# MAT 3007 – Optimization

## Convergence and Newton's Method

*Lecture 17* *July 16th*

Andre Milzarek SDS / CUHK-SZ

## Repetition

# Recap: Optimization Methods

One-Dimensional Problems:

- Bisection method: solve $f'(x) = 0$.
- Golden section method: does not require $f'$.

High-Dimensional Problems:

- General framework: Choose a descent direction $d^k$ and a stepsize $\alpha_k$ in each iteration.
- Gradient descent method: Choose $d^k = -\nabla f(x^k)$.
- Stepsize: We can use exact line search via applying golden section method. (Might not be very efficient in practice).
- The most commonly used method is backtracking line search.

# Backtracking / Armijo Line Search

Assume we have found a descent direction $d^k$ and we want to choose step size $\alpha_k$.

Let $\sigma, \gamma \in (0,1)$ be given. Choose $\alpha_k$ as the largest element in $\{1, \sigma, \sigma^2, \sigma^3, ...\}$ such that

$$f(x^k + \alpha_k d^k) - f(x^k) \leq \gamma \alpha_k \cdot \nabla f(x^k)^\top d^k.$$

- This condition is called Armijo condition.
- $\alpha_k$ can be determined after finitely many steps if $d^k$ is a descent direction.

Procedure:

1. Start with $\alpha = 1$.
2. If $f(x^k + \alpha d^k) \leq f(x^k) + \gamma\alpha \cdot \nabla f(x^k)^\top d^k$, choose $\alpha_k = \alpha$. Otherwise, set $\alpha = \sigma\alpha$ and repeat this step.

## Gradient Descent Method

1. Initialization: Select an initial point $x^0 \in \mathbb{R}^n$.

**For** $k = 0, 1, \ldots$:

2. Pick a stepsize $\alpha^k$ by a line search procedure (exact line search or backtracking) on the function

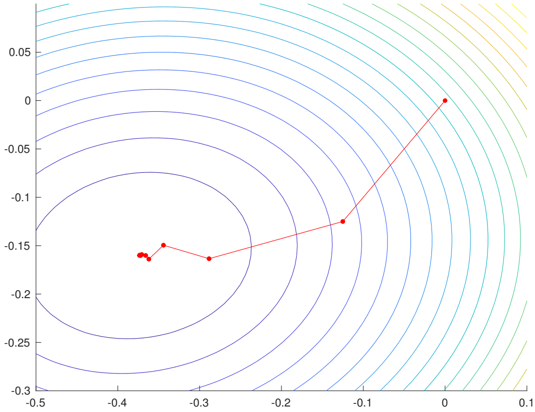
$$\phi(\alpha) = f(x^k - \alpha \nabla f(x^k)).$$

3. Set $x^{k+1} = x^k - \alpha_k \nabla f(x^k)$.
4. If $\|\nabla f(x^{k+1})\| \leq \varepsilon$, then STOP and $x^{k+1}$ is the output.

Minimize

$$f(x) = \exp(x_1 + x_2) + x_1^2 + 3x_2^2 - x_1x_2$$

using the gradient method with Armijo line search.



## Gradient Method: Convergence and Properties

We now derive and analyze different convergence properties of the gradient method.

Global Convergence:

- We show that the gradient method can find stationary points independent of the chosen initial point.
- We call such a property global convergence.

Local Convergence and Rate of Convergence:

- Under appropriate assumptions a rate of convergence can be established.

$\rightsquigarrow$ Guaranteed and quantifiable progress in each iteration.

We start with a definition of accumulation points.

**Definition: Accumulation Point**

A point $x$ is an accumulation point of $(x^k)_k$ if for every $\varepsilon > 0$, there are infinitely many numbers $k$ with $x^k \in B_\varepsilon(x)$.

We continue we several remarks:

- If $x$ is an accumulation point of $(x^k)_k$ then there exists a subsequence $(x^{k_\ell})_\ell$ that converges to $x$.
- If $(x^k)_k$ converges to some $x \in \mathbb{R}^n$, then $x$ is the unique accumulation point of $(x^k)_k$.
- A bounded sequence always possesses at least one accumulation point.

