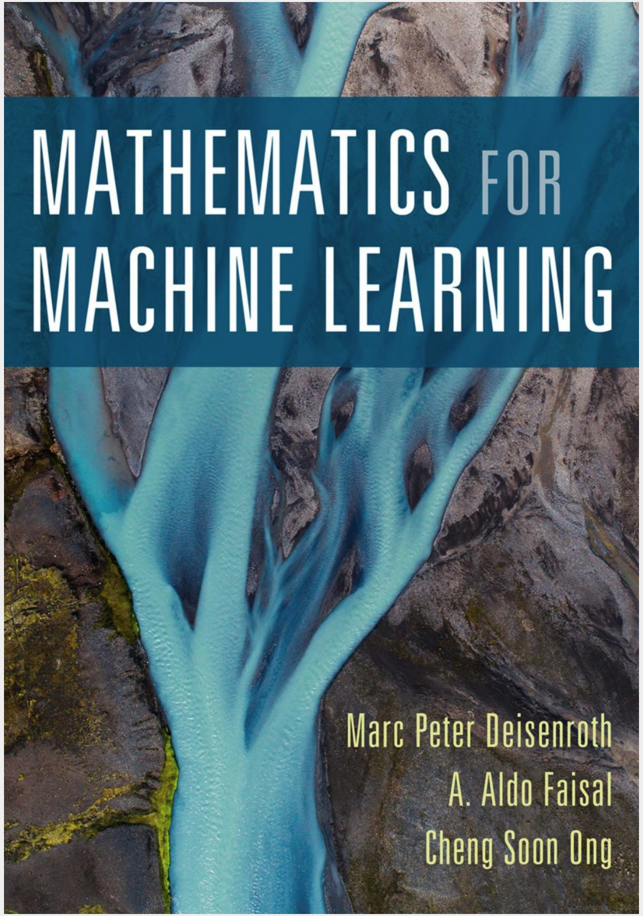
CS365
Foundations of Data Science

Vector Calculus and Optimization

Charalampos E. Tsourakakis
ctsourak@bu.edu

Chapters 5 and 7
Vector calculus



Plotting f:R²→R

Consider a vector p=[x,y].

- How do we plot functions of p such as the following:

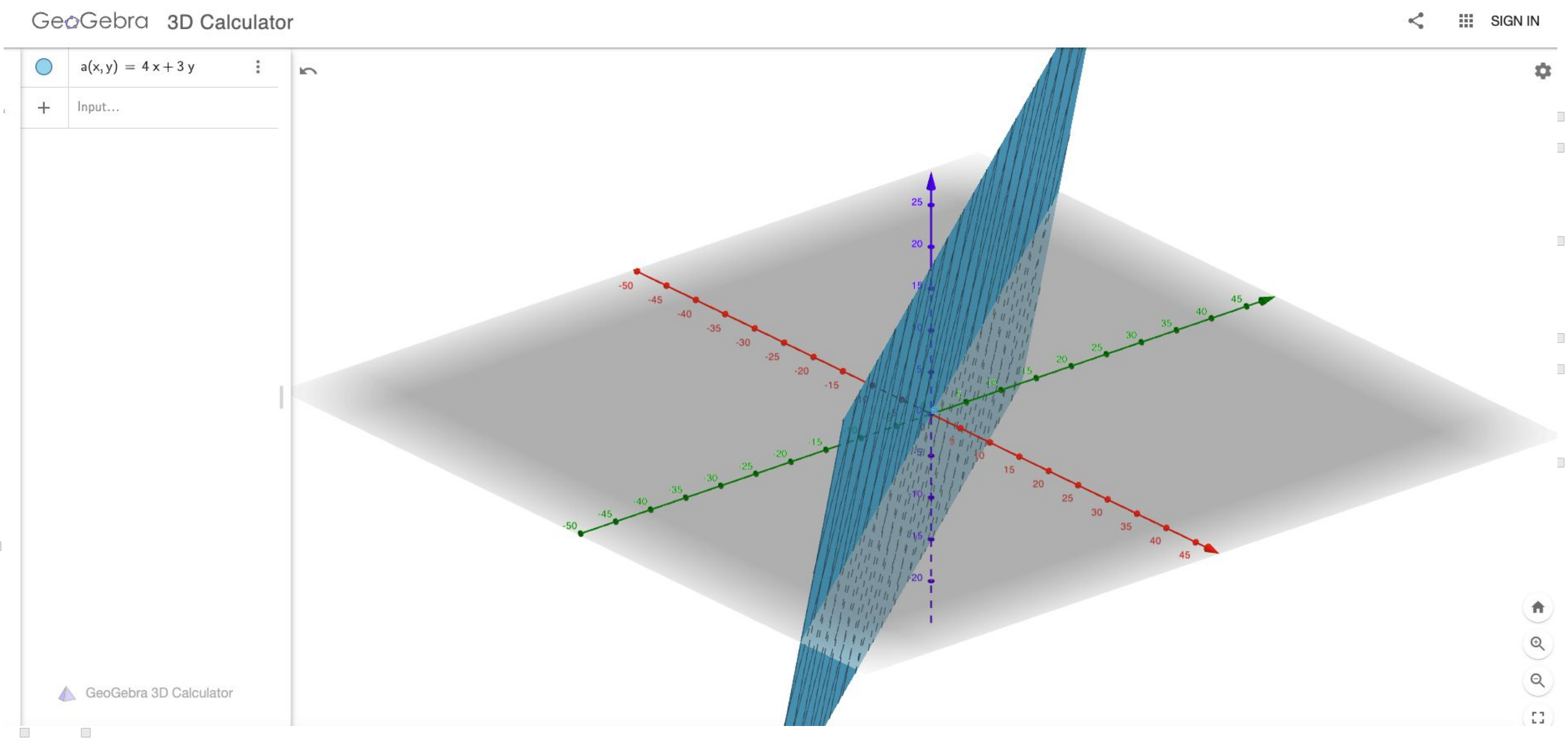
$$z = [4,3]p = 4x + 3y$$

$$z = p^T p = x^2 + y^2$$

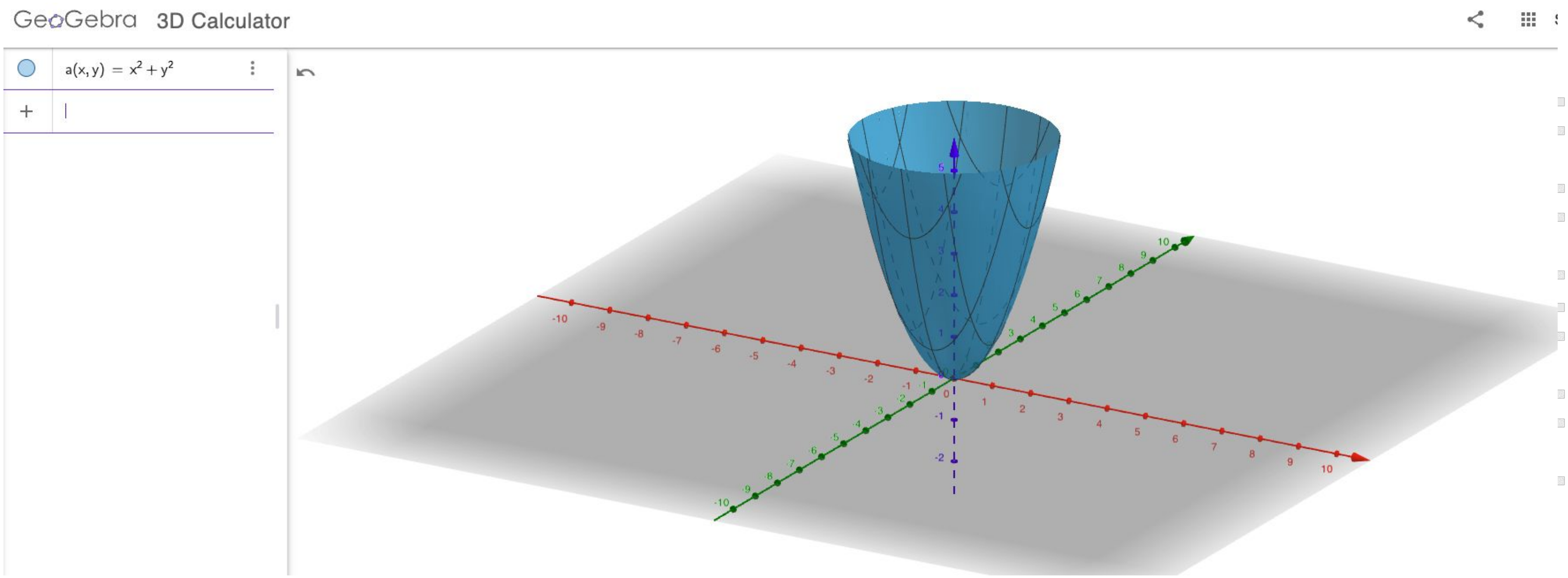
$$z = p^T A p = [x,\ y]\begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix} = -x^2 + y^2$$

$$z = p^T A p = [x,\ y]\begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix} = 2x^2 + y^2$$

