

AMA 505: Optimization Methods

Subject Lecturer: Ting Kei Pong

Lecture 2 Unconstrained Optimization Optimality conditions and gradient descent

Problem setting

Aim: Given $f \in C^1(\mathbb{R}^n)$, solve

$$\underset{x \in \mathbb{R}^n}{\text{Minimize}} \quad f(x)$$

Problem setting

Aim: Given $f \in C^1(\mathbb{R}^n)$, solve

$$\underset{x \in \mathbb{R}^n}{\text{Minimize}} \quad f(x)$$

- Derivative-free methods:
 - ★ Pattern search;
 - ★ Grid search;
 - ★ Nelder-Mead method; etc.

Problem setting

Aim: Given $f \in C^1(\mathbb{R}^n)$, solve

$$\underset{x \in \mathbb{R}^n}{\text{Minimize}} \quad f(x)$$

- Derivative-free methods:
 - ★ Pattern search;
 - ★ Grid search;
 - ★ Nelder-Mead method; etc.
- Derivative-based methods:
 - ★ Steepest descent;
 - ★ Newton's method (if f is C^2);
 - ★ quasi-Newton method; etc.

Problem setting

Aim: Given $f \in C^1(\mathbb{R}^n)$, solve

$$\underset{x \in \mathbb{R}^n}{\text{Minimize}} \quad f(x)$$

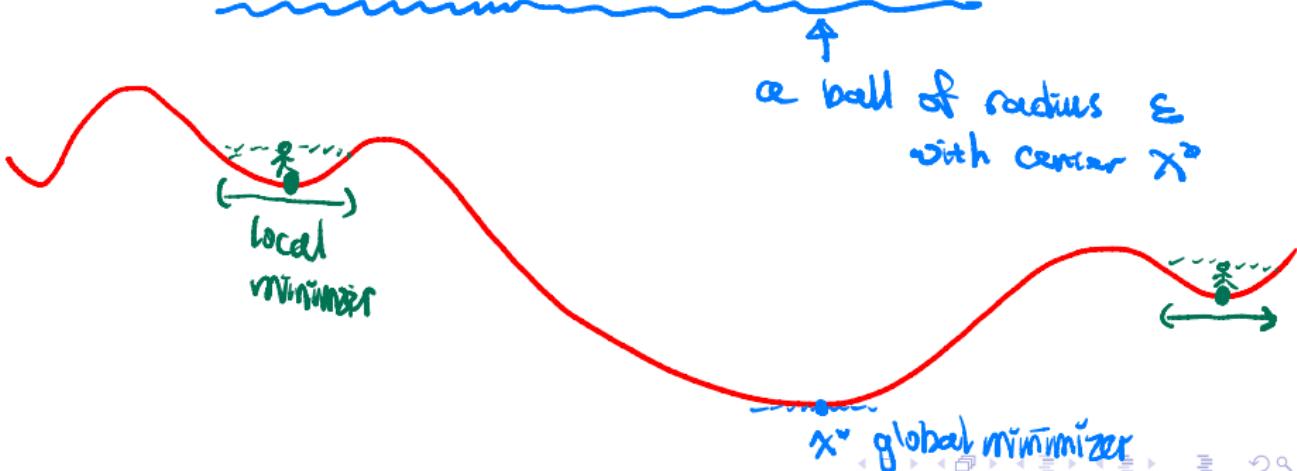
- Derivative-free methods:
 - ★ Pattern search;
 - ★ Grid search;
 - ★ Nelder-Mead method; etc.
- Derivative-based methods:
 - ★ Steepest descent;
 - ★ Newton's method (if f is C^2);
 - ★ quasi-Newton method; etc.

In general, finding global minimizers for f is NP-hard.

Minimizers

Definition:

- We say that x^* is a **global minimizer** of f if $f(x) \geq f(x^*)$ for all $x \in \mathbb{R}^n$.
- We say that x^* is a **local minimizer** of f if there exists $\epsilon > 0$ so that $f(x) \geq f(x^*)$ for all x satisfying $\|x - x^*\|_2 < \epsilon$.



Minimizers

Definition:

- We say that x^* is a **global minimizer** of f if $f(x) \geq f(x^*)$ for all $x \in \mathbb{R}^n$.
- We say that x^* is a **local minimizer** of f if there exists $\epsilon > 0$ so that $f(x) \geq f(x^*)$ for all x satisfying $\|x - x^*\|_2 < \epsilon$.

Remarks:

- Finding local minimizers is also NP-hard in general.
- In order to set a modest goal, we look at more properties of local minimizers.

1st-order necessary conditions

Theorem 2.1. Let $f \in C^1(\mathbb{R}^n)$ and suppose that x^* is a local minimizer of f . Then $\nabla f(x^*) = 0$.

n \leftarrow n equations

1st-order necessary conditions

Theorem 2.1. Let $f \in C^1(\mathbb{R}^n)$ and suppose that x^* is a local minimizer of f . Then $\nabla f(x^*) = 0$.

Proof: Fix any $h \in \mathbb{R}^n$. Then for all sufficiently small $t > 0$, there exists $\xi^t \in \{x^* + sth : s \in (0, 1)\}$ such that we have

$$f(x^*) \leq f(x^* + th) = f(x^*) + t[\nabla f(\xi^t)]^T h.$$

Hence,

$$[\nabla f(\xi^t)]^T h \geq 0.$$

Passing to the limit and noting that $\xi^t \rightarrow x^*$, we conclude that $[\nabla f(x^*)]^T h \geq 0$.

1st-order necessary conditions

Theorem 2.1. Let $f \in C^1(\mathbb{R}^n)$ and suppose that x^* is a local minimizer of f . Then $\nabla f(x^*) = 0$.

Proof: Fix any $h \in \mathbb{R}^n$. Then for all sufficiently small $t > 0$, there exists $\xi^t \in \{x^* + sth : s \in (0, 1)\}$ such that we have

$$f(x^*) \leq f(x^* + th) = f(x^*) + t[\nabla f(\xi^t)]^T h.$$

Hence,

$$[\nabla f(\xi^t)]^T h \geq 0.$$

Passing to the limit and noting that $\xi^t \rightarrow x^*$, we conclude that

$$[\nabla f(x^*)]^T h \geq 0.$$

$$0 \leq [\nabla f(x^*)]^T (-\nabla f(x^*)) = -\|\nabla f(x^*)\|_2^2$$

Set $h = -\nabla f(x^*)$ to obtain the desired conclusion.

2nd-order conditions

Definition: We say that x^* is a stationary point of f if $\nabla f(x^*) = 0$.

2nd-order conditions

Definition: We say that x^* is a stationary point of f if $\nabla f(x^*) = 0$.

Remarks: Local minimizers are stationary points.

2nd-order conditions

Definition: We say that x^* is a stationary point of f if $\nabla f(x^*) = 0$.

Remarks: Local minimizers are stationary points.

Theorem 2.2. Let $f \in C^2(\mathbb{R}^n)$.

1. If x^* is a local minimizer of f , then $\nabla^2 f(x^*) \succeq 0$.
2. If x^* is a stationary point of f and $\nabla^2 f(x^*) \succ 0$, then x^* is a local minimizer.

Proof: We first prove part 1. Fix any $h \in \mathbb{R}^n$. Then for all sufficiently small $t > 0$, there exists $\xi^t \in \{x^* + t\alpha h : \alpha \in (0, 1)\}$ such that

$$f(x^*) \leq f(x^* + th) = f(x^*) + t \underbrace{[\nabla f(x^*)]^T h}_{=0} + \frac{t^2}{2} h^T \nabla^2 f(\xi^t) h$$

2nd-order conditions

Definition: We say that x^* is a stationary point of f if $\nabla f(x^*) = 0$.

Remarks: Local minimizers are stationary points.

Theorem 2.2. Let $f \in C^2(\mathbb{R}^n)$.

