MATH 4211/6211 – Optimization

Gradient Method

Xiaojing Ye
Department of Mathematics & Statistics
Georgia State University

Consider $\boldsymbol{x}^{(k)}$ and compute $\boldsymbol{g}^{(k)} := \nabla f(\boldsymbol{x}^{(k)})$. Set descent direction to $\boldsymbol{d}^{(k)} = -\boldsymbol{g}^{(k)}$.

Now we want to find $\alpha \geq 0$ such that $\boldsymbol{x}^{(k)} - \alpha \boldsymbol{g}^{(k)}$ improves $\boldsymbol{x}^{(k)}$.

Define $\phi(\alpha) := f(\boldsymbol{x}^{(k)} - \alpha \boldsymbol{g}^{(k)})$, then ϕ has Taylor expansion:

$$f(\boldsymbol{x}^{(k)} - \alpha \boldsymbol{g}^{(k)}) = f(\boldsymbol{x}^{(k)}) - \alpha \|\boldsymbol{g}^{(k)}\|^2 + o(\alpha)$$

For α sufficiently small, we have

$$f(\boldsymbol{x}^{(k)} - \alpha \boldsymbol{g}^{(k)}) \leq f(\boldsymbol{x}^{(k)})$$

Gradient Descent Method (or Gradient Method):

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} - \alpha_k \boldsymbol{g}^{(k)}$$

Set an initial guess $\boldsymbol{x}^{(0)}$, and iterate the scheme above to obtain $\{\boldsymbol{x}^{(k)} : k = 0, 1, \dots\}$.

- $\boldsymbol{x}^{(k)}$: current estimate;
- $\boldsymbol{g}^{(k)} := \nabla f(\boldsymbol{x}^{(k)})$: gradient at $\boldsymbol{x}^{(k)}$;
- $\alpha_k \geq 0$: step size.

Steepest Descent Method: choose α_k such that

$$\alpha_k = \arg\min_{\alpha \geq 0} f(\boldsymbol{x}^{(k)} - \alpha \boldsymbol{g}^{(k)})$$

Steepest descent method is an *exact line search* method.

We will first discuss some properties of steepest descent method, and consider other (inexact) line search methods.

Proposition. Let $\{\boldsymbol{x}^{(k)}\}$ be obtained by steepest descent method, then

$$(\boldsymbol{x}^{(k+2)} - \boldsymbol{x}^{(k+1)})^\top (\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)}) = 0$$

Proof. Define $\phi(\alpha) := f(\boldsymbol{x}^{(k)} - \alpha \boldsymbol{g}^{(k)})$. Since $\alpha_k = \arg\min \phi(\alpha)$, we have

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

On the other hand, we have

$$\begin{aligned} \boldsymbol{x}^{(k+2)} &= \boldsymbol{x}^{(k+1)} - \alpha_{k+1} \boldsymbol{g}^{(k+1)} \\ \boldsymbol{x}^{(k+1)} &= \boldsymbol{x}^{(k)} - \alpha_k \boldsymbol{g}^{(k)} \end{aligned}$$

Therefore, we have

$$(\boldsymbol{x}^{(k+2)} - \boldsymbol{x}^{(k+1)})^\top (\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)}) = \alpha_{k+1} \alpha_k \boldsymbol{g}^{(k+1)^\top} \boldsymbol{g}^{(k)} = 0.$$

Proposition. Let $\{\boldsymbol{x}^{(k)}\}$ be obtained by steepest descent method and $\boldsymbol{g}^{(k)} \neq \mathbf{0}$, then $f(\boldsymbol{x}^{(k+1)}) < f(\boldsymbol{x}^{(k)})$

Proof. Define $\phi(\alpha) := f(\boldsymbol{x}^{(k)} - \alpha \boldsymbol{g}^{(k)})$. Then

$$\phi'(0) = -\nabla f(\boldsymbol{x}^{(k)} - 0\boldsymbol{g}^{(k)})^\top \boldsymbol{g}^{(k)} = -\|\boldsymbol{g}^{(k)}\|^2 < 0.$$

Since α_k is a minimizer, there is

$$f(\boldsymbol{x}^{(k+1)}) = \phi(\alpha_k) < \phi(0) = f(\boldsymbol{x}^{(k)}).$$

Stopping Criterion.

For a prescribed $\epsilon > 0$, terminate the iteration if one of the followings is met:

- $\|\boldsymbol{g}^{(k)}\| < \epsilon$;
- $|f(\boldsymbol{x}^{(k+1)}) - f(\boldsymbol{x}^{(k)})| < \epsilon$;
- $\|\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)}\| < \epsilon$.

More preferable choices using "relative change":

- $|f(\boldsymbol{x}^{(k+1)}) - f(\boldsymbol{x}^{(k)})|/|f(\boldsymbol{x}^{(k)})| < \epsilon$;
- $\|\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)}\|/\|\boldsymbol{x}^{(k)}\| < \epsilon$.

Example. Use steepest descent method for 3 iterations on

$$f(x_1, x_2, x_3) = (x_1 - 4)^4 + (x_2 - 3)^2 + 4(x_3 + 5)^4$$

with initial point $\boldsymbol{x}^{(0)} = [4, 2, -1]^\top$.

Solution. We will repeatedly use the gradient, so let's compute it first:

$$\nabla f(\boldsymbol{x}) = \begin{bmatrix} 4(x_1 - 4)^3 \\ 2(x_2 - 3) \\ 16(x_3 + 5)^3 \end{bmatrix}$$

We keep in mind that $\boldsymbol{x}^* = [4, 3, -5]^\top$.

In the **1st iteration**:

- Current iterate: $\boldsymbol{x}^{(0)} = [4, 2, -1]^\top$;