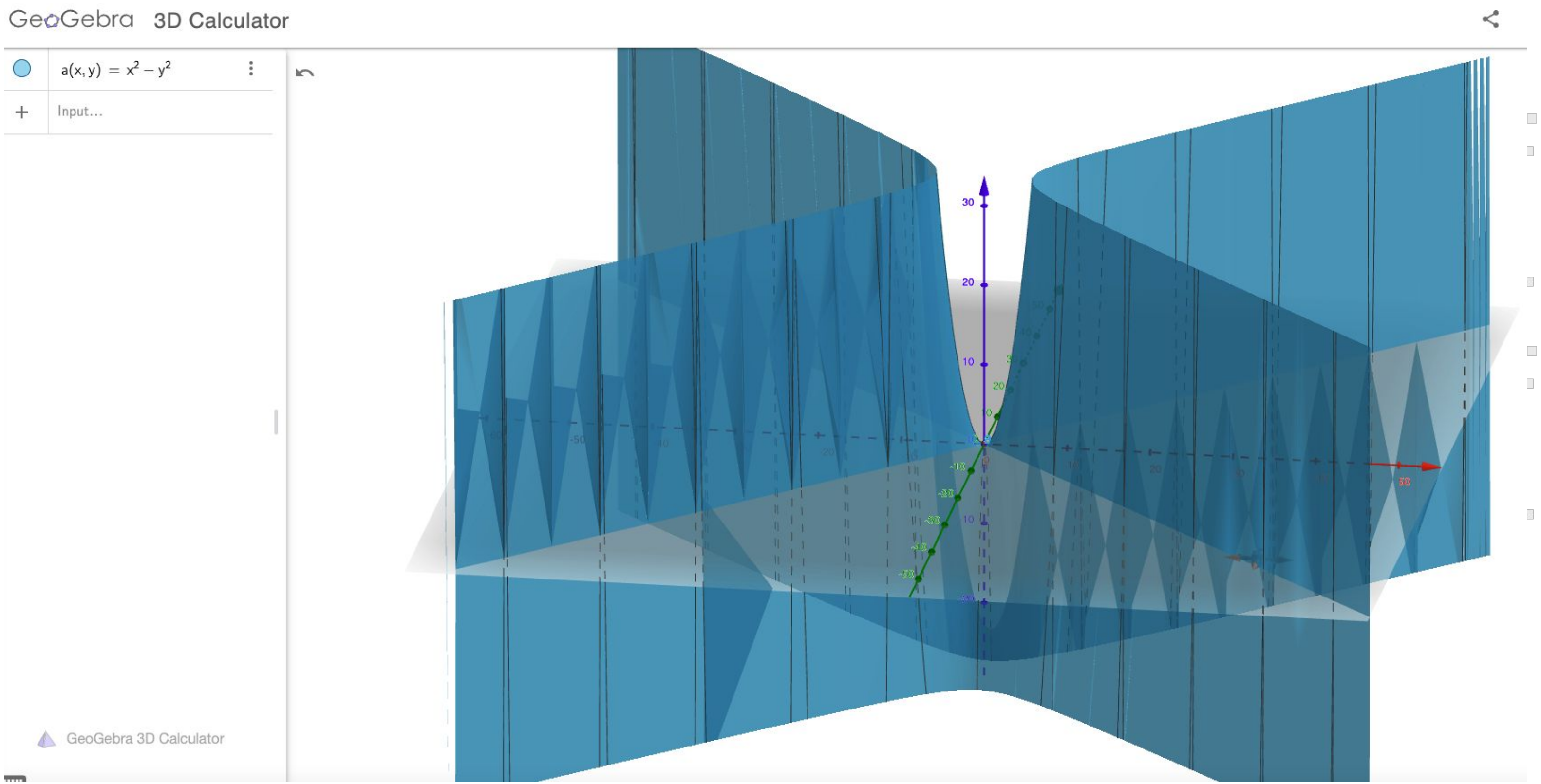
z=4x+3y


GeoGebra 3D Calculator
SIGN IN
a(x, y) = 4 x + 3 y
Input...
GeoGebra 3D Calculator

$z=x^2+y^2$

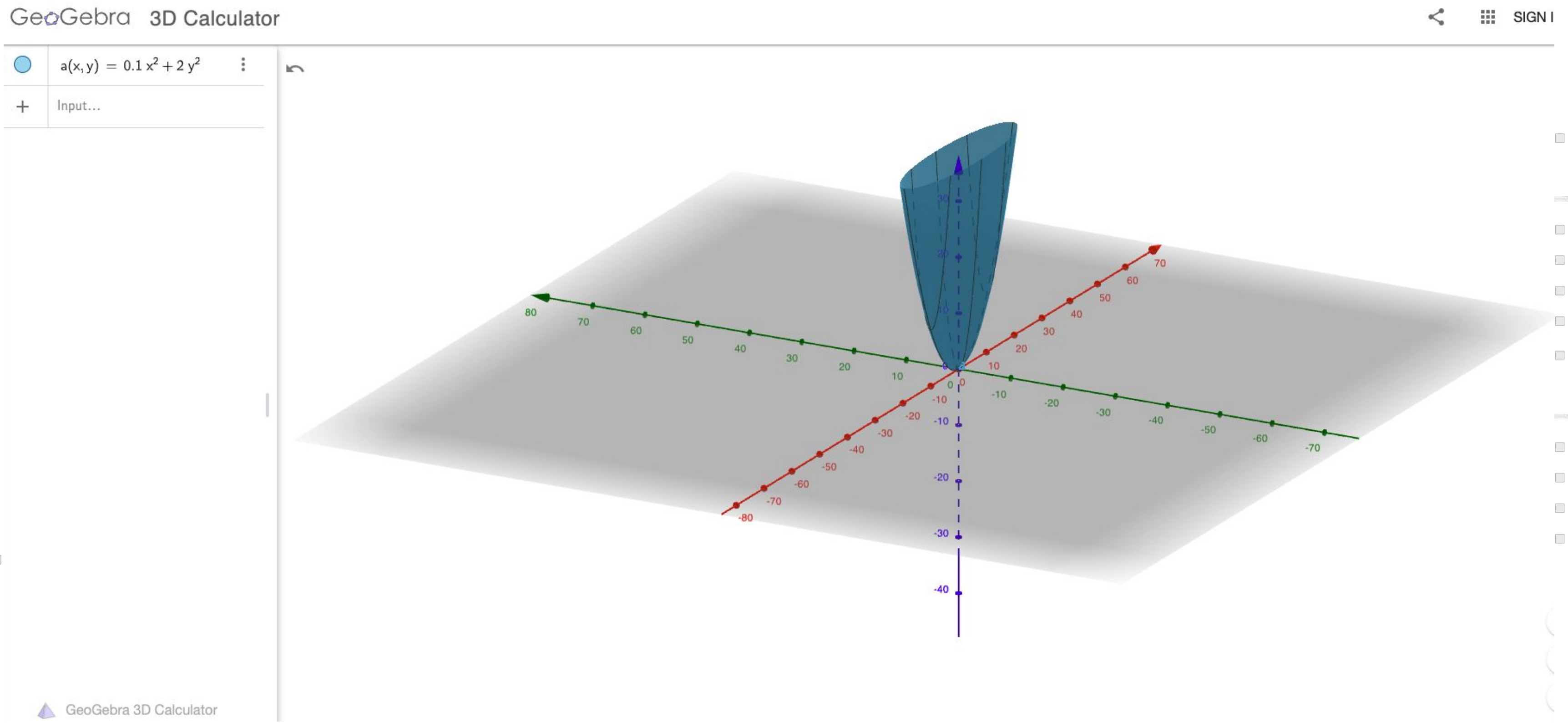


$z=x^2-y^2$



GeoGebra 3D Calculator
a(x, y) = x² − y²
Input...
30
20
10
GeoGebra 3D Calculator

$z=0.1x^2+2y^2$



GeoGebra 3D Calculator
SIGN I
a(x, y) = 0.1 x² + 2 y²
Input...
GeoGebra 3D Calculator

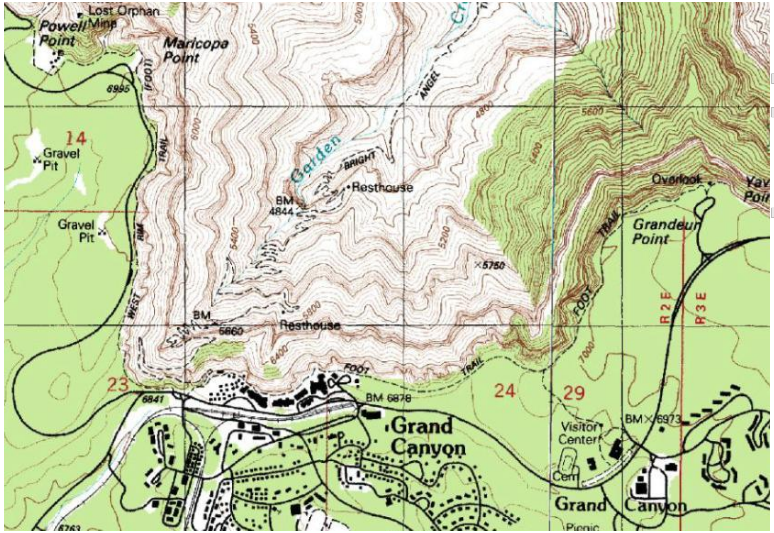
Level curves

The level curves of a function f of two variables x,y are the curves with equation

$$f(x, y) = c$$

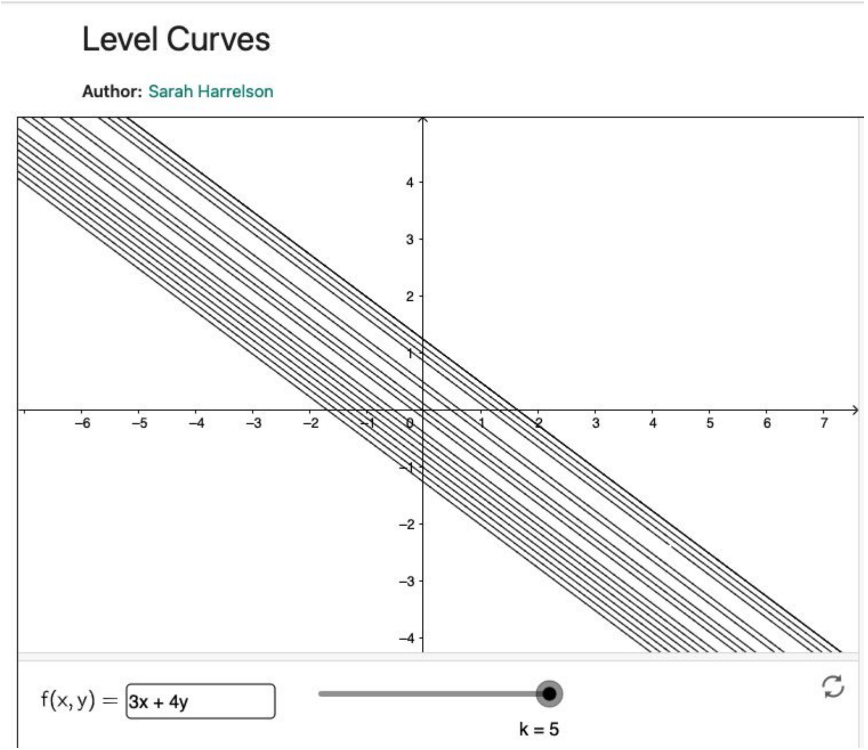
where c is a constant in the range of f.

