
EE1001

Foundations of Digital Techniques and Data Analysis

Essence of Optimization – Descent Methods



Instructor: Jie Chen

DEPARTMENT OF ELECTRONIC ENGINEERING

Gradient Descent Method

- Intuition
 - Iterations
 - Examples
 - Local minimum and global minimum
 - Gradient ascent method
 - Local maximum and global maximum
-

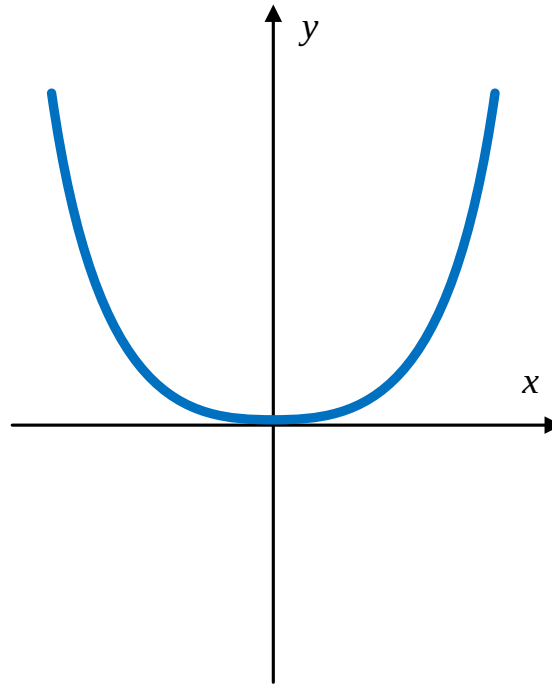
Cauchy and Gradient Descent Method



Louis Augustin Cauchy

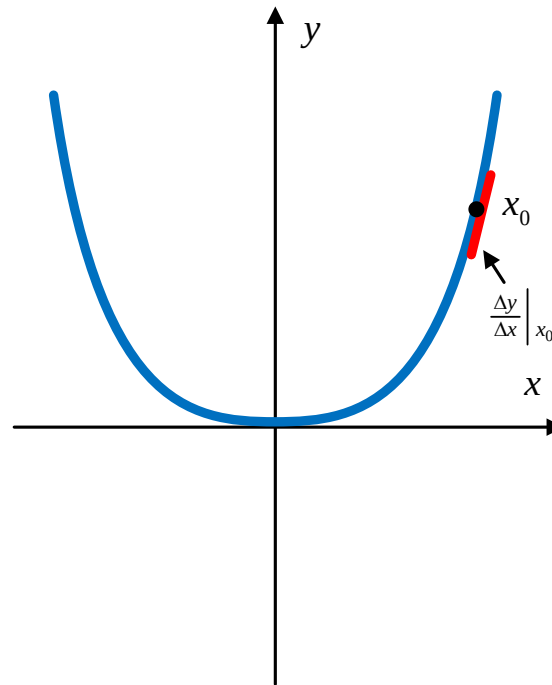
- Cauchy is motivated by astronomic calculations (to compute the orbit of a heavenly body) which, as everybody knows, are normally very voluminous
- In his book: *Methode generale pour la resolution des systemes d'equations simultanees*, 1847

Gradient Descent Method



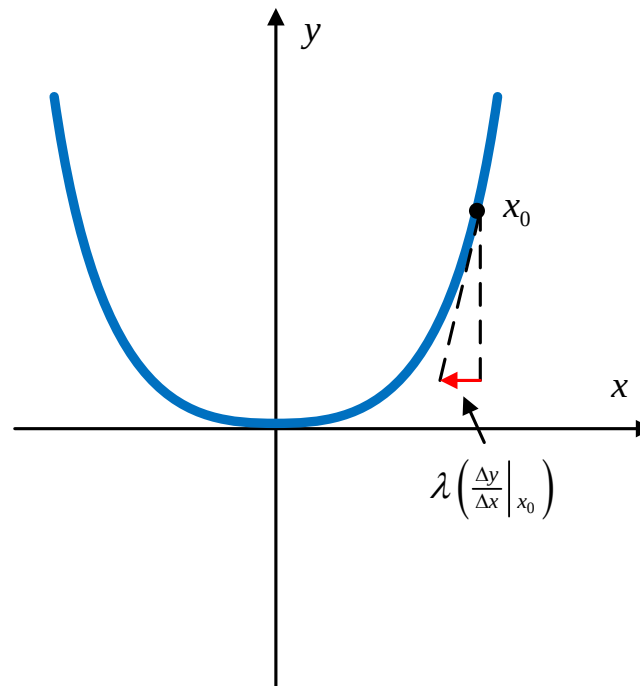
- Gradient descent method is a way to find a local minimum of a function

Gradient Descent Method



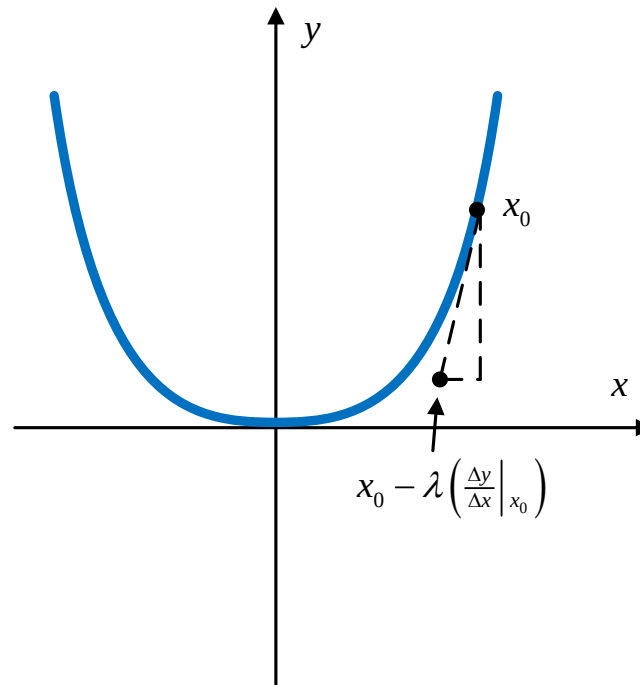
- Start with an initial guess of the solution: x_0

Gradient Descent Method



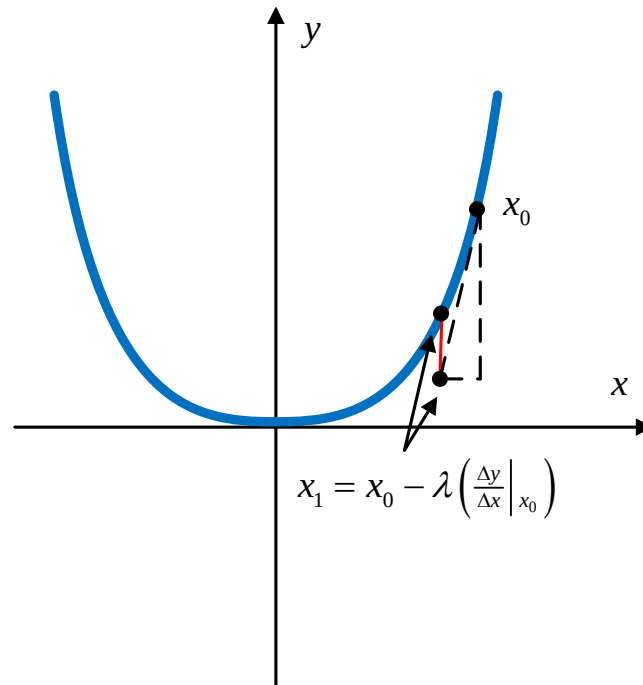
- Take the gradient of the function at that point
- Gradient at x_0 : $\frac{\Delta y}{\Delta x} \Big|_{x_0}$
- Step the solution in the opposite direction of the gradient: downhill

Gradient Descent Method



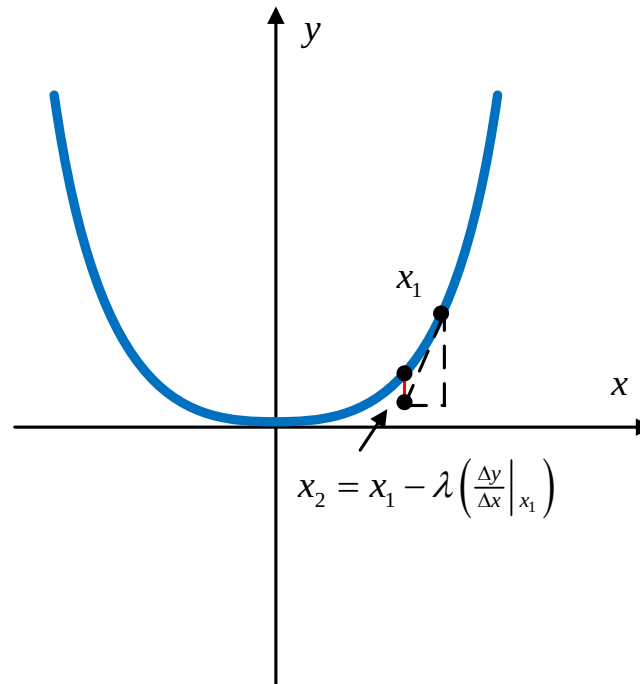
- Herein, the gradient is positive at x_0
- Step to the left, to $x_0 - \lambda \left(\frac{\Delta y}{\Delta x} \Big|_{x_0} \right)$
- λ is a positive constant

Gradient Descent Method



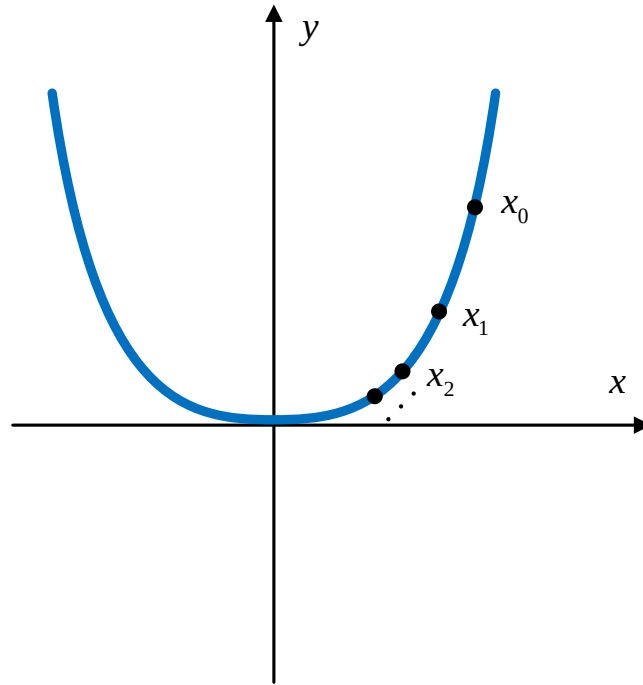
- Denote $x_1 = x_0 - \lambda \left(\frac{\Delta y}{\Delta x} \Big|_{x_0} \right)$, and repeat the process for x_1

