Convex Optimization

Lecture 1: Unconstrained Optimization for Differentiable Functions

Lecturer: *Dr*. Wan-Lei Zhao

Autumn Semester 2025

Contact: wlzhao@xmu.edu.cn

Outline

Syllabus

1. Optimization for Unconstrained Differentiable Functions (2)
2. Linear Programming
   a. Introduction (2)
   b. Simplex (4)
   c. Degenerated and Two-phase Simplex (2)
   d. Duality (2)
3. Quadratic Optimization
   a. Introduction (2)
   b. Convex Set (2)
   c. Convex Function and Convex Problem (3)
   d. Lagrange Multiplier and KKT (2)
   e. Dual of Lagrangian (2)
4. Portfolio Problem (2)
5. Support Vector Machine (2)
6. Integer Programming (2)

Assessment

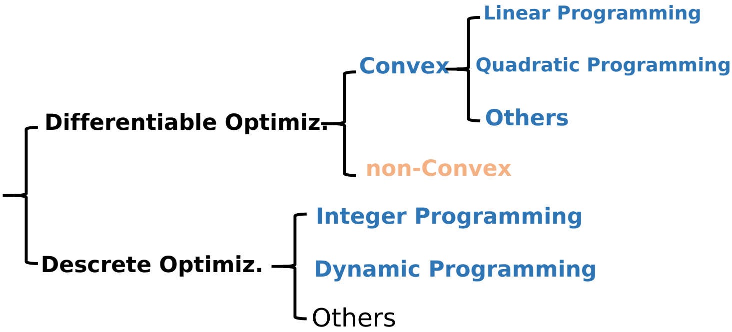
1. Assignment (10%)
   - Coding
   - Manual calc. exercises
2. Attendence (10%)
3. Middle exam (30%)
4. Final exam (50%)

$$Score = 0.1 \times \sum_i Assignment_i + 0.1 \times Attend. + 0.3 \times mid + 0.5 \times final$$

- No cheating!
- No bargaining!
- No chatGPT!

Covered Topics

- **Optimization**: seek for an optimal solution for a defined objective function
- Under/without constraints



Nothing happens in the universe that does not have a sense of either certain maximum or minimum.

—Leonhard Euler (1707 – 1783)

Outline

Derivatives on Function of Vector Variable

- Given $f(x) = a{\cdot}x_1 + b{\cdot}x_2$, $x = [x_1, x_2]^T \in R^2$
- Take partial derivative on $f(x)$, we have

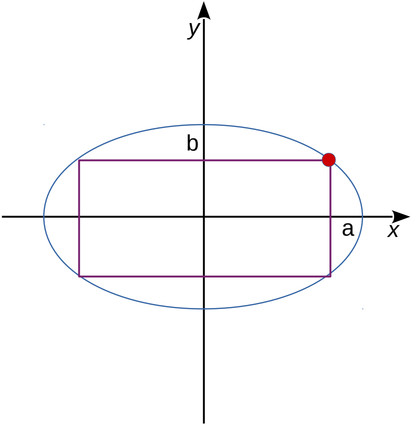
$$\frac{\partial f(x)}{\partial x_1} = a$$
$$\frac{\partial f(x)}{\partial x_2} = b$$

- The gradient of $f(x)$ is $[a, b]^T$ or $[a, b]$ depending on the context

Extreme Values of Univariat Function (1-1)

- Find out the inscribed rectangular of an ellipse that holds maximum area

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$$



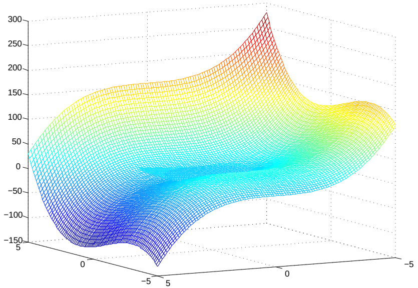
Extreme Values of Univariat Function (1-2)

- Find out the maximum point of the area function A(x)
  1. $A(x) = 4 * x * b * \sqrt{(1 - x^2/a^2)}$
  2. A'(x) = 0

```
syms a, b, x;
A = 4*b*x*(1 - x^2/a^2)^(1/2)
dA = diff(A,x)
solve(4*b*(1 - x^2/a^2)^0.5 - (4*b*x^2)/(a^2*(1 - x^2/a^2)^0.5)==0,x)
```