Constant elevation curves of Grand Canyon (source here)



Geogebra calculator

Online examples : https://www.geogebra.org/m/M2P4KsRe, see also desmos

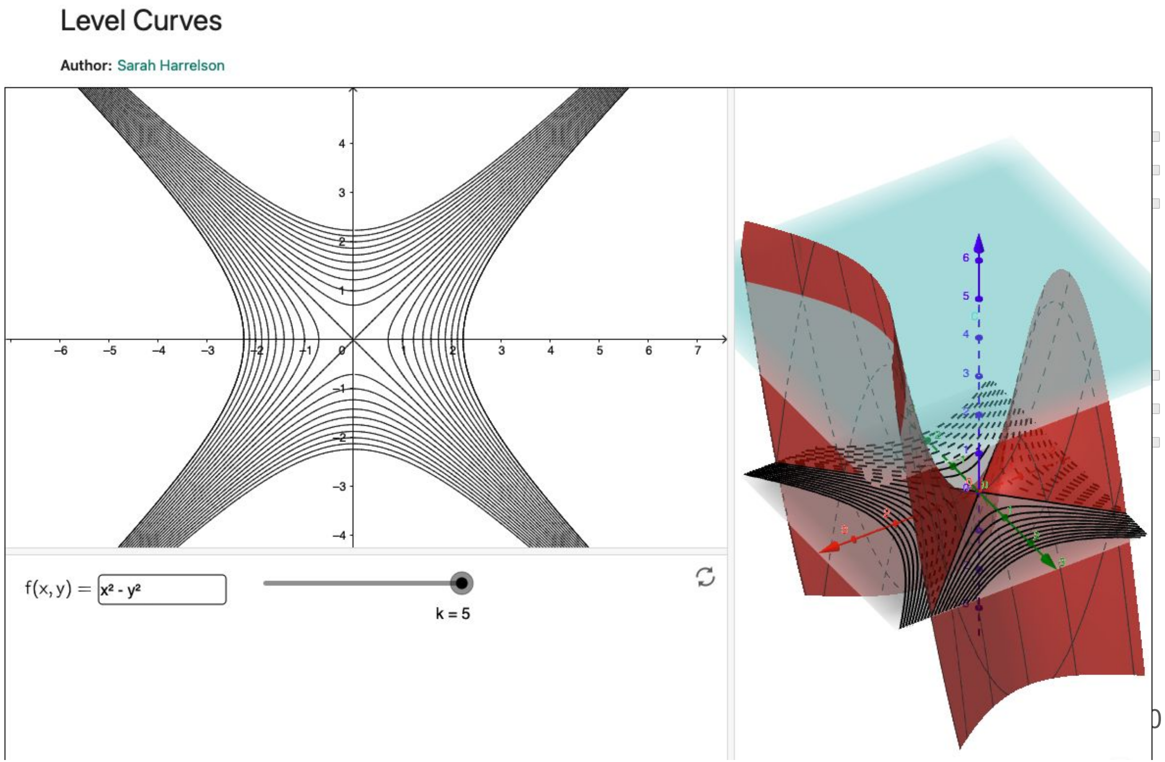


Level curves

Online examples : https://www.geogebra.org/m/M2P4KsRe, see also desmos

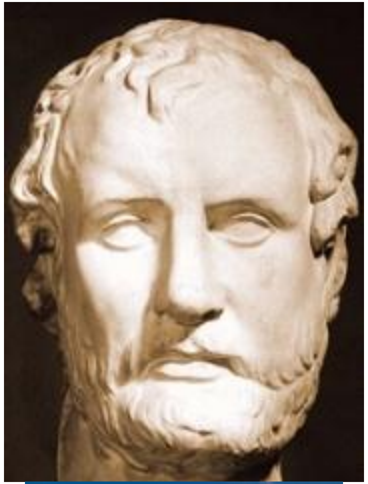
Hyperbolic paraboloid

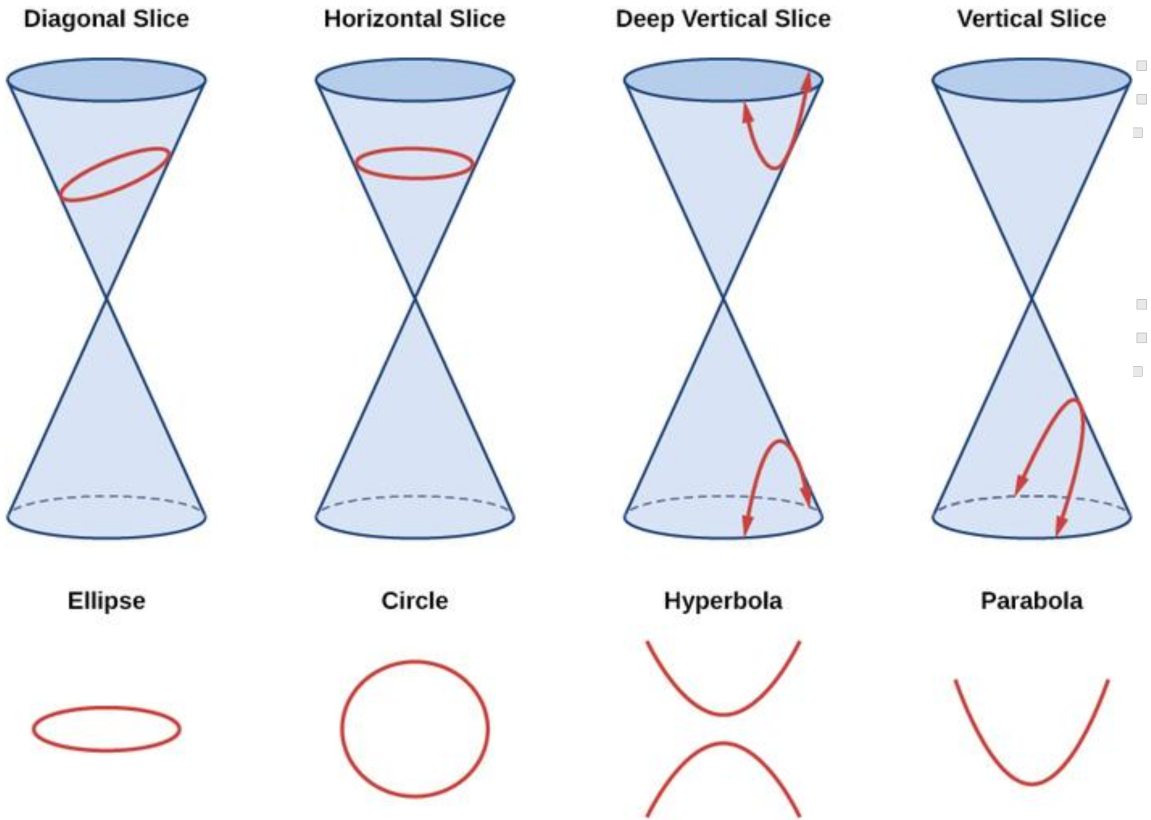
- Why is it called so?
- What would be an Ellipstic paraboloid?



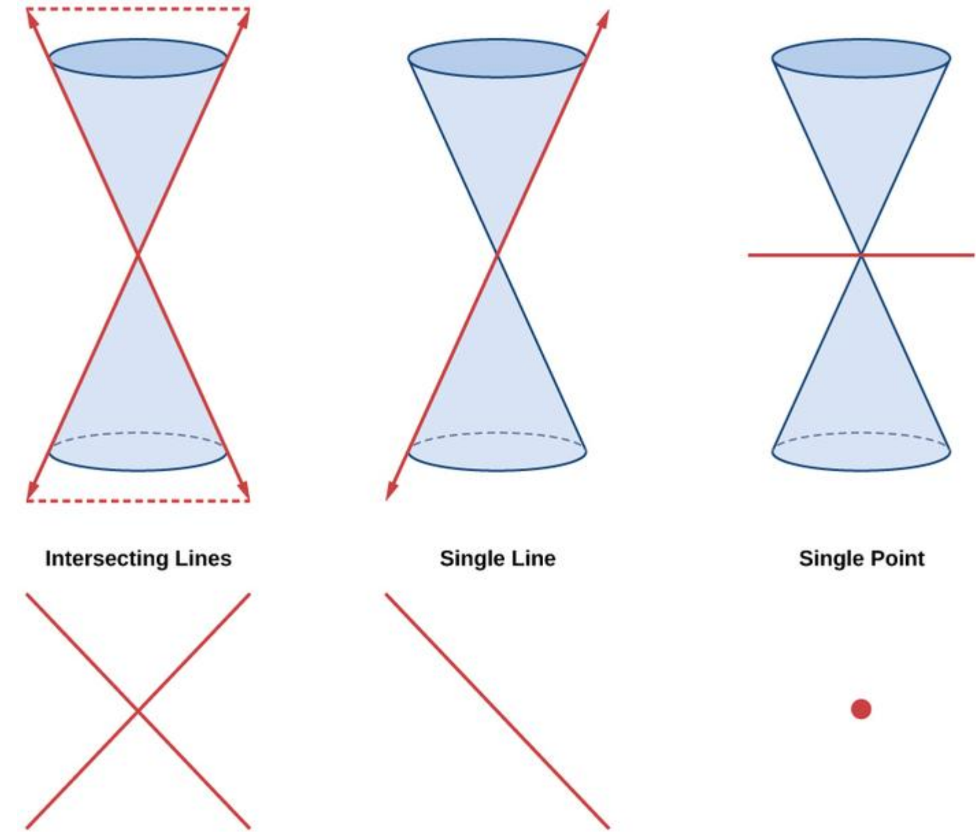
Conic sections

Menaechmus

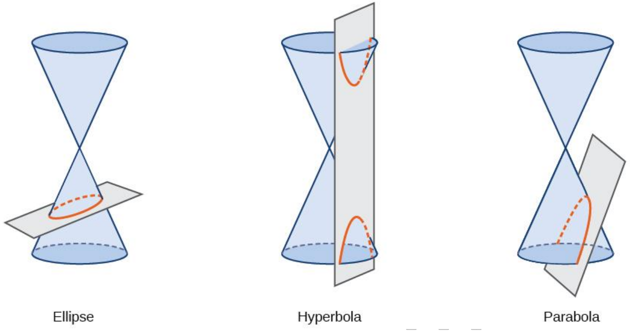




Conic sections



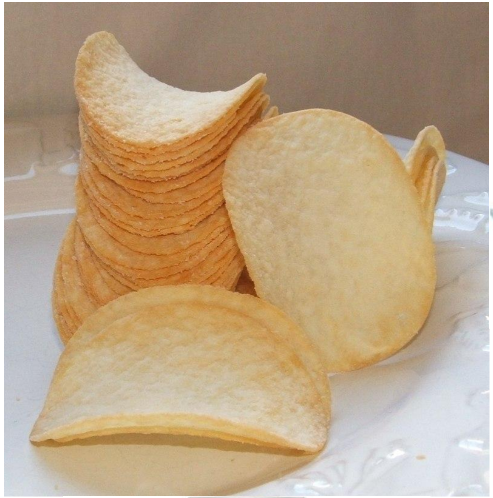
General form of conic sections

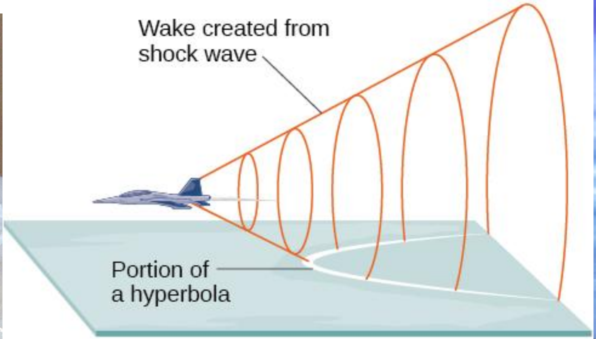


$$Ax^2+Bxy+Cy^2+Dx+Ey+F = 0$$

- Identify the values of A and C from the general form.
- If A and C are nonzero, have the same sign, and are not equal to each other, then the graph may be an ellipse.
- If A and C are equal and nonzero and have the same sign, then the graph may be a circle.
- If A and C are nonzero and have opposite signs, then the graph may be a hyperbola.
- If either A or C is zero, then the graph may be a parabola.