Examples:

- The sequence $(a_k)_k$ with $a_k = (-1)^k$ has the two accumulation points $a = +1$ and $a = -1$.
- The sequence

$$a_k := \begin{cases} k & k \text{ is odd}, \\ 0 & k \text{ is even}, \end{cases}$$

is not bounded. However, it has the accumulation point $a = 0$.

Let $f : \mathbb{R}^n \to \mathbb{R}$ be cont. diff. and let $(x^k)_k$ be generated by the gradient method for solving

$$\min_x \; f(x) \quad \text{s.t.} \quad x \in \mathbb{R}^n$$

with one of the following step size strategies:

- exact line search,
- Armijo line search (backtracking) with $\sigma, \gamma \in (0, 1)$.

Then, $(f(x^k))_k$ is nonincreasing and every accumulation point of $(x^k)_k$ is a stationary point of $f$.

- If $\nabla f(x^k) \neq 0$ for all $k$ (i.e., the method does not terminate after finitely many steps), then the acc. points of $(x^k)_k$ can only be local/global minima or saddle points!

Can we say more? What's the typical situation?

- If $f$ is a polynomial function of the variables $x_1, x_2, ..., x_n$ and $(x^k)_k$ is bounded, the whole sequence $(x^k)_k$ converges to a stationary point $x^*$ of $f$.

- Let $x^*$ be an acc. point of $(x^k)_k$ and suppose that the second order sufficient optimality conditions hold at $x^*$:

$\rightsquigarrow$ The sequence $(x^k)_k$ converges to the strict local min. $x^*$.

# Convergence: Additional Results and Comments

- If $\nabla f(x^k) \neq 0$ for all $k$ (i.e., the method does not terminate after finitely many steps), then the acc. points of $(x^k)_k$ can only be local/global minima or saddle points!

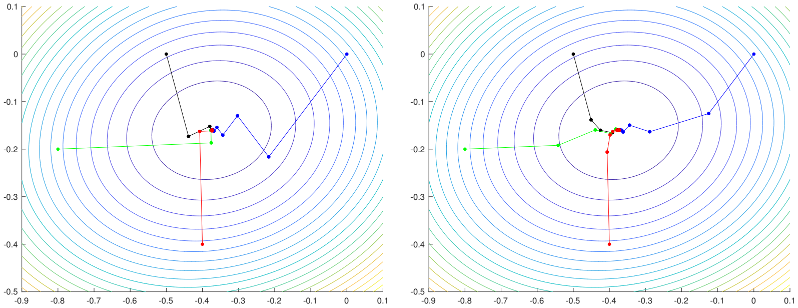
Can we say more? What's the typical situation?

- If $f$ is a polynomial function of the variables $x_1, x_2, ..., x_n$ and $(x^k)_k$ is bounded, the whole sequence $(x^k)_k$ converges to a stationary point $x^*$ of $f$.
- Let $x^*$ be an acc. point of $(x^k)_k$ and suppose that the second order sufficient optimality conditions hold at $x^*$:

$\rightsquigarrow$ The sequence $(x^k)_k$ converges to the strict local min. $x^*$.

We use the same function as example:

$$f(x) = \exp(x_1 + x_2) + x_1^2 + 3x_2^2 - x_1x_2$$




- Left: exact line search. Right: backtracking.

## Local Convergence and Rates

$\rightsquigarrow$ We require some additional properties to derive rates.

We need to assume that $\nabla f$ is Lipschitz continuous over $\mathbb{R}^n$:

$$\|\nabla f(x) - \nabla f(y)\| \le L\|x - y\|, \quad \forall\, x, y \in \mathbb{R}^n,$$

where $L > 0$ is the Lipschitz constant. The class of functions with Lipschitz gradient with constant $L$ is denoted by $C_L^{1,1}(\mathbb{R}^n)$ or $C_L^{1,1}$.