- Answer: $x = \frac{\sqrt{2}a}{2}$, viz $\arg\max_x A(x) = \frac{\sqrt{2}a}{2}$
- Max(A(x)) = 2*a*b

Extreme Values of multi-variable Function without Constraint (2-1)



- Given function:

$$f(x,y) = x^3 - y^3 + 3x^2 + 3y^2 - 9x \quad (1)$$

- Calculate its Hessian matrix

$$H = \begin{bmatrix} \frac{\partial f^2(x,y)}{\partial x^2} & \frac{\partial f^2(x,y)}{\partial x \partial y} \\ \frac{\partial f^2(x,y)}{\partial y \partial x} & \frac{\partial f^2(x,y)}{\partial y^2} \end{bmatrix} \quad (2)$$

Extreme Values of multi-variable Function without Constraint (2-1)

$$H = \begin{bmatrix} \frac{\partial f^2(x,y)}{\partial x^2} & \frac{\partial f^2(x,y)}{\partial x \partial y} \\ \frac{\partial f^2(x,y)}{\partial y \partial x} & \frac{\partial f^2(x,y)}{\partial y^2} \end{bmatrix}$$

$$\Downarrow$$

$$H = \begin{bmatrix} 6x+6 & 0 \\ 0 & 6-6y \end{bmatrix}$$

$$\Downarrow$$

$$\mid H \mid = (6x+6)(-6y+6) - 0 = -36xy - 36y + 36x + 36 \quad (3)$$

Extreme Values of multi-variable Function without Constraint (2-2)

- Given function:

$$f(x,y) = x^3 - y^3 + 3x^2 + 3y^2 - 9x \tag{4}$$

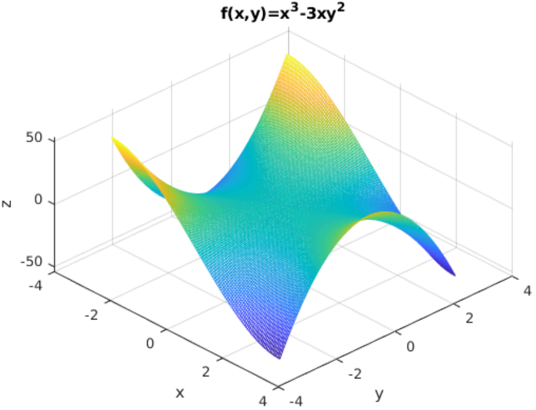
$$\begin{cases} \frac{\partial f(x,y)}{\partial x} = 3x^2 + 6x - 9 = 0, \Rightarrow x = -3, 1 \\ \frac{\partial f(x,y)}{\partial y} = -y^2 + 6y = 0, \Rightarrow y = 0, 2 \end{cases}$$

x	y	$a = f_{xx}$	$\mid H(x,y) \mid = a \cdot c - b^2$	f	Ext.
-3	0	-12	-72	27	uncertain
-3	2	-12	72	31	Ext. large
1	0	6	72	-5	Ext. small
1	2	6	-72	-1	uncertain

Outline

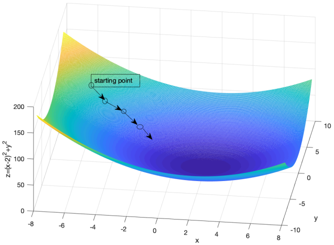
Why Gradient Descent?

- In practice, it is very hard to solve out a complex function

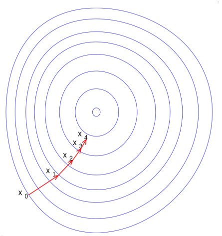


Why Gradient Descent?

- Given $f(x,\ y)$
- We want to find its minimum value



(a)



(b)

Derivative of a function and its tangent

$$g(x) = 2x^3 + 3x^2 - 12x + 7$$

- Given g(x), tangent at x_0 is defined as:

$$f(x) = g'(x_0)(x - x_0) + g(x_0)$$

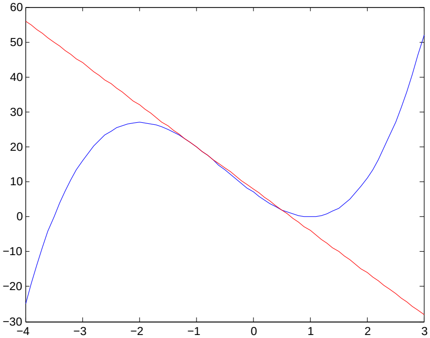


Figure: g(x) is curve in blue, tangent of g(x) at x=-1 is curve in red.

