Lecture 22: Algorithms for Unconstrained Optimization

Zizhuo Wang

Institute of Data and Decision Analytics (iDDA)
Chinese University of Hong Kong, Shenzhen

Nov 23, 2018

Announcement

- Homework 8 due next Wednesday (11/28)

Recap

We have introduced several methods for unconstrained optimization problems

For one-dimensional problems

- Bisection method: search for $f'(x) = 0$
- Golden section method: does not need $f'(x)$

For high-dimensional problems

- General framework: Choose search direction and stepsize in each iteration
- Gradient descent method: Choose direction as the negative of the gradient
- Stepsize: We may do an exact line search, which could use the golden section method. However, exact line search may not be very efficient in practice, an approximate search method may work better
- The most commonly used method is *backtracking line search*

Backtracking Line Search

Assume we have found $\mathbf{d}^k$ and we want to choose step size α_k.

1. We first choose a small $\alpha \in (0, 0.5)$. Also choose a constant $0 < \beta < 1$
2. Let $t = 1$
3. If $f(\mathbf{x}^k + t\mathbf{d}^k) \leq f(\mathbf{x}^k) + \alpha t \nabla f(\mathbf{x}^k)^T \mathbf{d}^k$, then choose $\alpha_k = t$. Otherwise, set $t = \beta t$ and repeat this step.

Why this works?

- We know by Taylor expansion, if t is sufficiently small, we must have

$$f(\mathbf{x}^k + t\mathbf{d}^k) \approx f(\mathbf{x}^k) + t\nabla f(\mathbf{x}^k)^T \mathbf{d}^k < f(\mathbf{x}^k) + \alpha t \nabla f(\mathbf{x}^k)^T \mathbf{d}^k$$

  Therefore, as long as t is small enough, the condition in Step 3 must be satisfied (remember $\nabla f(\mathbf{x}^k)^T \mathbf{d}^k = -||\nabla f(\mathbf{x}^k)||^2 < 0$)

Stopping Criterion for Gradient Descent Method

Remember that for local optimality, we need

$$\nabla f(\mathbf{x}) = 0$$

Since we don't know the optimal value, we use the gradient as the stopping criterion:

- We stop when $||\nabla f(\mathbf{x})|| < \epsilon$ for a pre-chosen ϵ

Theorem

Suppose $f(\mathbf{x})$ is convex and the smallest eigenvalue of $\nabla^2 f(\mathbf{x})$ is m. Then

$$||\mathbf{x} - \mathbf{x}^*|| \leq \frac{2}{m}||\nabla f(\mathbf{x})||$$

where $\mathbf{x}^*$ is the global minimum of $f(\mathbf{x})$

- Therefore, when $||\nabla f(\mathbf{x})||$ is small enough, the solution is guaranteed to be close to the optimal solution

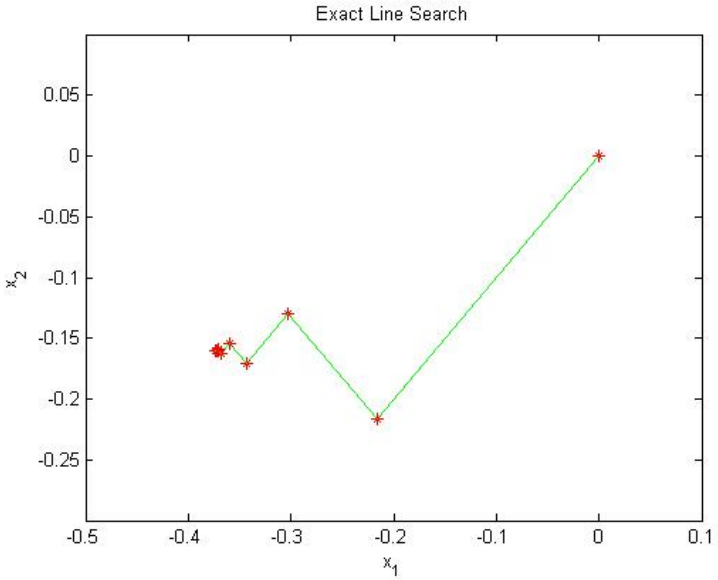
Gradient Descent Algorithm

Start with any point $\mathbf{x}^0$. Set $k = 0$ and stopping criterion $\epsilon > 0$