Examples:

- The linear function $f(x) := b^\top x + c$, $b \in \mathbb{R}^n$, $c \in \mathbb{R}$, is in $C_0^{1,1}$.
- Consider the quadratic function $f(x) := \frac{1}{2}x^\top Ax + b^\top x + c$:

$$\begin{aligned} \|\nabla f(x) - \nabla f(y)\| &= \|(Ax + b) - (Ay + b)\| \\ &= \|A(x - y)\| \le \|A\| \cdot \|x - y\|. \end{aligned}$$

  Hence, we have $f \in C_L^{1,1}$ with $L = \|A\|$.

# Remarks and Lipschitz Continuity via Hessian

Remarks:

- Here, the norm $\|A\|$ denotes the so-called spectral norm of $A$:

$$\|A\| = \sqrt{\lambda_{\max}(A^\top A)} = \max_{\|d\|=1} \|Ad\|$$

- If $f \in C_L^{1,1}$, we can also use constant stepsizes $\bar{\alpha} \in (0, \frac{2}{L})$.

If $f$ is twice continuously differentiable, then Lipschitz continuity of the gradient is equivalent to boundedness of the Hessian.

**Theorem: Lipschitz Continuity via Hessians**

Let $f$ be a twice cont. differentiable function. Then, the following two conditions are equivalent:

- $f \in C_L^{1,1}(\mathbb{R}^n)$.
- $\|\nabla^2 f(x)\| \le L$ for any $x \in \mathbb{R}^n$.

**Definition: Linear Convergence**

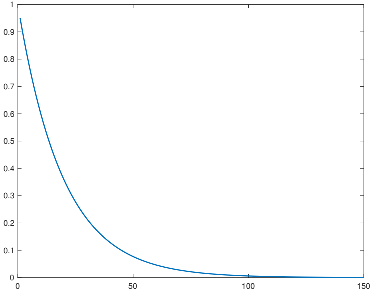
We say that $(x^k)_k$ converges linear with rate $\eta \in (0,1)$ to $x^* \in \mathbb{R}^n$ if there is $\ell \geq 0$ such that

$$\|x^{k+1} - x^*\| \leq \eta \cdot \|x^k - x^*\|, \quad \forall\ k \geq \ell.$$

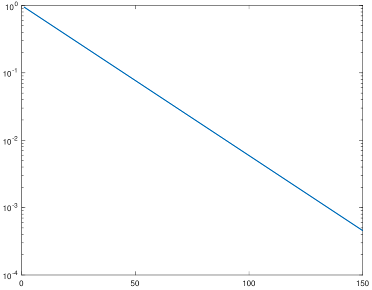
Example:

- Let $\eta \in (0,1)$ be given, then the sequence $(x^k)_k$ with $x^k := \eta^k$ converges linear to $x^* = 0$ with rate $\eta$. In fact, we have:

$$\frac{|x^{k+1} - x^*|}{|x^k - x^*|} = \frac{\eta^{k+1}}{\eta^k} = \eta, \quad \forall\ k \geq 0.$$




(a) Plot of $(0.95^k)_k$




(b) Logarithmic plot of $(0.95^k)_k$

- The plot in (b) shows convergence of the adjusted sequence $\tilde{x}^k = \log_{10}(0.95^k) = \log_{10}(0.95) \cdot k \approx -0.022 \cdot k$.
- The labels of the $y$-axis are given by $10^{\tilde{x}^k}$.
- In logarithmic plots, linear convergence corresponds to linear behavior with slope $\log_{10}(0.95)$.