Conic sections are foundational across disciplines!



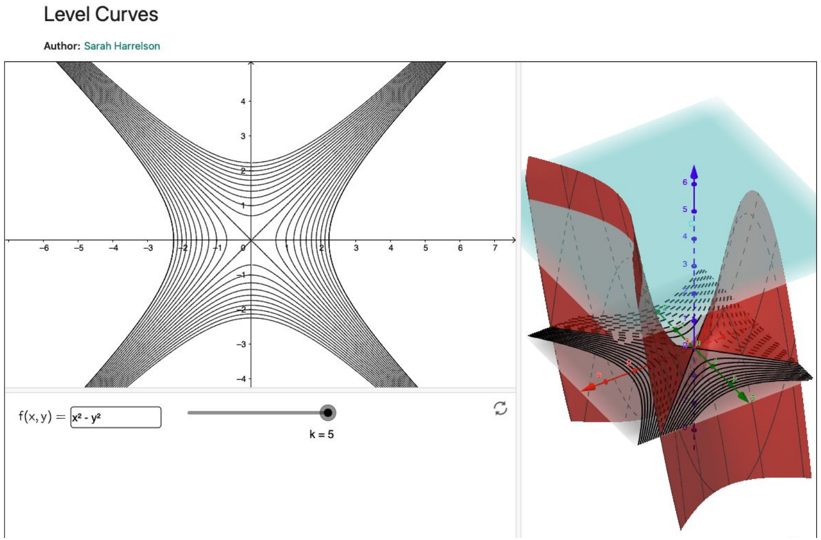


Examples

Conic Sections	Example
ellipse	$4x^2 + 9y^2 = 1$
circle	$4x^2 + 4y^2 = 1$
hyperbola	$4x^2 - 9y^2 = 1$
parabola	$4x^2 = 9y$ or $4y^2 = 9x$

Back to our hyperbolic paraboloid

Hyperbolic paraboloid



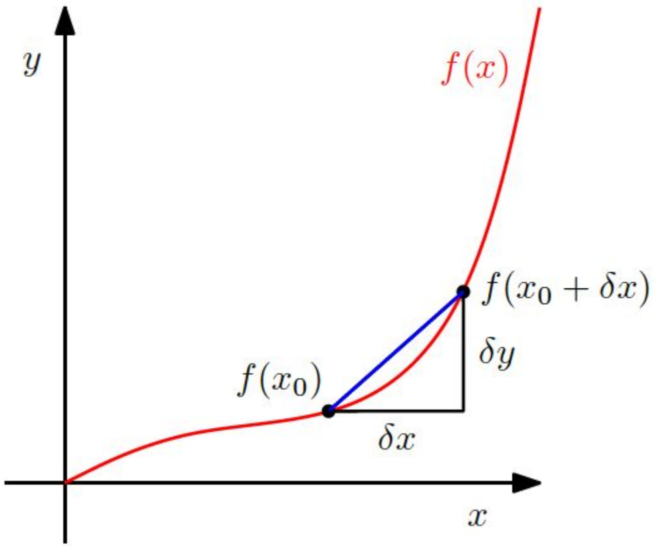
$$f(x,y) = x^2 - y^2 = 0 \Rightarrow (x-y)\cdot(x+y) = 0$$

$$f(x,y) = c \Rightarrow \frac{x^2}{c} - \frac{y^2}{c} = 1 \, (\text{Hyperbola!})$$

A refresher I: Single variable function

The difference quotient   computes the slope of the secant line through two points of y=f(x).

$$\frac{\delta y}{\delta x} = \frac{f(x+\delta x) - f(x)}{\delta x}$$



The idea of the derivative f′(x) is that it is the slope of the tangent line at x to the curve.

$$\frac{df}{dx} = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}$$

What is the derivative of d/dx(x^n)?

A refresher II: Single variable function

Product rule: $$(f(x)g(x))' = f'(x)g(x) + f(x)g'(x) \tag{5.29}$$

Quotient rule: $$\left(\frac{f(x)}{g(x)}\right)' = \frac{f'(x)g(x) - f(x)g'(x)}{(g(x))^2} \tag{5.30}$$

Sum rule: $$(f(x) + g(x))' = f'(x) + g'(x) \tag{5.31}$$

Chain rule: $$\big(g(f(x))\big)' = (g \circ f)'(x) = g'(f(x))f'(x) \tag{5.32}$$

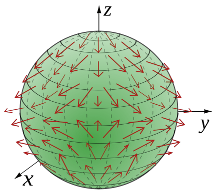
Source Chapter 5 https://mml-book.github.io/ (Mandatory reading)

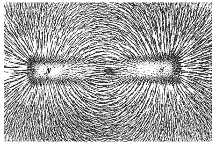
Matrix calculus

- Scalar field, a function f that maps vectors to reals $f : \mathbb{R}^n \rightarrow \mathbb{R}$

$$z = [4, 3]p = 4x + 3y$$
$$z = p^T p = x^2 + y^2$$

- Vector field, or vector valued functions $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$





- Functions of matrices f(A).

Gradient of a scalar field

- Partial derivative at $x=(x_1,..,x_n)$

$$\frac{\partial f}{\partial x_i} = \lim_{h\to 0} \frac{f(x_{1,} \ \ldots, x_i + h, \ldots, x_n) - f(x_{1,} \ \ldots, x_i, \ldots, x_n)}{h}, \ i = 1, \ldots, n$$

- We collect them at the row vector known as the gradient of the function ***f***

$$\nabla f(x) = \nabla_x f = \mathrm{grad} f = \begin{pmatrix} \frac{\partial f}{\partial x_1} & \frac{\partial f}{\partial x_2} & \cdots & \frac{\partial f}{\partial x_n} \end{pmatrix} \in \mathbb{R}^{1\times n}$$

Remark: the gradient collects the slopes in the positive x_i direction for all i=1..n.

Directional derivative

- Instead of computing the slopes in the positive x_i directions for all i=1..n, we can compute the derivative along any direction.
  - Directional derivative

$$\nabla_v f(x) = D_v f(x) = \lim_{h \to 0} \frac{f(x + hv) - f(x)}{h} = \nabla f(x) \cdot v$$

- **Exercise**
  Let $f(x,y)=x^2y$. Find the following:
  - The gradient of f
  - The gradient of f at (3,2)
  - The derivative of f in the direction of (1,2) at the point (3,2).

Demo

Hessian of a scalar field

If all second partial derivatives of f exist and are continuous over the domain of the function, then the Hessian matrix is a square matrix, usually defined and arranged as follows:

$$\mathbf{H}_f = \begin{bmatrix} \dfrac{\partial^2 f}{\partial x_1^2} & \dfrac{\partial^2 f}{\partial x_1\,\partial x_2} & \cdots & \dfrac{\partial^2 f}{\partial x_1\,\partial x_n} \\[2.2ex] \dfrac{\partial^2 f}{\partial x_2\,\partial x_1} & \dfrac{\partial^2 f}{\partial x_2^2} & \cdots & \dfrac{\partial^2 f}{\partial x_2\,\partial x_n} \\[2.2ex] \vdots & \vdots & \ddots & \vdots \\[2.2ex] \dfrac{\partial^2 f}{\partial x_n\,\partial x_1} & \dfrac{\partial^2 f}{\partial x_n\,\partial x_2} & \cdots & \dfrac{\partial^2 f}{\partial x_n^2} \end{bmatrix}$$

Example

- Compute the Hessian of f(x,y)=xy(x+y) at (1,1).

$$H_f(x,y) = \begin{pmatrix} 2y & 2(x+y) \\ 2(x+y) & 2x \end{pmatrix}, \; H_f(1,1) = \begin{pmatrix} 2 & 4 \\ 4 & 2 \end{pmatrix}$$

- The symmetry of H is not a coincidence; of f(x,y) is a twice continuously differentiable function, then

$$\frac{\partial^2 f}{\partial x \partial y} = \frac{\partial^2 f}{\partial y \partial x}$$

Taylor Series

Taylor polynomial f:R→R

The Taylor polynomial of degree n of f:R→R at x_0 is defined as

where $f^{(k)}(x_0)$ is the k-th derivative of f at x_0.

$$T_n(x) = \sum_{k=0}^{n} \frac{f^{(k)}(x_0)}{k!}(x - x_0)^k$$

Taylor series f:R⟶R

The Taylor series of of a smooth function f:R⟶R at x_0 is defined as

$$T_\infty(x) = \sum_{k=0}^{+\infty} \frac{f^{(k)}(x_0)}{k!}(x - x_0)^k$$

For x_0=0, we obtain Maclaurin series as a special instance of The Taylor series.

If $f(x) = T_\infty(x)$, then f is called analytic.

Examples

- Taylor polynomial T_6 for $f(x)=x^4$ evaluated at $x_0=1$

$$T_6(x) = 1 + 4(x-1) + 6(x-1)^2 + 4(x-1)^3 + (x-1)^4 + 0 = \ldots = x^4$$

- Taylor series for trigonometric functions

$$\cos(x) = \sum_{k=0}^{\infty} (-1)^k \frac{1}{(2k)!} x^{2k}$$

$$\sin(x) = \sum_{k=0}^{\infty} (-1)^k \frac{1}{(2k+1)!} x^{2k+1}$$

- https://en.wikipedia.org/wiki/Taylor_series

Taylor series $f: R^n \to R$

Example (whiteboard)

$$f(x) \approx f(x_0) + \nabla f(x_0)(x - x_0) + \frac{1}{2}(x - x_0)^T H_f(x_0)(x - x_0)$$

Chain rule

$$\frac{\partial}{\partial x}(g \circ f)(x) = \frac{\partial}{\partial x} g\left(f(x)\right) = \frac{\partial g}{\partial f}\frac{\partial f}{\partial x}$$

- Examples:
  Consider a function $f: R^2 \rightarrow R$ of two variables x_1, x_2. Furthermore, suppose that x_1, x_2 are functions of a variable t.

