1 Exact line search

1.1 What is line search

For most numerical methods, we find a direction that leads us to minimum. However, walking through that direction not necessarily leads to convergence. For example, steepest descent on the function $f(x) = x^2$ at $x = 1$ will yields an repeating numerical sequence \{1, -1, 1, -1, \ldots\}.

The main problem is on walking through that direction too much. Thus, we adopt some method to ensure that this won’t happen.

1.2 The exact line search

One the direction is decided, we find how much we have to walk. Exact line search tells to walks to the minimum through that direction. Let we examine an example.

1.2.1 Steepest descent with exact line search

Consider

$$f(x_1, x_2) = \frac{1}{2} (x_1^2 + \alpha x_2^2)$$

for some $\alpha > 1$. \hspace{1cm} (1)

The gradient is

$$\nabla f(x_1, x_2) = [x_1, \alpha x_2]^T.$$ \hspace{1cm} (2)

Thus, the steepest descent direction for some point $x_n = (x_{n1}, x_{n2})$

$$-\nabla f(x_{n1}, x_{n2}) = -[x_{n1}, \alpha x_{n2}]^T.$$ \hspace{1cm} (3)

The following is the most important part of line search procedure. We set

$$x_{n+1} = x_n - \eta \nabla f(x_n)$$

for some $\eta > 0$. \hspace{1cm} (4)
η is selected such that $f(x_{n+1})$ is the minima along the descent direction. We find this by letting

$$\theta(\eta) = f(x_n - \eta \nabla f(x_n)).$$

Then, solve

$$\theta'(\eta) = 0.$$  

(6)

(5)

to find the minimized $\eta$. After calculating,

$$\eta = \frac{x_{n_1}^2 + \alpha x_{n_2}^2}{x_{n_1}^2 + \alpha^2 x_{n_2}^2}.$$  

(7)

Then,

$$x_{n+1} = \left(\frac{\alpha x_{n_1} x_{n_2}^2(\alpha - 1)}{x_{n_1}^2 + \alpha^2 x_{n_2}^2}, \frac{-x_{n_1} x_{n_2}^2(\alpha - 1)}{x_{n_1}^2 + \alpha^2 x_{n_2}^2}\right).$$

(8)

This gives the updating rule. □

1.2.2 How fast is steepest descent with exact line search

Assume $x_n = (x_{n_1}, x_{n_2})$ and $\frac{x_{n_1}}{x_{n_2}} = 1$. Then, $|\frac{x_{n_1}}{x_{n_2}}| = 1, \alpha, 1, \alpha, \ldots$. This means the points approaches optimal slowly w.r.t $x_2$. (Small displacement on $x_2$ causes large displacement on function value.)

For example, we select $\alpha = 100$, $x_1 = (1, 1)$. Then,

$$x_2 = (0.9900, -0.0001), f(x_2) = 0.49$$

(9)

$$x_3 = (0.0097, 0.0097), f(x_3) = 0.48$$

(10)

For another example, we select $\alpha = 2$, $x_1 = (1, 1)$. Then,

$$x_2 = (0.4444, -0.1111), f(x_2) = 0.1111$$

(11)

$$x_3 = (0.0741, 0.0741), f(x_3) = 0.0082$$

(12)

The parameter $\alpha$ somehow has a lot to do with the convergence.

1.3 Exact line search on steepest descent’s convergence

We try to generalize what we discussed before. For all quadratic functions, we can W.L.O.G define it as

$$f(x) = c^T x + \frac{1}{2} x^T H x,$$

where $H$ is symmetric positive definite.

The preceding example has $H = [1, 0; 0, \alpha]$ and $c = 0$.

1.3.1 The conditional number

Definition 1 Conditional number The conditional number of a matrix $H$, denoted $\text{cond}(H)$, is its largest eigenvalue divided by its smallest eigenvalue.

You may have noticed that in the preceding example, the conditional number of $H$ is $\alpha$. We can guess the convergence rate is related to the conditional number of its Hessian, and the convergence is slow for large conditional numbers.
1.3.2 The Q-norm

To prove the convergence rate, we have to define in what norm we are in. Let me introduce the Q-norm we’ll use later.

**Definition 2 Norm** A norm \( \| \| \) over a vector space \( V \) must satisfy that for all \( u, v \in V \) and scalar \( a \)

\[
\| av \| = |a| \| v \| \\
\| v + u \| \leq \| v \| + \| u \| \\
\| v \| = 0 \text{ iff } v = 0
\]

(14) (15) (16)

**Definition 3 Q-norm** For a symmetric positive definite matrix \( Q \), we define the \( Q \)-norm of a vector \( x \) as

\[
\| x \|_Q = \left( x^T Q x \right)^{1/2}
\]

Theorem \( \infty \) Spectral theorem Every symmetric matrix \( Q \in \mathbb{R}^{n \times n} \) with eigenvalues \( \lambda_1, \lambda_2 \ldots \lambda_n \), can be transformed to a diagonal matrix \( D \) with \( D_{ii} = \lambda_i \) by some linear operator \( L \).