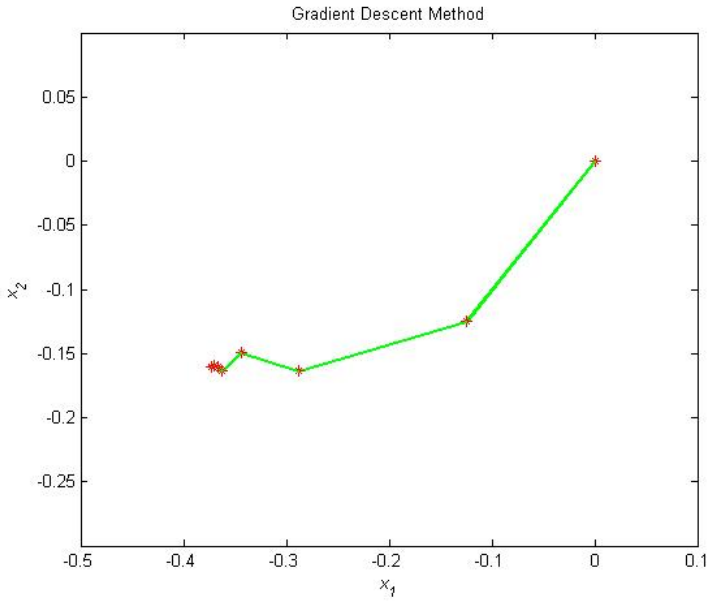
1. Check $||\nabla f(\mathbf{x}^k)||$. If $||\nabla f(\mathbf{x}^k)|| \leq \epsilon$, stop and output $\mathbf{x}^k$. Otherwise, continue to Step 2
2. Let $\mathbf{d}^k = -\nabla f(\mathbf{x}^k)$
3. Use either exact line search or backtracking line search to find α_k
4. Let $\mathbf{x}^{k+1} = \mathbf{x}^k + \alpha_k \mathbf{d}^k$, let $k = k + 1$. Go back to step 1.

Illustration

Minimize $f(x) = \exp(x_1 + x_2) + x_1^2 + 3x_2^2 - x_1 x_2$ using gradient method.



(a) Exact Line Search



(b) Backtracking Line Search

Another Example

Consider the least squares problem

$$\text{minimize}_{\mathbf{x}} \quad ||A\mathbf{x} - \mathbf{b}||_2^2$$

Use gradient method to solve this problem

Convergence Result of Gradient Descent Method

Theorem (See Boyd page 468)

Under some technical conditions on $f(\mathbf{x})$, in the gradient method

$$f(\mathbf{x}^k) - p^* \leq c^k (f(\mathbf{x}^0) - p^*)$$

where p^* is the objective value of a local minimizer, $c < 1$ is some constant.

Therefore, the gradient method is guaranteed to converge to a local minimizer.

Corollary

If $f(\mathbf{x})$ is convex, then the gradient method is guaranteed to converge to a global minimizer.

Definition

We call the convergence speed specified in the above theorem *linear convergence*

Linear Convergence

In the linear convergence,

$$f(\mathbf{x}^k) - f(\mathbf{x}^*) < L \cdot c^k$$

for some $c < 1$ and some constant L.

Therefore, in order to achieve ϵ optimal, one needs about $\log_c \epsilon$ iterations.

For example, if $c = 0.9$, then

Target accuracy ϵ	0.1	0.01	0.001	0.0001
Iterations needed	22	44	66	88

- We can see the reason of calling it linear convergence

Properties of Gradient Descent Method

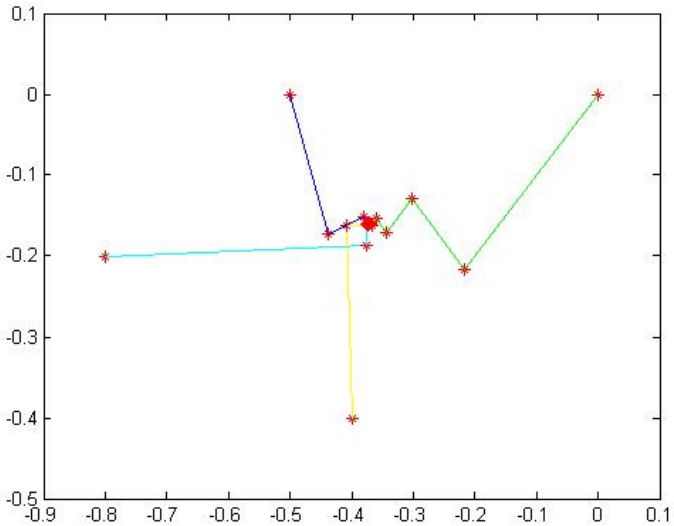
The convergence of gradient descent method doesn't depend on the initial point

