Chapter 8 Gradient Methods

1. Steepest Descent
2. Analysis of Gradient Methods

Optimization Problem

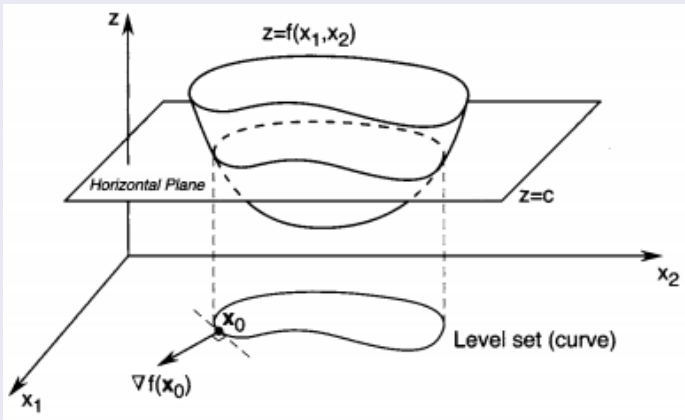
unconstrained optimization problem: $\min\{f(\boldsymbol{x}) \mid \boldsymbol{x} \in \mathbb{R}^n\}$

- The function $f : \mathbb{R}^n \to \mathbb{R}$ is multi-dimensional objective function.
- The vector $\boldsymbol{x} \in \mathbb{R}^n$ is n-dimensional decision variables.

Definition (level set, normal vector, and tangent vector)

Let $f : \mathbb{R}^n \to \mathbb{R}$ be function, and $c \in \mathbb{R}$ be a constant.

- level set of f w.r.t constant c:
$S_c(f) = \{\boldsymbol{x} \in \mathbb{R} \mid f(\boldsymbol{x}) = c\}$.
$\Longrightarrow \boldsymbol{x}_0 \in S_c(f)$ if $f(\boldsymbol{x}_0) = c$.
- tangent of f at $\boldsymbol{x}_0$:
$l(\boldsymbol{x}) = f(\boldsymbol{x}_0) + \nabla f(\boldsymbol{x}_0)^\top (\boldsymbol{x} - \boldsymbol{x}_0)$.
- normal of f at $\boldsymbol{x}_0$: $\nabla f(\boldsymbol{x}_0)$.



★ Given small displacement, f increases more in the direction of $\nabla f(\boldsymbol{x}_0)$ than in any other direction.

proof. Taylor series and the rate of increase $\langle \nabla f(\boldsymbol{x}_0), \boldsymbol{d} \rangle$ with $\|\boldsymbol{d}\| = 1$ and Cauchy inequality

Gradient Descent Methods

motivation for $\min\{f(\boldsymbol{x}) \mid \boldsymbol{x} \in \mathbb{R}^n\}$

Let $\boldsymbol{x}^{(0)}$ be an initial point, and a new point $\boldsymbol{x}^{(0)} - \alpha \nabla f(\boldsymbol{x}^{(0)})$.

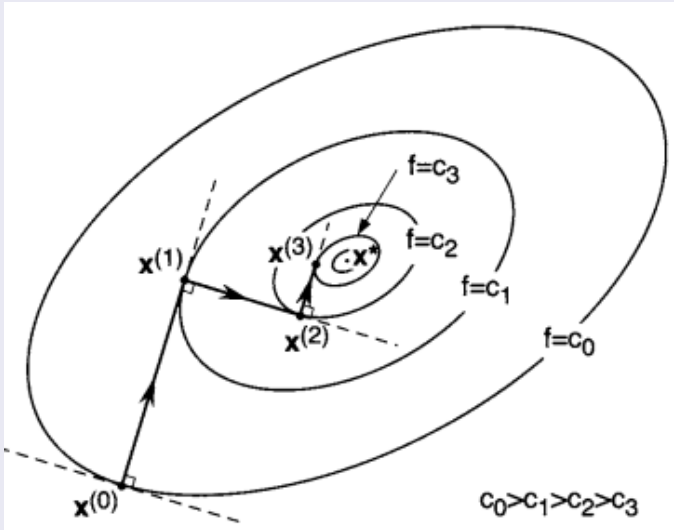
$$\xrightarrow{\text{Taylor's series}} f(\boldsymbol{x}^{(0)} - \alpha \nabla f(\boldsymbol{x}^{(0)})) = f(\boldsymbol{x}^{(0)}) - \alpha \|\nabla f(\boldsymbol{x}^{(0)})\|^2 + o(\alpha).$$

If $\nabla f(\boldsymbol{x}^{(0)}) \neq 0 \Longrightarrow f(\boldsymbol{x}^{(0)} - \alpha \nabla f(\boldsymbol{x}^{(0)})) < f(\boldsymbol{x}^{(0)}), \forall \alpha > 0.$

$\Longrightarrow \boldsymbol{x}^{(0)} - \alpha \nabla f(\boldsymbol{x}^{(0)})$ is an improvement over $\boldsymbol{x}^{(0)}$ as a minimizer.

gradient descent method for solving $\min\{f(\boldsymbol{x}) \mid \boldsymbol{x} \in \mathbb{R}^n\}$

- gradient descent method:
  - S0: given $\boldsymbol{x}^{(0)}$ and step size $\alpha_k > 0$,
  - S1: $\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} - \alpha_k \nabla f(\boldsymbol{x}^{(k)})$.
- steepest descent method:
  - S0: given $\boldsymbol{x}^{(0)}$,
  - S1: $\alpha_k = \arg\min_{\alpha > 0} f(\boldsymbol{x}^{(k)} - \alpha \nabla f(\boldsymbol{x}^{(k)}))$.
  - S2: $\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} - \alpha_k \nabla f(\boldsymbol{x}^{(k)})$.



