MA 5037: Optimization Methods and Applications
Chapter 4: The Gradient Method

Suh-Yuh Yang (楊肅煜)

Department of Mathematics, National Central University
Jhongli District, Taoyuan City 32001, Taiwan

syyang@math.ncu.edu.tw
http://www.math.ncu.edu.tw/∼syyang/

Descent direction methods

We consider the unconstrained minimization problem:

$$\min\{f(\boldsymbol{x}) : \boldsymbol{x} \in \mathbb{R}^n\},$$

where the objective function is *continuously differentiable* over $\mathbb{R}^n$. We will consider an iterative algorithm for finding stationary points of f. The iterative algorithm takes the form

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + t_k \boldsymbol{d}_k, \quad k = 0, 1, \cdots,$$

where $\boldsymbol{d}_k$ is the direction and t_k is the stepsize.

Definition: *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a continuously differentiable function. A vector $\mathbf{0} \neq \boldsymbol{d} \in \mathbb{R}^n$ is called a descent direction of f at $\boldsymbol{x}$ if the directional derivative $f'(\boldsymbol{x}; \boldsymbol{d}) < 0$.* (Note that $f'(\boldsymbol{x}; \boldsymbol{d}) = \nabla f(\boldsymbol{x})^\top \boldsymbol{d}$)

Descent property: *If $\boldsymbol{d}$ is a descent direction of f at $\boldsymbol{x}$, then $\exists\, \varepsilon > 0$ such that $f(\boldsymbol{x} + t\boldsymbol{d}) < f(\boldsymbol{x})$ for any $t \in (0, \varepsilon]$.* $\square$

Taking small enough steps along these descent directions lead to a decrease of the objective function.

Schematic descent direction method

Initialization: Pick $\boldsymbol{x}_0 \in \mathbb{R}^n$.
General step: For any $k = 0, 1, \cdots$, set

(a) Pick a descent direction $\boldsymbol{d}_k$.

(b) Find a stepsize t_k satisfying $f(\boldsymbol{x}_k + t_k \boldsymbol{d}_k) < f(\boldsymbol{x}_k)$.

(c) Set $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + t_k \boldsymbol{d}_k$

(d) If a stopping criterion is satisfied then stop, $\boldsymbol{x}_{k+1}$ is the output.

The descent direction method remains *"conceptual"* and cannot be implemented. Many details are missing in the above description:

- *What is the starting point $\boldsymbol{x}_0$?*
- *How to choose the descent direction $\boldsymbol{d}_k$?*
- *What stepsize should be taken t_k?*
- *What is the stopping criterion?*

Three popular choices of stepsize t_k

The process of finding t_k is called *line search,* since it is essentially a minimization procedure on the 1-D function $g(t) := f(\boldsymbol{x}_k + t\boldsymbol{d}_k)$.

- **constant stepsize:** $t_k = \bar{t}$ for any k.
- **exact line search:** t_k is a minimizer of f along the ray $\boldsymbol{x}_k + t\boldsymbol{d}_k$:

$$t_k \in \arg\min_{t \geq 0} f(\boldsymbol{x}_k + t\boldsymbol{d}_k).$$

- **backtracking:** The method requires three parameters: $s > 0$ (*not too small*), $\alpha, \beta \in (0, 1)$.

```
set   t_k ← s
while f(x_k) − f(x_k + t_k d_k) < −α t_k ∇f(x_k)^T d_k   do
      set   t_k ← β t_k
```

$$\texttt{while}\ f(\boldsymbol{x}_k) - f(\boldsymbol{x}_k + t_k\boldsymbol{d}_k) < -\alpha t_k \overbrace{\nabla f(\boldsymbol{x}_k)^\top \boldsymbol{d}_k}^{f'(\boldsymbol{x}_k;\boldsymbol{d}_k)}\ \texttt{do}$$

Therefore, the stepsize is chosen as $t_k = s\beta^{i_k}$, where i_k is the smallest nonnegative integer for which $(\star)$ is satisfied:

$$f(\boldsymbol{x}_k) - f(\boldsymbol{x}_k + s\beta^{i_k}\boldsymbol{d}_k) \geq -\alpha s\beta^{i_k}\nabla f(\boldsymbol{x}_k)^\top \boldsymbol{d}_k. \qquad (\star)$$

The third option is in a sense a compromise between the other twos.

Validity of the sufficient decrease condition $(\star)$

Theorem: *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a continuously differentiable function and $\boldsymbol{x} \in \mathbb{R}^n$. Assume that $\boldsymbol{0} \neq \boldsymbol{d} \in \mathbb{R}^n$ is a descent direction of f at $\boldsymbol{x}$ and let $\alpha \in (0,1)$. Then $\exists\, \varepsilon > 0$ such that for all $t \in [0, \varepsilon]$, we have*

$$f(\boldsymbol{x}) - f(\boldsymbol{x} + t\boldsymbol{d}) \geq -\alpha t \nabla f(\boldsymbol{x})^\top \boldsymbol{d}.$$

Proof: Since f is continuously differentiable it follows that

$$f(\boldsymbol{x} + t\boldsymbol{d}) = f(\boldsymbol{x}) + t\nabla f(\boldsymbol{x})^\top \boldsymbol{d} + o(t\|\boldsymbol{d}\|),$$

and hence

$$f(\boldsymbol{x}) - f(\boldsymbol{x} + t\boldsymbol{d}) = -\alpha t \nabla f(\boldsymbol{x})^\top \boldsymbol{d} - (1-\alpha) t \nabla f(\boldsymbol{x})^\top \boldsymbol{d} - o(t\|\boldsymbol{d}\|).$$

Since $\boldsymbol{d}$ is a descent direction of f at $\boldsymbol{x}$, we have

$$\lim_{t \to 0^+} \frac{-(1-\alpha) t \nabla f(\boldsymbol{x})^\top \boldsymbol{d} - o(t\|\boldsymbol{d}\|)}{t} = -(1-\alpha) \nabla f(\boldsymbol{x})^\top \boldsymbol{d} > 0.$$

Hence, $\exists\, \varepsilon > 0$ such that for all $t \in (0, \varepsilon]$, we have

$$-(1-\alpha) t \nabla f(\boldsymbol{x})^\top \boldsymbol{d} - o(t\|\boldsymbol{d}\|) > 0,$$

which implies the desired result. □

Example: exact line search for quadratic functions

Let $f(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x} + 2\boldsymbol{b}^\top \boldsymbol{x} + c$, where $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ is a symmetric positive definite matrix, $\boldsymbol{b} \in \mathbb{R}^n$, and $c \in \mathbb{R}$. Let $\boldsymbol{x} \in \mathbb{R}^n$ and let $\boldsymbol{d} \in \mathbb{R}^n$ be a descent direction of f at $\boldsymbol{x}$. The exact line search for the stepsize can be obtained by considering

$$\min_{t\geq 0}\{g(t) := f(\boldsymbol{x} + t\boldsymbol{d})\}.$$

By a direct computation, we have

$$\begin{aligned} g(t) = f(\boldsymbol{x} + t\boldsymbol{d}) &= (\boldsymbol{d}^\top \boldsymbol{A}\boldsymbol{d})t^2 + 2(\boldsymbol{d}^\top \boldsymbol{A}\boldsymbol{x} + \boldsymbol{d}^\top \boldsymbol{b})t + \boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x} + 2\boldsymbol{b}^\top \boldsymbol{x} + c \\ &= (\boldsymbol{d}^\top \boldsymbol{A}\boldsymbol{d})t^2 + 2(\boldsymbol{d}^\top \boldsymbol{A}\boldsymbol{x} + \boldsymbol{d}^\top \boldsymbol{b})t + f(\boldsymbol{x}). \end{aligned}$$