Gradient Descent Method



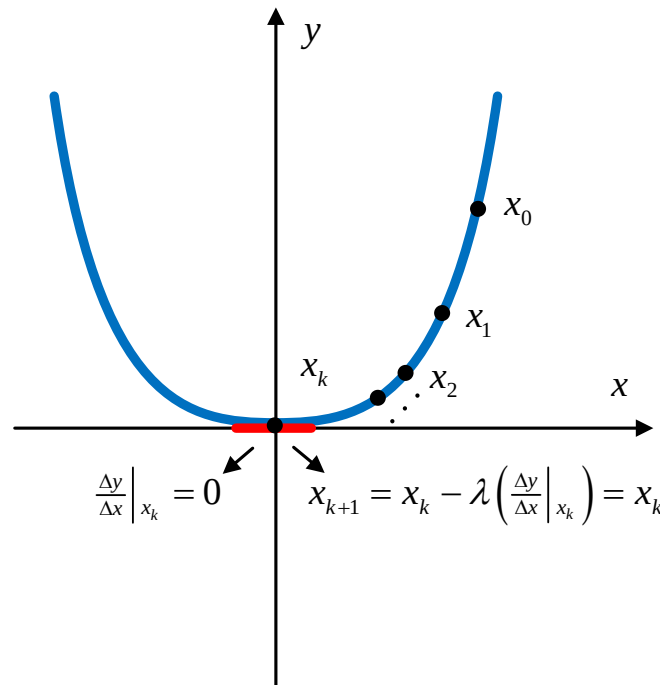
- By doing so, $x_2 = x_1 - \lambda \left(\frac{\Delta y}{\Delta x} \Big|_{x_1} \right)$ can be obtained

Gradient Descent Method



- By similar iterations, $x_3, x_4, \dots$ can be obtained

Gradient Descent Method



- The algorithm will eventually converge where the gradient is zero
- Gradient is zero: local minimum

Gradient Descent Method

Minimize $y = x^2$

- Take $x_0 = 1$, and $\lambda = 0.2$

- Gradient:

$$\frac{\Delta y}{\Delta x} = 2x$$

Gradient Descent Method

- Let $\lambda = 0.5$

- At $x_0 = 1$, $\left. \frac{\Delta y}{\Delta x} \right|_{x_0} = 2x_0 = 2$, and

$$x_1 = x_0 - \lambda \left(\left. \frac{\Delta y}{\Delta x} \right|_{x_0} \right) = 1 - 0.5 \times 2 = 0$$

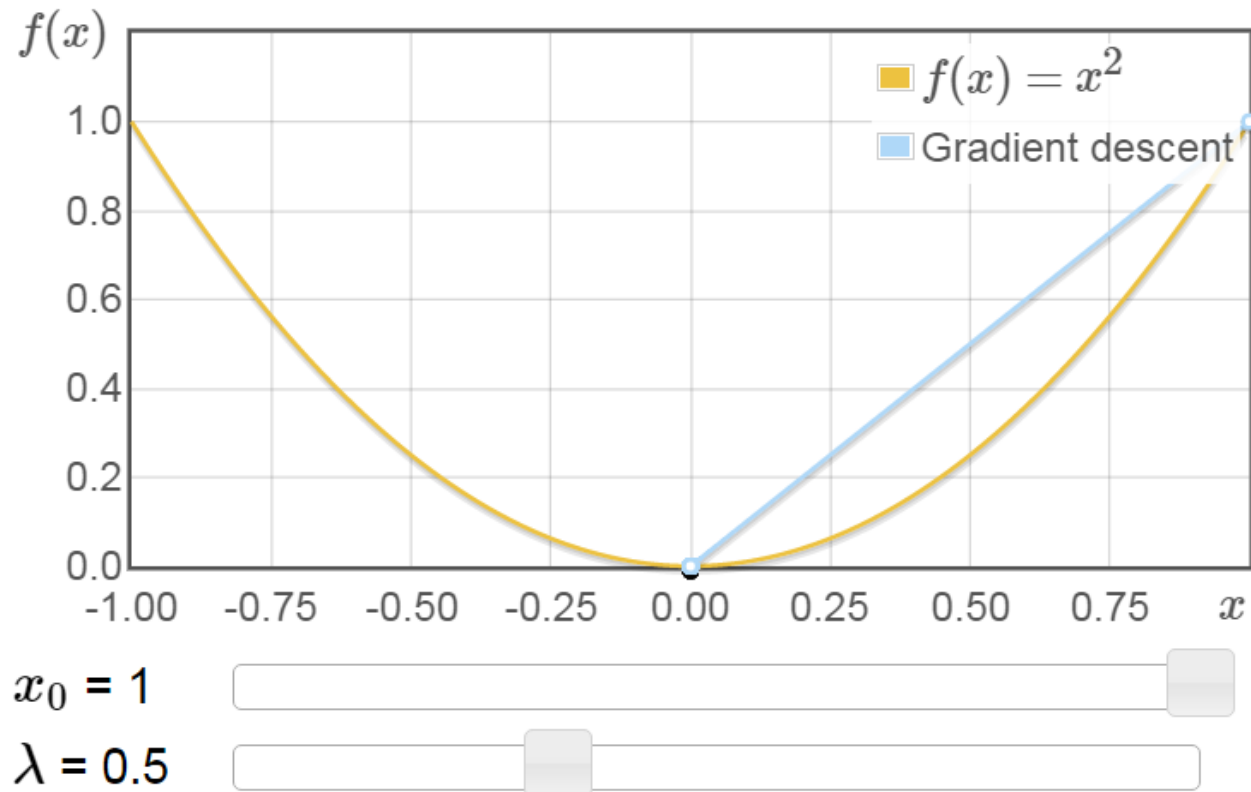
- At $x_1 = 0$, $\left. \frac{\Delta y}{\Delta x} \right|_{x_1} = 2x_1 = 0$, and

$$x_2 = x_1 - \lambda \left(\left. \frac{\Delta y}{\Delta x} \right|_{x_1} \right) = 0 - 0.5 \times 0 = 0 = x_1$$

- So $y = x^2$ reaches its minimum $y = 0$ at $x = 0$

- We are just being lucky to pick $\lambda = 0.5$

Gradient Descent Method



- The iterations as $k \rightarrow \infty$ when $x_0 = 1$ and $\lambda = 0.5$

Gradient Descent Method

- For other λ , things may be more complicated

- For instance, let $\lambda = 0.4$

- At $x_0 = 1$, $\left. \frac{\Delta y}{\Delta x} \right|_{x_0} = 2x_0 = 2$, and

$$x_1 = x_0 - \lambda \left(\left. \frac{\Delta y}{\Delta x} \right|_{x_0} \right) = 1 - 0.4 \times 2 = 0.2$$

- At $x_1 = 0.2$, $\left. \frac{\Delta y}{\Delta x} \right|_{x_1} = 2x_1 = 0.4$, and

$$x_2 = x_1 - \lambda \left(\left. \frac{\Delta y}{\Delta x} \right|_{x_1} \right) = 0.2 - 0.4 \times 0.4 = 0.04$$

Gradient Descent Method

- At $x_2 = 0.04$, $\left. \frac{\Delta y}{\Delta x} \right|_{x_2} = 2x_2 = 0.08$, and

$$x_3 = x_2 - \lambda \left(\left. \frac{\Delta y}{\Delta x} \right|_{x_2} \right) = 0.04 - 0.4 \times 0.08 = 0.008$$

- At $x_3 = 0.008$, $\left. \frac{\Delta y}{\Delta x} \right|_{x_3} = 2x_3 = 0.016$, and

$$x_4 = x_3 - \lambda \left(\left. \frac{\Delta y}{\Delta x} \right|_{x_3} \right) = 0.008 - 0.4 \times 0.016 = 0.0016$$

- At $x_4 = 0.0016$, $\left. \frac{\Delta y}{\Delta x} \right|_{x_4} = 2x_4 = 0.032$, and

$$x_5 = x_4 - \lambda \left(\left. \frac{\Delta y}{\Delta x} \right|_{x_4} \right) = 0.0016 - 0.4 \times 0.032 = 0.00064$$

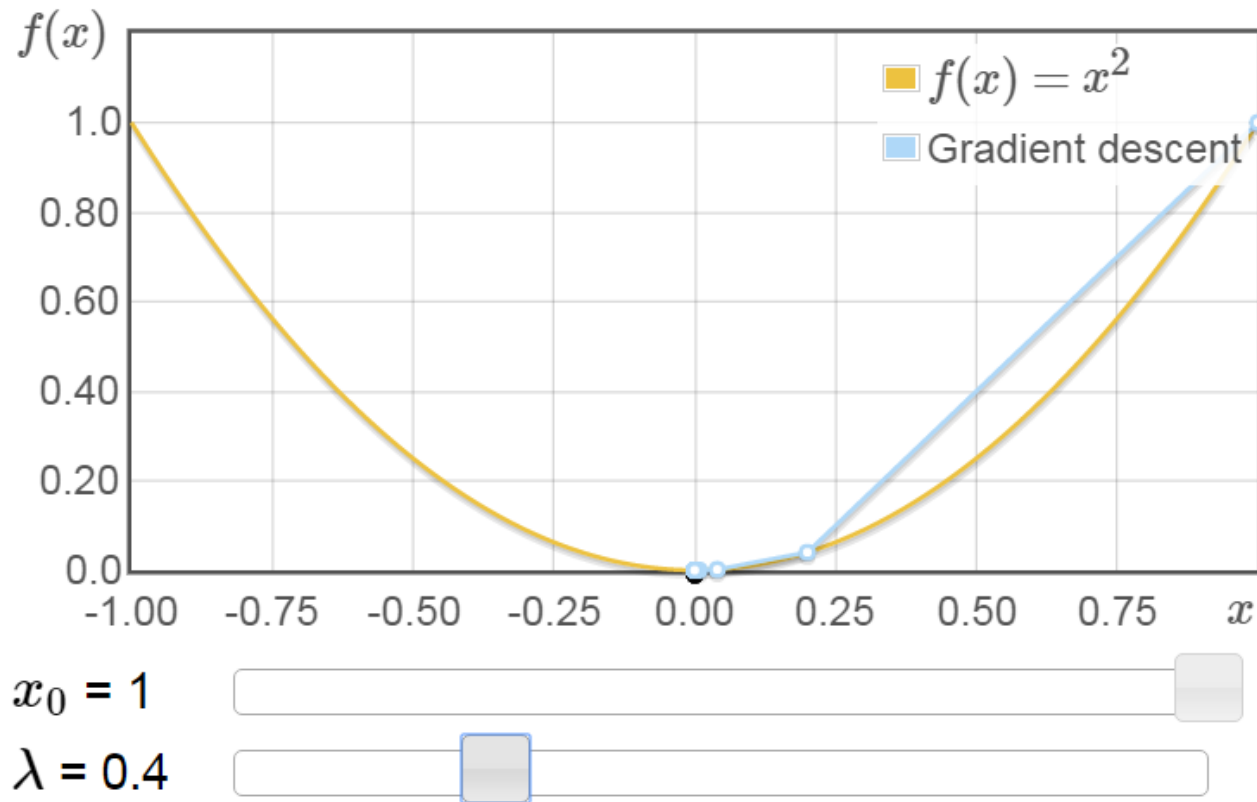
Gradient Descent Method

- Can you see the pattern?
- Indeed, it can be proven that $x_k = (0.2)^k$, $k = 0, 1, \dots$
- We already know that x^2 reaches its minimum when $x = 0$
- On the other hand, $x_k = (0.2)^k > 0$, for $k < \infty$
- It will literally never reach 0!