- No matter where it starts, it will always converge to a local minimizer (when $f(\cdot)$ is convex, it converges to a global minimizer)
- We call this property *global convergence* property.

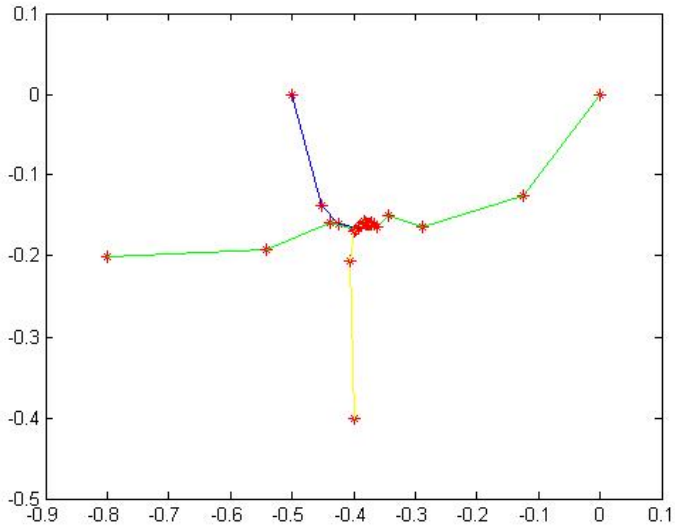
Global Convergence

We use the same function as example

$$f(\mathbf{x}) = \exp(x_1 + x_2) + x_1^2 + 3x_2^2 - x_1 x_2$$



(c) Exact Line Search



(d) Backtracking Line Search

More Properties of Gradient Descent Method

We have seen that when using exact line search, the directions between consecutive steps are perpendicular, i.e.,

$$(\mathbf{d}^{k+1})^T \mathbf{d}^k = 0$$

In fact, this is always true when using exact line search.

Why?

- If α_k is the minimizer of $f(\mathbf{x}^k + \alpha \mathbf{d}^k)$. Then the gradient of $f(\mathbf{x}^k + \alpha \mathbf{d}^k)$ with respect to α must be 0 at α_k, which means

$$\nabla f(\mathbf{x}^k + \alpha_k \mathbf{d}^k)^T \mathbf{d}^k = (\mathbf{d}^{k+1})^T \mathbf{d}^k = 0$$

Pros and Cons of Gradient Descent Method

Pros:

- Easy to understand and implement
- Only need to know the first-order (gradient) information
- Globally convergent, doesn't depend on the initial point

Cons:

- Convergence speed may not be fast enough: Linear convergence

Newton's Method

Next we study another method for unconstrained optimization:

- Newton's method

It has the following features:

- Converge much faster than the gradient method
- Require second-order information (second-order derivative)
- Sensitive to the initial point

Newton's Method: One Dimension

We want to minimize $f(x)$

- A necessary condition is $g(x) = f'(x) = 0$. We first try to find such points.

Newton's method is also an iterative method. At each point x^k.
We first approximate $g(x)$ using first-order Taylor expansion at x^k:

$$g(x) \approx g(x^k) + g'(x^k)(x - x^k)$$

We set the right-hand side to be 0 and solve it:

$$x = x^k - \frac{g(x^k)}{g'(x^k)}$$

We choose this to be x^{k+1}

- Here we assume $g'(x) \neq 0$ at each step

Illustration of Newton's Method to Find $g(x) = 0$

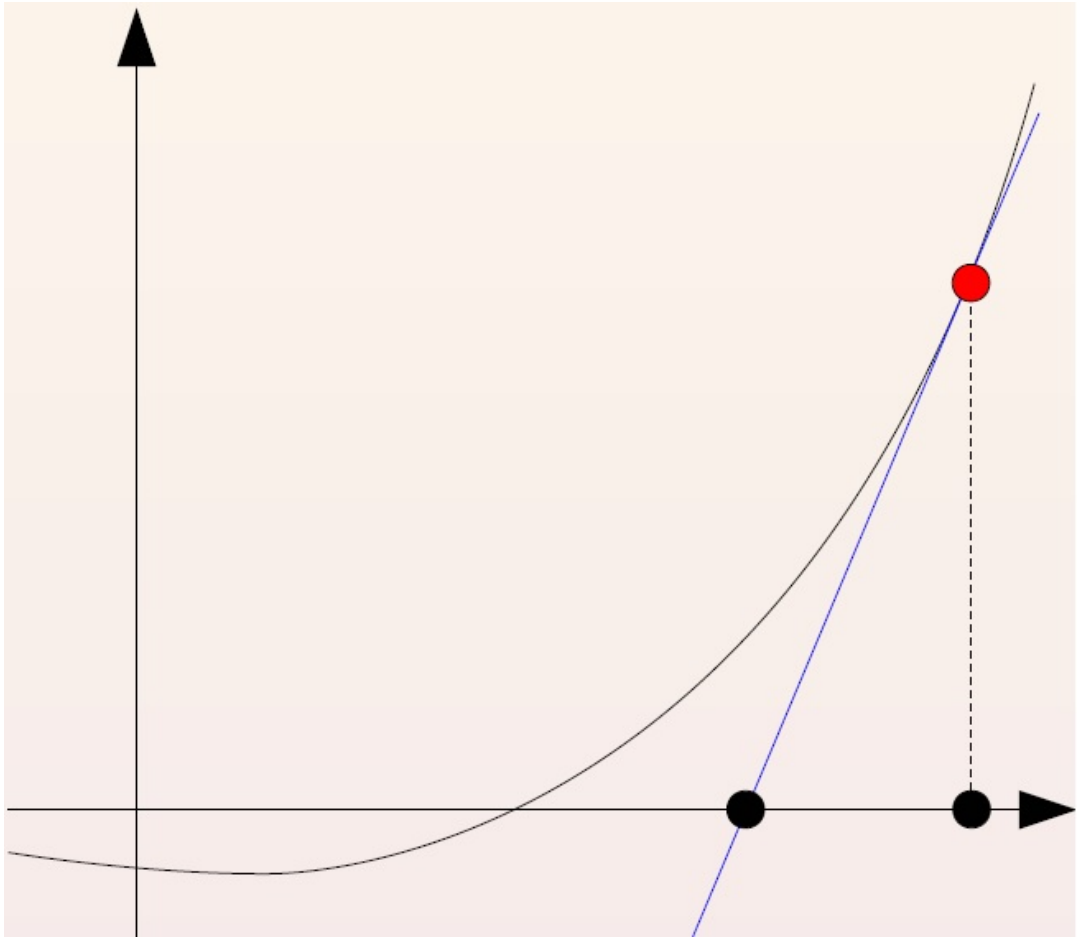


Illustration of Newton's Method to Find $g(x) = 0$

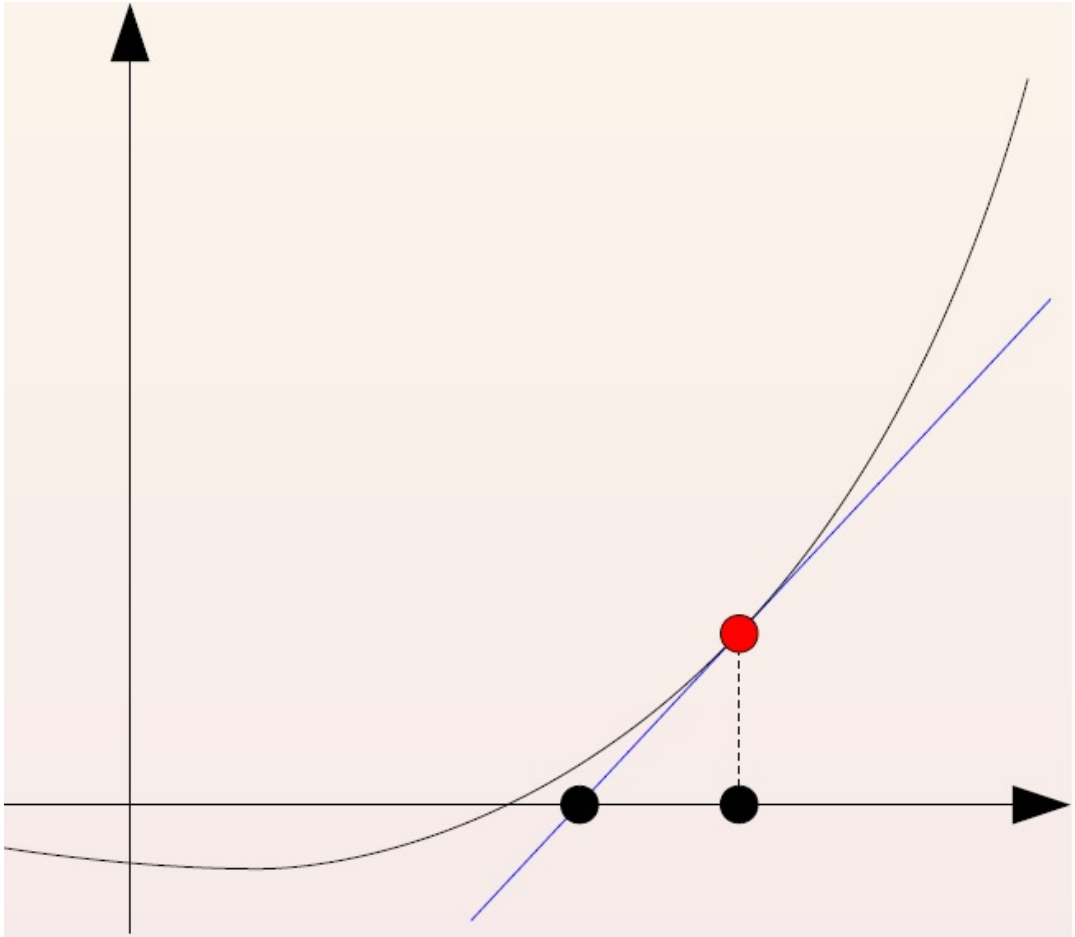
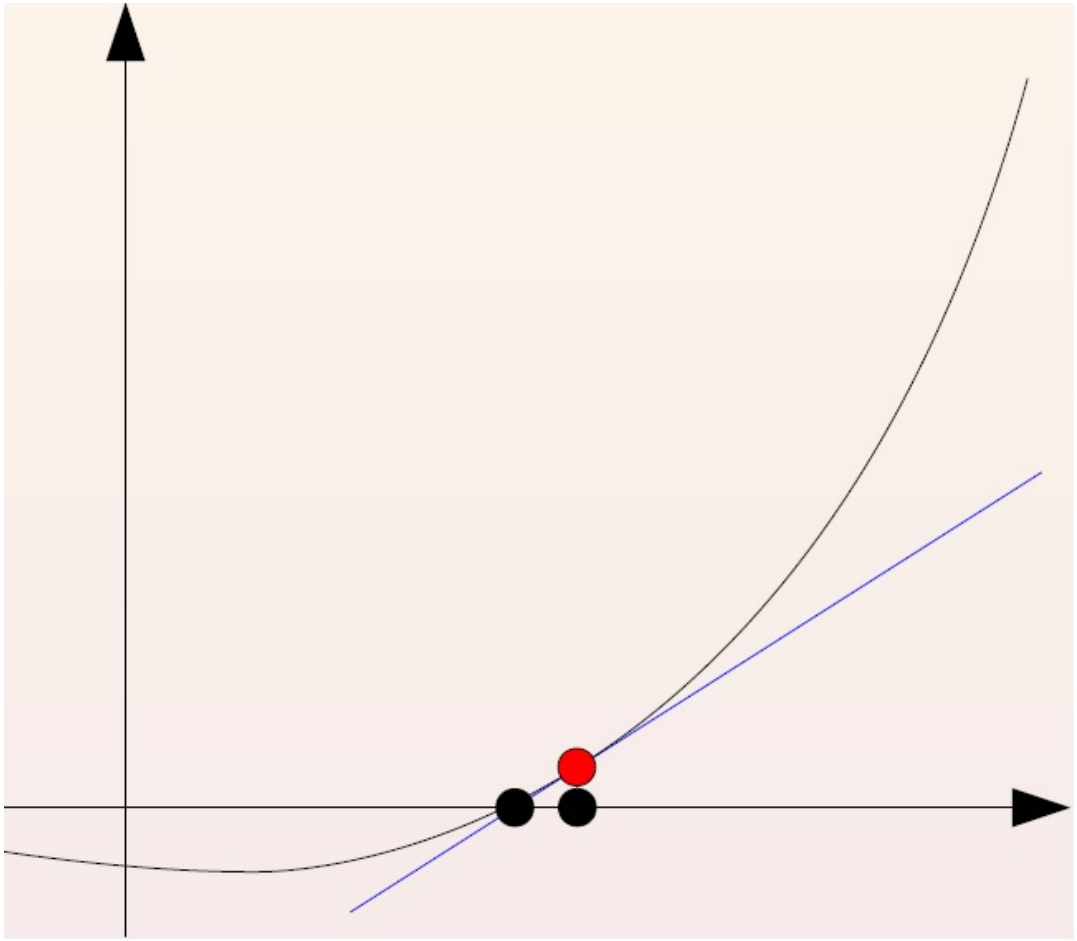


