Multivariable Calculus

Yuhan Liu

ECE 4200
2/7/20 Recitation

Outline

- Machine learning as optimization problem
- Partial derivative
- Gradient
  - Property of gradient
  - Hessian (second order derivative)
  - Application to optimization problems
- Jacobian matrix and chain rule
- Convex functions

References:

1. Convex function: http://ee364a.stanford.edu/lectures/functions.pdf
2. Multivariable calculus: http://sites.tufts.edu/andrewrosen/files/2012/02/Calc-III-Review.pdf
3. http://www.cs.cornell.edu/courses/cs6780/2019sp/lecture/03-erm.pdf

The slides are be credited to

1. https://igl.ethz.ch/teaching/tau/cg/cg2005/cg_ex6.ppt
2. http://www.robots.ox.ac.uk/~oval/
3. http://portal.unimap.edu.my/portal/page/portal30/Lecturer%20Notes/IMK/Semester%201%20Sidang%20Akademik%2020162017/EQT%20101/ZAINAB%20YAHYA/PD_1.pptx
4. http://macs.citadel.edu/zhangli/Courses-Taught/Fall2016/courses/math231/StewartCalcET8_14_05.ppt

Machine Learning as Optimization

- (X_i, y_i) i.i.d. from some joint distribution D (unknown),
- Loss function $l: R \times R \mapsto R^+$.
- Want to find a function $f \in H$ that hopefully recovers $y = f(X)$:

$$E_{(X,y)\sim D} l(f(X), y)$$

Machine Learning as Optimization

- Empirical Risk Minimization: in practice we only have access to i.i.d. data, so we minimize empirical loss instead:

$$\min_{f \in H} \frac{1}{n} \sum_{i=1}^{n} l(f(X_i), y_i)$$

- Sometimes f is parametrized by parameters $\theta \in R^d$, denoted as $f(x; \theta)$. New objective:

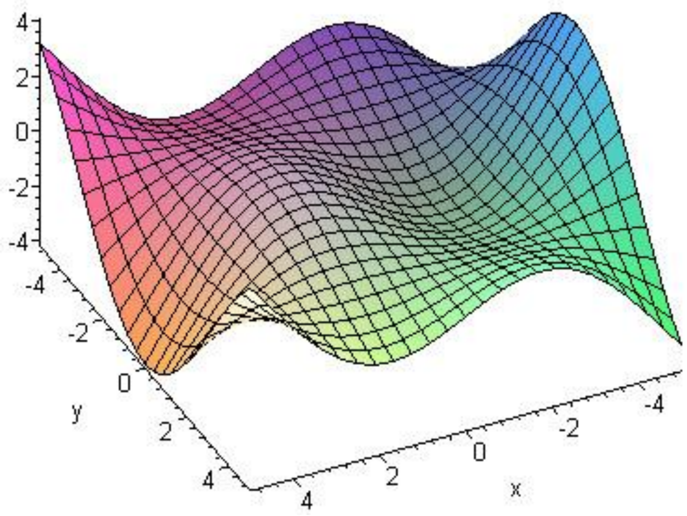
$$\min_{\theta \in R^d} \frac{1}{n} \sum_{i=1}^{n} l(f(X_i; \theta), y_i)$$

Optimization

- Typical optimization problem: $f: R^n \mapsto R$

$$\min f(x)$$
$$\text{s.t. } x \in \Omega \subseteq R^n$$

- n=1 is easy for f differentiable
- $n \geq 2$?

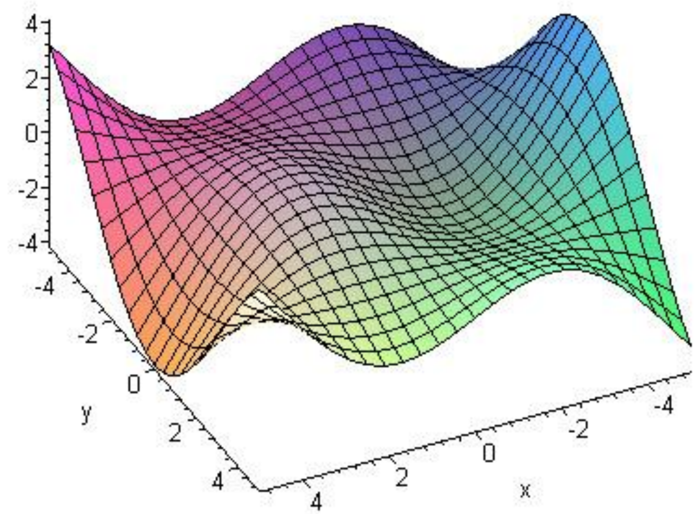


Basics

- Single variable, real valued function $f: R \mapsto R$
  - Derivative

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

- We want to generalize the notion of derivative to these cases:
  - Multivariable, real valued function $f: R^n \mapsto R$
  - Multivariable, vector valued function $\mathbf{f}: R^n \mapsto R^m$



Partial Derivative: Introduction

- Consider the multivariate function $f\left(x_1, x_2, \ldots, x_n\right)$ where $x_1, x_2, \ldots, x_n$ are independent variables.

- If we differentiate f with respect variable x_i, then we assume that
  i. x_i as a single variable
  ii. $x_1, x_2, \ldots x_{i-1}, x_{i+1}, \ldots, x_n$ as constants

PD: Definition

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

$$f_x(a,b) = \lim_{h \to 0} \frac{f(a+h,b) - f(a,b)}{h}$$

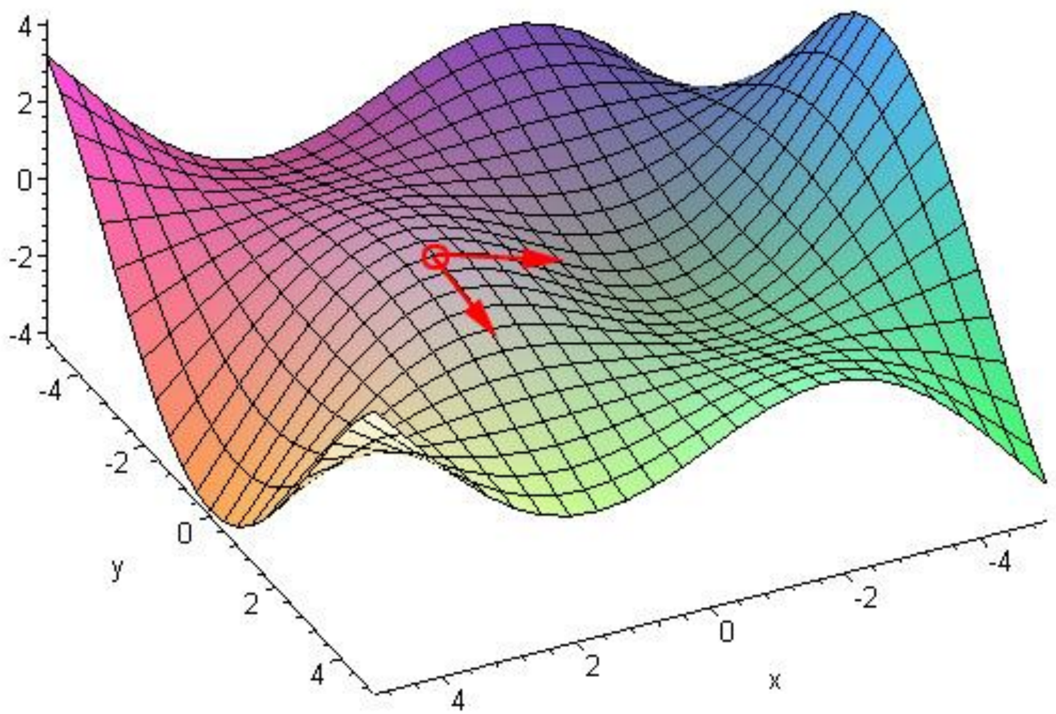
$$f_y(a,b) = \lim_{h \to 0} \frac{f(a,b+h) - f(a,b)}{h}$$

So basically just take the derivative of one (the subscript) given that the other one is a constant.