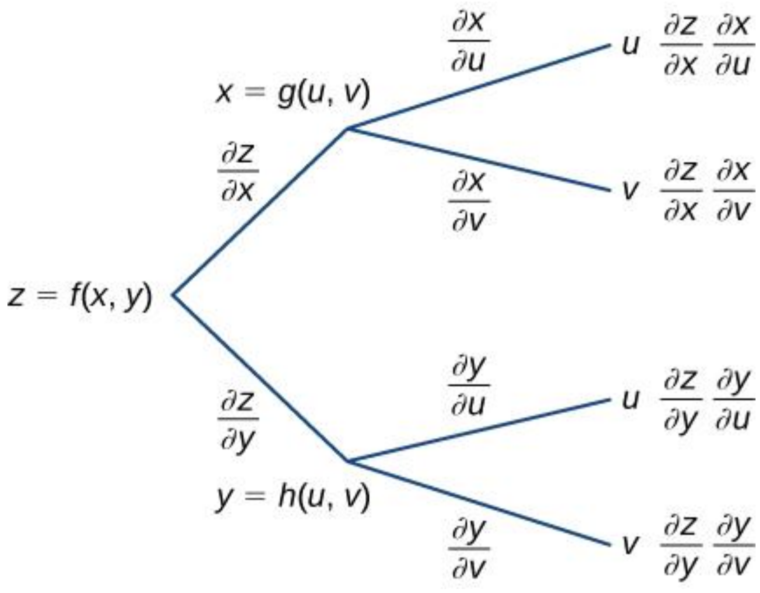
$$\frac{df}{dt} = \begin{bmatrix} \frac{\partial f}{\partial x_1} & \frac{\partial f}{\partial x_2} \end{bmatrix} \begin{bmatrix} \frac{\partial x_1(t)}{\partial t} \\ \frac{\partial x_2(t)}{\partial t} \end{bmatrix}$$

  Consider a function $f: R^2 \rightarrow R$ of two variables x_1, x_2. Furthermore, suppose that x_1, x_2 are functions of two variables s, t.

$$Let\, q = [s, t].\ \frac{df}{dq} = \begin{bmatrix} \frac{\partial f}{\partial x_1} & \frac{\partial f}{\partial x_2} \end{bmatrix} \begin{bmatrix} \frac{\partial x_1(s,t)}{\partial s} & \frac{\partial x_1(s,t)}{\partial t} \\ \frac{\partial x_2(s,t)}{\partial s} & \frac{\partial x_2(s,t)}{\partial t} \end{bmatrix}$$

Chain rule examples

$$Let\ z = f(x,y).\ \frac{dz}{d(u,t)} = \begin{bmatrix} \frac{\partial z}{\partial u} & \frac{\partial z}{\partial v} \end{bmatrix} = \begin{bmatrix} \frac{\partial z}{\partial x} & \frac{\partial z}{\partial y} \end{bmatrix} \begin{bmatrix} \frac{\partial x(u,v)}{\partial u} & \frac{\partial x(u,v)}{\partial v} \\ \frac{\partial y(u,v)}{\partial u} & \frac{\partial y(u,v)}{\partial v} \end{bmatrix}$$



Generalized chain rule

Let $z=f(x_1,\ldots,x_m)$ be a scalar field of m variables, each of which is a differential function of n independent variables $x_i=xi(t_1,..,t_n)$. Then,

$$\frac{\partial z}{\partial t_i} = \sum_{j=1}^{m} \frac{\partial z}{\partial x_j}\frac{\partial x_j}{\partial t_i} = \frac{\partial z}{\partial x_1}\frac{\partial x_1}{\partial t_i} + \ldots + \frac{\partial z}{\partial x_m}\frac{\partial x_m}{\partial t_i},\ i = 1,\ldots,n$$

Examples

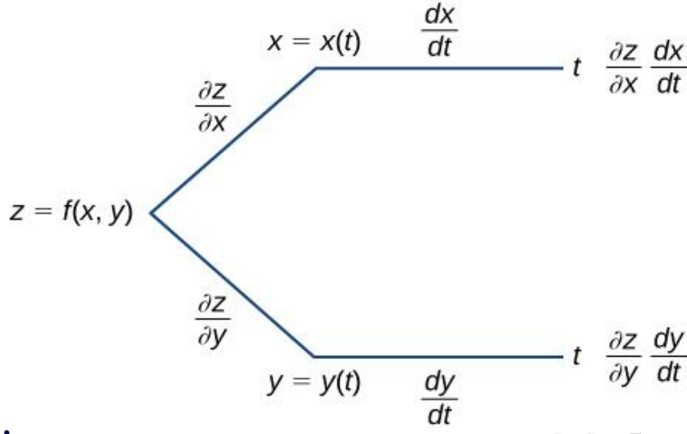
Calculate the derivative of z with respect to t, where

$$z = f(x, y) = x^2 - 3xy + 2y^2$$
$$x = x(t) = 3\sin(2t)$$
$$y = y(t) = 4\cos(2t)$$

Solution:

$$\frac{dz}{dt} = \frac{\partial z}{\partial x}\frac{dx}{dt} + \frac{\partial z}{\partial y}\frac{dy}{dt} = (2x - 3y)6\cos(2t) + (-3x + 4y)(-8\sin(2t)) =$$
$$= 6\cos(2t)(6\ \sin(2t) - 12\cos(2t)) - 8\sin(2t)(-9\sin(2t) + 16\ \cos(2t)) =$$
$$= \ldots = -46\sin(4t) - 72\ \cos(4t)$$

Examples



$f(x,y)=4x^2+3y^2$, $x(t)=\sin(t)$, $y(t)=\cos(t)$

We compute $\frac{\partial z}{\partial x} = 8x, \frac{\partial z}{\partial y} = 6y, \frac{dx}{dt} = \cos t, \frac{dy}{dt} = -\sin t.$

Now we apply the chain rule

$$\frac{dz}{dt} = \frac{\partial z}{\partial x}\frac{dx}{dt} + \frac{\partial z}{\partial y}\frac{dy}{dt} = 8x\cos t - 6y\sin t = 8\sin t\cos t - 6\,\cos t\sin t = 2\cos t\sin t$$

1st order derivatives of a vector field: Jacobian

$$f(x_1, \ldots, x_n) = \begin{bmatrix} f_1(x_1, \ldots, x_n) \\ \ldots \\ f_m(x_1, \ldots, x_n) \end{bmatrix}$$

The collection of all first-order derivatives of a vector field/vector-valued function $f: R^n \rightarrow R^m$ is called the Jacobian.

$$\boldsymbol{J} = \nabla_{\boldsymbol{x}} \boldsymbol{f} = \frac{\mathrm{d}\boldsymbol{f}(\boldsymbol{x})}{\mathrm{d}\boldsymbol{x}} = \begin{bmatrix} \frac{\partial \boldsymbol{f}(\boldsymbol{x})}{\partial x_1} & \cdots & \frac{\partial \boldsymbol{f}(\boldsymbol{x})}{\partial x_n} \end{bmatrix}$$

$$= \begin{bmatrix} \frac{\partial f_1(\boldsymbol{x})}{\partial x_1} & \cdots & \frac{\partial f_1(\boldsymbol{x})}{\partial x_n} \\ \vdots & & \vdots \\ \frac{\partial f_m(\boldsymbol{x})}{\partial x_1} & \cdots & \frac{\partial f_m(\boldsymbol{x})}{\partial x_n} \end{bmatrix},$$

Jacobian

Let $\begin{aligned} y_1 &= -2x_1 + x_2 \\ y_2 &= x_1 + x_2 \end{aligned}$. The Jacobian is simply $J = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} \\ \frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} \end{bmatrix} = \begin{bmatrix} -2 & 1 \\ 1 & 1 \end{bmatrix}$

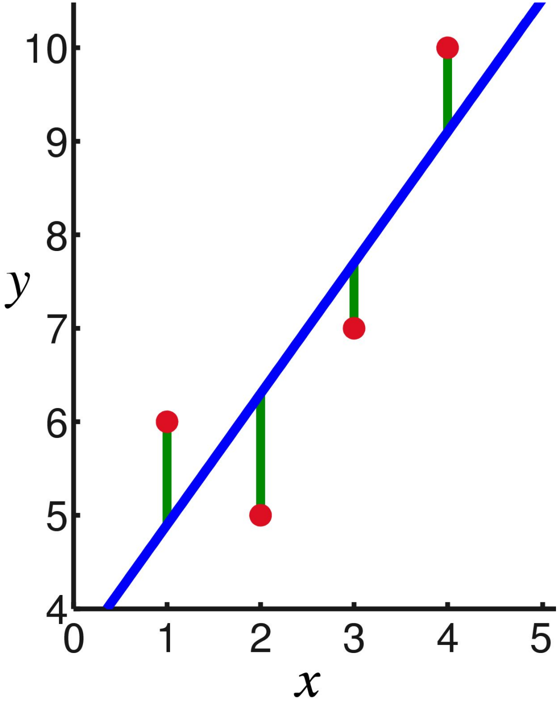
This example generalizes to the following. Let f(x)=Ax, where A is a mxn matrix, and x is an mx1 vector. Then,

$$\frac{df}{dx} = A$$

Gradient of a Least-Squares Loss in a Linear Model

Consider the linear model

$$y^{n\times 1} = \Phi^{n\times d}\theta^{d\times 1}$$
$$L(e) = \|e\|^2$$
$$e(\theta) = y - \Phi\theta$$



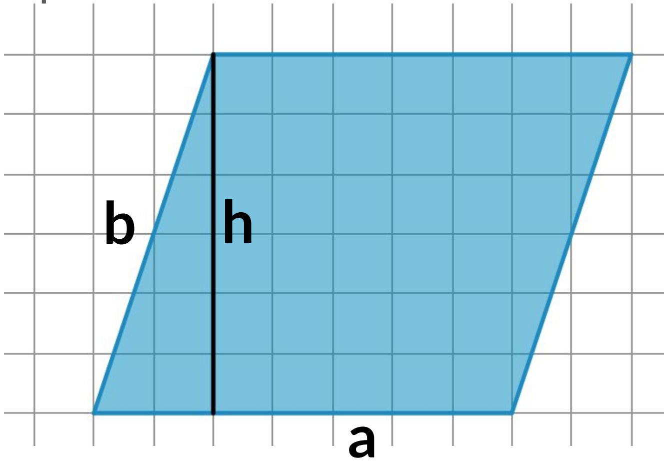
Let's prove that $\dfrac{\partial L}{\partial \theta} = -2\left(y^T - \theta^T\Phi^T\right)\Phi$

(whiteboard, see also example 5.11 here)

Parallelogram of maximum area

Find paralellogram of maximum area with a given perimeter.

$$\begin{aligned} \max_{a,b,h} \ & ah \\ 2a + 2b &= \ell \\ h &\le b \\ a, b, h &\ge 0 \end{aligned}$$



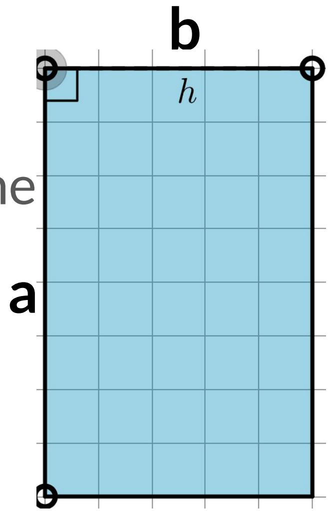
Clearly given a,b, h=b is an obvious solution.