1. If x^* is a local minimizer of f , then $\nabla^2 f(x^*) \succeq 0$.
2. If x^* is a stationary point of f and $\nabla^2 f(x^*) \succ 0$, then x^* is a local minimizer.

Proof: We first prove part 1. Fix any $h \in \mathbb{R}^n$. Then for all sufficiently small $t > 0$, there exists $\xi^t \in \{x^* + t\alpha h : \alpha \in (0, 1)\}$ such that

$$f(x^*) \leq f(x^* + th) = f(x^*) + t \underbrace{[\nabla f(x^*)]^T h}_{=0} + \frac{t^2}{2} h^T \nabla^2 f(\xi^t) h$$

Theorem 2.1

Taylor

local min

Hence, $h^T \nabla^2 f(\xi^t) h \geq 0$. Passing to the limit as $t \downarrow 0$, we obtain $h^T \nabla^2 f(x^*) h \geq 0$. Since this is true for any $h \in \mathbb{R}^n$, it follows that $\nabla^2 f(x^*) \succeq 0$.

2nd-order conditions cont.

Proof of Theorem 2.2 cont.: We now prove part 2. Since $\nabla^2 f(x^*) \succ 0$ and $f \in C^2(\mathbb{R}^n)$, there exists $\epsilon > 0$ so that $\nabla^2 f(y) \succ 0$ whenever $\|y - x^*\|_2 < \epsilon$.

2nd-order conditions cont.

Proof of Theorem 2.2 cont.: We now prove part 2. Since $\nabla^2 f(x^*) \succ 0$ and $f \in C^2(\mathbb{R}^n)$, there exists $\epsilon > 0$ so that $\nabla^2 f(y) \succ 0$ whenever $\|y - x^*\|_2 < \epsilon$.

Consider any nonzero h with $\|h\|_2 < \epsilon$. Then

$$\begin{aligned}f(x^* + h) &= f(x^*) + \int_0^1 \nabla f(x^* + th)^T h \, dt \\&= f(x^*) + \int_0^1 [\underbrace{\nabla f(x^* + th)^T h}_{\varphi(t)} - \underbrace{\nabla f(x^*)^T h}_{\varphi(0)}] \, dt \\&= f(x^*) + \int_0^1 th^T \nabla^2 f(x^* + \xi_t h) h \, dt\end{aligned}$$

for some $\xi_t \in [0, t] \subseteq [0, 1]$. Hence, $h^T \nabla^2 f(x^* + \xi_t h) h > 0$ and thus $f(x^* + h) \geq f(x^*)$.

Example 1

Example: Consider the function $f(x_1, x_2) = x_1^2 + (x_1 + 1)x_2^2$. Then

$$\nabla f(x) = \begin{bmatrix} 2x_1 + x_2^2 \\ 2x_2(x_1 + 1) \end{bmatrix}.$$

$\frac{\partial f}{\partial x_1}$ $\frac{\partial f}{\partial x_2}$

Hence, $\nabla f(x) = 0$ gives stationary points:

$$(0, 0), (-1, \sqrt{2}), (-1, -\sqrt{2}).$$

$$\left\{ \begin{array}{l} 2x_1 + x_2^2 = 0 \\ 2x_2(x_1 + 1) = 0 \end{array} \right.$$

Two cases

$$\begin{aligned} &x_2 = 0 \\ &\text{or} \\ &x_1 = -1 \end{aligned}$$

Example 1

Example: Consider the function $f(x_1, x_2) = x_1^2 + (x_1 + 1)x_2^2$. Then

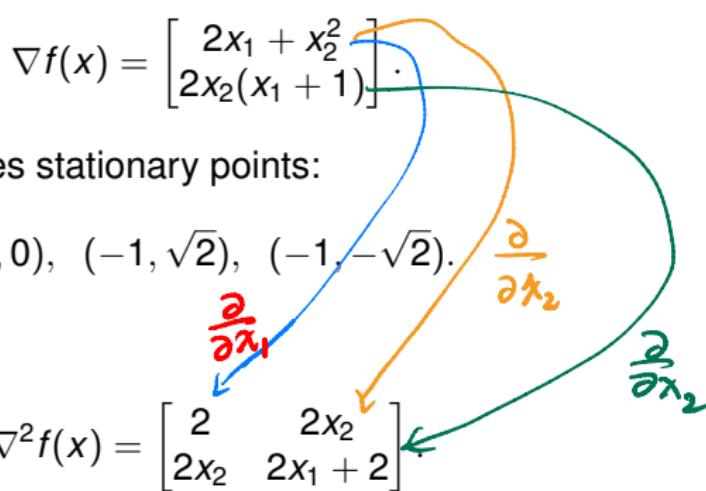
$$\nabla f(x) = \begin{bmatrix} 2x_1 + x_2^2 \\ 2x_2(x_1 + 1) \end{bmatrix}.$$

Hence, $\nabla f(x) = 0$ gives stationary points:

$$(0, 0), (-1, \sqrt{2}), (-1, -\sqrt{2}).$$

Next,

$$\nabla^2 f(x) = \begin{bmatrix} 2 & 2x_2 \\ 2x_2 & 2x_1 + 2 \end{bmatrix}$$



Example 1 cont.

Example cont.: Then

$$\nabla^2 f(0, 0) = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix} \succ 0 \Rightarrow (0, 0) \text{ is a local minimizer.}$$

$$\nabla^2 f(-1, \sqrt{2}) = \begin{bmatrix} 2 & 2\sqrt{2} \\ 2\sqrt{2} & 0 \end{bmatrix} \text{ is indefinite}$$

both +ve & -ve eigenvalues

$\Rightarrow (-1, \sqrt{2})$ is not a local minimizer nor maximizer.

↗ saddle point

$$\nabla^2 f(-1, -\sqrt{2}) = \begin{bmatrix} 2 & -2\sqrt{2} \\ -2\sqrt{2} & 0 \end{bmatrix} \text{ is indefinite}$$

$\Rightarrow (-1, -\sqrt{2})$ is not a local minimizer nor maximizer.

Q1: If $\nabla^2 f(x^*) \succeq 0$ and x^* is stationary, what can we say?

Cannot say anything

$$f(x) = x^4$$



local min
at $x^* = 0$

$$f(x) = -x^4$$



local max
at $x^* = 0$

$$f(x) = x^3$$

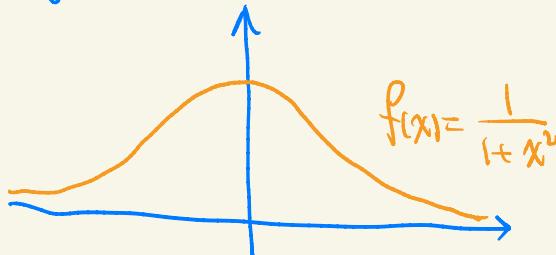


saddle pt at
 $x^* = 0$

Q2: If $f \in C^1(\mathbb{R})$, and it has ONE stationary point, is that a global minimizer? No  see this

What if we restrict ourselves to those f with $\inf f > -\infty$?

No.



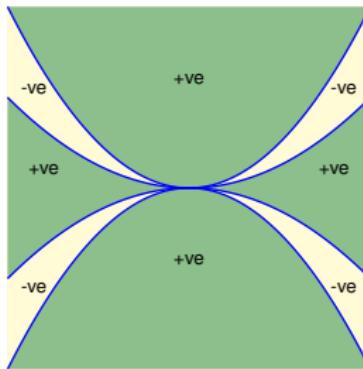
Example 2

skip

Example: Consider the function $f(x_1, x_2) = (x_2^2 - x_1^4) \left(x_2^2 - \frac{x_1^4}{4} \right)$ at the stationary point $(0, 0)$. Then for any $h \in \mathbb{R}^2 \setminus \{0\}$, there exists $t_0 > 0$ such that

$$f(th) > 0 \text{ for all } t \in (0, t_0).$$

However, $(0, 0)$ is not a local minimizer of f ! Note, however, that $\nabla^2 f(0, 0) = 0 \succeq 0$.