Since $g'(t) = 2(\boldsymbol{d}^\top \boldsymbol{A}\boldsymbol{d})t + 2\boldsymbol{d}^\top(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b})$ and $\nabla f(\boldsymbol{x}) = 2(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b})$, it follows that $g'(t) = 0$ if and only if

$$t = t^* := -\frac{\boldsymbol{d}^\top \nabla f(\boldsymbol{x})}{2\boldsymbol{d}^\top \boldsymbol{A}\boldsymbol{d}} > 0,$$

where since $\boldsymbol{d}$ is a descent direction of f at $\boldsymbol{x}$, $f'(\boldsymbol{x}; \boldsymbol{d}) = \boldsymbol{d}^\top \nabla f(\boldsymbol{x}) < 0$ and then $t^* > 0$.

In what direction f decreases most rapidly?

- Making an observation, for $n = 2$, we have
$$f'(\boldsymbol{x}_k; \boldsymbol{d}_k) = \langle \nabla f(\boldsymbol{x}_k), \boldsymbol{d}_k \rangle = \|\nabla f(\boldsymbol{x}_k)\| \|\boldsymbol{d}_k\| \cos \theta_k,$$
where θ_k is the angle between the vectors $\nabla f(\boldsymbol{x}_k)$ and $\boldsymbol{d}_k$. Therefore, f decreases most rapidly when $\theta_k = \pi$, i.e., in the direction of $-\nabla f(\boldsymbol{x}_k)$ whenever $\nabla f(\boldsymbol{x}_k) \neq \mathbf{0}$.
- *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a continuously differentiable function and let $\boldsymbol{x} \in \mathbb{R}^n$ be a nonstationary point, $\nabla f(\boldsymbol{x}) \neq \mathbf{0}$. Then an optimal solution of* $\min_{\boldsymbol{d} \in \mathbb{R}^n} \{f'(\boldsymbol{x}; \boldsymbol{d}) : \|\boldsymbol{d}\| = 1\}$ *is* $\boldsymbol{d} = -\frac{\nabla f(\boldsymbol{x})}{\|\nabla f(\boldsymbol{x})\|}$.

  Proof: By the Cauchy-Schwarz inequality, for $\|\boldsymbol{d}\| = 1$, we have
$$f'(\boldsymbol{x}; \boldsymbol{d}) = \nabla f(\boldsymbol{x}_k)^\top \boldsymbol{d} \geq -\|\nabla f(\boldsymbol{x})\| \|\boldsymbol{d}\| = -\|\nabla f(\boldsymbol{x})\| \quad \leftarrow \text{a lower bound}$$
  Taking $\boldsymbol{d} = -\frac{\nabla f(\boldsymbol{x})}{\|\nabla f(\boldsymbol{x})\|}$, we attain the lower bound. $\square$

The gradient method

- *In the gradient method, we take* $\mathbf{d}_k = -\nabla f(\boldsymbol{x}_k)$, *provided* $\nabla f(\boldsymbol{x}_k) \neq \mathbf{0}$.

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

- **The gradient method**

Input: Tolerance parameter $\varepsilon > 0$.

Initialization: Pick $\boldsymbol{x}_0 \in \mathbb{R}^n$ arbitrarily.

General step: For any $k = 0, 1, \cdots$, execute

(a) Pick a stepsize t_k by a line search procedure on the function

$$g(t) := f(\boldsymbol{x}_k - t\nabla f(\boldsymbol{x}_k)).$$

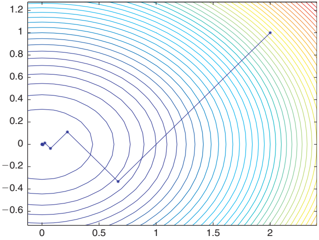
(b) Set $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - t_k \nabla f(\boldsymbol{x}_k)$.

(c) If $\|\nabla f(\boldsymbol{x}_{k+1})\| \leq \varepsilon$ then stop and $\boldsymbol{x}_{k+1}$ is the output.

Example

Consider the 2-D minimization problem $\min_{x,y}(x^2 + 2y^2)$ whose optimal solution is $(x, y) = (0, 0)$ with corresponding optimal value 0.

- MATLAB function: `gradient_method_quadratic(···)`
  For solving $\min_{\boldsymbol{x}\in\mathbb{R}^n}\{\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x} + 2\boldsymbol{b}^\top \boldsymbol{x}\}$, $\boldsymbol{A} \succ \boldsymbol{0}$, exact line search.



- MATLAB function: `gradient_method_constant(···)`
- MATLAB function: `gradient_method_backtracking(···)`
  In computational experience, backtracking does not have real disadvantages in comparison to exact line search!

The gradient method: zig-zag effect

The zig-zag effect: *Let $\{\boldsymbol{x}_k\}$ be the sequence generated by the gradient method with exact line search for solving a problem of minimizing a continuously differentiable function f. Then for any $k = 0, 1, 2, \ldots$*

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

Proof: By the definition of the gradient method, we have

$$\boldsymbol{x}_{k+1} - \boldsymbol{x}_k = -t_k \nabla f(\boldsymbol{x}_k), \quad \boldsymbol{x}_{k+2} - \boldsymbol{x}_{k+1} = -t_{k+1} \nabla f(\boldsymbol{x}_{k+1}).$$

Therefore, we wish to prove that $\nabla f(\boldsymbol{x}_k)^\top \nabla f(\boldsymbol{x}_{k+1}) = 0$. Since

$$g(t) := f(\boldsymbol{x}_k - t\nabla f(\boldsymbol{x}_k)),$$

we have

$$0 = g'(t_k) = -\nabla f(\boldsymbol{x}_k)^\top \nabla f(\boldsymbol{x}_k - t_k \nabla f(\boldsymbol{x}_k)).$$

That is,

$$-\nabla f(\boldsymbol{x}_k)^\top \nabla f(\boldsymbol{x}_{k+1}) = 0,$$

which is the desired result. □ *(see the figure on page 9)*

A quadratic minimization problem

Consider the simple quadratic minimization problem

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} \{f(\boldsymbol{x}) := \boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}\},$$

where $\boldsymbol{A} \in \mathbb{R}^{n\times n}$, $\boldsymbol{A} \succ \boldsymbol{0}$. The optimal solution is obviously $\boldsymbol{x}^* = \boldsymbol{0}$. The gradient method with exact line search takes the form

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + t_k\boldsymbol{d}_k, \quad \boldsymbol{d}_k = -\nabla f(\boldsymbol{x}_k) = -2\boldsymbol{A}\boldsymbol{x}_k, \quad t_k = \frac{\boldsymbol{d}_k^\top \boldsymbol{d}_k}{2\boldsymbol{d}_k^\top \boldsymbol{A}\boldsymbol{d}_k}.$$