# Theorem: Rates for Convex Problems

Let $f \in C_L^{1,1}$ and suppose there exists $\mu > 0$ such that

$$\mu\|d\|^2 \ \leq \ d^\top \nabla^2 f(x) d \ \ (\leq \ L\|d\|^2) \quad \forall\, d,\ \forall\, x.$$

Let $(x^k)_k$ be generated by the gradient method and let $x^*$ be the solution of $\min_x f(x)$. Then:

$(x^k)_k$ converges linearly to $x^*$

with rate $\eta = 1 - \frac{M\mu}{2}$ (see next slide $\rightsquigarrow M$) and it follows

$$f(x^k) - f(x^*) \leq \eta^k \cdot [f(x^0) - f(x^*)]$$

and

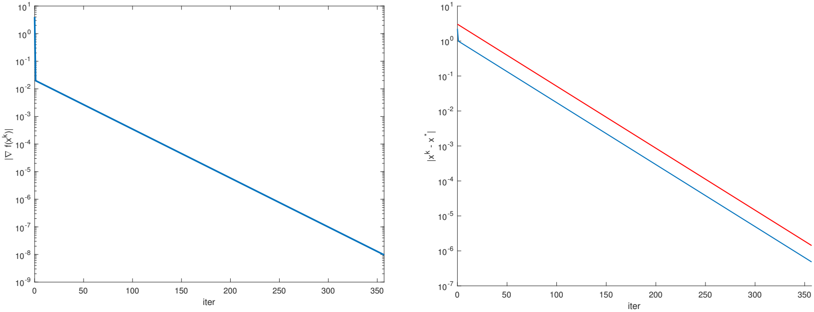
$$\|\nabla f(x^k)\| \leq \sqrt{\frac{L}{\mu}\eta^k} \cdot \|\nabla f(x^0)\|, \ \|x^k - x^*\| \leq \sqrt{\frac{L}{\mu}\eta^k} \cdot \|x^0 - x^*\|.$$

Remarks:

- The constant $M$ depends on the chosen line search procedure:

$$M = \begin{cases} \bar{\alpha}(1 - \frac{L\bar{\alpha}}{2}) & \text{constant step size: } \bar{\alpha} \in (0, \frac{2}{L}), \\ \frac{1}{2L} & \text{exact line search,} \\ \gamma \min\{1, \frac{2\sigma(1-\gamma)}{L}\} & \text{Armijo line search.} \end{cases}$$

- In the theorem a stronger notion of convexity is required – the so-called strong convexity.




- Gradient method with backtracking ($\gamma = \sigma = \frac{1}{2}$) for

$$\min_x \frac{1}{2} x^\top A x, \quad A = \begin{pmatrix} 2 & 0 \\ 0 & \frac{1}{50} \end{pmatrix}, \quad x^0 = \begin{pmatrix} 2 \\ 1 \end{pmatrix}.$$

  It holds $L = 2$, $\mu = \frac{1}{50}$ and the predicted rate is $\frac{799}{800} \approx 0.998$.

- Logarithmic plot of $(\|\nabla f(x^k)\|)_k$ and $(\|x^k - x^*\|)_k$. In red, the actual rate $\gamma \approx 0.96$ is shown.

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

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

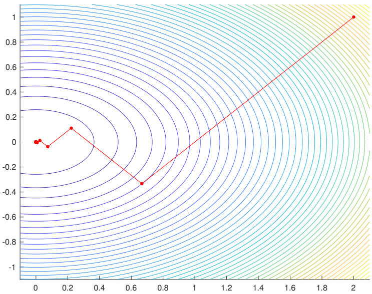
In fact, this is always true when using exact line search.

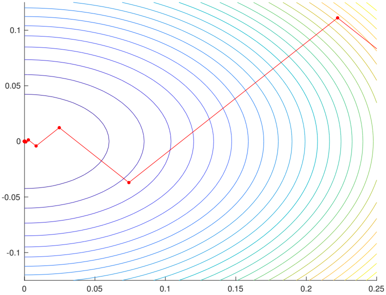
Why?