Gradient Descent Methods

Theorem (property of steepest descent method)

If $\{\boldsymbol{x}^{(k)}\}_{k=0}^{\infty}$ is a steepest descent sequence, then $\nabla f(\boldsymbol{x}^{(k+1)}) \perp \nabla f(\boldsymbol{x}^{(k)})$ or $(\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)}) \perp (\boldsymbol{x}^{(k)} - \boldsymbol{x}^{(k-1)})$.

proof. let $\varphi(\alpha) = f(\boldsymbol{x}^{(k)} - \alpha \nabla f(\boldsymbol{x}^{(k)}))$.

By definition of $\alpha_k \Longrightarrow \varphi'(\alpha_k) = 0 \Longrightarrow \langle \nabla f(\boldsymbol{x}^{(k+1)}), \nabla f(\boldsymbol{x}^{(k)}) \rangle = 0$.

By iterative scheme $\boldsymbol{x}^{(k)} \Longrightarrow \langle \boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)}, \boldsymbol{x}^{(k)} - \boldsymbol{x}^{(k-1)} \rangle = 0$

Theorem (property of steepest descent method)

If $\{\boldsymbol{x}^{(k)}\}_{k=0}^{\infty}$ is a steepest descent sequence, and if $\nabla f(\boldsymbol{x}^{(k)}) \neq 0$, then $f(\boldsymbol{x}^{(k+1)}) < f(\boldsymbol{x}^{(k)})$.

proof. By the definition of φ and Taylor series

$$\Longleftrightarrow \quad \varphi(\alpha_k) = \varphi(\alpha_0) + \varphi'(\alpha_0)\alpha_k + o(\alpha_k)$$

$$\Longleftrightarrow \quad f(\boldsymbol{x}^{(k+1)}) = f(\boldsymbol{x}^{(k)}) - \alpha_k \|\nabla f(\boldsymbol{x}^{(k)})\|^2 + o(\alpha_k).$$

If $\nabla f(\boldsymbol{x}^{(k)}) \neq 0$, then $f(\boldsymbol{x}^{(k+1)}) < f(\boldsymbol{x}^{(k)})$.

Gradient Descent Methods

- ★ steepest descent method render the searching direction in "zigzag" phenomenon.
- ★ the gradient descent method render the objective function value monotonically decrease.

stopping criterion: let $\varepsilon > 0$ is a prespecified threshold (e.g.,10^{-6})

- 1st-order necessary condition: $\|\nabla f(\boldsymbol{x}^{(k)})\| < \varepsilon$.
- difference of successive objective function values:
  $|f(\boldsymbol{x}^{(k+1)}) - f(\boldsymbol{x}^{(k)})| < \varepsilon$, $\frac{|f(\boldsymbol{x}^{(k+1)}) - f(\boldsymbol{x}^{(k)})|}{|f(\boldsymbol{x}^{(k)})|} < \varepsilon$, $\frac{|f(\boldsymbol{x}^{(k+1)}) - f(\boldsymbol{x}^{(k)})|}{\max\{1, |f(\boldsymbol{x}^{(k)})|\}} < \varepsilon$.
- difference of successive iterates:
  $\|\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)}\| < \varepsilon$, $\frac{\|\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)}\|}{\|\boldsymbol{x}^{(k)}\|} < \varepsilon$, $\frac{\|\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)}\|}{\max\{1, \|\boldsymbol{x}^{(k)}\|\}} < \varepsilon$.

Gradient Descent Methods

Example ($\min_{\boldsymbol{x}\in\mathbb{R}^3} f(\boldsymbol{x}) = (x_1-4)^4 + (x_2-3)^2 + 4(x_3+5)^4$)

Minimize f by steepest descent method with $\boldsymbol{x}^{(0)} = [4, 2, -1]^\top$.

Ans: The gradient of f is $\nabla f(\boldsymbol{x}) = [4(x_1-4)^3, 2(x_2-3), 16(x_3+5)^3]^\top$.

S1: $\because \nabla f(\boldsymbol{x}^{(0)}) = [0, -2, 1024]^\top$ and exact stepsize

$$\alpha_0 = \arg\min_{\alpha>0} f(\boldsymbol{x}^{(0)} - \alpha\nabla f(\boldsymbol{x}^{(0)}) \xrightarrow{\text{secant method}} \alpha_0 = 3.967\times 10^{-3}.$$

The new iterate $\boldsymbol{x}^{(1)} = \boldsymbol{x}^{(0)} - \alpha_0\nabla f(\boldsymbol{x}^{(0)} = [4.000, 2.008, -5.062]$.

S2: $\because \nabla f(\boldsymbol{x}^{(1)}) = [0.000, -1.984, -0.003875]^\top$ and exact stepsize

$$\alpha_1 = \arg\min_{\alpha>0} f(\boldsymbol{x}^{(1)} - \alpha\nabla f(\boldsymbol{x}^{(1)}) \xrightarrow{\text{secant method}} \alpha_1 = 0.500.$$

The new iterate $\boldsymbol{x}^{(2)} = \boldsymbol{x}^{(1)} - \alpha_1\nabla f(\boldsymbol{x}^{(1)} = [4.000, 3.000, -5.060]$.

