Jingmin Sun(sunj8) RIN:661849071
Chenyu Wu(wuc10) RIN:661672516

Optimization Lecture Notebook

Contents

1 Lecture 1

1.1 Three Part of Numerical Method:

- Input of data, initial guess of solution
- Update rule of iterate: How to renew the guess
- Stopping Condition

Example:

$$\begin{cases} 2x_1 + x_2 &= 3 \\ -x_1 + 2x_2 &= 1 \end{cases} \qquad x* = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

1. Input:

$$A = \begin{bmatrix} 2 & 1 \\ -1 & 2 \end{bmatrix} \qquad b = \begin{bmatrix} 3 \\ 1 \end{bmatrix} \qquad x^0 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

2. Suppose the current guess is x^{old}

$$\begin{cases} x_1^{new} &= (3 - x_2^{old})/2 \\ x_2^{new} &= (1 + x_1^{old})/2 \end{cases}$$

$$x^1 = \begin{bmatrix} 3/2 \\ 1/2 \end{bmatrix} \qquad x^2 = \begin{bmatrix} 5/4 \\ 5/4 \end{bmatrix} \qquad x^3 = \begin{bmatrix} 7/8 \\ 9/8 \end{bmatrix}$$

3. Stopping Condition:

$$||Ax^k - b|| \le tol$$

e.g. $tol = 10^{-8}$

1.2 Rate of Convergence

Definition 1.1 (Q-linear Convergence)**.** Let $x^k \to x^*$, we say the convergence is Q-linear if there is $r \in (0, 1)$, such that $\dfrac{||x^{k+1} - x^*||}{||x^k - x^*||} \le r$, for k sufficiently large.

Example:
$x^k = 1 + \frac{1}{2^k}$, $k = 1, 2, \cdots$ x^k : Q-linearly converge to $x^* = 1$

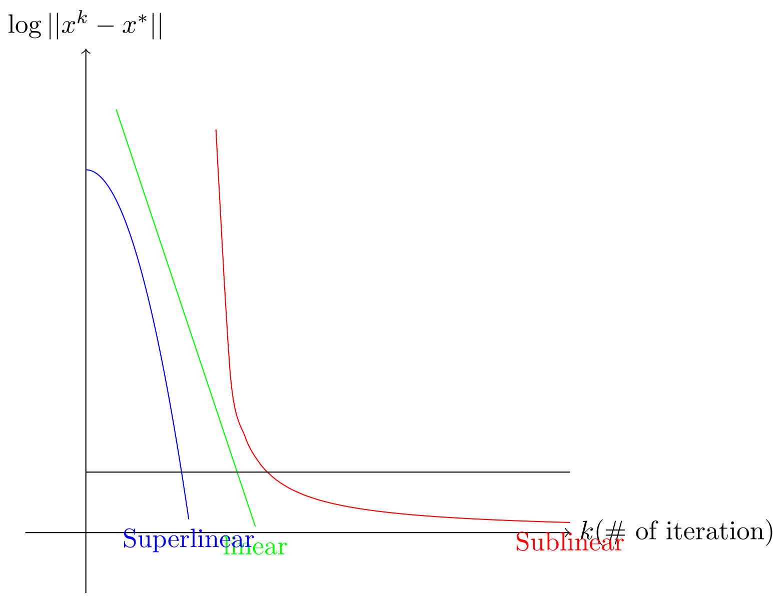
Definition 1.2 (Q-superlinear Convergence)**.** Let $x^k \to x^*$, we say the convergence is Q-superlinear if $\lim_{k\to\infty} \dfrac{||x^{k+1} - x^*||}{||x^k - x^*||} = 0$, for k sufficiently large.

Example: $x^k = 1 + k^{-k}$ $\quad k = 1, 2...$ 比linear快

Definition 1.3 (Q-sublinear Convergence)**.** Let $x^k \to x^*$, we say the convergence is Q-linear if

$$\lim_{k\to\infty} \frac{||x^{k+1} - x^*||}{||x^k - x^*||} = 1$$

Example: $x^k = 1 + \frac{1}{k} \quad k = 1, 2, ...$ 比linear 慢



达到相同的差值（比）， superlinear 需要的时间（iteration）最少。

Definition 1.4 (Q-quadratic convergence)**.** Let $x^k \to x^*$. we say the convergence is Q-quadratic if there is $M > 0$, such that

$$\frac{||x^{k+1} - x^*||}{||x^k - x^*||^2} \leq M \text{ For k sufficiently large}$$
$$\to ||x^{k+1} - x^*|| \leq M||x^k - x^*||^2$$
$$\to \frac{||x^{k+1} - x^*||}{||x^k - x^*||} \leq M||x^k - x^*|| \to 0$$

Definition 1.5 (R-linear convergence)**.** Let $x^k \to x^*$. we say the convergence s R-linear if $||x^k - x^*|| \leq v^k, \forall k \geq 1$, and $\{v^k\}$ is Q-linearly convergent to 0

Example:

$$x^k = \begin{cases} 1 + \frac{1}{2^k} & \text{if k is even} \\ 1 & \text{if k is odd} \end{cases}$$
$$\implies x^k \to x^* = 1$$
$$\implies \text{Check for Q linear:}$$
$$\frac{|x^{k+1} - x^*|}{|x^k - x^*|} \leq r \in (0, 1)$$
$$\forall \text{ odd } k, \frac{|x^{k+1} - x^*|}{|x^k - x^*|} = \infty$$
$$\therefore not\ Q - linear$$

$$\begin{aligned}
&\implies \text{Choose } v^k = \frac{1}{2^k} \\
&\text{Note } \frac{|v^{k+1} - 0|}{|v^k - 0|} = \frac{1}{2} \in (0,1) && \forall k = 1, 2, \cdots \\
&\text{and } |x^k - x^*| \leq v^k && \forall k = 1, 2, \cdots
\end{aligned}$$

Therefore, x^k is R-linearly convergent to $x^* = 1$.

Definition 1.6 (R-superlinear Convergence)**.** Let $x^k \to x^*$, we say the convergence is R-superlinear if $||x^k = x^*|| \leq v^k$, $\forall k \geq 1$, and $\{v^k\}$ is Q-superlinear convergent to 0.

Definition 1.7 (R-sublinear Convergence)**.** Let $x^k \to x^*$, we say the convergence is R-sublinear if $||x^k = x^*|| \leq v^k$, $\forall k \geq 1$, and $\{v^k\}$ is Q-sublinear convergent to 0.

Exercise:

$$\begin{cases} 2x_1 + x_2 & = 3 \\ -x_1 + 2x_2 & = 1 \end{cases} \qquad x* = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

If $x_0 = x^*$, 后面的iteration值不变
Let

$$\begin{cases} x_1^{k+1} = (3 - x_2^k)/2 \\ x_2^{k+1} = (1 + x_1^k)/2 \end{cases} \qquad x^0 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Q1: Is $\{x^k\}$ convergent to x^*?
Q2: What is the convergence speed?

$$\begin{aligned}
||x^{k+1} - x^*||^2 &= (x_1^{k+1} - 1)^2 + (x_2^{k+1} - 1)^2 \\
&= (\frac{3 - x_2^k}{2} - 1)^2 + (\frac{1 + x_1^k}{2} - 1)^2 \\
&= (\frac{1 - x_2^k}{2})^2 + (\frac{x_1^k - 1}{2})^2 \\
&= \frac{1}{4}||x^k - x^*||^2 && \forall k \\
&= \frac{1}{4^2}||x^{k-1} - x^*||^2 \\
&= \frac{1}{4^{k+1}}||x^0 - x^*||^2 && ask \to \infty
\end{aligned}$$

So $x^k \to x^*$. Also, note $\dfrac{||x^{k+1} - x^*||}{||x^k - x^*||} = \frac{1}{2} \in (0,1)$, so the convergence is Q-linear.

2 L2

2.1 Unconstraint Problem

$$\min_{x \in \mathbb{R}^n} f(x)$$

Example: Logistic Regression.

2.2 Vector norm

Assume $x \in \mathbb{R}^n$,

$$\begin{aligned}
||x||_1 &= \sum_{i=1}^{n} |X_n| \\
||x||_2 &= \sqrt{\sum_{i=1}^{n} x_i^2} && (= ||x||) \\
||x||_\infty &= \max_{1 \leq x \leq n} |x_i| \\
||x||_p &= \left(\sum_{x_i}^{n} |x_i|^p \right)^{1/p} && \text{for } p \geq 1
\end{aligned}$$

In $\mathbb{R}^n$, a function Φ defines a norm if

1. $\Phi(x) = 0 \iff x = \vec{0}$
2. $\Phi(\alpha x) = |\alpha| \Phi(x)$, $\forall \alpha \in \mathbb{R}, \forall x \in \mathbb{R}^n$
3. $\Phi(x + y) \leq \Phi(x) + \Phi(y), \forall x, y \in \mathbb{R}^n$

2.3 Matrix norm

$A \in \mathbb{R}^{m \times n}$
Let

$$\begin{aligned}
A &= \begin{bmatrix} | & | & & | \\ a_1 & a_2 & \cdots & a_n \\ | & | & & | \end{bmatrix} = \begin{bmatrix} - & \tilde{a_1} & - \\ - & \tilde{a_2} & - \\ - & \tilde{a_3} & - \end{bmatrix} \\
||A||_1 &= \max_{1 \leq j \leq n} ||a_j||_1 = \sup_{||x||_1 = 1} ||A_x||_1 \\
||A||_\infty &= \max_{1 \leq i \leq m} ||\tilde{a_i}||_1 = \sup_{||x||_\infty = 1} ||A_x||_\infty \\
||A||_2 &= \sigma_{max}(A) = \sup_{||x||_2 = 1} ||A_x||_2 \\
||A||_F &= \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} (a_{ij})^2} = \sqrt{\operatorname{tr}(A^* A)} = \sqrt{\sum_{i=1}^{\min\{m,n\}} \sigma_i^2} && \text{(Froberius)} \\
||A||_* &= \sum_{i=1}^{m} \sigma_i(A)
\end{aligned}$$

$\sigma(A)$ the singular value of A

2.4 Gradient

: Assume $f : \mathbb{R}^{\ltimes} \to \mathbb{R}$,

$$\nabla f(x) = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \cdots \\ \frac{\partial f}{\partial x_n} \end{bmatrix}$$

Example:

$$\begin{aligned} f(x) &= \frac{1}{2}x^T Qx + c^T x \qquad\qquad Q^T = Q \\ \nabla f(x) &= Qx + c \\ \text{verify for } Q &= \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} c = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \end{aligned}$$

Hessian Matrix:

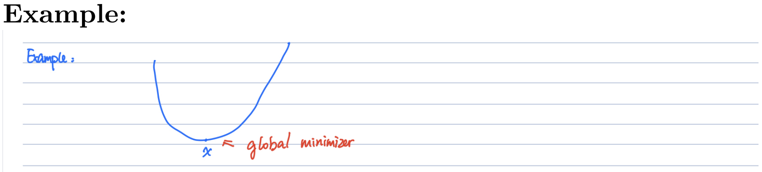
$$\nabla^2 f(x) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 x_n} \\ \cdots & \cdots & & \cdots \\ \frac{\partial^2 f}{\partial x_n x_1} & \frac{\partial^2 f}{\partial x_n x_2} & \cdots & \frac{\partial^2 f}{\partial x_n x_n} \end{bmatrix}$$

Example:

$$\begin{aligned} f(x) &= \frac{1}{2}x^T Qx + c^T x \qquad\qquad Q^T = Q \\ \nabla^2 f(x) &= Q \end{aligned}$$

Definition 2.1 (global minimizer)**.** x^* is called a global minimizer of f if $f(x^*) \leq f(x), \forall x \in \mathbb{R}^n$

Example:

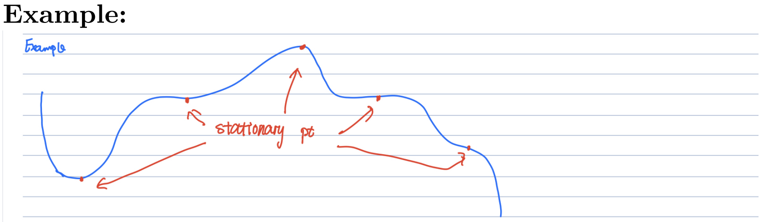


Definition 2.2 (Local minimizer)**.** x^* is called a locall minimizer of f if there exists $\delta > 0$ such that $f(x^*) \leq f(x), \forall x \in B_\delta(x^*) = \{x : ||x - x^*|| \leq \delta\}$

Example:

Definition 2.3 (Stationary point)**.** : A point is called stationary point of f, if $\nabla f(\bar{x}) = 0$

Example:



Definition 2.4 (Saddle point)**.** A point $\bar{x}$ is saddle point of f if $\nabla f(\bar{x}) = 0$, but $\bar{x}$ is neither a local minimizer or a local maximizer.

Theorem 2.1 (First-order necessary condition)**.** Assume f is a differentiable function if x^* is a local minimizer, then $\nabla f(x^*) = 0$

Remark:

1. $\nabla f(x^*) = 0$ does NOT imply x^* to be a local minimizer
2. If f is a convex function, then $\nabla f(x^*) = 0$ implies that x^* is a global minimizer.
   Every point of tangent line below $f(x) \implies$ convex function.

Example of convex function:

1. $f(x) = e^x$
2. $f(x) = -\log x$
3. $f(x)$ is a vector norm
4. $f(x) = g(Ax + b)$, where g is convex

Exercise
Let $f(x, y) = x^2 - 2xy + y^2 - 2$
Q1: Is $\begin{bmatrix} 1 \\ 0 \end{bmatrix}$ a local minimizer?

$$\nabla f(x, y) = \begin{bmatrix} 2x - 2y \\ -2x + 2y \end{bmatrix}$$

$$\text{so} \nabla f(1, 0) = \begin{bmatrix} 2 \\ -2 \end{bmatrix} \neq \vec{0}$$

and $\begin{bmatrix}1\\0\end{bmatrix}$ cannot be a local minimizer.

Q2: Is $\begin{bmatrix}1\\1\end{bmatrix}$ a local minimizer?

$$\nabla f(x,y) = \begin{bmatrix}2x-2y\\-2x+2y\end{bmatrix}$$
$$\text{so}\nabla f(1,1) = \begin{bmatrix}0\\0\end{bmatrix}$$

and f is convex, so $\begin{bmatrix}1\\1\end{bmatrix}$ is a global minimizer, and of course a local minimizer.

Theorem 2.2 (Second-Order Necessary Condition)**.** : Assume f is a twice-differentiable function, if x^* is a local minimizer, then $\nabla f(x^*) = 0$, $\nabla^2 f(x^*) \geq 0$

Exercise:
$f(x,y) = \cos(x,y)$
Is $\begin{bmatrix}0\\0\end{bmatrix}$ a local minimizer?

$$\nabla f(x,y) = \begin{bmatrix}-\sin(x+y)\\-\sin(x+y)\end{bmatrix}$$
$$\nabla f(0,0) = 0$$
$$\nabla^2 f(x,y) = \begin{bmatrix}-\cos(x+y) & -\cos(x+y)\\-\cos(x+y) & -\cos(x+y)\end{bmatrix}$$
$$\nabla^2 f(0,0) = \begin{bmatrix}-1 & -1\\-1 & -1\end{bmatrix} < 0$$

$\therefore \begin{bmatrix}0\\0\end{bmatrix}$ is not a local minimizer.

Theorem 2.3 (Second -Oder Sufficient Condition)**.** Assume f is twice-differentiable. If $\nabla f(x^*) = 0$, and $\nabla^2 f(x^*) > 0$, then x^* must be local minimizer.

Exercise:
Let
$$f(x) = \frac{1+(2-x)^2}{1+x^2}$$
find the local minimizer of f
Solution:

$$\begin{aligned}
f(x) &= \frac{1+(4-4x+x^2)}{1+x^2}\\
&= 1+4\frac{1-x}{1+x^2}\\
\therefore \min_x f(x) \iff & \min_x g(x) := \frac{1-x}{1+x^2}\\
g'(x) &= \frac{-(1+x^2)-2x(1-x)}{(1+x^2)^2}\\
&= \frac{x^2-2x-1}{(1+x^2)^2}
\end{aligned}$$

Solve

$$
\begin{aligned}
g'(x) &= 0 \\
x^2 - 2x - 1 &= 0 \\
x &= 1 + \sqrt{2} x = 1 - \sqrt{2} \\
g''(x) &= \frac{(2x-2)(1+x^2)^2 - (x^2-2x-1)(4x(1+x^2))}{(1+x^2)^4} \\
\therefore g''(1+\sqrt{2}) &> 0 \quad \text{local minimizer} \\
g''(1-\sqrt{2}) &< 0
\end{aligned}
$$

3 L3:steepest gradient descent

Problem:$\min_{x\in\mathbb{R}^n} f(x)$

3.1 searching method:

Two strategies: line search and trust region
Idea of line search:

- Find a descent direction p^k at k-th iteration.
- search a step size $\alpha_k > 0$ such that
$$f(x^k + \alpha_k p^k) < f(x^k)$$
- Update the iterate:
$$x^{k+1} = x^k + \alpha_k p^k$$

Idea of trust region:

- Determine the size of a search region at k-th iteration: $\{p \in \mathbb{R}^n : ||p|| \leq \Delta k\}$
- Approximation the model within the region
$$m(x^k + p) \approx f(x^k + p)$$
- solve a subproblem:
$$p^k = \arg\min_p m(x^k + p), \text{ s.t. } ||p|| \leq \Delta k$$
and let $x^{k+1} = x^k + p^k$

3.2 steepest gradient method

1. Search direction: $p^k = -\nabla f(x^k)$ Lemma: If $\nabla f(x^k) \neq 0$, then $< p^k, \nabla f(x^k) >< 0$ and there is $\bar{\alpha} > 0$, such that $f(x^k + \alpha p^k) < f(x^k), \forall \alpha \in (0, \bar{\alpha})$.
Proof: Assume $||\nabla^2 f(x)||_2 \leq \sigma, \forall x \in \mathbb{R}^n$
By Taylor expansion,
$$f(x^k + \alpha p^k) = f(x^k) + \alpha < p^k, \nabla f(x^k) > + \frac{\alpha^2}{2}(p^k)^T \nabla^2 f(\tilde{x}^2) p^k$$
for some pt $\tilde{x^k}$
Because $||\nabla^2 f(x)||_2 \leq \sigma, \forall x \in \mathbb{R}^n$ it holds
$$(p^k)^T \nabla^2 f(\tilde{x}^k) p^k \leq \sigma ||p^k||^2$$
Hence if $0 < \alpha\bar{\alpha} = -\frac{2<p^k, \nabla f(x^k)>}{\sigma ||p^k||^2}$
then $f(x^k + \alpha p^k) < f(x^k)$
2. search step size: 1. exact line search, 2. inexact line search

(a) exact line search

$$\alpha_k = \arg\min_{\alpha \geq 0} f(x^k + \alpha p^k)$$

Example:

$$f(x, y) = x^2 + xy + y^2 - x - y, \text{for} x, y \in \mathbb{R}$$

$$\text{text} \begin{bmatrix} x^0 \\ y^0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$\nabla f(x, y) = \begin{bmatrix} 2x + y - 1 \\ x + 2y - 1 \end{bmatrix}$$

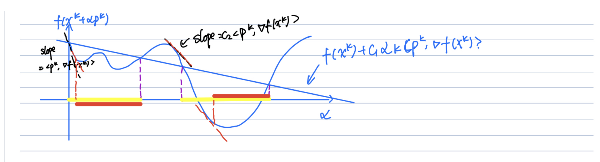
$$p^0 = -\nabla f(x^0, y^0) = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

$$f(\begin{bmatrix} x^0 \\ y^0 \end{bmatrix} + \alpha p^0) = f(x^0 + \alpha, y^0 + \alpha) = f(\alpha, \alpha) = 3\alpha^2 - 2\alpha$$

$$\alpha_0 = \arg\min_{\alpha \geq 0} 3\alpha^2 - 2\alpha \to \alpha_0 = \frac{1}{3}$$

(b) inexact line search
Armijo's condition: find $\alpha_k > 0$ such that

$$f(x^k + \alpha_k p^k) \leq f(x^k + c_1 \alpha_k < p^k, \nabla f(x^k) >)$$



where $c_1 \in (0, 1)$ is a constant. Wolfe conditions: search $\alpha_k > 0$ such that

$$f(x^k + \alpha_k p^k) \leq f(x^k) + c_1 \alpha_k < p^k, \nabla f(x^k) >$$

$$< \nabla f(x^k + \alpha_k p^k), p^k > \geq c_2 < p^k, \nabla f(x^k) >$$

where $0 < c_1 < c_2 < 1$

Theorem 3.1 (Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable Let p^k be a vector such that $< p^k, \nabla f(x^k) > < 0$)**.**

Assume $\phi(\alpha) = f(x^k + \alpha p^k)$ to be bounded below on $\alpha \geq 0$. Then there exist an interval of α such that the wolfe conditions are satisfied for any α on that interval.
Example:

$$f(x, y) = x^2 + xy + x^2 - x - y, \text{ for } x, y \in \mathbb{R}$$

$$\text{Let} \begin{bmatrix} x^0 \\ y^0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$\nabla f(x, y) = \begin{bmatrix} 2x + y - 1 \\ x + 2y - 1 \end{bmatrix}$$

$$p^0 = -\nabla f(x^0, y^0) = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

Q1: set $c_1 = \frac{1}{10}$
Does $\alpha = \frac{1}{2}$ satisfy the Armijo condition?

$$f(\begin{bmatrix} x^0 \\ y^0 \end{bmatrix} + \alpha p^0) = 3\alpha^2 - 2\alpha$$
$$< p^0, \nabla f(x^0, y^0) > = -2$$
$$f(x^0, y^0) = 0$$
$$\text{Check } 3\alpha^2 - 2\alpha \leq 0 + \alpha c_1(-2), \text{ for } \alpha = \frac{1}{2}?$$
$$-\frac{1}{4} < -\frac{1}{10}\surd$$

So $\alpha = \frac{1}{2}$ is acceptable.