Then we have

$$\begin{aligned}
f(\boldsymbol{x}_{k+1}) &= \boldsymbol{x}_{k+1}^\top \boldsymbol{A}\boldsymbol{x}_{k+1} = (\boldsymbol{x}_k + t_k\boldsymbol{d}_k)^\top \boldsymbol{A}(\boldsymbol{x}_k + t_k\boldsymbol{d}_k) \\
&= \boldsymbol{x}_k^\top \boldsymbol{A}\boldsymbol{x}_k + 2t_k\boldsymbol{d}_k^\top \boldsymbol{A}\boldsymbol{x}_k + t_k^2\boldsymbol{d}_k^\top \boldsymbol{A}\boldsymbol{d}_k \\
&= \boldsymbol{x}_k^\top \boldsymbol{A}\boldsymbol{x}_k - t_k\boldsymbol{d}_k^\top \boldsymbol{d}_k + t_k^2\boldsymbol{d}_k^\top \boldsymbol{A}\boldsymbol{d}_k \\
&= \boldsymbol{x}_k^\top \boldsymbol{A}\boldsymbol{x}_k - \frac{1}{4}\frac{(\boldsymbol{d}_k^\top \boldsymbol{d}_k)^2}{\boldsymbol{d}_k^\top \boldsymbol{A}\boldsymbol{d}_k} \\
&= \boldsymbol{x}_k^\top \boldsymbol{A}\boldsymbol{x}_k\left(1 - \frac{1}{4}\frac{(\boldsymbol{d}_k^\top \boldsymbol{d}_k)^2}{(\boldsymbol{d}_k^\top \boldsymbol{A}\boldsymbol{d}_k)(\boldsymbol{x}_k^\top \boldsymbol{A}\boldsymbol{A}^{-1}\boldsymbol{A}\boldsymbol{x}_k)}\right).
\end{aligned}$$

Kantorovich inequality

Since $\boldsymbol{d}_k = -2\boldsymbol{A}\boldsymbol{x}_k$, we have

$$f(\boldsymbol{x}_{k+1}) = \left(1 - \frac{(\boldsymbol{d}_k^\top \boldsymbol{d}_k)^2}{(\boldsymbol{d}_k^\top \boldsymbol{A}\boldsymbol{d}_k)(\boldsymbol{d}_k^\top \boldsymbol{A}^{-1}\boldsymbol{d}_k)}\right) f(\boldsymbol{x}_k).$$

Kantorovich inequality: *Let $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ and $\boldsymbol{A} \succ \boldsymbol{0}$. Then $\forall\, \boldsymbol{0} \neq \boldsymbol{x} \in \mathbb{R}^n$,*

$$\frac{(\boldsymbol{x}^\top \boldsymbol{x})^2}{(\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x})(\boldsymbol{x}^\top \boldsymbol{A}^{-1}\boldsymbol{x})} \geq \frac{4\lambda_{\max}(\boldsymbol{A})\lambda_{\min}(\boldsymbol{A})}{(\lambda_{\max}(\boldsymbol{A}) + \lambda_{\min}(\boldsymbol{A}))^2}.$$

Proof: Let $m := \lambda_{\min}(\boldsymbol{A}) > 0$ and $M := \lambda_{\max}(\boldsymbol{A}) > 0$. Then the eigenvalues of $\boldsymbol{A} + Mm\boldsymbol{A}^{-1}$ are $\lambda_i(\boldsymbol{A}) + \frac{Mm}{\lambda_i(\boldsymbol{A})}$, $i = 1, 2, \cdots, n$. The maximum value of the 1-D function $\varphi(t) = t + \frac{Mm}{t}$ on $[m, M]$ can be attained at $t = m$ and $t = M$ and the value is $M + m$. Therefore, the eigenvalues of $\boldsymbol{A} + Mm\boldsymbol{A}^{-1}$ are smaller than $M + m$. Thus

$$\boldsymbol{A} + Mm\boldsymbol{A}^{-1} \preceq (M + m)\boldsymbol{I},$$

which implies that

$$\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x} + Mm(\boldsymbol{x}^\top \boldsymbol{A}^{-1}\boldsymbol{x}) \leq (M + m)(\boldsymbol{x}^\top \boldsymbol{x}).$$

Using the inequality $\alpha\beta \leq \frac{1}{4}(\alpha + \beta)^2$, we obtain the desired result

$$(\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x})(Mm(\boldsymbol{x}^\top \boldsymbol{A}^{-1}\boldsymbol{x})) \leq \frac{1}{4}\left(\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x} + Mm(\boldsymbol{x}^\top \boldsymbol{A}^{-1}\boldsymbol{x})\right)^2 \leq \frac{(M + m)^2}{4}(\boldsymbol{x}^\top \boldsymbol{x})^2. \quad \square$$

Convergence rate analysis

Coming back to the convergence rate analysis of the gradient method on the quadratic minimization problem, we have

$$\begin{aligned} f(\boldsymbol{x}_{k+1}) &= \left(1 - \frac{(\boldsymbol{d}_k^\top \boldsymbol{d}_k)^2}{(\boldsymbol{d}_k^\top \boldsymbol{A}\boldsymbol{d}_k)(\boldsymbol{d}_k^\top \boldsymbol{A}^{-1}\boldsymbol{d}_k)}\right) f(\boldsymbol{x}_k) \\ &\le \left(1 - \frac{4Mm}{(M+m)^2}\right) f(\boldsymbol{x}_k) = \left(\frac{M-m}{M+m}\right)^2 f(\boldsymbol{x}_k), \end{aligned}$$

which implies *a linear rate* to the optimal value,

$$|f(\boldsymbol{x}_{k+1}) - 0| = f(\boldsymbol{x}_{k+1}) \le c f(\boldsymbol{x}_k) = c|f(\boldsymbol{x}_k) - 0| \text{ and } f(\boldsymbol{x}_k) \le c^k f(\boldsymbol{x}_0),$$

$$c := \left(\frac{M-m}{M+m}\right)^2 = \left(\frac{\chi - 1}{\chi + 1}\right)^2 < 1,\ \chi := \frac{M}{m} = \frac{\lambda_{\max}(\boldsymbol{A})}{\lambda_{\min}(\boldsymbol{A})}.$$

Definition: $\chi(\boldsymbol{A}) := \dfrac{\lambda_{\max}(\boldsymbol{A})}{\lambda_{\min}(\boldsymbol{A})}$ *is called the condition number of* $\boldsymbol{A}$.

Note: Although the condition number can be defined for general matrices, here we restrict ourselves to SPD real matrices.

Nonquadratic objective functions

- Matrices with large condition number are called *ill-conditioned.* Matrices with small condition number are called *well-conditioned.*
- The entire discussion until now was on the restrictive class of quadratic objective functions, where the Hessian matrix is constant, *but the notion of condition number also appears in the context of nonquadratic objective functions. In that case, it is well known that the rate of convergence of* $\boldsymbol{x}_k$ *to a given stationary point* $\boldsymbol{x}^*$ *depends on the condition number of* $\chi(\nabla^2 f(\boldsymbol{x}^*))$.
- We will not focus on these theoretical results, but will illustrate it on a well-known ill-conditioned problem, *the Rosenbrock function,* see next page.

The Rosenbrock function (control theory)

The Rosenbrock function is $f(x_1, x_2) := 100(x_2 - x_1^2)^2 + (1 - x_1)^2$.