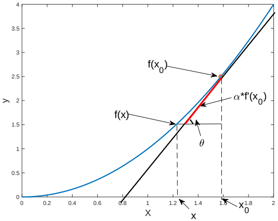
The gradient descent procedure

- Given we are going to find the minimum of $f(x)$, $x \in R^d$
- The learning rate is α

1. *Initialize*(x)
2. Repeat
   a. $x^+ := x - \alpha f'(x)$
   b. $x = x^+$

- Once we reaches the extreme, $f'(x) = 0$
- The procedure converges

Why Gradient Descent? (1)



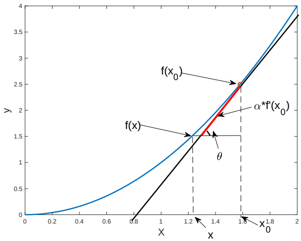
- Given $f(x)$, $x \in R^d$, $f'(x_0) = \frac{\Delta y}{\Delta x}$
- We want to find its extreme value
- According to Taylor expansion $f(x) \approx f(x_0) + (x - x_0)f'(x_0)$

$$f(x) \approx f(x_0) + (x - x_0)f'(x_0) \tag{5}$$

- According to the above figure, we have

$$f(x) = f(x_0) - f'(x_0)\Delta x \tag{6}$$

Why Gradient Descent? (2)



$$f(x) \approx f(x_0) + (x - x_0)f'(x_0)$$
$$f(x) = f(x_0) - f'(x_0)\Delta x$$

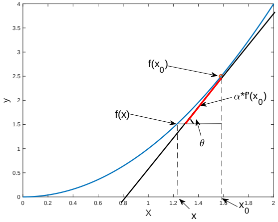
- Combining Eqn. 5 and Eqn. 6, we have

$$f(x_0)+(x - x_0)f'(x_0) = f(x_0) - f'(x_0)\Delta x \\ \Rightarrow x = x_0 - \Delta x \tag{6}$$

- Given α is the stepsize, $\Delta x = \alpha \cdot f'(x_0)$, we have

$$x = x_0 - \alpha \cdot f'(x_0) \tag{7}$$

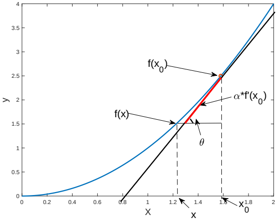
Why Gradient Descent? (3)



- Moreover, $\theta \rightarrow 0$, x drops the steepest

$$
\begin{aligned}
f(x_0)+(x-x_0)f'(x_0) &= f(x_0) - f'(x_0)\Delta x \\
\Longrightarrow x &= x_0 - \Delta x \\
\Longrightarrow x &= x_0 - \alpha f'(x_0)
\end{aligned} \tag{6}
$$

The General Steps in Gradient Descent

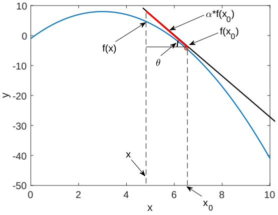


$$x = x_0 - \alpha f'(x_0) \tag{6}$$

1. *Initialize*(x)
2. Repeat
   a. $x^+ := x - \alpha f'(x)$
   b. $x = x^+$

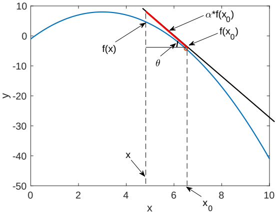
Why Gradient Ascent? (1)

- Given we want to find the maximum of $f(x)$, $x \in R^d$



$$f(x) = f(x_0) + (x - x_0)f'(x_0)$$
$$f(x) = f(x_0) + f'(x_0)\Delta x$$

Why Gradient Ascent? (2)