Example 2 cont.

Example cont.: Details: We show that for any $h \in \mathbb{R}^2 \setminus \{0\}$, there exists $t_0 > 0$ such that

$$f(th) > 0 \text{ for all } t \in (0, t_0).$$

Example 2 cont.

Example cont.: Details: We show that for any $h \in \mathbb{R}^2 \setminus \{0\}$, there exists $t_0 > 0$ such that

$$f(th) > 0 \text{ for all } t \in (0, t_0).$$

Case 1: $h = (h_1, h_2)$ for some $h_2 \neq 0$. Then

$$f(th) = t^4(h_2^2 - t^2h_1^4) \left(h_2^2 - t^2 \frac{h_1^4}{4} \right)$$

is positive for all sufficiently small $t > 0$.

Case 2: $h = (h_1, 0)$ for some $h_1 \neq 0$. Then

$$f(th) = (-t^4 h_1^4) \left(-t^4 \frac{h_1^4}{4} \right)$$

is positive for all $t > 0$.

Example 2 cont.

Example cont.: **Details**: We show that $(0, 0)$ is not a local minimizer.

Example 2 cont.

Example cont.: Details: We show that $(0, 0)$ is not a local minimizer.
For each $\epsilon \in (0, 1)$, consider

$$(x_1, x_2) = (\sqrt{3}\epsilon, 2\epsilon^2)$$

Then $\|(x_1, x_2)\|_2 < 3\epsilon$ and

$$f(x_1, x_2) = \epsilon^8(4 - 9) \left(4 - \frac{9}{4}\right) < 0.$$

Example 2 cont.

Example cont.: Details: We show that $(0, 0)$ is not a local minimizer.
For each $\epsilon \in (0, 1)$, consider

$$(x_1, x_2) = (\sqrt{3}\epsilon, 2\epsilon^2)$$

Then $\|(x_1, x_2)\|_2 < 3\epsilon$ and

$$f(x_1, x_2) = \epsilon^8(4 - 9) \left(4 - \frac{9}{4}\right) < 0.$$

Since $\epsilon > 0$ is arbitrary, we have shown that:

No matter how small we shrink the neighborhood $\{x : \|x\|_2 < 3\epsilon\}$, there is always a point in it such that f goes negative.

Thus, $(0, 0)$ is not a local minimizer.

Aim (revised)

Aim (Revised): Given $f \in C^1(\mathbb{R}^n)$:

- Find a stationary point of f (i.e., x^* so that $\nabla f(x^*) = 0$).
- Test whether it is a local minimizer by looking at $\nabla^2 f(x^*)$ if $f \in C^2(\mathbb{R}^n)$ and if Hessian is not too hard to compute.

Aim (revised)

Aim (Revised): Given $f \in C^1(\mathbb{R}^n)$:

- Find a stationary point of f (i.e., x^* so that $\nabla f(x^*) = 0$).
- Test whether it is a local minimizer by looking at $\nabla^2 f(x^*)$ if $f \in C^2(\mathbb{R}^n)$ and if Hessian is not too hard to compute.

First attempt: Solve $\nabla f(x) = 0$?

Aim (revised)

Aim (Revised): Given $f \in C^1(\mathbb{R}^n)$:

- Find a stationary point of f (i.e., x^* so that $\nabla f(x^*) = 0$).
- Test whether it is a local minimizer by looking at $\nabla^2 f(x^*)$ if $f \in C^2(\mathbb{R}^n)$ and if Hessian is not too hard to compute.

First attempt: Solve $\nabla f(x) = 0$? — Newton's method (when $f \in C^2(\mathbb{R}^n)$).

Newton's method

Let $x^0 \in \mathbb{R}^n$. For $k = 0, 1, 2, \dots$, update

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

Solve for d

$$\nabla^2 f(x^k) d = -\nabla f(x^k)$$



Newton direction

Aim (revised)

Aim (Revised): Given $f \in C^1(\mathbb{R}^n)$:

- Find a stationary point of f (i.e., x^* so that $\nabla f(x^*) = 0$).
- Test whether it is a local minimizer by looking at $\nabla^2 f(x^*)$ if $f \in C^2(\mathbb{R}^n)$ and if Hessian is not too hard to compute.

First attempt: Solve $\nabla f(x) = 0$? — Newton's method (when $f \in C^2(\mathbb{R}^n)$).

Newton's method

Let $x^0 \in \mathbb{R}^n$. For $k = 0, 1, 2, \dots$, update

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

Note:

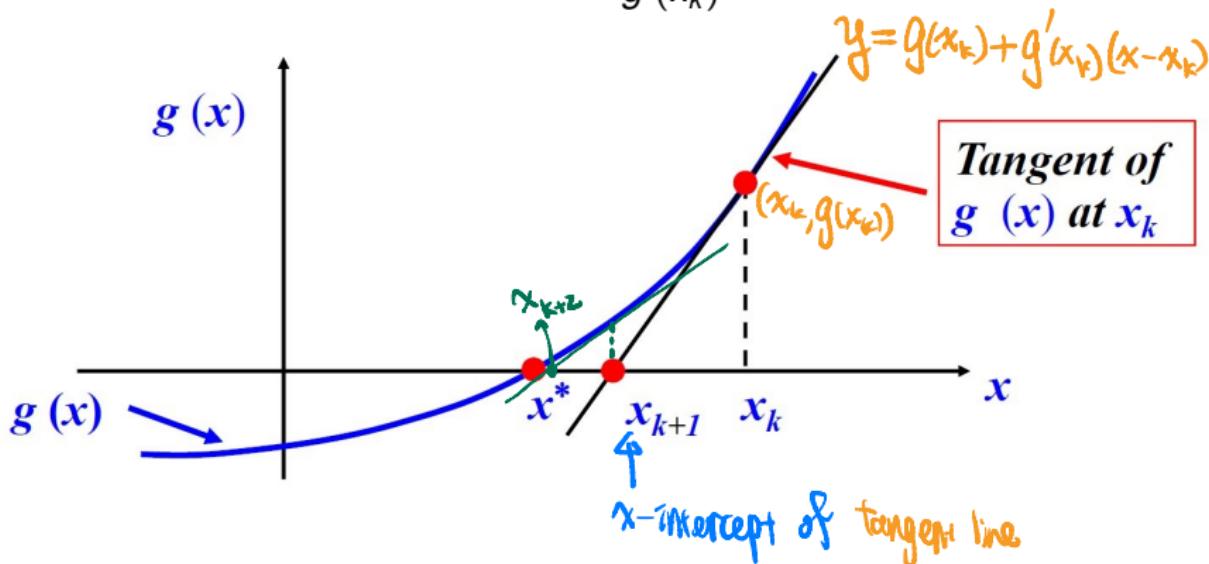
- The above iterates require that $\nabla^2 f(x^k)$ is invertible for each k . The method fails if $\nabla^2 f(x^k)$ is singular.
- In practice, computing $[\nabla^2 f(x^k)]^{-1} \nabla f(x^k)$ can be expensive.

Newton's method

$$f'(x) = 0$$

- In \mathbb{R} , to solve $g(x) = 0$ with $g \in C^1(\mathbb{R})$, the Newton's method takes the form

$$x_{k+1} = x_k - \frac{g(x_k)}{g'(x_k)}.$$



Newton's method cont.

Under certain conditions, Newton's method enjoys **fast local convergence**. For simplicity, we only state and prove the case for \mathbb{R} .

Theorem 2.3. (Quadratic convergence of Newton's method)

Let $g \in C^2(\mathbb{R})$ and x_* satisfies $g(x_*) = 0$ and $g'(x_*) \neq 0$. Then there exists $\epsilon > 0$ so that if $|x_0 - x_*| < \epsilon$, then the Newton's iterate $x_{k+1} = x_k - g(x_k)/g'(x_k)$ is well defined and there exists $M > 0$ so that

$$|x_{k+1} - x_*| \leq M|x_k - x_*|^2.$$

Newton's method cont.