Illustration of Newton's Method to Find $g(x) = 0$

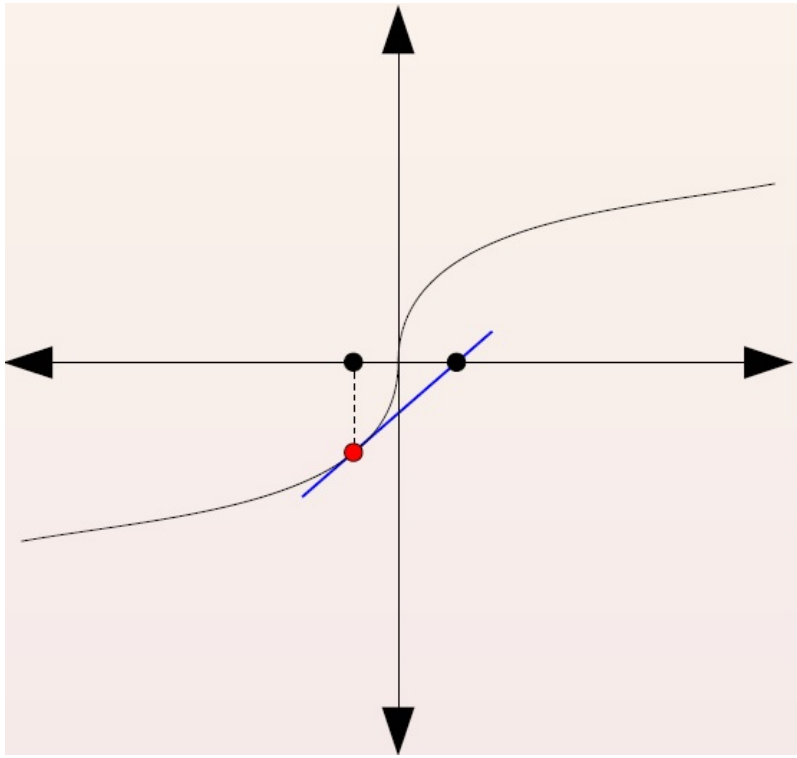


Another animation on Wikipedia

However..

Newton's method may not converge for some initial points.