3. Stopping Criterion:
   - option I: $||\nabla f(x^k)|| \leq$tol
   - option II: $||x^{k+1} - x^k|| \leq$tol
   - option III: $|f(x^{k+1} - f(x^k))| \leq$tol

3.3 Algorithm

Initialization;
Choose x^0 and $tol > 0$;
if $||p^k|| < tol$ **then**
  stop and return x^k
else
  while *the Armijo's condition does not hold* **do**
    $\sigma_k \alpha_k \to \alpha_k$
    $x^{k+1} = x^k + \alpha_k p^k$
  end
end

Algorithm 1: How to write algorithms

Theorem 3.2 (Convergence of Steepest gradient descent:)**.** Assume f is lower bounded and has Lipschitz conditions, gradient, namely, $\exists L > 0$, such that

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

Let $\{x^k\}$ be the sequence give by the steepest gradient descent with the step size α_k satisfying the Wolfe's conditions or satisfying the Armijo's condition and $\alpha_k \geq \alpha_{min} > 0, \forall k$. Then $||\nabla f(x^k)|| \to 0$

Remark: when $||\nabla f(x^k)|| \to 0$, if $\alpha_k \leq \alpha_{max} < \infty$,
then $||x^{k+1} - x^k|| \to 0$ and $|f(x^{k+1} - f(x^k))| \to 0$.
If further more, f is convex, then $f(x^k) \to f^*$ claim: if α_k is obtained y exact line search, then the Wolfe's conditions hold.
proof: Check the curvature condtiion:

$$< \nabla f(x^k + \alpha_k p^k), p^k > \geq c_2 < \nabla f(x^k), p^k >$$

Since $\alpha_k = \arg\min_{\alpha \geq 0} f(x^k + \alpha p^k) > 0$
we have the first order optimality condition

$$\begin{aligned} 0 &= \frac{d}{d\alpha} f(x^k + \alpha p^k)|_{\alpha=\alpha_k} \\ &= < \nabla f(x^k + \alpha_k p^k), p^k > \end{aligned}$$

3.4 Steepest gradient descent for quadratic minimization:

$$\min_{x \in \mathbb{R}^n} f(x) = \frac{1}{2} x^T Q x - b^T x$$

where Q is symmetric and positive definite

3.4.1 Derivation of the algorithm:

1. $$p^k = -\nabla f(x^k) = -(Qx^k - b) = b - Qx^k$$

2. exact line search:
$$\alpha_k = \arg\min_{\alpha \geq 0} f(x^k + \alpha p^k) = \frac{||p^k||}{(p^k)^T Q p^k}$$
let $\phi(\alpha) = f(x^k + \alpha p^k)$
Then the 1st-order opt. condition:
$$\begin{aligned} \phi'(\alpha_k) &= 0 \\ \text{Since } \phi'(\alpha) &= < \nabla f(x^k + \alpha p^k), p^k > \\ &= < Q(x^k + \alpha p^k) - b, p^k > \\ &= < Qx^k - b + \alpha Q p^k, p^k > \\ &= < -p^k + \alpha Q p^k, p^k > \end{aligned}$$
Then solve $\phi'(\alpha) = 0 \rightarrow \alpha_k = \frac{<p^k, p^k>}{<Qp^k, p^k>}$

3. Update $x^{k+1} = x^k + \alpha_k p^k$

3.4.2 Per-iteration complexity

:

$$n(2n-1) + n + n(2n-1) + 2n - 1 + 2n - 1 + 2n \approx 4n^2$$

total complexity to have a solution $\bar{x}$ such that

$$f(\bar{x}) - f(x^*) \leq \epsilon < 1$$

Goal: estimate k such that

$$f(x^k) - f(x^*) \leq \epsilon$$

3.4.3 Convergence rate analysis

claim1: $f(x^k+1)-f(x^*)=f(x^k)-f(x^*)-\frac{1}{2}\frac{||\nabla f(x^k)||^2}{\nabla f(x^k)^TQ\nabla f(x^k)}\cdot||\nabla f(x^k)||^2$
claim2: $f(x)-f(x^*)=\frac{1}{2}(x-x^*)^TQ(x-x^*)$ where $x^*=\arg\min_x f(x), i.e. Qx^*=b$
claim 2 is easy to verify. Let's prove claim 1 as follows:

$$
\begin{aligned}
&f(x^{k+1})-f(x^*)\\
&=\frac{1}{2}(x^k-\alpha_k\nabla f(x^k)-x^*)^TQ(x^k-\alpha_k\nabla f(x^k)-x^*)\\
&=\frac{1}{2}(x^k-x^*)^TQ(x^k-x^*)-\alpha_k\nabla f(x^k)^TQ(x^k-x^*)+\frac{1}{2}\alpha_k^2\nabla f(x^k)^TQ\nabla f(x^k)\\
&=\frac{1}{2}(x^k-x^*)^TQ(x^k-x^*)-\alpha_k||\nabla f(x^k)||^2+\frac{1}{2}\alpha^2\nabla f(x^k)^TQ\nabla f(x^k)\\
&=\frac{1}{2}(x^k-x^*)^TQ(x^k-x^*)-\frac{||\nabla f(x^k)||^2}{\nabla f(x^k)^TQ\nabla f(x^k)}||\nabla f(x^k)||^2+\frac{||\nabla f(x^k)||^2}{2\nabla f(x^k)^TQ\nabla f(x^k)}||\nabla f(x^k)||^2\\
&=\frac{1}{2}(x^k-x^*)^TQ(x^k-x^*)-\frac{||\nabla f(x^k)||^2}{2\nabla f(x^k)^TQ\nabla f(x^k)}||\nabla f(x^k)||^2
\end{aligned}
$$

Note $\nabla f(x^k)=Qx^k-b=Q(x^k-x^*)$
so $x^k-x^*=Q^{-1}\nabla f(x^k)$
Therefore, by claim ,

$$
\begin{aligned}
f(x^k)-f(x^*)&=\frac{1}{2}(x^k-x^*)^TQ(x^k-x^*)\\
&=\frac{1}{2}\nabla f(x^k)^TQ^{-1}\nabla f(x^k)
\end{aligned}
$$

so $||\nabla f(x^k)||^2=\frac{||\nabla f(x^k)||^2}{\frac{1}{2}\nabla f(x^k)^TQ^{-1}\nabla f(x^k)}[f(x^k)-f(x^*)]$
By claim 1:

$$
\begin{aligned}
f(x^{k+1}-f(x^*))&=f(x^k)-f(x^*)-\frac{1}{2}\frac{||\nabla f(x^k)||^2}{\nabla f(x^k)^TQ\nabla f(x^k)}\frac{||\nabla f(x^k)||^2}{\frac{1}{2}\nabla f(x^k)^TQ^{-1}\nabla f(x^k)}[f(x^k)-f(x^*)]\\
&=(1-\frac{1}{2}\frac{||\nabla f(x^k)||^2}{\nabla f(x^k)^TQ\nabla f(x^k)}\frac{||\nabla f(x^k)||^2}{\frac{1}{2}\nabla f(x^k)^TQ^{-1}\nabla f(x^k)})[f(x^k)-f(x^*)]\\
&\le(1-\frac{\lambda_{min}(Q)}{\lambda_max(Q)})[f(x^k)-f(x^*)]\theta=(1-\frac{\lambda_{min}(Q)}{\lambda_max(Q)})\in 0,1
\end{aligned}
$$

Where in the last inequality, we have used

$$
\begin{aligned}
V^TQV^T&\le\lambda_{max}(Q)||v||^2\\
V^TQ^{-1}V^T&\le\frac{1}{\lambda_{min}}(Q)||v||^2
\end{aligned}
$$

Therefore, $f(x^k)-f(x^*)$ is Q-linearly convergent to 0
To have $f(x^k)-f(x^*)\le\epsilon$
it suffices to have

$$\theta^k(f(x^0)-f(x^*))\le\epsilon$$

because $f(x^k)-f(x^*)\le\theta^k(f(x^0)-f(x^*))$
so

$$
\begin{aligned}
k&\ge ln\frac{\epsilon}{f(x^0)-f(x^*)}/ln(\theta)\\
&=ln\frac{f(x^0)-f(x^*)}{\epsilon}/ln(\frac{1}{\theta})
\end{aligned}
$$

4 L4

Newton's method and conjugate gradient

4.1 Newton's method for

$$\min_{x \in \mathbb{R}^n} f(x)$$

Where f is twice differentiable.

1. search direction: $p^k = -(\nabla^2 f(x^k))^{-1} \nabla f(x^k)$
2. Step size: $\alpha_k = 1$

Remark: this method only works locally.

Theorem 4.1 (Suppose $\nabla^2 f(x)$ is Lip-continuous in a neighborhood of a solution x^*, where $\nabla^2 f(x^*)$ is nonsingular. Let
If x^0 is sufficiently close to x^*, then $x^k \to x^*$

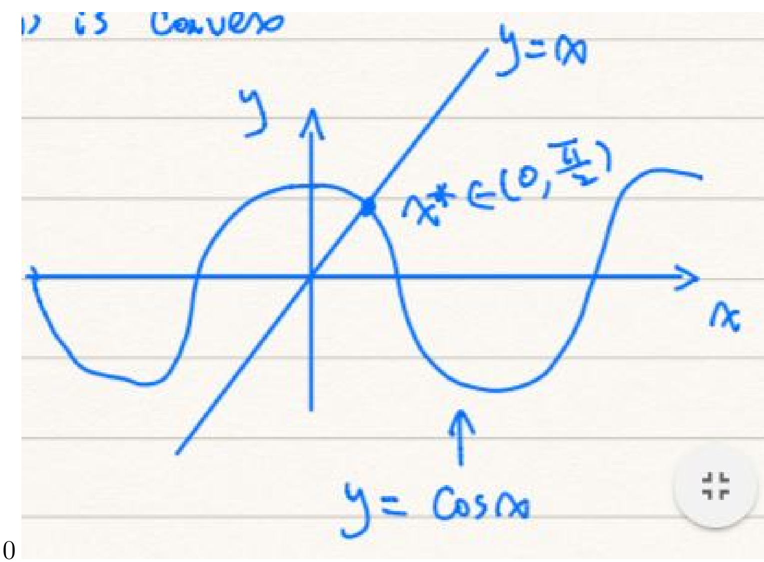
2. and the convergence of $\{x^k\}$ and $\{||\nabla f(x^k)||\}$ are both Q-quadratic.