Under certain conditions, Newton's method enjoys **fast local convergence**. For simplicity, we only state and prove the case for \mathbb{R} .

Theorem 2.3. (Quadratic convergence of Newton's method)

Let $g \in C^2(\mathbb{R})$ and x_* satisfies $g(x_*) = 0$ and $g'(x_*) \neq 0$. Then there exists $\epsilon > 0$ so that if $|x_0 - x_*| < \epsilon$, then the Newton's iterate $x_{k+1} = x_k - g(x_k)/g'(x_k)$ is well defined and there exists $M > 0$ so that

$$|x_{k+1} - x_*| \leq M|x_k - x_*|^2.$$

Note:

- This means that if x_0 is initialized sufficiently close to a **nice** solution, the Newton's method is well defined and converges very fast: **roughly doubling the number of correct digits every iteration.**

Newton's method cont.

Under certain conditions, Newton's method enjoys **fast local convergence**. For simplicity, we only state and prove the case for \mathbb{R} .

Theorem 2.3. (Quadratic convergence of Newton's method)

Let $g \in C^2(\mathbb{R})$ and x_* satisfies $g(x_*) = 0$ and $g'(x_*) \neq 0$. Then there exists $\epsilon > 0$ so that if $|x_0 - x_*| < \epsilon$, then the Newton's iterate $x_{k+1} = x_k - g(x_k)/g'(x_k)$ is well defined and there exists $M > 0$ so that

$$|x_{k+1} - x_*| \leq M|x_k - x_*|^2.$$

0.1
0.1²
0.1⁴

Note:

- This means that if x_0 is initialized sufficiently close to a **nice** solution, the Newton's method is well defined and converges very fast: **roughly doubling the number of correct digits every iteration**.
- Using a more **delicate** analysis, one can replace " $g \in C^2(\mathbb{R})$ " by " $g \in C^1(\mathbb{R})$ and $\exists L > 0$ with $|g'(x) - g'(y)| \leq L|x - y|$ for all x and y close to x_* ".

Newton's method cont.

Proof of Theorem 2.3: Since $g'(x_*) \neq 0$, there exist $\epsilon_1 > 0$ and $\delta > 0$ so that $|g'(x)| > \delta$ whenever $|x - x_*| \leq \epsilon_1$. Moreover, since g'' is continuous, there exists τ so that $\tau \geq |g''(x)|$ for these x .

Now, for each such x , by [Taylor's theorem](#), there exists ξ_x between x_* and x so that

$$0 = g(x_*) = g(x) + g'(x)(x_* - x) + 0.5g''(\xi_x)(x_* - x)^2.$$

This means

$$x - \frac{g(x)}{g'(x)} - x_* = \frac{g''(\xi_x)}{2g'(x)}(x_* - x)^2.$$

Thus

$$\left| x - \frac{g(x)}{g'(x)} - x_* \right| \leq \frac{\tau}{2\delta} |x_* - x|^2.$$

Hence, if $|x_0 - x_*| < \min\{\epsilon_1, \frac{2\delta}{\tau}\} =: \epsilon$, an induction shows that $|x_k - x_*| \leq \epsilon_1$ for all k and the desired inequality holds with $M = \frac{\tau}{2\delta}$.

Newton's method cont.

Applying Newton's method to $g(x) = x^3 - 3$ starting at $x_0 = 1.5$:

$$x_{k+1} = x_k - \frac{x_k^3 - 3}{3x_k^2}.$$

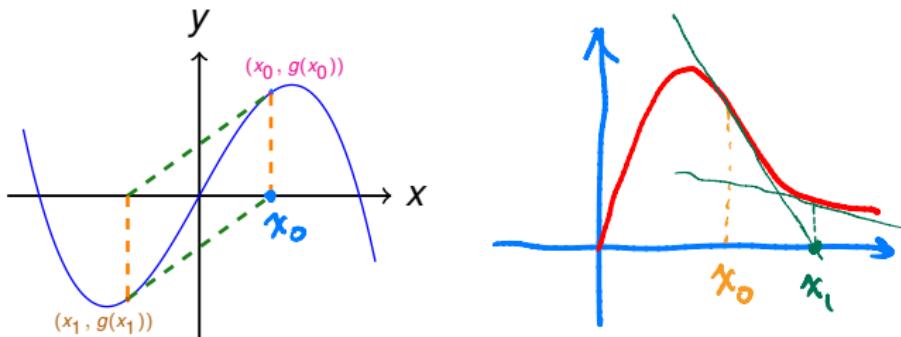
We have (in 10 s.f.)

x_1	1.444444444e+00
x_2	1.442252904e+00
x_3	1.442249570e+00
x_4	1.442249570e+00

Thus, $x_* = 1.4422$, rounded to 4 decimal places.

Failure of Newton's method

Failure of Newton's method: Besides failing when $g'(x_k) = 0$, if x^0 is too far away from x^* , Newton's method can also fail due to cycling:



Newton's method fails for $g(x) = x - x^3$, starting at $x_0 = \frac{1}{\sqrt{5}}$.

Steepest descent

- Instead of just solving for $\nabla f(x) = 0$, we take advantage of the **function values** of f .
- Since $f \in C^1(\mathbb{R}^n)$, we have

$$f(x + d) = f(x) + [\nabla f(x)]^T d + [\nabla f(\xi) - \nabla f(x)]^T d,$$

where $\xi \in \{x + td : t \in (0, 1)\}$.

Steepest descent

- Instead of just solving for $\nabla f(x) = 0$, we take advantage of the **function values** of f .
- Since $f \in C^1(\mathbb{R}^n)$, we have

$$f(x + d) = f(x) + [\nabla f(x)]^T d + [\nabla f(\xi) - \nabla f(x)]^T d,$$

where $\xi \in \{x + td : t \in (0, 1)\}$. Thus, if $\nabla f(x) \neq 0$ (i.e., x is not stationary) and we take $d = -\alpha \nabla f(x)$ for some $\alpha > 0$, then

$$f(x - \alpha \nabla f(x)) = f(x) - \alpha \|\nabla f(x)\|_2^2 - \underbrace{\alpha ([\nabla f(\xi) - \nabla f(x)]^T \nabla f(x))}_{\rightarrow 0 \text{ as } \alpha \rightarrow 0}.$$

Steepest descent

- Instead of just solving for $\nabla f(x) = 0$, we take advantage of the **function values** of f .
- Since $f \in C^1(\mathbb{R}^n)$, we have

$$f(x + d) = f(x) + [\nabla f(x)]^T d + [\nabla f(\xi) - \nabla f(x)]^T d,$$

where $\xi \in \{x + td : t \in (0, 1)\}$. Thus, if $\nabla f(x) \neq 0$ (i.e., x is not stationary) and we take $d = -\alpha \nabla f(x)$ for some $\alpha > 0$, then

$$f(x - \alpha \nabla f(x)) = f(x) - \alpha \|\nabla f(x)\|_2^2 - \underbrace{\alpha ([\nabla f(\xi) - \nabla f(x)]^T \nabla f(x))}_{\rightarrow 0 \text{ as } \alpha \rightarrow 0}.$$

Hence, for sufficiently small $\alpha > 0$, it holds that

$$\underline{f(x - \alpha \nabla f(x)) < f(x)}.$$



$$\nabla f(x^k - \alpha_k \nabla f(x^k))^T \nabla f(x^k) = 0$$

x^{k+1}

$$y(\alpha) = f(x - \alpha \nabla f(x)) \quad \text{Fix on } x$$

$$y'(\alpha_k) = 0$$

$$y(0) = f(x)$$

$$y'(x) = \nabla f(x - \alpha \nabla f(x))^T (-\nabla f(x))$$

$$\therefore y'(0) = -\nabla f(x)^T \nabla f(x) = -\|\nabla f(x)\|_2^2$$

If $\nabla f(x) \neq 0$, then

$$y'(0) < 0.$$

Steepest descent cont.