PD: Illustration

$$\frac{\partial f(x,y)}{\partial y}$$

$$\frac{\partial f(x,y)}{\partial x}$$



PD: Example

Write down all partial derivatives of the following function

$$f(x,y) = x^3 y^3 - 2x\cos(2y) + y^2 \ln x$$

PD: Example

Write down all partial derivatives of the following function

$$f(x,y) = x^3 y^3 - 2x\cos(2y) + y^2 \ln x$$

Solution

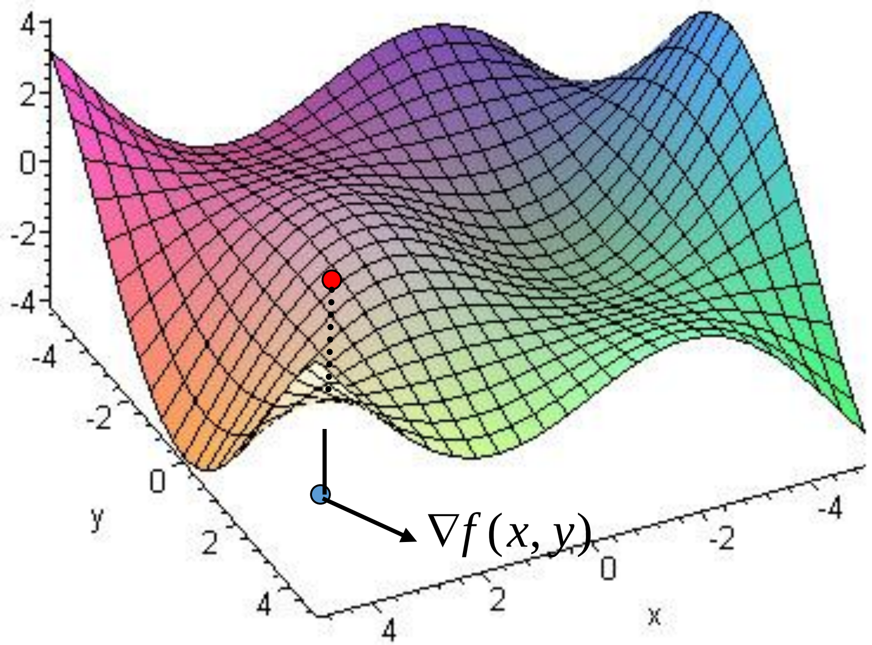
First order PD

$$\frac{\partial f}{\partial x} = 3x^2 y^3 - 2\cos(2y) + \frac{y^2}{x}$$

$$\frac{\partial f}{\partial y} = 3x^3 y^2 + 4x\sin(2y) + 2y\ln x$$

The Gradient: in $\boldsymbol{R^2}$

$$f : R^2 \to R \qquad \nabla f(x,y) := \begin{pmatrix} \dfrac{\partial f}{\partial x} & \dfrac{\partial f}{\partial y} \end{pmatrix}^{\mathsf{T}}$$



In the plane

The Gradient: Definition

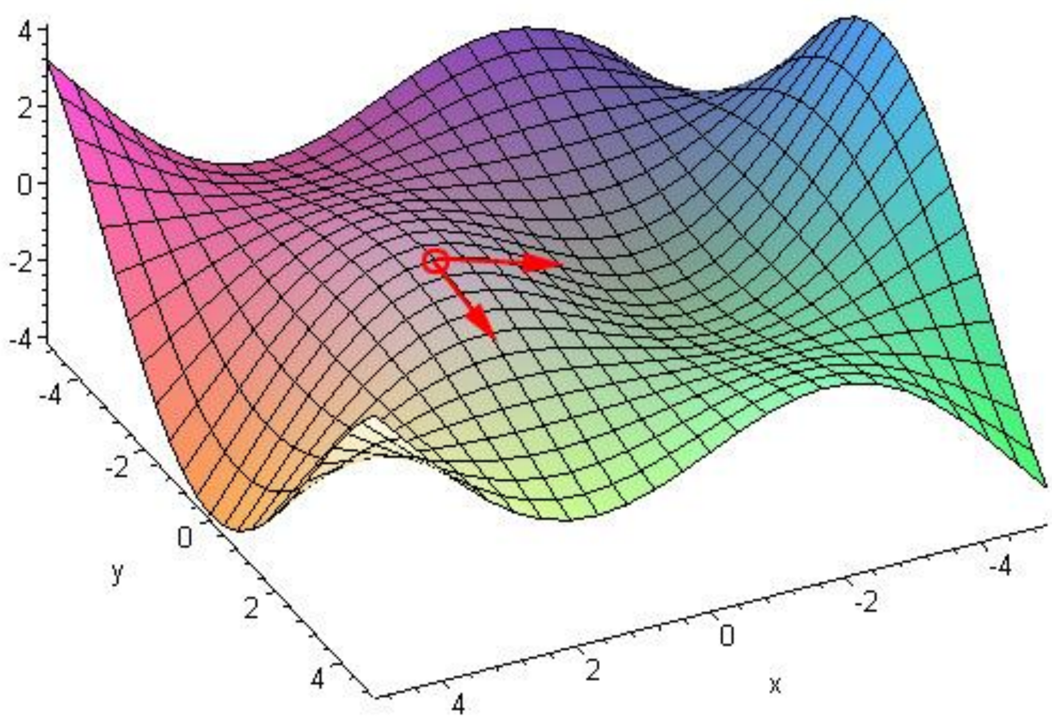
$$f : R^n \rightarrow R$$

$$\nabla f(x_1, \ldots, x_n) := \left(\frac{\partial f}{\partial x_1}, \ldots, \quad \frac{\partial f}{\partial x_n} \right)^{\mathsf{T}}$$