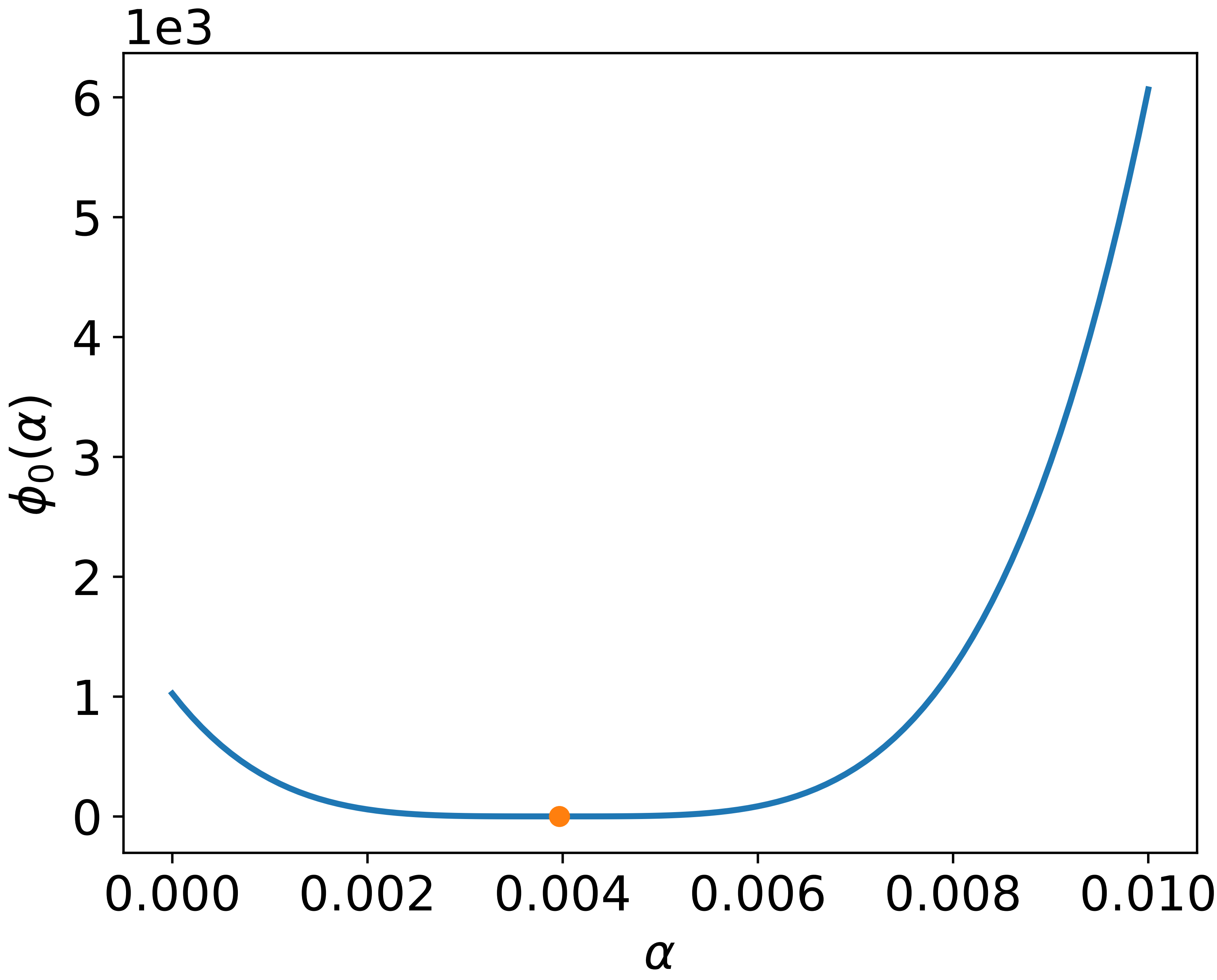
- Current gradient: $\boldsymbol{g}^{(0)} = \nabla f(\boldsymbol{x}^{(0)}) = [0, -2, 1024]^\top$;

- Find step size:

$$\begin{aligned}\alpha_0 &= \arg\min_{\alpha \geq 0} f(\boldsymbol{x}^{(0)} - \alpha \boldsymbol{g}^{(0)}) \\ &= \arg\min_{\alpha \geq 0} \Big(0 + (2 + 2\alpha - 3)^2 + 4(-1 - 1024\alpha + 5)^4\Big)\end{aligned}$$

  and use secant method to get $\alpha_0 = 3.967 \times 10^{-3}$.

- Next iterate: $\boldsymbol{x}^{(1)} = \boldsymbol{x}^{(0)} - \alpha_0 \boldsymbol{g}^{(0)} = \cdots = [4.000, 2.008, -5.062]^\top$.



1e3
6
5
4
3
2
1
0
$\phi_0(\alpha)$
0.000
0.002
0.004
0.006
0.008
0.010
α

In the **2nd iteration**:

- Current iterate: $\boldsymbol{x}^{(1)} = [4.000, 2.008, -5.062]^\top$;

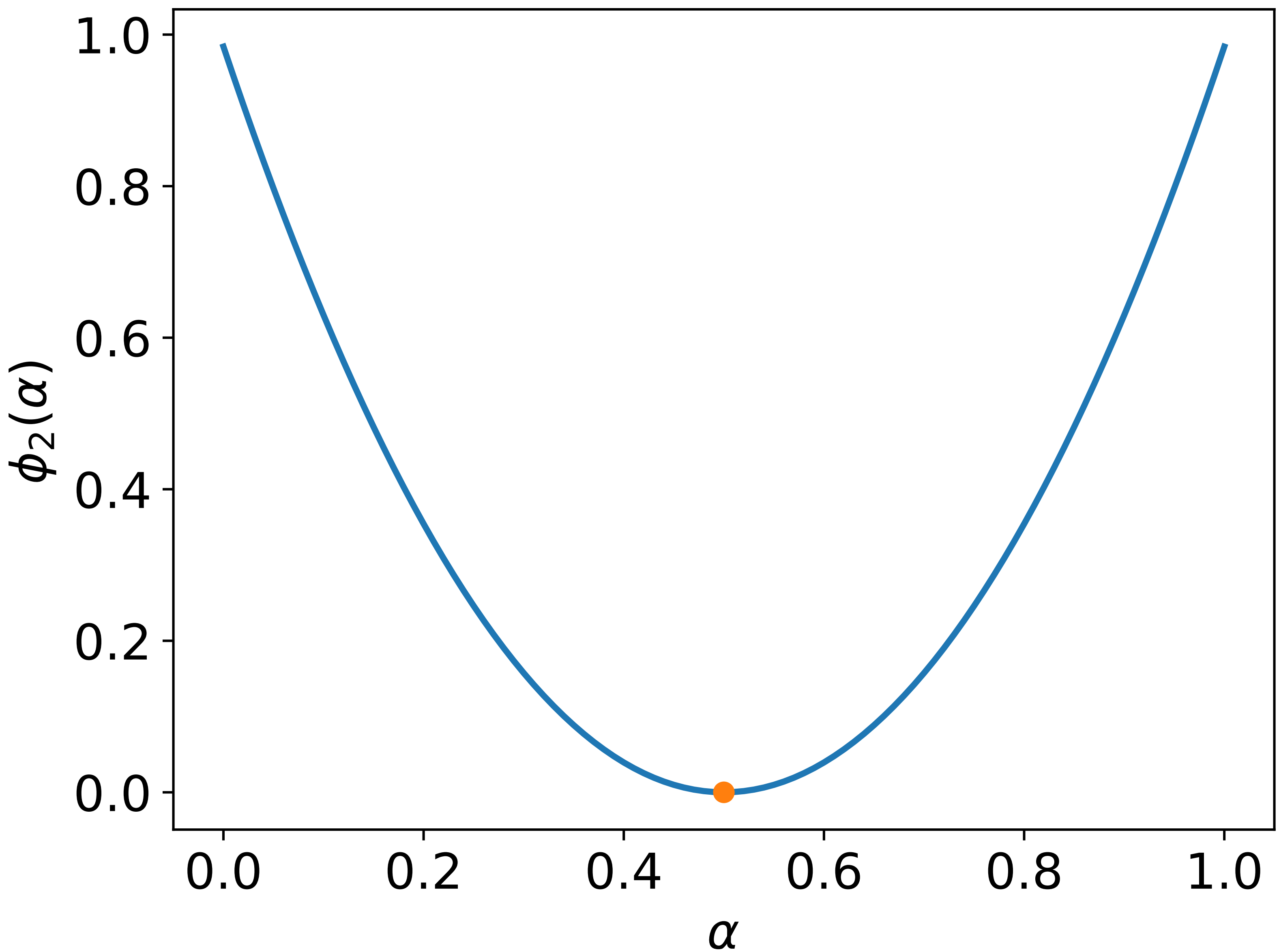
- Current gradient: $\boldsymbol{g}^{(1)} = \nabla f(\boldsymbol{x}^{(1)}) = [0.001, -1.984, -0.003875]^\top$;