- $-\nabla f(x)$ is called the **steepest descent direction**.
- A natural **greedy** algorithm is

Steepest descent with exact line search

Start at $x^0 \in \mathbb{R}^n$. For each $k = 0, 1, 2, \dots$,

- ★ Set $d^k = -\nabla f(x^k)$.

- ★ Pick α_k so that

the set of minimizers

$$\alpha_k \in \operatorname{Arg} \min_{\alpha \geq 0} \{f(x^k + \alpha d^k) : \alpha \geq 0\}. \quad (1)$$

- ★ Set $x^{k+1} = x^k + \alpha_k d^k$.

Note: The update $x^{k+1} = x^k + \alpha_k d^k$ is prototypical in optimization.

- d^k is called the **search direction**. In the above algorithm, $d^k = -\nabla f(x^k)$. *learning rate*
- α_k is called the **step size**. In the above algorithm, it is chosen according to the **exact line search** criterion (1).

Steepest descent cont.

- In Steepest descent with exact line search, it is implicitly assumed that a minimizer α_k exists for the exact line search subproblem (1).

If α_k exists and $\nabla f(x^k) \neq 0$, then $\alpha_k > 0$. Why? Hence, we have

$$0 = \frac{d}{d\alpha} f(x^k + \alpha d^k) \Big|_{\alpha=\alpha_k} = (d^k)^T \nabla f(x^{k+1}) = -(\nabla f(x^k))^T \nabla f(x^{k+1}).$$

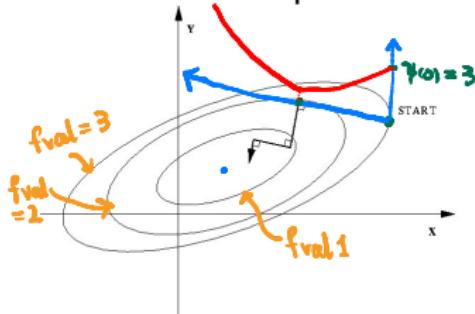
Steepest descent cont.

- In Steepest descent with exact line search, it is implicitly assumed that a minimizer α_k exists for the exact line search subproblem (1).
If α_k exists and $\nabla f(x^k) \neq 0$, then $\alpha_k > 0$. Why? Hence, we have

$$0 = \frac{d}{d\alpha} f(x^k + \alpha d^k) \Big|_{\alpha=\alpha_k} = (d^k)^T \nabla f(x^{k+1}) = -(\nabla f(x^k))^T \nabla f(x^{k+1}).$$

New direction \perp old direction: Creating zigzag path!

- Exact line search can be hard to perform.



Picture downloaded from <http://trond.hjorteland.com/thesis/node26.html>.

Armijo rule

In contrast to exact line search, usually **inexact line search** strategy is performed. One commonly used rule is:

Armijo rule:

Let $\sigma \in (0, 1)$, $x \in \mathbb{R}^n$ and $d \in \mathbb{R}^n$. Find $\alpha > 0$ so that

$$f(x + \alpha d) \leq f(x) + \alpha \sigma [\nabla f(x)]^T d.$$

Armijo rule

In contrast to exact line search, usually **inexact line search** strategy is performed. One commonly used rule is:

Armijo rule:

Let $\sigma \in (0, 1)$, $x \in \mathbb{R}^n$ and $d \in \mathbb{R}^n$. Find $\alpha > 0$ so that

$$f(x + \alpha d) \leq f(x) + \alpha\sigma[\nabla f(x)]^T d.$$

Definition. Let $f \in C^1(\mathbb{R}^n)$ and $x \in \mathbb{R}^n$. A $d \in \mathbb{R}^n$ is said to be a **descent direction** of f at x if

$$[\nabla f(x)]^T d < 0.$$

Armijo rule

In contrast to exact line search, usually **inexact line search** strategy is performed. One commonly used rule is:

Armijo rule:

Let $\sigma \in (0, 1)$, $x \in \mathbb{R}^n$ and $d \in \mathbb{R}^n$. Find $\alpha > 0$ so that

$$f(x + \alpha d) \leq f(x) + \alpha\sigma[\nabla f(x)]^T d.$$

Definition. Let $f \in C^1(\mathbb{R}^n)$ and $x \in \mathbb{R}^n$. A $d \in \mathbb{R}^n$ is said to be a **descent direction** of f at x if

$$[\nabla f(x)]^T \underbrace{d}_{< 0} < 0.$$

Examples: At an x that is **not stationary**, $\nabla f(x) \neq 0$

- $d = -\nabla f(x)$ is a descent direction; $\because \nabla f(x)^T (-\nabla f(x)) = -\|\nabla f(x)\|_2^2 < 0$
- More generally, if $D \succ 0$, then $d = -D\nabla f(x)$ is a descent direction.

$$\therefore \nabla f(x)^T (-D \nabla f(x)) = -\nabla f(x)^T D \nabla f(x) < 0$$

Armijo rule

In contrast to exact line search, usually **inexact line search** strategy is performed. One commonly used rule is:

Armijo rule:

Let $\sigma \in (0, 1)$, $x \in \mathbb{R}^n$ and $d \in \mathbb{R}^n$. Find $\alpha > 0$ so that

$$f(x + \alpha d) \leq f(x) + \alpha\sigma[\nabla f(x)]^T d.$$

Definition. Let $f \in C^1(\mathbb{R}^n)$ and $x \in \mathbb{R}^n$. A $d \in \mathbb{R}^n$ is said to be a **descent direction** of f at x if

$$[\nabla f(x)]^T d < 0.$$

Examples: At an x that is **not stationary**,

- $d = -\nabla f(x)$ is a descent direction;
- More generally, if $D \succ 0$, then $d = -D\nabla f(x)$ is a descent direction.

Is the Newton direction $-[\nabla^2 f(x)]^{-1}\nabla f(x)$ a descent direction?

Newton direction is in general not a descent direction.

Example

$$f(x) = \frac{1}{2}x^T \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix} x$$

$$\nabla^2 f(x) = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix} > 0$$

$$\begin{aligned} \therefore d &= -\nabla^2 f(x)^{-1} \nabla f(x) = -\begin{bmatrix} 1 & 0 \\ 0 & \frac{1}{2} \end{bmatrix} \nabla f(x) \\ &= \begin{bmatrix} 1 & 0 \\ 0 & \frac{1}{2} \end{bmatrix} \nabla f(x) \end{aligned}$$

$$\therefore \nabla f(x)^T d = \nabla f(x)^T \begin{bmatrix} 1 & 0 \\ 0 & \frac{1}{2} \end{bmatrix} \nabla f(x) > 0$$

when $\nabla f(x) \neq 0$

Armijo rule cont.

The next theorem shows that Armijo rule is not void.

Theorem 2.4:

Let $f \in C^1(\mathbb{R}^n)$, $x \in \mathbb{R}^n$, and $d \in \mathbb{R}^n$ be a **descent direction** at x .

Let $\sigma \in (0, 1)$. Then there exists $\alpha_1 > 0$ so that for all $\alpha \in [0, \alpha_1]$,

usually 10^{-4} $f(x + \alpha d) \leq f(x) + \alpha\sigma[\nabla f(x)]^T d$.

Armijo rule cont.

The next theorem shows that Armijo rule is not void.

Theorem 2.4:

Let $f \in C^1(\mathbb{R}^n)$, $x \in \mathbb{R}^n$, and $d \in \mathbb{R}^n$ be a descent direction at x .