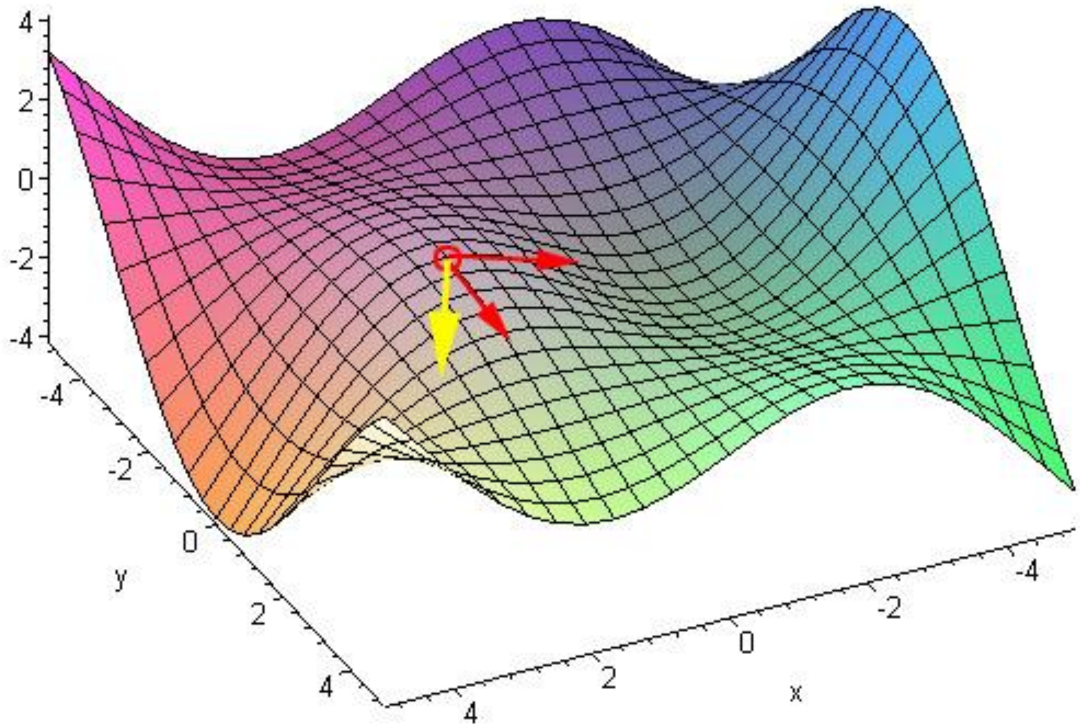
Directional Derivatives:
Along the Axes…

$$\frac{\partial f(x,y)}{\partial y}$$

$$\frac{\partial f(x,y)}{\partial x}$$



Directional Derivatives:
In general direction…



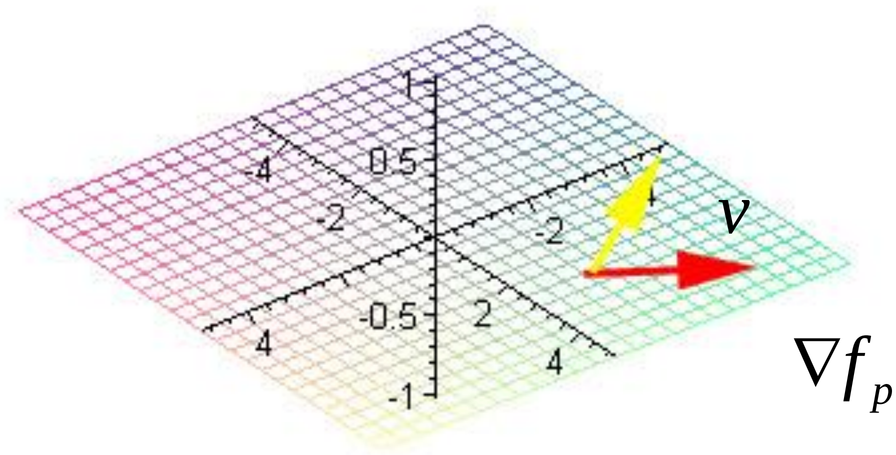
$$v = (v_x, v_y)$$

$$\|v\| = 1$$

$$\frac{\partial f(x,y)}{\partial \boldsymbol{v}} = \lim_{h \to \infty} \frac{f\left(x + hv_x\, , y + hv_y\right) - f(x,y)}{h}$$

The Gradient: Properties

$$\|v\| = 1 \qquad\qquad \frac{\partial f}{\partial \boldsymbol{v}}(\mathbf{p}) = \langle \nabla f_{\boldsymbol{p}} , \boldsymbol{v} \rangle$$



The Gradient: Properties

- **Proposition 1:**

$$\frac{\partial f}{\partial v}$$ **is maximal choosing** $$v = \frac{1}{\left\|\nabla f_p\right\|} \cdot \nabla f_p$$

is minimal choosing $$v = \frac{-1}{\left\|\nabla f_p\right\|} \cdot \nabla f_p$$

(intuition: the gradient points at the direction of greatest change rate)

The Gradient: Properties

- **Proposition 2**: let $f : R^n \longrightarrow R$ be differentiable around **p**,

if f has **local minimum** (maximum) at p, then

$$\nabla f_p = \bar{0}$$

(Intuitive: necessary for local min(max))

Gradient Descent

Quite simple algorithm. Goal:

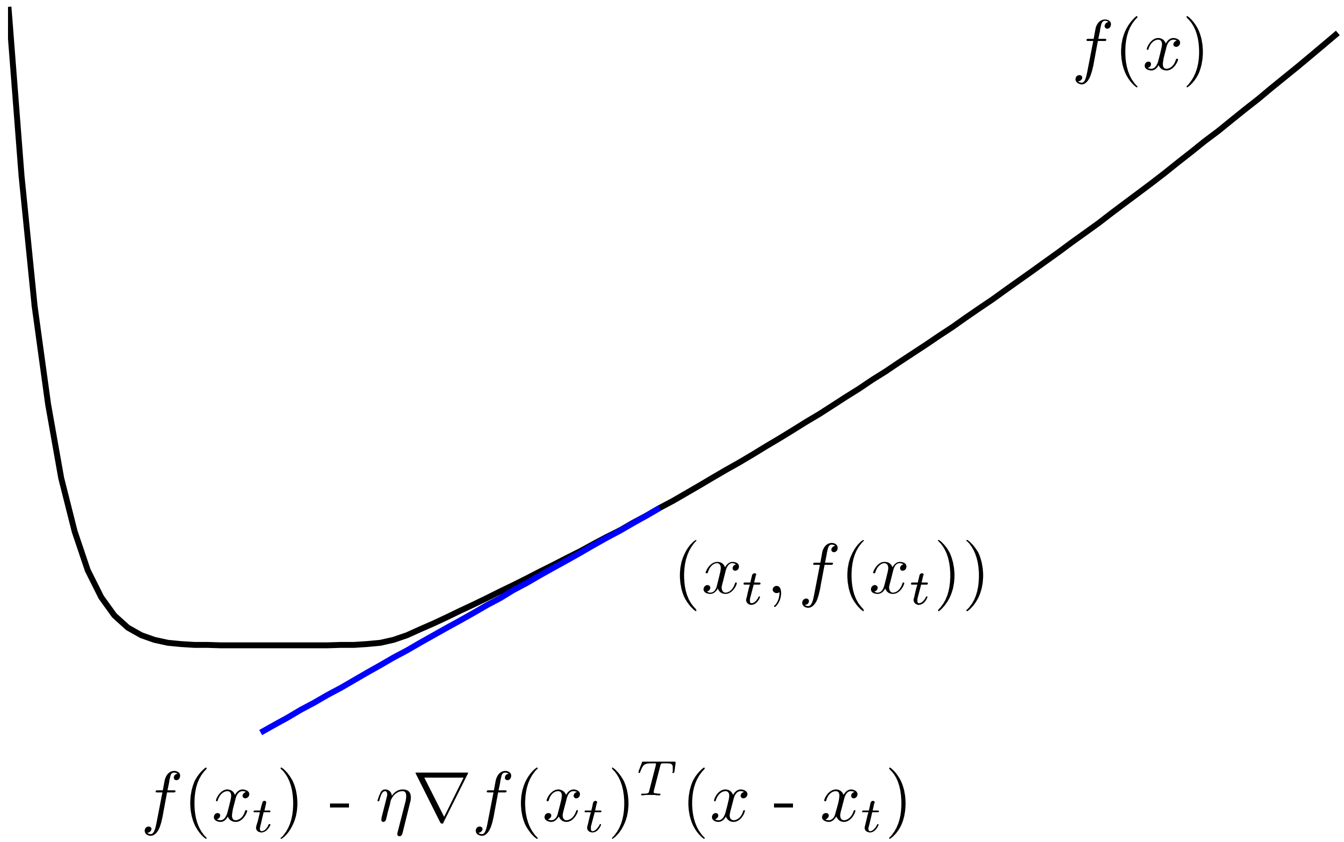
$$min_x f(x)$$

Just iterate

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

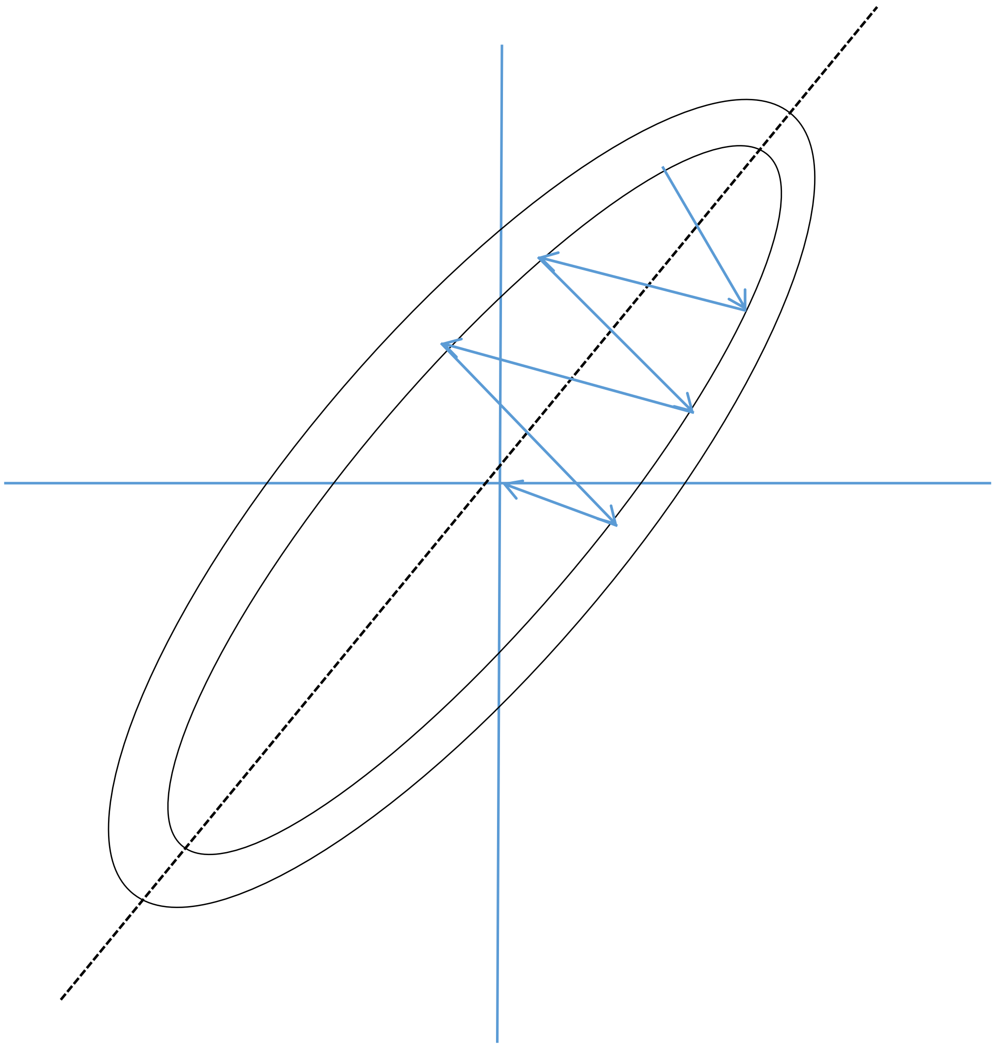
where η_t is the stepsize

Single Step Illustration



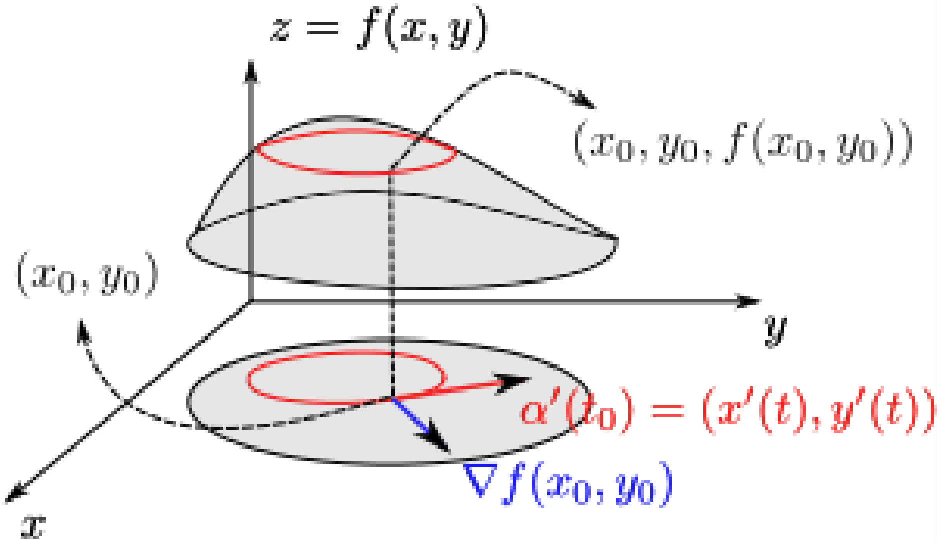
Imagine dropping a ball and let it roll down the slope

Full Gradient Descent Illustration



The Gradient: Properties

- **Proposition 3.** let $f: R^n \mapsto R$ be differentiable. At point x_0, let $f(x_0) = c$. The gradient $\nabla f(x_0)$ is orthogonal to the curve $f(x) = c$ (level curve)