- Find step size:

$$\begin{aligned}\alpha_1 &= \arg\min_{\alpha \geq 0} f(\boldsymbol{x}^{(1)} - \alpha \boldsymbol{g}^{(1)}) \\ &= \arg\min_{\alpha \geq 0} \Big(0 + (2.008 + 1.984\alpha - 3)^2 + 4(-5.062 + 0.003875\alpha + 5)^4\Big)\end{aligned}$$

  and use secant method to get $\alpha_1 = 0.500$.

- Next iterate: $\boldsymbol{x}^{(2)} = \boldsymbol{x}^{(1)} - \alpha_1 \boldsymbol{g}^{(1)} = \cdots = [4.000, 3.000, -5.060]^\top$.


1.0
0.8
0.6
0.4
0.2
0.0
$\phi_2(\alpha)$
0.0
0.2
0.4
0.6
0.8
1.0
α

In the **3rd iteration**:

- Current iterate: $\boldsymbol{x}^{(2)} = [4.000, 3.000, -5.060]^\top$;

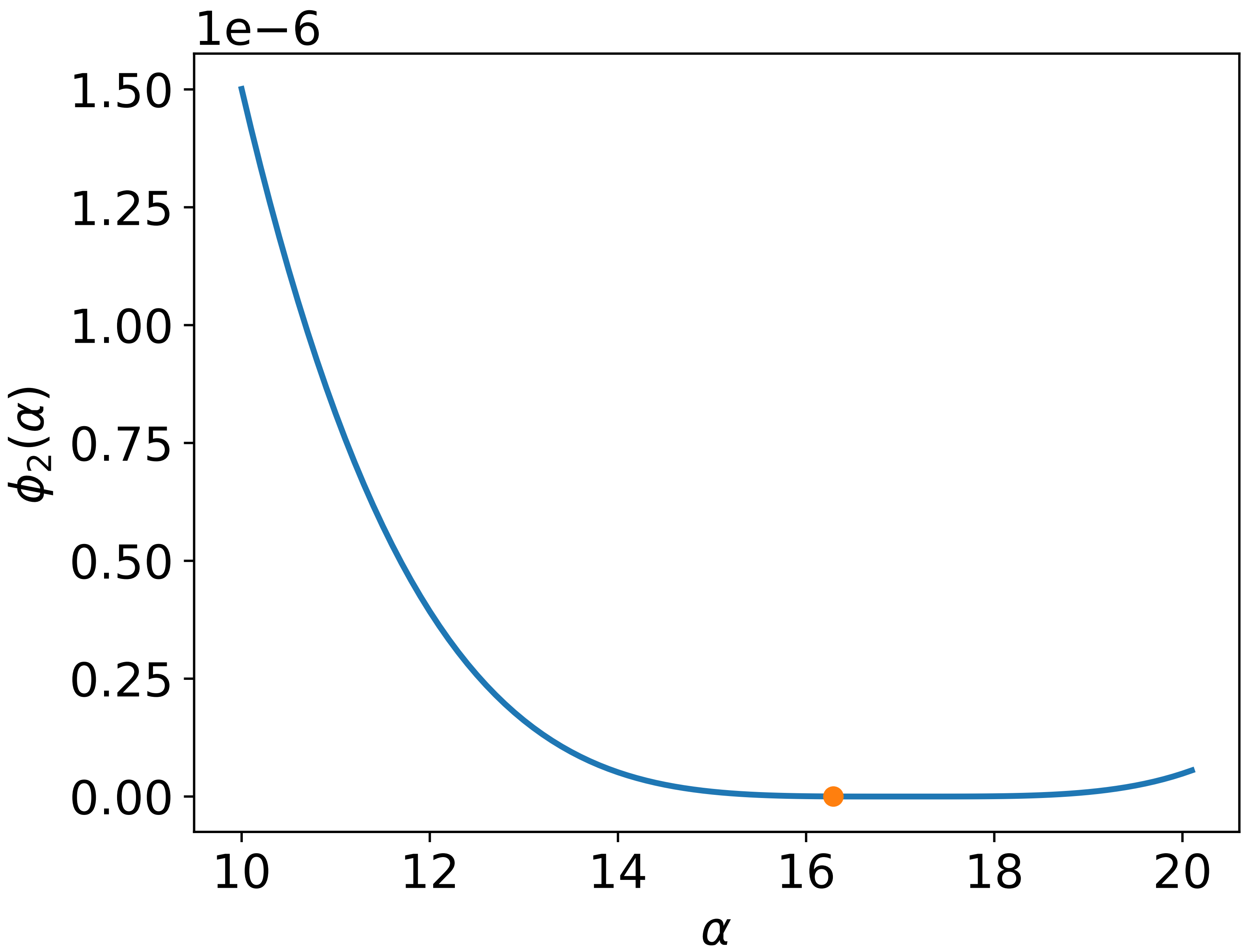
- Current gradient: $\boldsymbol{g}^{(2)} = \nabla f(\boldsymbol{x}^{(2)}) = [0.000, 0.000, -0.003525]^\top$;

- Find step size:

$$\begin{aligned} \alpha_2 &= \arg\min_{\alpha \ge 0} f(\boldsymbol{x}^{(2)} - \alpha \boldsymbol{g}^{(2)}) \\ &= \arg\min_{\alpha \ge 0} \left(0 + 0 + 4(-5.060 + 0.003525\alpha + 5)^4\right) \end{aligned}$$

  and use secant method to get $\alpha_2 = 16.29$.