S3: $\because \nabla f(\boldsymbol{x}^{(2)}) = [0.000, 0.000, -0.003525]^\top$ and exact stepsize

$$\alpha_2 = \arg\min_{\alpha>0} f(\boldsymbol{x}^{(2)} - \alpha\nabla f(\boldsymbol{x}^{(2)}) \xrightarrow{\text{secant method}} \alpha_2 = 16.29.$$

The new iterate $\boldsymbol{x}^{(3)} = \boldsymbol{x}^{(2)} - \alpha_2\nabla f(\boldsymbol{x}^{(2)} = [4.000, 3.000, -5.002]$.

★ The minimizer of $\boldsymbol{x}^* = [4, 3, -5]^\top$. It can be implement by C/C++/matlab/$\cdots$ programs.

Steepest Descent Methods on Quadratic Function

Definition (quadratic function)

$f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^\top \boldsymbol{Q}\boldsymbol{x} - \boldsymbol{b}^\top \boldsymbol{x}$, where $\boldsymbol{Q} \in \mathbb{R}^{n\times n}$ is a symmetric positive definite matrix, $\boldsymbol{b} \in \mathbb{R}^n$, and $\boldsymbol{x} \in \mathbb{R}^n$.

By 1st-order necessary condition: $0 = \nabla f(\boldsymbol{x}) = \boldsymbol{Q}\boldsymbol{x} - \boldsymbol{b}$, $\Longrightarrow$ unique minimizer: $\boldsymbol{x}^* = \boldsymbol{Q}^{-1}\boldsymbol{b}$.

Notations: let gradient of f at $\boldsymbol{x}^{(k)}$ be $\boldsymbol{g}^{(k)} = \nabla f(\boldsymbol{x}^{(k)})$ and Hessian of f be $\boldsymbol{F}(\boldsymbol{x}) = \nabla^2 f(\boldsymbol{x}^{(k)}) = \boldsymbol{Q}$.

steepest descent method for quadratic function

$$\alpha_k = \frac{(\boldsymbol{g}^{(k)})^\top \boldsymbol{g}^{(k)}}{(\boldsymbol{g}^{(k)})^\top \boldsymbol{Q}\boldsymbol{g}^{(k)}} = \frac{\|\boldsymbol{g}^{(k)}\|_2^2}{\|\boldsymbol{g}^{(k)}\|_{\boldsymbol{Q}}^2},$$
$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} - \alpha_k \boldsymbol{g}^{(k)}.$$

Question: why $\alpha_k = \frac{(\boldsymbol{g}^{(k)})^\top \boldsymbol{g}^{(k)}}{(\boldsymbol{g}^{(k)})^\top \boldsymbol{Q}\boldsymbol{g}^{(k)}}$?

Steepest Descent Methods on Quadratic Function

exact stepsize: $\alpha_k = \dfrac{(\boldsymbol{g}^{(k)})^\top \boldsymbol{g}^{(k)}}{(\boldsymbol{g}^{(k)})^\top Q \boldsymbol{g}^{(k)}}$

$\because\ \alpha_k = \arg\min_{\alpha>0} \varphi(\alpha) := f(\boldsymbol{x}^{(k)} - \alpha \boldsymbol{g}^{(k)})$,

By optimality condition of minimizer α_k, $\therefore \varphi'(\alpha) = 0$, i.e.,

$$0 = \varphi'(\alpha_k) = (\boldsymbol{g}^{(k)})^\top \nabla f(\boldsymbol{x}^{(k)} - \alpha \nabla \boldsymbol{g}^{(k)}) = (\boldsymbol{g}^{(k)})^\top (\boldsymbol{Q}\boldsymbol{x}^{(k)} - \alpha \boldsymbol{Q}\boldsymbol{g}^{(k)} - b)$$
$$= (\boldsymbol{g}^{(k)})^\top (\boldsymbol{g}^{(k)} - \alpha \boldsymbol{Q}\boldsymbol{g}^{(k)}) = (\boldsymbol{g}^{(k)})^\top \boldsymbol{g}^{(k)} - \alpha (\boldsymbol{g}^{(k)})^\top \boldsymbol{Q}\boldsymbol{g}^{(k)}$$

Example ($\min_{\boldsymbol{x}\in\mathbb{R}^2} f(\boldsymbol{x}) = x_1^2 + x_2^2$)

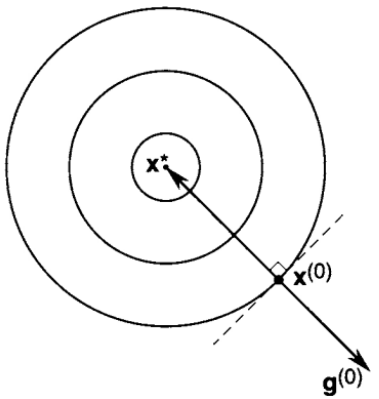
Minimize f by steepest descent method with any initial point $\boldsymbol{x}^{(0)} \in \mathbb{R}^2$.

Ans: $\because$ gradient of f is $\nabla f(\boldsymbol{x}) = 2\boldsymbol{x}$,

Hessian of f is $\nabla^2 f(\boldsymbol{x}) = Q = 2I$.