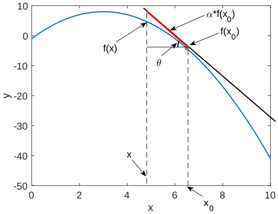


$$f(x) = f(x_0) + (x - x_0)f'(x_0)$$
$$f(x) = f(x_0) + f'(x_0)\Delta x$$

$$\Downarrow$$

$$x = x_0 + \alpha f'(x_0) \tag{6}$$

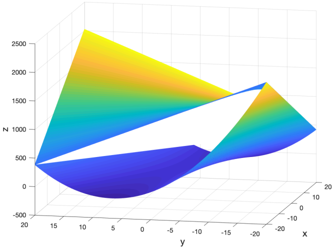
The General Steps in Gradient Descent



$$x = x_0 + \alpha f'(x_0) \tag{6}$$

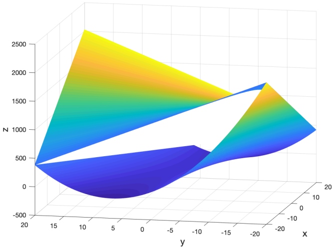
1. $\mathit{Initialize}(x)$
2. Repeat
   a. $x^+ := x + \alpha f'(x)$
   b. $x = x^+$

Minimum found by Gradient Descent (1)



$$f(x, y) = x^2 * y^3 + 3 * y^2 + 2 * x * y + x + 3; \tag{7}$$

Minimum found by Gradient Descent (2)



$$f(x, y) = x^2 * y^3 + 3 * y^2 + 2 * x * y + x + 3$$

$$Grad(f) = [2 * x * y^3 + 2 * y + 1, 3 * x^2 * y^2 + 6 * y + 2 * x]^T$$

Minimum found by Gradient Descent (3)

- Given learning rate $\alpha = 0.00005$

```
lr      = 0.00005;
xy    = rand(2,1);
iter = 0;
while iter < 500
    xy   = xy + lr*df(xy)';
    iter = iter+1;
end

function [fval]=f(xy)
   x = xy(1);  y = xy(2);
   fval = x.^2*y.^3+3*y.^2+2*x.*y+x+3;
end

function [dx, dy]=df(xy)
   x   = xy(1);  y = xy(2);
   dx = 2*x.*y.^3+2*y+1;
   dy = 3*x.^2.*y.^2+6*y+2*x;
end
```

Gradient with Momentum

- In order to speed up the convergence, or enhance the stability of gradient descent
- Gradient with Momentum is introduced

$$x_{t+1} = x_t - \alpha f'(x)$$

$$v_{t+1} = \beta v_t - \alpha f'(x)$$
$$x_{t+1} = x_t + v_{t+1}$$

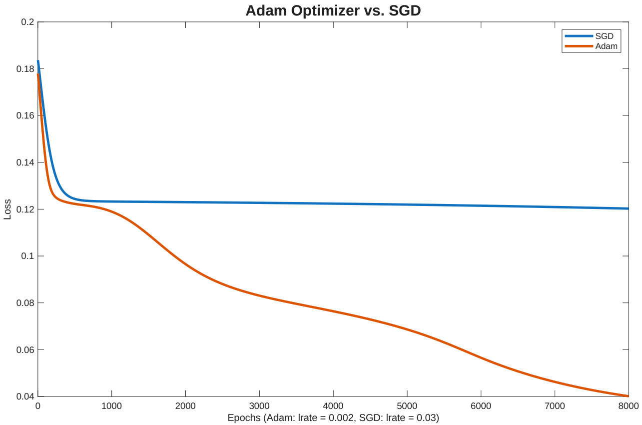
Adaptive moment (Adam) Optimizer

- Adaptive moment estimation performs pretty well[1]
- Widely adopted in various deep learning tasks

$$
\begin{aligned}
g_t &= f'(x) \\
m_t &= \beta_1 \cdot m_{t-1} + (1-\beta_1) \cdot g_t \\
v_t &= \beta_2 \cdot v_{t-1} + (1-\beta_2) \cdot g_t^2 \\
\hat{m}_t &= m_t / (1-\beta_1^t) \\
\hat{v}_t &= v_t / (1-\beta_2^t) \\
x_t &= x_{t-1} - \alpha \hat{m}_t / (\sqrt{\hat{v}_t} + \epsilon)
\end{aligned}
$$

$\beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 1e-8, \alpha = 0.002$

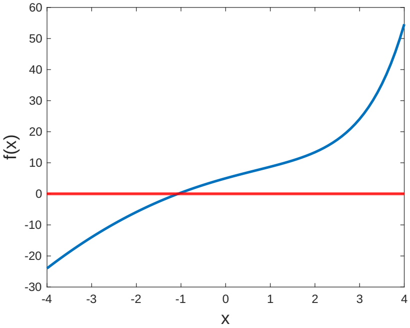
[1] ADAM: A Method for Stochastic Optimization, Diederik P. Kingma, Jimmy Lei Ba, ICLR 2015