- Next iterate: $\boldsymbol{x}^{(3)} = \boldsymbol{x}^{(2)} - \alpha_2 \boldsymbol{g}^{(2)} = \cdots = [4.000, 3.000, -5.002]^\top$.


1e−6
1.50
1.25
1.00
0.75
0.50
0.25
0.00
$\phi_2(\alpha)$
10
12
14
16
18
20
α

A quadratic function f of x can be written as

$$f(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}^\top \boldsymbol{x}$$

where $\boldsymbol{A}$ is not necessarily symmetric.

Note that $\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x} = \boldsymbol{x}^\top \boldsymbol{A}^\top \boldsymbol{x}$ and hence $\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x} = \frac{1}{2}\boldsymbol{x}^\top (\boldsymbol{A} + \boldsymbol{A}^\top)\boldsymbol{x}$ where $\boldsymbol{A} + \boldsymbol{A}^\top$ is symmetric.

Therefore, a quadratic function can always be rewritten as

$$f(x) = \frac{1}{2}\boldsymbol{x}^\top \boldsymbol{Q}\boldsymbol{x} - \boldsymbol{b}^\top \boldsymbol{x}$$

where $\boldsymbol{Q}$ is symmetric. In this case, the gradient and Hessian are:

$$\nabla f(\boldsymbol{x}) = \boldsymbol{Q}\boldsymbol{x} - \boldsymbol{b} \quad \text{and} \quad \nabla^2 f(\boldsymbol{x}) = \boldsymbol{Q}.$$

Now let's see what happens when we apply the steepest descent method to a quadratic function f:

$$f(x) = \frac{1}{2}\boldsymbol{x}^\top \boldsymbol{Q}\boldsymbol{x} - \boldsymbol{b}^\top \boldsymbol{x}$$

where $\boldsymbol{Q} \succ 0$.

At k-th iteration, we have $\boldsymbol{x}^{(k)}$ and $\boldsymbol{g}^{(k)} = \nabla f(\boldsymbol{x}^{(k)}) = \boldsymbol{Q}\boldsymbol{x}^{(k)} - \boldsymbol{b}$.

Then we need to find the step size $\alpha_k = \arg\min_\alpha \phi(\alpha)$ where

$$\phi(\alpha) := f(\boldsymbol{x}^{(k)} - \alpha \boldsymbol{g}^{(k)}) = \frac{1}{2}(\boldsymbol{x}^{(k)} - \alpha \boldsymbol{g}^{(k)})^\top \boldsymbol{Q}(\boldsymbol{x}^{(k)} - \alpha \boldsymbol{g}^{(k)}) - \boldsymbol{b}^\top(\boldsymbol{x}^{(k)} - \alpha \boldsymbol{g}^{(k)})$$

Solving $\phi'(\alpha) = -(\boldsymbol{x}^{(k)} - \alpha \boldsymbol{g}^{(k)})^\top \boldsymbol{Q}\boldsymbol{g}^{(k)} + \boldsymbol{b}^\top \boldsymbol{g}^{(k)} = 0$, we obtain

$$\alpha_k = \frac{(\boldsymbol{Q}\boldsymbol{x}^{(k)} - \boldsymbol{b})^\top \boldsymbol{g}^{(k)}}{\boldsymbol{g}^{(k)\top} \boldsymbol{Q}\boldsymbol{g}^{(k)}} = \frac{\boldsymbol{g}^{(k)\top} \boldsymbol{g}^{(k)}}{\boldsymbol{g}^{(k)\top} \boldsymbol{Q}\boldsymbol{g}^{(k)}}$$

Therefore, the steepest descent method applied to $f(x) = \frac{1}{2}\boldsymbol{x}^\top \boldsymbol{Q}\boldsymbol{x} - \boldsymbol{b}^\top \boldsymbol{x}$ with $\boldsymbol{Q} \succ 0$ yields

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} - \left(\frac{\boldsymbol{g}^{(k)\top} \boldsymbol{g}^{(k)}}{\boldsymbol{g}^{(k)\top} \boldsymbol{Q} \boldsymbol{g}^{(k)}} \right) \boldsymbol{g}^{(k)}$$

Several concepts about algorithms and convergence:

- **Iterative algorithm**: an algorithm that generates sequence $\boldsymbol{x}^{(0)}$, $\boldsymbol{x}^{(1)}$, $\boldsymbol{x}^{(2)}$,$\dots$, each based on the points preceding it.

- **Descent method**: a method/algorithm such that $f(\boldsymbol{x}^{(k+1)}) \leq f(\boldsymbol{x}^{(k)})$.

- **Globally convergent**: an algorithm that generates sequence $\boldsymbol{x}^{(k)} \to \boldsymbol{x}^*$ starting from ANY $\boldsymbol{x}^{(0)}$.

- **Locally convergent**: an algorithm that generates sequence $\boldsymbol{x}^{(k)} \to \boldsymbol{x}^*$ if $\boldsymbol{x}^{(0)}$ is sufficiently close to $\boldsymbol{x}^*$.

- **Rate of convergence**: how fast is the convergence (more later).

Now we come back to the convergence of the steepest descent applied to quadratic function $f(x) = \frac{1}{2}\boldsymbol{x}^\top \boldsymbol{Q}\boldsymbol{x} - \boldsymbol{b}^\top \boldsymbol{x}$ where $\boldsymbol{Q} \succ 0$.

Since $\nabla^2 f(\boldsymbol{x}) = \boldsymbol{Q} \succ 0$, f is strictly convex and only has a unique minimizer, denoted by $\boldsymbol{x}^*$.

By FONC, there is $\nabla f(\boldsymbol{x}^*) = \boldsymbol{Q}\boldsymbol{x}^* - \boldsymbol{b} = \boldsymbol{0}$, i.e., $\boldsymbol{Q}\boldsymbol{x}^* = \boldsymbol{b}$.

To examine the convergence, we consider

$$\begin{aligned} V(\boldsymbol{x}) &:= f(\boldsymbol{x}) + \frac{1}{2}\boldsymbol{x}^{*\top}\boldsymbol{Q}\boldsymbol{x}^* \\ &= \cdots \\ &= \frac{1}{2}(\boldsymbol{x}-\boldsymbol{x}^*)^\top \boldsymbol{Q}(\boldsymbol{x}-\boldsymbol{x}^*) \end{aligned}$$

(show this as an exercise).

Since $\boldsymbol{Q} \succ 0$, there is $V(\boldsymbol{x}) = 0$ iff $\boldsymbol{x} = \boldsymbol{x}^*$.