Gradient Descent Method

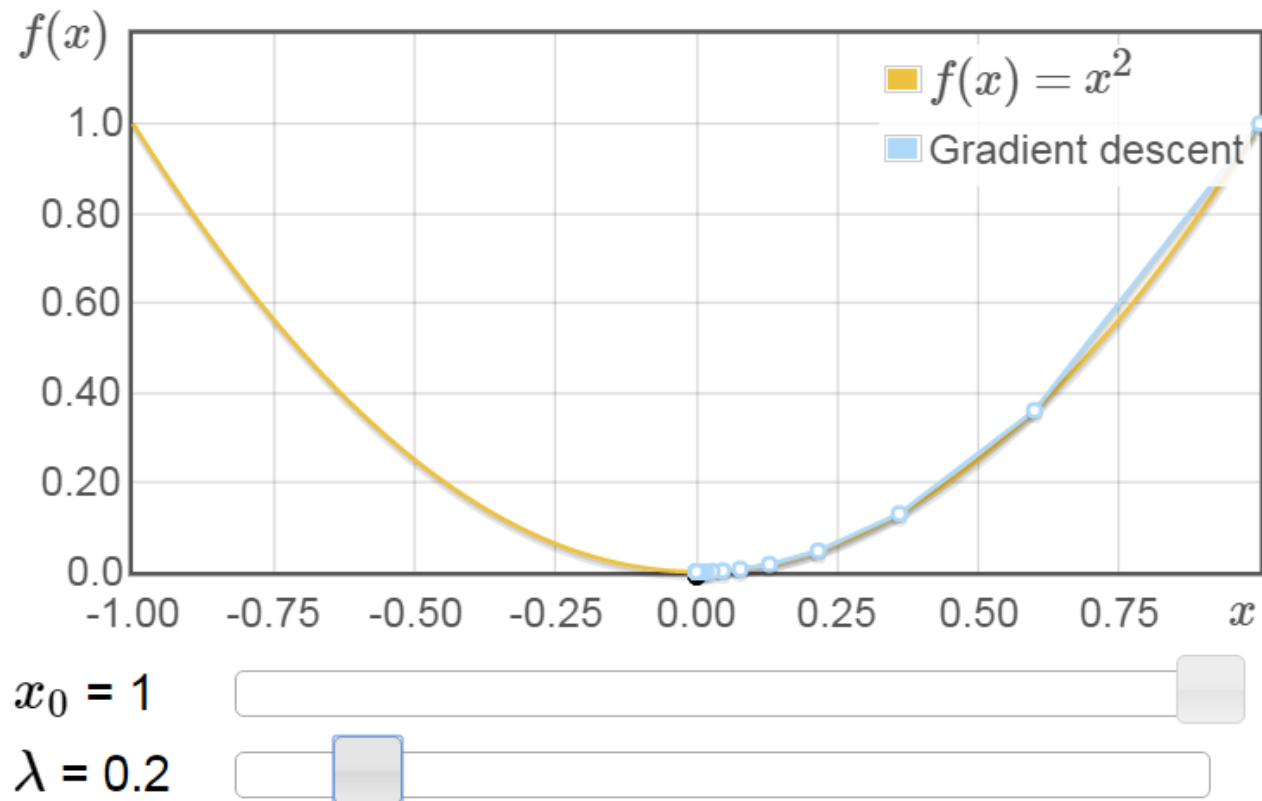
- So we need a threshold, a **tolerance**, a **stopping criteria** on the error for iterations to stop
- For instance, let the stopping criteria be $|x_k - x_{k-1}| \leq 10^{-3}$
- Then, the iterations stop at $k = 5$, since $|x_5 - x_4| = 0.00096 < 10^{-3}$, and $|x_{k+1} - x_k| > 10^{-3}$ for $k < 4$
- On the other hand, $x_5 = 0.00064 \neq 0$: this error is almost inevitable

Gradient Descent Method



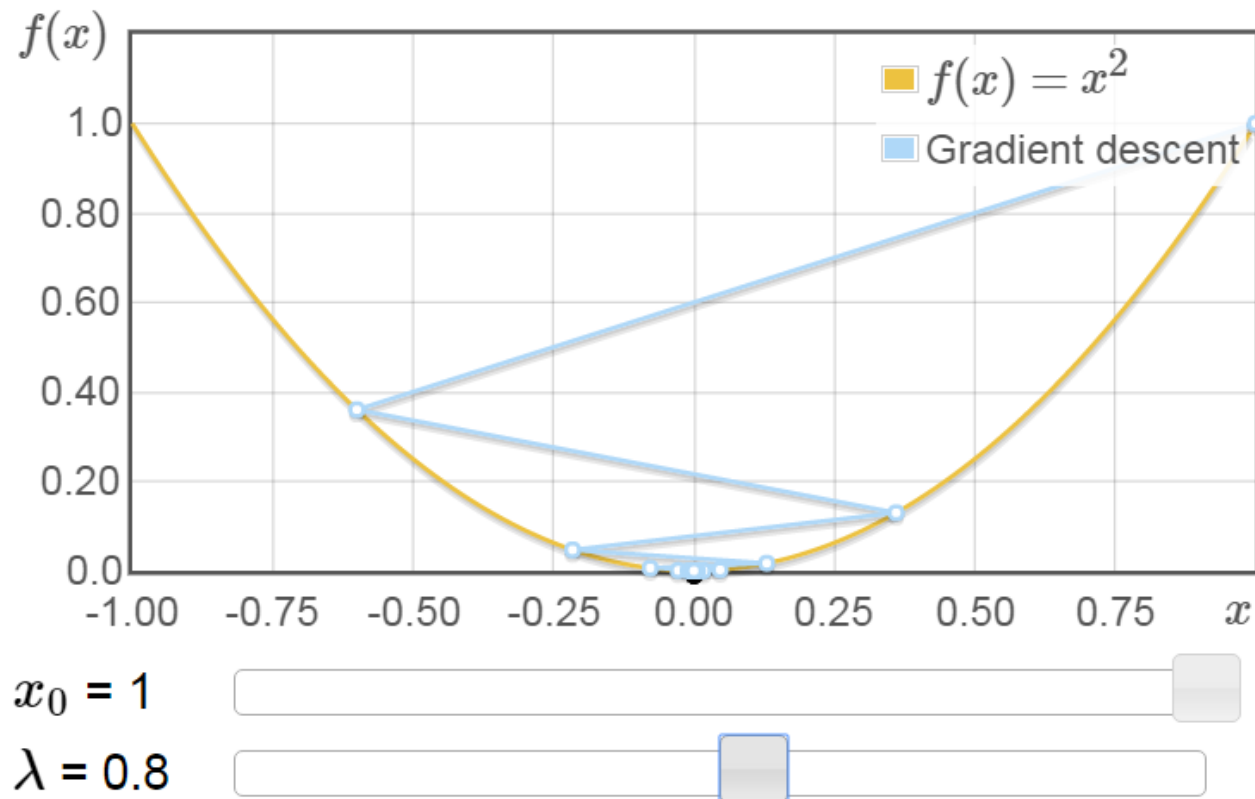
- The iterations as $k \rightarrow \infty$ when $x_0 = 1$ and $\lambda = 0.4$
- If the iterations are getting closer and closer to the minimum, the method is said to **converge**

Gradient Descent Method



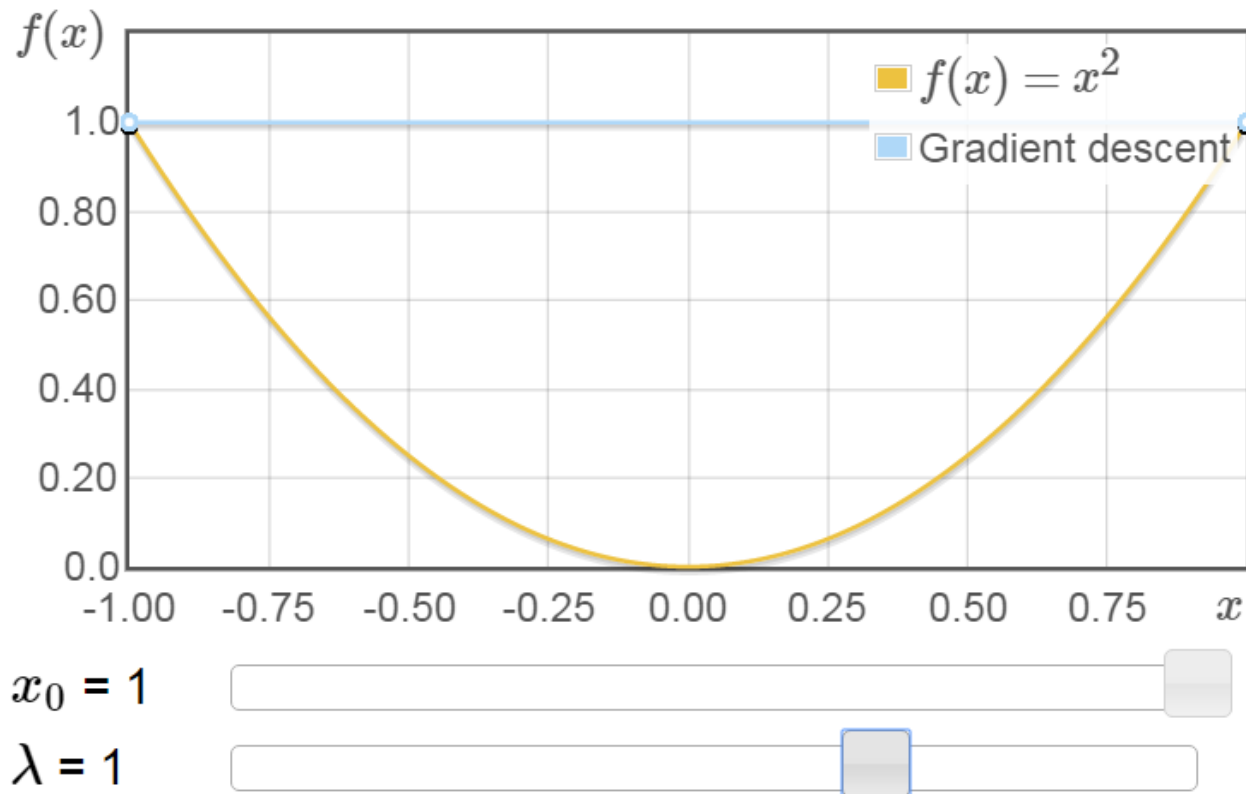
- The iterations as $k \rightarrow \infty$ when $x_0 = 1$ and $\lambda = 0.2$
- The convergence is slower when λ is too small

Gradient Descent Method



- The iterations as $k \rightarrow \infty$ when $x_0 = 1$ and $\lambda = 0.8$
- Oscillations may appear when λ is too big

Gradient Descent Method



- The iterations as $k \rightarrow \infty$ when $x_0 = 1$ and $\lambda = 1$
- May even not converge

Gradient Descent Method

- Let $\lambda = 1$

- At $x_0 = 1$, $\left. \frac{\Delta y}{\Delta x} \right|_{x_0} = 2x_0 = 2$, and

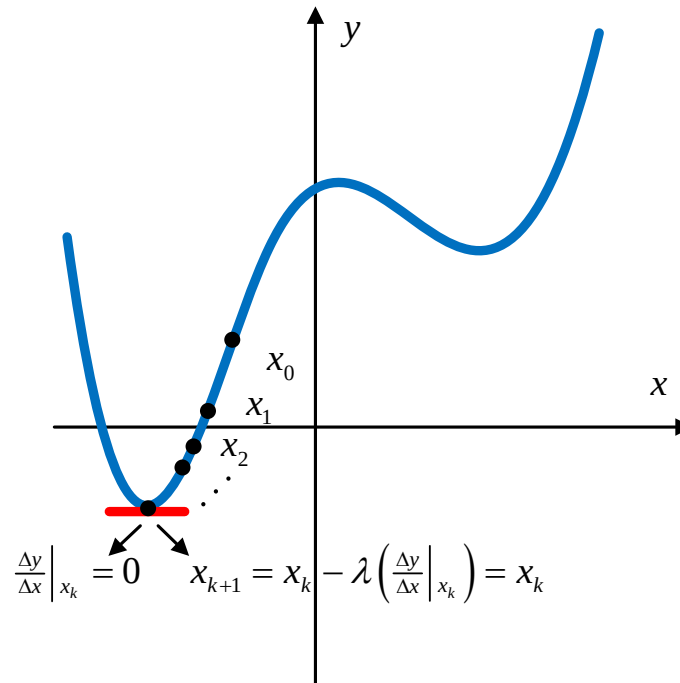
$$x_1 = x_0 - \lambda \left(\left. \frac{\Delta y}{\Delta x} \right|_{x_0} \right) = 1 - 1 \times 2 = -1$$