$\because\ \boldsymbol{g}^{(0)} = 2\boldsymbol{x}^{(0)} \Longrightarrow \alpha_0 = \dfrac{(\boldsymbol{g}^{(0)})^\top \boldsymbol{g}^{(0)}}{(\boldsymbol{g}^{(0)})^\top Q \boldsymbol{g}^{(0)}} = \dfrac{1}{2}$

$\therefore$ new iterate: $\boldsymbol{x}^{(1)} = \boldsymbol{x}^{(0)} - \alpha_0 \boldsymbol{g}^{(0)} = [0, 0]^\top$.



Steepest Descent Methods on Quadratic Function

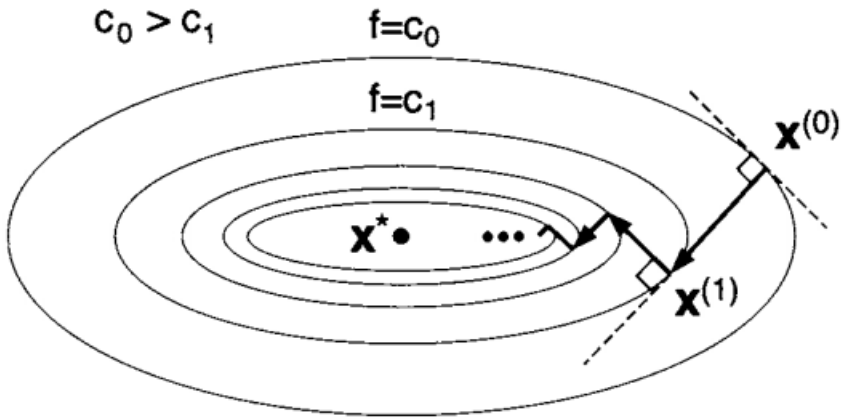
Example ($\min_{\boldsymbol{x}\in\mathbb{R}^2} f(\boldsymbol{x}) = \frac{x_1^2}{5} + x_2^2$)

Minimize f by steepest descent method with any initial point $\boldsymbol{x}^{(0)} \in \mathbb{R}^2$.

Ans: with any initial point $\boldsymbol{x}^{(0)}$,
implement steepest descent method

$$\begin{cases} \alpha_k = \frac{(\boldsymbol{g}^{(k)})^\top \boldsymbol{g}^{(k)}}{(\boldsymbol{g}^{(k)})^\top Q \boldsymbol{g}^{(k)}} = \frac{\|\boldsymbol{g}^{(k)}\|_2^2}{\|\boldsymbol{g}^{(k)}\|_{\boldsymbol{Q}}^2}, \\ \boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} - \alpha_k \boldsymbol{g}^{(k)}, \end{cases}$$

(e.g., in Matlab/C/C++)



Properties of steepest descent methods on quadratic function

i) function values are monotonically descent, i.e., $f(\boldsymbol{x}^{(k+1)}) \le f(\boldsymbol{x}^{(k)})$;

ii) globally convergent, $\boldsymbol{x}^{(k)} \xrightarrow{k\to\infty} \boldsymbol{x}^*$ a minimizer of f.

Convergence Analysis

- Goal: convergence analysis of steepest descent method on $\min_{\boldsymbol{x}\in\mathbb{R}^n} f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^\top \boldsymbol{Q}\boldsymbol{x} - \boldsymbol{b}^\top \boldsymbol{x}$, where $\boldsymbol{Q} \succ 0$, $\boldsymbol{b} \in \mathbb{R}^n$, and $\boldsymbol{x} \in \mathbb{R}^n$.
- Facts: $\boldsymbol{x}^*$ is minimizer (i.e., $\boldsymbol{Q}\boldsymbol{x}^* = \boldsymbol{b}$) $\Rightarrow f(\boldsymbol{x}^*) = -\frac{1}{2}(\boldsymbol{x}^*)^\top \boldsymbol{Q}\boldsymbol{x}^*$.
- Notations: $V(\boldsymbol{x}) = f(\boldsymbol{x}) - f(\boldsymbol{x}^*) = \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{x}^*\|_{\boldsymbol{Q}}^2 = \frac{1}{2}\|\boldsymbol{g}\|_{\boldsymbol{Q}^{-1}}^2$

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

Let $\{\boldsymbol{x}^{(k)}\}_{k=0}^{\infty}$ be a sequence generated by steepest descent method. Then

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

where $\boldsymbol{g}^{(k)} = \boldsymbol{Q}\boldsymbol{x}^{(k)} - \boldsymbol{b}$ and the parameter

$$\gamma_k = \begin{cases} 1, & \text{if } \boldsymbol{g}^{(k)} = 0, \\ \dfrac{\alpha_k \|\boldsymbol{g}^{(k)}\|_{\boldsymbol{Q}}^2}{\|\boldsymbol{g}^{(k)}\|_{\boldsymbol{Q}^{-1}}^2} \left(2\dfrac{\|\boldsymbol{g}^{(k)}\|_2^2}{\|\boldsymbol{g}^{(k)}\|_{\boldsymbol{Q}}^2} - \alpha_k\right), & \text{if } \boldsymbol{g}^{(k)} \neq 0. \end{cases}$$