- If $\alpha_k$ is the minimizer of $\phi(\alpha) = f(x^k + \alpha d^k)$. Then, $\phi'(\alpha_k) = 0$, which means:

$$0 = \phi'(\alpha_k) = \nabla f(x^k + \alpha_k d^k)^\top d^k = -(d^{k+1})^\top d^k.$$

# Example: Perpendicular Steps



# Discussion and Summary

Pros:

- Easy to understand and implement.
- Only need to know the first-order (gradient) information.
- Globally convergent, does not depend on the initial point.

Cons:

- Convergence speed may not be fast enough $\rightsquigarrow$ linear convergence.

Newton's Method

# 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).
- More sensitive to the initial point.

## Newton's Method – in $\mathbb{R}$

We want to minimize $f$:

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

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

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

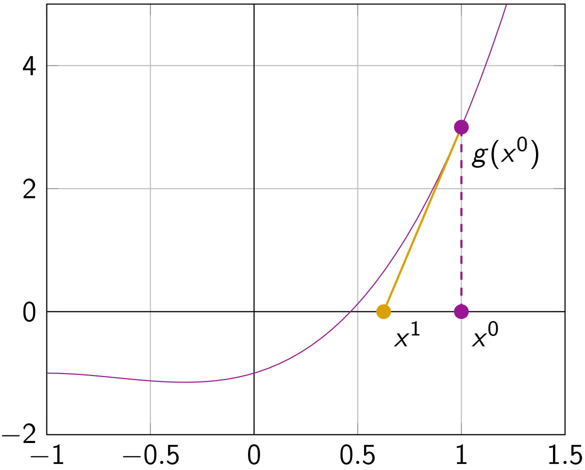
We set the right-hand side to be 0 and solve it:

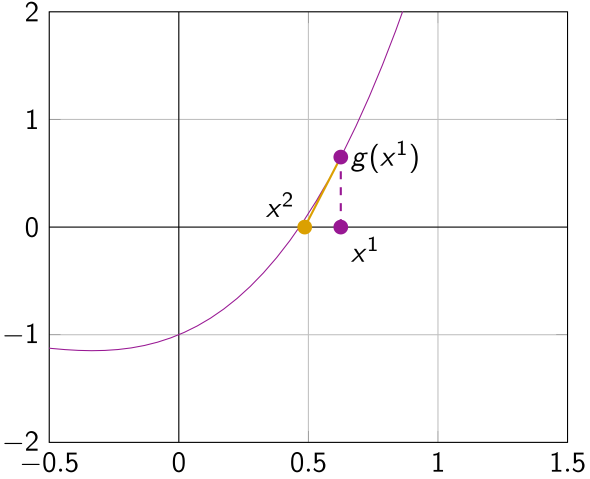
$$x = x^k - \frac{g(x^k)}{g'(x^k)}$$

We choose this $x$ as our next iterate $x^{k+1}$.

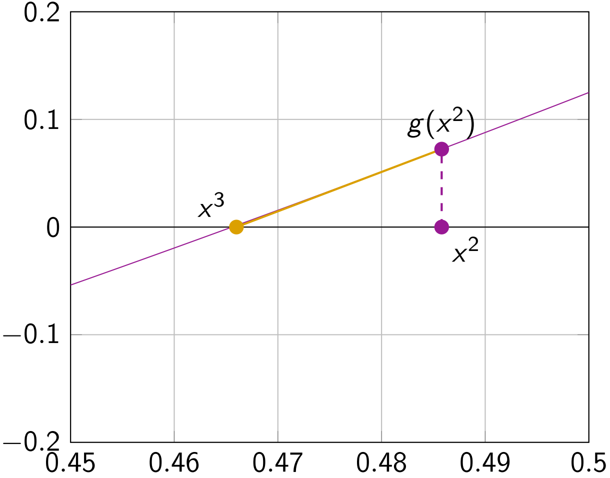
- Here we assume $g'(x) \neq 0$ at each step!