$\nabla^2 f(x)$ is Lip-continuous if there is $L > 0$ such that $||\nabla^2 f(x) - \nabla^2 f(y)|| \leq L||x - y||, \forall x, y \in \mathbb{R}^n$.
Remark: Assume $||\nabla^2 f(x) - \nabla^2 f(y)|| \leq L||x - y||$ and $||\nabla^2 f(x)^{-1}|| \leq 2||\nabla^2 f(x^*)^{-1}||$, for any $x, y \in Br(x^*) = \{x \in \mathbb{R}^n : ||x - x^*|| \leq r\}$
Then we can take $||x^0 - x^*|| \leq min(r, L||\nabla^2 f(x^*)^{-1}||)$

Example:

$$\min_{x \in \mathbb{R}} f(x) = 0.5x^2 - sin(x)$$

$$f'(x) = x - cos(x)$$
$$f''(x) = 1 + sin(x)$$



Note $f''(x) \geq 0$, so $f(x)$ is convex set $x - cos(x) = 0$

4.2 Conjugate gradient

Problem:

$$\min_{x\in\mathbb{R}^n} \frac{1}{2}x^TAx - b^Tx$$

where A is a symmetric positive definite matrix in $\mathbb{R}^{n\times n}$ and $b\in\mathbb{R}^n$

Definition 4.1 (Conjugate directions)**.** : Let A be a symmetric positive definite matrix in $\mathbb{R}^{n\times n}$. A set of nonzeros vectors $\{p^1, p^2, ..., p^m\}$ is said to be conjugate with respect to A or A- conjugate if

$$< p^i, Ap^j >= 0, \forall i \neq j$$

Example: If $A = \begin{bmatrix} a_1 & 0 & 0 & 0 & 0 \\ 0 & a_2 & 0 & 0 & 0 \\ 0 & 0 & . & 0 & 0 \\ 0 & 0 & 0 & . & 0 \\ 0 & 0 & 0 & 0 & a_n \end{bmatrix}$ where $a_i > 0, \forall i = 1, ..., n$ then the basis vectors $\{e_1, e_2, ..., e_n\}$ is A-conjugate. If $A = VDV^T$, where V is orthogonal (i.e. $V^TV = I$) and D is positive diagonal, then $\{v_1, v_2, ..., v_m\}$ is A-conjugate.

$$< v_i, Av_j >=< v_i, d_jv_j >= 0, \forall i \neq j$$

Lemma 4.2. If $\{p^1, p^2, ..., p^m\}$ is conjugate with respect to a symmetric position definite matrix A, then $p^1, p^2, ..., p^m$ are linear independent.

Proof
Assume $\sum_{i=1}^m \alpha_i p^i = 0$, for any $j \in \{1, 2, \cdots m\}$,

$$\begin{aligned}
0 &= \langle p^j, A\sum_{i=1}^m a_ip^i\rangle \\
&= \langle p^j, A(a_jp^j + \sum_{i\neq j} a_ip^i)\rangle \\
&= \alpha_j\langle p_j, Ap_j\rangle + \sum_{i\neq j}\alpha_i\langle p^i, Ap^i\rangle \\
&= \alpha_j\langle p_j, Ap_j\rangle
\end{aligned}$$

Since $\langle p^j, Ap^j\rangle > 0$, we have $\alpha_j = 0$.

Conjugate direction method

Let A be symmetric positive definite, given a set of A-conjugate vectors $\{p^1, p^2, \cdots p^n\}$

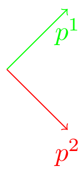
Initialization: x^1 be any vector in $\mathbb{R}^n$
For $k = 1, 2\cdots n$

$$x^{k+1} = x^k + \alpha_kp^k$$
$$\text{where } \alpha^k = \arg\min_\alpha f(x^k + \alpha p^k) = -\frac{\langle p^k, \nabla f(x^k)\rangle}{\langle p^k, Ap^k\rangle}$$

Theorem 4.3. Suppose $\{x^k\}$ is given by the conjugate direction method. Then:

$$\langle \nabla f(x^k), p^i \rangle = 0 \qquad \forall i = 1, 2, \cdots k-1$$
$$\text{and } x^{k+1} = \arg \min_{x \in x^k} f(x)$$
$$\text{where } x^k = x^1 + span\{p^1 \cdots p^k\}$$

Corollary 4.4. In at most n iteration, the conjugate direction will give the exact minimizer.



Define an inner product $\langle \cdot, \cdot \rangle$ as $\langle x, y \rangle = x^T A y$

Example:

$$A = \begin{bmatrix} 3 & 1 \\ 1 & 3 \end{bmatrix}, \qquad b = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$
$$\min_{x \in \mathbb{R}^2} f(x) = \frac{1}{2} x^T A x - b^T x$$
$$\text{Let } x^1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \qquad p^1 - \begin{bmatrix} 1 \\ 1 \end{bmatrix} \qquad p^2 = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$$

Check A-conjugacy:

$$\begin{aligned}
\langle p^1, Ap^2 \rangle &= \langle \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \begin{bmatrix} -2 \\ 2 \end{bmatrix} \rangle \\
&= 0 \\
\nabla f(x^1) &= Ax^1 - b \\
&= \begin{bmatrix} -1 \\ 0 \end{bmatrix} \\
\alpha_1 &= \frac{-\langle p^1, \nabla f(x^1) \rangle}{\langle p^1, Ap^1 \rangle} \\
&= -\frac{\langle \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \begin{bmatrix} -1 \\ 0 \end{bmatrix} \rangle}{\langle \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 4 \\ 4 \end{bmatrix} \rangle} \\
&= \frac{1}{8} \\
x^2 &= x^1 + \alpha_1 p^1 \\
&= \begin{bmatrix} 1/8 \\ 1/8 \end{bmatrix} \\
\nabla f(x^2) &= Ax^2 - b \\
&= \begin{bmatrix} 1/2 \\ 1/2 \end{bmatrix} - \begin{bmatrix} 1 \\ 0 \end{bmatrix} \\
&= \begin{bmatrix} -1/2 \\ 1/2 \end{bmatrix}
\end{aligned}$$

$$
\begin{aligned}
\alpha_2 &= -\frac{\langle p^2, \nabla f(x^2)\rangle}{\langle p^2, p^2\rangle} \\
&= \frac{-\langle \begin{bmatrix}-1\\1\end{bmatrix}, \begin{bmatrix}-1/2\\1/2\end{bmatrix}\rangle}{\langle \begin{bmatrix}-1\\1\end{bmatrix}, \begin{bmatrix}-2\\2\end{bmatrix}\rangle} \\
&= -\frac{1}{4} \\
x^3 &= x^2 + \alpha_2 p^2 \\
&= \begin{bmatrix}1/8, 1/8\end{bmatrix} - \frac{1}{4}\begin{bmatrix}-1\\1\end{bmatrix} \\
&= \begin{bmatrix}3/8\\-1/8\end{bmatrix} \\
\nabla f(x^3) &= Ax^3 - b \\
&= \begin{bmatrix}1\\0\end{bmatrix} - \begin{bmatrix}1\\0\end{bmatrix} \\
&= \begin{bmatrix}0\\0\end{bmatrix}
\end{aligned}
$$

Conjugate Gradient Method

$$\min_x f(x) = \frac{1}{2}x^T Ax - b^T x$$

where $A \in \mathbb{R}^{n\times n}$ is a symmetric positive definite matrix.
Denote $r = Ax - b = \nabla f(x)$
Question: If $\{p^1, p^2, \cdots p^{k-1}\}$ is A conjugate, and $\langle r^j, p^i\rangle = 0$, $\forall i < j \le k$, where $r^i = \nabla f(x^i)$, how to get p^k such that $\{p^1, p^2, \cdots p^k\}$ is A-conjugate?
$p^k = -r^k + \beta_k p^{k-1}$ and set β_k
Since we require $\{p^1, p^2, \cdots p^k\}$ to be A - conjugate, then $\langle p^k, Ap^{k-1}\rangle = 0$,
so $\langle -r^k + \beta_k p^{k-1}, Ap^{k-1}\rangle = 0 \implies \beta_k = \dfrac{\langle r^k, Ap^{k-1}\rangle}{\langle p^{k-1}, Ap^{k-1}\rangle}$
Claim: If $p^k = -r^k + \beta_k p^{k-1}$, then $\langle p^k, Ap^i\rangle = 0, \forall i = 1, 2, \cdots k-1$
Sketch the proof: for $i = 1, 2, \cdots k-2$,

$$
\begin{aligned}
\langle p^k, Ap^i\rangle &= \langle -r^k + \beta_k p^{k-1}, Ap^i\rangle \\
&= \langle -r^k, Ap^i\rangle + \beta_k\langle p^{k-1}, p^i\rangle \\
&= 0 + 0 \\
&= 0 && \text{Since } Ap^y \in span\{r^1, r^2, \cdots r^{i+1}\}
\end{aligned}
$$

Algorithm:

Input;
$A \in \mathbb{R}^{n\times n}$ (sym. pd.) and $b \in \mathbb{R}^n$, $tol > 0$;
Initialization;
Choose x^1, let $r^1 = Ax^1 - b$, set $p^1 = -r^1$;
for $k = 1, 2, \cdots$ **do**
$\alpha_k = -\dfrac{\langle r^k, p^k\rangle}{\langle p^k, Ap^k\rangle}$
$x^{k+1} = x^k + \alpha_k p^k$
$r^{k+1} = Ax^{k+1} - b$
if $||r^{k+1}|| < tol$ **then**
stop
else
$\beta^{k+1} = \dfrac{\langle r^{k+1}, Ap^k\rangle}{\langle p^k, Ap^k\rangle}$
$p^{k+1} = -r^{k+1} + \beta_{k+1}p^k$
end
end

Theorem 4.5. For the conjugate gradient method, it holds: $\langle r^k, r^i\rangle = 0$, $\forall i = 1, 2, \cdots k-1$

$$\begin{aligned} span\{r^1, \cdots r^k\} &= span\{r^1, Ar^1 \cdots A^{k-1}r^1\} \\ &= span\{p^1, p^2 \cdots p^k\} \\ \{p^1, p^2 \cdots p^k\} &\text{ is A conjugate} \end{aligned}$$

Check: $\langle r^2, r^1\rangle = 0$?
Note

$$\begin{aligned} r^2 &= A(x^1 + \alpha_1 p^1) - b \\ &= r^1 + \alpha A p^1 \\ \langle r^2, r^1\rangle &= \langle r^1 + \alpha_1 Ap^1, r^1\rangle \\ &= \langle r^1 - \alpha_1 Ar^1, r^1\rangle \\ \alpha_1 &= \frac{\langle r^1, p^1\rangle}{\langle p^1, Ap^1\rangle} \\ &= \frac{\langle r^1, r^1\rangle}{\langle r^1, Ar^1\rangle} \end{aligned}$$