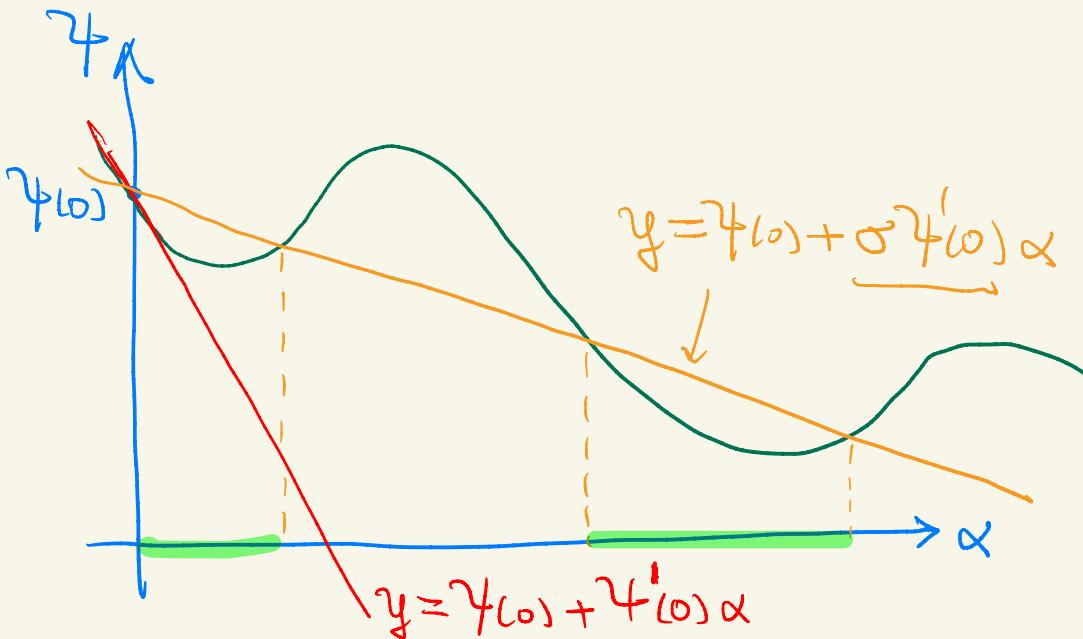
Let $\sigma \in (0, 1)$. Then there exists $\alpha_1 > 0$ so that for all $\alpha \in [0, \alpha_1]$,

$$f(x + \alpha d) \leq f(x) + \alpha \sigma [\nabla f(x)]^T d.$$

Proof: Since $f \in C^1(\mathbb{R}^n)$, we have for any $\alpha > 0$ that

$$\begin{aligned} f(x + \alpha d) &= f(x) + \alpha [\nabla f(x)]^T d + \alpha [\nabla f(\xi) - \nabla f(x)]^T d \\ &= f(x) + \sigma \alpha [\nabla f(x)]^T d + \alpha \left\{ (1 - \sigma) [\nabla f(x)]^T d + [\nabla f(\xi) - \nabla f(x)]^T d \right\}, \end{aligned}$$

where $\xi \in \{x + \alpha t d : t \in (0, 1)\}$. Since $(1 - \sigma) [\nabla f(x)]^T d < 0$ and $\lim_{\alpha \downarrow 0} [\nabla f(\xi) - \nabla f(x)]^T d = 0$, the green part is negative for all sufficiently small $\alpha > 0$.



$$y(\alpha) = f(x + \alpha d) \text{ and } \frac{\nabla f(x)^T d < 0}{\text{descent direction.}}$$

$$y(0) = f(x)$$

$$y'(\alpha) = \nabla f(x + \alpha d)^T d$$

$$y'(0) = \nabla f(x)^T d < 0$$

$$f(x + \alpha d) \leq f(x) + \sigma \nabla f(x)^T d \cdot \alpha$$

$$y(\alpha) \quad \boxed{y(0) + \sigma y'(0) \alpha}$$

Armijo rule cont.

How to execute Armijo rule in practice?

Armijo line search by backtracking:

Fix $\sigma \in (0, 1)$ and $\beta \in (0, 1)$. Given $x \in \mathbb{R}^n$, $d \in \mathbb{R}^n$ and $\bar{\alpha} > 0$.

Find the smallest nonnegative integer $j = j_0$ so that

$$f(x + \bar{\alpha}\beta^j d) \leq f(x) + \bar{\alpha}\beta^j \sigma [\nabla f(x)]^T d. \quad (2)$$

The stepsize generated is then $\bar{\alpha}\beta^{j_0}$.

β to the power j

Note:

- According to Theorem 2.4, if d is a descent direction, then (2) is satisfied for all sufficiently large j .
- In practice, one tests the validity of (2) for $j = 0, 1, 2, \dots$ **successively**. This is called **backtracking** because the stepsize $\bar{\alpha}\beta^j$ being tested keeps decreasing.
- The choice of $\bar{\alpha}$ is crucial for the efficiency of such scheme.

Convergence under Armijo rule

Theorem 2.5:

Let $f \in C^1(\mathbb{R}^n)$ with $\inf f > -\infty$. Let $\{\bar{\alpha}_k\} \subset \mathbb{R}$ satisfy $0 < \inf_k \bar{\alpha}_k \leq \sup_k \bar{\alpha}_k < \infty$, and fix $\sigma \in (0, 1)$ and $\beta \in (0, 1)$. Suppose $\{x^k\}$ is generated as

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

where

descent direction

- $d^k := -D_k \nabla f(x^k)$; here $\{D_k\}$ is a bounded sequence of positive definite matrices with $D_k - \delta I \succeq 0$ for some $\delta > 0$; indep of k.
- α_k is generated via the Armijo line search by backtracking with $x = x^k$, $d = d^k$ and $\bar{\alpha} = \bar{\alpha}_k$, and σ and β defined above.

Then any accumulation point of $\{x^k\}$ is a stationary point of f .

Any convergent subsequence $\lambda_{\min}(D_k) \geq \delta > 0$
(if exists) will converge

to a stationary point of f .

Convergence under Armijo rule

Theorem 2.5:

Let $f \in C^1(\mathbb{R}^n)$ with $\inf f > -\infty$. Let $\{\bar{\alpha}_k\} \subset \mathbb{R}$ satisfy $0 < \inf_k \bar{\alpha}_k \leq \sup_k \bar{\alpha}_k < \infty$, and fix $\sigma \in (0, 1)$ and $\beta \in (0, 1)$. Suppose $\{x^k\}$ is generated as

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

where

- $d^k := -D_k \nabla f(x^k)$; here $\{D_k\}$ is a **bounded** sequence of **positive definite matrices** with $D_k - \delta I \succeq 0$ for some $\delta > 0$;
- α_k is generated via the **Armijo line search by backtracking** with $x = x^k$, $d = d^k$ and $\bar{\alpha} = \bar{\alpha}_k$, and σ and β defined above.

Then any **accumulation point** of $\{x^k\}$ is a stationary point of f .

Remark:

- If x^k is non-stationary, then d^k is a descent direction.
- The condition $D_k - \delta I \succeq 0$ implies that for any $y \in \mathbb{R}^n$, we have $y^T(D_k - \delta I)y \geq 0$. Hence $y^T D_k y \geq \delta \|y\|_2^2$. $\Rightarrow \lambda_{\min}(D_k) \geq \delta$

Convergence under Armijo rule cont.

Proof sketch of Theorem 2.5: If x^k is a stationary point for some finite $k \geq 0$, then $x^l \equiv x^k$ whenever $l \geq k$ and we are done.

Assume that x^k is not stationary for each k . Then according to [Armijo line search by backtracking](#), we have $\alpha_k > 0$ for all k and

$$f(x^{k+1}) \leq f(x^k) + \sigma \alpha_k [\nabla f(x^k)]^T d^k.$$

Note that $[\nabla f(x^k)]^T d^k < 0$ for each k . Rearranging terms and summing from $k = 0$ to ∞ , we have

$$0 \leq -\sigma \sum_{k=0}^{\infty} \alpha_k [\nabla f(x^k)]^T d^k \leq f(x^0) - \inf f < \infty.$$

Thus,

$$\lim_{k \rightarrow \infty} \alpha_k [\nabla f(x^k)]^T d^k = 0. \quad (3)$$

Convergence under Armijo rule cont.