- At $x_1 = -1$, $\left. \frac{\Delta y}{\Delta x} \right|_{x_1} = 2x_1 = -2$, and

$$x_2 = x_1 - \lambda \left(\left. \frac{\Delta y}{\Delta x} \right|_{x_1} \right) = -1 - 1 \times (-2) = 1 = x_0$$

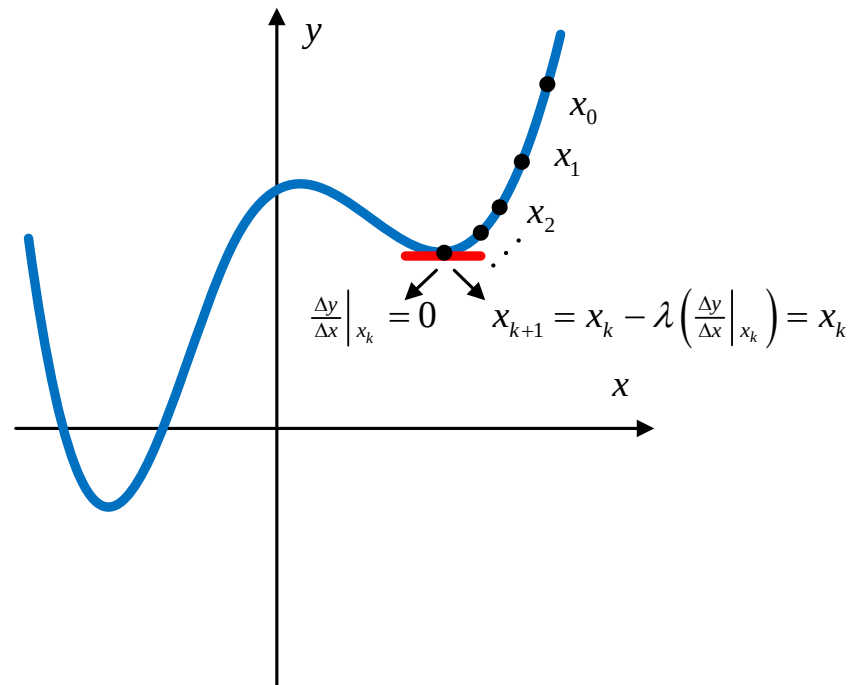
- It never converges: $x_3 = -1, x_4 = 1, x_5 = -1, x_6 = 1 \dots$

Global Minimum



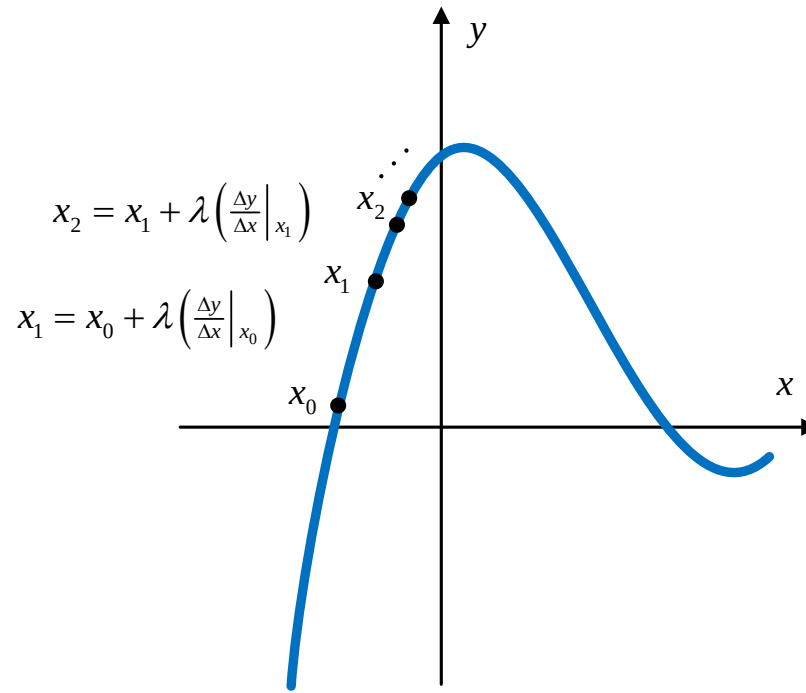
- What about when the function has more than one local minimum?
- May still find the global minimum

Local Minimum



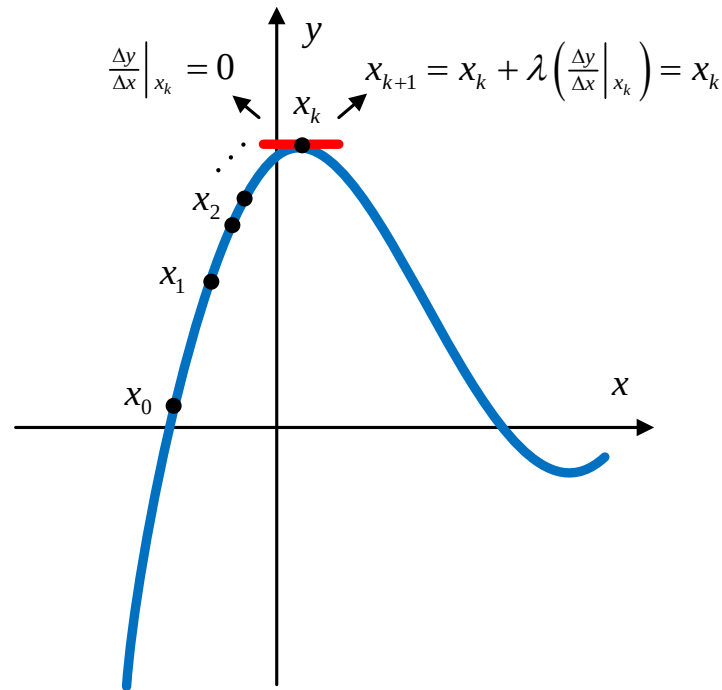
- May also get stuck in a local minimum
- Depends on the starting point x_0 and λ

Gradient Ascent Method



- Gradient ascent iteration: $x_{k+1} = x_k + \lambda \left(\frac{\Delta y}{\Delta x} \Big|_{x_k} \right)$, $k = 0, 1, \dots$
- To compare with gradient descent iteration: $x_{k+1} = x_k - \lambda \left(\frac{\Delta y}{\Delta x} \Big|_{x_k} \right)$, $k = 0, 1, \dots$
- λ is a positive constant in both cases

Local Maximum and Global Maximum



- This algorithm will converge where the gradient is zero
- Gradient is zero: local maximum
- May not be the global maximum

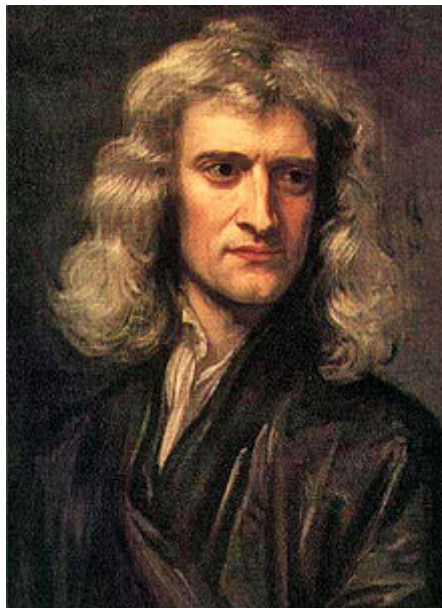
Procedures of Gradient Descent Method

- Start with an initial guess of the solution: x_0
- Take the gradient of the function x_0 : $\left. \frac{\Delta y}{\Delta x} \right|_{x_0}$
- Step the solution in the opposite direction of the gradient: to $x_0 - \lambda \left(\left. \frac{\Delta y}{\Delta x} \right|_{x_0} \right)$, λ is a positive constant
- Denote $x_1 = x_0 - \lambda \left(\left. \frac{\Delta y}{\Delta x} \right|_{x_0} \right)$, and repeat the process for x_1 , then $x_2 = x_1 - \lambda \left(\left. \frac{\Delta y}{\Delta x} \right|_{x_1} \right)$ can be obtained
- By similar iterations, $x_3, x_4, \dots$ can be obtained
- The algorithm will eventually converge where the gradient is zero

How to Choose λ

- In gradient descent method, $x_{k+1} = x_k - \lambda \left(\frac{\Delta y}{\Delta x} \Big|_{x_k} \right)$, λ is a given positive constant
- In the examples, it can be seen that different λ may lead to different convergence processes, and some even do not converge
- Are there better ways of choosing λ than trial-and-error?

Newton and Newton's Method



Isaac Newton

- In calculus, Newton's method is an iterative method for finding the roots of a function $y = f(x)$, i.e. solutions to the equation $y = f(x) = 0$
- In optimization, Newton's method is applied to the derivative of $y = f(x)$ to find the roots of the derivative, i.e., solutions to $\frac{\Delta y}{\Delta x} = 0$
- In his book: Methodus fluxionum et serierum infinitarum, 1664-1671

Newton's Method

- Intuition
- Iterations
- Examples
- Comparison with gradient descent method

Newton's Method for Optimization

- Taylor expansion (second order) of $y = f(x)$ (you will learn this in calculus):

$$f(a + b) \approx f(a) + b \left(\frac{\Delta y}{\Delta x} \Big|_a \right) + \frac{b^2}{2} \left(\frac{\Delta z}{\Delta x} \Big|_a \right)$$

where $z = \frac{\Delta y}{\Delta x}$

- Thus,

$$\begin{aligned} f(x_{k+1}) &= f(x_k + (x_{k+1} - x_k)) \\ &\approx f(x_k) + (x_{k+1} - x_k) \left(\frac{\Delta y}{\Delta x} \Big|_{x_k} \right) + \frac{(x_{k+1} - x_k)^2}{2} \left(\frac{\Delta z}{\Delta x} \Big|_{x_k} \right) \end{aligned}$$

where $a = x_k$, $b = x_{k+1} - x_k$, and $z = \frac{\Delta y}{\Delta x}$

Finding the Optimal x_{k+1}

■ What is the optimal x_{k+1} ?