One more efficient form of CG:
Claim1:

$$\begin{aligned} \alpha_k &= \frac{||r^k||^2}{\langle p^k, Ap^k\rangle} \\ \alpha_k &= -\frac{\langle r^k, p^k\rangle}{\langle p^k, Ap^k\rangle} \\ \text{Note } p^k &= -r^k + \beta_k p^{k-1} \\ \text{So } \langle r^k, p^k\rangle &= \langle r^k, -r^k + \beta_k p^{k-1}\rangle \\ &= -||r^k||^2 + \beta^k\langle r^k, p^{k-1}\rangle \end{aligned}$$

$$\begin{aligned}\therefore \alpha_k &= -\frac{\langle r^k, p^k\rangle}{\langle p^k, Ap^k\rangle}\\ &= \frac{||r^k||^2}{\langle p^k, Ap^k\rangle}\end{aligned}$$

Claim2:

$$\begin{aligned}
\beta_{k+1} &= \frac{||r^{k+1}||^2}{||r^k||^2}\\
\beta_{k+1} &= \frac{\langle p^{k+1}, Ap^k\rangle}{\langle p^k, Ap^k\rangle}\\
&= \frac{\langle r^{k+1}, \alpha_k Ap^k\rangle}{\langle p^k, \alpha_k Ap^k\rangle}\\
\textbf{Note } x^{k+1} &= x^k + \alpha_k p^k\\
\therefore Ax^{k+1} - b &= r^{k+1}\\
&= Ax^k + \alpha_k p^k - b\\
\therefore \alpha_k Ap^k &= r^{k+1} - r^k\\
\text{Then}\langle r^{k+1}, \alpha_k Ap^k\rangle &= \langle r^{k+1}, r^{k+1} - r^k\rangle\\
&= ||r^{k+1}||^2\\
\langle p^k, \alpha_k Ap^k\rangle &= \langle p^k, r^{k+1} - r^k\rangle\\
&= \langle p^k, -r^k\rangle\\
&= ||r^k||^2\\
\therefore \beta_{k+1} &= \frac{\langle p^{k+1}, Ap^k\rangle}{\langle p^k, Ap^k\rangle}\\
&= \frac{||r^{k+1}||^2}{||r^k||^2}
\end{aligned}$$

Algorithm:

Input;
$A \in \mathbb{R}^{n\times n}$ (sym. pd.) and $b \in \mathbb{R}^n$, $tol > 0$;
Initialization;
Choose x^1, let $r^1 = Ax^1 - b$, set $p^1 = -r^1$;
for $k = 1, 2, \cdots$ **do**
  $\alpha_k = \dfrac{||r^k||^2}{\langle p^k, Ap^k\rangle}$
  $x^{k+1} = x^k + \alpha_k p^k$
  $r^{k+1} = r^k + \alpha_k Ap^k$
  if $||r^{k+1}|| < tol$ **then**
    stop
  else
    $\beta^{k+1} = \dfrac{||r^{k+1}||^2}{||r^k||^2}$
    $p^{k+1} = -r^{k+1} + \beta_{k+1}p^k$
  end
end

Example :

$$A = \begin{bmatrix} 3 & 1 \\ 1 & 3 \end{bmatrix} \qquad b = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \qquad x^1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$r^1 = Ax^1 - b = -b = \begin{bmatrix} -1 \\ 0 \end{bmatrix} \qquad p^1 = -r^1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

k=1: $$\begin{aligned}
\alpha_1 &= \frac{||r^1||^2}{\langle p^1, Ap^1 \rangle} \\
&= \frac{1}{\langle \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 3 \\ 1 \end{bmatrix} \rangle} \\
&= \frac{1}{3} \\
x^2 &= x^2 + \alpha_1 p^2 \\
&= \frac{1}{3} \begin{bmatrix} 1 \\ 0 \end{bmatrix} \\
&= \begin{bmatrix} 1/3 \\ 0 \end{bmatrix} \\
r^2 &= r^1 + \alpha A p^1 \\
&= \begin{bmatrix} -1 \\ 0 \end{bmatrix} + \frac{1}{3} \begin{bmatrix} 3 \\ 1 \end{bmatrix} \\
&= \begin{bmatrix} 0 \\ 1/3 \end{bmatrix} \\
\beta_2 &= \frac{||r_2||^2}{||r_1||^2} \\
&= \frac{1/9}{1} \\
&= \frac{1}{9} \\
p^2 &= -r^2 + \beta_2 p^1 \\
&= -\begin{bmatrix} 0 \\ 1/3 \end{bmatrix} + 1/9 \begin{bmatrix} 1 \\ 0 \end{bmatrix} \\
&= \begin{bmatrix} 1/9 \\ -1/3 \end{bmatrix}
\end{aligned}$$

k=2: $$\begin{aligned}
\alpha_2 &= \frac{||r^2||^2}{\langle p^2, Ap^2 \rangle} \\
&= \frac{1/9}{\langle \begin{bmatrix} 1/9 \\ -1/3 \end{bmatrix}, \begin{bmatrix} 0 \\ -8/9 \end{bmatrix} \rangle} \\
x^3 &= x^2 + \alpha_2 p^2 \\
&= \begin{bmatrix} 1/3 \\ 0 \end{bmatrix} + 3/8 \begin{bmatrix} 1/9 \\ -1/3 \end{bmatrix} \\
&= \begin{bmatrix} 3/8 \\ -1/8 \end{bmatrix} \\
r^3 &= r^2 + \alpha_2 A p^2 \\
&= \begin{bmatrix} 0 \\ 1/3 \end{bmatrix} + 3/8 \begin{bmatrix} 0 \\ -8/9 \end{bmatrix} \\
&= \begin{bmatrix} 0 \\ 0 \end{bmatrix}
\end{aligned}$$

Convergence rate of CG

Let $\kappa(A) = \dfrac{\lambda_{max}(A)}{\lambda_{min}(A)}$ and x^* be the optional solution, then

$$||x^{k+1} - x^*||_A \leq \left(\frac{\sqrt{\kappa(A)} - 1}{\sqrt{\kappa(A)} + 1} \right)^{2k} ||x^1 - x^*||_A$$

$||x^{k+1} - x^*||_A$:R-linear
$||x^1 - x^*||_A$:Q-linear
where $||x||_A = \sqrt{x^T A x}$

Preconditioning:

$$\min_{x \in \mathbb{R}} f(x) = \frac{1}{2} x^T A x - b^T x \ \ (*)$$

where A is symmetric p.d.
(*) is required to solve:

$$\begin{aligned} Ax &= b \\ \iff (C^T)^{-1} Ax &= (C^T)^{-1} b \qquad\qquad C \in \mathbb{R}^{n \times n} \text{is nonsingular.} \\ \iff (C^T)^{-1} A C^{-1} C x &= (C^T)^{-1} b \\ \iff \hat{A}\hat{x} &= \hat{b} \\ \text{where } \hat{A} &= C^{-T} A C^{-1} \\ \hat{x} &= cx \\ \hat{b} &= C^{-T} b \\ \iff \min_{\hat{x}} \hat{f}(\hat{x}) &= \frac{1}{2} \hat{x}^T A \hat{x} - \hat{b}^T \hat{x} \end{aligned}$$

Algorithm:

Input;
$\hat{A} \in \mathbb{R}^{n \times n}$ (sym. pd.) and $\hat{b} \in \mathbb{R}^n$, $tol > 0$;
Initialization;
Choose $\hat{x}^1$, let $\hat{r}^1 = \hat{A}\hat{x}^1 - \hat{b}$, set $\hat{p}^1 = -\hat{r}^1$;
for $k = 1, 2, \cdots n$ **do**
  $\alpha_k = \dfrac{||\hat{r}^k||^2}{\langle \hat{p}^k, \hat{A}\hat{p}^k \rangle}$
  $\hat{x}^{k+1} = \hat{x}^k + \alpha_k \hat{p}^k$
  $\hat{r}^{k+1} = \hat{r}^k + \alpha_k \hat{A}\hat{p}^k (\nabla f(k+1))$
  if $||\hat{r}^{k+1}|| < tol$ **then**
    stop
  else
    $\beta^{k+1} = \dfrac{||\hat{r}^{k+1}||^2}{||\hat{r}^k||^2}$
    $\hat{p}^{k+1} = -\hat{r}^{k+1} + \beta_{k+1} \hat{p}^k$
  end
end

Suppose $\hat{x}^k$ is the k^{th} iteration of CG to preconditional problem, let

$$\begin{aligned}
x^k &= C^{-1}\hat{x}^k \\
\hat{r}^k &= \hat{A}\hat{x}^k - \hat{b} \\
\textbf{Also let } p^k &= C^{-1}\hat{p}^k, \qquad M = C^TC \\
||\hat{r}^k||^2 &= ||\hat{A}\hat{x}^k - \hat{b}||^2 \\
&= ||C^{-T}AC^{-1}\hat{x}^k - C^{-T}b||^2 \\
&= ||C^{-T}r^k||^2 \\
&= \langle C^{-T}r^k, C^{-T}r^k\rangle \\
&= (r^T)C^{-1}C^{-T}r^k \\
&= \langle r^k, C^{-1}C^{-T}r^T\rangle \\
&= \langle r^k, M^{-1}r^T\rangle \\
\langle \hat{p}^k, \hat{A}\hat{p}^k\rangle &= \langle \hat{p}^k, C^{-T}AC^{-1}\hat{p}^k\rangle \\
&= (\hat{p}^K)^TC^{-T}AC^{-1}\hat{p}^k \\
&= \langle C^{-1}\hat{p}^k, AC^{-1}\hat{p}^k\rangle \\
&= \langle p^k, Ap^k\rangle \\
\alpha_k &= \frac{\langle r^k, M^{-1}r^k\rangle}{\langle p^k, Ap^k\rangle} \\
\hat{r}^{k+1} &= \hat{r}^k + \alpha_k\hat{A}\hat{p}^k \\
C^{-T}AC^{-1}\hat{x}^{k+1} - \hat{b} &= C^{-T}AC^{-1}\hat{x}^k - \hat{b} + \alpha_kC^{-T}AC^{-1}\hat{p}^k \\
C^{-T}AC^{-1}Cx^{k+1} - C^{-T}b &= C^{-T}AC^{-1}Cx^k - C^{-T}b + \alpha_kC^{-T}AC^{-1}Cp^k \\
C^{-T}Ax^{k+1} - C^{-T}b &= C^{-T}Ax^k - C^{-T}b + \alpha_kC^{-T}Ap^k \\
Ax^{k+1} - b &= Ax^k - b + \alpha_kAp^k \\
\beta_{k+1} &= \frac{||\hat{r}^{k+1}||^2}{||\hat{r}^k||^2} \\
&= \frac{\langle r^{k+1}, M^{-1}r^{k+1}\rangle}{\langle r^k, M^{-1}r^k\rangle} \\
\hat{p}^{k+1} &= -\hat{r}^{k+1} + \beta_{k+1}\hat{p}^k \\
Cp^{k+1} &= -C^{-T}r^{k+1} + \beta_{k+1}Cp^k \\
p^{k+1} &= -M^{-1}r^{k+1} + \beta_{k+1}p^k
\end{aligned}$$