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



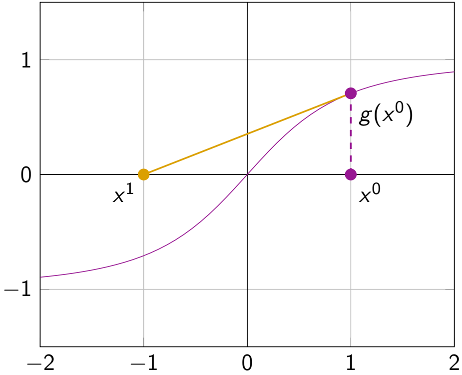

2
1
0
−1
−2
−0.5
0
0.5
1
1.5
$g(x^1)$
$x^2$
$x^1$

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



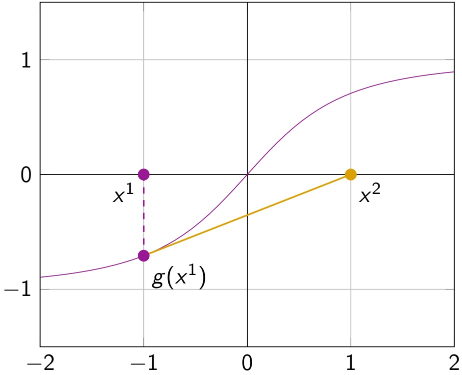
Newton's method may not converge for every initial point.

- Consider $g(x) = x/\sqrt{1+x^2}$. It has root $x = 0$.



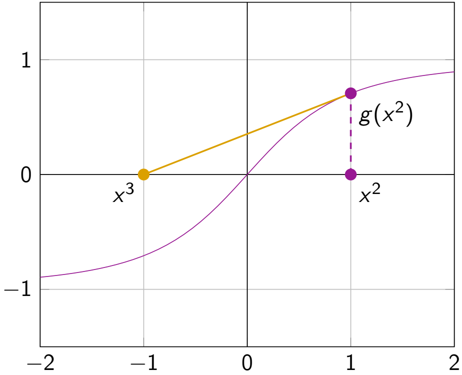
Newton's method may not converge for every initial point.

- Consider $g(x) = x/\sqrt{1+x^2}$. It has root $x = 0$.



Newton's method may not converge for every initial point.

- Consider $g(x) = x/\sqrt{1+x^2}$. It has root $x = 0$.



# Convergence of Newton's Method (1-D Case)

## Theorem: Convergence Newton's Method

If $g$ is twice cont. differentiable and $x^*$ is a root of $g$ 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.

Remember gradient descent method has linear convergence rate:

$$|x^{k+1} - x^*| \le \eta |x^k - x^*|.$$

Now, Newton's method has quadratic convergence rate:

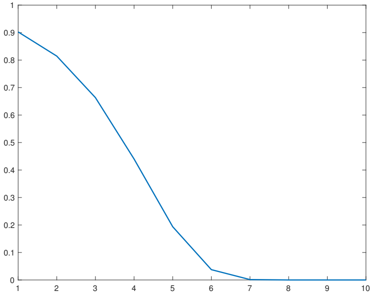
$$|x^{k+1} - x^*| \le C |x^k - x^*|^2.$$

Example:

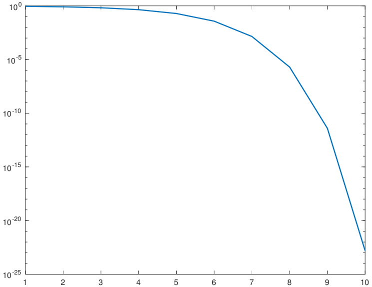
Let us set $\eta = C = 0.5$ and $|x^0 - x^*| = 0.5$. Then:

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

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




(a) Plot of $(0.95^{2^k})_k$




(b) Logarithmic plot of $(0.95^{2^k})_k$

- Quadratic convergence implies that the number of correct digits (i.e., the digits that coincide with the limit) double after each iteration.
- The logarithmic plot in (b) is similar to a quadratic function that opens downward.