- The optimal solution (global minimum) is $(x_1, x_2) = (1, 1)$ with corresponding optimal value 0.
- The gradient and Hessian of f are respectively

$$\begin{aligned}\nabla f(x_1, x_2) &= \begin{bmatrix} -400x_1(x_2 - x_1^2) - 2(1 - x_1) \\ 200(x_2 - x_1^2) \end{bmatrix}, \\ \nabla^2 f(x_1, x_2) &= \begin{bmatrix} -400x_2 + 1200x_1^2 + 2 & -400x_1 \\ -400x_1 & 200 \end{bmatrix}.\end{aligned}$$

- $(x_1, x_2) = (1, 1)$ is the unique stationary point and

$$\nabla^2 f(1, 1) = \begin{bmatrix} 802 & -400 \\ -400 & 200 \end{bmatrix}.$$

```
>> A = [802, -400; -400, 200];
>> cond(A)
ans = 2.5080e+003
```

A condition number of more than 2500 (ill-conditioned) should have severe effects on the convergence speed of the gradient method.

Sensitivity of solutions to linear systems

- We are given a linear system $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$, where $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ is symmetric, $\boldsymbol{A} \succ \boldsymbol{0}$ and $\boldsymbol{b} \in \mathbb{R}^n$. Then the solution is $\boldsymbol{x} = \boldsymbol{A}^{-1}\boldsymbol{b}$.
- We consider a perturbation $\boldsymbol{b} + \Delta\boldsymbol{b}$ in the RHS. The new solution is denoted by $\boldsymbol{x} + \Delta\boldsymbol{x}$, i.e., $\boldsymbol{A}(\boldsymbol{x} + \Delta\boldsymbol{x}) = \boldsymbol{b} + \Delta\boldsymbol{b}$. We have $\boldsymbol{x} + \Delta\boldsymbol{x} = \boldsymbol{A}^{-1}(\boldsymbol{b} + \Delta\boldsymbol{b}) = \boldsymbol{x} + \boldsymbol{A}^{-1}\Delta\boldsymbol{b}$. Then

$$\begin{aligned}
\frac{\|\Delta\boldsymbol{x}\|}{\|\boldsymbol{x}\|} &= \frac{\|\boldsymbol{A}^{-1}\Delta\boldsymbol{b}\|}{\|\boldsymbol{x}\|} \le \frac{\|\boldsymbol{A}^{-1}\|\,\|\Delta\boldsymbol{b}\|}{\|\boldsymbol{x}\|} = \frac{\lambda_{\max}(\boldsymbol{A}^{-1})\,\|\Delta\boldsymbol{b}\|}{\|\boldsymbol{x}\|} \\
&= \frac{1}{\lambda_{\min}(\boldsymbol{A})}\frac{\|\Delta\boldsymbol{b}\|}{\|\boldsymbol{x}\|} = \frac{1}{\lambda_{\min}(\boldsymbol{A})}\frac{\|\Delta\boldsymbol{b}\|}{\|\boldsymbol{A}^{-1}\boldsymbol{b}\|} \le \frac{1}{\lambda_{\min}(\boldsymbol{A})}\frac{\|\Delta\boldsymbol{b}\|}{\lambda_{\min}(\boldsymbol{A}^{-1})\,\|\boldsymbol{b}\|} \\
&= \frac{\lambda_{\max}(\boldsymbol{A})}{\lambda_{\min}(\boldsymbol{A})}\frac{\|\Delta\boldsymbol{b}\|}{\|\boldsymbol{b}\|} = \chi(\boldsymbol{A})\frac{\|\Delta\boldsymbol{b}\|}{\|\boldsymbol{b}\|}, \quad \text{where we have used}
\end{aligned}$$

$$\|\boldsymbol{A}^{-1}\boldsymbol{b}\| = \sqrt{\boldsymbol{b}^\top\boldsymbol{A}^{-2}\boldsymbol{b}} \ge \sqrt{\lambda_{\min}(\boldsymbol{A}^{-2})\|\boldsymbol{b}\|^2} = \lambda_{\min}(\boldsymbol{A}^{-1})\|\boldsymbol{b}\|.$$

- *We can therefore deduce that the sensitivity of the solution of the linear system to right-hand-side perturbations depends on the condition number of the coefficients matrix.*

Scaling for ill-conditioned problems

We consider the unconstrained minimization problem:

$$\min\{f(\boldsymbol{x}) : \boldsymbol{x} \in \mathbb{R}^n\}.$$

Let $\boldsymbol{S} \in \mathbb{R}^{n\times n}$ be a nonsingular matrix. Let $\boldsymbol{y} := \boldsymbol{S}^{-1}\boldsymbol{x}$. Then $\boldsymbol{x} = \boldsymbol{S}\boldsymbol{y}$ and we obtain the equivalent problem:

$$\min\{g(\boldsymbol{y}) := f(\boldsymbol{S}\boldsymbol{y}) : \boldsymbol{y} \in \mathbb{R}^n\}.$$

Since $\nabla_y g(\boldsymbol{y}) = \boldsymbol{S}^\top \nabla f(\boldsymbol{S}\boldsymbol{y}) = \boldsymbol{S}^\top \nabla f(\boldsymbol{x})$, the gradient method for solving $\min\limits_{\boldsymbol{y}\in\mathbb{R}^n} g(\boldsymbol{y})$ takes the form:

$$\boldsymbol{y}_{k+1} = \boldsymbol{y}_k - t_k \boldsymbol{S}^\top \nabla f(\boldsymbol{S}\boldsymbol{y}_k).$$

Multiplying $\boldsymbol{S}$ and letting $\boldsymbol{x}_k := \boldsymbol{S}\boldsymbol{y}_k$, we have

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - t_k \boldsymbol{S}\boldsymbol{S}^\top \nabla f(\boldsymbol{x}_k).$$

Defining $\boldsymbol{D} := \boldsymbol{S}\boldsymbol{S}^\top$, we obtain *the scaled gradient method* with scaling matrix $\boldsymbol{D}$:

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - t_k \boldsymbol{D} \nabla f(\boldsymbol{x}_k).$$

The scaled gradient

- The matrix $\boldsymbol{D} = \boldsymbol{S}\boldsymbol{S}^\top$ is positive definite (cf. Exercise 2.6). The direction $-\boldsymbol{D}\nabla f(\boldsymbol{x}_k)$ is a descent of f at $\boldsymbol{x}_k$ when $\nabla f(\boldsymbol{x}_k) \neq \boldsymbol{0}$ since

$$f'(\boldsymbol{x}_k; -\boldsymbol{D}\nabla f(\boldsymbol{x}_k)) = -\nabla f(\boldsymbol{x}_k)^\top \boldsymbol{D}\nabla f(\boldsymbol{x}_k) < 0.$$

- To summarize the above discussion, we have shown that the scaled gradient method with scaling matrix $\boldsymbol{D} \succ \boldsymbol{0}$ is equivalent to the gradient method employed on the function

$$g(\boldsymbol{y}) = f(\boldsymbol{D}^{1/2}\boldsymbol{y}),$$

where $\boldsymbol{y} := \boldsymbol{D}^{-1/2}\boldsymbol{x}$ ($\Longleftrightarrow \boldsymbol{x} = \boldsymbol{D}^{1/2}\boldsymbol{y}$). We note that the gradient and Hessian of g are given by

$$\begin{aligned} \nabla_y g(\boldsymbol{y}) &= \boldsymbol{D}^{1/2}\nabla f(\boldsymbol{D}^{1/2}\boldsymbol{y}) = \boldsymbol{D}^{1/2}\nabla_x f(\boldsymbol{x}), \\ \nabla_y^2 g(\boldsymbol{y}) &= \boldsymbol{D}^{1/2}\nabla^2 f(\boldsymbol{D}^{1/2}\boldsymbol{y})\boldsymbol{D}^{1/2} = \boldsymbol{D}^{1/2}\nabla_x^2 f(\boldsymbol{x})\boldsymbol{D}^{1/2}. \end{aligned}$$

The scaled gradient method

Input: Tolerance parameter $\varepsilon > 0$.