Algorithm:

Input;
$A \in \mathbb{R}^{n\times n}$ (sym. pd.) and $b \in \mathbb{R}^n$, $tol > 0$;
Initialization;
Choose x^1, let $r^1 = Ax^1 - b$, set $p^1 = -(C^TC)^{-1}r^1$;
for $k = 1, 2, \cdots n$ **do**
  Solve $C^TCy^k = r^k$, $y^k = M^{-1}r^k$
  Set $\alpha_k = \dfrac{\langle r^k, y^k\rangle}{\langle p, Ap^k\rangle}$
  Let $x^{k+1} = x^k + \alpha_k p^k$
  $r^{k+1} = r^k + \alpha_k Ap^k$
  Solve $C^TCy^{k+1} = r^{k+1}$ to have y^{k+1}
  Set $\beta_{k+1} = \dfrac{\langle r^{k+1}, y^{k+1}\rangle}{\langle r^k, y^k\rangle}$
  Let $p^{k+1} = -y^{k+1} + \beta_{k+1}p^k$
end

Remark: C needs ti be chosen such that $C^TCy = r$ is easy to solve.
Example: C is diagonal, triangular or sparse.
Nonlinear Gradient Method

$$\min_{x\in\mathbb{R}^n} f(x)$$

Algorithm: (Fletcher - Reeves Method)

Initialization;
Choose x^0, set $p^0 = -\nabla f(x^0)$;
for $k = 1, 2, \cdots$ **do**
  Choose α_k and set $x^{k+1} = x^k + \alpha_k p^k$
  $\beta_{k+1} = \dfrac{||\nabla f(x^{k+1})||^2}{||\nabla f(x^k)||^2}$
  Let $p^{k+1} = -\nabla f(x^{k+1}) + \beta_{k+1}p^k$
end

If α_k is by exact line search, i.e. $\alpha_k = \arg\min_\alpha f(x^k + \alpha p^k) := \Phi(\alpha)$ $\Phi'(\alpha) = 0$
Then $\langle \nabla f(x^k + \alpha_k p^k), p^k\rangle = 0$
So

$$\begin{aligned}
\langle p^{k+1}, \nabla f(x^{k+1})\rangle &= \langle -\nabla f(x^{k+1}) + \beta_{k+1}p^k, \nabla f(x^{k+1})\rangle \\
&= -||\nabla f(x^{k+1})||^2 + \beta_{k+1}\langle \nabla f(x^{k+1}), p^k\rangle \\
&= -||\nabla f(x^{k+1})||^2
\end{aligned}$$

Practically, α_k by inexact line search, such that the strong Wolfe Condition hold:

$$f(x^k + \alpha_k p^k) \le f(x^k) + c_1\alpha_k \nabla f(x^k)^T p^k$$
$$|\nabla f(x^k + \alpha_k p^k)^T p^k| \le c_2 |\nabla f(x^k)^T p^k|$$
$$\text{where } 0 < c_1 < c_2 < \frac{1}{2}$$

5 L5 Proximal Gradient Method

Proximal Gradient Method

$$\min_{x\in\mathbb{R}^n} f(x) + g(x)$$

where f is lip-differenciable, and g is "simply" differentiable function.
Example 1: Lasso Problem

$$\min_{x\in\mathbb{R}^n} \frac{1}{2}||Ax-b||^2 + \lambda||x||_1$$

Example 2:

$$\min_{x\in\mathbb{R}^n} \frac{1}{2}x^TAx - b^Tx + \sum_{i=1}^{n}\sigma_+(x_i)$$

$$\text{where } \sigma_+(x_i) = \begin{cases} 0 & \text{if } x_i \leq 0 \\ \infty & \text{if } x_i < 0 \end{cases}$$

Not Differentiable on boundary.

Definition 5.1 (Proximal Mapping)**.** Given a function g on $\mathbb{R}^n$, its proximal mapping $prox_g(\cdot) : \mathbb{R}^n \to \mathbb{R}^n$ is defined as

$$prox_g(y) = \arg \min_{x\in\mathbb{R}^n \frac{1}{2}||x-y||^2+g(x)} \quad \forall y \in \mathbb{R}^n$$

Example 1: Let $g(x) = |x|$, for $x \in \mathbb{R}$,

$$prox_g(y) = \arg\min_{x\in\mathbb{R}} \frac{1}{2}(x-y)^2 + |x| = F(x)$$

Note:

$$\frac{1}{2}(x-y)^2 + |x| = \begin{cases} \frac{1}{2}(x-y)^2 + x & \text{if } x > 0 \\ \frac{1}{2}(x-y)^2 - x & \text{if } x < 0 \end{cases}$$

For $x \geq 0$, the minimizer of $\frac{1}{2}(x-y)^2 + x$ is

$$\begin{aligned} \hat{x_1} &= \arg\min_{x\in\mathbb{R}} \frac{1}{2}(x-y)^2 + x \\ \implies \hat{x_+} - y + 1 &= 0 \\ \implies \hat{x_+} &= y - 1 \end{aligned}$$

Hence,

$$\begin{aligned} \hat{x_+^*} &= \arg\min_{x\geq 0} \frac{1}{2}(x-y)^2 + x \\ &= \max(0, y-1) \end{aligned}$$

Similarly,

$$\begin{aligned} \hat{x_-^*} &= \arg\min_{x\leq 0} \frac{1}{2}(x-y)^2 - x \\ &= \min(0, y+1) \end{aligned}$$

Therefore, when $y - 1 \geq 0$, $prox_g(y) = y - 1$
when $y + 1 \leq 0$, $prox_g(y) = y + 1$

when $y-1 \le 0$ and $y+1 \ge 0$, $prox_g(y) = 0$
$\therefore prox_g(y) = sign(y) \cdot \max(|y|-1, 0)$
where $sign(y) = \begin{cases} +1 & y > 0 \\ -1 & y < 0 \\ 0 & y = 0 \end{cases}$
Exercise: If $g(x) = \lambda|x|$ for $x \in \mathbb{R}$, where $\lambda > 0$, then $prox_g(y) = sign(y) \cdot \max(|y|-\lambda, 0)$
Example 2: Given $a \le b$, let $g(x) = \begin{cases} 0 & \text{if } x \in [a,b] \\ \infty & \text{otherwise} \end{cases}$

$$\begin{aligned} prox_g(y) &= \arg\min_{x\in\mathbb{R}} g(x) + \frac{1}{2}(x-y)^2 \\ &= \arg\min_{x\in[a,b]} \frac{1}{2}(x-y)^2 \\ &= \min(\max(a,y), b) \end{aligned}$$

Remark: Given $x \subseteq \mathbb{R}^n$, if $g(x) = \begin{cases} 0 & \text{if } x \in [a,b] \\ \infty & \text{otherwise} \end{cases}$, the $prox_g$ is also called projection onto x, denoted as $Proj_x$
Proximal gradient method for $\min_{x\in\mathbb{R}} f(x) + g(x)$
Algorithm:

Initialization;
x^0;
for $k = 1, 2, \cdots n$ **do**
$x^{k+1} = \arg\min_{x\in\mathbb{R}^n} f(x^k) + \langle \nabla f(x^k), x - x^k \rangle + \dfrac{1}{2\alpha_k}||x - x^k||^2 + g(x)$
$= prox_{\alpha_k g}(x^k - \alpha_k f(x^k))$
end

By the definition of x^{k+1},

$$\begin{aligned} x^{k+1} &= \arg\min_{x\in\mathbb{R}^n} f(x^k) + \langle \nabla f(x^k), x - x^k \rangle + \frac{1}{2\alpha_k}||x - x^k||^2 + g(x) \\ &= \arg\min_{x\in\mathbb{R}^n} \alpha_k \langle \nabla f(x^k), x - x^k \rangle + \frac{1}{2}||x - x^k||^2 + \alpha_k g(x) \\ &= \arg\min_{x\in\mathbb{R}^n} \frac{1}{2}||x - (x^k - \alpha_k \nabla f(x^k))||^2 - \frac{1}{2}||\alpha_k \nabla f(x^k)||^2 + \alpha_k g(x) \\ &= \arg\min_{x\in\mathbb{R}^n} \frac{1}{2}||x - (x^k - \alpha_k \nabla f(x^k))||^2 + \alpha_k g(x) \end{aligned}$$

How to choose α_k?
Choose α_k such that:

$$f(x^{k+1} \le f(x^k) + \langle \nabla f(x^k), x^{k+1-x^k} \rangle + \frac{1}{2\alpha_k}||x^{k+1} - x^k||^2 \quad (*)$$

Lemma 5.1. Assume ∇f is lipschitz continuous with constant L, i.e. $||\nabla f(x) - \nabla f(y)|| \le L||x-y||$, $\forall x, y \in \mathbb{R}$, thern we coan always choose $\alpha_k = \frac{1}{L}$ and (*) will hold.

Example:

$$\min_{x\in\mathbb{R}} \frac{1}{2}||Ax - b||^2 + \lambda||x||$$

Notice: $\nabla f(x) = A^T(Ax - b)$ is lip continuous with constant $L = ||A^T A|| = \lambda_{max}(A^T A)$
Update: $x^{k+1} = \arg\min_{x\in\mathbb{R}^n} \frac{1}{2}||x - (x^k - \alpha_k \nabla f(x^k))||^2 + \alpha_k \lambda ||x||_1$
where $\alpha_k = 1/L$
Let $y = x^k - \alpha_k \nabla f(x^k)$ and $\bar{\lambda} = \alpha_k \lambda = \lambda / L$
Then

$$\begin{aligned}
& x^{k+1} = \arg\min_{x\in\mathbb{R}^n} \frac{1}{2}||x - y||^2 + \bar{\lambda}||x||_1 \\
\iff\ & x_i^{k+1} = \arg\min_{x_i\in\mathbb{R}^n} \frac{1}{2}||x_i - y_i||^2 + \bar{\lambda}||x_i||_1 & \forall i = 1, 2, \cdots n \\
& = sign(y_i) \cdot \max(|y_i| - \bar{\lambda}, 0) & \forall i = 1, 2, \cdots n
\end{aligned}$$

Convergence result
Assume ∇f is lip continuous with constant $L > 0$, and g is convex, set $\alpha_k = \frac{1}{2}, \forall k$, then any cluster point of $\{x_k\}$ is a stationary point. If futhermore f is convex, then

$$F(x^k) - F(x^*) \leq \frac{L||x^0 - x^*||^2}{2k} \quad \forall k \geq 1$$

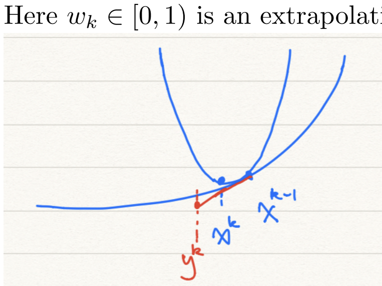
where F=f+g and x^* is one minimizer of F.
Accelerated proximal gradient

$$\min_{x\in\mathbb{R}^n} f(x) + g(x)$$

where f is lip-differentiable (i.e. ∇f is lip.cont.) and g is a "simple" function.
Algorithm:

Initialization;
x^0, set $y^1 = x^0$;
for $k = 1, 2, \cdots$ **do**
  $x^k = prox_{\alpha_k g}(y^k - \alpha_k \nabla(y^k))$
  $y^{k+1} = x^k + w_k(x^k - x^{k-1})$
end

Here $w_k \in [0, 1)$ is an extrapolation weight.