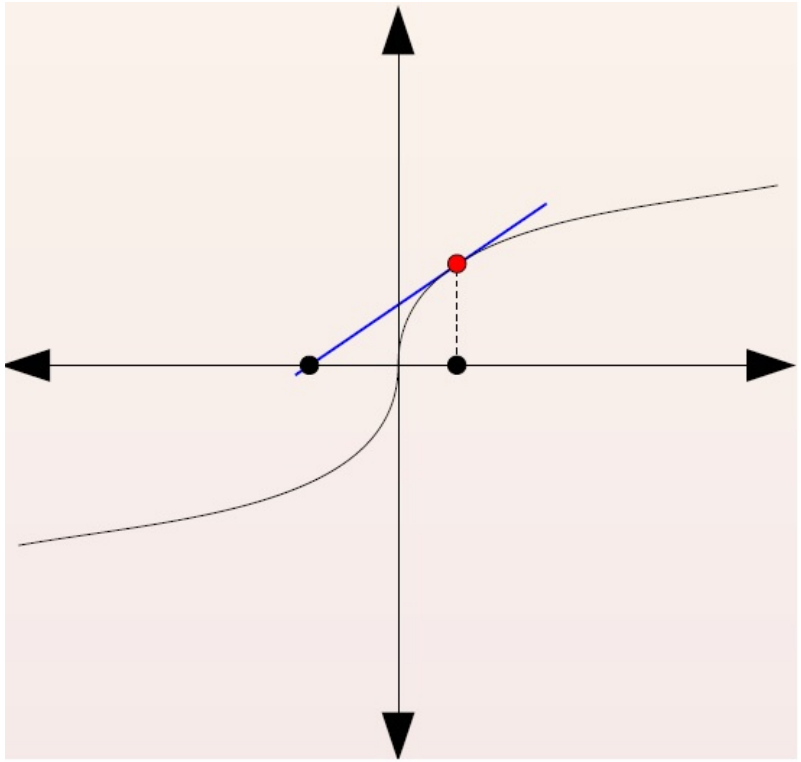
- Consider $g(x) = x^{1/3}$. It has root $x = 0$.



However..

Newton's method may not converge for some initial points.

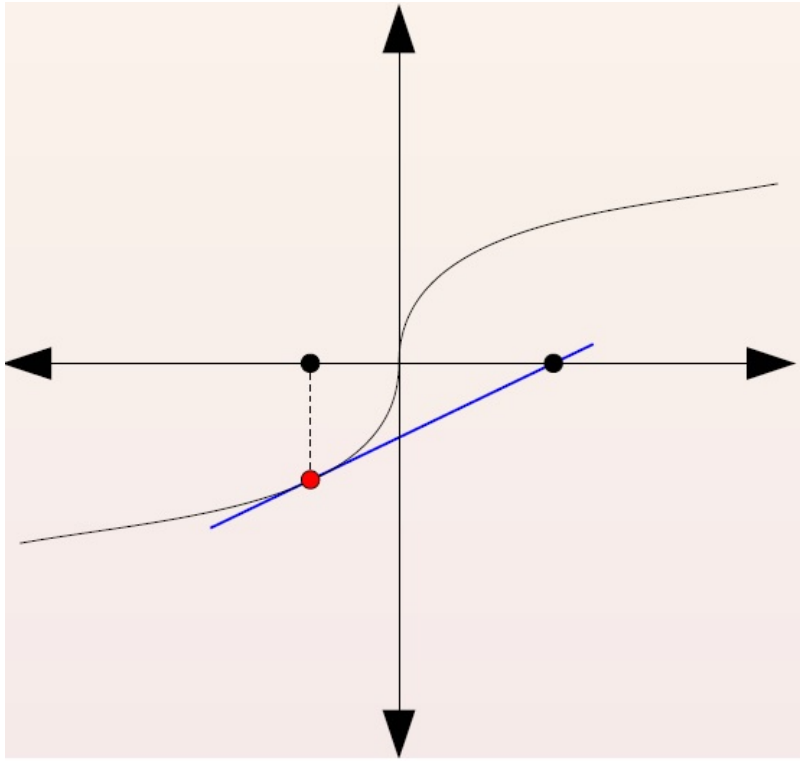
- Consider $g(x) = x^{1/3}$. It has root $x = 0$.



However..

Newton's method may not converge for some initial points.