Lemma. Let $\{\boldsymbol{x}^{(k)}\}$ be generated by the steepest descent method. Then

$$V(\boldsymbol{x}^{(k+1)}) = (1-\gamma_k)V(\boldsymbol{x}^{(k)})$$

where

$$\gamma_k = \begin{cases} 0 & \text{if } \|\boldsymbol{g}^{(k)}\| = 0 \\ \alpha_k \frac{\boldsymbol{g}^{(k)\top}\boldsymbol{Q}\boldsymbol{g}^{(k)}}{\boldsymbol{g}^{(k)\top}\boldsymbol{Q}^{-1}\boldsymbol{g}^{(k)}} \left(2\frac{\boldsymbol{g}^{(k)\top}\boldsymbol{g}^{(k)}}{\boldsymbol{g}^{(k)\top}\boldsymbol{Q}\boldsymbol{g}^{(k)}} - \alpha_k \right) & \text{if } \|\boldsymbol{g}^{(k)}\| \neq 0 \end{cases}$$

Proof. If $\|\boldsymbol{g}^{(k)}\| = 0$, then $\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)}$ and $V(\boldsymbol{x}^{(k+1)}) = V(\boldsymbol{x}^{(k)})$. Hence $\gamma_k = 0$.

If $\|\boldsymbol{g}^{(k)}\| \neq 0$, then

$$\begin{aligned} V(\boldsymbol{x}^{(k+1)}) &= \frac{1}{2}(\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^*)^\top \boldsymbol{Q}(\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^*) \\ &= \frac{1}{2}(\boldsymbol{x}^{(k)} - \boldsymbol{x}^* + \alpha_k \boldsymbol{g}^{(k)})^\top \boldsymbol{Q}(\boldsymbol{x}^{(k)} - \boldsymbol{x}^* + \alpha_k \boldsymbol{g}^{(k)}) \\ &= V(\boldsymbol{x}^{(k)}) - \alpha_k \boldsymbol{g}^{(k)\top} \boldsymbol{Q}(\boldsymbol{x}^{(k)} - \boldsymbol{x}^*) + \frac{1}{2}\alpha_k^2 \boldsymbol{g}^{(k)\top} \boldsymbol{Q} \boldsymbol{g}^{(k)} \end{aligned}$$

Therefore

$$\frac{V(\boldsymbol{x}^{(k)}) - V(\boldsymbol{x}^{(k+1)})}{V(\boldsymbol{x}^{(k)})} = \frac{\alpha_k \boldsymbol{g}^{(k)\top} \boldsymbol{Q}(\boldsymbol{x}^{(k)} - \boldsymbol{x}^*) - \frac{1}{2}\alpha_k^2 \boldsymbol{g}^{(k)\top} \boldsymbol{Q} \boldsymbol{g}^{(k)}}{V(\boldsymbol{x}^{(k)})}$$

Note that:

$$\begin{aligned}\boldsymbol{Q}(\boldsymbol{x}^{(k)}-\boldsymbol{x}^*) = \boldsymbol{Q}\boldsymbol{x}^{(k)}-\boldsymbol{b} &= \nabla f(\boldsymbol{x}^{(k)}) = \boldsymbol{g}^{(k)} \\ \boldsymbol{x}^{(k)}-\boldsymbol{x}^* &= \boldsymbol{Q}^{-1}\boldsymbol{g}^{(k)} \\ V(\boldsymbol{x}^{(k)}) &= \frac{1}{2}(\boldsymbol{x}^{(k)}-\boldsymbol{x}^*)^\top \boldsymbol{Q}(\boldsymbol{x}^{(k)}-\boldsymbol{x}^*) = \frac{1}{2}\boldsymbol{g}^{(k)\top}\boldsymbol{Q}^{-1}\boldsymbol{g}^{(k)}\end{aligned}$$

Then we obtain

$$\frac{V(\boldsymbol{x}^{(k)})-V(\boldsymbol{x}^{(k+1)})}{V(\boldsymbol{x}^{(k)})} = \frac{\alpha_k a - \frac{1}{2}\alpha_k^2 b}{\frac{1}{2}c} = \alpha_k \frac{b}{c}\left(2\frac{a}{b}-\alpha_k\right)$$

where

$$a := \boldsymbol{g}^{(k)\top}\boldsymbol{g}^{(k)}, \quad b := \boldsymbol{g}^{(k)\top}\boldsymbol{Q}\boldsymbol{g}^{(k)}, \quad c := \boldsymbol{g}^{(k)\top}\boldsymbol{Q}^{-1}\boldsymbol{g}^{(k)}$$

Now we have obtained $V(\boldsymbol{x}^{(k+1)}) = (1-\gamma_k)V(\boldsymbol{x}^{(k)})$, from which we have

$$V(\boldsymbol{x}^{(k)}) = \left[\prod_{i=0}^{k-1}(1-\gamma_i)\right] V(\boldsymbol{x}^{(0)})$$

Since $\boldsymbol{x}^{(0)}$ is given and fixed, we can see

$$\begin{aligned} V(\boldsymbol{x}^{(k)}) \to 0 \quad & \Longleftrightarrow \quad \prod_{i=0}^{k-1}(1-\gamma_i) \to 0 \\ & \Longleftrightarrow \quad -\sum_{i=0}^{k-1}\log(1-\gamma_i) \to +\infty \\ & \Longleftrightarrow \quad \sum_{i=0}^{k-1}\gamma_i \to +\infty \end{aligned}$$

We summarize the result below:

Theorem. Let $\left\{\boldsymbol{x}^{(k)}\right\}$ be generated by the gradient algorithm for a quadratic function $f(\boldsymbol{x}) = (1/2)\boldsymbol{x}^\top \boldsymbol{Q}\boldsymbol{x} - \boldsymbol{b}^\top \boldsymbol{x}$ (where $\boldsymbol{Q} \succ 0$) with step sizes α_k converges, i.e., $\boldsymbol{x}^{(k)} \to \boldsymbol{x}^*$ iff $\sum_{k=0}^{\infty} \gamma_k = +\infty$.

Proof. (Sketch) Use the inequalities

$$\gamma \leq -\log(1-\gamma) \leq 2\gamma$$

which hold for $\gamma \geq 0$ close to 0. Then use the squeeze theorem.

Rayleigh's inequality: given a symmetric $\boldsymbol{Q} \succ 0$, there is

$$\lambda_{\text{min}}(\boldsymbol{Q})\|\boldsymbol{x}\|^2 \le \boldsymbol{x}^\top \boldsymbol{Q}\boldsymbol{x} =: \|\boldsymbol{x}\|_{\boldsymbol{Q}}^2 \le \lambda_{\text{max}}(\boldsymbol{Q})\|\boldsymbol{x}\|^2$$

for any $\boldsymbol{x}$.

Here $\lambda_{\text{min}}(\boldsymbol{Q})$ ($\lambda_{\text{max}}(\boldsymbol{Q})$) are the minimum (maximum) eigenvalue of $\boldsymbol{Q}$.