Thus we get the following equivalent problem:

Parallelogram of maximum area

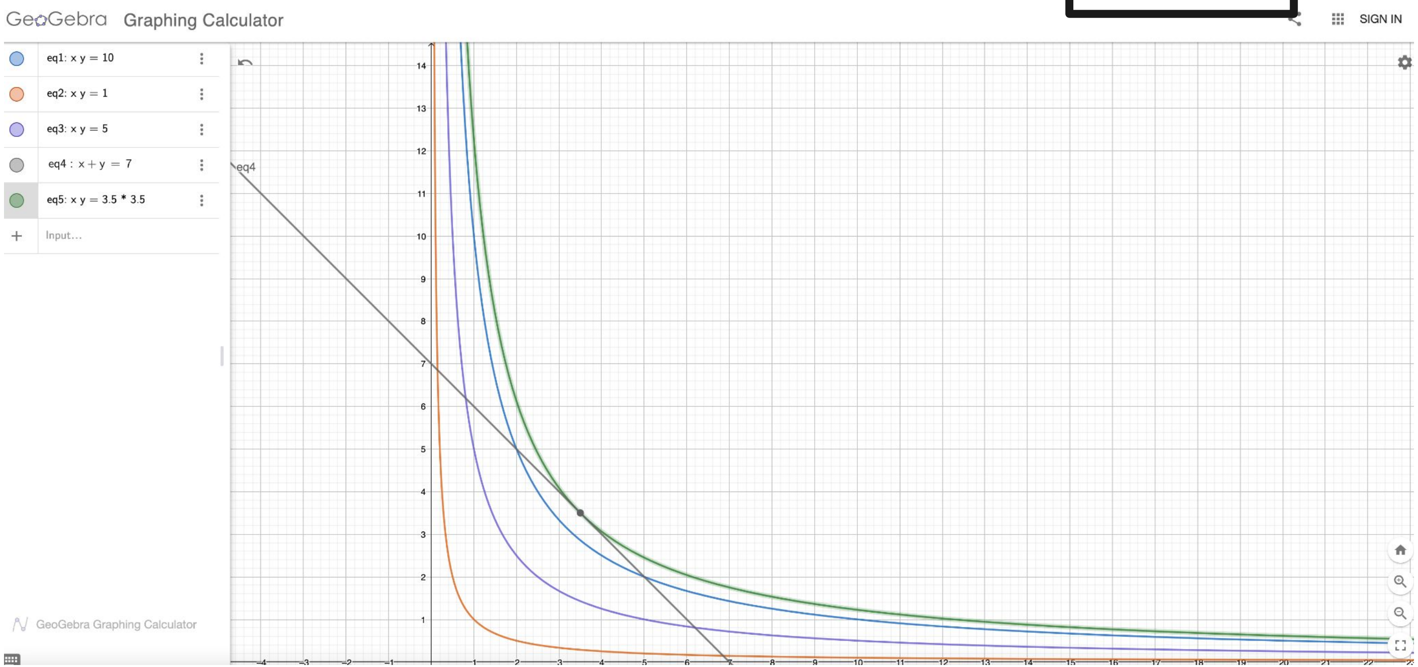
Find paralellogram of maximum area with a given perime

$$\max_{a,b} \; ab$$
$$2a + 2b = \ell$$
$$a, b \geq 0$$

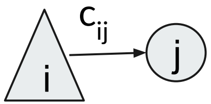


Optimal solution a=b=l/4 (h=b)

Optimal solution is square



Transportation problem



Minimize the cost of goods transported from

- a set of m sources to ..
- ... a set of n destinations
    - subject to the supply and demand of the sources and destination respectively

Given:

- $a_1,\dots,a_m$: units to transfer from sources
- $b_1,\dots,b_n$: units to receive by destinations
- c_{ij}: cost of transferring a unit from source **i** to destination **j**

Transportation problem

- Find the quantities xij to be transferred from *source* i to *destination* j for i=1,..,m, j=1,..,n.

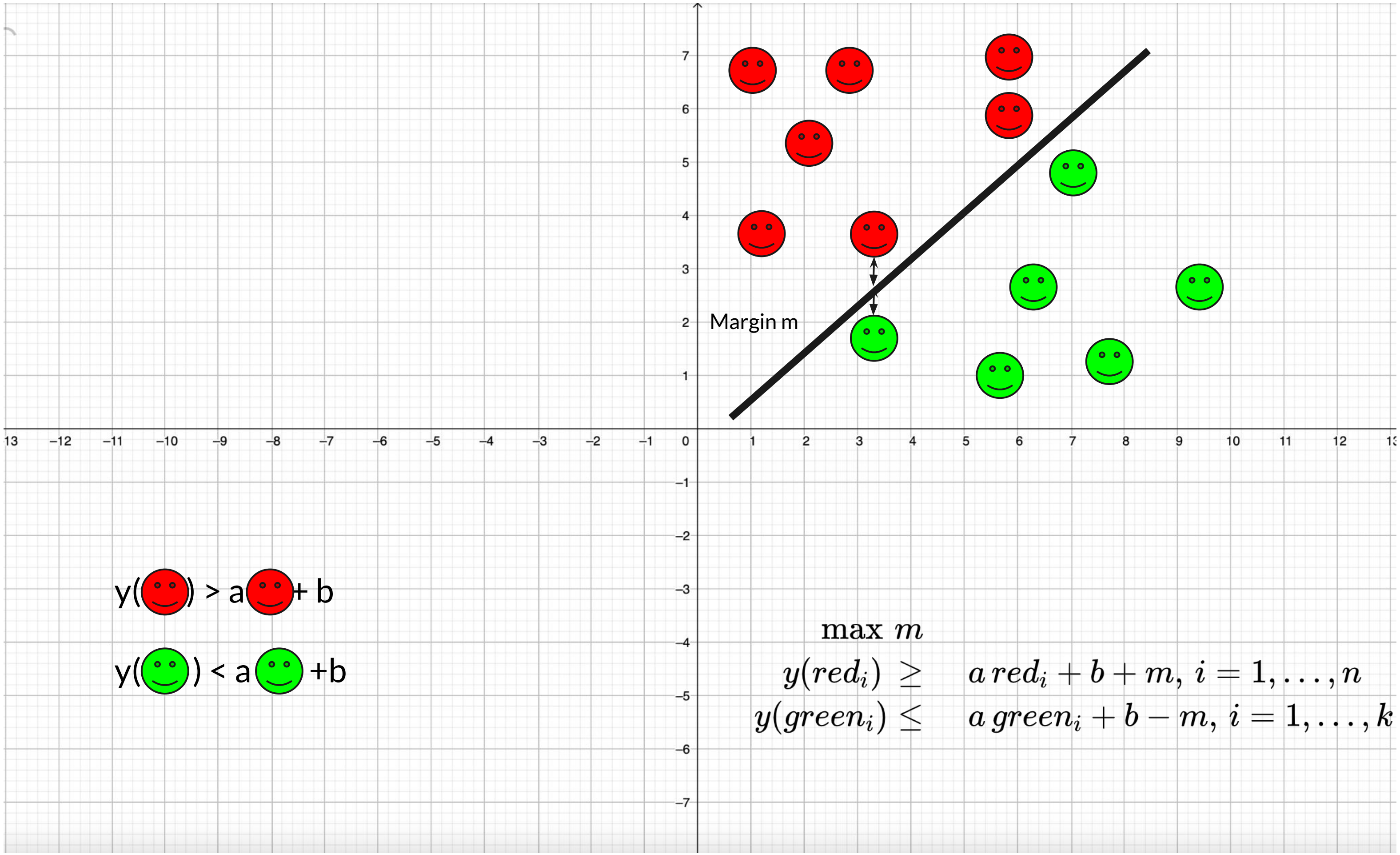
$$\min \sum_{i=1}^{m} \sum_{j=1}^{n} c_{ij} x_{ij}$$

$$\sum_{j=1}^{n} x_{ij} = a_i \,, \qquad i = 1, \ldots, m$$

$$\sum_{i=1}^{m} x_{ij} = b_j, \qquad j = 1, \ldots, n$$

$$x_{ij} \geq 0$$

A (not so) Toy ML problem



$$\max\ m$$
$$y(red_i) \geq a\,red_i + b + m,\ i = 1, \ldots, n$$
$$y(green_i) \leq a\,green_i + b - m,\ i = 1, \ldots, k$$

Minimization

$$\min_{x \in F} f(x)$$

Let $f: R^n \rightarrow R$.

- When $F = R^n$, the optimization is *unconstrained.*
- When $F = \{x \in \mathbb{R}^n : h(x) = 0, g(x) \leq 0\}$
  where $h: R^n \rightarrow R^m$, $g: R^n \rightarrow R^k$ are real functions
  the problem is called *constrained.*

But what does it mean to be a minimum? And why don't we talk about maximization?

Minimization

- Minimize f is equivalent to maximize -f.

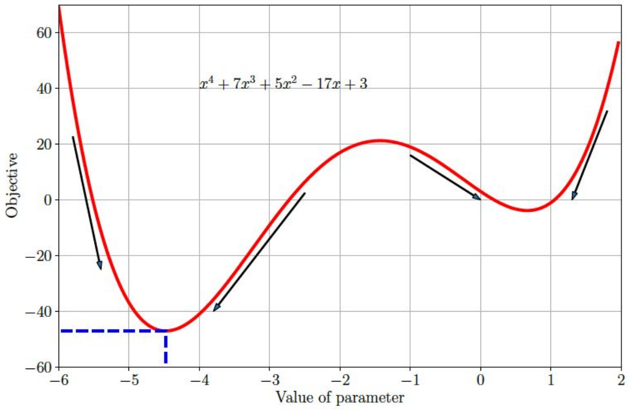
- **Definition:** A point x* is called a ***local minimum*** of f in F if there exist ε>0 such that $f(x) \geq f(x^\star)$ for all x in F such that $\|x - x^\star\| \leq \epsilon$.