Convergence Analysis

proof. assume $\boldsymbol{g}^{(k)} \neq 0$. Therefore,

$$
\begin{aligned}
& V(\boldsymbol{x}^{(k+1)}) - V(\boldsymbol{x}^{(k)}) \\
&= \tfrac{1}{2}\|\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^*\|_{\boldsymbol{Q}}^2 - \tfrac{1}{2}\|\boldsymbol{x}^{(k)} - \boldsymbol{x}^*\|_{\boldsymbol{Q}}^2 \\
&= \tfrac{1}{2}\|\boldsymbol{x}^{(k)} - \alpha_k \boldsymbol{g}^{(k)} - \boldsymbol{x}^*\|_{\boldsymbol{Q}}^2 - \tfrac{1}{2}\|\boldsymbol{x}^{(k)} - \boldsymbol{x}^*\|_{\boldsymbol{Q}}^2 \qquad [\because \text{iterative scheme}] \\
&= -\alpha_k (\boldsymbol{x}^{(k)} - \boldsymbol{x}^*)^\top \boldsymbol{Q} \boldsymbol{g}^{(k)} + \tfrac{\alpha_k^2}{2}\|\boldsymbol{g}^{(k)}\|_{\boldsymbol{Q}}^2 \\
&= -\alpha_k (\boldsymbol{g}^{(k)})^\top \boldsymbol{g}^{(k)} + \tfrac{\alpha_k^2}{2}\|\boldsymbol{g}^{(k)}\|_{\boldsymbol{Q}}^2 \\
&= \underbrace{\frac{\|\boldsymbol{g}^{(k)}\|_{\boldsymbol{Q}^{-1}}^2}{2}}_{V(\boldsymbol{x}^{(k)})} \underbrace{\frac{2}{\|\boldsymbol{g}^{(k)}\|_{\boldsymbol{Q}^{-1}}^2}\left(-\alpha_k \|\boldsymbol{g}^{(k)}\|_2^2 + \frac{\alpha_k^2}{2}\|\boldsymbol{g}^{(k)}\|_{\boldsymbol{Q}}^2\right)}_{\gamma_k}
\end{aligned}
$$

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

Convergence Analysis

Question: how to guarantee $\boldsymbol{x}^{(k)} \to \boldsymbol{x}^*$?

Lemma

Let $\{\boldsymbol{x}^k\}_{k=0}^{\infty}$ be generated by steepest descent method. Then, $\{\boldsymbol{x}^k\}_{k=0}^{\infty}$ converges to $\boldsymbol{x}^*$ for any initial point $\boldsymbol{x}^{(0)}$ if and only if $\sum_{k=0}^{\infty} \gamma_k = \infty$.

proof. $\because\ V(\boldsymbol{x}^{(k+1)}) = (1-\gamma_k)V(\boldsymbol{x}^{(k)}) \Longrightarrow V(\boldsymbol{x}^{(k+1)}) = \left(\prod_{i=0}^{k}(1-\gamma_i)\right) V(\boldsymbol{x}^{(0)})$

$$\therefore\ \boxed{\boldsymbol{x}^{(k)} \to \boldsymbol{x}^*} \iff \boxed{V(\boldsymbol{x}^{(k+1)}) \to 0} \iff \boxed{\prod_{i=0}^{k}(1-\gamma_i) \to 0}$$

$$\Updownarrow$$

$$\because \boxed{\sum_{i=0}^{\infty}\gamma_i = \infty} \iff \boxed{\sum_{i=0}^{\infty} -\log(1-\gamma_i) = \infty}$$

★ The last "⇔" is due to $-2x \le \log(1-x) \le -x$ in calculus.

Convergence Analysis

★ Question: how to guarantee $\sum_{k=0}^{\infty} \gamma_k = \infty$?

Lemma (Rayleigh's inequality)

Let $\lambda_{\max}(\boldsymbol{Q})$ and $\lambda_{\min}(\boldsymbol{Q})$ be the maximal and minimal eigenvalues of $\boldsymbol{Q} \succ 0$.

Then, $$\begin{cases} \lambda_{\min}(\boldsymbol{Q}) \le \dfrac{\|\boldsymbol{x}\|_{\boldsymbol{Q}}^2}{\|\boldsymbol{x}\|_2^2} \le \lambda_{\max}(\boldsymbol{Q}), \\ \dfrac{1}{\lambda_{\max}(\boldsymbol{Q})} \le \dfrac{\|\boldsymbol{x}\|_{\boldsymbol{Q}^{-1}}^2}{\|\boldsymbol{x}\|_2^2} \le \dfrac{1}{\lambda_{\min}(\boldsymbol{Q})}, \end{cases} \quad \forall \boldsymbol{x} \neq 0.$$

(proof: quadratic form in linear algebra.)

Lemma

For any $\boldsymbol{Q} \succ 0$ and $\boldsymbol{x} \neq 0$, $\dfrac{\lambda_{\min}(\boldsymbol{Q})}{\lambda_{\max}(\boldsymbol{Q})} \le \dfrac{\|\boldsymbol{x}\|_{\boldsymbol{Q}}^2 \|\boldsymbol{x}\|_{\boldsymbol{Q}^{-1}}^2}{\|\boldsymbol{x}\|_2^4} \le \dfrac{\lambda_{\max}(\boldsymbol{Q})}{\lambda_{\min}(\boldsymbol{Q})}$.