■ Denote

$$f(x_{k+1}) \approx f(x_k) + (x_{k+1} - x_k) \left(\frac{\Delta y}{\Delta x} \Big|_{x_k} \right) + \frac{(x_{k+1} - x_k)^2}{2} \left(\frac{\Delta z}{\Delta x} \Big|_{x_k} \right)$$

as

$$f(x_{k+1}) \approx f(x_k) + (x_{k+1} - x_k) c + \frac{(x_{k+1} - x_k)^2}{2} d$$

where $c = \frac{\Delta y}{\Delta x} \Big|_{x_k}$, $d = \frac{\Delta z}{\Delta x} \Big|_{x_k}$

■ Then,

$$f(x_{k+1}) \approx \frac{d}{2} x_{k+1}^2 + (c - dx_k) x_{k+1} + \frac{d}{2} x_k^2 - cx_k$$

Finding the Optimal x_{k+1}

- It is known from high school mathematics that $px^2 + qx + r$ reaches its minimum when $x = -\frac{q}{2p}$

- Thus,

$$f(x_{k+1}) \approx \frac{d}{2}x_{k+1}^2 + (c - dx_k)x_{k+1} + \frac{d}{2}x_k^2 - cx_k$$

reaches its minimum when $x_{k+1} = -\frac{(c-dx_k)}{d} = x_k - \frac{c}{d}$

- On the other hand,

$$x_{k+1} = x_k + \frac{c}{d} = x_k - \frac{\left.\frac{\Delta y}{\Delta x}\right|_{x_k}}{d} = x_k - \frac{1}{d} \left(\left.\frac{\Delta y}{\Delta x}\right|_{x_k} \right)$$

How to Choose λ

- Thus, in the iteration formula $x_{k+1} = x_k - \lambda \left(\frac{\Delta y}{\Delta x} \Big|_{x_k} \right),$

$$\lambda = \frac{1}{d} = \frac{1}{\frac{\Delta z}{\Delta x} \Big|_{x_k}}$$

where $z = \frac{\Delta y}{\Delta x}$

Example 1

Minimize $y = x^2$

- Gradient: $\frac{\Delta y}{\Delta x} = 2x$
- Thus, $z = \frac{\Delta y}{\Delta x} = 2x$, and $\frac{\Delta z}{\Delta x} = 2$
- So $\lambda = \frac{1}{\frac{\Delta z}{\Delta x}|_{x_k}} = 0.5$, and $x_{k+1} = x_k - \lambda \left(\frac{\Delta y}{\Delta x} \Big|_{x_k} \right) = x_k - 0.5 \left(\frac{\Delta y}{\Delta x} \Big|_{x_k} \right)$

Taking $x_0 = 1$ as Starting Point

- At $x_0 = 1$, $\left. \frac{\Delta y}{\Delta x} \right|_{x_0} = 2x_0 = 2$, and

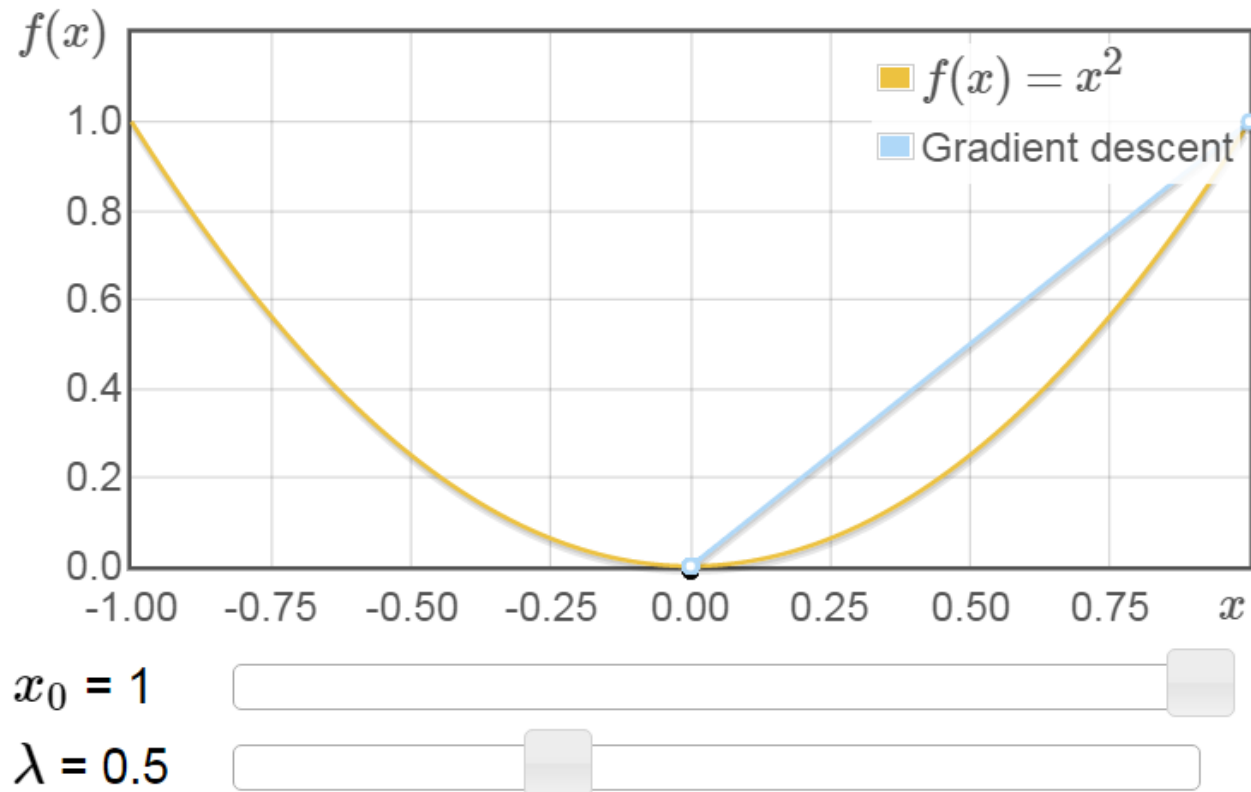
$$x_1 = x_0 - 0.5 \left(\left. \frac{\Delta y}{\Delta x} \right|_{x_0} \right) = 1 - 0.5 \times 2 = 0$$

- At $x_1 = 0$, $\left. \frac{\Delta y}{\Delta x} \right|_{x_1} = 2x_1 = 0$, and

$$x_2 = x_1 - 0.5 \left(\left. \frac{\Delta y}{\Delta x} \right|_{x_1} \right) = 0 - 0.5 \times 0 = 0 = x_1$$

- So $y = x^2$ reaches its minimum $y = 0$ at $x = 0$

Taking $x_0 = 1$ as Starting Point



- The iterations as $k \rightarrow \infty$ when $x_0 = 1$ and $\lambda = 0.5$

What about Other Starting Points?

- At $x_0 = 0.6$, $\left. \frac{\Delta y}{\Delta x} \right|_{x_0} = 2x_0 = 1.2$, and

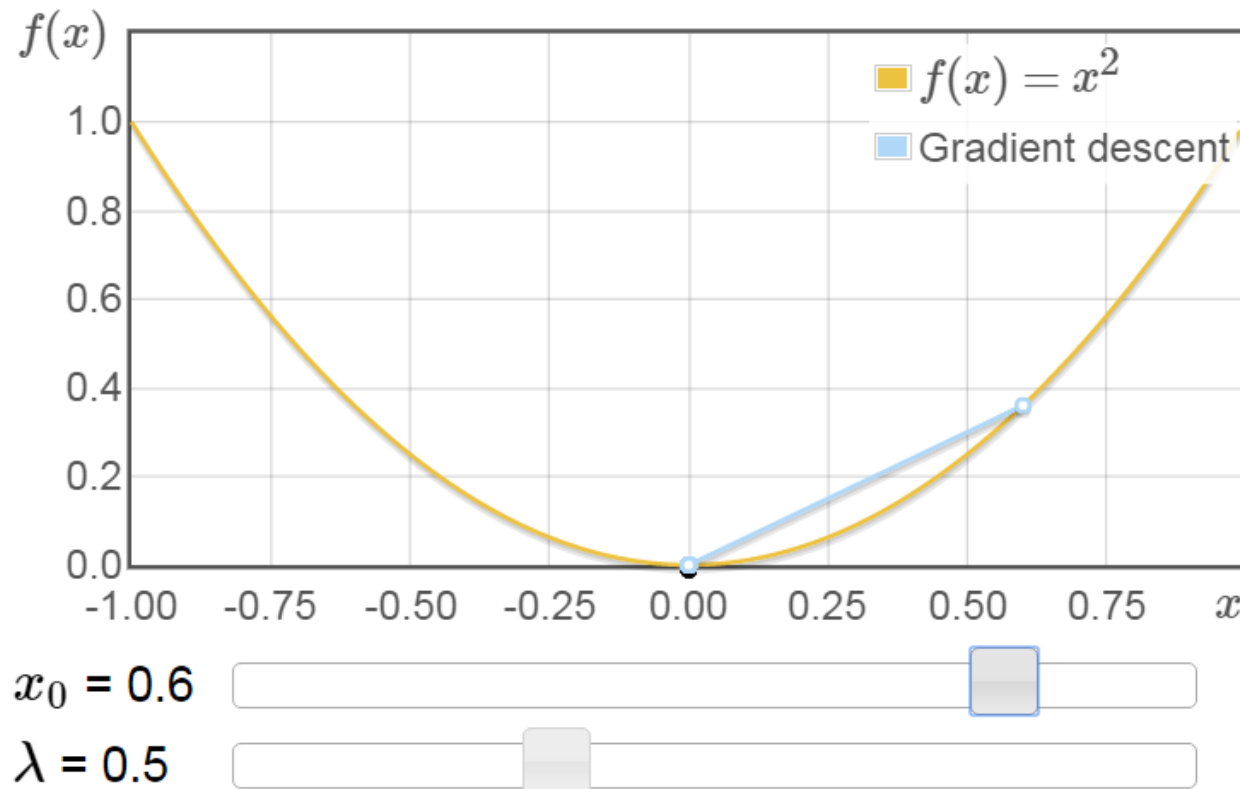
$$x_1 = x_0 - 0.5 \left(\left. \frac{\Delta y}{\Delta x} \right|_{x_0} \right) = 1 - 0.5 \times 1.2 = 0$$

- At $x_1 = 0$, $\left. \frac{\Delta y}{\Delta x} \right|_{x_1} = 2x_1 = 0$, and

$$x_2 = x_1 - 0.5 \left(\left. \frac{\Delta y}{\Delta x} \right|_{x_1} \right) = 0 - 0.5 \times 0 = 0 = x_1$$

- So $y = x^2$ reaches its minimum $y = 0$ at $x = 0$

Taking $x_0 = 0.6$ as starting point



- The iterations as $k \rightarrow \infty$ when $x_0 = 0.6$ and $\lambda = 0.5$

A General Conclusion

- For x^2 , Newton's method finds the minimum within one step for any starting point x_0
- More generally, for any $px^2 + qx + r$, Newton's method finds the minimum within one step for any starting point x_0
- The proof uses the fact that second order Taylor expansion of $y = f(x)$ is exact for $px^2 + qx + r$:

$$f(a + b) = f(a) + b \left(\frac{\Delta y}{\Delta x} \Big|_a \right) + \frac{b^2}{2} \left(\frac{\Delta^2 y}{\Delta x^2} \Big|_a \right)$$

- But it is not true for higher-order functions, such as x^4

Example 2

Minimize $y = x^4$

- Gradient: $\frac{\Delta y}{\Delta x} = 4x^3$
- Thus, $z = \frac{\Delta y}{\Delta x} = 4x^3$, and $\frac{\Delta z}{\Delta x} = 12x^2$
- So $\lambda = \frac{1}{\frac{\Delta z}{\Delta x}|_{x_k}} = \frac{1}{12x_k^2}$, and $\frac{1}{\frac{\Delta y}{\Delta x}|_{x_k}} = 4x_k^3$
- $x_{k+1} = x_k - \lambda \left(\frac{\Delta y}{\Delta x} \Big|_{x_k} \right) = x_k - \frac{1}{12x_k^2} \times 4x_k^3 = x_k - \frac{1}{3}x_k = \frac{2}{3}x_k$
- If $x_0 = 1$ is starting point: $x_0 = 1, x_1 = \frac{2}{3}, x_2 = \left(\frac{2}{3}\right)^2, \dots, x_k = \left(\frac{2}{3}\right)^k$, and as $k \rightarrow \infty, x_k \rightarrow 0$
- It is well known that the minimum is at $x = 0$

Comparison with Gradient Descent with Constant λ

Minimize $y = x^4$

- Gradient: $\frac{\Delta y}{\Delta x} = 4x^3$
- Thus, $z = \frac{\Delta y}{\Delta x} = 4x^3$, and $\frac{\Delta y}{\Delta x} \Big|_{x_k} = 4x_k^3$
- $x_{k+1} = x_k - \lambda \left(\frac{\Delta y}{\Delta x} \Big|_{x_k} \right) = x_k - 4\lambda x_k^3$
- Let $x_0 = 1$ be the starting point