Initialization: Pick $\boldsymbol{x}_0 \in \mathbb{R}^n$ arbitrarily.

General step: For any $k = 0, 1, \cdots$, execute

(a) Pick a scaling matrix $\boldsymbol{D}_k \succ \boldsymbol{0}$.

(b) Pick a stepsize t_k by a line search procedure on the function

$$h(t) := f(\boldsymbol{x}_k - t\boldsymbol{D}_k \nabla f(\boldsymbol{x}_k)).$$

(c) Set $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - t_k \boldsymbol{D}_k \nabla f(\boldsymbol{x}_k)$.

(d) If $\|\nabla f(\boldsymbol{x}_{k+1})\| \leq \varepsilon$ then stop and $\boldsymbol{x}_{k+1}$ is the output.

It is often beneficial to choose the scaling matrix differently at each iteration.

How to choose the D_k? damped Newton's method

- To accelerate the rate of convergence of $\{x_k\}$, which depends on the condition number of the scaled Hessian $D_k^{1/2}\nabla^2 f(x_k)D_k^{1/2}$. The scaling matrix is often chosen to make this scaled Hessian to be as close as possible to the identity matrix.
- When $\nabla^2 f(x_k) \succ \mathbf{0}$, we choose $D_k = (\nabla^2 f(x_k))^{-1}$ and the scaled Hessian becomes the identity matrix. *The resulting method is the so-called damped Newton's method:*

$$\boxed{x_{k+1} = x_k - t_k(\nabla^2 f(x_k))^{-1}\nabla f(x_k)}$$

  One difficulty associated with damped Newton's method is that it requires full knowledge of the Hessian.
- The term $(\nabla^2 f(x_k))^{-1}\nabla f(x_k)$ suggests that a linear system of the form $\nabla^2 f(x_k)d = \nabla f(x_k)$ needs to be solved at each iteration, which might be costly from a computational point of view.
- The simplest of all scaling matrices are diagonal matrices. A natural choice for diagonal elements is $D_{ii} = (\nabla^2 f(x_k))_{ii}^{-1}$.

The Gauss-Newton method

We consider the nonlinear least squares (NLS) problem:

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} \{g(\boldsymbol{x}) := \sum_{i=1}^{m} (f_i(\boldsymbol{x}) - c_i)^2\},$$

where $f_1, f_2, \cdots, f_m$ are continuously differentiable over $\mathbb{R}^n$ and $c_1, c_2, \cdots, c_m \in \mathbb{R}$. The problem can be reformulated as

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} \|F(\boldsymbol{x})\|^2,$$

where the vector-valued function F is given by

$$F(\boldsymbol{x}) := \begin{bmatrix} f_1(\boldsymbol{x}) - c_1 \\ f_2(\boldsymbol{x}) - c_2 \\ \vdots \\ f_m(\boldsymbol{x}) - c_m \end{bmatrix}.$$

The Gauss-Newton method (A linearization method):

Given the iterate $\boldsymbol{x}_k$, find

$$\boldsymbol{x}_{k+1} := \arg\min_{\boldsymbol{x}\in\mathbb{R}^n} \Big\{ \sum_{i=1}^{m} \Big(f_i(\boldsymbol{x}_k) + \nabla f_i(\boldsymbol{x}_k)^\top (\boldsymbol{x} - \boldsymbol{x}_k) - c_i \Big)^2 \Big\}.$$

The Gauss-Newton method (cont'd)

The minimization problem is essentially a linear LS problem:

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} \|\boldsymbol{A}_k\boldsymbol{x} - \boldsymbol{b}_k\|^2,$$

where

$$\boldsymbol{A}_k = \begin{bmatrix} \nabla f_1(\boldsymbol{x}_k)^\top \\ \nabla f_2(\boldsymbol{x}_k)^\top \\ \vdots \\ \nabla f_m(\boldsymbol{x}_k)^\top \end{bmatrix} := J(\boldsymbol{x}_k),$$

is the so-called Jacobian matrix and

$$\boldsymbol{b}_k = \begin{bmatrix} \nabla f_1(\boldsymbol{x}_k)^\top \boldsymbol{x}_k - f_1(\boldsymbol{x}_k) + c_1 \\ \nabla f_2(\boldsymbol{x}_k)^\top \boldsymbol{x}_k - f_2(\boldsymbol{x}_k) + c_2 \\ \vdots \\ \nabla f_m(\boldsymbol{x}_k)^\top \boldsymbol{x}_k - f_m(\boldsymbol{x}_k) + c_m \end{bmatrix} := J(\boldsymbol{x}_k)\boldsymbol{x}_k - F(\boldsymbol{x}_k).$$

The underlying assumption is that $J(\boldsymbol{x}_k)$ is of a full column rank; otherwise the minimization will not produce a unique minimizer.

The Gauss-Newton method (cont'd)

We write an explicit expression for the Gauss-Newton iterates (see Chapter 3)

$$\boldsymbol{x}_{k+1} = (J(\boldsymbol{x}_k)^\top J(\boldsymbol{x}_k))^{-1} J(\boldsymbol{x}_k)^\top \boldsymbol{b}_k.$$

The method can also be written as

$$\begin{aligned}
\boldsymbol{x}_{k+1} &= (J(\boldsymbol{x}_k)^\top J(\boldsymbol{x}_k))^{-1} J(\boldsymbol{x}_k)^\top \boldsymbol{b}_k \\
&= (J(\boldsymbol{x}_k)^\top J(\boldsymbol{x}_k))^{-1} J(\boldsymbol{x}_k)^\top (J(\boldsymbol{x}_k)\boldsymbol{x}_k - F(\boldsymbol{x}_k)) \\
&= \boldsymbol{x}_k - (J(\boldsymbol{x}_k)^\top J(\boldsymbol{x}_k))^{-1} J(\boldsymbol{x}_k)^\top F(\boldsymbol{x}_k).
\end{aligned}$$

The Gauss-Newton direction is therefore

$$\boldsymbol{d}_k = -(J(\boldsymbol{x}_k)^\top J(\boldsymbol{x}_k))^{-1} J(\boldsymbol{x}_k)^\top F(\boldsymbol{x}_k).$$

Noting that $\nabla g(\boldsymbol{x}) = 2J(\boldsymbol{x})^\top F(\boldsymbol{x})$, we can conclude that

$$\boldsymbol{d}_k = -\frac{1}{2}(J(\boldsymbol{x}_k)^\top J(\boldsymbol{x}_k))^{-1} \nabla g(\boldsymbol{x}_k)$$

meaning that the Gauss-Newton method is essentially a scaled gradient method with $t_k = 1$ *and the following positive definite scaling matrix:*

$$\boldsymbol{D}_k = \frac{1}{2}(J(\boldsymbol{x}_k)^\top J(\boldsymbol{x}_k))^{-1}.$$

The damped Gauss-Newton method

The method described so far is also called the pure Gauss-Newton method since no stepsize is really involved. To transform this method into a practical algorithm, *a stepsize is introduced, leading to the damped Gauss-Newton method.*

The damped Gauss-Newton method

Input: Tolerance parameter $\varepsilon > 0$.