In addition, we can get the min/max eigenvalues of $\boldsymbol{Q}^{-1}$:

$$\lambda_{\text{min}}(\boldsymbol{Q}^{-1}) = \frac{1}{\lambda_{\text{max}}(\boldsymbol{Q})} \qquad \text{and} \qquad \lambda_{\text{max}}(\boldsymbol{Q}^{-1}) = \frac{1}{\lambda_{\text{min}}(\boldsymbol{Q})}$$

Lemma. If $\boldsymbol{Q} \succ 0$, then for any $\boldsymbol{x}$, there is

$$\frac{\lambda_{\min}(\boldsymbol{Q})}{\lambda_{\max}(\boldsymbol{Q})} \le \frac{\|\boldsymbol{x}\|^4}{\|\boldsymbol{x}\|_{\boldsymbol{Q}}^2 \|\boldsymbol{x}\|_{\boldsymbol{Q}^{-1}}^2} \le \frac{\lambda_{\max}(\boldsymbol{Q})}{\lambda_{\min}(\boldsymbol{Q})}$$

Proof. By Rayleigh's inequality, we have

$$\lambda_{\min}(\boldsymbol{Q})\|\boldsymbol{x}\|^2 \le \|\boldsymbol{x}\|_{\boldsymbol{Q}}^2 \le \lambda_{\max}(\boldsymbol{Q})\|\boldsymbol{x}\|^2 \text{ and } \frac{\|\boldsymbol{x}\|^2}{\lambda_{\max}(\boldsymbol{Q})} \le \|\boldsymbol{x}\|_{\boldsymbol{Q}^{-1}}^2 \le \frac{\|\boldsymbol{x}\|^2}{\lambda_{\min}(\boldsymbol{Q})}$$

These imply

$$\frac{1}{\lambda_{\max}(\boldsymbol{Q})} \le \frac{\|\boldsymbol{x}\|^2}{\|\boldsymbol{x}\|_{\boldsymbol{Q}}^2} \le \frac{1}{\lambda_{\min}(\boldsymbol{Q})} \quad \text{and} \quad \lambda_{\min}(\boldsymbol{Q}) \le \frac{\|\boldsymbol{x}\|^2}{\|\boldsymbol{x}\|_{\boldsymbol{Q}^{-1}}^2} \le \lambda_{\max}(\boldsymbol{Q})$$

Multiplying the two yields the claim.

We can show the steepest descent method has α_k set to satisfy $\sum_k \gamma_k = +\infty$:

First recall that $\alpha_k = \frac{\boldsymbol{g}^{(k)^\top}\boldsymbol{g}^{(k)}}{\boldsymbol{g}^{(k)^\top}\boldsymbol{Q}\boldsymbol{g}^{(k)}}$.

Then there is

$$\begin{aligned}
\gamma_k &= \alpha_k \frac{\boldsymbol{g}^{(k)^\top}\boldsymbol{Q}\boldsymbol{g}^{(k)}}{\boldsymbol{g}^{(k)^\top}\boldsymbol{Q}^{-1}\boldsymbol{g}^{(k)}}\left(2\frac{\boldsymbol{g}^{(k)^\top}\boldsymbol{g}^{(k)}}{\boldsymbol{g}^{(k)^\top}\boldsymbol{Q}\boldsymbol{g}^{(k)}} - \alpha_k\right) \\
&= \frac{(\boldsymbol{g}^{(k)^\top}\boldsymbol{g}^{(k)})^2}{\boldsymbol{g}^{(k)^\top}\boldsymbol{Q}\boldsymbol{g}^{(k)}\boldsymbol{g}^{(k)^\top}\boldsymbol{Q}^{-1}\boldsymbol{g}^{(k)}} \\
&= \frac{\|\boldsymbol{g}^{(k)}\|^4}{\|\boldsymbol{g}^{(k)}\|_{\boldsymbol{Q}}^2\|\boldsymbol{g}^{(k)}\|_{\boldsymbol{Q}^{-1}}^2} \ge \frac{\lambda_{\min}(\boldsymbol{Q})}{\lambda_{\max}(\boldsymbol{Q})} > 0
\end{aligned}$$

Therefore $\sum_k \gamma_k = +\infty$.

Now let's consider the gradient method with fixed step size $\alpha > 0$:

Theorem. If the step size $\alpha > 0$ is fixed, then the gradient method converges if and only if

$$0 < \alpha < \frac{2}{\lambda_{\max}(\boldsymbol{Q})}$$

Proof. "$\Leftarrow$" Suppose $0 < \alpha < \frac{2}{\lambda_{\max}(\boldsymbol{Q})}$, then

$$\begin{aligned}
\gamma_k &= \alpha \frac{\|\boldsymbol{g}^{(k)}\|_{\boldsymbol{Q}}^2}{\|\boldsymbol{g}^{(k)}\|_{\boldsymbol{Q}^{-1}}^2} \left(2 \frac{\|\boldsymbol{g}^{(k)}\|^2}{\|\boldsymbol{g}^{(k)}\|_{\boldsymbol{Q}}^2} - \alpha \right) \\
&\geq \alpha \frac{\lambda_{\min}(\boldsymbol{Q}) \|\boldsymbol{g}^{(k)}\|^2}{\lambda_{\max}(\boldsymbol{Q}^{-1}) \|\boldsymbol{g}^{(k)}\|^2} \left(\frac{2}{\lambda_{\max}(\boldsymbol{Q})} - \alpha \right) \\
&= \alpha \lambda_{\min}^2(\boldsymbol{Q}) \left(\frac{2}{\lambda_{\max}(\boldsymbol{Q})} - \alpha \right) > 0
\end{aligned}$$

Therefore $\sum_k \gamma_k = \infty$ and hence GM converges.