Comparison with Gradient Descent with Constant λ

- Let $\lambda = 0.1$, then $x_{k+1} = x_k - 0.4x_k^3$
- $x_1 = x_0 - 0.4x_0^3 = 0.6$ ($x_1 =$ in Newton's method)
- $x_2 = x_1 - 0.4x_1^3 = 0.5136$ ($x_2 = 0.6667$ in Newton's method)
- $x_3 = x_2 - 0.4x_2^3 = 0.4594$ ($x_3 = 0.4444$ in Newton's method)
- $x_4 = x_3 - 0.4x_3^3 = 0.4206$ ($x_4 = 0.2963$ in Newton's method)
- $x_5 = x_4 - 0.4x_4^3 = 0.3909$ ($x_5 = 0.1975$ in Newton's method)
- $x_6 = x_5 - 0.4x_5^3 = 0.3670$ ($x_6 = 0.1317$ in Newton's method)

Not to Mention Divergence of Gradient Descent

- Let $\lambda = 1$, then $x_{k+1} = x_k - 4x_k^3$

- $x_1 = x_0 - 4x_0^3 = -3$

- $x_2 = x_1 - 4x_1^3 = 105$

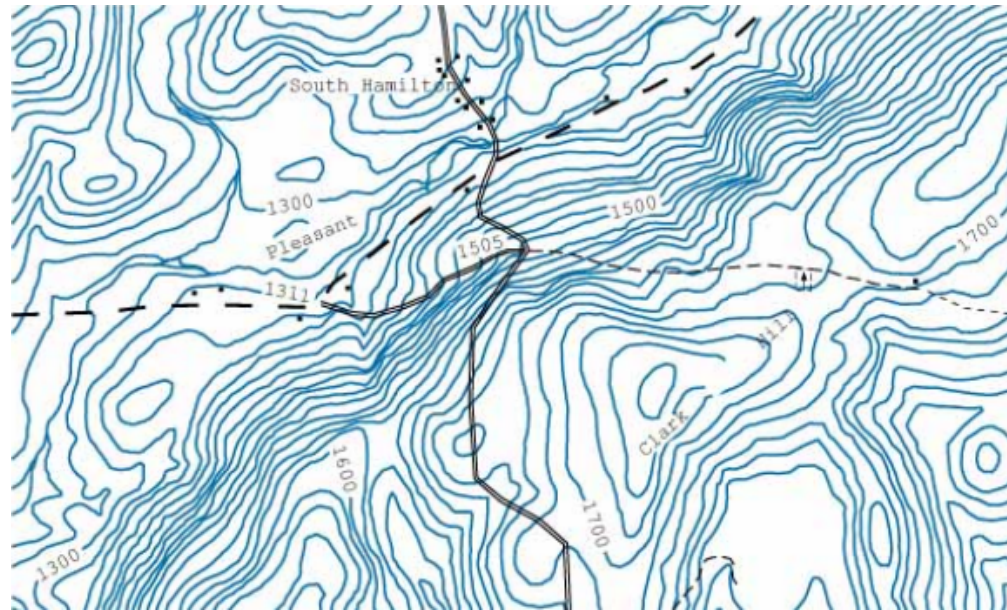
- $x_3 = x_2 - 4x_2^3 = -4630395$

- ...

Summary

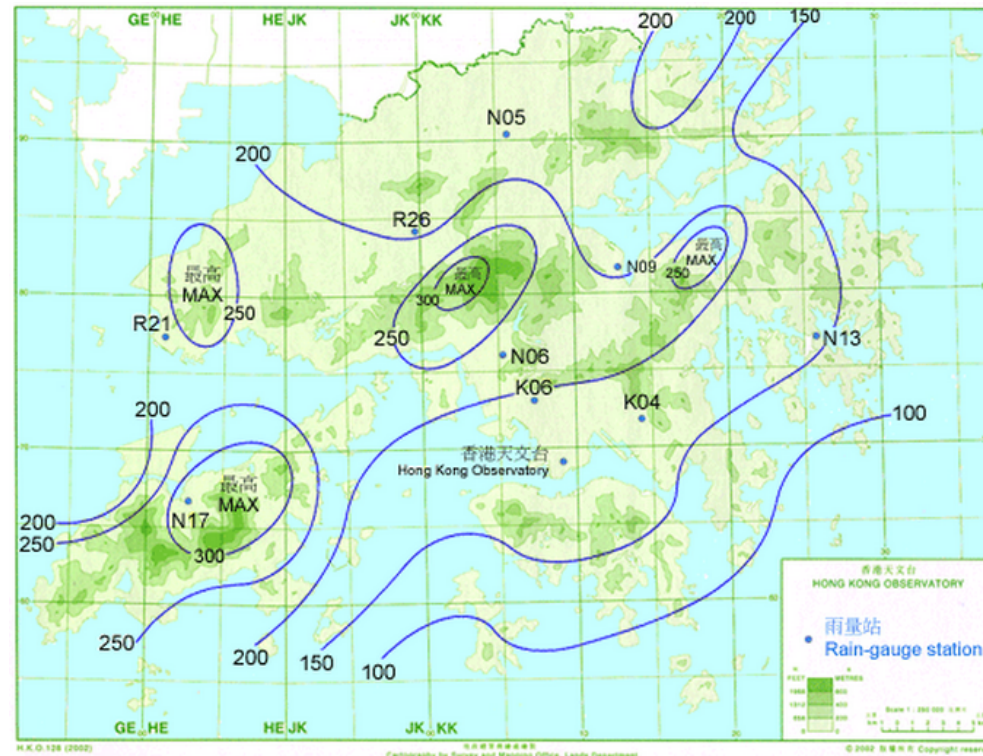
- Gradient descent with constant λ : easy to compute, but λ is difficult to choose
- Newton's method: λ is chosen in a systematic way, but needs more computation

Beyond



- Contour plot
- Points that are connected are at the same height
- Region around South Hamilton, New York

Contour Plot of Hong Kong

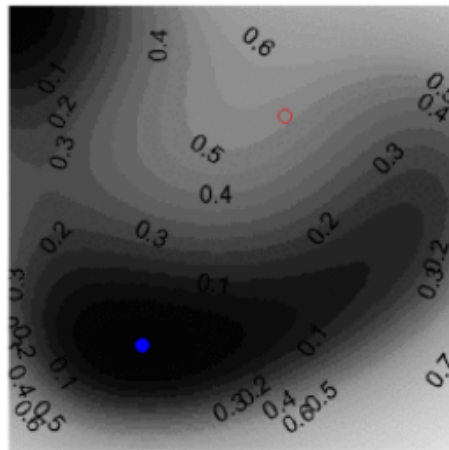


- Where are the local maximum and global maximum?
- Where are the local minimum and global minimum?

Illustrations of Optimization on Contour Plots

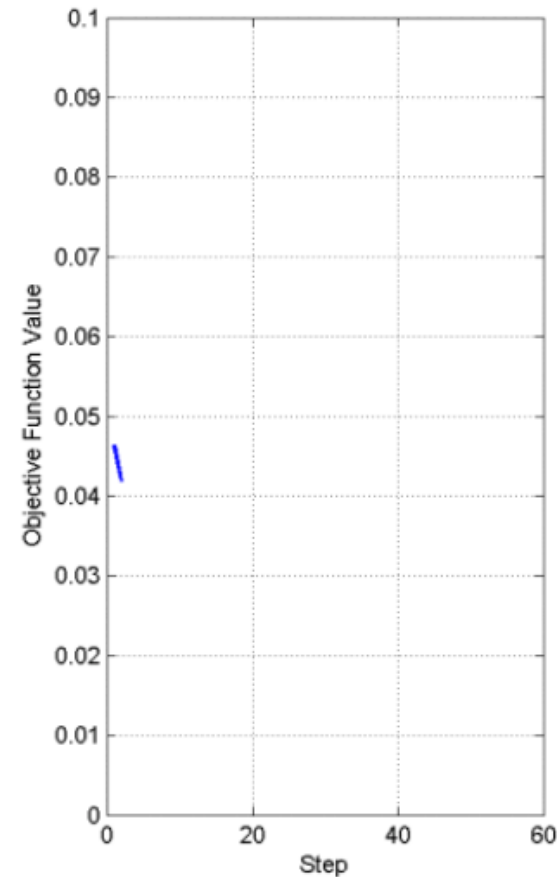
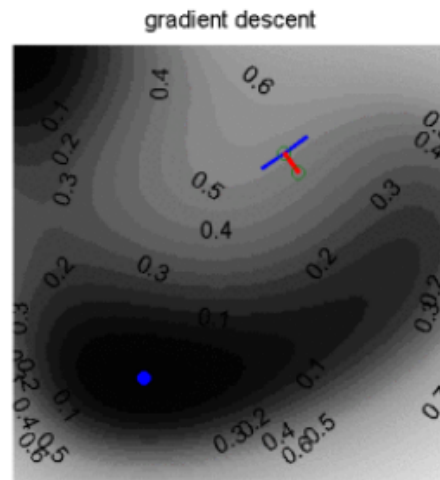
- Finding the minimum
- With the same starting point
- Gradient descent vs. Newton's method
- Question: is the minimum found local or global?

Gradient Descent



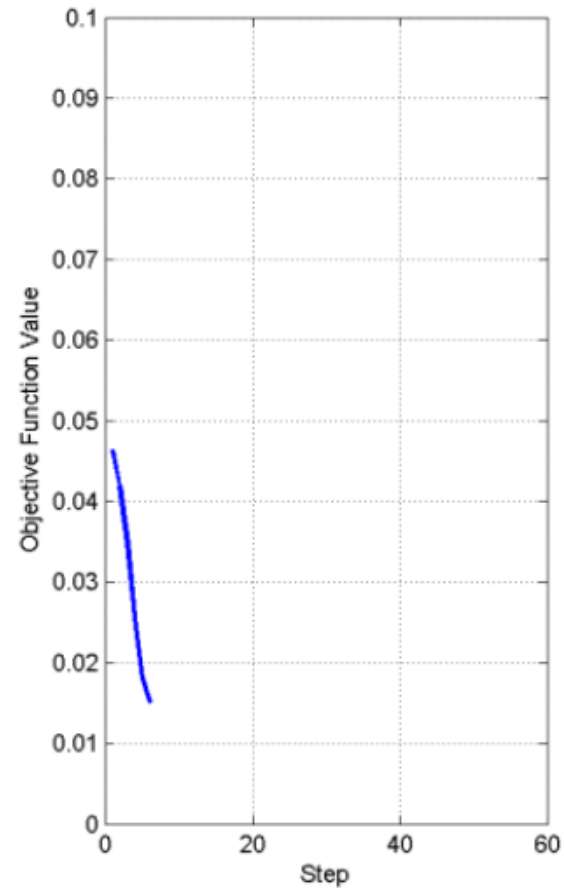
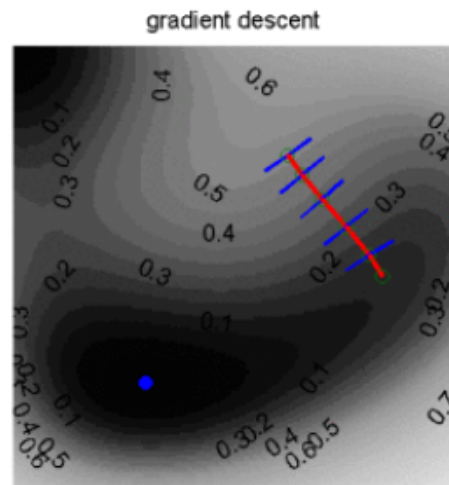
- Red circle: starting point
- Blue dot: minimum