Proof sketch of Theorem 2.5 cont.: Let \bar{x} be an accumulation point of $\{x^k\}$. By definition, there is a subsequence $\{x^{k_i}\}$ with $\lim_{i \rightarrow \infty} x^{k_i} = \bar{x}$.

If $\liminf_{i \rightarrow \infty} \alpha_{k_i} > 0$, then (3) implies

$$\lim_{i \rightarrow \infty} [\nabla f(x^{k_i})]^T d^{k_i} = 0.$$

Recall that $d^k = -D_k \nabla f(x^k)$ for some bounded sequence $\{D_k\}$. By passing to a further subsequence if necessary, we may assume that

(Bolzano-Weierstrass theorem is invoked)

$$\lim_{i \rightarrow \infty} D_{k_i} = D_*$$

for some matrix D_* . This implies

$$\begin{aligned} 0 &= -\lim_{i \rightarrow \infty} [\nabla f(x^{k_i})]^T D_{k_i} \nabla f(x^{k_i}) = -[\nabla f(\bar{x})]^T D_* \nabla f(\bar{x}) \\ &= -\lim_{i \rightarrow \infty} [\nabla f(\bar{x})]^T D_{k_i} \nabla f(\bar{x}) \leq -\delta \|\nabla f(\bar{x})\|_2^2. \end{aligned}$$

Thus, we have $\nabla f(\bar{x}) = 0$ as desired.

Convergence under Armijo rule cont.

Proof sketch of Theorem 2.5 cont.: Now it remains to consider the case that $\liminf_{i \rightarrow \infty} \alpha_{k_i} = 0$.

By passing to a further subsequence if necessary, we may assume that $\lim_{i \rightarrow \infty} \alpha_{k_i} = 0$.

Since $\inf_k \bar{\alpha}_k > 0$ and $\lim_{i \rightarrow \infty} \alpha_{k_i} = 0$, the Armijo line search by backtracking must have been invoked when i is sufficiently large.

Then for all large i

$$f(x^{k_i} + [\alpha_{k_i}/\beta]d^{k_i}) > f(x^{k_i}) + \sigma(\alpha_{k_i}/\beta)[\nabla f(x^{k_i})]^T d^{k_i}.$$

Convergence under Armijo rule cont.

Proof sketch of Theorem 2.5 cont.: Then

$$\frac{f(x^{k_i} + [\alpha_{k_i}/\beta]d^{k_i}) - f(x^{k_i})}{(\alpha_{k_i}/\beta)} > \sigma[\nabla f(x^{k_i})]^T d^{k_i}. \quad (4)$$

Recall that $d^k = -D_k \nabla f(x^k)$ for some bounded sequence $\{D_k\}$. By passing to a further subsequence if necessary, we may assume that

(Bolzano-Weierstrass theorem is invoked)

$$\lim_{i \rightarrow \infty} d^{k_i} = -D_* \nabla f(\bar{x}) =: d^*$$

for some $D_* := \lim_{i \rightarrow \infty} D_{k_i}$. Passing to the limit in (4), we have $[\nabla f(\bar{x})]^T d^* \geq \sigma[\nabla f(\bar{x})]^T d^*$. Since $\sigma \in (0, 1)$, this implies

$$\begin{aligned} 0 &\leq [\nabla f(\bar{x})]^T d^* = -[\nabla f(\bar{x})]^T D_* \nabla f(\bar{x}) \\ &= -\lim_{i \rightarrow \infty} [\nabla f(\bar{x})]^T D_{k_i} \nabla f(\bar{x}) \leq -\delta \|\nabla f(\bar{x})\|_2^2. \end{aligned}$$

Hence, $\nabla f(\bar{x}) = 0$ also in the case that $\liminf_{i \rightarrow \infty} \alpha_{k_i} = 0$.

Convergence under Armijo rule cont.

Some remarks on parameters:

- σ is chosen to be **small** so that (2) may be satisfied with a small number of backtracking steps: note that each backtracking requires an evaluation of $f(x^k + \alpha d^k)$, which adds to the **main computational cost**. A typical choice is $\sigma = 10^{-4}$.
- β is typically $\frac{1}{2}$.
- The choice of $\{\bar{\alpha}_k\}$ is **crucial**. Ideally, it should be chosen so that (2) may be satisfied with a small number of backtracking steps.

Possible choices are:

- $\star \bar{\alpha}_k \equiv 1$ for “Newton-like” directions.
 - $\star \bar{\alpha}_k = \max\{u, \min\{\ell, \alpha_{k-1}\}\}$, where u and ℓ are positive.
 - \star (Projected) **Barzilai-Borwein stepsize** (see the next lecture).
- One can terminate when $\|\nabla f(x^k)\|_2 \leq tol \cdot \max\{|f(x^k)|, 1\}$, i.e., when the gradient is small relative to the function value.
-
- The handwritten notes explain the third choice of $\bar{\alpha}_k$: "project α_{k-1} onto $[\ell, u]$ ". Two blue arrows point from the numbers 10^{-8} to the endpoints u and ℓ of the interval $[\ell, u]$.

Special case

$$\nabla f(x)$$

Find w

Corollary 2.1: (Steepest descent with constant stepsize)

Let $f \in C^2(\mathbb{R}^n)$ with $\inf f > -\infty$. Suppose that there exists $L > 0$ so that

$$L \geq \|\nabla^2 f(x)\|_2 \text{ for all } x.$$

Fix any $\gamma \in (0, 2)$ and consider the sequence generated as

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

Then any accumulation point of $\{x^k\}$ is a stationary point of f .

$$\text{Step size } \in \left(0, \frac{2}{L}\right)$$

Special case

Corollary 2.1: (Steepest descent with constant stepsize)

Let $f \in C^2(\mathbb{R}^n)$ with $\inf f > -\infty$. Suppose that there exists $L > 0$ so that

$$L \geq \|\nabla^2 f(x)\|_2 \text{ for all } x.$$

Fix any $\gamma \in (0, 2)$ and consider the sequence generated as

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

Then any **accumulation point** of $\{x^k\}$ is a stationary point of f .

Remark:

- Given L , the above algorithm can be written in one line.
- While the algorithm avoids line search (which can be costly), it can be potentially slow because the constant stepsize can be **too conservative** in making progress.

Special case cont.

Proof of Corollary 2.1: It suffices to show that if one sets $\sigma = 1 - \frac{\gamma}{2} \in (0, 1)$, $D_k = \frac{\gamma}{L} I$ and $\bar{\alpha}_k \equiv 1$ in [Theorem 2.5](#), then backtracking is not invoked in (2).

To this end, note that for each x , with $d := -\frac{\gamma}{L} \nabla f(x)$, there exists ξ such that

$$\begin{aligned}f(x + d) &= f(x) + \nabla f(x)^T d + \frac{1}{2} d^T \nabla^2 f(\xi) d \\&\leq f(x) + \nabla f(x)^T d + \frac{1}{2} \|d\|_2 \|\nabla^2 f(\xi) d\|_2 \\&\leq f(x) + \nabla f(x)^T d + \frac{1}{2} \|d\|_2 \|\nabla^2 f(\xi)\|_2 \|d\|_2 \\&\leq f(x) + \nabla f(x)^T d + \frac{L}{2} \|d\|_2^2 \\&= f(x) + \nabla f(x)^T d - \frac{\gamma}{2} \nabla f(x)^T d \\&= f(x) + (1 - \frac{\gamma}{2}) \nabla f(x)^T d \\&= f(x) + \sigma \nabla f(x)^T d.\end{aligned}$$

This shows that the Armijo rule is satisfied with $\alpha = 1$.