Second order PD

$$\frac{\partial^2 f}{\partial x^2} = \frac{\partial}{\partial x}\left(\frac{\partial f}{\partial x}\right)$$

$$= \frac{\partial}{\partial x}\left(3x^2y^3 - 2\cos(2y) + \frac{y^2}{x}\right)$$

$$= 6xy^3 - \frac{y^2}{x^2}$$

$$\frac{\partial^2 f}{\partial y^2} = \frac{\partial}{\partial y}\left(\frac{\partial f}{\partial y}\right)$$

$$= \frac{\partial}{\partial y}\left(3x^3y^2 + 4x\sin(2y) + 2y\ln x\right)$$

$$= 6x^3y + 8x\cos(2y) + 2\ln x$$

Second order PD **(mixed partial)**

$$\frac{\partial^2 f}{\partial x \partial y} = \frac{\partial}{\partial x}\left(\frac{\partial f}{\partial y}\right)$$

$$= \frac{\partial}{\partial x}\left(3x^3 y^2 + 4x\sin(2y) + 2y\ln x\right)$$

$$= 9x^2 y^2 + 4\sin(2y) + \frac{2y}{x}$$

$$\frac{\partial^2 f}{\partial y \partial x} = \frac{\partial}{\partial y}\left(\frac{\partial f}{\partial x}\right)$$

$$= \frac{\partial}{\partial y}\left(3x^2 y^3 - 2\cos(2y) + \frac{y^2}{x}\right)$$

$$= 9x^2 y^2 + 4\sin(2y) + \frac{2y}{x}$$

Hessian Matrix

- $f: R^n \mapsto R$ $\qquad \nabla f(x_1,\ldots,x_n) := \left(\frac{\partial f}{\partial x_1}, \ldots, \frac{\partial f}{\partial x_n} \right)$
- Hessian: "Gradient of the gradient"
- $\nabla^2 f = \begin{bmatrix} \frac{\partial^2 f}{(\partial x_1)^2} & \frac{\partial^2 f}{\partial x_2 \partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_n \partial x_1} \\ \frac{\partial^2 f}{\partial x_1 \partial x_2} & \frac{\partial^2 f}{(\partial x_2)^2} & & \frac{\partial^2 f}{\partial x_n \partial x_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_1 \partial x_n} & \frac{\partial^2 f}{\partial x_2 \partial x_n} & \cdots & \frac{\partial^2 f}{(\partial x_n)^2} \end{bmatrix}$

The Gradient: Example

$$f(x, y) = 0.5x^2 + xy + 2y^2$$

$$\nabla f(x, y) = (x + y, x + 4y)^T$$

$$\nabla^2 f(x, y) = \begin{bmatrix} 1 & 1 \\ 1 & 4 \end{bmatrix}$$

Taylor Approximation

- Recall the one dimensional case:
- $f(x) - f(x_0) \simeq f'(x_0)(x - x_0) + \frac{1}{2} f''(x_0)(x - x_0)^2$
- Similar for multivariable functions:

$$f(x) - f(x_0) \simeq \left(\nabla f(x_0)\right)^T (x - x_0) +$$
$$\frac{1}{2}(x - x_0)^T \nabla^2 f(x_0)(x - x_0)$$

One Variable Chain Rule

- If $y = f(x)$ and $x = g(t)$, where f and g are differentiable functions, then

$$\frac{dy}{dt} = \frac{dy}{dx}\,\frac{dx}{dy}$$

- For functions of more than one variable?

Jacobian Matrix

- Vector valued multivariable function $\mathbf{f}: R^n \mapsto R^m$
- $\mathbf{f}(x_1, \dots, x_n) = \left(f_1(x), \dots, f_m(x)\right)^T$
- Jacobian Matrix:
- $J = \frac{\partial(f_1, f_2, \dots, f_m)}{\partial(x_1, x_2, \dots, x_n)} = \begin{bmatrix} (\nabla f_1)^\top \\ \vdots \\ (\nabla f_m)^\top \end{bmatrix} \in R^{m \times n}$

The Chain Rule: General Case

- $\mathbf{g}: R^n \mapsto R^m$, $\mathbf{f}: R^m \mapsto R^k$. Let $\mathbf{z} = \mathbf{f}\big(\mathbf{g}(\mathbf{x})\big)$, then:

$$\frac{\partial(z_1,z_2,\ldots,z_k)}{\partial(x_1,x_2,\ldots,x_n)} = \frac{\partial(f_1,f_2,\ldots,f_k)}{\partial(g_1,g_2,\ldots,g_m)} \frac{\partial(g_1,g_2,\ldots,g_k)}{\partial(x_1,x_2,\ldots,x_n)}$$

The Chain Rule: Case 1

- Suppose that $z = f(x, y)$ is differentiable function of x and y. And $x = g(t)$ and $y = h(t)$ are both differentiable function of t. Then z is a differentiable function of t and

$$\boxed{\frac{dz}{dt} = \frac{\partial z}{\partial x}\frac{dx}{dt} + \frac{\partial z}{\partial y}\frac{dy}{dt}}$$

Example:

- If $z = x^2y + 3xy^4$, where $x = \sin 2t$ and $y = \cos t$, find dz/dt when $t = 0$.

- Solution:
- The Chain Rule gives

$$\frac{dz}{dt} = \frac{\partial z}{\partial x}\frac{dx}{dt} + \frac{\partial z}{\partial y}\frac{dy}{dt}$$

$$= (2xy + 3y^4)(2\cos 2t) + (x^2 + 12xy^3)(-\sin t)$$

The Chain Rule: Case 2

- We now consider the situation where $z = f(x, y)$ but each of x and y is a function of two variables s and t: $x = g(s, t)$, $y = h(s, t)$.

- Then z is indirectly a function of s and t and we wish to find $\partial z/\partial s$ and $\partial z/\partial t$.

The Chain Rule: Case 2

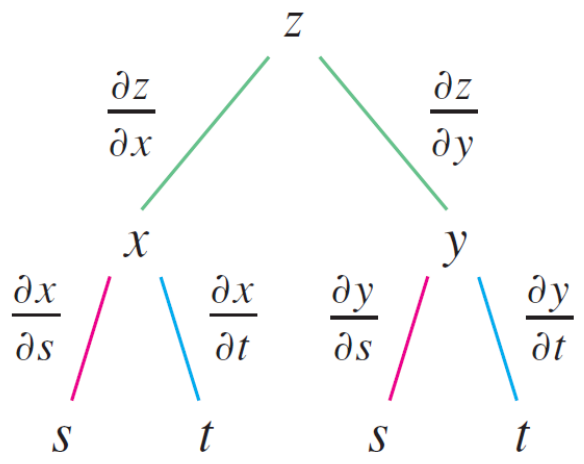
Suppose that $\mathrm{z} = f(x, y)$ is differentiable function of x and y , where f_x and f_y are continuous. And $\mathrm{x} = g(s, t)$ and $\mathrm{y} = h(s, t)$ are both differentiable function of t. Then z is a differentiable function of t and

$$\frac{\partial z}{\partial s} = \frac{\partial z}{\partial x}\frac{\partial x}{\partial s} + \frac{\partial z}{\partial x}\frac{\partial x}{\partial s}, \qquad \frac{\partial z}{\partial t} = \frac{\partial z}{\partial x}\frac{\partial x}{\partial t} + \frac{\partial z}{\partial y}\frac{\partial y}{\partial t}$$