Initialization: Pick $\boldsymbol{x}_0 \in \mathbb{R}^n$ arbitrarily.

General step: For any $k = 0, 1, \cdots$, execute

(a) Set $\boldsymbol{d}_k = -(J(\boldsymbol{x}_k)^\top J(\boldsymbol{x}_k))^{-1} J(\boldsymbol{x}_k)^\top F(\boldsymbol{x}_k)$.

(b) Set stepsize t_k by a line search procedure on the function

$$h(t) := g(\boldsymbol{x}_k + t\boldsymbol{d}_k).$$

(c) Set $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + t_k \boldsymbol{d}_k$.

(d) If $\|\nabla g(\boldsymbol{x}_{k+1})\| \le \varepsilon$ then stop and $\boldsymbol{x}_{k+1}$ is the output.

Lipschitz property of the gradient

We consider the following unconstrained minimization problem

$$\min\{f(\boldsymbol{x}) : \boldsymbol{x} \in \mathbb{R}^n\},$$

where the objective function f is *continuously differentiable.*

- **Definition:** *∇f is Lipschitz continuous over $\mathbb{R}^n \iff \exists\, L \geq 0$ s.t. $\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\| \leq L\|\boldsymbol{x} - \boldsymbol{y}\|, \forall\, \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$.*
- $C_L^{1,1}(\mathbb{R}^n)$ or $C_L^{1,1}$ or $C^{1,1}(\mathbb{R}^n)$ or $C^{1,1}$: the class of functions over $\mathbb{R}^n$ with Lipschitz gradient with constant L.
- $C_L^{1,1}(D)$: the set of all functions over $D \subseteq \mathbb{R}^n$ whose gradient satisfies the above Lipschitz condition for any $\boldsymbol{x}, \boldsymbol{y} \in D$.
- **Examples:**

  (1) Linear functions: given $\boldsymbol{a} \in \mathbb{R}^n, f(\boldsymbol{x}) = \boldsymbol{a}^\top \boldsymbol{x}$ is in $C_0^{1,1}$.

  (2) Quadratic functions: let $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ be symmetric, $\boldsymbol{b} \in \mathbb{R}^n$ and $c \in \mathbb{R}$. Then $f(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x} + 2\boldsymbol{b}^\top \boldsymbol{x} + c$ is in $C_L^{1,1}$, since

$$\begin{aligned} \|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\| &= 2\|(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}) - (\boldsymbol{A}\boldsymbol{y} + \boldsymbol{b})\| \\ &\leq 2\|\boldsymbol{A}\|\|\boldsymbol{x} - \boldsymbol{y}\| := L\|\boldsymbol{x} - \boldsymbol{y}\|. \end{aligned}$$

The Fundamental Theorem of Calculus (FTC)

The Fundamental Theorem of Calculus: *Let $f : [a,b] \to \mathbb{R}$ be a real-valued function.*

Part 1: *Let $f \in \mathcal{R}[a,b]$. Define $F(x) := \int_a^x f(t)dt$, $x \in [a,b]$. Then*
(i) $F(x)$ is continuous on $[a,b]$; (ii) $F'(x) = f(x)$ for $x \in (a,b)$ where f is continuous.

Part 2: *If $f' \in \mathcal{R}[a,b]$, then $\int_a^b f'(x)dx = f(b) - f(a)$.*

Application: Let $f : \mathbb{R}^n \to \mathbb{R}$ be a continuously differentiable function over $D \subseteq \mathbb{R}^n$. Let $\boldsymbol{x}, \boldsymbol{y} \in D$ and $[\boldsymbol{x}, \boldsymbol{y}] \subseteq D$. Define $g(t) := f((1-t)\boldsymbol{x} + t\boldsymbol{y})$ for $t \in [0,1]$. Using the chain rule and the FTC, we respectively obtain $g'(t) = \nabla f((1-t)\boldsymbol{x} + t\boldsymbol{y}) \cdot (\boldsymbol{y} - \boldsymbol{x})$ and

$$\begin{aligned} f(\boldsymbol{y}) - f(\boldsymbol{x}) &= g(1) - g(0) = \int_0^1 g'(t)dt = \int_0^1 \nabla f((1-t)\boldsymbol{x} + t\boldsymbol{y}) \cdot (\boldsymbol{y} - \boldsymbol{x})dt \\ &= \int_0^1 \langle \nabla f(\boldsymbol{x} + t(\boldsymbol{y} - \boldsymbol{x})),\ \boldsymbol{y} - \boldsymbol{x} \rangle dt. \end{aligned}$$

In addition, if f is twice continuously differentiable over $D \subseteq \mathbb{R}^n$, then

$$\begin{aligned} f_x(\boldsymbol{y}) - f_x(\boldsymbol{x}) &= \int_0^1 \nabla(f_x)\big(\boldsymbol{x} + t(\boldsymbol{y} - \boldsymbol{x})\big) \cdot (\boldsymbol{y} - \boldsymbol{x})dt, \\ f_y(\boldsymbol{y}) - f_y(\boldsymbol{x}) &= \int_0^1 \nabla(f_y)\big(\boldsymbol{x} + t(\boldsymbol{y} - \boldsymbol{x})\big) \cdot (\boldsymbol{y} - \boldsymbol{x})dt. \end{aligned}$$

That is, we have

$$\nabla f(\boldsymbol{y}) - \nabla f(\boldsymbol{x}) = \int_0^1 \nabla^2 f\big(\boldsymbol{x} + t(\boldsymbol{y} - \boldsymbol{x})\big)(\boldsymbol{y} - \boldsymbol{x})dt.$$

Boundedness of the Hessian

Theorem: *Let f be a twice continuously differentiable function over $\mathbb{R}^n$. Then*

$$f \in C_L^{1,1}(\mathbb{R}^n) \iff \|\nabla^2 f(\boldsymbol{x})\| \le L, \ \forall\, \boldsymbol{x} \in \mathbb{R}^n.$$

Proof: ($\Leftarrow$) By the fundamental theorem of calculus, $\forall\, \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$, we have

$$\nabla f(\boldsymbol{y}) - \nabla f(\boldsymbol{x}) = \int_0^1 \nabla^2 f(\boldsymbol{x} + t(\boldsymbol{y} - \boldsymbol{x}))(\boldsymbol{y} - \boldsymbol{x})dt = \int_0^1 \nabla^2 f(\boldsymbol{x} + t(\boldsymbol{y} - \boldsymbol{x}))dt\ (\boldsymbol{y} - \boldsymbol{x}).$$

Thus, we have

$$\begin{aligned} \|\nabla f(\boldsymbol{y}) - \nabla f(\boldsymbol{x})\| &\le \left\| \int_0^1 \nabla^2 f(\boldsymbol{x} + t(\boldsymbol{y} - \boldsymbol{x}))dt \right\| \|\boldsymbol{y} - \boldsymbol{x}\| \\ &\le \left(\int_0^1 \|\nabla^2 f(\boldsymbol{x} + t(\boldsymbol{y} - \boldsymbol{x}))\| dt \right) \|\boldsymbol{y} - \boldsymbol{x}\| \le L \|\boldsymbol{y} - \boldsymbol{x}\|. \end{aligned}$$

