Book)

Adam

Algorithm 8.7 The Adam algorithm

Require: Step size ϵ (Suggested default: 0.001)
Require: Exponential decay rates for moment estimates, ρ_1 and ρ_2 in $[0, 1)$. (Suggested defaults: 0.9 and 0.999 respectively)
Require: Small constant δ used for numerical stabilization (Suggested default: 10^{-8})
Require: Initial parameters $\boldsymbol{\theta}$
Initialize 1st and 2nd moment variables $\boldsymbol{s} = \boldsymbol{0}$, $\boldsymbol{r} = \boldsymbol{0}$
Initialize time step $t = 0$
while stopping criterion not met **do**
Sample a minibatch of m examples from the training set $\{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(m)}\}$ with corresponding targets $\boldsymbol{y}^{(i)}$.
Compute gradient: $\boldsymbol{g} \leftarrow \frac{1}{m}\nabla_{\boldsymbol{\theta}} \sum_i L(f(\boldsymbol{x}^{(i)}; \boldsymbol{\theta}), \boldsymbol{y}^{(i)})$
$t \leftarrow t + 1$
Update biased first moment estimate: $\boldsymbol{s} \leftarrow \rho_1 \boldsymbol{s} + (1 - \rho_1)\boldsymbol{g}$
Update biased second moment estimate: $\boldsymbol{r} \leftarrow \rho_2 \boldsymbol{r} + (1 - \rho_2)\boldsymbol{g} \odot \boldsymbol{g}$
Correct bias in first moment: $\hat{\boldsymbol{s}} \leftarrow \frac{\boldsymbol{s}}{1-\rho_1^t}$
Correct bias in second moment: $\hat{\boldsymbol{r}} \leftarrow \frac{\boldsymbol{r}}{1-\rho_2^t}$
Compute update: $\Delta\boldsymbol{\theta} = -\epsilon \frac{\hat{\boldsymbol{s}}}{\sqrt{\hat{\boldsymbol{r}}}+\delta}$ (operations applied element-wise)
Apply update: $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \Delta\boldsymbol{\theta}$
end while

$$g_t = \nabla_{\theta_t} J(\theta_t)$$

Adam: $\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{\hat{v}_t + \epsilon}} \hat{m}_t$,

where $\hat{m}_t = \frac{m_t}{1-\beta_1^t}$

and $\hat{v}_t = \frac{v_t}{1-\beta_2^t}$

$m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t$

and $v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t^2$

Using Adam when you are unfamiliar about optimization; while using momentum SGD when you are familiar about optimization

(Credits from Deep Learning Book)

Thank You!