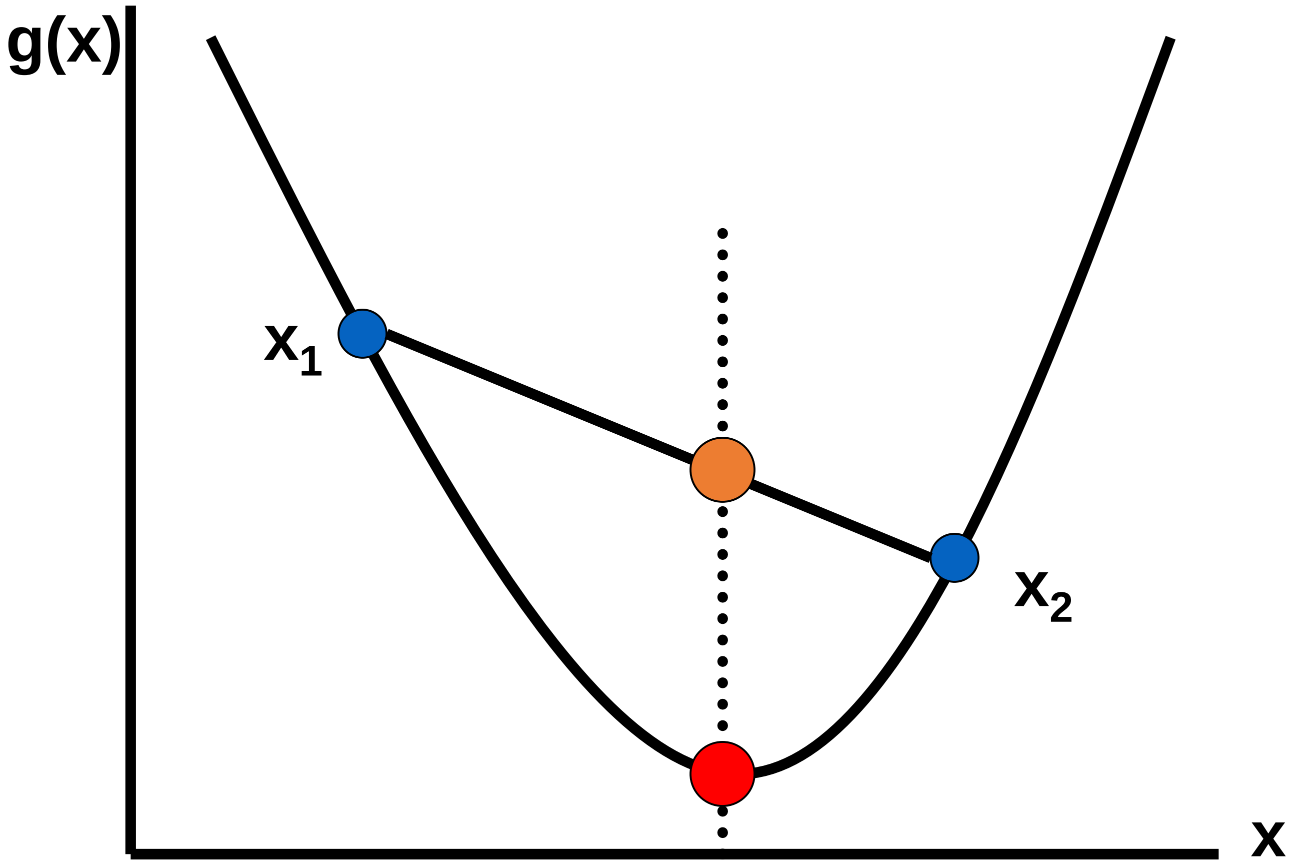
Case 2 of the Chain Rule contains three types of variables: *s* and *t* are **independent** variables, *x* and *y* are called **intermediate** variables, and *z* is the **dependent** variable.