Adam Optimizer vs. SGD
SGD
Adam
Loss
0.2
0.18
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0
1000
2000
3000
4000
5000
6000
7000
8000
Epochs (Adam: lrate = 0.002, SGD: lrate = 0.03)

Outline

The motivation

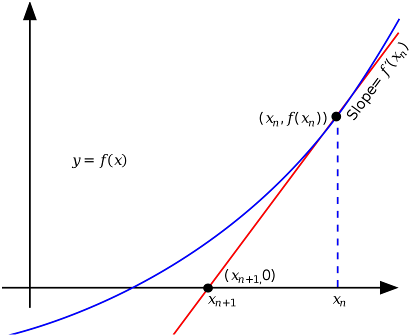
- We want to solve $f(x) = 0$, for instance $e^x - x^2 + 3x + 4 = 0$
- It is a little bit complicated



- We can solve it by an iterative procedure
- Which is known as Newton method

How it works? (1)

- We want to solve $f(x) = 0$, for instance $e^x - x^2 + 3x + 4 = 0$
- It is a little bit complicated

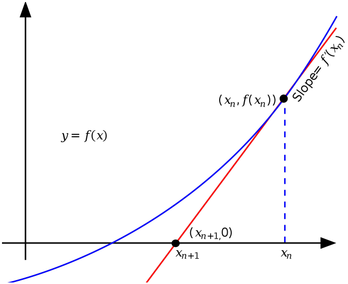


- Given we are at $(x_n, f(x_n))$, we wan to move to $(x_{n+1}, f(x_{n+1}))$,
- which is closer to $f(x) = 0$

How it works? (2)

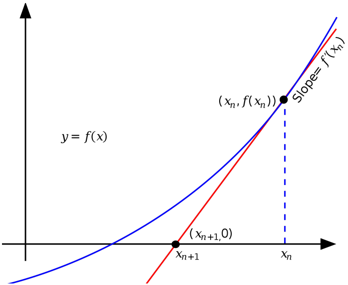
- The tangent line at $(x_n, f(x_n))$ is given as $y = f'(x_n)x + b$
- Since this line passes through $(x_n, f(x_n))$
- We have $y = f'(x_n)\cdot(x - x_n) + f(x_n)$
- We can easily find out its intersect point with x-axis

$$
\begin{aligned}
y = & f'(x_n)\cdot(x - x_n) + f(x_n) \\
\xrightarrow{y=0} & f'(x_n)\cdot(x - x_n) + f(x_n) \\
\Longrightarrow & x = \frac{x_n\cdot f'(x_n) - f(x_n)}{f'(x_n)} \\
\Longrightarrow & x = x_n - \frac{f(x_n)}{f'(x_n)}
\end{aligned}
$$



How it works? (3)

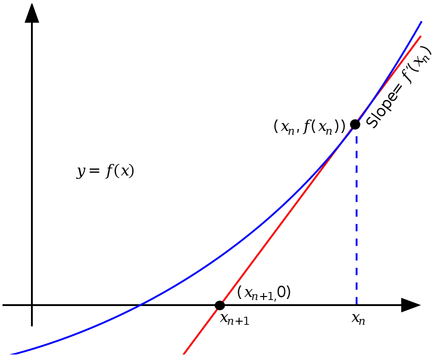
$$
\begin{aligned}
y &= f'(x_n)\cdot(x - x_n) + f(x_n) \\
&\xrightarrow{y=0} f'(x_n)\cdot(x - x_n) + f(x_n) \\
&\Longrightarrow x = \frac{x_n\cdot f'(x_n) - f(x_n)}{f'(x_n)} \\
&\Longrightarrow x = x_n - \frac{f(x_n)}{f'(x_n)}
\end{aligned}
$$



- $x = x_n - \frac{f(x_n)}{f'(x_n)}$ is the next iteration step
- The iteration continues until $f(x_n)$ reaches to 0

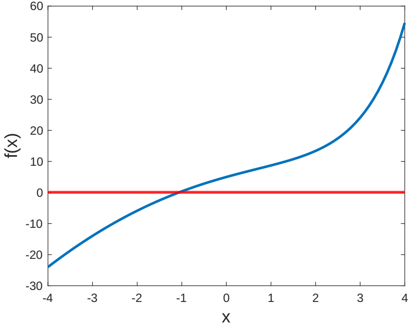
The Newton's method procedure

1. $x_n = x_o$
2. Repeat
   a. $x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$
   b. $x_n = x_{n+1}$
3. Until $f(x_n)$ close to 0