  If for all $x \neq x^\star, \|x - x^\star\| \leq \epsilon,$ $f(x) > f(x^\star)$ then x* is called ***strict local minimum.***

Minimization

- **Definition:** A point x* is called a ***global minimum*** of f in F if $f(x) \geq f(x^\star)$

If $f(x) > f(x^\star)$, for all $x \neq x^\star$ then x* is called ***strict global minimum.***



Does the minimum always exist?

What is the minimum of $f(x) = -0.5x + 4 \text{ where } 0 \leq x < 2$

- The minimum does not exist.
- Set $x=2-\varepsilon$, $\varepsilon>0$. What is f(x)?
- Now set $x=2-\varepsilon/2$, $\varepsilon>0$. What is f(x) now?

Sufficient conditions

Weierstrass theorem states that if $f:R^n \rightarrow R$ is continuous, and F is compact then f has a global minimum in F.

Theorem (1st order necessary conditions)

If $f:\mathbb{R}^n \to \mathbb{R}$ is continuous, differentiable function and x^* is a local minimum of f, then

$$\boxed{\nabla f(x^\star) = 0}$$

Remark: Necessary, but not sufficient.

Example: least squares

- A^{mxn} matrix (assume columns are independent)
- b^{mx1} vector

Least squares problem: Solve $\min_x ||Ax-b||^2$

Least squares

Question: Why can we invert $(A^T A)$?

$$f(x) = \|Ax - b\|^2 = (Ax-b)^T(Ax-b)$$
$$= x^T A^T A x - 2x^T A^T b + b^T b$$

$$\nabla f(x) = 2x^T A^T A - 2b^T A = 0 \Rightarrow$$
$$A^T A x = A^T b \Rightarrow x = (A^T A)^{-1} A^T b$$

$$A^T A x = 0 \Rightarrow x^T A^T A x = 0 \Rightarrow$$
$$\|Ax\|^2 = 0 \Rightarrow Ax = 0 \Rightarrow$$
$$x = 0\,(why?)$$

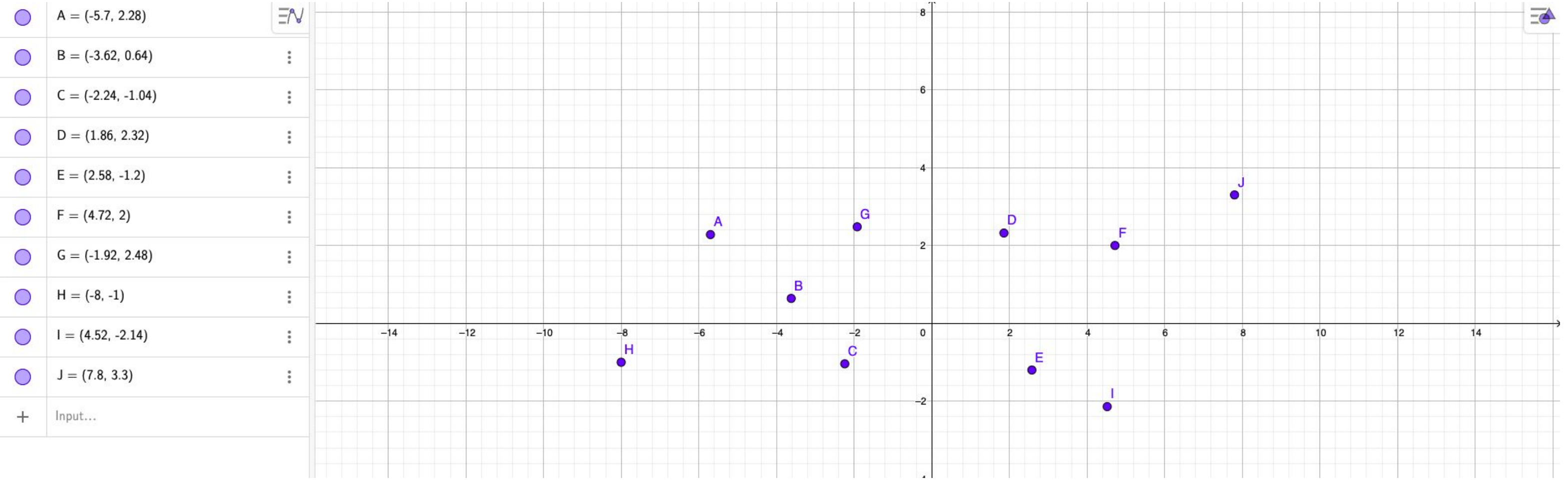
Normal equations

Turns out that this is the strict global minimum since f(x) is convex (to be discussed later)

Practice problem

What is the best-fit function of the following form that passes through the given points?

$$y = A\cos(x) + B\sin(x) + C\cos(2x) + D$$



Stationary points

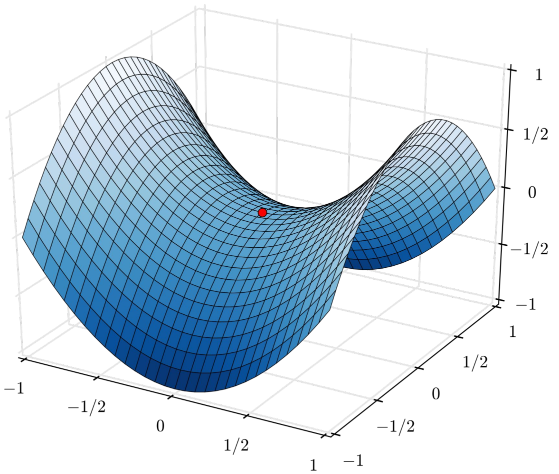
Consider the set of stationary points of f
These include:

$$D = \{x^\star \in \mathbb{R}^n : \nabla f(x^\star) = 0\}$$

- Local minima
- Local maxima
- Saddle points

How do we recognize the type of a stationary point? (More on next lecture, but for now...)

Saddle point



Let $f : \mathbb{R}^{n+m} \to \mathbb{R}$

The point (x*,y*) in R^{n+m} is a saddle point:

$$f(x^\star, y) \leq f(x^\star, y^\star) \leq f(x, y^\star)\, \forall x : \|x - x^\star\| \leq \epsilon, \forall y : \|y - y^\star\| \leq \epsilon$$

- For fixed y=y*, f has a local min at x*
- For fixed x=x*, f has a local max at y*

Theorem (2nd order necessary conditions)

If $f:R^n \rightarrow R$ is continuous, twice differentiable function and x^* is a local minimum of f, then

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

$$x^T \frac{\partial^2 f(x^\star)}{\partial x^2} x \geq 0 \quad \text{for all } x \in \mathbb{R}^n$$

Necessary but not sufficient

- The previous theorem provides necessary but not sufficient conditions.
- Let's see an example. Consider the following unconstrained minimization problem ($F=R^2$)

$$\min_{x_1, x_2} (x_1 - x_2)^2 + (x_1 + x_3)^3$$

Necessary but not sufficient

From the 1st order necessary condition we obtain

$$\nabla f(x^\star) = 0 \Rightarrow \left[2(x_1 - x_2) + 3(x_1 + x_2)^2,\ -2(x_1 - x_2) + 3(x_1 + x_2)^2\right] = [0, 0] \Rightarrow$$

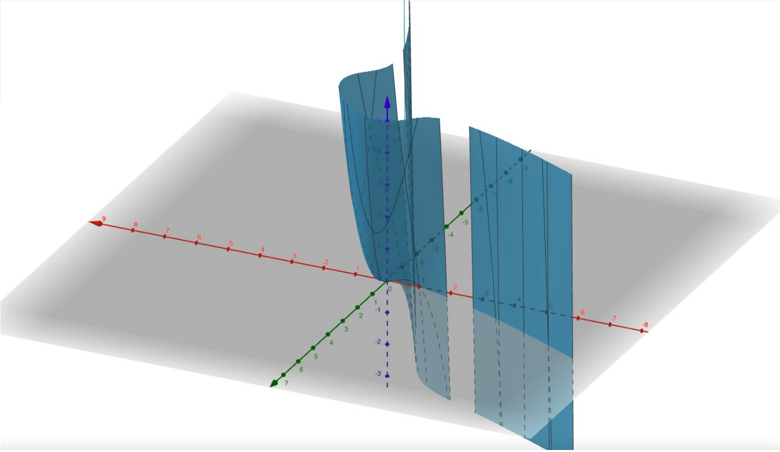
$$x_1 = 0,\ x_2 = 0 \Rightarrow x^\star = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Necessary but not sufficient

The Hessian of f is $H_f = \frac{\partial^2 f}{\partial x^2} = \begin{bmatrix} 2+6(x_1+x_2) & -2+6(x_1+x_2) \\ -2+6(x_1+x_2) & 2+6(x_1+x_2) \end{bmatrix}$

Thus, $H_f(x^\star) = \begin{bmatrix} 2 & -2 \\ -2 & 2 \end{bmatrix}$. The eigenvalues are 4,0, so the matrix is positive semidefinite. Another way to see this is as follows:

$$z^T H_f(x^\star) z = 2z_1^2 - 4z_1 z_2 + 2z_2^2 = 2(z_1 - z_2)^2, \ \forall z = \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} \in \mathbb{R}^2$$



Necessary but not sufficient

However, x^* is not a local minimum. Let's see why. Consider the all-ones eigenvector corresponding to the 0 eigenvalue, and consider moving from x^* in this direction, i.e., consider

$x = x^* + a[1,1]^T$ where $a<0$. Then the objective becomes $8a^3 < 0 = f(x^*)$

Theorem (2nd order sufficient conditions)

If $f: R^n \to R$ is continuous, twice differentiable function and x^* is a strict local minimum of f, then

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

$$x^T \frac{\partial^2 f(x^\star)}{\partial x^2} x > 0 \ \text{ for all } x \in \mathbb{R}^n$$

Gradient descent

Let's consider the linearization of f:R⟶R

$$f(x+\epsilon) = f(x) + \epsilon f'(x) + O(\epsilon^2)$$

Question: Assuming second-order terms are negligible, how would you choose ε to decrease the value of the function, i.e., f(x+ε)<=f(x)

$$f\left(x - \eta f'(x)\right) = f(x) - \eta\left(f'(x)\right)^2 + O\left(\eta^2\left(f'(x)\right)^2\right),\ \eta > 0$$

$$\boxed{x \leftarrow x - \eta f'(x),\ \eta > 0}$$

Example $f(x)=x^2$.