The Chain Rule

- To remember the Chain Rule, it’s helpful to draw the **tree diagram.**

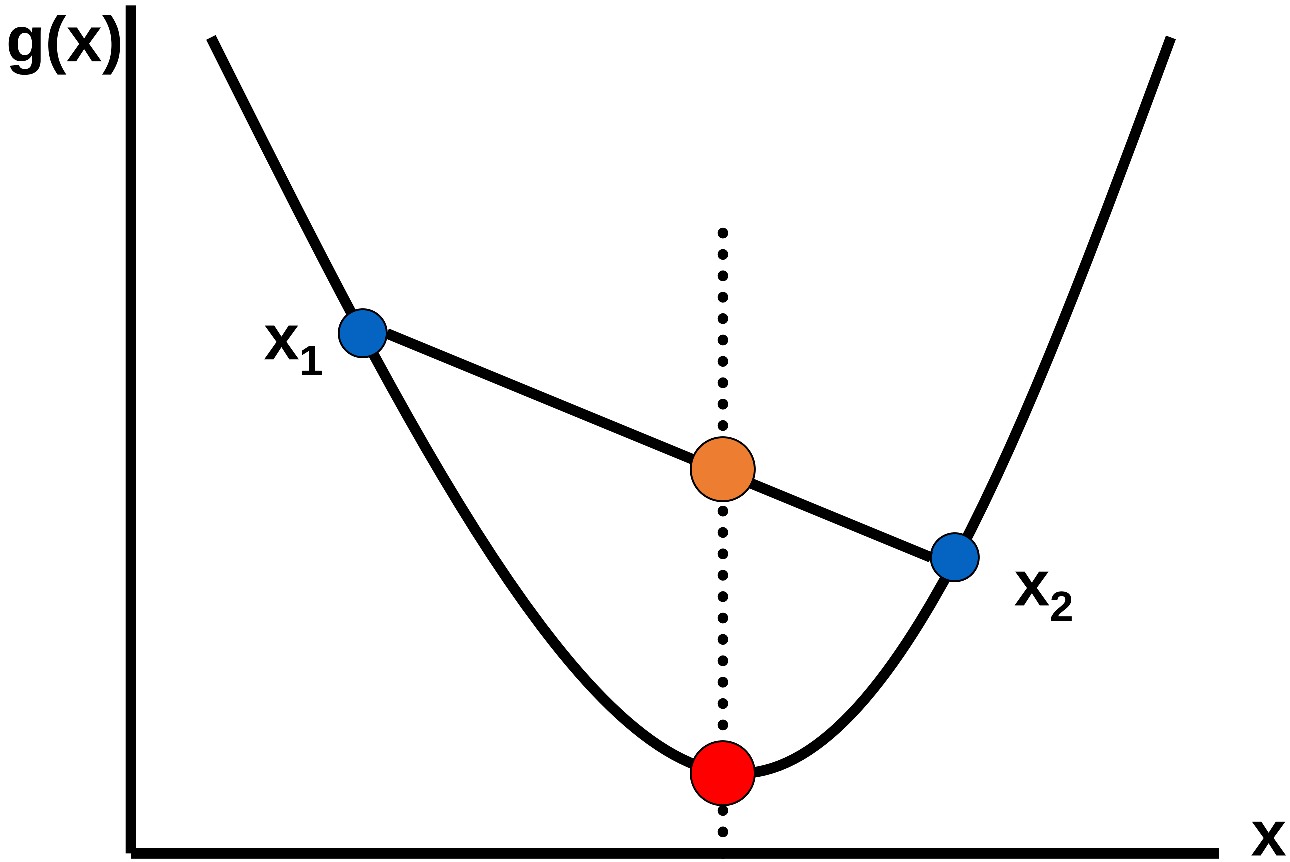


Convex Function



Orange point always lies above red point

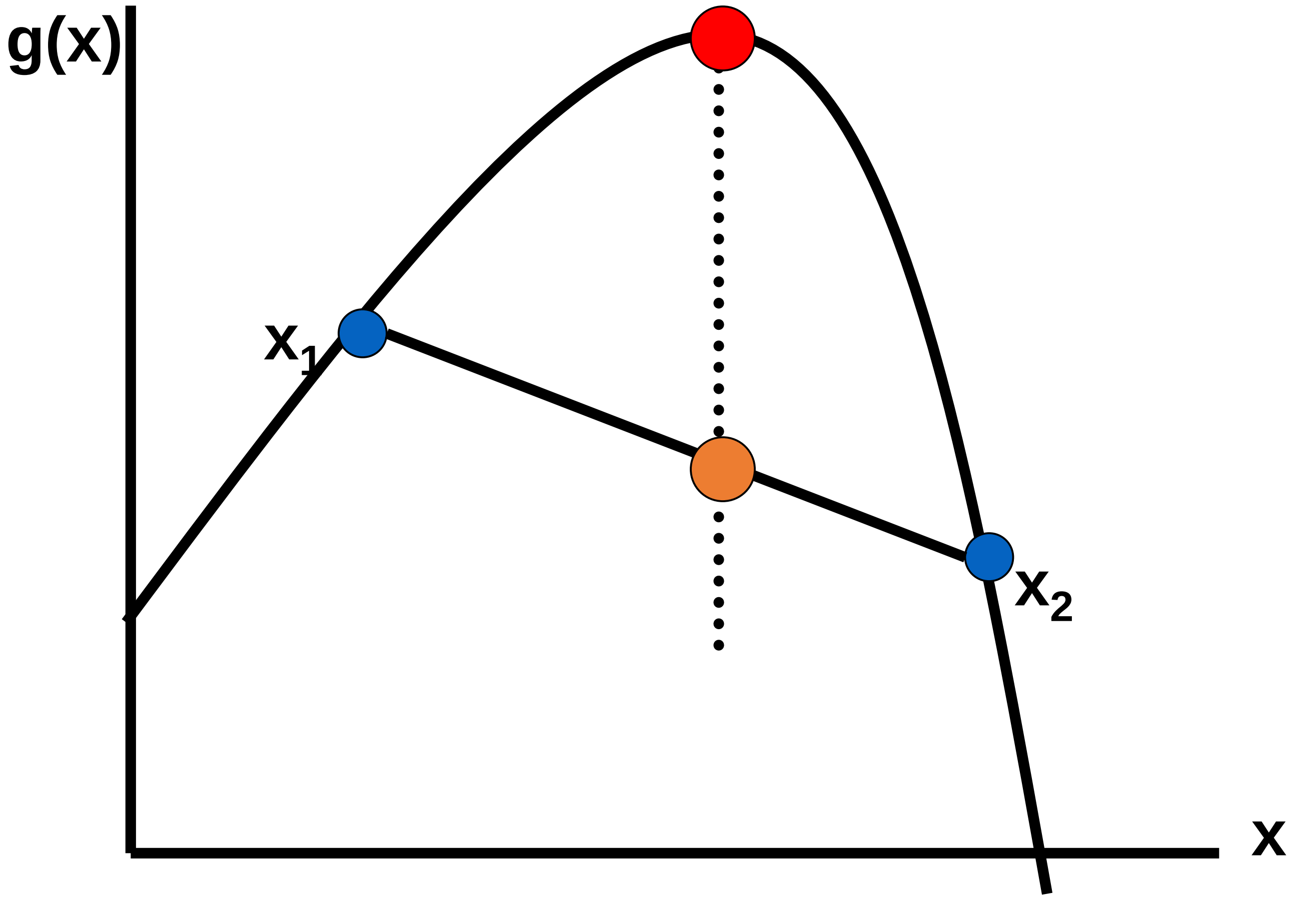
Convex Function: Definition



$$g(c\,\mathbf{x}_1 + (1 - c)\,\mathbf{x}_2) \leq c\,g(\mathbf{x}_1) + (1 - c)\,g(\mathbf{x}_2)$$

-g(.) is concave

Concave Function: Definition

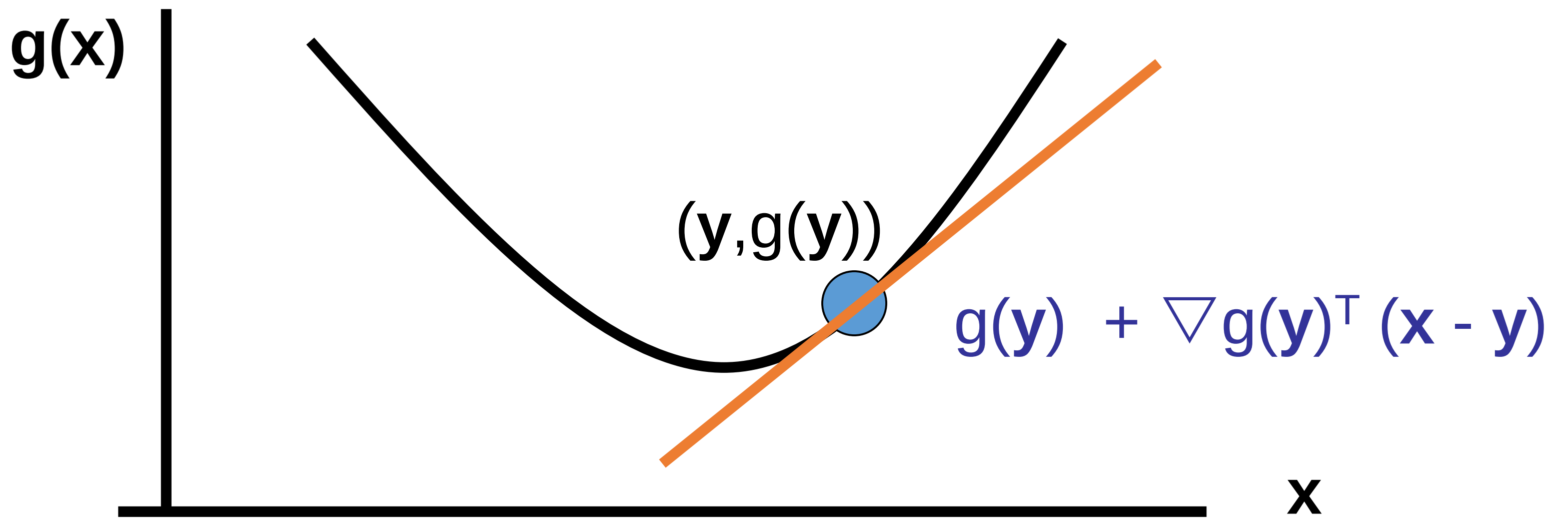


$$g(c\,\mathbf{x}_1 + (1 - c)\,\mathbf{x}_2) >= c\,g(\mathbf{x}_1) + (1 - c)\,g(\mathbf{x}_2)$$

Convex Function: Definition

Once-differentiable functions

$$g(\mathbf{y}) + \nabla g(\mathbf{y})^T (\mathbf{x} - \mathbf{y}) \leq g(\mathbf{x})$$



Twice-differentiable functions

$$\text{Hessian } H(\mathbf{x}) \succeq 0$$

Convex Function: Properties

If the functions f and g are convex (concave), then any *linear combination*

$$af + bg$$

where a, b are positive real numbers is also convex(concave).

Convex Function: Properties

If the function $u = g(x)$ is convex, and the function $y = f(u)$ is convex and non-decreasing, then the *composite function*

$$y = f(g(x))$$

is also convex.