($\Rightarrow$) By the fundamental theorem of calculus, $\forall\, \boldsymbol{d} \in \mathbb{R}^n$ and $\alpha > 0$, we have

$$\nabla f(\boldsymbol{x} + \alpha \boldsymbol{d}) - \nabla f(\boldsymbol{x}) = \int_0^\alpha \nabla^2 f(\boldsymbol{x} + t\boldsymbol{d})\ \boldsymbol{d}\ dt.$$

Thus, we have

$$\left\| \left(\int_0^\alpha \nabla^2 f(\boldsymbol{x} + t\boldsymbol{d})dt \right) \boldsymbol{d} \right\| = \|\nabla f(\boldsymbol{x} + \alpha \boldsymbol{d}) - \nabla f(\boldsymbol{x})\| \le \alpha L \|\boldsymbol{d}\|.$$

Dividing by α and taking the limit $\alpha \to 0^+$, we obtain $\|\nabla^2 f(\boldsymbol{x})\boldsymbol{d}\| \le L\|\boldsymbol{d}\|$, *where we have used the mean value theorem for definite integrals for each matrix component of* $\nabla^2 f(\boldsymbol{x} + t\boldsymbol{d})$.$\square$

The descent lemma

The following descent lemma is fundamental in convergence proofs of gradient-based methods.

The descent lemma: *Let $D \subseteq \mathbb{R}^n$ and $f \in C_L^{1,1}(D)$ for some $L > 0$. Then for any $\boldsymbol{x}, \boldsymbol{y} \in D$ satisfying $[\boldsymbol{x}, \boldsymbol{y}] \subseteq D$ it holds that*

$$f(\boldsymbol{y}) \leq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^\top (\boldsymbol{y} - \boldsymbol{x}) + \frac{L}{2} \|\boldsymbol{y} - \boldsymbol{x}\|^2.$$

Proof: By the fundamental theorem of calculus, we have

$$\begin{aligned} f(\boldsymbol{y}) - f(\boldsymbol{x}) &= \int_0^1 \langle \nabla f(\boldsymbol{x} + t(\boldsymbol{y} - \boldsymbol{x})), \boldsymbol{y} - \boldsymbol{x} \rangle dt \\ &= \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle + \int_0^1 \langle \nabla f(\boldsymbol{x} + t(\boldsymbol{y} - \boldsymbol{x})) - \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle dt. \end{aligned}$$

Therefore, we have

$$\begin{aligned} |f(\boldsymbol{y}) - f(\boldsymbol{x}) - \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle| &= \left| \int_0^1 \langle \nabla f(\boldsymbol{x} + t(\boldsymbol{y} - \boldsymbol{x})) - \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle dt \right| \\ &\leq \int_0^1 |\langle \nabla f(\boldsymbol{x} + t(\boldsymbol{y} - \boldsymbol{x})) - \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle| \, dt \\ &\leq \int_0^1 \|\nabla f(\boldsymbol{x} + t(\boldsymbol{y} - \boldsymbol{x})) - \nabla f(\boldsymbol{x})\| \|\boldsymbol{y} - \boldsymbol{x}\| dt \\ &\leq \int_0^1 tL \|\boldsymbol{y} - \boldsymbol{x}\|^2 dt = \frac{L}{2} \|\boldsymbol{y} - \boldsymbol{x}\|^2. \quad \square \end{aligned}$$

A sufficient decrease lemma

- Note that the proof of the descent lemma actually shows both upper and lower bounds on the function:

$$f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^\top (\boldsymbol{y} - \boldsymbol{x}) - \frac{L}{2}\|\boldsymbol{y} - \boldsymbol{x}\|^2 \le f(\boldsymbol{y}) \le f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^\top (\boldsymbol{y} - \boldsymbol{x}) + \frac{L}{2}\|\boldsymbol{y} - \boldsymbol{x}\|^2.$$

- **A sufficient decrease lemma:** *Suppose that* $f \in C_L^{1,1}(\mathbb{R}^n)$. *Then* $\forall\, \boldsymbol{x} \in \mathbb{R}^n$ *and* $t > 0$, *we have*

$$f(\boldsymbol{x}) - f(\boldsymbol{x} - t\nabla f(\boldsymbol{x})) \ge t\left(1 - \frac{Lt}{2}\right)\|\nabla f(\boldsymbol{x})\|^2.$$

Proof: By the descent lemma we have

$$\begin{aligned} f(\boldsymbol{x} - t\nabla f(\boldsymbol{x})) &\le f(\boldsymbol{x}) - t\|\nabla f(\boldsymbol{x})\|^2 + \frac{Lt^2}{2}\|\nabla f(\boldsymbol{x})\|^2 \\ &= f(\boldsymbol{x}) - t\left(1 - \frac{Lt}{2}\right)\|\nabla f(\boldsymbol{x})\|^2. \end{aligned}$$

The result then follows by simple rearrangement of terms. □

Sufficient decrease of the gradient method

Theorem: *Let $f \in C_L^{1,1}(\mathbb{R}^n)$. Let $\{\boldsymbol{x}_k\}_{k\geq 0}$ be the sequence generated by the gradient method for solving $\min_{\boldsymbol{x}\in\mathbb{R}^n} f(\boldsymbol{x})$ with one of the following stepsize strategies:*

(1) *constant stepsize $\bar{t} \in (0, \frac{2}{L})$,*

(2) *exact line search,*

(3) *backtracking procedure with parameters $s > 0$, $\alpha, \beta \in (0,1)$.*

Then we have

$$f(\boldsymbol{x}_k) - f(\boldsymbol{x}_{k+1}) \geq M \|\nabla f(\boldsymbol{x}_k)\|^2,$$

where

$$M := \begin{cases} \bar{t}\left(1 - \dfrac{\bar{t}L}{2}\right) & \text{constant stepsize,} \\ \dfrac{1}{2L} & \text{exact line search,} \\ \alpha \min\left\{s, \dfrac{2(1-\alpha)\beta}{L}\right\} & \text{backtracking.} \end{cases}$$

Proof: constant stepsize and exact line search

(1) **(constant stepsize)** By the sufficient decrease lemma, we immediately have

$$f(\boldsymbol{x}_k) - f(\boldsymbol{x}_{k+1}) \geq \bar{t}\left(1 - \frac{L\bar{t}}{2}\right)\|\nabla f(\boldsymbol{x}_k)\|^2 \geq 0 \quad \text{for } \bar{t} \in (0, \frac{2}{L}). \quad \square$$

Furthermore, if we wish to obtain the largest guaranteed bound on the decrease, then we seek the maximum of

$$\bar{t}\left(1 - \frac{L\bar{t}}{2}\right), \quad \forall\, \bar{t} \in (0, \frac{2}{L}).$$

One can show that this maximum is attained at $\bar{t} = \frac{1}{L}$ and we have

$$f(\boldsymbol{x}_k) - f(\boldsymbol{x}_{k+1}) = f(\boldsymbol{x}_k) - f\left(\boldsymbol{x}_k - \frac{1}{L}\nabla f(\boldsymbol{x}_k)\right) \geq \frac{1}{2L}\|\nabla f(\boldsymbol{x}_k)\|^2. \qquad (\star)$$