★ (stringent version) By Cauchy and Kantorovich inequalities:

For any $\boldsymbol{Q} \succ 0$ and $\boldsymbol{x} \neq 0$, $1 \le \dfrac{\|\boldsymbol{x}\|_{\boldsymbol{Q}}^2 \|\boldsymbol{x}\|_{\boldsymbol{Q}^{-1}}^2}{\|\boldsymbol{x}\|_2^4} \le \dfrac{[\lambda_{\max}(\boldsymbol{Q}) + \lambda_{\min}(\boldsymbol{Q})]^2}{4\lambda_{\max}(\boldsymbol{Q})\lambda_{\min}(\boldsymbol{Q})}$.

Convergence Analysis

Theorem

The sequence $\{\boldsymbol{x}^{(k)}\}$ generated by steepest descent method on minimizing quadratic function converges a minimizer $\boldsymbol{x}^*$ for any $\boldsymbol{x}^{(0)}$.

proof. assume $\boldsymbol{g}^{(k)} \neq 0$.

$$\because \quad \gamma_k = \frac{\alpha_k \|\boldsymbol{g}^{(k)}\|_{\boldsymbol{Q}}^2}{\|\boldsymbol{g}^{(k)}\|_{\boldsymbol{Q}^{-1}}^2} \left(2\frac{\|\boldsymbol{g}^{(k)}\|_2^2}{\|\boldsymbol{g}^{(k)}\|_{\boldsymbol{Q}}^2} - \alpha_k\right)$$

$$\xrightarrow{\alpha_k = \frac{\|\boldsymbol{g}^{(k)}\|_2^2}{\|\boldsymbol{g}^{(k)}\|_{\boldsymbol{Q}}^2}} \gamma_k = \frac{\|\boldsymbol{g}^{(k)}\|_2^4}{\|\boldsymbol{g}^{(k)}\|_{\boldsymbol{Q}}^2 \|\boldsymbol{g}^{(k)}\|_{\boldsymbol{Q}^{-1}}^2}.$$

$$\therefore \quad \frac{4\lambda_{\max}(\boldsymbol{Q})\lambda_{\min}(\boldsymbol{Q})}{[\lambda_{\max}(\boldsymbol{Q})+\lambda_{\min}(\boldsymbol{Q})]^2} \leq \gamma_k \leq 1$$

$$\iff \sum_{i=0}^{\infty} \gamma_k = \infty$$

$$\iff V(\boldsymbol{x}^{(k)}) \to 0$$

$$\iff \boldsymbol{x}^{(k)} \to \boldsymbol{x}^*.$$

Convergence Analysis

Notation: GD=gradient descent method.

Corollary (GD with fixed stepsize: $\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} - \alpha \boldsymbol{g}^{(k)}$)

The sequence $\{\boldsymbol{x}^{(k)}\}$ produced by GD with fixed stepsize on minimizing quadratic function converges to $\boldsymbol{x}^*$ for any $\boldsymbol{x}^{(0)}$ and $\alpha \in (0, \frac{2}{\lambda_{\max}(\boldsymbol{Q})})$.

proof. $\gamma_k = \frac{\alpha_k \|\boldsymbol{g}^{(k)}\|_{\boldsymbol{Q}}^2}{\|\boldsymbol{g}^{(k)}\|_{\boldsymbol{Q}^{-1}}^2} \left(2\frac{\|\boldsymbol{g}^{(k)}\|_2^2}{\|\boldsymbol{g}^{(k)}\|_{\boldsymbol{Q}}^2} - \alpha_k\right) \xrightarrow{\alpha_k \equiv \alpha} \gamma_k = \alpha \frac{\|\boldsymbol{g}^{(k)}\|_{\boldsymbol{Q}}^2}{\|\boldsymbol{g}^{(k)}\|_{\boldsymbol{Q}^{-1}}^2} \left(2\frac{\|\boldsymbol{g}^{(k)}\|_2^2}{\|\boldsymbol{g}^{(k)}\|_{\boldsymbol{Q}}^2} - \alpha\right).$

$\because \frac{\|\boldsymbol{g}^{(k)}\|_{\boldsymbol{Q}}^2}{\|\boldsymbol{g}^{(k)}\|_{\boldsymbol{Q}^{-1}}^2} = \frac{\|\boldsymbol{g}^{(k)}\|_{\boldsymbol{Q}}^2}{\|\boldsymbol{g}^{(k)}\|_2^2} \frac{\|\boldsymbol{g}^{(k)}\|_2^2}{\|\boldsymbol{g}^{(k)}\|_{\boldsymbol{Q}^{-1}}^2} \xrightarrow{\text{Rayleigh's inequality}} \frac{\|\boldsymbol{g}^{(k)}\|_{\boldsymbol{Q}}^2}{\|\boldsymbol{g}^{(k)}\|_{\boldsymbol{Q}^{-1}}^2} \geq [\lambda_{\min}(\boldsymbol{Q})]^2$

$\because \frac{\|\boldsymbol{g}^{(k)}\|_2^2}{\|\boldsymbol{g}^{(k)}\|_{\boldsymbol{Q}}^2} \geq \frac{1}{\lambda_{\max}(\boldsymbol{Q})}.$

$\therefore \gamma_k \geq \alpha[\lambda_{\min}(\boldsymbol{Q})]^2 \left(\frac{2}{\lambda_{\max}(\boldsymbol{Q})} - \alpha\right) \xrightarrow{\alpha \in (0, \frac{2}{\lambda_{\max}(\boldsymbol{Q})})} \text{Const} > 0$

$\therefore \sum_{i=0}^{\infty} \gamma_k = \infty \iff V(\boldsymbol{x}^{(k)}) \to 0 \iff \boldsymbol{x}^{(k)} \to \boldsymbol{x}^*.$

★ $\alpha \in (0, \frac{2}{\lambda_{\max}(\boldsymbol{Q})})$ is necessary and sufficient condition, i.e., any α beyond this interval can't guarantee the convergence (contradictory proof in textbook).