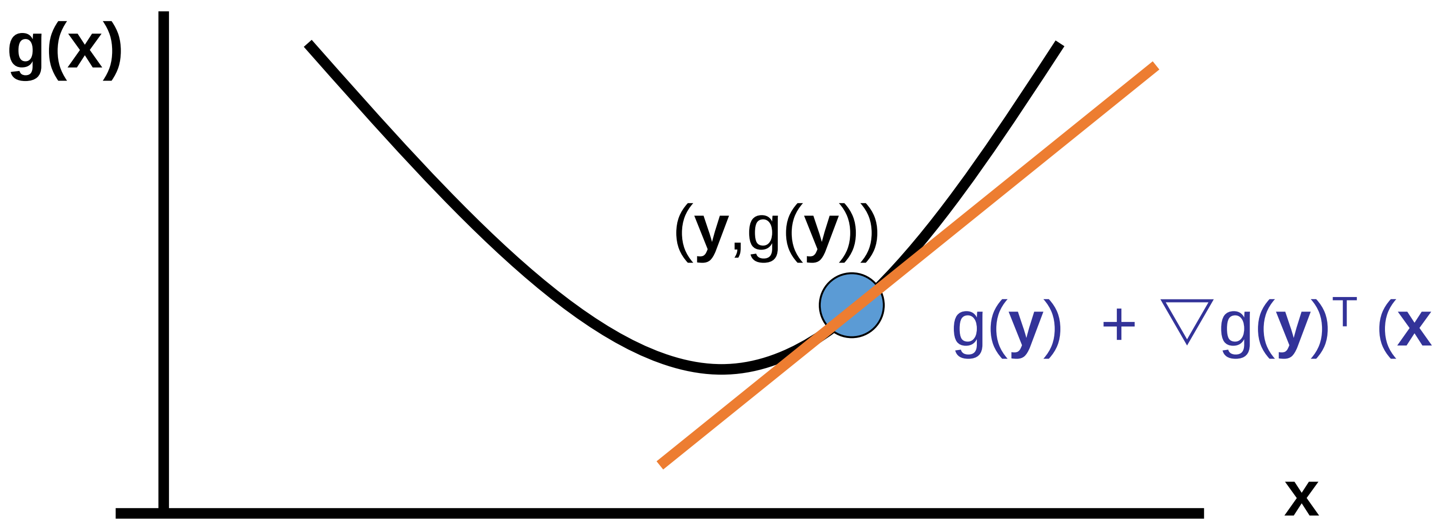
If the function $u = g(x)$ is concave, and the function $y = f(u)$ is convex and non-increasing, then the *composite function*

$$y = f(g(x))$$

is convex.

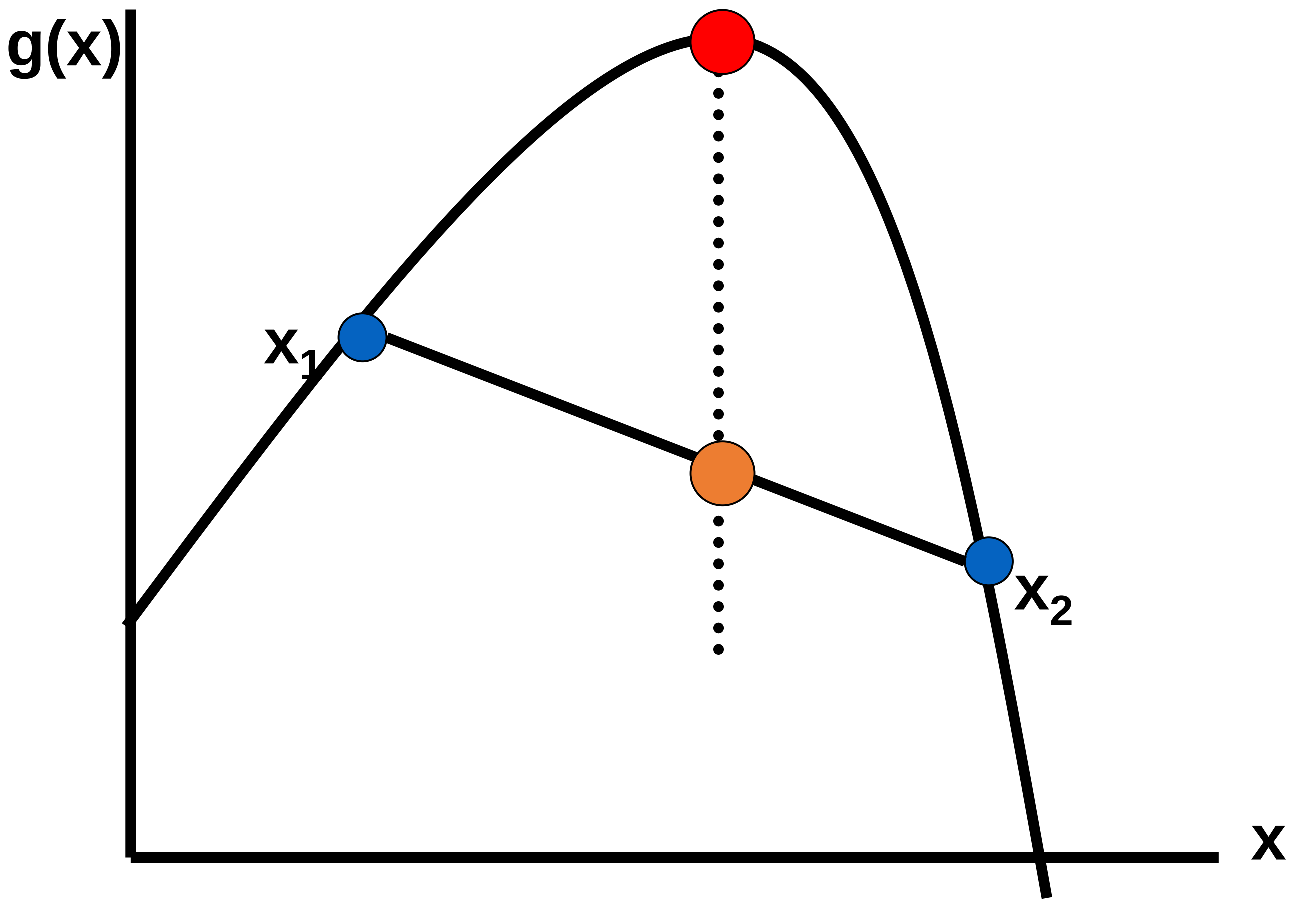
Convex Function: Properties

Any *local minimum* of a convex function defined on the interval [a,b] is also its *global minimum* on this interval.



Convex Function: Properties

Any *local maximum* of a concave function defined on the interval [a,b] is also its *global maximum* on this interval.



Examples of Convex Functions

Linear function $\mathbf{a}^{\mathrm{T}}\mathbf{x}$

$$\nabla f(x, y) = \mathbf{a}$$

$$\nabla^2 f(x, y) = \mathbf{0}$$

Quadratic functions $\mathbf{x}^{\mathrm{T}}\, \mathbf{Q}\, \mathbf{x}$ is convex when $\mathbf{Q} \succeq 0$

$$\nabla f(x, y) = 2\mathbf{Q}\mathbf{x}$$

$$\nabla^2 f(x, y) = 2\mathbf{Q}$$

Least square objective: $|A\mathbf{x} - \mathbf{b}|^2$

$$\nabla f(x, y) = 2A^{\mathrm{T}}(A\mathbf{x} - \mathbf{b})$$

$$\nabla^2 f(x, y) = 2A^{\mathrm{T}}A$$

Minimizing Convex Functions

Happens all the time …

Convex minimization has applications in a wide range of disciplines, such as automatic control systems, estimation and signal processing, communications and networks, electronic circuit design,[2] data analysis and modeling, finance, statistics (optimal experimental design),[3] and structural optimization.[4]