Theorem 5.2. Choose w_k in the following way:
$t_1 = 1$, $t_{k+1} = \dfrac{1 + \sqrt{1 + 4t_k^2}}{2}$, $w_k = \dfrac{t_k - 1}{t_{k+1}}$
Then if f and g are convex, we have $F(x^k) - F(x^*) \leq \dfrac{2L||x^0 - x^*||^2}{k^2}, \forall k \geq 1$
where L is Lip-constant of ∇f

Algorithm:(FISTA)

Initialization;

x^0, set $y^1 = x^0$,$t_1 = 1$;

for $k = 1, 2, \cdots$ **do**

$x^k = \arg\min_x \langle \nabla f(y^k), x - y^k \rangle + \dfrac{1}{2\alpha_k}||x - y^k||^2 + g(x)$

$prox_{\alpha_k g}(y^k - \alpha_k \nabla(y^k))$

$t_{k+1} = \dfrac{1 + \sqrt{1 + 4t_k^2}}{2}$

$w_k = \dfrac{t_k - 1}{t_{k+1}}$

$y^{k+1} = x^k + w_k(x^k - x^{k-1})$

end

Let

$$\begin{aligned}
f(x) &= \frac{1}{2}||Ax - b||^2 \\
&= \frac{1}{2}\langle Ax - b, Ax - b \rangle \\
&= \frac{1}{2}x^T A^T Ax - \frac{1}{2}b^T Ax - \frac{1}{2}A^T b)^T x + \frac{1}{2}||b||^2 \\
&= \frac{1}{2}x^T A^T Ax - (A^T b)^T x + \frac{1}{2}||b||^2 \\
\therefore \nabla f(x) &= A^T Ax - A^T b \\
&= A^T (Ax - b)
\end{aligned}$$

(Fact : If $h(x) = \dfrac{1}{2}x^T Qx - c^T x$, then $\nabla h(x) = Qx - c$)

6 L6 BB Method

Barzaili - Borwein Method

$$\min_{x\in\mathbb{R}^n} f(x)$$

where f is a differentiable function.
Recall: Steepest gradient descent

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

where α_k id obtained by line search

- Pros:

  Per-Update is "Cheap", global Convergence.

- Cons:

  Convergence is slow

Newton's Method

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

- Pros:

  Fast Convergence near solution

- Cons:

  Per-update is expensive, no global convergence

Idea of BB method: Choose α_k such that $\alpha_k\nabla f(x^k)$ "approximate" $\nabla^2 f(x^k)^{-1}\nabla f(x^k)$
Direction of BB Method: Consider

$$\min_{x\in\mathbb{R}^n} f(x) = \frac{1}{2}x^T Ax - b^T x$$

where A is symmetric and positive definite.
Let $s^{k-1} = x^k - x^{k-1}$ and $y^{k-1} = \nabla f(x^k) - \nabla f(x^{k-1})$
Then $As^{k-1} = y^{k-1}$ or $s^{k-1} = A^{-1}y^{k-1}$
i.e.$\nabla^2 f(x^k)s^{k-1} = y^{k-1}$ or $s^{k-1} = (\nabla^2 f(x^k))^{-1}y^{k-1}$
Option 1 to choose α_k
$\alpha_k^{-1}s^{k-1} \approx y^{k-1}$
by letting

$$\begin{aligned}
\alpha_k^{-1} &= \arg\min_{\beta} ||\beta s^{k-1} - y^{k-1}||^2 \\
&= \frac{\langle s^{k-1}, y^{k-1}\rangle}{||s^{k-1}||^2} \\
\frac{d}{d\beta}||\beta s^{k-1} - y^{k-1}||^2 &= \frac{d}{d\beta}||s^{k-1}\beta - y^{k-1}||^2 \\
&= 2(s^{k-1})^T(s^{k-1}\beta - y^{k-1}) \\
&= 0 \\
\implies \alpha_k &= \frac{||s^{k-1}||^2}{\langle s^{k-1}, y^{k-1}\rangle}
\end{aligned}$$

Option 2 to choose α_k
$s^{k-1} \approx \alpha_k y^{k-1}$

by letting

$$\begin{aligned}\alpha_k^{-1} &= \arg\min_{\beta} ||s^{k-1} - \beta y^{k-1}||^2 \\ &= \frac{\langle s^{k-1}, y^{k-1}\rangle}{||y^{k-1}||^2}\end{aligned}$$

Algorithm(BB method for $\min_{x\in\mathbb{R}^n} f(x)$):

Initialization;
x^0, $tol > 0$, set $x^1 = x^0 - \alpha_0 \nabla f(x^0)$;
where α_0 is obtained by line search;
for $k = 1, 2, \cdots$ **do**
  Let $s^{k-1} = x^k - x^{k-1}, y^{k-1} = \nabla f(x^k) - \nabla f(x^{k-1})$
  if $||\nabla f(x^k)|| < tol$ **then**
    Stop
  else
    Update: $x^{k+1} = x^k - \alpha_k \nabla f(x^k)$
    where $\alpha_k = \dfrac{||s^{k-1}||^2}{\langle s^{k-1}, y^{k-1}\rangle}$ or $\dfrac{\langle s^{k-1}, y^{k-1}\rangle}{||y^{k-1}||^2}$
  end
end

Remark: for the quadratic minimizer:

$$\min_{x\in\mathbb{R}^n} f(x) = \frac{1}{2}x^T Ax - b^T x$$

where $A > 0$

$$\begin{aligned}\langle s^{k-1}, y^{k-1}\rangle &= \langle x^k - x^{k-1}, A(x^k - x^{k-1})\rangle \\ &\geq \lambda_{min}(A)||x^k - x^{k-1}||^2\end{aligned}$$

In addition, $||\nabla f(x^k)||$ R-linearly converges to 0. (only for quadratic)
Remark : BB method may not converge even for smooth convex problem.
BB method with nonmonotone line search $\min_{x\in\mathbb{R}^n} f(x)$ (For convex, converge)

Algorithm:

Initialization;
$C_1 \in (0,1), \rho > 1, x^0, c_0 = f(x_0), Q_0 = 1, \eta \in [0,1]$
set $x^1 = x^0 - \alpha_0 \nabla f(x^0)$;
where $\alpha_0 > 0$ is obtained by line search;
for $k = 1, 2, \cdots$ **do**
    Let $s^{k-1} = x^k - x^{k-1}, y^{k-1} = \nabla f(x^k) - \nabla f(x^{k-1})$
    if $||\nabla f(x^k)|| < tol$ **then**
        Stop
    else
        Let $\alpha_k = \dfrac{||s^{k-1}||^2}{\langle s^{k-1}, y^{k-1}\rangle}$ or $\dfrac{\langle s^{k-1}, y^{k-1}\rangle}{||y^{k-1}||^2}$
        Update: $x^{k+1} = x^k - \alpha_k \nabla f(x^k)$
        while $f(x^{k-1}) > c_k - \alpha_k c_1 ||\nabla f(x^k)||^2$ **do**
            Let $\alpha_k \leftarrow \alpha_k / p$
            $x^{k+1} = x^k - \alpha_k \nabla f(x^k)$
        end
        Set $Q_{k+1} = \eta Q_k + 1$
        $c_{k+1} - \dfrac{\eta Q_k c_k + f(x^{k+1})}{Q_{k+1}}$
    end
end

If $\eta = 0 \rightarrow$ amijo
If $\eta \neq 0 \rightarrow$ modified amijo
Remark: If $\eta = 0$, then $c_k = f(x^k), \forall k = 0, 1, \cdots$ and while loop is doing back tracking by amijo's condition. $\implies$ 在 amijo 条件下充分下降
If $\eta = 1 \implies c_k = \sum_{j=0}^{k} \dfrac{f(x^j)}{k+1}$

7 Quasi-Newton Method

Problem

$$\min_{x\in\mathbb{R}^n} f(x)$$

where f is twice differentiable.
Recall: Newton's Method:

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

(BB method use $\alpha_k I$ to approximate the Hessian matrix)
Memory requirement: $O(n^2)$
Computational cost: $O(n^3)$
Idea of quasi-Newton method:
Find $B_k > 0$ to approximate $\nabla^2 f(x^k)$, or $H^k > 0$ to qpproximate $\nabla^2 f(x^k)^{-1}$
Recall: BB method:
Find α_k such that $\alpha_k I$ approxinate $\nabla^2 f(x^k)$ in the sense $\alpha_k^{-1}s^{k-1} \approx y^{k-1}$ or $s^{k-1} \approx \alpha_k y^{k-1}$
Derivation of Quasi-Newton method:
Suppose $B_k > 0$ qpproximate $\nabla^2 f(x^k)$
Let $p^k = -B_k^{-1}\nabla f(x^k)$
Update $x^{k+1} = x^k + \alpha_k p^k$
where $\alpha_k > 0$ id obtained by line search that the wolfe's condition hold.

$$\begin{aligned} f(x^k + \alpha_k p^k) &\le f(x^k) - c_1\alpha_k \nabla f(x^k)^T p^k \\ \langle p^k, \nabla f(x^k + \alpha_k p^k)\rangle &\ge c_2 \nabla f(x^k)^T p^k \\ \text{with } 0 < c_1 &< c_2 < 1 \end{aligned}$$

Question: How to get B_{k+1}?