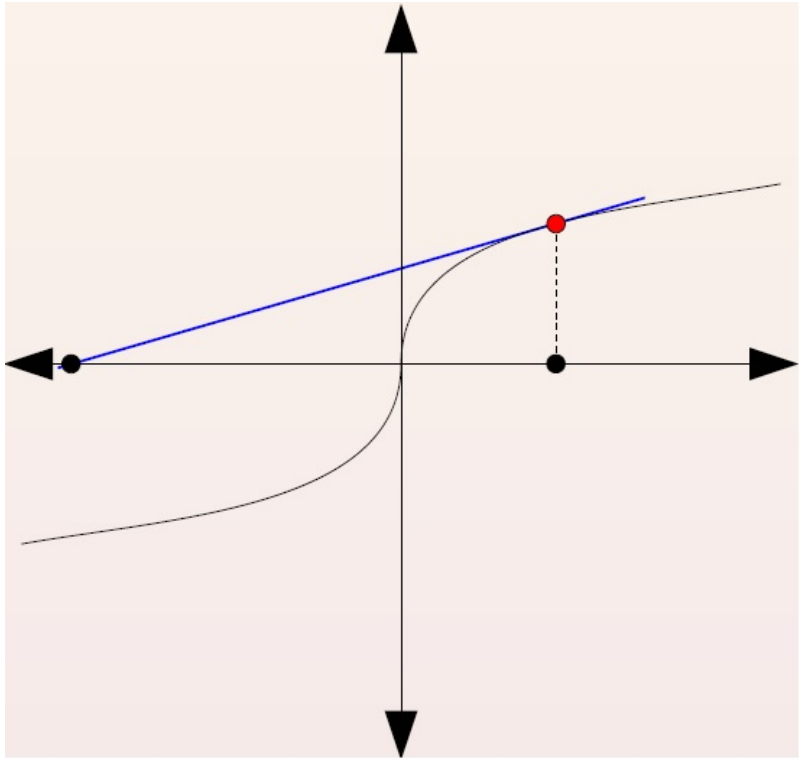
- Consider $g(x) = x^{1/3}$. It has root $x = 0$.



However..

Newton's method may not converge for some initial points.

- Consider $g(x) = x^{1/3}$. It has root $x = 0$.



Convergence of Newton's Method (for 1-Dimension Case)

Theorem

If $g(x)$ is twice continuously differentiable and x^* is a root of $g(x)$ at which $g'(x^*) \neq 0$, then provided that $|x^0 - x^*|$ is sufficiently small, the sequence generated by the Newton iterations:

$$x^{k+1} = x^k - \frac{g(x^k)}{g'(x^k)}$$

will satisfy

$$|x^{k+1} - x^*| \leq C|x^k - x^*|^2$$

with $C = \sup_x \frac{1}{2}|\frac{g''(x)}{g'(x)}|$.

- We call this convergence speed *quadratic convergence*

Linear Convergence vs Quadratic Convergence

Remember gradient descent method has linear convergence rate

$$f(x^{k+1}) - f(x^*) \leq C_1(f(x^k) - f(x^*))$$

Now Newton's method has quadratic convergence rate

$$|x^{k+1} - x^*| \leq C_2|x^k - x^*|^2$$

Let's set $C_1 = C_2 = 0.5$ and the first term to be 0.5. Then

Iteration	1	2	3	5
Gradient (linear conv.)	0.25	0.125	0.063	0.031
Newton (quadratic conv.)	0.125	0.0078	3×10^{-5}	1×10^{-19}

In order to achieve 1×10^{-19}, Newton's method needs 5 iterations, while the gradient method needs 64 iterations.

Back to the Optimization Problem

Remember $g(x) = f'(x)$ where $f(x)$ is the function we want to minimize.

Therefore, in terms of $f(\cdot)$, the Newton iteration can written as:

$$x^{k+1} = x^k - \frac{f'(x^k)}{f''(x^k)}$$

With proper conditions, this sequence of $\{x^k\}$ converges to a KKT point

- When $f(\cdot)$ is convex, it converges to the global minimizer (under proper conditions)

Connection to the Gradient Descent Method

In Newton's method:

$$x^{k+1} = x^k - \frac{f'(x^k)}{f''(x^k)}$$

Remember in the gradient descent method:

$$x^{k+1} = x^k - \alpha f'(x^k)$$

Therefore in the one dimension case, Newton's method simply specifies a unique step size in the gradient method (rather than doing line searches).

- In high-dimensional case, however, Newton's method will also alter the direction.

Another View of Newton's Method

Consider the function $f(x)$ we want to minimize. We first write the second-order Taylor expansion at current step x^k:

$$f(x) \approx f(x^k) + f'(x^k)(x - x^k) + \frac{1}{2} f''(x^k)(x - x^k)^2$$

What is the minimizer of the right hand side term?

- Solve the quadratic function, we get the minimizer is:

$$x^k - \frac{f'(x^k)}{f''(x^k)}$$

  which is exactly the next iterate in Newton's method.

- Therefore one can view Newton's method as first approximating the objective function by a quadratic function locally, then minimize that quadratic function.
- If the original objective function is a quadratic function, then Newton's method converges in 1 step.
- This idea is useful to study high-dimensional case.