Gradient descent

When $f: R^n \to R$, we use the gradient of f

$$x \leftarrow x - \eta(\nabla f(x))^T, \eta > 0$$

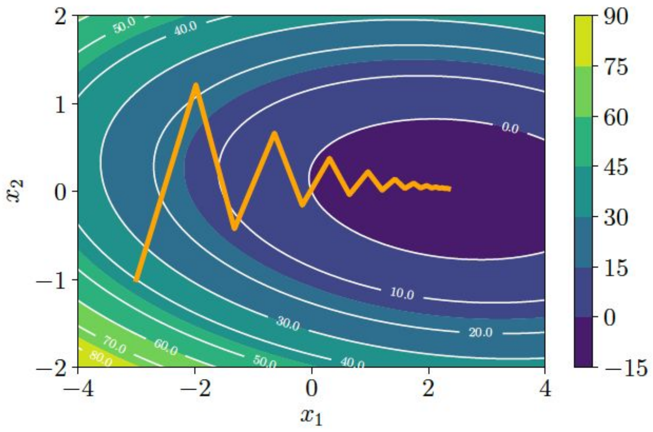
Example

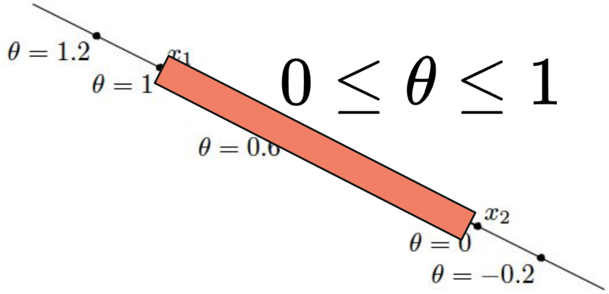
Consider a quadratic function in two dimensions

$$f\left(\begin{bmatrix}x_1\\x_2\end{bmatrix}\right)=\frac{1}{2}\begin{bmatrix}x_1\\x_2\end{bmatrix}^\top\begin{bmatrix}2&1\\1&20\end{bmatrix}\begin{bmatrix}x_1\\x_2\end{bmatrix}-\begin{bmatrix}5\\3\end{bmatrix}^\top\begin{bmatrix}x_1\\x_2\end{bmatrix}$$

with gradient

$$\nabla f\left(\begin{bmatrix}x_1\\x_2\end{bmatrix}\right)=\begin{bmatrix}x_1\\x_2\end{bmatrix}^\top\begin{bmatrix}2&1\\1&20\end{bmatrix}-\begin{bmatrix}5\\3\end{bmatrix}^\top.$$





Line

Suppose x_1,x_2 are two points in R^n. Points of the form

$$y = \theta x_1 + (1-\theta)x_2,\ \theta \in \mathbb{R}$$

form the line passing through x_1,x_2

Affine set

Definition: A set C is affine if the line through any two distinct points lines in C.

- The idea generalizes to more than two points. An affine combination of k points $x_1,...,x_k$ in C is $\theta_1 x_1 + \ldots + \theta_k x_k$ where $\theta_1 + \ldots + \theta_k = 1$

Claim: An affine set contains every affine combination of its points.

(induction on the number of points)

Affine sets - Prove the following:

1. The solution set $\{x | A^{m \times n} x^{n \times 1} = b^{m \times 1}\}$ is an affine set.

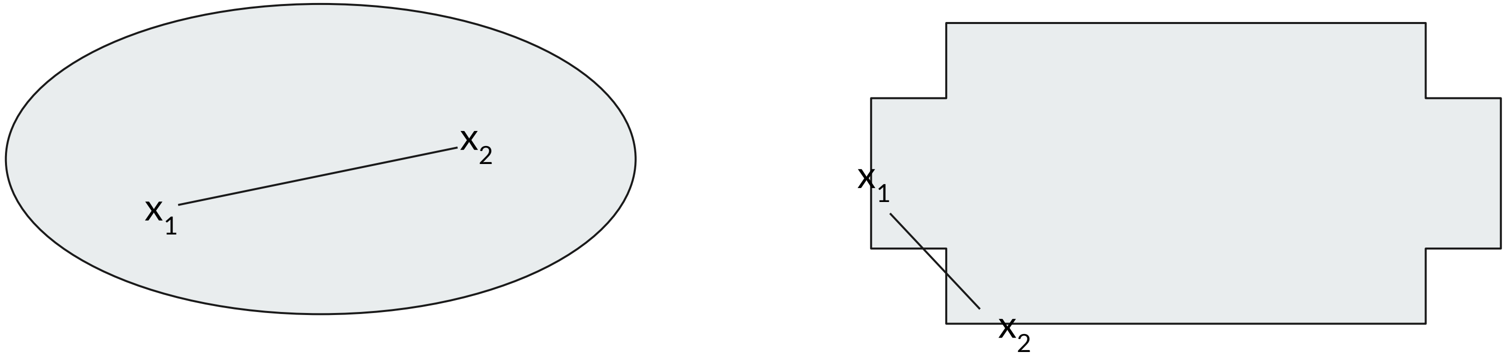
2. If C is an affine set, and x_0 is in C, then the set

$$V = C - x_0 = \{x - x_0 \mid x \in C\}$$

is a subspace.

(Proofs on whiteboard)

Convex vs non-convex set



A set C is convex if the line segment between any two points in C lies in C, i.e., for any x_1,x_2 in C and for any , $0 \leq \theta \leq 1$

$$\theta x_1 + (1 - \theta) x_2 \quad \in C$$

Hyperplanes

$$a^T x = b,$$
$$where \quad a \in \mathbb{R}^n, \; a \neq 0, b \in \mathbb{R}$$

- b offset of the hyperplane from 0

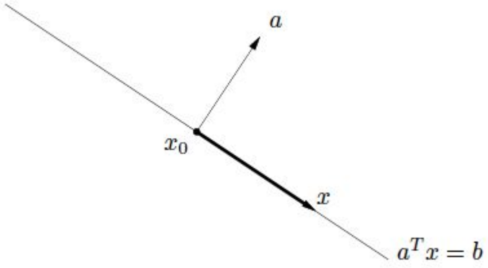


Figure 2.6 Hyperplane in $\mathbf{R}^2$, with normal vector a and a point x_0 in the hyperplane. For any point x in the hyperplane, $x - x_0$ (shown as the darker arrow) is orthogonal to a.

Halfspaces

- A hyperplane divides R^n into two halfspaces.
- Halfspaces are convex but not affine

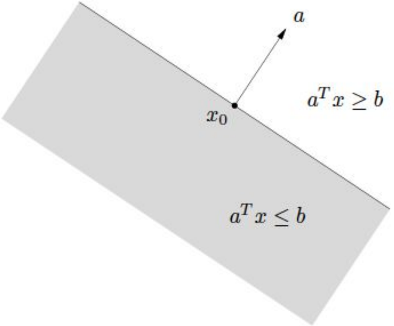


Figure 2.7 A hyperplane defined by $a^T x = b$ in $\mathbf{R}^2$ determines two halfspaces. The halfspace determined by $a^T x \geq b$ (not shaded) is the halfspace extending in the direction a. The halfspace determined by $a^T x \leq b$ (which is shown shaded) extends in the direction $-a$. The vector a is the outward normal of this halfspace.

Convex function

A function f:Rn→R is convex if its domain dom(f) is convex and if for all x,y in dom(f), and θ in [0,1] $f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta)f(y)$

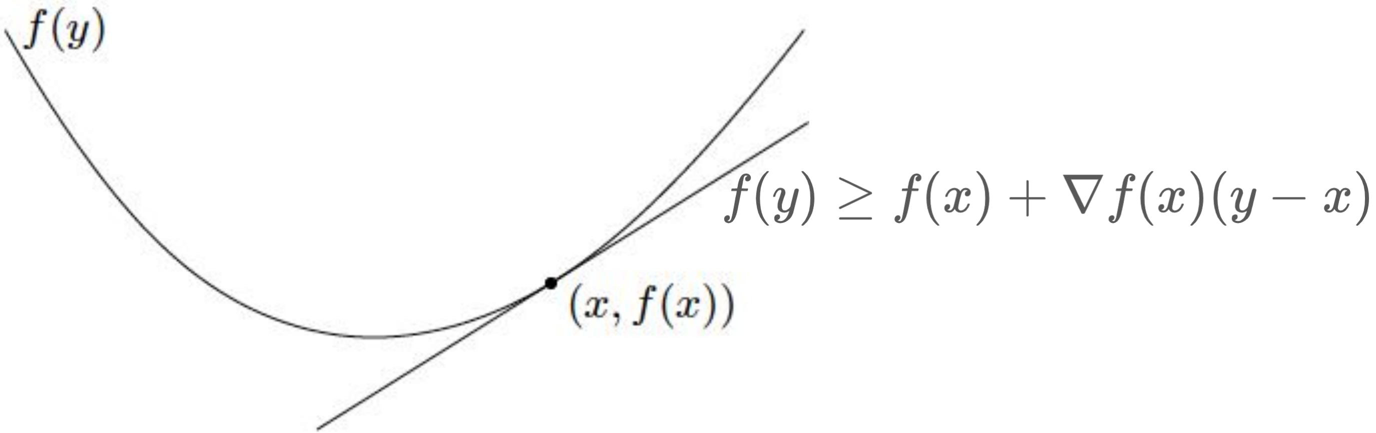
- It is strictly convex if the inequality is strict for all θ in (0,1).

- f is concave if -f is convex.

Figure 3.1 Graph of a convex function. The chord (*i.e.*, line segment) between any two points on the graph lies above the graph.

Convex function, 1st order condition

Suppose f is differentiable. Then f is convex if its domain is a convex set and $f(y) \geq f(x) + \nabla f(x)(y - x)$



Convex function, 2nd order condition

Assuming f is twice differentiable. f is convex iff f's domain is convex and the Hessian is positive semidefinite

$$x^T \frac{\partial^2 f(x^\star)}{\partial x^2} x \geq 0, \qquad \text{for all } x \in \mathbb{R}^n$$

Exercise: Prove that f(x,y)=x^2/y where x in R, and y>0 is convex.

Convex optimization

A constrained optimization problem is called a convex optimization problem if

$$\begin{aligned} &\min \; f(x) \\ &\text{subject to} \quad g_i(x) \leq 0\,,\; i = 1, \ldots m \\ &\qquad\qquad\quad a_i^T x - b_i = 0,\; j = 1, \ldots, p \end{aligned}$$

where f,gi's are convex functions.

Remark: the feasible set of a convex optimization problem is convex (why?)

Readings and Refs

Mandatory readings

[1] Chapters 5 and 7 https://mml-book.github.io/

Additional readings

[2] https://mathinsight.org/thread/multivar

[3] Libretexts in Math (conic sections), and multivariable calculus