Practice with Newton's method (1)

- Solve $e^x - x^2 + 3x + 4 = 0$



- $f'(x) = e^x - 2x + 3$
- Notice that $f(x)$ is defined by ourselves

Practice with Newton's method (2)

- $f(x) = e^x - x^2 + 3x + 4$
- $f'(x) = e^x - 2x + 3$

```
function [x] = newtonsolve()
   xn    = 6;
   fval = 8;
   while fval > 0.001 do
       fval    = f(xn);
       dfval =df(xn);
       xp      = xn - fval/dfval;
       xn      = xp;
   end
   x = xn;
end

function [fval]=f(x)
    fval=exp(x) - x.^2+3*x + 4;
end

function [dfval]=df(x)
    dfval=exp(x)-2*x+3;
end
```

Outline

The Motivation

- Given a twice differentiable function: $f : R \rightarrow R$
- We have following minimization problem

$$\min_{x \in R} f(x) \tag{8}$$

- We want to construct a sequence $\{x_k\}$ based on initial guess x_0
- They allow $f(x_0) > f(x_1) > \cdots f(x_k) > \cdots > f(x^*)$
- Each time, we are going to search the minimum value within the neighborhood of x_k
- So that we jump to the local minimum, namely x_{k+1}

How it works? (1)

- Given we are at x_k, t is a small value
- Then we expand $f(x_k)$ by *Taylor expansion*

$$f(x_k + t) \approx f(x_k) + f'(x_k)t + \frac{1}{2}f''(x_k)t^2 \tag{9}$$

- The next location x_{t+1} is defined to minimize $f(\cdot)$, $x_{k+1} = x_k + t$
- Given $f''(x_k)$ is positive, the minimum exists, and we can find it by

$$0 = \frac{d}{dt}(f(x_k) + f'(x_k)t + \frac{1}{2}f''(x_k)t^2) = f'(x_k) + f''(x_k)t$$
$$\Rightarrow t = -\frac{f'(x)}{f''(x)} \tag{10}$$

How it works? (2)

$$f(x_k + t) \approx f(x_k) + f'(x_k)t + \frac{1}{2}f''(x_k)t^2 \tag{11}$$

- The next location x_{t+1} is defined to minimize $f(\cdot)$, $x_{k+1} = x_k + t$
- Given $f''(x_k)$ is positive, the minimum exists, and we can find it by

$$0 = \frac{d}{dt}(f(x_k) + f'(x_k)t + \frac{1}{2}f''(x_k)t^2) = f'(x_k) + f''(x_k)t$$

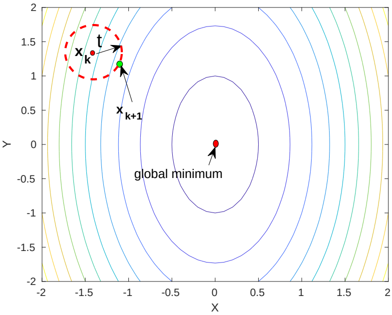
$$\Rightarrow t = -\frac{f'(x)}{f''(x)} \tag{12}$$

- Since $x_{k+1} = x_k + t$, we have

$$x_{k+1} = x_k - \frac{f'(x)}{f''(x)} \tag{13}$$

The Newton's method procedure for minimization

1. $x_k = x_0$
2. Repeat
   a. $x_{k+1} = x_k - \frac{f'(x_k)}{f''(x_k)}$
   b. $x_k = x_{k+1}$
3. Until $|\ f(x_k) - f(x_{k+1})\ |$ is close to 0



Extend Newton's method to Multiple Dimension Cases

- Given a twice differentiable function: $f : R^n \rightarrow R$
- We have following minimization problem

$$\min_{x \in R^n} f(x) \tag{14}$$

- The Hessian matrix for $f(x)$ is defined as $H_f(x) \in R^{d \times d}$

$$x_{k+1} = x_k - H_f(x)^{-1} f'(x) \tag{15}$$

- For stableness, we also introduce the step size α for the iteration

$$x_{k+1} = x_k - \alpha H_f(x)^{-1} f'(x) \tag{16}$$