# Back to the Optimization Problem

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

Therefore, in terms of $f$, the Newton iteration can written as:

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

- Under proper conditions, this sequence of $\{x^k\}$ converges to a stationary point of $f$.
- When $f$ is convex, it converges to the global minimizer (under appropriate assumptions).

One Newton step is given by:

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

A gradient descent step is given by:

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

Observation:

- In the 1-D case, Newton's method simply specifies a unique step size in the gradient method (rather than performing line searches).
- In the high-dimensional case, however, Newton's method will also alter the direction.

Consider the function $f$ 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 quadratic approximation?

- The minimizer is given by $(f''(x^k) > 0)$:

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

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

Interpretation:

- Newton's method build a quadratic approximation of $f$ locally. The Newton step then is the minimizer of this model.
- If the original objective function is quadratic, then Newton's method converges in one step.

## Newton's Method – in $\mathbb{R}^n$

We want to solve $\min_{x \in \mathbb{R}^n} f(x)$ with $f : \mathbb{R}^n \to \mathbb{R}$.

At $x^k$, we approximate the objective function by its second order Taylor expansion:

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

We minimize this quadratic approximation and get:

$$x = x^k - (\nabla^2 f(x^k))^{-1} \nabla f(x^k).$$

This motivates to define the search direction (Newton direction):

$$d^k = -(\nabla^2 f(x^k))^{-1} \nabla f(x^k).$$

In the gradient descent method, the direction is $-\nabla f(x^k)$.

- Newton's method refines the search direction by using the second-order information: $\nabla^2 f(x^k)$.

We can also consider the nonlinear equation $\nabla f(x) = 0$.

Using a Taylor expansion at $x^k$, we have

$$\nabla f(x) \approx \nabla f(x^k) + \nabla^2 f(x^k)(x - x^k) =: q_k(x).$$

The solution to $q_k(x) = 0$ is

$$x = x^k - (\nabla^2 f(x^k))^{-1} \nabla f(x^k)$$

which is also Newton's step.

- In these derivations, we assume that $\nabla^2 f(x)$ is invertible in the search region.

# Connection to Descent Directions

A vector $d$ is a descent direction if $\nabla f(x)^\top d < 0$.

- If we go a very small step in that direction, the objective value must be decreasing (due to Taylor's expansion).
- In the gradient descent method, we have $d = -\nabla f(x)$ and

$$\nabla f(x)^\top d = -\|\nabla f(x)\|^2 < 0.$$

# Newton's Step as a Descent Step

In Newton's method, we have

$$d = -(\nabla^2 f(x))^{-1} \nabla f(x).$$

Then, it holds that:

$$\nabla f(x)^\top d = -\nabla f(x)^\top (\nabla^2 f(x))^{-1} \nabla f(x).$$

- If $f$ is convex, then $\nabla^2 f(x)$ is positive semidefinite and we obtain $\nabla f(x)^\top d \leq 0$.
- If $\nabla^2 f(x)$ is positive definite, then $\nabla f(x)^\top d < 0$.

$\rightsquigarrow$ In this case, Newton's direction is a descent direction.

# Step Length

As we said earlier, Newton's method may not converge unless the starting point is close.

One way to ensure convergence is to again use a step size parameter $\alpha_k$ in

$$x^{k+1} = x^k + \alpha_k d^k$$

where $d^k = -(\nabla^2 f(x^k))^{-1}\nabla f(x^k)$ is Newton's direction.

- We can use backtracking line search to determine $\alpha_k$.

### The Newton Method

1. Initialization: Select an initial point $x^0 \in \mathbb{R}^n$.

**For** $k = 0, 1, \ldots$:

2. Compute the Newton direction $d^k$ which is the solution of the linear system
$$\nabla^2 f(x^k) d^k = -\nabla f(x^k).$$
3. Choose a step size $\alpha_k$ by backtracking line search and calculate $x^{k+1} = x^k + \alpha_k d^k$.
4. If $\|\nabla f(x^{k+1})\| \leq \varepsilon$, then STOP and $x^{k+1}$ is the output.

Questions?