Example

$$\nabla f(x) = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \end{bmatrix} = \begin{bmatrix} 2x_1 + 3x_2 \\ 3x_1 + 16x_2 \end{bmatrix}$$

Example: Let $f(x_1, x_2) = x_1^2 + 3x_1x_2 + 8x_2^2$.

- Show that $\|\nabla^2 f(x)\|_2 \leq 18$ for all x .
- Write down the general update formula and the first 2 iterations of the steepest descent with constant stepsize, starting at $(x_1, x_2) = (0, 1)$ and using $\gamma = 0.9$.

$$\nabla^2 f(x) = \begin{bmatrix} 2 & 3 \\ 3 & 16 \end{bmatrix}$$

Example

Example: Let $f(x_1, x_2) = x_1^2 + 3x_1x_2 + 8x_2^2$.

- Show that $\|\nabla^2 f(x)\|_2 \leq 18$ for all x .
- Write down the general update formula and the first 2 iterations of the steepest descent with constant stepsize, starting at $(x_1, x_2) = (0, 1)$ and using $\gamma = 0.9$.



Remarks: For a symmetric matrix A , it holds that

$$\|A\|_2 = \max\{|\lambda_{\max}(A)|, |\lambda_{\min}(A)|\}.$$

$$\sqrt{\lambda_{\max}(A^T A)} = \sqrt{\lambda_{\max}(A^2)}$$

e.g.
 $\{3, 1, -4\}$
are eigenvalues
of A .

square
 $\Rightarrow \{9, 1, 16\}$
sort
 $\Rightarrow \{3, 1, 4\}$

Example cont.

$$D = \begin{vmatrix} 2 & 3 \\ 3 & 16 \end{vmatrix}$$

Solution:

$$\nabla f(x) = \begin{bmatrix} 2x_1 + 3x_2 \\ 3x_1 + 16x_2 \end{bmatrix}, \quad \nabla^2 f(x) = \begin{bmatrix} 2 & 3 \\ 3 & 16 \end{bmatrix}.$$

The eigenvalues of $\nabla^2 f(x)$ are $9 \pm \sqrt{58}$. Hence

$$\|\nabla^2 f(x)\|_2 = 9 + \sqrt{58} \approx 16.62 < 18.$$

The iterative scheme is given by

$$x^{k+1} = \underbrace{x^k}_{\text{---}} - \frac{0.9}{18} \begin{bmatrix} 2x_1^k + 3x_2^k \\ 3x_1^k + 16x_2^k \end{bmatrix}.$$

Hence,

$$x^1 = \begin{bmatrix} 0 \\ 1 \end{bmatrix} - 0.05 \begin{bmatrix} 3 \\ 16 \end{bmatrix} = \begin{bmatrix} -0.15 \\ 0.2 \end{bmatrix},$$

$$x^2 = \begin{bmatrix} -0.15 \\ 0.2 \end{bmatrix} - 0.05 \begin{bmatrix} -0.3 + 0.6 \\ -0.45 + 3.2 \end{bmatrix} = \begin{bmatrix} -0.165 \\ 0.0625 \end{bmatrix}.$$

A chain rule

Let $h \in C^2(\mathbb{R}^m)$ and let $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$.

Define $f(x) := h(Ax - b)$. Then $f \in C^2(\mathbb{R}^n)$ and

$$\nabla f(x) = \underline{A^T \nabla h(Ax - b)} \text{ and } \nabla^2 f(x) = \underline{A^T \nabla^2 h(Ax - b) A}.$$

$$f(x+u) = h(\underline{\underline{Ax}} + \underline{\underline{Au}} - \underline{\underline{b}})$$

$$= h(Ax - b) + \nabla h(Ax - b)^T \underline{\underline{Au}} + \frac{1}{2} (\underline{\underline{Au}})^T \nabla^2 h(\underline{\underline{\xi}}) \underline{\underline{Au}}$$

$$\therefore \nabla f(x) = (\nabla h(Ax - u)^T A)^T$$

$$= A^T \nabla h(Ax - u)$$



A chain rule

Let $h \in C^2(\mathbb{R}^m)$ and let $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$.

Define $f(x) := h(Ax - b)$. Then $f \in C^2(\mathbb{R}^n)$ and

$$\nabla f(x) = A^T \nabla h(Ax - b) \text{ and } \nabla^2 f(x) = A^T \nabla^2 h(Ax - b) A.$$

In particular, if there exists L such that $L \geq \|\nabla^2 h(y)\|_2$ for all y , then

$$\begin{aligned}\|\nabla^2 f(x)\|_2 &\leq \|A^T\|_2 \|\nabla^2 h(Ax - b)\|_2 \|A\|_2 \\ &\leq L \|A^T\|_2 \|A\|_2 = L \lambda_{\max}(A^T A).\end{aligned}$$

$$\sqrt{\lambda_{\max}(AA^T)} = \sqrt{\lambda_{\max}(A^T A)}$$

Why the chain rule?

Sometimes $\|\nabla^2 h\|_2$ may be easier to estimate

e.g. $h(y) = \ln(1+y_1^2) + \ln(1+y_2^2)$

$$\nabla h(y) = \begin{bmatrix} \frac{\partial h}{\partial y_1} \\ \frac{\partial h}{\partial y_2} \end{bmatrix} = \begin{bmatrix} \frac{2y_1}{1+y_1^2} \\ \frac{2y_2}{1+y_2^2} \end{bmatrix}$$

$$\nabla^2 h(y) = \begin{bmatrix} S+h & \textcircled{1} \\ \textcircled{2} & S+h \text{ else} \end{bmatrix}$$

Diagonal \Rightarrow eigenvalues are just the diagonal entries.

i. Examine the diagonal to get $\|\nabla^2 h(y)\|_2$

Example

Example: Let $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$ and

$$h(y) = \sum_{i=1}^m \ln(1 + y_i^2).$$

Fix any $\mu > 0$ and consider the problem of minimizing

$$f(x) := h(Ax - b) + \frac{\mu}{2} \|x\|_2^2.$$

Discuss how the parameters can be chosen for implementing
steepest descent with constant stepsize.

Remark: The function h is related to **Cauchy** measurement noise in b ,
and $\mu > 0$ is a tuning parameter for ridge regression.

Example cont.

Solution: We first compute an upper estimate of $\|\nabla^2 h(y)\|_2$. Notice that $\nabla^2 h(y)$ is a diagonal matrix with the i th diagonal entry given by

$$\frac{\partial^2}{\partial y_i^2} \ln(1 + y_i^2) = \frac{\partial}{\partial y_i} \left(\frac{2y_i}{y_i^2 + 1} \right) = \frac{2(1 - y_i^2)}{(y_i^2 + 1)^2}$$

Thus,

$$|(\nabla^2 h(y))_{ii}| \leq \frac{2(1 + y_i^2)}{(y_i^2 + 1)^2} = \frac{2}{y_i^2 + 1} \leq 2.$$

Hence, $\|\nabla^2 h(y)\|_2 \leq 2$. Since $\nabla^2 f(x) = A^T \nabla^2 h(Ax - b)A + \mu I$, it follows that

$$\|\nabla^2 f(x)\|_2 \leq 2\lambda_{\max}(A^T A) + \mu.$$

Consequently, we can take $L = 2\lambda_{\max}(A^T A) + \mu$ and any $\gamma \in (0, 2)$ in the algorithm.

Example cont.

Remark on computation cost: Using flop counts, we can make the following observations.

- Note that computing Ax requires $m(2n - 1)$ flops. This is the dominant computation in computing f .
- Since $\nabla f(x) = A^T \nabla h(Ax - b)$, computing ∇f requires computing one Au and $A^T v$. The former can be saved during the computation of $f(x)$, the latter requires another $n(2m - 1)$ flops.
- Thus, if the algorithm in Theorem 2.5 is used with $D_k \equiv I$ and if no backtracking is invoked in the line search, then each iteration requires computing one Au and $A^T v$. Any additional backtrack step requires recomputing $f(x + \alpha d)$.