Gradient Descent



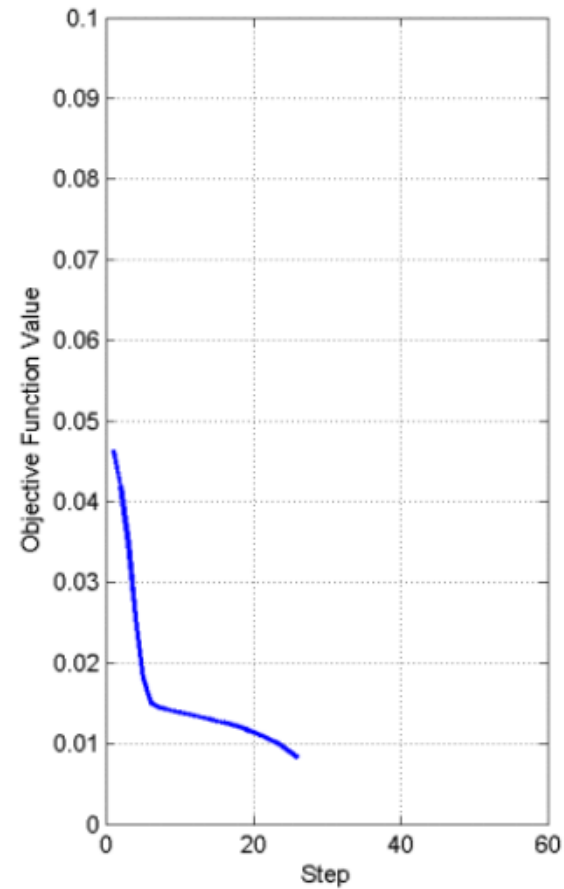
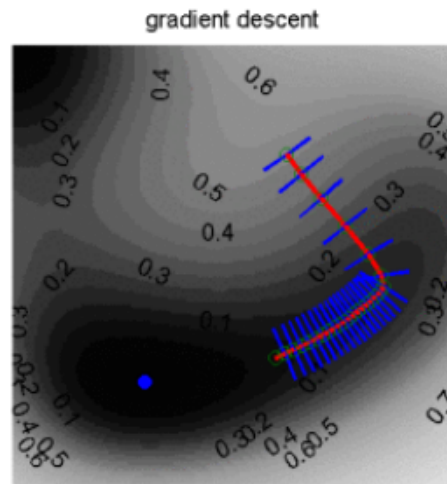
- Blue line (left figure): tangent line of the contour
- Red line: the path to go for this iteration

Gradient Descent



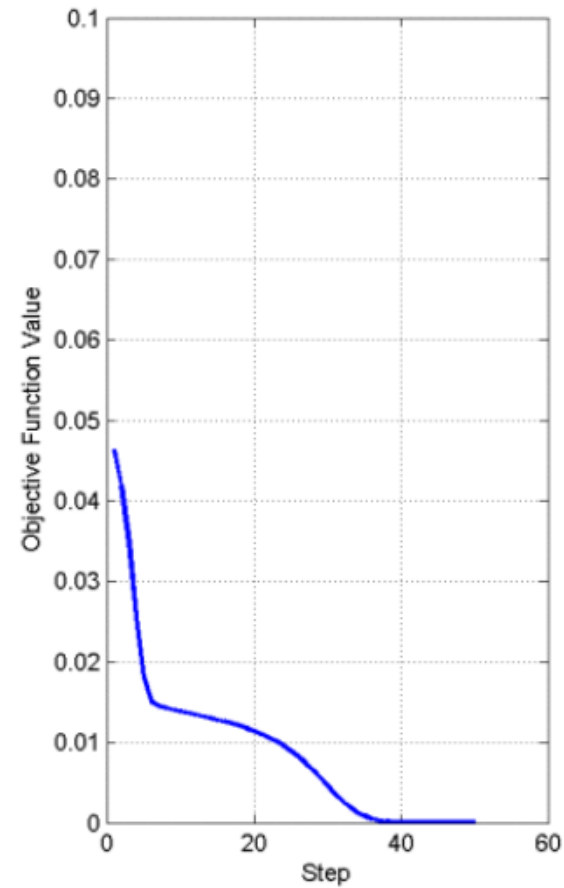
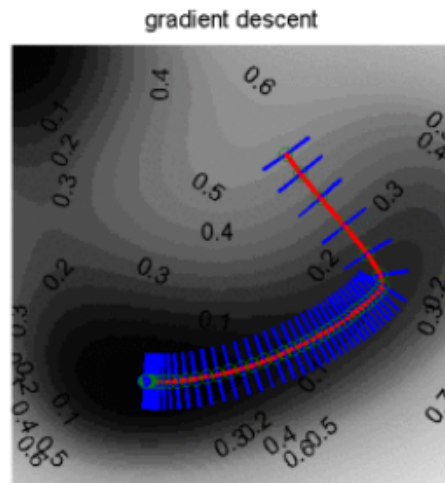
- One iteration after another
- Right figure: objective function value

Gradient Descent



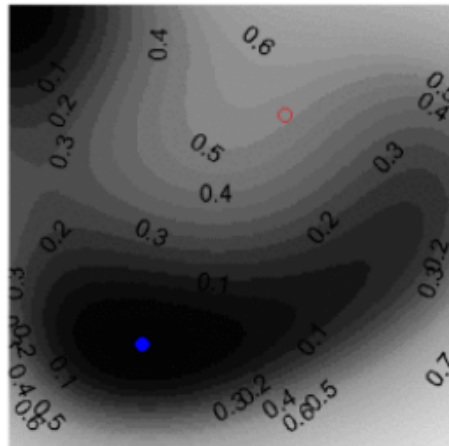
- Descent towards the minimum

Gradient Descent



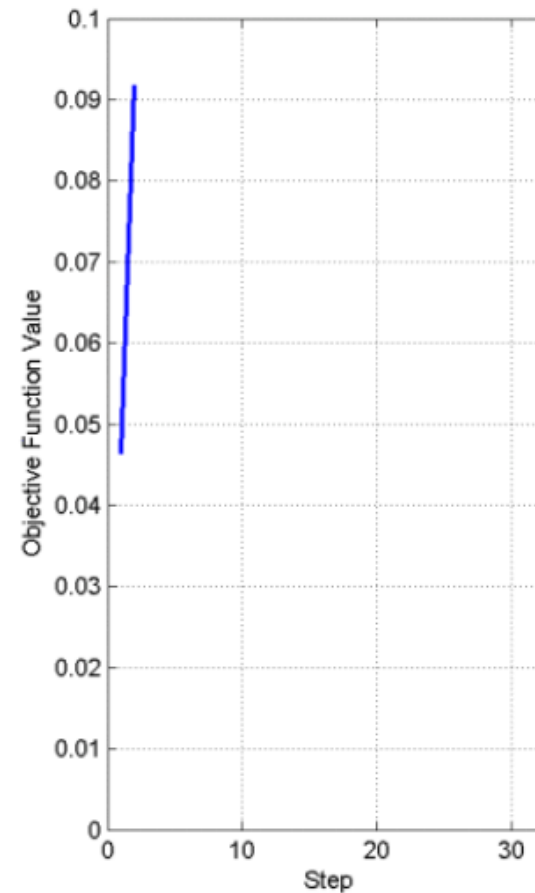
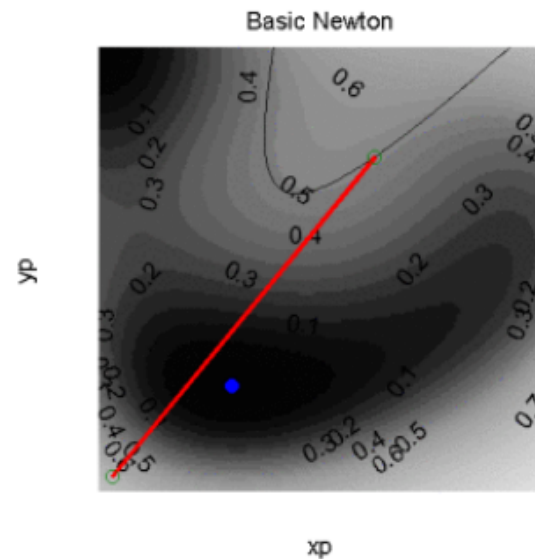
- Eventually, converge to the minimum

Newton's Method



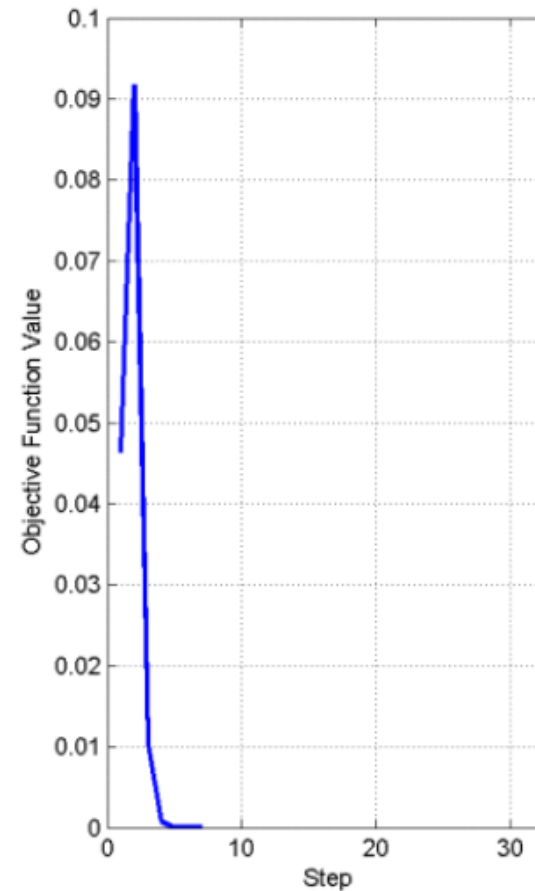
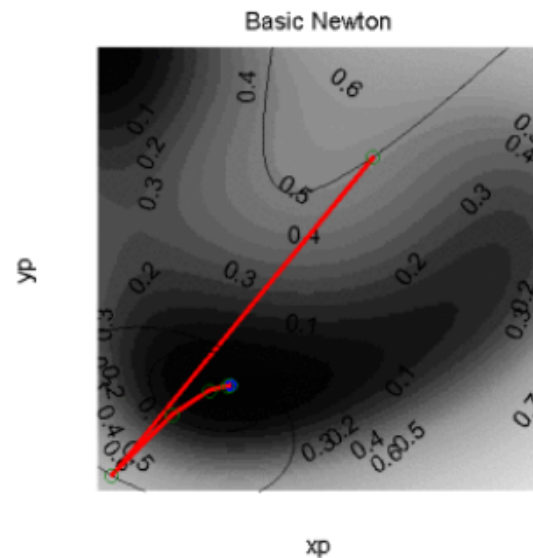
- Same starting point

Newton's Method



- One iteration
- Black line: contour plot approximation (second-order) at the starting point