Convergence Analysis

Example (GD with fixed stepsize: $\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} - \alpha \boldsymbol{g}^{(k)}$)

Minimize $f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^\top \begin{bmatrix} 4 & 2\sqrt{2} \\ 0 & 5 \end{bmatrix} \boldsymbol{x} + \boldsymbol{x}^\top \begin{bmatrix} 3 \\ 6 \end{bmatrix} + 24.$

Ans: $\boldsymbol{Q} = \frac{1}{2}\left(\begin{bmatrix} 4 & 2\sqrt{2} \\ 0 & 5 \end{bmatrix} + \begin{bmatrix} 4 & 2\sqrt{2} \\ 0 & 5 \end{bmatrix}^\top \right) = \begin{bmatrix} 4 & \sqrt{2} \\ \sqrt{2} & 5 \end{bmatrix}$

$\Longrightarrow \lambda_{\max}(\boldsymbol{Q}) = 12$

$\therefore$ GD with fixed stepsize $\alpha \in (0, \frac{1}{6})$ converges to $\boldsymbol{x}^*$.

Convergence Analysis

Theorem (convergence rate of steepest descent method)

If steepest descent method minimizes quadratic function, then
$V(\boldsymbol{x}^{(k+1)}) \leq \left(\frac{\kappa-1}{\kappa+1}\right)^2 V(\boldsymbol{x}^{(k)})$.

proof. $\because\ V(\boldsymbol{x}^{(k+1)}) = (1-\gamma_k)\, V(\boldsymbol{x}^{(k)})$ and $\frac{4\lambda_{\max}(\boldsymbol{Q})\lambda_{\min}(\boldsymbol{Q})}{[\lambda_{\max}(\boldsymbol{Q})+\lambda_{\min}(\boldsymbol{Q})]^2} \leq \gamma_k$.

$$V(\boldsymbol{x}^{(k+1)}) \leq \frac{[\lambda_{\max}(\boldsymbol{Q})-\lambda_{\min}(\boldsymbol{Q})]^2}{[\lambda_{\max}(\boldsymbol{Q})+\lambda_{\min}(\boldsymbol{Q})]^2} V(\boldsymbol{x}^{(k)})$$

$$\xrightarrow{\kappa=\frac{\lambda_{\max}(\boldsymbol{Q})}{\lambda_{\min}(\boldsymbol{Q})}} V(\boldsymbol{x}^{(k+1)}) \leq \frac{[\kappa-1]^2}{[\kappa+1]^2} V(\boldsymbol{x}^{(k)}).$$

★ $c = \frac{[\kappa-1]^2}{[\kappa+1]^2}$ is called convergent rate. [converge fast if $c \downarrow 0$; and vice verse]

★ one step convergent (i.e., $c = 0$) $\Longleftrightarrow\ \lambda_{\max}(\boldsymbol{Q}) = \lambda_{\min}(\boldsymbol{Q})$.

Convergence Rate

Definition (Let $\{\boldsymbol{x}^k\}_{k=0}^{\infty}$ be a sequence convergeing to $\boldsymbol{x}^*$)

The order of convergence of the sequence $\{\boldsymbol{x}^k\}_{k=0}^{\infty}$ is defined by

$$\begin{cases} p, & \text{if } \lim_{k\to\infty} \dfrac{\|\boldsymbol{x}^{(k+1)}-\boldsymbol{x}^*\|_2}{\|\boldsymbol{x}^{(k)}-\boldsymbol{x}^*\|_2^p} \in (0,\infty), \\ \infty, & \text{if } \lim_{k\to\infty} \dfrac{\|\boldsymbol{x}^{(k+1)}-\boldsymbol{x}^*\|_2}{\|\boldsymbol{x}^{(k)}-\boldsymbol{x}^*\|_2^p} = 0,\ \forall p>0 \end{cases}$$

★ The large the p is, the faster the convergence of $\{\boldsymbol{x}^k\}_{k=0}^{\infty}$

particular cases: let $\lim_{k\to\infty} \dfrac{\|\boldsymbol{x}^{(k+1)}-\boldsymbol{x}^*\|_2}{\|\boldsymbol{x}^{(k)}-\boldsymbol{x}^*\|_2^p} = \mu$

- sublinear: $p=1$ and $\mu=1$;
- linear: $p=1$ and $\mu\in(0,1)$.
- superlinear: $p=1$ and $\mu=0$ (also $p>1$);
- quadratic linear: $p=2$.