"⇒" Suppose GM converges but $\alpha \le 0$ or $\alpha \ge \frac{2}{\lambda_{\max}(\boldsymbol{Q})}$. Then if $\boldsymbol{x}^{(0)}$ is chosen such that $\boldsymbol{x}^{(0)} - \boldsymbol{x}^*$ is the eigenvector corresponding to the eigenvalue $\lambda_{\max}(\boldsymbol{Q})$ of $\boldsymbol{Q}$, we have

$$\begin{aligned}
\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^* &= \boldsymbol{x}^{(k)} - \alpha \boldsymbol{g}^{(k)} - \boldsymbol{x}^* \\
&= \boldsymbol{x}^{(k)} - \alpha(\boldsymbol{Q}\boldsymbol{x}^{(k)} - \boldsymbol{b}) - \boldsymbol{x}^* \\
&= \boldsymbol{x}^{(k)} - \alpha(\boldsymbol{Q}\boldsymbol{x}^{(k)} - \boldsymbol{Q}\boldsymbol{x}^*) - \boldsymbol{x}^* \\
&= (\boldsymbol{I} - \alpha \boldsymbol{Q})(\boldsymbol{x}^{(k)} - \boldsymbol{x}^*) \\
&= (\boldsymbol{I} - \alpha \boldsymbol{Q})^{k+1}(\boldsymbol{x}^{(0)} - \boldsymbol{x}^*) \\
&= (1 - \alpha \lambda_{\max}(\boldsymbol{Q}))^{k+1}(\boldsymbol{x}^{(0)} - \boldsymbol{x}^*)
\end{aligned}$$

Taking norm on both sides yields

$$\|\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^*\| = |1 - \alpha \lambda_{\max}(\boldsymbol{Q})|^{k+1} \|\boldsymbol{x}^{(0)} - \boldsymbol{x}^*\|$$

where $|1 - \alpha \lambda_{\max}(\boldsymbol{Q})| \ge 1$ if $\alpha \le 0$ or $\alpha \ge \frac{2}{\lambda_{\max}(\boldsymbol{Q})}$. Contradiction.

Example. Find an appropriate α for the GM with fixed step size α for

$$f(\boldsymbol{x}) = \boldsymbol{x}^\top \begin{bmatrix} 4 & 2\sqrt{2} \\ 0 & 5 \end{bmatrix} \boldsymbol{x} + \boldsymbol{x}^\top \begin{bmatrix} 3 \\ 6 \end{bmatrix} + 24$$

Solution. First rewrite f into the standard quadratic form with symmetric $\boldsymbol{Q}$:

$$f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^\top \begin{bmatrix} 8 & 2\sqrt{2} \\ 2\sqrt{2} & 10 \end{bmatrix} \boldsymbol{x} + \boldsymbol{x}^\top \begin{bmatrix} 3 \\ 6 \end{bmatrix} + 24$$

Then we compute the eigenvalues of $\boldsymbol{Q} = \begin{bmatrix} 8 & 2\sqrt{2} \\ 2\sqrt{2} & 10 \end{bmatrix}$:

$$|\lambda \boldsymbol{I} - \boldsymbol{Q}| = \begin{vmatrix} \lambda - 8 & -2\sqrt{2} \\ -2\sqrt{2} & \lambda - 10 \end{vmatrix} = (\lambda - 8)(\lambda - 10) - 8 = (\lambda - 6)(\lambda - 12)$$

Hence $\lambda_{\text{max}}(\boldsymbol{Q}) = 12$, and the range of α should be $(0, \frac{2}{12})$.

Convergence rate of steepest descent method:

Recall that applying SD to $f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^\top \boldsymbol{Q}\boldsymbol{x} + \boldsymbol{b}^\top \boldsymbol{x}$ with $\boldsymbol{Q} \succ 0$ yields

$$V(\boldsymbol{x}^{(k+1)}) \le (1-\kappa)V(\boldsymbol{x}^{(k)})$$

where $V(\boldsymbol{x}) := \frac{1}{2}(\boldsymbol{x}-\boldsymbol{x}^*)^\top \boldsymbol{Q}(\boldsymbol{x}-\boldsymbol{x}^*)$, and $\kappa = \frac{\lambda_{\mathsf{min}}(\boldsymbol{Q})}{\lambda_{\mathsf{max}}(\boldsymbol{Q})}$.

Remark. $\frac{\lambda_{\mathsf{max}}(\boldsymbol{Q})}{\lambda_{\mathsf{min}}(\boldsymbol{Q})} = \|\boldsymbol{Q}\|\|\boldsymbol{Q}^{-1}\|$ is called the **condition number** of $\boldsymbol{Q}$.

Order of convergence

We say $\boldsymbol{x}^{(k)} \to \boldsymbol{x}^*$ with order p if

$$0 < \lim_{k\to\infty} \frac{\|\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^*\|}{\|\boldsymbol{x}^{(k)} - \boldsymbol{x}^*\|^p} < \infty$$

It can be shown that $p \geq 1$, and the larger p is, the faster the convergence is.

Example.

- $\boldsymbol{x}^{(k)} = \frac{1}{k} \to 0$, then

$$\frac{|\boldsymbol{x}^{(k+1)}|}{|\boldsymbol{x}^{(k)}|^p} = \frac{k^p}{k+1} < \infty$$

if $p \le 1$. Therefore $\boldsymbol{x}^{(k)} \to 0$ with order 1.

- $\boldsymbol{x}^{(k)} = q^k \to 0$ for some $q \in (0, 1)$, then

$$\frac{|\boldsymbol{x}^{(k+1)}|}{|\boldsymbol{x}^{(k)}|^p} = \frac{q^{k+1}}{q^{kp}} = q^{k(1-p)+1} < \infty$$

if $p \le 1$. Therefore $\boldsymbol{x}^{(k)} \to 0$ with order 1.

Example.

- $\boldsymbol{x}^{(k)} = q^{2^k} \to 0$, then

$$\frac{|\boldsymbol{x}^{(k+1)}|}{|\boldsymbol{x}^{(k)}|^p} = \frac{q^{2^{k+1}}}{q^{p2^k}} = q^{2^k(2-p)} < \infty$$

if $p \leq 2$. Therefore $\boldsymbol{x}^{(k)} \to 0$ with order 2.

In general, we have the following result:

Theorem. If $\|\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^*\| = O(\|\boldsymbol{x}^{(k)} - \boldsymbol{x}^*\|^p)$, then the convergence is of order at least p.

Remark. Note that $p \geq 1$.

Descent method and line search

Given a descent direction $\boldsymbol{d}^{(k)}$ of $f : \mathbb{R}^n \to \mathbb{R}$ at $\boldsymbol{x}^{(k)}$ (e.g., $\boldsymbol{d}^{(k)} = -\boldsymbol{g}^{(k)}$), we need to decide the step size α_k in order to compute

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} + \alpha_k \boldsymbol{d}^{(k)}.$$

Exact line search computes α_k by solving for

$$\alpha_k = \arg\min_{\alpha} \phi_k(\alpha), \quad \text{where} \quad \phi_k(\alpha) := f(\boldsymbol{x}^{(k)} + \alpha \boldsymbol{d}^{(k)}).$$

Notice that $\phi : \mathbb{R}_+ \to \mathbb{R}$ and $\phi'(\alpha) = \nabla f(\boldsymbol{x}^{(k)} + \alpha \boldsymbol{d}^{(k)}) \boldsymbol{d}^{(k)}$. Hence we can use the secant method:

$$\alpha^{(l+1)} = \alpha^{(l)} - \frac{\alpha^{(l)} - \alpha^{(l-1)}}{\phi_k'(\alpha^{(l)}) - \phi_k'(\alpha^{(l-1)})} \phi_k'(\alpha^{(l)}).$$

with some initial guess $\alpha^{(0)}, \alpha^{(1)}$, and set α_k to $\lim_{l\to\infty} \alpha^{(l)}$.