**Theorem 1 Q-norm is a norm**

**Proof** For all \( u, v \in V \) and scalar \( a \). As \( Q \) is positive definite, \( v^T Q v \leq 0 \). Thus, \( (v^T Q v)^{1/2} \) is valid.

\[
\| av \|_Q = (a^2 v^T Q v)^{1/2} = |a| \| v \|_Q \tag{17}
\]

This completes the proof of the first rule.

Apply spectral theorem, transforming \( Q \) to a diagonal matrix \( D \). Thus, in this space,

\[
v^T Q u = \sum_i D_{ii} v_i u_i \leq \left( \sum_i (D_{ii})^{1/2} v_i^2 \right) \left( \sum_i (D_{ii})^{1/2} u_i^2 \right)^{1/2} = \| v \|_Q \| u \|_Q \tag{18}
\]

The \( \leq \) in the above formula is Cauchy’s inequality. Then,

\[
\| v + u \|_Q \leq (\| v \|_Q + \| u \|_Q + (v^T Q v + u^T Q u)^{1/2} \tag{19}
\]

\[
= (\| v \|_Q^{2} + \| u \|_Q^{2} + 2u^T Q v)^{1/2} \tag{20}
\]

\[
\leq (\| v \|_Q^{2} + \| u \|_Q^{2} + 2\| u \|_Q \| v \|_Q^{2})^{1/2} \tag{21}
\]

\[
= \| v \|_Q + \| u \|_Q \tag{22}
\]

The third equality comes from the symmetric of \( Q \). That is, \( v^T Q u = (v^T Q u)^T = u^T Q v \). The \( \leq \) comes from what we just derived.

\[
\| 0 \|_Q = 0 \text{ If } \| v \| = 0, \text{ then } v^T Q v = 0 \text{. By positive-definiteness, } v = 0.
\]

This completes the proof of the third rule.
1.3.3 Exact line search on this problem

Let’s back to the problem.

\[ f(x) = c^T x + \frac{1}{2} x^T H x , \text{ where } H \text{ is symmetric positive definite.} \]  

(25)

For any given \( x_n \), we’ll updated it to \( x_{n+1} \).

\[ \text{let } g_{x_n} = \nabla f(x_n) = c + Hx_n. \]  

(26)

Now, \( -g_{x_n} \) is the steepest descent direction.

\[ \text{let } \eta^*_n = \arg \min_{\eta > 0} \nabla f(x_n - \eta g_{x_n}) = g_{x_n}^T g_{x_n} / g_{x_n}^T H g_{x_n}. \]  

(27)

Then, \( \eta^*_n \) is the minima through the steepest descent direction. Finally, set

\[ x_{n+1} = x_n - \eta^*_n g_{x_n}. \]  

(28)

1.3.4 The convergence rate

Let \( x^* \) be the minimizer of \( f \). Define the error function \( e : \mathbb{R}^n \to \mathbb{R} \) as

\[ e(x) = \|x - x^*\|_H \]  

(29)

For a optimization sequence \( \{x_n\}_{n=1}^{\infty} \) We’ll try to evaluate the rate of decreasing, \( r = \frac{e(x_{n+1})}{e(x_n)} \). If \( r \) is small, then the minimization procedure is fast; if \( r \) is large, the minimization procedure is slow.

After calculation,

\[ e(x_{n+1}) = [1 - \frac{(g_n^T g_n)^2}{(g_n^T H g_n)(g_n^T H^{-1} g_n)}] e(x_n) \]  

(30)

\[ \frac{e(x_{n+1})}{e(x_n)} = [1 - \frac{(g_n^T g_n)^2}{(g_n^T H g_n)(g_n^T H^{-1} g_n)}] \leq (1 - \alpha)^2 (1 + \alpha)^2 \]  

(31)

The last \( \leq \), comes from the Kantorovich inequality, which will be stated below.

**Theorem 2 Kantorovich inequality**  Let \( H \) be an \( n \times n \), symmetric, positive definite matrix with conditional number \( \alpha \). Then, for any \( x \in \mathbb{R}^n \), we have

\[ \frac{(x^T x)^2}{(x^T H x)(x^T H^{-1} x)} \geq \frac{4\alpha}{(1 + \alpha)^2} \]  

(32)

**Proof** To prove this theorem, you need spectral theorem and the convexity. However, the calculation is somehow complex.