$x^{(k)} = \frac{1}{k}$

$\because\ x^* = \lim_{k\to\infty} x^{(k)} = 0$

$$\therefore\ \lim_{k\to\infty} \frac{|x^{(k+1)}-x^*|}{|x^{(k)}-x^*|^p} = \lim_{k\to\infty} \frac{k^p}{k+1} = \begin{cases} 0, & p<1, \\ 1, & p=1, \\ \infty & p>1, \end{cases} \Rightarrow \text{1-order and sublinear}$$

Convergence Rate

$x^{(k)} = \gamma^k$ with $\gamma \in (0,1)$

$$\because \ x^* = \lim_{k\to\infty} x^{(k)} = 0$$

$$\therefore \ \lim_{k\to\infty} \frac{|x^{(k+1)} - x^*|}{|x^{(k)} - x^*|^p} = \lim_{k\to\infty} \gamma^{k(1-p)+1} \Longrightarrow \begin{cases} 0, & p < 1 \\ \gamma, & p = 1 \\ \infty, & p > 1 \end{cases} \Rightarrow \text{1-order and linear}$$

$x^{(k)} = \gamma^{q^k}$ with $q > 1$ and $\gamma \in (0,1)$

$$\because \ x^* = \lim_{k\to\infty} x^{(k)} = 0$$

$$\therefore \ \lim_{k\to\infty} \frac{|x^{(k+1)} - x^*|}{|x^{(k)} - x^*|^p} = \lim_{k\to\infty} \gamma^{(q-p)q^k} \Longrightarrow \begin{cases} 0, & p < q \\ 1, & p = q \\ \infty, & p > 1 \end{cases} \Rightarrow q\text{-order}$$

$x^{(k)} = 1$

$$\because \ x^* = \lim_{k\to\infty} x^{(k)} = 1, \ \therefore \ \lim_{k\to\infty} \frac{|x^{(k+1)} - x^*|}{|x^{(k)} - x^*|^p} = 0, \ \forall p > 0 \Rightarrow \infty\text{-order}$$

Convergence Rate

Definition

$a = O(h)$: $\exists c$ such that $|a| \le c|h|$ for sufficiently small h.
$a = o(h)$: $\frac{a}{h} \to 0$ as $h \to 0$.

Definition

If $\|\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^*\|_2 = O(\|\boldsymbol{x}^{(k)} - \boldsymbol{x}^*\|_2^p)$, then $\boldsymbol{x}^{(k)} \to \boldsymbol{x}^*$ in $\ge p$-order.
If $\|\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^*\|_2 = o(\|\boldsymbol{x}^{(k)} - \boldsymbol{x}^*\|_2^p)$, then $\boldsymbol{x}^{(k)} \to \boldsymbol{x}^*$ in $> p$-order.

Example (GD for minimizing $f(x) = x^2 - \frac{x^3}{3}$ with $\alpha_k \equiv \frac{1}{2}$, $x^{(0)} = 1$)

$\because\ x^{(k+1)} = x^{(k)} - \frac{1}{2} f'(x^{(k)}) \Longrightarrow x^{(k+1)} = \frac{1}{2}(x^{(k)})^2 \Longrightarrow x^* = 0.$

$\therefore\ \lim\limits_{k\to\infty} \frac{|x^{(k+1)} - x^*|}{|x^{(k)} - x^*|^2} = \lim\limits_{k\to\infty} \frac{|x^{(k+1)}|}{|x^{(k)}|^2} = \frac{1}{2} \Longrightarrow$ 2-order

★ $x^{(k+1)} = \frac{1}{2}(x^{(k)})^2 \xrightarrow{\text{why?}} x^{(k+1)} = \left(\frac{1}{2}\right)^{2^{k-1}}$

Convergence Rate

Question: For $V(\boldsymbol{x}^{(k+1)}) = (1-\gamma_k)V(\boldsymbol{x}^{(k)})$, when $\gamma_k = 1$?

Lemma

For steepest descent method, $\gamma_k = 1$ iff $\boldsymbol{g}^{(k)}$ is an eigenvector of $\boldsymbol{Q}$.

proof. ($\Leftarrow$) if $\boldsymbol{g}^{(k)}$ is an eigenvector of $\boldsymbol{Q} \Longrightarrow \gamma_k = \frac{\|\boldsymbol{g}^{(k)}\|_2^4}{\|\boldsymbol{g}^{(k)}\|_{\boldsymbol{Q}}^2 \|\boldsymbol{g}^{(k)}\|_{\boldsymbol{Q}^{-1}}^2} = 1.$

($\Rightarrow$) if $\gamma_k = 1 \Longrightarrow V(\boldsymbol{x}^{(k+1)}) = 0 \Longrightarrow \boldsymbol{x}^{(k+1)} = \boldsymbol{x}^*$

$\therefore \boldsymbol{x}^* = \boldsymbol{x}^{(k)} - \alpha_k \boldsymbol{g}^{(k)}$

$\therefore \ \boldsymbol{Q}\boldsymbol{x}^* = \boldsymbol{Q}\boldsymbol{x}^{(k)} - \alpha_k \boldsymbol{Q}\boldsymbol{g}^{(k)}$, i.e.,

$\frac{1}{\alpha_k}\boldsymbol{g}^{(k)} = \boldsymbol{Q}\boldsymbol{g}^{(k)} \Longrightarrow \boldsymbol{g}^{(k)}$ is an eigenvector of $\boldsymbol{Q}$

★ (inversion of Lemma) if $\boldsymbol{g}^{(k)}$ isn't an eigenvector of $\boldsymbol{Q}$, then $\gamma_k \leq 1$.

homework

exercise in textbook: 8.4, 8.10, 8.16