In practice, it is not computationally economical to use exact line search.

Instead, we prefer **inexact line search**. That is, we do not exactly solve

$$\alpha_k = \arg\min_{\alpha} \phi_k(\alpha), \quad \text{where} \quad \phi_k(\alpha) := f(\boldsymbol{x}^{(k)} + \alpha \boldsymbol{d}^{(k)}),$$

but only require α_k to satisfy certain conditions such that:

- easy to compute in practice.
- guarantees convergence.
- performs well in practice.

There are several commonly used conditions for α_k:

- **Armijo condition:** let $\varepsilon \in (0,1)$, $\gamma > 1$ and

$$\phi_k(\alpha_k) \le \phi_k(0) + \varepsilon\alpha_k\phi_k'(0) \qquad \text{(so } \alpha_k \text{ not too large)}$$
$$\phi_k(\gamma\alpha_k) \ge \phi_k(0) + \varepsilon\gamma\alpha_k\phi_k'(0) \qquad \text{(so } \alpha_k \text{ not too small)}$$

- **Armijo-Goldstein condition:** let $0 < \varepsilon < \eta < 1$ and

$$\phi_k(\alpha_k) \le \phi_k(0) + \varepsilon\alpha_k\phi_k'(0) \qquad \text{(so } \alpha_k \text{ not too large)}$$
$$\phi_k(\alpha_k) \ge \phi_k(0) + \eta\alpha_k\phi_k'(0) \qquad \text{(so } \phi_k'(\alpha_k) \text{ not too small)}$$

- **Wolfe condition:** let $0 < \varepsilon < \eta < 1$ and

$$\phi_k(\alpha_k) \le \phi_k(0) + \varepsilon\alpha_k\phi_k'(0) \qquad \text{(so } \alpha_k \text{ not too large)}$$
$$\phi_k'(\alpha_k) \ge \eta\phi_k'(0) \qquad \text{(so } \phi_k \text{ not too steep at } \alpha_k\text{)}$$

  Strong-Wolfe condition: replaces the second condition with $|\phi_k'(\alpha_k)| \le \eta|\phi_k'(0)|$.

Backtracking line search

In practice, we often use the following backtracking line search:

Backtracking: choose initial guess $\alpha^{(0)}$ and $\tau \in (0, 1)$ (e.g., $\tau = 0.5$), then set $\alpha = \alpha^{(0)}$ and repeat:

1. Check whether $\phi_k(\alpha) \leq \phi_k(0) + \varepsilon\alpha\phi_k'(0)$ (first Armijo condition). If yes, then terminate.

2. Shrink α to $\tau\alpha$.

In other words, we find the smallest integer $m \in \mathbb{N}_0$ such that $\alpha_k = \tau^m\alpha^{(0)}$ satisfies the first Armijo condition $\phi_k(\alpha_k) \leq \phi_k(0) + \varepsilon\alpha_k\phi_k'(0)$.

Why line search guarantees convergence?

First, note that here by convergence we mean $\|\nabla f(\boldsymbol{x}^{(k)})\| \to 0$.

We take Wolfe condition and $\boldsymbol{d}^{(k)} = -\boldsymbol{g}^{(k)}$ for simplicity. Assume ∇f is L-Lipschitz continuous. Now

$$\begin{aligned} \boldsymbol{x}^{(k+1)} &= \boldsymbol{x}^{(k)} - \alpha_k \boldsymbol{g}^{(k)} \\ \phi_k(\alpha_k) &= f(\boldsymbol{x}^{(k+1)}) \\ \phi_k'(\alpha_k) &= -\nabla f(\boldsymbol{x}^{(k+1)}) \boldsymbol{g}^{(k)} \\ \phi_k(0) &= f(\boldsymbol{x}^{(k)}) \\ \phi_k'(0) &= -\nabla f(\boldsymbol{x}^{(k)}) \boldsymbol{g}^{(k)} \end{aligned}$$

Moreover, L-Lipschitz continuity of ∇f implies

$$\pm\langle \nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \rangle \le \|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\| \|\boldsymbol{x} - \boldsymbol{y}\| \le L \|\boldsymbol{x} - \boldsymbol{y}\|^2$$

for any $\boldsymbol{x}, \boldsymbol{y}$.

Claim. $\alpha_k \geq \frac{1-\eta}{L}$.

Proof of Claim. The second Wolfe condition $\phi_k'(\alpha_k) \geq \eta\phi_k'(0)$ implies $\phi_k'(\alpha_k) - \phi_k'(0) \geq (\eta - 1)\phi_k'(0)$, which is

$$-\langle \nabla f(\boldsymbol{x}^{(k+1)}) - \nabla f(\boldsymbol{x}^{(k)}), \boldsymbol{g}^{(k)} \rangle \geq (1-\eta)\|\boldsymbol{g}^{(k)}\|^2.$$

Note that $\boldsymbol{g}^{(k)} = \frac{\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)}}{\alpha_k}$, we know

$$-\langle \nabla f(\boldsymbol{x}^{(k+1)}) - \nabla f(\boldsymbol{x}^{(k)}), \boldsymbol{g}^{(k)} \rangle \leq \frac{L}{\alpha_k}\|\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)}\|^2 = L\alpha_k\|\boldsymbol{g}^{(k)}\|^2$$

Combining the two inequalities above yields the claim.

The first Wolfe condition (Armijo condition) implies

$$f(\boldsymbol{x}^{(k+1)}) \le f(\boldsymbol{x}^{(k)}) - \varepsilon\alpha_k \|\boldsymbol{g}^{(k)}\|^2 \le f(\boldsymbol{x}^{(k)}) - \frac{\varepsilon(1-\eta)}{L}\|\boldsymbol{g}^{(k)}\|^2.$$

Taking telescope sum yields

$$f(\boldsymbol{x}^{(K)}) \le f(\boldsymbol{x}^{(0)}) - \frac{\varepsilon(1-\eta)}{L}\sum_{k=0}^{K-1} \|\boldsymbol{g}^{(k)}\|^2.$$

which implies

$$\frac{\varepsilon(1-\eta)}{L}\sum_{k=0}^{K-1} \|\boldsymbol{g}^{(k)}\|^2 \le f(\boldsymbol{x}^{(0)}) - f(\boldsymbol{x}^{(K)}) < \infty$$

for any K (we assume f is bounded below). Notice that $\frac{\varepsilon(1-\eta)}{L} > 0$.

Therefore $\|\boldsymbol{g}^{(k)}\| = \|\nabla f(\boldsymbol{x}^{(k)})\| \to 0$.