(2) **(exact line search)** In the exact line search setting, $t_k \in \operatorname{argmin}_{t \geq 0} f(\boldsymbol{x}_k - t\nabla f(\boldsymbol{x}_k))$. By the definition of t_k we know that

$$f(\boldsymbol{x}_k - t_k\nabla f(\boldsymbol{x}_k)) \leq f(\boldsymbol{x}_k - \frac{1}{L}\nabla f(\boldsymbol{x}_k)).$$

Therefore, we have

$$\begin{aligned} f(\boldsymbol{x}_k) - f(\boldsymbol{x}_{k+1}) &= f(\boldsymbol{x}_k) - f(\boldsymbol{x}_k - t_k\nabla f(\boldsymbol{x}_k)) \\ &\geq f(\boldsymbol{x}_k) - f\left(\boldsymbol{x}_k - \frac{1}{L}\nabla f(\boldsymbol{x}_k)\right) \geq \frac{1}{2L}\|\nabla f(\boldsymbol{x}_k)\|^2, \end{aligned}$$

where the last inequality comes from $(\star)$. $\square$

Proof: backtracking

(3) **(backtracking)** In the backtracking setting we seek a small enough stepsize t_k for which we have

$$f(\boldsymbol{x}_k) - f(\boldsymbol{x}_k - t_k \nabla f(\boldsymbol{x}_k)) \geq \alpha t_k \|\nabla f(\boldsymbol{x}_k)\|^2, \quad \alpha \in (0,1).$$

We would like to find a lower bound on t_k. There are two options. Either $t_k = s$ (the initial value of the stepsize) or the stepsize t_k/β is not acceptable, i.e.,

$$f(\boldsymbol{x}_k) - f(\boldsymbol{x}_k - \frac{t_k}{\beta} \nabla f(\boldsymbol{x}_k)) < \alpha \frac{t_k}{\beta} \|\nabla f(\boldsymbol{x}_k)\|^2. \qquad (\star 1)$$

By the sufficient decrease lemma with $\boldsymbol{x} = \boldsymbol{x}_k$ and $t = \frac{t_k}{\beta}$, we have

$$f(\boldsymbol{x}_k) - f(\boldsymbol{x}_k - \frac{t_k}{\beta} \nabla f(\boldsymbol{x}_k)) \geq \frac{t_k}{\beta}\left(1 - \frac{L t_k}{2\beta}\right) \|\nabla f(\boldsymbol{x}_k)\|^2. \qquad (\star 2)$$

From $(\star 1)$ and $(\star 2)$, we obtain

$$\frac{t_k}{\beta}\left(1 - \frac{L t_k}{2\beta}\right) < \alpha \frac{t_k}{\beta} \iff t_k > \frac{2(1-\alpha)\beta}{L}.$$

Overall, we have

$$t_k \geq \min\left\{s, \frac{2(1-\alpha)\beta}{L}\right\}.$$

Finally, we obtain

$$f(\boldsymbol{x}_k) - f(\boldsymbol{x}_k - t_k \nabla f(\boldsymbol{x}_k)) \geq \alpha \min\left\{s, \frac{2(1-\alpha)\beta}{L}\right\} \|\nabla f(\boldsymbol{x}_k)\|^2. \quad \square$$

Convergence of the gradient method

Theorem: *Let $f \in C_L^{1,1}(\mathbb{R}^n)$. Let $\{\boldsymbol{x}_k\}_{k\geq 0}$ be the sequence generated by the gradient method for solving $\min_{\boldsymbol{x}\in\mathbb{R}^n} f(\boldsymbol{x})$ with one of the following stepsize strategies:*

(1) *constant stepsize $\bar{t} \in (0, \frac{2}{L})$,*

(2) *exact line search,*

(3) *backtracking procedure with parameters $s > 0$, $\alpha, \beta \in (0, 1)$.*

Assume that f is bounded below over $\mathbb{R}^n$, i.e., $\exists\, m \in \mathbb{R}$ such that $f(\boldsymbol{x}) > m$ for all $\boldsymbol{x} \in \mathbb{R}^n$. Then we have the following:

(a) *The sequence $\{f(\boldsymbol{x}_k)\}_{k\geq 0}$ is nonincreasing. In addition, for any $k \geq 0$, $f(\boldsymbol{x}_{k+1}) < f(\boldsymbol{x}_k)$ unless $\nabla f(\boldsymbol{x}_k) = \mathbf{0}$.*

(b) *$\nabla f(\boldsymbol{x}_k) \to \mathbf{0}$ as $k \to \infty$.*

Proof of the convergence theorem

(a) By the sufficient decrease of the gradient method, we have

$$f(\boldsymbol{x}_k) - f(\boldsymbol{x}_{k+1}) \geq M\|\nabla f(\boldsymbol{x}_k)\|^2, \quad (\star\star)$$

for some constant $M > 0$, and hence the equality $f(\boldsymbol{x}_k) = f(\boldsymbol{x}_{k+1})$ can hold only when $\nabla f(\boldsymbol{x}_k) = \mathbf{0}$.

(b) Since the sequence $\{f(\boldsymbol{x}_k)\}_{k\geq 0}$ is nonincreasing, and bounded below, it converges. Thus, in particular

$$f(\boldsymbol{x}_k) - f(\boldsymbol{x}_{k+1}) \to 0 \quad \text{as} \quad k \to \infty,$$

which combined with $(\star\star)$ implies $\|\nabla f(\boldsymbol{x}_k)\| \to 0$ as $k \to \infty$, according to the squeeze theorem. Therefore, we obtain

$$\nabla f(\boldsymbol{x}_k) \to \mathbf{0} \quad \text{as} \quad k \to \infty. \quad \square$$

"Rate of convergence" of gradient norms

Theorem: *Under the setting of the previous theorem, let f^* be the limit of the convergent sequence $\{f(\boldsymbol{x}_k)\}_{k\geq 0}$. Then for any $\ell = 0, 1, 2, \ldots$*

$$\min_{k=0,1,\cdots,\ell} \|\nabla f(\boldsymbol{x}_k)\| \leq \sqrt{\frac{f(\boldsymbol{x}_0) - f^*}{M(\ell+1)}},$$

where

$$M = \begin{cases} \bar{t}\left(1 - \frac{\bar{t}L}{2}\right) & \textit{constant stepsize,} \\ \frac{1}{2L} & \textit{exact line search,} \\ \alpha \min\{s, \frac{2(1-\alpha)\beta}{L}\} & \textit{backtracking.} \end{cases}$$

Proof: Summing the inequality $(\star\star)$ on the previous page over $k = 0, 1, \cdots, \ell$, we obtain the following inequality

$$f(\boldsymbol{x}_0) - f(\boldsymbol{x}_{\ell+1}) \geq M \sum_{k=0}^{\ell} \|\nabla f(\boldsymbol{x}_k)\|^2.$$

Since $f(\boldsymbol{x}_{\ell+1}) \geq f^*$, we can conclude that

$$f(\boldsymbol{x}_0) - f^* \geq M \sum_{k=0}^{\ell} \|\nabla f(\boldsymbol{x}_k)\|^2 \geq M(\ell+1) \min_{k=0,1,\cdots,\ell} \|\nabla f(\boldsymbol{x}_k)\|^2,$$

implying the desired result. $\square$