Newton's Method

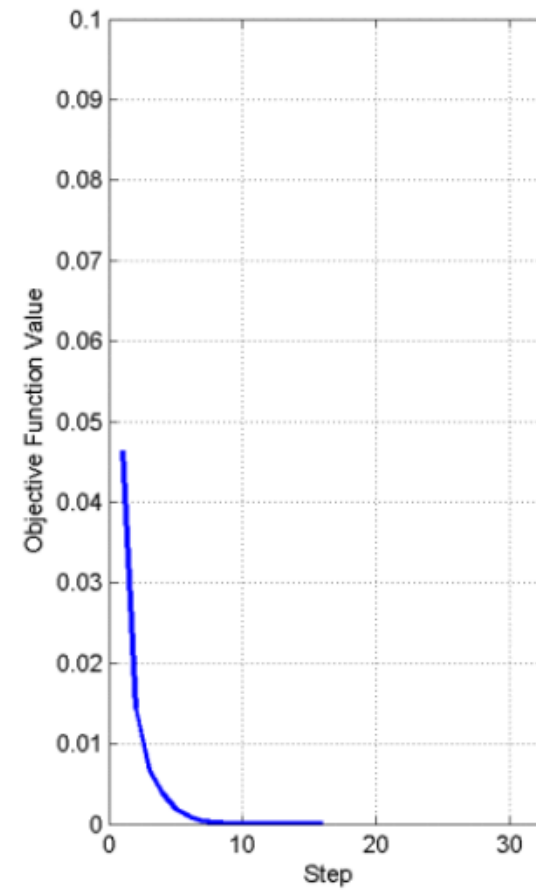
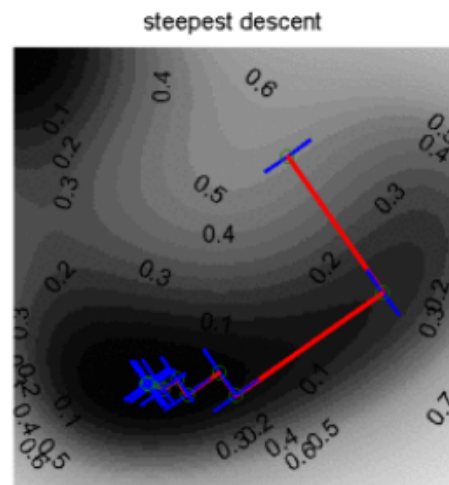


- Eventually, converge to the minimum
- Faster than gradient descent in this case

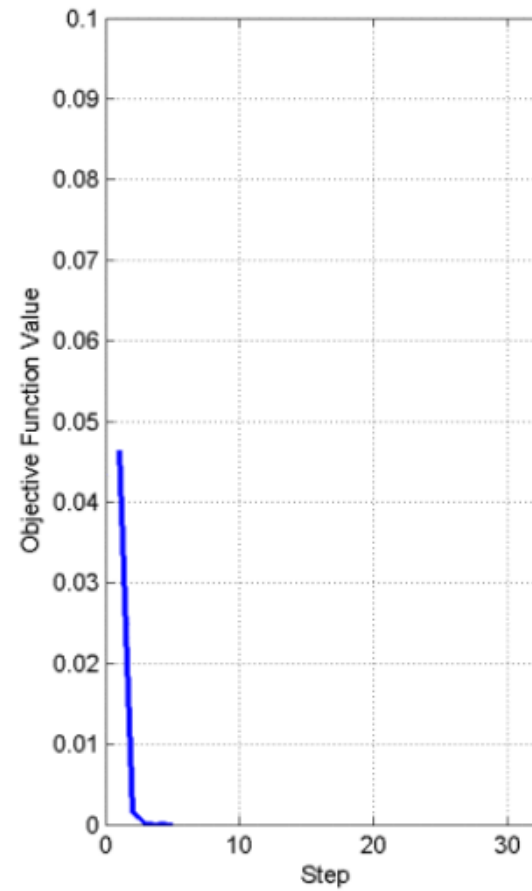
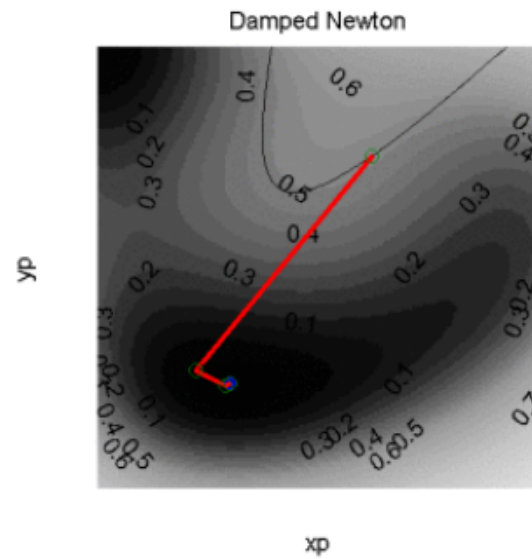
And More

- Check out more at <http://www.52ml.net/12647.html> with animations
- Steepest descent
- Damped Newton's method
- Genetic algorithm
- Levenberg-Marquardt method
- Gauss-Newton least squares
- Polak-Ribiere-Polyak (PRP) conjugate gradient method
- Only to give you a general impression; Google it for technical details

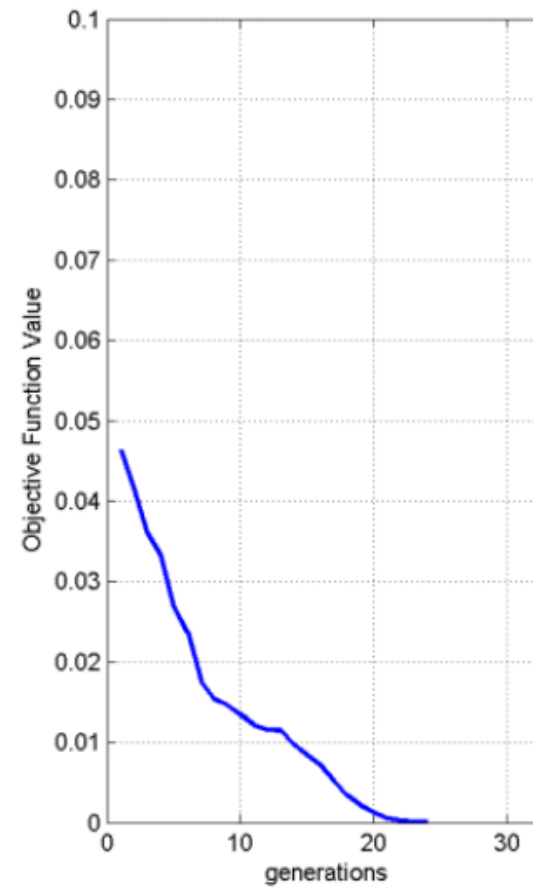
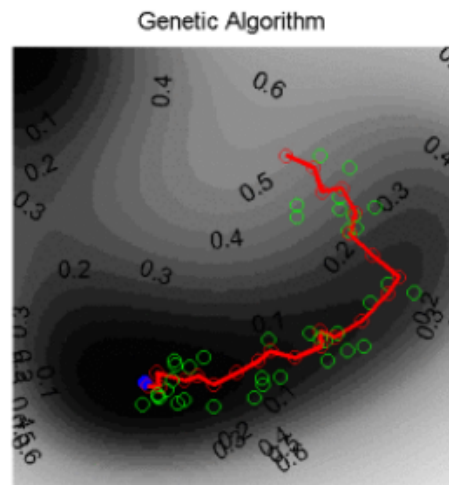
Steepest Descent



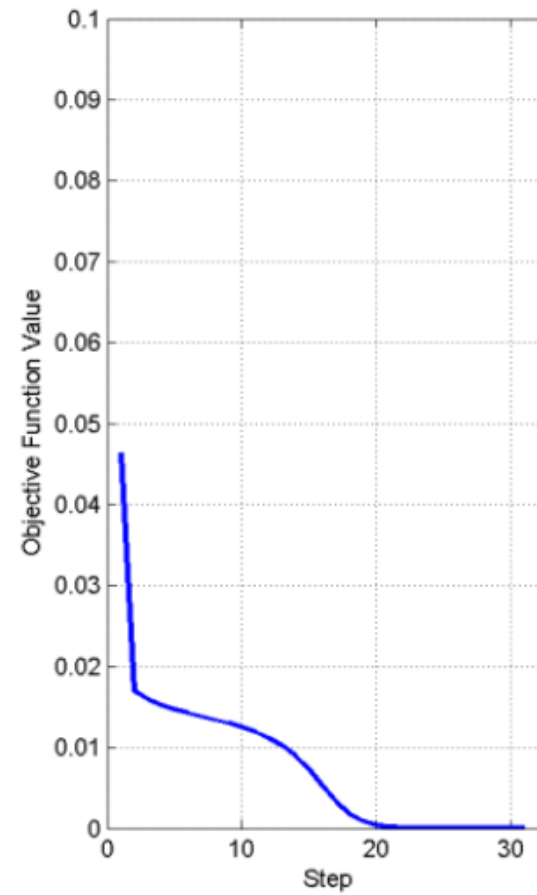
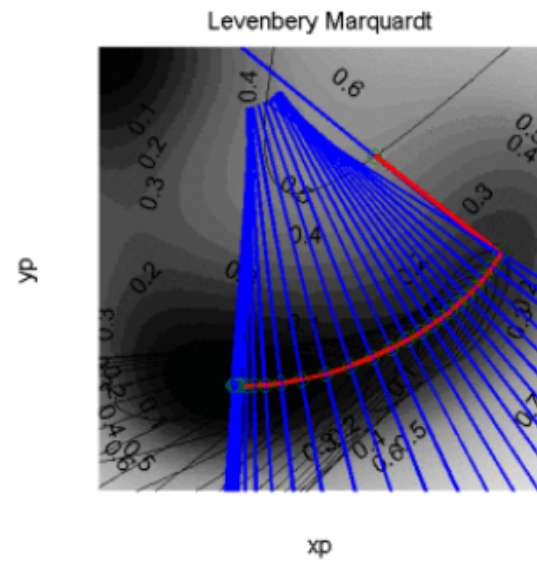
Damped Newton's Method



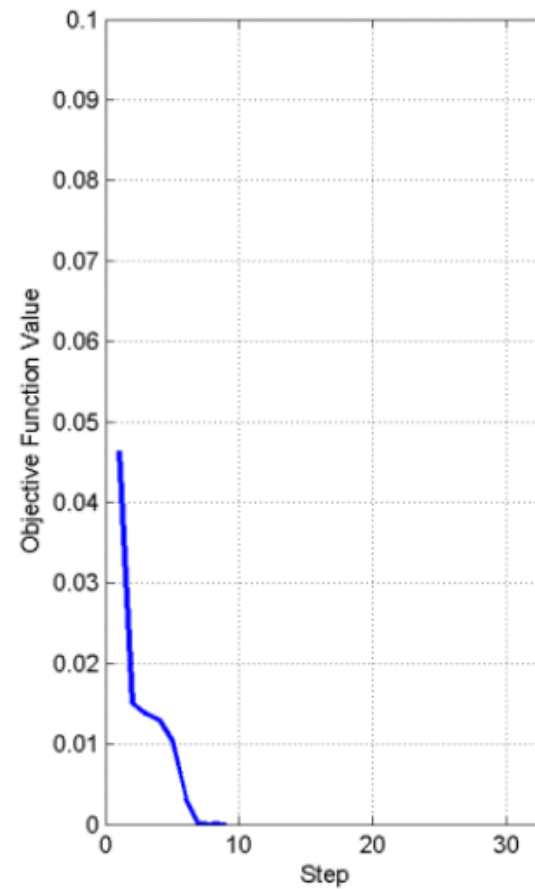
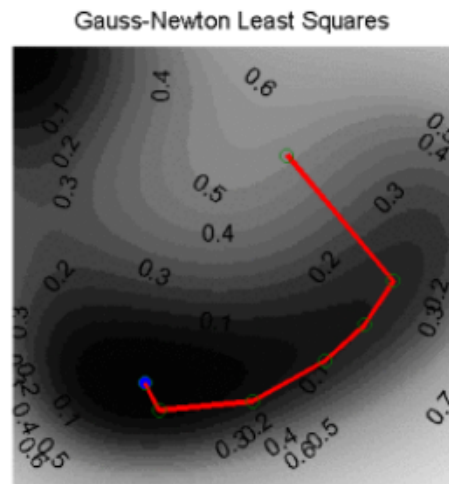
Genetic Algorithm



Levenberg-Marquardt Method



Gauss-Newton Least Squares



Polak-Ribiere-Polyak (PRP) Conjugate Gradient Method