7.1 DFP (Davidan - Fletcher -Powell) update for B

Set

$$B_{k+1} = \arg\min_B \{||B - B_k||_w^2, \text{such that } B > 0, Bs^k = y^k\} \quad (*)$$

where $w > 0$ such that $wy^k = s^k$ and $||A||_w = ||w^{1/2}Aw^{1/2}||_F$. ($||A||_F = \sqrt{\sum_i \sum_j (a_{ij}^2)}$)

Lemma 7.1. Suppose B is a solution of (*), then there is a vector $v \in \mathbb{R}^n$ such that

$$\begin{cases} w(B - B_k)w - s^k v^T - v(s^k)^T = 0 \\ Bs^k = y^k B > 0 \end{cases}$$

From

$$w(B - B_k)w - s^k v^T - v(s^k)^T = 0 \quad (1)$$
$$Bs^k = y^k \quad (2)$$

We have

$$\begin{cases} (B - B_k) - w^{-1}s^k v^T w^{-1} - w^{-1}v(s^k)^T w^{-1} = 0 \\ Bs^k = y^k B > 0 \end{cases}$$

Let $u = w^{-1}v$
Note $wy^k = s^k$, so $w^{-1}s^k = y^k$
Hence (1) becomes:

$$B - B_k - y^k u^T - u(y^k)^T = 0$$
$$\text{so } B = B_k + y^k u^T + u(y^k)^T$$

Plug B into (2)

$$B_k s^k + y^k u^T s^k + u(y^k)^T s^k = y^k \tag{3}$$

Let $\gamma^k = \dfrac{1}{(y^K)^T s^k}$
From (3):

$$u = \gamma_k(y^k - B_k s^k - y^k u^T s^k)$$

and

$$\begin{aligned}(s^k)^T B s^k + (s^k)^T y^k u^T s^k + (s^k)^T u (y^k)^T s^k &= (s^k)^T y^k \\ \implies u^T s^k &= \frac{(s^k)^T y^k - (s^k)^T B s^k}{2(s^k)^T y^k} \\ &= \frac{1 - \gamma_k (s^k)^T B s^k}{2}\end{aligned}$$

So

$$u = \gamma_k \left(y_k - B_k s^k - \frac{1 + \gamma_k (s^k)^T B_k s^k}{2} y^k \right)$$

Recall $B = B_k + y_k u^T + u(y_k)^T$
we have

$$\begin{aligned}B_{k+1} &= B_k + \gamma_k y^k (y^k - B_k s^k - \frac{1 - \gamma_k (s^k)^T B_k s^k}{2} y^k)^T + \gamma_k (y^k - B_k s^k - \frac{1 - \gamma_k (s_k)^T B_k s^k}{2} y^k)(y^k)^T \\ &= (I - \gamma_k y^k (s^k)^T) B^k (I - \gamma_k s^k (y^k))^T) + \gamma_k y^k (y^k)^T\end{aligned}$$

Claim1: $\gamma_k > 0$ if Wolfe's condition hold.
Note $\langle p^k, \nabla f(x^k + \alpha_k p^k) \rangle \geq c_2 \nabla f(x^k)^T p^k$
i.e. $\langle \dfrac{x^{k+1} - x^k}{\alpha_k}, \nabla f(x^{k+1}) \rangle \geq c_2 \langle \dfrac{x^{k+1} - x^k}{\alpha_k}, \nabla f(x^k) \rangle$
Since $s^k = x^{k+1} - x^k$, we have

$$\begin{aligned}\langle s^k, \nabla f(x^{k+1}) - \nabla f(x^k) \rangle &\geq (c_2 - 1) \langle s^k, \nabla f(x^k) \rangle \\ \implies \langle s^k, y^k \rangle &\geq (c_2 - 1) \langle s^k, \nabla f(x^k) \rangle \\ &= \alpha_k (c_2 - 1) \langle p^k, \nabla f(x^k) \rangle > 0\end{aligned}$$

Claim 2: $B_{k+1} > 0$
First $B^{k+1} \geq 0$
Scond, suppose there is $Z \in \mathbb{R}^n$ such thst $B_{k+1} z = 0$, then $z^T B_{k+1} z = 0$
$z^T (I - \gamma_k y^k (s^k)^T) B_k (I - \gamma_k s^k (y^k)^T) z + r^k z^T y^k (y^k)^T z = 0$

so $z^T(I-\gamma_k y^k(s^k)^T)B_k(I-\gamma_k s^k(y^k)^T)z = 0$and $r^k z^T y^k (y^k)^T z = 0$, which means $z^T B_k z = 0$ and $z^T y^k = 0$
so$z = 0$.
Implementation of DFP update for $B_{k+1}^{-1}\nabla f(x^{k+1})$
Denote

$$\begin{aligned} H_k &= B_k^{-1} \\ H_{k+1} &= B_{k+1}^{-1} \\ &= [(I-\gamma_k y^k(s^k)^T)B^k(I-\gamma_k s^k(y^k))^T) + \gamma_k y^k(y^k)^T]^{-1} \\ &= H_k - \frac{H_k y^k (y^k)^T H_k}{(y^k)^T H_k y^k} + \frac{s^k(s^k)^T}{(s^k)^T y^k} \end{aligned}$$

(This can directly verify from definition of inverse)
Algorithm (DFP method for $\min_{x\in\mathbb{R}^n} f(x)$):

Initialization: x^0 , $H_0 \succ 0$, $tol > 0$. **for** *k=0,1,2,*$\cdots$ **do**
  if $||\nabla f(x^k)|| \le tol$ **then**
    stop
  else

$$\begin{aligned} p^k &= -H_k \nabla f(x^k) \\ x^{k+1} &= x^k + \alpha_k p^k \end{aligned}$$

where α_k is obtained by line search such that the wolfe's conditions hold.

$$s^k = x^{k+1} - x^k \quad y^k = \nabla f(x^{k+1}) - \nabla f(x^k)$$

$$H_{k+1} = H_k - \frac{H_k y^k (y^k)^T H_k}{(y^k)^T H_k y^k} + \frac{s^k(s^k)^T}{(s^k)^T y^k}$$

  end
end

Remark: in the Wolfe's conditions, $c_1 = 10^{-4}$ and $c_2 = 0.9$ are commonly used. To choose H_0, one can run one steepest gradient descent to have two points $x^0, \hat{x^0}$, where

$$\hat{x^0} = x^0 - \hat{\alpha_0}\nabla f(x^0)$$

Then set $\hat{s^0} = \hat{x^0} - x^0, \hat{y^0} = \nabla f(\hat{x^0}) - \nabla f(x^0)$
and $H_0 = \frac{\langle \hat{s^0}, \hat{y^0}\rangle}{||\hat{y^0}||^2} I$ or $H_0 = \frac{||s^0||^2}{\langle \hat{s^0}, \hat{y^0}\rangle}$

7.2 BFGS(Broyden, Fletcher, Gddfarb, Shano)Method.

Idea: directly work on H matrix
Suppose we have $H_k \succ 0$
Find $H_{k+1} = \arg\min_H ||H - H_k||_W^2$ such that $H \succ 0$, $Hy^k = s^k$
Where $W \succ 0$, and $Ws^k = y^k$.
By Similar arguments as DFP:

$$H_{k+1} = (I - \gamma_k s^k (y^k)^T) H_k (I - \gamma_k y^k (s^k)^T) + \gamma_k s^k (s^k)^T$$

Where $\gamma_k = \frac{1}{(y^k)^T s^k}$

7.3 Convergence result about DFP and BFGS:

Assume f is twice differentiable, and there are $0 < \underline{\lambda} < \bar{\lambda} < +\infty$ such that

$$\underline{\lambda} \leq \lambda_{min}(\nabla^2 f(x)) \leq \lambda_{max}(\nabla^2 f(x)) \leq \bar{\lambda}, \forall x.$$

Let $\{x^k\}$ be generated by DFP or BFGS method. with any initial x^0 and $H_0 \succ 0$. Then $x^k \to x^*$, where x^* is the minimizer.
Furthermore, if $\nabla^2 f(x)$ is Lip-continuous, then BFGS method has R-super-linear convergence.

Observation: DFP and BFGS both perform rank-2 update to H matrix.

$$\begin{aligned} H_{k+1} &= (I - \gamma_k s^k (y^k)^T) H_k (I - \gamma_k y^k (s^k)^T) + \gamma_k s^k (s^k)^T \\ &= H_k - \gamma_k s^k (y^k)^T H_k - \gamma_k H_k y^k (s^k)^T + \gamma_k^2 (y^k)^T H_k y^k s^k (s^k)^T + \gamma_k s^k (s^k)^T \\ &= H_k - \gamma_k s^k [(y^k)^T H_k - \gamma_k (y^k)^T H_k y^k (s^k)^T - (s^k)^T] \end{aligned}$$

7.4 SRI method

$$\min_{x \in \mathbb{R}^n} f(x)$$

Suppose we have H_k that approximates $\nabla^2 f(x^k)^{-1}$
Let $H_{k+1} = H_k + \theta v v^T$ such that $H_{k+1} y^k = s^k$,
Where $\theta = 1$or -1
how to find v and θ
From $H_{k+1} y^k = s^k$, it follows

$$\begin{aligned} H_k y^k + \theta v v^T y^k &= s^k \\ \Leftrightarrow \theta v v^T y^k &= s^k - H_k y^k \\ \Rightarrow \theta (y^k)^T v v^T y^k &= (y^k)^T (s^k - H_k y^k) \end{aligned}$$

7.4.1 Case I:

$$(y^k)^T (s^k - H_k y^k) \neq 0$$

set $\theta = sign((y^k)^T (s^k - H_k y^k))$
Then $(v^T y^k)^2 = |(y^k)^T (s^k - H_k y^k)|$ So $(v^T y^k) = \pm |(y^k)^T (s^k - H_k y^k)|^{\frac{1}{2}}$
and $v = \pm |(y^k)^T (s^k - H_k y^k)|^{-\frac{1}{2}} (s^k - H_k y^k)$
Hence,

$$\begin{aligned} H_{k+1} &= H_k + \theta v v^T \\ &= H_k + \frac{(s^k - H_k y^k)(s^k - H_k y^k)^T}{(y^k)^T (s^k - H_k y^k)} \end{aligned}$$

7.4.2 Case II:

$s^k - H_k y^k = 0$
set $v = 0$
and $H_k k + 1 = H_k$

7.4.3 Case III:

$(y^k)^T(s^k - H_k y^k) = 0$ but $s^k - H_k y^k \neq 0$
no (θ, v) such that $\theta v v^T y^k = s^k - H_k y^k$
we just set $H_{k+1} = H_k$
Heuristic: if $|(y^k)^T(s^k - H_k y^k)| \leq \rho||y^k|| \cdot ||s^k - H_k y^k||$.
Then set $H_{k+1} = H_k$
Otherwise, set

$$H_{k+1} = H_k + \frac{(s^k - H_k y^k)(s^k - H_k y^k)^T}{(y^k)^T(s^k - H_k y^k)}$$

Observation: H_{k+1} is not guaranteed to be p.d. So $p^{k+1} = -H_{k+1}\nabla f(x^{k+1})$ may not be a descent directoin.
Exact line search should be used. i.e.

$$\alpha_{k+1} = \arg\min_{\alpha} f(x^{k+1} + \alpha p^{k+1})$$

or trust-region step:

$$p^{k+1} \leftarrow \arg\min_{p} \langle \nabla f(x^{k+1}), p \rangle + \frac{1}{2} p^T B_{k+1} p$$

$$s.t. ||p|| \leq \beta$$

Where B_{k+1} is the approximation of $\nabla^2 f(x^{k+1})$.
By the same derivation:

$$B_{k+1} = B_k + \frac{(y^k - B_k s^k)(y^k - B_k s^k)^T}{(s^k)^T(y^k - B_k s^k)}$$